THE EVOLUTION OF CHROMATIN FOLDING IN MAMMALS: A ROLE FOR CTCF

PhD Thesis

by
MATTEO VIETRI RUDAN

University College London

2015

DECLARATION

I, Matteo Vietri Rudan, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

To all my teachers

ABSTRACT

The organisation of the DNA inside eukaryotic cell nuclei is not random. Rather than being intermingled with one another, chromosomes are compacted in a hierarchical manner which is conserved throughout eukaryote evolution. Topological domains have recently emerged as key architectural building blocks of chromosomes in complex genomes. This thesis aimed to explore the evolutionary dynamics of chromatin architecture in order to shed light on its functional relevance and on how architectural proteins can shape it. Comparative Hi-C in liver cells from four mammalian species was used to characterise conservation and divergence of chromosomal structure within distantly related genomes. Results show that the modular organisation of chromosomes is robustly conserved in syntenic regions, and that domain structure is maintained during chromosomal rearrangements. This conservation is compatible with the evolution of the binding landscape of the architectural protein CTCF. Specifically, conserved CTCF sites are more often co-localised with cohesin, enriched at strong topological domain borders and bind to DNA motifs that are under strong selection. Interestingly, CTCF binding sites which are divergent between species are strongly correlated with divergence of internal domain structure. This divergence is likely driven by CTCF binding sequence changes, demonstrating how genome evolution can be linked directly with a continuous flux of local chromosome conformation changes. Finally, the architectural activity of CTCF is cohesin-dependent and the manner in which individual CTCF/cohesin sites choose interacting partners is dictated by the orientation of the CTCF binding motif.

TABLE OF CONTENTS

PUBLICATIONS ARISING FROM THE WORK IN THIS THESIS

Vietri Rudan, M., Barrington, C., Henderson, S., Ernst, C., Odom, D. T., Tanay, A., & Hadjur, S. (2015). Comparative Hi-C Reveals that CTCF Underlies Evolution of Chromosomal Domain Architecture. Cell Reports, 10(8), 1297–1309. http://doi.org/10.1016/j.celrep.2015.02.004

LIST OF FIGURES AND TABLES

«A scientist's work is 80% troubleshooting»
– Pablo Rodriguez-Viciana

«Nothing is 100% in biology»
– Suzana Hadjur

CHAPTER ONE:

INTRODUCTION

1.1 BACKGROUND

The folding of DNA in the eukaryotic nucleus.

The architecture of the nucleus and the organisation of chromosomes.

A fundamental problem in cell biology is how the genome can be packaged into a constrained nuclear space and how normal cellular processes such as transcription can be accommodated within this compact environment. The 46 chromosomes harboured in a human somatic cell, for example, sum up to around two metres of DNA, which must fit into a nucleus that can be as small as 6 μm in diameter. The way in which this extensive folding occurs has been intriguing scientists for a very long time. Early observations of nuclear organisation date back to the end of the XIX century when Carl Rabl and Theodor Boveri first introduced the concept of a "chromosome territory" (CT). This is the notion that chromosomes occupy a defined region within nuclear space when decondensed, rather than being randomly intermingled (Rabl, 1885; Boveri, 1909). The existence of CTs was first demonstrated in the early 1980s using laser-UV microirradiation experiments (Cremer et al., 1982) and it is now widely accepted that chromosomes are non-randomly organised within nuclear space (Foster and Bridger, 2005). In addition to occupying a defined territory, chromosomes also have preferred positions within nuclear space. Small, gene-dense, early replicating chromosomes tend to be located in the nuclear interior (Tanabe et al., 2002b), a feature of nuclear organisation conserved during evolution (Tanabe et al., 2002a; Tanabe et al., 2002b; Mayer et al., 2005; Neusser et al., 2007). These observations are consistent with a model whereby the interior of the nucleus is a more "active" environment, and the nuclear periphery is a more "repressive" environment (Lanctôt et al., 2007; Geyer et al., 2011).

The conserved nature of nuclear organisation implies a functional role for chromosomal architecture. Indeed, many studies point to a correlation between nuclear architecture and genome function. Some of the first evidence for this comes from the observation that functional sub-nuclear compartments exist within nuclear space. The most prominent of these, the nucleolus, associates the genomic regions that encode ribosomal RNAs (Wallace and Birnstiel, 1966). Cajal bodies have been implicated in the processing of nuclear RNA (Cajal, 1903; Nizami et al., 2010). Active genes have also been shown to co-localise with

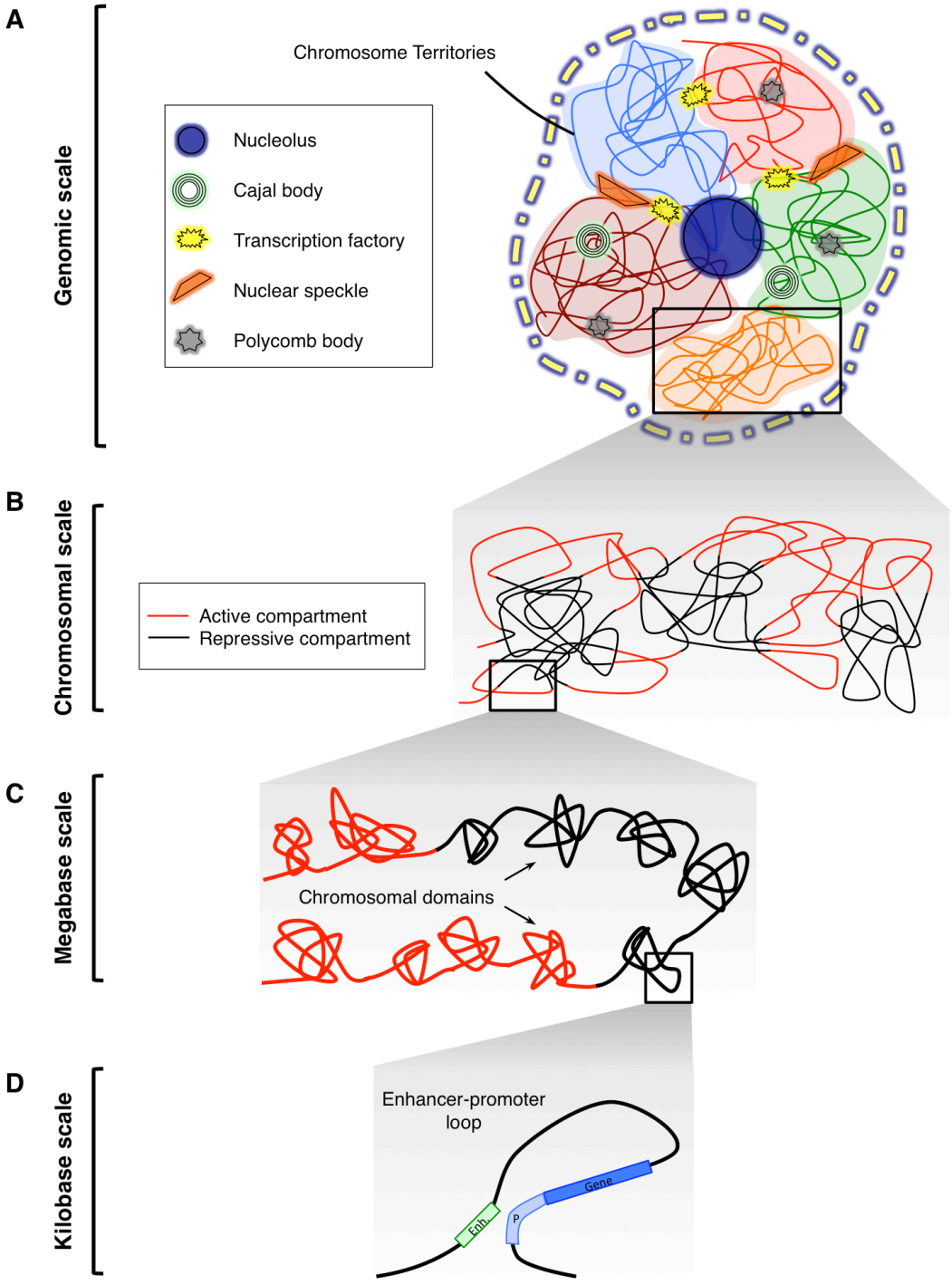


FIGURE 1 – An overview of nuclear architecture and the hierarchical nature of chromatin folding. (**A**) A schematic of the general architecture of the genome within the nucleus. (**B**) At the chromosomal scale, regions of chromatin cluster together to form active and repressive compartments. (**C**) At the megabase scale, the fibre forms a succession of highly self-interacting chromosomal domains. (**D**) Within each domain, individual loops (for example between enhancers and gene promoters) can be established.

nuclear foci containing RNA polymerase II, termed "transcription factories" (Iborra et al., 1996; Osborne et al., 2004), or around larger "speckles" enriched for factors involved in RNA splicing (Shopland et al., 2003; Moen et al., 2004). In *Drosophila* genes that are

regulated by the repressor complex Polycomb can cluster together into so-called "Polycomb bodies" (Messmer et al., 1992; Lanzuolo et al., 2007) that are important for the correct regulation of the *Hox* gene clusters (Bantignies et al, 2010). All these observations highlight how certain genomic activities can occur in specific places within the three-dimensional nuclear space (**FIGURE 1A**).

How the specific chromosomes engage with these functional sub-compartments is an area of intense research. It has been shown that regions that are gene-dense or have a high density of transcribed genes may "loop out" of their chromosome territories and interact with transcription factories present at the interface between chromosome territories in order to facilitate gene expression (Mahy et al., 2002; Fraser and Bickmore, 2007). Consistent with this, individual territories defined by "painting" whole chromosomes with fluorescently labelled probes appear to intermingle with one another (Branco and Pombo 2006).

Taken together, these observations describe a highly organised nucleus and strongly indicate a link between its architecture and the function of the genome. Whether similar organising principles apply also to the folding of chromatin at the chromosomal and sub-chromosomal level has been the object of intense investigation over the past few years.

Molecular approaches to study the 3D organisation of chromosomes in high resolution.

While microscopy-based methods have clearly shown the non-random organisation of chromsomes in nuclear space, the precise nature of chromosome compaction has remained a mystery due to the limited resolution and throughput of these approaches. To overcome these barriers, a number of molecular methods have been developed. These techniques, termed "Chromosome Conformation Capture" (or 3C) allow for the identification of contacts between genomic elements that are physically close together in nuclear space, but far apart on the linear chromosome and rely on the "proximity ligation principle" (Dekker et al., 2002). In brief, cells are first treated with formaldehyde to cross-link protein-protein and protein-DNA interactions, effectively "capturing" chromatin in its native three-dimensional (3D) configuration. DNA is then digested with restriction enzymes within the intact nucleus and subsequently re-ligated (**FIGURE 2A**). Fragments that are not contiguous in the linear sequence, but close in three-dimensional space will be re-ligated together more frequently than fragments that are spatially farther apart. Variations on this basic

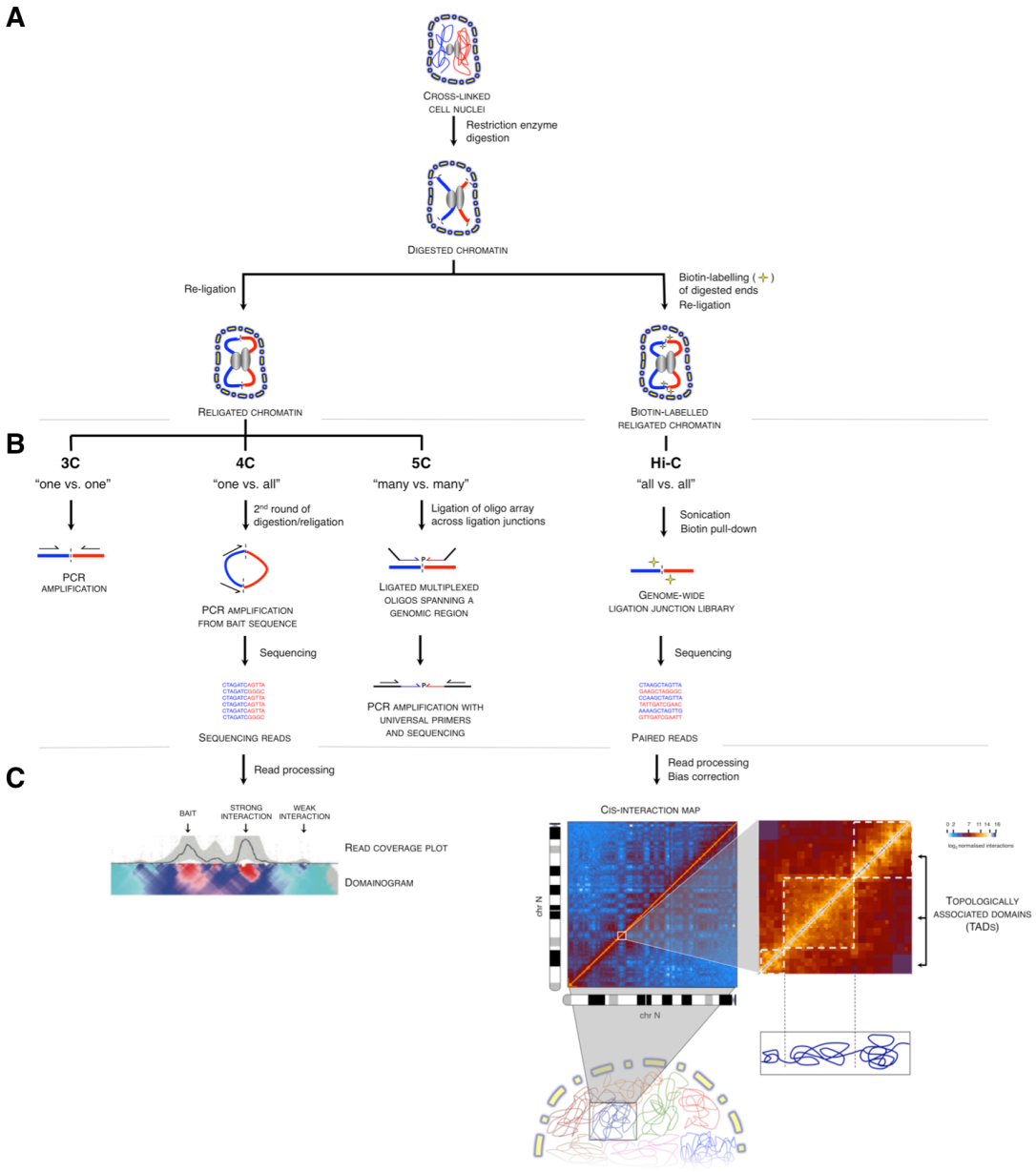


Figure 2 – The proximity ligation principle and the "C"-methodologies. (**A**) According to the proximity-ligation principle, after digestion of cross-linked chromatin, fragments that are close together in three-dimensional space will be re-ligated together more often than fragments that are more spatially separated. (**B**) Variations on the proximity ligation principle allow for the study of interactions between two specific loci (3C); all interactions from a single "viewpoint" (4C, also referred to as 4C-seq); interactions between fragments in a defined genomic region (5C); or unbiased interaction across the entire genome (Hi-C). (**C**) Data visualisation in 4C-seq (left) and Hi-C (right).

principle allow for the characterisation of interactions between two defined loci in the genome (3C) (Dekker et al., 2002), between one locus and the rest of the genome (4C) (Zhao et al., 2006; van de Werken et al., 2012), across a specific genomic region (5C) (Dostie et al., 2006) or across the whole genome in an unbiased fashion (Hi-C) (Lieberman-Aiden et al., 2009) (**Figure 2B**). The use of these methods has greatly

expanded our view of the three-dimensional architecture of chromatin and its relationship with genome function.

Chromosomes are organised into a series of hierarchical chromatin loops.

The high-resolution nature of C-methods has begun to revolutionise our understanding of chromosome structure, revealing that chromsomes are organised as a hierarchy of chromatin loops (**FIGURE 1B-D**).

In mammals, on a chromosomal scale, Hi-C data has revealed that chromatin clusters into two compartments (**FIGURE 1B**). In other words, chromatin belonging to one compartment will tend to interact more frequently with chromatin belonging to the same compartment rather than the other compartment. These two clusters can be functionally classified into either active (Transcription Start Sites (TSS)-rich, Lamin-interacting region-poor) or inactive (TSS-poor, Lamin-interaction-rich) groups (Lieberman-Aiden et al., 2009). The compartments show peculiar rates of decay of interactions with increasing genomic distance, suggesting that the underlying compaction of chromatin belonging to each cluster is different: loci separated by a certain linear distance belonging to the active compartment show a lower frequency of interactions than loci separated by the same linear distance belonging to the inactive compartment, pointing to the latter being more compacted (Lieberman-Aiden et al., 2009; Sexton et al., 2012).

The increased resolution due to deeper sequencing has further resolved domain clustering, identifying six compartments that correlated with several chromatin marks (Rao et al., 2014). Two of these compartments (named A1 and A2) are associated with active chromatin, with A2 having a later replication time, being less GC-rich and containing longer genes than A1. The remaining four compartments (B1 to B4) are associated with more inactive chromatin: B1 shows marks typical of facultative heterochromatin; B2 comprises most pericentromeric heterochromatin, is associated with the nuclear lamina and the nucleolus; B3 is also associated with the nuclear lamina, but not with the nucleolus; B4 was only identified on human chromosome 19 and contains genes of the KRAB-ZNF superfamily (Rao et al., 2014).

At a smaller scale, Hi-C maps have revealed that chromosomes are further subdivided into a series of "chromosomal domains" (also called "Topologically Associated Domains" or

TADs) (Dixon et al., 2012) (**FIGURE 1C**). Domains are characterised as highly self-interacting modules that are separated from each other by "domain boundaries", or elements across which contacts do not cross.

TAD boundaries are normally defined using one of two different statistics: the "directionality index" or the "insulation score".
The first method, developed by Dixon et al. (2012), is based on the reasoning that fragments located at the periphery of TADs will interact preferentially either downstream (for sites located "at the beginning" of a TAD) or upstream (for sites located "at the end" of a TAD); it follows that TAD boundaries will be found where the directionality index undergoes a significant shift from an upstream bias to a downstream bias.
The "insulation score", or "scaling factor", is a different statistic that is obtained by comparing the observed number of interactions between two fragments separated by a certain linear distance and the average number of interactions for fragments separated by that distance in the whole genome. For example, if two fragments separated by 100 kb are interacting with a frequency of 50, and the average interaction frequency for sites located 100 kb apart is 500, the distance of the two fragments is "scaled" to fit the average interaction/distance decay curve, (in this example it would be scaled to whatever distance has a genome-wide average interaction frequency of 50). Given that generally interaction frequency becomes lower the more two fragments are separated, the scaled distance in this toy example wil be >100 kb. Thus, a high "scaling factor" would mean that two fragments interact less frequently than expected, making them more likely to be located on opposite sides of a TAD boundary (Sexton et al., 2012). Despite the different approaches to the problem, the two methods give comparable results (Lajoie et al, 2015).

The observation that TADs exist in multiple species – human, mouse, *Drosophila* and yeast (Sexton et al, 2012; Dixon et al., 2012; Nora et al., 2012; Mizuguchi et al., 2014), in multiple cell types (Dixon et al., 2012; Phillips-Cremins et al., 2013; Sofueva et al., 2013; Rao et al., 2014) and even in individual cells (Nagano et al., 2014), implies that they represent a fundamental organizing principle of chromosomes. In fact, preliminary data suggests that domain organisation within syntenic regions is evolutionarily conserved to some extent between human and mouse (Dixon et al., 2012).

Functional importance of chromosomal domains.

Although it is still unclear what the functional role of chromomal domains is, it has been suggested that they could function to spatially restrict distal regulatory elements with gene promoters – facilitating the search for target genes. A step-change in our understanding of gene regulation came from the observation that transcriptional activation often involves long-range regulation from distal regulatory elements, for example enhancer-promoter interactions can take place between loci separated by distances ranging from tens of kilobases up to the megabase scale (Kleinjan and van Heyningen, 2005; Anderson and Hill, 2014). The mechanisms for how genes could be regulated from a distance have been the subject of intense research over the years, with the "looping model" being most prevalent (Ptashne, 1986). The influence of chromatin looping on gene regulation has been demonstrated at various genomic loci in many species, with one classic example being the β-globin gene cluster (Wijgerde et al, 1995; Tolhuis et al., 2002). High-level expression of the genes in this locus requires the presence of a cluster of DNAse hypersensitive sites known as the Locus Control Region (LCR) located 50 kb away. The use of 3C showed how in foetal liver, where the β-globin genes are strongly expressed, the locus adopts a specific three-dimensional conformation, whereby the promoters of the genes are brought into close proximity to the LCR and the intervening DNA is "looped out". Importantly, this configuration was not observed in the brain, where the genes are silent and the locus adopts a linear conformation, i.e. the interaction frequencies are directly correlated with the genomic distance (Tolhuis et al., 2002). Similar observations have been made at the T_H2 cytokine locus (Spilianakis et al, 2004) and at the α-globin locus (Vernimmen et al., 2007).

Indeed, some evidence suggests that TADs may indeed facilitate enhancer-promoter contacts. High-resolution studies (Phillips-Cremins et al., 2013; Sofueva et al., 2013) using C-methods have observed structures analogous to topological domains at scales smaller than TADs (sometimes termed sub-TADs or loops). While the larger domains appear constitutive, at this more local level the differences in chromatin structure across cell types become more prominent (Phillips-Cremins et al., 2013; Jin et al., 2014; Rao et al., 2014). These smaller structures have been linked with the formation of enhancer-promoter chromatin loops and the regulation of gene expression (**FIGURE 1D**): cell-type specific interactions at the sub-TAD scale correlated with cell-type specific gene expression patterns (Phillips-Cremins et al., 2013); disruption of contacts within TADs caused widespread gene deregulation (Sofueva et al., 2013); ablation of an intra-TAD loop

boundary caused alteration of the expression profiles of the neighbouring genes (Dowen et al, 2014). One study mapped interactions between enhancers and promoters in the sub-domain distance range (up to about 50 kb) and revealed how the genome architecture of a certain cell type can be poised to respond to external stimuli (Jin et al., 2014).

In support of their functional relevance, TADs have been correlated with epigenetic marks in *Drosophila* (Sexton et al., 2012) and have been linked to expression regulation in mouse. In one case, during X-inactivation the ratio between the expression of the non-coding transcript *Xist* or its antisense version *Tsix* dictates the activation state of the chromosome; The ratio of *Xist* and *Tsix* is modulated by antagonising elements clustered within two different TADs on either side of the *Xist/Tsix* locus (Nora et al., 2012). In another example, the gene coding for the developmental regulator Sonic HedgeHog (*Shh*) is regulated by an array of enhancers spread across 1 Mb of linear distance, that restrict its spatiotemporal expression during ontogenesis (Anderson and Hill, 2014); as evidence that domains facilitate transcriptional regulation, *Shh* and its cis-regulatory elements all fall within the same TAD (Anderson et al., 2014). While the two examples mentioned above strongly point to TADs as key elements for transcriptional regulation, they do not formally prove that they are necessary for it. A recent study investigated a series of chromosomal rearrangements that cause abnormal expression patterns for a series of neighbouring developmental genes, leading to aberrant distal limb development in human and mouse. Using a combination of CRISPR, 4C-seq, RNA-seq and in-situ RNA hybridisation the authors were able to show how the pathological phenotypes arised only when the rearrangements eliminated a TAD boundary, providing direct evidence for the role of TADs in the regulation of gene expression (Lupiáñez et al., 2015).

Consistent with the idea that TADs might restrict the range of enhancer-promoter interactions, Noordermeer et al. (2011) have shown that a β-globin LCR integrated in an ectopic location cannot contact the β-globin cluster located on the homologous chromosome and does not majorly alter the interaction profile of the recipient site before the integration. The LCR is thus unable to freely "search" for target genes anywhere in the genome, but rather has its mobility constrained by its chromosomal context.

Decades of work and multiple complementary approaches are beginning to reveal the nature of chromosome architecture at multiple scales and its impact on genome function.

In summary, chromsomes are non-randomly organised into chromosome territories that tend to be very stable in their large-scale framework. TADs act as modular units of CTs within which regulatory contacts can occur. Such enhancer and promoter contacts help to compact each chromatin fibre into its unique 3D shape. This kind of organisation has been preserved throughout evolution and appears intimately linked with the functionality of the genome.

What are TADs?

The hierarchical nature of chromatin architecture, with smaller DNA loops becoming included into progressively larger loops, may create some confusion as to how certain structures are defined. For example, when is a loop big enough to be called a TAD?

On the one hand, one may argue that the distinction between so-called "TADs" and "sub-TADs" is merely an artificial one, a product the progression from the lower resolution of the contact maps in the early studies (Dixon et al., 2012; Sexton et al., 2012) – that only allowed the identification of larger "TADs" – to the higher definition of the later "C" datasets (Phillips-Cremins et al., 2013; Sofueva et al., 2013; Rao et al., 2014) – that enabled the identification of smaller-scale "sub-TAD" structures. On the other hand, however, data from knockout- or CRISPR-based experiments points to some biological distinction between the two structural entities. Depletion of cohesin or CTCF seems to affect sub-TADs, while leaving TADs mostly intact (Sofueva et al., 2013; Zuin et al., 2013). Accordingly, CRISPR-mediated deletion of individual CTCF sites is able to eliminate boundaries at the sub-TAD level (Dowen et al., 2014; Narendra et al., 2015), while the deletion of larger DNA segments (~50-100 kb) is needed to ablate a TAD boundary (Nora et al., 2012; Lupiáñez et al., 2015). Hence, TADs appear to be markedly more stable than sub-TADs. A likley explanation for these data is that chromatin structures exist as "a hierarchy of block-like structures" (Lajoie et al., 2015) lying in a spectrum of scales and stabilities and that TADs and sub-TADs represent opposite ends of such spectrum.

The methods that are currently used to identify TADs (described above) are thresholded so that they can reliably identify the roughly megabase-sized blocks identified in the original study (Dixon et al., 2012). Given the complex nature of chromatin architecture and that the definition of TADs represents an artificial, albeit useful in some cases, simplification, I decided not to rely on TAD definition for the majority of the analyses conducted in this

thesis, preferring to focus on the quantification of interactions across whole sections of the contact maps or around loci of interest. In a few instances, however, TADs were defined to support my findings and help place them in the framework of the current literature.

The architectural protein CTCF.

Such a specific and evolutionarily conserved chromosomal organisation implies that mechanisms exist to establish these structures. The boundaries separating chromosomal domains in mammalian cells are enriched for CCCTC-Binding Factor (CTCF), cohesin proteins, housekeeping genes and Short Interspersed Nuclear Elements (SINEs) (Dixon et al., 2012). While the role of transcription and SINEs with respect to the establishment or maintenance of mammalian domain boundaries remains elusive, a growing body of evidence points to CTCF as a key factor in organising the architecture of the genome.

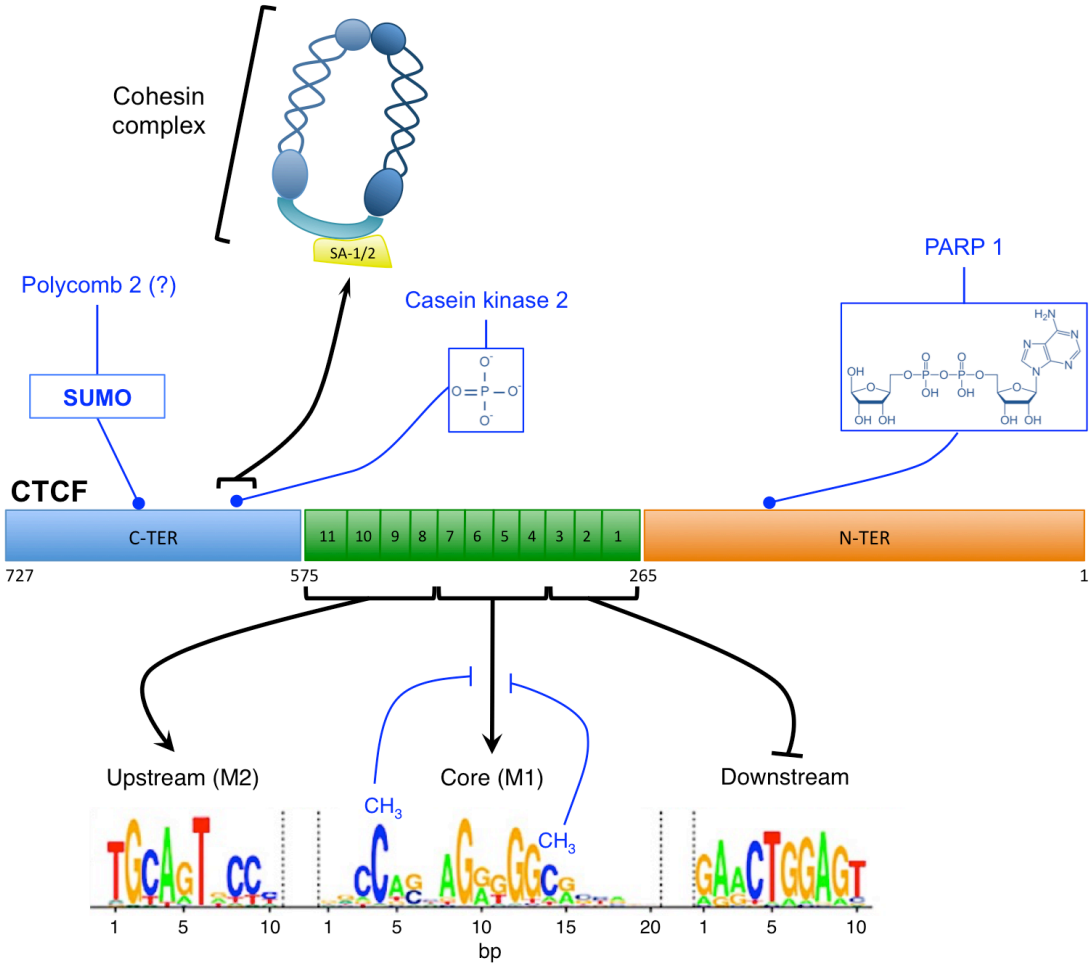
CTCF is a conserved DNA-binding protein.

The zinc finger protein CTCF is conserved across most of the animal evolutionary tree, but notably absent in yeast, derived nematodes such as *Caenorhabditis elegans*, fungi and plants (Heger et al., 2012). The 11 zinc fingers in the central region of CTCF bind in a combinatorial manner to a well-defined, information-rich DNA sequence motif (Filippova et al., 1996; Nakahashi et al., 2013). The sequence underneath the majority of CTCF sites contains a well conserved, non-palindromic ~20 bp core motif (also referred to as the M1 motif, Schmidt et al., 2012). This motif is unlike almost all known Transcription Factor (TF) motifs, in that it is longer and information-rich, making CTCF unique among DNA binding proteins. Up- and downstream of the core consensus sequence, shorter motifs (~ 10bp) have been described which are thought to support or destabilise CTCF binding to chromatin, respectively (Schmidt et al., 2012; Nakahashi et al., 2013) (**FIGURE 3**). At any single CTCF binding site, the core motif alone or together with one or both of the other modules can be present, with different outputs on the ability of the protein to bind there. A comparison of mouse strains that had single-nucleotide variants of individual CTCF sites revealed that the affinity of CTCF for a certain site is strongly correlated with its similarity to the core motif consensus sequence; when the upstream motif is present, the occupancy of CTCF is also directly correlated with the motif's similarity to its consensus; conversely, the downstream motif showed a negative correlation between similarity to its consensus and CTCF binding (Nakahashi et al., 2013). Notably, a subset of CTCF sites lack any

sequence motifs, suggesting the possibility of an alternative mechanism of recruitment for this factor. However, such sites have a markedly lower affinity when compared to sites containing the core motif (Nakahashi et al., 2013).

Chromatin binding landscape of CTCF.

CTCF binds to vertebrate genomes at tens of thousands of sites all along chromosome arms (Kim et al., 2007; Xie et al., 2007). Similarly to general transcription factors, CTCF binding is strongly correlated to gene density (Kim et al., 2007), with its greater share found in intergenic regions (~46%), but also in the gene bodies (~34%: 22% in introns and 12%



adapted from Xiao et al., 2011; Wang et al., 2012; Ong and Corces, 2014 and Nakahashi et al., 2013

FIGURE 3 – Schematic of the molecular relationships of CTCF. A C-terminal portion of CTCF can interact with the SA-2 subunit of the cohesin complex. CTCF can be post-translationally modified by SUMOylation, phosphorylation or Poly(ADP-ribos)ylation. CTCF interacts with DNA through a complex conserved motif using its 11 zinc-fingers. A core motif is present at most CTCF sites and is bound by the central zinc-fingers; an upstream motif can stabilise the binding, while a downstream DNA element can antagonise it. Note that CTCF is depicted in a reverse orientation to reflect the way it binds to the motif. Methylation of the DNA interferes with CTCF binding.

in exons) and near TSSs (~20%) (Kim et al., 2007). The binding of CTCF to chromatin can be regulated at multiple levels. First, binding of CTCF to its consensus can be regulated by DNA methylation. When the cytosine bases are methylated, the binding of CTCF is impaired (Bell and Felsenfeld, 2000; Hark et al., 2000; Engel et al., 2004; Wang et al., 2012). Bisulphite-sequencing analysis of a subset of CTCF sites revealed that DNA methylation of CpG residues influencing CTCF binding occurs predominantly at two specific positions of the core consensus (Wang et al, 2012).

Second, the presence of nucleosomes at a CTCF consensus motif can also interfere with binding (Kanduri et al., 2002), which normally happens in nucleosome-free regions. Third, post-translational modifications of the protein itself, such as poly(ADP-ribosyl)ation (Yu et al., 2004), SUMOylation (MacPherson et al., 2009) and phosphorylation by casein-kinase 2 (CK2) (Klenova et al., 2001; El-Kady et al., 2005) can affect CTCF function without interfering with its binding to DNA (**FIGURE 3**).

CTCF acts as a multifunctional chromatin looper.

CTCF was originally identified as a negative regulator of *Myc* expression (Lobanenkov et al., 1990) and has since been implicated in a variety of functions related to the organisation of chromatin (Phillips and Corces, 2009; Ong and Corces, 2014).

Classically, CTCF is known to function as an insulator, blocking communication between gene promoters and distal enhancers. This role was first described for CTCF in transgenic assays where an enhancer and a reporter gene were separated by a 1.2 kb DNA fragment containing a DNAse hypersensitive site located upstream of the chicken β-globin locus. The capacity of this element to function as an insulator was found to be dependent on a 42 bp sequence bound by CTCF (Bell et al., 1999). Interestingly, in the same study, the authors note how this sequence can work as an insulator only when placed between the enhancer and the promoter and that when a gene is placed between two copies of a larger 1.2 kb fragment containing the CTCF binding site it becomes protected from ectopic regulation. This result was corroborated when sequences within the mouse or human Imprinting Control Region (ICR) of the H19/Igf2 locus were placed between an enhancer and a reporter gene in a transgenic assay. The ICR region contains several CTCF binding sites and was able to act as an insulator in a CTCF-dependent manner. Moreover, methylation of the ICR – as seen on the paternal allele – was found to abolish both CTCF

binding and insulator activity from this DNA element, thus involving CTCF in the mechanism of genetic imprinting (Bell and Felsenfeld, 2000; Hark et al., 2000).

Over the years, a plethora of other candidate functions have been suggested for CTCF, including: activating transcription (Vostrov and Quitschke, 1997), participating in X-inactivation (Donohoe et al, 2007), forming a barrier between active and inactive epigenetic domains (Cuddapah et al., 2008), tethering chromatin domains to the nuclear Lamina (Guelen et al., 2008), regulating somatic recombination at the antigen receptor loci (Guo et al., 2011; Ribeiro de Almeida, 2011), regulating expression of gene clusters – such as the Hox (Moon et al., 2005; Kim et al., 2011), MHC (Majumder et al., 2008; Majumder et al., 2010) or Protocadherins (Hirayama et al., 2012) – influencing transcriptional pausing and splicing (Shukla et al., 2011), and facilitating enhancer-promoter interactions (Kuzmin et al., 2005; Sanyal et al., 2012).

This wide range of diverse functions may actually be reduced to a single one: CTCF mediates DNA loops and in so doing brings together loci that are separated by large linear distances. This observation has been first made at the β-globin locus, where the three-dimensional structure formed in erythroid cells (Tolhuis et al., 2002) is lost after depletion of CTCF or disruption of a CTCF binding site (Splinter et al., 2006). A parallel study reported that CTCF-mediated long-range chromatin interactions are responsible for looping of the imprinted *H19/Igf2* locus and that a mutation in a CTCF binding site at the maternal ICR enables *Igf2* to contact a group of enhancers from which it is normally kept separate (Kurukuti et al., 2006). Similar loops at the MHC class II locus were found to be dependent on CTCF and to influence the expression of this gene cluster (Majumder et al., 2008). The involvement of CTCF in chromatin looping has led to the hypothesis that the nature of the sites that are involved in each looping interaction could produce variable, context-dependent functional outcomes that may explain the array of functions that have been attributed to CTCF (Ong and Corces, 2014).

CTCF and cohesin collaborate to anchor chromatin loops.

The best-characterised mechanism by which CTCF can loop DNA is through anchoring of the cohesin complex (Parelho et al., 2008; Rubio et al., 2008; Stedman et al., 2008; Wendt et al., 2008). Cohesin is a deeply conserved protein complex assembled in a ring-like structure that is responsible for tethering sister chromatids together during mitosis by encircling two DNA filaments (Michaelis et al., 1997; Gruber et al, 2003). Three subunits

participate in the formation of the core ring: two "Structural Maintenance of Chromosome" family subunits, Smc1 and Smc3, and a klesin family subunit, Scc1 (Rad21 in vertebrates). Another component, Scc3 is also found in association with Scc1 and in vertebrates three isoforms of Scc3 exist: SA-1, SA-2 and SA-3. The core complex can associate with a variety of other proteins that influence its loading (Ciosk et al, 2000), association with chromatin through the cell cycle (Toth et al, 1999; Rankin et al., 2005; Kueng et al., 2006) and degradation (Uhlmann et al., 2000).

Aside from its classic role during DNA replication, cohesin is present on chromatin even in non-dividing cells (Wendt et al., 2008; Pauli et al., 2010) and mutations in cohesin subunits or regulatory factors can cause developmental abnormalities in humans (Krantz et al., 2004; Vega et al, 2005), suggesting that these proteins could contribute to gene regulation. In vertebrates, the involvement of a complex essential for sister chromatid cohesion in gene expression is made possible by its extensive co-localisation with CTCF (Parelho et al., 2008; Rubio et al., 2008; Stedman et al., 2008; Wendt et al., 2008). Depending on the cell type, up to 50-80% of CTCF sites can be co-occupied by cohesin (Parelho et al., 2008; Wendt et al., 2008). CTCF and cohesin have been shown to interact directly, whereby the C-terminal region of CTCF binds to the SA-1 or SA-2 subunits of cohesin (Xiao et al., 2011) (**FIGURE 3**).

Cohesin proteins are required to fulfill CTCFs insulator activity, as knockdowns of cohesin or CTCF interfered with CTCF's enhancer-blocking action in reporter assays and produce overlapping sets of de-regulated genes (Wendt et al., 2008). Cohesin was reported to be essential for anchoring chromatin loops from CTCF binding sites *in vivo* at the H19/Igf2 (Nativio et al, 2009) and at the IFNG loci (Hadjur et al., 2009). The ability of CTCF to loop DNA has indeed been shown to underlie several of the functions mentioned in the previous section. It is worth keeping in mind, though, that CTCF might also be able to promote looping of DNA via cohesin-independent mechanisms, for example homo-dimerisation (Yusufzai et al., 2004; Pant et al., 2004) or interaction with other proteins.

In this respect, a large number of CTCF interactors and binding partners have been described, albeit usually at candidate genomic loci and without a unifying theme. Among them: nucleophosmin (Yusufzai et al, 2004), Kaiso (Defossez et al., 2005), the Thyroid Hormone Receptor (Lutz et al., 2003) and the chromodomain helicase CHD8 (Ishihara et

al., 2006) have all been implicated in CTCF's insulator function; TAF3 has instead been involved in its role as a transcriptional regulator (Liu et al., 2011); transcription factor YY1 seems to participate in CTCF-mediated regulation of the IgH locus recombination (Guo et al, 2011) and in the activity of CTCF at the X-inactivation centre (Donohoe et al., 2007); regulator of tRNA transcription TFIIIC has been shown to colocalise with CTCF at multiple sites (Moqtaderi et al., 2010); stem cell transcription factor Oct4 can directly interact with CTCF (Donohoe et al, 2009) and interferes with the association of CTCF and cohesin at the HOXA cluster in mouse Embryonic Stem Cells (mESCs) (Kim et al., 2011); the DEAD-box RNA helicase p68 and its interacting RNA SRA have been shown to aid CTCF insulator function and stabilize its interaction with cohesin at the H19/Igf2 locus (Yao et al., 2010). It should be noted, however, that none of the other characterised CTCF binding partners shows the extensive genome-wide co-localisation that it has with cohesin, pointing to the other factors as relevant only in some context-dependent scenario (Weth and Renkawitz, 2011; Ong and Corces, 2014). Given the ubiquitous expression of CTCF and cohesin, it is tempting to speculate that the association of CTCF with different factors in different parts of the genome, together with the regulation of its binding to DNA, might create a complex and versatile spectrum of modulation to adapt CTCF's ability to loop chromatin to specific loci or differentiation stages. Finally, some proposed CTCF functions, for example its involvement in alternative splicing (Shukla et al., 2011), might also be independent of chromatin looping.

CTCF is implicated in global genome architecture.

Given the role of CTCF in chromatin looping at multiple loci in the genome, it was widely hypothesised that it could be involved in the global architecture of the genome.

Indeed, a comparison of CTCF binding profiles with Hi-C data has revealed a number of interesting insights and a new fundamental function for this protein in the partitioning of chromosomes into domains. CTCF is enriched at the boundaries of TADs, with most of them (~76%) containing a CTCF site (Dixon et al., 2012). The fraction of CTCF found within boundary regions, though, represented just 15% of the total CTCF binding sites, leaving the majority of this protein bound within domains and leading to the interpretation that presence of this factor alone is not sufficient to demarcate a large-scale domain boundary (Dixon et al., 2012). However, using a high-resolution approach on seven candidate genomic regions in mouse ESCs and Neural Progenitor Cells (mNPCs), Phillips-Cremins et al. (2013), show that CTCF is able to mediate looping at both the megabase and

sub-megabase scales by partnering with other "architectural proteins". In more detail, the authors postulate that CTCF and cohesin form TAD-level, larger-scale loops that appear to remain constant across cell types, while various combinations of cohesin, Mediator and CTCF organise intra-TAD architectures. Consistently, Sofueva et al. (2013) also reported an enrichment for sites co-occupied by CTCF and cohesin at domain boundaries at multiple scales. These results indicate that CTCF is globally involved in the organisation of chromatin structure both at the level of TAD demarcation and at a more local level, internally to domains.

Given this extensive correlation between CTCF, cohesin, and domain structure it is widely hypothesised that these proteins have a causal role for chromosomal organisation. Formal demonstration of this, though, has proven difficult because depletion of CTCF (Moore et al., 2012) or cohesin in embryos is lethal and even in a cell culture setting the essential role of cohesin during cell replication makes it an unsuitable target for deletion in cycling cells. In order to circumvent this, Sofueva et al. (2013) deleted a subunit of the cohesin complex in non-cycling astrocytes before performing Hi-C. This way, they observed a widespread decompaction of large-scale chromosomal domains and a concomitant increase in inter-domain contacts without a complete dissolution of the domain boundaries. These alterations in chromatin structure were accompanied by genome-wide deregulation of gene expression. Similarly, Zuin et al., prepared Hi-C libraries after depleting either cohesin or CTCF, leading in both cases to domain decompaction without complete loss of the boundaries. However, the authors report increased interactions between domains only after CTCF knockdown and that different sets of genes appear to be de-regulated in the two mutants, leading them to suggest that CTCF and cohesin have distinct roles in genome organisation.

Another study identified activating or repressive intra-TAD domains of chromatin bordered by a pair of CTCF/cohesin sites in mESC. Upon specific deletion of one of these borders, the domain is no longer constrained, causing both the contacts and the activating or repressive activity to spread to genes lying beyond the deleted CTCF/cohesin site (Dowen at al., 2014). Similarly, CTCF was found to partition the HOXA cluster into transcriptionally active and inactive domains in mNPCs and deletion of a CTCF binding site caused interactions, histone marks and gene activity to spread until the next CTCF site (Narendra et al., 2015).

All these observations are beginning to help understand the intimate relationship between CTCF, cohesin, chromatin architecture and the function of the genome.

The landscape of evolving genomes.

When compared with the vast majority of other transcription factors, the binding motif for CTCF is remarkably better conserved, with a very similar consensus found in organisms separated by over 500 million years of evolution, such as mammals and *Drosophila* (Heger at al, 2012).

This peculiarity of CTCF has prompted researchers to investigate the conservation of its binding profile. Schimdt et al. (2012) have elucidated a mechanism of CTCF binding site evolution by studying multiple mammalian lineages. By performing chromatin immunoprecipitation followed by sequencing (ChIP-seq) in hepatocytes of five different mammals, they were able to identify a subset of deeply shared CTCF binding sites across all five genomes. They also show how in several lineages CTCF binding sites localise within clade-specific SINE retrotransposons. The authors suggest that this embedding of CTCF binding sites within SINEs has driven their evolution and spread across the genome.

Given the involvement of CTCF in global chromatin architecture, this thesis explored how the changes in its binding profile might have affected the structure of chromatin in the context of evolving genomes. In order to delve deeper into this complex interplay, it is important to introduce the evolutionary forces acting on animal genomes and the consequences of their action.

An overview of the mutations causing genomic divergence.

The differences between the genomes of living organisms have been accumulating over the course of evolution by means of processes of mutation and non-random selection. Most of the variability between genomes resides in features such as size, number of chromosomes, order of genes along chromosomes, abundance and size of introns and amount of repetitive DNA (Alberts et al., 2008). The mutations that lead to such variability range include nucleotide substitutions, duplications, insertions, deletions, inversion and translocations and can range in size from a single nucleotide to events on a chromosomal or even genome-wide scale.

Differences in genome size are often the result of very large-scale mutational events such as whole-genome duplications. Mammalian genomes, for example, have undergone two rounds of tetraploidisation events close to the inception of their lineage (Ohno, 1973), but have since maintained a relatively uniform size of around 2-3 billion nucleotides. The genomic heterogeneity within this clade mostly stems from smaller-scale duplications and deletions, replication of transposon elements and chromosomal rearrangements (Alberts et al., 2008).

Duplication is a particularly remarkable kind of mutation, because it is thought to be one of the main mechanisms through which new gene functions can arise. When a gene is duplicated, one of the copies becomes in fact "free" to mutate while the other ensures that the original functionality is kept intact (Ohno, 1968). In some cases tandem gene duplications can even sprout the formation of gene clusters, arrays of related neighbouring genes that usually retain a common functionality and whose expression can be coordinated (Lawrence, 1999).

Mammalian genomes are rich in repetitive elements that possess the ability to copy themselves and spread to other genomic positions, through a process known as transposition. Numerous families of these transposons exist and the mechanisms by which they can move can vary slightly. The majority of transposable elements in mammals consists of retrotransposons, which can re-integrate into a target site through an RNA intermediate. The relevance of transposition in mammalian genome evolution is highlighted by the large fraction of the genomic sequence that these elements cover: for example the SINEs and Long Interspersed Nuclear Elements (LINEs) families of retrotransposons together make up for over 40% of the human genome (Kazazian, 2004; Alberts et al., 2008). Other than being a source of mutation *per se*, transposon activity can influence homologous recombination that can lead to chromosomal rearrangements or duplications/deletions (Kazazian, 2004).

Nucleotide substitutions are extremely important in the study of genome evolution. These events are the result of errors made by the DNA polymerase during the replication of the DNA. When nucleotide substitutions occur in regions that are not subjected to selection (i.e. they are evolutionarily neutral), they accumulate at relatively regular rates that allow the estimation of times of divergence between species based on how different their sequences

are. This principle has become known as the "Molecular Clock" (Zuckerkandl and Pauling, 1962; Ho and Duchene, 2014).

Chromosomal rearrangements are a heterogeneous class of mutations in which the structure of a chromosome becomes altered in various different ways. From an evolutionary standpoint, chromosomal rearrangements reshuffle gene order and are responsible for the heterogeneity in chromosome numbers and karyotypes observed across species (Ruiz-Herrera et al., 2012), and can produce extremely variable outcomes, even in related groups of species, as for example Rodents (Romanenko et al., 2012). This class of mutations is particularly relevant in the evolution of genome architecture.

Conserved synteny allows comparison of rearranged genomes.

The karyotypic differences among genomes could make comparisons between them challenging. The rate at which genomes evolve, though, is not the same in all their parts. For example, the coding sequence of genes changes much more slowly than the rest of the genome, due to the fact that alterations at those loci are likely to interfere directly with protein functionality and are therefore under strong purifying selection. A similar "resilience" to change can apply to gene regulatory elements. Other genomic portions such as intronic or intergenic sequences can instead become modified more easily, as in many cases they do not immediately endanger the fitness of the individual.

The existence of conserved DNA elements makes it possible to unambiguously identify such sequences in different organisms, even when their genomes have undergone radical changes. When two sequences can be identified across species as being derived from the same ancestral sequence, they are said to be "orthologous" and can be used as markers to match the corresponding portions of the rearranged genomes. When comparing two species' rearranged genomes, the linear conservation of a set of orthologous genes or sequences that can be identified only once in each genome is considered to be a region of conserved synteny; these regions represent the portions of the two genomes that can be compared to one another. The genomic space between the edges of two conserved synteny blocks represents an area where some rearrangement has taken place over the course of evolution and has been indicated as a "breakpoint" region (Murphy et al., 2005) (**FIGURE 4**). This principle forms the basis for how the sequenced genomes of different organisms can be aligned to one another and genome coordinates can be converted between species

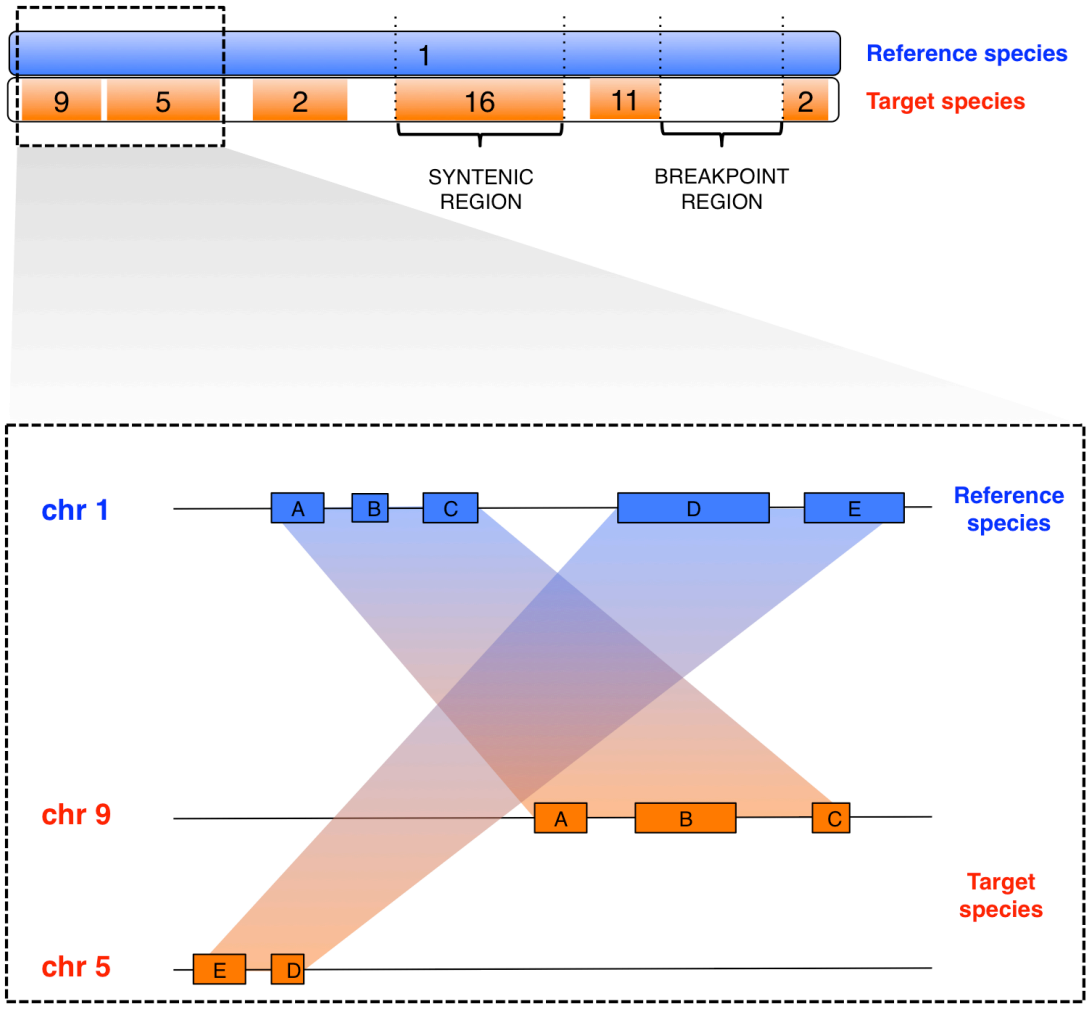


FIGURE 4 – A simplified view of synteny relationships between two species. Stretches of one-to-one orthologous markers (A-E boxes) are used to identify and map corresponding portions of two different genomes. Markers can be genes or any sequence that is sufficiently conserved and uniquely identifiable in both genomes.

(Blanchette et al., 2004; Ma et al, 2006). For simplicity, regions of conserved synteny will be referred to as "syntenic regions" or "syntenic blocks" throughout the text, acknowledging though that this is an improper, albeit common, use of the term (Passarge et al., 1999).

Three models for karyotypic evolution.

The synteny relationships between different genomes can shed light on how evolutionary forces have acted upon them and can help identify elements that have been protected from disruption because of their functional relevance. For these reasons, the comparison of syntenic blocks across multiple species has been an object of intense investigation for several decades.

Before the advent of genome sequencing, cytogenetic approaches dominated this field of research. Many studies employing these low-resolution methodologies, pioneered by Nadeau and Taylor (1984), were consistent with a model originally proposed by Ohno (1973) whereby the occurrence of chromosomal rearrangements would be a completely random phenomenon affecting the entire genome with the same likelihood ("Random Breakage Model").

The introduction of higher-resolution techniques through the analysis of whole genome sequences revealed that the distribution of breakpoints in vertebrates is actually not uniform in genomes and that "hotspot" regions undergo chromosomal rearrangement more frequently than others or might even be re-used during genomic evolution (Pevzner and Tesler, 2003; Zhao et al., 2004; Bourque et al., 2005; Murphy et al., 2005). These findings led to the postulation of the "Fragile Breakage Model" (Pevzner and Tesler, 2003), whereby breakpoints would accumulate in fragile regions of the genome that are more susceptible to break or undergo improper recombination.

A third model has more recently been suggested to explain the unequal distribution of breakpoint regions in vertebrate genomes, highlighting the existence of groups of genes that tend to be inherited together because of regulatory relationships between them. These genes are usually involved in fundamental biological processes such as regulation of transcription or embryonic development and can be under the control of several elements, that are often preserved through evolution as Highly Conserved Noncoding Regions (HCNR). Sometimes, one of these regulatory elements might become embedded in an unrelated ("bystander") gene, for example in an intron. Such an occurrence would create a functional association between the regulated gene(s) and the bystander, so that they would now form a Genomic Regulatory Block (GRB). Due to the presence of long-range chromatin interactions, GRB can span large genomic distances. A chromosomal breakage disrupting the GRB would interfere with the proper regulation of gene expression and be selected against (Becker and Lenhard, 2007; Kikuta et al., 2007; Irimia et al., 2012; Irimia et al., 2013). Differently from the other models described above, this one takes selection into account to explain the unequal distribution of breakages and can be referred to as a "Selective Breakage Model".

While the Random Breakage Model did not withstand the test of high-resolution studies, the debate still stands as to which model between the Fragile Breakage and the Selective Breakage best describes the uneven distribution of chromosomal rearrangements throughout the genome. Importantly, the two models are not mutually exclusive and some portions of the genome might be more prone to break, while others will encompass regulatory blocks that won't tolerate disruption.

The influence of 3D chromatin interactions on the occurrence of chromosomal rearrangements has not been thoroughly assessed. One study has described how regions that are contiguous in one genome (e.g. in mouse) but genomically distant in another (e.g. in human) tend to be significantly closer in 3D space in the latter than expected by chance. This can be explained by the fact that two regions that were ancestrally separated, but lied close to one another in 3D space, became adjacent after the occurrence of a breakpoint. Alternatively, two regions that were ancestrally contiguous became separated by a genomic rearrangement, but remained close spatially, possibly to ensure the correct functioning of the genome. The notion of conserved spatial proximity has been described as 3D synteny (Véron et al., 2011).

1.2 AIM OF THE WORK

Given the conservation of the principles governing nuclear architecture, the fact that CTs themselves are a collection of long-range interactions, and the relevance of chromosomal and chromatin spatial organisation for genome function, I hypothesise that the folding of chromatin at a more local scale has also been subjected to some level of selective pressure over the course of mammalian evolution. If so, this could imply that syntenic regions in different genomes fold in a similar way. Such conservation would further strengthen the link between chromatin architecture and genome function.

To test this hypothesis, I have generated high-resolution, unbiased genome-wide interaction maps using the Hi-C technique from the hepatocytes of four different mammalian species,. The contact maps will be used to quantitatively compare the interactions occurring within orthologous regions of conserved synteny or at loci that have undergone breakage and rearrangement over the course of mammalian evolution. This allows the investigation of the conservation of interactions within and across vertebrate lineages, providing a first in-depth study of the evolution of chromatin architecture at a high resolution.

Despite several lines of evidence pointing to CTCF as a key chromatin organiser, only 15% of its binding sites are enriched at large-scale domain boundaries (Dixon et al., 2012). On the other hand, a subset of CTCF sites that have been deeply conserved across mammals has been identified, but a clear function that would distinguish it from the less conserved binding loci has not been found yet (Schmidt et al., 2012). With this in mind, I hypothesise that conserved CTCF binding sites represent key architectural sites in the genome and that they engage in conserved long-range interactions.

To address this, I have integrated CTCF binding data with Hi-C chromatin interaction from the same mammalian species. This analysis aims to explore the relationship between the conservation of CTCF binding sites and its role in global genome architecture over the course of evolution.

CHAPTER TWO:

COMPARATIVE HI-C REVEALS THE EVOLUTION OF CHROMOSOME TOPOLOGIES

2.1 INTRODUCTION

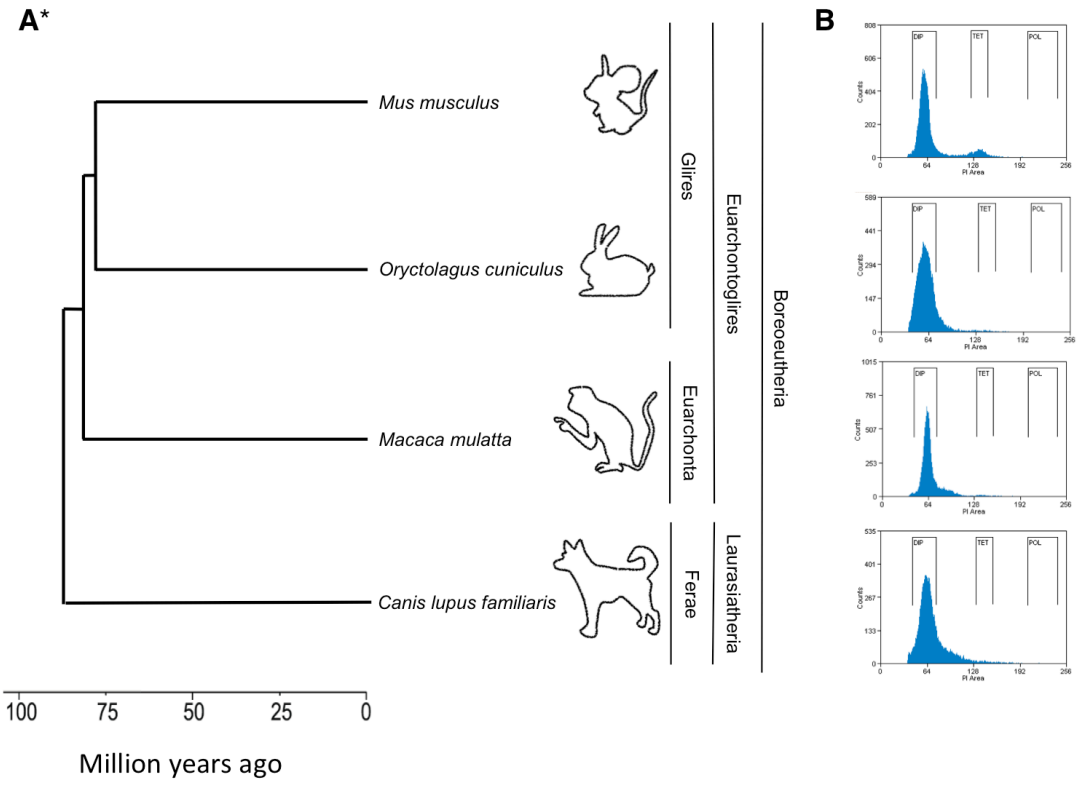
Over the course of evolution, genomes become increasingly modified due to the occurrence of various kinds of mutations and chromosomal rearrangements. For example, the non-coding portions of the human and mouse genomes have a sequence identity of 50% or less and their karyotypes have been largely re-shuffled (Mouse Genome Sequencing Consortium, 2002). The nuclear organisation of the genome has been shown to be an evolutionary conserved feature of eukaryotes (Tanabe et al., 2002a; Tanabe et al., 2002b, Mayer et al., 2005). Preliminary reports indicated that, despite their divergence, the high-order structure of chromatin is qualitatively conserved between human and mouse (Dixon et al. 2012). In order to assess the extent of chromatin topology conservation and divergence across the mammalian lineage, comparative chromosome conformation capture (Hi-C) was used to examine the architecture of the genome in primary liver cells from four different mammalian species.

In Hi-C (High-throughput Chromosome conformation capture) the genome is digested *in-nucleo* and the sticky ends of the restriction fragments are filled-in with a biotinylated nucleotide before re-ligation, so that the ligation junctions between two fragments can be enriched and characterised in an unbiased global fashion using paired-end sequencing (Lieberman-Aiden et al., 2009) (**FIGURE 2B**). *Cis*-interaction data obtained with this technique is normally represented as a heatmap in which one chromosome is plotted against itself and colour coding reflects the enrichment of observed contacts between two chromosomal coordinates with respect to an expected rate of interactions between the same two loci based on a probabilistic model (Yaffe and Tanay, 2011; see also MATERIALS & METHODS) (**FIGURE 2C**).

Liver cells were chosen as the source of material for Hi-C experiments. Several reasons led to this choice. First, liver tissue is relatively homogeneous, with an estimated 70-80% of its mass made by hepatocytes (Duncan, 2013). Second, in non-injury conditions, hepatocytes exist primarily in G_0, with fewer than 0.1% of them cycling at a given time (Fausto et al., 2003), ensuring that the chromatin interactions analysed are enriched for one cell cycle stage. Third, liver cells are easily attainable from most species of interest.

Liver cells were collected from mouse (*Mus musculus*, Mmus) rabbit *(Oryctolagus cuniculus*, Ocun), rhesus macaque (*Macaca mulatta*, Mmul) and dog (*Canis lupus familiaris*, Cfam),

separated from each other by 80-92 million years. Mouse and rabbit are closest to each other, and are included in the grandorder *Glires* (Most Recent Common Ancestor, MRCA, ~80 Mya); they are followed by macaque, grouped together with mouse and rabbit in the superorder *Euarchontoglires* (MRCA ~83 Mya). Dog belongs to the separate superorder *Laurasiatheria*, but to the same magnorder *Boreoeutheria* (MRCA ~92 Mya) as the other species (Meredith et al., 2011) (**FIGURE 5A**). Because the mouse genome is better assembled and annotated, it was selected as the reference species in all the comparisons described below.



**adapted from Meredith et al, Science, 2012*

FIGURE 5 – Phylogenetic relationships and liver cell ploidy of the species in this study. (**A**) All species are separated by 80-92 million years and their relative distance is reflected by the number of "nodes" separating them. The three groupings shown on the right are hierarchical (from left to right grandorder < superorder < magnorder). Data adapted from Meredith et al., 2012. (**B**) Shown are DNA content profiles obtained by staining ~1x10^6 liver cell nuclei from each of the species indicated. The fraction of diploid nuclei (DIP bar) appears dominant in all species accounting for 85% of the nuclei in mouse, 85% in rabbit, 82% in macaque, and 75% in dog. A subpopulation of tetraploid nuclei (TET bar) is only visible in mouse (10%). Nuclei with higher ploidies (POL bar) are negligible in all samples.

2.2 RESULTS

Liver cell nuclei from all species are mostly diploid.

Mammalian hepatocytes are known to exhibit polyploidy (Milne, 1909; Duncan, 2013). The fraction of polyploid hepatocytes increases with age (Enesco and Samborsky, 1983), and varies greatly between species (Vinogradov et al., 2001). Since ploidy can be a confounding factor in the interpretation of the genome-wide chromatin interaction profiles obtained with Hi-C, the DNA content of the liver cells from each of the four species of interest was profiled using propidium iodide staining followed by FACS analysis (**FIGURE 5B**). Liver populations analysed from rabbit, macaque and dog exhibited mostly diploid nuclei (with 85, 90 and 75% of diploid nuclei, respectively). In the mouse, the liver used for one of the Hi-C replicates had a sub-population of tetraploid nuclei (19%), but the majority (60%) were diploid (data not shown). The liver cells used to prepare the other mouse replicate were collected from younger animals and exhibited a stronger enrichment for diploid DNA content (85%). Details of the samples used are presented in **TABLE 1**.

	Strain	Age	Sex	Diploid nuclei %
mouse 1	C57/BL6	9 weeks	F	60
mouse 2	Outbred	2-4 weeks	M	85
rabbit	-	young adult	M	85
macaque	-	young adult	M	90
dog	-	young adult	M	75

TABLE 1 – Samples analysed in this thesis.

Syntenic regions exhibit a strongly correlated chromatin structure across multiple mammalian species.

Hi-C libraries were prepared from the collected liver cells using a modified version of the protocol first described in Lieberman-Aiden et al. (2009). The original protocol performed the re-ligation step in diluted conditions to favour the joining of fragments that are kept together by a cross-linked protein (Dekker et al., 2002). More recently, protocols have been adapted to re-ligate the digested ends directly within intact nuclei, yielding higher-quality libraries (Sofueva et al., 2013; Nagano et al., 2013; Rao et al., 2014). This version of the Hi-C protocol was applied to mouse liver samples that had been fixed by immersion in 10% formalin and had been subsequently kept at -80ºC. Single cell suspensions were prepared by Dounce homogenising the tissue and the Hi-C protocol was performed. Analysis of

digestion and re-ligation profiles on an agarose gel showed efficient enzymatic reactions but a progressive reduction in the amount of material remaining due to extensive lysis of the nuclei during the permeabilisation/digestion steps. A number of variations of the protocol were performed to prevent nuclear lysis including: removal of cytoplasmic lysis step, changing SDS concentration during permeabilisation, changing the enzymatic reaction incubation temperatures and fixing the cells after dissociation from the organ (**TABLE 2**). To ensure effective enzymatic reactions without nuclear lysis, liver cells required longer formaldehyde fixation of cells in suspension and harsher permeabilisation conditions (see MATERIALS & METHODS for details). Two replicates were prepared for each of the species.

Fixation conditions	Permeabilization conditions	Digestion conditions	Outcome
10% formalin (immersion)	1h, 0.3% SDS, 37°C, 800 rpm 1 h, 2% TX, 37°C, 800 rpm	36 h, 37°C, 800 rpm	Nuclei lysed during digestion
10% formalin (immersion)	1h, 0.6% SDS,37°C, 800 rpm 1 h, 4% TX, 37°C, 800 rpm	36 h, 37°C, 800 rpm	Nuclei lysed during digestion
10% formalin (immersion)	None	24 h, 37°C, 800 rpm	Failed ligation
10% formalin (immersion)	1h 0.1% SDS, 37°C, 800 rpm 1 h, 0.6% TX, 37°C, 800 rpm	24 h, 37°C, 800 rpm	Nuclei lysed during digestion
10% formalin (immersion)	1h 0.1% SDS, 37°C, 800 rpm 1 h, 0.6% TX, 37°C, 800 rpm	36 h, 37°C, 800 rpm	Nuclei lysed during digestion
10% formalin (immersion)	1h 0.1% SDS, 37°C, 800 rpm 1 h, 0.6% TX, 37°C, 800 rpm	24 h, 37°C, 800 rpm	Nuclei lysed during digestion
10% formalin (immersion) + 10 min 1% FA (after homogenisation)	1h, 0.3% SDS, 37°C, 800 rpm 1 h, 2% TX, 37°C, 800 rpm	48 h, 37°C, 800 rpm	Nuclei lysed during digestion
10% formalin (immersion)	1h 0.1% SDS, 37°C, 800 rpm 1 h, 0.6% TX, 37°C, 800 rpm	15 h, 37°C, 800 rpm	Nuclei lysed during permeabilization
10% formalin (immersion)	30 min 0.1% SDS, 37°C, 800 rpm 30 min, 0.6% TX, 37°C, 800 rpm	15 h, 37°C, 800 rpm	Nuclei lysed during permeabilization
10% formalin (immersion)	None	36 h, 37°C, 800 rpm	Nuclei lysed during digestion
10% formalin (immersion)	None	15 h, 37°C, 800 rpm	Nuclei lysed during digestion
10% formalin (immersion)	None	36 h, RT + 24 h, 37°C	Nuclei lysed during digestion at 37°@C
10% formalin (immersion)	None	72 h, RT	Inefficient digestion, failed ligation
10% formalin (immersion)	None	96 h, RT	Inefficient digestion, failed ligation
10% formalin (immersion)	None	96 h, RT (+ 5mM ATP)	Inefficient digestion, failed ligation
10 min 1% FA (after homogenisation)	1h 0.1% SDS, 37°C, 800 rpm 1 h, 0.6% TX, 37°C, 800 rpm	15 h, 37°C, 800 rpm	Successful

FA= Formadehyde TX = Triton X-100

TABLE 2 – List of Hi-C protocol variations attempted on mouse liver.

Due to sample availability, biological replicates were prepared only for mouse, while replicates for all other species were technical, prepared by independently processing different slices of the same liver. Digestion and ligation efficiencies were assessed by running de-crosslinked aliquots of the samples on an agarose gel. The efficiency of the digestion was also evaluated by qPCR amplification across a set of random *HindIII* restriction sites and quantifying the reduction in PCR amplification as compared to an undigested control. With the optimised protocol, digestion and ligation efficiencies were high in all libraries (**FIGURE 6A**). The fill-in reaction abolishes the *HindIII* restriction site and introduces a new *NheI* restriction site at ligation junctions (**FIGURE 6B**). For this reason, the efficiency of the fill-in reaction can be assessed by amplifying across ligation

junctions and digesting the amplicons with either *HindIII* or *NheI* and comparing the extent of digestion by each enzyme. A perfectly filled-in sample should be completely digested by *NheI*, but unaffected by *HindIII* digestion. Fill-in efficiency ranged between roughly 50 and 70% as assessed by visual inspection of the gels (**FIGURE 6C**).

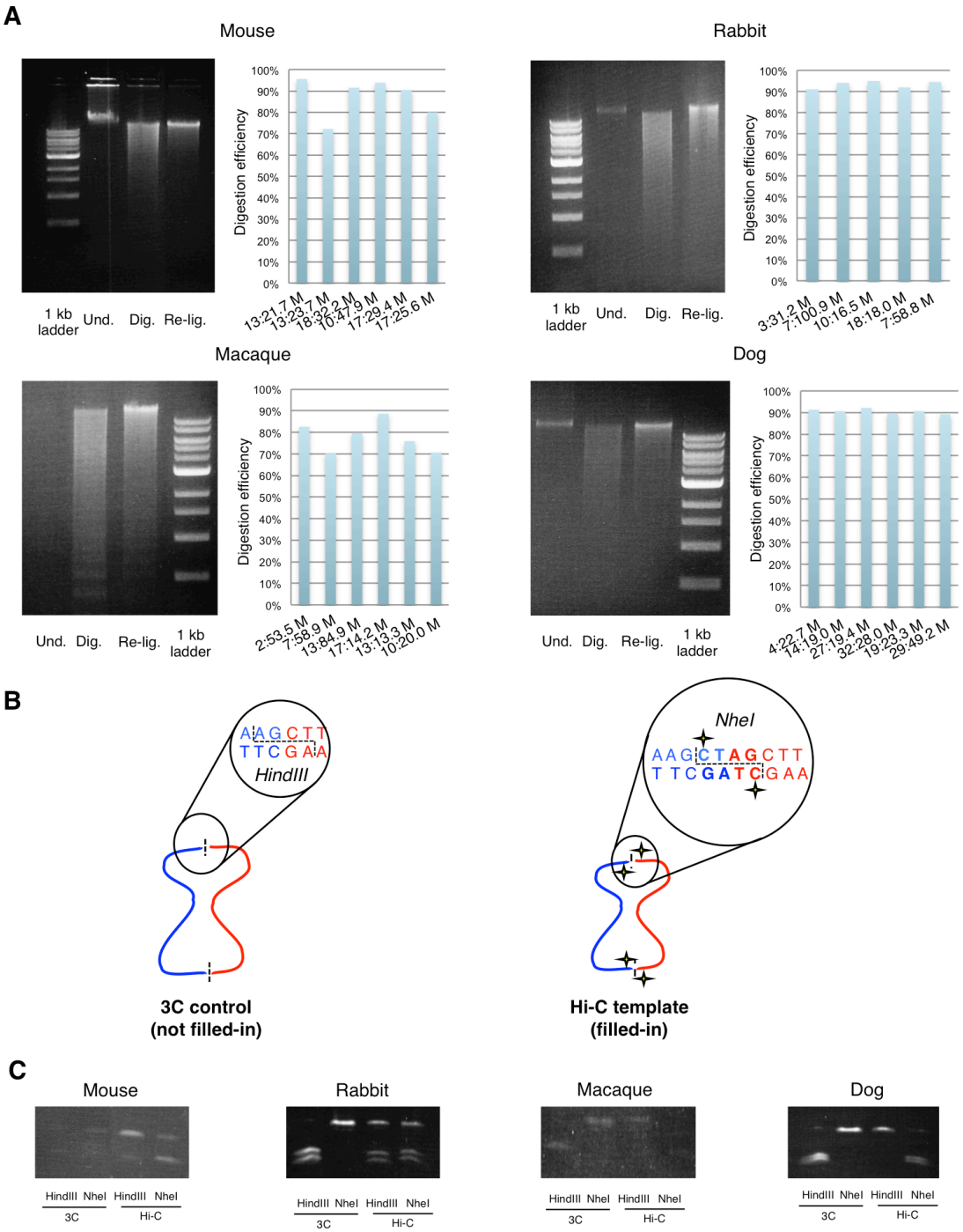


FIGURE 6 – Hi-C templates prepared from all four species exhibit good quality controls. (**A**) Digestion and re-ligation efficiencies were assessed by running sample aliquots on a 1% agarose gel (left panels). Digestion efficiency was also evaluated by qPCR amplification across random HindIII restriction sites (right panels). HindIII sites in the qPCR chart are labelled by their genomic location in megabases (chromosome number:megabase number). (**B**) Addition of the fill-in sequence abolishes a HindIII restriction site and introduces a NheI restriction site. (**C**) Fill-in efficiencies were assessed by digesting ligation junctions with both restriction enzymes and range between approximately 50-70% as estimated by visual assessment of the gels (see Methods for more details).

After paired-end sequencing, the reads were aligned to the genome of their species. According to the alignment, result each read pair was put into one of three groups: aligned on both ends (A/A); aligned on one end (A/NA); not aligned on any end (NA/NA). The macaque samples had a lower fraction of aligned reads, likely due to a less reliable genome assembly. Since the sequencing inserts are obtained after sonication, which breaks the DNA in random points, one of the reads in each pair could potentially include the ligation junction. In this case that read would not be alignable to the genome. In order to address this problem, the reads that could not be aligned within the A/NA group were scanned for the presence of the ligation junction sequence; this sequence was then computationally removed and the read was re-aligned to the genome, allowing for a partial "rescue" of the A/NA read pairs (work of Dr. Wen-Ching Chan). The results for these alignments are reported in **TABLE 3**.

Filtering and normalisation of the Hi-C ligation products was performed as before (Yaffe and Tanay, 2011; Sofueva et al., 2013). Since the maximum theoretical resolution for a Hi-C experiment is at the level of the restriction fragments that the genome is cut into, interactions were counted as pairwise combinations of the ends of these restriction fragments, rather than their exact genomic coordinates. Read pairs that could not be assigned to the ends of two different, non-contiguous restriction fragments were filtered out as non-informative. In greater detail, pairs mapping to the same restriction fragment were categorised as "self ligation" when they had divergent orientation (reflective of a

	mouse 1	mouse 2	rabbit 1	rabbit 2	macaque 1	macaque 2	dog 1	dog 2
Raw read pairs	131,334,081	324,765,565	181,156,401	115,120,956	180,143,310	112,727,383	282,580,527	712,887,067
Aligned read pairs	63,301,020	162,343,090	75,653,635	45,466,426	32,676,446	25,641,321	105,211,241	363,336,553
Rescued read pairs	19,898,516	44,139,288	16,626,788	22,050,771	17,952,456	12,424,566	69,764,180	27,513,813
Total aligned read pairs	76,015,342	206,482,372	88,019,178	66,757,608	53,599,765	38,670,891	174,975,417	390,850,354
PCR Duplicates	25,170,677	2,928,458	2,874,830	941,544	894,797	1,308,384	40,460,977	53,333,258
Input read pairs	47,144,425	197,312,560	66,241,698	52,803,348	50,690,957	36,576,244	129,064,213	315,832,399
Self ligation	442,905	1,056,625	522,035	210,083	452,547	151,069	982,539	2,991,629
No ligation	12,116,203	66,684,615	27,661,906	13,453,007	10,684,390	9,981,740	17,034,398	118,996,013
No restriction	1,067,072	6,985,010	1,130,703	1,118,412	1,289,545	796,994	1,568,334	15,466,420
Read out of range	5,712,658	27,018,747	12,979,239	5,568,518	7,727,899	4,912,362	9,046,971	129,127,351
Valid read pairs	27,805,587	95,567,563	23,947,815	32,453,328	30,536,576	20,734,079	100,431,971	49,250,986
Close cis pairs (<10k)	2,533,933	6,552,517	1,897,840	1,091,547	1,950,098	884,161	4,819,514	4,107,941
Far cis pairs (>10k)	19,253,297	64,163,993	18,063,113	23,971,218	23,050,587	15,503,551	63,785,704	31,154,569
Trans pairs	6,018,357	24,851,053	3,986,862	7,390,563	5,535,891	4,346,367	31,826,753	13,988,476
Unique f-end pairs	21,255,286	89,855,944	22,699,382	31,809,236	29,591,855	20,154,862	82,906,314	41,771,133
Close cis unique pairs (<10k)	1,655,513	3,793,030	1,433,306	919,631	1,587,718	795,528	2,795,560	2,522,890
Far cis unique pairs (>10k)	14,881,664	61,475,605	17,366,174	23,534,082	22,548,166	15,123,640	52,865,900	26,838,742
Trans unique pairs	4,718,109	24,587,308	3,899,902	7,355,523	5,455,971	4,235,694	27,244,854	12,409,501

TABLE 3 – Hi-C library stats. "Aligned read pairs" are pairs for which both ends could be successfully aligned. "Rescued read pairs" are pairs that could be aligned after computationally removing the ligation junction sequence from one of the reads. Discarded reads are highlighted in red.

circularised ligation fragment) or as "no ligation" when they had convergent orientation (reflective of a fragment that was digested but never re-ligated). Other non-informative products that were discarded included reads mapping very close to one another ("no restriction"). Given that all Hi-C libraries were prepared with a maximum sequencing insert size of 400 bp and all inserts should include a ligation junction, all reads should fall within 400 bp of the closest HindIII restriction site. For this reason, reads that mapped more than 500 bp away from the closest HindIII restriction site were also discarded as spurious ("segment length out of range"). A more lenient threshold (500 bp instead of 400 bp) was used to account for possible imprecisions in the size-selection process. A summary of the read statistics for each library is presented in **TABLE 3** and **FIGURE 7A**.

A high quality Hi-C library produces a larger fraction of cis- (intra-chromosomal) rather than trans- (inter-chromosomal) read pairs (Kalhor et al., 2011; Sofueva et al., 2013): based on the fact that interphase chromatin is organised into CTs, a genomic fragment within a CT will be more likely to interact with fragments from the same chromosome than from a different chromosome. In contrast, completely random ligations (from non-crosslinked material) result in mostly trans interactions (Kalhor et al, 2011). Cis-to-trans interaction ratios in libraries from all four species were comparable to published high-quality Hi-C datasets (Sofueva et al., 2013). The dog had a larger fraction of trans- interactions (~30%) compared to the others, possibly due to its higher number of chromosomes (n = 39) (**FIGURE 7B**).

A matrix of "observed counts" was defined as all detected pairwise interactions from the Hi-C library and was normalised to the "expected" numbers of reads based on a model accounting for technical biases in the experiment. The technical biases that such model takes into account include (1) the length of the interacting restriction fragments (which might influence the likelihood of two fragments to contact one another), (2) the fraction of GC at the fragment-ends (which might bias the PCR amplification) and (3) the mappability of the genomic region (which might bias the alignment step) (Yaffe and Tanay 2011) (**FIGURE 7C**). Consistent with other Hi-C studies (Sexton et al., 2012; Sofueva et al, 2013), in all libraries the number of cis-interactions decayed with increasing linear distance (**FIGURE 7D**). The log-ratios of observed and expected reads for different genomic window sizes ('resolutions') were plotted as the resultant Hi-C heatmap (see MATERIALS AND METHODS for details).

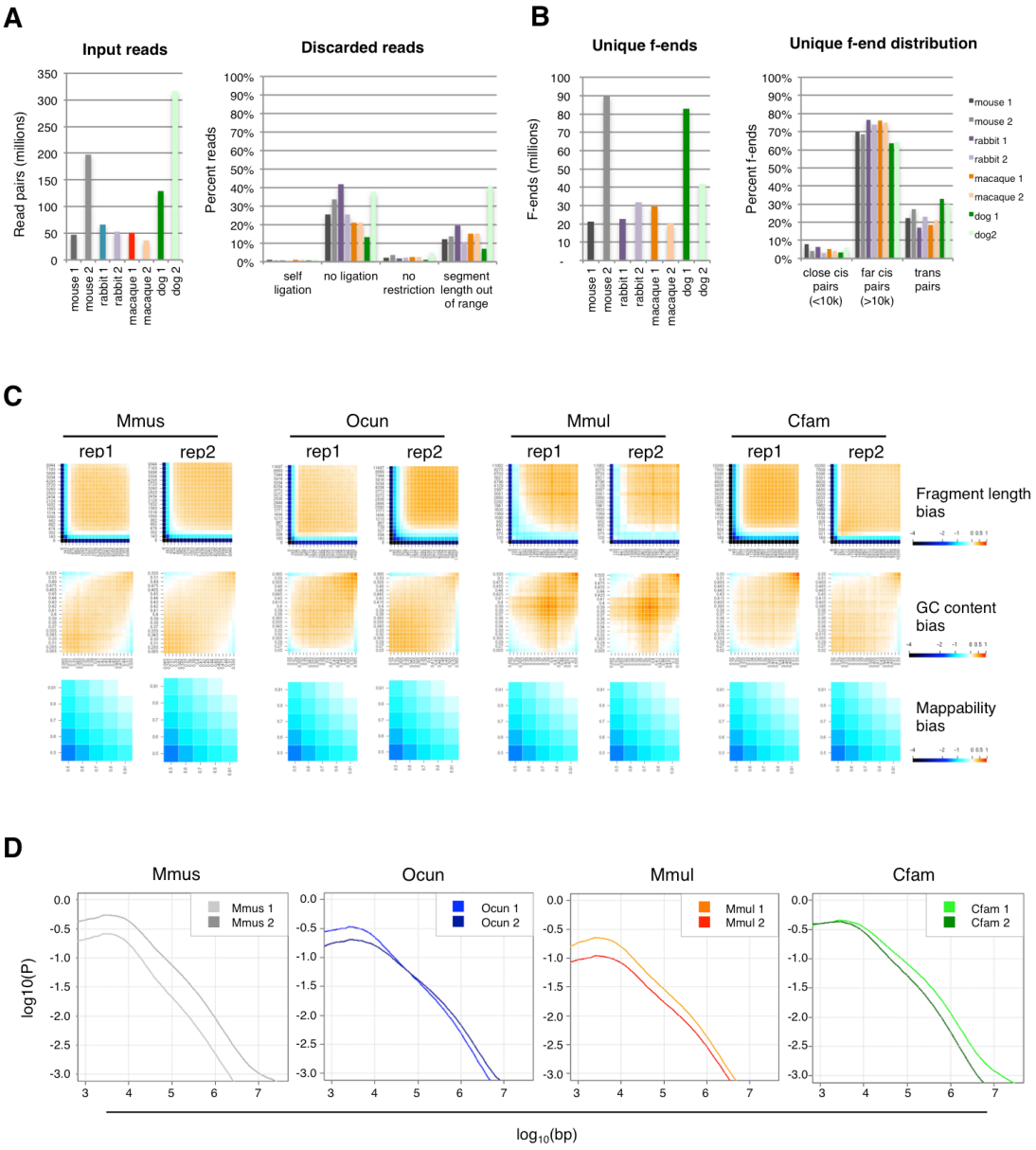


FIGURE 7 – Hi-C analysis pipeline stats. Hi-C reads were processed as previously described (Yaffe and Tanay, 2011). (**A**) Shown are the total number of read pairs inputted into the Hi-C analysis pipeline (left plot) and percentages of reads filtered out from further analysis (right plot). (**B**) Shown in the left plot is the number of unique fragment end (f-end) pairs that have been identified from the valid input read pairs and used to construct the Hi-C maps. The overall quality of all libraries is high, as indicated by the good ratio between intra- (cis) and inter- (trans) chromosomal interactions identified (right plot). (**C**) Correction matrices used to account for experimental biases introduced by the Hi-C protocol as previously described in Yaffe and Tanay (2011). (**D**) Decay of the number of interactions with linear genomic distance.

To quantify the correlation between replicate libraries, the maps were divided into sections or "bands', each one containing interactions in a specific distance range (e.g. 80-160 kb band). The number of contacts across the genome within a certain band was calculated, so as to generate a one-dimensional track (named "crossover track") (Sofueva et al. 2013) (**FIGURE 8A**). This approach allowed for a comprehensive description of domain structure

at multiple scales. Comparison of replicate libraries using this methodology revealed a strong correlation between them (**FIGURE 8B**). As a control, randomisation of the pairing between the bins of the two replicates led to a complete loss of correlation (**FIGURE 8C**)

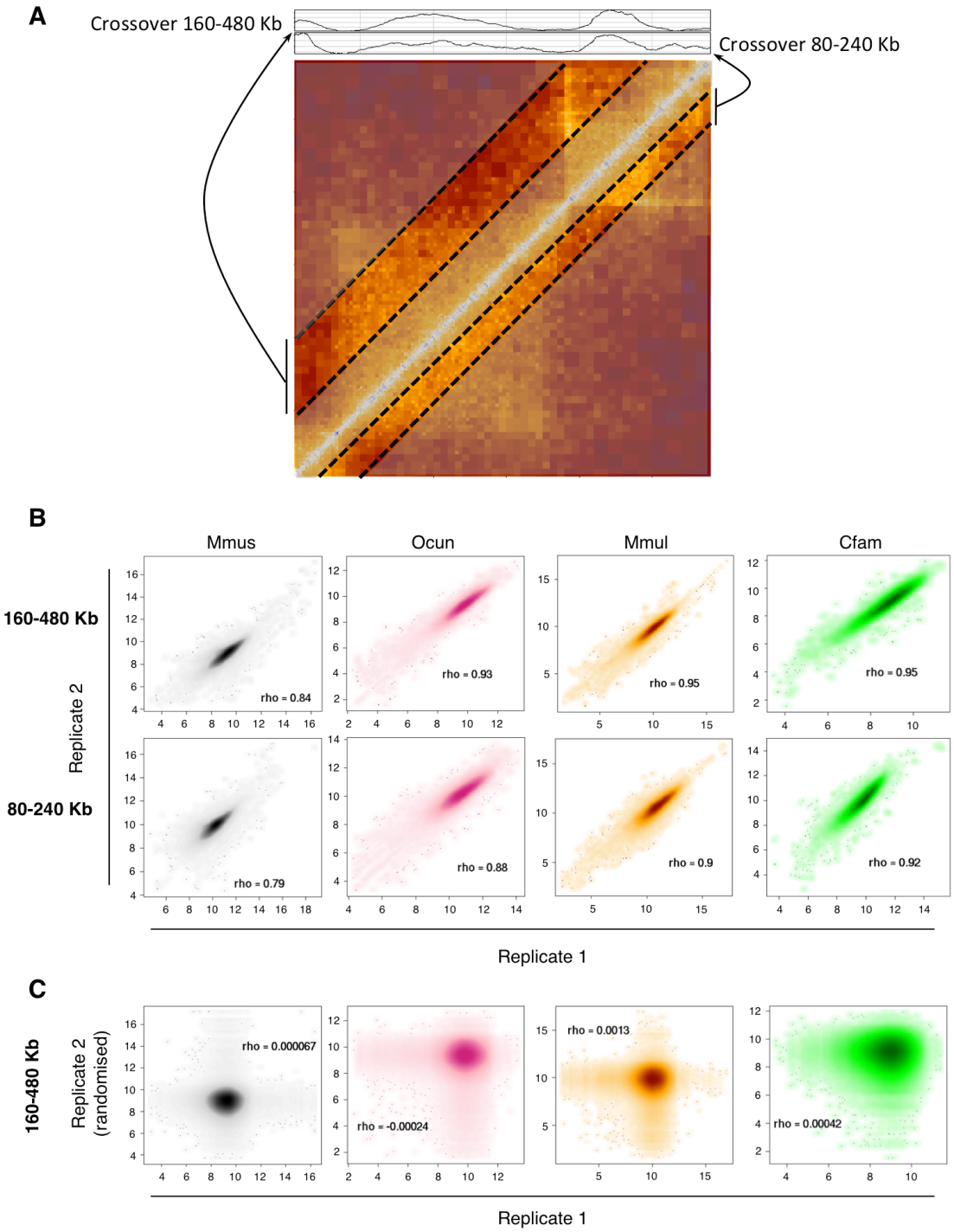


FIGURE 8 – Correlation of replicate Hi-C libraries. (**A**) Schematic of the quantification of contacts ("crossover") for fragments interacting at multiple distance ranges ("bands"). The two bands are drawn on different sides of the diagonal for clarity purposes. (**B**) Crossover contacts within two bands, 160-480 Kb (upper panels) and 80-240 Kb (lower panels), are compared for the replicate libraries in each species. Spearman correlation values are shown. Axes units are contact enrichments [log2(observed/expected)]. (**C**) A comparison between replicates at the 160-480 kb band after randomisation of the bin pairing provides a negative control for correlation.

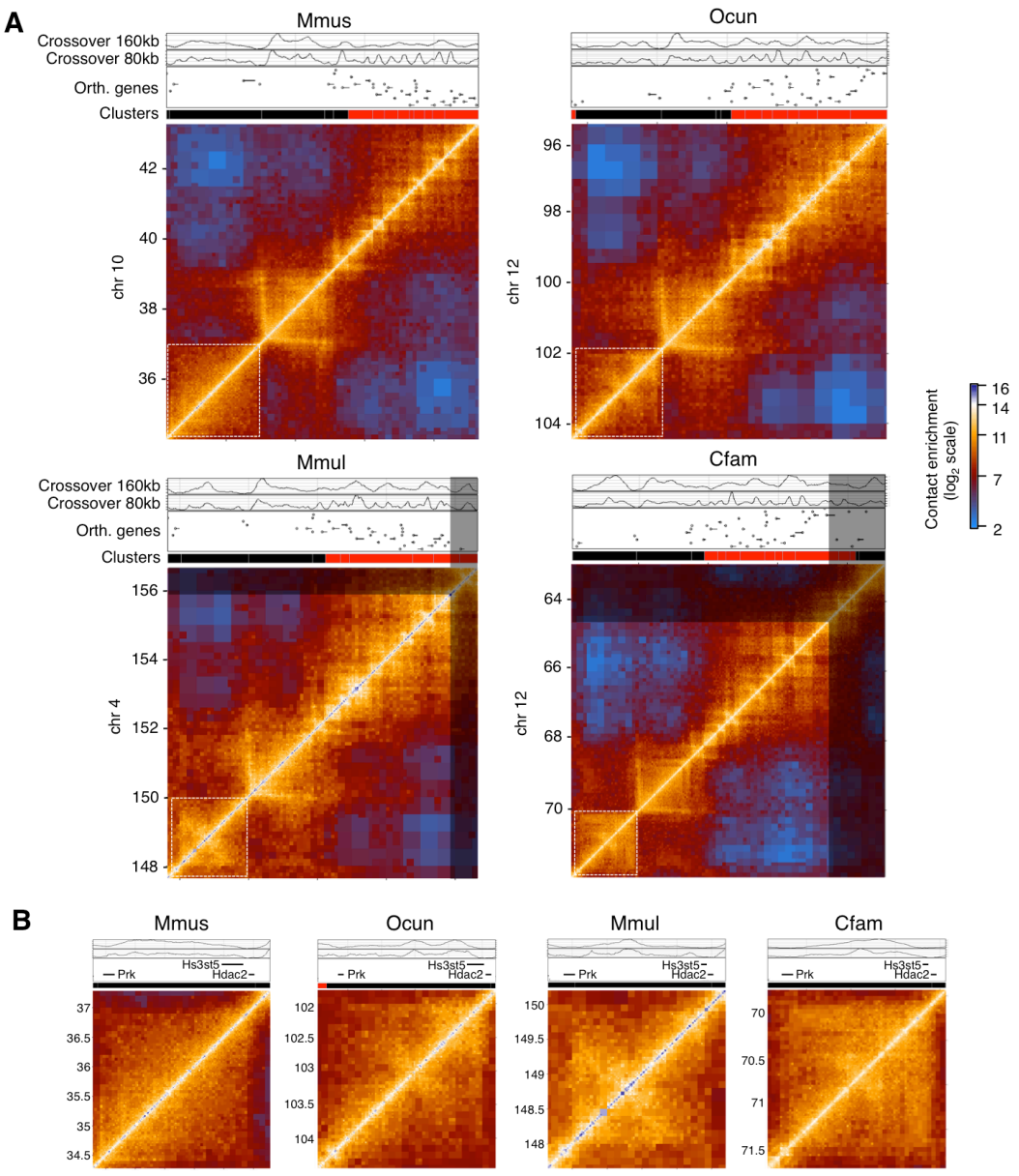


FIGURE 9 – Chromosomal domain structure is robustly conserved in mammals. (**A**) Representative Hi-C contact maps of a 9 Mb syntenic region from mouse (Mmus), rabbit (Ocun), macaque (Mmul) and dog (Cfam). Maps are colored according to technically corrected contact enrichment. For scale purposes, the same size region is shown for all species, but in the macaque and dog genome the syntenic region is reduced in size. The portion of the map that is outside of the synteny boundary is shaded out. Shown above each map is the quantification of contacts for distance bands of 160-480 Kb (Crossover 160 Kb) and of 80-240 Kb (Crossover 80 Kb), as well as a track of orthologous genes, highlighting their conserved distribution in chromosomal domains. (**B**) A zoom-in of the domain highlighted with a white dashed box in (**A**) reveals differences in its internal organization across species.

Evaluation of the overall topological structure within all species first indicated the integrity of their reference genome structures. A comparison of the maps in syntenic regions showed that the chromosome topologies in macaque, dog and rabbit are characterised by a chromosomal domain structure that is remarkably similar to the one inferred before for human and mouse (Dixon et al., 2012). For example, comparison of a 9 Mb syntenic region highlighted the extensive conservation of chromosomal structure between species (**FIGURE**

9A). The maps also suggested a higher degree of intra-domain divergence between species: for example the orthologous domain highlighted in **FIGURE 9B** appears to have a more homogeneous interaction pattern in mouse than in the other species.

The crossover tracks generated in rabbit, macaque and dog were converted from their genome coordinates to mouse. This way, for a certain band, each point in the mouse genome had the number of contacts in mouse, but also the number of contacts in the other

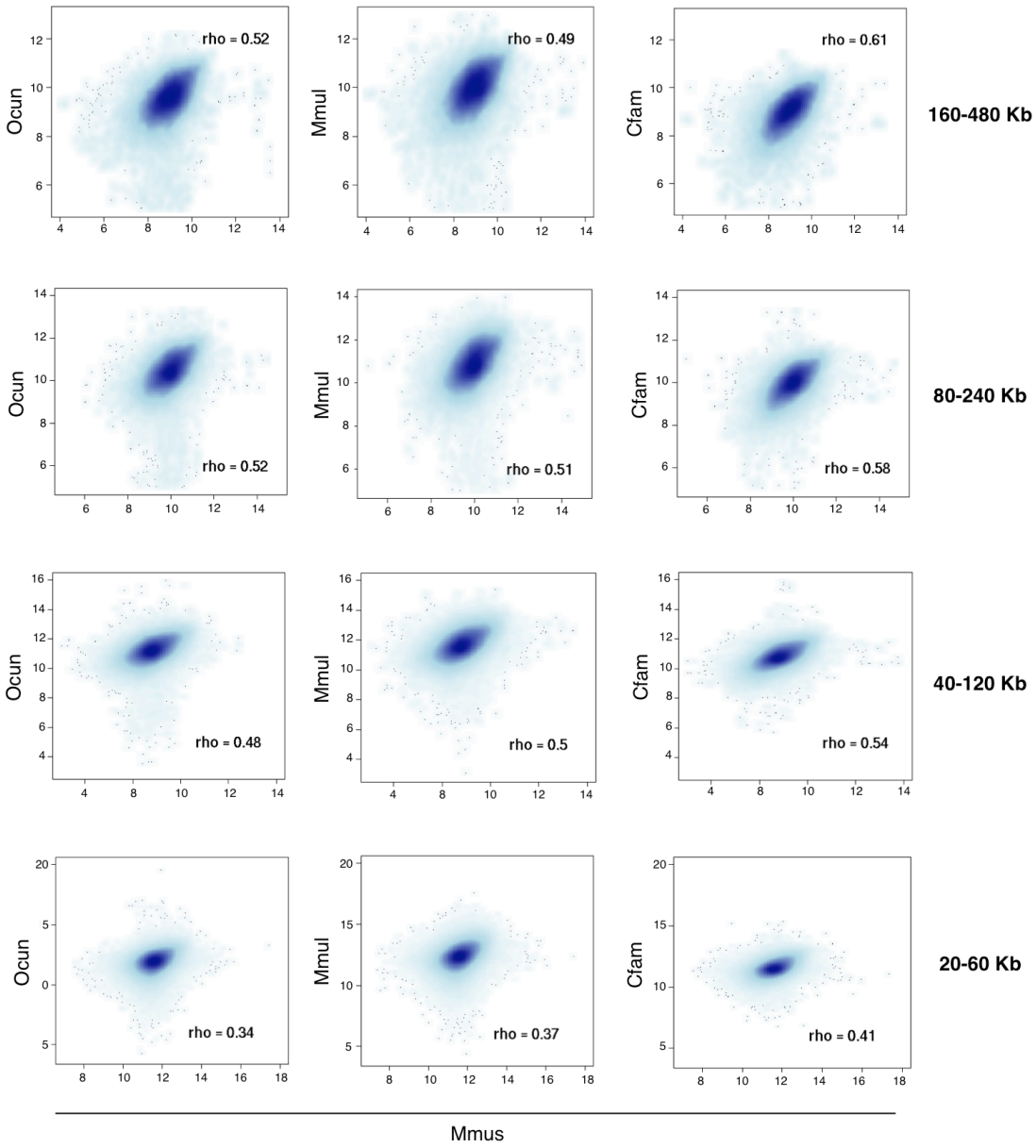


FIGURE 10 – Chromosomal domain structure is quantitatively correlated across mammals. Global quantification of the correlation between the maps. Genome-wide contact enrichments between elements separated by multiple distance ranges in rabbit, macaque and dog were lifted-over to the mouse genome. Spearman correlation values are shown inside the plots. Axes units are contact enrichments [$\log_2$(observed/expected)].

species. Of note, though, this comparison could only be made for the points in the genome that could be converted, i.e. the loci for which the level of synteny between the genomes was sufficient. For this particular analysis, these loci amount to 39%, 53% and 46% of the genome when converting from rabbit, macaque and dog, respectively. Pairwise comparisons of these tracks revealed extensive genome-wide interspecies correlation of chromosome structure that extended to multiple bands (**FIGURE 10**), although bands for closer-ranging interactions showed a reduced degree of correlation. Together, these data represent an extensive analysis of the evolution of chromosomal topologies within regions that did not go through substantial genome rearrangement, and indicate a strong conservation of chromatin structure over the course of mammalian evolution.

Domains maintain their integrity during chromosomal rearrangements.

Genomic regions that underwent large-scale rearrangements are difficult to analyse because such modifications produce disruption of synteny, making it difficult or impossible to compare the two genomes. In certain instances, such as some inversions or large-scale insertions/deletions, the synteny can be maintained in the sequences flanking the rearrangements, allowing for the tracing of the regions from one genome to another. Moreover, if the synteny markers (e.g. orthologous genes) are sufficiently dense, the edge of the rearrangement (the "breakpoint') can be defined with a good degree of precision.

Mmus coordinates (mm10)	Cfam coordinates (cfam3)	Type	Genes	Notes
chr4:108229724-116000000	chr15:8595758-16366034	insertion	Slc5a9_Trabd2b	Fig. 11
chr7:45789003-51827440	chr21:38170174-44208611	insertion	Ptpn5_Zdhhc13	Fig. 12
chr12:66437857-75437857	chr8:26500000-35500000	insertion	Frmd6_Actr10	Fig. 13
chr6:86659598-92810123	chr20:0-6280377	inversion	Efcc1_Klf15	data not shown
chr1:46755506-54785224	chr37:0-8819902	inversion	Slc39a10_Asnsd1	data not shown
chr1:149000000-157000000	chr7:12000000-21000000	inversion	Pla2g4a_Tor1aip2	data not shown

TABLE 4 – List of the chromosomal rearrangements analysed.

If chromosomal domains act as modular units (for example to regulate gene expression), then large-scale rearrangements would be expected to occur at domain borders, so as to maintain the integrity of these structures. With this in mind, a list of candidate rearranged regions was generated by comparing the distances between contiguous orthologous genes in the mouse and dog genomes. The list was then manually refined by looking at individual candidates on the Hi-C maps. This analysis uncovered a number of complex rearrangements between the mouse and dog genomes involving insertions and inversions

(**TABLE 4**). In each case the breakpoint of the rearrangement appeared to have occurred at the border between two chromosomal domains. This is shown in the Hi-C map from chromosome 15 in dog (**FIGURE 11**). Here we found two domains, one containing the *Slc5a9* gene and the other containing the *Trabd2b* gene (highlighted by red dots). Comparison of this region to the mouse genome revealed that a 2 Mb insertion occurred in the mouse genome that contains the *Skint* gene family, which are rapidly evolving and

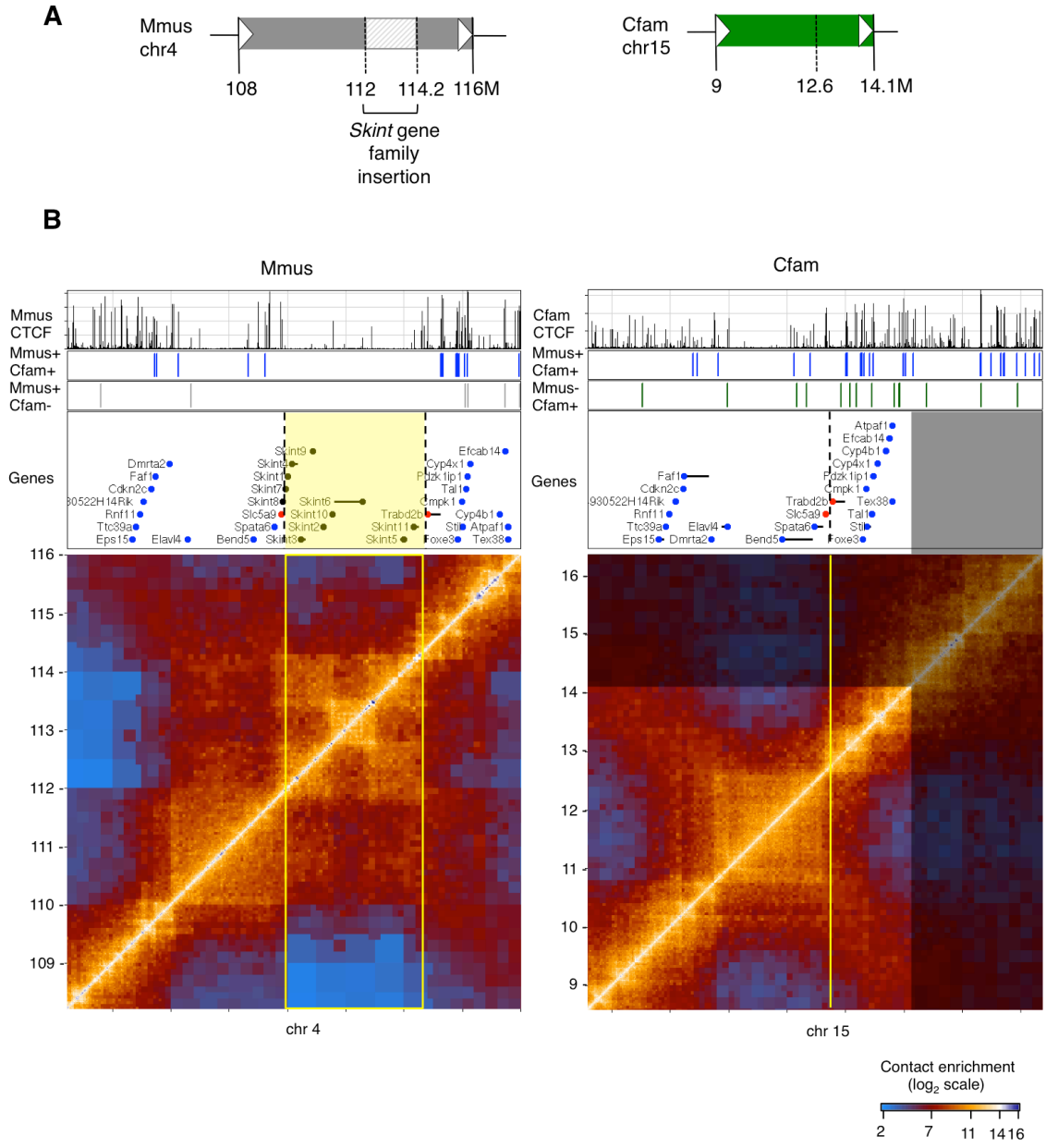


FIGURE 11 – Chromosomal domains evolve as modular units. (**A**) A schematic of the rearrangement between mouse (grey) and dog (green). (**B**) Hi-C maps from a syntenic region in mouse (left) and dog (right). Shown also are CTCF ChIP tracks, conserved and divergent sites and genes in the region. A cluster of non-orthologous Skint genes (black dots) has been inserted in mouse between the orthologous genes Slc5a9 and Trabd2b (highlighted as red dots). The inserted region (bordered in yellow) forms its own nested chromosomal domain structure, probably as a result of gene duplication events. Highlighted in blue are other orthologous genes in the region.

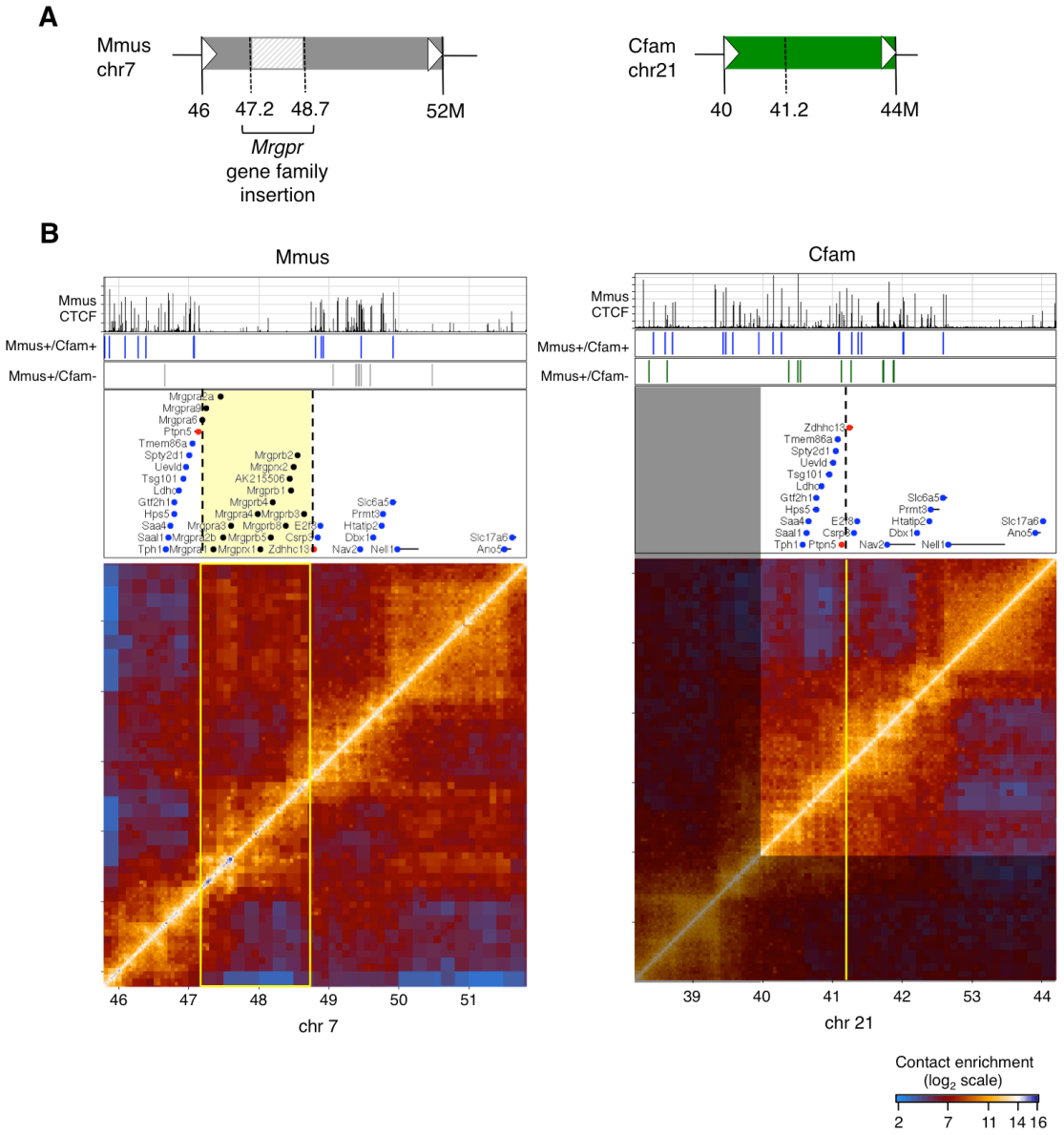


FIGURE 12 – Insertion of a gene cluster. (**A**) Schematic of the rearrangement. (**B**) Hi-C maps from a syntenic region in mouse (left) and dog (right). Shown also are CTCF ChIP tracks, conserved and divergent sites and genes in the region. A cluster of non-orthologous *Mrgpr* genes (black dots) has been inserted in mouse between the orthologous genes *Ptpn5* and *Zdhhc13* (highlighted as red dots). The inserted region (bordered in yellow) forms its own nested chromosomal domain structure, probably as a result of gene duplication events. Highlighted in blue are other orthologous genes in the region.

unique to the mouse lineage (Boyden et al., 2008). Remarkably, the insertion occurred directly between two neighbouring dog domains in such a way as to perfectly maintain their integrity. Also noteworthy is the fact that the inserted domain seems to be able to interact to an extent with the neighbouring domains, but does not prevent the *Slc5a9* domain and the *Trabd2b* domain from interacting with one another as they do in the dog genome, where they are positioned next to each other. A similar rearrangement event occurred at the *Mrgpr* gene cluster in the mouse genome (Dong et al., 2001), again maintaining the structure of the neighbouring domains (**FIGURE 12**). In this case, the inserted domain

seems to preferentially interact with a domain about 1 Mb downstream (marked by the *Nell1* gene), while "ignoring" its immediate neighbours. When a syntenic region was inserted or inverted, its internal global domain structure remained broadly unaltered,

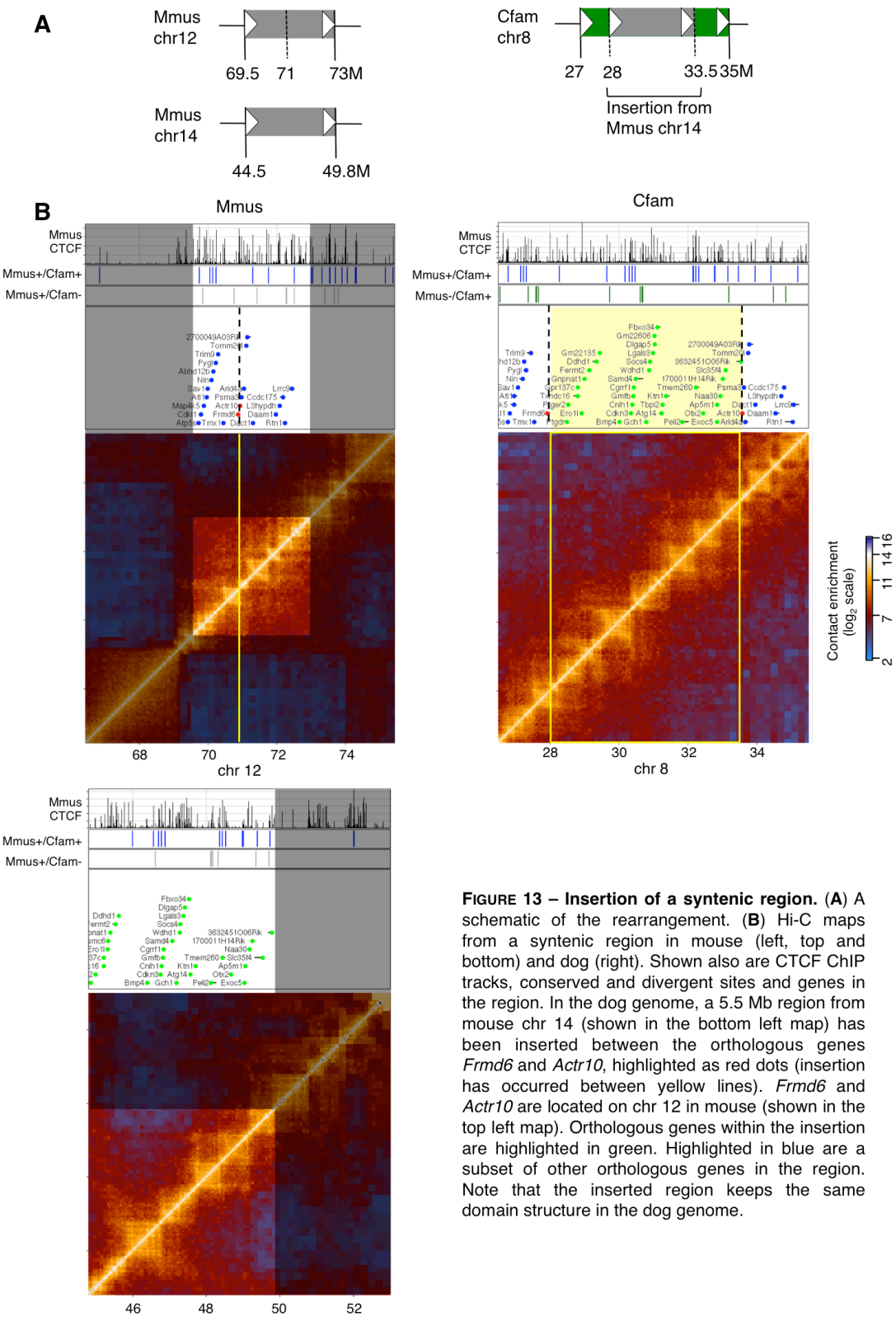


FIGURE 13 – Insertion of a syntenic region. (**A**) A schematic of the rearrangement. (**B**) Hi-C maps from a syntenic region in mouse (left, top and bottom) and dog (right). Shown also are CTCF ChIP tracks, conserved and divergent sites and genes in the region. In the dog genome, a 5.5 Mb region from mouse chr 14 (shown in the bottom left map) has been inserted between the orthologous genes *Frmd6* and *Actr10*, highlighted as red dots (insertion has occurred between yellow lines). *Frmd6* and *Actr10* are located on chr 12 in mouse (shown in the top left map). Orthologous genes within the insertion are highlighted in green. Highlighted in blue are a subset of other orthologous genes in the region. Note that the inserted region keeps the same domain structure in the dog genome.

irrespective of the new neighbouring context. This is evident in a large-scale 5.5 Mb insertion/deletion containing multiple domains: again, the domains on either side of the rearrangement have been maintained intact (**FIGURE 13**), and the inserted region shows a structure that is remarkably conserved when compared to its syntenic counterpart in the mouse genome.

Albeit sparse in number, these examples suggest that domains may function as modular units that are selected against breakage during genome rearrangements and that, when rearranged, can adapt to new surrounding environments. Further support for this hypothesis should come from a more systematic analysis of rearranged regions.

2.3 DISCUSSION

The comparison of syntenic regions across four mammalian species revealed a striking preservation of chromosomal architecture at multiple levels, which has been quantitatively analysed in this thesis.

As evidenced in **FIGURE 9**, this structural preservation is even resilient to resizing of the genome. Despite being organised on a large number of chromosomes, the dog genome is markedly smaller than those of the other three species, with around 2.4 billion nucleotides against the roughly 3 billion of mouse, rabbit and macaque. Such size difference is reflected in the syntenic region showed in **FIGURE 9A**, but the architecture is not altered with respect to the other genomes, exhibiting a kind of proportional "shrinking" of the domains, with orthologous genes still mapping to orthologous domains. This resilience points to a strong relevance of chromosomal architecture for the correct function of the genome.

Preliminary evidence for the conservation of chromosomal structure between the human and mouse genome has been provided in Dixon et al. (2012). There, the authors defined a statistic (termed directionality index) and identified domain borders using a Hidden Markov model based on it. To assess conservation they converted their mouse calls to the human genome and vice versa and measured the overlap of the borders, which ranged between 54 and 76% depending on the direction of the coordinate conversion used (human-to-mouse or mouse-to-human, respectively). While this approach can accurately identify conserved domain borders and provides a valid preliminary assessment of structural conservation, it is focused only on domain borders and inevitably relies on a

parameter threshold to define them. In the present study, the comparison of the Hi-C datasets from four mammals was done avoiding the use of thresholds for domain or border definition and rather comparing contacts across whole sections of the maps at multiple scales, thus painting a full picture of structural conservation between mammalian species.

The degree of correlation observed is also partially dependent on technical factors, namely the quality of the coordinate conversion and the resolution of the maps. A poor conversion will generate mis-paired regions that will not have a similar number of contacts because they are not syntenic, rather than evolutionarily divergent. The reliability of the conversion will in turn be dependent on the quality of the genome assemblies involved: among the species analysed here, mouse has the best assembled genome, followed in order by dog, rabbit and macaque, with the latter two being draft assemblies sequenced with a depth of 7.48X and 5.1X respectively (http://genome.ucsc.edu). For this reason, the correlation values in **FIGURE 10** should not be directly compared across species pairs.

The resolution of the maps, on the other hand, should be taken into consideration when examining cross-species map correlations at different distance bands. As evident from **FIGURE 10**, the correlation seems to decrease as the interactions become shorter range. Hi-C datasets will have a greater level of statistical noise when examining shorter-range interactions, because each point in a finer-resolution map (or in a narrower band) is built with less data than a larger block in a lower-resolution map (or in a wider band). This technical noise can contribute to the loss of correlation observed in **FIGURE 10**, but it is not possible to estimate the extent of such a contribution, which is likely mixed with a real biological effect. In this respect, it is important to note that a reduction of map correlation for intra-domain interactions is consistent with qualitative observations from the maps (**FIGURE 9B**) and other results (CHAPTER FOUR).

Although the aforementioned technical caveats should be indeed taken into account when interpreting the data, it is also interesting to point out how both of them would lead to an underestimation of the correlation in chromosome structure across the mammalian lineage, further stressing its remarkable conservation and functional importance.

CHAPTER THREE:

CTCF BINDING EVOLVES ACCORDING TO TWO DIFFERENT REGIMES

3.1 INTRODUCTION

The evolution of the binding sites for CTCF has been analysed across multiple mammalian lineages (Schmidt et al., 2012). Chromatin immunoprecipitation (ChIP-seq) was performed in hepatocytes of five different mammals and subsets of deeply shared, lineage-specific or species-specific CTCF binding sites were identified. This analysis suggests that CTCF binding sites can be embedded in retrotransposable elements and have been evolving by a copy-and-paste mechanism along the mammalian clade. In addition to the rapidly evolving CTCF binding sites, a highly conserved subset was also identified, however the functional relevance of these sites remains unknown.

Given the importance of CTCF in insulator function as well as its role in chromatin loop formation, I hypothesise that deeply conserved CTCF sites may have a key role in the organisation of chromosome structures which are themselves conserved across species.

To explore this, I re-analysed published CTCF binding data for mouse, dog and macaque (from Schmidt et al., 2012), aiming to generate and characterise high-confidence groups of conserved and divergent CTCF binding sites which could then be used in downstream comparisons with the Hi-C structures generated from the same species (see CHAPTER FOUR).

3.2 RESULTS

Multi-species comparison of CTCF ChIP-seq identifies conserved and divergent binding events.

Mouse, dog and macaque CTCF ChIP-seq profiles from primary liver cells (data from Schmidt et al, 2012) were analysed to investigate how conservation and divergence of the insulator binding landscape co-evolved with chromosomal topology. Analysis of replicate library correlation allowed the selection of peak calling thresholds (see also MATERIALS & METHODS for more details) (**FIGURE 14A**). As before, mouse was used as a reference species. ChIP-seq tracks from dog and macaque were converted to mouse genome coordinates and pairwise CTCF ChIP-seq comparisons against the mouse genome identified conserved or divergent CTCF binding sites within syntenic chromosomal regions (**FIGURE 14B-C**). These sites were heavily filtered to account for possible uncertainties

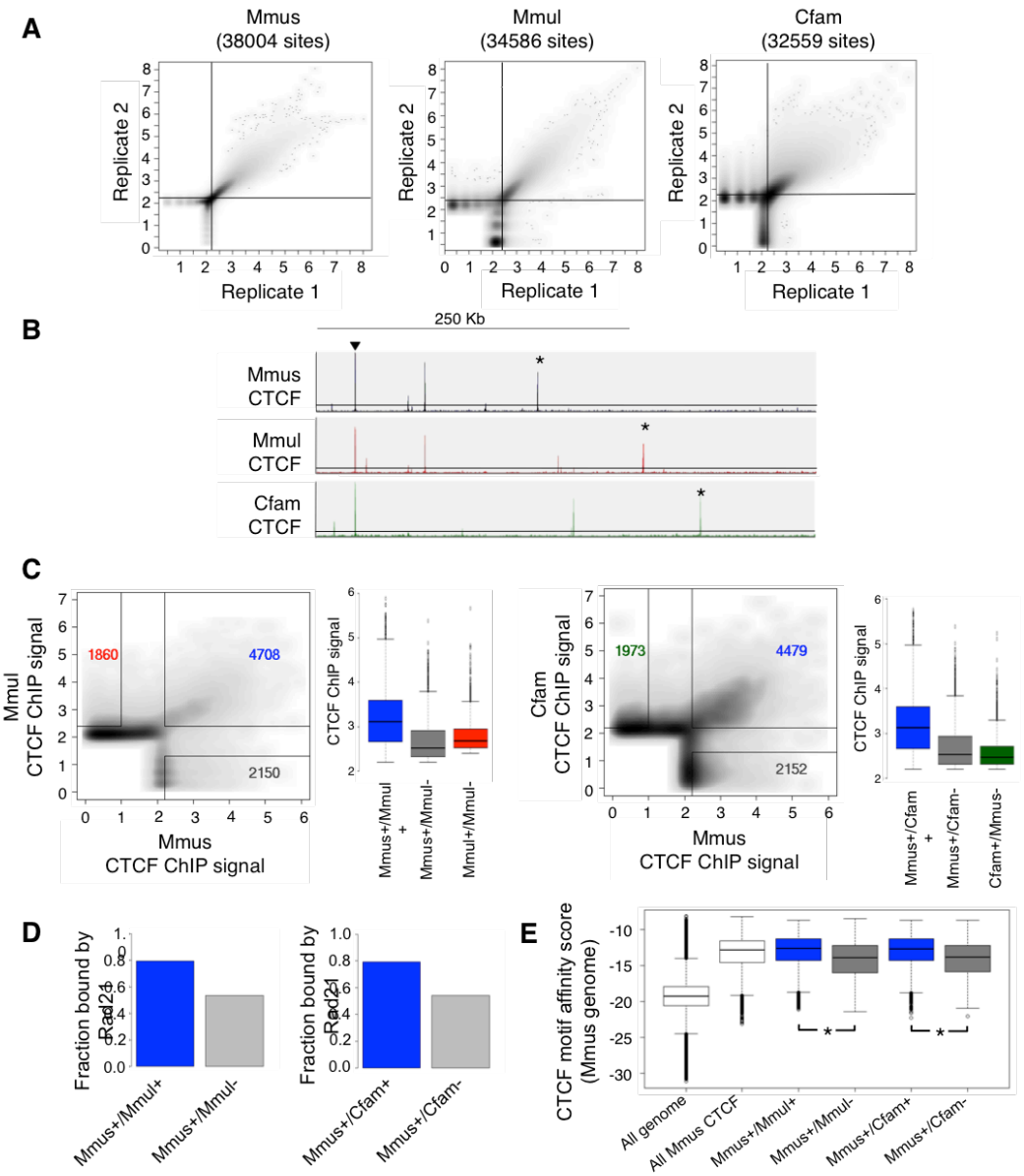


FIGURE 14 – Comparative ChIP-seq identifies groups of conserved and divergent CTCF binding sites. (**A**) Percentile-transformed CTCF ChIP-seq replicates from mouse (Mmus), macaque (Mmul) and dog (Cfam) were used to select thresholds for peak calling. (**B**) Representative CTCF ChIP-seq tracks on mouse (Mmus) chromosome 5; macaque (Mmul) and dog (Cfam) tracks are shown after liftOver to the mouse genome (mm10). Thresholds used are indicated as black lines. Sites conserved across all three species (arrowhead) or specific to a single species (asterisks) are indicated as examples. (**C**) Pairwise comparison of mouse CTCF ChIP and macaque (left panel) or dog (right panel) CTCF ChIP; the non-reference species' tracks were lifted-over, identifying conserved or divergent binding sites (see Methods). Axes are $-\log_{10}$(1-percentile) of the CTCF ChIP-seq read coverage. The distribution of signal intensities is also shown for the conservation groups identified. For divergent sites, intensity values refer to the genome where CTCF is bound. (**D**) Fraction of conserved (blue) and divergent (grey) CTCF found co-localising with cohesin subunit Rad21. Shown are comparisons of mouse against macaque (left plot) and against dog (right plot) (**E**) CTCF motif affinity for conserved binding sites (blue: Mmus+/Mmul+; Mmus+/Cfam+) or mouse-divergent (grey: Mmus+/Mmul-; Mmus+/Cfam-). In both pairwise comparisons, conserved sites have a stronger CTCF motif than mouse-specific sites (Kolmogorov-Smirnov test, $p < 2.2 \times 10^{-16}$). The distributions of motif affinity for all CTCF sites in mouse and for the entire mouse genome are shown as controls.

introduced by the coordinate conversion to the reference genome: sites falling within repeat elements or in regions of low overall synteny were excluded from the comparison. This approach produced reliable, high-confidence groups.

Interestingly, conserved CTCF sites exhibit higher signal intensities than divergent sites (**FIGURE 14C**). Co-localisation with cohesin complex protein Rad21 in mouse (ChIP-seq data from Faure et al., 2012) was observed in both classes, although the overlap was markedly more prominent for conserved CTCF binding sites in both pairwise comparisons (**FIGURE 14D**). The sequences below each peak in the different classes of binding sites were scored for similarity to the canonical CTCF consensus motif (see MATERIALS & METHODS for more details). This analysis revealed that the levels of motif affinity for conserved sites were overall higher than the level for the mouse-specific sites (**FIGURE 14E**).

CTCF binding evolution is correlated to changes in the underlying sequence.

To understand the relationship between sequence affinity and CTCF binding at conserved or divergent sites, changes in CTCF ChIP-seq signal were correlated with changes in CTCF sequence motif affinity between species. Remarkably, a direct association between sequence divergence and CTCF binding divergence was observed. Conserved CTCF binding sites showed overall higher motif affinities and a high degree of affinity conservation; conversely, motifs underlying the divergent sites were evolutionarily dynamic, and diverged in strong correlation to divergent binding intensity (**FIGURE 15A-B**). A few sites exhibited a strong difference in ChIP-seq signal, but a small difference in motif affinity; these sites were found to have a low affinity to the CTCF consensus even in the species where they are bound (**FIGURE 15C-D**), possibly indicating that the sequence is not the main driver of binding at those loci. The data show that in most cases when strong motifs in CTCF binding sites diverge, CTCF binding itself is concomitantly gained or lost.
Together, these results suggest that the CTCF insulator landscape is evolving under two regimes: the first involves a tight conservation of both sequence and binding landscape, and the second showing a dynamic interplay between divergence of specific DNA *cis*-elements and consequential evolution of the CTCF binding trait. The relatively direct influence of motif divergence on CTCF binding forms a potential link between sequence evolution and large-scale genome evolution.

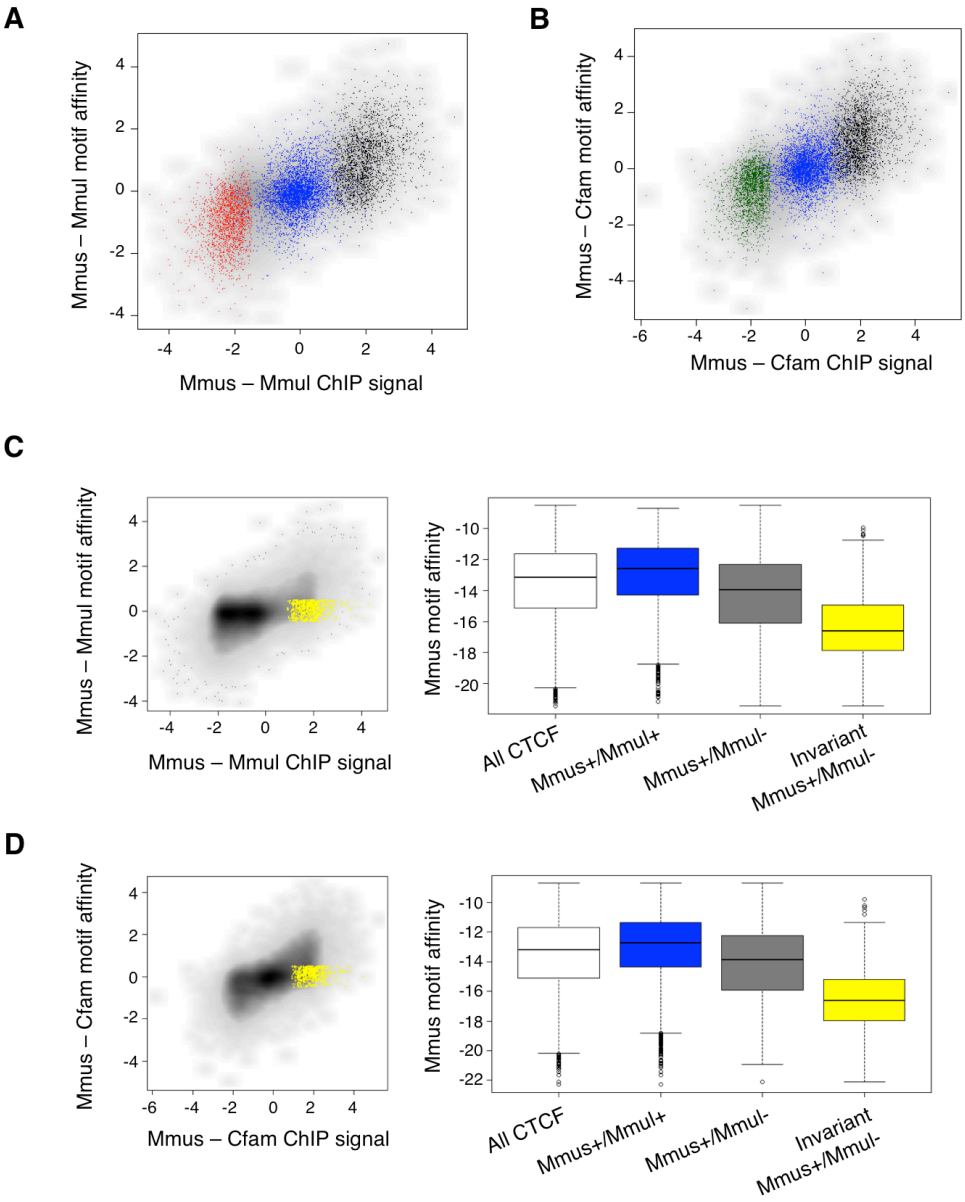


FIGURE 15 – Evolution of CTCF binding profiles in mammals. (**A**) Comparison of the interspecies difference in CTCF ChIP-seq signal against the difference in CTCF motif affinity in mouse vs. macaque. Scatterplots are highlighted for conserved (Mmus+/Mmul+, blue), mouse-divergent (Mmus+/Mmul-, black) and macaque-divergent sites (Mmus-/Mmul+, red). (**B**) Comparison of the interspecies difference in CTCF ChIP signal against the difference in CTCF motif affinity in mouse vs dog. Scatterplots are highlighted for conserved (Mmus+/Cfam+, blue), mouse-divergent (Mmus+/Cfam-, black) and dog-divergent sites (Mmus-/Cfam+, green) (**C**) Divergent sites which do not show changes in motif energy between macaque and mouse (yellow highlight) have a low affinity CTCF consensus sequence. (**D**) Same analysis as in (**C**), but for the mouse/dog comparison.

3.3 DISCUSSION

A stringent analysis of CTCF ChIP-seq from three species has allowed for the identification of two groups of CTCF binding sites with markedly different characteristics. A first group is composed of conserved CTCF sites (identified from pairwise comparisons) that are characterised by very strong signal intensities. This observation is in agreement with what was reported before for ultra-conserved CTCF binding sites (i.e. sites that are conserved among five mammalian species, Schmidt et al., 2012). Ultra-conserved sites display a striking binding stability and are even resilient to knock-down of CTCF (Schmidt et al., 2012). Although the stringent pairwise analysis presented in this thesis cannot be directly compared with the multi-species analysis in Schmidt et al., it is worth noting that 65% of the sites conserved between mouse and dog are also conserved in macaque and that indeed over 50% of the pairwise conserved sites identified here overlap with their ultra-conserved sites, making the two independent analyses consistent with each other and strengthening the reliability of the groups identified. The high binding intensity of the conserved sites is accompanied by high similarity to the CTCF core consensus motif across all genomes. This suggests a remarkable stability both in the binding of the protein at these sites and in the sequence underlying them.

A second group is composed of divergent CTCF binding sites (identified from pairwise comparisons) that exhibit weaker, likely more dynamic binding, accompanied by a lower average similarity to the CTCF core consensus motif. The divergence in the binding at these sites across species is remarkably mirrored by a divergence in the sequence that underlies them, confirming the importance of the sequence for the binding of CTCF: in other words when a site is no longer bound, the sequence underneath it has lost its similarity to the CTCF consensus motif. This likely means that weaker divergent CTCF binding sites, in sharp contrast to the conserved ones, represent a pool of plastic units that can be dynamically modified over the course of evolution: small deletions can eliminate sites, while duplication or transposition may cause their spread to different genomic locations; moreover, point mutations in the CTCF motif at these sites may directly increase or decrease their affinity for the protein and possibly modulate their functionality.

CHAPTER FOUR:

CTCF UNDERLIES HI-C DOMAIN STRUCTURE EVOLUTION IN THE MAMMALIAN GENOME

4.1 INTRODUCTION

CTCF binding is correlated with the topological architecture of mammalian chromosomes (Dixon et al., 2012; Phillips-Cremins et al., 2013; Sofueva et al., 2013; Zuin et al., 2013), and participates in long-range chromatin loops (Kurukuti et al, 2006; Splinter et al, 2006; Nativio et al., 2009; Hadjur et al., 2009), thereby underlying global contact insulation.
Considering the striking conservation of domain boundaries and that no clear functionality has been attributed to conserved CTCF binding sites, I hypothesise that the conservation of a CTCF binding site correlates with its presence at domain boundaries.

Furthermore, I reasoned that evolutionary studies could provide a platform for analysing the causal connection between CTCF and chromosome structure.
The involvement of CTCF and cohesin in the looping of the DNA has prompted researchers to ask what would happen when one of these proteins is no longer present. Studies where either protein was targeted for depletion have reported pervasive effects on chromatin organisation and gene regulation (Sofueva et al., 2013; Zuin et al, 2013), which render the dissection of the mechanisms of their action difficult. On the other hand, editing of individual candidate sites in candidate genomic settings has indeed shown an influence of CTCF on chromatin structure and transcription (Dowen et al., 2014; Narendra et al., 2015), but this approach is limited in scope due to the low-throughput nature of genome editing.
In this context, a high-confidence set of divergent CTCF binding sites can be viewed as an array of thousands of "naturally-occurring genome editing events" that can be characterised both at the sequence and chromosome topology levels, thus circumventing the problems associated with genetic perturbation experiments.

4.2 RESULTS

Conserved and divergent CTCF binding sites are differentially distributed within domains.

To investigate how the different classes of CTCF binding conservation correlate with the evolving chromosomal structure, the peak calls for conserved and divergent CTCF sites were visualised alongside the mouse, macaque and dog Hi-C contact maps. Initial visual

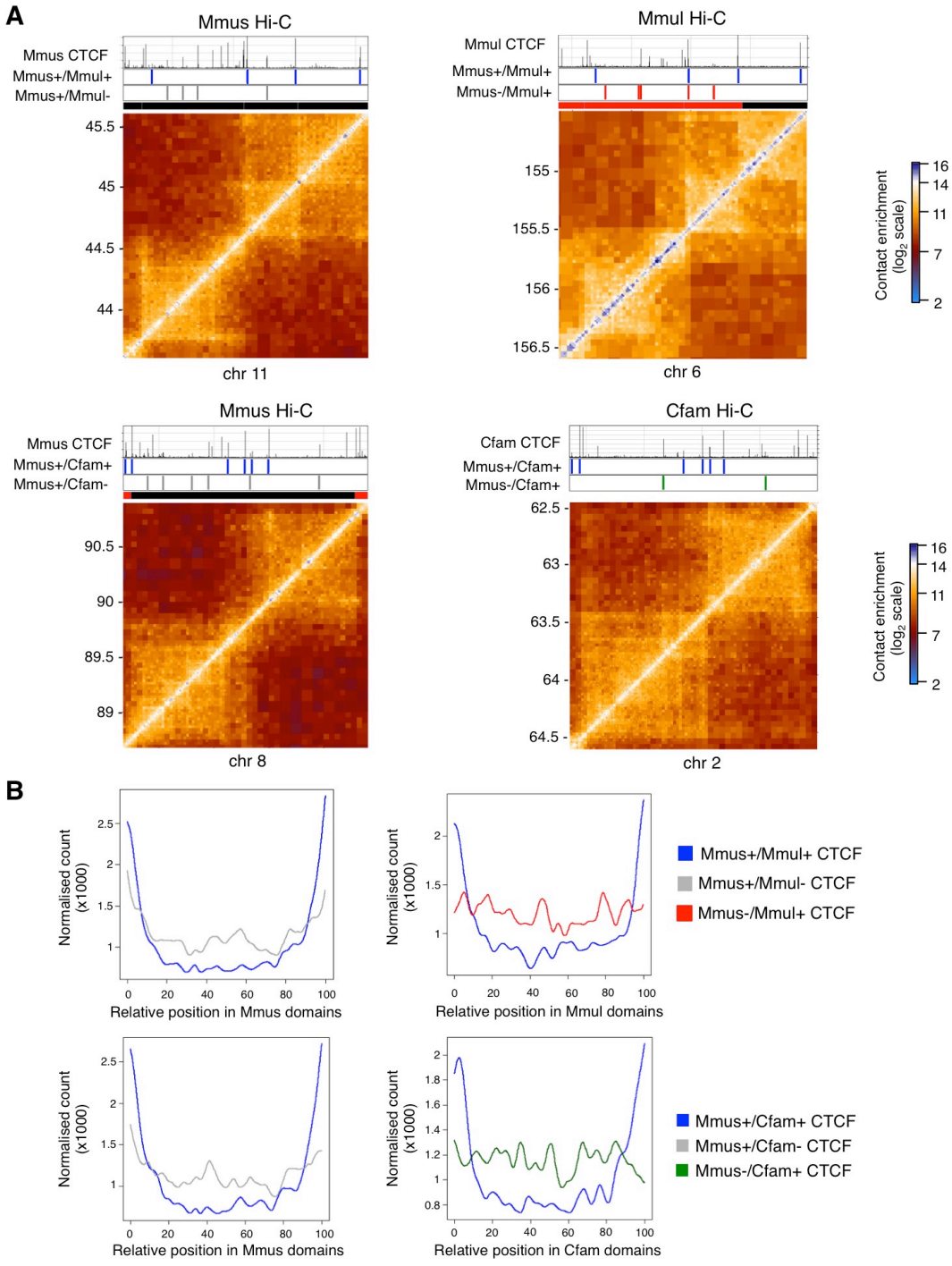


FIGURE 16 – CTCF conservation correlates with Hi-C structure. (**A**) Hi-C contact maps for a 2 Mb syntenic region in mouse and macaque (upper panels) and in mouse and dog (lower panels). Shown are the CTCF ChIP-seq tracks for each species as well as the conserved (blue) and divergent (grey for mouse; red for macaque, green for dog) CTCF sites. (**B**) Genome-wide relative position of conserved and divergent CTCF sites within domains. Data are shown separately for the mouse/macaque comparison (upper panels) and the mouse/dog comparison (lower panels). Counts are normalised by the total number of sites in each group.

inspection of the maps suggested that conserved CTCF binding sites are predominantly found at the borders of large-scale Hi-C domains while divergent CTCF sites are located internal to domains (**FIGURE 16A**). To assess if such pattern was observed genome-wide, the relative position of conserved and divergent CTCF sites within the domain they belonged to was determined across the genomes of all three species. As evident from **FIGURE 16B** the edges of the domains presented a striking enrichment of conserved sites, whereas the divergent sites appeared more evenly distributed across the domains. This result was observed in all species and in both pairwise comparisons. Notably, the borders occupied by the conserved CTCF sites shown in the examples in **FIGURE 16A** are themselves conserved across species. The fact that the CTCF sites are all located into broadly syntenic regions indicates that they are indeed marking conserved borders genome-wide.

CTCF binding conservation groups show differing insulation properties.

To further characterise chromosomal contacts around conserved and divergent CTCF sites, the average contact distribution around these sites was analysed globally, measuring "contact insulation" by quantifying the decrease in contact probability between multiple elements separated by a CTCF site (Sofueva et al., 2013). In this analysis, the number of interactions at a certain band are counted around a set of genomic intervals (e.g. conserved or divergent CTCF binding sites) and normalised to the number of contacts in the flanking regions, allowing for identification of local minimums that would represent insulating points (**FIGURE 17A**). The profiles thus obtained for each site are then averaged together, providing statistical robustness. Such an analysis can be done at multiple independent bands and then represented together in one multi-band contact insulation analysis. Analysis of contact insulation from mouse/macaque and mouse/dog conserved CTCF sites showed strong insulation profiles at all distance ranges, further supporting the idea that these conserved, high ChIP signal CTCF sites were co-occurring with large-scale domain borders (**FIGURE 17B**, left panels). In comparison, the lower-signal divergent sites showed a significantly weaker, more localised insulation profile, consistent with the enrichment of divergent sites within domains (**FIGURE 17B**, right panels). The same results were observed across all species. The strong levels of insulation observed at conserved CTCF sites are consistent with the general conservation of large-scale domain structure described in CHAPTER TWO. This contact insulation analysis also correlates well with the distribution of conserved and divergent CTCF with respect to chromosomal domains shown in **FIGURE**

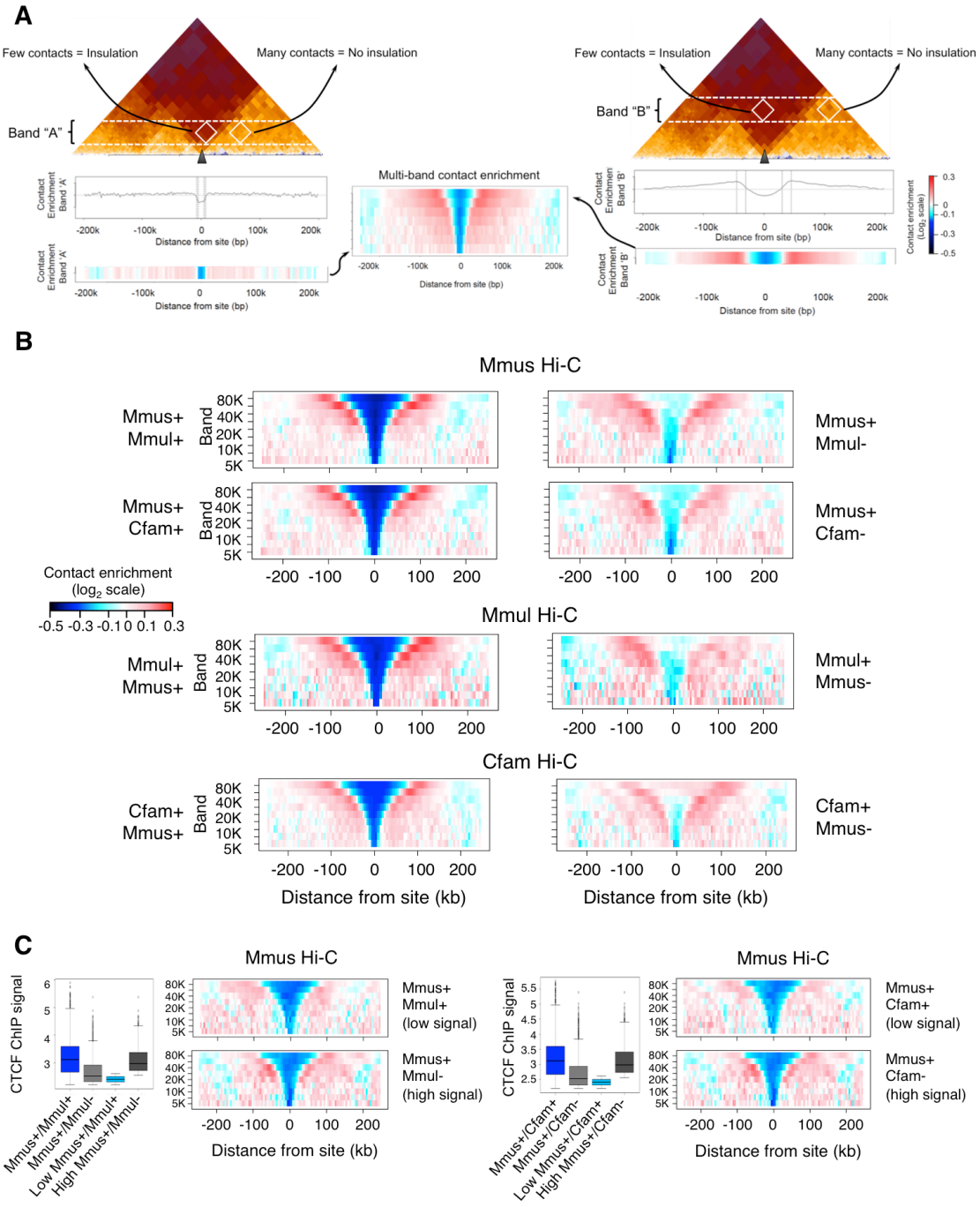


FIGURE 17 – CTCF conservation is correlated to its ability to mediate contact insulation. (**A**) A schematic of how contact insulation plots are generated. Note that for this illustration, the Hi-C maps have been turned by 45° with the diagonal now lying on the horizontal axis and representing the linear sequence of the DNA. (**B**) Average contact insulation analysis at conserved (left panels) and divergent (right panels) CTCF sites in mouse, macaque and dog Hi-C datasets. (**C**) Average contact insulation analysis at low signal conserved CTCF sites and high signal divergent sites in mouse. The left panels refer to the comparison against macaque, the right panels to the comparison against dog. Boxplots of the signal distribution of the low- and high- signal subsets are also shown compared to the signal distribution of all conserved and divergent sites. Boxplot axes are $-\log_{10}$(1-quantile) of the CTCF ChIP-seq read coverage.

16B, with the strongly insulating conserved sites mostly found at domain borders and the locally insulating divergent sites more evenly distributed across domains.

The conservation of CTCF sites is positively correlated to their ChIP signal intensity (**FIGURE 14C**). For this reason, it is possible that the different levels of contact insulation observed at conserved versus divergent CTCF sites are simply a reflection of their different ChIP signals. Dissecting the contribution of conservation or signal intensity to contact insulation is difficult, because these two variables are likely linked (see CHAPTER FIVE). To try to separate these two variables, I compared contact insulation at a subset of low-signal conserved CTCF sites with a subset of high signal divergent CTCF sites. In these two subsets, the ChIP signal distributions are reversed, with conserved sites having weaker signals and divergent sites having stronger signals. Despite this, the level of insulation observed at low-signal conserved sites was equivalent to the level of insulation at high-signal divergent sites, meaning that conservation is able to separate strongly from weakly insulating sites more effectively than ChIP signal alone (**FIGURE 17C**).

The results presented in **FIGURE 17B** were not altered when using a more stringent ChIP threshold for the definition of conserved and divergent CTCF sites (**FIGURE 18A-B**) or when divergent sites were further validated by stratifying for the presence or absence of the CTCF motif. Sites that contained a motif were the majority (1329 Mmus+/Mmul-, 1472 Mmus+/Cfam-, 1144 Mmul+/Mmus- and 1153 Cfam+/Mmus-, representing 61%, 68% 62% and 58% of the total divergent sites, respectively) and showed an insulation profile comparable to all divergent sites (**FIGURE 18C**). Interestingly, divergent sites for which a motif could not be identified also appeared to mediate some level of contact insulation (**FIGURE 18C**).

In summary, a strong correlation exists between the evolutionary dynamics of CTCF binding sites and mouse, macaque and dog chromosome topology, indicating the possibility of a direct link between insulator site divergence and the evolution of TAD structure.

Divergent CTCF binding drive structural change within large-scale domains.

In contrast to highly stable and strongly insulating sites, the data showed that species-specific CTCF sites were located primarily within domains and exhibited local contact

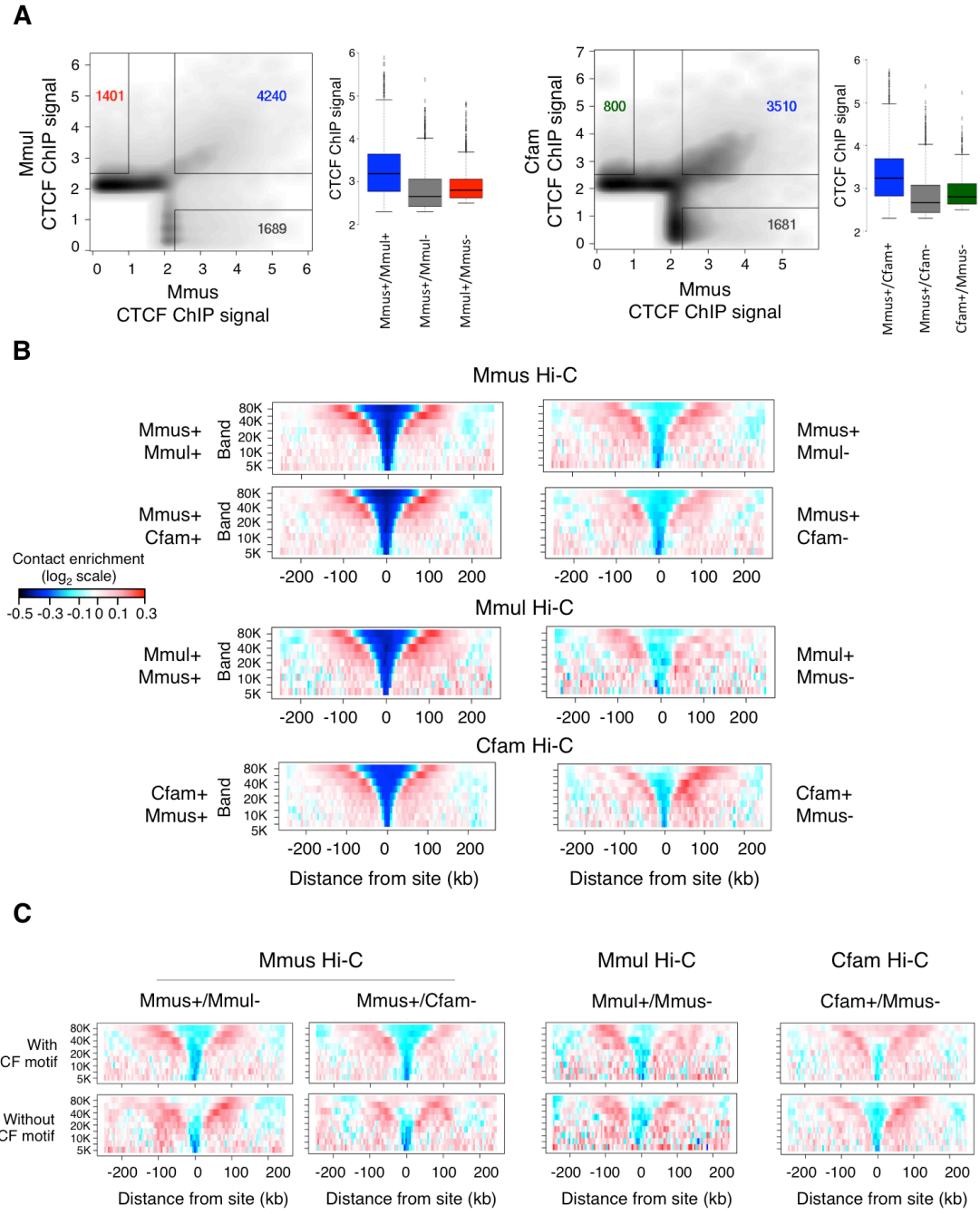


FIGURE 18 – Increased stringency in the definition of CTCF conservation groups does not influence the results. (**A**) Pairwise comparison of mouse CTCF ChIP and macaque (left panel) or dog (right panel) CTCF ChIP as shown in **FIGURE 14C**, but using increased signal thresholds to define conservation groups. Axes are $-\log_{10}$(1-percentile) of the CTCF ChIP-seq read coverage. The distribution of signal intensities is also shown for the conservation groups identified. For divergent sites, intensity values refer to the genome where CTCF is bound. (**B**) Average contact insulation analysis at conserved (left panels) and divergent (right panels) CTCF sites in mouse, macaque and dog Hi-C datasets using the groups defined in (**A**). (**C**) Average contact insulation analysis at divergent sites stratified for the presence of a CTCF motif in their sequence.

insulation (**FIGURE 16B** and **FIGURE 17B**, right panels). The pairwise divergent CTCF sites were analysed for their contact insulation profiles in the genome where CTCF binding was present and in the corresponding points of the genome where the binding was absent (**FIGURE 19**). For example, dog-specific CTCF sites (Mmus-/Cfam+) exhibited contact insulation specifically in the dog genome, whereas these same sites (or rather the corresponding syntenic loci) exhibited background levels of contact insulation when examined in the mouse Hi-C data (**FIGURE 19C**). The same observation held true for macaque comparisons (**FIGURE 19A**). To have a more quantitative assessment, the level of contact enrichment was calculated across individual CTCF binding sites. This approach yielded information about the level of contacts crossing a CTCF site and the corresponding site in another species: for conserved sites, the locus was occupied by CTCF in both species, whereas for divergent sites, the protein was not bound at one of the two loci. This way, it was possible to calculate the difference in the level of contact insulation at each pair of loci. Since a high degree of contact insulation means that there is low contact enrichment across a CTCF site, the data from **FIGURE 19A** and **C** suggest that sites bound by CTCF should have a smaller value of contact enrichment across them than sites where CTCF is not bound. For example, subtracting the contact enrichment mouse from the contact enrichment in dog for Mmus+/Cfam- CTCF sites should give a distribution shifted towards more negative values, with respect to Mmus-/Cfam+ CTCF sites. Indeed, comparing the distribution of pairwise changes in contact enrichment revealed that following CTCF binding site divergence a strongly significant difference at the lower (20 kb) scale, but not at the higher (80 kb) scale, suggesting that CTCF evolution prominently affects shorter-range interactions, while large scale domain structures are either less affected by changes in CTCF binding or their alteration is under strong negative selection and therefore not observed. This localised effect of divergent CTCF binding was particularly evident in the comparison against dog (**FIGURE 19D**), where at the higher band no significant difference in contact insulation was observed. The comparison with macaque, on the other hand, yielded similar results, but the difference in insulation at the higher band remained significant in this case, albeit to a much lower extent (**FIGURE 19B**). This data demonstrate a relationship between CTCF binding divergence and divergence of insulation structure and therefore points toward a role for CTCF in driving structural change in the genome.

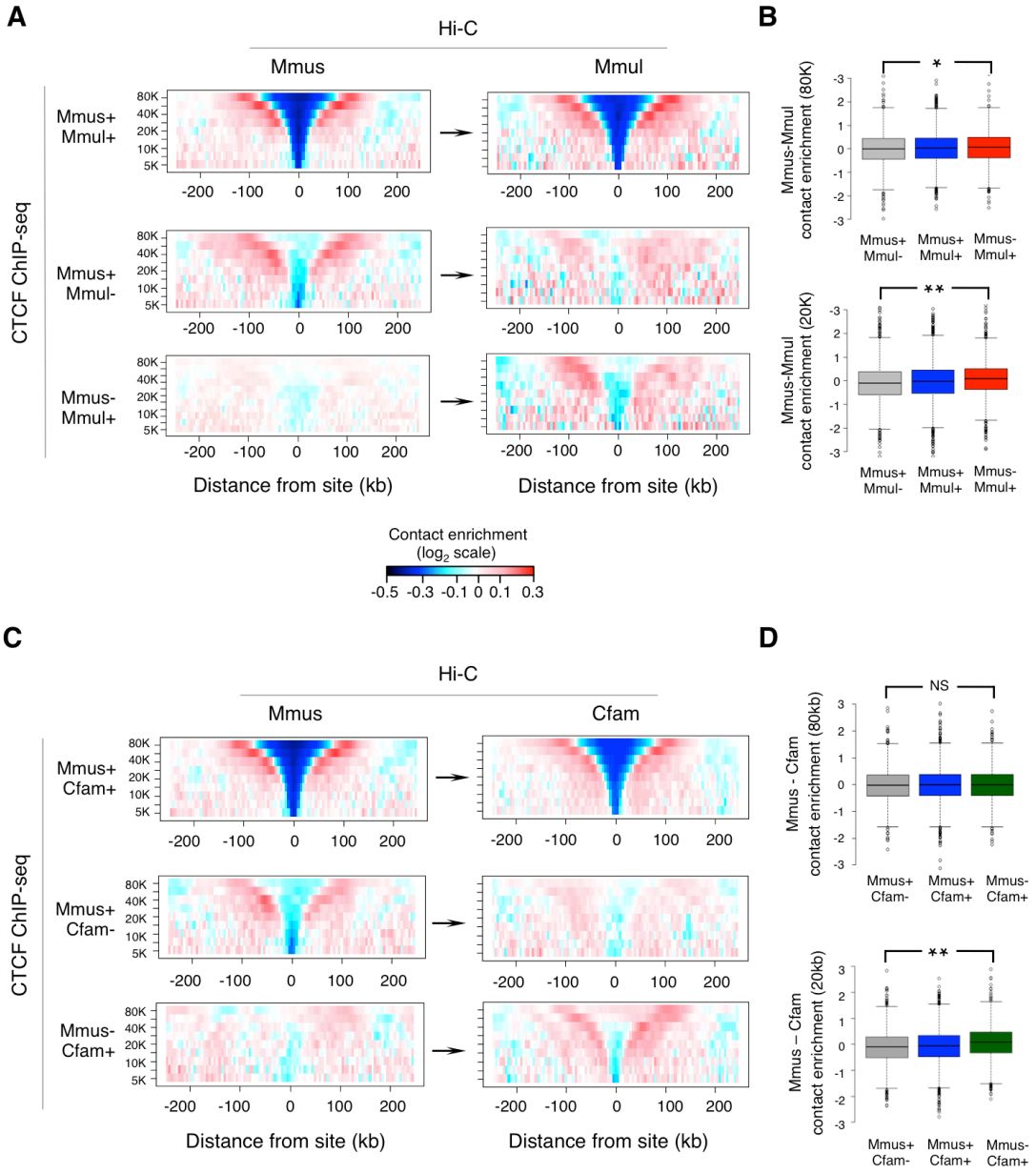


FIGURE 19 – Changes in CTCF binding correlate with changes in contact insulation. (**A**) Average contact insulation analysis at conserved (Mmus+/Mmul+) mouse-divergent (Mmus+/Mmul-) and macaque-divergent (Mmus-/Mmul+) CTCF sites in mouse (left panels) and macaque (right panels) Hi-C datasets. Arrows indicate liftOver of sites. (**B**) Distribution of the difference in contact insulation between mouse and macaque at conserved or divergent CTCF sites at 80-240 Kb (upper panel, 20 Kb band) or 20-60 Kb (lower panel, 80 Kb band) scales. Asterisks indicate a statistically significant difference (**Kolmogorov-Smirnov test, $p < 1x10^{-11}$, *Kolmogorov-Smirnov test, $p < 0.01$). (**C**) Average contact insulation analysis at conserved (Mmus+/Cfam+), mouse-divergent (Mmus/Cfam-) and dog-divergent (Mmus-/Cfam+) CTCF sites for the mouse (left panels) and the dog (right panels) Hi-C datasets. Arrows indicate liftOver of sites. (**D**) Same as (**B**), but for the mouse/dog comparison (NS: Kolmogorov-Smirnov test not significant).

Conserved CTCF sites interact with other conserved sites.

To ask if evolutionarily stable and flexible CTCF sites interact with one another and to further understand how they contribute to the evolution of chromosome domain structure, a high-resolution 4C-seq study was performed. In 4C-seq (Circular Chromosome

Conformation Capture-sequencing), PCR is used to amplify a specific fragment in the genome (e.g. one containing a transcription factor binding site) and paired-end sequencing is subsequently used to identify all possible partners of the "bait" fragment (van de Werken et al., 2012). Two rounds of digestion/re-ligation are employed in this technique to maximise the resolution and complexity of the fragment pool (**FIGURE 2B**). The sequencing results are normally visualised as a line graph showing the median read coverage and as a heatmap ("domainogram") using a range of sliding windows from 2 kb to 50 kb (**FIGURE 2C**).

Four 4C-seq viewpoints were designed to a series of neighbouring conserved CTCF binding sites bordering conserved domains as well as to a mouse-specific site. Bait primers were designed to both mouse and dog genomic locations. The results showed that each conserved CTCF site engages in very strong and specific interactions with other neighbouring conserved CTCF sites (**FIGURE 20A-B**). Remarkably, the specific interactions mediated by conserved sites in the mouse genome were themselves precisely conserved in the dog genome and reflect the underlying Hi-C domain structure. Moreover, a viewpoint designed to a mouse-specific site exhibited a smear of weak interactions that were contained within the mouse domain. Importantly, the mouse-divergent viewpoint had no prominent interactions in the dog genome, confirming the specificity of its interaction network.

Global analysis of Hi-C contacts between pairs of CTCF binding sites stratified according to genomic distances in *cis* (Sofueva et al., 2013) confirmed the 4C-seq observation systematically (**FIGURE 20C**). Consistent with the high-resolution 4C-seq profiles, Hi-C trends showed that conserved CTCF sites strongly contacted one another within the same domain even when separated by large distances. Divergent CTCF sites engaged in levels of contacts with other divergent sites that were similar to conserved-conserved up to about 100 kb, but became significantly weaker for longer distances. Importantly, little to no contact was observed in the mouse genome between dog-divergent sites. These results show that evolutionarily stable CTCF sites are engaged in strong, longer-range contacts with one another and suggest that in so doing, they create an interaction network that may support the conservation of domain structure. On the other hand, divergent CTCF sites are involved in weaker interactions within domains, perhaps reflecting the evolutionary plasticity of the binding sites themselves.

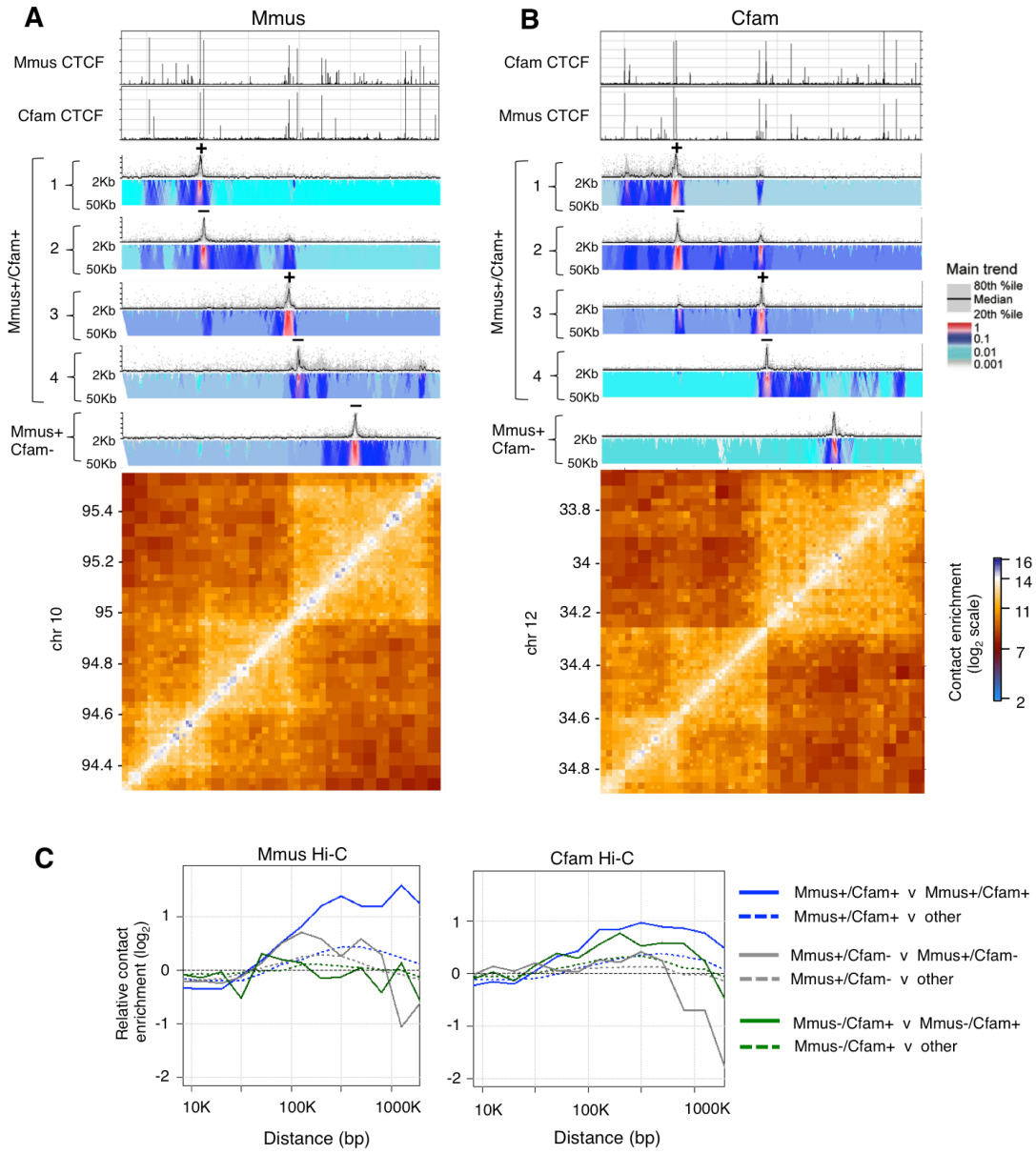


FIGURE 20 – Conserved CTCF sites engage in strong, directional interactions with other conserved sites. (**A**) 4C-seq analysis and Hi-C maps of a 1.2 Mb syntenic region in the mouse and (**B**) dog genomes. Shown are 4C-seq viewpoints designed to four conserved (Mmus+/Cfam+) CTCF binding sites proximal to Hi-C domain borders, and one 4C-seq viewpoint located at a mouse-divergent (Mmus+/Cfam-) CTCF site. The symbol above each 4C-seq bait indicates the strand (and orientation) of each viewpoint. Each 4C-seq experiment is represented by the median normalized 4C-seq coverage in a sliding window of 5 Kb (top) and a multi-scale domainogram indicating normalized mean coverage in windows ranging between 2 Kb and 50 Kb. (**C**) Relative intra-domain contact enrichment between CTCF sites (solid lines) as a function of distance when the two sites are < 5 Kb away from a CTCF binding site (solid lines) or where only one site is < 5 Kb away from a CTCF site (dotted lines). Shown are the relative contact enrichments within repressive domains for conserved (Mmus+/Cfam+, blue), mouse-divergent (Mmus+/Cfam-, grey) and dog-divergent (Mmus-/Cfam+, green) CTCF sites in mouse Hi-C (left) and dog Hi-C (right).

CTCF mediates directional interactions based on the orientation of its binding motif.

The prominent directional bias observed in the 4C-seq profiles of the conserved CTCF binding sites together with the fact that this factor binds to an asymmetrical consensus

motif suggests the possibility of a simple relationship between the orientation of its binding and the directionality of its interaction profile.

In order to investigate this possibility, the sequences underneath the five CTCF peaks profiled by 4C-seq were scanned for the presence of the CTCF consensus motif and each peak was annotated with the orientation of the motif underneath it. Orientation is indicated using the strand on which the consensus motif was lying. This analysis showed that peaks with the same motif orientation had corresponding biases in the direction of their

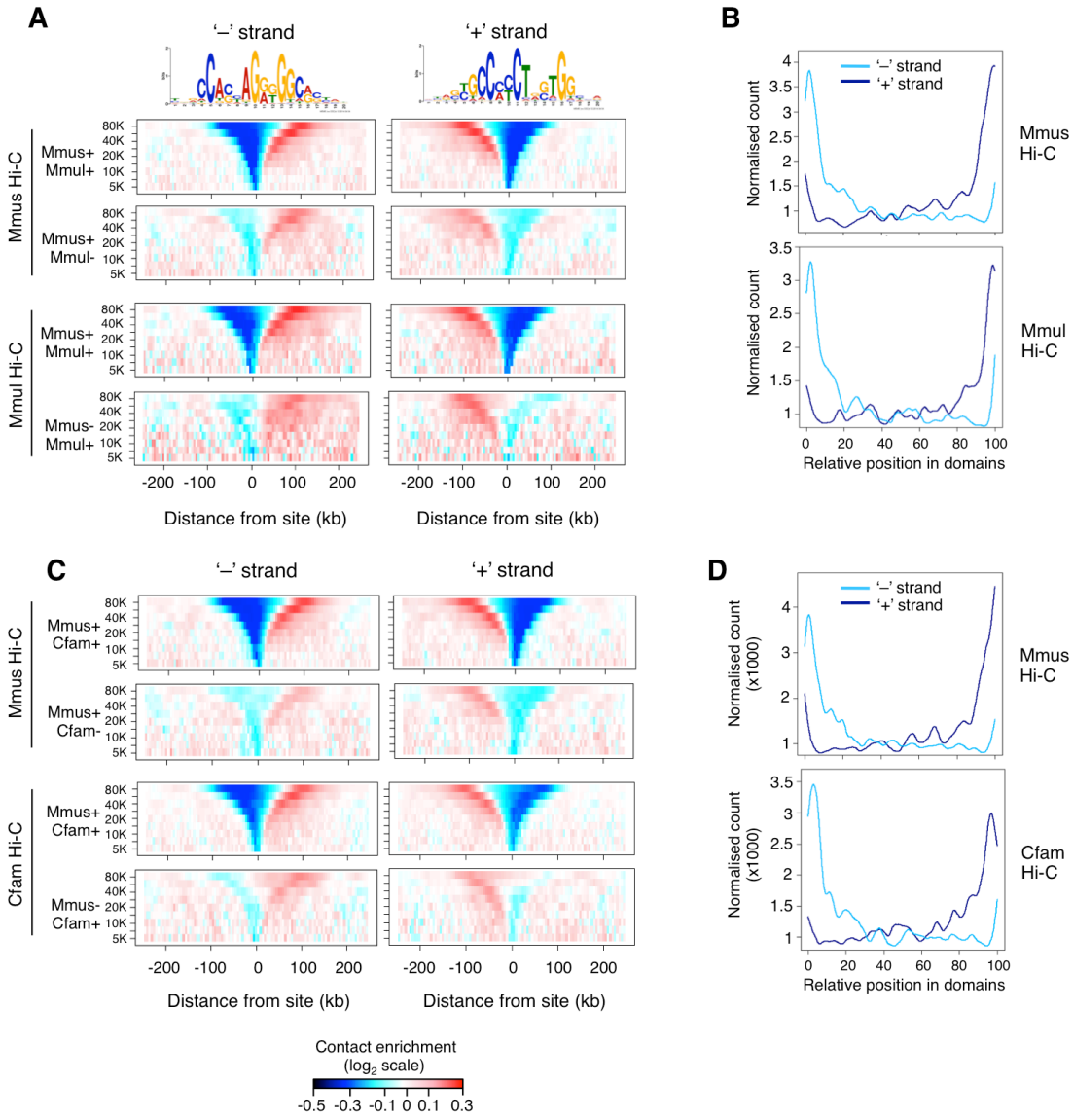


FIGURE 21 – CTCF dictates directional interactions based on the orientation of its binding motif. (**A**) Average contact insulation analysis in both mouse and macaque genomes for conserved (Mmus+/Mmul+) and divergent (Mmus +/Mmul- or Mmus-/Mmul+) CTCF binding sites grouped according to the orientation of their binding motif. The consensus motif shown was generated from mouse binding sites. (**B**) Genome-wide relative position within chromosomal domains of Mmus+/Mmul+ conserved CTCF sites grouped according to the orientation of their binding motif as above. (**C**) Average contact insulation analysis in both mouse and dog genomes for conserved (Mmus+/Cfam +) and divergent (Mmus+/Cfam- or Mmus-/Cfam+) CTCF binding sites grouped according to the orientation of their binding motif. (**D**) Genome-wide relative position within chromosomal domains of Mmus+/Cfam+ conserved CTCF sites grouped according to the orientation of their binding motif as above. Counts are normalised by the total number of sites in each group.

interactions (**FIGURE 20A-B**); specifically, all peaks lying on the "plus" stand interacted mostly upstream of the binding site, whereas peaks lying on the opposite strand interacted downstream of the binding site. Notably, even the divergent CTCF shows a weaker directional bias in its interaction profile in agreement with its binding motif orientation.

To determine if this was a genome-wide effect, the orientation of every conserved or divergent CTCF peak was assessed and insulation plots were generated separating binding sites lying on the two different strands. As can be seen in **FIGURE 21A** and **C** a striking asymmetry in the plots can now be observed, with sites lying on opposite strands having a mirrored insulation profile. In these plots, one side of the anchor points shows an enrichment in contacts (visualised in red in the heatmaps), whereas the other side exhibits a marked depletion of contacts (visualised in blue in the heatmaps). This can be interpreted as CTCF sites "folding the DNA" in one direction or the other depending on the orientation of their CTCF binding motif. Such contact asymmetry applies to both conserved and divergent CTCF sites, albeit with different intensities.

As conserved CTCF binding sites are found predominantly close to the borders of domains (**FIGURE 16B**), the segregation of the conserved CTCF binding sites according to their motif orientation was also able to effectively separate sites enriched at the "left" and "right" edges of the chromosomal domains in all species studied (**FIGURE 21B** and **D**). Consistent with these observations, the motif orientation for conserved CTCF binding sites was also found to be conserved in 94% of cases.

Cohesin is required for CTCF-dependent contact insulation.

The best-defined mechanism by which CTCF mediates the organisation of chromatin folding is through recruitment of the cohesin complex (Hadjur et al., 2009; Nativio et al., 2009). To investigate the possible influence of cohesin co-occupancy on the directionality of CTCF-mediated interactions, CTCF peaks in mouse were grouped based on the presence or absence of cohesin subunit Rad21 (22806 and 8246, respectively, see MATERIALS AND METHODS for details), and then further segregated according to the orientation of their CTCF binding motif. All mouse CTCF peaks were analysed (without the filters for the inter-species comparisons) to try and maximise the number of sites and improve the robustness of the result. This analysis was run only for mouse, as the Rad21 ChIP-seq (from Faure at al. 2012) was most reliable for this species. The insulation plots

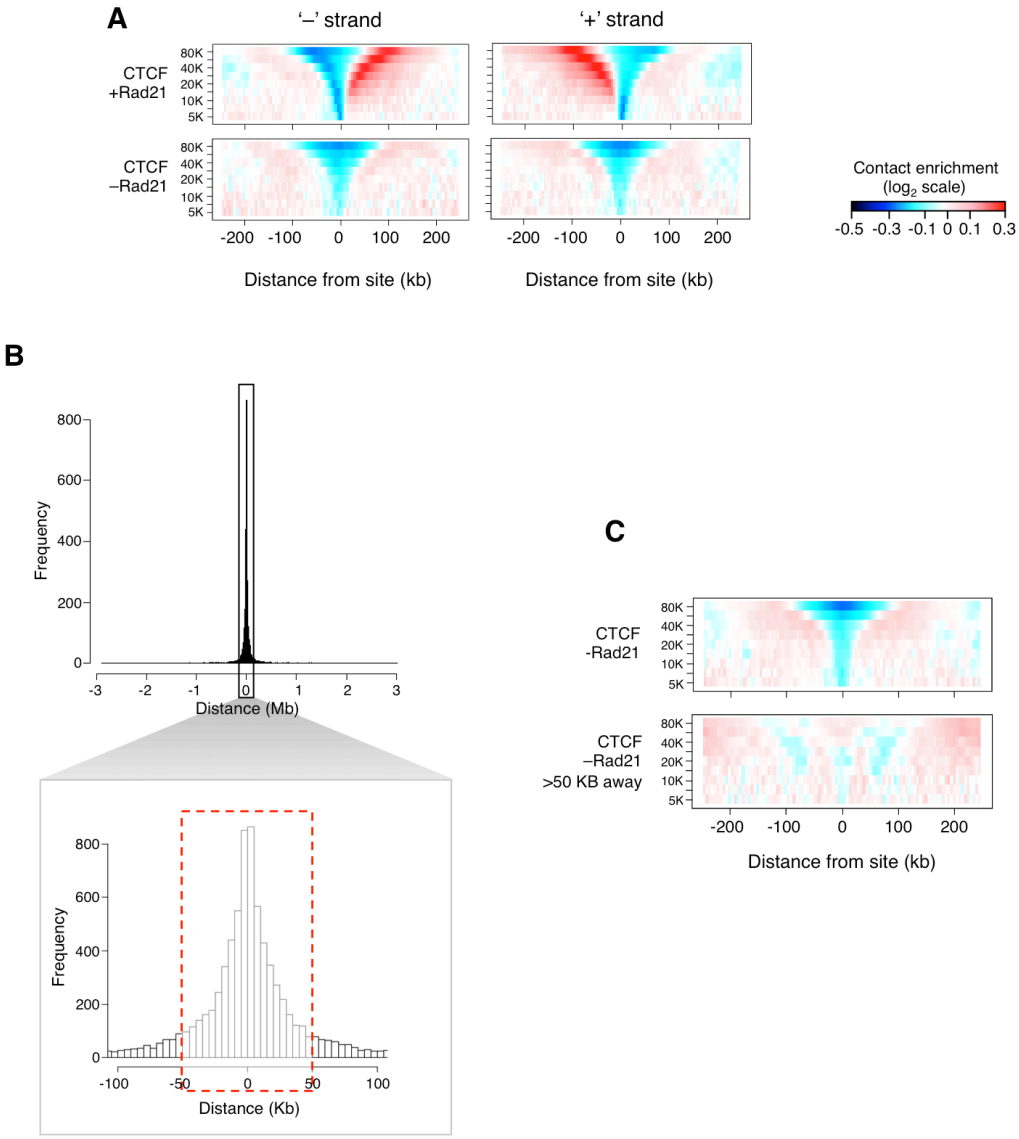


FIGURE 22 – CTCF-mediated insulation depends on the presence of cohesin. (**A**) Average contact insulation analysis for oriented mouse CTCF sites co-occupied by or without cohesin subunit Rad21. (**B**) Distribution of distances between CTCF sites without Rad21 and the closest CTCF/cohesin site. The full distribution is shown in the upper panel. The lower panel shows the sites that were excluded from subsequent insulation analysis due to their proximity to CTCF/cohesin (dashed red box). (**C**) Average contact insulation analysis for mouse CTCF sites without cohesin (upper panel), and the same plot after removal of sites located less than 50 Kb away from a CTCF/cohesin site (lower panel).

show the expected directional profile only for the CTCF sites co-occupied by Rad21, whereas when cohesin is missing, despite the insulation being still present, the asymmetry is lost (**FIGURE 22A**). Importantly, analysis of the distance distribution between CTCF sites without cohesin and CTCF/cohesin sites revealed that they were in most cases found in close proximity of one another (**FIGURE 22B**, upper panel). It was therefore possible that the insulation seen in the plots anchored at CTCF sites without cohesin was simply reflecting the activity of a nearby CTCF/cohesin site. To resolve this, CTCF sites lacking

cohesin that were within 50 kb of a CTCF/cohesin site were excluded from the insulation analysis (**FIGURE 22B**, lower panel). Remarkably, 1958 CTCF sites that do not co-localise with cohesin and are located more than 50 kb away from a CTCF/cohesin site displayed no contact insulation across them, but only two drops in contacts about 50-100 kb on either side of the anchor point (**FIGURE 22C**). Interestingly this is the expected distance of the closest CTCF/cohesin site for the majority of CTCF sites contributing to this plot.
Taken together, these results show a strong correlation between the directionality of CTCF-mediated interactions and the sequence motif underlying its binding sites and identify the cohesin complex as a required factor for CTCF-mediated asymmetric contact insulation.

4.3 DISCUSSION

By integrating the evolution of CTCF binding with chromosome topology maps, I was able to separate CTCF sites into two groups – those sites which were conserved and found near the edges of domains from those which were divergent and distributed within domains. Moreover, these sites exhibit markedly different abilities to mediate contact insulation across them, providing a possible functional basis for their conservation: while conserved sites exhibited a strong insulation profile, divergent sites appeared to be insulating more weakly. Interestingly, divergent sites that did not include a CTCF binding motif were also able to mediate a small level of contact insulation. Indeed, CTCF is known to bind sites that do not contain a motif (Nakahashi et al., 2013), so these sites may actually be occupied by the protein with low affinity and be involved in the definition of more dynamic chromatin structures.

Comparing contact insulation at divergent sites showed how gain/loss of CTCF binding leads to a parallel gain/loss of insulation. This observation strongly supports a causal connection between CTCF binding and chromosomal looping structures. The divergent CTCF sites might be able to engage in regulatory interactions between enhancers and promoters within domains. The presence or absence of CTCF at these sites might contribute to the differential regulation of gene expression across species.

Interestingly, the CTCF sites affecting structural divergence were primarily locally insulating. Instances of large-scale insulation points showing strong divergence among species, thus producing split or fused TADs, were not observed, at least in this analysis. This is consistent with domains being inherited as intact modules (**FIGURES 11-13**) and paints a picture where conserved large-scale domains provide a stable framework within which plastic local interactions can shape evolving genomes.

Going more into the molecular detail of how chromatin interactions are set up, the high-resolution 4C-seq data showed a marked asymmetry of the interactions coming from conserved CTCF sites (also observed in Sofueva et al., 2013). Examining the sequence underneath the binding sites genome-wide revealed a striking influence of the orientation of the CTCF binding motif on the directionality of the interactions. Notably, when studied genome-wide, this orientation-mediated asymmetry applied also to divergent sites. Looking carefully, the 4C-seq profile at the divergent site does in fact shows a slight asymmetry that is consistent with the orientation of its motif. The weaker directionality might be explained by the lower ChIP-seq signal at this site, suggesting that CTCF was possibly not bound in all of the cells in the population.

These results further strengthen a direct link between the DNA sequence and the architecture of the genome, suggesting that the sequence orientation alone might be sufficient to instruct the direction in with CTCF will fold the DNA. To confirm this, though, genome-editing approaches such as CRISPR-mediated deletion of a CTCF site followed by insertion of the same site with a reversed orientation will be needed.

Remarkably, CTCF sites that do not co-localise with cohesin and are located at least 50 kb from a CTCF/cohesin site appear unable to mediate contact insulation. The fact that the only insulation visible in this plots is localised where the CTCF/cohesin site closest to the anchor is found further points to CTCF/cohesin sites as major determinants in the definition of chromatin structure and the requirement of both factors for this function.

CHAPTER FIVE:

DISCUSSION AND FUTURE PERSPECTIVES

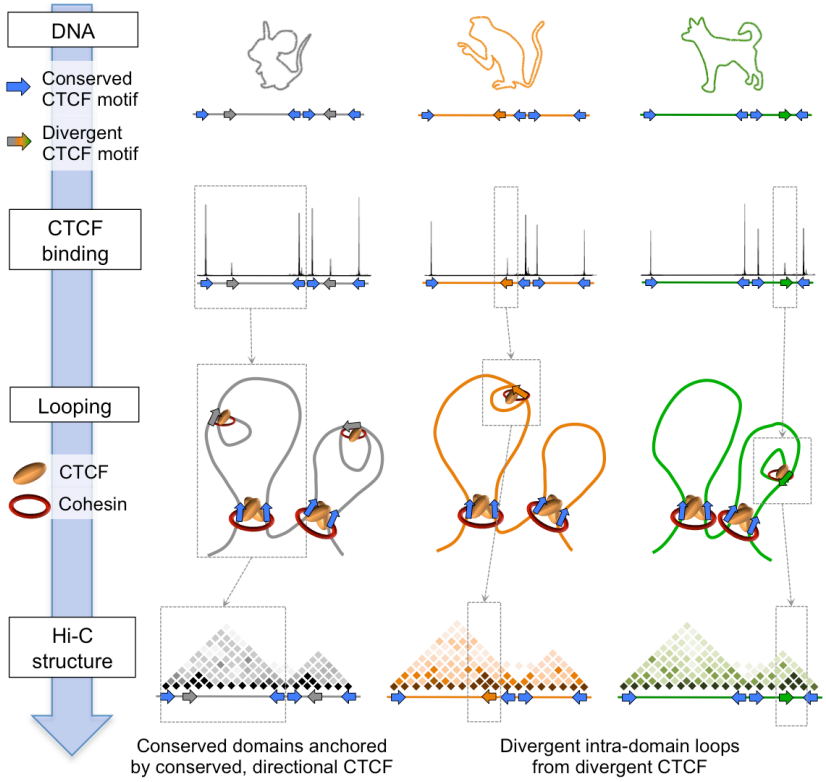


FIGURE 23 – Visual summary of the findings presented in this work.

This thesis explores the folding of chromosomes in the context of evolving genomes. By comparing Hi-C datasets prepared from the livers of four mammalian species, my results have revealed a remarkable level of conservation of chromosomal architecture across 100 million years of evolution. Structural divergence appears to be limited to local interactions within the interior of large-scale domains that tend to be reshuffled as intact modules during chromosomal rearrangements. Relating this structural information to the evolution of the binding of CTCF unveiled how these two phenomena are deeply intertwined. CTCF genomic distribution appears to have evolved under two different regimes, as it was possible to identify both a conserved population of sites characterised by strong binding and high affinity to the protein's consensus sequence, and a separate population of divergent, more plastic sites. Conserved CTCF binding sites are strongly enriched at large-scale domain borders and occur at strongly insulating loci in all mammalian genomes studied. Conversely, divergent CTCF sites are widely distributed within domains and

mediate weaker insulation that can be observed only in the genomes where the protein is bound to the DNA. A high-resolution analysis of candidate sites revealed that conserved CTCF binding sites tend to contact one another and that interactions anchored at such sites are themselves conserved. These interaction profiles exhibit a marked directionality bias that is dictated by the orientation of the DNA sequence motif underlying the site. The influence of CTCF motif orientation on the directionality of its interactions applies also to divergent sites, although in a much less prominent way. Notably, no contact insulation can be observed at mouse CTCF sites that are not co-occupied by cohesin. A visual summary of the findings reported in this thesis is presented in **FIGURE 23**.

These findings provide evidence for a direct link between the DNA sequence, binding of CTCF to its motif and the architecture of chromosomes in the context of evolving genomes and have strong implications on the study of CTCF and genome function, and on our understanding of the evolutionary dynamics in complex genomes.

The evolution of mammalian chromatin architecture

Large-scale domains are rearranged as intact modules during mammalian evolution.

The genomes of the species analysed in this study have largely rearranged karyotypes. Aside from the higher number of chromosomes in dog (39 pairs against the 20, 22 and 21 of mouse, rabbit and macaque, respectively), the rodent lineage has undergone an extensive karyotypic diversification (Romanenko et al., 2012) and mouse has a very derived karyotype. This diversity is the basis for the vast repertoire of chromosomal rearrangements that can be observed between these species. In most cases the reorganisation of the chromosome is too complex to analyse and can generate non-syntenic regions that prevent any comparison between genomes. In some cases, however, the rearrangement is simple enough to not introduce regions lacking synteny and allows for a reconstruction of the event that led to it. The nature of such events is normally that of large inversions or insertions/deletions.

Even though these occurrences are difficult to study systematically due to their context-dependent nature, my examination of candidate cases has yielded some interesting insights. Most importantly, I observed that rearrangements happened in such a way that the domains were kept intact. This fact could be consistent with either the "Fragile Breakage Model" or the "Selective Breakage Model".

According to the first explanation, the boundaries of large-scale domains would represent regions that are particularly prone to rearrangement. Indeed, domain boundaries might represent points of torsional stress for the DNA helix, which may be more likely to suffer a double-stranded breakage and subsequently be repaired by mechanisms such as Non-Homologous End Joining (NHEJ). NHEJ may aberrantly re-join a pair of broken ends without any annealing of homologous regions (Lieber et al., 2008). Similarly, domain boundaries are enriched for SINE retrotransposons (Dixon et al., 2012), which could encourage pairings between sequences that are homologous but non-allelic, that is, despite being similar in sequence they belong to different parts of the genome, a phenomenon known as Non-Allelic Homologous Recombination (NAHR) (Sasaki et al., 2010). Moreover, the presence of proteins such as CTCF strongly bound in the vicinity of boundaries might produce replication fork stalling at these sites, in turn rendering them more amenable to undergo DNA polymerase template switching (Lee et al., 2007) or, simply more prone to break.

Alternatively, the preservation of domains as intact units might stem from a strong negative selection on rearrangements that disrupt them, so that important intra-domain regulatory interactions can be preserved. In support of this, several studies are consistent with the idea that most interactions that are relevant for gene expression happen within domains: a significant example of this behaviour is the regulation of the Sonic HedgeHog gene (*Shh*), where different enhancers within a chromosomal domain affect the expression of the gene at different times or places during development (Anderson et al., 2014). Similarly, the characterisation of a subset of enhancer-promoter interactions and repressive regulatory domains in ES cells has also indicated that these gene-regulatory interactions tend to happen within domains (Dowen et al., 2014).

Selection may operate in certain cases to protect a domain boundary against breakage. The enrichment of housekeeping genes at domain boundaries (Dixon et al., 2012) and examples of gene regulation that involve neighbouring domains – as in the case of *Xist/Tsix* (Nora et al., 2012) – would point to negative selection for boundary-proximal rearrangements at these sites. In such instances, the boundary will still remain conserved because it is directly relevant to genome function, but it will tend not to be rearranged; in other words, the

boundary will separate the same two neighbouring TADs over evolutionary time. Notably, this action of selection would preserve multi-domain syntenic regions across species.
These evolutionary forces – increased fragility and the action of selection – are not mutually exclusive and likely operate at the same time with varying strengths at different locations to influence the karyotypic evolution of genomes.

Gene clusters can form their own domain and selectively interact with their surroundings.
Gene clusters are thought to emerge from tandem gene duplication events followed by diversification and functional integration of the related genes (Lawrence, 1999). The intermediate steps through which these groups of genes become co-regulated remain obscure.

The insertions of the *Skint* (**FIGURE 11**) and *Mrgpr* (**FIGURE 12**) gene clusters are likely the result of tandem gene duplication events. It is worth noting that these clusters make up their own domains and appear to form an elaborate internal network of interactions. In the case of the Skint cluster, for example, one large gene in the middle of the cluster (*Skint6*) forms its own domain that remains slightly more "isolated" from the rest of the cluster. This articulate internal organisation of a newly emerged cluster of genes argues for the view of domains as modules within which gene regulatory loops can occur. Moreover, the fact that tandemly duplicated genes are able to form a highly self-interacting chromosomal domain might encourage the establishment of novel regulatory contacts between them.

It is also interesting to look at how the inserted regions integrate into their "new" surroundings. In this respect, the data suggests a more context-dependent behaviour: while the *Skint* cluster engages in contacts with the domains on either side of it, the *Mgrpr* cluster seem to reach out to a domain about 1 Mb downstream. The recent identification of six compartments of domains clustering based on their epigenetic chromatin state (Rao et al, 2014) might suggest a rationale for this selectivity in inter-domain contacts: the histone modification patterns present in the inserted domain could direct its preference to interact with specific partner domains. In both the *Skint* and *Mrgpr* case, however, the interactions occurring between the domains flanking the inserted regions are maintained. The importance of these contacts between large-scale domains on genome function is still poorly understood, but these observations suggest that they have a degree of relevance since some of them have been preserved across millions of years of evolution.

The role of CTCF in genome architecture evolution

Two mechanisms may contribute to the conservation of CTCF sites.

CTCF binding sites can be classified based on their conservation across distantly related species (Schmidt et al., 2012; this work). Two possible models could explain the evolutionary stability of certain CTCF sites.

On the one hand, conserved sites might have a very strong functional importance and mutations that reduce the affinity of the protein, or create a very strong new binding site might simply be not well tolerated and thus powerfully selected against *a posteriori*. The specific and marked enrichment for conserved CTCF sites at the borders of large-scale chromosomal domains argues in favour of this mechanism, because it supports a role for these sites in the establishment or maintenance of borders, whose sudden emergence or disappearance is likely to be selected against, as discussed in the previous section.

Another explanation is that CTCF occupancy of a certain site influences its mutational rate *a priori*. The high ChIP-seq signal intensity of the conserved sites can be interpreted as a high frequency of binding at these sites, which would mean that CTCF is nearly always bound there; such high occupancy (and possibly the fact that the DNA is being looped) might reduce the likelihood of mutations happening in these sequences, thereby rendering them more stable in time; conversely, at divergent sites, less frequent binding would leave the sequence more exposed to change. Alternatively, these highly occupied sites might interfere with the process of retro-transposition that has been hypothesised to drive CTCF evolution (Shmidt et al., 2012). The RNA polymerase activity that is necessary for the early steps in this process might be influenced by presence of other proteins in its path and indeed CTCF has been suggested to promote RNA polymerase pausing (Shukla et al., 2011; Paredes et al., 2013). If this were true, retro-transposition of sites that are very-stably bound and mediate strong DNA looping would occur more rarely if at all. It should be emphasised however that the relationship between CTCF binding and RNA polymerase pausing has been described sparsely and only in protein-coding genes making this a highly speculative model.

Of note, the action of selection and the influence of protein occupancy on mutational rate are not mutually exclusive and may act in concert on the evolution of CTCF binding.

The evolution of CTCF binding and genome architecture are deeply intertwined.

Genome-wide chromatin architecture studies have identified a strong enrichment of CTCF/cohesin at domain boundaries (Dixon et al, 2012; Sofueva et al., 2013; Zuin et al., 2014). The vast majority of CTCF, though, binds away from such locations and more internal to domains. What may distinguish between CTCF sites bound at boundaries and those found internal to domain has remained elusive. In light of the remarkable preservation of large-scale chromatin structure across multiple mammalian species, I hypothesised that the presence of CTCF at these key structural loci could be related to the evolutionary conservation of its binding. My results have indeed identified conserved or divergent CTCF binding events that are differentially distributed relative to domains, with more conserved sites strongly enriched at domain boundaries.

The stability of larger domains is likely related to the presence of these strong, conserved CTCF binding at their borders and to the ability of these sites to form a conserved network of preferential interactions between them. Further, the presence of other proteins involved in chromatin insulation might also enhance the robustness of these boundaries. Recently, de-convolution of chromatin interaction data has also revealed an influence of internal domain contacts on the stability of the boundaries between domains (Giorgetti et al., 2014); in other words, the looping of DNA within a domain alone tends to "pull away" the chromatin fibre from the neighbouring domain, helping to create the boundary between them. The same study also indicated domains as very dynamic structures, raising questions about how they are formed and maintained and whether these are distinct or closely related processes.

The emergence of CTCF might have created a novel way to modulate DNA folding.

The role of transcription in the formation of chromosomal domains has been subject of debate. Domain boundaries are enriched for housekeeping genes (Dixon et al, 2012), leading to the hypothesis that a high level of transcriptional activity might play a role in the demarcation or in the maintenance of these structural hotspots. The results presented in this thesis reveal a direct correlation between chromatin structure and DNA sequences which act as a grid to recruit architectural proteins. These findings, however, do not rule out transcription as a contributor to the folding of the chromatin fibre.

The recent characterisation of chromosomal structure in fission yeast (*Schizosaccharomyces pombe*) (Mizugushi et al., 2014) makes room for some interesting speculation in this respect. In *S. pombe*, chromosomes are subdivided in smaller domain-like compartments (termed "globules", 50-100 kb in size), whose boundaries are dependent on the presence of cohesin and are found between convergent genes. Interestingly, yeast does not express CTCF and on chromosome arms cohesin has no known anchor to tether it to specific DNA sequences; in this context, transcription from convergent genes might dictate the position of a domain boundary by "pushing" cohesin rings in the space between said genes, insulating them into separate globules. A similar phenomenon has indeed been described in the budding yeast *Saccharomyces cerevisiae* (Legronne et al., 2004; Glynn et al, 2004).

In metazoan organisms, and more specifically in *Bilateria*, evolution of CTCF may have, at least in part, "decoupled" domain definition from transcription, providing a more versatile tool to fine-tune DNA folding without having to act on the potentially delicate balances of gene expression or gene positioning. This is consistent with the hypothesis that CTCF created a novel way by which genomic elements could be regulated, perhaps allowing the evolution of fundamental gene networks – for example the *Hox* cluster – and placing this factor at the centre-stage of the "Cambrian explosion" (~540 million years ago), when all major clades of bilateral animals appeared over a short evolutionary time (Heger et al., 2012). A fascinating scenario would be that this decoupling happened through CTCF serving as an anchor for cohesin already at the time of its emergence, near the root of *Bilateria*; an alternative model would have CTCF acting as a chromatin organiser in a cohesin-independent fashion in early bilaterians, for example by looping DNA through dimerisation, and would restrict the molecular association between CTCF and cohesin to vertebrates. Studying the relationship between CTCF and cohesin in other non-vertebrate bilaterian animals will help uncover its evolutionary history.

Irrespective of when the CTCF/cohesin "partnership" was established, the fact that CTCF binding can be modulated by simply acting on the sequence makes it valuable in the an context of evolution, allowing for an increased plasticity to adapt chromatin structures to the mutations shaping the evolving genome.
Further insights that will shed light on the evolution of chromatin architecture and the mechanisms underlying it may come from the mapping of chromatin interactions in

bilaterian organisms that have lost CTCF, such as *C. elegans*, and from the dissection of the interplay between cohesin, transcription and domain boundaries in this species.

Intra-domain interaction divergence might drive the evolution of gene regulatory networks.

While the influence of transcription on chromatin folding remains to be elucidated, DNA looping has a known impact on the transcriptional regulation of genes. Given that most enhancer-promoter contacts happen within domains at scales well below a megabase (Phillips-Cremins et al., 2013; Jin et al., 2014) I hypothesise that local interactions anchored at divergent CTCF sites might rewire the internal architecture of domains to potentially contribute to the emergence of new evolutionary phenotypes by establishing different networks of gene regulation.

The divergence of local contact insulation at divergent CTCF sites could for example underlie differential gene regulation across species. Although a preliminary analysis was consistent with this hypothesis, the lack of sufficiently high-resolution HiC maps, prevented me from robustly identify individual instances of gain/loss of short-range contacts. Such an analysis is also complicated by the fact that the expression of orthologous genes in adult primary liver is highly correlated across different species (Brawand et al., 2011), rendering the search for differentially expressed genes challenging. Accurate identification of candidates would require deep RNA-sequencing and perhaps modelling of the expected expression differences between species. Use of a tissue or cell type that can ensure a wider gene expression radiation in different mammals might provide also a more suitable system.

CTCF motif directionality may provide insight into complex assembly at domain boundaries.

One of the most exciting observations in my thesis is that the orientation of CTCF is able to dictate the directionality of the interactions it mediates. Strongly interacting CTCF sites are found to have motifs oriented in opposing directions. This result is consistent with early reports showing that the insulator activity of a CTCF binding site at the H19/Igf2 locus was dependent on orientation in transgenic assays (Bell and Felsenfeld, 2000; Hark et al., 2000). The directionality of CTCF binding sites may give some insight on how the anchoring of chromatin loops may be set up at the molecular level.

I have shown that two strong CTCF sites with opposite orientations anchor chromatin loops and thereby define chromosomal domains (**FIGURE 19**). CTCF has been shown to

bind cohesin through its C-terminal portion (Xiao et al., 2011) and is known to bind the consensus labelled as "minus strand" in this study in a reverse orientation (Nakahashi et al., 2013) (**FIGURE 3**). This would theoretically place the cohesin complex towards the outside of the domain, forcing most contacts to happen on the opposite side, towards the interior of the domain. A similar model has indeed been proposed by another paper which has described directionality of CTCF binding sites (Rao et al., 2014). However little is known about how the CTCF protein is folded in vivo, thus it will be important to validate this hypothesis using high-definition ChIP techniques such as ChIP-exo (Rhee and Pugh, 2011) and structural biology approaches.

Biochemical study of the complexes assembled at domain boundaries will also help to shed light on the possible presence of other factors and elucidate the exact stoichiometry of the molecules that participate in the formation of boundary structures. In this context, the way in which cohesin complexes are assembled on chromatin has been a long-standing open question (Zhang et al., 2009). The cohesin complex forms a proteinaceous ring that associates with DNA by encircling a filament. In yeast, it has been demonstrated that a single ring wraps around two DNA filaments to tether sister chromatids together, in what has become known as the "embrace" model (Huis in 't Veld et al., 2014). An alternative "handcuff" model has been proposed, in which two cohesin rings, each encircling one DNA filament, may interact with one another with an anti-parallel orientation (Zhang et al., 2008). This observation is reminiscent of the anti-parallel orientation I observed from conserved CTCF sites and might point to different molecular assemblies for cohesin in the case of sister chromatid cohesion or inter-chromosomal looping. The presence of a "handcuff" at intra-chromosomal loops might also resolve the issue of how the structure of chromatin is kept intact during S-phase: since in this model each cohesin ring is encircling only one filament, a DNA polymerase might simply pass through each ring to generate a new DNA molecule that is already cohesed and correctly looped. The study of the stoichiometry and composition of protein assemblies mediating chromatin loops are challenging because of the likely heterogeneity of CTCF/cohesin complexes present at different loci the genome and will require the development of efficient techniques to isolate and characterise complexes binding at specific genomic locations.

Interestingly, when I stratified divergent sites for the presence of the CTCF binding motif, an insulation profile was visible even in the absence of a motif (**FIGURE 18C**). This

observation raises the question of how insulation can be mediated in the absence of directionality. A simple technical explanation could be that the sequences underneath these sites is indeed sufficient to specify an orientation for the binding but the software used for the analysis is unable to detect it. Excluding this, CTCF may bind at these loci in either orientation, thereby making them even more dynamic in the interactions they might mediate. It is also possible that a different molecular mechanism altogether is operating at these sites to form chromatin structures without the need for oriented CTCF binding.

Regulation of the CTCF/cohesin interaction may be important for genome architecture.

CTCF sites that do not colocalise with cohesin are unable to mediate significant contact insulation (**FIGURE 21C**). Thus, certain CTCF sites are unable to bind to cohesin and are therefore unable to engage in the definition of chromatin structures. This result stresses the importance of studying the ways in which the association between CTCF and cohesin might be modulated.

Although the interaction between CTCF and the cohesin complex is well established, very little is known about its regulation. Various mechanisms could modulate the binding of CTCF to cohesin. These involve post-translational modifications of either protein, or may rely on the action of other factors. Indeed, the fact that the contact interface between CTCF and SA-1/2 includes a phosphorylation site (Xiao et al, 2011) and that phosphorylation at these residues by protein kinase CK2 interferes with CTCF-mediated repression of *c-myc* (Klenova et al., 2001; El-Kady and Klenova, 2005), makes this modification an intriguing candidate for taking part in the regulation of the interaction between CTCF and cohesin. Alternatively, a SUMOylation site on CTCF is also present in the C-terminal region (MacPherson et al., 2009) and may also play a role in modulating the interplay between the two proteins. Intriguingly, SUMOylation of CTCF strengthens CTCF-mediated repression of *c-myc* (MacPherson et al., 2009), the opposite effect than phosphorylation, raising the possibility of a cross-talk between multiple post-translational modifications to fine-tune the CTCF/cohesin interaction. In addition to PTMs regulating the interacton between CTCF/cohesin, it is possible that other factors intervene in the anchoring of cohesin by CTCF. Such a role has been described for stemness factor Oct4, which is able to directly interact with CTCF (Donohoe et al., 2009) and has been shown to interfere with the interaction between CTCF and cohesin at the HOXA cluster in mouse Embryonic Stem Cells (Kim et al., 2011). However the cell-type specific expression of

Oct4 and its limited co-localisation with CTCF (Chen et al., 2008) make it an unlikely candidate as a general regulator of the binding between cohesin and CTCF.

Conclusion

Taken together, the data presented in this thesis strongly support the functional relevance of chromosomal domains and reveal the complex interplay between the evolution of genomes, protein binding and chromatin architecture. These observations represent a further step towards the understanding of the evolutionary process and the regulation of genome functionality.

MATERIALS & METHODS

Liver homogenisation and fixation

The starting material for Hi-C or 4Cseq libraries was frozen liver from mouse, rabbit, macaque and dog, that had been fixed in 10% formalin for 20 min by immersion (see Schmidt et al., 2012). In some experiments, fresh mouse livers were used instead. Pieces of fixed liver (1-1.5 g) from each species were cut and processed with a 15 ml glass Dounce homogeniser [Wheaton cat#357544] using a loose pestle for 10-15 strokes, and a thick pestle for 10-15 strokes. The homogenised material was then filtered through a 70 μm nylon cell strainer to remove connective tissue and washed twice with 50 ml Dulbecco's Phosphate Buffered Saline (DPBS) [Life Technologies cat#14190-094] to remove debris, spinning down to collect the cells at 852 rcf for 5 min at 4ºC. Cells were counted using an Improved Neubauer haemocytometer [Hawksley cat#BS 748]. 1-5 x 10^7 liver cells were then fixed for a second time in 20 ml Fixation Buffer (1% Formaldehyde [Sigma-Aldrich cat#F8775-25ML], 750 μg/ml Fraction V Bovine Serum Albumin [Life Technologies cat#15260] in 50:50 DMEM/Ham's F12 [PAA cat#E15-816]) for 10-30 min at room temperature (see **TABLE 5** for details about each sample). The fixation reaction was quenched using 0.125 M Glycine [Sigma-Aldrich cat#G8898] for 5 min at room temperature. Samples were then spun at 1230 rcf for 5 min at 4ºC, washed twice with 10 ml DPBS and finally pelleted in 1x10^7 cells aliquots and stored at -80ºC. Mouse Hi-C libraries were prepared from fresh liver samples of biological replicates (9 week old C57/BL6 mouse and the pooled livers from 2-4 week old outbred mice). The libraries for the other three organisms were technical replicates.

Propidium iodide staining of hepatocytes

The cytoplasms of 1x10^6 formaldehyde-fixed liver cells from mouse, rabbit, macaque and dog were lysed by incubating for 30 min in a hypotonic buffer (10 mM Tris-HCl pH 8 [Gibco cat#15568-025], 10 mM NaCl [Sigma-Aldrich cat#S3014] 0.2% Igepal CA-640 [Sigma-Aldrich cat#18896], 1X Complete EDTA-Free protease inhibitors [Roche cat#11873580001]) on ice. The DNA in the nuclei was subsequently stained by adding 2 ml PI staining buffer (100 μg/ml propidium iodide [Sigma-Aldrich cat#81845] in DPBS) with addition of 5 μl 20 mg/ml PureLink RNAse A [Life Technologies cat#12091-021] and 0.05% Triton-X-100 [Sigma-Aldrich cat#T8787] to a pellet of cells while gently vortexing and incubating for 60 min on ice. After incubation, cells were pelleted and resuspended in 1 ml PI staining buffer without Triton-X-100, run on a MoFlo cell sorter and finally analysed using Summit 4.3 software [Beckman Coulter].

High-throughput mapping of chromatin interactions via Hi-C

A modified version of the Hi-C protocol described in Sofueva et al. (2013) was used to map genome-wide chromatin interaction in an unbiased way. A range of different conditions were attempted to optimise the protocol for primary liver cells. The protocol that was finally applied to the Hi-C libraries can be summarised as follows (see **TABLE 5** for details about each sample):

Template preparation and quality controls:

Hepatocyte cytoplasms were lysed by incubating with Hi-C Lysis Buffer (10mM Tris-HCl pH8, 10mM NaCl, 0.2% Igepal CA-640, 1X Complete EDTA-Free protease inhibitors) for 30 min on ice and the nuclei were exposed. The sample was transferred to Protein LoBind Tubes [Eppendorf cat#022431081]. The nuclei were permeabilised with 0.1%-0.6% sodium dodecyl sulfate (SDS) [Fisher BioReagents cat#BP1311-1] in 500 µl 1X NEBuffer 2 [New England Biolabs cat#B7002], incubating for 1 h at 37ºC with 800 rpm shaking, and the reaction was subsequently quenched by adding Triton X-100 to a concentration of 0.67%-4% and incubating again for 1 h at 37ºC with 800 rpm shaking. The nuclei were digested for 24-72h with 1500 U HindIII [New England Biolabs cat#R0104] in 500 µl 1X NEBuffer 2. Digestion efficiency was assessed by running de-crosslinked sample aliquots taken before and after digestion on a 1% agarose gel and by qPCR amplification of the aliquots across randomly selected HindIII cutting sites. One half of the sample (the "Hi-C template") underwent fill-in of the digested ends with a biotin-labelled nucleotide. The other half of the sample was not filled-in and served as a "3C control" for fill-in and re-ligation efficiency checks. To change the buffer after digestion, the Hi-C template was quick-spun for 10 sec and washed once with 1ml 1X NEBuffer2. The fill-in reaction was set up in 200µl 1X NEBuffer 2 containing 15 µM of each dATP, dGTP, dTTP [Life Technologies cat#18252-015, 18254-011, 18255-018], 15 µM of Biotin-14-dCTP [Life Technologies cat#19518-018] and 25 U DNA Polymerase I, Large (Klenow) Fragment [New England Biolabs cat#M0210]. Reaction was incubated for 45 min at 37ºC, inverting the tube every 15 min to ensure proper mixing. Following quick-spinning for 10 sec and washing twice with 100 µl 1X T4 Ligase Buffer [New England Biolabs cat#B0202], both the Hi-C template and the 3C control were re-ligated by incubating overnight at 16ºC with 2000 cohesive-end U T4 DNA Ligase [New England Biolabs cat#M0202L] in 100 µl 1X T4 Ligase Buffer. Re-ligation was checked by running a de-crosslinked aliquot of the re-

ligated material on a 1% agarose gel. The sample was de-crosslinked by incubating overnight at 65ºC with 40 µl 10 mg/ml Proteinase K [Bioline cat#BIO-37-037] in 400µl 1X TE Buffer pH 8.0 (10 mM Tris-HCl pH 8.0 [Life Technologies cat#15568-025], 1 mM EDTA [Life Technologies cat#15575-038]). It was then incubated for 30 min at 37ºC with 10 µl 20 mg/ml PureLink RNAse A and purified by phenol/chloroform - chloroform extraction and ethanol precipitation. The sample was finally re-suspended in 100 µl Milli-Q water and quantified using the Qubit DNA HS assay kit [Life Technologies cat#Q32851] with a Qubit 1.0 fluorometer [Life Technologies cat#Q32857]. To assess fill-in efficiency, randomly selected ligation junctions were amplified via PCR in both the "3C control" and the "Hi-C template". The primers for ligation junction amplification are designed flanking HindIII cutting sites at known interacting loci or flanking two neighbouring but non-adjacent HindIII cutting sites. The PCR products were then digested using HindIII or NheI [New England Biolabs cat#R0131] and run on a pre-cast Novex 6% TBE polyacrylamide gel [Life Technologies cat#EC62652BOX].

	Fixation conditions	Permeabilization conditions	Digestion conditions
mouse 1	$1x10^7$ cells 10 min in 20ml FB (after homogenisation)	1h 0.1% SDS, 37°C, 800 rpm 1 h, 0.67% TX, 37°C, 800 rpm	20 h, 37°C, 800 rpm
mouse 2	$2x10^7$ cells 15 min in 20 ml FB (after homogenisation)	1h 0.6% SDS, 37°C, 800 rpm 1 h, 4% TX, 37°C, 800 rpm	20 h, 37°C, 800 rpm
rabbit 1	10 min 10% formalin (by immersion) $1x10^7$ cells 30 min in 20 ml FB (after homogenisation)	1h 0.1% SDS, 37°C, 800 rpm 1 h, 0.67% TX, 37°C, 800 rpm	20 h, 37°C, 800 rpm
rabbit 2	10 min 10% formalin (by immersion) $4x10^7$ cells 30 min in 20 ml FB (after homogenisation)	30 min 0.3% SDS, 65°C, 800 rpm 30 min 0.3% SDS, 37°C, 800 rpm 1 h, 2% TX, 37°C, 800 rpm	20 h, 37°C, 800 rpm
macaque 1	10 min 10% formalin (by immersion) $5x10^7$ cells 30 min in 20 ml FB (after homogenisation)	1h 0.3% SDS, 37°C, 800 rpm 1 h, 2% TX, 37°C, 800 rpm	48 h, 37°C, 800 rpm
macaque 2	10 min 10% formalin (by immersion) $6x10^7$ cells 30 min in 40 ml FB (after homogenisation)	30 min 0.3% SDS, 65°C, 800 rpm 30 min 0.3% SDS, 37°C, 800 rpm 1 h, 2% TX, 37°C, 800 rpm	20 h, 37°C, 800 rpm
dog 1	10 min 10% formalin (by immersion) $1x10^7$ cells 30 min in 20 ml FB (after homogenisation)	1h 0.1% SDS, 37°C, 800 rpm 1 h, 0.67% TX, 37°C, 800 rpm	20 h, 37°C, 800 rpm
dog 2	10 min 10% formalin (by immersion) $1x10^7$ cells 30 min in 20 ml FB (after homogenisation)	1h 0.1% SDS, 37°C, 800 rpm 1 h, 0.67% TX, 37°C, 800 rpm	48 h, 37°C, 800 rpm

FB= Fixation Buffer TX = Triton X-100 o/n = overnight

TABLE 5 – Hi-C protocol conditions used for each sample.

Library preparation:

The biotinylated nucleotides were removed from the un-ligated ends by splitting the Hi-C template in ~5µg aliquots and incubating each with 5U T4 DNA Polymerase [New England Biolabs cat#M0203], 100 µg/ml BSA [New England Biolabs cat#B9001], 100 µM dATP, 100 µM dGTP, in 100 µl 1X NEBuffer 2 for 2 h at 12ºC. The reaction was stopped by adding Ultrapure EDTA pH 8 [Life Technologies cat#15575-038] to a concentration of 10 mM. The aliquots were pooled back together, purified by phenol/chloroform – chloroform extraction (see above) and finally re-suspended in 100 µl Tris Low EDTA

Buffer (TLE: 10 mM Tris pH 8, 100 μM EDTA). The sample was quantified using the Qubit fluorometer. The sample was split into ~4μg aliquots that are diluted in 120μl TLE Buffer each, before sonication with Covaris S220 focused ultra-sonicator. Sonication parameters were: 20% Duty Cycle, 5 Intensity, 200 cycles/burst, 140 sec. The sonication profile was checked by loading 1 μl from one aliquot on a DNA 1000 kit [Agilent Technologies cat#5067-1504] and running on the Agilent 2100 Bioanalyzer. Most of the fragments should be in the 150-400 bp range. The fragment ends were repaired by incubating each sonicated aliquot in 168μl 1X T4 DNA Ligase Buffer containing 250 μM dNTP, 15 U T4 DNA Polymerase [New England Biolabs cat#M0203], 50 U T4 Polynucleotide Kinase (PNK) [New England Biolabs cat#M0201] and 5 U DNA Polymerase I, Large (Klenow) Fragment. The reaction was incubated for 30 min at 22ºC. The sample was purified using the MinElute PCR Purification kit [Qiagen cat#28004] using one column for each ~4 μg DNA aliquot. Each column was eluted in 30 μl TLE Buffer. 10 μl 10X Orange G were added to each sample aliquot and the whole sample is run on a 1.5% Ultrapure Agarose [Life Technologies cat#16500] TBE gel. The gel was stained with SYBR Green I nucleic acid gel stain [Life Technologies cat#S7563] and observed using a Dark Reader Transilluminator [Clare Chemical]. Fragments of size 200-400 bp were cut out of the gel, and purified using the Qiaquick Gel Extraction kit [Qiagen cat#28706]. Each column was eluted in 18 μl TLE Buffer, the aliquots were pooled and the sample was quantified using the Qubit fluorometer. Ligation junctions were purified via pull-down of the biotinylated nucleotides using 50 μl Dynabeads MyOne Streptavidin C1 [Life Technologies cat#65001] per 2.5 μg of DNA. The beads were first washed twice with 400 μl Tween Buffer (TWB: 5mM Tris-HCl pH 8, 500 μM EDTA pH 8, 1 M NaCl [Sigma-Aldrich cat#S3014], 0.05% Tween 20 [Bio-Rad cat#170-6531]), and subsequently incubated with the Hi-C template DNA in 360 μl Binding Buffer (BB: 5mM Tris-HCl pH 8, 500 μM EDTA pH 8, 1 M NaCl) for 1 h rotating at room temperature. The beads were then washed twice with BB, once with 1X NEBuffer 2 and finally re-suspended in 50 μl 1X NEBuffer 2. All washing steps were incubated 3 min rotating at room temperature. Adenines overhangs were added to the 3" end of the fragments by incubating the beads in 100 μl 0.5X NEBuffer 2 containing 100 μM dATP and 5 U Klenow (3'→ 5" exo minus) [New England Biolabs cat#M0212] for 45 min at 37ºC. Inserts were ligated to adaptor oligonucleotides for paired end sequencing. Illumina adapters were added to the beads in a ratio of 6 pmol/μg of DNA and were incubated with 1200 cohesive-end units of T4 DNA ligase in 50μl 1X T4 DNA Ligase Buffer. The reaction was left rotating at room

temperature overnight. The un-ligated adapters were then removed by washing several times with TWB, BB and 1X NEBuffer 2. Beads were finally re-suspended in 50 μl NEBuffer 2. A 10-15 cycle PCR reaction using the Herculase II fusion DNA polymerase [Agilent Technologies cat#600675] was set up to amplify the sample and to attach paired end sequencing primers (PE PCR primer 1.0 and 2.0) [Illumina] to the ends. The PCR reaction was then cleaned up using Agencourt AMPure beads [Beckman Coulter cat#A63880] and the sample was purified from the beads. 1 μl of the amplified sample was loaded on a DNA 1000 kit and run on an Agilent 2100 Bioanalyzer to check size and amount of the material. 75-100 base pairs from each end of the library inserts were sequenced using one lane on the Illumina Hi-seq platform per library.

Hi-C interaction matrix generation and domain calling

Sequencing reads were aligned to the mouse (mm10), rabbit (oryCun2), macaque (rheMac2) and dog (canFam3) genome assemblies using Bowtie 0.12.9 (Langmead et al., 2009). The parameters used for the alignment allowed a maximum of three mismatches and strictly one alignment per read. Processing of the aligned reads and normalisation of the interaction matrices were performed as previously described (Yaffe and Tanay, 2011; Sofueva et al., 2013; see also CHAPTER 1.2). Interaction matrices for each library were generated displaying seven different resolutions simultaneously (12,500, 25,000, 50,000, 100,000, 250,000, 500,000 and 1,000,000 bp).

Domains were identified and clustered as described in Sexton et al. (2012) with the modification that scaling factors were inferred using fends 100 - 400 kb apart, to account for the lower resolution of the mouse map compared to the *Drosophila* map. Domain borders were called using the 95% percentile of the scaling track as before (Sofueva et al., 2013). A domain-level map was partitioned into two clusters and clusters were assigned as passive/active according to LaminB MEF data, as before. For the rabbit, macaque and dog genome, the LaminB MEF track for mouse was lifted over to the corresponding genome to label domain clusters.

ChIP-seq analysis

CTCF ChIP-seq data for mouse, macaque and dog livers was obtained from Schmidt et al. (2012). Rad21 ChIP-seq data for mouse liver was obtained from (Faure et al., 2012). Mouse, macaque and dog ChIP-seq reads were mapped using bowtie (Langmead et al.,

2009) with parameters -n 2 -a -m 1 --best --strata. Alignment was followed by extension of sequenced tags to 300 bp fragments and pile-up into 50 bp bins. ChIP-seq coverage was normalised by computing the distribution of pile-up coverage on 50 bp bins, and transforming each coverage value v into $-\log_{10}(1\text{-quantile}(v))$; this latter values is referred to as "ChIP-seq signal" in this thesis. To define binding sites, a simple threshold was applied to the sum of values from two biological replicates for each CTCF dataset. Thresholding of CTCF ChIP-seq data was based on replicate correlation (**FIGURE 14A**) and signal-to-noise ratio of the tracks in the genome browser (**FIGURE 14B**). A replicate was not available for Rad21 ChIP data from mouse and this data was also thresholded. Thresholds used were as follows: mouse CTCF=2.2, macaque CTCF=2.4, dog CTCF=2.2, mouse Rad21=2.3. CTCF sites that were co-occupied by Rad21 were defined as having a CTCF ChIP-seq signal of >2.2 and a Rad21 ChIP-seq signal >2.3; however, to account for the increased difficulty in the identification of the lack of a protein and adopt a stringent approach, CTCF sites not co-occupied by Rad21 (**FIGURE 21**) were defined as having a CTCF ChIP-seq signal >2.2 and a Rad21 ChIP-seq signal <2. Application of more stringent or more lenient thresholds did not change the results. Binding site width was standardised at 200 bp and the ChIP-seq intensity for each site was calculated as the maximum value across the 200 bp. The relative distribution of CTCF within TADs (**FIGURE 16B**) was calculated as the distance of each CTCF site from the centre of its domain. Half the size of the domain was added to convert it to a measure of distance from the edge of the domain and this value was subsequently divided by the size of the domain.

Interspecies comparison of CTCF sites

Macaque and dog CTCF ChIP-seq libraries were converted to mouse genome coordinates using the liftOver tool from UCSC. To reduce the chance of inaccurate liftOver, a number of filters were implemented: sites within low mappability regions or repeats were excluded, as were sites with insufficient level of synteny within a window of 100 kb. To estimate mappability, each genome was broken into 50 bp bins and the whole-genome sequence was split into artificial reads and then mapped back to the genome. For each 50 bp bin, the mappability score was then defined to be the portion of artificial reads mapped uniquely to that bin. To estimate the level of synteny in the 100 kb around a CTCF site, the mappability tracks for macaque and dog were liftOver-ed to the mouse genome and all bins for which liftOver was not possible were converted to zeroes. The liftOver-ed tracks were

subsequently smoothed over 100 kb and CTCF sites falling in regions below the top quartile of such smoothed tracks were excluded from all subsequent analysis.

CTCF binding energy function

A CTCF DNA-binding energy function from the Cortex CTCF binding sites (ENCODE Cortex CTCF mouse, GSM769019, (Shen et al., 2012)) was used to profile all genomes for their similarity to the CTCF consensus motif. Given a set of genomic sites, we compute for each site the maximal energy value within a 200 bp window centered on the point.

Crossover analysis

Crossover analysis was performed as described in Sofueva et al., 2013. Briefly, for a given contact band, the average number of bias-corrected interactions is computed for a group of CTCF sites and the 250 kb upstream and downstream of them with a 1 kb sliding window. Each one of the averaged values is normalised for the local number of contacts to account for regional fluctuations in interaction density. Averaging multiple sites strongly increases the statistical robustness of the results, even at smaller bands, as each point of the resulting plot will represent thousands of interactions. The bands used were 5-7.5, 7.5-11.25, 10-15, 15-22.5, 20-30, 30-45, 40-60, 60-90 and 80-120 kb.

Distal Contact analysis

To calculate the average interaction profiles for a group of genomic landmarks, *Hind*III fragment ends were grouped into classes by associating each end with a genomic element located within 5 kb and then grouping all fragment ends associated with an element of the same class. For the mouse, macaque and dog genomes, three classes of CTCF sites (conserved, divergent-present, divergent-absent) and TSS sites were defined. These classes were further divided to sites within active or passive Hi-C domains. The remaining fragment ends (not classified given other landmarks) were defined as the background.

4C-seq

Preparation of 4C samples, libraries, sequencing analysis and normalisation were all performed as previously described (Sofueva et al., 2013). Hepatocyte cytoplasms were lysed by incubating with Hi-C Lysis Buffer (10mM Tris-HCl pH8, 10mM NaCl, 0.2% Igepal CA-640, 1X Complete EDTA-Free protease inhibitors) for 30 min on ice and the nuclei were exposed. The sample was transferred to Protein LoBind Tubes [Eppendorf

cat#022431081]. The nuclei were permeabilised with 0.1%-0.6% sodium dodecyl sulfate (SDS) [Fisher BioReagents cat#BP1311-1], incubating for 1 h at 37ºC with 800 rpm shaking, and the reaction was subsequently quenched with 0.67%-4% Triton X-100, incubating again for 1 h at 37ºC with 800 rpm shaking. The nuclei were digested for 48h with 750 U DpnII [New England Biolabs cat#R0543] in 500 µl 1X NEBuffer DpnII [New England Biolabs cat#B0543]. Digestion efficiency was assessed by running de-crosslinked sample aliquots taken before and after digestion on a 1% agarose gel. Samples were re-ligated by incubating overnight for two nights at 16ºC with 2000 cohesive-end U T4 DNA Ligase [New England Biolabs cat#M0202] in 100 µl 1X T4 Ligase Buffer [New England Biolabs cat#B0202]. Re-ligation was checked by running a de-crosslinked aliquot of the re-ligated material on a 1% agarose gel. The sample was de-crosslinked by incubating overnight at 65ºC with 40 µl 10 mg/ml Proteinase K [Bioline cat#BIO-37-037] in 400µl 1X TE Buffer pH 8.0 (10 mM Tris-HCl pH 8.0 [Life Technologies cat#15568-025], 1 mM EDTA [Life Technologies cat#15575-038]). It was then incubated for 30 min at 37ºC with 10 µl 20 mg/ml PureLink RNAse A and purified by phenol/chloroform - chloroform extraction and ethanol precipitation. The sample was finally re-suspended in 50 µl Milli-Q water and quantified using the Qubit DNA HS assay kit [Life Technologies cat#Q32851] with a Qubit 1.0 fluorometer [Life Technologies cat#Q32857]. A second digestion was set up using aliquots of 6µg of material and incubating them overnight with 120 U Csp6I [Thermo Scientific cat#ER0211] in 300 µl 1X Buffer B [Thermo Scientific cat#BB5] at 37ºC shaking at 300 rpm. Reaction was stopped by de-activating the enzyme at 65ºC for 20 min. Digestion was checked by running a de-crosslinked aliquot of the digested material on a 1% agarose gel. The DNA was subsequently purified through phenol/chloroform – chloroform extraction and re-suspended in 50 µl Milli-Q water. A second ligation was set up at this point in dilute conditions, incubating the samples with 1600 U T4 DNA Ligase [New England Biolabs cat#M0202] in 6 ml 1X T4 Ligase Buffer [New England Biolabs cat#B0202] overnight at 16 ºC.

The libraries thus obtained wera amplified via PCR using primers specific to the genome coordinates of interest (viewpoint). Primer sequences were chosen to viewpoint sites which were as close as possible to CTCF ChIP-seq peaks. Mouse primers were designed according to the genome-wide 4C primer database from van de Werken et al., 2012. For dog primers, a similar database was generated for the regions of interest. PCRs were performed using Expand High Fidelity PCR System [Roche cat#11732650001] for 30

cycles. The PCR reaction was cleaned up using Agencourt AMPure beads [Beckman Coulter cat#A63880] and the sample was purified from the beads. 1 µl of the amplified sample was loaded on a DNA 1000 kit and run on an Agilent 2100 Bioanalyzer to check size and amount of the material. Material from multiple viewpoints was subsequently pooled, diluted to 10 pM and sequenced using an Illumina MiSeq platform.

4C-seq primers used in this study

Mouse 4C-seq primers (mm10)			
Viewpoint	**CTCF peak**	**Reading primer**	**Non-reading primer**
Fig. 19A Mmus+/Cfam+ 1	chr10:94609250	CCATCTGTTTGAACAAGATC	CAAGAGAGAGTGGAAACAGG
Fig. 19A Mmus+/Cfam+ 2	chr10:94623583	AGTCAGATGGAATGCAGATC	CTAGATACAGCAATCAGCCC
Fig. 19A Mmus+/Cfam+ 3	chr10:94958324	ATTGCTTTCTCTGGTTGATC	AGTCACTCCTGCTCCTGTAA
Fig. 19A Mmus+/Cfam+ 4	chr10:94991353	GTTTCTGTTGGTTCACGATC	AAGCATTGTCCTACGTGATT
Fig. 19A Mmus+/Cfam-	chr10:95218005	CTACTCTGGCTTCTATGATC	CCCTTCCCTTCTATGTTTCT

Dog 4C-seq primers (canFam3)			
Viewpoint	**CTCF peak**	**Reading primer**	**Non-reading primer**
Fig. 19B Mmus+/Cfam+ 1	chr15:34606369	GCTCTTGCTCTAAACTGATC	TGGACCTCACCTCTCCTA
Fig. 19B Mmus+/Cfam+ 2	chr15:34596229	TGAGGTCCAGCAGAGATC	GTCGCATCACTTACTGGG
Fig. 19B Mmus+/Cfam+ 3	chr15:34269944	CTCCACTGAGCATTAAGATC	GCGGGATAGTTCTTTTCTCT
Fig. 19B Mmus+/Cfam+ 4	chr15:34244336	CTTATGTGCTCCTCCAGATC	AATCATATGCCTCCTCCTCT
Fig. 19B Mmus+/Cfam-	chr15:33989305	AAAGTAATCCCACCCAGATC	CTGAAGGAAACAACAATGTCA

REFERENCES

Alberts, B., Johnson, A., Lewis, J., Raff, M., & Roberts, K. (2008). *Molecular Biology of the Cell (5th edn.). New York, NY: Garland Science*.

Anderson, E., & Hill, R. E. (2014). Long range regulation of the sonic hedgehog gene. Current Opinion in Genetics & Development, 27, 54–59.

Anderson, E., Devenney, P. S., Hill, R. E., & Lettice, L. A. (2014). Mapping the Shh long-range regulatory domain. Development (Cambridge, England), 141(20), 3934–3943.

Bantignies, F., Roure, V., Comet, I., Leblanc, B., Schuettengruber, B., Bonnet, J., et al. (2011). Polycomb-dependent regulatory contacts between distant Hox loci in Drosophila. *Cell*, *144*(2), 214–226.

Becker, T. S., & Lenhard, B. (2007). The random versus fragile breakage models of chromosome evolution: a matter of resolution. *Molecular Genetics and Genomics : MGG*, *278*(5), 487–491.

Bell, A. C., & Felsenfeld, G. (2000). Methylation of a CTCF-dependent boundary controls imprinted expression of the Igf2 gene. *Nature*, *405*(6785), 482–485.

Bell, A. C., West, A. G., & Felsenfeld, G. (1999). The protein CTCF is required for the enhancer blocking activity of vertebrate insulators. *Cell*, *98*(3), 387–396.

Blanchette, M., Kent, W. J., Riemer, C., Elnitski, L., Smit, A. F. A., Roskin, K. M., et al. (2004). Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Research*, *14*(4), 708–715.

Bourque, G., Zdobnov, E. M., Bork, P., Pevzner, P. A., & Tesler, G. (2005). Comparative architectures of mammalian and chicken genomes reveal highly variable rates of genomic rearrangements across different lineages. *Genome Research*, *15*(1), 98–110.

Boveri, T. (1909). *Die Blastomerenkerne von Ascaris megalocephala und dieTheorie der Chromosomenindividualitit. Arch. Zellforschung. 3: 181-268*. Boveri1813Arch. Zellforschung.

Boyden, L. M., Lewis, J. M., Barbee, S. D., Bas, A., Girardi, M., Hayday, A. C., et al. (2008). Skint1, the prototype of a newly identified immunoglobulin superfamily gene cluster, positively selects epidermal gammadelta T cells. *Nature Genetics*, *40*(5), 656–662.

Brawand, D., Soumillon, M., Necsulea, A., Julien, P., Csárdi, G., Harrigan, P., et al. (2011). The evolution of gene expression levels in mammalian organs. *Nature*, *478*(7369), 343–348.

Cajal, S. R. (1903). Un sencillo metodo de coloracion seletiva del reticulo protoplasmatico y sus efectos en los diversos organos nerviosos de vertebrados e invertebrados. Trab. Lab. Invest. Biol.(Madrid).

Chen, X., Xu, H., Yuan, P., Fang, F., Huss, M., Vega, V. B., et al. (2008). Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. *Cell*, *133*(6), 1106–1117.

Ciosk, R., Shirayama, M., Shevchenko, A., Tanaka, T., Toth, A., & Nasmyth, K. (2000). Cohesin's binding to chromosomes depends on a separate complex consisting of Scc2 and Scc4 proteins. Molecular Cell, 5(2), 243–254.

Cremer, T., Cremer, C., Schneider, T., Baumann, H., Hens, L., & Kirsch-Volders, M. (1982). Analysis of chromosome positions in the interphase nucleus of Chinese hamster cells by laser-UV-microirradiation experiments. *Human Genetics*, *62*(3), 201–

209.

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

Defossez, P.-A., Kelly, K. F., Filion, G. J. P., Pérez-Torrado, R., Magdinier, F., Menoni, H., et al. (2005). The human enhancer blocker CTC-binding factor interacts with the transcription factor Kaiso. *The Journal of Biological Chemistry*, *280*(52), 43017–43023.

Dekker, J., Rippe, K., Dekker, M., & Kleckner, N. (2002). Capturing chromosome conformation. *Science (New York, NY)*, *295*(5558), 1306–1311.

Dixon, J. R., Selvaraj, S., Selvaraj, S., Yue, F., Kim, A., Li, Y., et al. (2012). Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, *485*(7398), 376–380.

Dong, X., Han, S., Zylka, M. J., Simon, M. I., & Anderson, D. J. (2001). A diverse family of GPCRs expressed in specific subsets of nociceptive sensory neurons. *Cell*, *106*(5), 619–632.

Donohoe, M. E., Silva, S. S., Pinter, S. F., Xu, N., & Lee, J. T. (2009). The pluripotency factor Oct4 interacts with Ctcf and also controls X-chromosome pairing and counting. *Nature*, *460*(7251), 128–132.

Donohoe, M. E., Zhang, L.-F., Xu, N., Shi, Y., & Lee, J. T. (2007). Identification of a Ctcf cofactor, Yy1, for the X chromosome binary switch. *Molecular Cell*, *25*(1), 43–56.

Dostie, J., Richmond, T. A., Arnaout, R. A., Selzer, R. R., Lee, W. L., Honan, T. A., et al. (2006). Chromosome Conformation Capture Carbon Copy (5C): a massively parallel solution for mapping interactions between genomic elements. Genome Research, 16(10), 1299–1309.

Dowen, J. M., Fan, Z. P., Hnisz, D., Ren, G., Abraham, B. J., Zhang, L. N., et al. (2014). Control of cell identity genes occurs in insulated neighborhoods in mammalian chromosomes. *Cell*, *159*(2), 374–387.

Duncan, A. W. (2013). Aneuploidy, polyploidy and ploidy reversal in the liver. *Seminars in Cell & Developmental Biology*, *24*(4), 347–356.

El-Kady, A., & Klenova, E. (2005). Regulation of the transcription factor, CTCF, by phosphorylation with protein kinase CK2. *FEBS Letters*, *579*(6), 1424–1434.

Enesco, H. E., & Samborsky, J. (1983). Liver polyploidy: influence of age and of dietary restriction. *Experimental Gerontology*, *18*(1), 79–87.

Engel, N., West, A. G., Felsenfeld, G., & Bartolomei, M. S. (2004). Antagonism between DNA hypermethylation and enhancer-blocking activity at the H19 DMD is uncovered by CpG mutations. *Nature Genetics*, *36*(8), 883–888.

Faure, A. J., Schmidt, D., Watt, S., Schwalie, P. C., Wilson, M. D., Xu, H., et al. (2012). Cohesin regulates tissue-specific expression by stabilizing highly occupied cis-regulatory modules. *Genome Research*, *22*(11), 2163–2175.

Fausto, N., & Campbell, J. S. (2003). The role of hepatocytes and oval cells in liver regeneration and repopulation. *Mechanisms of Development*, *120*(1), 117–130.

Filippova, G. N., Fagerlie, S., Klenova, E. M., Myers, C., Dehner, Y., Goodwin, G., et al. (1996). An exceptionally conserved transcriptional repressor, CTCF, employs different combinations of zinc fingers to bind diverged promoter sequences of avian and mammalian c-myc oncogenes. *Molecular and Cellular Biology*, *16*(6), 2802–2813.

Finlan, L. E., Sproul, D., Thomson, I., Boyle, S., Kerr, E., Perry, P., et al. (2008). Recruitment to the nuclear periphery can alter expression of genes in human cells. *PLoS Genetics*, *4*(3), e1000039.

Foster, H. A., & Bridger, J. M. (2005). The genome and the nucleus: a marriage made by evolution. Genome organisation and nuclear architecture. *Chromosoma*, *114*(4), 212–229.

Fraser, P., & Bickmore, W. (2007). Nuclear organization of the genome and the potential for gene regulation. *Nature*, *447*(7143), 413–417.

Geyer, P. K., Vitalini, M. W., & Wallrath, L. L. (2011). Nuclear organization: taking a position on gene expression. *Current Opinion in Cell Biology*, *23*(3), 354–359.

Gruber, S., Haering, C. H., & Nasmyth, K. (2003). Chromosomal cohesin forms a ring. *Cell*, *112*(6), 765–777.

Guelen, L., Pagie, L., Brasset, E., Meuleman, W., Faza, M. B., Talhout, W., et al. (2008). Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions. *Nature*, *453*(7197), 948–951.

Guo, C., Gerasimova, T., Hao, H., Ivanova, I., Chakraborty, T., Selimyan, R., et al. (2011). Two forms of loops generate the chromatin conformation of the immunoglobulin heavy-chain gene locus. *Cell*, *147*(2), 332–343.

Hadjur, S., Williams, L. M., Ryan, N. K., Cobb, B. S., Sexton, T., Fraser, P., et al. (2009). Cohesins form chromosomal cis-interactions at the developmentally regulated IFNG locus. *Nature*, *460*(7253), 410–413.

Hark, A. T., Schoenherr, C. J., Katz, D. J., Ingram, R. S., Levorse, J. M., & Tilghman, S. M. (2000). CTCF mediates methylation-sensitive enhancer-blocking activity at the H19/Igf2 locus. *Nature*, *405*(6785), 486–489.

Heger, P., Marin, B., Bartkuhn, M., Schierenberg, E., & Wiehe, T. (2012). The chromatin insulator CTCF and the emergence of metazoan diversity. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(43), 17507–17512.

Hirayama, T., Tarusawa, E., Yoshimura, Y., Galjart, N., & Yagi, T. (2012). CTCF is required for neural development and stochastic expression of clustered Pcdh genes in neurons. *Cell Reports*, *2*(2), 345–357.

Ho, S. Y. W., & Duchêne, S. (2014). Molecular-clock methods for estimating evolutionary rates and timescales. *Molecular Ecology*, *23*(24), 5947–5965.

Huis in 't Veld, P. J., Herzog, F., Ladurner, R., Davidson, I. F., Piric, S., Kreidl, E., et al. (2014). Characterization of a DNA exit gate in the human cohesin ring. Science (New York, NY), 346(6212), 968–972.

Iborra, F. J., Pombo, A., Jackson, D. A., & Cook, P. R. (1996). Active RNA polymerases are localised within discrete transcription "factories" in human nuclei. *Journal of Cell Science*, *109*(6), 1427–1436.

Irimia, M., Maeso, I., Roy, S. W., & Fraser, H. B. (2013). Ancient cis-regulatory constraints and the evolution of genome architecture. *Trends in Genetics : TIG*, *29*(9), 521–528.

Irimia, M., Tena, J. J., Alexis, M. S., Fernandez-Miñan, A., Maeso, I., Bogdanovic, O., et al. (2012). Extensive conservation of ancient microsynteny across metazoans due to cis-regulatory constraints. *Genome Research*, *22*(12), 2356–2367.

Ishihara, K., Oshimura, M., & Nakao, M. (2006). CTCF-dependent chromatin insulator is linked to epigenetic remodeling. Molecular Cell, 23(5), 733–742.

Jin, F., Li, Y., Dixon, J. R., Selvaraj, S., Ye, Z., Lee, A. Y., et al. (2013). A high-resolution map of the three-dimensional chromatin interactome in human cells. *Nature*, *503*(7475), 290–294.

Kazazian, H. H. (2004). Mobile elements: drivers of genome evolution. *Science (New York, NY)*, *303*(5664), 1626–1632.

Kikuta, H., Laplante, M., Navratilova, P., Komisarczuk, A. Z., Engström, P. G., Fredman, D., et al. (2007). Genomic regulatory blocks encompass multiple neighboring genes and maintain conserved synteny in vertebrates. *Genome Research*, *17*(5), 545–555.

Kim, T. H., Abdullaev, Z. K., Smith, A. D., Ching, K. A., Loukinov, D. I., Green, R. D., et al. (2007). Analysis of the vertebrate insulator protein CTCF-binding sites in the human genome. *Cell*, *128*(6), 1231–1245.

Kim, Y. J., Cecchini, K. R., & Kim, T. H. (2011). Conserved, developmentally regulated mechanism couples chromosomal looping and heterochromatin barrier activity at the homeobox gene A locus. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(18), 7391–7396.

Klenova, E. M., Chernukhin, I. V., El-Kady, A., Lee, R. E., Pugacheva, E. M., Loukinov, D. I., et al. (2001). Functional phosphorylation sites in the C-terminal region of the multivalent multifunctional transcriptional factor CTCF. *Molecular and Cellular Biology*, *21*(6), 2221–2234.

Krantz, I. D., McCallum, J., DeScipio, C., Kaur, M., Gillis, L. A., Yaeger, D., et al. (2004). Cornelia de Lange syndrome is caused by mutations in NIPBL, the human homolog of Drosophila melanogaster Nipped-B. *Nature Genetics*, *36*(6), 631–635.

Kueng, S., Hegemann, B., Peters, B. H., Lipp, J. J., Schleiffer, A., Mechtler, K., & Peters, J.-M. (2006). Wapl controls the dynamic association of cohesin with chromatin. Cell, 127(5), 955–967.

Kurukuti, S., Tiwari, V. K., Tavoosidana, G., Pugacheva, E., Murrell, A., Zhao, Z., et al. (2006). CTCF binding at the H19 imprinting control region mediates maternally inherited higher-order chromatin conformation to restrict enhancer access to Igf2. *Proceedings of the National Academy of Sciences of the United States of America*, *103*(28), 10684–10689.

Kuzmin, I., Geil, L., Gibson, L., Cavinato, T., Loukinov, D., Lobanenkov, V., & Lerman, M. I. (2005). Transcriptional regulator CTCF controls human interleukin 1 receptor-associated kinase 2 promoter. *Journal of Molecular Biology*, *346*(2), 411–422.

Lanctôt, C., Cheutin, T., Cremer, M., Cavalli, G., & Cremer, T. (2007). Dynamic genome architecture in the nuclear space: regulation of gene expression in three dimensions. *Nature Reviews. Genetics*, *8*(2), 104–115.

Langmead, B., Trapnell, C., Pop, M., & Salzberg, S. L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology*, *10*(3), R25.

Lanzuolo, C., Roure, V., Dekker, J., Bantignies, F., & Orlando, V. (2007). Polycomb response elements mediate the formation of chromosome higher-order structures in the bithorax complex. *Nature Cell Biology*, *9*(10), 1167–1174.

Lawrence, J. (1999). Selfish operons: the evolutionary impact of gene clustering in

prokaryotes and eukaryotes. Current Opinion in Genetics & Development, 9(6), 642–648.

Lee, J. A., Carvalho, C. M. B., & Lupski, J. R. (2007). A DNA replication mechanism for generating nonrecurrent rearrangements associated with genomic disorders. *Cell*, *131*(7), 1235–1247.

Lemaitre, C., Zaghloul, L., Sagot, M.-F., Gautier, C., Arneodo, A., Tannier, E., & Audit, B. (2009). Analysis of fine-scale mammalian evolutionary breakpoints provides new insight into their relation to genome organisation. *BMC Genomics*, *10*(1), 335.

Lieber, M. R. (2008). The mechanism of human nonhomologous DNA end joining. *The Journal of Biological Chemistry*, *283*(1), 1–5.

Lieberman-Aiden, E., van Berkum, N. L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., et al. (2009). Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science (New York, NY)*, *326*(5950), 289–293.

Liu, Z., Scannell, D. R., Eisen, M. B., & Tjian, R. (2011). Control of embryonic stem cell lineage commitment by core promoter factor, TAF3. *Cell*, *146*(5), 720–731.

Lobanenkov, V. V., Nicolas, R. H., Adler, V. V., Paterson, H., Klenova, E. M., Polotskaja, A. V., & Goodwin, G. H. (1990). A novel sequence-specific DNA binding protein which interacts with three regularly spaced direct repeats of the CCCTC-motif in the 5'-flanking sequence of the chicken c-myc gene. *Oncogene*, *5*(12), 1743–1753.

Lutz, M., Burke, L. J., LeFevre, P., Myers, F. A., Thorne, A. W., Crane-Robinson, C., et al. (2003). Thyroid hormone-regulated enhancer blocking: cooperation of CTCF and thyroid hormone receptor. *The EMBO Journal*, *22*(7), 1579–1587.

Ma, J., Zhang, L., Suh, B. B., Raney, B. J., Burhans, R. C., Kent, W. J., et al. (2006). Reconstructing contiguous regions of an ancestral genome. *Genome Research*, *16*(12), 1557–1565.

MacPherson, M. J., Beatty, L. G., Zhou, W., Du, M., & Sadowski, P. D. (2009). The CTCF insulator protein is posttranslationally modified by SUMO., *29*(3), 714–725.

Mahy, N. L., Perry, P. E., & Bickmore, W. A. (2002). Gene density and transcription influence the localization of chromatin outside of chromosome territories detectable by FISH. *The Journal of Cell Biology*, *159*(5), 753–763.

Majumder, P., & Boss, J. M. (2010). CTCF controls expression and chromatin architecture of the human major histocompatibility complex class II locus. *Molecular and Cellular Biology*, *30*(17), 4211–4223.

Majumder, P., Gomez, J. A., Chadwick, B. P., & Boss, J. M. (2008). The insulator factor CTCF controls MHC class II gene expression and is required for the formation of long-distance chromatin interactions. *The Journal of Experimental Medicine*, *205*(4), 785–798.

Mayer, R., Brero, A., Hase, von, J., Schroeder, T., Cremer, T., & Dietzel, S. (2005). Common themes and cell type specific variations of higher order chromatin arrangements in the mouse. *BMC Cell Biology*, *6*(1), 44.

Meister, P., Towbin, B. D., Pike, B. L., Ponti, A., & Gasser, S. M. (2010). The spatial dynamics of tissue-specific promoters during C. elegans development. *Genes & Development*, *24*(8), 766–782.

Meredith, R. W., Janečka, J. E., Gatesy, J., Ryder, O. A., Fisher, C. A., Teeling, E. C., et al. (2011). Impacts of the Cretaceous Terrestrial Revolution and KPg extinction on mammal diversification. *Science (New York, NY)*, *334*(6055), 521–524.

Messmer, S., Franke, A., & Paro, R. (1992). Analysis of the functional role of the Polycomb chromo domain in Drosophila melanogaster. *Genes & Development*, *6*(7), 1241–1254.

Michaelis, C., Ciosk, R., & Nasmyth, K. (1997).Ccohesins: chromosomal proteins that prevent premature separation of sister chromatids. *Cell*, *91*(1), 35–45.

Milne, L. S. (1909). The histology of liver tissue regeneration. *The Journal of Pathology and Bacteriology*, *13*(1), 127–160.

Mizuguchi, T., Fudenberg, G., Mehta, S., Belton, J.-M., Taneja, N., Folco, H. D., et al. (2014). Cohesin-dependent globules and heterochromatin shape 3D genome architecture in S. pombe. *Nature*, *516*(7531), 432–435.

Moen, P. T., Johnson, C. V., Byron, M., Shopland, L. S., la Serna, de, I. L., Imbalzano, A. N., & Lawrence, J. B. (2004). Repositioning of muscle-specific genes relative to the periphery of SC-35 domains during skeletal myogenesis. *Molecular Biology of the Cell*, *15*(1), 197–206.

Monahan, K., Rudnick, N. D., Kehayova, P. D., Pauli, F., Newberry, K. M., Myers, R. M., & Maniatis, T. (2012). Role of CCCTC binding factor (CTCF) and cohesin in the generation of single-cell diversity of protocadherin-α gene expression. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(23), 9125–9130.

Moon, H., Filippova, G., Loukinov, D., Pugacheva, E., Chen, Q., Smith, S. T., et al. (2005). CTCF is conserved from Drosophila to humans and confers enhancer blocking of the Fab-8 insulator. *EMBO Reports*, *6*(2), 165–170.

Moore, J. M., Rabaia, N. A., Smith, L. E., Fagerlie, S., Gurley, K., Loukinov, D., et al. (2012). Loss of maternal CTCF is associated with peri-implantation lethality of Ctcf null embryos. *PloS One*, *7*(4), e34915.

Moqtaderi, Z., Wang, J., Raha, D., White, R. J., Snyder, M., Weng, Z., & Struhl, K. (2010). Genomic binding profiles of functionally distinct RNA polymerase III transcription complexes in human cells. *Nature Structural & Molecular Biology*, *17*(5), 635–640.

Mouse Genome Sequencing Consortium, Waterston, R. H., Lindblad-Toh, K., Birney, E., Rogers, J., Abril, J. F., et al. (2002). Initial sequencing and comparative analysis of the mouse genome. Nature, 420(6915), 520–562.

Murphy, W. J., Larkin, D. M., Everts-van der Wind, A., Bourque, G., Tesler, G., Auvil, L., et al. (2005). Dynamics of mammalian chromosome evolution inferred from multispecies comparative maps. *Science (New York, NY)*, *309*(5734), 613–617.

Nadeau, J. H., & Taylor, B. A. (1984). Lengths of chromosomal segments conserved since divergence of man and mouse. *Proceedings of the National Academy of Sciences*, *81*(3), 814–818.

Nagano, T., Lubling, Y., Stevens, T. J., Schoenfelder, S., Yaffe, E., Dean, W., et al. (2013). Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. *Nature*, *502*(7469), 59–64.

Nakahashi, H., Kwon, K.-R. K., Resch, W., Vian, L., Dose, M., Stavreva, D., et al. (2013). A genome-wide map of CTCF multivalency redefines the CTCF code. *Cell Reports*, *3*(5), 1678–1689.

Nativio, R., Wendt, K. S., Ito, Y., Huddleston, J. E., Uribe-Lewis, S., Woodfine, K., et al. (2009). Cohesin is required for higher-order chromatin conformation at the imprinted

IGF2-H19 locus. *PLoS Genetics*, *5*(11), e1000739.

Neusser, M., Schubel, V., Koch, A., Cremer, T., & Müller, S. (2007). Evolutionarily conserved, cell type and species-specific higher order chromatin arrangements in interphase nuclei of primates. *Chromosoma*, *116*(3), 307–320.

Nizami, Z., Deryusheva, S., & Gall, J. G. (2010). The Cajal Body and Histone Locus Body. *Cold Spring Harbor Perspectives in Biology*, *2*(7), a000653–a000653.

Nora, E. P., Lajoie, B. R., Schulz, E. G., Giorgetti, L., Okamoto, I., Servant, N., et al. (2012). Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature*, *485*(7398), 381–385.

Noordermeer, D., de Wit, E., Klous, P., van de Werken, H., Simonis, M., Lopez-Jones, M., et al. (2011). Variegated gene expression caused by cell-specific long-range DNA interactions. Nature Cell Biology, 13(8), 944–951.

Ohno, S. (1973). Ancient Linkage Groups and Frozen Accidents. *Nature*, *244*(5414), 259–262.

Ohno, S., Wolf, U., & Atkin, N. B. (1968). Evolution from fish to mammals by gene duplication. *Hereditas*.

Ong, C.-T., & Corces, V. G. (2011). Enhancer function: new insights into the regulation of tissue-specific gene expression. *Nature Reviews. Genetics*, *12*(4), 283–293.

Ong, C.-T., & Corces, V. G. (2014). CTCF: an architectural protein bridging genome topology and function., *15*(4), 234–246.

Osborne, C. S., Chakalova, L., Brown, K. E., Carter, D., Horton, A., Debrand, E., et al. (2004). Active genes dynamically colocalise to shared sites of ongoing transcription. *Nature Genetics*, *36*(10), 1065–1071.

Pant, V., Kurukuti, S., Pugacheva, E., Shamsuddin, S., Mariano, P., Renkawitz, R., et al. (2004). Mutation of a single CTCF target site within the H19 imprinting control region leads to loss of Igf2 imprinting and complex patterns of de novo methylation upon maternal inheritance., *24*(8), 3497–3504.

Parelho, V., Hadjur, S., Spivakov, M., Leleu, M., Sauer, S., Gregson, H. C., et al. (2008). Cohesins functionally associate with CTCF on mammalian chromosome arms. *Cell*, *132*(3), 422–433.

Passarge, E., Horsthemke, B., & Farber, R. A. (1999). Incorrect use of the term synteny. *Nature Genetics*.

Pevzner, P., & Tesler, G. (2003). Human and mouse genomic sequences reveal extensive breakpoint reuse in mammalian evolution. *Proceedings of the National Academy of Sciences of the United States of America*, *100*(13), 7672–7677.

Phillips, J. E., & Corces, V. G. (2009). CTCF: master weaver of the genome. Cell, 137(7), 1194–1211.

Phillips-Cremins, J. E., Phillips-Cremins, J. E., Sauria, M. E. G., Sanyal, A., Gerasimova, T. I., Lajoie, B. R., et al. (2013). Architectural protein subclasses shape 3D organization of genomes during lineage commitment. *Cell*, *153*(6), 1281–1295.

Ptashne, M. (1986). Gene regulation by proteins acting nearby and at a distance. *Nature*, *322*(6081), 697–701.

Rabl, C. (1885). *Uber Zelltheilung. Morphol. Jahrb. 10, 214-330*. Rabl21410Morphol. Jahrb.

Rankin, S., Ayad, N. G., & Kirschner, M. W. (2005). Sororin, a substrate of the anaphase-promoting complex, is required for sister chromatid cohesion in vertebrates. Molecular Cell, 18(2), 185–200.

Rao, S. S. P., Huntley, M. H., Durand, N. C., Stamenova, E. K., Bochkov, I. D., Robinson, J. T., et al. (2014). A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping. *Cell*, *159*(7), 1665–1680.

Rhee, H. S., & Pugh, B. F. (2011). Comprehensive genome-wide protein-DNA interactions detected at single-nucleotide resolution. *Cell*, *147*(6), 1408–1419.

Ribeiro de Almeida, C., Stadhouders, R., de Bruijn, M. J. W., Bergen, I. M., Thongjuea, S., Lenhard, B., et al. (2011). The DNA-binding protein CTCF limits proximal Vκ recombination and restricts κ enhancer interactions to the immunoglobulin κ light chain locus. *Immunity*, *35*(4), 501–513.

Romanenko, S. A., Perelman, P. L., Trifonov, V. A., & Graphodatsky, A. S. (2012). Chromosomal evolution in Rodentia. *Heredity*, *108*(1), 4–16.

Rubio, E. D., Reiss, D. J., Welcsh, P. L., Disteche, C. M., Filippova, G. N., Baliga, N. S., et al. (2008). CTCF physically links cohesin to chromatin. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(24), 8309–8314.

Ruiz-Herrera, A., Castresana, J., & Robinson, T. J. (2006). Is mammalian chromosomal evolution driven by regions of genome fragility? *Genome Biology*, *7*(12), R115.

Ruiz-Herrera, A., Farré, M., & Robinson, T. J. (2012). Molecular cytogenetic and genomic insights into chromosomal evolution. *Heredity*, *108*(1), 28–36.

Sanyal, A., Lajoie, B. R., Jain, G., & Dekker, J. (2012). The long-range interaction landscape of gene promoters. *Nature*, *489*(7414), 109–113.

Sasaki, M., Lange, J., & Keeney, S. (2010). Genome destabilization by homologous recombination in the germ line. *Nature Reviews. Molecular Cell Biology*, *11*(3), 182–195.

Schmidt, D., Schwalie, P. C., Wilson, M. D., Ballester, B., Gonçalves, Â., Kutter, C., et al. (2012). Waves of retrotransposon expansion remodel genome organization and CTCF binding in multiple mammalian lineages. *Cell*, *148*(1-2), 335–348.

Schoenfelder, S., Sexton, T., Chakalova, L., Cope, N. F., Horton, A., Andrews, S., et al. (2010). Preferential associations between co-regulated genes reveal a transcriptional interactome in erythroid cells. *Nature Genetics*, *42*(1), 53–61.

Sexton, T., Yaffe, E., Kenigsberg, E., Bantignies, F., Leblanc, B., Hoichman, M., et al. (2012). Three-dimensional folding and functional organization principles of the Drosophila genome. *Cell*, *148*(3), 458–472.

Shen, Y., Yue, F., McCleary, D. F., Ye, Z., Edsall, L., Kuan, S., et al. (2012). A map of the cis-regulatory sequences in the mouse genome. *Nature*, *488*(7409), 116–120.

Shopland, L. S., Johnson, C. V., Byron, M., McNeil, J., & Lawrence, J. B. (2003). Clustering of multiple specific genes and gene-rich R-bands around SC-35 domains: evidence for local euchromatic neighborhoods. *The Journal of Cell Biology*, *162*(6), 981–990.

Shukla, S., Kavak, E., Gregory, M., Imashimizu, M., Shutinoski, B., Kashlev, M., et al. (2011). CTCF-promoted RNA polymerase II pausing links DNA methylation to splicing. *Nature*, *479*(7371), 74–79.

Sofueva, S., Yaffe, E., Chan, W.-C., Georgopoulou, D., Vietri Rudan, M., Mira-Bontenbal, H., et al. (2013). Cohesin-mediated interactions organise chromosomal domain architecture. *The EMBO Journal*, *32*(24), 3119–3129.

Spilianakis, C. G., & Flavell, R. A. (2004). Long-range intrachromosomal interactions in the T helper type 2 cytokine locus. Nature Immunology, 5(10), 1017–1027.

Splinter, E., Heath, H., Kooren, J., Palstra, R.-J., Klous, P., Grosveld, F., et al. (2006). CTCF mediates long-range chromatin looping and local histone modification in the beta-globin locus. *Genes & Development*, *20*(17), 2349–2354.

Stedman, W., Kang, H., Lin, S., Kissil, J. L., Bartolomei, M. S., & Lieberman, P. M. (2008). Cohesins localise with CTCF at the KSHV latency control region and at cellular c-myc and H19/Igf2 insulators. *The EMBO Journal*, *27*(4), 654–666.

Tanabe, H., Habermann, F. A., Solovei, I., Cremer, M., & Cremer, T. (2002a). Non-random radial arrangements of interphase chromosome territories: evolutionary considerations and functional implications. *Mutation Research*, *504*(1-2), 37–45.

Tanabe, H., Müller, S., Neusser, M., Hase, von, J., Calcagno, E., Cremer, M., et al. (2002b). Evolutionary conservation of chromosome territory arrangements in cell nuclei from higher primates. *Proceedings of the National Academy of Sciences*, *99*(7), 4424–4429.

Tolhuis, B., Palstra, R.-J., Splinter, E., Grosveld, F., & De Laat, W. (2002). Looping and interaction between hypersensitive sites in the active beta-globin locus. *Molecular Cell*, *10*(6), 1453–1465.

Toth, A., Ciosk, R., Uhlmann, F., Galova, M., Schleiffer, A., & Nasmyth, K. (1999). Yeast cohesin complex requires a conserved protein, Eco1p(Ctf7), to establish cohesion between sister chromatids during DNA replication. Genes & Development, 13(3), 320–333.

Uhlmann, F., Wernic, D., Poupart, M. A., Koonin, E. V., & Nasmyth, K. (2000). Cleavage of cohesin by the CD clan protease separin triggers anaphase in yeast. Cell, 103(3), 375–386.

van de Werken, H. J. G., Landan, G., Holwerda, S. J. B., Hoichman, M., Klous, P., Chachik, R., et al. (2012). Robust 4C-seq data analysis to screen for regulatory DNA interactions. *Nature Methods*, *9*(10), 969–972.

Vega, H., Waisfisz, Q., Gordillo, M., Sakai, N., Yanagihara, I., Yamada, M., et al. (2005). Roberts syndrome is caused by mutations in ESCO2, a human homolog of yeast ECO1 that is essential for the establishment of sister chromatid cohesion. *Nature Genetics*, *37*(5), 468–470.

Vernimmen, D., De Gobbi, M., Sloane-Stanley, J. A., Wood, W. G., & Higgs, D. R. (2007). Long-range chromosomal interactions regulate the timing of the transition between poised and active gene expression. The EMBO Journal, 26(8), 2041–2051.

Véron, A. S., Lemaitre, C., Gautier, C., Lacroix, V., & Sagot, M.-F. (2011). Close 3D proximity of evolutionary breakpoints argues for the notion of spatial synteny. *BMC Genomics*, *12*(1), 303.

Vinogradov, A. E., Anatskaya, O. V., & Kudryavtsev, B. N. (2001). Relationship of hepatocyte ploidy levels with body size and growth rate in mammals. *Genome / National Research Council Canada = Génome / Conseil National De Recherches Canada*, *44*(3), 350–360.

Vostrov, A. A., & Quitschke, W. W. (1997). The zinc finger protein CTCF binds to the

APBbeta domain of the amyloid beta-protein precursor promoter. Evidence for a role in transcriptional activation. *The Journal of Biological Chemistry*, *272*(52), 33353–33359.

Wallace, H., & Birnstiel, M. L. (1966). Ribosomal cistrons and the nucleolar organiser. *Biochimica Et Biophysica Acta (BBA) - Nucleic Acids and Protein Synthesis*, *114*(2), 296–310.

Wang, H., Maurano, M. T., Qu, H., Varley, K. E., Gertz, J., Pauli, F., et al. (2012). Widespread plasticity in CTCF occupancy linked to DNA methylation. *Genome Research*, *22*(9), 1680–1688.

Wendt, K. S., Yoshida, K., Itoh, T., Bando, M., Koch, B., Schirghuber, E., et al. (2008). Cohesin mediates transcriptional insulation by CCCTC-binding factor. *Nature*, *451*(7180), 796–801.

Weth, O., & Renkawitz, R. (2011). CTCF function is modulated by neighboring DNA binding factors. *Biochemistry and Cell Biology = Biochimie Et Biologie Cellulaire*, *89*(5), 459–468.

Wijgerde, M., Grosveld, F., & Fraser, P. (1995). Transcription complex stability and chromatin dynamics in vivo. Nature, 377(6546), 209–213.

Xiao, T., Wallace, J., & Felsenfeld, G. (2011). Specific sites in the C terminus of CTCF interact with the SA2 subunit of the cohesin complex and are required for cohesin-dependent insulation activity. *Molecular and Cellular Biology*, *31*(11), 2174–2183.

Xie, X., Mikkelsen, T. S., Gnirke, A., Lindblad-Toh, K., Kellis, M., & Lander, E. S. (2007). Systematic discovery of regulatory motifs in conserved regions of the human genome, including thousands of CTCF insulator sites. *Proceedings of the National Academy of Sciences of the United States of America*, *104*(17), 7145–7150.

Yaffe, E., & Tanay, A. (2011). Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterise global chromosomal architecture. *Nature Genetics*, *43*(11), 1059–1065.

Yao, H., Brick, K., Evrard, Y., Xiao, T., Camerini-Otero, R. D., & Felsenfeld, G. (2010). Mediation of CTCF transcriptional insulation by DEAD-box RNA-binding protein p68 and steroid receptor RNA activator SRA. *Genes & Development*, *24*(22), 2543–2555.

Yu, W., Ginjala, V., Pant, V., Chernukhin, I., Whitehead, J., Docquier, F., et al. (2004). Poly(ADP-ribosyl)ation regulates CTCF-dependent chromatin insulation. *Nature Genetics*, *36*(10), 1105–1110.

Yusufzai, T. M., Tagami, H., Nakatani, Y., & Felsenfeld, G. (2004). CTCF tethers an insulator to subnuclear sites, suggesting shared insulator mechanisms across species. *Molecular Cell*, *13*(2), 291–298.

Zhang, N., Kuznetsov, S. G., Sharan, S. K., Li, K., Rao, P. H., & Pati, D. (2008). A handcuff model for the cohesin complex. The Journal of Cell Biology, 183(6), 1019–1031.

Zhang, N., & Pati, D. (2009). Handcuff for sisters: a new model for sister chromatid cohesion. Cell Cycle (Georgetown, Tex.), 8(3), 399–402.

Zhao, S., Shetty, J., Hou, L., Delcher, A., Zhu, B., Osoegawa, K., et al. (2004). Human, mouse, and rat genome large-scale rearrangements: stability versus speciation. *Genome Research*, *14*(10A), 1851–1860.

Zhao, Z., Tavoosidana, G., Sjölinder, M., Göndör, A., Mariano, P., Wang, S., et al. (2006). Circular chromosome conformation capture (4C) uncovers extensive networks of epigenetically regulated intra- and interchromosomal interactions. Nature Genetics,

38(11), 1341–1347.

Zuckerkandl, E., & Pauling, L. (1962). *Molecular disease, evolution and genetic heterogeneity*.

Zuin, J., Dixon, J. R., van der Reijden, M. I. J. A., Ye, Z., Kolovos, P., Brouwer, R. W. W., et al. (2014). Cohesin and CTCF differentially affect chromatin architecture and gene expression in human cells. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(3), 996–1001.

ACKNOWLEDGEMENTS

The work presented in this thesis has been funded by Cancer Research UK, without whose generous contribution I would have probably died of starvation or exposure, not being able to afford a living in London out of my own savings. Thank you for believing in me.

This thesis would also not exist if it weren't for the help and the support of some extraordinary people, to whom I owe my gratitude.

I have to thank Sue first and foremost for letting me work in her group, taking on the responsability of a PhD student at the difficult time when she was just starting her lab. Sue has been the beacon of this doctoral work, always providing support, guidance and motivation. She created an ideal, relaxed lab environment in which ideas could easily thrive and I cannot be grateful enough for that.

My thank you goes also to Duncan Odom, who provided all the non-mouse samples and incepted this project together with Sue and me; Bianca Schmitt and Christina Ernst from Duncan's lab, who helped me with sequencing and some ChIP-seq samples; Stephen Henderson at the Bill Lyons Centre for Bioinformatics, who initiated me to CTCF motif analysis using MEME.

Thanks to Amos Tanay and the people in his lab, particularly Eitan Yaffe, Pedro Olivares and Yaniv Lubling, for their invaluable help in giving me an understanding of bioinformatics and Hi-C data analysis.

Dimitra, my inseparable travelling companion (or in her words "my twin") throughout my PhD, with whom I shared all of the (frequent) frustrations and (rare) joys of the doctoral experience. She has always been there to listen to and support me (in her very own "dramatic" way), even at times that were difficult for her, something that I deeply appreciate and that makes her a precious friend to me. Thank you, really.

I want to thank all my fellow members of the Hadjur lab. Chris, for the hours spent in front of the terminal trying to figure out the latest cryptical error and for his patience in answering all my "how to do this basic R thing?" questions; Sevil, for the countless times I consulted with her about my agarose gel profiles and for all her advice on Hi-C; Wen-Ching, for his support in the early days of my Hi-C analysis; José and Andrew for the help, the laughs and the good times we shared.

Aside from the science, my PhD has given me the opportunity to meet some great people that I am now happy to call my friends. Thank you Enrique, Manolo and Valenti for the beers, the laughs and the little big dramas of our Friday nights between the pub and the Living Room. Thanks to Kate and Sabine for all the talks and the fun we had together, and for your support in and out of the lab; to Valentina for the wonderful time we shared; to all the other people at CI that have made the time of my PhD a good one: you know who you are, and you have my thank you.

I have to thank Gabba and Sara, my flatmates, and Michela, our "almost-flatmate", with whom I shared four of the best years of my life, made of singing, cinema, good food and great times. Fellows Road will be in my heart forever. I also need to credit Michela for crafting the beautiful image you can see in the title page (Michela Papini – Designer – www.michelapapini.com).

Thanks to my "historical friends", the solid ground on which I walk every day: AGR, Picio, Prisco, Morelli, my friends from uni. Close or far I know that you're always there for me and I don't take that for granted.

Last but certainlty not least, I want to thank my family, my cousin Marco, my aunt Marina and especially my mum.

Mum, I know how difficult it has been for you when I left home and country, but despite that you have never failed to support my choices, even while you were so sad you couldn't stop the tears. I cannot express how important this is for me and how brave and strong I think you are. You are the best in the world and you are the unwaivering force of my life.

```
>: 4 + 8 + 15 + 16 + 23 + 42 = ...
```