Chromosome topology underlying factors:
studies on a model gene locus
and
an exemplary DNA looping protein

Sjoerd Holwerda

ISBN: 978-90-393-6026-2

Printing: Off Page
Cover: Picture was kindly provided by and reproduced with permission of 3Dimka.

The research described in this thesis was performed at the Hubrecht Institute of the Royal Dutch Academy of Arts and Sciences (KNAW), within the framework of the Graduate School of Cancer Genomics and Developmental Biology in Utrecht, The Netherlands.

Chromosome topology underlying factors: studies on a model gene locus and an exemplary DNA looping protein

Chromosoom topologie onderliggende factoren: studies aan een model gen locus en een prototype DNA looping eiwit (met een samenvatting in het Nederlands)

Proefschrift

ter verkrijging van de graad van doctor aan de Universiteit Utrecht
op gezag van de rector magnificus,
prof.dr. G.J. van der Zwaan,
ingevolge het besluit van het college voor promoties
in het openbaar te verdedigen

op

donderdag 10 oktober des ochtends te 10.30 uur

door

Sjoerd Johannes Bastiaan Holwerda
geboren op 28 februari 1981
te Utrecht

Promotor: Prof. dr. W.L. de Laat

Als je gaat, ver ga je niet
Je blijft altijd herinnerd worden
Op het netvlies, in het hart
Wie er gaat dat doet er niet

Toe maar, ga je gang en doe zoals
Je wilt, dan blijft het leven leuk
En interessant. Wat zoek je dan?
Het komt echt, hoe het je toevalt.

Veranderingen vinden plaats.
Zo was het in den beginne, zo zij het thans en voor immer
tot in de eeuwen der eeuwen.
Ongemerkt traag of snel als de wind

De vraag die brandt, het antwoord rookt
Wat is 't belang,
Van de weg die je loopt
De vlam of zijn walm,
De laatste, zal waaien met de wind

Contents

Scope of this thesis

Scope of this thesis

Topology of the genome plays a role in transcription regulation. Nuclear organization can be studied at the level of chromosomal positioning relative to transcriptional favorable or restrictive nuclear environments. At a smaller scale, transcriptional regulation can be studied at the level of regulatory elements that contact gene promoters to potentially enhance transcription. The relation between transcription and nuclear organization is described and studied in this thesis on both levels.

The relation between gene regulation and chromatin topology is described in **Chapter 1**. Here examples of prototype gene loci are used to explain that transcription can be regulated over large distances and that chromatin looping plays an important role in the regulation gene expression. The role for certain proteins involved in chromatin loops and the positioning of genes relative to the nuclear lamina are also discussed.

The role of a DNA binding protein, CTCF, in transcriptional regulation is highlighted in **Chapter 2**. The genome wide association of CTCF with regulatory elements connects this protein to gene regulation through mediating long range chromatin loops between such elements and genes. We discuss the distribution of CTCF binding, its association with other protein complexes and its potential role at boundary forming sequences.

The role of gene positioning for transcription of the IgH locus is investigated in **Chapter 3**. The results that we obtained confirmed some general nuclear organizational 'rules'. On the ohter hand, we show that the structure of the genome and transcription can be uncoupled. We have used an allele specific 4C-seq method to show the structural similarities and differences between two functionally different IgH alleles in B cells.

In **Chapter 4**, we describe a high-resolution 4C-seq method which can be used to identify interactions between regulatory regions and target promoters. This technique is applied in **Chapter 5**.

In **Chapter 5** we have analyzed the role of the earlier discussed DNA binding protein, CTCF. Analysis of CTCF binding and its ability to form chromatin loops in three different tissues sheds light on the complex role of CTCF in the mediating chromatin loops.

In **Chapter 6**, the relation between nuclear organization and transcription regulation is discussed. Here, the author of this thesis ventilates his view on the importance of nuclear organization for the function of our genome: deciphering the information that is encrypted in DNA and transforming this information into messenger RNA.

1
Chromatin loops, gene positioning and gene expression

Frontiers in Genetics 2012;3:217

1

Chromatin loops, gene positioning and gene expression

Sjoerd Holwerda[1], Wouter de Laat[1,*]

1. Hubrecht Institute-KNAW & University Medical Center Utrecht, Utrecht, The Netherlands
* Corresponding author: Email: w.delaat@hubrecht.eu

ABSTRACT

Technological developments and intense research over the last years have led to a better understanding of the three-dimensional structure of the genome and its influence on genome function inside the cell nucleus. We will summarize topological studies performed on four model gene loci: the alpha- (α) and beta-globin (β-globin) gene loci, the antigen receptor loci, the imprinted H19-Igf2 locus and the Hox gene clusters. Collectively, these studies show that regulatory DNA sequences physically contact genes to control their transcription. Proteins set up the three-dimensional configuration of the genome and we will discuss the roles of the key structural organizers CTCF and cohesin, the nuclear lamina and the transcription machinery. Finally, genes adopt non-random positions in the nuclear interior. We will review studies on gene positioning and propose that most inter-chromosomal DNA contacts will have little impact on the identity of cell populations but can be important for single cells.

Introduction

Only a few percent of the 3.2 billion basepairs of our genome is coding sequence. The remainder is intronic and inter-genic sequences, long considered to be junk DNA, but now realized to contain hundreds of thousands of sequence modules with the potential to regulate gene expression (1). This greatly outnumbers the ~25.000 genes that we carry in our genome. For the great majority of regulatory sites we do not know though whether they really exert a function in vivo and, if so, to which target gene they direct their activity. Studies into the shape of our genome provided evidence that regulatory DNA sequences can control transcription over distance by physically contacting target genes via chromatin looping. Initially such work was primarily done on individual gene loci. We will highlight findings on some of the most studied model gene systems, including the α- and β-globin gene loci, the immunoglobulin and other antigen receptor gene loci, the imprinted H19-Igf2 locus and the Hox gene clusters. Collectively, these studies showed how local DNA topology can change dynamically in time and place to accommodate developmental gene expression. It also uncovered some of the trans-acting factors that fold the chromatin. We will discuss the role of the nuclear lamina, CTCF, cohesin and RNA polymerase II (RNAPII), being currently the most intensively studied general organizers of chromosome topology. Collectively, all studies emphasize the relationship between genome structure and genome function. Consensus seems to have reached now for shape being crucial for function within the ~1 megabase (Mb) scale. Here, regulatory sequences need to physically get in contact with genes to control their transcription. Beyond this level of organization, it is not as obvious how relevant the nuclear position and/or genomic environment of genes will be. Studies manipulating the nuclear location of genes start to provide insight in this and will be discussed. Finally, we propose that the probabilistic nature of nuclear positioning implies that we need to move from cell population-based to single cell studies to understand how remote genomic sequences can influence each other's function.

Functionally relevant DNA interactions between genes and regulatory sequences

The realization that sequence information required for proper gene expression may sometimes reside at a large chromosomal distance away from the gene body came from observations in patients, showing that the deletion of sequences away from the β-globin genes proper caused thalassemia (2). For a long time, the mechanisms behind long range gene activation remained enigmatic. Although still not entirely understood it is now clear that it involves physical contacts between such remote regulatory sequences and the genes that they control. This discovery relied mostly on the development of chromosome conformation capture (3C) technology, a method invented ten years ago (3) that allows quantitative measurements of DNA contact frequencies between pairs of selected genomic sites. Here, we will highlight observations made by 3C technology on four gene clusters (the globin gene loci, the antigen receptor loci, the imprinted H19-Igf2 locus and the Hox gene loci) that serve as model systems for varying types of gene regulation.

The α- and β-globin loci

Early evidence for chromatin looping being involved in mammalian gene regulation comes from studies on the β-globin locus. This is perhaps unsurprising as the globin loci have always been the subject of intense gene expression studies: their misregulation underlies thalassemia and the α- and β-globin genes serve as model systems to study developmental gene regulation. As pointed out, the observation that the deletion of sequences away from, but not affecting, the genes proper caused thalassemia (4) first suggested that gene transcription was controlled by remote regulatory sequences. A series of remote regulatory sites were then demonstrated to exist in these loci, the most important ones in the β-globin locus collectively referred to as a Locus Control Region (LCR). The LCR controls expression of multiple β-globin genes which are arranged on the chromosome in order of their timed expression during development: embryonic β-globin genes are closest to and adult genes are furthest away from the LCR (**Figure 1a**). Proximity on the linear DNA template therefore clearly matters, but the exact mode of LCR action over distance long remained elusive. Three-dimensional proximity was implicated in transcription regulation when it was found that linear proximity is no longer important when two genes are positioned together at a large distance from the LCR (5,6). In 2002, first direct evidence for chromatin looping and spatial contacts between the LCR and an active β-globin gene was obtained, in studies using RNA TRAP (7) and 3C technology (8). 3C technology in particular appeared extremely useful for further investigations on the topology of the β-globin locus. The three-dimensional configuration of the β-globin locus was found to dynamically follow the changes in gene expression that occur during development and during red blood cell differentiation. LCR-gene contacts are not detectable in tissue where the globins are inactive. During development, the LCR switches its contacts from embryonic to adult β-globin genes to ensure their activation at the appropriate developmental stage (9). Proteins were shown to set up the chromatin loops in the locus. Transcription factors such as EKLF, GATA1 and Ldb1, that are important for proper globin gene expression and that bind to both the LCR and gene promoter regions, all appear necessary for stable LCR-gene interactions (10-12). Another transcription factor, CTCF, forms chromatin loops between binding sites surrounding the locus (**Figure 1a**). These CTCF-mediated loops precede LCR-gene contacts during red blood cell maturation (9). The spatial entity formed in red blood cells as a consequence of LCR-gene and CTCF-mediated DNA interactions was referred to as an Active Chromatin Hub (8).
An outstanding question is whether gene activity follows locus conformation or vice versa. The inhibition of transcription was found to not change the chromatin loops, suggesting that function follows structure in the β-globin locus (13,14). More direct evidence that transcriptional enhancement is a consequence of

looping has recently been provided. Ldb1 requires GATA1 for recruitment to the β-globin promoter, but binds to the LCR in a GATA1 independent manner. In an elegant assay employing artificial zinc fingers (ZF) in GATA1-null cells, the tethering of ZF-Ldb1 to the β-globin promoter was shown to induce LCR-gene contacts and chromatin looping, and to activate β-globin gene expression. Without the LCR, loops were absent and gene expression was not activated (15). This data supports the idea that looping towards target genes is crucial for distal enhancers to activate transcription. Interestingly, a truncated version of Ldb1 composed of only its self-association domain was already sufficient to induce chromatin looping and activate transcription initiation, suggesting that Ldb1 multimerization may stabilize contacts between remote globin DNA sequences.
Similar to the β-globin locus, the mammalian α-globin genes are controlled by distal enhancer elements (16-18). Active histone marks and erythroid-specific transcription factors are present at the locus before the occupancy by RNAPII is measurable (19), suggesting that there is a role for these factors in recruitment of RNA polymerases to the α-globin gene promoters. Looping of the key enhancer elements to the α-globin promoters, with intervening DNA sequences looping out, has been demonstrated (20,21). Timing of looping coincides with the binding of the pre-initiation complex and elongation factors (20). Protein factors like GATA1, Ldb1 and Sp/XKLF also bind to the α-globin genes and regulatory sequences, and can be expected to perform similar roles in chromatin looping and transcription regulation as seen for β-globin.

Antigen receptor gene loci

The immunoglobulin loci, which are active in B cells, and the T cell receptor (TCR) loci that are active in T cells, generally stretch over large chromosomal regions of up to 3 Mb and are subdivided into different regions (V, D, J and C) that each contain multiple gene segments. Particularly the V region is often extremely large. DNA rearrangement via V(D)J recombination is required to combine the different gene segments and assemble a functional antigen receptor that is unique in every B or T cell (22). The RAG proteins carry out V(D)J recombination and need to physically hold together two target sequences to cut and paste them together (23). The 3D topology of the antigen receptor loci therefore must play a role in their regulation. 3D FISH studies were originally performed to search for topological features of the recombining loci. Indeed it was shown that the two ends of the receptor loci spatially come together prior to rearrangement (24,25). The simultaneous visualization of intervening sequences then allowed demonstrating that locus contraction was not just a consequence of compaction but the result of chromatin looping, with intervening sequences looping out (26-28). Multiple proteins including Pax5, YY1, CTCF, cohesin, and ikaros have been implicated in the spatial organization of these gene loci. Initial evidence for this was based on the observation that their depletion reduced contraction of the locus and lead to altered usage of the V genes during recombination (26,27,29-31). More recently, 3C-based evidence was provided for looping between CTCF and cohesin bound chromatin sites across the antigen receptor loci (**Figure 1b**). Long-range chromatin interactions with three regulatory sequences in particular, the 3' regulatory region (3'RR), the Eμ-intronic enhancer and the recently discovered intergenic control region 1 (IGCR1), seem important for proper rearrangement of the IgH locus. These loops may facilitate the inclusion of distal V genes, thereby enhancing the diversity of choice in usage of coding V elements during V(D)J recombination (32-36). Additionally, CTCF and cohesin may regulate chromatin accessibility and transcription in suβ-regions of the loci, thereby directing the recombination machinery. As was pointed out, while multiple proteins that shape the conformation of the antigen receptor loci are known now, there is as yet no evidence that they act directly to promote synapsis between distal gene segments (37). Whether such activity exists, or whether the overall spatial structure of the antigen receptor loci is already sufficient to direct such interactions and warrant usage of the full repertoire of gene segments, remains to be investigated.

H19/Igf2 locus

The H19/Igf2 locus is an imprinted locus, with the H19 gene being expressed from the maternal and the Igf2 gene from the paternal allele. Both genes are under the control of a shared enhancer located on one side of the locus, 3' of the H19 gene. The targeting of this enhancer to either one of the genes is determined by an imprinting control region (ICR) located in between Igf2 and H19 (ICR) (38-41). This ICR, which contains multiple CTCF binding sites, is methylated when paternally inherited and unmethylated when derived from the mother (40,41). CTCF can only bind to the unmethylated, hence the maternally inherited, ICR (42,43).
Using an elegant approach that involved the site-specific integration of ectopic Gal-binding sites near the ICR it was shown that the ICR separates the H19 and the Igf2 gene in different chromatin compartments (44). Because of the distinct capacity to bind CTCF, ICR contacts differ between the alleles such that enhancers are enabled to contact the Igf2 gene on the paternal allele but not on the maternal allele (44). Subsequent studies based on 3C technology came to similar but not identical conclusions (45). Whereas one study reported bi-allelic interactions between the ICR and the enhancers (Kurukuti et al. 2006), another reported this interaction to be specific for the maternal allele. This study also showed that the CTCF-bound ICR promiscuously contacted enhancers and promoters, suggesting that such contacts are important for insulators to block effective enhancer-promoter communication (Yoon et al. 2007). In addition to its insulator function, the ICR appears required to initiate H19 gene expression: upon deletion of the four CTCF binding sites in the ICR, H19 transcripts were hardly detectable in the early embryo (Engel et al. 2006). In summary, studies on the H19/Igf2 locus confirm that gene competition for a shared enhancer involves competition for physical promoter-enhancer interactions (**Figure 1c**). Moreover, they show that insulators bound by CTCF can hamper this interaction, possibly by physically competing for these contacts.

3D Organization of the Hox genes

When it comes to developmental gene regulation, the Hox gene clusters are among the most fascinating gene clusters. In mammals, four of these clusters are present (HoxA-D), each containing roughly a dozen genes that are expressed during development in a temporal and spatial manner that is co-linear with their genomic context (46). The HoxD gene cluster, but also other Hox clusters, is flanked on both sides by large gene-poor chromosomal regions. The Hox genes encode for transcription factors and are important for body axis formation as well as proper formation of the extremities. Correct spatiotemporal expression along the body axis appears controlled within the gene cluster proper, independent of surrounding gene sequences. As was shown by 4C technology, here the genes show little specific interactions with surrounding sequences, but fold into a distinct active and inactive compartment. When moving posteriorly along the axis, the number of genes contained within the active compartment increases, in agreement with their progressive activation and corresponding change of histone modifications (47). It was suggested that this topological separation can mediate the temporal expression pattern of the HoxD genes. In the extremities, in this case the developing limb bud, a different mechanism of transcriptional control is in place, with a correspondingly different 3D conformation of the gene cluster. The HoxD genes depend on distinct long-range regulatory sequences for their expression in the proximal and distal parts of the limb bud (**Figure 1d**). These sequences are present in the gene-poor regions located on the telomeric and centromeric side of the gene cluster, respectively (48,49). The active, much more than the inactive, HoxD genes loop towards these sides to contact the regulatory DNA sequences. Based on the DNA contact profiles of the active HoxD13 gene, as generated by 4C technology, new enhancers were identified in the gene desert that showed correct

spatiotemporal reporter gene expression in transgenic mice (50). The emerging picture from these studies is that Hox gene expression in the limb bud is under the control of a complex regulatory landscape with many enhancers spread over hundreds of kilobases of flanking DNA working in concert (50). This picture seems confirmed by a recent high-resolution FISH study, which also revealed that further fine-tuning of the contacts between HoxD genes and flanking regulatory sequences takes place along the anterior-posterior axis of the limb bud (Williamson et al. 2012). A 5C analysis of the HoxA gene cluster in human primary fibroblasts taken from different anatomical sites revealed yet another dimension of Hox gene regulation. Contacts were identified with a site 5' of the cluster that expresses a long intergenic non-coding RNA

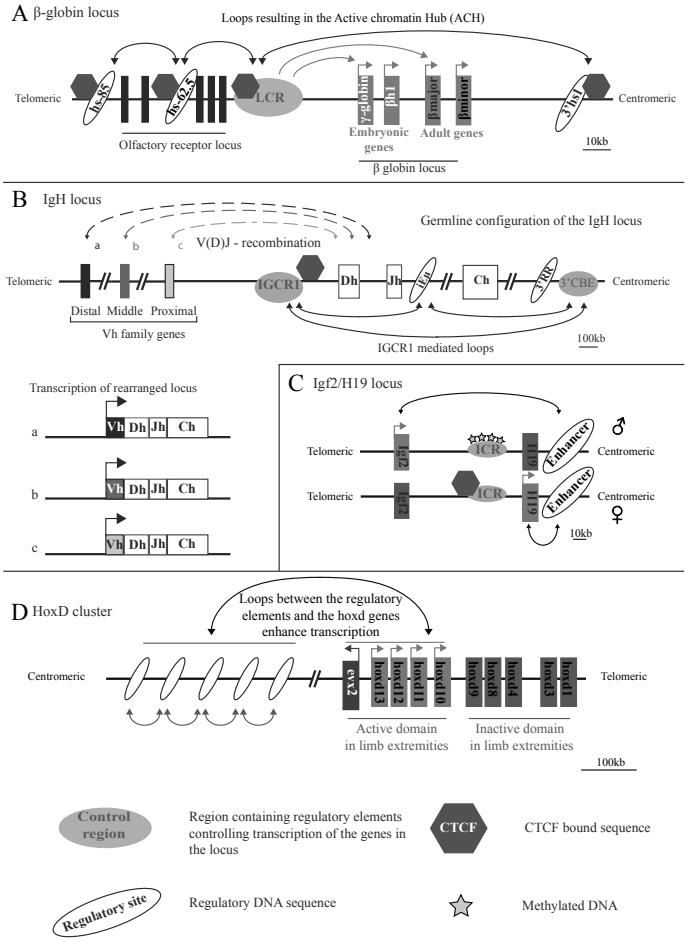


Figure 1. Long-range transcriptional regulation at model gene loci. (A) At the active β-globin locus, LCR-gene contacts and interactions between flanking CTCF sites set up an Active Chromatin Hub (ACH). (B) The IGCR1 contacts the 3' regulatory region and the intronic enhancer of the IgH locus in pro-B cells. Inclusion of the distal V-genes is influenced by the presence of the IGCR1. (C) CTCF blocks the interaction of the Igf2/H19 enhancer with the Igf2 gene on the maternal allele. Methylation of the ICR prevents CTCF binding and enables Igf2 expression from the paternal allele. (D) A 'regulatory archipelago' controls the expression of the hoxd13-hoxd10 genes over distance in limb extremities.

(lincRNA), named HOTTIP (51). HOTTIP RNA was reported to recruit proteins (WDR5) necessary to modify the histones and activate transcription of the genes contacted by the lincRNA locus (51). Thus, proper spatiotemporal Hox gene expression appears to be controlled by a very complex network of proximal and distal regulatory sequences that loop in a developmentally controlled manner towards specific Hox genes to physically confront them with activating protein and RNA molecules.

The overall shape of the 3D genome

The initial 3C studies discussed above focussed on individual genes and gene clusters, highlighting the functional importance of local chromatin loops and uncovering proteins that determine the topology of these gene loci (52). However, the genome is structurally organized also beyond the level of individual gene clusters. Original evidence that overall chromatin in the nucleus is not organized in a random fashion and that nuclear organization is related to transcriptional activity comes from microscopy observations. It showed the separation of densely packed inactive chromatin and loosely packed active chromatin and demonstrated that chromosomes occupy individual chromosome territories (53,54). It also demonstrated that larger chromosomes tend to occupy more peripheral positions in the nucleus, while smaller ones often reside more in the nuclear interior. A recurrent theme in nuclear organization is that folding and positioning follow probabilistic rules. Thus, a given chromosome will have a preferred nuclear position, but this does not imply that it occupies this exact position in every cell (55). In other words: all genomes in a population of cells can be expected to fold according to the same probabilistic rules, yet every single cell likely has a different genome structure. Thanks to the development of more genome-wide versions of 3C technology (56,57), the underlying, probabilistic, rules for genome folding are now rapidly being uncovered.

The most dominant force shaping the 3D genome seems the spatial separation between active and inactive chromatin. First observed under the microscope as a general feature of nuclear organization, it was then confirmed to also be relevant for the folding of individual chromosome segments (58) and, at much higher resolution, for the genomic environments of individual genes (59). The latter observation made by 4C technology for a few selected chromosomal sites was confirmed to apply to regions across the genome by recent Hi-C studies. In Hi-C, all versus all interactions of the genome are mapped, with the resolution of contact maps depending on the depth of sequencing, the size of the genome and the complexity of the sample analyzed (60-63). Hi-C studies showed that chromosomes are subdivided into topological domains that cover 0.2 - 1Mb. The domains mark chromosomal regions within which DNA contacts are confined. They generally demarcate regions with a defined gene density and activity, and with corresponding chromatin accessibility, histone modifications and replication timing. Preferred contacts among two types of topological domains are seen, the active and inactive topological domains, with the separation of active and inactive chromatin in the nucleus as a consequence (60-64). In Drosophila in particular, an additional domain type hallmarked by the association of polycomb group (PcG) proteins is observed, which also shows preferred contacts with other PcG-bound topological domains (65,66). Marks for active chromatin (DNase I sensitivity, H3K4me1 and –me3, RNAPII) were enriched for regions showing also interchromosomal DNA contacts (61,62), suggesting that open and active chromatin most easily reaches out of the chromosome territory. Boundaries of the domains were found enriched for CTCF, H3K4me1, transcriptional start sites (TSSs) and housekeeping genes, tRNA genes and SINE elements (61,63,65). Interestingly, during cellular differentiation the topological domains appear to largely remain intact and structural changes mostly occur within the domains, suggesting that the domain boundaries are largely conserved between cell types (63) (**Figure 2**). The active and inactive compartments each seem to organize themselves independently. This was shown in studies on the active and inactive X chromosome in mammalian female cells, where the inactive X chromosome showed normal contacts between active chromatin regions but was found to specifically lack long-range contacts between inactive chromatin domains. Interestingly, these latter contacts were restored when the non-coding RNA Xist, which coats the inactive X chromosome, was deleted, implicating a role also for non-coding RNA in chromosome topology (67).

Whether RNA plays a general role in the topological organization of chromosomes remains to be demonstrated. Proteins, however, are known to shape the configuration of the genome inside the cell.

Nuclear lamina proteins, CTCF, cohesin and RNAPII are best recognized as general organizers of the 3D genome and will be discussed below.

Proteins shaping the genome

Lamins and the nuclear periphery

The nuclear periphery of mammalian cells is known to be enriched for inactive chromatin and to correlate with relatively low gene expression levels (68-71). The inner part of the nuclear membrane is coated with a protein network called the nuclear lamina. Lamina-associated domains (LADs) were identified across the genome based on an elegant approach called DamID, which takes advantage of DNA adenine methylase (DAM) fused in this case to lamin B1, a component of the nuclear lamina. Chromosomal regions spanning 0.1 – 10 megabases were identified (LADs) that interact with the nuclear periphery (72). Characterization of the genomic content enriched in LADs showed that they are generally gene poor, transcriptionally inactive, depleted for active transcription marks such as RNAPII and active histone marks. At LAD borders, promoters transcribing away from LADs are found enriched, as well as CTCF binding sites (72). Dynamic interaction of the genome with the nuclear lamina was seen during neural differentiation of embryonic stem cells (ESCs). Some, but certainly not all, regions in the genome that were transcriptionally activated or repressed during this process changed their association to the nuclear lamina accordingly (73). Furthermore, mis-expressed genes were correlated with a change in nuclear localization of these genes in cells carrying disease related lamin A mutations (74). Recently, mapping of the lamin A-interacting genes showed that lamin A is similarly involved in anchoring silent genes to the nuclear lamina. Intriguingly though, depletion of lamin A changed the nuclear positioning of the lamin A bound genes but was not enough to change the

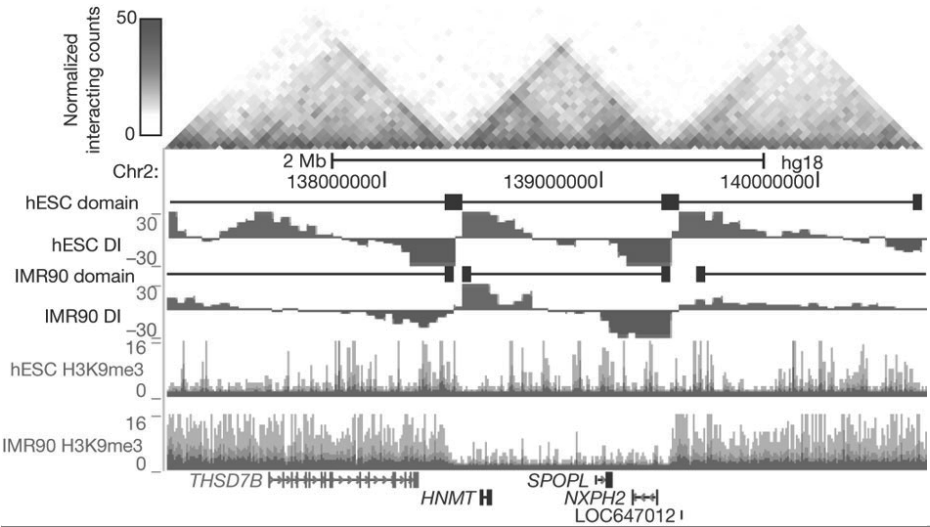


Figure 2. Topological boundaries can act as barriers for spreading of heterochromatin. The two-dimensional heat map shows the Hi-C interaction frequency in human ES cells. Underneath is indicated the directionality index (DI) in hESCs and IMR90 cells. The DI is a Hi-C measure showing a site's preference to engage in unidirectional contacts with downstream (red) or upstream (green) sequences. Borders of the topological domains are defined by a change in the directionality of interactions (transition from green to red). The UCSC Genome Browser shots show the distribution of H3K9me3, a measure for heterochromatin formation. Note that in IMR90 cells heterochromatin stops at the topological boundaries. Reprinted by permission from Macmillan Publishers Ltd: Nature (Dixon et al. 2012), copyright (2012)

expression of these genes (75). Oppositely, as discussed below, the artificial tethering of genes to the nuclear lamina sometimes, but not always, leads to their silencing. Clearly, the nuclear lamina is involved in the spatial organization of the genome in a manner that at least reflects transcriptional activity. To what extent a peripheral positioning also determines gene activity still remains to be investigated.

CTCF

CTCF is probably the best characterized structural organiser of the genome to date. From the first description of the protein (76), it has been shown to be a versatile protein having direct transcriptional effects (77-79) as well as effects on transcription over distance (80). The approximately 40,000 CTCF binding sites in the human and murine genome preferentially locate to intergenic regions and show high conservation between different cell types (81-84). CTCF is ubiquitously expressed and an essential protein (85). It has a well established role in chromatin folding at the β-globin locus, and in chromatin folding and gene expression at the H19/Igf2 locus and the antigen receptor loci, as described above. Also at other loci, including the human major histocompatibility complex class II locus and the Kcnq5 gene, CTCF-mediated chromatin loops were found involved in gene regulation (86-88). At a more genome-wide scale, CTCF binding sites were found enriched at borders between the topological domains identified by Hi-C (61,63) as well as at LAD borders (72), further hinting at an important role for this protein in organizing the 3D structure of chromosomes. Interest in the protein was raised even further when cohesin was found to co-occupy genomic sites with, and be positioned by, CTCF (see below) (89-91).
ChIA-PET is a technology that combines chromatin immunoprecipitation (ChIP) with a 3C approach, to direct DNA topology studies specifically to the genomic sites that are bound by a protein of interest (92). ChIA-PET was applied to CTCF to study its DNA interactome (93). Mostly intrachromosomal and a few interchromosomal interactions between CTCF bound sequences were identified, with the intrachromosomal loop sizes ranging from 10-200 kb. The loops appeared to serve different purposes (**Figure 3**). They can isolate an active chromatin region from surrounding inactive chromatin or bring together enhancers and promoters in a single loop. Yet other loops formed by CTCF seem to isolate undefined chromatin from a flanking active and inactive chromosomal region (93). Only a few percent of the total number of CTCF sites was found engaged in loop formation. This suggests that ChIA-PET only uncovers the tip of the topological iceberg. Alternatively, the majority of CTCF-bound sites is not involved in long-range chromatin interactions. If the latter is true, it would be interesting to understand what determines whether a CTCF binding site is engaged or not in a chromatin loop.

Cohesin

Cohesin is a multiprotein complex that forms a ring-like structure which captures and holds together the two DNA double-strand helices of sister chromatids after DNA replication. The discovery that cohesin binds to CTCF binding sites also in G1 phase of the cell cycle suggested that it has an additional role besides keeping sister chromatids together. Without CTCF, cohesin still binds to chromatin but is no longer found at specific locations along the chromosome arms, suggesting that CTCF positions cohesin on the chromatin (89-91). Given its shape and function, cohesin was obviously considered an attractive protein for chromatin loop formation (94). Indeed, cohesin was found to mediate chromatin looping at CTCF binding sites in several loci including the immunoglobulin locus (31), the interferon gamma locus (95), the HoxA locus (96), the MHC class II locus (97), the β-globin locus (84,98) and the H19/Igf2 locus (99). Interestingly, at several sites bound by CTCF across different cell-types, cohesin association was found to

differ in a cell-dependent manner, with topological changes and altered gene expression changing accordingly (96,98). This suggests that possibly the co-recruitment of additional factors like cohesin determines whether a given CTCF binding site is engaged in a chromatin loop in a given cell type. A CTCF-independent role for cohesin in transcription regulation was also demonstrated, in a study that revealed cohesin and estrogen receptor co-binding near upregulated genes upon estrogen treatment of MCF-7 cells (100). Cohesin binding was enriched at sites demonstrated by ChIA-PET to form ER-mediated loops (92), suggesting that cohesin may help ER to mediate transcriptional responses via long-range DNA interactions (100). A further CTCF-independent role of cohesin was observed in embryonic stem cells, where cohesin association was detected at sites bound by mediator and RNAPII, but not CTCF (101).

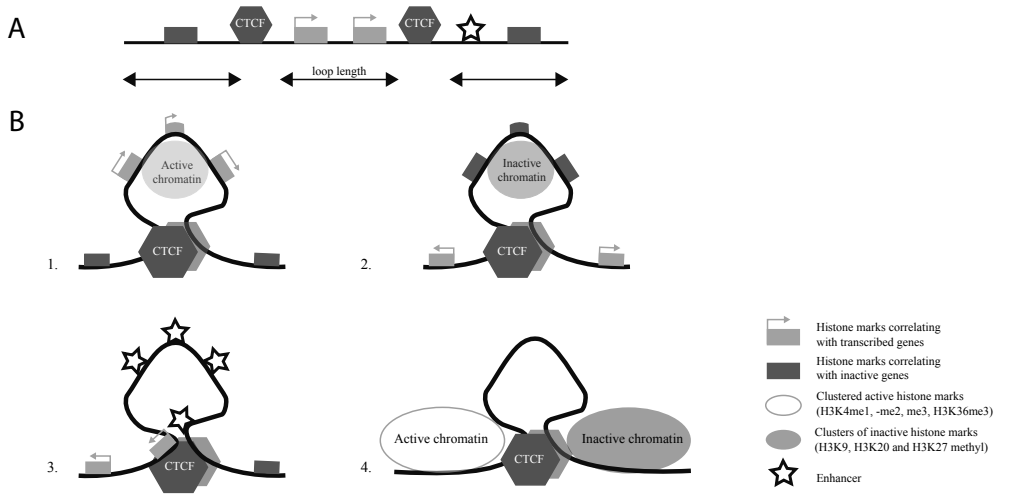


Figure 3. CTCF flanks chromatin marked by specific histone modifications. (A) Linear representation of a chromosomal region with active and inactive genes, CTCF binding sites and an enhancer (for explanation of symbols, see bottom figure). (B) ChIA-PET reveals different chromatin loops formed by CTCF (Handoko et al. 2011): CTCF loops demarcate regions (1) with active chromatin marks, (2) with inactive chromatin marks, (3) with enhancers and promoters, (4) with undefined chromatin surrounded by regions with opposing chromatin signatures.

Enhancer promoter interactions of tissue specific genes were shown by 3C technology to be mediated by the interaction with mediator and the cohesin loading factor, Nipbl. Cohesin and mediator together share distinct genomic sites in different tissues, unlike the shared binding sites between CTCF and cohesin which seem largely conserved between cell types (101). Thus, cohesin may have CTCF-dependent and –independent roles in chromosome topology and gene regulation during development (100,101).

RNA pol II

Transcription, and in particular the nuclear localization of RNA polymerase, has always been considered an attractive candidate to shape the three-dimensional genome (102). It may explain why active chromatin comes together in the nuclear space. Clusters of RNAPII, termed transcription factories, have been identified in the nucleus by electron microscopy and immunofluoresence (103-106). It is difficult to assess the number of factories per cell as this appears to differ between cell-types and is also dependent on the microscopy method used (107). The concept assumes that genes need to migrate to pre-existing protein factories where multiple genes are transcribed simultaneously. In a more extreme model there may even be dedicated transcription factories that contain specific combinations of transcription factors and therefore

need to be visited by defined categories of co-regulated genes (108,109). Does form indeed follow function, as suggested by these models? Not all observations necessarily support this idea. Live cell imaging with fluorescently tagged RNAPII so far has not provided convincing evidence for the existence of transcription factories (110,111), nor for movement of genes upon transcriptional activation (111). Inhibition of transcription caused most RNA polymerase to dissociate from active genes, yet had no appreciable impact on their contacts with other active genes, as assessed by 4C technology, nor interfered with enhancer-gene contacts (14). The recent demonstration that loop formation in the β-globin locus precedes transcriptional activation also suggests that function follows form (15). Possibly, shape and function both influence each other. It was proposed that initiating RNA polymerases that are close together in the nuclear space may aggregate to form the observed transcription factories. This is easiest envisioned to happen between genes that are proximal on the linear chromosome, as these per definition are close together in the nuclear space, rather than involving genes searching for distant co-regulated genes (112). Indeed, a ChIA-PET study focussing on chromatin loops formed between RNAPII-bound chromatin sites recently demonstrated the clustering of active gene promoters that neighbour each other on the chromosomes (113).
ChIA-PET enables an unbiased genome-wide assessment of contacts formed by the genomic sites bound by a protein of interest. Remarkably, for all proteins studied so far, ChIA-PET primarily identifies local contacts between sites close together on the linear chromosome. On the one hand this probably emphasizes the importance of local chromatin loops for the expression of genes involved in these loops. On the other hand it raises the question: how important is the position of a gene relative to other chromosomal regions elsewhere in the genome? So far, mostly microscopy studies have tried to address this.

Gene positioning in the cell nucleus

One of the earliest studies that followed the positioning of individual genes focussed on the Ikaros proteins, required for the development of cells of the lymphoid lineage (68,69). Highly expressed lymphoid genes like CD45 and CD19 were not found associated with Ikaros in B cells, but stage-specific genes showed differential association with Ikaros during differentiation (68). When bound by Ikaros, these genes were found to be silenced and repositioned to pericentromeric heterochromatin (PCH). It was proposed that PCH-association facilitated heritable gene silencing during B cell differentiation (68,69). Subsequently, also other genes were found to occupy particular nuclear locations in relation to their status of transcription, and again this has been studied most notably for the forementioned model gene loci. The IgH locus, for example, was found to adopt a peripheral position in cells not transcribing the gene. When active in B cells, it adopts a more internal nuclear position (24). In mature B cells, the non-productive IgH allele as reported to be frequently associated with PCH, perhaps to ensure its silencing (26,70). Repositioning of loci to PCH is also important during lineage choice in T cells (114,115), where repositioning of the CD8 locus to PCH is seen in $CD4^+$ T cells and vice versa. Here localization was stated to be predictive for the developmental state of the T cell (114). Localization of inactive genes to the nuclear periphery was also found for the human CFTR locus (71,116) and the casein cluster in mammary glands (117).
Similar observations were done on the β-globin locus. During erythroid maturation, which is accompanied by LCR-mediated transcriptional activation, the locus was observed to move from the periphery to the interior. Expression at the periphery was found, but it occurred more frequently in the nuclear interior, and the inward movement was dependent on the LCR (118). Whereas one study reported preferred clustering of the active β-globin genes with other active erythroid genes (108), two other studies did not find this (59,119). A different type of movement was observed for the Hox gene clusters. Induction of Hox gene expression influenced the position of the Hoxb1 and Hoxb9 genes relative to their chromosome territories (CTs) (120). Expression was associated with a position more outside of the chromosome territory. This

nuclear organization was dynamic as hoxb1 and –b9 could be repositioned in different stages of differentiation, in agreement with their transcriptional state (120,121). Similarly, Hoxd genes were looped outside their CT in the tailbud of e9.5 mice (122). In the forelimb bud, where Hoxd9 is also expressed (123), no looping out of the CT for this gene is found (122). Moreover, neighbouring genes that are dragged along outside the CT not necessarily show bystander upregulation of gene expression (124,125). Thus, these studies show that genes can, but don't need to move away from their CT and that looping out of the CT is not sufficient for gene activation.

To better understand the consequences of nuclear repositioning, tethering experiments can be done. These are based on the genomic integration of repeats of DNA binding sites (often bacterial LacO or TetO sequences) and the simultaneous expression in eukaryotic cells of cognate bacterial proteins (LacR or TetR) fused to a protein of interest. Fusion to fluorescent GFP enables following the genomic integration sites in live cell imaging studies (126,127) and revealed that individual gene loci show limited movement during the interphase of mammalian cells (128). Recruitment of transcriptional activators caused locus decondensation concomitant with increased transcription and histone acetylation, but neither was required to maintain the decondensed chromatin state (127,129-131). The targeting of heterochromatin protein 1 (HP1) to a non-heterochromatic locus reduced gene expression, induced locus condensation and resulted in local H3K9me3 modifications, indicative of heterochromatin formation (132,133).

Several studies used fusions of lamina components to address the consequences of recruitment to the nuclear periphery. In one study, which also enabled simultaneous visualization of nascent transcripts, the association of lamin B1 to a reporter locus caused repositioning, but only after cell division. Here, the kinetics of gene activation were similar to that at internal locations, indicating that loci maintain their transcriptional competence at the nuclear periphery (134). In another study however, repositioning through the recruitment of emerin (EMD) was found to be accompanied by reporter gene silencing (135). A third study measured chromosome-wide gene expression differences after tethering of the chromosome to the inner nuclear membrane. A few genes, some nearby and some at great distance from the integrated LacO cassettes, showed repressed transcription, but expression was not incompatible with peripheral location (136). Interestingly, in a recent study it was demonstrated that the ectopic integration of LAD sequences can also reposition surrounding chromosomal regions to the periphery, and negatively influences the expression of surrounding genes (137)/. GAGA motifs were found enriched in LADs and demonstrated to be responsible for peripheral recruitment. They are targets for the transcriptional repressor cKrox and the associated HDAC3 and Lap2b proteins, which were found to be necessary for peripheral recruitment (137). Collectively, these studies suggest that nuclear compartmentalization and gene expression are coupled, but also emphasize the probabilistic nature of nuclear organization: genes positioned at the periphery of the cell nucleus do not necessarily lose their capacity to be transcribed, but appear more susceptible to transcriptional repression than at more internal nuclear positions.

Concluding remarks

Over the last years research has made major progress in understanding the relationship between structure and function of the genome. Studies on model gene systems such as those discussed here have shown that local DNA interactions between regulatory sites and genes are important for transcriptional control. In mammals, such regulatory interactions can take place over chromosomal distances as large as a megabase. Transcription factors bound to these chromatin sites seem responsible for setting up the chromatin loops in chromosomal segments. Others, such as CTCF, appear capable to modify chromatin topology such that it hampers these interactions. Beyond this local scale of structural organization, genome folding seems to follow more probabilistic rules. Active and inactive chromatin separate, some chromosomal regions have an

increased chance of being at the periphery than others, and, when assayed across large cell populations, all individual gene loci appear to have many different contact partners. Together this suggests that the exact genome conformation will differ from cell to cell. As a consequence, a given contact between two dispersed genomic regions will only occur in a subset of cells. If this contact influences the expression of the associated genes, this may not have an impact on the entire cell population, but can be important for the individual cells involved, as was shown recently (138). To study the functional consequences of cell to cell differences in genome conformation we therefore probably need to analyse form and function at the single cell level, with the exciting possibility to discover that the overall shape of our genome can determine cell fate decisions of individual cells.

Acknowledgements

This work was financially supported by grant no. 935170621 from the Dutch Scientific Organization (NWO) and a European Research Council Starting Grant (209700, '4C') to WdL.

References

1. Shen, Y., Yue, F., McCleary, D.F., Ye, Z., Edsall, L., Kuan, S., Wagner, U., Dixon, J., Lee, L., Lobanenkov, V.V. *et al.* (2012) A map of the cis-regulatory sequences in the mouse genome. *Nature*.
2. Kleinjan, D.A. and van Heyningen, V. (2005) Long-range control of gene expression: emerging mechanisms and disruption in disease. *Am J Hum Genet*, **76**, 8-32.
3. Dekker, J., Rippe, K., Dekker, M. and Kleckner, N. (2002) Capturing chromosome conformation. *Science*, **295**, 1306-1311.
4. Van der Ploegh, L.H., Konings, A., Oort, M., Roos, D., Bernini, L. and Flavell, R.A. (1980) gamma-beta-Thalassaemia studies showing that deletion of the gamma- and delta-genes influences beta-globin gene expression in man. *Nature*, **283**, 637-642.
5. Hanscombe, O., Whyatt, D., Fraser, P., Yannoutsos, N., Greaves, D., Dillon, N. and Grosveld, F. (1991) Importance of globin gene order for correct developmental expression. *Genes Dev*, **5**, 1387-1394.
6. Dillon, N., Trimborn, T., Strouboulis, J., Fraser, P. and Grosveld, F. (1997) The effect of distance on long-range chromatin interactions. *Mol Cell*, **1**, 131-139.
7. Carter, D., Chakalova, L., Osborne, C.S., Dai, Y.F. and Fraser, P. (2002) Long-range chromatin regulatory interactions in vivo. *Nat Genet*, **32**, 623-626.
8. Tolhuis, B., Palstra, R.J., Splinter, E., Grosveld, F. and de Laat, W. (2002) Looping and interaction between hypersensitive sites in the active beta-globin locus. *Mol Cell*, **10**, 1453-1465.
9. Palstra, R.J., Tolhuis, B., Splinter, E., Nijmeijer, R., Grosveld, F. and de Laat, W. (2003) The beta-globin nuclear compartment in development and erythroid differentiation. *Nat Genet*, **35**, 190-194.
10. Drissen, R., Palstra, R.J., Gillemans, N., Splinter, E., Grosveld, F., Philipsen, S. and de Laat, W. (2004) The active spatial organization of the beta-globin locus requires the transcription factor EKLF. *Genes Dev*, **18**, 2485-2490.
11. Vakoc, C.R., Letting, D.L., Gheldof, N., Sawado, T., Bender, M.A., Groudine, M., Weiss, M.J., Dekker, J. and Blobel, G.A. (2005) Proximity among distant regulatory elements at the beta-globin locus requires GATA-1 and FOG-1. *Mol Cell*, **17**, 453-462.
12. Song, S.H., Hou, C. and Dean, A. (2007) A positive role for NLI/Ldb1 in long-range beta-globin locus control region function. *Mol Cell*, **28**, 810-822.
13. Mitchell, J.A. and Fraser, P. (2008) Transcription factories are nuclear subcompartments that remain in the absence of transcription. *Genes Dev*, **22**, 20-25.
14. Palstra, R.J., Simonis, M., Klous, P., Brasset, E., Eijkelkamp, B. and de Laat, W. (2008) Maintenance of long-range DNA interactions after inhibition of ongoing RNA polymerase II transcription. *PLoS One*, **3**, e1661.
15. Deng, W., Lee, J., Wang, H., Miller, J., Reik, A., Gregory, P.D., Dean, A. and Blobel, G.A. (2012) Controlling long-range genomic interactions at a native locus by targeted tethering of a looping factor. *Cell*, **149**, 1233-1244.
16. Gourdon, G., Sharpe, J.A., Wells, D., Wood, W.G. and Higgs, D.R. (1994) Analysis of a 70 kb segment of DNA containing the human zeta and alpha-globin genes linked to their regulatory element (HS-40) in transgenic mice. *Nucleic Acids Res*, **22**, 4139-4147.
17. Higgs, D.R., Sharpe, J.A. and Wood, W.G. (1998) Understanding alpha globin gene expression: a step towards effective gene therapy. *Semin Hematol*, **35**, 93-104.

18. Sharpe, J.A., Summerhill, R.J., Vyas, P., Gourdon, G., Higgs, D.R. and Wood, W.G. (1993) Role of upstream DNase I hypersensitive sites in the regulation of human alpha globin gene expression. *Blood*, **82**, 1666-1671.
19. Anguita, E., Hughes, J., Heyworth, C., Blobel, G.A., Wood, W.G. and Higgs, D.R. (2004) Globin gene activation during haemopoiesis is driven by protein complexes nucleated by GATA-1 and GATA-2. *EMBO J*, **23**, 2841-2852.
20. Vernimmen, D., De Gobbi, M., Sloane-Stanley, J.A., Wood, W.G. and Higgs, D.R. (2007) Long-range chromosomal interactions regulate the timing of the transition between poised and active gene expression. *EMBO J*, **26**, 2041-2051.
21. Vernimmen, D., Marques-Kranc, F., Sharpe, J.A., Sloane-Stanley, J.A., Wood, W.G., Wallace, H.A., Smith, A.J. and Higgs, D.R. (2009) Chromosome looping at the human alpha-globin locus is mediated via the major upstream regulatory element (HS -40). *Blood*, **114**, 4253-4260.
22. Jung, D. and Alt, F.W. (2004) Unraveling V(D)J recombination; insights into gene regulation. *Cell*, **116**, 299-311.
23. Schatz, D.G. and Ji, Y. (2011) Recombination centres and the orchestration of V(D)J recombination. *Nat Rev Immunol*, **11**, 251-263.
24. Kosak, S.T., Skok, J.A., Medina, K.L., Riblet, R., Le Beau, M.M., Fisher, A.G. and Singh, H. (2002) Subnuclear compartmentalization of immunoglobulin loci during lymphocyte development. *Science*, **296**, 158-162.
25. Fuxa, M., Skok, J., Souabni, A., Salvagiotto, G., Roldan, E. and Busslinger, M. (2004) Pax5 induces V-to-DJ rearrangements and locus contraction of the immunoglobulin heavy-chain gene. *Genes Dev*, **18**, 411-422.
26. Roldan, E., Fuxa, M., Chong, W., Martinez, D., Novatchkova, M., Busslinger, M. and Skok, J.A. (2005) Locus 'decontraction' and centromeric recruitment contribute to allelic exclusion of the immunoglobulin heavy-chain gene. *Nat Immunol*, **6**, 31-41.
27. Sayegh, C.E., Jhunjhunwala, S., Riblet, R. and Murre, C. (2005) Visualization of looping involving the immunoglobulin heavy-chain locus in developing B cells. *Genes Dev*, **19**, 322-327.
28. Jhunjhunwala, S., van Zelm, M.C., Peak, M.M., Cutchin, S., Riblet, R., van Dongen, J.J., Grosveld, F.G., Knoch, T.A. and Murre, C. (2008) The 3D structure of the immunoglobulin heavy-chain locus: implications for long-range genomic interactions. *Cell*, **133**, 265-279.
29. Liu, H., Schmidt-Supprian, M., Shi, Y., Hobeika, E., Barteneva, N., Jumaa, H., Pelanda, R., Reth, M., Skok, J. and Rajewsky, K. (2007) Yin Yang 1 is a critical regulator of B-cell development. *Genes Dev*, **21**, 1179-1189.
30. Reynaud, D., Demarco, I.A., Reddy, K.L., Schjerven, H., Bertolino, E., Chen, Z., Smale, S.T., Winandy, S. and Singh, H. (2008) Regulation of B cell fate commitment and immunoglobulin heavy-chain gene rearrangements by Ikaros. *Nat Immunol*, **9**, 927-936.
31. Degner, S.C., Wong, T.P., Jankevicius, G. and Feeney, A.J. (2009) Cutting edge: developmental stage-specific recruitment of cohesin to CTCF sites throughout immunoglobulin loci during B lymphocyte development. *J Immunol*, **182**, 44-48.
32. Guo, C., Yoon, H.S., Franklin, A., Jain, S., Ebert, A., Cheng, H.L., Hansen, E., Despo, O., Bossen, C., Vettermann, C. *et al.* (2011) CTCF-binding elements mediate control of V(D)J recombination. *Nature*, **477**, 424-430.
33. Guo, C., Gerasimova, T., Hao, H., Ivanova, I., Chakraborty, T., Selimyan, R., Oltz, E.M. and Sen, R. (2011) Two forms of loops generate the chromatin conformation of the immunoglobulin heavy-chain gene locus. *Cell*, **147**, 332-343.
34. Degner, S.C., Verma-Gaur, J., Wong, T.P., Bossen, C., Iverson, G.M., Torkamani, A., Vettermann, C., Lin, Y.C., Ju, Z., Schulz, D. *et al.* (2011) CCCTC-binding factor (CTCF) and cohesin influence the genomic architecture of the Igh locus and antisense transcription in pro-B cells. *Proc Natl Acad Sci U S A*, **108**, 9566-9571.
35. Seitan, V.C., Hao, B., Tachibana-Konwalski, K., Lavagnolli, T., Mira-Bontenbal, H., Brown, K.E., Teng, G., Carroll, T., Terry, A., Horan, K. *et al.* (2011) A role for cohesin in T-cell-receptor rearrangement and thymocyte differentiation. *Nature*, **476**, 467-471.
36. Ribeiro de Almeida, C., Stadhouders, R., de Bruijn, M.J., Bergen, I.M., Thongjuea, S., Lenhard, B., van Ijcken, W., Grosveld, F., Galjart, N., Soler, E. *et al.* (2011) The DNA-binding protein CTCF limits proximal Vkappa recombination and restricts kappa enhancer interactions to the immunoglobulin kappa light chain locus. *Immunity*, **35**, 501-513.
37. Seitan, V.C. and Merkenschlager, M. (2012) Cohesin and chromatin organisation. *Curr Opin Genet Dev*, **22**, 93-100.
38. Leighton, P.A., Saam, J.R., Ingram, R.S., Stewart, C.L. and Tilghman, S.M. (1995) An enhancer deletion affects both H19 and Igf2 expression. *Genes Dev*, **9**, 2079-2089.
39. Thorvaldsen, J.L., Duran, K.L. and Bartolomei, M.S. (1998) Deletion of the H19 differentially methylated domain results in loss of imprinted expression of H19 and Igf2. *Genes Dev*, **12**, 3693-3702.
40. Bartolomei, M.S., Webber, A.L., Brunkow, M.E. and Tilghman, S.M. (1993) Epigenetic mechanisms underlying the imprinting of the mouse H19 gene. *Genes Dev*, **7**, 1663-1673.
41. Ferguson-Smith, A.C., Sasaki, H., Cattanach, B.M. and Surani, M.A. (1993) Parental-origin-specific epigenetic modification of the mouse H19 gene. *Nature*, **362**, 751-755.
42. Hark, A.T., Schoenherr, C.J., Katz, D.J., Ingram, R.S., Levorse, J.M. and Tilghman, S.M. (2000) CTCF mediates methylation-sensitive enhancer-blocking activity at the H19/Igf2 locus. *Nature*, **405**, 486-489.
43. Bell, A.C. and Felsenfeld, G. (2000) Methylation of a CTCF-dependent boundary controls imprinted expression of the Igf2 gene. *Nature*, **405**, 482-485.
44. Murrell, A., Heeson, S. and Reik, W. (2004) Interaction between differentially methylated regions partitions the

imprinted genes Igf2 and H19 into parent-specific chromatin loops. *Nat Genet*, **36**, 889-893.
45. Kurukuti, S., Tiwari, V.K., Tavoosidana, G., Pugacheva, E., Murrell, A., Zhao, Z., Lobanenkov, V., Reik, W. and Ohlsson, R. (2006) CTCF binding at the H19 imprinting control region mediates maternally inherited higher-order chromatin conformation to restrict enhancer access to Igf2. *Proc Natl Acad Sci U S A*, **103**, 10684-10689.
46. Kmita, M. and Duboule, D. (2003) Organizing axes in time and space; 25 years of colinear tinkering. *Science*, **301**, 331-333.
47. Noordermeer, D., Leleu, M., Splinter, E., Rougemont, J., De Laat, W. and Duboule, D. (2011) The dynamic architecture of Hox gene clusters. *Science*, **334**, 222-225.
48. Spitz, F., Gonzalez, F. and Duboule, D. (2003) A global control region defines a chromosomal regulatory landscape containing the HoxD cluster. *Cell*, **113**, 405-417.
49. Gonzalez, F., Duboule, D. and Spitz, F. (2007) Transgenic analysis of Hoxd gene regulation during digit development. *Dev Biol*, **306**, 847-859.
50. Montavon, T., Soshnikova, N., Mascrez, B., Joye, E., Thevenet, L., Splinter, E., de Laat, W., Spitz, F. and Duboule, D. (2011) A regulatory archipelago controls Hox genes transcription in digits. *Cell*, **147**, 1132-1145.
51. Wang, K.C., Yang, Y.W., Liu, B., Sanyal, A., Corces-Zimmerman, R., Chen, Y., Lajoie, B.R., Protacio, A., Flynn, R.A., Gupta, R.A. *et al.* (2011) A long noncoding RNA maintains active chromatin to coordinate homeotic gene expression. *Nature*, **472**, 120-124.
52. Splinter, E. and de Laat, W. (2011) The complex transcription regulatory landscape of our genome: control in three dimensions. *EMBO J*, **30**, 4345-4355.
53. Branco, M.R. and Pombo, A. (2006) Intermingling of chromosome territories in interphase suggests role in translocations and transcription-dependent associations. *PLoS Biol*, **4**, e138.
54. Joffe, B., Leonhardt, H. and Solovei, I. (2010) Differentiation and large scale spatial organization of the genome. *Curr Opin Genet Dev*, **20**, 562-569.
55. Bolzer, A., Kreth, G., Solovei, I., Koehler, D., Saracoglu, K., Fauth, C., Muller, S., Eils, R., Cremer, C., Speicher, M.R. *et al.* (2005) Three-dimensional maps of all chromosomes in human male fibroblast nuclei and prometaphase rosettes. *PLoS Biol*, **3**, e157.
56. de Wit, E. and de Laat, W. (2012) A decade of 3C technologies: insights into nuclear organization. *Genes Dev*, **26**, 11-24.
57. Dostie, J. and Bickmore, W.A. (2012) Chromosome organization in the nucleus - charting new territory across the Hi-Cs. *Curr Opin Genet Dev*, **22**, 125-131.
58. Shopland, L.S., Lynch, C.R., Peterson, K.A., Thornton, K., Kepper, N., Hase, J., Stein, S., Vincent, S., Molloy, K.R., Kreth, G. *et al.* (2006) Folding and organization of a contiguous chromosome region according to the gene distribution pattern in primary genomic sequence. *J Cell Biol*, **174**, 27-38.
59. Simonis, M., Klous, P., Splinter, E., Moshkin, Y., Willemsen, R., de Wit, E., van Steensel, B. and de Laat, W. (2006) Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4C). *Nat Genet*, **38**, 1348-1354.
60. Lieberman-Aiden, E., van Berkum, N.L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B.R., Sabo, P.J., Dorschner, M.O. *et al.* (2009) Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, **326**, 289-293.
61. Yaffe, E. and Tanay, A. (2011) Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture. *Nat Genet*, **43**, 1059-1065.
62. Kalhor, R., Tjong, H., Jayathilaka, N., Alber, F. and Chen, L. (2012) Genome architectures revealed by tethered chromosome conformation capture and population-based modeling. *Nat Biotechnol*, **30**, 90-98.
63. Dixon, J.R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., Hu, M., Liu, J.S. and Ren, B. (2012) Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, **485**, 376-380.
64. Nora, E.P., Lajoie, B.R., Schulz, E.G., Giorgetti, L., Okamoto, I., Servant, N., Piolot, T., van Berkum, N.L., Meisig, J., Sedat, J. *et al.* (2012) Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature*, **485**, 381-385.
65. Sexton, T., Yaffe, E., Kenigsberg, E., Bantignies, F., Leblanc, B., Hoichman, M., Parrinello, H., Tanay, A. and Cavalli, G. (2012) Three-dimensional folding and functional organization principles of the Drosophila genome. *Cell*, **148**, 458-472.
66. Tolhuis, B., Blom, M., Kerkhoven, R.M., Pagie, L., Teunissen, H., Nieuwland, M., Simonis, M., de Laat, W., van Lohuizen, M. and van Steensel, B. (2011) Interactions among Polycomb domains are guided by chromosome architecture. *PLoS Genet*, **7**, e1001343.
67. Splinter, E., de Wit, E., Nora, E.P., Klous, P., van de Werken, H.J., Zhu, Y., Kaaij, L.J., van Ijcken, W., Gribnau, J., Heard, E. *et al.* (2011) The inactive X chromosome adopts a unique three-dimensional conformation that is dependent on Xist RNA. *Genes Dev*, **25**, 1371-1383.
68. Brown, K.E., Guest, S.S., Smale, S.T., Hahm, K., Merkenschlager, M. and Fisher, A.G. (1997) Association of transcriptionally silent genes with Ikaros complexes at centromeric heterochromatin. *Cell*, **91**, 845-854.
69. Brown, K.E., Baxter, J., Graf, D., Merkenschlager, M. and Fisher, A.G. (1999) Dynamic repositioning of genes in the nucleus of lymphocytes preparing for cell division. *Mol Cell*, **3**, 207-217.
70. Skok, J.A., Brown, K.E., Azuara, V., Caparros, M.L., Baxter, J., Takacs, K., Dillon, N., Gray, D., Perry, R.P., Merkenschlager, M. *et al.* (2001) Nonequivalent nuclear location of immunoglobulin alleles in B lymphocytes. *Nat*

Immunol, **2**, 848-854.
71. Zink, D., Amaral, M.D., Englmann, A., Lang, S., Clarke, L.A., Rudolph, C., Alt, F., Luther, K., Braz, C., Sadoni, N. *et al.* (2004) Transcription-dependent spatial arrangements of CFTR and adjacent genes in human cell nuclei. *J Cell Biol*, **166**, 815-825.
72. Guelen, L., Pagie, L., Brasset, E., Meuleman, W., Faza, M.B., Talhout, W., Eussen, B.H., de Klein, A., Wessels, L., de Laat, W. *et al.* (2008) Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions. *Nature*, **453**, 948-951.
73. Peric-Hupkes, D., Meuleman, W., Pagie, L., Bruggeman, S.W., Solovei, I., Brugman, W., Graf, S., Flicek, P., Kerkhoven, R.M., van Lohuizen, M. *et al.* (2010) Molecular maps of the reorganization of genome-nuclear lamina interactions during differentiation. *Mol Cell*, **38**, 603-613.
74. Mewborn, S.K., Puckelwartz, M.J., Abuisneineh, F., Fahrenbach, J.P., Zhang, Y., MacLeod, H., Dellefave, L., Pytel, P., Selig, S., Labno, C.M. *et al.* (2010) Altered chromosomal positioning, compaction, and gene expression with a lamin A/C gene mutation. *PLoS One*, **5**, e14342.
75. Kubben, N., Adriaens, M., Meuleman, W., Voncken, J.W., van Steensel, B. and Misteli, T. (2012) Mapping of lamin A- and progerin-interacting genome regions. *Chromosoma*.
76. Lobanenkov, V.V., Nicolas, R.H., Adler, V.V., Paterson, H., Klenova, E.M., Polotskaja, A.V. and 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**, 1743-1753.
77. Filippova, G.N., Fagerlie, S., Klenova, E.M., Myers, C., Dehner, Y., Goodwin, G., Neiman, P.E., Collins, S.J. and Lobanenkov, V.V. (1996) An exceptionally conserved transcriptional repressor, CTCF, employs different combinations of zinc fingers to bind diverged promoter sequences of avian and mammalian c-myc oncogenes. *Mol Cell Biol*, **16**, 2802-2813.
78. Vostrov, A.A. and 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. *J Biol Chem*, **272**, 33353-33359.
79. Yang, Y., Quitschke, W.W., Vostrov, A.A. and Brewer, G.J. (1999) CTCF is essential for up-regulating expression from the amyloid precursor protein promoter during differentiation of primary hippocampal neurons. *Journal of neurochemistry*, **73**, 2286-2298.
80. Bell, A.C., West, A.G. and Felsenfeld, G. (1999) The protein CTCF is required for the enhancer blocking activity of vertebrate insulators. *Cell*, **98**, 387-396.
81. Kim, T.H., Abdullaev, Z.K., Smith, A.D., Ching, K.A., Loukinov, D.I., Green, R.D., Zhang, M.Q., Lobanenkov, V.V. and Ren, B. (2007) Analysis of the vertebrate insulator protein CTCF-binding sites in the human genome. *Cell*, **128**, 1231-1245.
82. Barski, A., Cuddapah, S., Cui, K., Roh, T.Y., Schones, D.E., Wang, Z., Wei, G., Chepelev, I. and Zhao, K. (2007) High-resolution profiling of histone methylations in the human genome. *Cell*, **129**, 823-837.
83. Chen, X., Xu, H., Yuan, P., Fang, F., Huss, M., Vega, V.B., Wong, E., Orlov, Y.L., Zhang, W., Jiang, J. *et al.* (2008) Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. *Cell*, **133**, 1106-1117.
84. Hou, C., Dale, R. and Dean, A. (2010) Cell type specificity of chromatin organization mediated by CTCF and cohesin. *Proc Natl Acad Sci U S A*, **107**, 3651-3656.
85. Heath, H., Ribeiro de Almeida, C., Sleutels, F., Dingjan, G., van de Nobelen, S., Jonkers, I., Ling, K.W., Gribnau, J., Renkawitz, R., Grosveld, F. *et al.* (2008) CTCF regulates cell cycle progression of alphabeta T cells in the thymus. *EMBO J*, **27**, 2839-2850.
86. Majumder, P., Gomez, J.A., Chadwick, B.P. and 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. *J Exp Med*, **205**, 785-798.
87. Majumder, P. and Boss, J.M. (2010) CTCF controls expression and chromatin architecture of the human major histocompatibility complex class II locus. *Mol Cell Biol*, **30**, 4211-4223.
88. Ren, L., Wang, Y., Shi, M., Wang, X., Yang, Z. and Zhao, Z. (2012) CTCF mediates the cell-type specific spatial organization of the Kcnq5 locus and the local gene regulation. *PLoS One*, **7**, e31416.
89. Wendt, K.S., Yoshida, K., Itoh, T., Bando, M., Koch, B., Schirghuber, E., Tsutsumi, S., Nagae, G., Ishihara, K., Mishiro, T. *et al.* (2008) Cohesin mediates transcriptional insulation by CCCTC-binding factor. *Nature*, **451**, 796-801.
90. Parelho, V., Hadjur, S., Spivakov, M., Leleu, M., Sauer, S., Gregson, H.C., Jarmuz, A., Canzonetta, C., Webster, Z., Nesterova, T. *et al.* (2008) Cohesins functionally associate with CTCF on mammalian chromosome arms. *Cell*, **132**, 422-433.
91. Rubio, E.D., Reiss, D.J., Welcsh, P.L., Disteche, C.M., Filippova, G.N., Baliga, N.S., Aebersold, R., Ranish, J.A. and Krumm, A. (2008) CTCF physically links cohesin to chromatin. *Proc Natl Acad Sci U S A*, **105**, 8309-8314.
92. Fullwood, M.J., Liu, M.H., Pan, Y.F., Liu, J., Xu, H., Mohamed, Y.B., Orlov, Y.L., Velkov, S., Ho, A., Mei, P.H. *et al.* (2009) An oestrogen-receptor-alpha-bound human chromatin interactome. *Nature*, **462**, 58-64.
93. Handoko, L., Xu, H., Li, G., Ngan, C.Y., Chew, E., Schnapp, M., Lee, C.W., Ye, C., Ping, J.L., Mulawadi, F. *et al.* (2011) CTCF-mediated functional chromatin interactome in pluripotent cells. *Nat Genet*, **43**, 630-638.
94. Nasmyth, K. and Haering, C.H. (2009) Cohesin: its roles and mechanisms. *Annu Rev Genet*, **43**, 525-558.
95. Hadjur, S., Williams, L.M., Ryan, N.K., Cobb, B.S., Sexton, T., Fraser, P., Fisher, A.G. and Merkenschlager, M.

(2009) Cohesins form chromosomal cis-interactions at the developmentally regulated IFNG locus. *Nature*, **460**, 410-413.
96. Kim, Y.J., Cecchini, K.R. and Kim, T.H. (2011) Conserved, developmentally regulated mechanism couples chromosomal looping and heterochromatin barrier activity at the homeobox gene A locus. *Proc Natl Acad Sci U S A*, **108**, 7391-7396.
97. Majumder, P. and Boss, J.M. (2011) Cohesin regulates MHC class II genes through interactions with MHC class II insulators. *J Immunol*, **187**, 4236-4244.
98. Chien, R., Zeng, W., Kawauchi, S., Bender, M.A., Santos, R., Gregson, H.C., Schmiesing, J.A., Newkirk, D.A., Kong, X., Ball, A.R., Jr. *et al.* (2011) Cohesin mediates chromatin interactions that regulate mammalian beta-globin expression. *J Biol Chem*, **286**, 17870-17878.
99. Nativio, R., Wendt, K.S., Ito, Y., Huddleston, J.E., Uribe-Lewis, S., Woodfine, K., Krueger, C., Reik, W., Peters, J.M. and Murrell, A. (2009) Cohesin is required for higher-order chromatin conformation at the imprinted IGF2-H19 locus. *PLoS Genet*, **5**, e1000739.
100. Schmidt, D., Schwalie, P.C., Ross-Innes, C.S., Hurtado, A., Brown, G.D., Carroll, J.S., Flicek, P. and Odom, D.T. (2010) A CTCF-independent role for cohesin in tissue-specific transcription. *Genome Res*, **20**, 578-588.
101. Kagey, M.H., Newman, J.J., Bilodeau, S., Zhan, Y., Orlando, D.A., van Berkum, N.L., Ebmeier, C.C., Goossens, J., Rahl, P.B., Levine, S.S. *et al.* (2010) Mediator and cohesin connect gene expression and chromatin architecture. *Nature*, **467**, 430-435.
102. Fraser, P. and Bickmore, W. (2007) Nuclear organization of the genome and the potential for gene regulation. *Nature*, **447**, 413-417.
103. Jackson, D.A., Hassan, A.B., Errington, R.J. and Cook, P.R. (1993) Visualization of focal sites of transcription within human nuclei. *EMBO J*, **12**, 1059-1065.
104. Jackson, D.A., Iborra, F.J., Manders, E.M. and Cook, P.R. (1998) Numbers and organization of RNA polymerases, nascent transcripts, and transcription units in HeLa nuclei. *Mol Biol Cell*, **9**, 1523-1536.
105. Iborra, F.J., Pombo, A., Jackson, D.A. and Cook, P.R. (1996) Active RNA polymerases are localized within discrete transcription "factories' in human nuclei. *J Cell Sci*, **109 (Pt 6)**, 1427-1436.
106. Grande, M.A., van der Kraan, I., de Jong, L. and van Driel, R. (1997) Nuclear distribution of transcription factors in relation to sites of transcription and RNA polymerase II. *J Cell Sci*, **110 (Pt 15)**, 1781-1791.
107. Osborne, C.S., Chakalova, L., Brown, K.E., Carter, D., Horton, A., Debrand, E., Goyenechea, B., Mitchell, J.A., Lopes, S., Reik, W. *et al.* (2004) Active genes dynamically colocalize to shared sites of ongoing transcription. *Nat Genet*, **36**, 1065-1071.
108. Schoenfelder, S., Sexton, T., Chakalova, L., Cope, N.F., Horton, A., Andrews, S., Kurukuti, S., Mitchell, J.A., Umlauf, D., Dimitrova, D.S. *et al.* (2010) Preferential associations between co-regulated genes reveal a transcriptional interactome in erythroid cells. *Nat Genet*, **42**, 53-61.
109. Xu, M. and Cook, P.R. (2008) Similar active genes cluster in specialized transcription factories. *J Cell Biol*, **181**, 615-623.
110. Kimura, H., Sugaya, K. and Cook, P.R. (2002) The transcription cycle of RNA polymerase II in living cells. *J Cell Biol*, **159**, 777-782.
111. Zobeck, K.L., Buckley, M.S., Zipfel, W.R. and Lis, J.T. (2010) Recruitment timing and dynamics of transcription factors at the Hsp70 loci in living cells. *Mol Cell*, **40**, 965-975.
112. Razin, S.V., Gavrilov, A.A., Pichugin, A., Lipinski, M., Iarovaia, O.V. and Vassetzky, Y.S. (2011) Transcription factories in the context of the nuclear and genome organization. *Nucleic Acids Res*, **39**, 9085-9092.
113. Li, G., Ruan, X., Auerbach, R.K., Sandhu, K.S., Zheng, M., Wang, P., Poh, H.M., Goh, Y., Lim, J., Zhang, J. *et al.* (2012) Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation. *Cell*, **148**, 84-98.
114. Merkenschlager, M., Amoils, S., Roldan, E., Rahemtulla, A., O'Connor, E., Fisher, A.G. and Brown, K.E. (2004) Centromeric repositioning of coreceptor loci predicts their stable silencing and the CD4/CD8 lineage choice. *J Exp Med*, **200**, 1437-1444.
115. Collins, A., Hewitt, S.L., Chaumeil, J., Sellars, M., Micsinai, M., Allinne, J., Parisi, F., Nora, E.P., Bolland, D.J., Corcoran, A.E. *et al.* (2011) RUNX transcription factor-mediated association of Cd4 and Cd8 enables coordinate gene regulation. *Immunity*, **34**, 303-314.
116. Ballester, M., Kress, C., Hue-Beauvais, C., Kieu, K., Lehmann, G., Adenot, P. and Devinoy, E. (2008) The nuclear localization of WAP and CSN genes is modified by lactogenic hormones in HC11 cells. *J Cell Biochem*, **105**, 262-270.
117. Kress, C., Kieu, K., Droineau, S., Galio, L. and Devinoy, E. (2011) Specific positioning of the casein gene cluster in active nuclear domains in luminal mammary epithelial cells. *Chromosome Res*, **19**, 979-997.
118. Ragoczy, T., Bender, M.A., Telling, A., Byron, R. and Groudine, M. (2006) The locus control region is required for association of the murine beta-globin locus with engaged transcription factories during erythroid maturation. *Genes Dev*, **20**, 1447-1457.
119. Brown, J.M., Green, J., das Neves, R.P., Wallace, H.A., Smith, A.J., Hughes, J., Gray, N., Taylor, S., Wood, W.G., Higgs, D.R. *et al.* (2008) Association between active genes occurs at nuclear speckles and is modulated by chromatin environment. *J Cell Biol*, **182**, 1083-1097.
120. Chambeyron, S., Da Silva, N.R., Lawson, K.A. and Bickmore, W.A. (2005) Nuclear re-organisation of the Hoxb

complex during mouse embryonic development. *Development*, **132**, 2215-2223.
121. Chambeyron, S. and Bickmore, W.A. (2004) Chromatin decondensation and nuclear reorganization of the HoxB locus upon induction of transcription. *Genes Dev*, **18**, 1119-1130.
122. Morey, C., Da Silva, N.R., Perry, P. and Bickmore, W.A. (2007) Nuclear reorganisation and chromatin decondensation are conserved, but distinct, mechanisms linked to Hox gene activation. *Development*, **134**, 909-919.
123. Tarchini, B. and Duboule, D. (2006) Control of Hoxd genes' collinearity during early limb development. *Dev Cell*, **10**, 93-103.
124. Noordermeer, D., Branco, M.R., Splinter, E., Klous, P., van Ijcken, W., Swagemakers, S., Koutsourakis, M., van der Spek, P., Pombo, A. and de Laat, W. (2008) Transcription and chromatin organization of a housekeeping gene cluster containing an integrated beta-globin locus control region. *PLoS Genet*, **4**, e1000016.
125. Morey, C., Kress, C. and Bickmore, W.A. (2009) Lack of bystander activation shows that localization exterior to chromosome territories is not sufficient to up-regulate gene expression. *Genome Res*, **19**, 1184-1194.
126. Robinett, C.C., Straight, A., Li, G., Willhelm, C., Sudlow, G., Murray, A. and Belmont, A.S. (1996) In vivo localization of DNA sequences and visualization of large-scale chromatin organization using lac operator/repressor recognition. *J Cell Biol*, **135**, 1685-1700.
127. Tumbar, T., Sudlow, G. and Belmont, A.S. (1999) Large-scale chromatin unfolding and remodeling induced by VP16 acidic activation domain. *J Cell Biol*, **145**, 1341-1354.
128. Chubb, J.R., Boyle, S., Perry, P. and Bickmore, W.A. (2002) Chromatin motion is constrained by association with nuclear compartments in human cells. *Curr Biol*, **12**, 439-445.
129. Ye, Q., Hu, Y.F., Zhong, H., Nye, A.C., Belmont, A.S. and Li, R. (2001) BRCA1-induced large-scale chromatin unfolding and allele-specific effects of cancer-predisposing mutations. *J Cell Biol*, **155**, 911-921.
130. Nye, A.C., Rajendran, R.R., Stenoien, D.L., Mancini, M.A., Katzenellenbogen, B.S. and Belmont, A.S. (2002) Alteration of large-scale chromatin structure by estrogen receptor. *Mol Cell Biol*, **22**, 3437-3449.
131. Chen, D., Belmont, A.S. and Huang, S. (2004) Upstream binding factor association induces large-scale chromatin decondensation. *Proc Natl Acad Sci U S A*, **101**, 15106-15111.
132. Verschure, P.J., van der Kraan, I., de Leeuw, W., van der Vlag, J., Carpenter, A.E., Belmont, A.S. and van Driel, R. (2005) In vivo HP1 targeting causes large-scale chromatin condensation and enhanced histone lysine methylation. *Mol Cell Biol*, **25**, 4552-4564.
133. Hathaway, N.A., Bell, O., Hodges, C., Miller, E.L., Neel, D.S. and Crabtree, G.R. (2012) Dynamics and memory of heterochromatin in living cells. *Cell*, **149**, 1447-1460.
134. Kumaran, R.I. and Spector, D.L. (2008) A genetic locus targeted to the nuclear periphery in living cells maintains its transcriptional competence. *J Cell Biol*, **180**, 51-65.
135. Reddy, K.L., Zullo, J.M., Bertolino, E. and Singh, H. (2008) Transcriptional repression mediated by repositioning of genes to the nuclear lamina. *Nature*, **452**, 243-247.
136. Finlan, L.E., Sproul, D., Thomson, I., Boyle, S., Kerr, E., Perry, P., Ylstra, B., Chubb, J.R. and Bickmore, W.A. (2008) Recruitment to the nuclear periphery can alter expression of genes in human cells. *PLoS Genet*, **4**, e1000039.
137. Zullo, J.M., Demarco, I.A., Pique-Regi, R., Gaffney, D.J., Epstein, C.B., Spooner, C.J., Luperchio, T.R., Bernstein, B.E., Pritchard, J.K., Reddy, K.L. *et al.* (2012) DNA sequence-dependent compartmentalization and silencing of chromatin at the nuclear lamina. *Cell*, **149**, 1474-1487.
138. Noordermeer, D., de Wit, E., Klous, P., van de Werken, H., Simonis, M., Lopez-Jones, M., Eussen, B., de Klein, A., Singer, R.H. and de Laat, W. (2011) Variegated gene expression caused by cell-specific long-range DNA interactions. *Nat Cell Biol*, **13**, 944-951.

2
CTCF: the protein, the binding partners, the binding sites and their chromatin loops

Philos Trans R Soc Lond B Biol Sci. 2013 May 6;368(1620)

CTCF: the protein, the binding partners, the binding sites and their chromatin loops

Sjoerd Holwerda[1] and Wouter de Laat[1,*]

1 Hubrecht Institute-KNAW & University Medical Center Utrecht, Utrecht, The Netherlands
* Corresponding author: Email: w.delaat@hubrecht.eu

ABSTRACT

CTCF has it all. The transcription factor binds to tens of thousands of genomic sites, some tissue-specific, others ultra-conserved. It can act as a transcriptional activator, repressor and insulator and it can pause transcription. CTCF binds at chromatin domain boundaries, at enhancers and gene promoters and inside gene bodies. It can attract many other transcription factors to chromatin, including tissue-specific transcriptional activators, repressors, cohesin and RNA polymerase II and it forms chromatin loops. Yet, or perhaps therefore, CTCF's exact function at a given genomic site is unpredictable. It appears to be determined by the associated transcription factors, by the location of the binding site relative to the transcriptional start site of a gene, and by the site's engagement in chromatin loops with other CTCF binding sites, enhancers or gene promoters. Here, we will discuss genome-wide features of CTCF binding events, as well as locus-specific functions of this remarkable transcription factor.

Introduction

CTCF is a ubiquitously expressed and an essential protein (1) and is, in many ways, an exceptional transcription factor. It was first described as a transcriptional repressor (2) but was also found to act as a transcriptional activator (3,4). Most strikingly, it harbours insulator activity: when positioned in between an enhancer and gene promoter, it can block their communication and prevent transcriptional activation (5-7). Systematic chromatin immunoprecipitation experiments combined with high-throughput sequencing (ChIP-seq) have been performed to map CTCF binding events across the genome in many tissues of different species (8-10). They show that the genome is covered with a myriad of CTCF binding sites. More than most other transcription factors CTCF appears to bind to intergenic sequences, often at a distance from the transcriptional start site (TSS) (11). CTCF was one of the first proteins demonstrated to mediate chromatin looping between its binding sites (12,13). Further evidence for its role in the organization of genome structure comes from observations that it frequently binds to boundaries between chromosomal regions that occupy distinct locations in the nucleus, to boundaries between regions with different epigenetic signatures and/or different transcriptional activities, and to boundaries between recently identified topological domains, which are spatially defined chromosomal units within which sequences preferentially interact with each other (14,15). Here, we will discuss studies on CTCF and evaluate its function in genome folding and gene expression.

CTCF at the beta-globin and the H19-Igf2 locus, a short history

Functions of the versatile DNA binding protein CTCF were initially explored at individual loci, in particular at the β-globin locus and the imprinted H19-Igf2 locus. The chicken beta-globin (β-globin) locus carries a DNaseI hypersensitive site (5'HS4) at its 5' side that separates the locus from neighbouring heterochromatin and this site was found capable of blocking enhancer activity (16). CTCF was subsequently demonstrated to be responsible for this insulator activity of 5'HS4 (5). The human and mouse -globin loci also locate inside large chromosomal regions of inactive chromatin and are similarly flanked by CTCF binding sites (17,18). These were suspected to form a barrier for incoming heterochromatin, but their deletion did not lead to closing or inactivation of the β-globin locus (12,19). The application of chromosome conformation capture (3C) technology enabled the demonstration that the β-globin CTCF sites physically interact with each other. They form large chromatin loops encompassing the β-globin main regulatory element, the locus control region (LCR), and its genes. These loops are erythroid-specific and are formed in erythroid progenitor cells, prior to LCR-mediated high expression of the β-globin genes (**Figure 1a**) (12,20). It was speculated that the CTCF loops can facilitate subsequent spatial interactions between the LCR and its target genes, but evidence for this is still lacking.
Another locus historically important for CTCF's reputation as an interesting transcription factor is the imprinted H19/Igf2 locus. The locus contains a differentially methylated region (DMR) that is known as the imprinting control region (ICR), located in between the H19 and the Igf2 genes. The ICR determines that H19 is active on the maternal allele and that Igf2 is transcribed from the paternal allele (21,22). CTCF entered stage here when it was found to bind to the ICR in a methylation-dependent manner: the binding of CTCF to the unmethylated maternal ICR prevents shared enhancers near the H19 gene from reaching across and activating Igf2. On the paternal allele CTCF cannot exert its insulator activity as DNA methylation prevents its binding to the ICR (**Figure 1b**) (6,23). Again, chromatin loops are formed and seem important for ICR functioning (24-26). Allele-specific chromatin loops with both enhancers and promoters are formed by the maternal, CTCF-bound, ICR, suggesting that such contacts may underlie CTCF-mediated insulator activity (26). Collectively, the early studies on CTCF functioning at the β-globin and the H19-Igf2 locus revealed that the protein can interfere with promoter-enhancer communication. They also showed that CTCF can form chromatin loops between its binding sites, and perhaps also with other regulatory sequences.

CTCF binds across the genome to chromatin boundaries, enhancers and gene promoters

The systematic mapping of genome-wide binding sites by ChIP revealed that CTCF binds to tens of thousands of genomic sites (10,11,27). Association to roughly one-third of these sites is relatively conserved across different cell types (9). An inter-species comparison between CTCF-binding profiles in the liver of five mammalian organisms uncovered approximately 5000 sites that are ultra-conserved between the species and tissues. These appear to be the high affinity binding sites, suggesting that differences in affinity could be related to the strength of conservation (8). The activation of retro-elements has produced species-specific expansions of CTCF binding sites and this form of genome evolution is still highly active in mammals (8). Classification of CTCF binding sites based on a consensus motif score lead to similar conclusions: high-occupancy sites appear to be conserved across cell-types, while low-occupancy sites are more tissue-restricted (28).

The CTCF consensus binding sequence contains CpG and can therefore be subject to DNA methylation. CTCF is able to bind to methylated DNA sequences *in vitro* (29), but preferentially binds to unmethylated

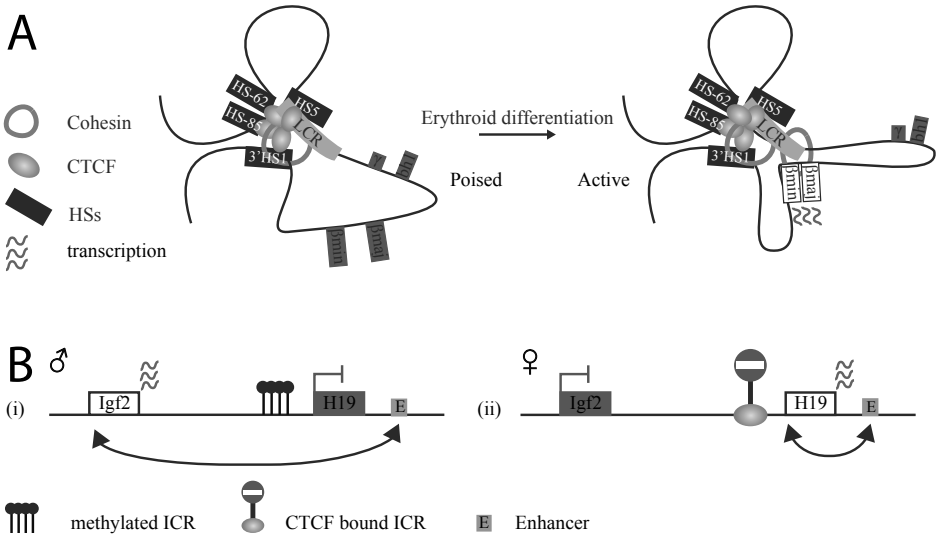


Figure 1. CTCF, chromatin loops and transcription regulation at selected gene loci. (A) Genes at the β-globin locus are under control of the LCR. CTCF binding sites interact to create a chromatin hub with a loop encompassing the LCR and the β-globin genes. Upon erythroid differentiation, erythroid-specific transcription factors and cohesin enable the formation of an active chromatin hub (ACH) in which the LCR contacts the genes and enhances their expression. (B) Imprinted expression of the H19 and Igf2 genes is mediated by methylation dependent binding of CTCF at the imprinted control region (ICR). (i) On the paternal allele, methylation of the ICR prevents CTCF binding and allows expression of the Igf2 gene mediated by contacts between the distal enhancer (E) and the Igf2 promoter. (ii) CTCF binding at the ICR blocks communication between the Igf2 gene and the distal enhancer resulting in expression of the H19 gene from the maternal allele.

sequences, as seen also at the H19-Igf2 locus. In fact, DNA methylation appears to play a role in some of the tissue-specific binding events of CTCF (9). Moreover, CTCF can influence DNA methylation by forming a complex with two enzymes related to DNA methylation: poly(ADP-ribose) polymerase 1 (PARP1) and the ubiquitously expressed DNA (cytosine-5)-methyltransferase 1 (DNMT1). CTCF activates PARP1, which then can add ADP-ribose groups to DNMT1 to inactivate this enzyme, with maintenance of methyl-free CpGs as the result (30-32).

A portion of CTCF binding sites is found enriched at transitions between active chromatin (high in H2K5Ac) and inactive chromatin domains (high in H3K27me3) (27,33). This seems particularly true for retro-transposed CTCF binding sites (8). CTCF sites frequently flank so-called lamina associated domains (LADs). LADs are chromosomal regions associated with the lamin-based protein network that coats the inner side of the nuclear envelope; these chromosomal regions tend to be transcriptionally inactive (14). Its presence at LAD boundaries suggests that CTCF helps organizing the three-dimensional structure of chromatin. In *Drosophila*, the knockdown of CTCF leads to decreased levels of H3K27me3 inside inactive domains, indicating that CTCF binding at boundaries is required for the maintenance of repression (34). Association of CTCF follows the re-setting of active and inactive domains during cellular differentiation, further suggesting that it functions to separate different chromatin states (33). Some of the LADs also dynamically change during cellular differentiation (35), but whether CTCF binds to the borders of these differential LADs is currently unclear.

Although CTCF binding is often found distal from TSSs, it does show a strong correlation with gene density (**Figure 2a,b**) (11). Indeed, evidence for a direct role of CTCF in transcription regulation already came from early studies on individual genes (3,36). Genome wide, a portion of CTCF sites co-localizes with the promoter-specific H3K4me3 mark and another part coincides with the enhancer mark H3K4me1 (27). CTCF binding events at promoters tend to be conserved across tissues, while CTCF binding to enhancers is more tissue-restricted (10).

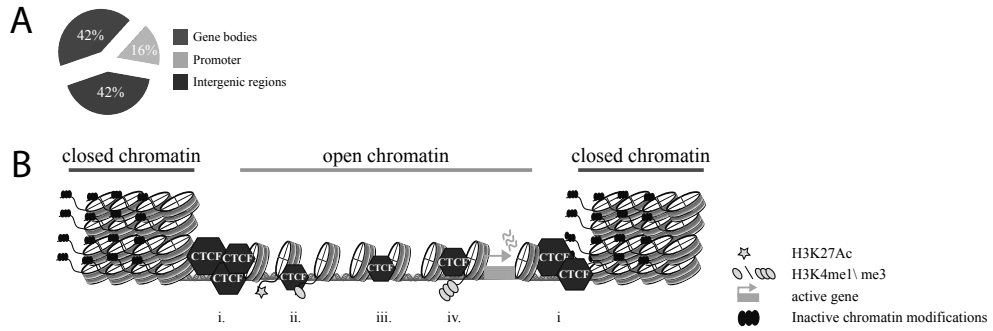


Figure 2. A versatile role for CTCF in chromatin biology. (A) Functional categories of CTCF binding sites across the genome, adopted from (82). (B) (i) CTCF binding sites at boundaries separate active and inactive chromatin domains. CTCF binding to enhancer-like sequences (ii) and gene promoters (iv) can facilitate looping between these sequences. CTCF binding in between enhancers and gene promoters (iii) can block the interaction between an enhancer and its target promoter.

CTCF and cohesin share DNA binding sites

An unanticipated observation was the co-localization of cohesin with many of the chromosomal binding sites of CTCF (37-41). Cohesin has always been associated with DNA replication and sister chromatid cohesion during the S, G2 and M phase of the cell cycle (42). It is a protein complex that contains members of a family of 'structural maintenance of chromosomes' ('SMCs') proteins. The complex forms a ring-like protein structure that is thought to embrace two DNA helices. Surprisingly at the time, cohesin was found to bind chromatin also in post-mitotic cells, with half of its binding sites overlapping with CTCF sites (37-39). Cohesin association to these sites is dependent on the presence of CTCF: without CTCF, cohesin still binds to chromatin but is no longer found at specific sequences. In contrast, CTCF does not rely on cohesin for finding its DNA binding sites. One possibility is that bound CTCF serves as a roadblock or barrier to position a sliding cohesin molecule on the chromatin template (37-41).

Its cell-cycle independent association to DNA suggests that cohesin has an additional role in gene regulation. Given its capacity to hold together two sister chromatids, cohesin is obviously also attractive as a looping factor. Indeed, at the H19-Igf2 locus, cohesin was shown to be important for CTCF-mediated chromatin loop formation and proper regulation of Igf2 transcription (43). Similarly, at the interferon gamma (IFNG) locus, depletion of cohesin was found to disrupt chromatin loops between regulatory DNA sequences and cause a reduction in IFNG expression (44). Also at the β-globin locus cohesin has been implicated in chromatin looping, not only between the flanking CTCF sites but also between the LCR enhancer region and the downstream β-globin target genes (45). Conditional deletion of cohesin in thymocytes was shown to disrupt the formation of regulatory chromatin loops in the T cell receptor-a (TCRa) locus, with reduced transcription and impaired V(DJ) rearrangement as a consequence (46). Pairwise comparison between two cell types revealed that it is mostly the CTCF-independent cohesin binding events that show cell-type specificity. At these sites, cohesin is often found co-localized with mediator and RNA polymerase II (RNAPII), indicating a CTCF-independent function at enhancer sequences. Consistent with this, these genomic sites were often found to be close to actively transcribed genes (47) and to be co-occupied by tissue-specific transcription factors (48,49). Collectively this shows that CTCF and cohesin have shared and independent functions at regulatory sequences in the genome. Cohesin can form chromatin loops during interphase. However, whether this occurs through its embracement of two DNA double helices still awaits formal proof.

CTCF and other binding partners

CTCF performs multiple roles, and in agreement the protein shares chromatin binding sites with many other factors (50-52). Co-association events such as those with the histone deacetylase SIN3 (53), the thyroid hormone receptor (54), nucleophosmin (55), Kaiso (56) and the DEAD-box RNA helicase p68 with associated non-coding RNA (57) have been implicated in its insulator function. Interestingly, the p68 RNA-protein complex appears required for positioning cohesin at the CTCF sites of the H19-Igf2 ICR (57). In addition, CTCF co-occupies sites with the transcription factors FOXA1 and the estrogen receptor (ER). These sites tend to locate near ER-responsive genes, suggesting that CTCF facilitates their transcriptional activation (58). Furthermore, CTCF recruits the basal transcription factor TAF3 to intergenic sites in embryonic stem cells (ESCs), where TAF3-dependent chromatin loop formation was shown to activate gene transcription (59). In a study that monitored RNAPII tracking along long tumor necrosis factor-alpha (TNFalpha) responsive genes, pausing of RNAPII was observed at CTCF- and cohesin bound sites (60). This pausing can serve to incorporate weak exons and therefore facilitate alternative splicing (61). Thus, intra- and intergenic CTCF sites can have many different roles.

CTCF function at individual gene loci

Given its diverse activities, it seems necessary to zoom in on individual loci to understand CTCF's local function. At the proto-oncogene Myb locus, CTCF binding occurs in the first intron of the gene, where it inhibits RNAPII elongation. Transcriptional pausing by CTCF can be overcome by upstream enhancers that bind tissue-specific transcriptional activators and loop towards the Myb promoter (62). At the major histocompatibility complex class II (MHCII) locus, CTCF and cohesin binding to, and looping between, upstream sequences precedes transcriptional activation. Upon binding of the MHCII trans-activator CIITA to the promoter sequences, loops are induced between them and the various CTCF sites, resulting in increased expression of MHCII genes (63,64).

CTCF also binds to many sites across the immunoglobulin and T cell receptor antigen receptor gene loci. In conditional CTCF knockout mice, V gene usage in the Igk light chain locus was found to be altered, with increased recombination with proximal and reduced recombination with distal V segments. This was accompanied by corresponding changes in germline transcription at these locations, suggesting that CTCF, like cohesin (46), mediates gene usage of the antigen receptor loci via the local regulation of germline transcription (65). In this model, germline transcription increases accessibility of the region, which facilitates their selection for recombination (66). CTCF depletion does not always result in aberrant gene expression. Using the same conditional knock-out mice (1) Hoxd gene expression in the developing limb bud was unaltered after knockout of CTCF (67). Hoxd gene expression in the limb bud is under the control of many distant regulatory sequences that physically loop towards the genes (68). Unaltered Hoxd expression in the absence of CTCF suggests that these enhancer-promoter loops are not influenced by CTCF binding to sites in and around the locus. This raises the question whether CTCF has any impact on the three-dimensional topology of the locus. While Hoxd expression was not affected, CTCF depletion did cause massive cell death in the limb, showing that CTCF is critical for the transcriptional regulation of other genes involved in cellular homeostasis (67).

A final locus that is interesting to discuss is the protocadherin-a (PCDHa) cluster. This cluster encodes neuronal specific transmembrane proteins that are mono-allelically expressed and thought to be involved in the recognition and diversification of neurons. Expression of the cluster is under control of a downstream enhancer that influences the expression of the twelve isoforms, each of which is having its own alternative promoter (69). Interestingly, the enhancer as well as each individual promoter has a binding site for CTCF.

Expression of the isoforms is reduced upon conditional CTCF knock-out in post-mitotic neurons. This suggests that long-range interactions are part of the regulatory process that controls transcription of these genes (69-71).
Collectively, these studies support the idea that chromatin-bound CTCF can attract many different transcription factors in a tissue- and genomic context-specific manner. Its exact function at a given genomic site is probably determined by these associated transcription factors, by the location of this site relative to the transcriptional start site of a gene, and by its engagement in chromatin loops with other CTCF binding sites, enhancers or gene promoters.

Genome wide chromatin loops mediated by CTCF

A computational intersection of the genomic binding sites of CTCF (assessed by genome-wide ChIP) with a genome-wide DNA contact map generated by Hi-C (72) suggested that CTCF is involved in chromatin interactions between and within chromosomes across the genome (73). Chromatin interaction analysis with paired-end tag sequencing (ChIA-PET) combines ChIP with a 3C approach and was developed to study genome-wide DNA interactions mediated by a protein of interest (74). When targeted to CTCF, ChIA-PET uncovered roughly 1,500 intra-chromosomal and around 300 inter-chromosomal interactions mediated by this protein (13). Subsequent clustering of the regions (10-200 kb) encompassed by the intra-chromosomal loops was done based on the distribution of histone marks. This showed that CTCF loops can contain active chromatin separated from inactive chromatin outside the loops, and vice versa. CTCF can also capture enhancers and promoters together in a chromatin loop (13). Only a fraction of the roughly ~40,000 CTCF binding sites was found to participate in the roughly 1,500 CTCF mediated loops. This implies that either not all interactions mediated by CTCF have been identified, or that most CTCF sites are not engaged in the formation of loops.
The latter may well be true, since 5C technology showed that most CTCF sites across 1% of the genome do not participate in chromatin loops, no matter whether they are co-occupied by cohesin or not. CTCF bound sequences were often skipped by gene promoters making contacts with enhancers or with other CTCF sites even further away (75).
The recent availability of large genome-wide DNA interaction datasets (15,72) facilitates the assessment of CTCF's impact on chromosome topology. Sequences close on the chromosome to CTCF binding sites were

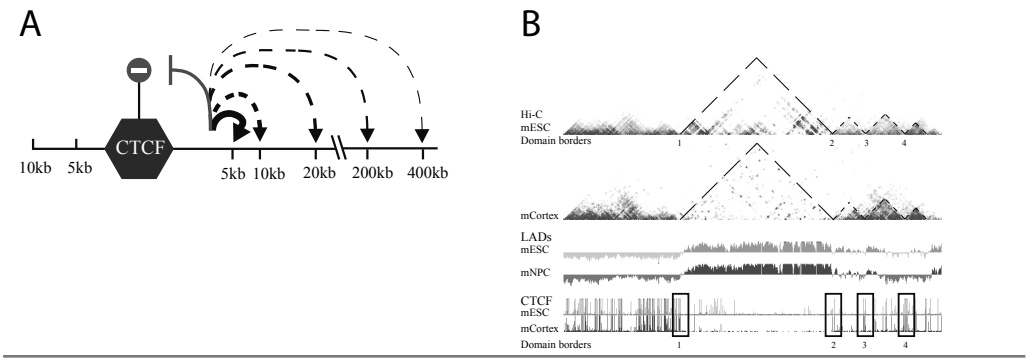


Figure 3. CTCF acts as an insulator by hampering DNA contacts across its binding sites. (A) CTCF hampers sequences within 10 kilobases of its binding sites to reach across (76). Decreased thickness of lines indicates a decrease in interaction probability. (B) CTCF binding sites are often found at the borders of topological domains. Top show Hi-C data in ESCs and in cortex, with color-coded contact frequencies between sequences on mouse chromosome 12. Triangles reveal and highlight topological domains (15), being chromosomal regions within which sequences preferentially interact with each other. Middle: LAD data (14, 35), showing that LAD boundaries can coincide with the borders of topological domains. Bottom: CTCF ChIP-seq profiles, showing clusters of CTCF binding sites at the borders (10). Note that such CTCF clusters also exist elsewhere, particularly in non-LADs. Region shown: chr12: 112.3-119.3 Mb (mm9).

shown to be biased in their DNA contacts: they interacted with other sequences on the same side of the CTCF site more than with sequences across this site (**Figure 3a**) (76). The same was previously shown for a different insulator protein in *Drosophila*: its binding to a site prevented flanking sequences to physically contact each other across this site (77). Interestingly, this may provide an explanation for how insulators function: they can prevent spatial DNA contacts across the insulating sequence. In a particularly detailed genome-wide DNA contact study topological domains were defined; they are chromosomal regions of on average 1 Mb in size, within which sequences preferentially interact with each other (15,78). A strong conservation of topological domains was seen between tissues and even between species, suggesting that these domains do not contribute themselves to the specific identity of cells. Interestingly, CTCF binding sites were enriched in 20kb windows surrounding the boundaries of these domains (**Figure 3b**), re-emphasizing its role as a chromatin organizer. In one case it was shown that disruption of a boundary led to intermingling of topological domains and caused misregulated expression of the genes involved (78). Unlike the topological domains themselves, contacts within the domains do change during differentiation. Also here CTCF appears to play a role, probably to accomodate developmental changes in gene expression (79-81).

Concluding remarks

Despite being subject of intense research, CTCF manages to remain a mysterious transcription factor. It binds to many thousands of sites across the genome, where it can interact with a plethora of other transcription factors. It is often found engaged in chromatin loops, sometimes with and sometimes without the involvement of cohesin. It can form chromatin loops with other CTCF binding sites, but also with enhancer and promoter sequences. CTCF binds to sequences outside and away from genes, but also inside the gene body where it appears capable of pausing the sliding polymerase molecule. Finally, CTCF binding sites still actively jump around as retrotransposable sequences, giving diversity to the CTCF binding landscape between different mammalian species.

We believe that the unifying theme which may explain the many, and sometimes opposing, functional consequences of CTCF association to chromatin is probably its ability to form chromatin loops. Depending on the sequences encompassed in the loops and those excluded from the loops, chromatin shaped by CTCF may facilitate or hamper 3D contacts between enhancers and target genes, with different outcomes for transcription. Many questions still remain though: Why do some CTCF sites form a chromatin loop and others not? To what extent does this rely on co-associated protein factors? How does the protein manage to interact with so many other transcription factors when bound to chromatin? One possibility is that CTCF serves as a roadblock for chromatin-scanning transcription factors that somehow get trapped when encountering the bound protein. What is the relevance of CTCF-mediated interchromosomal contacts? Does CTCF block enhancer-promoter communication by preventing 3D DNA contacts? Or does insulation involve the physical interaction of the insulator sequence with both enhancers and promoters? Answers to these questions are needed to enable predicting whether a given CTCF binding event will be functionally irrelevant, will cause transcriptional activation or repression, will interfere with transcriptional activation or will create a chromatin boundary.

Acknowledgements

We would like to thank Patrick Wijchers and Peter Krijger for useful comments. This work was financially supported by grant no. 935170621 from the Dutch Scientific Organization (NWO) and a European Research Council Starting Grant (209700, '4C') to WdL.

References

1. Heath, H., Ribeiro de Almeida, C., Sleutels, F., Dingjan, G., van de Nobelen, S., Jonkers, I., Ling, K.W., Gribnau, J., Renkawitz, R., Grosveld, F. *et al.* (2008) CTCF regulates cell cycle progression of alphabeta T cells in the thymus. *The EMBO journal*, **27**, 2839-2850.
2. Lobanenkov, V.V., Nicolas, R.H., Adler, V.V., Paterson, H., Klenova, E.M., Polotskaja, A.V. and 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**, 1743-1753.
3. Klenova, E.M., Nicolas, R.H., Paterson, H.F., Carne, A.F., Heath, C.M., Goodwin, G.H., Neiman, P.E. and Lobanenkov, V.V. (1993) CTCF, a conserved nuclear factor required for optimal transcriptional activity of the chicken c-myc gene, is an 11-Zn-finger protein differentially expressed in multiple forms. *Molecular and cellular biology*, **13**, 7612-7624.
4. Filippova, G.N., Fagerlie, S., Klenova, E.M., Myers, C., Dehner, Y., Goodwin, G., Neiman, P.E., Collins, S.J. and Lobanenkov, V.V. (1996) An exceptionally conserved transcriptional repressor, CTCF, employs different combinations of zinc fingers to bind diverged promoter sequences of avian and mammalian c-myc oncogenes. *Molecular and cellular biology*, **16**, 2802-2813.
5. Bell, A.C., West, A.G. and Felsenfeld, G. (1999) The protein CTCF is required for the enhancer blocking activity of vertebrate insulators. *Cell*, **98**, 387-396.
6. Hark, A.T., Schoenherr, C.J., Katz, D.J., Ingram, R.S., Levorse, J.M. and Tilghman, S.M. (2000) CTCF mediates methylation-sensitive enhancer-blocking activity at the H19/Igf2 locus. *Nature*, **405**, 486-489.
7. Recillas-Targa, F., Pikaart, M.J., Burgess-Beusse, B., Bell, A.C., Litt, M.D., West, A.G., Gaszner, M. and Felsenfeld, G. (2002) Position-effect protection and enhancer blocking by the chicken beta-globin insulator are separable activities. *Proceedings of the National Academy of Sciences of the United States of America*, **99**, 6883-6888.
8. Schmidt, D., Schwalie, P.C., Wilson, M.D., Ballester, B., Goncalves, A., Kutter, C., Brown, G.D., Marshall, A., Flicek, P. and Odom, D.T. (2012) Waves of retrotransposon expansion remodel genome organization and CTCF binding in multiple mammalian lineages. *Cell*, **148**, 335-348.
9. Wang, H., Maurano, M.T., Qu, H., Varley, K.E., Gertz, J., Pauli, F., Lee, K., Canfield, T., Weaver, M., Sandstrom, R. *et al.* (2012) Widespread plasticity in CTCF occupancy linked to DNA methylation. *Genome research*, **22**, 1680-1688.
10. Shen, Y., Yue, F., McCleary, D.F., Ye, Z., Edsall, L., Kuan, S., Wagner, U., Dixon, J., Lee, L., Lobanenkov, V.V. *et al.* (2012) A map of the cis-regulatory sequences in the mouse genome. *Nature*.
11. Kim, T.H., Abdullaev, Z.K., Smith, A.D., Ching, K.A., Loukinov, D.I., Green, R.D., Zhang, M.Q., Lobanenkov, V.V. and Ren, B. (2007) Analysis of the vertebrate insulator protein CTCF-binding sites in the human genome. *Cell*, **128**, 1231-1245.
12. Splinter, E., Heath, H., Kooren, J., Palstra, R.J., Klous, P., Grosveld, F., Galjart, N. and de Laat, W. (2006) CTCF mediates long-range chromatin looping and local histone modification in the beta-globin locus. *Genes & development*, **20**, 2349-2354.
13. Handoko, L., Xu, H., Li, G., Ngan, C.Y., Chew, E., Schnapp, M., Lee, C.W., Ye, C., Ping, J.L., Mulawadi, F. *et al.* (2011) CTCF-mediated functional chromatin interactome in pluripotent cells. *Nature genetics*, **43**, 630-638.
14. Guelen, L., Pagie, L., Brasset, E., Meuleman, W., Faza, M.B., Talhout, W., Eussen, B.H., de Klein, A., Wessels, L., de Laat, W. *et al.* (2008) Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions. *Nature*, **453**, 948-951.
15. Dixon, J.R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., Hu, M., Liu, J.S. and Ren, B. (2012) Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, **485**, 376-380.
16. Chung, J.H., Whiteley, M. and Felsenfeld, G. (1993) A 5' element of the chicken beta-globin domain serves as an insulator in human erythroid cells and protects against position effect in Drosophila. *Cell*, **74**, 505-514.
17. Farrell, C.M., West, A.G. and Felsenfeld, G. (2002) Conserved CTCF insulator elements flank the mouse and human beta-globin loci. *Molecular and cellular biology*, **22**, 3820-3831.
18. Bulger, M., Schubeler, D., Bender, M.A., Hamilton, J., Farrell, C.M., Hardison, R.C. and Groudine, M. (2003) A complex chromatin landscape revealed by patterns of nuclease sensitivity and histone modification within the mouse beta-globin locus. *Molecular and cellular biology*, **23**, 5234-5244.
19. Bender, M.A., Byron, R., Ragoczy, T., Telling, A., Bulger, M. and Groudine, M. (2006) Flanking HS-62.5 and 3' HS1, and regions upstream of the LCR, are not required for beta-globin transcription. *Blood*, **108**, 1395-1401.
20. Palstra, R.J., Tolhuis, B., Splinter, E., Nijmeijer, R., Grosveld, F. and de Laat, W. (2003) The beta-globin nuclear

compartment in development and erythroid differentiation. *Nature genetics*, **35**, 190-194.
21. Bartolomei, M.S., Webber, A.L., Brunkow, M.E. and Tilghman, S.M. (1993) Epigenetic mechanisms underlying the imprinting of the mouse H19 gene. *Genes & development*, **7**, 1663-1673.
22. Ferguson-Smith, A.C., Sasaki, H., Cattanach, B.M. and Surani, M.A. (1993) Parental-origin-specific epigenetic modification of the mouse H19 gene. *Nature*, **362**, 751-755.
23. Bell, A.C. and Felsenfeld, G. (2000) Methylation of a CTCF-dependent boundary controls imprinted expression of the Igf2 gene. *Nature*, **405**, 482-485.
24. Murrell, A., Heeson, S. and Reik, W. (2004) Interaction between differentially methylated regions partitions the imprinted genes Igf2 and H19 into parent-specific chromatin loops. *Nature genetics*, **36**, 889-893.
25. Kurukuti, S., Tiwari, V.K., Tavoosidana, G., Pugacheva, E., Murrell, A., Zhao, Z., Lobanenkov, V., Reik, W. and Ohlsson, R. (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**, 10684-10689.
26. Yoon, Y.S., Jeong, S., Rong, Q., Park, K.Y., Chung, J.H. and Pfeifer, K. (2007) Analysis of the H19ICR insulator. *Molecular and cellular biology*, **27**, 3499-3510.
27. Barski, A., Cuddapah, S., Cui, K., Roh, T.Y., Schones, D.E., Wang, Z., Wei, G., Chepelev, I. and Zhao, K. (2007) High-resolution profiling of histone methylations in the human genome. *Cell*, **129**, 823-837.
28. Essien, K., Vigneau, S., Apreleva, S., Singh, L.N., Bartolomei, M.S. and Hannenhalli, S. (2009) CTCF binding site classes exhibit distinct evolutionary, genomic, epigenomic and transcriptomic features. *Genome biology*, **10**, R131.
29. Stadler, M.B., Murr, R., Burger, L., Ivanek, R., Lienert, F., Scholer, A., van Nimwegen, E., Wirbelauer, C., Oakeley, E.J., Gaidatzis, D. *et al.* (2011) DNA-binding factors shape the mouse methylome at distal regulatory regions. *Nature*, **480**, 490-495.
30. Yu, W., Ginjala, V., Pant, V., Chernukhin, I., Whitehead, J., Docquier, F., Farrar, D., Tavoosidana, G., Mukhopadhyay, R., Kanduri, C. *et al.* (2004) Poly(ADP-ribosyl)ation regulates CTCF-dependent chromatin insulation. *Nature genetics*, **36**, 1105-1110.
31. Guastafierro, T., Cecchinelli, B., Zampieri, M., Reale, A., Riggio, G., Sthandier, O., Zupi, G., Calabrese, L. and Caiafa, P. (2008) CCCTC-binding factor activates PARP-1 affecting DNA methylation machinery. *The Journal of biological chemistry*, **283**, 21873-21880.
32. Zampieri, M., Guastafierro, T., Calabrese, R., Ciccarone, F., Bacalini, M.G., Reale, A., Perilli, M., Passananti, C. and Caiafa, P. (2012) ADP-ribose polymers localized on Ctcf-Parp1-Dnmt1 complex prevent methylation of Ctcf target sites. *The Biochemical journal*, **441**, 645-652.
33. Cuddapah, S., Jothi, R., Schones, D.E., Roh, T.Y., Cui, K. and 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**, 24-32.
34. Van Bortle, K., Ramos, E., Takenaka, N., Yang, J., Wahi, J.E. and Corces, V.G. (2012) Drosophila CTCF tandemly aligns with other insulator proteins at the borders of H3K27me3 domains. *Genome research*, **22**, 2176-2187.
35. Peric-Hupkes, D., Meuleman, W., Pagie, L., Bruggeman, S.W., Solovei, I., Brugman, W., Graf, S., Flicek, P., Kerkhoven, R.M., van Lohuizen, M. *et al.* (2010) Molecular maps of the reorganization of genome-nuclear lamina interactions during differentiation. *Molecular cell*, **38**, 603-613.
36. Vostrov, A.A. and 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**, 33353-33359.
37. Wendt, K.S., Yoshida, K., Itoh, T., Bando, M., Koch, B., Schirghuber, E., Tsutsumi, S., Nagae, G., Ishihara, K., Mishiro, T. *et al.* (2008) Cohesin mediates transcriptional insulation by CCCTC-binding factor. *Nature*, **451**, 796-801.
38. Rubio, E.D., Reiss, D.J., Welcsh, P.L., Disteche, C.M., Filippova, G.N., Baliga, N.S., Aebersold, R., Ranish, J.A. and Krumm, A. (2008) CTCF physically links cohesin to chromatin. *Proceedings of the National Academy of Sciences of the United States of America*, **105**, 8309-8314.
39. Parelho, V., Hadjur, S., Spivakov, M., Leleu, M., Sauer, S., Gregson, H.C., Jarmuz, A., Canzonetta, C., Webster, Z., Nesterova, T. *et al.* (2008) Cohesins functionally associate with CTCF on mammalian chromosome arms. *Cell*, **132**, 422-433.
40. Stedman, W., Kang, H., Lin, S., Kissil, J.L., Bartolomei, M.S. and Lieberman, P.M. (2008) Cohesins localize with CTCF at the KSHV latency control region and at cellular c-myc and H19/Igf2 insulators. *The EMBO journal*, **27**, 654-666.
41. Xiao, T., Wallace, J. and 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**, 2174-2183.
42. Nasmyth, K. and Haering, C.H. (2009) Cohesin: its roles and mechanisms. *Annual review of genetics*, **43**, 525-558.
43. Nativio, R., Wendt, K.S., Ito, Y., Huddleston, J.E., Uribe-Lewis, S., Woodfine, K., Krueger, C., Reik, W., Peters, J.M. and Murrell, A. (2009) Cohesin is required for higher-order chromatin conformation at the imprinted IGF2-H19 locus. *PLoS genetics*, **5**, e1000739.
44. Hadjur, S., Williams, L.M., Ryan, N.K., Cobb, B.S., Sexton, T., Fraser, P., Fisher, A.G. and Merkenschlager, M. (2009) Cohesins form chromosomal cis-interactions at the developmentally regulated IFNG locus. *Nature*, **460**,

410-413.
45. Chien, R., Zeng, W., Kawauchi, S., Bender, M.A., Santos, R., Gregson, H.C., Schmiesing, J.A., Newkirk, D.A., Kong, X., Ball, A.R., Jr. *et al.* (2011) Cohesin mediates chromatin interactions that regulate mammalian beta-globin expression. *The Journal of biological chemistry*, **286**, 17870-17878.
46. Seitan, V.C., Hao, B., Tachibana-Konwalski, K., Lavagnolli, T., Mira-Bontenbal, H., Brown, K.E., Teng, G., Carroll, T., Terry, A., Horan, K. *et al.* (2011) A role for cohesin in T-cell-receptor rearrangement and thymocyte differentiation. *Nature*, **476**, 467-471.
47. Kagey, M.H., Newman, J.J., Bilodeau, S., Zhan, Y., Orlando, D.A., van Berkum, N.L., Ebmeier, C.C., Goossens, J., Rahl, P.B., Levine, S.S. *et al.* (2010) Mediator and cohesin connect gene expression and chromatin architecture. *Nature*, **467**, 430-435.
48. Schmidt, D., Schwalie, P.C., Ross-Innes, C.S., Hurtado, A., Brown, G.D., Carroll, J.S., Flicek, P. and Odom, D.T. (2010) A CTCF-independent role for cohesin in tissue-specific transcription. *Genome research*, **20**, 578-588.
49. Faure, A.J., Schmidt, D., Watt, S., Schwalie, P.C., Wilson, M.D., Xu, H., Ramsay, R.G., Odom, D.T. and Flicek, P. (2012) Cohesin regulates tissue-specific expression by stabilizing highly occupied cis-regulatory modules. *Genome research*.
50. Wallace, J.A. and Felsenfeld, G. (2007) We gather together: insulators and genome organization. *Current opinion in genetics & development*, **17**, 400-407.
51. Zlatanova, J. and Caiafa, P. (2009) CTCF and its protein partners: divide and rule? *Journal of cell science*, **122**, 1275-1284.
52. Lee, B.K. and Iyer, V.R. (2012) Genome-wide studies of CCCTC-binding factor (CTCF) and cohesin provide insight into chromatin structure and regulation. *The Journal of biological chemistry*, **287**, 30906-30913.
53. Lutz, M., Burke, L.J., Barreto, G., Goeman, F., Greb, H., Arnold, R., Schultheiss, H., Brehm, A., Kouzarides, T., Lobanenkov, V. *et al.* (2000) Transcriptional repression by the insulator protein CTCF involves histone deacetylases. *Nucleic acids research*, **28**, 1707-1713.
54. Lutz, M., Burke, L.J., LeFevre, P., Myers, F.A., Thorne, A.W., Crane-Robinson, C., Bonifer, C., Filippova, G.N., Lobanenkov, V. and Renkawitz, R. (2003) Thyroid hormone-regulated enhancer blocking: cooperation of CTCF and thyroid hormone receptor. *The EMBO journal*, **22**, 1579-1587.
55. Yusufzai, T.M., Tagami, H., Nakatani, Y. and Felsenfeld, G. (2004) CTCF tethers an insulator to subnuclear sites, suggesting shared insulator mechanisms across species. *Molecular cell*, **13**, 291-298.
56. Defossez, P.A., Kelly, K.F., Filion, G.J., Perez-Torrado, R., Magdinier, F., Menoni, H., Nordgaard, C.L., Daniel, J.M. and Gilson, E. (2005) The human enhancer blocker CTC-binding factor interacts with the transcription factor Kaiso. *The Journal of biological chemistry*, **280**, 43017-43023.
57. Yao, H., Brick, K., Evrard, Y., Xiao, T., Camerini-Otero, R.D. and Felsenfeld, G. (2010) Mediation of CTCF transcriptional insulation by DEAD-box RNA-binding protein p68 and steroid receptor RNA activator SRA. *Genes & development*, **24**, 2543-2555.
58. Ross-Innes, C.S., Brown, G.D. and Carroll, J.S. (2011) A co-ordinated interaction between CTCF and ER in breast cancer cells. *BMC genomics*, **12**, 593.
59. Liu, Z., Scannell, D.R., Eisen, M.B. and Tjian, R. (2011) Control of embryonic stem cell lineage commitment by core promoter factor, TAF3. *Cell*, **146**, 720-731.
60. Wada, Y., Ohta, Y., Xu, M., Tsutsumi, S., Minami, T., Inoue, K., Komura, D., Kitakami, J., Oshida, N., Papantonis, A. *et al.* (2009) A wave of nascent transcription on activated human genes. *Proceedings of the National Academy of Sciences of the United States of America*, **106**, 18357-18361.
61. Shukla, S., Kavak, E., Gregory, M., Imashimizu, M., Shutinoski, B., Kashlev, M., Oberdoerffer, P., Sandberg, R. and Oberdoerffer, S. (2011) CTCF-promoted RNA polymerase II pausing links DNA methylation to splicing. *Nature*, **479**, 74-79.
62. Stadhouders, R., Thongjuea, S., Andrieu-Soler, C., Palstra, R.J., Bryne, J.C., van den Heuvel, A., Stevens, M., de Boer, E., Kockx, C., van der Sloot, A. *et al.* (2012) Dynamic long-range chromatin interactions control Myb proto-oncogene transcription during erythroid development. *The EMBO journal*, **31**, 986-999.
63. Majumder, P. and Boss, J.M. (2010) CTCF controls expression and chromatin architecture of the human major histocompatibility complex class II locus. *Molecular and cellular biology*, **30**, 4211-4223.
64. Majumder, P. and Boss, J.M. (2011) Cohesin regulates MHC class II genes through interactions with MHC class II insulators. *J Immunol*, **187**, 4236-4244.
65. Ribeiro de Almeida, C., Stadhouders, R., de Bruijn, M.J., Bergen, I.M., Thongjuea, S., Lenhard, B., van Ijcken, W., Grosveld, F., Galjart, N., Soler, E. *et al.* (2011) The DNA-binding protein CTCF limits proximal Vkappa recombination and restricts kappa enhancer interactions to the immunoglobulin kappa light chain locus. *Immunity*, **35**, 501-513.
66. Seitan, V.C. and Merkenschlager, M. (2012) Cohesin and chromatin organisation. *Current opinion in genetics & development*, **22**, 93-100.
67. Soshnikova, N., Montavon, T., Leleu, M., Galjart, N. and Duboule, D. (2010) Functional analysis of CTCF during mammalian limb development. *Developmental cell*, **19**, 819-830.
68. Montavon, T., Soshnikova, N., Mascrez, B., Joye, E., Thevenet, L., Splinter, E., de Laat, W., Spitz, F. and Duboule, D. (2011) A regulatory archipelago controls Hox genes transcription in digits. *Cell*, **147**, 1132-1145.
69. Kehayova, P., Monahan, K., Chen, W. and Maniatis, T. (2011) Regulatory elements required for the activation and

repression of the protocadherin-alpha gene cluster. *Proceedings of the National Academy of Sciences of the United States of America*, **108**, 17195-17200.
70. Monahan, K., Rudnick, N.D., Kehayova, P.D., Pauli, F., Newberry, K.M., Myers, R.M. and Maniatis, T. (2012) Role of CCCTC binding factor (CTCF) and cohesin in the generation of single-cell diversity of protocadherin-alpha gene expression. *Proceedings of the National Academy of Sciences of the United States of America*, **109**, 9125-9130.
71. Hirayama, T., Tarusawa, E., Yoshimura, Y., Galjart, N. and Yagi, T. (2012) CTCF is required for neural development and stochastic expression of clustered Pcdh genes in neurons. *Cell reports*, **2**, 345-357.
72. Lieberman-Aiden, E., van Berkum, N.L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B.R., Sabo, P.J., Dorschner, M.O. *et al.* (2009) Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, **326**, 289-293.
73. Botta, M., Haider, S., Leung, I.X., Lio, P. and Mozziconacci, J. (2010) Intra- and inter-chromosomal interactions correlate with CTCF binding genome wide. *Molecular systems biology*, **6**, 426.
74. Fullwood, M.J., Liu, M.H., Pan, Y.F., Liu, J., Xu, H., Mohamed, Y.B., Orlov, Y.L., Velkov, S., Ho, A., Mei, P.H. *et al.* (2009) An oestrogen-receptor-alpha-bound human chromatin interactome. *Nature*, **462**, 58-64.
75. Sanyal, A., Lajoie, B.R., Jain, G. and Dekker, J. (2012) The long-range interaction landscape of gene promoters. *Nature*, **489**, 109-113.
76. Yaffe, E. and Tanay, A. (2011) Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture. *Nature genetics*, **43**, 1059-1065.
77. Comet, I., Schuettengruber, B., Sexton, T. and Cavalli, G. (2011) A chromatin insulator driving three-dimensional Polycomb response element (PRE) contacts and Polycomb association with the chromatin fiber. *Proceedings of the National Academy of Sciences of the United States of America*, **108**, 2294-2299.
78. Nora, E.P., Lajoie, B.R., Schulz, E.G., Giorgetti, L., Okamoto, I., Servant, N., Piolot, T., van Berkum, N.L., Meisig, J., Sedat, J. *et al.* (2012) Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature*, **485**, 381-385.
79. Lin, Y.C., Benner, C., Mansson, R., Heinz, S., Miyazaki, K., Miyazaki, M., Chandra, V., Bossen, C., Glass, C.K. and Murre, C. (2012) Global changes in the nuclear positioning of genes and intra- and interdomain genomic interactions that orchestrate B cell fate. *Nature immunology*.
80. Lan, X., Witt, H., Katsumura, K., Ye, Z., Wang, Q., Bresnick, E.H., Farnham, P.J. and Jin, V.X. (2012) Integration of Hi-C and ChIP-seq data reveals distinct types of chromatin linkages. *Nucleic acids research*, **40**, 7690-7704.
81. Lin, Y.C., Benner, C., Mansson, R., Heinz, S., Miyazaki, K., Miyazaki, M., Chandra, V., Bossen, C., Glass, C.K. and Murre, C. (2012) Global changes in the nuclear positioning of genes and intra- and interdomain genomic interactions that orchestrate B cell fate. *Nature immunology*, **13**, 1196-1204.

2

3
Allelic exclusion of the immunoglobulin heavy chain locus is independent of its nuclear localization in mature B cells

Nucleic Acids Res. 2013 Jun 7

Allelic exclusion of the immunoglobulin heavy chain locus is independent of its nuclear localization in mature B cells

Sjoerd J.B Holwerda[1], Harmen J.G van de Werken[1#], Claudia Ribeiro de Almeida[2#], Ingrid M. Bergen[2], Marjolein J.W. de Bruijn[2], Marjon J. Verstegen[1], Marieke Simonis[1], Erik Splinter[1#], Patrick J. Wijchers[1], Rudi W. Hendriks[2*], Wouter de Laat[1*]

[1] Hubrecht Institute-KNAW & University Medical Center Utrecht, Utrecht, 3584 CT, The Netherlands
[2] Department of Pulmonary Medicine, Erasmus MC Rotterdam, Rotterdam, 3000 CA, The Netherlands

[#] Current addresses: [Harmen J.G van de Werken], Department of Cell Biology, Erasmus MC Rotterdam, PO Box 2040, 3000 CA Rotterdam, The Netherlands; [Claudia Ribeiro de Almeida], Sir William Dunn School of Pathology, University of Oxford, South Parks Road, Oxford OX1 3RE, UK; [Erik Splinter], Cergentis B.V, Padualaan 8, 3584 CH Utrecht, The Netherlands

[*] Corresponding authors: Email:
WL: w.delaat@hubrecht.eu
RW: r.hendriks@erasmusmc.nl

ABSTRACT

In developing B cells the immunoglobulin heavy chain (IgH) locus is thought to move from repressive to permissive chromatin compartments to facilitate its scheduled rearrangement. In mature B cells, maintenance of allelic exclusion has been proposed to involve recruitment of the non-productive IgH allele to pericentromeric heterochromatin. Here we used an allele-specific chromosome conformation capture combined with sequencing (4C-seq) approach to unambigously follow the individual IgH alleles in mature B lymphocytes. Despite their physical and functional difference, productive and non-productive IgH alleles in B cells and unrearranged IgH alleles in T cells share many chromosomal contacts and largely reside in active chromatin. In brain, however, the locus resides in a different, repressive environment. We conclude that IgH adopts a lymphoid-specific nuclear location that is however unrelated to maintenance of allelic exclusion. We additionally find that in mature B cells - but not in T cells – the distal V_H regions of both IgH alleles position themselves away from active chromatin. This, we speculate, may help to restrict enhancer activity to the productively rearranged V_H promoter element.

Introduction

B and T lymphocytes express a large repertoire of antigen receptors that safeguard the robustness of our adaptive immune response. Lymphocyte development uniquely relies on scheduled genomic rearrangement of V (variable), D (diversity) and J (joining) gene segments in the antigen receptor loci (1-3).
The murine *IgH* locus spans nearly ~3 Mb, with upstream ~150 functional V_H segments spread over ~2.4 Mb, followed by D_H and J_H segments and a ~200 kb constant (C_H) gene region. V(D)J recombination,

initiated by the recombination activating gene-1 (Rag1) and Rag2 proteins, is regulated at three different levels: (i) cell lineage-specificity, (ii) temporal order within a lineage and (iii) allelic exclusion, which is the mechanism that guarantees that only one receptor is expressed per lymphocyte (2-4). The *IgH* locus contains many *cis*-regulatory elements, including the intergenic control region 1 (IGCR1), the intronic enhancer Eμ, and the downstream 3' regulatory region (3'RR), which are involved in the regulation of the of V(D)J recombination (5-7) and class switch recombination (8). Chromosome topology and nuclear location have been implicated in the control of V(D)J recombination and allelic exclusion (3,9-11). In the early pro-B stage the *IgH* locus adopts a central position in the nuclear interior and chromatin looping mediates physical proximity of both ends of the locus (12,13), facilitating recombination of distal V_H genes (13-16). Succesfull D_H-to-J_H recombination on both alleles is followed by productive V_H to D_HJ_H recombination on only one allele. Prohibition of further rearrangement of the other allele, called allelic exclusion, is thought to be controlled by multiple (partly) redundant and successive mechanisms (17). In pre-B cells, upon successful V(D)J rearrangement both *IgH* loci decontract and the non-productive allele is seen to relocate to pericentromeric heterochromatin (PCH) (15). No heterochromatin tethering was observed in early pro-B cells prior to rearrangement, nor in resting splenic B cells, suggesting that mono-allelic recruitment to heterochromatin is developmentally controlled (18). Only upon activation of splenic B cells, mono-allelic *IgH* recruitment to PCH appears to re-occur (18). Mono-allelic expression was reported to take place preferentially from the non-associated allele, suggesting that recruitment to heterochromatin helps to maintain silencing of the non-productive *IgH* allele (18). In contrast with these findings it has also been reported that activated splenic B cells transcribe both *IgH* alleles (19). To what extent the two *IgH* alleles in mature B cells differ therefore remains unclear.

While FISH enables studying locus positioning at the single cell level, it is limited in throughput and provides relatively low resolution spatial information. Chromosome conformation capture (3C) technology (20) has been applied to study *IgH* locus conformation in more detail. 3C revealed two major contacts in the unrearranged *IgH* locus, one between Eμ and 3'RR, and the other between Eμ and IGCR1 (5,21). The CCCTC-binding factor CTCF (22) and cohesin were implicated in these loops, which appear to create a topological subdomain that covers the region from 3'RR to IGCR1 (5,21). The proximal and distal V_H region also adopt distinct topological substructures that then merge with the 3' domain to maximize D_HJ_H contacts with the full V_H gene repertoire (16,23). Thus, in early B cell development *IgH* topology ensures that proximal and distal V_H genes have equal opportunites to interact with Eμ. In mature B cells that have completed V(D)J recombination, however, the chromatin structure of *IgH* is expected to be different, since promiscuous interactions of Eμ with numerous upstream V_H promoters may interefere with accurate and efficient transcription from the functionally rearranged V_H promoter.

In this study, we characterized the structural properties and genomic environments of the productive and non-productive *IgH* allele separately. We applied allele-specific 4C-seq (24,25) to compare at high resolution the chromatin configuration of the productive and non-productive *IgH* alleles in mature B cells, as well as the unrearranged *IgH* alleles in T cells and non-lymphoid cells. We also evaluated *IgH* nuclear positioning, as determined by the genomic contacts formed by these alleles.

Results

Both IgH alleles are transcribed and positioned similarly in resting and activated splenic B cells

To obtain a pure B cell population from spleen, we used the cell sorting strategy described in (18) but included additional markers to further exclude non-B cells (see materials and methods). Because recruitment of non-productive *IgH* alleles to PCH was observed in activated but not in resting B cells (18),

we performed 4C-Seq experiments both in resting and anti-CD40-activated splenic B cells. Proper activation of splenic B cells was verified by gene profiling, demonstrating upregulation of key genes including *Aicda*, *IL-5R*, *CD44*, *Fas*, *c-Myc* and *cyclin D2* (**Suppl. Table 2**).
First, we tested *IgH* expression by RNA-FISH with a BAC-probe spanning the Cμ-J_H-D region. Both in resting and in activated B cells, we detected biallelic expression in ~75-80% of cells (**Figure 1A, 1B**). RNase-treated cells showed no signals, demonstrating that we were measuring RNA (**Suppl. Figure 1**). These results supported previously published work in which biallelic expression (19,30,31), active chromatin marks and RNA polymerase II binding (30) were found to be associated with both productive and non-productive *IgH* alleles in mature B cells.
We then asked whether B cell stimulation is accompanied by allele-specific repositioning of the non-productive *IgH* locus to PCH, as previously reported (18). We performed DNA FISH and measured *IgH* distances relative to PCH, as stained for with a g-satellite probe. To discriminate productive from non-productive alleles, we visualized both *IgH* ends and assumed that the most contracted locus represented the productive allele (15). In resting B cells none of the *IgH* alleles showed striking PCH proximity (**Figure 1C, 1D**). In stimulated B cells, the non-productive locus appeared a bit more frequently near PCH than the productive allele, but physical contacts within the 300 nm range were rare for both alleles (**Figure 1E, 1F**). These microscopy studies therefore indicate that in resting and activated splenic B cells both *IgH* loci are transcribed and that none of the two *IgH* loci are closely associated with PCH.

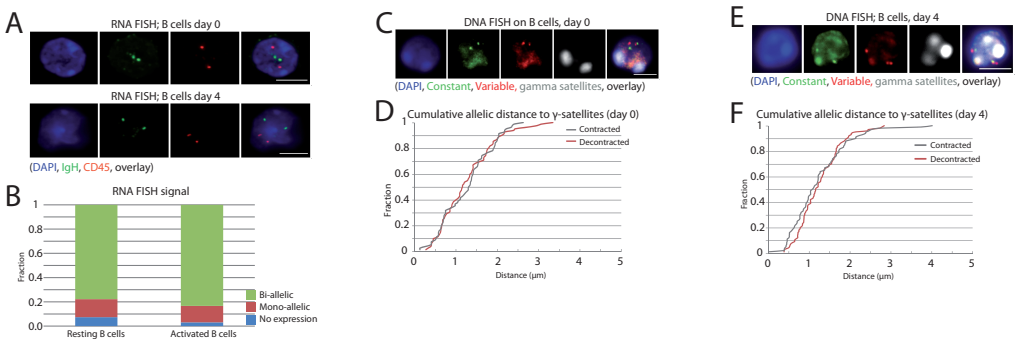


Figure 1. Biallelic expression and comparable nuclear positioning of the two IgH loci relative to PCH in resting and activated B cells. (A) Representative picture of RNA FISH in resting (upper row) and activated B cells (lower row). (B) Quantification of the RNA FISH data plotted as the percentage of IgH signals in CD45+ cells on the Y-axis. A minimum of 50 cells were analyzed per cell type. Cells with >2 RNA signals for IgH were excluded from the analysis. (C) Representative picture of DNA FISH in resting B cells. (D) Cumulative frequency of the minimal distance of IgH signals to the γ-satelite FISH signal in resting B cells. Contraction is defined as the minimal distance between two different probes on the IgH locus. The distance of the IgH signals to γ-satelites in μm is depicted on the X-axis. (E) Representative picture of DNA FISH in activated B cells. (F) Cumulative frequency of the minimal distance of the IgH signals to the γ-satelite FISH signal in activated B cells. FISH pictures represent several Z-stacks projected on top of each other, scale bar in overlays depicts 4μm. Images were collected using a Leica DM6000 B microscope equipped with a 100X objective, Leica DFC360 FX camera, taking z-steps of 0.2μm. Leica application suite 2.6.0 software was used both for image collection and deconvolution.

Allele-specific 4C-seq strategy to analyze productive and non-productive IgH alleles

To identify cells exclusively expressing the paternally or maternally derived *IgH* allele, we took advantage of IgM allotype differences: B cells from FVB or C57BL/6 mice produce heavy chains of the IgM^a or IgM^b allotype, respectively, which differ in a single amino acid (32). We employed allotype-specific antibodies in fluorescence activated cell sorting (FACS) to sort separate pools of IgM^a and IgM^b expressing splenic B cells from (FVB[IgM^a] X C57BL/6[IgM^b]) F1 mice. Both resting and α-CD40-activated splenic B cell fractions were sorted into two populations: one with cells carrying a productive FVB[IgM^a] and a non-productive C57BL/6 allele, and another with cells carrying a productive C57BL/6[IgM^b] and a non-productive FVB allele (**Figure 2A**).

To independently analyze the topology of these two functionally and physically different IgH alleles, we used allele-specific 4C-seq technology. 4C-seq enables the generation of genome-wide DNA contact profiles of a chromosomal sequence of interest, called the 'viewpoint' (24,25). In this allele-specific 4C-seq variant, we took advantage of C57BL/6 or FVB haplotype-specific single nucleotide polymorphisms (SNPs). We designed a strategy based on paired-end (PE) sequencing, whereby PE1 analyzes 4C ligation products and therefore identifies DNA contact partners, while PE2 reads a SNP inside the 'viewpoint' fragment and therefore links the PE1 contact profile to either the C57BL/6 or the FVB allele. Thus, the paired-end 4Cseq

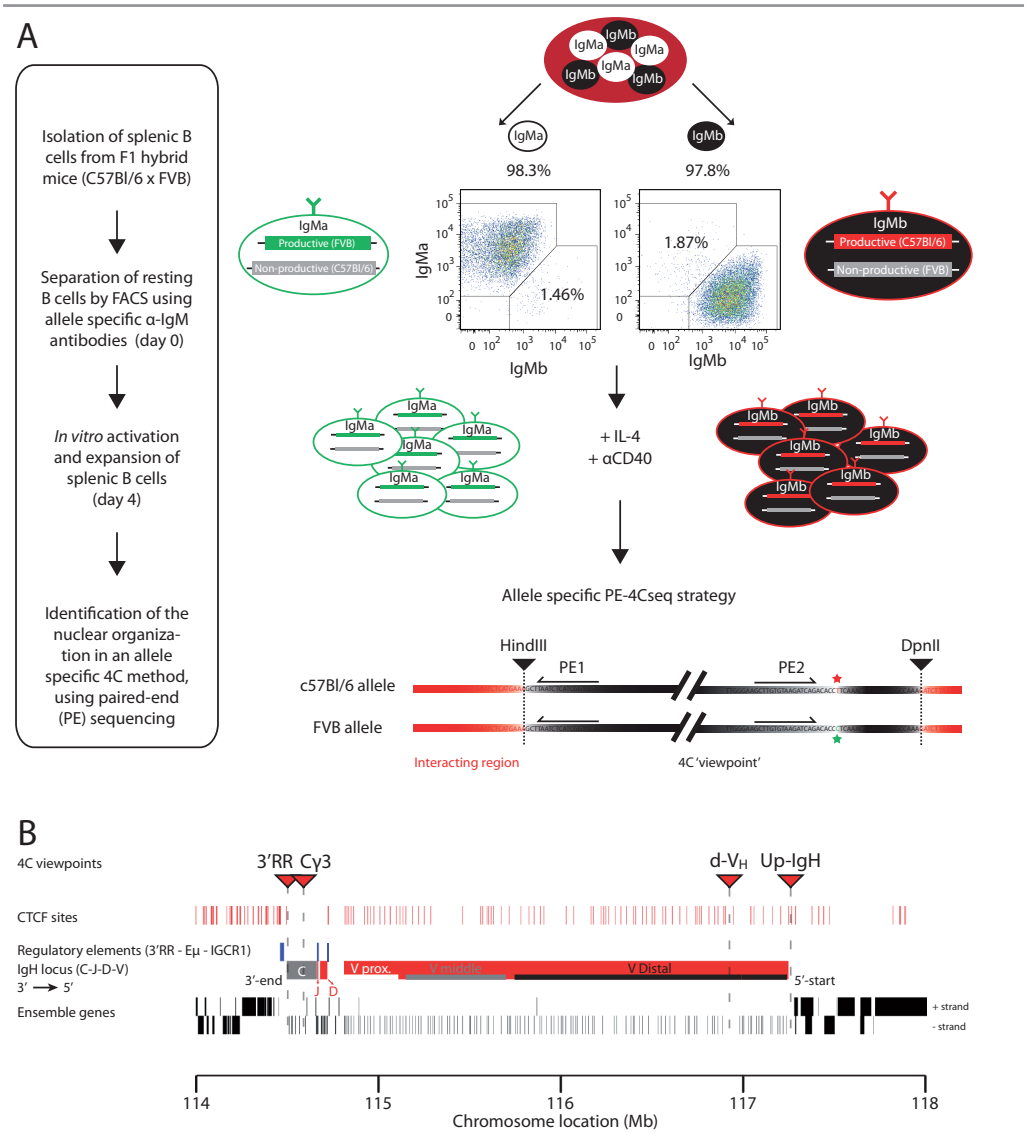


Figure 2. Allele specific 4C-seq strategy allowing separate analysis of productive and non-productive IgH alleles in B cells. (A) Schematic view of the experimental approach. The purity of the IgMa/IgMb-allotype-sorted populations is depicted above the FACS plots. The allele specific 4C-seq strategy shows a schematic 4C 'viewpoint'. The restriction sites are depicted in black triangles, the red and green stars highlight the SNP between the C57BL/6 and the FVB allele, respectively. The 4C viewpoint (black) with its captured interactions (red) is not drawn to scale. (B) Schematic view of the IgH locus highlighting the 4C viewpoints (red triangles), CTCF sites in mature B cells (GEO accession: GSM672402), regulatory elements (blue bars), the C region (grey), the JH, D and VH regions (red), and the Ensemble genes (black bars) on mouse chromosome 12 (mm9). The proximal (Vprox), middle (Vmiddle) and distal (Vdistal) region of the variable region are indicated with white, grey and black bars, respectively. The locus is drawn to scale. Mb= Megabase.

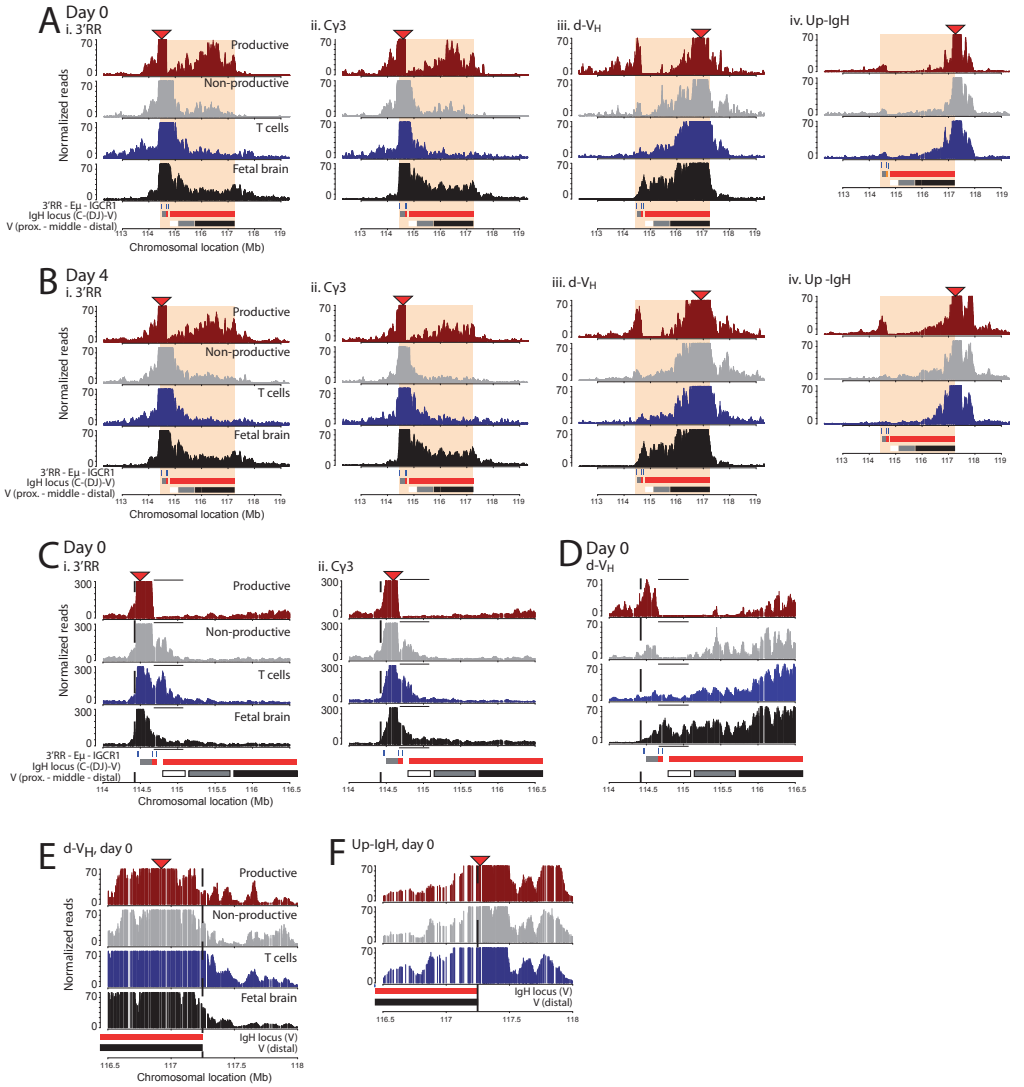


Figure 3. 4C-seq profiles reveal allele specific contacts in B cells and high-resolution definition of the topological domain spanning the IgH locus. (A) Contact profiles in resting B cells for the C57BL/6 allele across the IgH locus looking from four different 4C viewpoints depicted by red triangles, from left to right: 3'RR (i), Cγ3 (ii), d-VH (iii) and Up-IgH (iv). Per viewpoint the different tissues are plotted (from top to bottom): B cell productive allele (red) and non-productive allele (grey), T cell (blue) and fetal brain (black). A schematic representation showing the regulatory elements (blue), the C (grey) and V(red) regions of the IgH locus and the different regions within the V-region (proximal – middle – distal), as well as the chromosomal location in Megabase (Mb) is given at the bottom. The topological domain is depicted in shaded red. The Y-axis represents the normalized captured sequencing reads analyzed with a running median of 21 HindIII fragments, in arbitrary units. (B) Contact profiles in activated B cells. (C) Zoom of the 3'-border (dotted line) of the topological domain looking from the d-VH 4C viewpoint in resting splenic B cells. The black line above each track indicates the region where the productive allele in B cells has only few interactions compared to the other alleles. (D) Zooms of the 3'-border (dotted line) of the topological domain looking from the 3'RR (i) and the Cγ3 (ii) 4C viewpoints (red triangles). (E) Zoom of the 5'-border (dotted line) of the topological domain looking from the d-VH (i) and the Up-IgH (ii) 4C viewpoints (red triangles) in resting splenic B cells.

strategy (PE-4Cseq) enables independent but simultaneous analysis of both alleles, which is different from a previously developed method that employs single-end 4Cseq (SE-4Cseq) to analyze only one of the two alleles in a cell population (25). Three PE-4Cseq viewpoints were designed: 3'RR, near the upstream 3' regulatory region; Cγ3, inside the Cγ3-region and d-V_H, in the distal-V_HJ558 region. A fourth viewpoint, Upstream-IgH (Up-IgH), at the 5' end of IgH just beyond the most distal V_H gene (**Figure 2B**), was used

for allele-specific analysis based on SE-4Cseq (**Suppl. Figure 2**). The B cell populations studied consist of cells with differently rearranged IgH alleles. Three of the four 4Cseq viewpoints reside outside the V(D)J rearranged part of IgH and therefore enable DNA contact assessment of both alleles independent of their rearrangement. Only the d-V_H viewpoint resides just inside the distal-V_HJ558 region and may therefore miss a few rearranged alleles.

Topology of the IgH locus

We generated DNA contact profiles in resting and aCD40-activated splenic B cells, resting and aCD3-activated splenic T cells as well as fetal brain cells (serving as a non-lymphoid control). All 4C-seq profiles showed the typical contact distribution expected from polymer physics, with high contact frequencies between sequences close on the linear chromosome and with intrachromosomal captures being preferred over interchromosomal contacts (**Suppl. Figure 3**) (29). C57BL/6- and FVB-specific 4C-seq profiles were essentially identical; we show C57BL/6-specific profiles, unless specified as in **Suppl. Fig 4**.
In all cell types and for all three internal IgH viewpoints, the most abundant local contacts appeared confined to the ~3 Mb IgH locus (**Figure 3A, 3B**). In B lymphocytes, contacts made between the two ends of the locus were particularly frequent (**Figure 3D**), which probably reflects close linear proximity as a consequence of V_H-to-DJ_H recombination. Rapid drops in contact frequencies suggestive of structural boundaries were seen at either end of the locus. The 3' IgH boundary is best appreciated by the 4C-seq plots of the two closest viewpoints, 3'RR and Cγ3. Both showed loss in contact frequencies just beyond the 3'RR (**Figure 3A, 3B, 3C**). The 5' IgH boundary is evident from the contact profiles of the two 5' viewpoints. d-V_H, inside the IgH locus, preferentially captured IgH sequences but shows a clear reduction in contacts just beyond the last V_H gene (**Figure 3A, 3B, 3E**). By contrast, Up-IgH, just outside the IgH locus, showed a strong preference to capture sequences further away from IgH and did not make frequent contacts within IgH (**Figure 3A, 3B, 3F**). These findings suggested that the IgH locus forms a spatially distinct entity in mature B cells, as was described for pro-B cells (16), that is similar to the previously described topological domains identified by Hi-C (33-36) (**Suppl. Figure 5**). This structural organization was identified in cycling and non-cycling B and T lymphocytes, as well as in fetal brain cells.

Recombination and chromatin looping in the IgH domain

Further inspection of 4C-seq profiles revealed additional tissue-specific and allele-specific structural features of the IgH locus. The most distinct conformation was adopted by the productive allele in B cells, whereby the 3' viewpoints showed a complex landscape of frequent contacts across the middle and distant V regions (**Figure 3A, 3B**). Strong peaks in this landscape did not appear to cover specific locations, e.g. they did not coincide with the regulatory Pax5-binding PAIR elements (37). The 3'RR and Cγ3 viewpoints showed very few interactions with the ~0.3 Mb region containing the D segments, the IGCR1 element and proximal V_H genes (**Figure 3A, 3B, 3C**). Also, when looked from d-V_H, extensive loss of 4C signals in resting and stimulated B cells was seen across a large region containing the very proximal V_H genes (**Figure 3A, 3B, 3D**). Signals were strongly reduced but not completely absent (see **Suppl. Figure 6** for underlying raw 4C-seq data). When looked from the same distal viewpoint (**Figure 3C**), reduced signal across the proximal V region was also seen at the non-productive IgH allele. However, 3'RR and C*γ*3 show frequent interactions with non-rearranged sequences in this proximal ~0.3 Mb region (**Figure 3C**). These findings would be in agreement with frequenct DJ_H-configurations on the non-productive allele (present in ~50% of B cells (38-41)), as well as frequent rearrangement to the very proximal V_H7183 family. The latter is conceivable, since it is known that V_H7183-family genes are preferentially rearranged, but selected against cellularly because

of their incompetence to form a pre-BCR (42,43). Interestingly, the IgH locus in T cells structurally best resembles the non-productive IgH allele in B cells, indicating close proximity of the 3'RR and Em regions and the proximal V_H genes in T cells (**Figure 3A-D**).

Collectively, these results validate that our approach truly analyzes the productive and non-productive allele separately, show that the locus forms a large topological domain in all cell types analyzed, and confirm that large scale chromosomal rearrangements take place specifically at the productive allele.

The IgH locus is in a similar chromatin compartment in B and T lymphocytes

Microscopy studies have suggested that the IgH locus switches between positions inside the cell nucleus, involving recruitment to nuclear periphery or PCH, in a cell-type and allele-specific manner (12,18). We reasoned that such different locations should result in different chromatin environments, which can be assessed based on the long-range intra- and interchromosomal contacts measured by 4C-seq.

Chromosome-wide contact profiles revealed preferred contacts with specific regions across chromosome 12 (the chromosome that contains IgH), both in resting and stimulated B lymphocytes (shown for 3'RR in

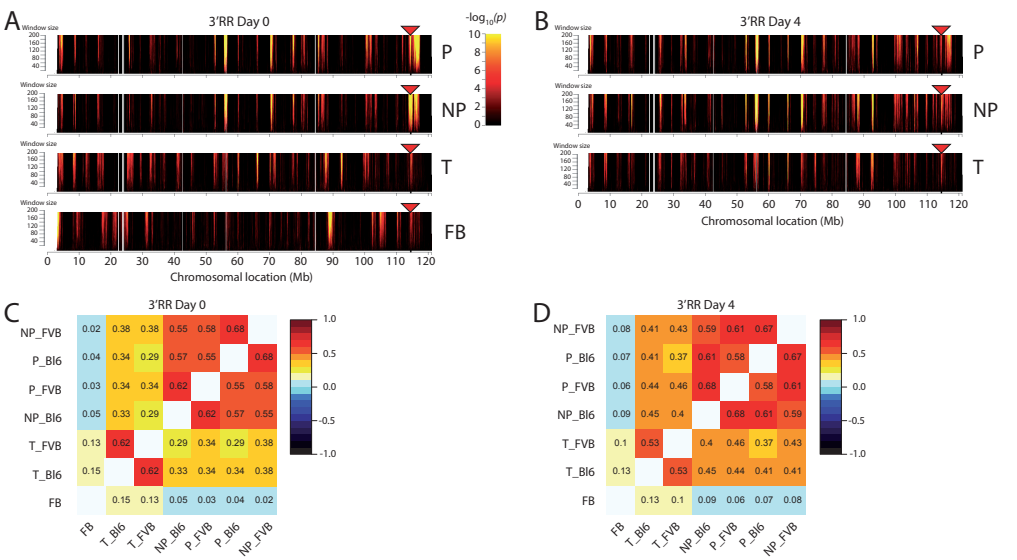


Figure 4. The IgH locus shows similar long range contacts along chromosome 12 between the different alleles in lymphocytes. (A) Domainograms showing chromosome-wide interaction profiles looking from the 3'RR 4C viewpoint (red triangle) on the IgH locus in resting B cells, from top to bottom: B cell productive (P), B cell non-productive (NP), T cell (T) and fetal brain allele (FB). Significance of the interactions is indicated by the range in color used in the domainogram as depicted in the legend: black is low significance (P = 1) and yellow represents high significance (P = 10-10) of interaction. Window size of the running window analysis is depicted on the Y-axis. (B) Chromosome-wide interaction profiles looking from the 3'RR 4C viewpoint (red triangle) in activated B cells. (C) Correlation plot of the interactions looking from the 3'RR 4C viewpoint in resting B cells (window size 21). The numbers represent the Spearman rank correlation coefficient, colors range from linear anti-correlation (black) to linear correlation (dark red). Tissues and separate alleles are coded as follows: B cell non-productive (NP), B cell productive (P), T cell (T), fetal brain (FB), C57BL/6 allele (Bl6) and FVB allele (FVB). (D) Correlation plot of the interactions looking from the 3'RR 4C viewpoint in activated B cells (window size 21).

Figure 4A and 4B, respectively; for FVB profiles see **Suppl. Figure 7**). DNA-FISH was performed to validate these results (**Suppl. Figure 8**). Correlation plots of the 4C results show that the productive and non-productive IgH alleles in B cells were engaged in very similar intra-chromosomal contacts, as apparent from all viewpoints (**Figure 4C, 4D; Suppl. Figures 9, 10**) and that many of these regions were also contacted in T cells. This is surprising, since the locus is thought to be differentially positioned in B and non-B cells (12). By contrast, in a non-lymphoid tissue, fetal brain, the IgH locus clearly made different contacts that even appeared mutually exclusive between brain and lymphocytes (**Figure 4A**). Not only intra-chromosomal, but also inter-chromosomal contacts corresponded between the productive and non-productive allele in B cells and those made by IgH in T cells, while the locus formed entirely distinct trans-contacts in fetal brain (**Suppl. Figure 11**).

Taken together, these data suggest that irrespective of its transcriptional or recombinational state, IgH is positioned in a similar chromatin compartment in splenic B and T lymphocytes, which is different from its chromatin environment in brain tissue.

The 5′ and 3′ end of the IgH locus are in different chromatin environments

Reported 3D FISH analyses indicated that - when recruited to heterochromatin - the IgH locus was oriented in such a way that the distal V_HJ558 gene family was positioned closer to the g-satellite cluster than the proximal V_H7183 or Cγ1 genes (15). It is therefore conceivable that 5' and 3' ends of IgH show different chromosome-wide contact profiles. Correlation plots of the total interactions in cis between the four individual viewpoints indeed demonstrated that in B lymphocytes chromosomal contacts formed by the 3' end of IgH (3'RR and Cγ3 viewpoints) were very different from those formed by its 5' end (the d-V_H viewpoint) (**Figure 5A** for activated B cells; **Suppl. Figure 12** for resting B cells). In these correlation analyses the productive and non-productive IgH allele did not differ. Intriguingly, contacts made by the region just upstream of IgH resembled those of the 3' viewpoints more than those made by its linearly close neighbor viewpoint d-V_H (**Figure 5A**). By contrast, in both T cells and brain cells, chromosomal contacts formed across the entire IgH locus were quite similar, with no exceptional profile seen for the distal V region (**Figure 5B-5C**).
Thus, specifically in B cells the IgH locus shows remarkable flexibility with the distal V_H regions of both IgH alleles being in a chromatin environment that is significantly different from either the 3'end of IgH or the upstream IgH flanking region.

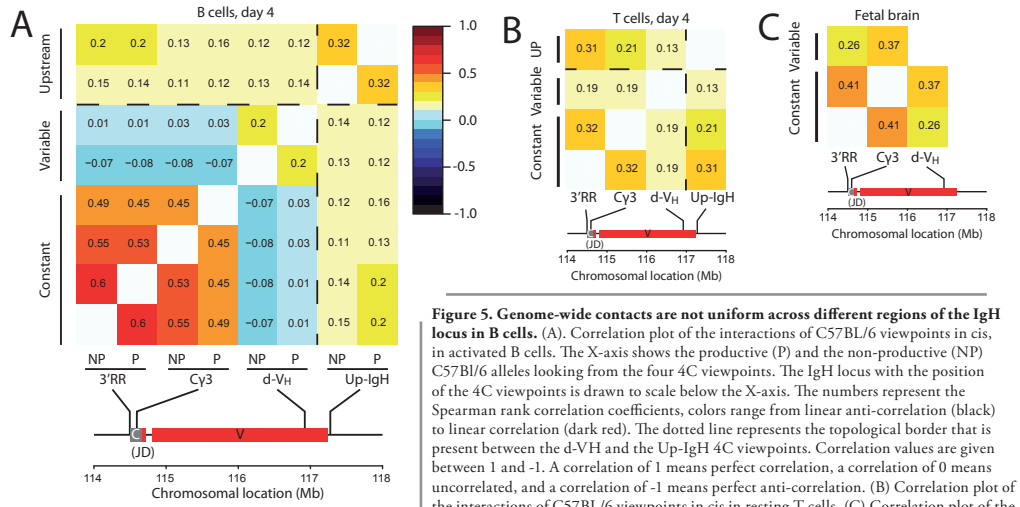


Figure 5. Genome-wide contacts are not uniform across different regions of the IgH locus in B cells. (A). Correlation plot of the interactions of C57BL/6 viewpoints in cis, in activated B cells. The X-axis shows the productive (P) and the non-productive (NP) C57Bl/6 alleles looking from the four 4C viewpoints. The IgH locus with the position of the 4C viewpoints is drawn to scale below the X-axis. The numbers represent the Spearman rank correlation coefficients, colors range from linear anti-correlation (black) to linear correlation (dark red). The dotted line represents the topological border that is present between the d-VH and the Up-IgH 4C viewpoints. Correlation values are given between 1 and -1. A correlation of 1 means perfect correlation, a correlation of 0 means uncorrelated, and a correlation of -1 means perfect anti-correlation. (B) Correlation plot of the interactions of C57BL/6 viewpoints in cis in resting T cells. (C) Correlation plot of the interactions made in fetal brain cells.

The distal V_H region of both IgH alleles is positioned away from active chromatin in B cells

To further characterize the *IgH* chromatin environments, we analyzed the transcriptional activity of regions contacted by *IgH*. We first analyzed the transcriptomes of our resting and stimulated splenic B and T cells in more depth. Hierarchical clustering confirmed the specific expression of B and T cell genes in the corresponding cell types and showed the upregulation of cell-cycle genes after stimulation (**Suppl. Figure**

13 and **Suppl. Tables 3-11**). We compared 4C-seq data with matched transcriptome data to analyze the transcriptional activity in contacted regions. In resting and stimulated splenic B cells, intra- and interchromosomal regions contacted by the 3' part of *IgH* showed relatively high transcriptional activity, when compared with non-contacted parts of the genome (**Figure 6A**). Surprisingly, this was true not only

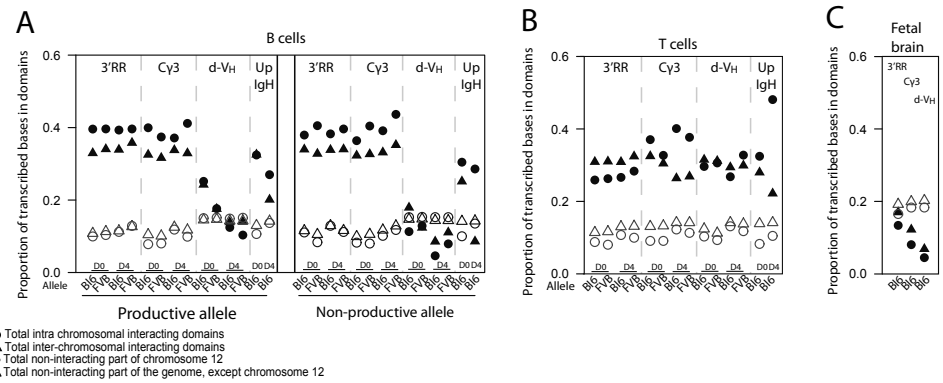


Figure 6. Genome-wide contacts of CH and VH regions in B cells show differences in transcriptional activity. (A) Quantification of the transcriptional activity of the interacting domains in cis (filled circles) and trans (filled triangles) contacted by four 4C viewpoints (indicated at the top) from the productive allele (red) and non-productive allele (grey) in resting and activated B cells. Viewpoints are separated by a dotted line, from left to right: 3'RR, Cγ3, d-VH and Up-IgH. Values for resting cells (D0) and activated cells (D4) are shown per viewpoint. Open circles and open triangles represent the transcriptional activity of the non-interacting regions in cis and trans, respectively. The proportion of transcribed bases in the interacting domains is depicted on the Y-axis. (B) Quantification of the transcriptional activity of the interacting domains in resting and activated T cells for the four different viewpoints. (C) Quantification of the transcriptional activity of the interacting domains in fetal brain for three different viewpoints.

for the productive, but also for the non-productive *IgH* allele in B cells, and even for the inactive *IgH* locus in T cells (**Figure 6A and 6B**). By contrast, regions contacted by *IgH* in brain were relatively inactive, when compared with the remainder of the genome (**Figure 6C**). Thus, these data show that the 3' part of the *IgH* locus switches from an inactive chromatin environment in brain cells to an active compartment in B and T lymphocytes, where it resides irrespective of its recombination and transcriptional status.
Flexibility of the *IgH* locus became further apparent when we analyzed the chromosomal contacts formed by the *d-V_H* and *Up-IgH* viewpoints. In resting and stimulated T cells, as well as in brain cells, chromosomal contacts were similar in transcriptional activity no matter whether they were assessed from the 3' or the 5' side of the locus (**Figure 6B, 6C**). These findings indicate that the entire *IgH* locus positions itself as a single entity in an inactive environment in brain cells and in an active compartment in T cells. Surprisingly, this was not the case in B cells: *d-V_H* did not necessarily contact active chromatin, as it did in T cells, but located to a more 'neutral' chromatin environment with chromosomal regions that were not different in transcriptional output from the remainder of the genome. Strikingly, this was only seen for *d-V_H*, since the *Up-IgH* viewpoint again contacted active chromosomal regions. This specific positioning was observed both for the productive and the non-productive *IgH* locus, and both in resting and stimulated B cells.
Thus, in mature B cells the two ends of both *IgH* loci position themselves in different chromatin environments, with the distal V_H region being positioned away from active chromosomal parts.

Discussion

The specificity of B cell responses during infection relies on extensive antibody diversity whereby each mature B cells bears a single unique type of B cell receptor. Monospecificity of B lymphocytes is ensured by allelic exclusion during V(D)J recombination events in B cell development, which results in the generation

of mature B cells with one productively and one non-productively rearranged *IgH* allele. Different mechanisms are thought to regulate mono-allelic *IgH* expression in mature B cells. In particular, mono-allelic recruitment to PCH was proposed to contribute to the maintenance of silencing of the non-productive *IgH* allele (15,18). On the other hand, it was reported that in proliferating splenic B cells more than half of the *IgH* alleles are located at the nuclear periphery, whereby a ~1 Mb distal V_H region regularly colocalizes with the nuclear lamina (44). In 3D-FISH studies the *IgH* locus was found to be peripheral also in non-B cells, whereby in EL-4 T cells both *IgH* alleles were associated with the nuclear lamina but not with PCH (12,44). The proposed model of mono-allelic recruitment to PCH in mature B cells (15,18) was further challenged by the observation that both productive and nonproductive *IgH* alleles are transcribed in activated splenic B cells (19,30).

Whereas FISH enables the analysis of locus positioning at the single-cell level, it is limited in throughput and provides relatively low-resolution spatial information. In our study, we employed an allele-specific 4C-seq strategy, based on IgM^a/IgM^b allotypes and the C57BL/6 or FVB-specific SNPs they contain. Using this strategy, we analyzed in great detail the genome-wide contacts made by the functionally different *IgH* alleles in splenic B cells, as well as in T cells and non-lymphoid cells. In contrast with published microscopy observations (15,18), we found that (i) the two physically different *IgH* alleles in splenic B cells occupy the same chromatin compartments; (ii) the overall chromatin environment of *IgH* is very similar in B and T cells; (iii) in mature B cells but not in T cells the distal V_H regions of both *IgH* alleles position themselves away from active chromatin; and (iv) that these features of the *IgH* locus do not differ between resting and activated lymphocytes.

We confirmed our 4C-seq-based results by microscopy measurements, which did not uncover pronounced differences in PCH proximity between productive and non-productive *IgH* alleles in cycling splenic B cells. Remarkably, our 4C-seq analyses (**Figure 6B**) also demonstrated that in peripheral resting and activated T cells both *IgH* alleles are localized to an active compartment. The identified localization in T cells of *IgH* (which is not transcribed in T cells) to active chromatin would be in line with our earlier finding that transcription per se is not necessary to maintain a gene in an active chromatin environment (45), but would be in disagreement with its reported localization near the nuclear lamina. Although the lamina has traditionally been associated with gene silencing (12,44), this region of the nucleus does not exclusively harbor silent genes (46-48).

The finding that productive and non-productive *IgH* alleles occupy very similar chromatin environments in resting and cycling splenic B cells adds to the list of similarities between the two: both *IgH* alleles are transcribed in activated splenic B cells, carry similar active chromatin marks, display equivalent RNA polymerase II loading after B cell stimulation (although mRNA of nonproductively rearranged alleles is rapidly degraded by non-sense mediated RNA decay) (19,30), and manifest comparable frequencies of transcription-rate dependent somatic hypermutation in germinal center B cells (49). We conclude that maintenance of allelic exclusion is therefore controlled independent of nuclear location.

It is conceivable that frequent interactions of numerous upstream V_H promoters with E_m may interfere with accurate and efficient transcription of the functionally rearranged V_H gene that is required in mature B cells and especially in plasma cells in which *IgH* is highly expressed. The observed recruitment of distal V_H regions present on both productive and non-productive *IgH* alleles to a less active chromatin environment (**Figure 6A**) may thus help to silence upstream V_H promoters and restrict enhancer activity to the productively rearranged V_H promoter element. Accordingly, non-recombined upstream V_H segments are thought to be inaccessible, since sense/antisense transcription of these V_H segments ceases when *IgH* allelic exclusion is established and is no longer detectable in mature B cells (50).

In summary, using an allele-specific 4C-seq strategy we analyzed genome-wide contacts made by the productively and non-productively rearranged *IgH* alleles in splenic B cells and the essentially unrearranged *IgH* locus in T cells. Different from published microscopy observations, we find that the overall chromatin environment of these three *IgH* types is very similar, except that distal V_H regions are in an active chromatin environment in T cells and in a less active chromatin environment in B cells. While it shows that maintenance of allelic exclusion in mature B cells does not depend on nulear positioning, these data do not necessarily suggest that we need to reconsider the importance of nuclear IgH positioning for allelic exclusion during rearrangement in early B cell development. Future allele-specific analyses in pro-B en pre-B cells should reveal the dynamics of chromatin environment during V(D)J recombination events at different stages of B cell development in the bone marrow.

Availability
Illumina sequencing data have been submitted to the GEO database and are accessible under accession number: GSE47129

Supplementary data
Supplementary Data are available at NAR online: Supplementary Tables 1-13, Supplementary Figures 1-13, Supplementary Methods, and Supplementary References [51-58].

Funding
This work was financially supported by grant no. 935170621 from the Dutch Scientific Organization (NWO) and a European Research Council Starting Grant (209700, '4C') to WdL.

Acknowledgements
We thank E. de Wit for analyzing and plotting the B cell HiC data, T. van Ravesteyn for excellent assistance confirming SNPs and S. Yuvaraj for lymphocyte cultures.

Materials and Methods

Separation & stimulation of IgM^a and IgM^b expressing B cells
Magnetic activated cells sorting (MACS): mature resting B cells were purified using streptavidin-coupled magnetic beads (Miltenyi Biotec), by negative selection using the following biotinylated antibodies: CD5, CD43, CD138, CD11b, Gr-1 and TER-119 (**Suppl. Table 1.1**).
Fluorescence activated cells sorting (FACS): Separation of B cell populations based on IgM allotype expression was done by FACS sorting of MACS-purified fractions of resting B cells using antibodies to B220 and CD19, in conjunction with allotype-specific antibodies for IgM^a [FVB] and IgM^b [C57BL/6], for details see **Suppl. Table 1.3.**
In-vitro activation: Purified (resting) B cells were *in-vitro* activated, as described (18) for 4 days using αCD40 coated plates (20ug/ml; BD Biosciences) and IL-4 (IL-4 50 ng/ml; Peprotech). For further details see supplemental data.

4C template preparation & mapping
FACS sorted cells were used for 4C template preparation. Cells were fixed and lysed as described (26) using HindIII (Roche) as a first cutter and DpnII (New England Biolabs) as a second cutter. An allele specific strategy for single-end 4C-sequencing was used as described (25,27), where restriction fragment length polymorphisms between the C57Bl/6 and the FVB genome are exploited. Primers for the single-end 4C-seq

experiments were designed around a SNP that creates an extra DpnII restriction enzyme site on the FVB allele. Consequently, only the C57Bl/6 allele will be analyzed using this strategy (**Suppl. Fig. 2**). For the allele specific paired-end 4C sequencing (PE-4Cseq), primers were designed such that one of the selected primers read a SNP (P2) (28) whereas the other primer (P1) read into the captured sequence ligated to the 'bait'-fragment in the 4C procedure. Consequently, simultaneous analysis of two alleles is possible.
The single-end data was mapped, allowing no mismatches, to a database of 4C-seq fragments-ends generated from the mm9/NCBI m37 version of the mouse genome (29). The paired-end sequencing data was first split based on the SNPs (C57Bl/6 vs FVB) detected in the second read (PE2) of the read-pair and subsequently the first read of the pair (PE1) was mapped as single-end data.
Significant genomic contacts, visualized by domainograms, were identified based on described algorithms (29). For further details see supplemental data.

RNA-FISH and DNA-FISH

MACS sorted cells were used for RNA and DNA FISH experiments. RNA & DNA FISH experiments were performed as described (27), with minor adjustments. Briefly, for DNA FISH denaturing of the DNA in the cells on slides was done for 10' @ 80°C in 50% formamide / 2X SSC after which a denatured probe was applied to the slide for overnight hybridization at 42°C followed by post hybridization washes and microscopic analysis. For further details see supplemental data.
For detailed description of mice, B and T cell isolation, lymphoid cell culture, separation of IgM^a and IgM^b expressing B cells, 4C-seq procedures, RNA expression analysis, RNA-FISH and DNA-FISH: see **Suppl. Materials and Methods.**

Supplementary Materials and Methods

Mice

Spleens were harvested from F1 hybrid mice (C57Bl/6 x FVB) at 8 to 14 weeks. Fetal brain was isolated from 14.5 dpc mouse embryos. Single cell suspensions, for fetal brains, were prepared by filtration through a 40µm cell strainer (BD biosciences). Mice were bred and kept at specified pathogen free conditions in the Erasmus MC experimental animal facility. All experimental protocols have been reviewed and approved by the Erasmus MC Committee of animal experiments.

Total B and T cell isolation, cell culture and separation of IgM^a and IgM^b expressing B cells

Splenic cell suspensions were prepared and sorted using magnetic beads as previously described (51). Erythrocytes were lysed using Gey's solution. B cells were purified using streptavidin-coupled magnetic beads (Miltenyi Biotec), by negative selection using the following biotinylated antibodies: CD5, CD43, CD138, CD11b, Gr-1 and TER-119 (**Suppl. Table 1.1**). Purified (resting) B cells were *in-vitro* activated, as described in Ref. (18) for 4 days using αCD40 coated plates (20ug/ml; BD Biosciences) and IL-4 (IL-4 50 ng/ml; Peprotech). Splenic T cells were isolated by negative selection using antibodies against B220, NK1.1, CD11b, GR-1 and Ter119 (**Suppl. Table 1.2**). Purified T cells were *in-vitro* activated for 4 days using aCD3 (10µg/ml), aCD28 (10µg/ml) and IL-2 (20µg/ml; R&D systems).
Separation of B cell populations based on IgM allotype expression was done by FACS sorting of MACS-purified fractions of resting B cells (as described above)using antibodies to B220 and CD19, in conjunction with allotype-specific antibodies for IgM^a [FVB] and IgM^b [C57BL/6], as detailed in **Suppl. Table 1.3.**

Mapping 4C single-end and paired-end sequencing data

The single-end data was mapped, allowing no mismatches, to a database of 4C-seq fragments-ends

generated from the mm9/NCBI m37 version of the mouse genome (29). The paired-end sequencing data, however, was first split based on the SNPs (C57Bl/6 vs FVB) detected in the second read (PE2) of the read-pair and subsequently the first read of the pair (PE1) was mapped as single-end data. If the first read of a pair was at the site of the second restriction enzyme, fragment-ends were used that contain both restriction sites, so called non-blind fragment-ends (26). No mismatches were allowed in the reads.

4C-seq analysis

4C domainograms were generated as described previously (29,52).

Correlation plots were generated according to the following scheme:

Reads corresponding to *trans* fragment-ends are presumed to reflect single ligation events and are therefore first binarized (29), after which a running window statistic (window size of 500 fragment-ends) was calculated (4C profile). We subsequently, calculated the pairwise Spearman's rank correlation coefficient of the 4C profiles. For the *cis* data, we used a window size of 21 fragment-ends and a z-score was calculated with a running background window of 3001 fragment-ends (25). All the z-scores within 2 Mb of the viewpoint were removed before the pairwise Spearman's rank correlation (r_s) coefficients were calculated. The correlation matrices were clustered based on Euclidean distance of 1-r_s using a hierarchical clustering algorithm with complete linkage. All computations and plotting of the resulting heat maps were carried out with R (53).

Interacting domains were determined according to the following scheme:

To determine which domains are interacting with the viewpoint of interest, we applied a False Discovery Rate (FDR) control (29). In brief, the FDR was calculated after the data was binarized and a binomial one-tailed test was applied on a running window of 100 fragment-ends in *cis* and a background window of 3001. In *trans*, however, the binarized data was transformed with a running window of 500 fragment-ends and a background window based on the number of unique fragment-ends in each chromosome. We permuted (N >=100) the dataset of each chromosome and calculated the FDR based on the *p*-value statistic of the binomial test. Consecutive fragment-ends with a *q*-value < 0.01 and having a maximum gap size of 25 (cis) and 250 (trans) fragment-ends were assigned as a single interacting domain.

RNA extraction, labeling and microarray hybridization

Expression analysis was done in B and T cells on Affymetrix microarrays (mouse gene st1.0 arrays). RNA was isolated using Trizol® according to manufacturer's instructions. RNA was labeled according to Affymetrix protocol instructions and hybridized to three independent microarrays per cell type per timepoint according to Affymetrix protocol. Expression data for fetal brain was used from (24).

RNA expression analysis

Normalized gene expression values were calculated using robust multi-array average (RMA) (54). The limma package (55) from BioConductor (56) was used for the analysis of differential gene expression in T- and B-cells. Holm's method was used to adjust p-values for multiple testing. Probe sets with an adjusted p-value < 0.05 were selected and their relative expression was plotted in a heat map. The probe sets, as well as, the experiments were clustered after the pairwise Pearson's correlation coefficient (*r*) matrix and the dissimilarity matrix (1-*r*) were calculated. The dissimilarity matrices were used to calculate the Euclidean distances and, subsequently, used for hierarchical clustering with complete linkage. To all 9 clusters, names were assigned based on the expression pattern across the 4 cell types. A Gene Ontology (GO) enrichment analysis was carried out on Entrez genes corresponding to the differentially expressed probe sets. The complete differential expressed gene set and the gene set for each cluster were analyzed with the GO-stats package (57) using a hyper geometrical test for over-representation of genes and a *p*-value threshold-value of 0.001.

Genes that have probe sets on the both Affymetrix gene expression arrays were used to compare the

properties of 4C-seq domains. For each gene the median value of the corresponding probe set was assigned, after the gene expression dataset was normalized by subtracting its median value. The gene expression distribution showed a clear bimodal distribution for each RNA-expression experiment. Therefore, the median value of the gene expression dataset was used to discriminate between active or transcribed genes and non-active genes.

RNA-FISH and DNA-FISH

Cells were spotted on poly-L-lysine coated coverslips. RNA FISH experiments were performed as described (27). Nuclei with double CD45 signals were, exclusively used to count IgH RNA signals. DNA FISH experiments were performed as described (27) with minor changes. Denaturing of the DNA in the cells on slides was done for 10' @ 80°C in 50% formamide/2X SSC followed by two washes with 2xSSC at 4°C, after which a denatured probe was applied to the slide for overnight hybridization at 42°C followed by post hybridization washes and microscopic analysis. The following BAC clones were used as probes: For the constant region of the IgH locus, RP23-109B20; for the variable region of the IgH locus, RP24-376H17 (kindly provided by M. Busslinger); for CD45, RP24-371I24; for γ-satelites a pBleuscript plasmid containing 8 copies of repetitive y-satelite sequence was used as described (18,58) (kindly provided by J.A. Skok). For the verification of the 4C data the following BAC clones were used. For two interactions in B cells RP23-215D11 and RP23-432F6 were used. For interactions in fetal brain, RP23-69J9 was used. Generation of the probes was done as described (25). Specificity of the probes was verified on metaphase spreads of mouse ES cells (data not shown). 3D distance measurements were done for at least 50 nuclei per data point, using Image J software. The active allele is defined as the least contracted allele. The other allele is defined as the inactive allele. Distances towards peri-centromeric heterochromatin are measured as the minimal distance of the IgH locus towards the edge of the γ-satellite signal.

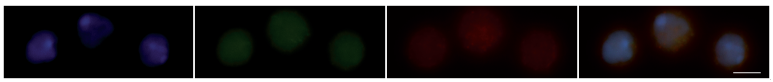
References

1. Jung, D., Giallourakis, C., Mostoslavsky, R. and Alt, F.W. (2006) Mechanism and control of V(D)J recombination at the immunoglobulin heavy chain locus. *Annual review of immunology*, **24**, 541-570.
2. Perlot, T. and Alt, F.W. (2008) Cis-regulatory elements and epigenetic changes control genomic rearrangements of the IgH locus. *Advances in immunology*, **99**, 1-32.
3. Bossen, C., Mansson, R. and Murre, C. (2012) Chromatin topology and the regulation of antigen receptor assembly. *Annual review of immunology*, **30**, 337-356.
4. Vettermann, C. and Schlissel, M.S. (2010) Allelic exclusion of immunoglobulin genes: models and mechanisms. *Immunological reviews*, **237**, 22-42.
5. Guo, C., Yoon, H.S., Franklin, A., Jain, S., Ebert, A., Cheng, H.L., Hansen, E., Despo, O., Bossen, C., Vettermann, C. *et al.* (2011) CTCF-binding elements mediate control of V(D)J recombination. *Nature*, **477**, 424-430.
6. Serwe, M. and Sablitzky, F. (1993) V(D)J recombination in B cells is impaired but not blocked by targeted deletion of the immunoglobulin heavy chain intron enhancer. *The EMBO journal*, **12**, 2321-2327.
7. Sakai, E., Bottaro, A., Davidson, L., Sleckman, B.P. and Alt, F.W. (1999) Recombination and transcription of the endogenous Ig heavy chain locus is effected by the Ig heavy chain intronic enhancer core region in the absence of the matrix attachment regions. *Proceedings of the National Academy of Sciences of the United States of America*, **96**, 1526-1531.
8. Pinaud, E., Marquet, M., Fiancette, R., Peron, S., Vincent-Fabert, C., Denizot, Y. and Cogne, M. (2011) The IgH locus 3' regulatory region: pulling the strings from behind. *Advances in immunology*, **110**, 27-70.
9. Kenter, A.L., Feldman, S., Wuerffel, R., Achour, I., Wang, L. and Kumar, S. (2012) Three-dimensional architecture of the IgH locus facilitates class switch recombination. *Annals of the New York Academy of Sciences*, **1267**, 86-94.
10. Chaumeil, J. and Skok, J.A. (2012) The role of CTCF in regulating V(D)J recombination. *Current opinion in immunology*, **24**, 153-159.
11. Del Blanco, B., Garcia, V., Garcia-Mariscal, A. and Hernandez-Munain, C. (2011) Control of V(D)J Recombination through Transcriptional Elongation and Changes in Locus Chromatin Structure and Nuclear Organization. *Genetics research international*, **2011**, 970968.
12. Kosak, S.T., Skok, J.A., Medina, K.L., Riblet, R., Le Beau, M.M., Fisher, A.G. and Singh, H. (2002) Subnuclear

compartmentalization of immunoglobulin loci during lymphocyte development. *Science*, **296**, 158-162.
13. Fuxa, M., Skok, J., Souabni, A., Salvagiotto, G., Roldan, E. and Busslinger, M. (2004) Pax5 induces V-to-DJ rearrangements and locus contraction of the immunoglobulin heavy-chain gene. *Genes & development*, **18**, 411-422.
14. Sayegh, C.E., Jhunjhunwala, S., Riblet, R. and Murre, C. (2005) Visualization of looping involving the immunoglobulin heavy-chain locus in developing B cells. *Genes & development*, **19**, 322-327.
15. Roldan, E., Fuxa, M., Chong, W., Martinez, D., Novatchkova, M., Busslinger, M. and Skok, J.A. (2005) Locus 'decontraction' and centromeric recruitment contribute to allelic exclusion of the immunoglobulin heavy-chain gene. *Nature immunology*, **6**, 31-41.
16. Jhunjhunwala, S., van Zelm, M.C., Peak, M.M., Cutchin, S., Riblet, R., van Dongen, J.J., Grosveld, F.G., Knoch, T.A. and Murre, C. (2008) The 3D structure of the immunoglobulin heavy-chain locus: implications for long-range genomic interactions. *Cell*, **133**, 265-279.
17. Brady, B.L. and Bassing, C.H. (2011) Differential regulation of proximal and distal Vbeta segments upstream of a functional VDJbeta1 rearrangement upon beta-selection. *J Immunol*, **187**, 3277-3285.
18. Skok, J.A., Brown, K.E., Azuara, V., Caparros, M.L., Baxter, J., Takacs, K., Dillon, N., Gray, D., Perry, R.P., Merkenschlager, M. *et al.* (2001) Nonequivalent nuclear location of immunoglobulin alleles in B lymphocytes. *Nature immunology*, **2**, 848-854.
19. Daly, J., Licence, S., Nanou, A., Morgan, G. and Martensson, I.L. (2007) Transcription of productive and nonproductive VDJ-recombined alleles after IgH allelic exclusion. *The EMBO journal*, **26**, 4273-4282.
20. Dekker, J., Rippe, K., Dekker, M. and Kleckner, N. (2002) Capturing chromosome conformation. *Science*, **295**, 1306-1311.
21. Degner, S.C., Verma-Gaur, J., Wong, T.P., Bossen, C., Iverson, G.M., Torkamani, A., Vettermann, C., Lin, Y.C., Ju, Z., Schulz, D. *et al.* (2011) CCCTC-binding factor (CTCF) and cohesin influence the genomic architecture of the Igh locus and antisense transcription in pro-B cells. *Proceedings of the National Academy of Sciences of the United States of America*, **108**, 9566-9571.
22. Ribeiro de Almeida, C., Stadhouders, R., Thongjuea, S., Soler, E. and Hendriks, R.W. (2012) DNA-binding factor CTCF and long-range gene interactions in V(D)J recombination and oncogene activation. *Blood*, **119**, 6209-6218.
23. Guo, C., Gerasimova, T., Hao, H., Ivanova, I., Chakraborty, T., Selimyan, R., Oltz, E.M. and Sen, R. (2011) Two forms of loops generate the chromatin conformation of the immunoglobulin heavy-chain gene locus. *Cell*, **147**, 332-343.
24. Simonis, M., Klous, P., Splinter, E., Moshkin, Y., Willemsen, R., de Wit, E., van Steensel, B. and de Laat, W. (2006) Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4C). *Nature genetics*, **38**, 1348-1354.
25. Splinter, E., de Wit, E., Nora, E.P., Klous, P., van de Werken, H.J., Zhu, Y., Kaaij, L.J., van Ijcken, W., Gribnau, J., Heard, E. *et al.* (2011) The inactive X chromosome adopts a unique three-dimensional conformation that is dependent on Xist RNA. *Genes & development*, **25**, 1371-1383.
26. van de Werken, H.J., Landan, G., Holwerda, S.J., Hoichman, M., Klous, P., Chachik, R., Splinter, E., Valdes-Quezada, C., Oz, Y., Bouwman, B.A. *et al.* (2012) Robust 4C-seq data analysis to screen for regulatory DNA interactions. *Nature methods*, **9**, 969-972.
27. Chaumeil, J., Le Baccon, P., Wutz, A. and Heard, E. (2006) A novel role for Xist RNA in the formation of a repressive nuclear compartment into which genes are recruited when silenced. *Genes & development*, **20**, 2223-2237.
28. Frazer, K.A., Eskin, E., Kang, H.M., Bogue, M.A., Hinds, D.A., Beilharz, E.J., Gupta, R.V., Montgomery, J., Morenzoni, M.M., Nilsen, G.B. *et al.* (2007) A sequence-based variation map of 8.27 million SNPs in inbred mouse strains. *Nature*, **448**, 1050-1053.
29. van de Werken, H.J., de Vree, P.J., Splinter, E., Holwerda, S.J., Klous, P., de Wit, E. and de Laat, W. (2012) 4C technology: protocols and data analysis. *Methods in enzymology*, **513**, 89-112.
30. Tinguely, A., Chemin, G., Peron, S., Sirac, C., Reynaud, S., Cogne, M. and Delpy, L. (2012) Cross talk between immunoglobulin heavy-chain transcription and RNA surveillance during B cell development. *Molecular and cellular biology*, **32**, 107-117.
31. Eberle, A.B., Herrmann, K., Jack, H.M. and Muhlemann, O. (2009) Equal transcription rates of productively and nonproductively rearranged immunoglobulin mu heavy chain alleles in a pro-B cell line. *RNA*, **15**, 1021-1028.
32. Stall, A.M. and Loken, M.R. (1984) Allotypic specificities of murine IgD and IgM recognized by monoclonal antibodies. *J Immunol*, **132**, 787-795.
33. Dixon, J.R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., Hu, M., Liu, J.S. and Ren, B. (2012) Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, **485**, 376-380.
34. Nora, E.P., Lajoie, B.R., Schulz, E.G., Giorgetti, L., Okamoto, I., Servant, N., Piolot, T., van Berkum, N.L., Meisig, J., Sedat, J. *et al.* (2012) Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature*, **485**, 381-385.
35. Lin, Y.C., Benner, C., Mansson, R., Heinz, S., Miyazaki, K., Miyazaki, M., Chandra, V., Bossen, C., Glass, C.K. and Murre, C. (2012) Global changes in the nuclear positioning of genes and intra- and interdomain genomic interactions that orchestrate B cell fate. *Nature immunology*, **13**, 1196-1204.
36. Zhang, Y., McCord, R.P., Ho, Y.J., Lajoie, B.R., Hildebrand, D.G., Simon, A.C., Becker, M.S., Alt, F.W. and Dekker, J. (2012) Spatial organization of the mouse genome and its role in recurrent chromosomal translocations. *Cell*, **148**, 908-921.
37. Ebert, A., McManus, S., Tagoh, H., Medvedovic, J., Salvagiotto, G., Novatchkova, M., Tamir, I., Sommer, A.,

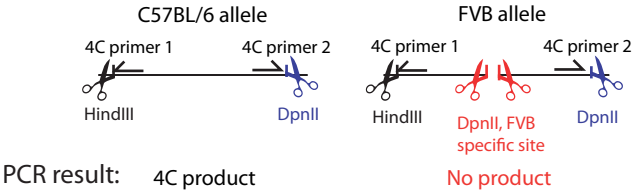
Jaritz, M. and Busslinger, M. (2011) The distal V(H) gene cluster of the Igh locus contains distinct regulatory elements with Pax5 transcription factor-dependent activity in pro-B cells. *Immunity*, **34**, 175-187.
38. Ehlich, A., Martin, V., Muller, W. and Rajewsky, K. (1994) Analysis of the B-cell progenitor compartment at the level of single cells. *Current biology : CB*, **4**, 573-583.
39. ten Boekel, E., Melchers, F. and Rolink, A. (1995) The status of Ig loci rearrangements in single cells from different stages of B cell development. *International immunology*, **7**, 1013-1019.
40. Rajewsky, K. (1996) Clonal selection and learning in the antibody system. *Nature*, **381**, 751-758.
41. Melchers, F., ten Boekel, E., Yamagami, T., Andersson, J. and Rolink, A. (1999) The roles of preB and B cell receptors in the stepwise allelic exclusion of mouse IgH and L chain gene loci. *Seminars in immunology*, **11**, 307-317.
42. Malynn, B.A., Yancopoulos, G.D., Barth, J.E., Bona, C.A. and Alt, F.W. (1990) Biased expression of JH-proximal VH genes occurs in the newly generated repertoire of neonatal and adult mice. *The Journal of experimental medicine*, **171**, 843-859.
43. Kawano, Y., Yoshikawa, S., Minegishi, Y. and Karasuyama, H. (2006) Pre-B cell receptor assesses the quality of IgH chains and tunes the pre-B cell repertoire by delivering differential signals. *J Immunol*, **177**, 2242-2249.
44. Yang, Q., Riblet, R. and Schildkraut, C.L. (2005) Sites that direct nuclear compartmentalization are near the 5' end of the mouse immunoglobulin heavy-chain locus. *Molecular and cellular biology*, **25**, 6021-6030.
45. Palstra, R.J., Simonis, M., Klous, P., Brasset, E., Eijkelkamp, B. and de Laat, W. (2008) Maintenance of long-range DNA interactions after inhibition of ongoing RNA polymerase II transcription. *PloS one*, **3**, e1661.
46. Hewitt, S.L., High, F.A., Reiner, S.L., Fisher, A.G. and Merkenschlager, M. (2004) Nuclear repositioning marks the selective exclusion of lineage-inappropriate transcription factor loci during T helper cell differentiation. *European journal of immunology*, **34**, 3604-3613.
47. Kumaran, R.I. and Spector, D.L. (2008) A genetic locus targeted to the nuclear periphery in living cells maintains its transcriptional competence. *The Journal of cell biology*, **180**, 51-65.
48. Finlan, L.E., Sproul, D., Thomson, I., Boyle, S., Kerr, E., Perry, P., Ylstra, B., Chubb, J.R. and Bickmore, W.A. (2008) Recruitment to the nuclear periphery can alter expression of genes in human cells. *PLoS genetics*, **4**, e1000039.
49. Delpy, L., Sirac, C., Le Morvan, C. and Cogne, M. (2004) Transcription-dependent somatic hypermutation occurs at similar levels on functional and nonfunctional rearranged IgH alleles. *J Immunol*, **173**, 1842-1848.
50. Bolland, D.J., Wood, A.L., Johnston, C.M., Bunting, S.F., Morgan, G., Chakalova, L., Fraser, P.J. and Corcoran, A.E. (2004) Antisense intergenic transcription in V(D)J recombination. *Nature immunology*, **5**, 630-637.
51. Kil, L.P., de Bruijn, M.J., van Nimwegen, M., Corneth, O.B., van Hamburg, J.P., Dingjan, G.M., Thaiss, F., Rimmelzwaan, G.F., Elewaut, D., Delsing, D. *et al.* (2012) Btk levels set the threshold for B-cell activation and negative selection of autoreactive B cells in mice. *Blood*, **119**, 3744-3756.
52. de Wit, E., Braunschweig, U., Greil, F., Bussemaker, H.J. and van Steensel, B. (2008) Global chromatin domain organization of the Drosophila genome. *PLoS genetics*, **4**, e1000045.
53. R Development Core Team. (2010). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing.
54. Irizarry, R.A., Bolstad, B.M., Collin, F., Cope, L.M., Hobbs, B. and Speed, T.P. (2003) Summaries of Affymetrix GeneChip probe level data. *Nucleic acids research*, **31**, e15.
55. Smyth, G.K., Michaud, J. and Scott, H.S. (2005) Use of within-array replicate spots for assessing differential expression in microarray experiments. *Bioinformatics*, **21**, 2067-2075.
56. Gentleman, R.C., Carey, V.J., Bates, D.M., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J. *et al.* (2004) Bioconductor: open software development for computational biology and bioinformatics. *Genome biology*, **5**, R80.
57. Falcon, S. and Gentleman, R. (2007) Using GOstats to test gene lists for GO term association. *Bioinformatics*, **23**, 257-258.
58. Brown, K.E., Guest, S.S., Smale, S.T., Hahm, K., Merkenschlager, M. and Fisher, A.G. (1997) Association of transcriptionally silent genes with Ikaros complexes at centromeric heterochromatin. *Cell*, **91**, 845-854.

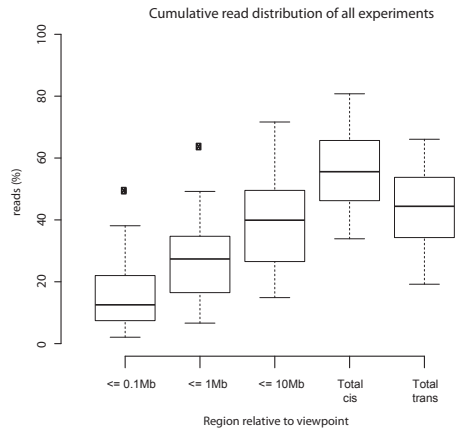
Supplemental Figures and tables



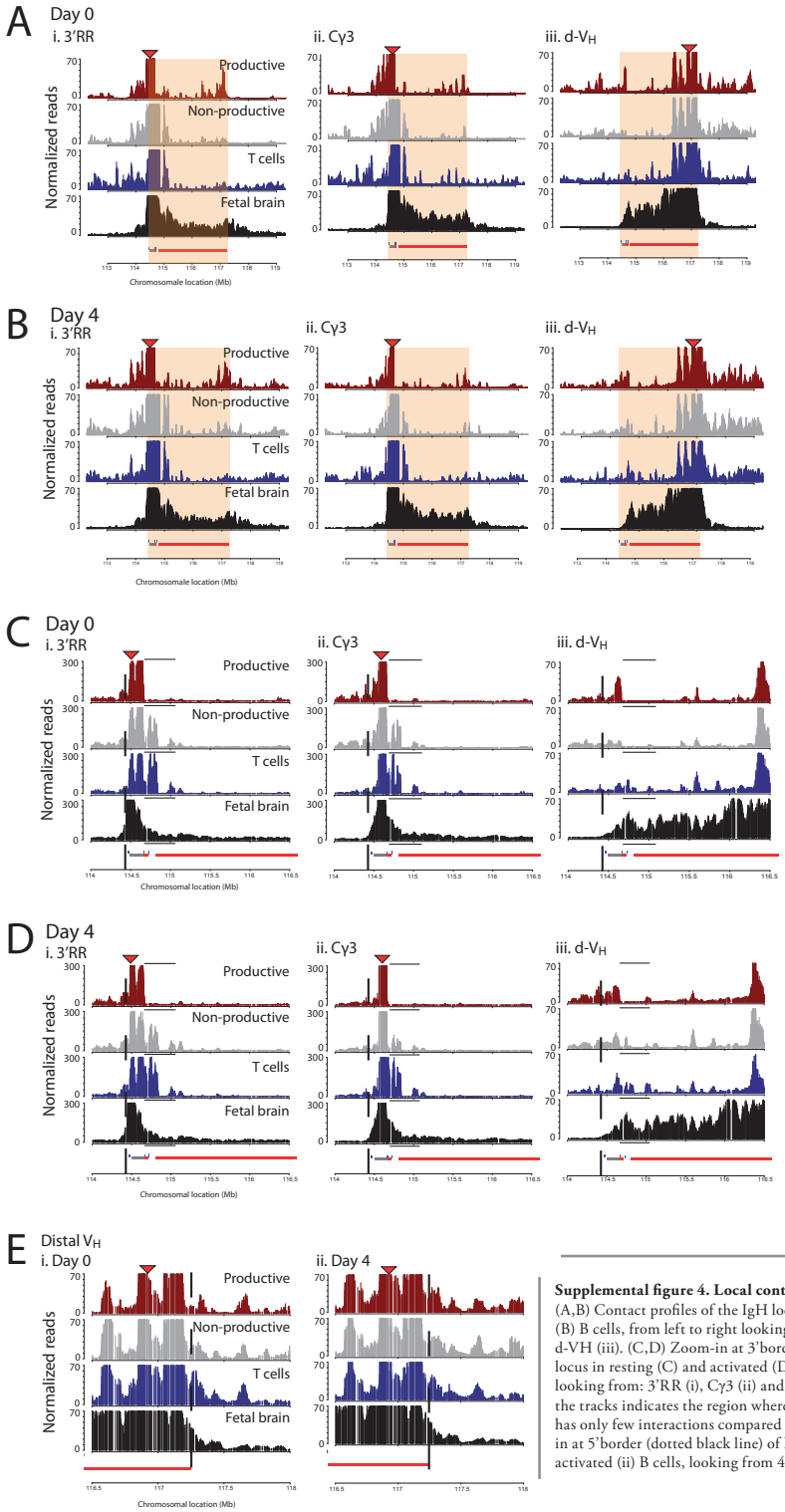
Supplemental figure 1. Lack of RNA signal after RNase treatment. Representative RNA FISH picture for RNase treated resting B cells.

Supplemental figure 2. Single end allele specific 4C strategy. Schematic representation of a 4C HindIII/DpnII fragment of a C57Bl/6 and a FVB allele. A SNP between the two alleles creates an additional DpnII restriction site (red scissor pair) in the FVB allele which prohibits the formation of a 4C PCR product using the inverse PCR primer 1 and 2.

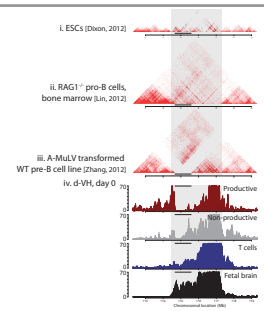




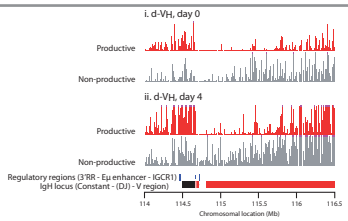
Supplemental figure 3. Quality control of the 4C experiments. Cumulative read distribution of all experiments plotted as the percentage of captured reads in different binned regions. Distance of the regions are relative to the viewpoints.



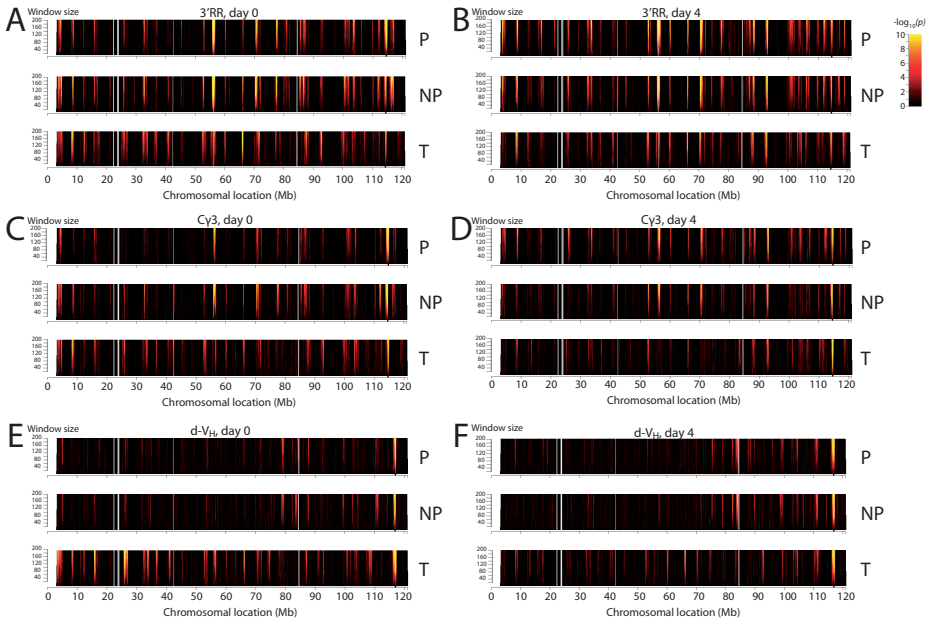
Supplemental figure 4. Local contact profiles at the FVB allele. (A,B) Contact profiles of the IgH locus in resting (A) and activated (B) B cells, from left to right looking from: 3'RR (i), Cγ3 (ii) and d-VH (iii). (C,D) Zoom-in at 3'border (dotted black line) of IgH locus in resting (C) and activated (D) B cells, from left to right looking from: 3'RR (i), Cγ3 (ii) and d-VH (iii). The black line above the tracks indicates the region where the productive allele in B cells has only few interactions compared to the other alleles. (E) Zoom-in at 5'border (dotted black line) of IgH locus in resting (i) and activated (ii) B cells, looking from 4C viewpoint d-VH.



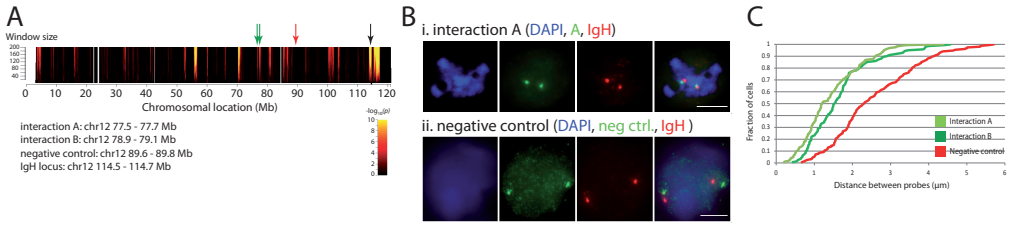
Supplemental figure 5. Topological and recombination borders defined by HiC and 4C. Contacts around the IgH locus are shown from top to bottom as Hi-C data in ESCs (i), RAG1-/- pro B cells (ii) and a G1 arrested pre-B cell line (iii). 4C data (iv) for the IgH locus looking from the d-VH 4C viewpoint in resting B cells. Top to bottom: the productive (red) and non-productive B cell allele (blue), T cells (blue) and fetal brain (black). The black bar depicts the region that is recombined in the preB cell line (iii) and on the productive allele (iv). The shaded region depicts the topological domain spanning the IgH locus.



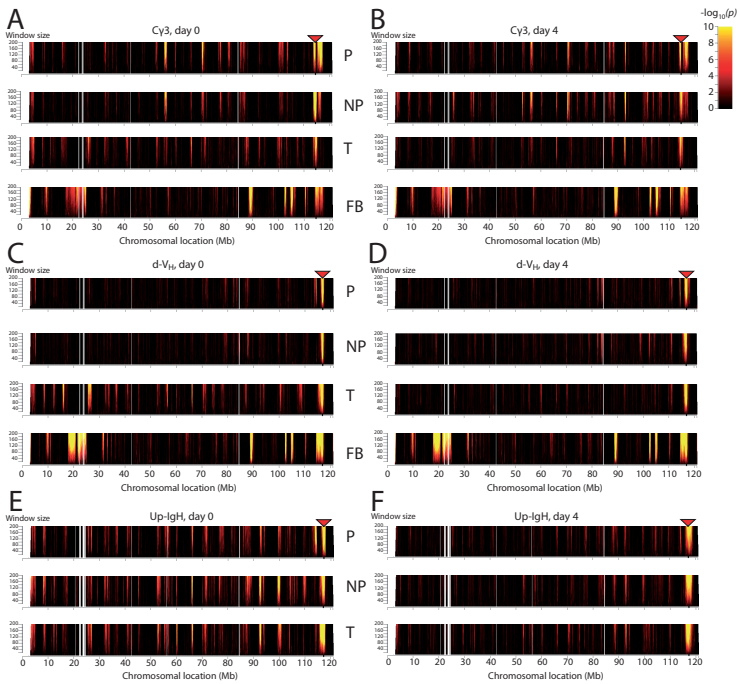
Supplemental figure 6. Difference in 4C contacts at the proximal V region. Raw 4C contact data looking from the d-VH 4C viewpoint in resting (i) and activated (ii) B cells. Contacts for the productive and the non-productive B cell alleles are shown in red and grey, respectively. Purple colored 4C signals depict saturation of the signal given the maximum value of the Y-axis.



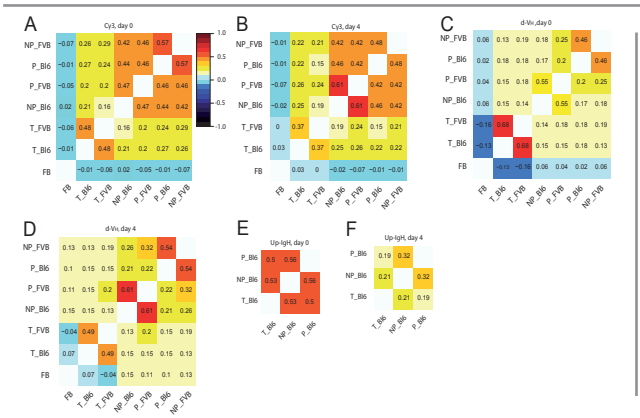
Supplemental figure 7. Cis contacts of the IgH locus at the FVB allele. (A-F) Domainograms showing chromosome-wide interactions of the FVB allele, in resting (A) or activated B cells (B) looking from the 3'RR 4C viewpoint; in resting (C) or activated B cells (D) looking from the Cγ3 viewpoint; in resting (E) or activated B cells (F) looking from the d-VH viewpoint. Different alleles are depicted below each other, from top to bottom: P = productive allele in B cells, NP = non-productive allele in B cells, T = T cells, FB = fetal brain. Red arrows depict the 4C viewpoints. Significance of the interactions is indicated by the range in color used in the domainogram as depicted in the legend: black is low significance (P = 1) and yellow represents high significance (P = 10-10) of interaction.



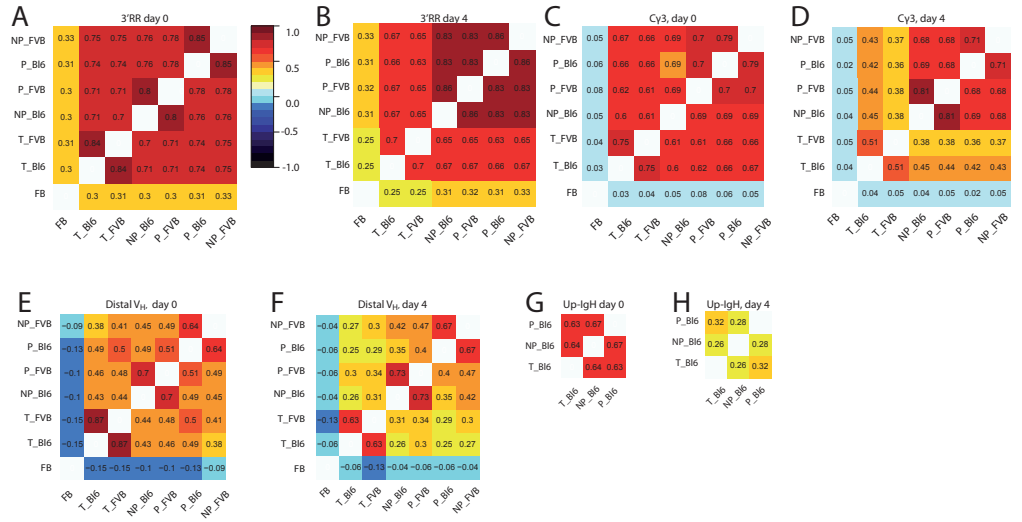
Supplemental figure 8. Verification of 4C by DNA FISH. (A) Schematic representation of the probe positions. Green arrows depict the positions of the probes used to verify the 4C interaction, the red arrow depicts the position of the probe used as a negative control, the black arrow depicts probe that is used to visualize the IgH locus. (B) Representative pictures of the DNA FISH, showing an interaction A (i) and the negative control (ii). Scale bar represents 4µm. (C) Cumulative fractions of cells plotted as a function of the distance between the test probe and the probe for the IgH locus. A minimum of 50 cells was counted per probe set.



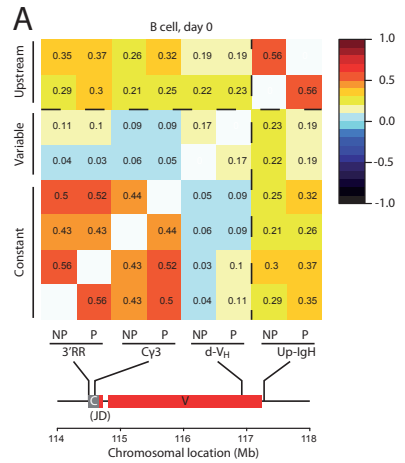
Supplemental figure 9. Cis contacts of the IgH locus at the C57BL/6 allele. (A-F) Domainograms showing chromosome-wide interactions of the C57Bl/6 allele, in resting (A) or activated B cells (B) looking from the Cγ3 4C viewpoint; in resting (C) or activated B cells (D) looking from the d-VH viewpoint; in resting (E) or activated B cells (F) looking from the Up-IgH viewpoint. Different alleles are depicted below each other, from top to bottom: P = productive allele in B cells, NP = non-productive allele in B cells, T = T cells, FB = fetal brain. Red arrows depict the 4C viewpoints. Significance of the interactions is indicated by the range in color used in the domainogram as depicted in the legend: black is low significance (P = 1) and yellow represents high significance ($P = 10^{-10}$) of interaction.

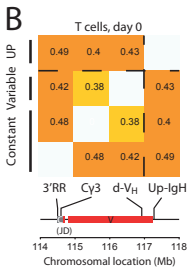


Supplemental figure 10. Correlation of cis contacts. (A,B) Correlation plots of the interactions looking from the Cγ3 viewpoint in resting (A) or activated (B) B cells (window size 21). The numbers represent the Spearman rank correlation coefficient, colors range from linear anti-correlation (black) to linear correlation (dark red). Cell types and the alleles are depicted along the X and Y-axis; B cells productive (P), B cell non-productive (NP), T cell (T), fetal brain (FB), C57BL/6 allele (Bl6) and FVB allele (FVB). (C,D) Correlation plots of the interactions looking from the d-VH viewpoint in resting (C) and activated (D) B cells (window size 21). (E,F) Correlation plots of the interactions looking from the Up-IgH viewpoint in resting (E) and activated (F) B cells.

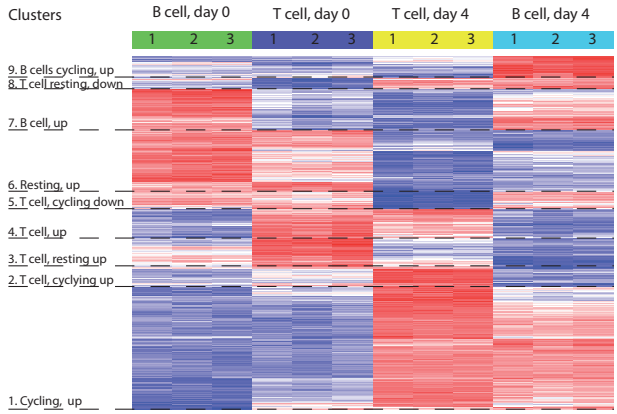


Supplemental figure 11. Correlation of trans contacts. (A,B) Correlation plot of the interactions looking from the 3'RR viewpoint in resting (A) and activated (B) B cells (window size 21). The numbers represent the Spearman rank correlation coefficient, colors range from linear anti-correlation (black) to linear correlation (dark red). Cell types and the alleles are depicted along the X and Y-axis; B cells productive (P), B cell non-productive (NP), T cell (T), fetal brain (FB), C57BL/6 allele (Bl6) and FVB allele (FVB). (C,D) Correlation plot of the interactions looking from the Cγ3 viewpoint in resting (C) and activated (D) B cells. (E,F) Correlation plot of the interactions looking from the d-VH viewpoint in resting (E) and activated (F) B cells. (G,H) Correlation plot of the interactions looking from the Up-IgH viewpoint in resting (G) and activated (H) B cells.





Supplemental figure 12. (A). Correlation plot of the interactions of C57BL/6 viewpoints in cis, in resting B cells. The X-axis shows the productive (P) and the non-productive (NP) C57Bl/6 alleles looking from the four 4C viewpoints. The IgH locus with the position of the 4C viewpoints is drawn to scale below the X-axis. The numbers represent the Spearman rank correlation coefficients, colors range from linear anti-correlation (black) to linear correlation (dark red). The dotted line represent the topological border that is present between the d-VH and the Up-IgH 4C viewpoints. (B) Correlation plot of the interactions of C57BL/6 viewpoints in cis in resting T cells.



Supplemental figure 13. Clustering of gene expression data in resting and activated B and T cells. Hierarchical clustering of microarray data in resting B and T cells. On the horizontal axis the tissues and the microarray experiments are depicted, highlighted in colors. The numbers refer to the biological triplo experiments per cell type. The vertical axis shows the names of the 9 different clusters. The name of a cluster depicts the tissue and the up/down regulation of the genes that define the specific cluster.

Supplemental tables

Table S1.1 MACS on B cells

Name	Company	Product number
CD5	BD	553019
CD43	ebioscience	13-0431-85
CD138	BD	553713
CD11b	BD	553309
Ly-6G (GR-1)	ebioscience	13-5931-85
TER-119	ebioscience	13-5921-85

Table S1.2 MACS on T cells

Name	Company	Product number
CD45R/B220	BD	553085
NK-1.1	BD	553163
CD11b	BD	553309
GR-1	ebioscience	13-5931-85
TER-119	ebioscience	13-5921-85

Table S1.3 Allotypic FACS on B cells

Name	Company	Product number
CD19 PerCP-Cy5.5	ebioscience	45-0193-82
CD45R (B220) FITC	ebioscience	11-0452-86
IgMa	BD	553517
IgMb	BD	553519
Streptavidin APC	ebioscience	17-4317-82

Supplementary table 1. Antibodies used in the MACS on T (S1.1) and B cells (S1.2) and the subsequent FACS where B cells were separated based on the haplotypic origin of their B cell receptor (S1.3).

EntrezID	gene symbols	gene description	Fold Change (D4/ D0)	P value
11628	Aicda	activation-induced cytidine deaminase	79.34	0
16192	Il5ra	interleukin 5 receptor, alpha	15.45	0
12505	Cd44	CD44 antigen	5.06	0
14102	Fas	Fas (TNF receptor superfamily member 6)	7.31	0
12444	Ccnd2	cyclin D2	7.06	0
17869	Myc	myelocytomatosis oncogene	1.99	0

Supplementary table 2. Gene expression changes between (D4 / D0) confirm in-vitro activation of resting, splenic B cells. Absolute fold change is calculated as the average expression, of three independent microarray experiments per cell type, of the activated B cells over the resting B cells.

Supplemental table 3a

Upregulated genes in cycling B and T cells (day 4)

GOBPID	Pvalue	OddsRatio	ExpCount	Count	Size	Term
GO:0000278	0	8.54506	23	131	493	mitotic cell cycle
GO:0048285	0	11.77863	15	107	313	organelle fission
GO:0000279	0	9.07895	17	106	377	M phase
GO:0051301	0	9.54178	16	101	344	cell division
GO:0022402	0	6.97876	17	87	394	cell cycle process
GO:0007067	0	12.09882	8	61	175	mitosis
GO:0006807	0	2.26305	219	370	4649	nitrogen compound metabolic process
GO:0071840	0	2.32487	165	300	3515	cellular component organization or biogenesis
GO:0007059	0	13.50008	5	38	98	chromosome segregation
GO:0006281	0	5.56945	15	66	321	DNA repair
GO:0006139	0	2.1394	192	320	4185	nucleobase-containing compound metabolic process
GO:0051325	0	6.95954	9	47	190	interphase
GO:0071842	0	2.23124	119	219	2601	cellular component organization at cellular level
GO:0006323	0	10.96241	5	33	96	DNA packaging
GO:0007017	0	5.67163	11	49	235	microtubule-based process
GO:0045786	0	5.03255	13	54	282	negative regulation of cell cycle
GO:0051276	0	3.44368	27	80	586	chromosome organization
GO:0044249	0	1.95967	171	275	3740	cellular biosynthetic process
GO:0010564	0	6.69623	7	34	144	regulation of cell cycle process
GO:0006310	0	6.07653	8	36	160	DNA recombination
GO:0006260	0	5.90243	7	34	158	DNA replication
GO:0006259	0	6.31407	6	31	145	DNA metabolic process
GO:0009161	0	18.29423	2	16	34	ribonucleoside monophosphate metabolic process
GO:0044281	0	2.29552	54	110	1205	small molecule metabolic process
GO:0007049	0	4.24271	11	40	284	cell cycle
GO:0006270	0	41.00556	1	12	18	DNA-dependent DNA replication initiation
GO:0009124	0	19.27525	1	15	31	nucleoside monophosphate biosynthetic process
GO:0051321	0	5.47345	7	30	144	meiotic cell cycle
GO:0009126	0	19.16923	1	14	29	purine nucleoside monophosphate metabolic process
GO:0006950	0	1.90543	94	158	1987	response to stress
GO:0044238	0	2.6557	32	73	960	primary metabolic process
GO:0006695	0	14.0138	2	15	37	cholesterol biosynthetic process
GO:0000724	0	12.19019	2	16	43	double-strand break repair via homologous recombination
GO:0031570	0	7.64185	3	20	74	DNA integrity checkpoint
GO:0000082	0	6.40178	4	22	93	G1/S transition of mitotic cell cycle
GO:0009059	0	1.66461	158	231	3356	macromolecule biosynthetic process
GO:0016125	0	5.79801	5	23	105	sterol metabolic process
GO:0006333	0	6.76234	4	20	81	chromatin assembly or disassembly
GO:0007010	0	2.39708	33	71	699	cytoskeleton organization
GO:0006261	0	10.41903	2	15	45	DNA-dependent DNA replication
GO:0031023	0	8.32886	3	17	59	microtubule organizing center organization
GO:0007093	0	7.56455	3	18	67	mitotic cell cycle checkpoint
GO:0000070	0	14.5914	1	12	29	mitotic sister chromatid segregation
GO:0019320	0	6.99212	3	18	71	hexose catabolic process
GO:0007076	0	54.44075	1	8	11	mitotic chromosome condensation
GO:0042180	0	2.31844	33	70	709	cellular ketone metabolic process
GO:0019752	0	2.34955	32	68	680	carboxylic acid metabolic process
GO:0006753	0	2.07829	47	89	1002	nucleoside phosphate metabolic process
GO:0006334	0	8.80246	2	15	50	nucleosome assembly
GO:0009259	0	2.28875	34	70	717	ribonucleotide metabolic process
GO:0009152	0	7.16838	3	17	66	purine ribonucleotide biosynthetic process
GO:0043412	0	1.70518	107	165	2282	macromolecule modification
GO:0019538	0	1.58871	164	231	3479	protein metabolic process
GO:0006793	0	1.78464	80	130	1703	phosphorus metabolic process
GO:0006096	0	7.70006	3	15	55	glycolysis
GO:0034453	0	11.17363	2	12	34	microtubule anchoring
GO:0090407	0	3.28595	12	34	249	organophosphate biosynthetic process
GO:0009056	0	1.78054	79	128	1707	catabolic process
GO:0034502	0	15.73039	1	10	23	protein localization to chromosome
GO:0034622	0	2.66294	19	46	407	cellular macromolecular complex assembly
GO:0033043	0	2.58747	21	48	436	regulation of organelle organization

Supplementary table 3-11. Representation of the gene ontology analysis that was done on the 9 clusters as identified in the gene expression analysis of the resting and activated B and T cells (see supplemental figure 10). Supplementary table 3-11 depict the list of biological processes that were found in the GO analysis for gene clusters 1-9, respectively. The lists of genes are sorted based on the P-values, from small to large.

Supplemental Figure 3b (continued from Suppl. table 3a)

Upregulated genes in cycling B and T cells (day 4)

GOBPID	Pvalue	OddsRatio	ExpCount	Count	Size	Term
GO:0031572	0	11.84848	1	11	30	G2/M transition DNA damage checkpoint
GO:0000090	0	122.23425	0	6	7	mitotic anaphase
GO:0071824	0	5.36608	4	18	87	protein-DNA complex subunit organization
GO:0009168	0	23.43048	1	8	15	purine ribonucleoside monophosphate biosynthetic process
GO:0044260	0	1.53359	181	242	4444	cellular macromolecule metabolic process
GO:0022607	0	1.83568	61	102	1287	cellular component assembly
GO:0009411	0	5.14166	4	18	90	response to UV
GO:0048519	0	1.56085	137	194	2918	negative regulation of biological process
GO:0016053	0	3.08562	12	32	247	organic acid biosynthetic process
GO:0046483	0	1.88903	51	88	1084	heterocycle metabolic process
GO:0051656	0	5.66519	3	16	74	establishment of organelle localization
GO:0046040	0	61.11381	0	6	8	IMP metabolic process
GO:0051988	0	61.11381	0	6	8	regulation of attachment of spindle microtubules to kinetochore
GO:0009163	0	11.35775	1	10	28	nucleoside biosynthetic process
GO:0000079	0	6.37934	3	14	59	regulation of cyclin-dependent protein kinase activity
GO:0031145	0	23.78761	1	7	13	anaphase-promoting complex-dependent proteasomal ubiquitin-dependent protein catabolic process
GO:0032465	0	12.25477	1	9	24	regulation of cytokinesis
GO:1901135	0	1.84115	51	87	1087	carbohydrate derivative metabolic process
GO:0016052	0	4.11627	6	20	120	carbohydrate catabolic process
GO:0007088	0	5.49526	3	15	71	regulation of mitosis
GO:0044237	0	2.90157	12	31	396	cellular metabolic process
GO:0051052	0	3.25354	9	26	191	regulation of DNA metabolic process
GO:0043067	0	1.81461	52	88	1114	regulation of programmed cell death
GO:0032392	0	8.03611	2	11	39	DNA geometric change
GO:0006006	0	3.43743	8	24	168	glucose metabolic process
GO:0033261	0	9.29069	2	10	32	regulation of S phase
GO:0009113	0	101.74945	0	5	6	purine nucleobase biosynthetic process
GO:0022900	0	4.30136	5	18	104	electron transport chain
GO:0050000	0	13.60354	1	8	20	chromosome localization
GO:0005996	0	3.08079	10	27	208	monosaccharide metabolic process
GO:0070925	0	4.62463	4	16	87	organelle assembly
GO:0009058	0	4.76804	4	15	94	biosynthetic process
GO:0009165	0	2.96522	10	27	216	nucleotide biosynthetic process
GO:0008219	0	1.68084	67	104	1417	cell death
GO:0006461	0	1.97653	33	61	707	protein complex assembly
GO:0008652	0	4.43642	4	16	90	cellular amino acid biosynthetic process
GO:0072528	0	11.65891	1	8	22	pyrimidine-containing compound biosynthetic process
GO:0007018	0	3.79169	6	19	122	microtubule-based movement
GO:0006268	0	50.87196	0	5	7	DNA unwinding involved in replication
GO:0006177	0	Inf	0	4	4	GMP biosynthetic process
GO:0006189	0	Inf	0	4	4	'de novo' IMP biosynthetic process
GO:0006694	0.00001	3.71906	6	19	124	steroid biosynthetic process
GO:0046034	0.00001	2.64874	13	31	273	ATP metabolic process
GO:0009628	0.00001	2.0513	28	53	592	response to abiotic stimulus
GO:0007127	0.00001	5.32392	3	13	63	meiosis I
GO:0071777	0.00001	20.36685	1	6	12	positive regulation of cell cycle cytokinesis
GO:0016310	0.00001	2.64182	12	30	274	phosphorylation
GO:0055114	0.00001	1.93898	33	59	705	oxidation-reduction process
GO:0009142	0.00001	5.58014	3	12	56	nucleoside triphosphate biosynthetic process
GO:0051289	0.00001	5.58014	3	12	56	protein homotetramerization
GO:0006163	0.00001	1.83307	40	69	858	purine nucleotide metabolic process
GO:0008608	0.00001	34.13519	0	5	8	attachment of spindle microtubules to kinetochore
GO:0009112	0.00001	9.5999	1	8	25	nucleobase metabolic process
GO:0090307	0.00001	17.45635	1	6	13	spindle assembly involved in mitosis
GO:0006760	0.00002	11.88993	1	7	19	folic acid-containing compound metabolic process
GO:0031324	0.00002	1.67276	58	90	1224	negative regulation of cellular metabolic process
GO:0008637	0.00002	5.22311	3	12	59	apoptotic mitochondrial changes
GO:0010212	0.00002	4.1545	4	15	89	response to ionizing radiation
GO:0006974	0.00002	3.68908	5	17	118	response to DNA damage stimulus

Supplementary table 3-11. Representation of the gene ontology analysis that was done on the 9 clusters as identified in the gene expression analysis of the resting and activated B and T cells (see supplemental figure 10). Supplementary table 3-11 depict the list of biological processes that were found in the GO analysis for gene clusters 1-9, respectively. The lists of genes are sorted based on the P-values, from small to large.

Supplemental Table 4.

Upregulated in cycling T cells (day 4)

GOBPID	Pvalue	OddsRatio	ExpCount	Count	Size	Term
GO:0016265	0	3.89883	12	37	1422	death
GO:0010740	0	7.27351	3	18	349	positive regulation of intracellular protein kinase cascade
GO:0031347	0	7.35463	2	15	283	regulation of defense response
GO:0006954	0	7.31752	2	15	297	inflammatory response
GO:0030003	0	6.6614	3	16	333	cellular cation homeostasis
GO:0051179	0	2.59032	32	63	3942	localization
GO:0043067	0	3.7224	9	29	1114	regulation of programmed cell death
GO:0055082	0	4.69064	5	21	624	cellular chemical homeostasis
GO:0006950	0	2.97687	16	40	1987	response to stress
GO:0035556	0	3.25275	12	33	1463	intracellular signal transduction
GO:0048584	0	3.73271	8	26	981	positive regulation of response to stimulus
GO:0008284	0	4.51236	5	20	613	positive regulation of cell proliferation
GO:0032101	0	5.99022	3	15	343	regulation of response to external stimulus
GO:0007204	0	8.7299	1	11	173	elevation of cytosolic calcium ion concentration
GO:0050801	0	4.3531	5	20	634	ion homeostasis
GO:0055065	0	6.1002	3	14	313	metal ion homeostasis
GO:0032943	0	7.93851	2	11	189	mononuclear cell proliferation
GO:0072507	0	6.3833	2	13	277	divalent inorganic cation homeostasis
GO:0050864	0	12.98149	1	8	86	regulation of B cell activation
GO:0006810	0	2.42389	26	51	3180	transport
GO:0051094	0	4.02704	6	20	682	positive regulation of developmental process
GO:0023056	0	3.84812	6	21	751	positive regulation of signaling
GO:0009966	0	2.93753	13	32	1548	regulation of signal transduction
GO:0042110	0	5.99174	2	13	294	T cell activation
GO:0002376	0	3.10477	10	28	1264	immune system process
GO:2000026	0	3.24303	9	26	1117	regulation of multicellular organismal development
GO:0050728	0	14.44165	1	7	68	negative regulation of inflammatory response
GO:0060548	0	4.0246	5	19	645	negative regulation of cell death
GO:0070887	0	3.1964	9	26	1132	cellular response to chemical stimulus
GO:0042592	0	3.19028	9	26	1134	homeostatic process
GO:0060401	0	13.3441	1	7	73	cytosolic calcium ion transport
GO:0043066	0	3.99656	5	18	612	negative regulation of apoptotic process
GO:0030334	0	4.91109	3	14	384	regulation of cell migration
GO:0009893	0	2.6637	15	34	1812	positive regulation of metabolic process
GO:0010647	0	3.62818	6	20	752	positive regulation of cell communication
GO:0001525	0	5.26738	3	13	332	angiogenesis
GO:0050670	0	8.22345	1	9	148	regulation of lymphocyte proliferation
GO:0010035	0	5.65657	2	12	285	response to inorganic substance
GO:0032940	0	4.0272	5	17	571	secretion by cell
GO:0070663	0.00001	7.88069	1	9	154	regulation of leukocyte proliferation
GO:0002694	0.00001	5.51312	2	12	292	regulation of leukocyte activation
GO:0051209	0.00001	15.31893	0	6	55	release of sequestered calcium ion into cytosol
GO:0051282	0.00001	15.31893	0	6	55	regulation of sequestering of calcium ion
GO:0002237	0.00001	6.6408	2	10	202	response to molecule of bacterial origin
GO:0045785	0.00001	8.94427	1	8	121	positive regulation of cell adhesion
GO:0006874	0.00001	5.86852	2	11	251	cellular calcium ion homeostasis
GO:0050776	0.00001	4.92181	3	13	354	regulation of immune response
GO:0070838	0.00001	5.3202	2	12	302	divalent metal ion transport
GO:0051270	0.00001	4.45585	3	14	421	regulation of cellular component movement
GO:0030890	0.00001	20.06598	0	5	36	positive regulation of B cell proliferation
GO:0033993	0.00001	20.06598	0	5	36	response to lipid
GO:0051241	0.00001	5.08844	3	12	315	negative regulation of multicellular organismal process
GO:0040012	0.00001	4.36809	4	14	429	regulation of locomotion
GO:0002700	0.00001	13.1634	1	6	63	regulation of production of molecular mediator of immune response
GO:0051238	0.00001	18.84789	0	5	38	sequestering of metal ion
GO:0001819	0.00002	6.79357	1	9	177	positive regulation of cytokine production
GO:0051247	0.00002	3.46968	6	18	699	positive regulation of protein metabolic process
GO:0044093	0.00002	3.01306	8	22	993	positive regulation of molecular function
GO:0097285	0.00002	5.20821	2	11	281	cell-type specific apoptotic process
GO:0032800	0.00002	29.10512	0	4	21	receptor biosynthetic process
GO:0061298	0.00002	29.10512	0	4	21	retina vasculature development in camera-type eye
GO:0050867	0.00003	6.3724	2	9	188	positive regulation of cell activation
GO:0043065	0.00003	3.80005	4	15	527	positive regulation of apoptotic process
GO:0048731	0.00003	2.16325	24	44	2928	system development
GO:0001568	0.00004	3.9226	4	14	475	blood vessel development
GO:0051707	0.00004	3.9226	4	14	475	response to other organism
GO:0002673	0.00004	14.8021	0	5	47	regulation of acute inflammatory response
GO:0009894	0.00004	3.68174	4	15	543	regulation of catabolic process
GO:0042327	0.00005	3.86248	4	14	482	positive regulation of phosphorylation
GO:0010942	0.00005	3.64622	5	15	548	positive regulation of cell death
GO:0071248	0.00005	14.1278	0	5	49	cellular response to metal ion
GO:0009653	0.00005	2.39548	15	31	1793	anatomical structure morphogenesis
GO:0010562	0.00005	3.80414	4	14	489	positive regulation of phosphorus metabolic process
GO:0045124	0.00006	22.48446	0	4	26	regulation of bone resorption
GO:0051251	0.00006	6.54889	1	8	162	positive regulation of lymphocyte activation
GO:0048519	0.00006	2.21842	20	38	2603	negative regulation of biological process
GO:0007260	0.00006	13.51214	0	5	51	tyrosine phosphorylation of STAT protein
GO:0048771	0.00006	7.70614	1	7	121	tissue remodeling
GO:0008285	0.0001	3.78556	4	13	454	negative regulation of cell proliferation

Supplementary table 3-11. Representation of the gene ontology analysis that was done on the 9 clusters as identified in the gene expression analysis of the resting and activated B and T cells (see supplemental figure 10). Supplementary table 3-11 depict the list of biological processes that were found in the GO analysis for gene clusters 1-9, respectively. The lists of genes are sorted based on the P-values, from small to large.

Supplemental Table 5.

Upregulated genes in resting T cells (day 0)

GOBPID	Pvalue	OddsRatio	ExpCount	Count	Size	Term
GO:0006793	0	2.9123	20	48	1703	phosphorus metabolic process
GO:0016310	0	2.98903	16	42	1427	phosphorylation
GO:0045321	0	4.42376	6	23	510	leukocyte activation
GO:0002764	0	7.92075	2	12	149	immune response-regulating signaling pathway
GO:0001817	0	4.84298	4	17	339	regulation of cytokine production
GO:0050790	0	2.67374	17	39	1451	regulation of catalytic activity
GO:0034097	0	4.89087	4	16	315	response to cytokine stimulus
GO:0006952	0	3.34561	8	25	724	defense response
GO:0002520	0	3.65381	7	22	581	immune system development
GO:0019221	0	6.4876	2	12	179	cytokine-mediated signaling pathway
GO:0002684	0	4.13096	5	18	418	positive regulation of immune system process
GO:0006955	0	3.49269	7	22	606	immune response
GO:0007243	0	3.27326	8	24	707	intracellular protein kinase cascade
GO:0030097	0	3.71138	6	20	517	hemopoiesis
GO:0048584	0	2.86865	11	29	981	positive regulation of response to stimulus
GO:0030155	0.00001	4.97898	3	13	249	regulation of cell adhesion
GO:0002253	0.00001	5.84812	2	11	180	activation of immune response
GO:0042110	0.00001	4.53053	3	14	294	T cell activation
GO:0023052	0.00001	1.87004	56	85	4850	signaling
GO:0007154	0.00001	1.84918	57	86	4960	cell communication
GO:0071310	0.00001	2.84771	10	25	841	cellular response to organic substance
GO:2000503	0.00001	130.99315	0	3	5	positive regulation of natural killer cell chemotaxis
GO:0030098	0.00001	4.95649	3	12	230	lymphocyte differentiation
GO:0032651	0.00002	18.34005	0	5	29	regulation of interleukin-1 beta production
GO:0036211	0.00002	2.11517	25	47	2202	protein modification process
GO:0051716	0.00002	1.79937	61	90	5346	cellular response to stimulus
GO:0031325	0.00002	2.24246	19	39	1699	positive regulation of cellular metabolic process
GO:0050727	0.00002	5.73627	2	10	166	regulation of inflammatory response
GO:0031401	0.00003	3.22161	6	19	559	positive regulation of protein modification process
GO:0032729	0.00003	16.29971	0	5	32	positive regulation of interferon-gamma production
GO:0072678	0.00004	26.97812	0	4	17	T cell migration
GO:0050789	0.00004	1.72679	98	128	8561	regulation of biological process
GO:0019220	0.00004	2.689	10	24	849	regulation of phosphate metabolic process
GO:0048247	0.00005	25.0498	0	4	18	lymphocyte chemotaxis
GO:0032496	0.00005	5.19828	2	10	182	response to lipopolysaccharide
GO:0002429	0.00005	8.05497	1	7	84	immune response-activating cell surface receptor signaling pathway
GO:0035556	0.00006	2.8586	8	21	756	intracellular signal transduction
GO:0051247	0.00006	2.84293	8	21	699	positive regulation of protein metabolic process
GO:0030335	0.00006	4.60741	3	11	225	positive regulation of cell migration
GO:0002697	0.00006	5.0501	2	10	187	regulation of immune effector process
GO:0006140	0.00006	3.51825	5	15	401	regulation of nucleotide metabolic process
GO:0048585	0.00007	2.81708	8	21	705	negative regulation of response to stimulus
GO:0001932	0.00007	2.80009	8	21	709	regulation of protein phosphorylation
GO:0002758	0.00008	9.45982	1	6	62	innate immune response-activating signal transduction
GO:0032612	0.00008	12.93914	0	5	39	interleukin-1 production
GO:0010820	0.00008	52.38904	0	3	8	positive regulation of T cell chemotaxis
GO:0009607	0.00008	3.16072	6	17	506	response to biotic stimulus
GO:2000401	0.00009	20.62601	0	4	21	regulation of lymphocyte migration
GO:0051272	0.0001	4.36	3	11	237	positive regulation of cellular component movement
GO:0045089	0.0001	7.2086	1	7	93	positive regulation of innate immune response

Supplementary table 3-11. Representation of the gene ontology analysis that was done on the 9 clusters as identified in the gene expression analysis of the resting and activated B and T cells (see supplemental figure 10). Supplementary table 3-11 depict the list of biological processes that were found in the GO analysis for gene clusters 1-9, respectively. The lists of genes are sorted based on the P-values, from small to large.

Supplemental Table 6

Upregulated genes in resting and cycling T cells (day 0 and day 4)

GOBPID	Pvalue	OddsRatio	ExpCount	Count	Size	Term
GO:0002429	0	23.37377	1	16	84	immune response-activating cell surface receptor signaling pathway
GO:0002682	0	5.36871	7	32	659	regulation of immune system process
GO:0002521	0	7.52742	4	23	334	leukocyte differentiation
GO:0048583	0	3.37403	21	57	1983	regulation of response to stimulus
GO:0050852	0	55.12212	0	9	25	T cell receptor signaling pathway
GO:0006955	0	5.43217	7	30	606	immune response
GO:0002764	0	11.90977	2	16	149	immune response-regulating signaling pathway
GO:0010646	0	3.62956	15	45	1398	regulation of cell communication
GO:0048534	0	5.26883	6	27	554	hemopoietic or lymphoid organ development
GO:0002253	0	9.64278	2	16	180	activation of immune response
GO:0001775	0	7.89431	3	18	266	cell activation
GO:0023051	0	3.06449	19	49	1798	regulation of signaling
GO:0002696	0	8.99327	2	15	179	positive regulation of leukocyte activation
GO:0050865	0	6.44446	3	19	313	regulation of cell activation
GO:0048522	0	2.52074	32	64	2933	positive regulation of cellular process
GO:0045058	0	60.06281	0	5	13	T cell selection
GO:0050850	0	31.55611	0	6	24	positive regulation of calcium-mediated signaling
GO:0033077	0	12.72067	1	8	68	T cell differentiation in thymus
GO:0046632	0	12.72067	1	8	68	alpha-beta T cell differentiation
GO:0001816	0	4.57853	4	17	382	cytokine production
GO:0051249	0	9.52811	1	9	103	regulation of lymphocyte activation
GO:0046649	0	7.94776	1	10	144	lymphocyte activation
GO:0009967	0	3.56267	7	22	663	positive regulation of signal transduction
GO:0019932	0	7.80898	1	10	133	second-messenger-mediated signaling
GO:0009968	0	3.74291	6	20	549	negative regulation of signal transduction
GO:0046651	0	17.34349	0	6	40	lymphocyte proliferation
GO:0070661	0	16.84175	0	6	41	leukocyte proliferation
GO:0045060	0.00001	46.89461	0	4	12	negative thymic T cell selection
GO:0042102	0.00001	11.45831	1	7	65	positive regulation of T cell proliferation
GO:0010941	0.00001	2.70158	12	30	1154	regulation of cell death
GO:0042981	0.00001	2.72538	12	29	1103	regulation of apoptotic process
GO:0007166	0.00001	2.06487	33	58	3077	cell surface receptor signaling pathway
GO:0043368	0.00001	41.68192	0	4	13	positive T cell selection
GO:0006796	0.00001	2.34619	18	38	1703	phosphate-containing compound metabolic process
GO:0050870	0.00001	13.31399	1	6	50	positive regulation of T cell activation
GO:0048519	0.00001	2.04386	31	55	2918	negative regulation of biological process
GO:0050670	0.00003	6.18264	2	9	148	regulation of lymphocyte proliferation
GO:0006464	0.00003	2.11141	24	44	2202	cellular protein modification process
GO:0042035	0.00004	8.51135	1	7	85	regulation of cytokine biosynthetic process
GO:0043412	0.00004	2.08616	25	45	2282	macromolecule modification
GO:0030217	0.00004	14.86076	0	5	39	T cell differentiation
GO:0070663	0.00004	5.92494	2	9	154	regulation of leukocyte proliferation
GO:0045621	0.00005	10.12288	1	6	62	positive regulation of lymphocyte differentiation
GO:0016265	0.00005	2.32191	15	32	1422	death
GO:0046640	0.00005	23.4375	0	4	20	regulation of alpha-beta T cell proliferation
GO:0050856	0.00005	23.4375	0	4	20	regulation of T cell receptor signaling pathway
GO:0043068	0.00007	3.21905	6	17	532	positive regulation of programmed cell death
GO:0032673	0.00007	22.05767	0	4	21	regulation of interleukin-4 production
GO:2000026	0.00007	2.47036	12	27	1117	regulation of multicellular organismal development
GO:0002460	0.00008	5.50396	2	9	165	adaptive immune response based on somatic recombination of immune receptors built from immunoglobulin superfamily domains
GO:0051716	0.00008	1.7509	57	83	5346	cellular response to stimulus
GO:0018108	0.00008	4.88173	2	10	206	peptidyl-tyrosine phosphorylation
GO:0032946	0.00009	7.37186	1	7	97	positive regulation of mononuclear cell proliferation
GO:0046631	0.00009	20.47476	0	4	23	alpha-beta T cell activation
GO:0006917	0.00009	4.07976	3	12	295	induction of apoptosis
GO:0042107	0.00009	7.29047	1	7	98	cytokine metabolic process
GO:0050857	0.0001	46.67073	0	3	9	positive regulation of antigen receptor-mediated signaling pathway

Supplementary table 3-11. Representation of the gene ontology analysis that was done on the 9 clusters as identified in the gene expression analysis of the resting and activated B and T cells (see supplemental figure 10). Supplementary table 3-11 depict the list of biological processes that were found in the GO analysis for gene clusters 1-9, respectively. The lists of genes are sorted based on the P-values, from small to large.

Supplemental Table 7.

Genes down-regulated in cycling T cells (day 4)

GOBPID	Pvalue	OddsRatio	ExpCount	Count	Size	Term
GO:0002684	0	9.33713	2	14	418	positive regulation of immune system process
GO:0050776	0	9.22173	2	12	354	regulation of immune response
GO:0019884	0	60.9019	0	5	25	antigen processing and presentation of exogenous antigen
GO:0019886	0	106.97778	0	4	13	antigen processing and presentation of exogenous peptide antigen via MHC class II
GO:0002252	0	8.51482	2	11	346	immune effector process
GO:0048002	0	40.58017	0	5	35	antigen processing and presentation of peptide antigen
GO:0042113	0	12.56908	1	8	168	B cell activation
GO:0002504	0	74.04615	0	4	17	antigen processing and presentation of peptide or polysaccharide antigen via MHC class II
GO:0006959	0	21.40022	0	6	75	humoral immune response
GO:0032943	0	11.09858	1	8	189	mononuclear cell proliferation
GO:0002757	0	13.07724	1	7	140	immune response-activating signal transduction
GO:0045089	0	16.95668	0	6	93	positive regulation of innate immune response
GO:0050670	0	12.33011	1	7	148	regulation of lymphocyte proliferation
GO:0070663	0	11.82313	1	7	154	regulation of leukocyte proliferation
GO:0045087	0.00001	11.03123	1	7	177	innate immune response
GO:0006950	0.00001	3.32145	9	23	1987	response to stress
GO:0050868	0.00001	19.6029	0	5	67	negative regulation of T cell activation
GO:0050867	0.00002	9.58513	1	7	188	positive regulation of cell activation
GO:0016064	0.00004	14.80627	0	5	87	immunoglobulin mediated immune response
GO:0002449	0.00005	10.07323	1	6	152	lymphocyte mediated immunity
GO:0051251	0.00007	9.42258	1	6	162	positive regulation of lymphocyte activation
GO:0002460	0.00008	9.24335	1	6	165	adaptive immune response based on somatic recombination of immune receptors built from immunoglobulin superfamily domains
GO:0002695	0.0001	12.00902	0	5	106	negative regulation of leukocyte activation

Supplementary table 3-11. Representation of the gene ontology analysis that was done on the 9 clusters as identified in the gene expression analysis of the resting and activated B and T cells (see supplemental figure 10). Supplementary table 3-11 depict the list of biological processes that were found in the GO analysis for gene clusters 1-9, respectively. The lists of genes are sorted based on the P-values, from small to large.

Supplemental Table 8.
Down regulated genes in cylcing B and T cells (day 4)

GOBPID	Pvalue	OddsRatio	ExpCount	Count	Size	Term
GO:0080090	0	2.22066	84	149	3828	regulation of primary metabolic process
GO:0051171	0	2.24085	67	123	3017	regulation of nitrogen compound metabolic process
GO:0051252	0	2.18627	56	104	2533	regulation of RNA metabolic process
GO:0035556	0	2.5116	32	71	1463	intracellular signal transduction
GO:0032774	0	2.02833	56	99	2550	RNA biosynthetic process
GO:0036211	0	2.09467	49	89	2202	protein modification process
GO:0048534	0	3.16594	12	35	554	hemopoietic or lymphoid organ development
GO:0001775	0	3.01635	13	35	579	cell activation
GO:0010627	0	3.04833	12	33	539	regulation of intracellular protein kinase cascade
GO:0018193	0	2.8953	13	34	583	peptidyl-amino acid modification
GO:0009966	0	2.39564	21	46	1009	regulation of signal transduction
GO:0034645	0	1.77418	72	112	3271	cellular macromolecule biosynthetic process
GO:0016310	0	2.09951	31	60	1427	phosphorylation
GO:0010556	0	2.06894	34	63	1622	regulation of macromolecule biosynthetic process
GO:0031326	0	2.03705	34	63	1650	regulation of cellular biosynthetic process
GO:0002684	0	3.06491	9	26	418	positive regulation of immune system process
GO:0002694	0	3.5598	6	21	292	regulation of leukocyte activation
GO:0048523	0	1.77365	58	92	2629	negative regulation of cellular process
GO:0044093	0	2.2335	22	45	993	positive regulation of molecular function
GO:0006357	0.00001	2.10206	26	50	1173	regulation of transcription from RNA polymerase II promoter
GO:0043065	0.00001	2.69569	12	29	527	positive regulation of apoptotic process
GO:0090304	0.00001	1.65593	79	115	3560	nucleic acid metabolic process
GO:0050790	0.00001	1.97792	32	58	1451	regulation of catalytic activity
GO:0034641	0.00001	1.59581	102	141	4607	cellular nitrogen compound metabolic process
GO:0031329	0.00001	2.75824	11	27	479	regulation of cellular catabolic process
GO:0046649	0.00001	2.85054	9	25	429	lymphocyte activation
GO:0010467	0.00001	1.61989	83	119	3763	gene expression
GO:0006955	0.00001	2.49783	13	31	606	immune response
GO:0010942	0.00001	2.58366	12	29	548	positive regulation of cell death
GO:0043123	0.00002	5.41205	2	11	103	positive regulation of I-kappaB kinase/NF-kappaB cascade
GO:0009893	0.00002	1.83199	40	67	1812	positive regulation of metabolic process
GO:0009058	0.00002	1.58493	92	128	4152	biosynthetic process
GO:0001932	0.00002	2.3387	16	34	709	regulation of protein phosphorylation
GO:0051270	0.00002	2.77891	9	24	421	regulation of cellular component movement
GO:0044267	0.00002	1.66711	63	94	2834	cellular protein metabolic process
GO:0050793	0.00002	1.92292	31	55	1406	regulation of developmental process
GO:0043547	0.00003	3.47679	5	17	240	positive regulation of GTPase activity
GO:0050789	0.00003	1.53812	148	187	7155	regulation of biological process
GO:0006355	0.00003	1.99843	27	49	1287	regulation of transcription, DNA-dependent
GO:0046578	0.00003	3.31387	6	18	266	regulation of Ras protein signal transduction
GO:0030099	0.00003	3.43009	5	17	243	myeloid cell differentiation
GO:0043067	0.00003	2.01835	25	46	1114	regulation of programmed cell death
GO:0018107	0.00003	7.35395	1	8	57	peptidyl-threonine phosphorylation
GO:0008360	0.00004	5.50974	2	10	92	regulation of cell shape
GO:0031328	0.00004	1.96503	27	49	1220	positive regulation of cellular biosynthetic process
GO:0050863	0.00004	3.92551	4	14	176	regulation of T cell activation
GO:0002274	0.00004	4.92743	2	11	112	myeloid leukocyte activation
GO:0010557	0.00004	1.99068	25	46	1128	positive regulation of macromolecule biosynthetic process
GO:0006911	0.00005	16.00203	0	5	19	phagocytosis, engulfment
GO:0033124	0.00005	3.15889	6	18	278	regulation of GTP catabolic process
GO:0002521	0.00006	2.9121	7	20	334	leukocyte differentiation
GO:0006793	0.00006	1.78879	38	62	1703	phosphorus metabolic process
GO:0019220	0.00007	2.11592	19	37	849	regulation of phosphate metabolic process
GO:0010035	0.00007	3.07492	6	18	285	response to inorganic substance
GO:0048584	0.00007	2.03192	22	41	981	positive regulation of response to stimulus
GO:0001817	0.00007	2.86568	7	20	339	regulation of cytokine production
GO:0002697	0.00007	3.67375	4	14	187	regulation of immune effector process
GO:0045619	0.00009	4.90825	2	10	102	regulation of lymphocyte differentiation
GO:0051247	0.00009	2.21722	15	32	699	positive regulation of protein metabolic process
GO:0051592	0.00009	6.31914	1	8	65	response to calcium ion

Supplementary table 3-11. Representation of the gene ontology analysis that was done on the 9 clusters as identified in the gene expression analysis of the resting and activated B and T cells (see supplemental figure 10). Supplementary table 3-11 depict the list of biological processes that were found in the GO analysis for gene clusters 1-9, respectively. The lists of genes are sorted based on the P-values, from small to large.

Supplemental Table 9.
Upregulated genes in B cells (resting and cycling cells)

GOBPID	Pvalue	OddsRatio	ExpCount	Count	Size	Term
GO:0006955	0	5.54409	10	47	606	immune response
GO:0002764	0	10.73164	3	22	149	immune response-regulating signaling pathway
GO:0050853	0	36.75176	0	11	29	B cell receptor signaling pathway
GO:0002253	0	7.67997	3	20	180	activation of immune response
GO:0002504	0	52.97945	0	8	17	antigen processing and presentation of peptide or polysaccharide antigen via MHC class II
GO:0002252	0	6.09514	4	21	238	immune effector process
GO:0006793	0	2.52925	29	63	1703	phosphorus metabolic process
GO:0016310	0	2.60126	24	55	1427	phosphorylation
GO:0042113	0	10.75273	1	12	82	B cell activation
GO:0002429	0	10.02646	1	12	84	immune response-activating cell surface receptor signaling pathway
GO:0002460	0	6.5165	3	16	165	adaptive immune response based on somatic recombination of immune receptors built from immunoglobulin superfamily domains
GO:0050670	0	6.82801	3	15	148	regulation of lymphocyte proliferation
GO:0019886	0	50.77437	0	6	13	antigen processing and presentation of exogenous peptide antigen via MHC class II
GO:0070663	0	6.5312	3	15	154	regulation of leukocyte proliferation
GO:0009617	0	4.35393	6	22	332	response to bacterium
GO:0065009	0	2.24815	31	62	1855	regulation of molecular function
GO:0001816	0	3.93316	6	23	382	cytokine production
GO:0002684	0	4.3122	5	20	310	positive regulation of immune system process
GO:0051704	0	2.96015	12	31	681	multi-organism process
GO:0002694	0	4.23661	5	19	292	regulation of leukocyte activation
GO:0050864	0	9.90706	1	9	64	regulation of B cell activation
GO:0048002	0	14.83906	1	7	35	antigen processing and presentation of peptide antigen
GO:0048584	0	2.52086	17	38	981	positive regulation of response to stimulus
GO:0002449	0	5.62432	3	13	152	lymphocyte mediated immunity
GO:0002520	0	3.00022	10	27	581	immune system development
GO:0002281	0	32.80366	0	5	14	macrophage activation involved in immune response
GO:0030889	0	32.80366	0	5	14	negative regulation of B cell proliferation
GO:0032496	0	5.0197	3	14	182	response to lipopolysaccharide
GO:0019884	0	18.69454	0	6	25	antigen processing and presentation of exogenous antigen
GO:0035556	0	2.19489	25	49	1463	intracellular signal transduction
GO:0043547	0	4.31741	4	16	240	positive regulation of GTPase activity
GO:0030811	0	3.84343	5	18	302	regulation of nucleotide catabolic process
GO:0006954	0	3.52679	6	20	365	inflammatory response
GO:0010604	0.00001	2.07783	28	53	1673	positive regulation of macromolecule metabolic process
GO:0071900	0.00001	3.78024	5	17	289	regulation of protein serine/threonine kinase activity
GO:0002758	0.00001	8.80901	1	8	62	innate immune response-activating signal transduction
GO:0034110	0.00001	22.70545	0	5	18	regulation of homotypic cell-cell adhesion
GO:0043549	0.00001	2.92217	9	24	526	regulation of kinase activity
GO:0002366	0.00001	5.48318	2	11	131	leukocyte activation involved in immune response
GO:0065008	0.00001	1.91848	35	61	2092	regulation of biological quality
GO:0002224	0.00001	10.12706	1	7	48	toll-like receptor signaling pathway
GO:0009607	0.00002	3.3421	6	19	374	response to biotic stimulus
GO:0071216	0.00002	5.43757	2	11	132	cellular response to biotic stimulus
GO:0016064	0.00002	6.87373	1	9	87	immunoglobulin mediated immune response
GO:0010543	0.00002	19.67598	0	5	20	regulation of platelet activation
GO:0036211	0.00002	1.88354	37	63	2202	protein modification process
GO:0009894	0.00002	2.82386	9	24	543	regulation of catabolic process
GO:0043085	0.00002	3.06905	7	21	455	positive regulation of catalytic activity
GO:0051347	0.00002	3.3937	6	18	339	positive regulation of transferase activity
GO:0033124	0.00002	3.68376	5	16	278	regulation of GTP catabolic process
GO:0031663	0.00002	12.2417	1	6	35	lipopolysaccharide-mediated signaling pathway
GO:0006952	0.00002	3.34908	6	18	359	defense response
GO:0010033	0.00002	2.07564	24	45	1403	response to organic substance
GO:0045577	0.00003	17.35934	0	5	22	regulation of B cell differentiation
GO:0045089	0.00003	6.38073	2	9	93	positive regulation of innate immune response
GO:0051241	0.00003	3.44564	5	17	315	negative regulation of multicellular organismal process
GO:0030890	0.00003	11.83302	1	6	36	positive regulation of B cell proliferation
GO:0031323	0.00003	1.68074	65	96	3867	regulation of cellular metabolic process
GO:0043406	0.00003	4.98154	2	11	143	positive regulation of MAP kinase activity
GO:0044270	0.00003	2.46658	13	29	751	cellular nitrogen compound catabolic process
GO:0032655	0.00003	11.45071	1	6	37	regulation of interleukin-12 production
GO:0032946	0.00004	6.08941	2	9	97	positive regulation of mononuclear cell proliferation
GO:0030098	0.00004	3.89427	4	14	230	lymphocyte differentiation
GO:0006470	0.00005	4.76344	3	11	149	protein dephosphorylation
GO:0019220	0.00005	2.33068	14	31	849	regulation of phosphate metabolic process
GO:0045576	0.00005	10.4387	1	6	40	mast cell activation
GO:0080090	0.00006	1.65177	65	94	3828	regulation of primary metabolic process
GO:0045860	0.00007	3.34655	5	16	304	positive regulation of protein kinase activity
GO:0001782	0.00007	14.04984	0	5	26	B cell homeostasis
GO:0046700	0.00007	2.39772	13	28	743	heterocycle catabolic process
GO:0045321	0.00007	4.53924	3	11	166	leukocyte activation
GO:0030097	0.00007	2.69974	9	22	517	hemopoiesis
GO:0001932	0.00008	2.42023	12	27	709	regulation of protein phosphorylation
GO:0023051	0.00009	1.87098	30	52	1798	regulation of signaling
GO:0012501	0.00009	2.00401	23	42	1345	programmed cell death
GO:0050867	0.00009	4.07922	3	12	188	positive regulation of cell activation
GO:1901292	0.00009	2.39834	12	27	715	nucleoside phosphate catabolic process
GO:0002700	0.00009	7.40859	1	7	63	regulation of production of molecular mediator of immune response
GO:0002699	0.00009	5.40966	2	9	108	positive regulation of immune effector process
GO:0032844	0.0001	3.56046	4	14	250	regulation of homeostatic process
GO:0042454	0.0001	2.87925	7	19	418	ribonucleoside catabolic process

Supplementary table 3-11. Representation of the gene ontology analysis that was done on the 9 clusters as identified in the gene expression analysis of the resting and activated B and T cells (see supplemental figure 10). Supplementary table 3-11 depict the list of biological processes that were found in the GO analysis for gene clusters 1-9, respectively. The lists of genes are sorted based on the P-values, from small to large.

Supplemental table 10
Downregulated genes in T cells (Day 0)

GOBPID	Pvalue	OddsRatio	ExpCount	Count	Size	Term
GO:0006281	0,00003	6,35353	2	9	321	DNA repair
GO:0008152	0,00003	2,39456	45	65	9211	metabolic process
GO:0044257	0,00004	6,15297	2	9	331	cellular protein catabolic process
GO:0006511	0,00008	6,34617	1	8	283	ubiquitin-dependent protein catabolic process
GO:0043632	0,0001	6,14214	1	8	292	modification-dependent macromolecule catabolic process

Supplementary table 3-11. Representation of the gene ontology analysis that was done on the 9 clusters as identified in the gene expression analysis of the resting and activated B and T cells (see supplemental figure 10). Supplementary table 3-11 depict the list of biological processes that were found in the GO analysis for gene clusters 1-9, respectively. The lists of genes are sorted based on the P-values, from small to large.

Supplemental Table 11.
Upregulated genes in cycling B cells (day 4)

GOBPID	Pvalue	OddsRatio	ExpCount	Count	Size	Term
GO:0002376	0.00001	2.9463	10	26	1264	immune system process

Supplementary table 3-11. Representation of the gene ontology analysis that was done on the 9 clusters as identified in the gene expression analysis of the resting and activated B and T cells (see supplemental figure 10). Supplementary table 3-11 depict the list of biological processes that were found in the GO analysis for gene clusters 1-9, respectively. The lists of genes are sorted based on the P-values, from small to large.

Supplemental table 12. 4C sequencing primers

Supplemental table 12a. 4C single-end sequencing primers

4C viewpoint	Primer type	Primer sequence
Upstream IgH	Primer 1	AATGATACGGCGACCACCGAACACTCTTTCCCTACACGACGCTCTTCCGATCTTCAAGGCTAATGAAGCTT
Upstream IgH	Primer 2	CAAGCAGAAGACGGCATACGATCAGCGGGGATGTAGAGC

Supplemental table 12b. 4C paired-end sequencing primers

4C viewpoint	Primer type	Primer sequence
3'RR_4C capture	PE 1	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTCCGAGGCCTTACAAGCTT
3'RR_SNP	PE 2	CAAGCAGAAGACGGCATACGAGATCGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATCTCTAACTTTGGCCCAGAAATC
IgG3_4C capture	PE 1	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTAAGCCTTTCTAAGGCAGATC
IgG3_SNP	PE 2	CAAGCAGAAGACGGCATACGAGATCGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATCTGGGACATTAGAGACATAGCTG
Distal V_4C capture	PE 1	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTGGTGATTTGCATTTCATAGATC
Distal V_SNP	PE 2	CAAGCAGAAGACGGCATACGAGATCGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATCTCGGGCCTCATATCAGCTA

Supplementary table 12. Sequence of the 4C primers that were used in the single end 4Cseq experiment (S12a) and the paired-end 4Cseq experiment (S12b). The Illumina adaptors are depicted in gray.

Supplemental table 13. Chromosomal locations of the SNPs (mm9)

4C viewpoint name	Chr	Position	C57Bl/6 sequence	FVB sequence
3RR	12	114495014	T	C
Cγ3	12	114594500	C	T
Distal V	12	119923316	A	G

Supplementary table 13. Genomic positions of the SNPs that are used to assign a 4C capture to the C57Bl/6 or the FVB allele in the paired-end 4Cseq strategy.

4
Robust 4C-seq data analysis to screen for regulatory DNA interactions

Nature Methods. 2012 Oct;9(10):969-72

Robust 4C-seq data analysis to screen for regulatory DNA interactions

Harmen J.G. van de Werken[1,4], Gilad Landan[2,4], Sjoerd J.B. Holwerda[1], Michael Hoichman[2], Petra Klous[1], Ran Chachik[2], Erik Splinter [1], Christian Valdes Quezada[3], Yuva Öz[1], Britta A.M. Bouwman[1], Marjon J.A.M. Verstegen[1], Elzo de Wit[1], Amos Tanay[2,5] & Wouter de Laat[1,5].

[1] Hubrecht Institute-KNAW & University Medical Center Utrecht, Uppsalalaan 8, 3584 CT Utrecht, The Netherlands.
[2] Department of Computer Science and Applied Mathematics and Department of Biological Regulation, Weizmann Institute, Rehovot 76100, Israel.
[3] Instituto de Fisiología Celular, Departamento de Genética Molecular, Universidad Nacional Autónoma de México, Apartado Postal 70-242, México, D.F. 04510, Mexico
[4] These authors contributed equally.
[5] Corresponding authors:
AT: Email: amos.tanay @weizmann.ac.il.
WdL: Email: w.delaat @hubrecht.eu.

ABSTRACT

Regulatory DNA elements can control expression of distant genes via physical interactions. Here, we present a cost-effective methodology and computational analysis pipeline for robust characterization of the physical organization around selected promoters and other functional elements using Chromosome Conformation Capture combined with high-throughput sequencing (4C-seq) data. Our approach can be multiplexed and routinely integrated with other functional genomics assays to facilitate physical characterization of gene regulation.

Results

Recent systematic efforts to map chromatin features along chromosomes(1) have identified hundreds of thousands of putative regulatory sites in the human and mouse genomes. With an estimated 25,000 genes per respective genome, this suggests that multiple sites regulate each gene, and that the great majority of regulatory interactions occur over long chromosomal distances. Understanding gene regulation in complex eukaryotic genomes is therefore dependent on detailed mapping of physical contacts between genomic elements such as promoters and enhancers. Such mapping has been revolutionized through the advent of Chromosome Conformation Capture (3C) technology(2) and 3C-based methods(3). However, existing strategies either provide low genome-wide resolution (4-6), or are designed for biased mapping of specific regulatory interactions(7). A cost effective and high-resolution methodology to identify and quantify interactions between selected genomic sites and unknown regulatory sequences is still not available, preventing incorporation of physical considerations into most studies of gene regulation. Here we present a modified version of 4C technology(8,9) to provide a solution to this problem.
Our modified high resolution 4C-seq protocol **(Fig. 1a)** involves two rounds of DNA digestion with four basepair specificity restriction enzymes. After crosslinking to capture genomic interactions, primary enzyme digestion with a four basepair rather than six basepair cutter (10) increases the pool of fragment ends that can be analyzed by over ten fold. This greatly enhances the statistical power of the 4C analysis, enabling

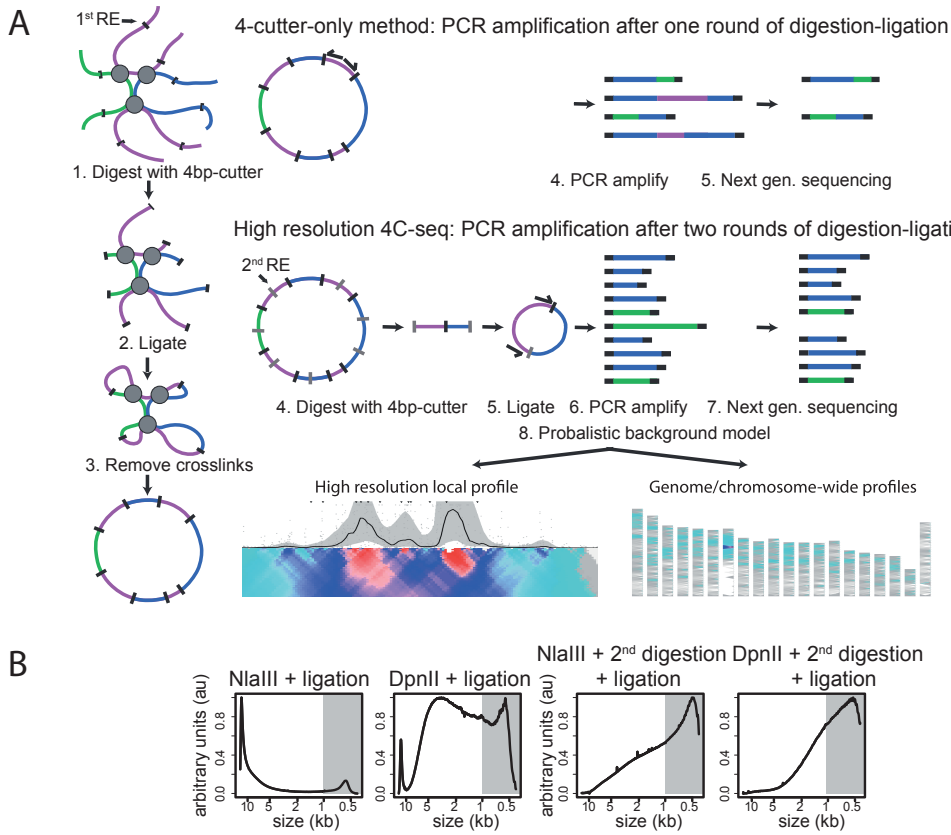


Figure 1. 4C-seq. (A). A diagram showing the importance of including two rounds of digestion and ligation in high resolution 4C-seq. (B). Graphical representation of DNA size distribution after first (left two panels) and second (right two panels) round of digestion and ligation. The shaded area indicates DNA fragment sizes that are PCR amplifiable and can be sequenced (see also Supplementary Fig. 1 and 2).

robust identification of specific interactions based on many ligation events rather than one or two ligation junctions. Subsequent ligation of crosslinked restriction fragments typically results in long (> 2 kb, **Fig. 1b, Supplementary Fig. 1)** DNA concatamers containing multiple (often > 10) restriction fragments, which presumably were all components of a single crosslinked chromatin aggregate. The size of these concatamers precludes efficient amplification and sequencing. We therefore include a second round of digestion using a different four basepair cutter (in contrast to other published 4C strategies(11)) using a different 4-bp cutter. This greatly increases the complexity of the amplified library and the robustness of contact profiles, and is crucial for the reproducibility of the method. After ligation-induced circularization of the short fragments we performed inverse PCR using primers designed to target a primary restriction fragment of interest (the 'viewpoint fragment') and to amplify its ligated partners. By using primers that include dangling Illumina adapter sequences(10), the inverse PCR products are immediately ready for sequencing, with the sequence read consisting of the forward inverse PCR primer, the restriction site and the near end of the ligated fragment. Inverse PCR is applied to approximately 500,000 cells per experiment, theoretically interrogating one million ligation products per viewpoint. A typical experiment therefore requires high throughput sequencing of no more than 1-2 million reads.

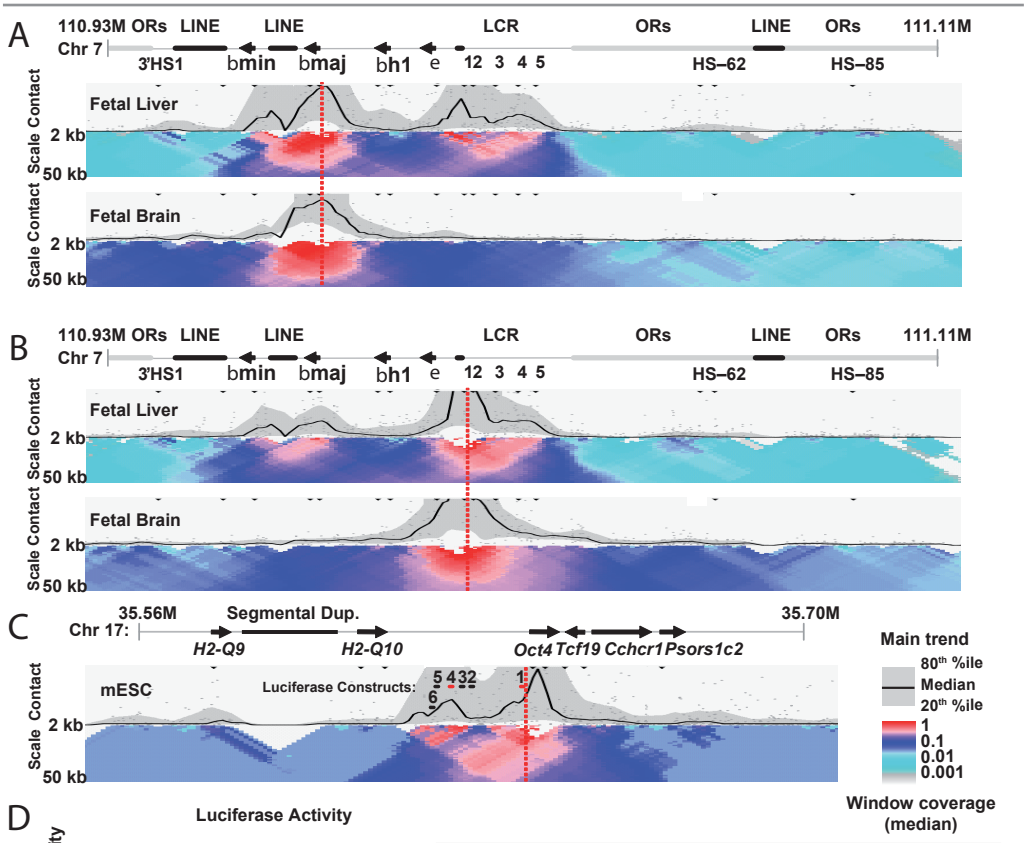


Figure 2. Promoter-LCR interactions in the active β-globin cluster. (A). Contact profiles of the active (top; Fetal Liver) and inactive (bottom; Fetal Brain) βmaj gene, using a 5 kb window size in main trend (B). Contact profiles for the HS2 element within the LCR in the active (top; Fetal Liver) and inactive state (bottom; Fetal Brain), using a 5 kb window size in main trend (C) Predicting novel promoter-enhancers contact for the Oct4 promoter in ES cells, using a 7 kb window size in main trend panel Sequences capable of driving luciferase activity are annotated with red lines, and cloned sequences that are incapable are shown in black. (D) Luciferase reporter assays demonstrating enhancer activity in ES cells of the upstream site found to interact with Oct4 (error bar indicates standard error of the mean; n = 9 (independent experiments, each analyzed by triplicate PCR) for Oct4 -17 kb, for all other regions n = 2 and individual data points are indicated with O). Asterisk indicates p-value ≤ 0.05 (one-tailed Student's t-test on log-transformed data)

We developed an analysis pipeline for mapping and normalization of 4C-seq data **(Supplementary Fig. 2)** that uses two complementary strategies. We take advantage of the rich fragment pool (about 6–8 fragment ends per 1 kb) to generate statistically robust semi-quantitative contact maps in the 10 kb – 1 Mb region surrounding the viewpoint, spanning intensities within 3–4 orders of magnitude. To all regions more remote from the viewpoint we apply a statistical enrichment approach **(Supplementary Fig. 2d)**, using an estimated probabilistic background model to compute expected total 4C coverage for fragment ends in genomic windows, and comparing the observed number of sequenced ligation products to the expected coverage. In both strategies, we analyze contacts at multiple scales(12) using the high resolution restriction site grid to quantify contact intensities in genomic windows varying in size from as little as few kilobases to as much as several megabases (Online Methods).

To validate the approach, we first applied high resolution 4C-seq to the well characterized β-globin locus **(Fig. 2).** Chromosomal contacts within the β-globin cluster were previously mapped using 3C (13) and 5C technologies(14) and contacts with regions elsewhere in the genome were mapped with 4C-on-chip(8). The

Hbb-b1 gene, or bmaj, is highly active in fetal liver cells, and its tissue specific upregulation depends on a locus control region (LCR) composed of five hypersensitive sites (HS1–5) located ~30 kb upstream of the gene. We generated contact profiles using viewpoints next to the bmaj gene and the LCR's HS2 element in fetal liver and fetal brain control. We obtained contact intensities for ~1000 fragment ends in the 150 kb β-globin domain, compared to a few dozen in previously described 3C, 5C or low resolution 4C-chip datasets. This level of detail allows for the unbiased identification of the LCR **(Fig. 2a)**. The data suggest the gene-proximal side (HS1–2) as the most prominently interacting part of the LCR **(Supplementary Fig. 3)**. A domain of preferred contacts exists that extends from the beginning of the LCR to the most distal active β-globin gene (bmin). A reciprocal profile generated using a viewpoint positioned at HS2 shows preferred contacts with the active bmaj and bmin genes and demarcates the same domain in fetal liver. The defined domain and preferred contacts therein are absent in fetal brain, where the locus is inactive **(Fig. 2b)**. CTCF sites flanking the β-globin locus were previously shown to interact with each other(15). Our experiments confirm these interactions, **(Supplementary Fig. 4)** identify a new interacting CTCF site (3'HS2) **(Supplementary Fig. 5)** and reveal that the single topological entity identified by 3C technology actually separates into two hierarchical structures: a chromatin loop that brings together CTCF sites flanking the locus and the regulatory interactions between the LCR and β-globin genes. We also applied our method to the a-globin locus(16-18). The data confirm known long-range interactions and show that the a-globin locus adopts a defined domain topology only when active **(Supplementary Fig. 6 and 7)**.

4C-seq can also be used to study inter-chromosomal and remote intrachromosomal interactions with high accuracy. Multi-scale analysis quantifies the intensity of cis-interactions between a viewpoint and chromosomal domains that vary in size between 10 kb and 5 Mb **(Supplementary Fig. 8)**. We used a statistical model to estimate the probability of observing contacts between each fragment end and the viewpoint, and computed the ratios between the expected and observed number of contacts in genomic windows of variable sizes. Comparison of the full chromosomal contact maps for different β-globin viewpoints that are within 50 kb of each other reveals highly consistent profiles in fetal liver. Remote (inter- or intra-) chromosomal contacts therefore appear to reflect the preferred neighborhood of larger chromosomal domains rather than specific interactions between regulatory sites. However, contact profiles are remarkably different in fetal brain cells. We observed similar cell-type dependent configurations for other tissue-specific genes **(Supplementary Fig. 8)**, as well as for interchromosomal contacts formed by β-globin viewpoints **(Supplementary Fig. 9)**. We note that, despite superior library complexity and resolution, the contact specificities for interchromosomal interactions are still modest, with the maximal enrichment detected as 6-fold over the background (compared to 1000-fold or more for local looping interactions). Still, the fact that genome-wide contact profiles are highly reproducible between different viewpoints in the same locus shows that the technique and its associated analysis also accurately identify contacts with the remainder of the genome.

We next examined two uncharacterized genes that are active in distinct tissues. Oct4 is a key pluripotency gene expressed in embryonic stem cells (ESCs). A 4C-seq profile from a viewpoint positioned near the TSS of Oct4 **(Fig. 2c)** reveals specific novel contacts with a domain located approximately 17 kb upstream of the TSS, at a larger distance from the promoter than the known distal enhancer (–2 kb) and proximal enhancer (–1 kb)(19). A luciferase reporter assay demonstrated that this segment drives increased reporter gene expression specifically in ESCs **(Fig. 2d)**. Four other surrounding sequences without obvious contacts with the Oct4 promoter did not show this activity in ESCs, despite enrichment for H3K4me1 at some of these segments **(Supplementary Fig. 10)**.

Satb1 is a gene that is highly active in thymocytes. Using 4C-seq, we find that the Satb1 TSS contacts a preferred gene-poor chromosomal domain in thymocytes that spans over 600 kb upstream and includes multiple contact hotspots **(Supplementary Fig. 11a)** (see also raw data in **Supplementary Fig. 12a–c**). The domain was identified reciprocally when assaying contacts from a remote putative contacting element 470 kb away from the gene. In fetal brain, where the Satb1 is ~20-fold less active (**Supplementary Fig. 13**), this

interaction domain is completely absent, with only weak residual contact noticeable with an upstream element 650 kb upstream (**Supplementary Fig. 12d**). The same is true in ESCs (**Supplementary Fig. 11a**) that express Satb1 at even lower levels (~500 fold difference compared to thymocytes). These results suggest that the large Satb1 proximal domain acts as a regulatory scaffold that facilitates the high expression of Satb1 in T cells. To examine the regulatory activity of the contact hotspots, we tested a series of distal sites in a luciferase reporter assay. Contacting elements at a distance of 253 kb, 470 kb and 649 kb from the TSS were found to significantly boost reporter gene expression in lymphoid cells (**Supplementary Fig. 11b**).
It is important to determine how an effective promoter viewpoint should be positioned relative to the TSS when screening for promoter-enhancer interactions. We therefore studied contact profiles derived from viewpoints positioned within a few kilobases around the Satb1 and Oct4 TSS. The data show that chromosomal contacts around TSSs can be surprisingly position-specific (**Supplementary Fig. 14 and 15**). For enhancer screening purposes, positioning the viewpoint immediately upstream of the TSS seems to be preferred. Controls using viewpoints further upstream and downstream of the TSS can be used to provide more data on the regional architecture of the TSS-enhancer domain.
Physical or topological domains around active genes were recently demonstrated globally in Hi-C experiments(4-6) suggesting that effective mapping of functionally relevant contacts within these domains requires high resolution strategies as presented here. Compared to (semi-)quantitative 3C, which is currently the method of choice to assay contacts between nearby regulatory sequences, the integrated 4C-seq strategy and pipeline is easier (one primer pair, no control template needed), much more robust (in the β-globin locus we analyze nearly 1000 independent ligation events, compared to 15–20 junctions analyzed in a typical 3C experiment) and unbiased to pre-chosen genomic partners. The analysis pipeline, and a genome-wide
4C primer database, can be downloaded from (http://compgenomics.weizmann.ac.il/tanay/?page_id=367).

ACKNOWLEDGEMENTS

We would like to thank R. Palstra, E. Yaffe and other members of our labs for help and input, G. Geeven for testing the 4C-seq pipeline and H. T. Timmers (Molecular Cancer Research, Netherlands Proteomics Centre, University Medical Center Utrecht, The Netherlands) and J. P. P. Meijerink (Department of Pediatric Oncology and Hematology, Erasmus Medical Center-Sophia Children's Hospital, Rotterdam, The Netherlands) for providing cells. This work was financially supported by grants from the MODHEP FP7 CP to WdL and AT, by grants from the Dutch Scientific Organization (NWO) (91204082 and 935170621), a European Research Council Starting Grant (209700, '4C') and by InteGeR FP7 Marie Curie Initial Training Networks (ITN) (contract number PITN-GA-2007-214902) to WdL and by the Israeli Science Foundation integrated technologies grant to AT.

AUTHOR CONTRIBUTIONS

H.W. designed, performed and analyzed experiments and wrote the manuscript; E.S. helped design experiments, G.L. helped design and analyze experiments and wrote the manuscript; P.K.;S.H; B.B., M.V, Y.O & C.Q. performed experiments; M.H., R.C, E.W. helped analyzing the experiments, A.T and W.d.L. designed experiments, supervised the project and wrote the manuscript.

COMPETING INTERESTS STATEMENT

The authors declare no competing financial interests.

OTHER STATEMENTS

All experiments were conducted under the approval of the animal care committee of the KNAW (Netherlands Royal Academy of Arts and Sciences).

Accesssion codes: 4C-seq data is available on GEO: GSE40420

Methods

General considerations with respect to modeling and interpreting 4C-seq experiments.

The raw experimental readout from a single 4C-seq experiment consists of 0.5 to 3 million reads containing the ligation junctions between the viewpoint fragment end (where the forward PCR primer is positioned) and any other primary restriction fragment end. To map the latter, we developed a specialized 4C-seq genome mapping algorithm that controls for sequencing errors and non-unique sequences while considering the high coverage (10^2–10^4x) of fragment ends that are proximal to the viewpoint fragment (Experimental Procedures). The number of reads mapped by the algorithm to each fragment end defines the experiment's *coverage profile* and represents the intensity of contacts between each restriction fragment and the viewpoint fragment, combined with multiple stochastic and systematic noise factors. The challenge in 4C-seq analysis is to create a robust scheme to normalize these noise factors while enabling maximal resolution of the derived contact profiles.

Interchromosomal and remote intrachromosomal contacts are observed with small probability throughout the genome in an overall spatially uncorrelated fashion (**Supplementary Fig. 2a**). In such remote regions, the observation of multiple reads for the same fragment end is not more informative than a single read (**Supplementary Fig. 2a–b**) and can be assumed to represent stochastic amplification events. In support of this, we find that long-range and interchromosomal contact profiles from replicate experiments are reproducible at the level of larger genomic regions (> 100 kb), but not at the level of the single fragment. This is similar to the behavior of Hi-C matrices that show a typical background contact probability of 0.001% between any two fragment ends(5,20). Conclusions on remote contacts in such settings are therefore obtained by pooling together statistics from hundreds of fragment ends (for 4C) or dozens of thousands of fragment pairs (in Hi-C), such that a sufficient number of reads can be aggregated. In contrast, local contacts between the 4C viewpoint and fragments in its chromosomal vicinity (i.e. < 1 Mb distance), are recovered with high probability, and the number of reads mapping to individual fragment ends in this region is highly informative (**Supplementary Fig. 2c**). The effective resolution of a 4C experiment therefore relies on the proper quantitative normalization of the coverage profile near the viewpoint.

The computational normalization and visualization of 4C-seq contact profiles is therefore achieved using two complementary strategies. A quantitative approach to contact intensity is applied in the region surrounding the viewpoint, generating a normalized contact profile that represents intensities within 3–4 orders of magnitude. A statistical enrichment approach (Methods, **Supplementary Fig. 2d**) is applied to all other regions, using an estimated probabilistic background model to compute expected total 4C coverage for fragment ends in selected genomic windows, and comparing the expected coverage to the observed number of sequenced ligation products. In both strategies, analysis is done using an approach that simultaneously captures interactions at multiple scales, reminiscent of the previously published domainograms(12). The high resolution restriction site grid provides the opportunity to examine contact intensities in genomic windows varying in size from as little as few kbs to as much as several megabases. For quantification of contact strength, we analyze and visualize medians of normalized coverage (in the viewpoint's vicinity) or enrichment of observed vs. expected number of reads (for the rest of the genome). Such quantities describe contact intensity rather than statistical significance (i.e. *p*-value) of non-background behavior. This approach prevents the bias toward larger genomic windows (more data points) that is commonly introduced when using p-values to visualize contact intensity. Instead, p-values are used only when testing the robustness of the intensity statistics. This unbiased approach ensures that statistics at different genomic scales are directly comparable and interpretable.

Preparation of 4C-template.

4C templates were prepared essentially as described previously(21). In brief, primary tissue was isolated, single cells suspensions were made, chromatin was cross-linked with 2% formaldehyde for 10 minutes at room temperature, nuclei were isolated and crosslinked DNA digested with a primary restriction enzyme recognizing a four basepair restriction site. This was followed by proximity ligation after which crosslinks were removed. A secondary restriction enzyme digestion was performed with a four basepair restriction enzyme recognizing a different sequence than the primary enzyme, followed again by proximity ligation. Typically, 200 ng of the resulting 4C template was used for the subsequent PCR reaction, of which 16 (total: 3.2 mg 4C template) were pooled and purified for next generation sequencing. The PCR-products were purified using two columns per sample of the High Pure PCR Product Purification Kit (Roche #11732676001). The kit separates the PCR products that are larger than 120 bp from the adapter-containing primers (~75 bp; ~40 bp). Similar results were obtained with products from a single PCR reaction (200 ng template).

4C-seq primer design.

4C primer pairs carry additional 5' overhangs composed of the adapter sequences (obtained from Illumina technical support) necessary for Illumina single read sequencing (GA-II and Hi-seq 2000). The strategy therefore produces sequencing reads (36-mers) composed of the 4C primer sequence (20 nucleotides, specific to a given viewpoint), followed by 16 nucleotides that identify a captured sequence. The reading primer always hybridizes to, and ends at the 3' side of, the entire first restriction recognition site. This design ensures analysis of primary ligation events only and provides sufficient sequence information to unambiguously identify most captured sequences (mapability of 16-mers directly adjacent to a given four basepair site is 68%, using *Nla*III and *Dpn*II as restriction enzyme combination). The non-reading primers, with a size between 18 to 27 bp, were designed at a distance of ≤120 bp from the secondary restriction site. PCR primers were designed taking into account the following additional rules. The viewpoint fragment preferably had a size of at least 500 bp to allow efficient cross-linking to other DNA fragments. The fragment end (the nucleotide sequences of the viewpoint fragment between the primary and secondary restriction site to which both 4C primers hybridize) was at least 300 bp and preferably more than 350 bp, to allow efficient circularization during the second ligation step. Primer3(22) was used to find the optimal primer pair for a given viewpoint fragment, with the following adaptations to the default settings: optimal temperature of 55 °C, minimum of 45 °C and maximum of 65 °C, GC content between 35 and 65 %. Primers were checked against the mouse genome with megablast(23) (settings -p 88.88 -W 12 -e 1 -F T) requiring primers on the reading side to be matched uniquely in the genome, and primers in the non-reading side to have a maximum of three perfectly matching Blast high-scoring segment pairs (HSP). Both primers were also required to have fewer than 30 HSPs with an identity of at least 88.88%(16/18 bp). **Supplementary Table 1** shows all the primers used in this study. Moreover, a genome-wide 4C primer database was constructed that implements all rules and can be downloaded from our website (http://compgenomics.weizmann.ac.il/tanay/?page_id=367).

Mapping and filtering sequence reads.

Supplementary Table 2 describes all experiments included in this work and provides statistics on their read counts. At least 10–20 different 4C experiments (different viewpoints, or the same viewpoint but barcoded)

were mixed and sequenced simultaneously in one Illumina GA-II or HiSeq 2000 lane. Sequence tags generated by the 4C-seq procedure are prefixed by the 4C reading primer that includes the restriction site sequences. We therefore separate multiplexed 4C-seq libraries according to the prefix and extract their suffixes for further processing. The algorithm for mapping of suffixes to the genome was designed given the following main considerations:
- Valid suffixes should begin at a primary restriction site and continue with its downstream sequence (i.e. should be mappable to one of the experiment's fragment ends).
- The expected coverage profile is highly non-uniform, and fragment ends that are proximal to the viewpoint are likely to be covered dozens to thousands of times. The number of reads mapped to each fragment end represents significant information in the viewpoint region (< 1 Mb), but very little information out of this region (**Supplementary Fig. 2a–b**).
- Even though the sequencing error rate is low, ligation juntions occur thousands of times may give rise to dozens of copies with particular mismatches. When such mismatches create variants that are mappable to the genome (by chance), significant coverage of remote fragment ends may be incorrectly inferred.
A special-purpose mapping algorithm based on these considerations is described next. We assume sequence tags of length L_{tag} (read length minus the length of the primer and restriction site) are given, and that base calling probabilities are provided for each tag. We define the precision of each sequence tag as the product of estimated base calling probabilities. The probability of any specific mismatch is computed by multiplying the base error probabilities at the variable positions. The algorithm proceeds along the following steps:

1) Constructing a fragment end index: all restriction sites in the genome are identified and the L_{tag} bp sequences both upstream and downstream are indexed using a hash table.
2) Computing interim coverage for well-separated fragment ends: a fragment end is classified as well separated if all other fragment-ends of size L_{tag} in the genome differ from it in at least two positions. We regard sequence tags with precision > 0.9 that map perfectly to well-separated fragment ends as unambiguous and compute an interim coverage profile (denoted $interim_j$) for each such fragment end j, by looking it up in the hash table and summing the total precision of tags mapped to it.
We note that the precision threshold is adjustable and may need to be changed for longer reads with low base calling probabilities towards the end of the read.
3) Computing fragment-end mapping weights: we define the mapping weight of each well-separated fragment end j as its interim coverage computed in the previous step. For all other fragment ends, the mapping weight equals the distance-weighted geometric mean of interim coverage on well separated fragment ends in a window of size W around the fragment end i:

$$wg(i) = \exp\left(1/Z(W,i) \sum_{\substack{d(i,j)<W \\ j\ well\ sep.}} \left(1 - \frac{d(i,j)}{W}\right) \log\left(interim_j\right)\right)$$

Where:

$$Z(W,i) = \sum_{\substack{d(i,j)<W \\ j\ well\ sep.}} \left(1 - \frac{d(i,j)}{W}\right)$$

and d(i,j) is the genomic distance between fragment ends i and j, $interim_j$ is the coverage of well-separated fragment end j, and W is 200 kb or the minimal size for which $Z(W,i) \geq 6$ if $Z(200\ kb, j) < 6$. This ensures that non-well-separated fragment ends that are extremely distant from well-separated fragment ends utilize

a larger window and are weighted using a robust geometric mean.
4) Computing the coverage profile. Given the mapping weights defined above, each read is distributed among its potential originating fragment ends according to the read's sequence and base calling probabilities, and the mapping weights of fragment ends with the same or similar sequence. Specifically, given a sequence tag s, the mapping vote for a fragment end i with sequence fes(i) is computed as:

$$vote = \frac{1}{Z}\Pr(fes(i)|s) * wgt(i)$$

where Pr(fes(i)|s) is the probability that the read s originated from the fragment end with sequence fes(i), calculated using the read's base calling errors. The normalization factor Z is computed by summing over all fragment ends with sequences that are not too far from the read sequence:

$$Z = \sum_{\Pr(fes(i)|s)>s} Pr(fes(i)|s) * wgt(i)$$

Here we used an epsilon value of 0.0001, but this can be modified according to read length and overall sequence quality as it significantly affects the algorithm's run time performance (by determining the number of fragment ends the algorithm must examine per read). Moreover, fragment ends with sequences that appear more than five times in the genome, and their associated reads, are filtered out. The final coverage profile is defined by distributing each read among fragment ends according to the mapping votes. In general, while the mapping algorithm takes full account of non-unique fragment ends (which, as described above, are resolved in a way that can also affect the expected coverage of unique fragment ends), the analysis included in the present work is based on coverage statistics for unique fragment ends only.
Our C++ implementation of the above mapping algorithm can complete mapping of a two million read experiment within approximately 10 minutes on a single core of a linux machine using up to six GB of RAM. We note that longer reads can further reduce mapping ambiguity and allow application of standard mapping algorithms for 4C-seq.

Construction of a background model for remote intra- and inter- chromosomal contacts.

Several steps in the 4C-seq experimental protocol are prone to systematic biases that may influence the distribution of coverage inferred by the mapping algorithm described above. Of these, factors affecting ligation and amplification efficiency include the restriction fragment length, its G/C content, and the size of the 4C amplification product as determined by the linear genomic distance between the primary restriction site and the nearest secondary restriction site. Broad enrichment patterns (such as those presented in **Supplementary Fig. 8 and 9**) represent relatively mild contact preferences (2–3 fold enrichment) of large genomic windows. Since some of the potential sources of bias in coverage are distributed non-uniformly in the genome and may differ significantly between large genomic windows, it is particularly important to normalize raw coverage before studying global contact trends.
We define the fragment length associated with each fragment end as the distance between the two primary restriction sites forming the fragment. This length is binned into the ranges (0–50, 50–100, 100–200, 200–300, 300–400, 400–500, > 500) and is denoted by fl(i). We define the fragment end length as the distance between the primary restriction site and the nearest secondary restriction site, binned into the ranges (0–100, 100–500, > 500) and denoted by fe(i). Furthermore, we define the variable bl(i) as indicating whether a fragment end belongs to a fragment that lacks a secondary restriction site. Such fragment ends are denoted as blind and are expected to generate longer 4C products than non-blind fragment ends (see Supplementary **Fig 2e**). More specifically, the 4C-seq products that map to blind fragment ends are dependent on the location of a secondary restriction site in another ligated fragment (which is part of the longer concatamer

generated by the initial 3C procedure). Given these parameters, and empirical coverage profile cov(j) (defined as 1 if fragment end j is covered by a 4C product and 0 otherwise), we estimate the background probability of coverage for a fragment end i as:

$$expcov(i) = \frac{\#\{j \in fends\ s.t.\ cov(j) = 1, fe(j) = fe(i), fl(j) = fl(i), bl(j) = fl(i)\}}{\#\{j \in fends\ s.t.\ fe(j) = fe(i), fl(j) = fl(i), bl(j) = bl(i)\}}$$

Some considerations involving model estimation should be noted:
- The model is estimated from fragment ends that are unique in the genome (see description of mapping strategy above). Normalization and computation of statistics in windows can be done using the non-unique fragment ends as well.
- Only the binary (yes/no) coverage profile cov(j) is used by the model. Multiple-read coverage was shown empirically not to be more informative than single-read coverage for remote interactions (**Supplementary Fig. 2a**).
- Separate background models are calculated for intra- and inter-chromosomal contacts as the chromosomal territory effect results in remote intra-chromosomal interactions at least 3–5 fold more enriched than inter-chromosomal interactions, even at a distance of > 100 Mb from the viewpoint. A region of 10 Mb around the viewpoint is discarded when estimating the model for intra-chromosomal contacts since coverage in this region cannot be assumed a-priori to be uniform.

Computing contact enrichment values for multi-scale windows.

Analysis and visualization of long range contacts is done by computing enrichment over genomic windows:

$$lr(o, e) = log2\left(\frac{o + \max(0, prior - e)}{\max(e, prior)}\right)$$

where the summation is performed over all fragment ends in the genomic range [x,y] and the lr function is defined as a regularized log ratio

$$oe(x, y) = lr\left(\sum_{i \in [x,y]} cov(i), \sum_{i \in [x,y]} expcov(i)\right)$$

and *prior* is a regularization parameter that we empirically set to 4. This approach is appropriate for quantifying the intensity of contact enrichment across different window sizes. It does not guarantee statistical significance, but we suggest that *p*-values not be used to explore contact intensities, but rather only to confirm hypotheses on patterns of such contacts. When required, p-values for rejecting the background model can be easily generated using empirical likelihood ratio tests or normal approximation of the log likelihood distribution over genomic windows of a given size.

Normalization of quantitative contact profiles in the region proximal to the viewpoint.

For analysis of proximal contacts, the inherent extreme variability in contact intensities in the viewpoint region undermines the uniformity assumption that allows the construction of the complex and parameter-rich background model described above. Empirical analysis suggests that the single most important factor affecting read count in the viewpoint region is whether a fragment is blind (lacking a second restriction site) or non-blind (see **Supplementary Fig. 2c**). Different genomic windows have variable ratios of blind and non-blind fragments, as determined by the frequency of the secondary restriction sites within the

window. This ratio may be coupled with various genomic features, like the regional G/C content, gene density and repeat density. As a result, without proper normalization, regions that are rich in blind fragments will be biased toward lower 4C-seq coverage (**Supplementary Fig. 2e**). Such lower 4C-seq coverage may in turn be misinterpreted to be correlated with features that are associated with gene regulation. One must therefore normalize blind and non-blind 4C-seq coverage in a quantitative fashion that is robust to the variable density of fragment end types and to the biased amplification of the ligation products involving these ends.

The normalization scheme for the contact profiles in the region proximal to the viewpoint is given a set of quantitative coverage profiles and assumes a-priori that all of the profiles represent the same distribution of contact intensities. Distinct profiles are generated from the blind and non-blind fragments in the same experiment, and optionally from replicate experiments, or even from two experiments using different first restriction enzymes, assuming the viewpoints are located within a very short genomic distance. The algorithm then performs several steps to combine the profiles:

- Fragment ends are mapped back to genomic coordinates (using bins of 16 bp in our current implementation). Each fragment end is mapped to the center of the fragment, such that the 3' and 5' fragment ends are mapped to the same genomic bin. Non-unique fragment ends are masked.
- We perform linear interpolation of the 3' and 5' coverage at fragment centers to generate coverage values for the remaining genomic bins. This is done to generate a fixed number of data points for each genomic interval, preventing systematic biases that can be caused by non-uniform distributions of blind or non-blind fragments. This results in four coverage profiles per track: 5' non-blind, 3' non-blind, 5' blind, and 3' blind.
- One interpolated profile is selected and all other profiles are quantile-normalized to match its distribution. We then project the normalized interpolated profiles back onto the fragment end space.
- Resulting profiles are combined by direct summation. The maximum median for all windows of size 5 kb (or as determined by the user) is identified and all medians are scaled by it. All depicted median values thus represent enrichment relative to the maximum attainable 5 kb median value.
- Medians of normalized coverage for running windows of size 5 kb are generated (to plot the contact intensity trend). The 20^{th} and 80^{th} percentiles are also computed and depicted (percentiles can be determined by the user). For visualization purposes, median trends are weakly smoothed (using means of three consecutive data points), while the 20^{th} and 80^{th} percentile trends are more vigorously smoothed (using means of seven consecutive data points).
- Medians are also calculated for sliding windows (2 kb – 50 kb) of linearly increasing size, displayed as color-coded multi-scale diagrams, with values representing enrichment relative to the maximum attainable 12 kb median value.

It is also possible to use statistics other than the median to view contact profiles near the viewpoint. These include mean, geometric mean, and variations that allow truncation of extreme values. Non-median statistics also support the use of standard deviations in place of percentiles. The options are supported by the pipeline but were not used in the analysis reported here.

We note that although empirical analysis indicates that additional factors beyond the blind/non-blind distinction (e.g., fragment length) may be correlated with systematic coverage biases in the region proximal to the viewpoint, such biases are small compared to the dynamic range of the typical contact profile (which spans 3 to 4 orders of magnitude). Therefore these additional factors cannot be effectively used for further normalization in the region proximal to the viewpoint (which contains a limited number of fragment ends).

Luciferase assays

The Oct4-1.5kb (1955 bp), Oct4-17kb (1084 bp), Oct4-20kb (1700 bp), Oct4-21kb (1523 bp), Satb1-649kb (1406 bp), amplicons were first cloned into Clone Jet vector (fermentas) with blunt end protocol, followed

by subcloning into the TATA box containing pGL4.10(luc2) plasmid (promega), while Oct4-13kb (1400 bp), Oct4-15kb (1504 bp), Satb1-648kb (1097 bp), Satb1-470kb (1119 bp), Satb1-253kb (1108 bp) were directly cloned into TATA box containing pGL4.10(luc2) plasmid using the In-Fusion HD cloning kit (Clontech).
Mouse ES (mES) cells were grown in buffalo rat liver cell-conditioned medium combined with DMEM GlutaMAX (Gibco) containing 15% fetal bovine serum (Invitrogen), leukemia inhibitory factor (LIF), 2-β-mercaptoethanol, and non-essential amino acids (Invitrogen), as previously described(24). Each experiment was performed with 2 ng renilla luciferase plasmid, 100 ng (or amounts corrected for plasmid size) pGL4.10 plasmid, and an unrelated 'stuffer' plasmid up to 800 ng per well of a 24 well plate (Gibco). mES Cells were transfected using Lipofectamin 2000 Transfection Reagent according to the manufacturer's instructions. Since primary mouse lymphocytes are difficult to culture and transfect we used the human lymphoid Jurkat cell line, which expresses Satb1 at reasonably high levels, as a surrogate system to test enhancer activity of the selected sites around the SatB1 gene. The Jurkat cells were grown in RPMI 1640 medium (GIBCO) with 10% fetal calf serum and 1% pen/strep at 37°C and 5% CO_2. Each experiment was performed with 500 ng pGL4.10 plasmid and 10 ng renilla plasmid. Jurkat cells were transfected through electroporation. 24 hours after transfection, dual luciferase reporter assays (Promega) were carried out with a Centro XS3 Microplate Luminometer LB960 (Berthold Technologies) according to the DLR kit protocol (Promega).

Chromatin immunoprecipitation

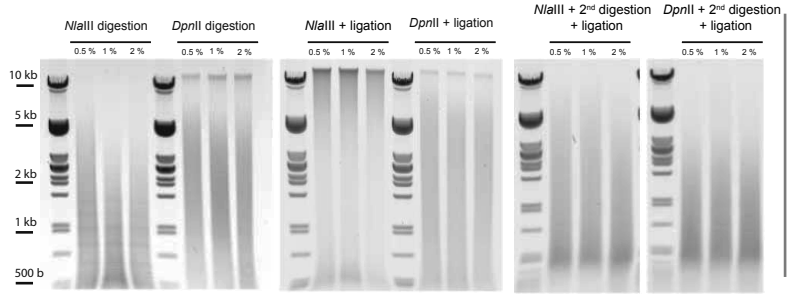
ChIP analysis on fetal liver cells was performed according to standard procedures, as previously described(25).

References

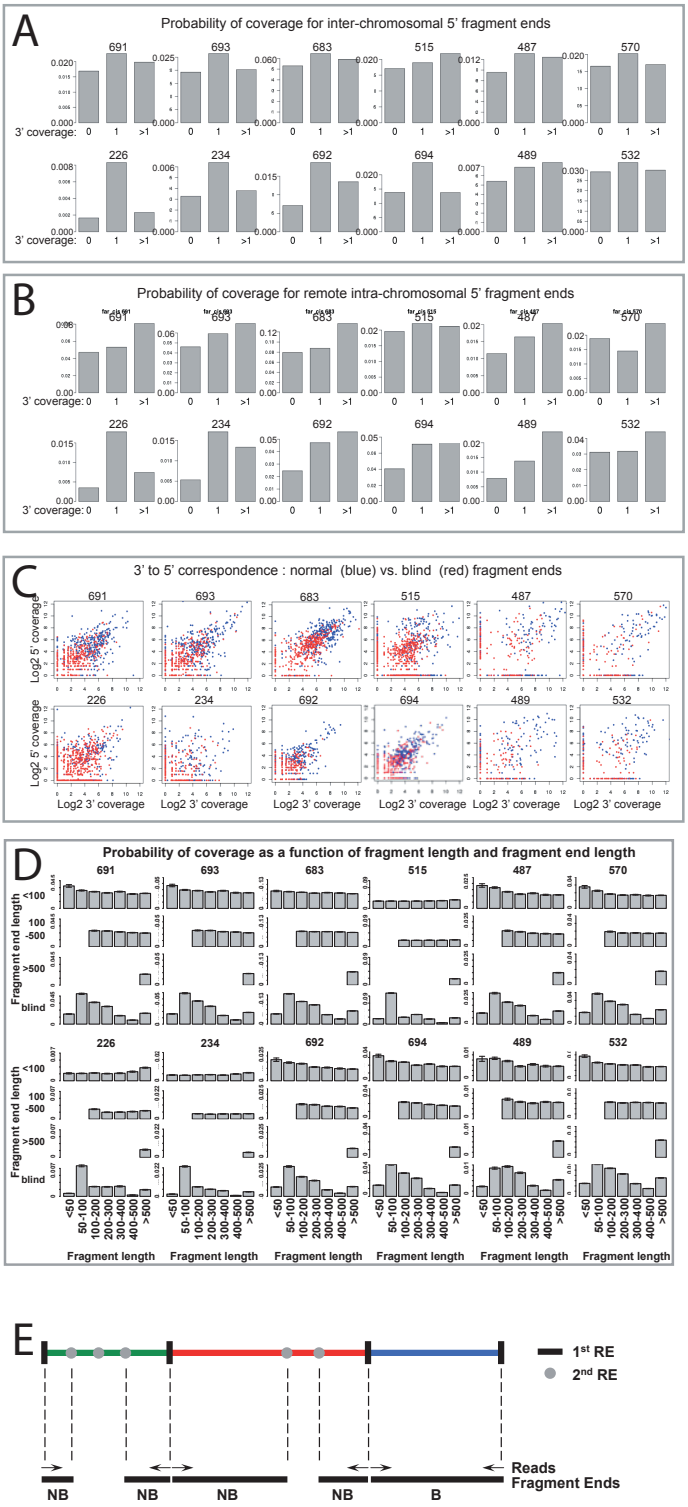
1. Shen, Y., Yue, F., McCleary, D.F., Ye, Z., Edsall, L., Kuan, S., Wagner, U., Dixon, J., Lee, L., Lobanenkov, V.V. *et al.* (2012) A map of the *cis*-regulatory sequences in the mouse genome. *Nature*, **488**, 116-120.
2. Dekker, J., Rippe, K., Dekker, M. and Kleckner, N. (2002) Capturing chromosome conformation. *Science*, **295**, 1306-1311.
3. de Wit, E. and de Laat, W. (2012) A decade of 3C technologies: insights into nuclear organization. *Genes Dev*, **26**, 11-24.
4. Lieberman-Aiden, E., van Berkum, N.L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B.R., Sabo, P.J., Dorschner, M.O. *et al.* (2009) Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, **326**, 289-293.
5. Sexton, T., Yaffe, E., Kenigsberg, E., Bantignies, F., Leblanc, B., Hoichman, M., Parrinello, H., Tanay, A. and Cavalli, G. (2012) Three-dimensional folding and functional organization principles of the Drosophila genome. *Cell*, **148**, 458-472.
6. Dixon, J.R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., Hu, M., Liu, J.S. and Ren, B. (2012) Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, **485**, 376-380.
7. Li, G., Ruan, X., Auerbach, R.K., Sandhu, K.S., Zheng, M., Wang, P., Poh, H.M., Goh, Y., Lim, J., Zhang, J. *et al.* (2012) Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation. *Cell*, **148**, 84-98.
8. Simonis, M., Klous, P., Splinter, E., Moshkin, Y., Willemsen, R., de Wit, E., van Steensel, B. and de Laat, W. (2006) Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4C). *Nat Genet*, **38**, 1348-1354.
9. Zhao, Z., Tavoosidana, G., Sjolinder, M., Gondor, A., Mariano, P., Wang, S., Kanduri, C., Lezcano, M., Sandhu, K.S., Singh, U. *et al.* (2006) Circular chromosome conformation capture (4C) uncovers extensive networks of epigenetically regulated intra- and interchromosomal interactions. *Nat Genet*, **38**, 1341-1347.
10. Splinter, E., de Wit, E., Nora, E.P., Klous, P., van de Werken, H.J., Zhu, Y., Kaaij, L.J., van Ijcken, W., Gribnau, J., Heard, E. *et al.* (2011) The inactive X chromosome adopts a unique three-dimensional conformation that is dependent on Xist RNA. *Genes Dev*, **25**, 1371-1383.

11. Lower, K.M., Hughes, J.R., De Gobbi, M., Henderson, S., Viprakasit, V., Fisher, C., Goriely, A., Ayyub, H., Sloane-Stanley, J., Vernimmen, D. *et al.* (2009) Adventitious changes in long-range gene expression caused by polymorphic structural variation and promoter competition. *Proc Natl Acad Sci U S A*, **106**, 21771-21776.
12. de Wit, E., Braunschweig, U., Greil, F., Bussemaker, H.J. and van Steensel, B. (2008) Global chromatin domain organization of the Drosophila genome. *PLoS Genet*, **4**, e1000045.
13. Tolhuis, B., Palstra, R.J., Splinter, E., Grosveld, F. and de Laat, W. (2002) Looping and interaction between hypersensitive sites in the active β-globin locus. *Mol Cell*, **10**, 1453-1465.
14. Dostie, J., Richmond, T.A., Arnaout, R.A., Selzer, R.R., Lee, W.L., Honan, T.A., Rubio, E.D., Krumm, A., Lamb, J., Nusbaum, C. *et al.* (2006) Chromosome Conformation Capture Carbon Copy (5C): a massively parallel solution for mapping interactions between genomic elements. *Genome Res*, **16**, 1299-1309.
15. Splinter, E., Heath, H., Kooren, J., Palstra, R.J., Klous, P., Grosveld, F., Galjart, N. and de Laat, W. (2006) CTCF mediates long-range chromatin looping and local histone modification in the β-globin locus. *Genes Dev*, **20**, 2349-2354.
16. Vernimmen, D., De Gobbi, M., Sloane-Stanley, J.A., Wood, W.G. and Higgs, D.R. (2007) Long-range chromosomal interactions regulate the timing of the transition between poised and active gene expression. *Embo J*, **26**, 2041-2051.
17. Zhou, G.L., Xin, L., Song, W., Di, L.J., Liu, G., Wu, X.S., Liu, D.P. and Liang, C.C. (2006) Active chromatin hub of the mouse α-globin locus forms in a transcription factory of clustered housekeeping genes. *Mol Cell Biol*, **26**, 5096-5105.
18. Baù, D., Sanyal, A., Lajoie, B.R., Capriotti, E., Byron, M., Lawrence, J.B., Dekker, J. and Marti-Renom, M.A. (2011) The three-dimensional folding of the α-globin gene domain reveals formation of chromatin globules. *Nat Struct Mol Biol*, **18**, 107-114.
19. Yeom, Y.I., Fuhrmann, G., Ovitt, C.E., Brehm, A., Ohbo, K., Gross, M., Hubner, K. and Scholer, H.R. (1996) Germline regulatory element of Oct-4 specific for the totipotent cycle of embryonal cells. *Development*, **122**, 881-894.
20. Yaffe, E. and Tanay, A. (2011) Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture. *Nat Genet*, **43**, 1059-1065.
21. Simonis, M., Kooren, J. and de Laat, W. (2007) An evaluation of 3C-based methods to capture DNA interactions. *Nat Methods*, **4**, 895-901.
22. Rozen, S. and Skaletsky, H. (2000) Primer3 on the WWW for general users and for biologist programmers. *Methods Mol Biol*, **132**, 365-386.
23. Zhang, Z., Schwartz, S., Wagner, L. and Miller, W. (2000) A greedy algorithm for aligning DNA sequences. *J Comput Biol*, 7, 203-214.
24. Di Stefano, B., Buecker, C., Ungaro, F., Prigione, A., Chen, H.H., Welling, M., Eijpe, M., Mostoslavsky, G., Tesar, P., Adjaye, J. *et al.* (2010) An ES-like pluripotent state in FGF-dependent murine iPS cells. *PLoS One*, **5**, e16092.
25. Kooren, J., Palstra, R.J., Klous, P., Splinter, E., von Lindern, M., Grosveld, F. and de Laat, W. (2007) β-globin active chromatin Hub formation in differentiating erythroid cells and in p45 NF-E2 knock-out mice. *J Biol Chem*, **282**, 16544-16552.

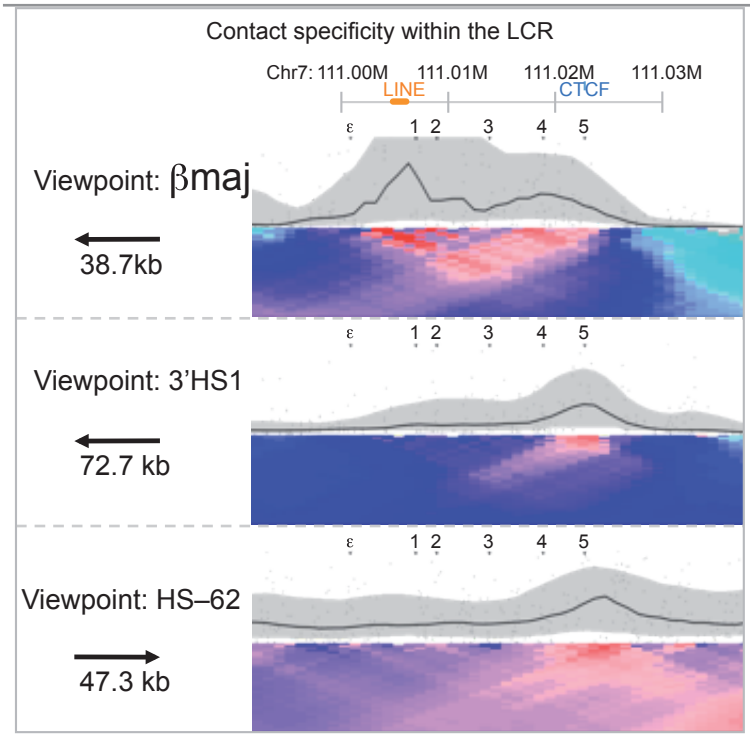
Supplemental figures and tables.



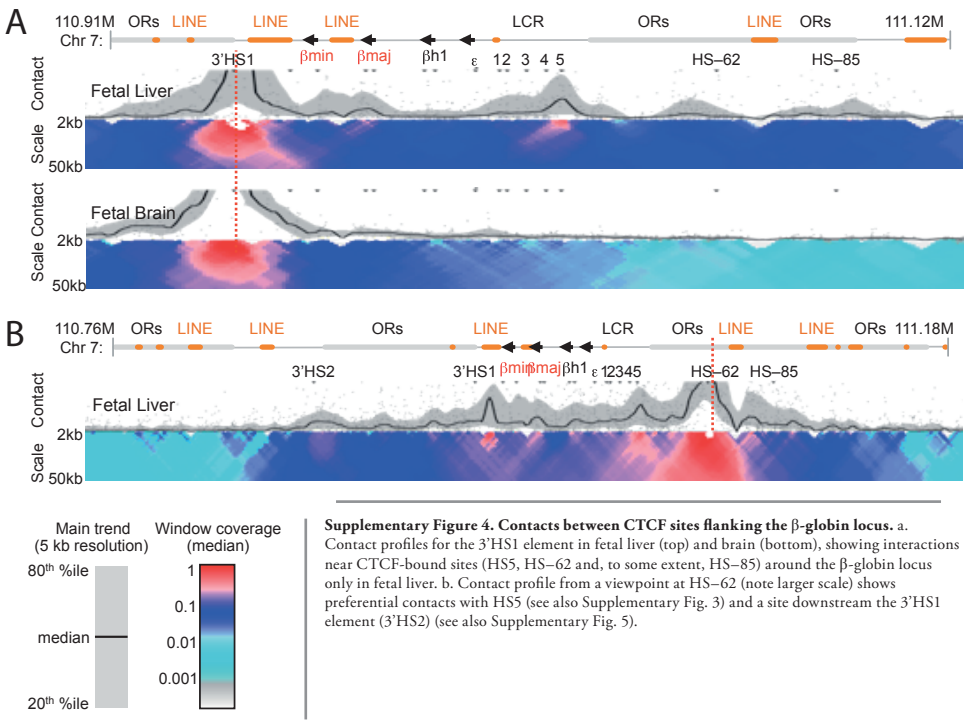
Supplementary Figure 1. DNA products at different stages of the 4C procedure, separated by gel electrophoresis. Left: products after first digestion with NlalII and DpnII. Middle: DNA products after first ligation. Right: DNA products after second round of digestion and ligation. Percentages on top of the lanes indicate fixation conditions (10 minutes with 0.5, 1.0 or 2.0 % formaldehyde). Graphical representation of the DNA size distributions is given in Fig. 1b.



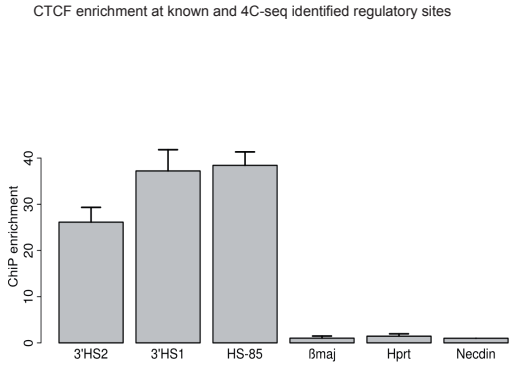
Supplementary Figure 2. Key considerations in the normalization of 4C-seq coverage profiles. A. Information in inter-chromosomal reads. Shown are the probabilities of covering a 5' fragment end (with 1 or more reads) conditioned on a) no coverage on the coupled 3' fragment end, b) coverage of exactly one read for the coupled 3' fragment end, or c) more than one read covering the coupled 3' fragment end. Each graph represents data for one 4C-seq experiment (indexed as described in Supplementary Table 2). For the analysis here, only fragments for which both ends are unique in the genome were considered, and fragments in the chromosome of the viewpoints were discarded. While a modest increase in the coverage probability is observed when conditioning on coverage for the other fragment end, the data show that in most cases, inter-chromosomal contacting fragments are spurious. The data also show that covering an inter-chromosomal fragment more than once almost never contributes additional information (in marked contrast to the situation in the densely covered viewpoint vicinity, as shown below). b. Information in remote intra-chromosomal reads. As above, but considering only the chromosome of the viewpoint and ignoring fragment ends within 10 Mb of the viewpoint. While in this regime there usually appears some added value to coverage greater than 1, this added value rarely represent more than a 1% increase in probability of 5' coverage. c. Quantitative 4C-seq coverage in the viewpoint vicinity. Shown are scatter plots depicting the correspondence between raw coverage of 5' and coupled 3' fragment ends in a region of 200 kb around the 4C viewpoint. Data for blind fragments (fragment lacking a second restriction site) and non-blind fragment (fragments containing at least one second restriction site) are plotted in red and blue, respectively. The analysis reflects significant quantitative correlation across 3 orders of magnitude (1 to over 1000), but also reveals a systematic bias lowering coverage for blind fragments. Our normalization scheme for the region proximal to the viewpoint takes this into account when generating the combined contact profile. d. Fragment end length and fragment length coverage biases. Shown are inferred model parameters for inter-chromosomal contacts (probability of coverage given a range of fragment length and fragment end length) for some of the experiments presented. Statistics from blind fragment ends are depicted separately, as these are normalized as an independent group. e. Depiction of non-blind and blind fragment ends. Shown is a model locus indicating several non-blind and one blind fragment end. Blind fragment ends are typically longer than non-blind fragment ends, and therefore on average more difficult to amplify.



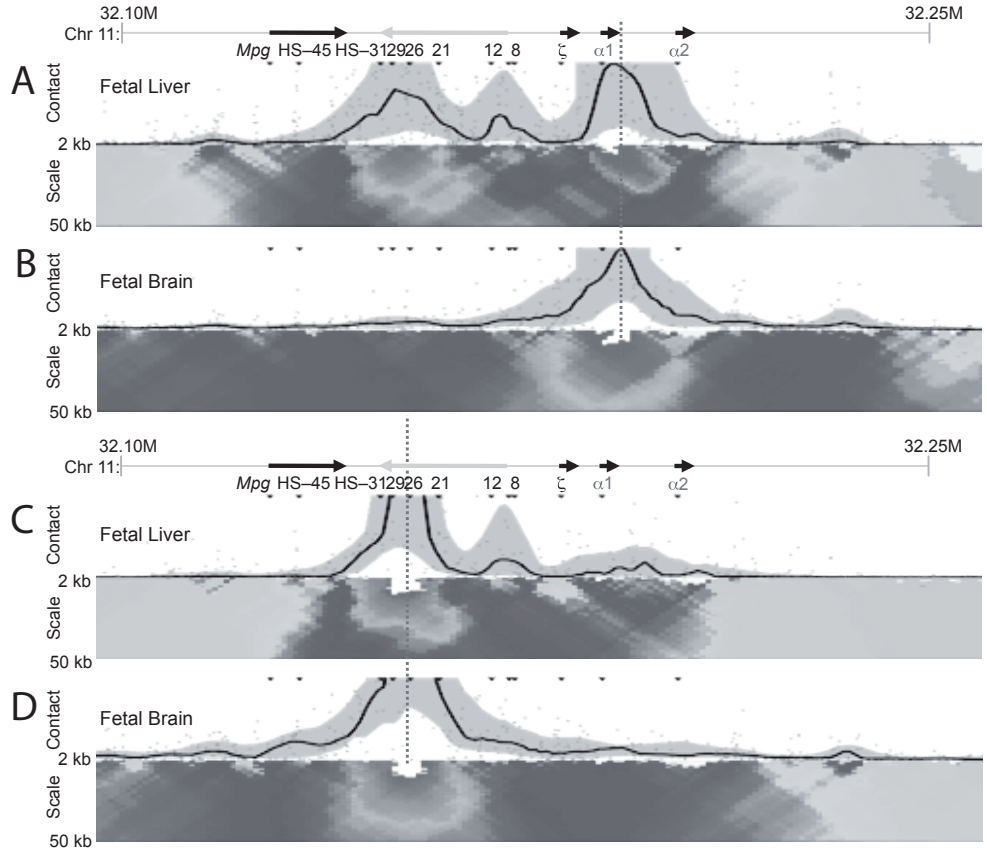
Supplementary Figure 3. Contact specificity within the LCR. Shown are zoom-ins focused on the LCR, illustrating the specificity of interaction between the gene promoter and the proximal portion of the LCR, as compared to the specific interactions between 3'HS1 and HS–62 with the distal portion of the LCR. Note that the upper profile suggests that a LINE element downstream of the LCR may participate in the βmaj interaction.



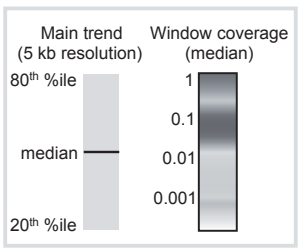
Supplementary Figure 4. Contacts between CTCF sites flanking the β-globin locus. a. Contact profiles for the 3'HS1 element in fetal liver (top) and brain (bottom), showing interactions near CTCF-bound sites (HS5, HS–62 and, to some extent, HS–85) around the β-globin locus only in fetal liver. b. Contact profile from a viewpoint at HS–62 (note larger scale) shows preferential contacts with HS5 (see also Supplementary Fig. 3) and a site downstream the 3'HS1 element (3'HS2) (see also Supplementary Fig. 5).



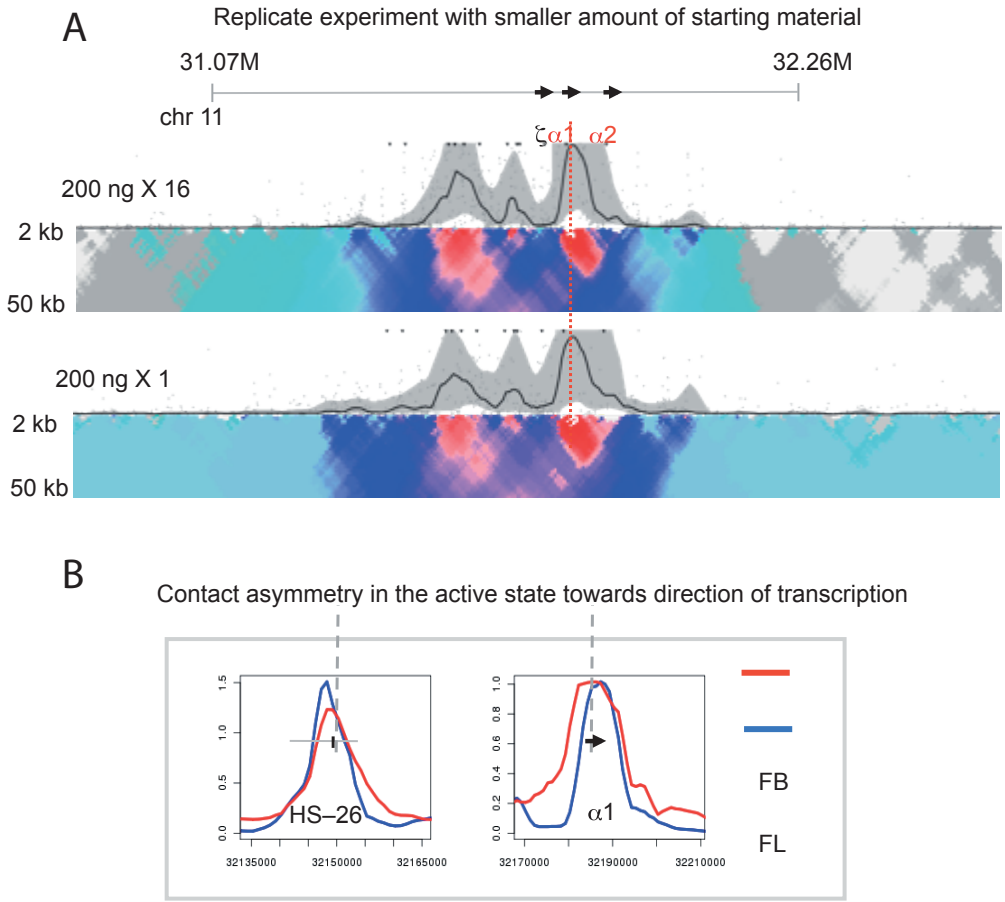
Supplementary Figure 5. 4C-seq identifies a new CTCF site downstream of the β-globin locus. Chromatin immunoprecipitation of CTCF in fetal liver cells. Six loci were tested: three identified using high resolution 4C-seq from HS–62, which contains a CTCF binding site (3'HS1, 3'HS2 and HS–85, n = 3 replicates), a negative control in the β-globin locus (βmaj, n = 2) and a negative control outside the locus (Hprt, n = 3). Necdin is used to normalize enrichment results within each experiment (error bars indicate standard error of the mean).



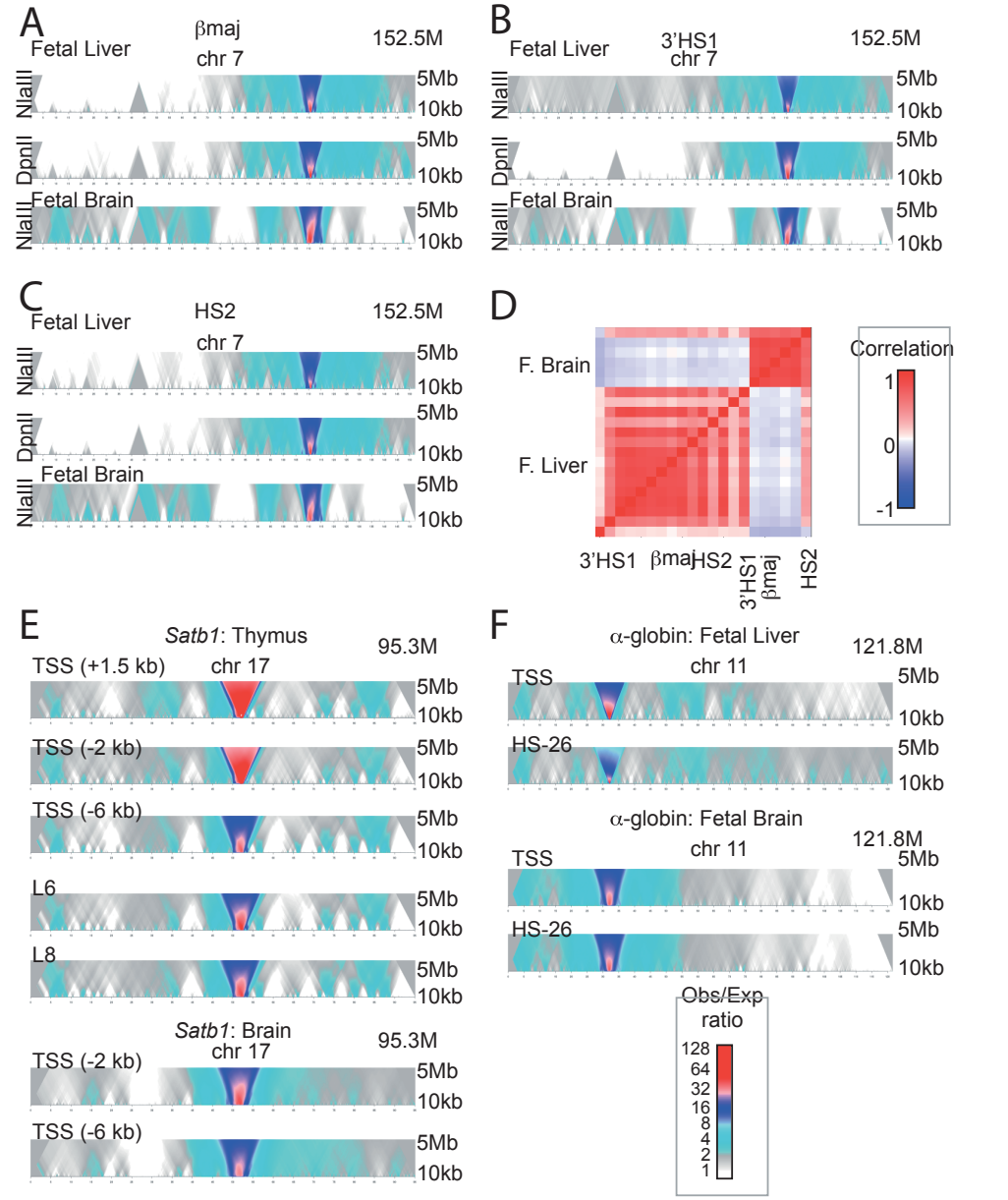
Supplementary Figure 6. Reciprocal promoter-enhancer contacts in the α-globin cluster. Background information. Previous studies identified several DNase hypersensitive sites upstream of the α-globin genes Hba-a1 (or α1) and Hba-a2 (α2). Of these, HS–26 is the most critical site for transcription control(26,27). Using semi-quantitative 3C technology(16,17) and a low-complexity 4C strategy(11), HS–26 was previously shown to physically interact with the active mouse α-globin genes.
In this figure, marked by 'Contact', normalized contact intensities (gray dots) and their running median trends (black line) are depicted. Medians are computed for 5 kb windows and the gray band displays the 20–80% percentiles for these windows. Below the profile, marked by 'Scale', we depict gray-coded medians of normalized contact intensities for multiple scales (from 2 kb (top row) to 50 kb (bottom)). Note that medians are readily comparable between different window sizes. Key features in the α-globin cluster (genes, hypersensitive sites) are depicted on top, and dashed red lines indicate viewpoint location. a. Contact profile from a viewpoint at the



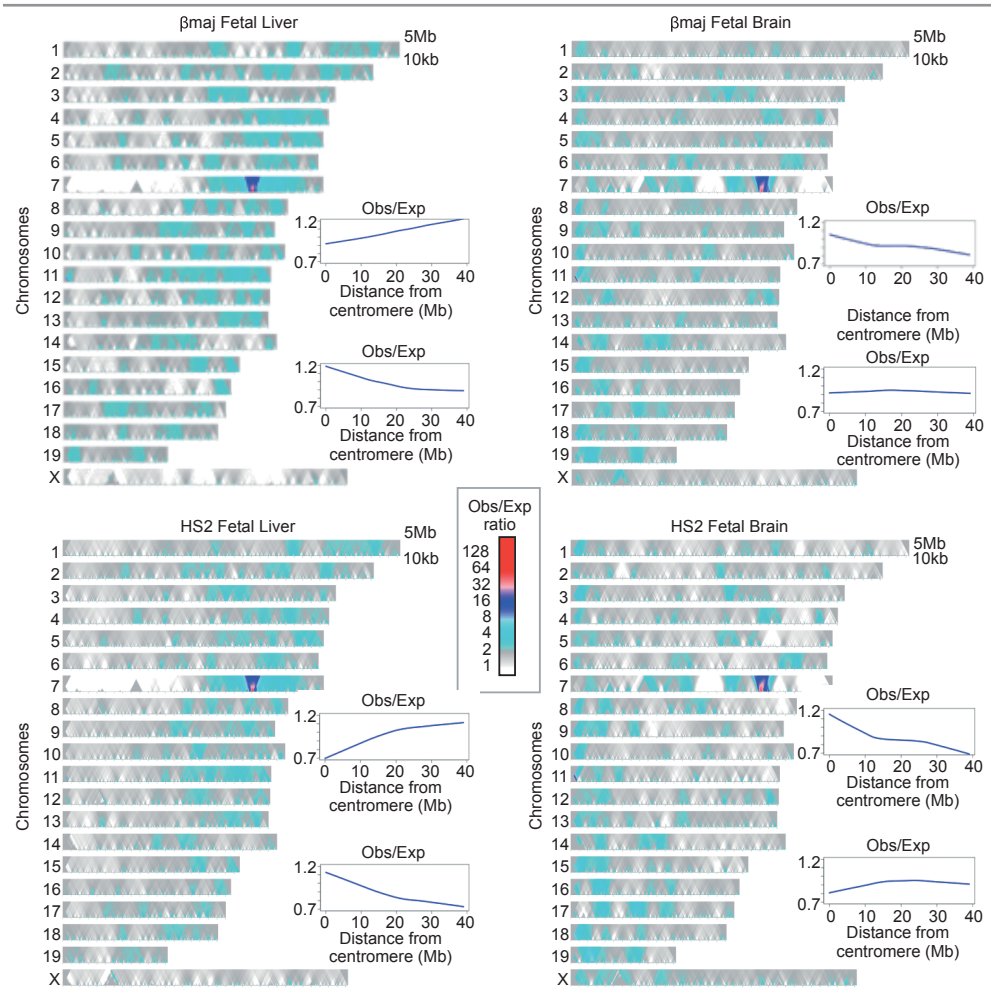
Transcriptional Start Site (TSS) of the active **(continued from previous page)** α1 globin gene in fetal liver. Note its preferential contacts with HS–26, HS–8 and the domain that demarcates the two α-globin genes and upstream regulatory sequences (color coded multiscale view). b. Contact profile from a viewpoint at the TSS of the inactive α1 globin gene in fetal brain. The remote contacts, as well as the domain organization that limits preferred contacts to the region spanning the regulatory sequences and α-globin genes, are completely absent in, confirming that these topological features are specific to the active state. c. Contact profile from a viewpoint near the HS–26 enhancer in the active locus (fetal liver), demonstrating reciprocal contact with the α-globin genes. d. Contact profile from a viewpoint near the HS–26 enhancer in the inactive locus (fetal brain), showing shows that domain organization and specific long-range contacts are not observed when the genes are inactive.
Explanatory note. The plots in Supplementary Fig. 6a and 6c may suggest that the intensity of the enhancer-anchored contact is weaker than the reciprocal promoter-anchored contact. It needs to be emphasized that the plotted contact intensities represent relative 4C-seq read counts for 5kb genomic windows, and reciprocal peak intensities therefore do not need to be identical. For example, when more than one alternative enhancer is present in the 5 kb window around HS–26, and the α1 gene promoter is engaged in specific contacts with these enhancers (possibly in a mutually exclusive fashion), the resulting window statistics may be higher when assaying from the promoter toward the enhancer region than it will be when assaying from any particular enhancer element toward the promoter.



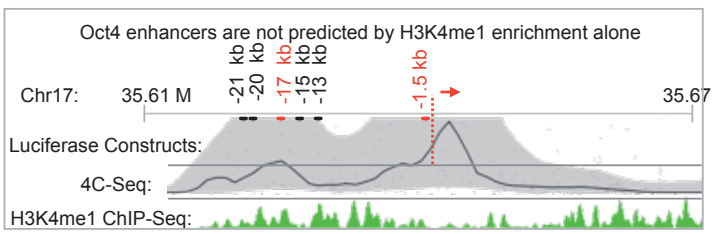
Supplementary Figure 7. Further characterizing promoter-enhancer contacts in the α-globin locus with smaller amounts of template. a. Replicate experiments with variable amounts of starting material for the Hba-a1 viewpoint. Shown are two profiles generated for the Hba-a1 viewpoint using 3.2 µg 4C template (top) and 0.2 µg 4C template (bottom), demonstrating excellent agreement between replicate experiments, even when using less starting material (with a minor loss in definition). b. Asymmetric contact preference around viewpoints in the active state. Median coverage in windows of 10 kb near the promoter (right) and enhancer (left) viewpoints, comparing the active FL and inactive FB states. The active state displays an asymmetric trend with a downstream preference around the a1 TSS, while the inactive state behaves symmetrically, supporting the lack of preferential architecture.



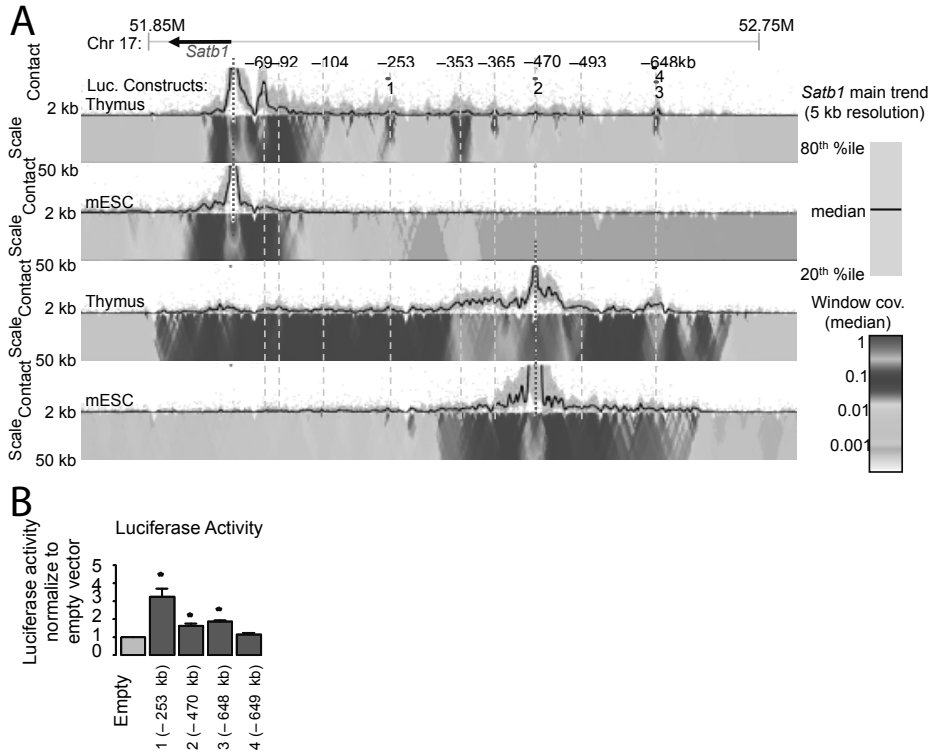
Supplementary Figure 8. Activity-dependent long-range intra-chromosomal contacts. (A-C), Contact enrichment values (ratio between observed and expected number of contacts) for windows of size 10 kb (upper row) to 5 Mb are color coded. For each of three sites in the β-globin cluster (A - 3'HS1, B - βmaj, C- HS2) we depict multi-scale profiles derived from two different fragment pools and conditions (upper – Fetal Liver (FL)/NlaIII, middle – FL/DpnII, lower Fetla Brain (FB)/NlaIII). (D). Correlation among 22 contact profiles derived from viewpoints in the β-globin region. The matrix depicts spearman correlations between enrichment values in 50 kb windows, omitting the 20 Mb surrounding the β-globin cluster. A remarkable separation between active (Fetal Liver) and inactive (Fetal Brain) derived profiles is evident. (E). Similar to a–c, but showing data from Satb1 viewpoints. Note that a highly similar chromosomal contact map is derived from viewpoints within the Satb1 active domain, even if these are 500 kb away on the linear chromosome. (F). Same as a, but comparing chromosomal interactions for different sites in the α-globin cluster.



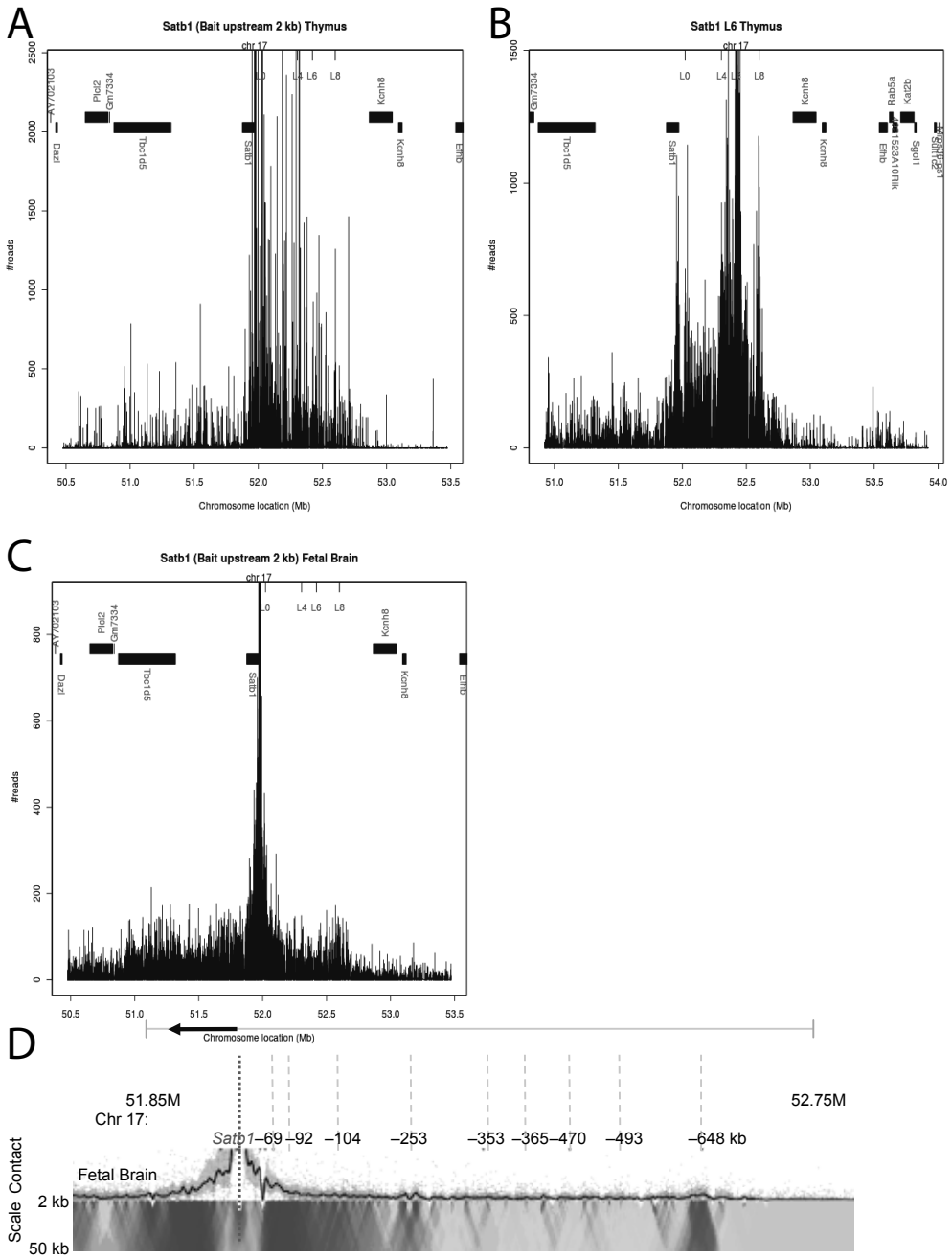
Supplementary Figure 9. Activity dependent whole-genome contexts. Shown are color-coded contact enrichment values (ratios between observed and expected number of contacts) for windows of size 10 kb (bottom row) to 5 Mb (upper row). Data is generated for all chromosomes, but the background model for the viewpoint chromosome is estimated from intra-chromosomal contacts only, while the background model for other chromosomes is estimated from inter-chromosomal contacts only. Data is shown for two viewpoints in the β-globin cluster (βmaj – upper, HS2 – lower) using nuclei in which the cluster is either active (fetal liver) or inactive (fetal brain). A global change in 3D chromosomal context is observed, with the active cluster biased towards telomeric contacts, and the inactive state biased towards centromeric contacts. This is summarized by the trend plots showing average enrichment (excluding intra-chromosomal contacts) as a function of the distance from the centromere (upper) or telomere (lower), averaged over all chromosomes.



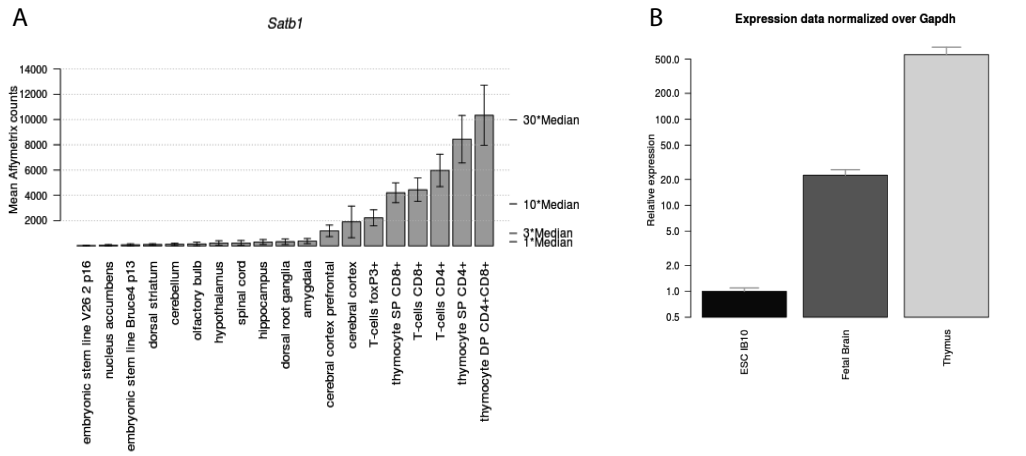
Supplementary Figure 10. Oct4 promoter contacts with a regulatory site are not predicted by H3K4me1 enrichment alone. Shown is a zoom-in on the Oct4 gene promoter and its upstream interacting domain. Sequences tested in luciferase enhancer assays are indicated, with those driving luciferase expression in red and those that do not in black. Mouse ESC H3K4me1 ChIP-Seq data from the ENCODE project(28) (Bing Ren GSM769009) is shown below, demonstrating that functional enhancers may lack H3K4me1 (e.g. 1.5 kb upstream of the TSS), and that sequences occupied by H3K4me1 need not necessarily function as enhancers (e.g. 13 kb upstream of the TSS).



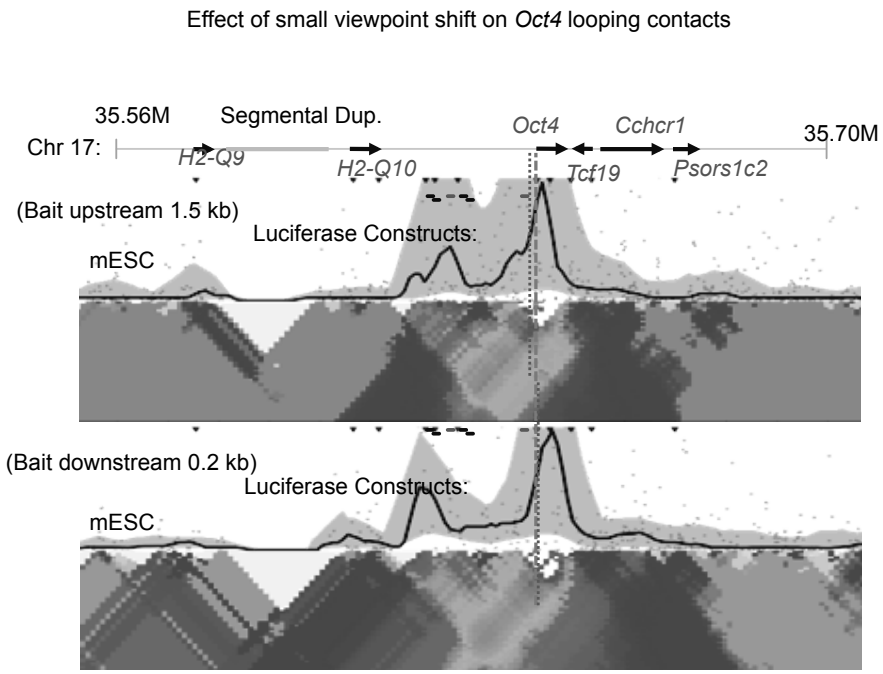
Supplementary Figure 11. Predicting novel promoter-enhancers contacts. (A) Contact profiles for the TSS of Satb1 in thymocytes (top row), the TSS of Satb1 in mESCs (second row), an interacting site 470 kb upstream of the TSS in thymocytes (third row) and the same site in mESCs (bottom row) are shown. A large domain with multiple sites of long range contact is demonstrated specifically in thymocytes, where the gene is highly active. (B) Luciferase reporter assays demonstrating enhancer activity in Jurkat cells of three sites found interacting with Satb1. Asterisk indicates p-value ≤ 0.05 (one-tailed Student's t-test on log-transformed data) with n = 3 independent experiments (each analyzed by triplicate PCR) and error bars indicate standard error of the mean.



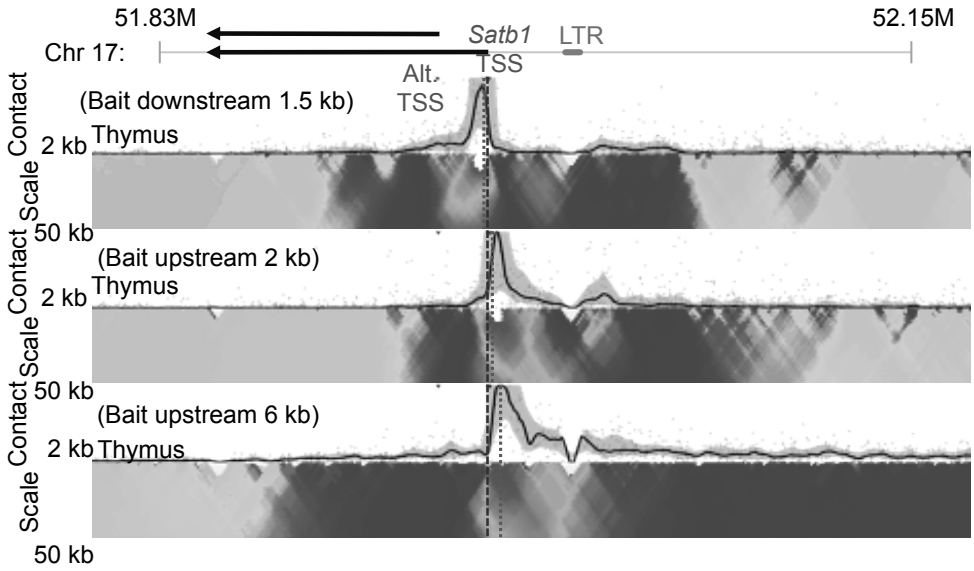
Supplementary Figure 12. Raw data of Satb1 4C-seq profiles. Viewpoints used: (A) Transcription start site (TSS) of Satb1 in thymocytes. (B) Interacting site 470 kb upstream of the TSS in thymocytes. (C) The TSS of Satb1 in fetal brain. Selection of multiple sites of long range contacts (–69 kb (L0), –365 kb (L4), –470 kb (L6) and –649 kb (L8), respectively) are indicated by lines at the top of the graphs, genes and their corresponding gene symbols are shown by black rectangles. (D) Contact profile for the TSS of Satb1 in fetal brain.



Supplementary Figure 13. Expression levels of Satb1 in different tissues. (A) Gene activity chart of Satb1 in various tissues measured by Affymetrix arrays(29). Median value calculated on the data of all tissues; horizontal lines indicate 1, 3, 10 and 30 times the median value. (B) Satb1 expression in mouse Embryonic stem cells (ESC), Fetal Brain and Thymus, using Q-PCR on cDNA and normalized over Gapdh expression levels. Error bars indicate standard error of mean. Results were similar after normalizing over Hprt expression levels and with other Satb1 reverse primer (data not shown).



Supplementary Figure 14. Oct4 TSS viewpoint variation. Shown are two profiles generated using viewpoints located 1.5 kb upstream and 0.2 kb downstream of the Oct4 TSS. While a chromosomal loop is observed in both profiles, the downstream viewpoint contacts peaks toward the outer part of the loop and shows much weaker contact with the luciferase-tested enhancer element revealed by studying contacts from the upstream viewpoint. This suggests that near the active TSS, small offsets in viewpoint positioning may influence the resulting contact profile significantly, partly explaining why lower resolution 3C profiles are often ambiguous or confusing.



Supplementary Figure 15. Satb1 TSS viewpoint variation. Three contact profiles generated using viewpoints 6 kb upstream, 2 kb upstream and 1.5 kb downstream the Satb1 TSS. Remarkable differences in contact intensities are observed, with the downstream viewpoint forming stronger contacts with the gene body, and the upstream viewpoint remaining effectively confined to the immediate TSS –69 kb chromosomal loop.

Supplemental Tables

Supplemental table 1. 4C-seq primers used in this study

Locus	First RE	Second RE	Reading primer	Non-reading primer
HS–26	DpnII	Csp6I	ATGACCAGTGACCCAAGATC	TCTCACAAGCATCTACTGCA
Hba-a1	DpnII	Csp6I	AAGGAATTTGCCCTATGATC	TGTTTGATGGGTATACGTCA
Oct4–1.5kb	NlaIII	Csp6I	CTGTCTGCTCCTACACCATG	AAGTCTTGTGTGAGGGGATT
Oct4–0.2kb	DpnII	Csp6I	GGCGGACATGGGGAGATC	CATGCCATCACTGTCTGTAC
Satb1–1.5kb	NlaIII	Csp6I	AGAACTTTCTCTGAGGCATG	CCTCAGTCCCTTGCATTC
Satb1–2kb	NlaIII	DpnII	GCCAGGAAGAGAAACTCATG	AAGTTTATAGGCACCCACTCT
Satb1–6kb	DpnII	Csp6I	AGGGAAGGAACTCAGGGATC	AAATCTATGGGACAGTCTGAAC
Satb1–470kb	DpnII	Csp6I	AAGGAACCAGCTACCTGATC	AGCCCTATAAGATGCCCTAC
3'HS1	DpnII	Csp6I	AAACCCTAGTTGACCTGATC	CAGCCAATTAATTACCTACCA
Hbb-b1	DpnII	Csp6I	AATGTGAGGAGCAACTGATC	TTAGGCTGCTGGTTGTCTAC
HS2	DpnII	Csp6I	TTTGAGCCCCTCCTTTGATC	ACAACCAGAGGAAACACATC
HS–62	NlaIII	DpnII	CATTCCCCACAGTGTTCATG	TCACCGGAGAAGTTTTCTAA

Supplemental table 2. 4C-seq profiles included in this work

Exp. ID	Viewpoint loc	First RE	Second RE	Tissue	Total reads	Mapped +filtered reads	%Mapped +fi
570	HS–26	DpnII	Csp6I	FL	791015	682227.6	0.86
693	HS–26	DpnII	Csp6I	FB	589740	423397.7	0.72
487	Hba-a1	DpnII	Csp6I	FL	2827655	1124590	0.4
508	Hba-a1	DpnII	Csp6I	FL	1542039	967509	0.63
488	Hba-a1	DpnII	Csp6I	FB	1435446	1237728	0.86
509	Hba-a1	DpnII	Csp6I	FB	1921317	1541778	0.8
656	Hba-a1_(200	DpnII	Csp6I	FL	340743	240890.3	0.71
234	Oct4–1.5kb	NlaIII	DpnII	ES	2143860	847676.5	0.4
629	Oct4–0.2kb	DpnII	Csp6I	ES	73366	48245.77	0.66
226	Satb1–1.5kb	NlaIII	Csp6I	Thymus	3592697	948852.5	0.26
535	Satb1–2kb	NlaIII	DpnII	FB	1608292	1007366	0.63
515	Satb1–2kb	NlaIII	DpnII	Thymus	2024806	801790.1	0.4
683	Satb1–6kb	DpnII	Csp6I	Thymus	2085294	1625742	0.78
643	Satb1–470kb	DpnII	Csp6I	Thymus	1413939	938948.4	0.66
498	3'HS1	DpnII	Csp6I	FL	1039678	557981.1	0.54
524	3'HS1	DpnII	Csp6I	FL	1760948	884850.5	0.5
685	3'HS1	DpnII	Csp6I	FB	420331	287686.2	0.68
489	Hbb-b1	DpnII	Csp6I	FL	1109806	915288.3	0.82
510	Hbb-b1	DpnII	Csp6I	FL	1246503	480811.5	0.39
692	Hbb-b1	DpnII	Csp6I	FB	351348	280834.5	0.8
532	HS2	DpnII	Csp6I	FL	1810899	1274804	0.7
553	HS2	DpnII	Csp6I	FL	2144733	1601760	0.75
694	HS2	DpnII	Csp6I	FB	863838	721173.5	0.83
206	HS–62	NlaIII	DpnII	FL	1164645	652512.5	0.56

Supplemental table 3. Primers used to amplify and clone candidate enhancers

Primer name	Primer sequence
Oct4-1.5kb fw:	GCAAGAAACTGGATCAGATG
Oct4-1.5kb rev:	TCCAGTCCTCCAGAGTTTAG
Oct4-13kb fw:	TTCCCAGAAATCCACGCTAC
Oct4-13kb rev:	CAATGGGTGTTCCAGGCTAT
Oct4-15kb fw:	GCCGATTAGGGTAAACCATC
Oct4-15kb rev:	TAGGGGAGGACTCTGCAGTT
Oct4-17kb fw:	TGATTTCCAATTCCCTTTTGTT
Oct4-17kb rev:	AAAGTTGGGCAGGGCTGTAT
Oct4-20kb fw:	AATGGTAGGAGGCACCTTCA
Oct4-20kb rev:	GGGACCATGAATTTGTGACC
Oct4-21kb fw:	TTTCCCAAACCAGGAGTTTTT
Oct4-21kb rev:	CCAGCATGGAGTCACAAGAA
Satb1-253kb fw:	CACAACATTAAAGAATATATTTTCAACTG
Satb1-253kb rev:	TTGTCTTCATTTCATCAGA
Satb1-470kb fw:	GGGTAGCAACACAAGCCATT
Satb1-470kb rev:	CGGGATTTACATGGGAACTG
Satb1-648kb fw:	AACTCTGTCATCAGGAAATAAG
Satb1-648kb rev:	AGCTACTTCACCATAGA
Satb1-649kb fw:	TATGGCATCTTTAGGCTGAT
Satb1-649kb rev:	ATCAAAAGGCTTGAACAAAA

Supplemental table 4. Primers used for qPCR

For ChIP quantification	
3'HS1-fw	AATCAGTGGAACACTTCTGC
3'HS1-rev	GTCTCAGGTTGTCAACTAAAGC
3'HS2-fw	CAGAACGTCAAGAAATCTAATAAA
3'HS2-rev	CCTCACATTTCTCTGGTGTC
HS-85–fw	GAGACTAAGTAATTCACCATGGG
HS-85-rev	GGATCTATCTTGATTGTCCTCC
Hprt-fw	AGCCTAAGATGAGCGCAAGT
Hprt-rev	ATGGCCACAGGACTAGAACA
Ndn-fw	GGTCCTGCTCTGATCCGAAG
Ndn-rev	GGGTCGCTCAGGTCCTTACTT
For cDNA quantification	
Satb1-fw	AGTGCCCCCTTTCACAGAG
Satb1-rev	TGGTTGGCACCTTGCTGGGA
Satb1-rev2	TGCTGCTGAGACATTTGCAT
Hprt-fw	AGCCTAAGATGAGCGCAAGT
Hprt-rev	ATGGCCACAGGACTAGAACA
Gapdh-fw	TTCACCACCATGGAGAAGGC
Gapdh-rev	GGCATGGACTGTGGTCATGA

5
Functional analysis of the role of CTCF in mediating chromatin flexibility

Work in progress

Functional analysis of the role of CTCF in mediating chromatin flexibility

Sjoerd J.B. Holwerda[1], Elzo de Wit[1], Harmen J.G. van de Werken[1,2], Patrick J. Wijchers[1], Michal Mokry[1,3], Edwin Cuppen[1,4], Wouter de Laat[1]

1 Hubrecht Institute-KNAW & University Medical Center Utrecht, Uppsalalaan 8, 3584 CT Utrecht, The Netherlands.
2 Current affiliation: Department of Cell Biology, Erasmus MC Rotterdam, PO Box 2040, 3000 CA Rotterdam, The Netherlands
3 Current affiliation: Laboratory of Pediatric Gastroenterology, Wilhelmina Children's Hospital, University Medical Centre, Lundlaan 6, 3584 EA Utrecht, The Netherlands
4 Department of Medical Genetics, UMC Utrecht, Universiteitsweg 100, 3584 GG Utrecht, The Netherlands

ABSTRACT

Nuclear organization has been linked to gene expression and can be a contributing factor to the function of the genome. The formation of chromatin loops can, for instance, facilitate communication between enhancer elements and gene promoters. The best-studied protein that is associated with the formation of chromatin loops, both in relation to gene expression and boundary formation is CCCTC binding protein (CTCF). Here, we applied high-resolution 4C-seq to more than one hundred CTCF binding sites (CTCFbs) across different cell types to characterize looping behavior. We find that ~70% of the selected sites is involved in looping, with strong binding sites more often forming loops than weaker binding sites. Looping nearly always occurs to another CTCFbs, but not necessarily to the nearest neighbor. Looping efficiency is not dramatically increased by the co-association of cohesin and also not by the presence of H3K27Ac marked at surrounding nucleosomes, which typically represent enhancer sites. Similarly typed CTCF sites however do show a preference to contact each other. Cellular differentiation is accompanied by the disappearance and the de novo formation of CTCF loops. Interestingly, this happens predominantly near tissue-specific genes, both between tissue-specific and tissue-invariant CTCFbs, strongly suggesting that looping behavior and interaction partner selection of CTCFbs is determined by tissue-specific changes in the repertoire of co-associated factors. Our data provide high-resolution insight into the looping behavior of CTCFbs and their impact on the flexibility of surrounding chromatin.

Introduction

The DNA binding capacity of CTCF is enabled by eleven zinc-fingers that can be used in different combinations to recognize a variety of DNA sequences (1,2). The various binding sites of CTCF share a consensus sequence containing the CCCTC motif (3), hence the name. CTCF is one of the few mammalian transcription factors described so far that can act as a boundary: it can prevent regulatory elements such as enhancers to act across its binding site (4-6).
Chromatin immunoprecipitation experiments of CTCF followed by high throughput sequencing (ChIPseq) in multiple tissues and cell lines revealed tens of thousands of CTCFbs in the mammalian genome (7-10). Their location provides further support for a boundary function of CTCF genome-wide (11,12). CTCFbs for example often mark transitions between active and inactive chromatin domains, as defined by the distribution of histone marks (11), or between domains that associate with the nuclear lamina and those

that adopt more internal nuclear positions (LADs) (12). CTCFbs also often demarcate so-called topological domains (13). The boundaries of these domains hamper DNA contacts between two flanking domains and thus subdivide chromosomes into distinct topological entities (13-16). Possibly, the mere binding of CTCF to chromatin is sufficient to create such a boundary. Alternatively, CTCF performs this function through its capacity to form chromatin loops (17). Interestingly, there are many more CTCFbs than chromatin boundaries (~40.000 sites versus e.g. ~2200 topological boundaries). Also, it was found that not all CTCFbs form chromatin loops (18). This raises an intriguing question: what characteristics of a CTCFbs determine its ability to act as borders or enable them to be engaged in long-range chromatin loops?
CTCF is capable of bringing together, or at least stabilizing contacts between, distant chromosomal binding sites (19,20). These loops may prevent interactions between enhancers and promoters if spatially separating them into distinct chromatin loops or oppositely, they can facilitate contacts between enhancers and promoters when encompassing them in the same chromatin loop (17). The latter seems to be true for CTCFbs that co-associate with cohesin (21-24). Indeed, a subset of CTCFbs is found to be shared with cohesin (25,26). Cohesin is a protein complex that indirectly binds DNA and is involved in the formation of chromatin loops, possibly by exploiting its ring-like structure to embrace DNA helices (26-29). It not only functions to hold together sister chromatids after DNA replication, but also appears to affect gene expression (25,26,30,31). Cohesin can exert these roles independent of CTCF (25,26,30,31) although CTCFbs co-associated with cohesin are also found to bind gene promoters and their distal regulatory elements (25). 'Mirrored' co-association of CTCF and cohesin at these regulatory elements and gene promoters mediated interactions between the two elements (26). However, the role of CTCF in the formation of such transcriptionally interesting loops at tissue specifically expressed genes is succinct.
Here we applied high resolution 4C-seq to hundreds of CTCF sites (1) to better understand what features in general determine the looping capacity of a CTCFbs in ESCs and (2) to investigate if CTCF can play a role in tissue specific gene expression through its involvement in long-range chromatin loops. Experiments were performed in embryonic stem cell (ESC) lines, in neural progenitor cells (NPCs) obtained through in vitro differentiation of ESCs (32) and in primary mouse fetal livers (FL). This setup allowed us to compare the 4C contact profiles for constitutive and tissue specific CTCFbs between these tissues. Moreover, CTCF sites co-associated with cohesin and or sites overlapping with enhancers, marked by H3K27ac, were included in the analysis, to assess the role of these chromatin features of CTCFbs in their ability to form loops or chromatin boundaries.

Results

Mapping and characterization of tissue specific and cell type invariant CTCF binding events during differentiation

To study the role of CTCF in chromatin folding we first mapped and characterized its binding behavior in three different cell types, two embryonic stem cell (ESC) lines, E14 and WW6, from different genetic backgrounds, and a E14 derived neural progenitor cell (NPC) (32) (**Figure 1a**). Sequential ChIP followed by high throughput sequencing was used to determine CTCF binding locations, resulting in binding profiles with high signal-to-noise ratio (33). CTCF peaks with a minimum of 25 tags per peak were called positive, revealing roughly 40.000 CTCF binding sites (CTCFbs) per cell type, comparable to previously published results (8). In ESCs 10-15% of the CTCFbs showed specific binding in one of the two ESC types (**Figure 1b**), whereas 20-25% of the CTCFbs showed tissue specific binding between ESCs and NPCs (**Figure 1c**). As expected (9), CTCF has a preference to bind in the vicinity of genes but is not exclusively enriched there, as a substantial proportion of CTCFbs binds at intergenic regions (**Figure 1d**).

Inspection of the tissue specific CTCFbs (**Figure 1e**) reveals that they are distributed similarly as the conserved binding sites, with no obvious preference for gain or loss of binding sites at either promoter, intragenic or intergenic regions. Analysis of ChIPseq tag scores confirmed previous observations that the most prominent CTCF peaks are conserved throughout cell types whereas tissue specific events show less prominent peaks (**Figure 1e**) (10,34). The fact that the majority of CTCFbs is conserved between pluripotent cells and differentiated cells is also in line with published data suggesting that CTCF mostly has a tissue invariant function (10).

Characterization of CTCF sites engaged in long range looping in embryonic stem cells

CTCF has been described as a general looping factor (reviewed in (35)), and is found involved in long range chromatin looping at many individual gene loci (19,20,36-38). Although it is widely accepted that CTCF engages in chromatin loops, it has also been shown that not all CTCF sites do this (18). The features that determine whether or not CTCFbs participate in loops are unknown. To start to explore this, we made use of 4C sequencing (39) to create high-resolution DNA contact maps of a series of CTCFbs. We defined several categories of CTCFbs, including (1) CTCF sites that are co-associated with cohesin (Rad21, a member of the cohesin complex) but that lack H3K27Ac marks (40), (2) CTCF sites that carry the active enhancer mark H3K27Ac (41) but lack cohesin, (3) CTCF sites carrying both cohesin and H3K27Ac, and (4) CTCF sites that have none of these features (CTCF only sites). In total, we generated 4C contact profiles for 126 different CTCF sites in E14 ESCs. We scored for the presence or absence of chromatin loops by visual inspection of the contact maps, considering windows of 450kb on either side of a viewpoint. We only scored and characterized one interaction partner per 4C viewpoint, being the most prominent interaction in the particular window. Examples of 4C profiles with and without chromatin loops are provided in **Figure 2a** and throughout this chapter. In a majority of cases analyzed (~70-75%) we found CTCF sites to be engaged in a loop (**Figure 2b**). Looping efficiency was not different between any of the four categories of CTCF sites, suggesting that co-associated cohesin and/or the presence of H3K27ac per se is not sufficient to impose loop formation (**Figure 2b**). Looping capacity was slightly influenced by the binding strength of CTCF, such that overall, CTCF sites involved in long-range contacts showed significant but not dramatically higher ChIPseq scores than sites not involved in looping (**Figure 2c)**. Per category, this however was only apparent for the cohesin carrying CTCF sites and not for the others (**Figure 2c)**. Moreover, independent of which category of CTCFbs was considered, no threshold for ChIPseq scores could be defined that strictly separates the looped from non-looped CTCF sites. We conclude that the level of CTCF occupancy at a site contributes to its ability to form chromatin loops but the level of occupancy alone is not sufficient to predict loop formation.

The ability of a CTCF site to become engaged in a specific long-range interaction must also rely on the availability of suitable DNA contact partners in cis. To further search for factors that can explain a CTCF site's involvement in looping, we therefore analyzed the density of CTCF binding events in a 900 kb window around each CTCF site of interest. Looping and non-looping CTCF sites were surrounded by domains with similar CTCFbs density, showing that the wider genomic context is not the discerning factor (**Figure 2d**). We then asked whether the presence of a strong CTCF binding site (arbitrarily defined as sites with a CTCF ChIPseq scores of >120) in the neighborhood influenced looping ability. Highly occupied CTCF sites are indeed more often found in the vicinity of looped versus non-looped CTCF sites, but again they are neither necessary nor sufficient to induce chromatin loops (**Figure 2e**). We finally asked whether the categories of CTCF sites showed preferred interaction partners. We note that for nearly all analyzed CTCF sites (95%) the most obvious chromatin loop formed was with another CTCF site (**Figure 2f & 2g**). All four categories of CTCF sites seem to have a preference to form contacts with category 3 type of CTCF sites (CTCF+cohesin+H3K27Ac), which is particularly true for sites belonging to this same category. Preferred homotypic interactions are also observed among cohesin shared sites (category 1) and among

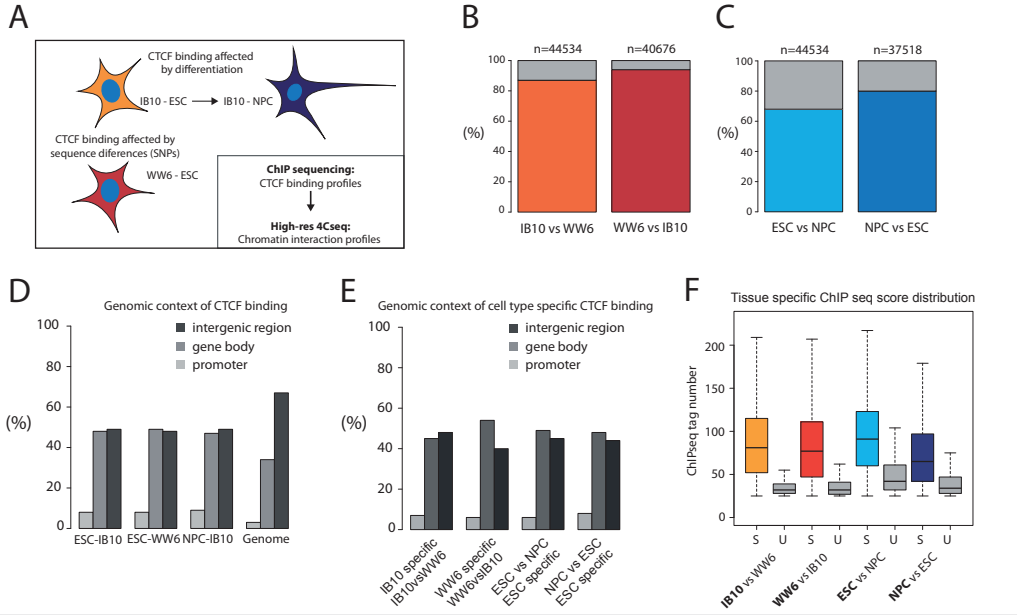


Figure 1 Characterization of CTCF binding events between pluripotent and differentiated cells. (A) Cell systems used to characterize CTCF binding and create 4C contact profiles. E14 ESCs were differentiated, in an in-vitro assay as described 32. (B) CTCF binding characteristics between E14 (129/ Ola) and WW6 (129/ C57bl/6/ SJL) ESCs. The orange/red or grey bars depict the percentage of shared or unique sites between the tissues, respectively. The numbers on top of the bars depict the total number of sites per cell type. (C) CTCF binding characteristics between E14 ESCs and NPCs. Blue bars depict the percentage of conserved sites between the two cell lines. (D) Distribution of all CTCF binding sites per cell type across the genome, divided over intergenic regions, promoters and gene bodies. (E) Same as for (D), but for tissue specific CTCF sites, per comparison. (F) ChIPseq tag counts for shared (S; colored bars) and unique (U; grey bars) CTCF sites per compared cell type pairs. Bold font depicts the cell type for which the bars specify the characteristics.

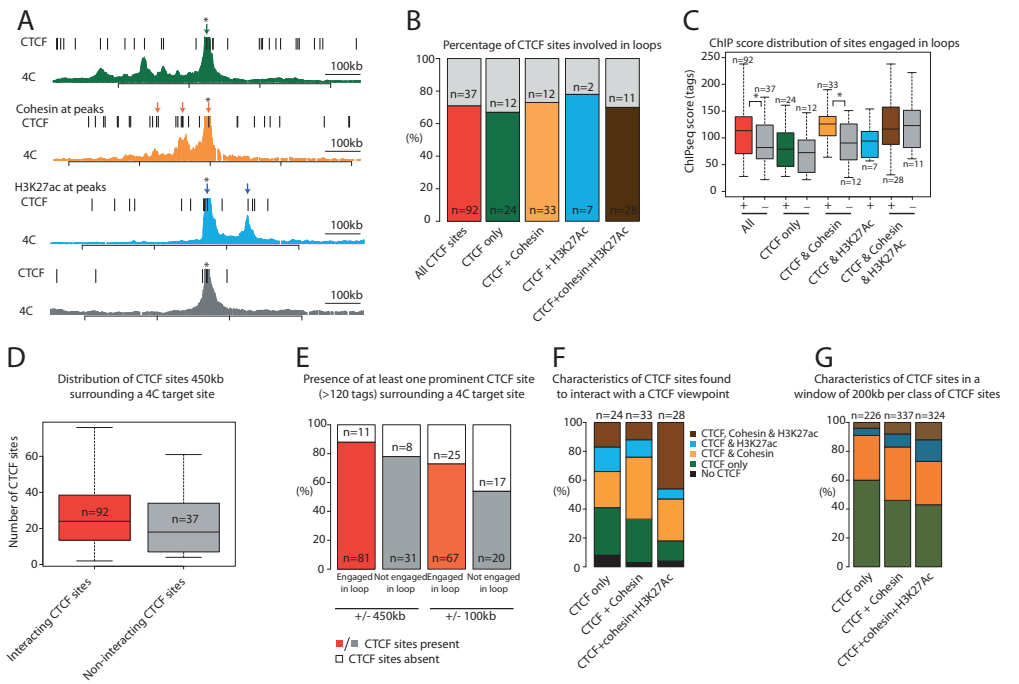


Figure 2 Characteristics of CTCF sites engaged in chromatin loops. (A) 4C contact profiles (4C) for viewpoints with different characteristics: exclusive binding CTCF (green track), CTCF co-associated with cohesin (orange track), CTCF associated with H3K27ac (blue track) and an example where CTCF is not engaged in looping (grey track). Above the tracks CTCF sites (black lines) and either cohesin (blue arrow) or H3K27ac (red arrow) are depicted.

(continued from previous page) Vertical bars below the track are drawn every 200kb, length of the total track is 900kb. The graphs are centred around the 4C viewpoint (asterisk). (B) Distribution of CTCF sites involved in loops per category. Colored bars depict CTCF sites engaged in a loop, grey bars depict sites not involved in a loop. Numbers depict the number of CTCFbs per category. (C) CTCFbs ChIP-seq score distribution per category. The X-axis depicts the viewpoint categories for CTCFbs engaged in a loop (+) or not engaged in a loop (-). The Y-axis depicts the number of tags counted per CTCF peak. The asterisks depict a significant difference ($p<0.05$) as tested using the Wilcoxon test. The median per set is depicted as a black bar in each box. (D) Boxplot showing the number of CTCF sites flanking the a 'viewpoint CTCFbs' engaged in a loop (red box) or viewpoints not involved in a loop (grey box) within a window of 900kb (450kb on either side). The median per set is depicted as a black bar in each box. (E) Percentage of CTCF sites with least one prominent CTCF site (>120 tag counts) within a window of 900kb (red bars) or within a window of 200kb (orange bars). The distribution around sites engaged in a loop is depicted in the colored bars, distribution around sites not engaged in a loop in grey bars. White bars indicate the number of times that a 'viewpoint CTCFbs' is not flanked by another prominent site in the appropriate window. (F) The distribution of CTCF sites interacting with 'viewpoint CTCFbs'. The X-axis depicts the categories of CTCF viewpoints used that are engaged in a loop. The Y-axis depicts the proportion of each category of CTCF sites are in contact with the 'viewpoint CTCFbs'. (G) Distribution of CTCF sites with different characteristics (Y-axis) for all CTCF viewpoints with a certain characteristic (X-axis) in a window of 200kb surrounding the CTCF sites. The X-axis depict the classes of CTCFbs that served as viewpoints, the Y-axis depict the sites that are found within a window of 200kb surrounding each class of 'CTCF viewpoints'.

'CTCF-only' sites (category 4), but there is no exclusive combination of interacting sites and all categories can interact with each other. The latter strongly suggests that cohesin does not mediate loop formation through the embracement of two double-strand helices.

Chromatin interactions mediated by CTCF frequently bypass directly neighboring CTCF sites

CTCF has classically been described as an insulator binding protein meaning that it can form a boundary that blocks enhancer-promoter communication when bound in between these genomic features (4,6,42). Genome-wide, CTCF has been assigned a boundary function due to its preferred association to sites of transition between various types of chromatin domains (7,11,12). Collectively, this suggests that CTCF can hamper DNA contacts across its binding site and thereby may define chromosomal domains within which sequences preferentially interact (13,15). Not all CTCF sites however localize to boundaries. Also, the spacing that we find between interacting CTCF sites is larger than that between CTCF sites in general (**Figure 3a**), suggesting that CTCF sites not always contact their nearest neighbor. To further address this, we investigated how often and which type of 'nearest neighbor' CTCF site marked the most prominent interaction partner of our selected CTCF binding sites. In 31% of the cases, the closest CTCFbs appeared to be the preferred 'interactor'. In most instances however (69%), we found the most prominent loop with a site more distal on the chromosome (**Figure 3c,d**). In these situations, the 'skipped' CTCFbs is apparently outcompeted by a more distal site for preferred interactions. We first asked whether low ChIP-seq scores explain why these sites are ignored. Their ChIPseq scores however did not deviate much from those at the engaged site, further emphasizing that occupancy is not a good predictor of loop formation (**Figure 3b**). A further comparison between 'skipped' and contacted CTCF sites revealed that category 3 (CTCF+cohesin+H3K27Ac) sites are clearly more often engaged in contacts than being skipped. Vice versa, 'CTCF only' sites tend to be more often ignored than contacted (**Figure 3e & 3f**). Thus, the co-association of cohesin and H3K27Ac may not be sufficient for a CTCF site to form chromatin loops, the presence of such site close to another CTCF site, and in particular close to another type 3 CTCF site, does increase the chance of efficient loop formation (**Figure 3e & 3f**).

Differential binding behavior of CTCF between ES cell lines is underlined by single nucleotide polymorphisms (SNPs) and can be associated with differential loop formation

CTCF is an essential protein (43), making it difficult to discern direct from indirect effects when knocking-out or -down the protein. To circumvent this problem and nevertheless address the role of CTCF in chromatin looping, we wished to compare looping behavior of sites that in different ES lines differentially bind CTCF due to naturally occurring single nucleotide polymorphisms (SNPs). For this, we searched for differential CTCF binding sites the E14 and the WW6 ES cell lines (which both express the CTCF protein

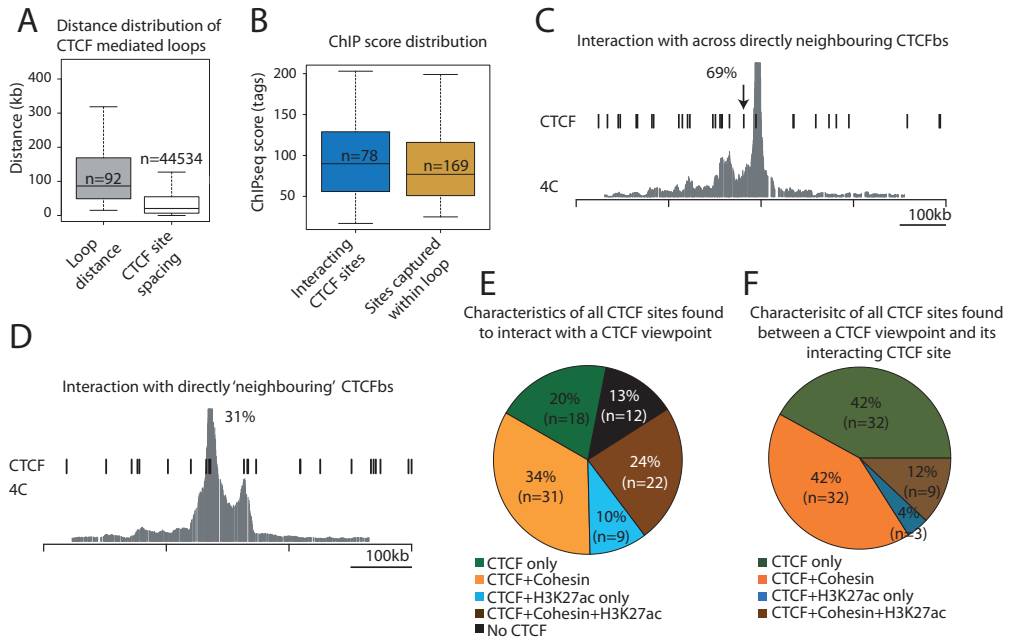


Figure 3. CTCF mediated loops across neighbouring CTCF sites. (A) Boxplot showing the distance distribution CTCF mediated loops (grey box), CTCFbs spacing in the genome (white box). The median per set is depicted as a black bar in each box. (B) Boxplot of the distribution of the ChIP-seq score of CTCFbs interacting with a 4C viewpoint (blue box) and the total number of CTCFbs that are located in between a 4C viewpoint and an interacting CTCFbs (orange box). The median per set is depicted as a black bar in each box. (C) 4C contact profile showing an example of a CTCF mediated loop across a flanking CTCF site. Black bars depict CTCF sites, the arrow depicts the 'skipped' CTCF site and the number depicts the percentage of sites where loops are found across the nearest CTCF site. Vertical bars below the track are drawn every 200kb, the total genomic length of the track is 900kb. (D) 4C contact profile showing an example of a CTCF mediated loop with the nearest flanking CTCF site. (E) Pie chart of the characteristics of contacted CTCF sites sites, categorized by co-occupancy of cohesin, H3K27ac or both or contacts made where CTCF binding was not found. Numbers in the pie chart depict the percentage and the actual number (between brackets) per category. (F) Pie chart showing the characteristics of CTCFbs found in between all CTCF viewpoints and the interacting sequences, categorized as in (E). Numbers in the pie chart depict the percentage and the actual number (between brackets) per category.

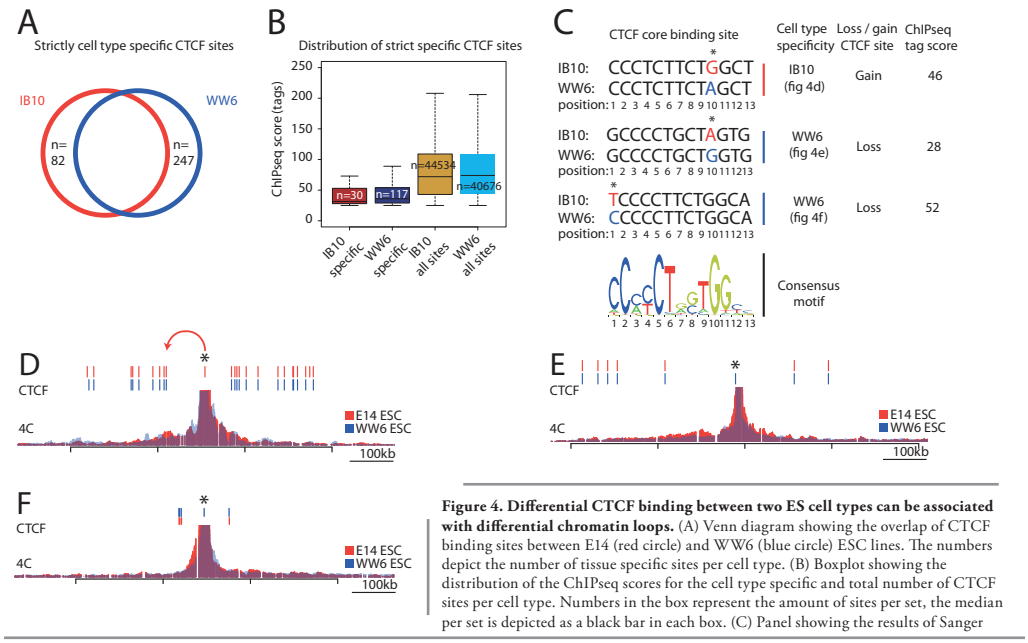


Figure 4. Differential CTCF binding between two ES cell types can be associated with differential chromatin loops. (A) Venn diagram showing the overlap of CTCF binding sites between E14 (red circle) and WW6 (blue circle) ESC lines. The numbers depict the number of tissue specific sites per cell type. (B) Boxplot showing the distribution of the ChIPseq scores for the cell type specific and total number of CTCF sites per cell type. Numbers in the box represent the amount of sites per set, the median per set is depicted as a black bar in each box. (C) Panel showing the results of Sanger

(Continued from previous page) sequenced CTCFbs at the core binding motifs. SNPs that were found in tissue specific binding events per cell type that cause a gain or loss of CTCF binding are depicted by an asterisk. The consensus motif is defined by JASPAR, the positions corresponding to the positions of the consensus motif are depicted below the sequences. ChIPseq tag scores per CTCFbs are depicted. (D) 4C contact profiles plotted on top of each other for E14 (red track) and WW6 (blue track) ESCs. Black bars above the track depict CTCFbs, the asterisk depicts the differential CTCF binding that is used as viewpoint. The red arrow depicts the differential loop. Vertical bars below the track are drawn every 200kb, the total genomic length shown is 900kb. (E), (F) Similar to (D), except different tissue specific CTCFbs are used as 4C viewpoint.

at normal, wild type, levels). Stringent thresholds were set to define cell type specific binding events, implying that sites should have no reads in the one and clearly accumulated reads (25 or more) in the other cell line. This resulted in the detection of 30 and 117 cell type specific binding events in the IB10 or WW6 cell line, respectively (**Figure 4a**). A first analysis of these sites revealed that they showed lower ChIPseq scores than the average CTCFbs found elsewhere in the genome (**Figure 4b**). To study looping capacity of these sites we selected the differential binding sites that have no other CTCFbs within 20 kb at either side, thus to prevent that nearby CTCF binding sites influence the contact profiles and obscure possible differential loops. To validate that these selected sites are true differential binding sites, ChIP results were analyzed by qPCR (data not shown). This showed that, despite the stringent selection criteria, most selected sites did bind CTCF in both ES lines. Three truly cell type specific CTCFbs were picked for further investigation by 4C-seq. In all three cases Sanger sequencing confirmed the presence of a SNP in the core binding site of CTCF, disrupting the consensus sequence in the cell line that failed to show CTCF binding (**Figure 4c**). Only for one site, some differential looping towards a neighboring CTCF was seen when bound by CTCF (**Figure 4d**). In two other cases differential binding of CTCF had no or non-interpretable effects on contacts formed by the site (**Figure 4e & 4f**). Thus, despite our efforts, the two genetically different ES lines were not very helpful in understanding the role of CTCF in chromatin loop formation.

Tissue specific chromatin loops mediated by CTCF during differentiation

ChIPseq studies have shown the tissue specific binding behavior of CTCF (10,30,31) and in some cases tissue specific chromatin loops between CTCFbs have been reported (44,45). Here we followed 111 CTCF sites for loop formation upon *in vitro* differentiation of ESCs to NPCs (32). A characterization of loops formed in NPCs showed that, similar to ESCs, they generally span a distance larger than the average spacing distance between CTCF sites in the genome (**Figure 5a**). Additionally, CTCFbs involved in loops in NPCs also show a trend to have a higher ChIPseq score than sites not engaged in a loop (**Figure 5b**). Comparing the looping behavior between cell type invariant and tissue specific CTCFbs showed that tissue specific loops occurred not only at tissue-specific CTCFbs but also between tissue-invariant CTCFbs (**Figure 5c, e-h)**. The latter strongly suggests that tissue-specific proteins dictate whether a given CTCF site is engaged in looping or not. The tissue specific sites that were engaged in loops were characterized by a relatively low ChIPseq score (**Figure 5d**), showing that particularly for this category of sites binding strength is not predictive for loop formation.

CTCF mediated loops around tissue specifically expressed gene loci

Many tissue specific cis-regulatory elements exist in the genome (46), which potentially have a role in transcription. The role of CTCF mediated chromatin looping in tissue specific gene expression has been studied for several gene loci in using 3C technology (19,44,47,48). To further understand the role of CTCF in setting up chromatin structure between regulatory sequences bound by CTCF around tissue specific genes we systematically followed 30 CTCF sites flanking 14 highly tissue specific genes in three different tissues: ESCs, NPCs and FL tissue (see **supplementary table 2**). Remarkably, 27 of these 30 sites (90%) were conserved between the three tissues, yet 20 (2/3rd) showed tissue-specific loop formation (**Figure 6a**).

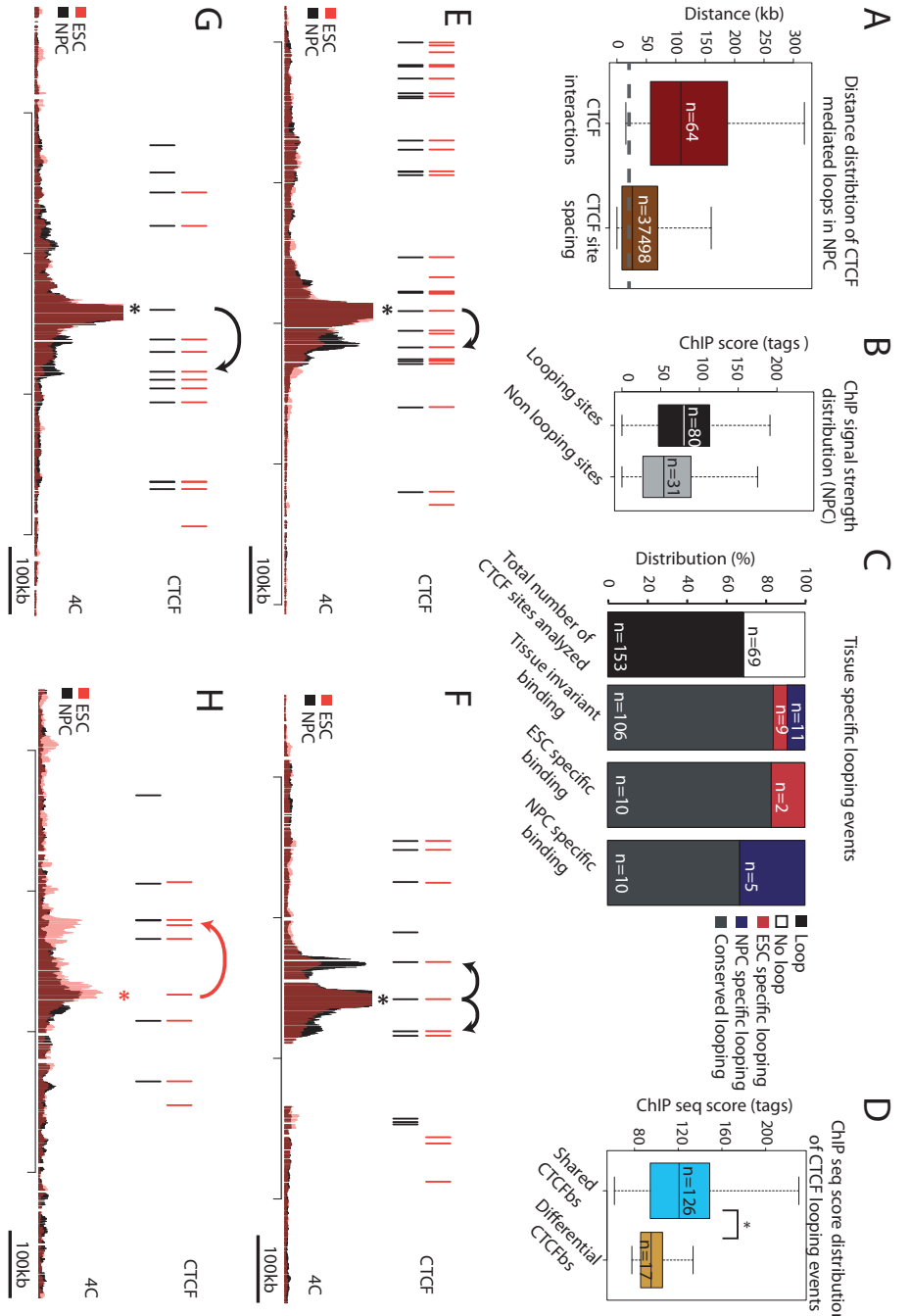


Figure 5. Tissue specific CTCF binding can be associated with tissue specific chromatin structure. (A) Boxplot showing the distribution of the distance found between two interacting sites in CTCF mediated loops (red box) and between CTCFbs in the genome (brown box) in NPCs. The numbers in the boxes represent the number of CTCFbs per category, the median per set is depicted as a black bar in each box. The dotted line represents the median of CTCFbs spacing in ESCs. (B) Boxplot of the distribution of ChIPseq tag score of CTCFbs viewpoints involved in mediating loops (black box) versus sites not involved in loops (grey box). The numbers in the boxes represent the number of CTCFbs per category, the median per set is depicted as a black bar in each box. (C) Barplots showing the percentage of looping events scored between E14 ESCs and NPCs. Each bar represents a category as depicted on the X-axis. The first bar represents the total number of CTCFbs analysed between ESCs and NPCs, the rest of the graph describes the looping events that are found between CTCFbs per category (shown on the X-axis). The Y-axis shows the percentage of sites that are engaged in conserved loops (grey), ESC specific loops (red) or NPC

(continued from previous page) specific loops (blue). The numbers in the bars represent the number of loops per category. (D) Boxplot of the distribution of ChIPseq tag scores for CTCF viewpoints that are engaged in loops at CTCFbs conserved between ESCs and NPCs (blue box) and CTCF viewpoints that are engaged in tissue specific loops (orange box). The asterisk depicts a significant difference ($p<0.05$) as tested using the Wilcoxon test. The numbers in the boxes represent the number of CTCFbs per category, the median per set is depicted as a black bar in each box. (E) 4C contact profiles for ESC (red profile) and NPC (black profile) plotted on top of each other. CTCFbs are depicted above the track by black bars. The arrow depicts a tissue specific interaction in NPCs, the asterisk depicts the CTCFbs that is used as a viewpoint. Vertical bars below the track are drawn every 200kb, total genomic length shown is 900kb. (F), (G) and (H) Similar to (E) except that the contact profiles for different CTCFbs are shown.

The majority of tissue specific loops (75%) were exclusive to one tissue type, being always the tissue expressing this gene (**Figure 6b & supplementary table 2**). Three examples of 'true-tissue specific loops' between tissue specific H3K27ac regulatory sequences at the Sox2 locus (**Figure 6c**), the Olig1 and -2 locus (**Figure 6d**) and the β-globin locus (**Figure 6e**) illustrate that tissue invariant CTCFbs can mediate tissue specific loops around gene loci. This provides further evidence that CTCF loop formation depends on the co-association of tissue-specific proteins at the CTCF binding sites. In cases where loops were found in two out of three of the tissues, it was most often (4/5) shared between the more related, *in-vitro* cultured ESCs and NPCs. They appeared at or around genes active in ESC (*Sox2* and *Nanog*) and NPC (*Pcdh*-α and *Slc1a3*). Only at the ESC-tissue specific gene *Klf4*, we found a loop was shared between ESCs and fetal liver that was absent in NPCs (**supplementary table 2**). We currently do not know yet what co-associated factors dictate the formation of these tissue-specific chromatin loops. Likely candidates to be tested are the well-known pluripotency factors OCT4, NANOG and SOX2 in ESCs, and the erythroid-specific transcription factors GATA1, EKLF and LDB1 in fetal liver. Indeed, the looping sites invariably show H3K27Ac, but in 75% of the cases this mark is also associated in tissues where the sites are not engaged in long-range interactions. Not all tissue specifically expressed genes show cell type specific chromatin loops as conserved looping between CTCFbs was seen around e.g. the Slc4a1 gene locus (**Figure 6f**). A final interesting situation is observed around the protocadherin-α (*Pcdh*-α) gene cluster, which is exclusively expressed in NPCs. Here, a CTCF site located inside a downstream enhancer showed looping to all alternative transcriptional start sites of the *Pcdh*-α gene. Interestingly, all these sites also contain CTCFbs. CTCF was previously reported to be required for the correct expression of the *Pcdh*-α genes (49) and for the formation of loops at the *Pcdh*-α locus (50). Our data confirm these loops but show this topological configuration already pre-exists in ESCs in which CTCF already binds to all these sites but *Pcdh*-α is silent. In fetal liver loops are absent and *Pcdh*-α is repressed (**Figure 6g**).

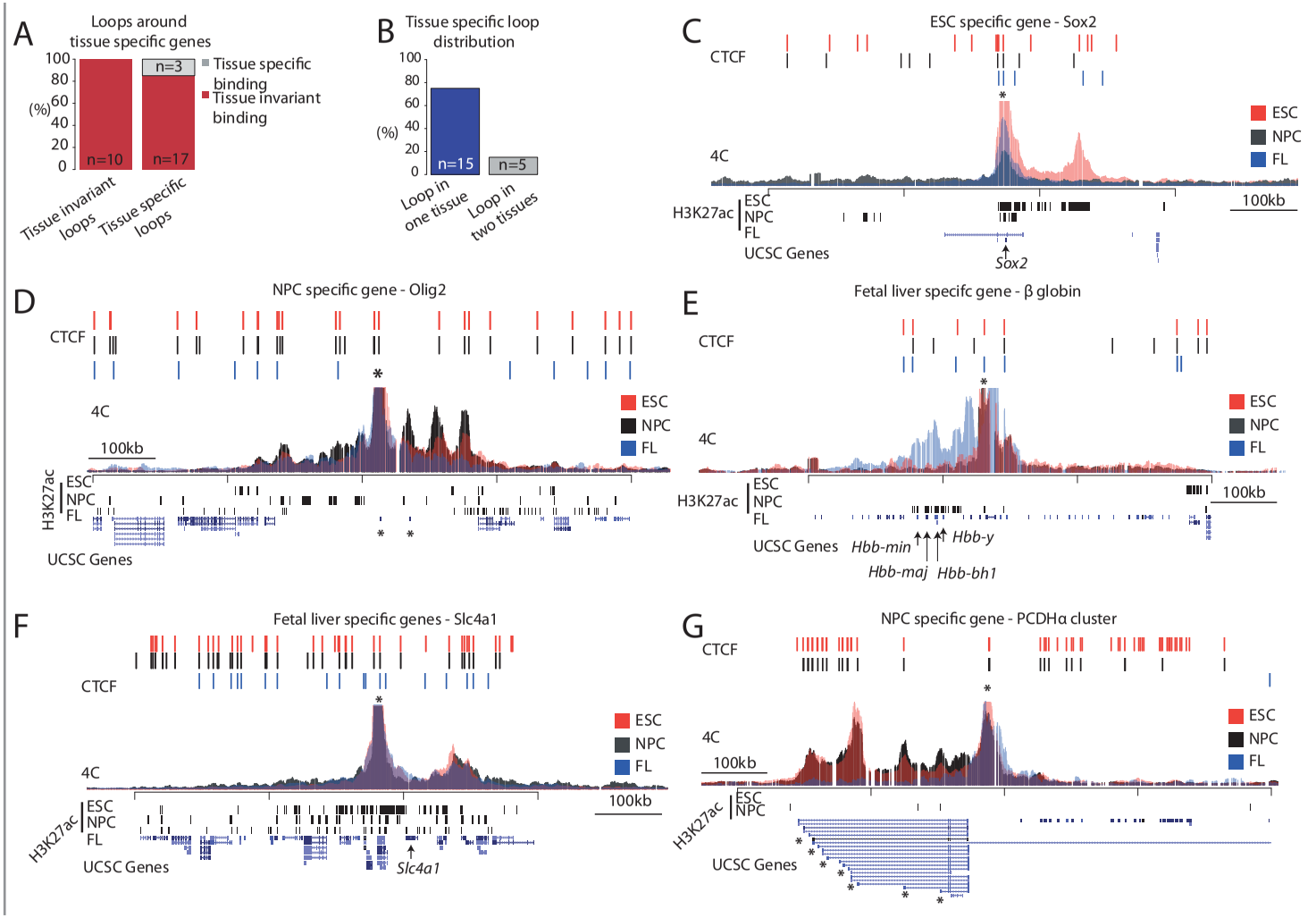


Figure 6 CTCF mediated loops at tissue specific gene loci act on enhancer promoter interactions. (A) Barplot showing the distribution of CTCF sites at conserved loops that are formed in all three tissues (green bar) versus CTCF sites found at loops that are

(continued from previous page) differential in at least one tissue (red bar). Numbers depict the number of CTCFbs found per category. (B) Barplot showing the percentage 0f tissue specific loops exclusive to one tissue (blue bar) and found in two tissues (grey bar). (C) 4C contact profile for a CTCFbs near the Sox2 gene for ESC (red), NPC (grey) and FL (blue) plotted on top of each other. Above the track, CTCF sites, per tissue, are plotted as vertical bars, the viewpoint is depicted by an asterisk. Below the profiles, a USCS genome browser track shows the peaks for H3K27ac in three tissues (from top to bottom: ESC, NPC and FL) are depicted in vertical bars, UCSC genes are plotted in blue bars. An asterisk below the tracks depicts the tissue specific gene. Vertical bars below the track are drawn every 200kb, total genomic length shown is 900kb. (D), (E), (F) and (G) Similar to (C) except that the contact profiles for different CTCFbs at tissue specifically expressed genes are shown: Olig2, the β-globin locus, Slc4a1 and protocadherin-α cluster, respectively.

Discussion

The role of CTCF in mediating long range interactions in the genome has been shown at some individual gene loci (reviewed in (35)) using 3C (19,20,38,44,45,51) and 4C technology (37,50). Genome-wide, chromatin interaction analysis with paired-end tag sequencing (ChIA-PET) has shown that CTCF binds at the basis of chromatin loops that mark transitions between opposing histone marks, separating active from inactive genes and bringing enhancer elements in close proximity with gene promoters (17). Furthermore, HiC studies have associated CTCF with boundaries across which chromatin flexibility is limited (15) and link CTCF to borders of topological domains (13). Here, we have used high-resolution 4C-seq (52) to profile in total 277 contact maps of CTCFbs in ESCs, NPCs, and FL and showed the involvement of CTCFbs in loops at single CTCFbs resolution.

Our data show that the majority (~70%) of CTCF sites are involved in long-range loops which means that genome-wide roughly 10.000-15.000 CTCF mediated loops potentially exist, assuming that no more than two CTCFbs mediate a loop. These numbers are comparable to the previously estimated number of CTCF mediated loops (17), but analyze the interactions in a more robust way. ChIA-PET analyses a limited number of ligation events between a CTCF mediated loops (5 to 100 read pairs per interaction) whereas in our 4C-seq experiments we analyzed 10.000 to 500.000 ligation events per experiment. ChIA-PET only identifies chromatin loops between the strongest binding sites, but here we show that also weaker binding sites can form loops, and that the level of occupancy only partially explains involvement in long-range contacts. ChIA-PET relies on the ChIP step and hence cannot analyse sites when they are no longer bound by the protein of interest. Differential binding sites can therefore not be followed during differentiation, nor can contacts be assessed upon depletion of the investigated protein. 4C-seq is not dependent on protein association and enabled us to analyse the consequence of CTCF binding in different tissues at conserved and differential CTCF binding sites.

Overlap in binding between CTCF and cohesin (21-24) suggested that at these sites both proteins could be involved in chromatin loops. We find that CTCF binding sites that are co-bound by cohesin prefer to interact with each other. Nevertheless, these sites also productively form loops with CTCF sites lacking cohesin. An important observation from our work is that tissue specific loop formation can be mediated from conserved CTCFbs. The tissue specific nature of these loop can be co-associated with enhancer sequences. This finding highly suggests that CTCF is involved in tissue transcriptional regulation through association with tissue specific enhancer sequences and perhaps with other tissue specific transcription factors. In agreement with this hypothesis, it has been reported that CTCFbs that a member of the transcription factor II D (TFIID), TAF3 (53) showed looping with promoters of TAF3 responsive genes. This looping behavior was shown to be dependent on TAF3 (53). CTCF has in fact been found co-associated with a plethora of transcription factors (54), fueling the idea that CTCF functions as a 'docking-site' (55) for other proteins to bind. Transcription factor accumulation at these sites may not only maintain or stabilize the protein-DNA interactions (28), but may also underlie loop formation of these sites. We are currently investigating this possibility by performing 4C-seq experiments on CTCFbs co-associated with ESC and fetal liver specific transcription factors. For factors that are dependent on CTCF for DNA binding this could mean that the interactions will require CTCF for their maintenance.

We assayed CTCF mediated loops at the *Pcdh*-α cluster, of which the genes are exclusively expressed in neuronal tissues. Here we found that loops are formed between the enhancer with several target promoters in ESCs and NPCs but not in fetal liver. Our results show that looping of regulatory sequences does not predict the transcriptional outcome of a gene, as was also shown for other gene loci (26,56). These 'preformed' loops in ESCs showed bivalent histone marks while in fetal liver CTCFbs involved in the chromatin loops seem to lack enhancer marks. This suggests that at the *Pcdh*-α locus, formation of loops is under the influence of the chromatin state of the locus. In the case of genes that show preformed loops between the promoter and its target enhancer, it will be interesting to ask what the molecular mechanism is that triggers the start of transcription after a cell has been differentiated to a state where transcription of that particular gene is required.

Material and methods

Cell culture and isolation of fetal liver tissue

E14Tg2A (E14) ES cells (IB10 cells) were cultured in BRL-conditioned Dulbecco's Modified Eagle Medium (DMEM with High Glucose, GlutaMAX™, Pyruvate; Life technologies) supplemented with 10% fetal calf serum (FCS), non-essential amino acids (NEAA) (Life technologies), 1000 U/ml leukemia inhibitory factor (LIF) and 2-mercaptoethanol. NPCs were kindly provided by B. van Steensel. NPCs were derived from E14 ESCs as described in (32) and were cultured on gelatin (0.1%) coated plates in DMEM/F12 medium (GIBCO-BRL) with addition of N2-supplement (life technologies), human basic FGF (20ng/ml; peprotech), murine EGF (20ng/ml; peprotech) and penicillin-streptomycin (P/S) (50U/ml - 50ug/ml) (Life technologies). WW6 ES cells, kindly provided by A. Skoultchi, were cultured on irradiated mouse embryonic fibroblasts in DMEM supplemented with 10% FCS, LIF, NEAA and P/S. Fetal livers (FL) were isolated from e14.5 embryos by dissection. Single cell suspensions for FL were obtained by filtration through a cell strainer (BD biosciences).

ChIP-reChIP sequencing

ChIP was performed as described in the Millipore protocol (www.millipore.com) using an antibody against CTCF (aCTCF) (Millipore; 07-729) with slight modifications. Briefly, cells were fixed in 1% formaldehyde/10% FCS/PBS for 10 minutes at room temperature (RT). Cells were lysed in cell lysis buffer (10 mM Tris-HCl pH 8.0; 10mM NaCl; 0.2% NP40; 10mM Na-Butyrate; Proteinase inhibitor (1X)) for 10 minutes at 4°C, subsequently nuclei were lysed in nuclei lysis buffer (50mM Tris-HCl pH 8.0; 10mM EDTA; 1% SDS; 10mM Na-butyrate; Proteinase inhibitor (1X)) for 10 minutes at 4°C. Crosslinked chromatin of 3-5x10^7 cells was sheared to a length of 500-800 base pairs using the Covaris S2 applying the following settings: duty cycle = 20%, intensity = 10, cycles per burst = 1000. Depending on the cell type, shearing was done for 10 to 15 minutes. After dilution of SDS to 0.15% using dilution buffer (0.01% SDS; 1.1% Triton X100; 1.2mM EDTA; 167mM NaCl; 16.7mM Tris-HCl, pH 8.0), chromatin was precipitated overnight (O/N) at 4°C, with 1µg aCTCF per million cells. The chromatin-αCTCF complex was precipitated using protein-G coupled agarose beads (Millipore), blocked in 0.1% BSA. After washing of the chromatin-beads complex, the chromatin was eluted from the beads in 1% SDS/0.1M $NaHCO_3$ for 15 minutes at RT. A second precipitation step was done after diluting the SDS to 0.15% using dilution buffer. Half the amount of antibody was used in the 2nd precipitation step, as described (33). Eluted chromatin was de-crosslinked for 6 hours at 65°C in the presence of 200mM NaCl, and treated with protK for 1 hour at 45°C. DNA was isolated using phenol extraction and ethanol precipitation.

Library preparation, SOLiD sequencing and mapping

Library preparation and SOLiD sequencing were done as described in (57). SOLiD sequence tags were mapped to the mm9 genome with the MAQ package (58) (V0.7.1, options: -c -n 3, -e 170).

Peak calling

MACS (59) was used to call peaks from the SOLiD data (V1.4.2, options: bw 300, -p 1e-10).
Tissue specific peaks used in the associative analysis (**Figure 2**) relative to genomic features are called above a threshold of 25 sequencing tags per peak and no peak called by MACS on the compared cell line. CTCF sites were extended with 300bps on each side to calculate the overlap between two tissues.
For calling differential peaks that were used as viewpoints in 4C analysis between E14 and WW6 ESCs (**Figure 4**) stringent criteria were used. Briefly, all peaks that had a tag count below 25 were not considered. Furthermore, a CTCFbs was only called tissue specific allowing zero reads in the cell type that was used in the comparison. Finally, all tissue specific CTCFbs were validated to be real by at least two independent ChIP-qPCR experiments (data not shown).

SNP calling on ESC specific CTCF sites

SNPs were called positive when present in two independently isolated samples of genomic DNA, by Sanger sequencing.

Association of CTCF sites with genomic features

For characterization of binding of CTCF sites relative to gene promoters and gene bodies the three CTCF data sets are compared to Entrez gene IDs. Promoters are defined as genomic regions 2500bp upstream and 500bp downstream of a transcription start site of a protein coding Entrez Gene ID. The gene body of a gene is defined as the transcription start site to the transcription end site of a gene. The rest of the genome is defined as intergenic.

4C template preparation

4C template was prepared as described in (52)with modifications. Briefly, cells were fixed using 2% formaldehyde/10%FCS/PBS for 10' at RT. Cells were lysed for 10' on ice (50mM Tris-HCl pH 7.5, 150mM NaCl, 5mM EDTA, 0.5% NP40 substitute, 1% Triton-X100, 1X proteinase inhibitors). Chromatin was digested in the context of the nucleus using the restriction enzymes DpnII or NlaIII (choice of restriction enzymes was dependent on the viewpoint) as first restriction enzyme, followed by ligation. Ligated chromatin was de-crosslinked in the presence of proteinase K at 65°C O/N and subsequently treated with RNase A. Genomic DNA was extracted by phenol/chloroform extraction and ethanol precipitation and subsequently digested O/N using the restriction enzymes Csp6I (in combination with 1st restriction enzymes DpnII or NlaIII) or using DpnII (combination with 1st restriction enzyme NlaIII) and ligated under diluted conditions (16°C, O/N) favoring intra-molecular ligations. Genomic precipitated and cleanup over Qiagen columns.

4C PCR, mapping and analysis of 4C data

The 4C PCR was performed with primers that are extended at the 5'-side with the single-end Illumina adaptors. Adaptor P5 is attached to the primers that read into the captured 4C fragments, this primer was designed on top of a restriction site (Fw) whereas the other 4C primers were extended with the Illumina P7 adaptor (Rev). High throughput sequencing is done on the Illumina Genome analyzer or Illumina HiSeq depending on the availability of the technique at the time. Reads were mapped, allowing no mismatches, to a database of 4C-seq fragment-ends generated from the mm9/NCBI m37 version of the mouse genome (52). Interaction profiles were calculated from the non-blind fragments (52)using a running mean with a window of 31 fragments. The profiles are normalized to the total amount of reads in cis.

Extraction of published data for CTCF, Rad21 and H3K27ac

ChIP sequencing data for CTCF in FL was taken from (46), GEO accession: GSE29218. The positions of CTCF sites covering the β-globin locus were based on previous ChIP-qPCR results . ChIP sequencing data for Rad21 in ESCs was taken from (40), GEO accession: GSE24030. ChIP sequencing data for the histone acetylation mark H3K27ac for ESC and NPC was taken from (41), GEO accession: GSE24165.

Acknowlegdements

We would like to thank B. Alako for peak calling of CTCFbs in ESCs and NPCs and G. Geeven for mapping part of the 4C data. M. Verstegen and P. Klous for excellent technical assistance.
This work was financially supported by grant no. 935170621 from the Dutch Scientific Organization (NWO) and a European Research Council Starting Grant (209700, '4C') to WdL.

References

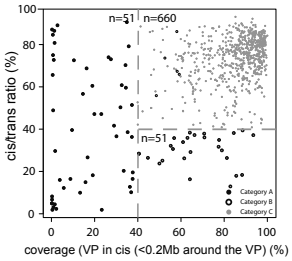
1. Klenova, E.M., Nicolas, R.H., Paterson, H.F., Carne, A.F., Heath, C.M., Goodwin, G.H., Neiman, P.E. and Lobanenkov, V.V. (1993) CTCF, a conserved nuclear factor required for optimal transcriptional activity of the chicken c-myc gene, is an 11-Zn-finger protein differentially expressed in multiple forms. *Molecular and cellular biology*, **13**, 7612-7624.
2. Filippova, G.N., Fagerlie, S., Klenova, E.M., Myers, C., Dehner, Y., Goodwin, G., Neiman, P.E., Collins, S.J. and Lobanenkov, V.V. (1996) An exceptionally conserved transcriptional repressor, CTCF, employs different combinations of zinc fingers to bind diverged promoter sequences of avian and mammalian c-myc oncogenes. *Molecular and cellular biology*, **16**, 2802-2813.
3. Lobanenkov, V.V., Nicolas, R.H., Adler, V.V., Paterson, H., Klenova, E.M., Polotskaja, A.V. and 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**, 1743-1753.
4. Bell, A.C., West, A.G. and Felsenfeld, G. (1999) The protein CTCF is required for the enhancer blocking activity of vertebrate insulators. *Cell*, **98**, 387-396.
5. Farrell, C.M., West, A.G. and Felsenfeld, G. (2002) Conserved CTCF insulator elements flank the mouse and human beta-globin loci. *Molecular and cellular biology*, **22**, 3820-3831.
6. Recillas-Targa, F., Pikaart, M.J., Burgess-Beusse, B., Bell, A.C., Litt, M.D., West, A.G., Gaszner, M. and Felsenfeld, G. (2002) Position-effect protection and enhancer blocking by the chicken beta-globin insulator are separable activities. *Proceedings of the National Academy of Sciences of the United States of America*, **99**, 6883-6888.
7. Barski, A., Cuddapah, S., Cui, K., Roh, T.Y., Schones, D.E., Wang, Z., Wei, G., Chepelev, I. and Zhao, K. (2007) High-resolution profiling of histone methylations in the human genome. *Cell*, **129**, 823-837.
8. Chen, X., Xu, H., Yuan, P., Fang, F., Huss, M., Vega, V.B., Wong, E., Orlov, Y.L., Zhang, W., Jiang, J. *et al.* (2008) Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. *Cell*, **133**, 1106-1117.
9. Kim, T.H., Abdullaev, Z.K., Smith, A.D., Ching, K.A., Loukinov, D.I., Green, R.D., Zhang, M.Q., Lobanenkov, V.V. and Ren, B. (2007) Analysis of the vertebrate insulator protein CTCF-binding sites in the human genome. *Cell*, **128**, 1231-1245.
10. Schmidt, D., Schwalie, P.C., Wilson, M.D., Ballester, B., Goncalves, A., Kutter, C., Brown, G.D., Marshall, A., Flicek, P. and Odom, D.T. (2012) Waves of retrotransposon expansion remodel genome organization and CTCF binding in multiple mammalian lineages. *Cell*, **148**, 335-348.

11. Cuddapah, S., Jothi, R., Schones, D.E., Roh, T.Y., Cui, K. and 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**, 24-32.
12. Guelen, L., Pagie, L., Brasset, E., Meuleman, W., Faza, M.B., Talhout, W., Eussen, B.H., de Klein, A., Wessels, L., de Laat, W. *et al.* (2008) Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions. *Nature*, **453**, 948-951.
13. Dixon, J.R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., Hu, M., Liu, J.S. and Ren, B. (2012) Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, **485**, 376-380.
14. Lieberman-Aiden, E., van Berkum, N.L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B.R., Sabo, P.J., Dorschner, M.O. *et al.* (2009) Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, **326**, 289-293.
15. Yaffe, E. and Tanay, A. (2011) Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture. *Nat Genet*, **43**, 1059-1065.
16. Nora, E.P., Lajoie, B.R., Schulz, E.G., Giorgetti, L., Okamoto, I., Servant, N., Piolot, T., van Berkum, N.L., Meisig, J., Sedat, J. *et al.* (2012) Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature*, **485**, 381-385.
17. Handoko, L., Xu, H., Li, G., Ngan, C.Y., Chew, E., Schnapp, M., Lee, C.W., Ye, C., Ping, J.L., Mulawadi, F. *et al.* (2011) CTCF-mediated functional chromatin interactome in pluripotent cells. *Nat Genet*, **43**, 630-638.
18. Sanyal, A., Lajoie, B.R., Jain, G. and Dekker, J. (2012) The long-range interaction landscape of gene promoters. *Nature*, **489**, 109-113.
19. Splinter, E., Heath, H., Kooren, J., Palstra, R.J., Klous, P., Grosveld, F., Galjart, N. and de Laat, W. (2006) CTCF mediates long-range chromatin looping and local histone modification in the beta-globin locus. *Genes & development*, **20**, 2349-2354.
20. Kurukuti, S., Tiwari, V.K., Tavoosidana, G., Pugacheva, E., Murrell, A., Zhao, Z., Lobanenkov, V., Reik, W. and Ohlsson, R. (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**, 10684-10689.
21. Wendt, K.S., Yoshida, K., Itoh, T., Bando, M., Koch, B., Schirghuber, E., Tsutsumi, S., Nagae, G., Ishihara, K., Mishiro, T. *et al.* (2008) Cohesin mediates transcriptional insulation by CCCTC-binding factor. *Nature*, **451**, 796-801.
22. Rubio, E.D., Reiss, D.J., Welcsh, P.L., Disteche, C.M., Filippova, G.N., Baliga, N.S., Aebersold, R., Ranish, J.A. and Krumm, A. (2008) CTCF physically links cohesin to chromatin. *Proceedings of the National Academy of Sciences of the United States of America*, **105**, 8309-8314.
23. Parelho, V., Hadjur, S., Spivakov, M., Leleu, M., Sauer, S., Gregson, H.C., Jarmuz, A., Canzonetta, C., Webster, Z., Nesterova, T. *et al.* (2008) Cohesins functionally associate with CTCF on mammalian chromosome arms. *Cell*, **132**, 422-433.
24. Stedman, W., Kang, H., Lin, S., Kissil, J.L., Bartolomei, M.S. and Lieberman, P.M. (2008) Cohesins localize with CTCF at the KSHV latency control region and at cellular c-myc and H19/Igf2 insulators. *The EMBO journal*, **27**, 654-666.
25. Faure, A.J., Schmidt, D., Watt, S., Schwalie, P.C., Wilson, M.D., Xu, H., Ramsay, R.G., Odom, D.T. and Flicek, P. (2012) Cohesin regulates tissue-specific expression by stabilizing highly occupied cis-regulatory modules. *Genome research*.
26. Demare, L.E., Leng, J., Cotney, J., Reilly, S.K., Yin, J., Sarro, R. and Noonan, J.P. (2013) The genomic landscape of cohesin-associated chromatin interactions. *Genome research*.
27. Nasmyth, K. and Haering, C.H. (2009) Cohesin: its roles and mechanisms. *Annual review of genetics*, **43**, 525-558.
28. Hadjur, S., Williams, L.M., Ryan, N.K., Cobb, B.S., Sexton, T., Fraser, P., Fisher, A.G. and Merkenschlager, M. (2009) Cohesins form chromosomal cis-interactions at the developmentally regulated IFNG locus. *Nature*, **460**, 410-413.
29. Seitan, V.C., Hao, B., Tachibana-Konwalski, K., Lavagnolli, T., Mira-Bontenbal, H., Brown, K.E., Teng, G., Carroll, T., Terry, A., Horan, K. *et al.* (2011) A role for cohesin in T-cell-receptor rearrangement and thymocyte differentiation. *Nature*, **476**, 467-471.
30. Kagey, M.H., Newman, J.J., Bilodeau, S., Zhan, Y., Orlando, D.A., van Berkum, N.L., Ebmeier, C.C., Goossens, J., Rahl, P.B., Levine, S.S. *et al.* (2010) Mediator and cohesin connect gene expression and chromatin architecture. *Nature*, **467**, 430-435.
31. Schmidt, D., Schwalie, P.C., Ross-Innes, C.S., Hurtado, A., Brown, G.D., Carroll, J.S., Flicek, P. and Odom, D.T. (2010) A CTCF-independent role for cohesin in tissue-specific transcription. Genome research, 20, 578-588.
32. Ying, Q.L. and Smith, A.G. (2003) Defined conditions for neural commitment and differentiation. Methods in enzymology, 365, 327-341.
33. Hatzis, P., van der Flier, L.G., van Driel, M.A., Guryev, V., Nielsen, F., Denissov, S., Nijman, I.J., Koster, J., Santo, E.E., Welboren, W. et al. (2008) Genome-wide pattern of TCF7L2/TCF4 chromatin occupancy in colorectal cancer cells. Molecular and cellular biology, 28, 2732-2744.
34. Essien, K., Vigneau, S., Apreleva, S., Singh, L.N., Bartolomei, M.S. and Hannenhalli, S. (2009) CTCF binding site classes exhibit distinct evolutionary, genomic, epigenomic and transcriptomic features. Genome biology, 10, R131.
35. Phillips, J.E. and Corces, V.G. (2009) CTCF: master weaver of the genome. Cell, 137, 1194-1211.

36. Murrell, A., Heeson, S. and Reik, W. (2004) Interaction between differentially methylated regions partitions the imprinted genes Igf2 and H19 into parent-specific chromatin loops. Nature genetics, 36, 889-893.
37. Ribeiro de Almeida, C., Stadhouders, R., de Bruijn, M.J., Bergen, I.M., Thongjuea, S., Lenhard, B., van Ijcken, W., Grosveld, F., Galjart, N., Soler, E. et al. (2011) The DNA-binding protein CTCF limits proximal Vkappa recombination and restricts kappa enhancer interactions to the immunoglobulin kappa light chain locus. Immunity, 35, 501-513.
38. Majumder, P. and Boss, J.M. (2010) CTCF controls expression and chromatin architecture of the human major histocompatibility complex class II locus. Molecular and cellular biology, 30, 4211-4223.
39. van de Werken, H.J., de Vree, P.J., Splinter, E., Holwerda, S.J., Klous, P., de Wit, E. and de Laat, W. (2012) 4C technology: protocols and data analysis. Methods in enzymology, 513, 89-112.
40. Nitzsche, A., Paszkowski-Rogacz, M., Matarese, F., Janssen-Megens, E.M., Hubner, N.C., Schulz, H., de Vries, I., Ding, L., Huebner, N., Mann, M. et al. (2011) RAD21 cooperates with pluripotency transcription factors in the maintenance of embryonic stem cell identity. PloS one, 6, e19470.
41. Creyghton, M.P., Cheng, A.W., Welstead, G.G., Kooistra, T., Carey, B.W., Steine, E.J., Hanna, J., Lodato, M.A., Frampton, G.M., Sharp, P.A. et al. (2010) Histone H3K27ac separates active from poised enhancers and predicts developmental state. Proceedings of the National Academy of Sciences of the United States of America, 107, 21931-21936.
42. Hark, A.T., Schoenherr, C.J., Katz, D.J., Ingram, R.S., Levorse, J.M. and Tilghman, S.M. (2000) CTCF mediates methylation-sensitive enhancer-blocking activity at the H19/Igf2 locus. Nature, 405, 486-489.
43. Heath, H., Ribeiro de Almeida, C., Sleutels, F., Dingjan, G., van de Nobelen, S., Jonkers, I., Ling, K.W., Gribnau, J., Renkawitz, R., Grosveld, F. et al. (2008) CTCF regulates cell cycle progression of alphabeta T cells in the thymus. The EMBO journal, 27, 2839-2850.
44. Hou, C., Dale, R. and Dean, A. (2010) Cell type specificity of chromatin organization mediated by CTCF and cohesin. Proceedings of the National Academy of Sciences of the United States of America, 107, 3651-3656.
45. Junier, I., Dale, R.K., Hou, C., Kepes, F. and Dean, A. (2012) CTCF-mediated transcriptional regulation through cell type-specific chromosome organization in the beta-globin locus. Nucleic acids research.
46. Shen, Y., Yue, F., McCleary, D.F., Ye, Z., Edsall, L., Kuan, S., Wagner, U., Dixon, J., Lee, L., Lobanenkov, V.V. et al. (2012) A map of the cis-regulatory sequences in the mouse genome. Nature.
47. Chien, R., Zeng, W., Kawauchi, S., Bender, M.A., Santos, R., Gregson, H.C., Schmiesing, J.A., Newkirk, D.A., Kong, X., Ball, A.R., Jr. et al. (2011) Cohesin mediates chromatin interactions that regulate mammalian beta-globin expression. The Journal of biological chemistry, 286, 17870-17878.
48. Ren, L., Wang, Y., Shi, M., Wang, X., Yang, Z. and Zhao, Z. (2012) CTCF mediates the cell-type specific spatial organization of the Kcnq5 locus and the local gene regulation. PloS one, 7, e31416.
49. Hirayama, T., Tarusawa, E., Yoshimura, Y., Galjart, N. and Yagi, T. (2012) CTCF is required for neural development and stochastic expression of clustered Pcdh genes in neurons. Cell reports, 2, 345-357.
50. Guo, Y., Monahan, K., Wu, H., Gertz, J., Varley, K.E., Li, W., Myers, R.M., Maniatis, T. and Wu, Q. (2012) CTCF/cohesin-mediated DNA looping is required for protocadherin alpha promoter choice. Proceedings of the National Academy of Sciences of the United States of America, 109, 21081-21086.
51. Guo, C., Yoon, H.S., Franklin, A., Jain, S., Ebert, A., Cheng, H.L., Hansen, E., Despo, O., Bossen, C., Vettermann, C. et al. (2011) CTCF-binding elements mediate control of V(D)J recombination. Nature, 477, 424-430.
52. van de Werken, H.J., Landan, G., Holwerda, S.J., Hoichman, M., Klous, P., Chachik, R., Splinter, E., Valdes-Quezada, C., Oz, Y., Bouwman, B.A. et al. (2012) Robust 4C-seq data analysis to screen for regulatory DNA interactions. Nature methods, 9, 969-972.
53. Liu, Z., Scannell, D.R., Eisen, M.B. and Tjian, R. (2011) Control of embryonic stem cell lineage commitment by core promoter factor, TAF3. Cell, 146, 720-731.
54. Holwerda, S.J., van de Werken, H.J., Ribeiro de Almeida, C., Bergen, I.M., de Bruijn, M.J., Verstegen, M.J., Simonis, M., Splinter, E., Wijchers, P.J., Hendriks, R.W. et al. (2013) Allelic exclusion of the immunoglobulin heavy chain locus is independent of its nuclear localization in mature B cells. Nucleic acids research.
55. Martinez, S.R. and Miranda, J.L. (2010) CTCF terminal segments are unstructured. Protein science : a publication of the Protein Society, 19, 1110-1116.
56. Eijkelenboom, A., Mokry, M., de Wit, E., Smits, L.M., Polderman, P.E., van Triest, M.H., van Boxtel, R., Schulze, A., de Laat, W., Cuppen, E. et al. (2013) Genome-wide analysis of FOXO3 mediated transcription regulation through RNA polymerase II profiling. Molecular systems biology, 9, 638.
57. Mokry, M., Hatzis, P., de Bruijn, E., Koster, J., Versteeg, R., Schuijers, J., van de Wetering, M., Guryev, V., Clevers, H. and Cuppen, E. (2010) Efficient double fragmentation ChIP-seq provides nucleotide resolution protein-DNA binding profiles. PloS one, 5, e15092.
58. Li, H., Ruan, J. and Durbin, R. (2008) Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome research, 18, 1851-1858.
59. Zhang, Y., Liu, T., Meyer, C.A., Eeckhoute, J., Johnson, D.S., Bernstein, B.E., Nusbaum, C., Myers, R.M., Brown, M., Li, W. et al. (2008) Model-based analysis of ChIP-Seq (MACS). Genome biology, 9, R137.

5

Supplemental Figures & tables



Supplemental Figure 1. 4Cseq quality scores.
Quality control plot 52 for 4Cseq profiles that were performed using four cutter restriction enzymes in the first and second restriction step by the author of this thesis over the course of 5 years at the Hubrecht Institute. The X-axis depicts the coverage of the restriction frag-ends in a window of 0.2Mb around the viewpoint (0.1Mb +/- the 4C viewpoint). The Y-axis depicts the cis/trans ratios of the uniquely mapped frag-ends. Experiments that categorized as "C" (gray) are of high quality and are can be used for scientific purposes, the rest of the experiments should not be considered for such purposes.

Category A	reads	unique reads	unique reads in cis	coverage cis (%)	cis/trans ratio (%)
average	414,197	56,968	14,499	17	41
min.	80	13	11	0	2
max.	1,220,318	317,484	148,290	38	91
total	21,124,056	2,905,380	739,467		
Category B					
average	693,169	141,802	42,082	67	30
min.	38,212	3,740	2,518	41	13
max.	1,507,997	514,431	162,953	94	80
total	18,715,571	3,828,645	1,136,226		
Category C					
average	723,984	226,957	169,384	86	74
min.	36,537	7,979	5,236	40	41
max.	3,689,502	1,332,996	1,020,162	100	93
total	477,829,248	149,791,460	111,793,594		

Category	Conditions
A	cis coverage < 40%
B	cis coverage > 40% AND cis/trans < 40% OR # cis uniq reads< 500
C	cis coverage AND cis/trans > 40%

Supplemental table 1. Sequencing reads and coverage of 4C experiments.
Numbers of bases sequenced, coverage in cis and cis/trans ratio's for all 4C experiments performed using four cutter restriction enzymes in the first and second restriction step by the author of this thesis over the course of 5 years at the Hubrecht Institute.

Chromosome	Start position viewpoint	Gene	tissue	Differential binding site (y/n)	Loop (y/n)	Differential loop (y/n)	Loop exclusive to one tissue	Loop exclusive to two tissues	Loop conserved in specific tissue ESC&NPC (EN); ESC&FL (EF); NPC&FL (NF)
3	34548270	Sox2 #1	ESC	n	y	y	1	0	E
3	34548270	Sox2 #1	NPC	n	n	y	1	0	E
3	34548270	Sox2 #1	FL	n	n	y	1	0	E
3	34658794	Sox2 #2	ESC	n	y	y	0	1	EN
3	34658794	Sox2 #2	FL	n	n	y	0	1	EN
3	34660483	Sox2 #2	NPC	n	y	y	0	1	EN
4	55527980	KLF4 #1	ESC	n	y	n	0	0	All
4	55527980	KLF4 #1	NPC	n	y	n	0	0	All
4	55527980	KLF4 #1	FL	n	y	n	0	0	All
4	55543823	KLF4 #2	ESC	n	y	y	0	1	EF
4	55543823	KLF4 #2	FL	n	y	y	0	1	EF
4	55543823	KLF4 #2	NPC	n	n	y	0	1	EF
4	134435805	Rhd #1	ESC	n	n	y	1	0	F
4	134435805	Rhd #1	NPC	n	n	y	1	0	F
4	134435814	Rhd #1	FL	n	y	y	1	0	F
4	134502243	Rhd #2	ESC	n	y	n	0	0	All
4	134502243	Rhd #2	FL	n	y	n	0	0	All
4	134502243	Rhd #2	NPC	n	y	n	0	0	All
6	122583758	Dppa3 #1	ESC	y	y	y	1	0	E
6	122583758	Dppa3 #1	FL	y	n	y	1	0	E
6	122583758	Dppa3 #1	NPC	y	n	y	1	0	E
6	122618964	Dppa3 #2	ESC	n	y	y	1	0	E
6	122618964	Dppa3 #2	FL	n	n	y	1	0	E
6	122618964	Dppa3 #2	NPC	n	n	y	1	0	E
6	122670095	Nanog	ESC	n	y	y	0	1	EN
6	122670095	Nanog	NPC	n	y	y	0	1	EN
6	122670095	Nanog	FL	n	n	y	0	1	EN
7	111061651	B-globin_hs-62.5	FL	y	y	y	1	0	F
7	111061651	B-globin_hs-62.5	ESC	y	n	y	1	0	F
7	111061651	B-globin_hs-62.5	NPC	y	n	y	1	0	F
8	87425935	Klf1 #1	ESC	n	y	n	0	0	All
8	87425935	Klf1 #1	FL	n	y	n	0	0	All
8	87425935	Klf1 #1	NPC	n	y	n	0	0	All
8	87436702	Klf1 #2	FL	n	y	y	1	0	F
8	87436702	Klf1 #2	ESC	n	n	y	1	0	F
8	87436702	Klf1 #2	NPC	n	n	y	1	0	F
8	87500029	Klf1 #3	FL	n	y	y	1	0	F
8	87500029	Klf1 #3	ESC	n	n	y	1	0	F
8	87500029	Klf1 #3	NPC	n	n	y	1	0	F
8	125946511	Tubb3	ESC	n	y	n	0	0	All
8	125946511	Tubb3	FL	n	y	n	0	0	All
8	125946511	Tubb3	NPC	n	y	n	0	0	All
10	57463988	Fabp7 #1	NPC	n	y	y	1	0	N
10	57463988	Fabp7 #1	ESC	n	n	y	1	0	N
10	57463988	Fabp7 #1	FL	n	n	y	1	0	N

Supplemental table 2. Additional information to 4C experiments for CTCF sites located around tissue specific genes.
Table depicting certain characteristics of CTCFbs used as 4C viewpoints to analyse chromatin interactions of CTCFbs around tissue specific genes in figure 6.

Supplemental table 2 - continued from previous page

Chromosome	Start position viewpoint	Gene	tissue	Differential binding site (y/n)	Loop (y/n)	Differential loop (y/n)	Loop exclusive to one tissue	Loop exclusive to two tissues	Loop conserved in specific tissue ESC&NPC (EN); ESC&FL (EF); NPC&FL (NF)
10	57515889	Fabp7 #2	NPC	n	y	y	1	0	N
10	57515889	Fabp7 #2	ESC	n	n	y	1	0	N
10	57515889	Fabp7 #2	FL	n	n	y	1	0	N
10	57568589	Fabp7 #3	ESC	n	y	n	0	0	All
10	57568589	Fabp7 #3	FL	n	y	n	0	0	All
10	57568589	Fabp7 #3	NPC	n	y	n	0	0	All
11	32138087	alpha Globin #1	FL	n	y	y	1	0	F
11	32138087	alpha Globin #1	ESC	n	n	y	1	0	F
11	32138087	alpha Globin #1	NPC	n	n	y	1	0	F
11	32224410	alpha Globin #2	FL	n	y	y	1	0	F
11	32224410	alpha Globin #2	ESC	n	n	y	1	0	F
11	32224410	alpha Globin #2	NPC	n	n	y	1	0	F
11	102160656	Slc4a1 #1	ESC	n	y	n	0	0	All
11	102160656	Slc4a1 #1	FL	n	y	n	0	0	All
11	102160656	Slc4a1 #1	NPC	n	y	n	0	0	All
11	102273033	Slc4a1 #2	FL	n	y	y	1	0	F
11	102273033	Slc4a1 #2	ESC	n	n	y	1	0	F
11	102273033	Slc4a1 #2	NPC	n	n	y	1	0	F
11	102755095	Gfap #1	FL	n	y	n	0	0	All
11	102755095	Gfap #1	ESC	n	y	n	0	0	All
11	102755095	Gfap #1	NPC	n	y	n	0	0	All
11	102762586	Gfap #2	FL	y	y	n	0	0	All
11	102762586	Gfap #2	ESC	y	y	n	0	0	All
11	102762586	Gfap #2	NPC	y	y	n	0	0	All
11	102854422	Gfap #3	NPC	n	y	y	1	0	N
11	102854422	Gfap #3	ESC	n	n	y	1	0	N
11	102854422	Gfap #3	FL	n	n	y	1	0	N
15	8577349	Slc1a3	NPC	n	y	y	0	1	EN
15	8577349	Slc1a3	ESC	n	y	y	0	1	EN
15	8577349	Slc1a3	FL	n	n	y	0	1	EN
16	91224690	Olig2 #1	NPC	n	y	y	1	0	N
16	91224690	Olig2 #1	ESC	n	n	y	1	0	N
16	91224690	Olig2 #1	FL	n	n	y	1	0	N
16	91314856	Olig2 #2	NPC	n	y	y	1	0	N
16	91314856	Olig2 #2	ESC	n	n	y	1	0	N
16	91314856	Olig2 #2	FL	n	n	y	1	0	N
17	50376256	Dazl #1	NPC	n	y	n	0	0	All
17	50376256	Dazl #1	ESC	n	y	n	0	0	All
17	50376256	Dazl #1	FL	n	y	n	0	0	All
17	50430446	Dazl #2	ESC	n	y	n	0	0	All
17	50430446	Dazl #2	FL	n	y	n	0	0	All
17	50430446	Dazl #2	NPC	n	y	n	0	0	All
18	37377867	Pcdh-a	ESC	y	y	y	0	1	EN
18	37377867	Pcdh-a	NPC	y	y	y	0	1	EN
18	37377867	Pcdh-a	FL	y	n	y	0	1	EN

6
General discussion

General discussion

Sjoerd J.B. Holwerda, Wouter de Laat

Introduction

The subject of investigation in this thesis is a very stable 'molecule' that resides in the nucleus of a cell, DNA. We distinguish eu- and heterochromatin in the cell nucleus and know that active genes mostly reside in euchromatin whereas inactive genes are mostly found in heterochromatic DNA. Interestingly, the position of euchromatin and heterochromatin seems to be non-random: Heterochromatic DNA preferentially resides at the edges of the nucleus and around the centromeric regions of the chromosomes, whereas euchromatic DNA adopts a more central position in the nucleus. At the level of individual chromosomes a non-random distribution was reported for gene poor chromosomes versus gene rich chromosomes following a similar pattern: gene rich chromosomes preferentially reside in the interior of the nucleus while gene poor chromosomes adopt a more peripheral position in the nucleus (1-3). These observations suggest that the position of a chromosome could be related to the activity of its coding elements, genes, and thus to the function of the genome and form the basis of a model where activity of genes is associated with their position. However, to what extent transcriptionally favourable or restrictive 'compartments' in the genome dictate or determine the expression of genes is currently subject of heavy debates and investigations.

Gene activity in relation to its position in the nucleus

Several examples have been described where activity of genes is correlated with their position in the nucleus (4,5). For instance, the constitutively expressed leukocyte marker CD45 and the B cell specific gene CD19 both adopt a position away from heterochromatin associated regions in B cells, whereas T cell specific genes CD4 and CD8 which are not expressed in B lymphocytes are found to associate with these heterochromatic regions (4). The phenomenon where inactive genes are associated with heterochromatin was shown to be a dynamic process. Upon mitogenic stimulation of resting B and T cells inactive genes could be repositioned to centromeric regions in a tissue specific manner whereas active genes did not (5).
An interesting question arising from the observations made in these studies is:
Does relocation of a gene to a transcriptionally repressive environment dictate transcriptional silencing?
The answer is, counter-intuitively, no.

Relocation of genomic loci from an 'active' to an 'inactive' environment can enable transcriptional silencing (6,7) but is not necessarily sufficient for silencing (7,8). Experiments where a locus was tethered to the nuclear lamina showed that for a gene dense region, within a window of 1Mb, a few genes were down regulated. On the other hand, transcription of many other genes on the same stretch of DNA was unperturbed (7). In another study, relocation of a reporter construct under the control of a Tet-inducible minimal CMV promoter showed that after tethering this construct to the lamina, Tet-inducible transcription could still take place (8). The model that emerges from these studies is that the environment where a gene resides in can have an influence on the transcriptional output of a gene but does not define its transcriptional state. The work described in Chapter 3 of this book supports this model. For the IgH locus, we found that although not transcribed in T cells, the locus does not occupy an inactive chromatin

environment, which is the case for this locus in fetal brain tissue. Rather, in T cells the IgH locus resides in a nearly identical genomic environment as its highly active equivalent in B cells. Thus, we show that transcription can be uncoupled from nuclear location. Uncoupling of transcription from genome structure was also found at the inactive X-chromosome (9). A non-coding RNA Xist coats the inactive X-chromosome and influences its topology and transcription (10-12). Upon deletion of Xist, the topology of the inactive X chromosome adopts a more 'open' conformation, resembling that of the active X chromosome, but transcription of the genes at the inactive X chromosome is kept silent (9). Thus, while adopting an 'open' chromatin state, transcription of the genes residing in this open conformation is not activated (9,13). We propose that relocation of a gene locus to repressive or active environments can be a mechanism by which a cell reinforces silencing or activation of genes but that association with these nuclear sub structures or volumes is not sufficient or needed for the activation or silencing of genes.

The role of long-range interactions in transcription

In general interactions between active genes and between inactive genes are preferred over interactions between active and inactive genes (14,15). An open question remains whether there is a functional reason for particular active or inactive genes to come together in nuclear space.

There are claims in the literature that enhancers on one chromosome can interact with and activate target genes located on different chromosomes (16,17). Olfactory receptor promoters were, for instance, found to form inter-chromosomal interactions with an enhancer element (16) that supposedly controlled the transcription of many olfactory receptor genes in trans. Deletion of the particular enhancer in a subsequent study however did not affect the expression of any of the genes in trans, but was only of influence of genes located in the direct vicinity of the H-enhancer, on the same chromosome (18). This demonstrated that trans-activation is not a mechanism for olfactory receptor gene expression control. In an experiment in our lab, an ectopic human β-globin enhancer (hLCR) integrated on chromosome 8 was found to be able to contact its natural target gene on chromosome 7 in a small percentage (2-3%) of cells. As a consequence an upregulation of this gene was observed only in this particular (small) fraction of cells where the inter-chromosomal interaction occurs, in contrast to cells where this interaction was not found (19). This exemplifies that inter-chromosomal interactions between regulatory elements can occur within the nucleus but that they occur with relatively low frequency compared to cis-interactions. This phenomenom is also what we expect based on analysis of 4C data: Local interactions are favoured over long-range cis interactions, whereas trans-interactions are sparsely present.

Having said this, trans-interactions made between different genes remain a popular subject to study in relation to transcription regulation. It has been observed that tissue specific, co-expressed genes co-localize at locations in the nucleus and are transcribed in 'hot-spots' where a tissue specific transcription factor, Klf1, accumulates (20). This study showed that the interactions and transcription of the interacting genes were dependent on Klf1, suggesting that these genes specifically come together at 'Klf1 hotspots', to be transcribed. Whether, in general interactions are dependent on transcription is questionable. A 4C study showed that both local interactions within the β-globin locus and the global nuclear environment of the locus remained unchanged upon inhibition of transcription (21). These two studies seem to contradict each other. In one case transcription is coupled to genome structure (20), where in the other study maintenance of the structure does not require ongoing transcription (21). In the first case an enrichment step for RNA polymerase II was done, selecting for active genes. Whether and to what extent inter- and intra-chromosomal contacts between genes underly the subsequent transcription of these genes awaits further investigation.

Although not necessarily needed for transcription, intra and inter-chromosomal interactions seem to require tissue specific transcription factors (TFs) (20,22) and importantly some TFs can play a direct role in the formation of long-range DNA interactions (22). Notably, not all TF binding sites are involved in setting up a genome-wide network between TF bound sites. Specifically for Nanog, an embryonic stem cell factor, clusters characterized by high Nanog binding density form such interactions (22). Potentially, the high degree of protein binding is underlined by other factors such as active histone marks and high gene density. This extremely 'open' conformation at certain sites in the genome could provide a high degree of chromatin flexibility to these sites compared to their surroundings. In this way, within the pool of active chromatin, specific sites in the genome could be characterized by an extra high degree of chromatin flexibility. These 'ultra'-sensitive sites potentially have an increased probability or affinity to interact with other 'ultra'-sensitive sites in the genome and thus could be able to interact with each other in intra- and inter-chromosomal interactions. The need of these interactions for transcription to occur, again, remains elusive.

In general, we believe that nuclear organization and interactions between genomic regions made in the nucleus are mostly determined by the probability of a locus or chromatin region to interact with other gene loci. This probability is highly related to the chromatin state of surrounding sequences.
On a genome wide level, active regions preferentially interact with active regions whereas inactive regions interact preferentially with inactive regions. This general feature found in 3C-based technologies reflects microscopic observations where heterochromatic regions cluster at the lamina and around centromeres and active regions reside in the centre of the nucleus. The fact that in general between all inactive regions and all active regions interactions are found is a consequence of the population based average that is measured in 4C and HiC assays. We believe that very specific long-range and inter-chromosomal interactions directly related to transcription are rare and that the chromatin environment of a locus determines the freedom of a certain sequence to roam the nuclear interior, not vice versa (23).

Local cis-interactions between regulatory sites and gene promoters at specific gene loci often do find each other efficiently and potentially in every cell (24). At the β-globin locus the locus control region (LCR) can enhance expression of the globin genes. Hyper sensitive sites present at the LCR are shown be involved in setting up an active chromatin hub (25) which encompasses the LCR, the globin genes and other regulatory elements surrounding the locus (25-29). Expression of the globin genes and concomitant looping of the LCR to the genes in the locus is dependent on GATA-1, an erythroid specific TF (30). Moreover, transcription of the β-globin genes could be induced by mere looping of the LCR to the β-major genes, in GATA1$^{-/-}$ cells (24). Here, tethering of the LCR to the β-major gene in the globin locus using zinc-fingers was required for the expression of the genes in the locus. This study showed, for the first time, that transcription was caused by the formation of a mere loop (24). In this particular case, mere coming together of regulatory elements was required for transcription of the target genes. The need for interaction of regulatory sites with their target promoters, in cis, to upregulate transcription is expected to be a general mechanism valid for more loci. However, loops between regulatory sites and promoters does not necessarily result in transcription. In some cases the structure between enhancers and promoters seems to be preformed, in the absence of transcription (31) perhaps to increase the efficiency of transcription of target genes to a certain stimulus.

4C-seq, ChIA-PET, HiC and future perspectives

To understand the role of local chromatin loops in the genome we have studied the behaviour of a general looping factor in the genome, CTCF, using 4C-seq. This DNA binding protein was shown to be involved in mediating long-range loops between its binding sites genome wide by Chromatin interaction analysis with paired-end tag sequencing (ChIA-PET) (32). A limitation of ChIA-PET compared to 4C-seq is that it only

allows studying sites in the genome bound by a protein, due to the chromatin immunoprecipitation step involved (32). 4C-seq however, allowed us to compare the consequences for DNA topology in different tissues in the presence and absence of CTCF. From these experiments, we conclude that tissue specific CTCFbs can be associated with differential chromatin loops between different tissues, something that would not have been possible using ChIA-PET. Another advantage of 4Cseq over ChIA-PET is the depth of which interactions between CTCFbs are analysed. ChIA-PET experiments will analyse a limited number of ligation events per interaction (ranging from 2 to 46 ligation events) while in the 4C-seq experiments done in this thesis we have analysed on average 11.242 distinct ligation events per experiment (Chapter 5; **supplemental table 1**). The complexity of the 4C-seq data provides robustness and reliability over ChIA-PET data, reflected by the poor reproducibility between biological replicates between two ChIA-PET experiments (32) and high reproducibility between 4C-seq experiments (Chapter 4).

Studying the topology of the complete genome, i.e. all possible interactions genome wide is enabled by a method called HiC (15). Studies using HiC have associated CTCF with topological structures in the genome (33,34). Enrichment of CTCF binding around the borders of these topological domains, suggests that CTCF bound sequences could have a function at these physical boundaries in the genome (34). Although HiC enables studying all interactions in the genome at the same time, it provides relatively poor resolution compared to 4C-seq. HiC provides a resolution of 40kb, at a sequencing depth of more than 100 million reads (34), while with high-resolution 4C-seq based on the analysis of only 1 million reads (35) we could show borders of topological domains at single CTCFbs resolution (**Figure 1**). To appoint border functions to given CTCFbs, 4C-seq is therefore preferred over HiC due to its higher resolution. However, when analysing the overall genome-wide chromatin topology, HiC is the preferred technique to use. The relatively poor resolution obtained using HiC will improve with deeper sequencing of the samples and with the use of restriction enzymes that cut more frequently than the '6-cutters' that are used standardly in HiC experiments. Innovation in high-throughput sequencing technology will, in the future, probably make it feasible to get to a resolution of a few kb using the HiC technique.

A challenge in the field of nuclear organization will be to show whether interactions can take place between more than 2 genomic sites at the same time. At the moment 3C related techniques analyse the presence of an interaction between two restriction fragments, while it could be possible that more than 2 genomic sequences interact with each other at the same time. Future techniques, allowing sequencing over multiple ligation events in one '3C circle' potentially could visualize the presence of for example multiple genes in a chromatin hub mediated by binding sites of a tissue specific transcription factor. This could potentially answer questions about mutual exclusivity versus synergy between interactions among genes and regulatory sequences, the next major challenge in studying genome structure.

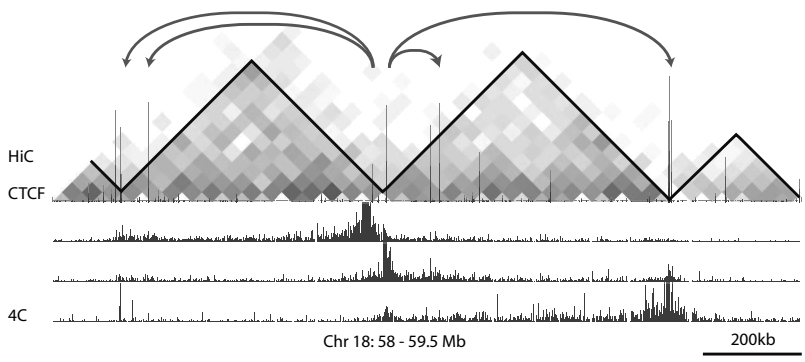


Figure 1. 4C-seq visualizes the CTCF site associated with a topological border and interactions between topological domain borders. HiC data in mouse ESCs is depicted in red [Dixon, 2012; obtained from http://bioinformatics-renlab.ucsd.edu/rentrac/] showing topological domains that are highlighted by the black lines. CTCF binding profiles are show at the same height as the HiC data in blue. The lower three tracks show 4Cseq data at and around topological borders. The region shown spans 1.5Mb.

References

1. Croft, J.A., Bridger, J.M., Boyle, S., Perry, P., Teague, P. and Bickmore, W.A. (1999) Differences in the localization and morphology of chromosomes in the human nucleus. *The Journal of cell biology*, **145**, 1119-1131.
2. Cremer, M., von Hase, J., Volm, T., Brero, A., Kreth, G., Walter, J., Fischer, C., Solovei, I., Cremer, C. and Cremer, T. (2001) Non-random radial higher-order chromatin arrangements in nuclei of diploid human cells. *Chromosome research : an international journal on the molecular, supramolecular and evolutionary aspects of chromosome biology*, **9**, 541-567.
3. Belmont, A.S., Dietzel, S., Nye, A.C., Strukov, Y.G. and Tumbar, T. (1999) Large-scale chromatin structure and function. *Current opinion in cell biology*, **11**, 307-311.
4. Brown, K.E., Guest, S.S., Smale, S.T., Hahm, K., Merkenschlager, M. and Fisher, A.G. (1997) Association of transcriptionally silent genes with Ikaros complexes at centromeric heterochromatin. *Cell*, **91**, 845-854.
5. Brown, K.E., Baxter, J., Graf, D., Merkenschlager, M. and Fisher, A.G. (1999) Dynamic repositioning of genes in the nucleus of lymphocytes preparing for cell division. *Molecular cell*, **3**, 207-217.
6. Reddy, K.E., Light, E.D., Rivera, D.J., Kisslo, J.A. and Smith, S.W. (2008) Color Doppler imaging of cardiac catheters using vibrating motors. *Ultrasonic imaging*, **30**, 247-250.
7. Finlan, L.E., Sproul, D., Thomson, I., Boyle, S., Kerr, E., Perry, P., Ylstra, B., Chubb, J.R. and Bickmore, W.A. (2008) Recruitment to the nuclear periphery can alter expression of genes in human cells. *PLoS genetics*, **4**, e1000039.
8. Kumaran, R.I. and Spector, D.L. (2008) A genetic locus targeted to the nuclear periphery in living cells maintains its transcriptional competence. *The Journal of cell biology*, **180**, 51-65.
9. Splinter, E., de Wit, E., Nora, E.P., Klous, P., van de Werken, H.J., Zhu, Y., Kaaij, L.J., van Ijcken, W., Gribnau, J., Heard, E. *et al.* (2011) The inactive X chromosome adopts a unique three-dimensional conformation that is dependent on Xist RNA. *Genes & development*, **25**, 1371-1383.
10. Chaumeil, J., Le Baccon, P., Wutz, A. and Heard, E. (2006) A novel role for Xist RNA in the formation of a repressive nuclear compartment into which genes are recruited when silenced. *Genes & development*, **20**, 2223-2237.
11. Clemson, C.M., Hall, L.L., Byron, M., McNeil, J. and Lawrence, J.B. (2006) The X chromosome is organized into a gene-rich outer rim and an internal core containing silenced nongenic sequences. *Proceedings of the National Academy of Sciences of the United States of America*, **103**, 7688-7693.
12. Dietzel, S., Schiebel, K., Little, G., Edelmann, P., Rappold, G.A., Eils, R., Cremer, C. and Cremer, T. (1999) The 3D positioning of ANT2 and ANT3 genes within female X chromosome territories correlates with gene activity. *Experimental cell research*, **252**, 363-375.
13. Holwerda, S.J., van de Werken, H.J., Ribeiro de Almeida, C., Bergen, I.M., de Bruijn, M.J., Verstegen, M.J., Simonis, M., Splinter, E., Wijchers, P.J., Hendriks, R.W. *et al.* (2013) Allelic exclusion of the immunoglobulin heavy chain locus is independent of its nuclear localization in mature B cells. *Nucleic acids research*.
14. Simonis, M., Klous, P., Splinter, E., Moshkin, Y., Willemsen, R., de Wit, E., van Steensel, B. and de Laat, W. (2006) Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4C). *Nature genetics*, **38**, 1348-1354.
15. Lieberman-Aiden, E., van Berkum, N.L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B.R., Sabo, P.J., Dorschner, M.O. *et al.* (2009) Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, **326**, 289-293.
16. Lomvardas, S., Barnea, G., Pisapia, D.J., Mendelsohn, M., Kirkland, J. and Axel, R. (2006) Interchromosomal interactions and olfactory receptor choice. *Cell*, **126**, 403-413.
17. Spilianakis, C.G., Lalioti, M.D., Town, T., Lee, G.R. and Flavell, R.A. (2005) Interchromosomal associations between alternatively expressed loci. *Nature*, **435**, 637-645.
18. Fuss, S.H., Omura, M. and Mombaerts, P. (2007) Local and cis effects of the H element on expression of odorant receptor genes in mouse. *Cell*, **130**, 373-384.
19. Noordermeer, D., de Wit, E., Klous, P., van de Werken, H., Simonis, M., Lopez-Jones, M., Eussen, B., de Klein, A., Singer, R.H. and de Laat, W. (2011) Variegated gene expression caused by cell-specific long-range DNA interactions. *Nature cell biology*, **13**, 944-951.
20. Schoenfelder, S., Sexton, T., Chakalova, L., Cope, N.F., Horton, A., Andrews, S., Kurukuti, S., Mitchell, J.A., Umlauf, D., Dimitrova, D.S. *et al.* (2010) Preferential associations between co-regulated genes reveal a transcriptional interactome in erythroid cells. *Nature genetics*, **42**, 53-61.
21. Palstra, R.J., Simonis, M., Klous, P., Brasset, E., Eijkelkamp, B. and de Laat, W. (2008) Maintenance of long-range DNA interactions after inhibition of ongoing RNA polymerase II transcription. *PloS one*, **3**, e1661.
22. de Wit, E., Bouwman, B.A.M., Zhu, Y., Klous, P., Splinter, E., Verstegen, M.J.A.M., Krijger, P.H.L., Festuccia, N., Nora, E.P., Welling, M. *et al.* (2013) The pluripotent genome in three dimensions is shaped around pluripotency factors. *Nature, in press*.
23. Krijger, P.H. and de Laat, W. (2013) Identical cells with different 3D genomes; cause and consequences? *Current opinion in genetics & development*, **23**, 191-196.
24. Deng, W., Lee, J., Wang, H., Miller, J., Reik, A., Gregory, P.D., Dean, A. and Blobel, G.A. (2012) Controlling long-range genomic interactions at a native locus by targeted tethering of a looping factor. *Cell*, **149**, 1233-1244.
25. Tolhuis, B., Palstra, R.J., Splinter, E., Grosveld, F. and de Laat, W. (2002) Looping and interaction between

hypersensitive sites in the active beta-globin locus. *Molecular cell*, **10**, 1453-1465.
26. Drissen, R., Palstra, R.J., Gillemans, N., Splinter, E., Grosveld, F., Philipsen, S. and de Laat, W. (2004) The active spatial organization of the beta-globin locus requires the transcription factor EKLF. *Genes & development*, **18**, 2485-2490.
27. Patrinos, G.P., de Krom, M., de Boer, E., Langeveld, A., Imam, A.M., Strouboulis, J., de Laat, W. and Grosveld, F.G. (2004) Multiple interactions between regulatory regions are required to stabilize an active chromatin hub. *Genes & development*, **18**, 1495-1509.
28. Palstra, R.J., Tolhuis, B., Splinter, E., Nijmeijer, R., Grosveld, F. and de Laat, W. (2003) The beta-globin nuclear compartment in development and erythroid differentiation. *Nature genetics*, **35**, 190-194.
29. Splinter, E., Heath, H., Kooren, J., Palstra, R.J., Klous, P., Grosveld, F., Galjart, N. and de Laat, W. (2006) CTCF mediates long-range chromatin looping and local histone modification in the beta-globin locus. *Genes & development*, **20**, 2349-2354.
30. Vakoc, C.R., Letting, D.L., Gheldof, N., Sawado, T., Bender, M.A., Groudine, M., Weiss, M.J., Dekker, J. and Blobel, G.A. (2005) Proximity among distant regulatory elements at the beta-globin locus requires GATA-1 and FOG-1. *Molecular cell*, **17**, 453-462.
31. Eijkelenboom, A., Mokry, M., de Wit, E., Smits, L.M., Polderman, P.E., van Triest, M.H., van Boxtel, R., Schulze, A., de Laat, W., Cuppen, E. *et al.* (2013) Genome-wide analysis of FOXO3 mediated transcription regulation through RNA polymerase II profiling. *Molecular systems biology*, **9**, 638.
32. Handoko, L., Xu, H., Li, G., Ngan, C.Y., Chew, E., Schnapp, M., Lee, C.W., Ye, C., Ping, J.L., Mulawadi, F. *et al.* (2011) CTCF-mediated functional chromatin interactome in pluripotent cells. *Nature genetics*, **43**, 630-638.
33. Yaffe, E. and Tanay, A. (2011) Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture. *Nature genetics*, **43**, 1059-1065.
34. Dixon, J.R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., Hu, M., Liu, J.S. and Ren, B. (2012) Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, **485**, 376-380.
35. van de Werken, H.J., Landan, G., Holwerda, S.J., Hoichman, M., Klous, P., Chachik, R., Splinter, E., Valdes-Quezada, C., Oz, Y., Bouwman, B.A. *et al.* (2012) Robust 4C-seq data analysis to screen for regulatory DNA interactions. *Nature methods*, **9**, 969-972.

&
Addendum

Nederlandse samenvatting
(voor niet ingewijden)

Elke cel in ons lichaam bevat dezelfde genetische informatie, die ligt opgeslagen in de celkern in een lange streng van nucleotiden, het DNA. Het DNA in het muizengenoom, het model organisme gebruikt in dit proefschrift, beslaat grofweg $1.87e^9$ nucleotiden en is verdeeld over twee allelen. De helft van de informatie, een allel, wordt verkregen via de moeder en de andere helft, het andere allel, via de vader. Genen in het DNA, bevatten de informatie die vertaald kan worden naar aminozuren, de bouwstenen voor eiwitten. De vertaling van DNA naar eiwitten gebeurt indirect: DNA wordt eerst vertaald naar RNA (transcriptie), pas daarna het RNA omgezet kan worden in eiwitten (translatie). Eiwitten vormen de functionele 'units' in ons lichaam waardoor een bepaalde cel zijn functie kan uitoefenen. Niet elke cel in ons lichaam gebruikt echter dezelfde eiwitten. Als gevolg zal elke cel in ons lichaam specifieke informatie uit ons DNA transcriberen (vertalen naar RNA) en niet alle coderende elementen uit ons DNA 'gebruiken'.
Gedurende de laatste decennia is het duidelijk geworden dat de functie van het genoom gerelateerd is de 'compactheid' van DNA en aan de positie van DNA in de celkern. De complexiteit van de vouwing van DNA is een gevolg van de verhouding tussen de totale lengte van DNA, 2m, en de diameter van een gemiddelde zoogdier celkern, 10µm. Om de lange strengen DNA in een celkern te kunnen passen wordt het genoom compact 'gevouwen' met behulp van eiwitten. Het complex dat gevormd wordt, chromatine, bestaat in twee vormen: eu-chromatine en hetero-chromatine. Eu-chromatine is de meest 'open' vorm van de twee en wordt voornamelijk geassocieerd met actieve transcriptionele processen. Hetero-chromatine is de compacte vorm van DNA en wordt over het algemeen geassocieerd met inactieve genen.
Niet alleen vouwing maar ook positie in de celkern speelt een rol bij de functie van het genoom. Zo worden bepaalde gebieden in de celkern geassocieerd met activatie of repressie van transcriptie. Een voorbeeld van een transcriptioneel repressief gebied in de celkern zijn de nucleaire lamina, aan de periferie van de kern, waar zich voornamelijk hetero-chromatisch DNA bevindt. Transcriptioneel actieve genen bevinden zich, in de eu-chromatische vorm, voornamelijk centraler in de celkern. Verschil in de expressie van een gen tussen cel typen kan dan ook correleren met een verschil in locatie of 'compactheid' van dit gen. In dit proefschrift staat de relatie tussen nucleaire organisatie en transcriptie centraal.

In **hoofdstuk 3** bestuderen we deze relatie aan de hand van een model gen locus dat specifiek tot expressie komt in B lymphocyten, het 'immunoglobulin heavy chain' (IgH) locus. Het eiwit waarvoor dit gen codeert is de B cel receptor, een extra-cellulair membraan gebonden eiwit dat anti-genen herkent en een belangrijke rol speelt in het immuunsysteem. Bij de expressie van dit gen, wat essentieel is voor B cel ontwikkeling, speelt nucleaire organisatie een belangrijke rol. Om er zeker van te zijn dat elke cel een enkele receptor op zijn membraan tot expressie brengt wordt de receptor mono-allelisch tot expressie gebracht. De 'keuze', welk allel hierbij wordt gebruikt wordt op een willekeurige manier gemaakt. Een serie DNA recombinaties op beide allelen gaat vooraf aan dit 'keuze' moment. Het moment wordt bepaald wanneer op één van de beide allelen volledige recombinatie plaatsvindt tijdens de ontwikkeling van de B lymphocyten. Complete recombinatie resulteert in productief transcript op één enkel allel, het productieve allel. Het niet-productieve allel wordt uitgesloten van eiwit productie in een proces dat allelische exclusie wordt genoemd. Allelische exclusie van het niet-productieve allel is lang geassocieerd geweest met transcriptionele repressie en relocatie van dit allel naar een repressieve omgeving in de celkern. In **hoofdstuk 3** van dit proefschrift gebruiken we dit dynamische model systeem en vergelijken de nucleaire structuur van het productieve en het non-productieve allel van het IgH locus. Hierbij maken we gebruik van 'Chromosome Conformation Capture on Chip' techniek (4C) die ons in staat stelt de nucleaire omgeving van een gen van interesse te bestuderen. DNA sequenties die in de directe nabijheid van het gen van interesse liggen kunnen worden 'gevangen' en geïdentificeerd. Vervolgens kan in kaart worden gebracht met welke delen van het genoom

een bepaald gen in contact staat. In dit hoofdstuk presenteren we een variant op de 4C techniek waarbij we de genomisch omgeving van beide allelen voor het eerst apart kunnen bestuderen. Hierbij vinden we verschillen in nucleaire organisatie tussen het productieve en niet-productieve allel die op een verschil in recombinatie wijzen tussen beide allelen. Ondanks de verschillen in organisatie van het locus zelf is de genomische omgeving van beide allelen vrijwel identiek. Bovendien vinden we geen verschil in lokalisatie van de twee allelen en concluderen dat allelische exclusie onafhankelijk is van de nucleaire organisatie in een matuur stadium van B cel ontwikkeling. Revisie van een lang bestaand mechanisme, waarbij nucleaire re-organisatie en de bijbehorende inhibitie van transcriptie wel van belang werd geacht voor allelische exclusie is het resultaat van dit onderzoek.

Nucleaire organisatie kan ook op een kleinere schaal beschreven worden. DNA-DNA interacties worden dan bestudeerd in relatie tot de regulatie van transcriptie door zogenaamde 'regulatoire elementen'. Het genoom bestaat voor slecht 3% uit genen en dus voor 97% uit DNA dat niet codeert voor een eiwit. Dit deel van het genoom, dat lang als 'junk-DNA' werd bestempeld, blijkt een plethora aan regulatoire elementen te bevatten die betrokken kunnen zijn bij de regulatie van genexpressie. Deze regulatoire elementen kunnen o.a. worden geïdentificeerd doordat ze gekenmerkt worden door modificaties aan chromatine eiwitten waar het DNA omheen gewonden is, histonen. Deze histon eiwitten kunnen modificaties bevatten die een functionele betekenis hebben voor een DNA sequentie. Zo is de tri-methylatie van lysine 36 op histon eiwit 3 (H3K36me3) kenmerkend voor actieve transcriptie en is de acetylatie van lysine 27 op histon eiwit 3 (H3K27ac) kenmerkend voor een regulatoir element wat transcriptie positief kan beinvloeden, een enhancer. Histon eiwitten kunnen het DNA dus van een 'epigenetische' code worden voorzien, die gebruikt kan worden om deze regulatoire elementen te bestuderen. Het genoom bevat honderd duizenden van deze regulatoire elementen die op grote afstand van een gen kunnen liggen. Distaal gelegen enhancers kunnen genexpressie beinvloeden door fysiek contact te maken met een gen promoter. Op deze manier worden lussen gemaakt in het DNA die functioneel gerelateerde stukken DNA dicht bij elkaar brengen en ontstaat er een 3D structuur op lokale schaal.

Om DNA-DNA contacten tussen een gen en zijn distal gelegen regulatoire elementen te kunnen visualiseren is een techniek ontwikkelt die beschreven staat in **hoofdstuk 4**, hoge-resolutie 4C. Deze techniek kan worden gebruikt om op een kost-effectieve manier, interacties tussen promotoren van genen en hun potentiële regulatoire elementen te detecteren. Het analyse pakket wat wordt gepresenteerd staat een robuuste analyse en kwantificatie toe van deze interacties tussen een gen en zijn potentiële 'enhancer'. De methode wordt gevalideerd op het α- en β-globine gen locus en identificeert nieuwe enhancers op het Oct4 en SatB1 locus.

DNA-DNA interacties worden genoomwijd geassocieerd met eiwitten die DNA binden op de interacterende DNA sequenties. Een DNA bindend eiwit wat dit soort 3D structuren kan vormen is CTCF. CTCF is een model eiwit dat betrokken is bij het vormen van loops tussen bindingsplekken van dit eiwit. In **hoofdstuk 5** bestudeer ik de rol van CTCF bij de formatie van lange afstands interacties tussen CTCF gebonden DNA sequenties. Bindingsplekken van CTCF kunnen in kaart worden gebracht met behulp van de *Chromatine Immuno Precipitatie* (ChIP) techniek waarbij DNA gebonden aan het eiwit wordt geïsoleerd. Vervolgens kan door 'high-throughput sequencing' van het gebonden DNA (ChIP-seq) de locatie bepaald worden van de bindingsplekken van een eiwit. (op dezelfde manier kunnen modificaties aan histonen in kaart worden gebracht). We vinden dat CTCF grofweg op 40.000 plekken in het genoom bindt. Door middel van hoge-resolutie 4C-sequencing analyseren we de genomische contacten van honderden van dit soort eiwit bindingsplekken en laten we zien dat niet alle CTCFbs betrokken zijn bij de formatie van lange afstands interacties. Deze observatie werpt de vraag op: *Welke eigenschappen van een bepaalde bindingsplek bevorderen de formatie van een bepaalde loop?*

In een poging deze vraag te beantwoorden hebben we de CTCF bindingsplekken een interactie vertonen met een andere CTCF bindingsplek gekarakteriseerd op basis van de co-occupatie van CTCF met een ander DNA bindend eiwit dat betrokken kan zijn bij 'loop' formatie, Cohesin en de co-associatie met enhancer sequenties (gekarakteriseerd door de H3K27ac modificatie). We classificeren CTCF bindingsplekken op basis van deze co-associatie waarbij we CTCF/Cohesin en CTCF/H3K27ac overlappende sequenties onderscheiden. Alhoewel we voornamelijk homo-typische interacties tussen dezelfde klasses CTCF bindingsplekken vinden, worden interacties tussen verschillende klasses ook vaak aangetroffen. Het feit dat we interacties tussen regulatoire elementen die CTCF/H3K27ac sequencties onderliggen vinden, suggereert dat CTCF betrokken kan zijn bij genexpressie door de formatie van chromatine loops tussen deze elementen.
Om de rol van CTCF gemediëerde 'loops' te begrijpen analyseren we de interacties van CTCF bindingsplekken rond celtype specifiek geëxpresseerde genen in verschillende weefsels. Een verrassende waarneming is dat geconserveerde CTCF bindingsplekken betrokken kunnen zijn bij DNA-DNA interacties tussen regulatoire elementen op een celtype specifieke manier. Op dit moment onderzoeken we de mogelijkheid dat CTCF, in associatie met celtype specifieke transcriptie factoren, een bijdrage kan leveren aan de specifieke 3D organisatie van chromatine. Een potentiëel, zeer interessante, uitkomst van deze analyse zou kunnen zijn dat CTCF de posities in het genoom markeert waar celtype specifieke factoren samen kunnen binden met een lange afstands interactie tot gevolg. Dit soort interacties tussen genen en regulatoire elementen zouden belangrijk kunnen zijn voor de weefsel specifieke expressie van deze genen.

In dit proefschrift beschrijven we analyses die gedaan zijn om de relatie tussen de 3D structuur van DNA en transcriptie te begrijpen. We hebben daarbij gebruik gemaakt van technieken om de positie van een bepaald gen locus te linken aan de activiteit van dit locus. Daarbij ontkrachten we het belang van nucleaire organisatie voor allelische exclusie en laten zien dat celtype specifieke genomische omgeving van DNA niet bepalend is voor de transcriptionele activiteit van een locus. Andere factoren dan de 3D structuur van het DNA zullen in veel gevallen in samenspel met de nucleaire organisatie van een gen locus bepalend zijn voor transcriptie een gen.
Een van deze factoren kan de binding van transcriptie factoren zijn. Hoe de interactie tussen een regulatoir element en een gen beinvloedt kan worden door eiwit binding is onderzocht door gebruik te maken van hoge resolutie 4C-sequencing. Analyse van de genomische interacties van honderden CTCF bindingsplekken is van belang geweest bij een mogelijke ontrafeling van het mechanisme hoe CTCF betrokken is het opzetten van celtype specifieke interacties. Onze hypothese is dat dit in samenspel zal gaan met celtype specifieke transcriptie factoren.
Recente ontdekkingen hebben laten zien dat cel specifieke transcriptie factoren een rol hebben bij het opzetten en/of stabiliseren van een genoom-wijde 3D organisatie waarbij specifiek genen die deze eiwitten binden ruimtelijk samenkomen in de kern. Verschillende technieken maken het mogelijk om dit soort phenomenon te bestuderen. Voor de analyse van interacties tussen regulatoire elementen is 4C-seq op dit moment superieur. Deze techniek zal in de nabije toekomst waarschijnlijk plaatsmaken voor een gerelateerde techniek, HiC, waarbij alle interacties die gemaakt worden in het genoom tegelijk in kaart worden gebracht.

Curriculum Vitea

Sjoerd Johannes Bastiaan Holwerda is geboren op 28 februari 1981 te Utrecht in Nederland. In 1999 behaald hij zijn VWO diploma aan het Petrus Canisius College te Alkmaar. Zijn pre-academische carrière ving aan met een studie Biotechnologie aan de Wageningen Universiteit waar hij zijn propadeutisch examen met succes heeft afgerond. In 2002 begon hij met een studie Life Sciences & Chemistry aan de Hogeschool van Utrecht, waar hij in 2005 is afgestudeerd. Na deze opleiding is hij begonnen als research analist op het ErasmusMC te Rotterdam in de groep experimentele chirurgische oncologie onder leiding van Dr. Timo ten Hagen. In 2006 is hij begonnen aan de top master Systems Biology aan de Vrije Universiteit te Amsterdam welke met succes is afgerond in 2008. In juli 2008 is hij zijn academische carrière gestart aan het Hubrecht Instituut te Utrecht, in de groep Biomedical Genomics onder begeleiding van Prof. Wouter de Laat. De resultaten van het onderzoek, verricht tijdens zijn PhD opleiding, staan beschreven in dit proefschrift.

Published work

Holwerda SJ , van de Werken HJ, Ribeiro de Almeida C, Bergen IM, de Bruijn MJW, Verstegen MJ, Simonis M, Splinter E, Wijchers PJ, Hendriks RW, de Laat W. *Allelic exclusion of the immunoglobulin heavy chain locus is independent of its nuclear localization in mature B cells.* Nucleic Acids Res. 2013 Aug 1; 141(14) PMID: 23748562

Medvedovic J, Ebert A, Tagoh H, Tamir IM, Schwickert T, Novatchkova M, Sun Q, Huis in 't Veld PJ, Guo C, Yoon HS, **Holwerda SJ**, de Laat W, Cogné M, Shi Y, Alt FW, Busslinger M. *Flexible long-range loops in the VH gene region of the Igh locus that likely facilitate the generation of a diverse antibody repertoire.* Immunity 2013 in press

Holwerda SJ, de Laat W. *CTCF: the protein, the binding partners, the binding sites and their chromatin loops.* Philos Trans R Soc Lond B Biol Sci. 2013 May 6; 368(1620) PMID: 23584178

Holwerda S, de Laat W. *Chromatin loops, gene positioning, and gene expression.* Front Genet. 2012 Oct 17;3 217 PMID: 23087710

van de Werken HJ*, Landan G*, **Holwerda SJ**, Hoichman M, Klous P, Chachik R, Splinter E, Valdes-Quezada C, Oz Y, Bouwman BA, Verstegen MJ, de Wit E, Tanay A, de Laat W. *Robust 4C-seq data analysis to screen for regulatory DNA interactions.* Nat Methods. 2012 Oct; 9(10) PMID: 22961246

van de Werken HJ, de Vree PJ, Splinter E, **Holwerda SJ,** Klous P, de Wit E, de Laat W. *4C technology: protocols and data analysis.* Methods Enzymol. 2012; 513 PMID: 22929766

Unpublished work

Holwerda SJ, de Wit E, van de Werken HJ, Wijchers PJ, Mokry M, Cuppen E, de Laat W. *Functional analysis of the role of CTCF in mediating chromatin flexibility.* Work in progress

Dankwoord

Het was me er eentje! De afgelopen periode als AIO heeft vooral in het teken gestaan van veel lol in het lab met een groep mensen die een soort (professionele) familie wordt in de loop van de tijd. Deze familiaire 'soap' op het Hubrecht instituut die me heeft helpen blijven lachen zal me altijd bijblijven als een leuke omgeving waarbij veel mensen voorbij gekomen zijn die ik wil bedanken worden voor hun bijdrage aan deze periode op allerlei manieren.

Allereerst **Wouter**, bedankt voor je begeleiding in de 5 jaar dat ik onderdeel uit het mogen maken van je groep. Jij hebt een ideale wetenschappelijke omgeving geschapen waarin ik met veel plezier en voldoening mijn 'werk' heb kunnen doen. Je hebt me de tijd gegunt en mogelijkheden geboden om me te ontwikkelen en ik wil je bedanken voor je geduld, kritische blik en het vertrouwen dat je me hebt gegeven. We hebben vooral in mijn laatste jaar 'onze' successen behaald en ik hoop dat we binnenkort het laatste deel van mijn AIO tijd kunnen afronden in de vorm van 'het CTCF paper'. Nogmaals bedankt voor de leerzame tijd in een fantastische groep!

Groep de Laat, pipetteren zal nooit meer hetzelfde zijn zonder jullie!
Het begon in Rotterdam waar ik als onervaren AIO ik tussen de 'chromatine-experts' veel heb kunnen leren. **Robert-Jan**, **Marieke** en **Daan** het was kort maar gezellig, bedankt.
Met een (relatief) kleine groep begonnen we in Utrecht. **Erik** jij bent in meerdere opzichten een 'nestor' voor me geweest. Je hebt niet alleen heel veel kennis en kunde in de moleculaire biologie op me overgebracht maar ook je positieve houding als persoon maakt je de perfecte persoon om mee samen te werken. Het CTCF project, waar we aan gewerkt hebben, gaat snel zijn 'einde' vinden.
Harmen, never a dull moment! Jouw brede mening over alles maakt elke lunch tot een waar genot. Jij bent mede verantwoordelijk voor een groot gedeelte van mijn boek. Onze samenwerking heeft vanaf het begin tot het eind standgehouden en heeft geresulteerd in mooie publicaties! Bedankt daarvoor en succes in Rotterdam en Bergschenhoek.
Petra, mijn 'FISH' en 4C leermeesteres, vriendin en paranimf. Je bent samen met Erik verantwoorlijk geweest voor mijn basis kennis van alle chromatine gerelateerde technieken in het lab. Ook buiten het lab hebben we veel gemeen: Dansen, lachen en biertjes drinken houden we erin, P! Gaaf dat je aan mijn zijde komt (hebt ge-) staan op 10 oktober.
Paranimf aan mijn andere zijde, **Patrick**. Jij bent als post-doc en collega een mentor geweest in het lab, dank je wel daarvoor. Je bent ook een vriend geworden buiten het lab. In tegenstelling tot onze werkwijzen die nogal van elkaar verschillen, ligt onze muzieksmaak vrij dicht bij elkaar. Buiten de techno waarmee we (in onze ogen) het lab verrijkten, hebben we ook de live hard gedanst op menig deuntje met als (perssonlijk) hoogtepunt 'The Advent'. Harder wordt het niet meer denk ik! Ik hoop dat je binnenkort je eigen groep mag leiden. Dat zal je zeer goed staan!
Yunnie, my lab 'sister'. We started around the same time and have shared a lot of lab frustrations and laughter together. I will never forget your KTV version of 'Kom van dat dak af!'. Thanks for the great times and good luck with finishing your project. Go get them!
Paula, heel veel plezier met je nieuwe gezin straks. Dank je voor je vrolijkheid in het lab en succes met de afronding van je PhD.
Yuva, my office neighbour. I remember our first cloning project together, luckily we have gained some expertise now. Good luck finalizing your PhD and let's keep in touch.
Elzo, je droge humor en scherpe kritische blik zijn uniek in het lab. Het is dan ook aangenaam en nuttig (goede combi) om met je samen te werken op het CTCF project wat hopelijk snel tot een mooie publicatie komt.

Blaise, our collaboration ended halfway my PhD period, when you moved to the UK. Thanks for everything you've done for the IgH and CTCF project.
Marjon, samenwerken met jou is net zo gemakkelijk als leuk. Bedankt voor je hulp in het lab, waar je zelfs op het moment dat ik dit schrijf mee bezig bent.
Geert, eindelijk iemand om mee over de Eredivisie te praten. Dat werd tijd na 3,5 jaar! Tussen de bioinformatica tips door was er vaak tijd voor een pilske. Let's keep it up!
Peter, de tweede 'wet-lab' post-doc. Altijd klaar met een kritische noot waar ik persoonlijk veel aan heb gehad. Je kennis over B cellen heeft ook zeker geholpen bij afronding van het IgH project. Thanks.
Britta, ga zo door! Dan wordt wel wat ;).
Carlo, mijn andere kantoor maat. Veel plezier en succes in je AIO tijd.
Christian, compadre! I enjoyed working with you. We will meet just before you start again as a post-doc in the lab. Looking forward to seeing you.
Thomas, de beste student die ik ooit begeleidt heb. Bedankt voor je werk, onze samenwerking was ging soepel. Succes op het NKI!
Mehmet, Cergentis pipetboy. Laat me weten als je een BMW koopt, dan stap ik een keer bij je in.
Monique, Cergentis pipetgirl. Het was grappig om een oude HLO klasgenoot tegen te komen in het laatste half jaar.

Buiten ons eigen lab is het Hubrecht instituut een plek geweest waar tussen de wetenschap door veel tijd is voor ontspanning die ik met veel mensen gedeeld heb. **Maartje**, altijd lekker om even slap mee te ouwehoeren en hard te lachen. Fiets ze! **Dan and Ira**, thanks showing me the Russian countryside. I enjoyed it so much. **Emily**, 'beetje kut' will stay my favourite translation of 'een aansteker'. **Bart**, even ontspannen tijdens een rondje rennen om de kromme Rijn is toch net anders dan een rondje over de 'Koning van Spanje'. Wat een hel, maar toch bedankt voor de tip. Ik had dat niet willen missen! Succes in the USofA. **Frank** en **Arnout** , alweer een tijdje uit Utrecht. Samen met Erik en Bart goed voor wat 'pap in de benen' op de fiets. Maar beide afzonderlijk ook voor een 'houten kop'. Succes in de States allebei.
Degenen die de het 'kantoor-leven' opleuken ook heel erg bedankt! **Alyson**, **Paul, Tam, Linda en Yvonne.** Het was me een waar genoegen om met jullie het kantoor te delen. **Alyson**, thanks for sharing the office with the 'chromatin-nerds'. **Paul**, de broodnuchtere Limburger, toch al zijn er momenten dat je minder nuchter bent ;). Respect dat je het binnen 4 jaar je project af hebt. Ik baal nog steeds een heel klein beetje dat je me eruit hebt gelopen in Gulpen. **Tam**, gek eigenlijk dat we elkaar nooit de tent uit hebben gevochten. Mijn ode aan jou komt zo.
Naast mijn directe collega's in het lab wil ik ook alle medewerkers van het Hubrecht Instituut bedanken die het een aangename plek maken om te werken en veel (niet wetenschappelijke) taken uit handen nemen: **Stieneke** (weefselkweek Koningin), de mannen van de **facilitaire dienst**, **technische dienst & IT**, de **financiele afdeling**, **personeel en organisatie, dierenverzorging** en **media keuken**.

Er zijn ook mensen buiten het Hubrecht instituut die hebben bijgedragen aan de wetenschappelijke inhoud van dit boek. Op het ErasmusMC, **Rudi** en de mensen in jouw groep, **Claudia**, **Ingrid**, **Marjolein** en **Yuvaraj**. Zonder de vele FACS experimenten die er gedaan zijn en de microarray studie aan de B en T cellen was hoofdstuk 3 en de publicatie er nooit geweest. Bedankt voor het werk en de tijd die jullie hebben genomen hiervoor.

Vrienden en familie uiteraard ook bedankt voor de nodige ontspanning tussen de noeste arbeid door.
Jeroen, altijd gezellig. Arnhem, Nijmegen, Amsterdam of Berlijn. Ont- of inspanning gaan nog altijd hand in hand. Volgend jaar maar weer een 'klein' rondje rennen. Thanks man.
Sjoerd, we hebben veel gelachen en beleefd. Jammer genoeg hebben te veel mensen te kort van je kunnen genieten. Het heeft helaas zo moeten zijn. Bedankt voor de vette tijd.

Maurice, ongemerkt heb ik veel aan het meditatie weekend in Klein Zundert gehad. Relativeren en 'zijn' is het wel voor mij, met mate dan ;).
Thanks **Leendert** en **Bart**, bedankt voor de 'opvang' in Arnhem.

Mijn lieve schoonfamilie. Je hebt ze niet voor het kiezen maar ik mag niet klagen ;).
Wim en **Joy**; **Patricia**, **Peter** & **Maroussia**; **Roby** en **Rinske**, bedankt voor alle gezelligheid. De deur staat altijd open het verre buitenland voor jullie. Special thanks aan **Joy**, voor de spoedcursus 'Indesign' en voor de hulplijn die ik af-en-toe heb ingezet. Zonder jouw professionele hulp had ik een stuk harder moeten zweten op de lay-out. Thanks a lot.

Femke, lief 'klein' zusje. Ik hoop dat ik ooit tegen iemand kan zeggen dat je een nieuwe 'langnek' hebt opgegraven. Het begon allemaal met: 'Vliegde ik?. Nee je valde'. Zet hem op tussen de baarden en stoffige oude mannen! We zijn straks een soort buren, in ieder geval wat taal betreft zitten we heel dicht bij elkaar in de buurt. X.
Jippe en **Ellen**, lieve broer en (schoon)zus. Huisje boompje en binnenkort een oomzeggertje. Ik kijk naar hem uit, Sinterklaas-avond zal nooit meer hetzelfde zijn :). Dank voor alle gezelligheid en warmte.
Thomas en **Natalia**, lieve broer en (schoon)zus. Een te gekke bruiloft achter de rug in Polen. De wijn en wodka die over is komt wel een keer op, Tam en ik komen wel helpen. X.
Eelco en **Tia**, lieve broer en (schoon)zus in China. Skype will do, for the moment. De momenten met jullie zijn altijd kort en krachtig. Ik kijk weer naar uit volgend jaar. *Tsjaidien!*
Lot en **Rob**, lieve zus en zwager. Jullie wonen lekker dichtbij in 't Stadsie. Jullie 'geestelijke' volwassenheid is iets waar ik bij elk bezoek weer iets uithaal voor mezelf. Rob, fietsen met de wind mee is het 'nieuwe' fietsen voor me. Wat een uitvinding.
Hans en **Dayenne**, broer en (schoon)zus. Helaas weinig contact, maar jullie zitten in mijn hart.

Heit en **Mem**, bedankt voor jullie onvoorwaardelijke steun en liefde. Zonder die twee dingen was dit boek er zeker niet geweest.

Tamara, skat. Je hebt mijn hele carriere op de voet gevolgd en ik de jouwe. Vier jaar lang hebben we een kantoor gedeeld, meestal fietsten we samen naar werk toe en ik kwam je de hele dag tegen op het lab. Dat is gelukkig nooit gaan vervelen. Je bent ook nog de leukste huistgenoot die er is! Ik heb daarom ook heel veel zin om met je te verhuizen naar Basel en te kijken of er wat te doen is daar. Bedankt voor de vrolijkheid en liefde die je geeft. Ik hou van je. X.