KEEPING THE GENOME IN SHAPE: A ROLE FOR PROTEIN AND RNA

Erik Splinter

Leescommissie:
Prof. dr. E. Cuppen
Prof. dr. R.F. Ketting
Prof. dr. B. van Steensel
Prof. dr. H.T.M. Timmers

The research described in this thesis was performed at the ErasmusMC, within the framework of the graduate school Medisch-Genetisch Centrum zuid-west Nederland (MGC) in Rotterdam and in the Hubrecht Institute of the royal Netherlands Academy of Arts and Sciences (KNAW), within the framework of the graduate school of Cancer Genomics and Developmental Biology in Utrecht, the Netherlands

Financial support by the J.E. Jurriaanse Stichting and Roche Diagnostics for publication of this thesis is gratefully acknowledged.

ISBN: 978-90-393-5651-7

Layout & printing: Off Page, Amsterdam

KEEPING THE GENOME IN SHAPE: A ROLE FOR PROTEIN AND RNA

Het in vorm houden van het genoom: een rol voor eiwit en RNA

(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
dinsdag 25 oktober 2011 des middags te 12.45 uur

door

Erik Cornelis Splinter

geboren op 21 mei 1980 te Leiderdorp

Promotoren:
Prof. dr. W.L. de Laat
Prof. dr. F.G. Grosveld

CONTENTS

SCOPE OF THIS THESIS

Gene regulatory information is primarily encoded in the DNA sequence. However, over the past decade also the three dimensional folding of the DNA fiber inside the nucleus has been implicated in regulating gene expression. This thesis describes experiments aimed at unraveling the shape of the genome, studying the relationship between genome structure and function and identifying the factors responsible for genome folding.

Chapter 1 provides a general introduction into the various mechanisms by which gene expression is regulated and how this relates to the spatial organization of the genome. This is continued more specifically in **Chapter 2**, where we describe the identification of regulatory DNA elements and discuss how genome folding relates to the functioning of these elements. This provides context for the experiments presented in the later chapters of this thesis.

Part of the research conducted focused on the development and adaptation of novel technologies to investigate nuclear organization. An overview of these technologies and of closely related methods can be found in Chapter 1 (3C based methods). A detailed description of the principles behind 3C (to study locus wide organization) and 4C (to study chromosome wide conformations) as well as practical protocols for their application are provided in **Chapter 3** and **Chapter 4** respectively.

Chapter 5 describes the application of the 3C method to the mouse β-globin locus. In this study, both by using cells carrying conditional CTCF alleles and taking advantage of cells carrying a targeted mutation of a CTCF site within the globin locus, we investigated the role of CTCF in dictating chromatin interactions. **Chapter 6** continues this study by using higres-4C-seq to determine the spatial organization of the mouse β-globin locus. Highres-4C-seq is a novel variant of 4C technology that allows the detection of interactions between specific regulatory elements with high resolution. This chapter builds on the data presented in Chapter 5 and provides novel insights into the role of CTCF mediated loops in dictating spatial contacts and regulating gene expression in the β-globin locus.

Next we switch from investigating locus wide conformations, to interrogating the spatial organization of a whole chromosome. The spatial organization of both the active and inactive female mouse X-chromosome was determined, of which the results are described in **Chapter 7**. To be able to discriminate between the active and inactive X-chromosome in female mammalian cells we developed an allele specific 4C-seq approach. Based on the presence of single nucleotide polymorphisms (SNPs), this method enabled separating the contacts made by the two chromosomes. Moreover, by studying cells depleted for the noncoding RNA Xist, this method was used to elucidate the role of this RNA molecule in dictating chromosome conformation.

Overall this thesis provides detailed information concerning the 3C and 4C methods that were used to investigate the spatial organization of DNA inside the nucleus. These studies ranged from investigating locus wide to chromosome wide organization. Moreover, CTCF and the noncoding RNA Xist were identified as factors involved in chromatin folding, each acting at a different level of organization, and providing novel insights into the mechanisms by which DNA is organized inside the nucleus of a cell.

1

INTRODUCTION

Introduction

Over 200 cell types exist within the human body, each being different in morphology and function, yet all containing the same genome. Regulation programs acting on the approximately 25.000 genes found in a typical mammalian genome drive the specification of cells during development. Also during later stages of life correct gene expression is required, as defects can lead to diseases such as cancer. To ensure correct spatial and temporal expression of genes different control mechanisms are in place acting on different levels of gene regulation. In this chapter we will discuss various levels of gene regulation by introducing transcription, chromatin and nuclear organization. Also insight is provided into the 3C, 4C and related technologies that can be used to study the relationship between gene transcription and chromosome topology.

Transcription and packaging of DNA into chromatin

Transcription is the process that retrieves the information stored in our genome to generate the biomolecules that support life. The key transcription enzyme is RNA polymerase, a molecule that comes in different flavors and which is capable to embrace and travel along a DNA strand to read its nucleotide sequence and synthesize a complementary RNA molecule. RNA molecules, in turn, can perform specific functions themselves or be translated into proteins that give structure and function to the cell. Some basal transcription may occur throughout the genome, but recruitment of RNA polymerases to DNA is not random. The transcription machinery is preferentially directed to the start of genes and other relevant sequences that need to be transcribed. Recruitment of RNA polymerase is facilitated by so-called transcription factors, which bind to specific DNA sequences. These binding or affinity sites, also called *cis*-regulatory DNA elements, can locate close to genes on the linear DNA template (like promoters) or away from genes (as is often seen for enhancers). In Chapter 2 we will discuss how these elements can be discovered and how they operate in the context of the folded genome. Here, we will first discuss how DNA is packaged into the cell nucleus.

Each mammalian nucleus, having a diameter of only 5-10μm, stores a DNA fiber with a total length of approximately 2 meters. To be able to fit the negatively charged DNA polymer it needs to be properly folded and compacted. This is accomplished by packaging the DNA into a structure called chromatin. Chromatin can be defined as the DNA and all the proteins that bind to it. First the DNA helix is wrapped 1.75 turns around a histone octamer forming a nucleosome, which is the basic building block of chromatin (Luger et al. 1997). The histone octamer consists of two copies of four different histone proteins (H2A, H2B, H3 and H4), each containing a globular DNA-binding domain and an N-terminal tail that protrudes from the nucleosome. Nucleosomes are connected via a small stretch of linker DNA that varies between 10 to 60bp in size (Smith et al. 1983; Luger et al. 1997), to form the 'beads on a string' or so called '10nm fiber' (Oudet et al. 1975). Arrays of nucleosomes are subsequently organized by linker histones belonging to the histone H1 family (Horn and Peterson 2002). The presence of histone H1 is anti correlated with transcriptional activity. Although the exact mechanism remains elusive, inhibition of transcription by histone H1 is likely to function by stabilizing the position of the nucleosome on the DNA, thereby shielding the affinity sites from regulatory proteins (Pennings et al.

1994). The level of chromatin organization beyond the 10nm fiber is less clear. Currently, mainly based on *in vitro* data, two models exist describing two different options of how the 10nm fiber is further compacted into a 30nm fiber: the solenoid and two-start helix zigzag model. The solenoid model proposes the coiling of the 10nm fiber around a central axis having 6-7 nucleosomes per turn. The two-start helix zigzag model predicts a zigzagging of two nucleosomes that in turn coil into a helix (reviewed by (Tremethick 2007)). Despite intense efforts to unravel this structure it must be mentioned that the 30nm fiber is not generally observed in mammalian nuclei and its relevance therefore remains to be determined (Horowitz-Scherer and Woodcock 2006; Fussner et al. 2011).

Transcriptional regulation via chromatin modifications

The packaging of DNA into chromatin prohibits transcription factors and other DNA metabolizing proteins from accessing the DNA and therefore has consequences not only for the regulation of transcription but also of processes like DNA replication and DNA repair. Thus, it is therefore not surprising that tight regulation mechanisms are in place to orchestra DNA accessibility in its chromatin environment. Nucleosome positioning is found highly organized, especially at the promoter regions where important regulatory elements reside. A passive force acting on nucleosome positioning is the affinity of the core histones for certain DNA sequences (Segal et al. 2006) or the presence of a genomic barrier on the DNA, like the DNA binding protein CTCF (Fu et al. 2008). Also, active, ATP dependent reorganization of nucleosome positioning takes place by sliding nucleosomes along the DNA fiber (Bowman 2010). The equilibrium of these different forces acting on the nucleosome will eventually determine its precise position (reviewed by (Segal and Widom 2009)). Besides moving or evicting nucleosomes, the core histone proteins can be substituted with variants to execute specific functions. For example histone H3 is replaced by CENP-A at centromeric regions (Palmer et al. 1991) and H2A can be substituted with macro-H2A, a variant related to gene silencing (Changolkar et al. 2010).

Not only the distribution of nucleosomes along the DNA fiber, but also modifications made to the chromatin proteins and to the DNA itself can modulate expression. DNA methylation primarily occurs at CpG dinucleotides. This modification acts in the repression of gene transcription by preventing the transcription machinery to associate with the DNA and by recruitment of repressive histone modifiers (reviewed by (Jones and Takai 2001; Deaton and Bird 2011)). In the vertebrate genome approximately 70% of all annotated gene promoters contain CpG's (Saxonov et al. 2006), showing that many genes can potentially be controlled by DNA methylation. The current idea is that gene silencing preceeds DNA methylation. This is based for example on experiments performed on the mammalian X-chromosome that is silenced and DNA methylated in somatic female cells: during the X inactivation process CpG methylation was found to follow gene silencing and to be required to prevent gene reactivation only after X inactivation (Csankovszki et al. 2001).

Like DNA, also histone proteins can be chemically modified. Various modifications are possible among which are: methylation, acetylation, ubiquitination and SUMO-ylation. Most modifications are placed on the histone tails that protrude from the nucleosome, and each has their specific function.

Those investigated in the scope of this thesis involve the methylation and acetylation of histone H3. Methylation of histone H3K9 and H3K27 are both related to gene repression, while H3K4 methylation and the acetylation of several lysines of H3 correlate with active gene transcription. Like the other chromatin modifications, histone modifications enable the binding of specific regulatory proteins, or they function indirectly, for instance by loosening the histone-DNA interaction, as in the case of histone acetylation (reviewed by (Kouzarides 2007)).

Taken together, the packaging of DNA into chromatin is necessary to fold our genome in the small cell nucleus, and simultaneously restricts the accessibility of DNA. DNA binding proteins can disrupt the nucleosome fiber when finding their cognate DNA binding sequence. The modification both of DNA and of histones can facilitate, or prohibit, the access of transcription factors to DNA.

Transcription coordination in the nuclear space

Also beyond the level of the chromatin fiber and its first orders of packaging, DNA is not randomly organized in the nucleus; each chromosome for example occupies its distinct space called a chromosome territory (Cremer et al. 2001). Within these territories, the chromosomal sequences fold to bring together regions of similar gene density and gene activity. This was known from microscopy studies, but demonstrated in more detail by novel genomics approaches based on 3C technology. Based on the detection of co-localizing DNA sequences, these methods allow the investigation of the spatial organization of DNA in the nucleus (a more detailed description and overview of these technologies can be found below). By applying 4C on the β-globin locus and comparing its genomic environment when it was transcriptionally active and inactive, a switch in environment was detected. When inactive the locus resides in an inactive environment, while active it mainly contacts other active genes (Simonis et al. 2006). This finding is in agreement with a study using Hi-C where it was shown that gene dense, and transcriptionally active regions segregate in nuclear space from the inactive, gene poor regions (Lieberman-Aiden et al. 2009). Also earlier work, investigating the relative positioning of gene loci using FISH, showed the clustering of active genes, leading the authors to propose a need for genes to co-localize for their transcription regulation (Osborne et al. 2004). This is an interesting concept of gene regulation, which is observed for the ribosomal genes that are transcribed by RNA polymerase I in the nucleolus. The nucleolus can be defined as a dedicated nuclear substructure that is newly formed after each cell division and is dependent on RNAPI activity for its integrity (Carmo-Fonseca et al. 2000). Likewise, RNAPII regulated genes could be transcribed in so called 'transcription factories'. Evidence for the existence of such factories is based on the observation of RNAPII foci in fixed cells (Iborra et al. 1996). Depending on the cell type up to 8000 of such foci were detected that were calculated to consist of about eight polymerases (Jackson et al. 1998; Pombo et al. 1999; Martin and Pombo 2003). However, the examination of living cells expressing a fluorescent version of RNAPII did not result in the detection of these foci, but showed a more homogeneous distribution pattern (Kimura et al. 2002). Moreover, it was suggested by others that the clustering of active genes may be more related to their location within decondensed chromatin and their association with SC35-enriched

speckles (Brown et al. 2008), also known as splicing speckles. Splicing speckles are aggregates of factors involved in the processing of RNA transcripts, but a clear mechanism by which these speckles would organize the genome remains elusive. Investigating the causal relationship between transcription and DNA folding is intrinsically difficult (Nunez et al. 2009; Grimaud and Becker 2010). For illustration, similarly the timing of DNA replication shows a high correlation with gene transcription (Schubeler et al. 2002). This could be explained by the increased accessibility of transcribed DNA for the replication machinery, but also the reverse might be true (Gilbert et al. 2010; Ryba et al. 2010). An experiment that in part was able to distinguish between the cause and consequence of chromosome organization and gene transcription is investigating the role of the noncoding RNA Xist in organizing the folding of the inactive X-chromosome in female neural precursor cells (Chapter 7). Upon depletion of Xist, we found the randomly organized inactive sequences of the inactive X to re-fold into an organized structure similar as was found on the active X. The observed segregation of 'potentially active' and inactive sequences occurred without gene re-activation or promoter de-methylation. Although we cannot exclude an increased RNAPII recruitment to the inactive X in absence of Xist, we show that chromosome folding is not dictated by the clustering of actively transcribed genes or DNA methylation (Splinter et al. 2011).

Apart from the nucleolus, transcription factories and splicing speckles, also nuclear substructures are present that affect the positioning and contacts of inactive loci. Polycomb (PcG) bodies consisting of PRC1 complexes are present that are involved in maintaining the H3K27me3 heterochromatin mark (Spector 2006). PcG repressed *Hox* genes in *Drosophila* were reported to be recruited to PcG bodies and PcG was found to mediate long range chromosome interactions between PcG repressed genes (Bantignies et al. 2011; Tolhuis et al. 2011). Also the affinity of heterochromatin for the nuclear lamina plays a role in defining nuclear topology (Misteli 2004). This can be observed by the relative positioning of whole chromosomes in the nucleus by FISH (Bourgeois et al. 1985; Croft et al. 1999; Zhang et al. 2007) or in more detail by DamID (Pickersgill et al. 2006; Guelen et al. 2008). Future experiments will reveal how the diversity and often redundant relationships between all these factors impacting on chromosome topology contribute to the regulation of gene expression.

Here we have provided a brief introduction into the packaging of DNA in the nucleus and how this relates to gene expression. This is continued in chapter 2, where we discuss the relation between the functioning of cis-regulatory elements and folding of the DNA in more detail. But first an overview of the major 3C(-based) methods is provided that have been instrumental in investigating chromosome organization.

3C based methods

Investigating the three dimensional organization of the genome has classically been based on microscopy. Techniques such as electron microscopy (EM) or fluorescent in situ hybirdisation (FISH) have proven powerful tools in determining the position and organization of DNA elements inside the nucleus. With the development of Chromosome Conformation Capture (3C) (Dekker et al. 2002) and 3C derivative methods like 4C (Simonis et al. 2006), 5C (Dostie et al. 2006), Hi-C (Lieberman-Aiden et al. 2009),

ChIP-loop (Horike et al. 2005), ChIA-pet (Fullwood et al. 2009) the toolbox to investigate chromosome topology has greatly expanded (reviewed by (van Steensel and Dekker 2010)). All C-based methods, of which an overview is given in Figure 1, aim to detect physical interactions between distant DNA fragments and the chromatin loops that are consequently formed (reviewed by (Simonis et al. 2007)). In short, the nuclear organization is captured by fixating cells with formaldehyde. Subsequently the chromatin is digested with a restriction enzyme creating aggregates of DNA fragments, originally co-localizing in nuclear space. The fragments are then ligated after which the cross-links are reversed and the DNA is purified. At this stage the ligation junctions can be analyzed by PCR to detect interactions between specific fragments (**3C**-technology), or the 'interaction-library' can be further processed for more advanced analysis. **4C** technology provides a measure for all fragments that were in close proximity to a fragment of

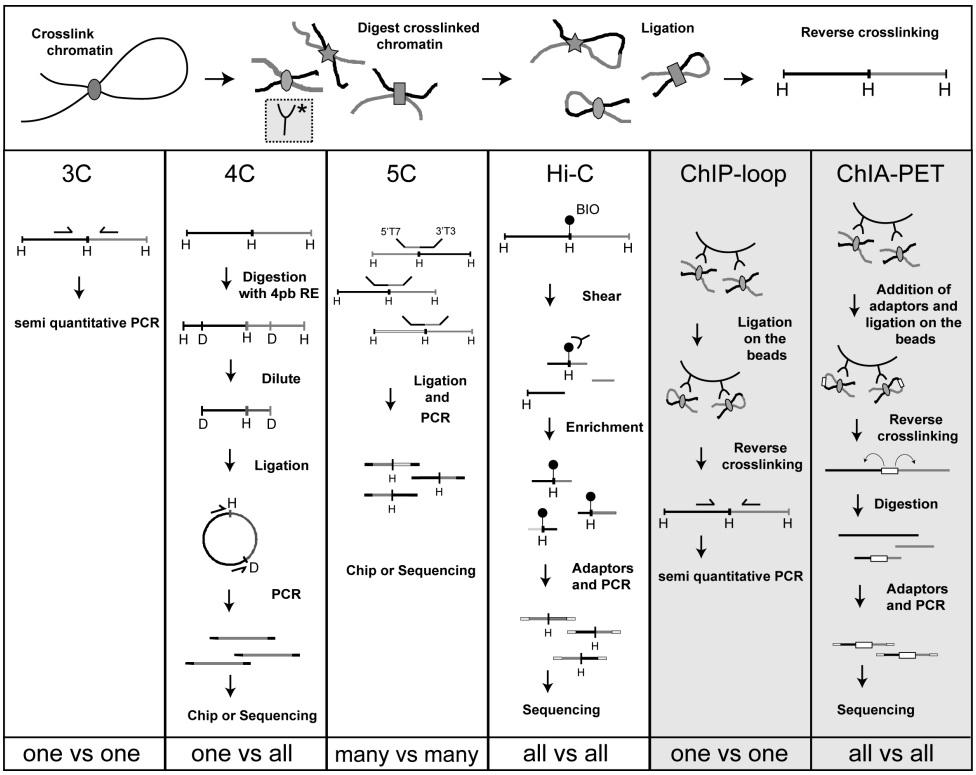


Figure 1: An overview on the methodology of the major 3C based technologies. The horizontal panel on top shows the steps that are shared between the technologies, namely: cross linking the cells with formaldehyde, digestion of the cross linked chromatin to form aggregates of co-localizing sequences, proximity ligation followed by reverse cross linking. When conducting the ChIP-loop or ChIA-PET assay, prior to the ligation step an additional chromatin immuno precipitation is performed. This is indicated by the schematic antibody drawn in the grey box. The vertical panels each depict a different variant of the 3C method depicting the subsequent steps of the method ending by the way interactions are measured; PCR, Chip or Sequencing. The most important difference between the different methods is the throughput by which interactions are detected. This is plotted below each method ranging from: one vs one to all vs all.

choice. This is achieved by further processing the interaction-library with a restriction enzyme followed by a ligation step. The now generated DNA circles allow the amplification of the interacting fragments which can be analyzed either by dedicated microarrays or next generation sequencing. The **5C** method is preferred when aiming to generate a 3D interaction model of selected loci. This method simultaneously interrogates ligation events of multiple restriction fragments by using combinations of fragment specific primers that are ligated while hybridized on the interaction-library. After amplification with 5'T7 and a 3'T3 oligo's, interaction frequencies are determined by next generation sequencing. **Hi-C** generates the most complex data set by interrogating the ligation frequency of all DNA fragments present in the nucleus. Due to this complexity and the amount of sequencing required, at this moment the resolution by which interactions are detected is around 1 Mb in mammalian cells, making this method especially suited for a global analysis of genome topology. In contrast to the previously mentioned 3C-based methods, Hi-C requires an immuno-precipitation (IP) to enrichment for the ligation junctions. For this purpose a biotin labeled nucleotide is incorporated during the preparation of the interaction-library, by which the junctions are retrieved after library fragmentation. Adaptors are added enabling a PCR amplification step and the junctions are analyzed by next generation sequencing.

In order to direct the analysis of spatial interactions between fragments engaged in transcription, or bound by a specific transcription factor, an IP-step was included in the basic 3C protocol, resulting in the development of the **ChIP-loop assay** and **ChIA-pet**. The ChIP-loop assay interrogates ligation frequencies between specific fragments (like 3C), while ChIA-pet, like Hi-C, generates interaction profiles that theoretically are based on questioning all fragments present in the genome. Although potentially interesting, the addition of the IP could introduce biases that are difficult to control for. For example the affinity of the antibody could be much stronger for aggregates where both DNA fragments are bound by the transcription factor of interest compared to those with a single bound fragment. Thus far, the data generated by these methods indeed show preferential interactions between DNA fragments occupied by the same transcription factor (Horike et al. 2005; Cai et al. 2006; Kumar et al. 2007; Fullwood et al. 2009; Schoenfelder et al. 2009), which could potentially be the result of such a precipitation bias. Future experiments will determine the value of these applications in uncovering chromosome topology.

2

THE COMPLEX TRANSCRIPTION REGULATORY LANDSCAPE OF OUR GENOME: CONTROL IN THREE DIMENSIONS

Erik Splinter and Wouter de Laat

Hubrecht Institute-KNAW & University Medical Center Utrecht, Uppsalalaan 8, 3584 CT Urecht, The Netherlands

Accepted for publication in *EMBO Journal*

ABSTRACT

The non-coding part of our genome contains sequence motifs that can control gene transcription over distance. Here, we discuss functional genomics studies that uncover and characterize these sequences across the mammalian genome. The picture emerging is of a genome being a complex regulatory landscape. We explore the principles that underlie the wiring of regulatory DNA sequences and genes. We argue transcriptional control over distance can be understood when considering action in the context of the folded genome. Genome topology is expected to differ between individual cells, and this may cause variegated expression. High resolution three-dimensional genome topology maps, ultimately of single cells, are required to understand the cis-regulatory networks that underlie cellular transcriptomes.

INTRODUCTION: GENE REGULATION BY REMOTE DNA ELEMENTS

Over 200 cell types exist within the human body, each being different in morphology and function, yet all containing the same genome. These differences are driven by cell-specific gene regulatory programs. Their proper execution not only relies on the protein-coding parts of genes, but also on that of non-coding sequences within and surrounding the genes. This was first realized when a deletion outside the β-globin gene was found to cause aberrant expression of the otherwise intact gene in a thalassaemia patient (Van der Ploeg et al. 1980). It is now well recognized that sequence alterations in the non-coding part of our genome, previously known as 'junk DNA', frequently drive the deregulation of critical genes to cause congenital and somatically acquired diseases (Kleinjan and van Heyningen 2005). These alterations can locate at large distances from the relevant genes. One extreme example is where point mutations in the DNA located ~1Mb away from the *Sonic Hedgehog* (*SHH*) gene interfere with its regulation. This has been linked to preaxial polydactyly, a condition where patients suffer from limb malformations (Lettice et al. 2003). Chromosomal rearrangements may remove, mutate or separate distant regulatory sequences from genes, but may also lead to so called 'position effects', where genes are placed adjacent to new regulatory sequences that drive their up- or downregulation. This is seen for example in B-cell lymphomas and T-cell leukemias, when translocations juxtapose strong regulatory sequences of antigen receptor loci to proto-oncogenes. How frequent diseases are caused by genetic changes outside gene bodies is unknown, because most large-scale screening efforts only focus on exome integrity. Based on an examination of the database of genome wide association studies (GWAS) (Hindorff et al. 2009), mutations in non-coding regions were estimated to contribute to a staggering 40% of disease cases (Visel et al. 2009b). No matter the exact frequency in disease, studies like these clearly underscore the importance of cis-linked regions in regulating gene expression and controlling normal development. Current research efforts in functional genomics therefore focus on uncovering the full regulatory potential of our genome.

Here, we will review studies that aim to identify and assign function to the regulatory DNA sequences of the mammalian genome

and discuss how regulatory DNA elements and genes are wired to properly execute cell-specific transcription regulatory programs.

Identifying and classifying regulatory DNA sites in the genome

Regulatory DNA sequences in the genome can be identified through various methods. Perhaps the most thorough assay is to screen for DNaseI hypersensitive sites (HSs) (Crawford et al. 2004; Dorschner et al. 2004). HSs typically are sites where the nucleosome fiber is locally disrupted, presumably through the action of DNA-binding proteins and associated factors. Any given human cell type appears to have ~100,000 or more HSs (Boyle et al. 2008). This includes promoter sequences of active genes, but the majority is located away from transcriptional start sites (TSS) in non-coding DNA.

To understand their function, we can test and categorize their activity. This is traditionally done using property defining assays, often in vitro, in which the isolated sequence element is placed in plasmids carrying a reporter gene. When transfected into cells, they may then activate or repress transcription, or neutralize transcriptional activation when placed between an activator and gene promoter. Accordingly, they are classified as an enhancer, silencer or insulator. A more laborious but often more revealing assay is to stably integrate an isolated site with a reporter gene in the genome, and measure transcriptional activity. By doing this in transgenic animals, one can also determine an element's tissue-specific activity and define other properties not appreciable in plasmid-based assays. Boundary elements, for example, are defined based on their property to protect against position effects when placed around an integrated transgene (Recillas-Targa et al. 2002), an ability not well appreciated using a plasmid-based assay. Similarly, locus control regions (LCRs), which often behave as enhancers in plasmid-based assays, have the additional capacity to confer tissue-specific, position-independent and copy-number dependent expression on linked reporter genes when stably integrated in transgenic animals (Grosveld et al. 1987). This defining property is not shared by classic enhancers and can only be appreciated when tested in transgenics.

While the categorization of DNA elements is useful, an activity picked up for a DNA motif in a reporter assay may not be relevant or detectable at its natural chromosomal context (Dillon and Sabbattini 2000). This has been demonstrated for insulators, enhancers and LCRs, (Epner et al. 1998; Bender et al. 2006; Splinter et al. 2006; Ahituv et al. 2007) (Table 1). Thus, DNA sequences may intrinsically harbor specific activities, and hence be classified accordingly, but the linear and, as we will argue below, three-dimensional chromosomal context often determines whether they exert this activity or not.

With 100,000 or more HSs per cell-type, mammalian genomes are emerging as highly complex regulatory landscapes. Alternative strategies for the exposure and classification of regulatory sites confirm this idea. Genome wide chromatin immunoprecipitation approaches (ChIP-chip and ChIP-seq) have been developed to identify specific classes of regulatory sites based on their unique chromatin signature. For example, enhancers were shown to have high H3K4me1 and promoters high H3K4me3 levels (Heintzman et al. 2007). Based on their H3K4me1 mark in two human cell types, ~55,000 potential enhancer elements were uncovered (Heintzman et al. 2009). Nearly 80% of the enhancers were unique for the one or the other cell type, despite the two sharing 85% of their

Table 1: Property defining activities measured for DNA sequences in reporter assays may not be detectable at their natural chromosomal context. [1] Note that the β-globin LCR is not necessary for maintenance of open chromatin in mice, but does appear necessary for this in humans (Forrester et al. 1990).

Genomic site	Definition	Property defining reporter assay	Effect of genomic site deletion
β-globin: HSs upstream of genes	LCR (Grosveld et al, 1987)	Confers position independent, copy-number dependent transgene expression in mice	No heterochromatinization, basal non-activated gene expression (Epner et al, 1998)[1]
CTCF sites flanking β-globin locus	Insulators (Farrell et al, 2002)	Block enhancer activity and shield reporter genes in plasmid-based assays	No heterochromatinization, no transcription changes measurable (Splinter et al, 2006; Bender et al, 2006)
Ultra-conserved elements close to Dmrt3, Rcn1, Pola1 and Sox3	Enhancers (Ahituv et al, 2007)	Tissue-specific reporter gene activation in transgenic mice	No transcription changes measurable (Ahituv et al, 2007)

active genes. This underscores the idea that enhancers are tissue-specific elements acting on tissue-specific genes. At the same time it raises the question whether all these sites are actively involved in gene regulation. A further sub-classification of H3K4me1 enhancers is now made based on the shared presence of p300 protein or acetylated H3K27. Both signatures appear to separate the active from the poised pool of enhancers. In the case of p300, a co-activator protein that can acetylate H3K27, several thousands (instead of tens of thousands) active enhancers were identified in various primary mouse tissues (Visel et al. 2009a). Similar numbers of active enhancers were found based on the H3K27Ac profile (Creyghton et al. 2010; Rada-Iglesias et al. 2011). Collectively these studies show that a given mammalian cell type contains thousands or more regulatory sites. With 200 different cell types, this confirms that our genome harbors a complex regulatory landscape (Fig. 1).

The mammalian genome: a complex regulatory landscape

The term 'regulatory landscape' was first coined to describe the complex organization of regulatory sequences around the Hox loci (Spitz et al. 2003). A recent study provided further insight into the complexity of mammalian gene regulation (Ruf et al. 2011). In this study, hundreds of transgenic mice were generated, each carrying a transposable reporter cassette containing a LacZ reporter gene driven by a minimal promoter sequence (50 bp of the human β-globin gene) inserted at a random location in the genome. Each was then analyzed for LacZ expression in E11.5 embryos, and categorized according to tissue-specificity of expression. The minimal promoter by itself was insufficient to drive reporter gene expression, yet nearly 60% of the transgenics showed reporter activity, demonstrating that the majority of genomic locations harbor activating potential. Among the activated transgenics, only a very small percentage showed ubiquitous expression, while the great majority (>95%) demonstrated restricted, tissue-specific expression (Fig.2). Often, but

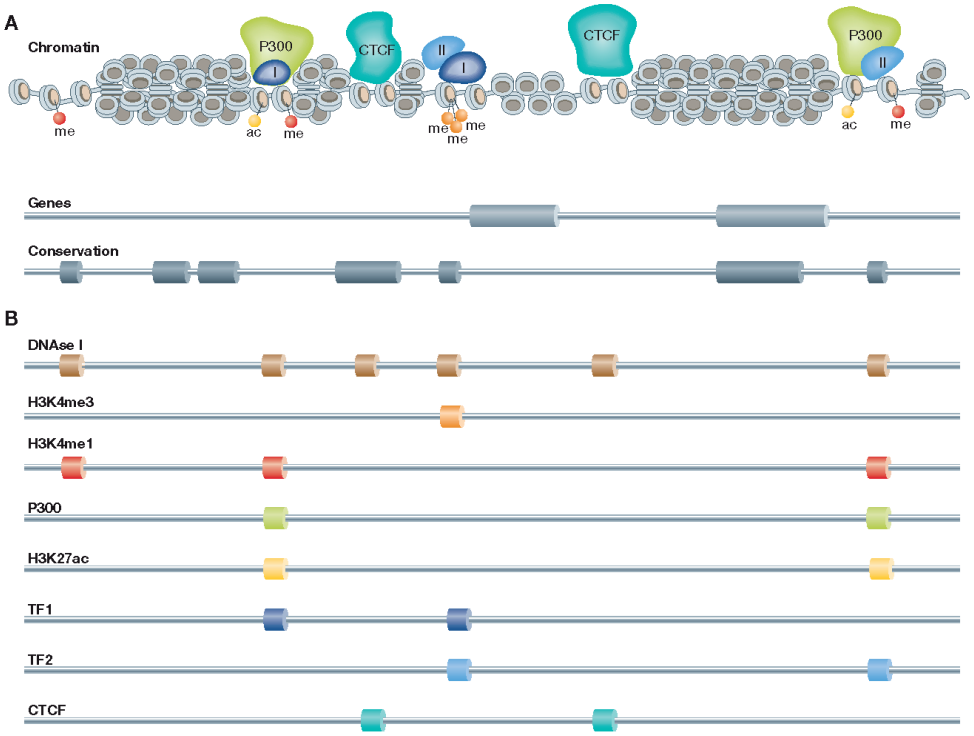


Figure 1: The complex regulatory landscape of the genome. A: DNA packaged into chromatin with regulatory proteins P300, Transcription Factors I and II, CTCF, and histone modifications present. Underneath the chromatin fiber the position of genes and conserved elements, representing information that can be found in databases (e.g. UCSC, Ensembl). **B:** The identification of accessible chromatin (DNAseI profile) and various chromatin marks (ChIP-seq) that are associated with regulatory elements yields a picture of a complex regulatory landscape.

not always, reporter expression followed that of the nearest endogenous gene. However, tissue-specific expression was also found for integration sites near genes with widespread expression patterns. This argues that also the expression of so-called housekeeping genes is modulated in a tissue-specific manner. The flanking transposon sites further enabled the investigators to locally hop around the reporter cassette and assess the regulatory impact at multiple locations within one chromosomal domain. Essentially all possible outcomes were found. Sometimes transposition over only a few kilobases resulted in clearly distinct expression patterns, whereas in other examples integration sites separated by hundreds of kilobases gave essentially the same expression profile (Ruf et al. 2011). This led the authors to propose that the genome harbors a 'regulatory jungle': chromosomal regions contain many regulatory sites that can activate gene expression over large distances and others that counteract this activity.

How are all these linearly organized activities orchestrated such that genes are expressed at the right time and place? Genetic studies that manipulated the non-coding part of the genome to understand transcriptional control have delineated some primary rules of promoter-enhancer engagement that must

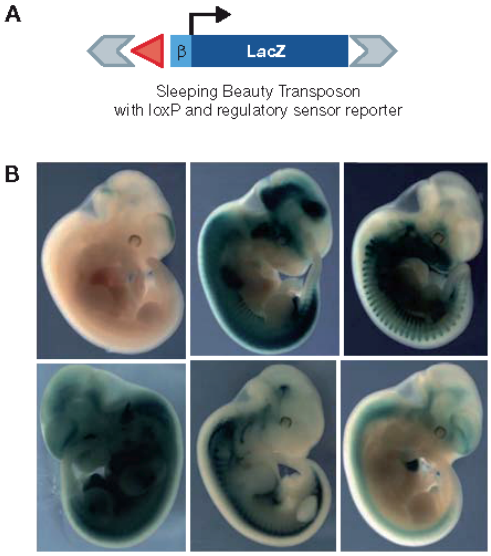


Figure 2: Random integration of a reporter cassette reveals the regulatory potential of the mouse genome. A: Schematic representation of the randomly integrated reporter cassette consisting of the LacZ gene driven by the beta-globin minimal promoter. **B:** Examples of LacZ expression patterns found in the different transgenic lines which, depending on the integration site, range from global to a highly restricted expression pattern (Reproduced with permission from Ruf et al, 2011).

be considered when addressing this question. So far, most of these studies focused on gene clusters such as the Hox and globin loci, aiming to understand how transcription of their individual genes is coordinated in time and space.

Rules of engagement in a complex regulatory landscape: (1) linear proximity.

It is often assumed that enhancers act on the nearest genes in *cis*, and in many cases this is correct. Proximity on the linear DNA template, or genomic order, is a major determinant of selectivity: first-come, first-served (Fig. 3, rule I). This was first demonstrated in plasmid-based competition assays, which showed that proximal genes have an advantage over distal genes to be activated by a shared enhancer (de Villiers et al. 1983; Wasylyk et al. 1983). The same principle was then shown to also apply to transcription regulation in the chromosomal context (Hanscombe et al. 1991; Dillon et al. 1997). Interestingly, order is no longer important when two competing genes are positioned close together at a large distance from a shared enhancer (Heuchel et al. 1989; Dillon et al. 1997). While linear proximity to a regulatory site is often a good predictor of target genes, many examples exist where enhancers ignore the nearest genes and specifically act on genes further away (de Laat and Grosveld 2003). The previously mentioned limb bud-specific enhancer of the *SHH* gene is one such example, present in an intron of the *Lmbr* gene, but acting on the *SHH* gene one megabase away. Why do enhancers not always stick to the 'first-come, first-served' rule, but sometimes act instead on more distal genes?

Rules of engagement in a complex regulatory landscape: (2) promoter specificity

Enhancer reporter assays in cell lines (in vitro) and mice (in vivo) usually analyze the capacity of test sequences to activate one and the same minimal or general promoter (Banerji et al. 1981; Visel et al. 2008; Ruf et

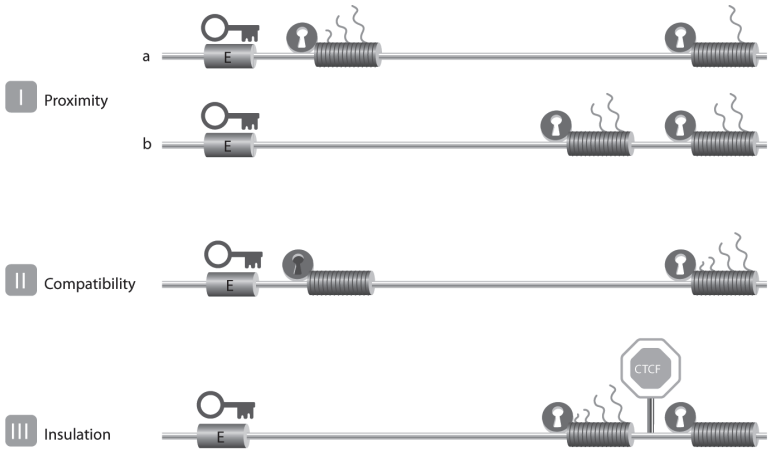


Figure 3: The three rules of engagement that dictate enhancer-promoter interactions. I: Proximity. (**a**) When multiple genes (cylinders) are compatible with (open lock) and relatively close to a shared enhancer (E), the most proximal gene is preferentially activated over the distal gene (represented by the number of transcripts originating from the gene). (**b**) This competitive advantage disappears when both genes are located far away from the shared enhancer. **II**: Compatibility. Enhancers ignore the 'first comes, first served' rule when the proximal promoter is incompatible (closed lock) with the enhancer. Result: activation of the distal gene. **III**: Insulation. The presence of CTCF can block enhancer function across its binding site and prevent a compatible gene from being activated by the enhancer.

al. 2011). The assumption is that enhancers and promoters show little specificity and will always act on each other. While this may be true for artificial constructs where the enhancer and minimal promoter are close together, it appears more complex in the context of the genome where regulatory sites do show target specificity. There are numerous examples of enhancers interacting with just a subset of equally nearby target promoters, and there are different possible explanations for this. Sometimes an enhancer is thought to ignore a gene because its promoter is not accessible in the tissue where the enhancer is active. In the β-globin locus for example, the LCR exclusively acts on the β-globin genes and totally ignores the nearby olfactory receptor (OR) genes, even when potentially interfering insulator sites are disrupted (Splinter et al. 2006). The LCR also ignores the nearby fetal globin genes to exclusively act on more distal adult β-globin genes at later stages of development. Inactivity, or promoter inaccessibility, may therefore be a reason for regulatory sites to skip nearby genes (Fig. 3, rule II).

Selective gene activation can also be explained by promoter competition, whereby the activation of one promoter precludes the activation of another equidistant promoter. This was shown clearly in Drosophila transgenic embryo assays, where different enhancers were demonstrated to prefer distinct classes of promoters, depending on the presence of certain core promoter elements (Ohtsuki et al. 1998; Butler and Kadonaga 2001).

Promoter competition also occurs between mammalian genes. It manifests itself when the deletion of one or more genes impacts on the expression of remaining neighboring genes, or when the deletion of a regulator causes downregulation of multiple genes. These phenomena have

been described at the Hox (Spitz et al. 2003; Tschopp et al. 2009) and globin gene clusters (Forrester et al. 1990; Hanscombe et al. 1991; Epner et al. 1998; Sabatino et al. 1998; Lower et al. 2009). The implication of gene competition is two-fold: the first is that regulatory sites exist in the genome that can act on multiple endogenous genes. Evidence exist that the number of genes controlled by a given regulatory site probably depends on its chromosomal context. At its natural location, the β-globin LCR activates maximally two or three β-globin genes at any given developmental stage. However, when tested without globin genes at a defined new location in the genome it was found to activate 6-7 genes up to 150 kb away in *cis* (Noordermeer et al. 2008), and two genes in trans (see below) (Noordermeer et al. 2011). A second implication of gene competition is that enhancer-promoter interactions are, at least for some time, mutually exclusive. Convincing evidence for this was obtained in single cell experiments that measured the ongoing transcriptional activity of LCR-competing β-globin genes in cells with traceable transcriptional history. The LCR was demonstrated to activate only a single gene at the time, but over time to dynamically flip-flop between competing globin genes (Wijgerde et al. 1995). This work provided insight into the mechanism behind competition between clustered genes for shared regulators.

In a genome described as a 'regulatory jungle' (Ruf et al. 2011) one might expect that regulators sometimes also 'fortuitously' activate more than their presumed target gene. Indeed, bystander activation has been observed in several cases. The B-cell specific human immunoglobulin-beta (Igbeta) gene (or CD79b) is highly expressed, but presumably not functional, in pituitary due to its linear proximity to the LCR that is acting on the more distal growth hormone (hGH) gene in this tissue (Cajiao et al. 2004). A similar phenomenon was recently found at the human α-globin locus that is located in a gene-dense chromosomal region with many housekeeping genes. When a 500 kb region around the locus was analyzed for gene expression levels, a gene called NME4, 300 kb apart, was found to be upregulated specifically in red blood cells where the α-globin genes are active. Subsequent analysis showed that the α-globin regulatory sequences were responsible for this, and that NME4 competes with the α-globin genes for these enhancers (Lower et al. 2009). Collectively, these data show that tissue-specific enhancers not always exclusively activate transcription of their real target genes but sometimes also that of unrelated genes nearby in *cis*. Based on arguments explained below, we predict this bystander activation will be seen even more often when transcriptome analysis is performed at the single cell level rather than the cell population level.

Rules of engagement in a complex regulatory landscape: (3) insulators can block enhancer-promoter interactions.

The original idea that led to the discovery of insulator elements was that the genome may be partitioned in physically separate chromatin domains that each have their own independent regulatory activities. If true, the assumption was that boundaries must exist that prevent regulatory cross-talk between these domains. To test this hypothesis, two types of assays were developed. One investigated the ability of sequence elements flanking a reporter gene to overcome gene repression when integrated into heterochromatin (Kellum and Schedl 1991). Another analyzed whether an element can

block enhancer activity when positioned in between the enhancer and a gene promoter (Kellum and Schedl 1992). With the discovery of more and more regulatory sites in the mammalian genome that ignore nearby genes to act specifically on much more distal genes, the idea that regulatory activities are strictly separated along the linear chromosome template had to be adjusted (Dillon and Sabbattini 2000; de Laat and Grosveld 2003). Nevertheless, the assays were proven to be very useful to identify an intriguing new class of regulatory sites known as 'insulators'. In mammals, one protein in particular is associated with insulator activity: CTCF (Chung et al. 1993; Bell et al. 1999). In *in vitro* reporter assays, CTCF bound to DNA often acts as an enhancer blocker (Fig. 3, rule III). *In vivo*, CTCF binding sites are found next to genes that are active in an otherwise repressive chromatin surrounding, such as the human and mouse β-globin genes (Farrell et al. 2002) and near genes on the silenced X chromosome that escape X inactivation (Filippova et al. 2005). CTCF also acts as an allele-specific enhancer-blocker to mediate imprinted gene expression at the H19-Igf2 locus (Bell and Felsenfeld 2000; Hark et al. 2000). Interestingly, a single CTCF site can function as an insulator, but the introduction of a second CTCF site in between an enhancer and promoter often alleviates the enhancer blocking effect (Cai and Shen 2001; Muravyova et al. 2001). ChIP-Seq experiments were carried out to obtain a genome-wide picture of CTCF binding sites. An estimated 20,000 to 87,000 CTCF binding sites exist in the human genome (Barski et al. 2007; Kunarso et al. 2010), meaning that on average a CTCF binding site is present every 35-155kb. Most of these sites (50-90%) appear conserved between different cell types (Barski et al. 2007; Chen et al. 2008; Cuddapah et al. 2009; Kunarso et al. 2010). They are enriched at boundaries between repressive (H3K27me3-rich) and active (H3K27me3-poor) chromatin (Cuddapah et al. 2009), and at the borders between so-called lamin-associated domains (LADs) and non-LADs, being chromosomal regions that preferentially locate to the periphery or the interior of the cell nucleus (Guelen et al. 2008). Furthermore, genes separated by a CTCF site show a markedly reduced correlation in gene expression (Xie et al. 2007). Collectively these studies demonstrate that, while not all CTCF sites will act as insulators, clearly their location, and that of possible other insulators, must be taken into account when considering the wiring of regulatory DNA networks in the genome. Interestingly, substantial overlap exists between the genomic binding sites of CTCF and cohesin (Parelho et al. 2008; Wendt et al. 2008), suggesting that cohesin, a protein complex that can hold two DNA helices together may assist CTCF in its function to separate regulatory activities.

In summary, genetic experiments have uncovered three rules of enhancer-promoter engagement: linear proximity matters, enhancer and promoter need to be compatible, and some sequences exist that can block enhancer activity. To understand the molecular mechanisms behind these rules we need to know how regulatory sites exert activities over distance.

Chromatin looping and spatial interactions between regulatory sites

The idea that DNA topology can play an important role in long-range gene activation was originally based on observations on bacterial and phage repressor proteins, like the Gal, AraC, and λ repressor proteins. They were found to only function when

homo-multimerized and bound to separated operator sequences and this was shown by electron microscopy to result in looping out of the intervening DNA fiber (Ptashne 1986). The first direct evidence in eukaryotes for spatial interactions between regulatory sites and their target genes was provided by two independent studies on the mouse β-globin locus. One study involved the use of RNA-TRAP to demonstrate a chromatin loop between an active β-globin gene and HS2 of the LCR (Carter et al. 2002). The other study applied 3C technology and showed that not only HS2 but also other β-globin regulatory sites participate in the interactions with the genes (Tolhuis et al. 2002). The spatial chromatin entity they formed was called an Active Chromatin Hub (ACH) (Fig. 4). 3C (chromosome conformation capture) technology turned out to be the method of choice for unraveling the three dimensional regulatory circuitries at gene loci. The technique relies on cross-linking and ligation of spatially proximal DNA sequences, which are then identified and quantified with PCR strategies (Dekker et al. 2002). In subsequent studies on the β-globin locus it was demonstrated that during development the genes dynamically switch their physical interactions with the LCR in relation to their change in gene expression (Palstra et al. 2003) and that transcription factors mediate these long-range DNA interactions (Drissen et al. 2004; Vakoc et al. 2005). 3C also provided evidence for chromatin loops between distal enhancers and target genes at other loci, including the interleukin (Spilianakis and Flavell 2004), α-globin (Vernimmen et al. 2007) and immunoglobulin loci (Liu and Garrard 2005), and showed that chromatin loops can also be involved in gene repression (Horike et al. 2005; Comet et al. 2006; Tiwari et al. 2008; Bantignies et al. 2011; Comet et al. 2011). In a recent study, both microscopy and 3C data showed that the distal limb bud enhancer of *SHH* loops to the gene (Amano et al. 2009).

Collectively, these and other studies firmly establish that regulatory sites act on genes via chromatin looping. This knowledge argues

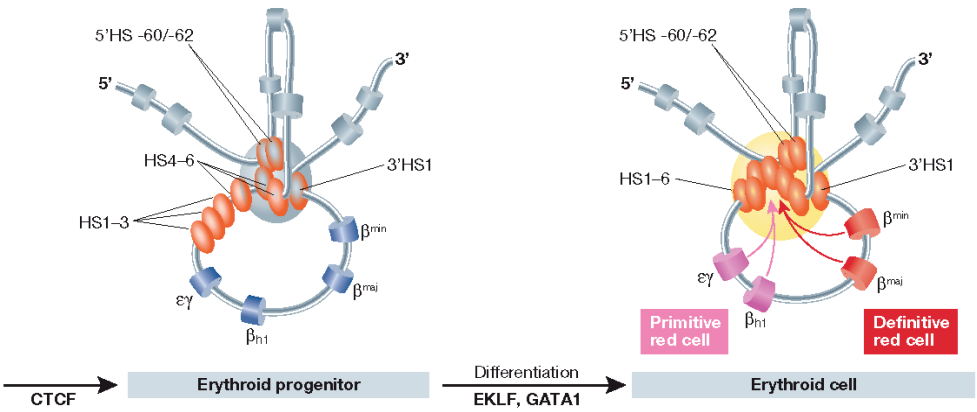


Figure 4: A schematic representation of the three dimensional structure of the mouse β-globin locus during differentiation. In erythroid progenitors a 'Chromatin Hub' (CH) is formed by the clustering of the CTCF sites that surround the β-globin locus. Upon differentiation the remaining hypersensitive sites of the LCR and the globin gene that is activated participate in this clustering to form an 'Active Chromatin Hub' (ACH). Hypersensitive sites are represented by the ovals and genes by the cylinders. In grey the Olfactory Receptor genes that surround the locus are depicted. A developmental switch in gene usage occurs between primitive and definitive red cells.

that we must know the topological constraints of chromatin in order to understand how transcription is controlled in the regulatory jungle of the genome. Question therefore is: how do separated genomic sites find each other?

Rules of engagement in three-dimensions: proximity matters

When we ignore all biological activity exerted by DNA-binding proteins and associated factors, a chromatin-packed chromosome fiber is expected to fold and behave essentially as a polymer with a given flexibility. This implies that random collisions will take place between pairs of sites present on a chromosome, with contact probabilities that exponentially decrease with increasing site separation on the linear template (Rippe et al. 1995). This is exactly what is measured in vivo by Hi-C technology (Lieberman-Aiden et al. 2009; Duan et al. 2010; Tanizawa et al. 2010) (Fig. 5a). Hi-C is a high-throughput genomic variant of 3C that analyzes interactions between all genomic sites. A smooth, continuous and exponential decline in contact probability is seen with increased site separation when Hi-C measurements for all pairs of sites are plotted in relation to their genomic distance. Beyond certain distances the curves plateau, indicating that contact probabilities then no longer depend on the amount of intervening DNA. For human chromosomes, this seems to happen only beyond 90 megabases of DNA or more. Of relevance, the data also show that two given sites on one chromosome, no matter their linear site separation, are much more likely to contact each other

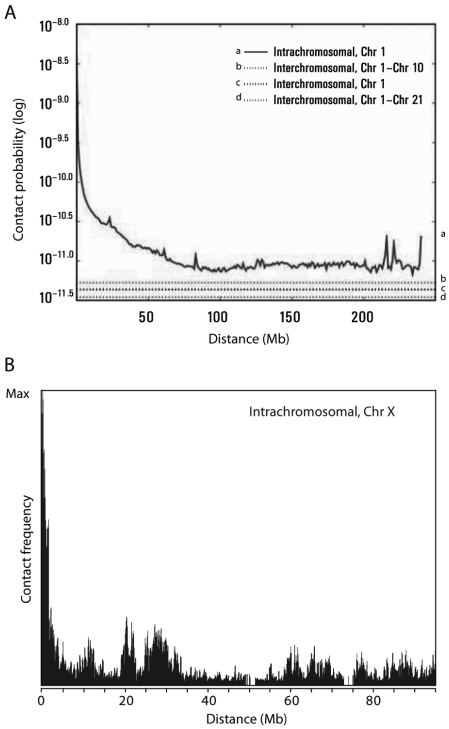


Figure 5: The relationship between contact frequencies (spatial distance) and the position of sequences on the DNA template. A: Hi-C experiments, showing the average contact probabilities between all pairs of sequences (within and between chromosomes). The average profile shows that with an increasing linear distance the interaction frequencies drop exponentially (Reproduced with permission from Lieberman-Aiden et al, 2009). **B:** 4C experiment that measures interaction frequencies for an individual locus ('viewpoint') versus all other genomic sequences. The same global anti-correlation between distance and contact frequency is seen, but on top specific interactions (peaks) are observed at sequences that are preferentially contacted by the viewpoint.

than to contact a given site on a different chromosome (Lieberman-Aiden et al. 2009).

Thus, Hi-C data confirm what was predicted from polymer physics: any given site in the genome has the ability to contact neighboring sequences, the chance of which decreases exponentially with increasing site separation on the linear template. This very basic concept of chromosome flexibility can be assumed to underlie the first rule of engagement: linear proximity matters. Without activities interfering with local chromatin flexibility, any regulatory site is much more likely to contact a gene nearby than a gene far away on the chromosome. The genetic observation that linear order no longer matters when two similar genes are close to each other but far apart from a shared enhancer (Dillon et al. 1997) is also in line with a polymer-like behavior of chromatin: both will have a roughly equal chance of colliding into the enhancer.

Rules of engagement in three-dimensions: affinity matters

While polymer physics dictates, and Hi-C data shows, that collisions between linearly proximal sequences will be frequent, they will be a-specific and equally transient for all pairs of sites unless two sites have increased affinity. This can be mediated for example via the proteins associated to them. Upon collision of two such sites, a stabilization of the contact will occur, which will be measurable as a chromatin loop. This is what is seen by 4C technology (Fig. 5b). 4C technology is a different high-throughput version of 3C technology that produces high resolution DNA contact maps for individual genomic sites. A 4C profile typically shows the same anti-correlation between contact probability and genomic distance as seen by Hi-C, but the curves are no longer smooth: peaks and valleys of contacts are imposed on it (Simonis et al. 2006; Simonis et al. 2007). A peak in a 4C contact profile identifies a preferred interaction site and reveals a chromatin loop between this site and the selected target sequence ('viewpoint'). Since 3C-based methods provide average topology impressions of populations of cells, such a loop must have been present in a relevant proportion of the cells the moment they were fixed. In several instances it was demonstrated that the stability of such chromatin loops indeed relies on the associated proteins (Drissen et al. 2004; Vakoc et al. 2005; Splinter et al. 2006).

One can speculate that the formation and stabilization of a given chromatin loop has two important consequences: one is that for the duration of the interaction the two sites involved may not be available for direct interactions with other genomic sites. The second is that this structure imposes constraints on the flexibility of the intervening and surrounding genomic sites. The latter is expected to be relevant for insulator function, as will be discussed below. The first would clearly be relevant for promoter competition: as long as an enhancer is stably engaged in an interaction with one gene, it cannot interact with another gene (Wijgerde et al. 1995). Thus, the genetically deduced rule of engagement stating that enhancer-promoter compatibility is essential may well be dependent on the affinities between enhancer- and promoter-bound proteins and their ability to stably form a chromatin loop. The outcome of promoter competition, i.e. the number of genes activated and their eventual transcript levels, will therefore depend on their order on the linear template, their relative distance to and their affinity for the enhancer (Fig. 6).

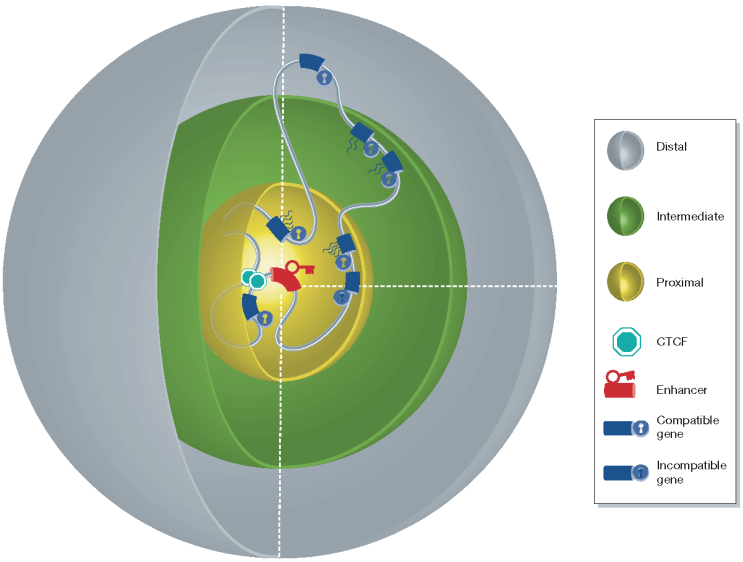


Figure 6: Spatial distance, enhancer-promoter compatibility and sites of insulation are the main determinants of enhancer mediated gene expression. This figure represents a three dimensional model of chromatin 'in action', explaining all 3 rules of engagement by considering regulatory action in three dimensions. Genes proximal to the enhancer in 3D space (in the yellow sphere) have a high probability to be contacted, but activation depends on compatibility and whether or not a gene is isolated from the enhancer (CTCF sites). Compatible genes further away in 3D (in the green sphere) have a decreased probability to be contacted and activated, while compatible genes that too far (grey sphere) will not be activated.

Rules of engagement in three dimensions: insulators and DNA flexibility

If chromatin looping between enhancers and promoters is critical for gene activation, would insulators exert their blocking activity also by acting on DNA topology? An increasing body of evidence suggests they do. A role for insulators in organizing genome topology was first appreciated from microscopy observation in Drosophila, showing that insulator sequences separated in the genome come together in the nuclear space in so-called insulator bodies (Gerasimova et al. 2000). Direct looping between neighboring insulator sites was first demonstrated by 3C technology for CTCF sites at the H19-Igf2 locus and the β-globin locus (Kurukuti et al. 2006; Splinter et al. 2006). Binding of CTCF to an internal site of the imprinted H19-Igf2 locus was shown to create a loop encompassing the Igf2 gene that presumably isolated the gene from enhancers outside the loop (Kurukuti et al. 2006). At the β-globin locus, flanking CTCF sites were demonstrated to form a large chromatin loop containing both enhancers and genes that was further folded later during differentiation into smaller loops to bring stage-specific genes in closer proximity to their enhancers (Splinter et al. 2006). Insulators not only block the activating effects of enhancers, they can also interfere with the spreading of repression mediated by polycomb group proteins (Sigrist and Pirrotta 1997; Mallin et al. 1998). It was noticed that chromatin coating by polycomb proteins will stop at an insulator site, but may continue into more

downstream regions beyond a second insulator site (Comet et al. 2006). The abrupt block at the first site suggests that insulators act as a physical barrier. The continued spreading beyond the second site suggests these downstream regions are brought into spatial proximity of the upstream source of polycomb protein, which indeed was shown to happen as a consequence of loop formation between the two intervening insulator sites (Comet et al. 2011). These and other detailed transgenic studies (e.g. (Maksimenko et al. 2008)), together with genomic evidence that CTCF sites often locate to borders of spatially separated chromatin domains (Guelen et al. 2008), have established that insulator sequences strongly influence chromosome topology. They induce chromatin loops that will affect the flexibility of neighboring and intervening chromatin sites. The impact hereof on gene regulation can be positive, negative or neutral, depending on the regulatory sites they spatially bring together or separate. Further evidence for this was recently provided in a study that performed large-scale mapping of chromatin loops between CTCF sites (Handoko et al. 2011). We expect that within chromosomal segments at any given time enhancer-promoter combination will compete with interspersed CTCF sites for chromatin loop formation, the outcome of which again will depend on linear proximity and affinity between interacting sites.

DNA interactions and gene expression in single cells

For genes that are crucial for cellular identity correct expression must be controllable in all relevant cells. Transcription occurs in bursts, as was shown in life cell imaging studies (Chubb et al. 2006). Chromatin looping is thought to bring additional DNA binding sites for relevant transcription factors in close proximity to gene promoters; this will increase the local concentration of these factors and consequently transcription efficiency will increase (Droge and Muller-Hill 2001; de Laat and Grosveld 2003). Whether enhancers increase the frequency or amplitude of transcription bursts is not yet known, but it is relevant to ask what the maximum amount of intervening DNA is that can be bridged by an enhancer to contact and activate a gene in every single cell. This question is pertinent also because overall DNA topology is thought to be relatively stable during interphase, both at the level of the genome (Chubb et al. 2002; Gerlich et al. 2003) as well as at that of individual chromosomes (Muller et al. 2010).

The previously mentioned study that analyzed ectopically integrated LCRs in the mouse genome aimed to get insight into the capacity of regulatory sites to reach and activate genes over ultra-long distances in individual cells. By 4C technology it was shown that the LCR had little impact on overall genome topology, as the many long-range genomic contacts made by the integration site were similar with and without the integrated LCR (Noordermeer et al. 2011). Transcriptome analysis then demonstrated that none of the contacted genes benefitted from the presence of the new LCR, with the exception of two endogenous mouse β-globin genes located on another chromosome. They were upregulated not in every single cell but exclusively in so-called 'jackpot' cells that showed the interchromosomal interaction. The work uncovered a number of principles behind long-range gene activation. Firstly, it showed that the ability of regulatory sites to search the nuclear interior for preferred target genes is severely limited by their chromosomal context. This is likely to be true for all genomic positions, although the degree of

constraint may vary between sites. Secondly, the data showed that promoter-enhancer compatibility is essential to transform spurious contacts between enhancers and promoters into productive interactions that drive gene activation; of all contacted genes, only two natural target genes of the LCR increased their transcription. Thirdly, the data demonstrated that relatively stable cell-specific genome conformations can induce variegated gene expression. Cells that have their genome folded such that it brings together a regulatory site with a responsive gene can have different transcript levels of this gene than otherwise identical cells (Noordermeer et al. 2011). The phenomenon was referred to as 'Spatial Effect Variegation' (SEV). SEV may happen in *trans*, as shown in this artificial situation, but may also work in *cis*. Possibly, previous work already provided one example for SEV in *cis*. The interaction over 1 megabase between the limb bud enhancer and *SHH* was shown to occur in the expressing limb bud cells in a manner not dependent on the enhancer per se: its deletion abrogated expression without changing the DNA configuration (Amano et al. 2009).

We envision that indeed beyond a certain chromosomal distance regulatory sites will not be able to independently find a target gene, but for contact instead depend on the topological constraints imposed by the remainder sequences of the chromosome. The overall genome topology may or may not bring these sites in relative proximity in single cells. Depending on the nature of the interacting regions, transcription may be activated or repressed in these cells, causing variegated expression across the cell population. It is tempting to speculate that such SEV, when acting on key developmental regulators, may provide cells with a mechanism to make autonomous cell fate decisions, without the need for external signaling.

Future prospects

Functional genomics strives to assign function to all the relevant sites of the genome. Over the last decade, large-scale efforts have produced chromosome maps with the linear distribution of nucleosomes, transcription factor binding sites and chromatin modifications. Collectively these studies demonstrated that what was previously known as junk DNA in fact appears a regulatory jungle. In order to understand the laws of the jungle, linear information must now be converted into spatial relationships. For this, highly detailed 3D topology maps need to be generated for all regulatory sites individually. They are expected to reveal the function of regulatory sites in gene deserts, to uncover the cis-regulatory networks of individual genes and to distinguish the functionally active from inactive regulatory sites. Detailed DNA interaction maps should also uncover the 3D chromosome scaffolds created by insulator sequences, and how the one regulatory site is hampered and the other facilitated in its action by the 3D configuration. Finally, this eventually needs to be done at the single cell level, as we predict an unknown portion of the genome will only reveal its regulatory potential at the level of the individual cell.

Acknowledgements

We would like to thank Patrick Wijchers and Elzo de Wit for useful comments. This work was financially supported 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 ITN (contract number PITN-GA-2007-214902).

3

3C TECHNOLOGY: ANALYZING THE SPATIAL ORGANISATION OF GENOMIC LOCI IN VIVO

Erik Splinter, Frank Grosveld and Wouter de Laat

Department of Cell Biology and Genetics, ErasmusMC, PO Box 1738, 3000 DR Rotterdam, The Netherlands

Adapted from *Methods in Enzymology* 2004 **375**: 493-507.

INTRODUCTION

The spectacular advances in light microscopy have re-established the technique at the forefront of cell biological research. Amongst a wealth of knowledge in many areas of interest, it has provided new insight into how chromosomes and chromosomal regions are organised in the context of the nucleus. For example, chromosomes were found to occupy distinct but fluctuating nuclear territories, while loci on these chromosomes move rapidly within a restricted small volume of this territory. Changes in the transcriptional status of a locus sometimes, but not always, alter its nuclear positioning, as measured against large nuclear landmarks such as centromeres and the nuclear membrane. However, optical constraints set limits to what can be resolved by light/fluorescence microscopy. Thus, it is not possible as yet to visualize the structural organization of a single gene locus that spans, for example, 200 kilobases of genomic DNA. However, intricate structural organisations are to be expected at this level of resolution, for example in cases where enhancers or other transcriptional regulatory elements communicate with distant promoters located *in cis*. Two novel independently developed assays allow to gain insight in the spatial organisation of such genomic loci *in vivo*. One assay, called RNA-TRAP, was developed by Carter et al. and involves targeting of horseradish peroxidase (HRP) to nascent RNA transcripts, followed by quantitation of HRP-catalyzed biotin deposition on chromatin nearby (Carter et al. 2002). The other technique, developed by Dekker et al. was named 3C technology (Chromosome Conformation Capture) and involves quantitation of cross-linking frequencies between two DNA restriction fragments as a measure of their frequency of interaction in the nuclear space (Dekker et al. 2002). Originally applied to the structural organisation of yeast chromosomes, Tolhuis et al. adopted 3C technology to analyse the conformation of a 200 kb region spanning the mouse b-globin gene cluster in its active and inactive transcriptional state (Tolhuis et al. 2002). Here, we discuss a detailed protocol for 3C analysis in mammalian cells, evaluate interpretation of results and discuss the advantages and disadvantages over RNA-TRAP.

Principles of 3C technology

An outline of 3C technology is provided in figure 1. In brief, cells (or isolated nuclei) are treated with formaldehyde to cross-link proteins to neighbouring other proteins and DNA. The resulting DNA-protein network is then subjected to cleavage by a restriction enzyme, which is followed by ligation at low DNA concentration. Under such conditions, ligations between cross-linked DNA fragments, which is intramolecular, is strongly favored over ligations between random fragments, which is intermolecular. After ligation, the cross-links are reversed and ligation products are detected and quantified by polymerase chain reaction (PCR) across the newly ligated ends of fragments. The cross-linking frequency of two specific restriction fragments, as measured by the amount of corresponding ligation product, is proportional to the frequency with which these two genomic sites are close to each other in space. Thus 3C analysis provides information about the spatial organisation of chromosomes or chromosomal regions *in vivo*. Below, we will discuss each step of this procedure in detail.

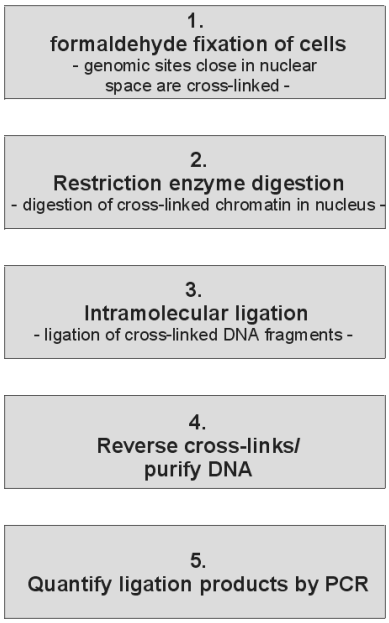


Figure 1. Outline of 3C technology.

Description of 3C technology

In vivo formaldehyde fixation of cells

Formaldehyde is an excellent cross-linking agent to study the composition and structure of chromatin *in vivo*: it reacts with amino and imino groups of proteins and nucleic acids to form protein-protein and protein-nucleic acid cross-links, but does not react with free double-stranded DNA. Formaldehyde cross-links bridge relative short distances (2Å), selecting intimate interactions, and can easily be reversed under mild conditions (Solomon and Varshavsky 1985; Orlando et al. 1997; Jackson 1999).

1. *Fixation is performed by incubating (tumbling) 1 x 10^7 cells in 10 ml of DMEM/10% FCS supplemented with 2% formaldehyde, for 10 minutes at room temperature (RT). The reaction is stopped by the addition of glycine (0.125M final concentration) and transfered to 4^0C.*

All amounts and volumes mentioned in this protocol refer to 1 x 10^7 cells of starting material. In our hands, this number of cells yields enough template for at least 600 PCR reactions, which should be sufficient to get an estimate of the structural organisation of, for example, a 200 kb genomic region. We find that cross-linking with 2% formaldehyde for 10 minutes at room temperature is sufficiently stringent to detect long-range interactions between β-globin genomic sites separated over 100 kb. However, the stringency of fixation required to detect a given interaction will depend on the frequency and stability of this interaction, and therefore other loci may require different fixation conditions for 3C analysis. In the original protocol, Dekker and co-workers reported a loss of PCR-signal when formaldehyde fixation is carried out on intact *Saccharomyces cerevisiae* cells and therefore they perform fixation on isolated

nuclei (Dekker et al. 2002). We prefer to keep manipulation of mammalian cells before fixation to a minimum and add formaldehyde directly to intact living cells; the PCR signals we obtain are reproducible and quantifiable. When working with a mixed cell population, it is important to remember that 3C technology provides an estimate of the average conformation of a genomic region as it is present in all cells included in the analysis. This implies that if the structure of a locus is to be correlated to, for example, its transcriptional activity, the percentage of non-expressing cells present in the sample needs to be sufficiently low to not obscure the detection of possible transcription-specific interactions. We routinely use populations in which at least 80% of the cells display the phenotype of interest. Cell sorting may be required to obtain a good representation of a single cell-type, particularly when working with mixtures of cell-types in a tissue. Formaldehyde-fixed cells can easily be sorted based on cellular markers. We have good experience with sorting fixed erythroid cells from mouse adult bone marrow, using Ter-119 as an erythroid-specific cellular marker. Sorting was done either by Fluorescence Activated Cell Sorting (FACS) or magnetic beads, both increasing the representation of Ter-119$^+$ cells from ~20% to >90%, but with magnetic beads being much faster in processing large amounts of cells.

If working with tissues, it is important to have a homogeneous single-cell suspension before carrying out fixation, since the presence of cell aggregates will seriously affect the efficiency and reproducibility of cross-linking. Soft tissues like the 14.5dpc mouse fetal liver and brain can initially be disrupted with a yellow tip in 100 µl of DMEM/10% FCS, whereas tissues like bone marrow first need to be crushed with a pestle and mortar (or cut and flushed) before taking up cells in medium. Next, all cells should be passed through a cell strainer to get homogeneous single-cell suspensions.

Lysis of cells

2. *Cells are pelleted by centrifugation at 320g for 8 min at 4^{0}C and washed once in 10 ml of cold PBS. The pellet is resuspended in 5 ml of cold lysis buffer (10 mM Tris pH 8, 10 mM NaCl, 0.2% NP-40, protease inhibitors), lysis is allowed to proceed for 10 min at 4^{0}C and nuclei are collected by centrifugation at 600g for 5 min at 4^{0}C. Pellets of nuclei are either frozen via liquid nitrogen and stored at -80^{0}C, or directly processed for digestion.*

To confirm the presence of isolated nuclei after lysis, one can stain proteins and DNA with e.g. Methylgrün-Pyronin and analyze the cells under a light microscope: all cells should be lysed.

DNA digestion in intact nuclei

3. *Take up pellet of nuclei in 0.5ml of 1.2x standard restriction buffer (will be 1x after addition of SDS and Triton X-100; the appropriate buffer depends on the restriction enzyme of choice). Transfer to 1.5 ml tube and add 7.5 µl of 20% SDS to a final concentration of 0.3%; shake for 1 hour at 37^{0}C. Add 50 µl of 20% Triton X-100 to a final concentration of 1.8% and shake for another hour at 37^{0}C. Add 400U of highly concentrated restriction enzyme and digest while shaking overnight at 37^{0}C.*

The conditions in step 3 were designed by Dekker et al. to allow restriction enzyme digestion of DNA in the context of the cross-linked nucleus, a critical event in the assay (Dekker et al. 2002). SDS serves to remove any non-cross-linked proteins from the DNA, Triton X-100 is added to sequester

SDS and allow subsequent digestion. We also find that addition of both substances is necessary for any digestion to occur, but others have reported conditions not requiring SDS and Triton X-100 for *PstI* to digest DNA in nuclei (Leach et al. 2003). The percentages of SDS and Triton X-100 given here were obtained after carefully titrating each component. They represent optimal conditions for *BglII*, *HindIII* and *EcoRI* digestion, but other enzymes, such as *BamHI*, *SpeI*, *PstI* and *NdeI*, do not work (optimally) under these conditions. Dialysing the lysed nuclei to reduce the concentration of SDS and then adding Triton-X100 to complex what is left may be an option to get these and other enzymes to be fully active (M. Spivakov and N. Dillon, personal communication).

The restriction enzyme of choice depends on the locus to be analyzed. It preferably isolates potentially interesting DNA sequences as small (<2 kb) discrete fragments, since larger fragments tend to display more background cross-linking. Preliminary experiments should be done to optimize digestion conditions. De-cross-linking DNA immediately after exposure to restriction enzyme, and running it on an ethidium-bromide stained agarose gel shows whether an enzyme has cut or not. Southern blot analysis and quantification needs to be done to determine the percentage of cleavage. An example is given in Figure 2. Digestion efficiencies of fetal liver cells that were subjected to different concentrations of formaldehyde (0.5%, 1% and 2%, respectively) are compared. Efficiency of

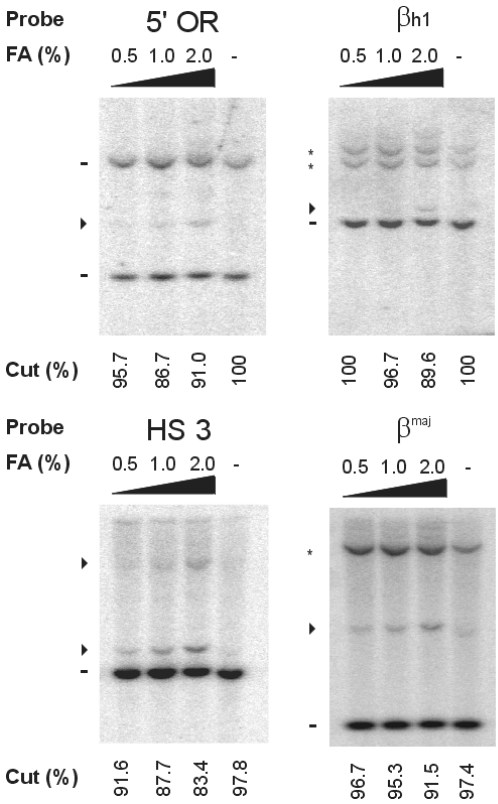


Figure 2. Digestion efficiency depends on fixation and is uniform throughout a locus. Southern blots show that in 14.5 dpc fetal livers digestion efficiency of cross-linked chromatin depends on formaldehyde concentration and is equal near non-transcribed genes (5'OR and βh1), a transcribed gene (βmajor, with DNaseI hypersensitive promoter) and a DNaseI hypersensitive site (HS3). Yield of specifically cut fragments is shown (percentages). Arrowheads depict partial digests and asterisks (expected) cross-hybridization signals with other β-like globin genes.

cleavage by *HindIII* is already high (~85%) at 2% formaldehyde, and increases to nearly 100% with decreasing amounts of formaldehyde. Thus, formaldehyde fixation accounts for the fact that DNA is not digested to completion. Digestion efficiency is independent of local chromatin configuration, since cutting works equally well at an inactive Olfactory Receptor gene (5'OR), an inactive globin gene (βh1), a DNaseI hypersensitive site (HS3) and an active globin gene (with a DNaseI hypersensitive promoter: βmajor) (Fig.2). In brain cells, that do not express any of these genes, we find similar digestion efficiency at all of these sites (data not shown). These important controls show that there is no bias in the assay due to prefered cleavage of one site over another. Once conditions have been defined for a given enzyme in a given cell-type, they appear to be applicable to other cell-types as well. Thus, in our hands *BglII*, *HindIII* and *EcoRI* cleave equally well in mouse fetal liver and brain cells, mouse adult bone marrow cells, mouse erythroleukemia (MEL) cells (an established tissue-culture line) and primary rat schwann cells (M. Ghazvini, personal communication) (data not shown).

Intramolecular ligation

4. *The next day (day 2), 40 µl of 20% SDS is added to a final concentration of 1.3% and the solution is incubated for 20 minutes at 65^{0}C to inactivate the restriction enzyme. The solution is transfered to a 50 ml tube and diluted by adding 7 ml of 1x standard ligation buffer. 375 µl of 20% TX-100 is added to a final concentration of 1% and the solution is incubated for1 hour at 37^{0}C to sequester SDS. 100U of T4 ligase (from highly concentrated stock) is added and DNA is ligated for 4-5 hours at 16^{0}C, followed by 30 min ligation at RT.*

At low DNA concentration (here ~3.7 ng/µl), ligation between cross-linked DNA fragments (i.e. intra-molecular ligation) is highly favored over ligation between non-cross-linked DNA fragments (i.e. intermolecular or random ligation), because the kinetics of essentially a mono-molecular reaction is much faster than that of a bi-molecular reaction. Indeed, when we fix cells with 2% formaldehyde, we never detect ligation products between random restriction fragments (e.g. coming from different chromosomes), whereas DNA fragments normally located *in cis* within 20 kilobases (and often more) on a chromosome can easily be observed (data not shown).

Reversal of cross-links and purification of DNA

Cross-links are reversed by overnight incubation at 65^{0}C in the presence of Proteinase K (300 µg total). The next day (day 3) 300 µg of RNase is added and RNA is degraded for 30-45 min at 37^{0}C, followed by phenol extraction and ethanol precipitation of DNA (including a 70% ethanol wash). DNA pellets are resolved in 150 µl 10 mM Tris (pH7.5). DNA is now ready to be analyzed by PCR.

Quantitative PCR analysis of cross-linking frequencies

Primer design for PCR.

A number of issues have to be considered for the design of primers for the quantitative PCR-analysis of ligation products. First, primers should hybridize to unique DNA sequences and one therefore needs to carefully check for the presence of repetitive DNA sequences near restriction sites of interest to avoid using a repetitive DNA primer. Second, for reproducible and quantifiable PCR signals, the size of the PCR products should be kept small. We try to design

all our PCR-primers within 50-100 bp from the restriction sites analyzed, yielding PCR-products of 100-200 bp in size. Third, differences in the Tm of primers should be kept to a minimum (<2^{0}C) to allow simultaneous analysis of all possible primer combinations. Primers can be designed at either end of a given restriction fragment.

Quantification of cross-linking frequencies. 3C technology involves quantifying ligation frequencies of restriction fragments by PCR, which gives a measure of their cross-linking frequencies. We quantify the formation of PCR products by scanning their signal intensities after separation on ethidium-bromide stained agarose gels, using a Typhoon 9200 imager (Molecular Dynamics). For this analysis to be quantitative, the amount of DNA template added per PCR reaction should critically be in the range that shows linear PCR product formation. This needs to be determined for each new template by testing PCR product formation on serial dilutions of template, using multiple primer combinations. We typically find that for most primer combinations ~20 ng of DNA template is in the range of linear PCR product formation, using 36 cycles of PCR-amplification (Fig. 3). The fact that such a large number of PCR cycles is needed shows that only very little ligation product is present in (and amplified from) a vast excess of genomic DNA. This may explain why quantitation of SYBR Green incorporation by real-time PCR sofar failed to provide us with satisfying results. Measuring incorporation of ^{32}P-labeled nucleotides in PCR products, a third method to quantify PCR-products, generally requires less cycles of amplification but otherwise involves the same principles as measuring EtBr-incorporation. Both techniques allow the quantification of the correct size fragment, while real-time PCR also measures the incorporation of dye in background products, primer-dimers, etc. Given the number of PCR reactions often involved in 3C analysis, we prefer the use of the non-radioactive method that measures intercalated ethidium-bromide.

Controls

Correct interpretation of data obtained by 3C technology critically depends on several controls. We mentioned before that it is important to exclude a bias in the assay due to prefered restriction enzyme digestion of one site over the other; this can be checked by Southern analysis of DNA directly after digestion.

A quantitative comparison of signal intensities of PCR-products obtained with different primer sets is only valid after correction for PCR amplification efficiency of each primer set. Thus, a control template is required in which all possible ligation products are present in equimolar amounts. In yeast, this template was obtained by digesting and randomly ligating non-cross-linked genomic DNA (Dekker et al. 2002). For mammalian cells, with a genome one hundred times the size of the yeast genome, we found that random ligation of two specific loci is too rare an event to be detected by PCR. One therefore needs to enrich for ligation products of interest. This can be done by taking a BAC, PAC or YAC carrying an artificial chromosome that covers the genomic region to be analyzed (Fig.4a). Digestion and ligation of the artificial chromosome yields a mixture that contains all possible ligation products of interest in equimolar amounts. An alternative method is to mix equimolar amounts of DNA fragments that span each of the restriction sites to be analyzed (Fig. 4a). Such fragments can often most easily be obtained by PCR. The approach involves gel purifying

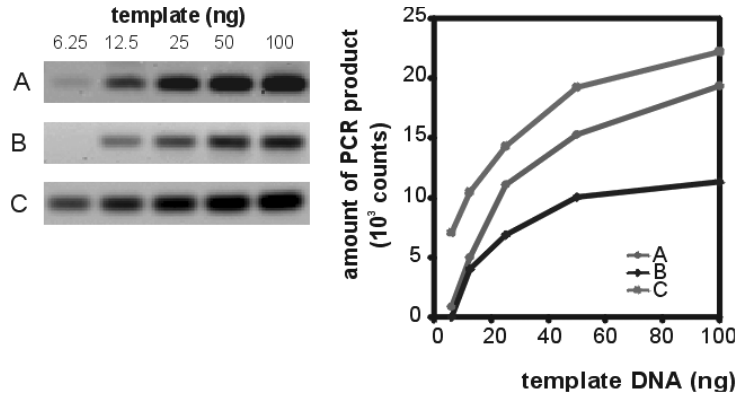


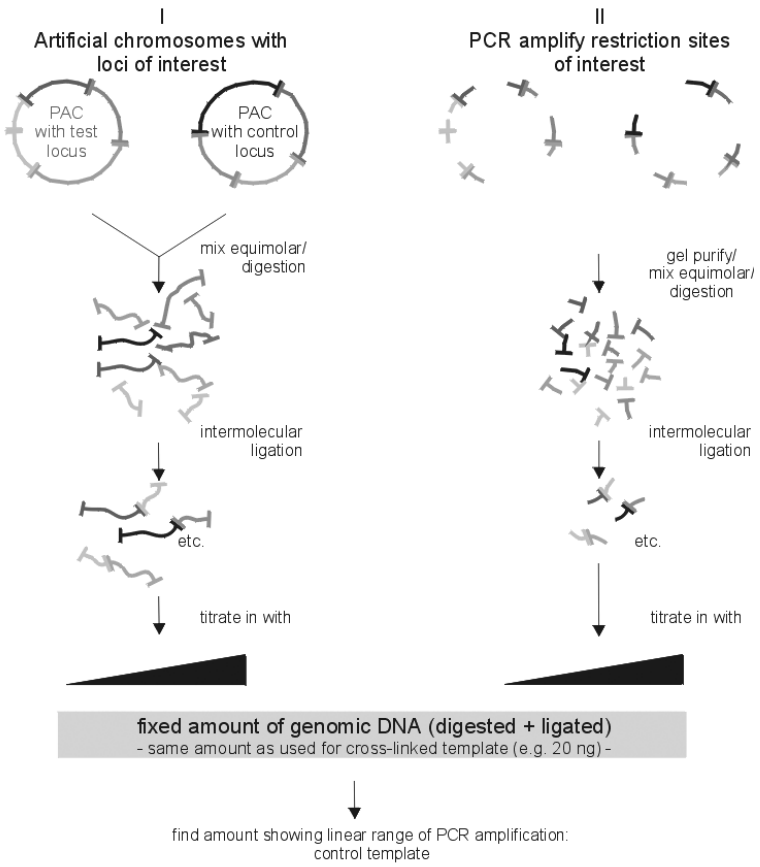
Figure 3. Determination of the linear range of PCR amplification of cross-linked DNA template. Examples are shown of titrations of the cross-linked DNA template cut with EcoRI (A) and HindIII (B, C) that was obtained from 14.5 dpc fetal livers. PCR products were separated on 2% agarose gels stained with ethidium- bromide and scanned with a Typhoon 9200 imager (Molecular Dynamics). Graph shows the quantitation of the PCR products. Primersets analyze ligation between restriction fragments containing the active βmajor gene and HS2 (A), the inactive εγ gene and HS-60 (B), and two sites ~7 kb apart in the XPB (component of the basal transcription factor TFIIH) locus, which is taken as a control (C).

these fragments and carefully determining their concentration before mixing them in equimolar amounts, followed by restriction enzyme digestion and ligation. Clearly, the latter is much more laborious and intrinsically less accurate than working with artificial chromosomes.

The PCR amplification efficiency of a primer set is also influenced by the amount of genomic DNA present in the PCR reaction. Thus for correct comparison, the random ligation template described above should be mixed with an amount of genomic DNA similar to that used for the test samples (~20 ng, see above). To completely mimic conditions, we mix the random ligation template with digested and re-ligated genomic DNA. A serial dilution of this template in a fixed amount of genomic DNA is carried out to determine the proper control template concentration. By taking the ratio of signal obtained by quantitative PCR on cross-linked template versus control template, one corrects for differences in amplification efficiency between primer sets and also for differences in signal intensities due to the size of PCR products. These controls are absolutely essential for a quantitative, or even a qualitative, interpretation of the PCR data and thus for any conclusion on the spatial organisation of a locus.

If different cross-linking frequencies have to be measured in different tissues, an internal standard is required that accounts for variations in efficiency of cross-linking and ligation. This is done by analysing the cross-linking frequency of a locus that is unrelated and that can reasonably be assumed to adopt a similar spatial organisation in the different tissues. Gene loci that express at similar levels in the tissues analysed are thought to meet this criterion. Previously, we analysed cross-linking frequencies between two fragments in the transcribed part of the calreticulin locus (CalR) with the restriction sites analysed ~1.5 kilobases apart. The CalR

A



B Relative cross-linking frequency *X* between two given fragments (fr1+fr2):

$$X(\text{fr1+fr2}) = \frac{[\text{A(fr1+fr2)/ A(c1+c2)}]_{\text{tissue}}}{[\text{A(fr1+fr2)/ A(c1+c2)}]_{\text{control template}}}$$

corrects for: PCR efficiency

corrects for:
- quality and quantity of template

Figure 4. Controls in 3C technology. (A) Two methods (I and II) to prepare the random ligation control template required to correct for differences in PCR amplification efficiency. Perpendicular bars on PACs (or YACs, BACs) and on DNA fragments indicate restriction sites. Different colours indicate different restriction fragments. PCR analysis is done with primers that amplify across the ligated sites. **(B)** Equation used to calculate relative cross-linking frequency. A: measured signal intensity of PCR product. The relative cross-linking value of 1 arbitrarily corresponds to the cross-linking frequency between two control fragments (c1 + c2) (fragments e.g. present in XPB or CalR; see text).

locus is embedded in an area of ubiquitously expressed genes and expresses at similar levels in the tissues analysed in that particular study (14.5dpc fetal brain and liver) (Tolhuis et al. 2002). However, other loci with well defined expression levels and patterns may be equally suitable. By normalising, within a tissue, each cross-linking frequency to the cross-linking frequency observed between the control fragments, one can correct for differences in quality and amount of template.

Finally, adding the control locus to the random ligation mix (either as an artificial chromosome or as DNA fragments, in equimolar amounts to the locus that is analysed), and routinely analysing its ligation efficiency, provides a loading control, to correct for differences in the amount of control template between experiments.

The equation used to calculate the relative cross-linking frequency is given in figure 4b. As a result of this normalisation, the "cross-linking frequency" value 1 arbitrarily corresponds to the cross-linking frequency between the control fragments (e.g. CalR fragments).

Interpretation of data obtained by 3C technology

As pointed out originally by Dekker et al., measuring cross-linking efficiency by the formation of ligation products largely depends on the frequency with which two genomic sites interact (Dekker et al. 2002). They showed that contributions of other parameters, such as local protein concentrations or a favourable geometry of the cross-linked intermediate, are minor. By fitting their data to polymer models, the authors interpreted the relationship between cross-linking frequency and genomic site separation to give an estimate of the three-dimensional organisation of chromosome III of *S. Cerevisiae* (Rippe 2001). For a number of reasons, we are still reluctant to interpret data obtained by 3C analysis in a strictly quantitative manner. In the first place, we find that additional parameters, e.g. the fragment size, notably affect the cross-linking efficiency. Particularly difficult to interpret are cross-linking frequencies measured for large fragments carrying multiple *cis*-regulatory elements, which are each likely to be engaged in unique and different interactions. Clearly, the specificity of measuring interactions increases with smaller fragments containing such regulatory sites as isolated entities. Second, changes in fixation conditions differentially affect cross-linking frequencies. Figure 5 shows the cross-linking frequencies obtained with 2%, 1%, and 0.5% formaldehyde for three neighbouring restriction fragments. In each PCR reaction the same amount of template was used and all products are in the linear range of PCR-amplification. The data show that nearby fragments require a lower percentage of formaldehyde to reach maximum cross-linking efficiency than more distal fragments. Third, interactions between distal DNA elements are thought to be dynamic, while the measurements represent steady-state average levels. Thus, short-lived but important interactions may score much lower than more long-lived interactions. Finally, proper modelling would require incorporation of an estimate of the packing ratio of the chromatin fibre. However, even for the extensively studied β-globin locus it is impossible to say if chromatin is folded as a 10nm fibre, a 30nm fibre or that it adopts yet another conformation. Moreover chromatin folding is certainly not uniform along the entire locus. Thus, for many loci in higher eukaryotes it is presently difficult to give a reliable estimate of the packing ratio of the chromatin fibre. We

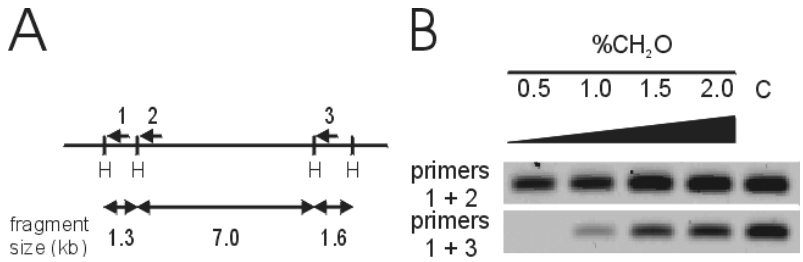


Figure 5. Changes in the fixation condition differentially affect the efficiency of cross-linking between different sites. A. Schematic presentation of part of the XPB locus, showing HindIII restriction sites (H) at that site. Primers used to analyze ligation products are indicated by arrows and are numbered. **B.** PCR analysis of ligation products formed after fixation with 0.5, 1, 1.5, and 2% formaldehyde. In all cases, 20 ng of template was used (in linear range of amplification, see text and figure 3). Under the conditions used, ligation between the neighbouring fragments 1 and 2 is saturated at 1.5% aand 2%, whereas ligation between fragments 1 and 3 is not. Control (C) was obtained with the random ligation control template (see text); intensities of these signals are a measure of the efficiency of amplification of each primerset.

currently prefer to describe data obtained by 3C technology in a qualitative manner, rather than interpreting the relationship between cross-linking frequency and genomic site separation in terms of real distances (Tolhuis et al. 2002). Measured interactions become particularly meaningful if they can be correlated to a phenotype, e.g. if they occur in a transcriptionally active locus and not in the inactive one.

3C technology and RNA TRAP

RNA TRAP was designed to identify DNA sequences that are in close proximity to actively transcribed genes (Carter et al. 2002). It applies fluorescent in situ hybridisation (FISH) technology to target horseradish peroxidase (HRP) to nascent RNA transcripts, followed by HRP-catalysed biotin deposition on chromatin nearby. After affinity purification, the relative abundance of biotinylated DNA sequences is determined by quantitative PCR analysis and is taken as a measure of proximity to the labelled nascent transcript. Although both 3C technology and RNA TRAP were only developed recently and undoubtedly will evolve further, it is useful to summarise the advantages and disadvantages of the two techniques.

A disadvantage of the 3C technology is that the resolution is restricted by the occurrence of restriction sites. Ideally, (potential) regulatory DNA sequences are analyzed by 3C as isolated entities present on small DNA fragments. Frequently cutting enzymes like Sau3AI or NlaIII, which on average cleave once every 256 basepairs, often yield such fragments, but inevitably will also cut at unfavourable positions. Multiple restriction enzymes can be included in the 3C analysis of a locus, but each enzyme requires its own PCR primers.

Another drawback of 3C technology is the difficulty to discriminate between directly and indirectly bound DNA fragments. Whereas in RNA TRAP the DNA fragment closest to the nascent transcript may shield other chromatin from the deposition of biotin, in 3C technology a nearby cross-linked DNA fragment will bring in other cross-linked DNA fragments, which DNA ends all can be ligated to the restriction site of interest. RNA TRAP may therefore be better to find directly interacting genomic sites, while 3C technology

gives a more general picture of DNA fragments nearby the site of interest. To optimise the specificity of ligation in 3C technology, one has to find cross-linking conditions that minimise the average number of DNA fragments present per cross-linked protein-DNA aggregate without loosing the interactions of interest.

RNA TRAP is dependent on nascent transcripts coming from actively transcribed genes, which is both an advantage and a disadvantage. The advantage is that non-expressing cells do not contribute to the measured values, meaning that genomic site interactions with the transcriptionally active locus can be picked up even in mixed cell populations. However, to determine whether such interactions are specific for the transcriptionally active status of this locus one needs to know its silent conformation, which is an important control that cannot be checked by RNA TRAP. Moreover, it may be difficult to obtain satisfying results by RNA TRAP for genes with moderate transcription rates. The globin genes, which were used to develop RNA TRAP, are transcribed very efficiently, and these genes are exceptionally good targets for visualising nascent transcripts (Wijgerde et al. 1995). Most other genes, however, are not transcribed so actively, and visualising ongoing transcription here is more troublesome. It remains to be determined what the minimal density of nascent transcripts is that is required for RNA TRAP. Another disadvantage of RNA TRAP is that even in a transcriptionally active locus it does not allow the analysis of the spatial organisation near sites other than the transcribed gene itself.

DNA TRAP, which would involve hybridisation to specific DNA sequences rather than to nascent RNA transcripts, may solve the above issues, but will undoubtedly be more difficult to develop (Carter et al. 2002). It requires specific binding of probes to chromatinized double-stranded DNA without denaturing and changing its higher order structure and hybridisation to single copy genomic DNA is much less sensitive than to a multi-copy RNA transcript. DNA TRAP will also circumvent the problem of RNA transcripts being dispersed along the transcribed region of the gene. This dispersed localisation of label could be particularly inconvenient when applying RNA TRAP to genes much larger than the globin genes, which are only 1.5 kb in size. Also, DNA TRAP, like the 3C-methodology, would have the potential to detect specific spatial interactions between transcriptional regulatory elements like promoters and enhancers, which cannot be done by RNA TRAP.

Despite their current limitations, both 3C technology and RNA/DNA-TRAP clearly are important new tools that undoubtedly will reveal exciting new insight in the structural and functional organisation of chromatin in the living nucleus, at a level presently not possible to analyse with microscopes.

ACKNOWLEDGEMENTS

We would like to thank Bas Tolhuis and Robert-Jan Palstra for their contribution to setting up 3C technology in our laboratory. We thank Mhernaz Ghazvini, M. Spivakov and Niall Dillon for sharing unpublished results. This work is supported by NWO (the Netherlands Organisation for Scientific Research) to WdL as part of the Innovational Research Incentives Scheme and by NWO and EC grants to FG.

4

DETERMINING LONG-RANGE CHROMATIN INTERACTIONS FOR SELECTED GENOMIC SITES USING 4C-SEQ TECHNOLOGY: FROM FIXATION TO COMPUTATION

Erik Splinter*, Elzo de Wit*, Harmen J.G. van der Werken, Petra Klous and Wouter de Laat

*Contributed equally to this work.

Hubrecht Institute-KNAW & University Medical Centre Utrecht, Utrecht 3584 CT, The Netherlands

Submitted for publication

SUMMARY

Chromosome Conformation Capture (3C) and 3C-based technologies are constantly evolving in order to probe nuclear organization with higher depth and resolution. One such method is 4C-technology that allows the investigation of the nuclear environment of a locus of choice. The use of Illumina next generation sequencing as a detection platform for the analysis of 4C data has further improved the sensitivity and resolution of this method. Here we provide a step-by-step protocol for 4C-seq, describing the procedure from the initial template preparation until the final data analysis, interchanged with background information and considerations.

1. INTRODUCTION

The role of genome topology in gene regulation is subject to extensive investigation. It is becoming increasingly clear that on top of the linear organization of genomes, chromosomes are organized in non-random three-dimensional (3D) structures. The technical advancement of existing methods and the development of new technologies have contributed greatly to the understanding of the underlying principles of chromosome topology. The 3C method (Dekker et al. 2002; Tolhuis et al. 2002) can be used to study chromosome folding with high resolution. 3C-technology is based on fixing the 3D conformation of the genome using formaldehyde, subsequent restriction enzyme (RE) digestion and ligation of the cross-linked DNA fragments. The frequency of ligation of two given DNA fragments, detected by PCR, is a direct measurement on the 3D contact frequency between those fragments. Due to the low-throughput nature of the method, the original 3C protocol is limited in its scope. We (Simonis et al. 2006) and others (Dostie et al. 2006; Lieberman-Aiden et al. 2009) have adapted the 3C protocol to make it ready for the genomics era. 4C-technology analyzes all fragments (*captures*) interacting with a fragment of choice (*viewpoint*), resulting in a high resolution interaction profile for a locus of interest (Figure 1) (Simonis et al. 2006; Zhao et al. 2006). Here a step-by-step protocol is provided, covering all practical issues from the selection of a model system and template preparation to the details of data analysis for uncovering long-range *cis* -and *trans* interactions. Moreover, background information is provided and considerations that need to be taken into account when interpreting 4C data are discussed.

2. 4C template preparation

2.1 Selection of a homogeneous model system

Before starting a 4C experiment it is important to realize that 4C technology, like all 3C -and other population based methods such as chromatin IP (ChIP), is based on the simultaneous analysis of thousands of cells. In 4C, one diploid cell maximally contributes two data points. This is because for most viewpoints there are only two alleles per cell that each can form a single ligation junction. To generate reproducible interaction profiles, thousands of ligation events need to be analyzed per viewpoint. 4C therefore provides an impression of the average conformation measured across many cells. As genome topologies differ between individual cells, the analysis should aim to identify those interactions that are shared between proportions of cells within the analyzed population.

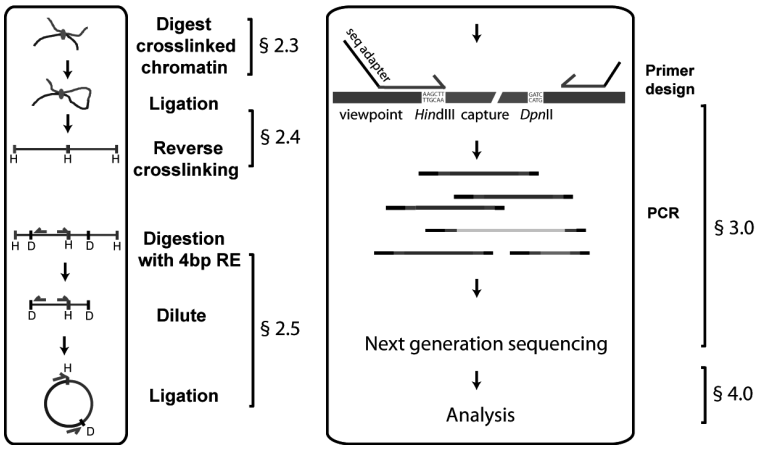


Figure 1. Outline of the 4C-seq procedure. Cross-linked chromatin is digested with *Hin*dIII (H) followed by proximity ligation. Subsequently cross-links are reversed and DNA fragments are trimmed using *Dpn*II (D) and circularized. 4C-primers are designed on the viewpoint fragment allowing the PCR amplification of captured fragments and analysis using Illumina sequencing. The paragraphs describing the different steps of the protocol are also indicated.

Vice versa, this implies that DNA interactions predicted by 4C need to be interpreted in a probabilistic manner: they will occur more often than others in the population of cells, but not necessarily need to be present in a single selected cell. Another consideration is that if one wishes to correlate specific topological features to the expression status or other functional characteristics of a gene it is important to apply 4C technology only to relatively homogeneous cell populations in which the gene of interest (and surrounding loci, see below) behaves similar in the majority of cells analyzed. For this reason, tissue culture cells are ideal for 4C analysis. Some primary tissues such as thymus and liver are also primarily composed of a single cell type. For other tissues, selection methods like FACS can be used to isolate a defined population of cells. If analysis needs to be directed to single alleles, for example in cases of mono-allelically expressed genes, one can use genetically hybrid cell populations and take advantage of single nucleotide polymorphisms (SNPs), as we showed in a study that analyzed topogical features of the active and inactive X chromosome in female mammalian cells (Splinter et al. 2011).

2.1.1 Step by step protocol

Collection of cells

Tissue culture:

- Suspension cells: proceed to step 1.
- Adherent cells can both be formaldehyde treated (step 2-3) and scraped from the culture dish, or first collected, using e.g. Trypsin, before proceeding to step 1.

Primary tissue:

For efficient fixation, a (viable) single cell preparation of the tissue of interest is required. To facilitate this process the use of collagenase and/or a cell strainer is advised, but incubation conditions have to be optimized empirically. For reference: a 14.5dpc fetal brain is dispersed by 0.00625% collagenase treatment in 250µl 10%FCS/PBS for 45min at 37 °C, followed by the use of a 40µm cell strainer

(BD Falcon #352340). For the disruption of 'softer' tissues like 14.5dpc fetal liver or thymus that mainly produce circulating cells (red blood and T-cells, respectively), collagenase treatment can be omitted and the use of a cell strainer is sufficient.

2.2 Formaldehyde cross-linking

Formaldehyde is able to cross-link proteins to proteins and proteins to DNA, but not DNA to DNA directly, and acts over a distance as small as 3Å. An often raised concern is that formaldehyde may create a bias because of preferential cross-linking between DNA sites bound by transcription factors and e.g. PolII. If anything the bias appears the other way around; as explained previously (Simonis et al. 2007), regulatory sequences bound by transcription factors are less efficiently crosslinked to proteins than the remainder of the genome. FAIRE (formaldehyde assisted isolation of regulatory elements) is an assay that takes advantage of this fact to identify potential regulatory sites in the genome (Giresi et al. 2007). With regard to 4C results, experiments performed using active and inactive regions showed that they make long-range contacts with similar aptitude (Simonis et al. 2006; Lieberman-Aiden et al. 2009) (Splinter et al. 2011).

2.2.1 Step by step protocol

Fixation and cell lysis

All the steps in this protocol are optimized for using $1x10^7$ cells.

1. Count cells and centrifuge 5 min, 280g at RT.
2. Discard the supernatant and resuspend the pellet in 5ml PBS/10% FCS.
3. Add 5ml 4% formaldehyde in PBS/10% FCS (2% final concentration formaldehyde) and incubate for 10 minutes at RT while tumbling.
4. Add 1.425ml 1M glycine (final concentration 0.125M), mix and put tubes immediately on ice to quench the cross-linking reaction. Note that glycine is not in excess and quenching will not be complete, therefore directly proceed to step 5.
5. Centrifuge 8 min, 400g at 4°C and remove all the supernatant.
6. Resuspend pellet in 1 ml cold lysis buffer (50mM Tris-HCl pH7.5, 150mM NaCl, 5mM EDTA, 0.5% NP-40, 1% TX-100 and 1X Complete protease inhibitors (Roche #11245200) and incubate 10 minutes on ice.
7. Determine the efficiency of cell lysis:
   Mix 3µl of cells with 3µl of Methyl Green-Pyronin staining (Sigma #HT70116) on a microscope slide and overlay with a coverslip. Assess the lysis efficiency using a microscope. Cytoplasm stains pink and the nuclei stains blue/green. When cell lysis is incomplete, douncing can be applied to increase efficiency.
   Note: cell lysis is an important step in the protocol as failure of lysis can hamper digestion efficiency.
8. Centrifuge 5 min, 750g at 4°C and carefully remove all supernatant. At this point nuclei can be stored for later use (proceed to step 9) or the protocol is continued directly (proceed to step 10).
9. Storing the nuclei at -80°C:
   - **9.1** Resuspend nuclei pellets in lysis buffer and transfer to a 1.5ml safe lock tube.
   - **9.2** Centrifuge 2 min, 540g at 4°C.
   - **9.3** Remove the supernatant, freeze the pellet in liquid nitrogen and store at -80°C.
10. Resuspend the pellet in 450µl Milli-Q and continue with step 11.

2.3 Digestion

4C technology is based on the detection of proximity ligation of restriction fragments generated in the context of cross-linked chromatin. As most restriction enzymes fail to exert their function under these conditions, the choice of restriction enzyme and its efficiency are crucial in generating a data set of considerable quality. In our hands, *Hin*dIII, *Eco*RI and *Bgl*II are the preferred restriction enzymes as they cut well in this procedure, but others have also used other enzymes, e.g. *Bam*HI (Nativio et al. 2009). Although these enzymes typically digest with >85% efficiency, efficiencies may vary between samples and cell types used. Therefore special attention is required to monitor this step in the 4C procedure. Separating digested and undigested chromatin samples on an agarose gel should routinely be performed as an initial quality control. Note however, that this only provides a rough estimate of the performance of the restriction enzyme. It may suffice as a quality check when a researcher routinely performs 4C technology and is experienced with the method. When setting up the assay though, qPCR analysis, using primer sets across restriction sites, is recommended for precise determination of digestion efficiency. Moreover, by inquiring multiple restriction sites this analysis can be used to also determine any potential restriction enzyme bias.

A further consideration regarding the choice of restriction enzyme is the frequency and distribution of restriction sites in the locus of interest. We use the following criteria for selecting a viewpoint. First, the fragment of choice should have a minimum size of 500bp and is preferentially located such that it covers the promoter or another sequence element of interest. If this is not possible, directly neighboring restriction fragments can be taken as viewpoints and their interaction profiles can be interpreted as an approximation of the profile of interest. Second, to be able to design a specific amplification primer, at least one of the primers, but preferably both, should be unique in the genome and match the criteria of a good quality primer (see section 3.1 below). For each experiment, the choice of a restriction enzyme depends on its capacity to create the viewpoint that best meets these criteria. In the protocol described here, *Hin*dIII was selected as the enzyme of choice, but *Hin*dIII can obviously be replaced by any of the other mentioned restriction enzymes. The choice for a 6bp recognition enzyme as presented here has consequences for the resolution and distances over which interactions can be identified. On average after digestion 4kb fragments are produced. As we are interested in recurrent interactions, represented by the presence of multiple captures in a particular area, a running window algorithm is applied resulting in the identification of interacting areas between 100 and 200kb in size. This is ideal for studying long-range *cis* -and *trans* interactions, for the identification of more defined contacts like promoter-enhancer interactions however, an alternative strategy should be applied (van der Werken In Prep.)

2.3.1 Step by step protocol

Digestion

11. Add 60µl of 10X restriction buffer B (supplied with *Hin*dIII).
12. Place the tube at 37°C and add 15 µl 10% SDS.
13. Incubate 1hr at 37°C while shaking at 900 RPM using an Eppendorf Thermomixer.
14. Add 75µl 20% Triton X-100.
15. Incubate 1hr at 37°C while shaking at 900 RPM.
16. Take a 5µl aliquot of the sample as the 'Undigested control' and store at 4°C until

used in step 21.

17. Add 200U *Hin*dIII (Roche #11274040001); incubate 4 hrs at 37°C while shaking at 900 RPM.

18. Add 200U *Hin*dIII; incubate O/N at 37°C while shaking at 900 RPM.

19. Add 200U *Hin*dIII; incubate 4 hrs at 37°C while shaking at 900 RPM.

20. Take a 5µl aliquot of the sample as the 'Digested control'.

21. Determination of the digestion efficiency:

- **21.1** Add 90µl 10mM Tris-HCl pH 7.5 to the 5µl samples from step 16 and 20.
- **21.2** Add 5µl Prot K (10 mg/ml Roche #03115836001) and incubate for 4 hours at 65°C.
- **21.3** Add 100µl Phenol-Chloroform (Sigma) to the samples and mix vigorously.
- **21.4** Spin for 10 min, 16400g at RT.
- **21.5** Transfer water phase to a clean tube and load ~ 20µl on a 0.6% agarose gel. Alternatively, Q-PCR analysis can be used to more precisely determine digestion efficiency, using multiple primer sets each spanning a different restriction site. This step is highly recommended when 4C is applied for the first time.
- **21.6** If digestion is OK proceed with step 22, otherwise repeat step 18, 20 and 21.

2.4 Ligation

For reasons of clarity, most cartoons illustrating 3C or 4C methodology depict crosslinked aggregates of two restriction fragments secured by a single protein (Tolhuis et al. 2002; Simonis et al. 2006; Lieberman-Aiden et al. 2009; Figure1). However, this does not reflect the nature of aggregates formed in the test-tube. The presence of nucleosomes is sufficient to bridge between DNA fragments. This allows the formation of multiple crosslinks, not only between two DNA fragments but also between other DNA fragments in close proximity, resulting in the formation of extensively branched chromatin aggregates ('hairballs'). This is reflected by the ability to detect multiple restriction fragments in one *4C-circle* (Zhao et al. 2006; Vernimmen et al. 2007) and can also be appreciated from the sizes of DNA circles formed after ligation (notice the extreme shift in the molecular weight of DNA before (step 21) and after ligation seen by gel electrophoresis (step 28) (Figure 2). The seizing of DNA circles after ligation of crosslinked DNA prohibits efficient PCR amplification directly from this template, necessitating the further processing steps described below. The shift seen after gel electrophoresis does however serve as a good indication of efficient ligation.

2.4.1 Step by step protocol

Ligation

22. Heat-inactivate the restriction enzyme by incubating 20 min. at 65°C and continue with step 23. Alternatively, when the restriction enzyme is not sensitive to heat inactivation (as is the case for e.g. *Bgl*II), continue with step 22.1.

- **22.1** Add 80µl 10% SDS and incubate 30 min. at 65°C.
- **22.2** Transfer the sample to a 50ml Falcon tube and add 5.4ml Milli-Q
- **22.3** Add 700µl 10X Ligase buffer (10X: 660mM Tris-HCl pH 7.5, 50mM $MgCl_2$, 10mM DTT, 10mM ATP)
- **22.4** Add 375µl 20% TX-100 and incubate 1hr 37°C
- **22.5** Continue with step 26.

23. Transfer the sample to a 50ml Falcon tube.

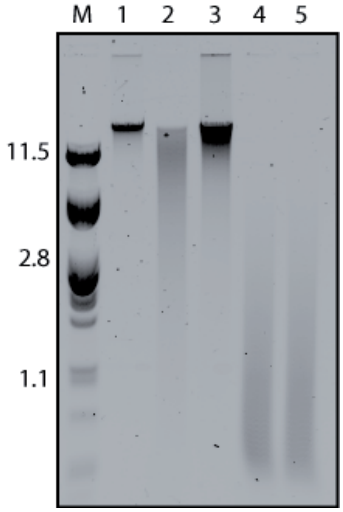


Figure 2. Restriction and ligation of cross-linked chromatin. Undigested (lane 1), HindIII digested (lane 2) and ligated (lane 3) DNA fragments collected during steps 16, 20 and 27 of the protocol respectively are separated on a 1.5% agarose gel. M indicates the λ*Pst*I marker. Please note the typical high molecular weight products (>10kb) formed after ligation.

24. Add 5.7ml Milli-Q.
25. Add 700µl 10X Ligase buffer (see step 22.3).
26. Add 50U T4 DNA Ligase (Roche, #799009), mix by swirling and incubate O/N at 16°C.
27. Take a 100µl aliquot of the sample as the 'Ligation control'.
28. Determine ligation efficiency:
  28.1 Add 5µl Prot K (10mg/ml) and incubate for 4 hours at 65°C.
  28.2 Add 100µl Phenol-Chloroform to the sample and mix vigorously.
  28.3 Spin 10 min, 16400g at RT.
  28.4 Transfer water phase to a clean tube and load ~ 20µl on a 0.6% agarose gel next to the 'digestion control' from step 20.
  28.5 If ligation is OK, proceed with step 29. If not, add fresh ATP (final concentration of 1mM) and add new ligase: repeat step 26-28.

Reverse cross-linking and precipitation

29. Add 30µl Prot K (10mg/ml) and reverse cross-links O/N at 65°C.
30. Add 30µl RNase A (10mg/ml, Roche #10109169001) and incubate 45 minutes at 37°C.
31. Add 7ml Phenol-Chloroform, mix vigorously.
32. Centrifuge 15 min, 3270g at RT.
33. Transfer the aqueous phase to a new 50ml Falcon tube and add:
- 7ml Milli-Q
- 1.5ml 2M NaAC pH 5.6
- 7µl Glycogen (20mg/ml, Roche #10901393001)
- 35ml 100% EtOH.

  Increasing the volume twice before precipitation can prevent the co-precipitation of DTT from the ligase buffer and therefore results in a sample with higher purity.

34. Mix and incubate at -80°C until the sample is frozen solid.
35. Spin 20 min, 8346g at 4°C.
36. Remove the supernatant and add 10 ml cold 70% ethanol.
37. Centrifuge 15 min, 3270g at 4°C.
38. Remove the supernatant and briefly dry the pellet at room temperature.
39. Dissolve the pellet in 150µl 10mM Tris-HCl pH 7.5 at 37°C.

40. Continue with step 41 or store sample at -20°C.

2.5 Trimming the circles: second round of restriction enzyme digestion

PCR amplification of the large molecules (>10kb) containing multiple restriction fragments, formed after the initial ligation step, is highly inefficient and biased towards the small DNA circles that are present in low abundance (data not shown). This is solved by trimming the large molecules with an additional restriction and ligation step (Figure 1). It is important that a RE is selected cutting with high frequency, *viz*. a 4bp recognizing RE, to create PCR amplifiable fragments. The choice of the second RE is, like the first RE, restricted by quality criteria. First, when working with mammalian model systems the RE of choice should not to be blocked by CpG methylation. Also, to facilitate the subsequent ligation, the use of a RE that leaves cohesive ends is recommended. Furthermore, to prevent a digestion bias towards C-G or A-T rich regions the recognition sequence of the RE is recommended to possess a mixture of all four nucleotides. Finally, because the efficiency and capacity of self-ligation is in part determined by the size of the DNA fragment itself (Rippe et al. 1995), the position of the recognition site relative to the *Hin*dIII site has to be taken into account. Ideally the 4bp RE site is located ~500nt away from the *Hin*dIII site and a minimal distance of 250nt is preferred (Rippe et al. 1995). Obviously, the ability to design a specific PCR primer near the 4bp RE site is also important, although the positioning of this primer is less strict than that of the *Hin*dIII primer (Figure 1 and section 3.1 below). We will focus in this protocol on *Dpn*II as the second RE. However replacing *Dpn*II with *Nla*III or *Csp*6I (*Cvi*QI) should give similar results.

2.5.1 Step by step protocol

Second Digestion

41. To 150µl 3C sample (~$1x10^7$ cells) add:
- 50µl 10X *Dpn*II restriction buffer
- Milli-Q to 500µl
- 50U *Dpn*II (New England Biolabs #R0543S)

42. Incubate O/N at 37°C.

43. Take a 5µl aliquot of the sample as the "Digestion control".

44. Determine digestion efficiency:

44.1 Add 95µl 10mM Tris-HCl pH 7.5 to the 5µl sample from step 43.

44.2 Load ~20µl on a 0.6% agarose gel next to the 'ligation control' from step 28.

44.3 If digestion is OK, proceed with step 45. If not, add fresh restriction enzyme and repeat step 42-44. Alternatively the sample can be re-purified to facilitate efficient digestion.

Second Ligation and purification

45. Inactivate enzyme by incubating at 65°C for 25 minutes and continue with step 46. If not heat sensitive, the restriction enzyme can be inactivated by sample purification. Continue with step 45.1.

45.1 Add 500µl Phenol-Chloroform and mix vigorously

45.2 Spin 10 min, 16400g at RT.

45.3 Transfer the aqueous phase to a fresh tube and add 50µl 2M NaAc pH 5.6 and 950µl 100% EtOH

45.4 Incubate at -80°C until completely frozen

45.5 Spin 20min 16400g at 4°C

45.6 Remove supernatant and add 150µl cold 70% ethanol.

45.7 Spin 10min 16400g at 4°C

45.8 Resuspend the pellet in 500µl 10mM Tris-HCl pH 7.5

46. Transfer sample to a 50ml tube and add:
- 12.1ml Milli-Q
- 1.4 ml 10X Ligation buffer (see step 22.3)
- 100U T4 DNA Ligase

47. Ligate O/N at 16°C.

48. Add: 0.7ml 2M NaAC pH 5.6, 7µl Glycogen (1mg/ml) and 35ml 100% EtOH. Mix well.

49. Store at ~80°C until completely frozen.

50. Spin 45 min, 8346g at 4°C

51. Remove the supernatant and add 10ml cold 70% ethanol.

52. Spin 15 min, 3270g at 4°C.

53. Remove the supernatant and briefly dry the pellet at RT.

54. Dissolve the pellet in 150µl 10mM Tris-HCl pH 7.5 at 37°C.

55. Purify samples with the QIAquick PCR purification kit (Qiagen #28104). Use 3 columns per sample; binding capacity is 10µg DNA per column. Elute columns with 50µl 10mM Tris-HCl pH 7.5 and pool samples.

56. Measure concentration using the Nanodrop spectrophotometer and run a serial dilution of 0.125, 0.25, 0.5 and 1µl sample on a 2% agarose gel in order to estimate the concentration compared to a reference sample, e.g. phage-λ DNA.

57. The 4C template is now ready for PCR and can be stored at -20°C or continued with directly in step 58.

3. Illumina sequencing

The 4C-seq strategy described here is compatible with Illumina GAII high-throughput sequencing. For Illumina sequencing the DNA libraries must contain two unique adaptor sequences that allow the sample DNA to attach to oligos that are immobilized to the flow-cell. On this flow-cell an initial PCR takes place, generating clusters. In a single-end sequencing experiment, sequencing will start from the reading adaptor. In contrast to library preparation for most other sequencing experiments where the sequencing adapters are added by ligation, in a 4C-seq experiment the adapters are introduced as part of the primers used for the 4C-PCR (Figure 1 and 3B). For the nucleotide composition of these adapters Illumina technical support should be consulted. As the Illumina adapters are introduced as 5' overhangs to the PCR primers that hybridize to the viewpoint, the first nucleotides of a sequencing read will match the sequence of one of these primers. Typically, we direct sequencing to the primer (the 'reading primer') that hybridizes near the primary restriction enzyme site of interest (here: the *Hin*dIII site; Figure 3B). Because a typical PCR primer has a length of 20nt, the first 20 nucleotides of all reads are taken by the viewpoint-specific primer sequences. The minimal read length provided by Illumina is 36 nucleotides, in which case 16nt (sequence $result_{36nt}$ – $primer_{20nt}$) are left to identify the captured fragment. This would not be sufficient for unambiguous mapping to the whole genome, but for a 4C-seq experiment this actually is sufficient. Reason is that the captured sequence necessarily needs to be one of the sequences that directly flank a primary restriction site (here: *Hin*dIII) in the genome. Thus, by reducing the complexity of the genome to only those sequences that directly flank the recognition sequences of an enzyme, we create a so-called 'fragment end library' which is used as input for the mapping procedure. Although mapping is performed against this 'reduced genome' sequencing 36nt does require the reading primer to include the entire *Hin*dIII site. Sequencing longer read lengths (e.g. 52nt) allows a more flexible positioning of the reading primer. Figure 4 shows an overview of the percentage of unique sequences for various lengths of fragment end

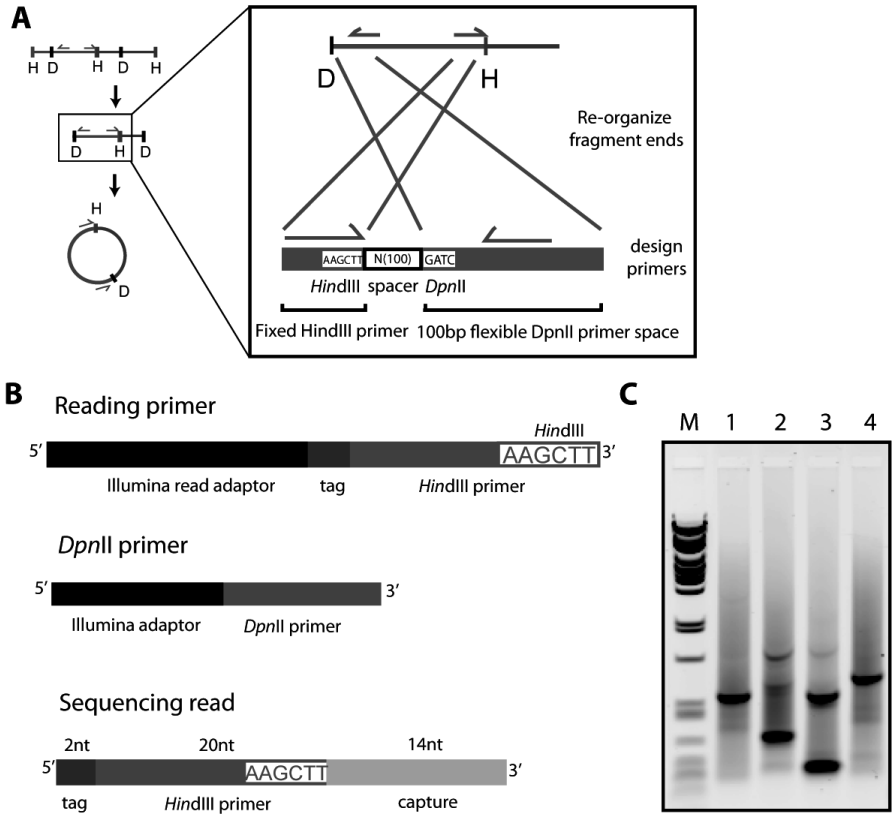


Figure 3. 4C-seq primer design and PCR products. 4C primer design is such that the primers point outwards on the viewpoint fragment. To facilitate correct design using Primer3, the viewpoint fragment-ends are re-organized (**A**). The HindIII primer position is fixed, while 100bp flexibility is allowed for DpnII primer design. The 4C primers used in the 4C procedure consist of different domains of which the classifications are indicated (**B**). Important to note is the inclusion of the *Hin*dIII restriction motive in the reading primer when reading 36nt during the sequencing reaction. This is illustrated by examing a typical sequence read from a 4C-seq experiment. In this example the first two nucleotides are reserved for a barcode followed by the sequence of the reading primer used in the amplification of the circles. The last nucleotides of the reading primers invariably contain the restriction site, often *Hin*dIII. In a typical 36nt sequencing run we will be left with sequences of 16nt (36nt – $primer_{20nt}$ or 36nt – $primer_{18\text{-}20nt}$ – $barcode_{2nt}$), which also shows the downside of using barcodes although this can be solved by running longer sequencing run (e.g. 50nt). Contrary to other sequencing experiments, however, 16nt sequences are sufficient for mapping because the reference genome is reduced. In panel **B** the PCR products of four (lane 1-4) different 4C experiments separated on a 1.5% agarose gel are shown. M indicates the λ*Pst*I marker. Please note the difference in length of the most abundant products in each lane, which can be used to verify 4C primer specificity.

libraries, from which it becomes clear that fragment ends longer than 14nt (read length 36nt) do not significantly increase the resolution of the experiment.

The sequence of the reading primer serves as a barcode and can be used to distinguish between 4C experiments directed to different viewpoints. They therefore allow the sequencing of multiple experiments in one Illumina sequencing lane. If one wishes to combine experiments that analyze the same viewpoint under different conditions (such as different

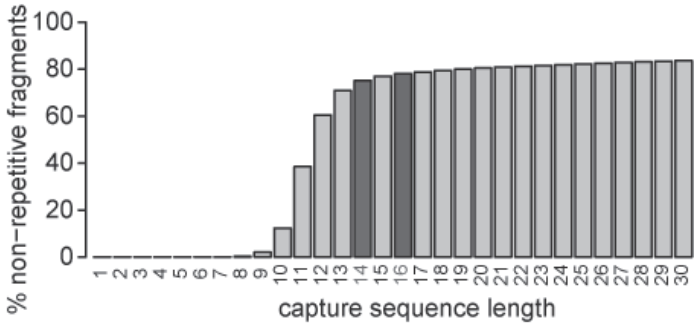


Figure 4. The fraction of unique fragment ends increases with its length in nucleotides. The length of a capture is typically 16nt (right dark grey bar) when using a 20bp reading primer omitting the tag sequence, resulting in ~80% unique potential captures. However including the 2nt tag, resulting in a 14nt capture (left dark grey bar), does not decrease the fraction of unique fragment ends (e.g. potential captures) dramatically.

cell types), a 2nt barcode can be added to the 5' end of the reading primer. The use of this additional barcode decreases the read length of the capture by 2nt, but this does not dramatically decrease the mappability of the captures (Figure 4). Typically 12-15 experiments are pooled within one sequencing lane, generating ~1million usable reads per experiment which is sufficient to generate a genome wide interaction profile. Note that the image analysis software in the Illumina GAII expects diversity in nucleotide content for every cycle and cannot (yet) handle a homogeneous nucleotide distribution for the clusters on the Illumina flow-cell. However, when sequencing the PCR product from a single 4C experiment the first 20nt originate from the 4C primer and will therefore be redundant. Therefore experiments with different 4C viewpoints should be pooled to accommodate limitations of the sequencing software. A remaining important consideration then concerns the simultaneous sequencing of the *Hin*dIII restriction sites, which can also cause errors in the Illumina software. By using reading primers with different lengths, for instance by using barcodes, this potential problem can be circumvented. Alternatively, experiments using a different primary RE, such as *Hin*dIII and *Eco*RI, can be pooled.

3.1 Primer design

As mentioned above, the reading primer needs to be designed on top of the 6nt restriction site (Figure 3A). Typically we aim to design primers around 20nt in length, with a TM of 55 °C, CG-content of 30-70% and limit the poly-N stretches to 4nt. The position of the primer located at the *Dpn*II end of the fragment is ideally positioned close to this restriction site in order to minimize PCR product size, but is otherwise flexible in its position. We aim to limit the position of the primer within 100 bp of the *Dpn*II site. Primer design is done using Primer3 (Rozen and Skaletsky 2000) <http://frodo.wi.mit.edu/primer3/>. Note that in contrast to regular PCR, where the primers point towards each other, in a 4C experiment the primers should face outward (inverse PCR). Automated primer design can be achieved by selecting the 100bp that flank the relevant *Dpn*II side and placing it in silico on the start of the fragment separated by a stretch of Ns (Figure 3A). By restricting the reading primer to the *Hin*dIII site, the correct reverse primer can be selected.

Although the success rate of the primer design is high, it is advisable to test the functionality of primer combinations. This can be tested on a 4C template and should result in a spectrum of differently sized PCR products, appearing as a 'smear' on an agarose gel, reflecting the multitude of different sized captures (Figure 3C). A control PCR on purified but otherwise untreated genomic DNA should not give this smear, but at most a few unspecific products. This test, although essential, does not reveal primer specificity *per se*. An indication of specificity is supplied by the presence of one, or often two, abundant bands within the smear of PCR products. These bands represent the 'undigested' and 'self-ligation' products, which are unique in size for every viewpoint used (Figure 3C). The 'undigested' product results from the ~10-20% undigested (or re-ligated) viewpoint fragment with the adjacent fragment on the linear genomic template. The 'self-ligation' product originates from the circularization of the *Hin*dIII fragment. Indeed, these are nearly always the most abundant sequencing products, together typically comprising ~20% (and sometimes more) of the sequence reads. When possible, this can be reduced by selecting a viewpoint that has both 'undigested' and 'self-ligated' products of >1kb in size.

3.2 PCR reaction

It is important to use a DNA polymerase that amplifies all PCR products equally, with as little bias for fragment length or abundance as possible. In our hands, the Expand Long Template polymerase (Roche) adheres most strongly to these criteria (Simonis et al. 2006), although this does not exclude other polymerases to function similarly.

After testing the 4C primers for their functionality and specificity, linear range of amplification is determined by using a dilution series of template. Ideally, a 4C primer-set with proven efficiency is included and can also serve as a control for 4C template quality. These control PCRs can be performed using a limited amount of template. For a representative sequencing dataset, a sufficient amount of ligation products must be analyzed. As mentioned, each cell can at most contribute two data points to the dataset. And most interactions will, irrespective of their biological significance, occur between DNA fragments relatively close on the linear DNA template (Dekker et al. 2002; Tolhuis et al. 2002; Simonis et al. 2006; Lieberman-Aiden et al. 2009). To be able to also robustly analyze interactions between fragments separated >1Mb or on different chromosomes (*trans*), many data points are required. This is accomplished by applying the 4C PCR to sufficient genome equivalents, which in our case means that we perform 16 PCR reactions on 200ng 4C template each, representing in total ~1 million ligation events per experiment.

3.2.1 Step by step protocol

PCR

58. Determine linear range of amplification by performing a PCR using template dilutions of 12.5, 25, 50 and 100ng 4C template. A typical 25µl PCR reaction consist of:

- 2.5µl 10X PCR buffer 1(supplied with the Expand Long Template Polymerase)
- 0.5µl dNTP (10mM)
- 35pmol forward primer (1.5µl of a 1/7 dilution from a 1µg/µl 20nt primer stock)
- 35pmol reverse primer (1.5µl of a 1/7 dilution from a 1µg/µl 20nt primer stock)
- 0.35µl Expand Long Template Polymerase (Roche #11759060001)
- X µl Milli-Q to a total volume of 25µl

A typical 4C-PCR program: 2' 94 °C; 10" 94 °C; 1' 55 °C; 3' 68 °C; 29x repeat; 5' 68 °C; ∞ 12 °C. The concentration of primers used in a 4C-PCR is typically three times higher than a regular PCR as this often facilitates the efficiency of amplification.

59. Separate 15µl PCR product on a 1.5% agarose gel and quantify to asses linear amplification and template quality.

60. Determine the functionality of the adaptor primers by comparing them with the 'short' primers from step 58. Note the volume of the adaptor primers is corrected for their length difference by using 4.5µl and 3µl of a 1/7 diluted 1µg/µl stock solution of the ~75nt reading primer and the ~40nt reverse primer. The adaptor primers should cause a shift in PCR product length which should be visible when separated and compared on a 1.5% agarose gel.

61. When satisfied about the quality and quantity of the PCR product generated using the adaptor primers, 4C template for sequencing is prepared as follows:
- 80µl 10X PCR buffer 1
- 16µl dNTP (10mM)
- 1.12nmol 75nt reading primer (24µl reading primer of a 1µg/µl 75nt primer stock)
- 1.12nmol 40nt reverse primer (16µl reverse primer of a 1µg/µl 40nt primer stock)
- typically 3.2µg 4C template
- 11.2µl Expand Long Template polymerase
- Milli-Q water till 800ul total

  Mix and separate into 16 reactions of 50µl before running the PCR

62. Collect and pool the 16 reactions. Purify the sample using the High Pure PCR Product Purification Kit (Roche #11732676001), which effectively separates between the non-used adaptor primers (~75nt) and the PCR product (>120nt). Use minimal two columns per 16 reactions.

63. Determine sample quantity and purity using the Nanodrop-spectrophotometer. Typically the yield resides between 10 and 20µg with A260/A280 ~1.85 and A260/A230 >1.5. Sample purity is important to control to prevent complications during the sequencing procedure. If absorption ratios deviate re-purification is advised.

64. Quality is determined by separation of 300ng purified PCR product on a 1.5% agarose gel.

65. Combine 4C PCR products of different experiments in preferred ratios for sequencing.

4. Sequence analysis

Like other quantitative sequencing methods, sequence reads from a 4C-seq data must be aligned to the genome. However, this process is not as straightforward as in a ChIP-seq experiment, for example. Our lab has built a custom pipeline for the analysis of 4C-seq data, of which the details and pitfalls will be outlined below.

4.1 Mapping of sequence data

The processed sequenced reads are compared to the fragment end library using custom perl scripts. Please note that MAQ (Li et al. 2008) cannot be used for the mapping procedure due to restrictions in the number of records allowed in the reference genome file. However, because we do not allow mismatches in the mapping procedure we can fairly quickly map sequence reads to the reduced genome. The mapping output takes into consideration whether the fragment contains a site for the second restriction enzyme (*Dpn*II) or not, in the latter case the fragment (and corresponding fragment

ends) are labeled *blind* fragments. Furthermore it is important to label whether fragment ends are unique in the genome, because non-unique sequences should be interpreted with extreme caution.

Post-processing

Feature-rich data files from the mapping procedure are ultimately compressed in wiggle track format (WIG) or GFF files that can be used in online genome browsers such as UCSC (Kent et al. 2002), or stand-alone browsers such as IGV (Robinson et al. 2011). Although these browsers can be used for visual inspection of the data, statistical analysis is crucial for any meaningful interpretation of the data.

4.2 Statistical analysis

Statistical analysis of 4C-seq data deals with the identification of regions that exhibit a higher sequence capture frequency than expected. For statistical analysis we advise using the R programming language (Team 2010). It is important to note that we transform the data to unique coverage (> 1 reads per fragment end is set to 1) to avoid possible PCR artifacts. Especially for long-range and *trans* interactions this is justified, because these captures are likely to originate from a single ligation event. This is reflected by the observation that coverage decreases to 5-10%, which means that the chance of observing the same read twice is between 0.25% and 1%. These percentages are negligible in our opinion. Note, that also sequences surrounding the viewpoint do possess also quantitative information which is ignored by this analysis. However, when interested in local interactions (within 2Mb from the viewpoint), it is recommended to apply an alternative 4C-seq approach better suited for this purpose (van der Werken et al. in prep). Below, the analysis to identify *cis* -and *trans*-interactions is described separately.

4.2.1 Identifying cis-interactions

Figure 5 explains the statistical procedure followed for the analysis of cis-interactions. A more formal definition is given below. Not taking biology into account, DNA folding is comparable to the behavior of any polymer, where interactions decline as a function of the distance to the viewpoint (i.e. high coverage close to the viewpoint and low at larger distances). To extract interacting regions with potential biological relevance we normalize the observed coverage for this background coverage. To this end, for a given window of fragment ends *i*, of size *w*, z-scores are calculated based on the relative unique coverage in the background window *W* ($p_{W,I} = cov_I/W$, where cov_I is the number of unique fragment ends covered in window *I*). *I* is chosen such that I»i and i runs from ⌈w/2⌉ until N-⌊w/2⌋, where N is the number of fragment ends on the chromosome. Window i spans i-⌊w/2⌋ until i+⌊w/2⌋ for odd values of w and (i-w/2)+1 until i+w/2 for even values of w. In general I = i, except for values where I < W and I > N-W, there I = W or I = N-W, respectively. Estimators for the mean and standard deviation, μ and σ, are calculated following the binomial distribution, for every window i given a window size w:

$$\mu_{w,i} = w \cdot p_W \,, \sigma_{w,i} = w \cdot p_{W,i} \cdot (1 - p_{W,i})$$

We use the relative unique coverage in window *w*, p_w, to calculate the z-score:

$$z_{w,i} = \frac{p_{w,i} - \mu_{w,i}}{\sigma_{w,i}}$$

As a rule of thumb, window sizes w=100 and W=3000 give reliable results. To identify regions of non-random 4C signals (i.e. contacted regions) the false discovery rate (FDR) is employed. To this end we randomly

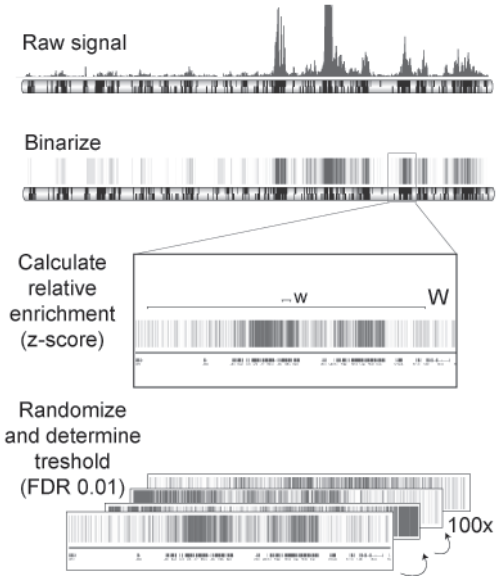


Figure 5. Outline of the 4C-seq analysis. After trimming the primer sequence from the raw reads the captures are mapped against the 'reduced genome'. Reads are made binary and the relative enrichment (z-score) is calculated using a sliding window of 100 compared to a 'background window' of 3000 fragment ends. Data permutation is subsequently used to determine a threshold with 0.01 FDR. Windows exceeding this threshold are scored 'interacting region'.

permuted the dataset 100 times and determined the threshold z-score at which the FDR was 0.01. Windows with a z-score above the FDR are scored as 'interacting region'.

4.2.2 Identifying trans-interactions

Although the analysis for identifying *trans* interactions is similar to the method for picking up *cis*-interactions, there are some important differences. Firstly there is no skewed data distribution, as for the cis chromosome around the viewpoint. Therefore no normalization is required, meaning we can use the raw coverage values as input for the FDR calculation. Second, because sequence captures on *trans* chromosomes are much less frequent than on the *cis* chromosome, a larger window size is required to pick up *trans*-interactions. For *trans* interactions an FDR threshold of 0.01 is determined based on 100 random permutations of the data for each chromosome. Typically, a window size of 500 unique fragment ends is used; windows that exceed the threshold are scored as *trans* interactions.

4.3 Visualization

For visualization and downstream analysis there are two options: 1) analysis of the discrete data, i.e. significantly contacted regions and 2) quantitative information, i.e. raw read counts or z-scores. The two options vary in their representation and will be discussed below.

4.3.1 Discrete data visualization

Tracks can be created for visualization in the UCSC genome browser or IGV by storing the position of the contacted region in BED files. The format of which is [chromosome,start,end] separated by tabs. An example of a custom visualization tool is the arachnogram shown in Figure 6A. Here the contacted regions are shown as branching out from the viewpoint, somewhat resembling a spider.

For *trans* interactions we use the excellent Circos tool (Krzywinski et al. 2009). Circos plots show the genome in a circular fashion, which compresses the visualization in an intuitive fashion (Figure 6B). However,

because of the wide range of applications and settings available to the researcher, setting up the correct configuration files for Circos is not a manner of point-and-click. We therefore provide example files that can be used for initially setting up Circos to analyze *trans* interactions, which are available on request.

4.3.2 Quantitative data visualization

In addition to visualizing the raw reads in the UCSC genome browser or in IGV, a multi-scale analysis analogous to the domainogram analysis (de Wit et al. 2008) can be performed. In this analysis z-scores (z_w) are calculated following the definition given in paragraph (4.2.1). z_w is calculated over a range of different values for w (w = 2..200) and stored in a matrix. For the visualization the z-score are transformed to probability scores based on the normal distribution and subsequently $\log_{10}$ transformed. In Figure 6C an example 4C-seq experiment is shown with a black-red-yellow color scheme.

4.4 Integration with external data sources

In order to assign biological meaning to the regions identified in a 4C-seq experiment it is often helpful to compare the identified regions with genomic data from other high-throughput sources. This is, however, one of the most precarious steps and one must be aware of confounding factors that can be present in the data. Here we will discuss some of the common pitfalls in data integration.

4.4.1 Scales of datasets

In Figure 7 an inventory of genome-scale methods (Simonis et al. 2006; Guelen et al. 2008; Lieberman-Aiden et al. 2009; Gilbert 2010; Lister et al. 2011) is shown and the approximate resolution at which they operate. The overview illustrates that there can be several orders of magnitude difference in size between the various genomic features. Although it is tempting to compare ChIP-seq data with raw 4C data obtained with the presented 4C strategy, we strongly advise against such an analysis. Although

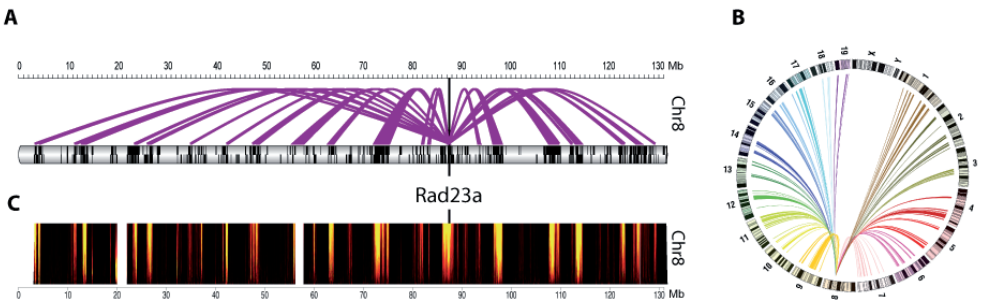


Figure 6. An overview of different visualization methods. For intuitive visualization, Cis-interacting regions can be depicted using the archanogram, or 'spiderplot'(**A**). In this example purple lines plotted above a representation of the *cis* chromosome connect the viewpoint with its distal interacting regions. The black bars within the chromosome represent the location of genes. The use of Circos plots when depicting *trans*-interactions provides a compact representation of interactions engaged with other chromosomes (**B**). Chromosomes are aligned in the outer ring while colored lines connect the viewpoint with its interacting regions on other chromosomes. To appreciate the significance of interactions 4C interactions can also be visualized using a domainogram (**C**). This visualizes a multi-scale analysis (y-axis) in which the significance of interaction is depicted by a color range, in this example ranging from black (not significant) to yellow (most significant).

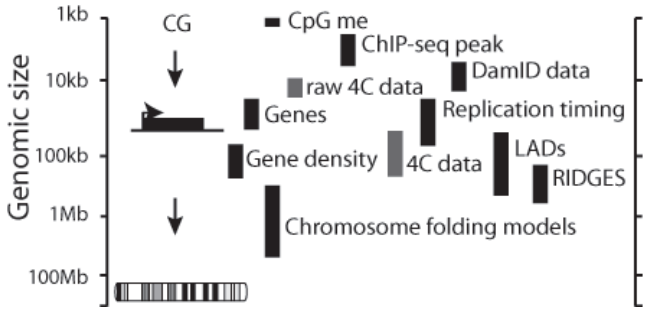


Figure 7. Different genomic features differ in their genomic size. When analyzing 4C interaction data it's tempting to directly compare these data to data of other genomic features available. However considerations have to be made when doing so as different genomic features describe different genomic scales.

there have been reports of restriction fragments specifically contacting each other over large distances (Ling et al. 2006; Hakim et al. 2009), it is our experience that interactions involve multiple restriction fragments, i.e. domains of interaction. Therefore a single restriction fragment end is often meaningless in a 4C analysis (and in fact not reproducible between biological replicate experiments), especially when this fragment is located more than a couple of Mbs away from the viewpoint. Rather one should apply windowed approaches as discussed above, extracting regions on the chromosome that are classified as interacting regions. As shown in Figure 7 these regions can often be directly compared to Lamin Associated domains (LADs) (Guelen et al. 2008) or Regions of IncreaseD Gene Expression (RIDGEs) (Caron et al. 2001; Versteeg et al. 2003). On the other hand comparisons can be made to ChIP-seq or DNA methylation data, but densities inside and outside 4C contacts should be calculated. However, in these analyses care must be taken to identify possible confounding factors in order to prevent over-interpretation of trivial observations.

4.4.2 Organization of the genome and 4C data interpretation

In this section at the hand of some practical examples a few points of concern are discussed when using analysis tools not specifically tailored for the interpretation of 4C data.

For instance, when 4C interacting regions from an active gene are compared to H3K4me3 ChIP-seq peaks, which mark the start of poised and active genes (Heintzman et al. 2007), it is possible to find a strong enrichment in these regions. However, in this respect it is important to point out that 3C-derived methods have the tendency to segregate the genome into gene-rich and gene-poor regions. It is therefore essential to correct for transcription start site density when interpreting these results, because a difference in gene density is often a more parsimonious explanation for the observed correlation.

Genome analysis has revealed that genes in the genome are non-randomly organized. Co-regulated or evolutionary duplicated genes often, and genes with similar function sometimes, are found clustered in the genome (Lee and Sonnhammer 2003; Pal and Hurst 2003; de Wit and van Steensel 2009). This is an important observation that has

consequences for the analysis of 4C data. The example below will concern GO enrichment, but is equally valid for co-expressed genes. Many GO enrichment tools rely on testing whether a set of genes with a given annotation is enriched among the target population using some statistical test (e.g. Fisher exact or hypergeometric test) (Huang da et al. 2009). How this might go wrong when directly applied on 4C data is illustrated by the following example. If a contacted region fortuitously overlaps with a cluster of olfactory receptor genes, this class of genes would show a highly significant enrichment. Although, this association is obviously present, the detected level of enrichment is biased and might lead to overstated conclusions.

A simple way to correct for the organization of the genome is to apply a circular permutation test (Figure 5). In this analysis, the identified regions are shifted along the chromosome in a given number of steps (typically 1000) and the underlying genomic structure is kept intact. By recalculating the enrichment of a given class of genes for every permutation a background distribution can be determined. Comparison of the original enrichment score to the background distribution allows for the calculation of a nominal p-value. Circular permutation is an important tool when one wants to determine enrichment in genomic domains of features that have a possible non-random distribution in the genome.

ACKNOWLEDGMENTS

The authors would like to thank Marieke Simonis and Patrick Wijchers for their contribution to the development of 4C-technology. This work was financially supported by grants from the Dutch Scientific Organization (NWO) to EdW (700.10.402, 'Veni') and to WdL (91204082 and 935170621), a FP7 Marie Curie ITN (PITN-GA-2007-214902) and a European Research Council Starting Grant (209700, '4C') to WdL.

5

CTCF MEDIATES LONG-RANGE CHROMATIN LOOPING AND LOCAL HISTONE MODIFICATION IN THE BETA-GLOBIN LOCUS

Erik Splinter[1], Helen Heath[1], Jurgen Kooren, Robert-Jan Palstra, Petra Klous, Frank Grosveld, Niels Galjart and Wouter de Laat

[1] These authors contributed equally

Dept. of Cell Biology and Genetics, Erasmus MC, PO Box 1738, 3000 DR Rotterdam, The Netherlands

Adapted from *Genes & Development* 2006 **20**(17): 2349-2354.

ABSTRACT

CTCF binds sites around the mouse β-globin locus that spatially cluster in the erythroid cell nucleus. We show that both conditional deletion of CTCF and targeted disruption of a DNA-binding site destabilize these long-range interactions, cause local loss of histone acetylation and gain of histone methylation, apparently without affecting transcription at the locus. Our data demonstrate that CTCF is directly involved in chromatin architecture and regulates local balance between active and repressive chromatin marks. We postulate that throughout the genome, relative position and stability of CTCF-mediated loops determine their effect on enhancer-promoter interactions, with gene insulation as one possible outcome.

INTRODUCTION

Chromatin insulators are DNA sequences that confer autonomous expression on genes by protecting them against inadvertent signals coming from neighboring chromatin. CTCF is the prototype vertebrate protein exhibiting insulator activity (Defossez and Gilson 2002; Recillas-Targa et al. 2002; West et al. 2002), that can act as an enhancer-blocker or as a barrier against repressive forces from nearby heterochromatin *in vitro* (Defossez and Gilson 2002; Recillas-Targa et al. 2002; West et al. 2002). *In vivo*, CTCF binds to the imprinting control region of the *H19*/insulin-like growth factor (*Igf2*) locus, where it acts as a methylation-sensitive enhancer-blocker (Bell and Felsenfeld 2000; Hark et al. 2000; Fedoriw et al. 2004). Moreover, CTCF binding sites have been found and its insulator activity has been anticipated at the imprinting center that determines choice of X inactivation (Chao et al. 2002), at boundaries of domains that escape X inactivation (Filippova et al. 2005) and at sites flanking CTG/CAG repeats at the *DM1* locus (Filippova et al. 2001; Cho et al. 2005). CTCF was first defined as an insulator protein when it was found to be required for the enhancer-blocking activity of a hypersensitive site 5' of the chicken β-globin locus (5'HS4) (Bell et al. 1999). A similar CTCF-dependent insulator site was subsequently found at the 3' end of the locus and both sites coincide with erythroid-specific transitions in DNase I sensitivity of chromatin (Saitoh et al. 2000). Such observations suggested that CTCF partitions the genome in physically distinct domains of gene expression. The molecular mechanism underlying CTCF's insulating activity is still unknown.

CTCF-binding sites also flank the human and mouse β-globin locus (Fig. 1A), which contain a number of developmentally regulated, erythroid-specific β-globin genes and an upstream locus control region (LCR) required for high β-globin expression levels. In mice, three CTCF-binding sites have been identified upstream (HS-85, HS-62 and HS5) and one downstream (3'HS1) of the locus (Farrell et al. 2002; Bulger et al. 2003). Previously we have applied chromosome conformation capture (3C) technology (Dekker et al. 2002) to study long-range DNA interactions between these and other sites in the β-globin locus. In erythroid cells, the CTCF-binding sites (including HS-85, see below) were found to participate in spatial interactions between the LCR and the active β-globin genes, and collectively form

an Active Chromatin Hub (ACH) (Tolhuis et al. 2002). No such long-range DNA interactions were detected in non-erythroid cells. However, in established I/11 erythroid progenitor cells that do not yet show activated β-globin gene expression, contacts between the LCR and the genes are absent, but long-range DNA interactions already exist between the hypersensitive sites that contain CTCF-binding sites (Palstra et al. 2003). Here, we investigated the involvement of CTCF in the formation of these loops.

5 RESULTS AND DISCUSSION

β-globin locus conformation in erythroid cells with reduced levels of CTCF protein

To investigate the role of CTCF in the formation of chromatin loops, we analyzed β-globin DNA contacts in cells lacking the CTCF protein. Analysis was focused on E12.5 erythroid progenitor cells because they can be expanded *ex vivo* (Dolznig et al. 2001), lack stable LCR-gene contacts and therefore best reveal the interactions between outer hypersensitive sites. Chromatin immunoprecipitation (ChIP) experiments revealed that CTCF was bound *in vivo* to cognate sites in the β-globin locus in these cells (Fig. 1F), while the protein was absent from HS5 and 3'HS1 in brain cells not showing these loops (Supplemental Fig. 1). Since CTCF-*null* mice die early during embryogenesis, a conditional knock-out mouse model was generated by inserting two lox sites upstream and downstream of the first and last coding exon of CTCF, respectively. To delete CTCF, fetal liver cells were isolated from lox/lox E12.5 embryos, cultured under conditions which select erythroid progenitors, and infected with a replication-deficient lentivirus expressing Cre recombinase (Fig. 1B). Heterozygous (lox/wt) cells from littermate embryos underwent the same treatment and served as controls. Cre recombination resulted in nearly 100% deletion of targeted CTCF alleles, with a reduction in mRNA and protein levels to 2-3% and 10-25%, respectively, in lox/lox as compared to wild-type (Fig. 1C-E). CTCF binding to cognate sites in the β-globin locus was reduced but not completely abolished in lox/lox cells, as demonstrated by ChIP (Fig. 1F). To investigate β-globin locus conformation in these cells by 3C technology, we used a novel taqman probe-based quantitative polymerase chain reaction (Q-PCR) strategy to accurately quantify 3C ligation efficiencies (Supplementary Fig. 2). We found that the structure of the β-globin locus in wild-type and lox/wt E12.5 progenitor cells was essentially the same as previously observed in I/11 progenitor cells (data not shown), with long-range interactions between the CTCF-binding sites HS-85, HS-62/60, HS4/5 and 3'HS1 (Fig. 1G). In lox/lox cells containing lower levels of CTCF protein, however, clearly reduced DNA-DNA interaction frequencies were observed specifically between the sites that normally bind CTCF (Fig. 1G-H). This is true for all combinations of binding sites, except for the interaction between 3'HS1 and HS-62 (but see below). The results demonstrated that CTCF is required for long-range DNA-DNA interactions between cognate binding sites in the β-globin locus. Gene expression analysis revealed the same, low, levels of expression for all β-globin genes in wild-type versus lox/lox progenitor cells (Supplementary Fig. 3). Moreover, we did not find activation of any of the mouse olfactory genes immediately surrounding the β-globin locus (MOR5B1-3

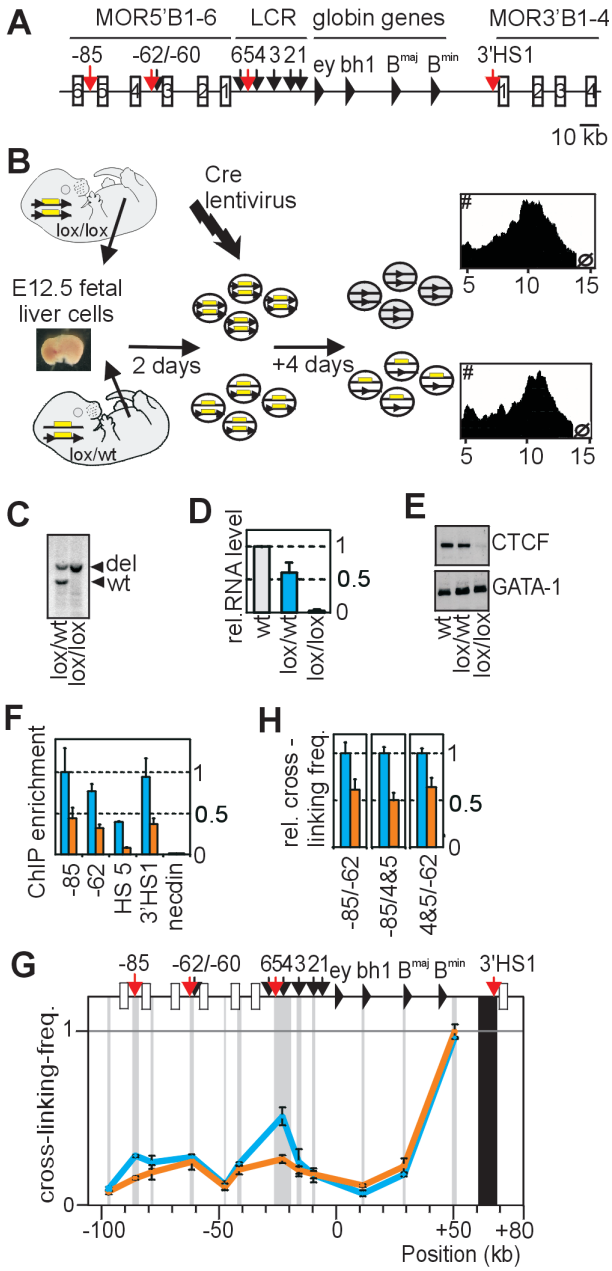


Figure 1. Deleting CTCF in primary erythroid progenitors reduces the frequency of interactions between cognate binding sites in the β-globin locus. **(A)** Schematic presentation of the mouse β-globin locus. DNase I hypersensitive sites (arrows) and CTCF-binding sites (red) are indicated. **(B)** Strategy to delete CTCF. Plots show a similar cellular size distribution for homozygous and heterozygous conditional knock-out cells. **(C)** Southern blot analysis showing complete deletion of CTCF conditional knock-out alleles. **(D)** CTCF mRNA levels (as determined by quantitative RT-PCR) in untreated wild-type (level set to 1) and Cre-recombined lox/wt and lox/lox cells. Standard deviation is indicated. **(E)** Western blot analysis of CTCF protein and GATA-1 protein (control, stripped and re-hybridized blot). **(F)** CTCF ChIP analysis. Lox/lox (orange) versus control cells (blue) **(G)** 3C analysis, demonstrating reduced interaction frequencies between 3'HS1, HS4/5 and HS-85 in lox/lox cells (orange), as compared to control cells (blue). **(H)** 3C analysis, demonstrating reduced interaction frequencies in lox/lox cells between the other CTCF-binding sites in the β-globin locus.

and MOR3B1-4) (data not shown). Hence, the reduction of CTCF protein to low levels had no appreciable effect on gene expression at, or around, the β-globin locus in erythroid cells representing a differentiation stage prior to LCR-mediated gene activation.

Long-range interactions of 3'HS1 containing nucleotide changes that disrupt CTCF-binding

The structural changes in the β-globin locus that we observed in cells with deleted CTCF may be a direct consequence of reduced protein-binding to the locus, or could be caused by secondary pathways that fail to act on the locus in the absence of sufficient CTCF. To investigate this, we disrupted CTCF-binding locally by changing 4 conserved nucleotides in the core CTCF-binding site of the endogenous 3'HS1 (Supplementary Fig. 4). Bandshift assays confirmed that these alterations completely abolished CTCF-binding *in vitro* (Supplementary Fig. 5). Targeting was performed in ES cells that were established from a cross between the two inbred strains 129 and C57BL/6 (B6) and was directed to the B6 allele. Two additional, non-conserved, nucleotides were changed 70 basepairs downstream of the core CTCF-binding site to allow allele-specific analysis of CTCF-binding to 3'HS1 by chromatin immunoprecipitation (ChIP). Moreover, an extra HindIII restriction site was introduced ~850 bp downstream of the CTCF-binding site, which enabled us to exclusively analyze DNA interactions of the targeted 3'HS1 by 3C. An independent control ES line was generated containing the extra HindIII site with the normal 3'HS1. In each cell line the neomycin selection cassette was removed by transient expression of Cre recombinase, leaving behind a single lox site immediately downstream of the newly introduced HindIII site (Supplementary Fig. 4).

Definitive erythroid progenitors were generated from the ES cells *in vitro* (Carotta et al. 2004)to analyze the consequences of the targeted nucleotide changes in erythroid cells. We established two such ES-EP cell lines, ES-EP(Δ3'HS1) (or 'Δ' and the control line ES-EP(c) (or 'c') (Fig. 2A). After validation of the cells as a model system for erythroid differentiation (Supplemental Fig. 6) (Carotta et al. 2004), we analyzed CTCF-binding to mutated and wild-type 3'HS1 *in vivo*. In the control line ES-EP(c), CTCF bound strongly and equally well to 3'HS1 on both alleles. In ES-EP(Δ3'HS1) however, binding to 3'HS1 on the non-targeted 129 allele was the same as in ES-EP(c), but binding to the mutated 3'HS1 on the targeted B6 allele was completely abolished (Fig. 2B). Thus, the change of 4 nucleotides in the core CTCF-binding site prevented binding of CTCF to 3'HS1 also *in vivo*. Next we analyzed whether disruption of CTCF-binding at 3'HS1 affected its long-range DNA interactions in the β-globin locus. The extra HindIII restriction site introduced downstream of 3'HS1 was used to focus 3C analysis exclusively on the targeted B6 allele. In undifferentiated ES-EP(c), the wild-type β-globin B6 allele formed a chromatin hub typically observed in normal erythroid progenitor cells, with 3'HS1 interacting with HS4/5, HS-62 and HS-85 (Fig. 2C). In undifferentiated ES-EP(Δ3'HS1) however, the mutated 3'HS1 showed a dramatic drop in interaction frequencies with all these DNA elements, to levels similar to those observed in non-expressing fetal brain cells (Fig. 2C). Thus, disruption of CTCF-binding to 3'HS1 severely destabilized the large chromatin loop containing the LCR and the globin genes in erythroid progenitor cells. The fact that the interaction with HS-62 was lost upon targeted disruption of CTCF-binding to 3'HS1, but not in our conditional CTCF knock-out

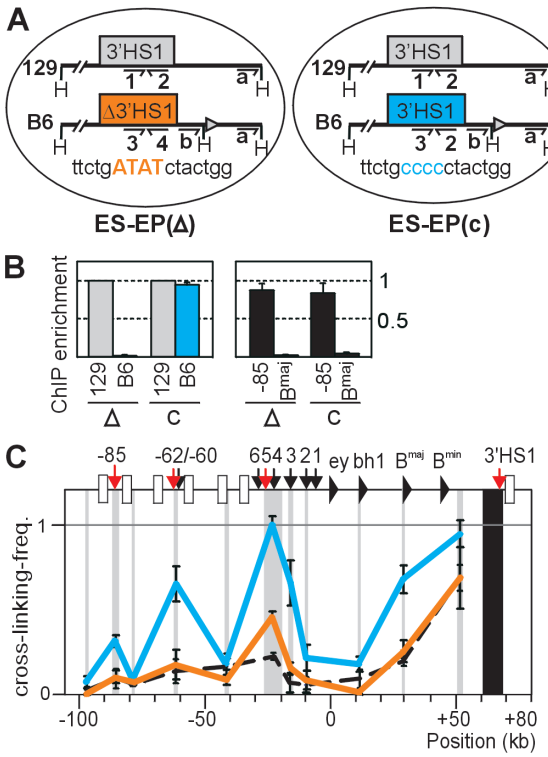


Figure 2. Targeted nucleotide changes in 3′HS1 disrupt CTCF-binding and reduce frequency of long-range 3′HS1 interactions. **(A)** Erythroid progenitor cell lines derived *in vitro* from ES cells. ES-EP(Δ3′HS1) harbors 4 targeted nucleotide changes in the core CTCF-binding site of 3′HS1 on the B6 allele (orange). ES-EP(c) contains wild-type 3′HS1 on the B6 allele (blue). The non-targeted, intact, 129 allele is in grey. For ChIP, each 3′HS1 CTCF-binding site can be analyzed with a unique primer pair (#1-4). An extra HindIII site targeted downstream of 3′HS1 allows exclusive 3C analysis of B6 allele (with 3C-primer 'b'). **(B)** ChIP on undifferentiated ES-EP cell lines, with antibody against CTCF. Left: 3′HS1 alleles in the two ES-EP lines (Δ and c); right: positive (HS-85) and negative (βmajor) controls. **(C)** 3C analysis with primer 'b' (see above). Note that interaction frequencies with mutated 3′HS1 (orange) are reduced compared to wild-type 3′HS1 (blue). Black hatched line: 3′HS1 interactions in fetal brain, analyzed with primer 'a' and plotted for comparison.

experiments, suggests that this interaction is more resistant than others to the reduction of levels of CTCF protein.

Expression of β-globin and surrounding olfactory receptor genes in the absence of CTCF-mediated chromatin loops

Since the large, CTCF-dependent, loops are formed only in human or mouse cells that are committed to, or do highly express the β-globin genes (Palstra et al. 2003), we investigated the relationship between these loops and transcriptional regulation in detail. First, we analyzed whether CTCF at 3′HS1 serves as an enhancer-blocker that prevents the inappropriate activation of downstream mouse olfactory receptor genes (MORs) by the β-globin LCR in erythroid cells, as suggested previously (Farrell et al. 2002). For this, we compared mRNA levels of the MOR3′B1-4 genes between differentiated

ES-EP(c) and ES-EP(Δ3'HS1) cells when the LCR is fully active. We found no inappropriate activation of any of the downstream MORs, nor of MOR5'B3, in the differentiated ES-EP(Δ3'HS1) cells (data not shown) and we concluded that insulator activity of CTCF at 3'HS1 is not required for blocking LCR-mediated activation of downstream MOR genes in ES-EP cells. Noteworthy, previously it was found that deletion of the complete HS5 from the endogenous locus also had no effect on expression of the surrounding MOR genes (Bulger et al. 2003). We envision that the transcription factor environment in erythroid cells does not support olfactory receptor genes to be activated. Next, we analyzed whether the CTCF-dependent loops influence β-globin gene expression. Upon erythroid differentiation, the LCR forms stable contacts with the active β-globin genes and strongly up-regulates their transcription rate (Carter et al. 2002; Tolhuis et al. 2002). We reasoned that a shared presence on one chromatin loop anchored by CTCF in progenitor cells would decrease the spatial distance between LCR and genes, which may facilitate their productive interaction later during differentiation. If true, absence of such a pre-existing loop could possibly result in a delay of full β-globin gene activation. To test this, we compared the kinetics of LCR-mediated gene activation between the individual alleles of differentiating ES-EP(c) and ES-EP(Δ3'HS1) cells. Two sets of βmajor intron primers were designed that allowed independent analysis of ongoing transcription from the B6 allele and 129 allele (data not shown). ES-EP(c) and ES-EP(Δ3'HS1) cells were induced to undergo synchronous differentiation and RNA was collected at various time intervals. As expected, βmajor transcription rates increased strongly upon differentiation. However, at each given stage of differentiation, we detected the same gene activity between the 129 and B6 allele, both in ES-EP(c) and ES-EP(Δ3'HS1) cells (Fig. 3A). Thus, the CTCF-dependent chromatin loop with 3'HS1 that topologically defines the β-globin locus in erythroid cells does not detectably influence the expression kinetics or levels of the β-globin and nearby OR genes. This was also true for the embryonic β-globin genes εy and βH1, which were off in both cell lines before and after differentiation (data not shown).

Establishment of LCR-gene contacts in the absence of a pre-existing loop with 3'HS1

The unaltered β-globin gene expression patterns from the targeted allele in ES-EP(Δ3'HS1) cells suggested that in the absence of a pre-existing chromatin loop LCR-gene contacts can still be established normally upon erythroid differentiation. To test this, we searched for 129/B6 polymorphisms near restriction sites in the LCR that would allow allele-specific 3C analysis. This resulted in the design of a taqman probe for a HindIII fragment encompassing HS4 and HS5 that signals exclusively from the B6 allele (data not shown). Although HS4 and HS5 are not prime candidates in the LCR to directly contact the genes, this relatively large HindIII fragment was previously shown to be representative of the complete LCR, since it displayed a prominent peak of interaction with the β-major gene upon erythroid differentiation (Tolhuis et al. 2002). In both differentiated ES-EP(c) and ES-EP(Δ3'HS1) cell lines, we found identical locus-wide interaction frequencies for HS4/5 between the B6 alleles containing either wild-type or mutated 3'HS1, and both showed a strong peak of interaction with the β-major gene (Fig. 3B). This demonstrated indeed that a pre-existing

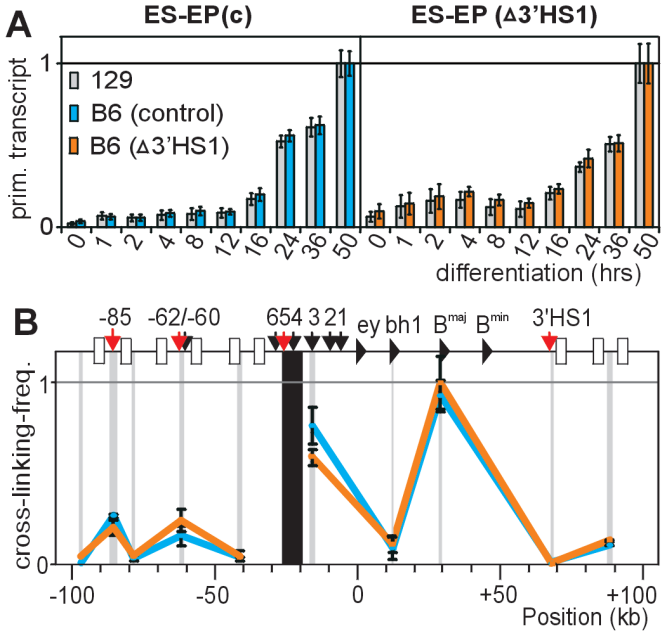


Figure 3. Targeted nucleotide changes in 3'HS1 do not affect β-globin gene expression. **(A)** Ongoing βmajor transcription as measured by quantitative RT-PCR, using 129 and B6 specific primers against intron 2 of the βmajor gene. X-axis: hours after induction of differentiation. Error bars represent standard-error of mean. **(B)** Locus-wide, B6-specific, analysis of interaction frequencies with HS4/5 after differentiation. Note that primer 'a' was used near 3'HS1, which on the B6 allele analyzes a small (0.5 kb) fragment downstream of (i.e. not containing) 3'HS1, whereas on the 129 allele this primer would analyze an ~8kb fragment encompassing 3'HS1. The dramatic drop in interaction frequencies shows that analysis is restricted to the B6 allele.

loop between upstream sites and 3'HS1 is dispensable for the establishment of stable LCR-gene contacts later during erythroid differentiation. Such a conclusion is in agreement with transgenic experiments showing full β-globin expression from constructs lacking 3'HS1 (Strouboulis et al. 1992).

Histone modifications in the absence of CTCF binding

3'HS1 was previously shown to be present in, and close to the 3' border of an open chromatin domain spanning ~145 kb around the β-globin locus in erythroid cells (Bulger et al. 2003). Within this domain, a large (~15kb) region of highly repetitive DNA is present approximately 3kb upstream of 3'HS1, that cannot be analyzed for nuclease sensitivity, but likely is packed into compact chromatin. To further investigate this, we analyzed histone modifications at and directly around 3'HS1 in ES-EP cells. Using an antibody that recognizes both di-methyl H3K9 and di-methyl H3K27, we found that these repressive marks were abundantly present on both sides of 3'HS1, but not inside 3'HS1 (Fig. 4A). Conversely, acetylation of histone H3, a mark for open chromatin, was clearly enriched at 3'HS1 but not, or much less, at sites surrounding the hypersensitive site (Fig. 4C). These data argued against the existence of a large open chromatin domain

extending across 3'HS1 and suggested that 3'HS1 forms an isolated entity of open chromatin. To address whether CTCF plays a role in the establishment of this pattern, we performed ChIP on ES-EP(c) and ES-EP(Δ3'HS1) cells and used allele-specific primer pairs to compare modifications at 3'HS1 on targeted, versus non-targeted alleles. In the control cell line, we found identical levels of di-meH3K9/K27 at 3'HS1 on the two alleles. In ES-EP(Δ3'HS1) however, loss of CTCF binding was accompanied by an increase of di-meH3K9/K27 and concomitant decrease of AcH3 at 3'HS1 (Fig. 4B-C). We found no indication for spreading of the methyl mark into the locus, not locally (compare levels of enrichment between Fig.4A and 4B), nor at the βmajor gene, which locates more inside the locus (Fig.4B, analyzed by allele-specific primers). In fact, also AcH3 levels at βmajor were similar for the targeted and non-targeted allele in ES-EP(Δ3'HS1) cells, two observations that were fully in agreement with our finding that βmajor gene expression was not affected by disrupted CTCF binding to 3'HS1 (Fig. 3A).

We considered the possibility that spreading of di-meH3K9/K27 into the locus requires disruption of CTCF binding to more sites than just 3'HS1. To investigate this, we compared di-meH3K9/K27 levels in E12.5 wild-type versus lox/lox conditional CTCF knock-out progenitor cells, the latter containing reduced amounts of CTCF (Fig. 1). We found that loss of CTCF binding to 3'HS1, HS5, HS-62 and HS-85 coincided with locally increased amounts of di-meH3K9/K27, while modification levels elsewhere in the locus appeared unaffected (Fig. 4D). Since CTCF-binding to β-globin sites was reduced but not absent in lox/lox cells, this leaves open the possibility that residual CTCF amounts prevent spreading of di-meH3K9/K27 into the locus. We concluded that CTCF regulates the balance between active and repressive chromatin modifications at its binding sites and we propose that CTCF-mediated acetylation of histones prevents their methylation. Mechanistically, CTCF may directly attract histone acetyl-transferases (HATs), although current evidence for this interaction is lacking. Alternatively, CTCF-mediated chromatin looping brings binding sites into spatial proximity with HATs bound elsewhere to the DNA (de Laat and Grosveld 2003). The observation that CTCF-binding was required for histone acetylation is interesting because previously these two events were claimed to be uncoupled (Recillas-Targa et al. 2002). Our data do not support the generality of boundaries demarcating expression domains, but fit better with the concept that genes maintain autonomous expression profiles mostly through their unique ability to productively interact with positive regulatory elements (Dillon and Sabbattini 2000; de Laat and Grosveld 2003).

CTCF organizes higher-order chromatin structure

We have presented two independent lines of evidence which together firmly establish that CTCF functions in the formation of chromatin loops; removal of most CTCF protein, as well as targeted disruption of a CTCF-binding site, resulted in destabilization of long-range contacts between cognate binding sites in the β-globin locus. CTCF is critical for the looped conformation present in erythroid progenitor cells, but is dispensable for LCR-gene contacts established later during differentiation. We, and others, previously have shown that the latter contacts depend on the transcription factors EKLF and GATA-1 (Drissen et al. 2004; Vakoc et al. 2005). Together, these studies begin to

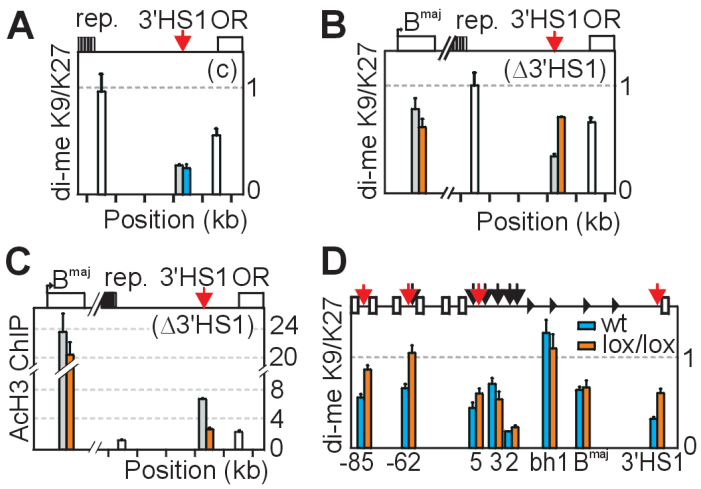


Figure 4. Histone modifications in the absence of CTCF binding. **(A)** ChIP enrichment for di-meH3K9/K27 in undifferentiated control ES-EP(c) cells on the 129-allele (grey), B6-allele (blue) or on both alleles (white). Note that values in 4*A-C* were normalized to input and therefore also white bars represent enrichment per allele. **(B, C)** ChIP enrichment for di-meH3K9/K27 (*B*) and acetylated histone H3 (*C*) in undifferentiated ES-EP(Δ3'HS1) cells. Orange bars: B6-allele, grey bars: 129-allele. Note increased methylation (*B*) and decreased acetylation (*C*) only at mutated 3'HS1 on B6-allele **(D)** ChIP enrichment for di-meH3K9/K27 at the β-globin locus in control (blue) and conditional CTCF knockout (lox/lox) (orange) E12.5 erythroid progenitors cells. Note that reduced levels of CTCF cause an increase in di-meH3K9/K27 specifically at the CTCF-binding sites HS-85, HS-62, HS5 and 3'HS1.

delineate the factors that act sequentially to form a functional β-globin ACH in differentiated erythroid cells where β-globin genes are fully expressed. Based on the observations that CTCF-dependent chromatin loops are tissue-specific and evolutionary conserved between mouse and man, it seemed reasonable to expect that these loops would play a role in gene expression. Such function may exist but is beyond our current detection limits. An alternative view is that evolutionary selection against sites forming chromatin loops within a gene locus positions them outside the β-globin locus, without necessarily being selected to act, positively or negatively, on gene expression (Dillon and Sabbattini 2000).

We hypothesize that CTCF also organizes higher-order chromatin structure at other gene loci and we predict that such chromatin loops facilitate communication between genes and regulatory elements but can also lead to the exclusion of interactions between elements. In terms of transcriptional regulation, the final outcome of such chromatin loops will depend on the position of CTCF-binding sites relative to other regulatory elements and the genes, the concentration of the trans-acting factors involved, and the affinities of the (long-range) interactions. In *Drosophila*, a limited 3C analysis previously provided indications for a loop between the *scs* and *scs'* enhancer blocking elements (Blanton et al. 2003). Moreover, insulator proteins like suppressor of Hairy-wing (Su[Hw]) and Modifier of mdg4 2.2 (Mod[mdg4]2.2) have been found to coalesce into large foci, called insulator bodies. These bodies preferentially localize at the nuclear periphery and are hypothesized to bring together distant

insulator sites, with intervening chromatin fibers looped out to form isolated expression domains (Gerasimova et al. 2000). Our observations made on the CTCF protein provide high-resolution insight into the nature of such loops in mammals. It will be interesting to see if CTCF forms chromatin loops through multimerization of CTCF molecules bound to distinct DNA elements (Yusufzai and Felsenfeld 2004), or whether this loop formation also involves other factors. Similarly, future experiments should provide insight whether CTCF-dependent chromatin looping occurs at a defined physical structure in the nucleus (Dunn et al. 2003; Yusufzai and Felsenfeld 2004; Yusufzai et al. 2004) or whether the base of such loops has a more fluid nature.

MATERIALS AND METHODS

Generation of conditional CTCF knock-out mice and CTCF antibody

Targeting constructs and strategy for the generation of conditional CTCF knock-out mice as well as the polyclonal antibody generated against CTCF is described in detail as part of a study that addresses the role of CTCF in T cell development (Heath et al. 2008).

Lentivirus production and infection

Cre-lentivirus was produced by transient transfection of 293T cells according to standard protocols (Zufferey et al. 1997). 293T cells were transfected with a 3:1:4 mixture of psPAX-2, pMD2G-VSVG (kind gifts of D. Trono, University of Geneva) and a transfer vector construct that is essentially as pRRLsin.sPPT.CMV.GFP.Wpre (Follenzi et al. 2002) but with CMV-Cre instead of CMV-GFP, using poly(ethylenimine) (PEI). After 24 hours medium was refreshed and virus-containing medium was harvested 48 and 72 hours after transfection. After filtration through a 0.45µm cellulose acetate filter, the virus stock was concentrated 1000 times by centrifugation at 19.4K rpm for 2 hours at 10^0C in a SW28 rotor. Virus stocks were stored at -80^0C. Virus activity/functionality was tested by serial dilutions on primary mouse embryonic fibroblasts (MEFs) containing loxP sites, which were scored for recombination after 4 days of infection by Southern blotting. Fetal livers were isolated from E12.5 embryos, resuspended in FCS with 10% DMSO by repeated pipetting and stored in liquid nitrogen until genotyping of embryos was completed. Per experiment, cells from three fetal livers of the same genotype were pooled and cultured as described (Dolznig et al. 2001). After 2 days of culturing, cells were infected by adding Cre-lentivirus to medium and centrifugation of cell culture plates for 55 minutes at 2.5K rpm (37^0C). Cre-mediated recombination efficiency was analyzed by standard Southern en western blotting techniques (antibody used to detect GATA-1: #N6, Santa-Cruz). CTCF RNA levels were analyzed by quantitive RT-PCR (see below).

3C Analysis

3C analysis was performed essentially as described (Splinter et al. 2004), using HindIII as the restriction enzyme. Quantitative real-time PCR (Opticon I, MJ Research) was performed with Platinum Taq DNA Polymerase (Invitrogen) and double-dye oligonucleotides (5'FAM + 3'TAMRA) as probes, using the following cycling conditions: 94^0C for 2 min and 44 cycles of 15 s at 94^0C and 90 s at 60^0C.

Chromatin Immunoprecipitation (ChIP)

ChIP was performed as described in the Upstate protocol (http://www.upstate.com), except that cells were cross-linked at 2% formaldehyde for 5 minutes at room temperature. Quantitative real-time PCR (Opticon I, MJ Research) was performed using SYBR Green (Sigma) and Platinum Taq DNA Polymerase (Invitrogen), under the following cycling conditions: 94^{0}C for 2 min, 44 cycles of 30 s at 94^{0}C, 60 s at 55^{0}C, 15 s at 72^{0}C and 15 s at 75^{0}C (during which measurements are taken). Enrichment was calculated relative to Necdin and values were normalized to input measurements. Antibodies used: Anti-acetyl-Histone H3 (#06-599, Upstate); anti-dimethyl Histone H3 K9/K27 (ab7312, Abcam).

Analysis of gene expression

Total RNA was isolated from cultured fetal liver cells or 0.5-1 x 10^6 of ES-EPs at the indicated time points using TRIzol reagent (Invitrogen) according to the manufacturer's instructions. 1 μg of RNA was treated for 1 hour at room temperature with amplification grade DNAseI (Invitrogen) to remove genomic DNA contamination. An aliquot was used as a no RT control. cDNA synthesis was performed using Superscript II RNase H-Reverse transcriptase (Invitrogen) according to the manufacturer's instructions using 200ng random hexamers as primers. Quantification of primary transcripts was performed on Opticon II real-time PCR machines (MJ research) using Platinum Taq (Invitrogen) and SYBR Green (Sigma), using the following PCR program: 2 min 94^{0}C, 45 cycles of 30 sec 94^{0}C, 1 min 62^{0}C, 15 sec 72^{0}C and 15 sec 75^{0}C (during which measurements are taken), followed by 10 minutes chain extension and a melting curve. Expression was normalized against HPRT expression levels.

Targeting nucleotide changes to 3'HS1 in ES cells

The 3'HS1 targeting constructs were based on a 5.6 kb BamHI-EcoRV isolated from BAC RP23-370E12 (BACPAC Resources) (Supplementary Fig. 1). Site-directed mutagenesis was performed on an internal 683 bp NdeI-NdeI fragment (coordinates: 67033-67716, see Supplementary Fig. 1), using QuikChange® Multi Site-Directed Mutagenesis Kit (Stratagene). Oligonucleotide used to change CTCF-binding site: CGGAAATCAGCGGAACACTTCTGATATCTACTGGTATGCAACAGG.

Oligonucleotide used to change 2 nucleotides 70 bp downstream of the core CTCF-binding site: CAGTTTATCCCAGTTACGTTTAGTTGACAACCTGAGAC. Before reintroduction into the BamHI-EcoRV targeting vector, the complete NdeI fragment was sequenced to confirm that only targeted nucleotides were changed. A TK-NEO cassette flanked by head-to-tail oriented loxP sites and containing a HindIII site immediately upstream one of the loxP sites was inserted as an XbaI-SpeI fragment into the AvrII site at position 68251 (Supplementary Fig. 1). For selection against random integration events, diphtheria toxin (DTA) (Yu et al. 2000) was cloned outside the region of homology. ES cells for targeting were isolated from 129xB6 F1 blastocysts and transfected with the SalI linearized targeting construct by electroporation. Clones scored positive for homologous recombination at 3'HS1 by Southern blot hybridization were transiently transfected with a CMV-Cre construct containing a PGK-puromycin selection cassette, followed the next day by a 40 hours selection on medium containing 2 μg/ml of puromycin. Surviving clones were analyzed by Southern blotting for successful Cre-mediated deletion of the neomycin selection cassette and by

PCR analysis for the absence of Cre. Positive clones were selected for *in vitro* differentiation into ES-EP cells.

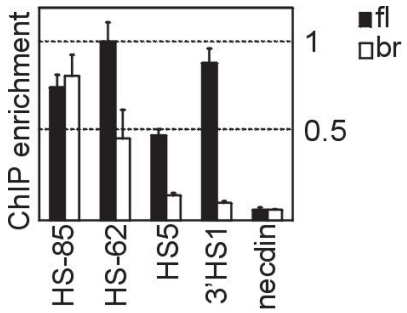
In vitro differentiation of ES cells into ES-EPs and characterization of ES-EPs.

Differentiation of ES cells into ES-EPs, expansion of ES-EPs and *in vitro* differentiation of ES-EPs into eryhrocytes was performed as described (Carotta et al. 2004) (Epo was a kind gift from Ortho-Biotech), except that embryoid bodies were formed in 4000-6000 hanging drop cultures (~200 ES cells/drop) that were pooled and disrupted after 6 days of culturing to generate ES-EPs. After 2 to 3 weeks of cultivation a homogeneous erythroid progenitor population was obtained (Supplemental Fig. 6).

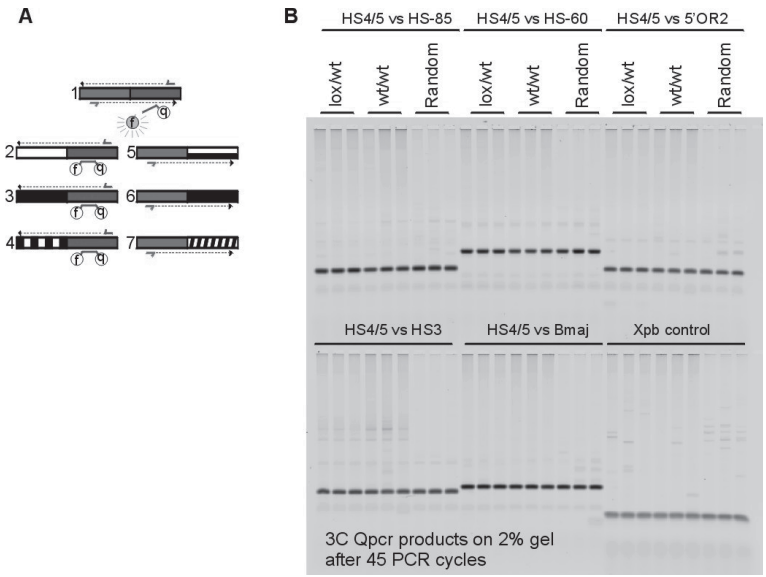
ACKNOWLEDGEMENTS

We thank A. de Wit, T. van Dijk and M. von Lindern for help with experiments. This work was supported by a grant from the division of Earth and Life Sciences (ALW) of the Netherlands Organization for Scientific Research (NWO), through the ESF EuroDYNA programme to N.G., EU grants LSHG-CT-2004-503433 (Epigenetic) and LSHM-CT-2003-504468 (Cells into Organs) and NWO grant 912.03.009 to F.G. and NWO grants 016-006-026 and 912-04-082 to W.d.L.

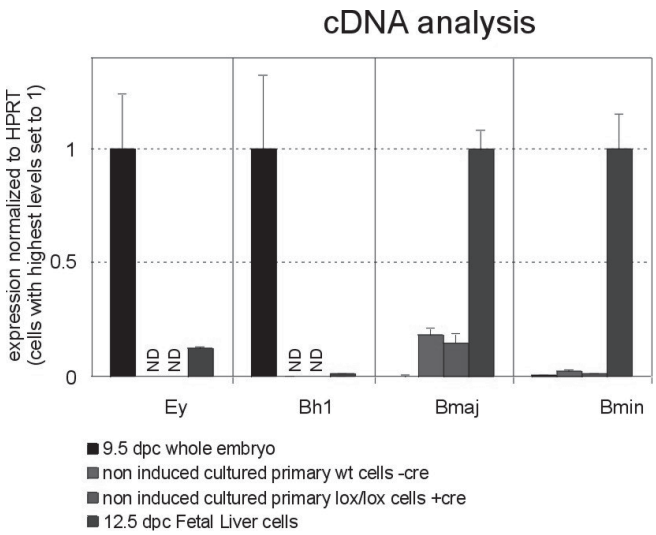
SUPPLEMENTAL FIGURES



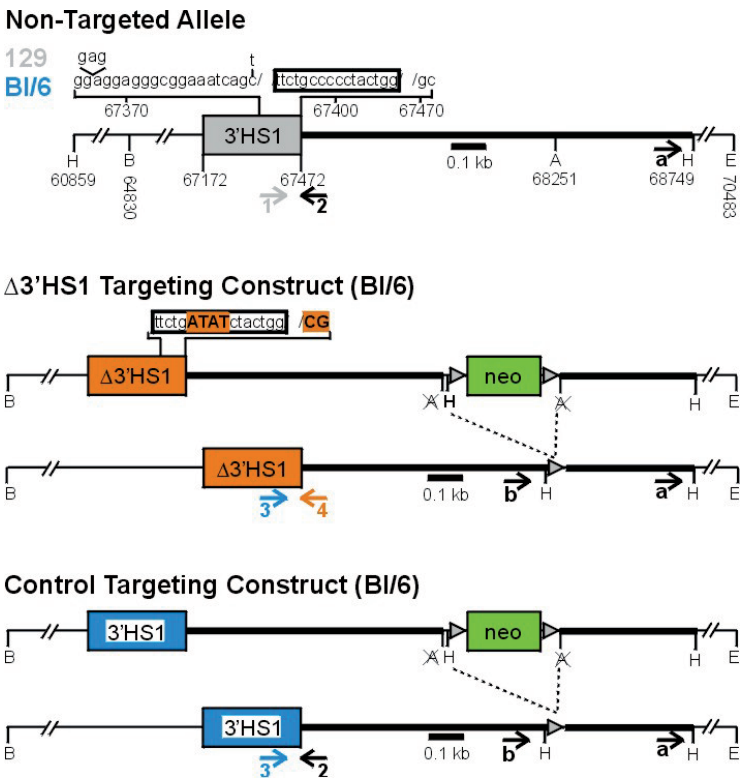
Supplemental Figure 1. ChIP with anti-CTCF antibody on E14.5 fetal liver (black) and brain (white bars). Data were normalized against input and expressed as enrichment over necdin (highest level set to 1).



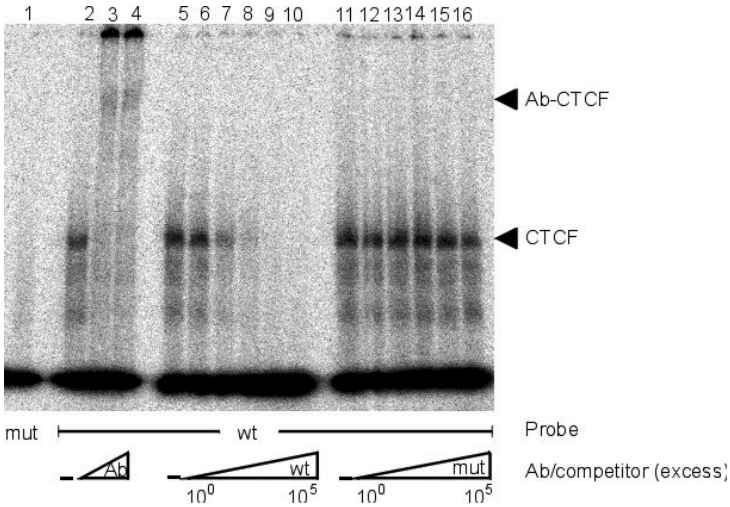
Supplemental Figure 2. (A) Q-PCR analysis of ligation frequencies obtained by 3C. The approach entails a primer-probe combination that is specific for a particular restriction fragment (blue), with the probe hybridizing to the opposite strand as compared to the PCR primer. A second PCR primer hybridizes to the fragment (orange) of which one wants to quantify its interaction. The primers/probe configuration guarantees that the probe only signals upon extension of the second primer across the ligated junction (# 1), which is important given the great variety of junctions (e.g. #2-7) formed with each fragment. f (fluorescent group) and q (quencher). **(B)** Examples of PCR products obtained after 45 cycles of QPCR, analyzed on a 2% agarose gel.



Supplemental Figure 3. Expression of all β-globin genes in cultured wild-type versus lox/lox progenitor cells. Messenger RNA levels were normalized to *HPRT* mRNA levels. For comparison, levels in primitive (E9.5) and definitive red blood cells (E12.5 fetal liver) are indicated and set to 1. ND: Not Detectable.



Supplemental Figure 4. Targeting strategy. Top: wild-type (non-targeted) allele. Nucleotide polymorphisms between 129 and B6 used to design allele-specific ChIP-primers #1 (129) and #3 (B6) are shown. Box indicates core CTCF-binding site. Coordinates are relative to the start of most upstream β-globin gene (εy). H: HindIII, B: BamHI, A: AvrII, E: EcoRV. Sequence between 67172 and 68749 is drawn to scale (see scale bar). 3C primer 'a' (is primer +68(3'HS1)) is used to analyze the 7.9 kb (60859-68749) wild-type HindIII fragment containing 3'HS1. Middle: targeting construct for Δ3'HS1. Targeted nucleotide changes in core CTCF-binding site (CCCC to ATAT) are indicated in orange. Downstream nucleotide changes (GC to CG) allow the design of ChIP-primer #4 that is specific for this targeted allele. The extra HindIII site upstream of the neomycin selection cassette flanked by two loxP sites is maintained after Cre-mediated deletion. This new HindIII site creates an allele-specific HindIII fragment around 3'HS1 that can be analyzed by primer 'b'. This fragment is 7.4 kb in size, not much smaller than the 7.9 kb fragment analyzed by primer 'a' on the non-targeted allele. On the targeted allele primer 'a' now analyzes a small (0.5kb) HindIII fragment downstream of 3'HS1. Bottom: targeting construct for the control ES-EP line (ES-EP(c)). 3'HS1 is untouched, but the extra HindIII site is introduced at the same position as before.

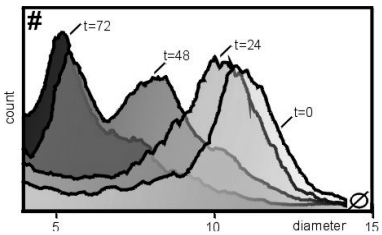


Supplemental Figure 5. Nucleotide changes in the core binding site of 3'HS1 effectively disrupt CTCF binding *in vitro*. Gel mobility shift assay with nuclear protein extracts from fetal livers (all lanes) (Wall et al. 1988) ^{32}P-labeled probe used: mutant (lane 1), wild-type (lane 2-16). Lane 3: +1 µl anti-CTCF antibody, lane 4: +2 µl anti-CTCF antibody, lane 6-10: competition with increasing amounts of unlabeled wild-type probe, lane 12-16: competition with increasing amounts of unlabeled mutant probe. Probe sequences: 3'HS1-wt-S: CGGAAATCAGTGGAACACTTCTGCCCCCTACTGGTATG-CAACAGG, 3'HS1-wt-AS: TCCTGTTGCATACCAGTAGGGGGCAGAAGTGTTCCACTGATTTCCG, 3'HS1-mut-S: CGGAAATCAGTGGAACACTTCTGATATCTACTGGTATGCAACAGG, 3'HS1-mut-AS: TCCTGTTGCATACCAGTAGATATCAGAAGTGTTCCACTGATTTCCG.

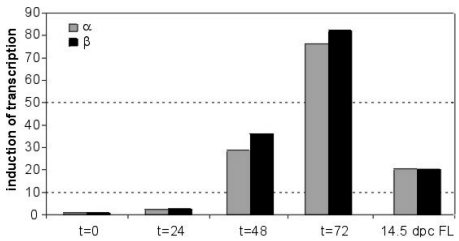
A.

Marker	ES-EP(D3'HS1)	ES-EP(c)
CD117 (%)	74	74
CD71/Ter119 (%)	98	100

B.



C.



Supplemental Figure 6. Characterization of ES-EP cells (A) ES-EPs express markers characteristic for proerythroblasts. **(B)** Synchronous *in vitro* differentiation of ES-EP cells. Cell size (μm) was measured at various intervals after induction of differentiation. The plots for the Δ and c lines are identical. **(C)** Globin mRNA expression during *in vitro* differentiation of ES-EP cells. Quantitative RT-PCR was performed to measure α- and β-globin mRNA levels relative to HPRT. Relative expression at t = 0 hours was set to 1. The fact that globin expression levels are higher in differentiated EP-EPs than in E14.5 fetal liver is attributed to the fetal liver being composed of cells at various stages of differentiation.

6

SPATIAL INTERACTION DOMAINS IN THE β-GLOBIN LOCUS FACILITATE CORRECT GENE EXPRESSION

Erik Splinter, Elzo de Wit, Harmen van de Werken, Dylan Mooijman and Wouter de Laat

Hubrecht Institute-KNAW & University Medical Centre Utrecht, Utrecht 3584 CT, The Netherlands

Work in progress

ABSTRACT

Spatial organization of chromatin in the nucleus of the cell plays an important role in regulating gene expression. CTCF is a DNA binding factor that influences long-range gene regulation presumably by its ability to form chromatin loops. In this study we applied a novel high resolution 4C strategy to investigate in detail the consequences of the chromatin loop that is formed by CTCF around the mouse β-globin locus. This loop encompasses the locus control region (LCR), the major regulatory sequence of the locus, as well as its two embryonic and two adult genes. We demonstrate that sequences within the loop preferentially contact each other while showing reduced contact frequencies with sites outside the loop. Vice versa, DNA sites outside the loop are hampered in their contacts with sequences inside the loop, but contact more easily sites across the loop. This situation is reversed, at least partially, when the locus carries a disrupted CTCF binding site that prohibits loop formation. As a consequence, the LCR now also more easily contacts the proximal embryonic genes at the expense of interactions with the distal adult genes. We show that this results in the aberrant upregulation of embryonic globin gene expression during the initial stages of adult erythroid differentiation. Thus, CTCF sets up a three-dimensional chromatin scaffold that impact on the flexibility of surrounding DNA sequences. Contacts between transcription regulatory sites and genes may benefit from or be hampered by this topology, with transcriptional activation and repression as the possible outcomes.

INTRODUCTION

DNA is highly compacted in the cell nucleus. The DNA helix is wrapped around an octet of histone proteins forming a 10 nanometer fiber. This fiber is further compacted by additional wrapping and looping forming higher order structures. A complex relationship exists between the packaging of the DNA and its functioning (Misteli 2007) and a growing body of evidence suggests a role for higher order chromatin organization in the regulation of gene expression (Fraser and Bickmore 2007). The relationship between DNA organization and gene expression is probably most extensively studied at the β-globin gene locus. The mouse β-globin locus consists of four developmentally regulated genes, namely the embryonic genes *εy* and *βh1* and the adult genes *βmaj* and *βmin* (Figure 1A). In the developing mouse embryo a transcriptional switch is made from embryonic to adult gene expression around twelve days post copulation. The high level of β-globin gene expression, as measured during erythroid differentiation, is controlled by a distal enhancer: the locus control region (LCR). The LCR consists of 5 or 6 hypersensitive sites (HSs), each with a specific function. For example; HS2 and HS3 contain binding sites for the erythroid transcription factors EKLF and Gata1 and were found crucial for gene activation. On the other hand, HS5 contains a CCCTC-Factor binding site (CTCFbs) and can act as an enhancer blocker in *in vitro* reporter assays (Bell et al. 1999; Farrell et al. 2002). Its role in the context of the endogenous β-globin locus is less clear though (Wai et al. 2003; Bender et al. 2006). Surrounding the locus several other CTCFbs are present which were found

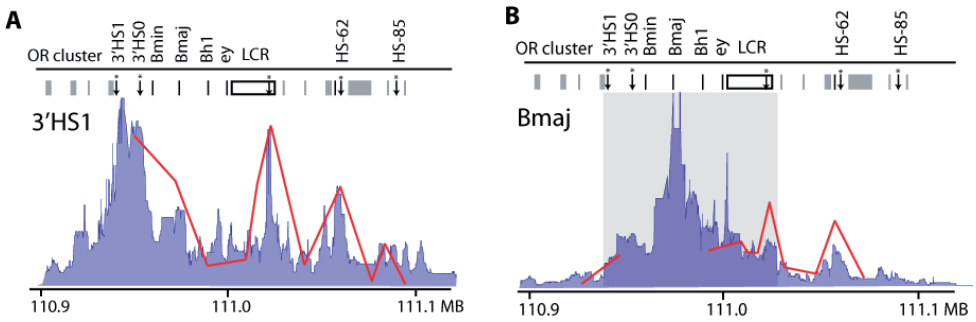


Figure 1: Highres-4C-seq results recapitulate spatial organization of the β-globin locus in high detail. A: The 3'HS1 interaction profile generated by highres-4C is depicted in blue. The red line indicates the previously generated 3C interactions profiles for comparison. Above these profiles the regulatory elements of the β-globin locus are plotted: grey bars, olfactory receptor genes surrounding the globin locus; black bars, globin genes; arrows, DNaseI hypersensitive sites; asterisks, CTCFbs. A scalebar indicating the genomic location of the interaction profile on mouse chromosome 7 is plotted below the graph. **B**: Similar to (A), except here the interaction profiles for β*maj* are shown. The preferred interaction domain of β*maj* is indicated by the gray area depicted behind the interaction profile.

to contact each other forming a chromatin hub in erythoid progenitor cells (Palstra et al. 2003). Upon differentiation the LCR upregulates globin gene expression by directly contacting one of the globin genes with the intervening sequences looping out (Carter et al. 2002; Tolhuis et al. 2002).

CTCF is an eleven zinc finger containing transcription factor that is highly conserved in higher eukaryotes. Since its discovery it has been assigned many different functions among which are activation/repression of transcription, insulation, X inactivation and imprinting (Phillips and Corces 2009). This diversity in function could be explained by the variety in binding sequences that may recruit different co-factors via the use of different combinations of its eleven zinc fingers. CTCF ChIP-seq experiments have indentified 20.000-40.000 CTCF bs of which many were shared between cell types and between human and mouse (Barski et al. 2007; Kim et al. 2007; Chen et al. 2008; Cuddapah et al. 2009; Kunarso et al. 2010). Interestingly, a subset of binding sites could be allocated to transitions of active (H2AK5) and inactive (H3K27me3) chromatin (Cuddapah et al. 2009) or found at the borders of lamina associated domains (LADs) (Guelen et al. 2008), suggesting a role for CTCF in segregating active and inactive chromatin. However, the mechanism behind these functions remains enigmatic.

Previously we were able to show a crucial role for CTCF in mediating loop formation between CTCFbs in the globin locus by using 3C technology (Chapter 5). Upon targeted mutation of the 3'HS1 CTCFbs we found a lack of participation of the 3'HS1 in the clustering of CTCFbs that surround the β-globin locus. However, an effect on the expression of *βmaj*, or the olfactory neuron (OR) genes surrounding the globin locus was not detected. Accordingly, when others depleted the CTCFbs at HS-62 and 3'HS1 they also did not find a detectable effect on *βmaj* gene expression (Bender et al. 2006), leaving no indication on the functionality of the CTCF mediated loops in regulating β-globin gene expression. Interestingly, CTCF ChIP-seq experiments performed by

others in embryonic stem cells (Chen et al. 2008) identified an additional CTCFbs in the β-globin locus, which we confirmed to also bind CTCF in erytroid cells (data not shown) and named 3'HS0 for convenience (Figure 1A). The presence of this additional CTCFbs site could, by means of redundancy, well be responsible for the absence of aberrant gene expression measured in the two studies.

Here we used a novel 4C-seq strategy (highres-4C-seq) that takes advantage of frequently cutting restriction enzymes ('four-cutters') to provide high resolution DNA topology maps (van de Werken et al. *in prep*). This strategy is particularly suited for studying local (<~2Mb) interactions between specific regulatory elements. We applied it to the β-globin locus with and without a flanking CTCFbs to investigate in detail the impact of chromatin loops formed by this protein.

RESULTS

Highres-4C-seq results recapitulate the spatial organization of the β-globin locus.

To investigate the spatial conformation of the β-globin locus in high detail we applied highres-4C-seq on cultured, unstimulated primary erythroblasts. These cells represent a progenitor stage of differentiation where contacts between CTCFbs surrounding the globin locus have been established, but LCR driven globin expression is absent (Dolznig et al. 2001; Palstra et al. 2003; Splinter et al. 2006). Upon stimulation of these cells with biological relevant factors the in vivo terminal differentiation process is recapitulated, including strong upregulation of globin gene expression and eventually enucleation (Dolznig et al. 2001). We previously found by 3C technology that interactions between the flanking CTCFbs are already strong in erythroid progenitors, while robust LCR-gene interactions are only formed upon differentiation of these cells. Here, essentially similar results were obtained when 4C was applied to 3'HS1 (Figure 1A) and *βmaj* (Figure 1B). From 3'HS1, both 3C and 4C profiles showed clear interactions between CTCFbs, although the specific participation of the HS-85, as appreciated from the 4C profile seems limited when compared to its surrounding interactions. Looking from *βmaj*, the moderate 3C interactions previously found were reproducibly identified by 4C. Thus, 4C provides essentially similar but much more detailed DNA interaction profiles for selected sites.

4C interaction profiles reveal domains of high frequency collisions within the β-globin locus

Close inspection of the detailed 4C profiles reveals topological features not appreciable by 3C technology. *βmaj* seems to make most frequent contacts with sequences located within the region flanked by the CTCFbs 3'HS1 and HS5 (Figure 1B; grey shade). This was true for *βmaj*, but also for *βmin* (Figure 2A), HS2 of the LCR (Figure 2B) and *βh1* (data not shown); they all showed sharp transitions in contact frequency beyond a domain demarcated by 3'HS1 and HS5. Vice versa, sequences outside are limited in their ability to contact sequences inside this domain (Figure 2C, D). CTCFbs themselves however, as shown for both 3'HS1 and HS-62, are found to preferentially contact each other, with contacts between CTCFbs not necessarily being limited to those directly flanking each other on the linear chromosome

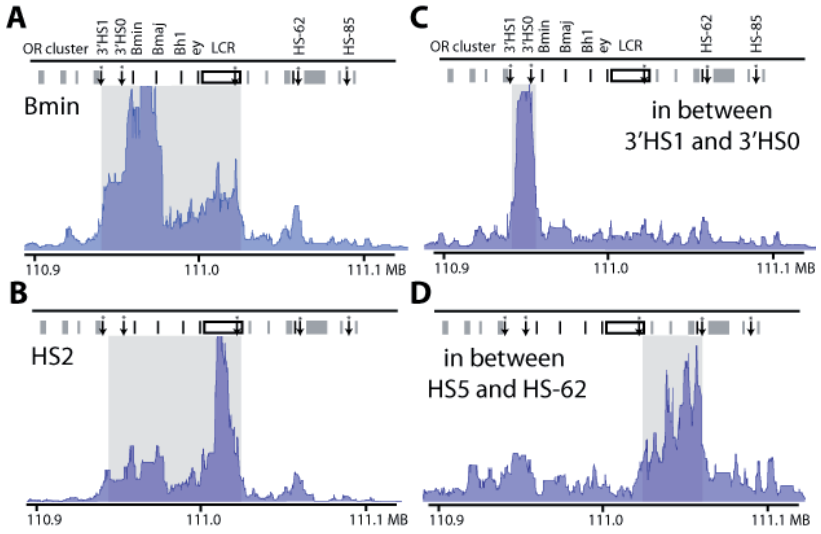


Figure 2: 4C interaction profiles reveal a domain organization within the β-globin locus demarcated by CTCF. Highres-4C interaction profiles for β*min* (**A**), HS2 of the LCR (**B**), a viewpoint in between 3'HS1 and 3'HS0 (**C**) and a viewpoint in between HS5 and HS-62 (**D**) are depicted. The preferred interaction domain of each viewpoint is indicated by the gray area behind the interaction profile.

template (Figure 1A and 3A, B). From these data a picture emerges of CTCF forming chromatin loops that demarcate domains in which DNA sequences preferentially interact with each other. At the *β*-globin locus, one such spatial interaction domain extends from the CTCFbs at HS5 to that of 3'HS1, encompassing the genes and the LCR that controls their expression.

Mutation of the 3'HS1-CTCFbs disrupts loop formation within the β-globin locus while not affecting long-range cis interactions

To further investigate the function of the chromatin loop formed by CTCF around the *β*-globin locus, we took advantage of a previously generated transgenic mouse line carrying the targeted mutation of the 3'HS1 CTCFbs. The mutant 3'HS1 fails to bind CTCF and no longer participates in loop formation in the β-globin locus (Chapter 5). As was found previously by 3C, the 4C strategy clearly revealed a reduction in contact frequency between the 3'HS1 and the other CTCFbs in the β-globin locus. This is appreciated both from the interaction profile of 3'HS1 as from that of HS-62 (Figure 3A, B; arrows). Next we determined the role of the mutated 3'HS1 CTCFbs in organizing distal (>2Mb) interactions in *cis* by applying regular (6x4) 4C-seq (using *Hin*dIII as the first RE). Comparing the interaction profile of the 3'HS1 itself and the HS-85, representing other CTCFbs in the β-globin locus, we found the long-range interaction profiles to be highly similar (Figure 3C) indicative of these sites to reside in a similar genomic environment. The high similarity between the genomic environments of these sites is further supported by the high correlation found between the two experiments when calculating the spearman's ρ (Figure 3D, blue curve). Collectively the data show that CTCF

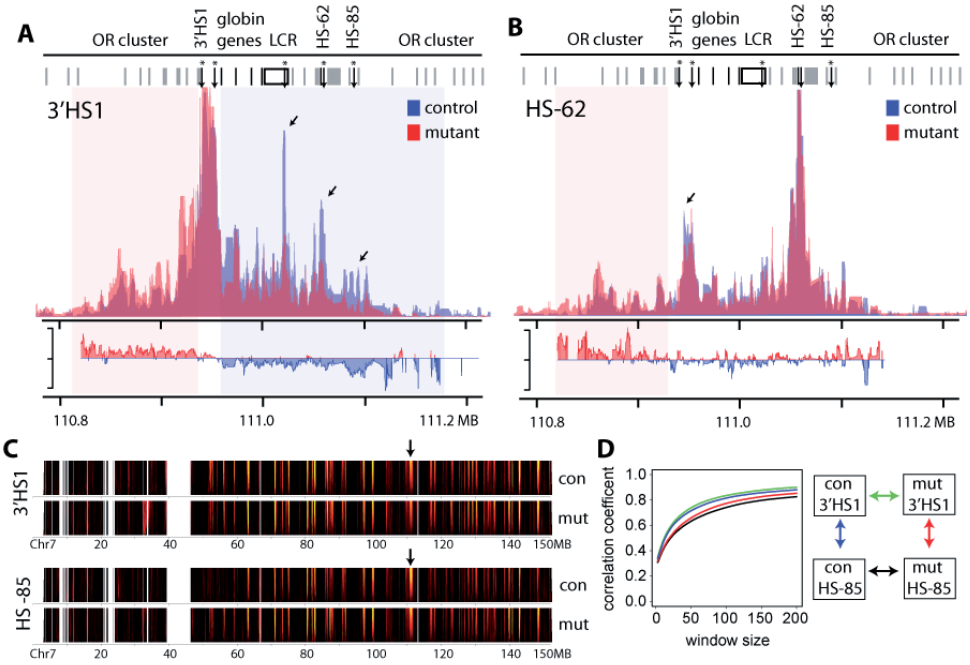


Figure 3: Mutation of the 3'HS1 CTCFbs disrupts loop formation within the β-globin locus while not affecting long-range interactions in *cis*. Highres-4C interaction profiles for 3'HS1 (**A**) and HS-62 (**B**) are plotted for both the control (blue) and the cell carrying the mutant 3'HS1 CTCFbs (red). The log-ratio is depicted below each graph to appreciate differences between the control and mutant 3'HS1 datasets. The scale bar represents values of 4 to -4, ranging from mutant to control preferred interactions respectively. The shift in preferred interactions is highlighted by the red and blue shaded areas behind the interaction profiles. **C**: Long-range *cis* interactions for 3'HS1 (top two panels) and HS-85 (bottom two panels) are depicted in domainograms (explained in detail in chapter 4). The arrow points toward the location of the viewpoint. For each viewpoint the interactions found in both the control (con) and mutant cells (mut) are shown. Significance of interaction is depicted by the range in color from black to yellow, indicating a p-value ranging from 10^{0} to 10^{-10} respectively. **D**: The similarity between the long-range interactions found in the different conditions is calculated via the Spearman's ρ. The scheme on the right depicts the colors representing the different comparisons that were made.

binding to 3'HS1 is necessary for the formation of a chromatin loop around the β-globin locus, but not for positioning this site or the rest of locus with respect to sequences elsewhere in the genome.

Distortion of the 3'HS1 CTCFbs disrupts domain organization within the β-globin locus

We reasoned that the formation, and disruption, of a chromatin loop between distant DNA sequences will impact on the flexibility of the chromatin fiber within and around the loop. To investigate this, we first focused on the DNA interaction profiles of the 3'HS1 and HS-62 (Figure 3A, B). Both viewpoints displayed an increased ability to interact with sequences across 3'HS1 when this site was mutated. For 3'HS1, this clearly occurs at the expense of contacts with sequences inside the globin locus (Figure 3A, B and Figure 4). We then selected viewpoints within and outside the chromatin loop formed by CTCF. The shift in spatial contacts was also observed for

βmaj, *βmin*, for sequences 7 and 12kb downstream 3'HS1, for HS2 of the LCR, which is located ~75kb away from the 3'HS1 and for two viewpoints located far upstream and downstream of the globin locus (Figure 4A-D and data not shown). Sequences on either side of 3'HS1 all showed increased contact probabilities with sequences across 3'HS1 when CTCF binding and loop formation was disrupted. It shows that the CTCF mediated loop limits local chromatin flexibility: it causes sequences within the loop to preferentially contact each other rather than sequences outside the loop, and vice versa. Taken together these data argue for a role of CTCF at 3'HS1 in directing spatial interactions of sequences located within and surrounding the β-globin locus.

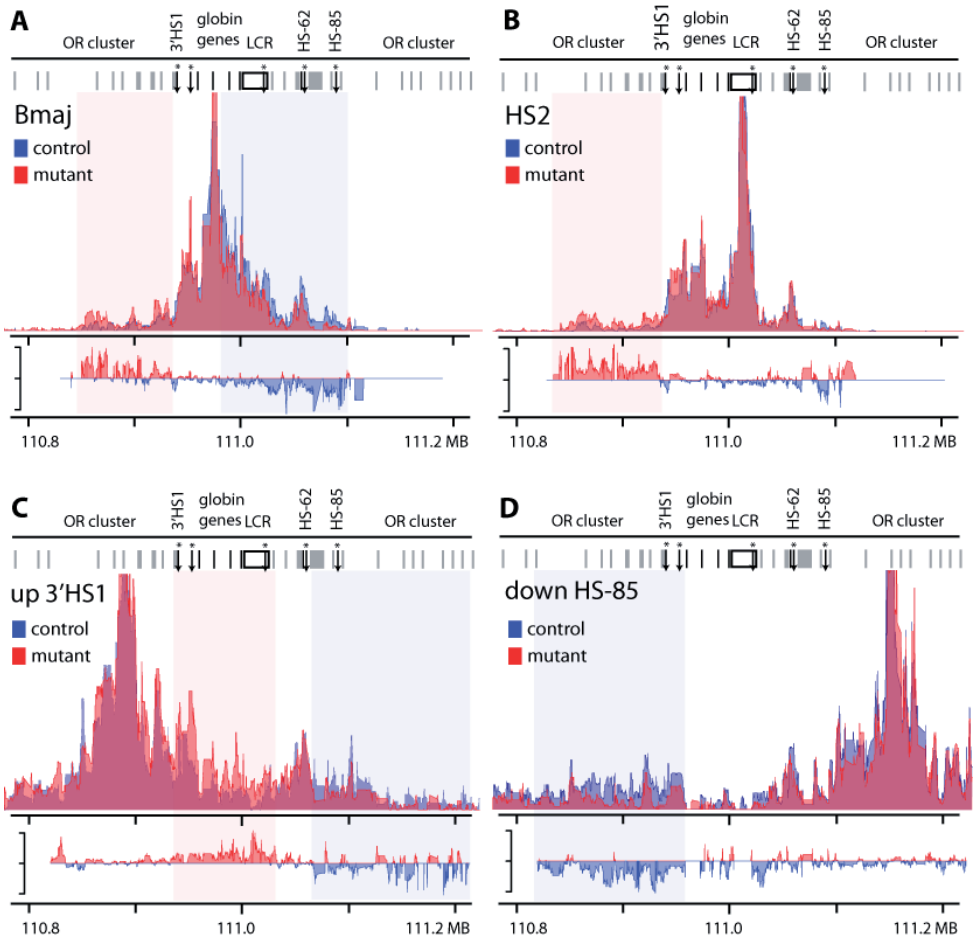


Figure 4: Targeted mutation of the 3'HS1 CTCFbs disrupts domain organization. Highres-4C interaction profiles for β*maj* (**A**), HS2 of the LCR (**B**), a viewpoint upstream 3'HS1 (**C**) and a viewpoint downstream HS-85 (**D**) are plotted for both the control (blue) and the cell carrying the mutant 3'HS1 CTCFbs (red). The log-ratio is depicted below each graph to appreciate differences between the control and mutant 3'HS1 datasets. The scale bar represents values of 4 to -4, ranging from mutant to control preferred interactions respectively. The shift in preferred interactions is highlighted by the red and blue shaded areas behind the interaction profiles.

CTCF mediated loop formation affects β-globin gene competition during early stages of differentiation.

The limitation in spatial flexibility imposed by the CTCF mediated loops has consequences for the contact frequencies measured from viewpoints within the globin locus. Although this change in preferred interaction partners is relatively small, as judged from the 4C interaction profile, it might be relevant for correct gene regulation. Previously, we asked whether the CTCF mediated loop that is present in erythroid progenitors helps the subsequent formation of LCR-gene interactions later during erythroid differentiation. We therefore tested whether messenger RNA levels of the adult *βmaj* gene accumulated earlier in timed differentiation series in wild-type versus mutant cells lacking the CTCF loop. No difference was found though Based on the 4C experiments which showed that without CTCF bound to 3'HS1, in adult progenitor cells the LCR tends to interact more with the proximal embryonic and less with the distal adult β-globin genes, we asked whether this impacts on their expression regulation. For this, we designed cDNA primers specific for embryonic *βh1* and the adult *βmaj* and *βmin* globin genes and compared the relative expression of these genes in mutant and control primary erythroblast during a timed differentiation assay. Upon differentiation, in both cell populations each globin gene was found to be slightly upregulated (2-5 fold) during the early time points of 16 and 24 hours compared to their expression before induction (Figure 5B). Full induction of the adult globin genes was seen after 40 hours of differentiation. Remarkably, during the early time points a significant increase of *βh1* expression was measured in the mutant versus wild-type cells (Figure 5B). Thus the mutation of four nucleotides that disrupt CTCF

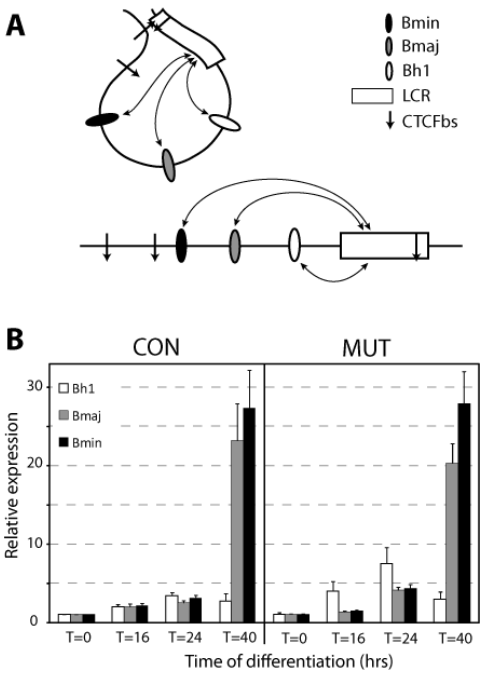


Figure 5: CTCF mediated loop formation affects β-globin gene competition during early stages of differentiation. A: A simplified model of the β-globin locus shows a decrease of spatial distance between the LCR and adult globin genes when a loop is formed between the CTCFbs HS5 of the LCR and the 3'HS1. **B**: The relative expression levels of β*h1*, β*maj* and β*min* during a timed differentiation assay. Expression was measured in 3'HS1control -and mutant erythroid cells and were normalized to *Rnh1* expression.

binding at a site ~75 kb away from the LCR causes upregulation of a gene that is ~55 kb more proximal to the LCR. We conclude that disentangling the pre-established CTCF loop encompassing the LCR and all globin genes provide a proximal gene with an advantage to contact the LCR and compete for its activity. This competitive advantage disappeared at late stages of differentiation. Then, the adult genes reach full expression levels and apparently no longer allow the embryonic genes to compete for the LCR.

DISCUSSION

The application of highres-4C-seq has proven informative in uncovering the 3D topology of the β-globin locus. Not only did this technology provide high detailed interaction maps, it provided the ability to appreciate specific interactions in the context of their surrounding sequences. As such, a novel function of the CTCF mediated chromatin loops in the β-globin locus was detected. We find that spatial collisions of DNA sequences preferentially occur between sequences located within a CTCF mediated chromatin loop while this is strongly reduced between different loops. Previously, using 5C technology, others have shown a globular organization of the α-globin locus (Bau et al. 2010) and the HoxA cluster (Wang et al. 2011). Although these studies did not provide data concerning a mechanism dictating 3D organization, this does suggest that the spatial separation of genomic sequences by the formation of chromatin loops/domains is a feature commonly found within the genome (Sanyal et al. 2011).

The need for an enhancer to contact its target gene in nuclear space to execute its function has been well established (Carter et al. 2002; Tolhuis et al. 2002; Palstra et al. 2003; Liu and Garrard 2005; Vernimmen et al. 2007; Amano et al. 2009). Therefore it is likely that any modulation that affects the ability of sequences to contact each other, either by increasing or decreasing the interaction frequency, could potentially affect the levels of gene transcription. The role of CTCF in dictating spatial chromatin flexibility is quite remarkable. Classically, CTCF is known to act as a boundary element, blocking the spreading of heterochromatin into euchromatic regions (Bell and Felsenfeld 1999) and for its ability to block an enhancer from activating its target gene. The latter has been investigated extensively in reporter constructs (Maksimenko et al. 2008) and to limited extend in transgenic studies involving naïve gene loci (Hou et al. 2008). Despite these and other studies investigating CTCF functioning, a precise mechanism of how CTCF is able to execute these functions remained indecisive. Our observations could provide an explanation, as we link the 'classic' functions of CTCF to its ability to induce chromatin loops, showing that this limits the spatial flexibility of surrounding sequences. CTCF-mediated loops in the β-globin locus seem to function in several ways. First, they restrict the flexibility of globin sequences to engage in contacts beyond the 3'HS1 and vice versa. This restriction possibly shields genes located outside the globin locus to be influenced by the LCR. Vice versa, it may prevent regulatory elements located beyond the 3'HS1 to affect globin gene expression. Second, the CTCF loop decreases the spatial proximity between the LCR and especially the adult globin genes *βmaj* and *βmin*, allowing them to more efficiently compete with linearly closer embryonic genes for LCR activity.

This is reflected by the preferential upregulation of *βh1* by the LCR in the early stages of differentiation of cells lacking the CTCF-loop. Next to *βh1*, expression of *εy,* not tested so far, is also interesting to investigate as this gene is closest to the LCR on the linear template. An additional experiment that needs to be performed is to check at early stages of differentiation whether indeed βh1-LCR contacts are more pronounced in mutant compared to wild-type cells. This could provide further evidence for our model that *βh1* upregulation is caused by its increased ability to contact the LCR in the absence of a CTCF-loop that drags along the competing adult globin genes towards the LCR.

Studying the effects of the mutated 3'HS1 on the domain organization of the globin locus relatively small changes were detected. Possibly, this is due to redundant actions of 3'HS0. But even so, effects of the 3'HS1 mutation are measurable, suggesting that the concerted action of both the 3'HS1 and 3'HS0 CTCFbs is required for organizing the spatial interaction domains within the locus. We predict that both the observed effects on spatial organization and gene expression become more pronounced when the 3'HS1 and 3'HS0 are mutated simultaneously.

More generally, it will be interesting to test if all CTCFbs across the genome are able to induce loop formation, or that this ability is limited to only a subset of binding sites. This is to be expected as CTCF binding affinity will differ from site to site and a subset of sites was found to also attract cohesin (Wendt et al. 2008) which is known to mediate chromatin interactions (Hadjur et al. 2009; Nativio et al. 2009; Chien et al. 2011). It will be interesting to determine the role of Cohesin in this spatial organization by determining its binding sites in erythroid progenitors and by studying the spatial organization of the locus in absence of Cohesin.

Taken together we have shown, by using highres-4C-seq, that CTCF spatially organizes the β-globin locus by mediating chromatin loops. High contact frequencies were measured within each loop, while collisions between elements located on different loops were found to be limited. This finding points towards a crucial role of CTCF in dictating the spatial flexibility of regulatory elements within the β-globin locus. By doing so, CTCF causes the aberrant upregulation of *βh1,* confirming the involvement of chromatin flexibility in regulating gene expression. We propose that the many functions assigned to CTCF can often be explained by CTCF's ability to dictate spatial interactions between regulatory elements via the formation of chromatin loops. The functional consequences of these loops will depend on the order of the regulatory elements on the linear DNA template and the balance of affinity and stability of the different interactions between these elements.

METHODS

Cell culture

Fetal livers from homozygous control -and mutant 3'HS1 (Splinter et al. 2006) 12.5dpc mouse embryos were isolated and cultured as previously described (Drissen et al. 2004; Splinter et al. 2006) to select for erythroblasts.

Highres-4C-seq

4C experiments were basically performed as described in Chapter 4, except when highres-4C-seq was applied. For this application we selected *Nla*III as the first restriction enzyme and *Dpn*II as the second. No

linearization was applied. The primers used in the highres-4C PCR are: upstream 3'HS1, F: #4388-GCAAGTATGGTGCACACATG, R: #4389-CTTGTTTATTTCACTACGTC; 3'HS1, F: #2717-ATGCCTCTTACCTC-CATG, R: #2716-GGATAAACTGTAT-CACCACC; in between 3'HS1 and 3'HS0, F: #3695-T GCATCTTGTTTGCATG, R: #3705-CTTTTCCAAAACTCAAAACAC; *βmin*, F: #3689-TCAGAAACA-GACATCATG, R: #3707-CATAA-GAAAATTTGCCTGTTC; *βmaj*, F: #3581-TTTCTCAGTTTGAGTGCATG, R: #3576-GAAGCCTGATTCCGTAGAG; *βh1*, F: #4390-TTGAAGCATAGGCAAACATG, R: #4391-TTGACCAATAGCCTCAGAGT; HS2, F:#3684-CTGTACGTTGTATTACATG, R: #3573-ATCAATGGTCAGCGTTTTAG; in between HS5 and HS-62, F: #4392-GGATGGTCAACTACCACATG, R: #4393-AGGGTTCTGAGTCTTTGGTT; HS-62, F: #3681-ATTCCCCACAGT-GTTCATG, R: #3577-TCACCGGA-GAAGTTTTCTAA; downstream HS-85, F: #4396-TGTGTCACAGGGAAGT-CATG, R: #4397-TTTCTGTGATATTGT-TCAGGTG. The primers used in the regular 4C PCR are: 3'HS1, F: #2746-AACCTAGTC-GATAAGCTT, R: #2403-GGATAAACT-GTATCACCACC; HS-85, F: #2406-ACCAACCCAGAAGAAAGCTT, R: #2407-ACTTGCCAGATGTTTACAGC. The 4C data presented are the raw mapped reads, allowing one mismatch, which are smoothened using a running median algorithm using a window of 29 fragment ends. Domainogram and Spearman correlation analysis were performed as described elsewhere (de Wit et al. 2008; chapter 7)

Expression analysis

RNA was extracted using Trizol® following manufacturers protocol (Invitrogen), followed by DNase treatment (Promega). cDNA was prepared by standard reverse transcriptase conversion using oligodT primers (Promega). Expression levels were determined using Quantitative PCR via sybergreen incorporation (MIQ, BIO-RAD). The PCR program consisted of 10' 94°C, 10' 94°C, 30' 62°C, 30' 72°C, 40 cycles from step 2, 10' 72°C followed by the generation of a melt curve. Primers used for this analysis: *βh1*, F: #970-TGGACAACCTCAAGGAGAC, R: #971-AGTAGAAAGGACAATCAC-CAAC; *βmaj*, F: #3251-CTCACATTTGCT-TCTGACATAG, R: #4310-TCACTT-TCTTGCCATGGGCC; *βmin*, F: #3247-TACGTTTGCTTCTGATTCTG, R: #4309-GAGGCTGTCCAAGTGATCAG; *Rnh1*, F: #1102-TCCAGTGTGAG-CAGCTGAG, R: #1103-TGCAGGCACT-GAAGCACCA, which were tested for specificity. Expression analysis was performed on cells from two individual embryos coming from both a control and mutant 3'HS1 litter. Globin expression values were normalized to *Rnh1* expression. Both embryo combinations (two times control vs mutant erythroid progenitor) showed *βh1* upregulation. For reasons of simplicity the data presented is the average of three independent PCRs applied on each embryo combinations.

7

THE INACTIVE X CHROMOSOME ADOPTS A UNIQUE THREE-DIMENSIONAL CONFORMATION THAT IS DEPENDENT ON XIST RNA

Erik Splinter[1], Elzo de Wit[1], Elphège P. Nora[2], Petra Klous[1], Harmen J.G. van de Werken[1], Yun Zhu[1], Lucas J.T. Kaaij[1], Wilfred van IJcken[3], Joost Gribnau[4], Edith Heard[2] and Wouter de Laat[1]

[1]Hubrecht Institute-KNAW & University Medical Center Utrecht, Utrecht 3584 CT, The Netherlands
[2]Mammalian Developmental Epigenetics Group, Institut Curie CNRS UMR3215 INSERM U934, Paris F-75248, France
[3]Erasmus Center of Biomics and [4]Department of Reproduction and Development, Erasmus Medical Center, Rotterdam 3015 GE, The Netherlands

Adapted from *Genes & Development* 2011 Jul 1;25(13):1371-83.

ABSTRACT

3D-topology of DNA in the cell nucleus provides a level of transcription regulation beyond the sequence of the linear DNA. To study the relationship between transcriptional activity and spatial environment of a gene, we have used allele-specific 4C-technology to produce high-resolution topology maps of the active and inactive X-chromosomes in female cells. We found that loci on the active X form multiple long-range interactions, with spatial segregation of active and inactive chromatin. On the inactive X, silenced loci lack preferred interactions, suggesting a unique random organization inside the inactive territory. However, escapees, among which is *Xist*, are engaged in long-range contacts with each other, enabling identification of novel escapees. Deletion of *Xist* results in partial re-folding of the inactive X into a conformation resembling the active X, without affecting gene silencing or DNA methylation. Our data point to a role for Xist RNA in shaping the conformation of the inactive X-chromosome at least partially independent of transcription.

INTRODUCTION

The spatial organization of DNA in the cell nucleus is non-random and provides opportunities to facilitate DNA metabolic processes like transcription and replication. Each chromosome in a mammalian nucleus occupies a distinct spatial territory in the nuclear space, with the larger and gene-poor chromosomes adopting a more peripheral location, and smaller, gene-rich chromosomes a more internal position inside the nucleus (Croft et al. 1999; Bolzer et al. 2005). Individual chromosomal segments also have preferred genomic neighbors in nuclear space, resulting in the spatial segregation of active from inactive chromatin and setting up the three-dimensional (3D) structure of chromosomes (Simonis et al. 2006; Lieberman-Aiden et al. 2009). It has been argued that the nuclear positioning of genomic regions follows probabilistic rules and will therefore differ from cell to cell. It is dependent not only on the properties of the region itself, but also on the characteristics of proximal regions on the linear chromosome template (Misteli 2001; de Laat and Grosveld 2007). Although factors have been identified that are involved in organizing and maintaining loops between enhancers and genes (Drissen et al. 2004; Kagey et al. 2010), little is known about factors involved in higher order chromosome folding. The fact that active genes come together in nuclear space strongly suggests that transcription shapes the 3D organization of the genome (Osborne et al. 2004; Schoenfelder et al. 2009; Papantonis et al. 2010). However, transcription inhibition studies often fail to produce significant conformational changes (Tumbar et al. 1999; Palstra et al. 2008; Muller et al. 2010). Aside from transcriptional activity, spatially segregated active and inactive chromatin domains differ in epigenetic marks (such as DNA methylation) as well as in bound trans-acting proteins. Because all these factors may impact on chromosome topology and vice versa, separating cause and effect remains challenging.

Mono-allelically expressed gene loci provide a useful model system to study the relationship between genome topology,

transcriptional activity and chromatin modifications. They are identical in DNA sequence, co-exist in the same cell and experience the same trans-acting environment, yet their chromatin composition, bound transcription factors and expression status is completely different. Provided one can distinguish the two alleles, mono-allelically expressed genes offer a unique opportunity to assess the impact of transcriptional activity and chromatin composition on nuclear and chromosomal organization. One of the most extreme and particularly intriguing examples of mono-allelic gene expression is found on the mammalian female X chromosomes. In mammalian cells, one X chromosome is inactivated to achieve dosage compensation between male and female cells. Random X chromosome inactivation (XCI) takes place during early embryonic development and is initiated by upregulation of the *Xist* gene on the future inactive X chromosome (X_i) (Senner and Brockdorff 2009; Barakat and Gribnau 2010). *Xist* encodes an untranslated RNA and its accumulation on the X chromosome *in cis* creates a silent nuclear compartment that excludes RNA polymerase II and associated transcription factors (Chaumeil et al. 2006). Upon accumulation of Xist RNA, various proteins involved in silencing are recruited to the X chromosome, including Polycomb group (PcG) protein complexes PRC2 and PRC1. Along with this recruitment a change in chromatin features is observed, with depletion of histone modifications linked with gene activity such as H3K9ac and an increase of heterochromatin marks such as H3K27me3 and H4K20me1. Subsequently the X_i becomes late replicating, incorporates histone variants such as macro-H2A, and promoter sequences undergo CpG methylation, the final outcome of which is that most genes on the X_i are stably silenced. Interestingly, some genes can escape from inactivation (Disteche 1995; Yang et al. 2010). Conditional deletion of *Xist* shows that, once established, Xist RNA no longer seems to be required to maintain XCI (Csankovszki et al. 1999; Wutz and Jaenisch 2000). Intriguingly, after establishment of XCI, loss of Xist does compromise PRC2 recruitment and macro-H2A incorporation, but silenced genes remain inactive (Csankovszki et al. 1999; Kohlmaier et al. 2004; Zhang et al. 2007; Pullirsch et al. 2010) presumably because epigenetic marks such as CpG methylation remain. A possibility not explored yet is that the 3D organization of the inactive X chromosome is important for maintenance of gene silencing. Indeed, although Xist RNA appears to result in spatial reorganization of the inactive X chromosome as seen under the microscope, it is not clear whether this is relevant for the initiation or maintenance of the inactive state.

The topology of the X chromosome and its position in the nucleus may well play a role in its expression status in various dosage compensation strategies. For example, in *Drosophila* dosage compensation is achieved by upregulating the male X chromosome via the dosage compensation complex (DCC). This upregulation is accompanied by male-specific folding of the X chromosome, which is dictated by the clustering of high affinity binding sites (HAS) for the DCC (Grimaud and Becker 2009). Such a folding pattern was suggested to promote an efficient distribution of the DCC on the male X chromosome. Similarly, chromosome topology might be important in contributing to the spread of gene silencing via Xist RNA along the entire X chromosome in *cis* in mammalian cells. In the mouse, *Xist* is located on the acrocentric X chromosome approximately 100 Mb away from the centromere and ~65 Mb away

from its telomere. Nothing is known about how Xist RNA spreads or coats the X chromosome. It may proceed linearly along the chromatin fiber, and/or diffuse to X-linked regions that are in spatial proximity to the site of Xist RNA production. The latter can be expected to greatly enhance the efficiency of this process at both the initiation and maintenance phase of XCI.

Most of our current knowledge of mammalian X chromosome folding is based on fluorescence *in situ* hybridization (FISH) studies interrogating the position of genes in relation to the chromosomal territory (Dietzel et al. 1999; Chaumeil et al. 2006; Clemson et al. 2006). It was found that the core of the Xist territory is composed mainly of repetitive sequences, while genic sequences reside more on the edge of the territory. During XCI, genes that are inactivated are relocated from the edge to occupy a more internal position, while escapees remain looped out or at the outer edge of the Xist domain, in contact with the transcription machinery (Chaumeil et al. 2006). Xist was found to be crucial for this relocalization as Xist RNA lacking the critical repeat A region did accumulate on the X in *cis* but was unable to induce gene silencing (Wutz et al. 2002) or gene relocalization (Chaumeil et al. 2006).

Although microscopy-based studies have been informative to study 3D genome organization, more recently developed techniques based on Chromosome Conformation Capture (3C) (Dekker et al. 2002) have enabled a much more detailed and comprehensive view of chromatin folding and chromosome organization in the nucleus. While 3C analyzes interactions between single, selected DNA fragments (one-versus-one), adapted versions of 3C allow for increased throughput analyses, with 4C analyzing interactions between one-versus-all (Simonis et al. 2006), 5C analyzing many-versus-many interactions (Dostie et al. 2006) and HiC analyzing all-versus-all (Lieberman-Aiden et al. 2009; Rodley et al. 2009; Duan et al. 2010). All C-methods are based on fixing the 3D genome inside living cells, digesting the DNA and ligating the cross-linked fragments to each other. By quantifying ligation products a measure of co-localization frequency can be obtained, either by PCR, micro-arrays or next generation sequencing (NGS). 4C and HiC are the two methods for generating genome-wide interaction profiles. For a given throughput, HiC provides a low resolution 3D map of all genomic interactions, while 4C gives a highly detailed interaction map for a single locus.

Here, we designed a strategy to allele-specifically direct 4C technology to active X chromosome (X_a) -and X_i-associated gene loci in differentiated female mouse neural precursor cells (NPCs), to gain detailed insight into the folding of the X chromosome in its euchromatic or heterochromatic state. We provide the first high resolution interaction maps of the active and inactive X chromosomes and demonstrate that the X_a and X_i fold very differently, with inactive loci on the X_i being unique in having lost their preference to co-localize with a defined subset of other chromosomal loci. This random organization of inactive genes within the inactive X chromosome territory is in sharp contrast to the more defined positions of escapees, which we find preferentially co-localizing and at the periphery of the X_i domain. In fact, the 4C interaction profiles of escapees allowed novel escape genes to be identified in the cell type studied. By deleting the *Xist* locus in the same cells we demonstrate that *Xist* depletion results in partial refolding of the X_i to a structure that looks more reminiscent of the X_a's conformation. This change

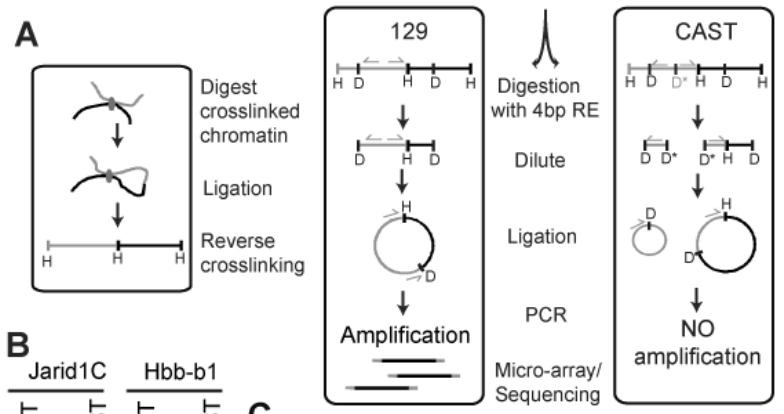
in chromosomal organization is not accompanied by overt changes in transcriptional activity or DNA methylation on the X_i. Our data demonstrate how a long non-coding RNA impacts on chromosome folding in a manner that is at least partially independent of transcription and DNA methylation.

RESULTS

Allele-specific 4C analysis discriminates between the active and inactive X chromosome in NPCs

We applied 4C technology to enable a detailed study of the conformational differences between the active and inactive X chromosome in female cells. 4C technology allows for screening the genome in an unbiased manner for DNA regions that interact in the nuclear space with a locus of choice. Based on validation of interaction frequencies of >100 pairs of loci, both within and between chromosomes, by high-resolution cryo-FISH (Simonis et al 2006; Palstra et al 2008; data not shown) we have previously demonstrated that this strategy robustly identifies contacting chromosomal regions. To direct the analysis specifically to either the active or inactive X chromosome, we used clonal F1 cell lines that had inactivated either the *Mus musculus* (SVJ129) or the *Mus musculus castaneus* (CAST) X chromosome (see below). Single nucleotide polymorphisms (SNPs) creating allele-specific restriction sites were used for the exclusive degradation of 4C template of one of the alleles, thereby enabling amplification and analysis only of the other allele (Fig. 1A, B and Supplemental Fig. S1A-C, further details in Methods). Dedicated microarrays (Simonis et al. 2006) as well as next generation sequencing (NGS) were used in combination with 4C for the analysis of DNA interactions (see Methods). The two strategies yielded highly similar results (Supplemental Fig. S1D-E), but 4C in combination with NGS, or 4C-seq, offers several advantages, the most important being the higher putative coverage and theoretically an unlimited dynamic range, providing increased sensitivity and better resolution. Thus, 4C-seq provides highly detailed interaction maps for selected genomic loci.

In vitro differentiation of female ESCs is accompanied by the random inactivation of one the X chromosomes, in a process that recapitulates the molecular events of *in vivo* X inactivation [reviewed by Barakat and Gribnau 2010). XCI is completed at the stage when ~7 day old embryoid bodies (EBs) are formed. As EBs contain mixtures of cells from all germ layers and 4C (and other 3C-based methods) provide an approximation of the average chromatin structure present in all cells analyzed, we continued differentiation until neural precursor cells (NPCs) were obtained (Supplemental Fig. S2A). NPCs show clear Xist RNA domains (Supplemental Fig. S2B) that are highly enriched in H3K27me3 (Supplemental Fig. S2C). Moreover, DNA methylation of promoters was observed on ~50% of the X-linked alleles analyzed (Supplemental Fig. S2D). Together, this is indicative of complete XCI. XCI is random, but once established, it is stably propagated to daughter cells. We confirmed this by selecting NPC clones from single cells, which exclusively inactivated either their X^{129} or the X^{CAST} chromosome (Supplemental Fig. S2E). Using these NPC clones we were thus able to direct 4C analysis specifically to the active or inactive X chromosome.



Viewpoint	Expression		Gene density
	Xa	Xi	
Slitrk4 (61.5MB)	Inactive	Inactive	low
MeCP2 (71.3MB)	Active	Inactive	high
Pcdh11x (117.7MB)	Active	Inactive	low
Jarid1C (148.7MB)	Active	Active	high

Figure 1: Outline of the allele-specific 4C approach to interrogate X chromosome folding. **A**: Schematic outline of the allele-specific 4C approach. RFLPs were identified and used to direct the 4C analysis to either the 129SVJ or the CAST allele. **B**: 4C PCR products using allele-specific primers (*Jarid1C,* 129SVJ specific) and primers not designed around a RFLP (*Hbb-b1*) separated on an agarose gel. **C**: Four genes with different characteristics, representing different genomic environments were chosen as viewpoints for the 4C analysis. An alternative allele-specific 4C approach and details concerning the use of next generation sequencing to analyze the 4C data can be found in Supplemental Fig. S1.

Shared long-range interactions between active and between inactive genes present on the active X chromosome

To understand conformational changes imposed on the X chromosome by the non-coding RNA Xist, we first analyzed the active X chromosome. Four genomic loci, or 'viewpoints', were initially chosen because they are embedded in different prototypical chromosomal contexts, with respect to gene activity, gene density or their ability to escape XCI (Fig. 1C). We started by investigating the nuclear environment of two active X-linked genes, *MeCP2* and *Jarid1C,* which are located in different gene-dense regions of the X chromosome. Both genes were engaged in many DNA interactions in *cis*, across the entire X chromosome (Fig. 2A), as well as in *trans* (Fig. 2B). Each gene was also found to contact the other gene, despite being separated almost 80 Mb on the linear DNA template. In fact, the two genes shared many of their interacting regions in *cis*, as well as in *trans* (Fig. 2C, and Supplemental Table S1). Characterization of the *cis*-interacting regions revealed that they were on average 490kb in size and enriched for other active genes that also locate in gene-dense regions of the X (Supplemental Fig. S3A, B). Furthermore, contacted regions show an enrichment of SINE repeats and a depletion of

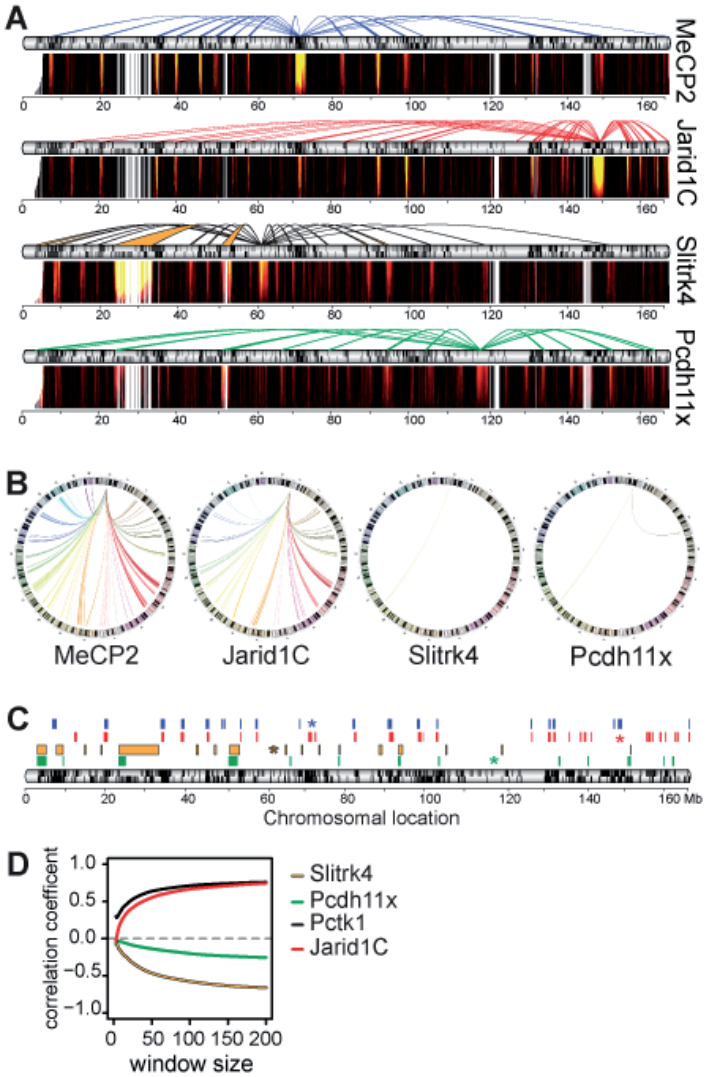


Figure 2: Long-range interactions engaged by the different viewpoints separate active from inactive chromatin on the active X chromosome. A: Spiderplots combined with domainograms depict long-range interactions identified with the four different viewpoints on the active X chromosome (detailed explanation can be found in chapter 4). Significant interactions are depicted in the spiderplot by the different colored lines, each color representing a different viewpoint. Small black bars below the spiderplot represent genes located on the X chromosome. Significance of interaction is indicated by the range in color used in the domainogram below each spiderplot; black is low (p=1), yellow represents high significance ($p=10^{-10}$) of interaction. **B**: Interactions with other chromosomes are depicted in Circos plots, each line represents a *trans*-interaction. Chromosomes are plotted round the circle. Colors indicate the chromosomes that were contacted. **C**: Cis interactions from (A) merged in one figure for comparison. Interactions are represented by the colored bars (*MeCP2*: blue, *Jarid1C*: red, *Slitrk4*: gold and *Pcdh11x*: green). The colored asterisk indicates the position of each viewpoint on the X chromosome. **D**: Spearman correlation calculated comparing 4C profiles of *Pctk1*^{Xa}, *Slitrk4*^{Xa}, *Pcdh11x*^{Xa} and *Jarid1C*^{Xa} to *MeCP2*^{Xa}. Further characterization of the identified interacting regions can be found in Supplemental Fig. S3.

LINE repeats, which is typical for gene-dense regions (Supplemental Fig. S3C). Regions located on other chromosomes that were contacted by the two genes showed similar characteristics (data not shown).

Two other genes analyzed, *Pcdh11x* and *Slitrk4*, were located in gene-poor regions, with *Pcdh11x* being active and *Slitrk4* inactive in NPCs. Both were found to be engaged in many specific long-range interactions in *cis* (Fig. 2A), but formed few specific contacts with other chromosomes (Fig. 2B). In contrast to *MeCP2* and *Jarid1C*, interactions with the inactive *Slitrk4* were mostly with other inactive, gene-poor sequences on the X. Correspondingly, these regions were enriched in LINEs and depleted for SINEs. Surprisingly, *Pcdh11x* which is thought to be an active gene [Mikkelsen et al 2007], contacted regions with similar characteristics of inactivity and in fact shared many of its interacting partners with the inactive *Slitrk4* gene (Fig. 2C). However, when analyzed by RNA FISH, *Pcdh11x* was found to be active in only 13% of cells, implying that this gene is in fact inactive in most cells analyzed at a given time by 4C. For comparison, *MeCP2* and *Jarid1C* loci were detected as active by RNA FISH in >95% of the cells (data not shown). Collectively, the data show that irrespective of its activity and chromosomal context each gene is engaged in many specific long-range DNA interactions across its chromosome. Active genes far apart on the chromosome, located in gene dense regions, interact with each other and share a distinct set of interactions with other active genes in *cis* and in *trans*. 'Inactive' genes separated on the chromosome and located in gene poor regions similarly share interactions, but now with other inactive regions. This spatial segregation of active from inactive chromatin is also illustrated by the anti-correlation found for 4C profiles of inactive genes versus active genes (Fig. 2C, D). The applied spearman's rank correlation analysis values similarity between the interaction profiles of two experiments by comparing the calculated z-scores (depicted in the domainograms). Active, gene-dense regions clearly show more interchromosomal DNA interactions than the more inactive gene-poor regions. This is in agreement with gene-dense active chromatin being more often looped out or at the periphery of the chromosome territory than inactive regions (Mahy et al. 2002a).

Inactive genes are randomly positioned inside the territory of the inactive X chromosome

Having established an understanding of X_a conformation of the active X in NPCs, we next focused on the inactive X. First we confirmed that the two genes analyzed above, *MeCP2* and *Pcdh11x*, were indeed both subject to silencing on the X_i in NPCs (Supplemental Fig. S2E and data not shown). To our surprise, 4C revealed that the two genes showed a near complete loss of specific long-range contacts on the X_i (Fig. 3A). The same was found for *Slitrk4* which is inactive on both X_i and X_a in NPCs. Whereas captured sequences on the X_a tend to cluster at specific regions, indicative of the formation of specific contacts, the long-range captured sequences of all three silenced loci on the X_i are distributed much more randomly. This was not due to limited restriction digestion efficiency, which was similar for active and inactive genes across the X_i (data not shown). It was also not caused by an inability of the X_i-linked inactive genes to reach outside their local chromatin structure, as compared to their counterpart genes on the X_a they all capture a relatively higher number of sequences over large (> 1 Mb) versus short

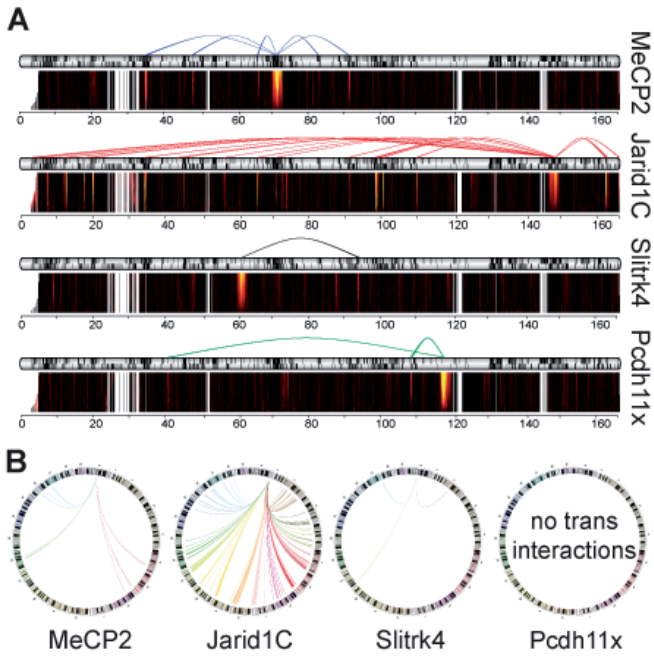


Figure 3: Allele-specific 4C analysis reveals dramatic changes in chromosome conformation on the inactive X chromosome. A: Spiderplots combined with domainograms depict long-range interactions identified on the inactive X chromosome. **B**: Trans interactions of the different viewpoints are depicted in Circos plots, each line represents a *trans*-interaction. Detailed analysis of the distribution of captured sequences on the X_i can be found in Supplemental Fig. S4.

distances (< 1 Mb) in *cis* (Supplemental Fig. S4B). The random distribution of long-range captured sequences on the X_i suggests that inactivated loci on the X_i no longer reside in preferred 3D genomic neighborhoods. This loss of specific contacts has not been observed in previous 4C or HiC studies (Simonis et al. 2006; Lieberman-Aiden et al. 2009), and in this study was exclusive for inactive loci on the X_i, since inactive loci on the X_a (Fig. 2A) as well as on autosomes (data not shown) did show normal interactions.

Escaping genes cluster and locate to the periphery of the Xist domain

Unlike silenced genes, *Jarid1C*, a gene well known to escape the XCI process, formed multiple long-range interactions on the X_i (Fig. 3A). Moreover, *Jarid1C*, and other escapees (see below), were the only genes on the X_i to show many specific interactions with regions on other chromosomes. This is consistent with the previous observation of repositioning of silenced chromatin to a more internal position of the X_i compared to escapees, which are positioned at the periphery or outside of the Xist RNA domain (Dietzel et al. 1999; Chaumeil et al. 2006). As active genes are known to be able to come together in nuclear space (Simonis et al 2006; Lieberman-Aiden et al 2009; Fig. 2A), we wondered whether contacts in *cis* were made with other genes escaping XCI. Indeed, the few genes known to escape XCI, such as *Utx*, *Eif2s3x, Xist* and *MidI*, the latter located in the pseudoautosomal region, were

located within regions contacted by *Jarid1C*^{Xi} (Fig. 4A). This prompted us to investigate if we could identify other escapees based on the 4C interaction profile of *Jarid1C* on the X_i. Employing a database of SNPs between 129/SvJ and CAST (Frazer et al. 2007), we tested eight other regions contacted by *Jarid1C*^{Xi} for the presence of escapees by cDNA analysis (see Methods for details). All of them contained at least one escapee, whereas all analyzed genes residing in non-contacted regions did not escape XCI (Fig. 4A and Supplemental Table S2).

One of the most distal contacts made by *Jarid1C* is with *Pctk1*, a gene located in a gene-dense area ~128 Mb away on the linear

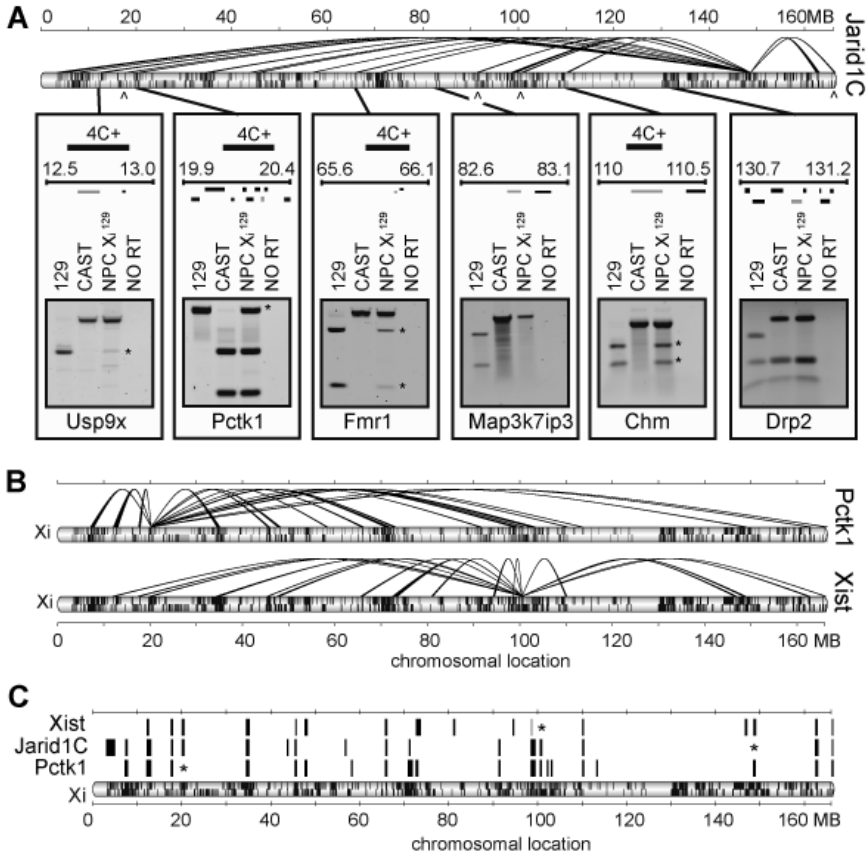


Figure 4: Escaping genes cluster in nuclear space. Jarid1C^{Xi} interacting regions contained other escapees. **A**: Genes known to escape XCI, (from left to right) *Utx*, *Eif2s3x*, *Xist* and *Midl* are indicated by arrowheads below the chromosome. The location and size of these regions are represented by the scaled black bars. The gel pictures show the result of the allele-specific cDNA analysis performed on the indicated genes (grey) located within the Jarid1C^{Xi} interacting regions. Other genes present in the interrogated region are drawn in black. The asterisk indicates the PCR product of transcripts originating from the X_i. *Map3k7ip3* and *Drp2* locate outside Jarid1C^{Xi} interacting regions and were found silenced on X_i. **B**: Spiderplot depicting long-range interactions of *Pctk1*^{Xi} (top) and *Xist*^{Xi} (bottom). **C**: Bars, representing long-range *cis* interactions on the X_i, from *Xist*, *Jarid1C* and *Pctk1* are plotted for comparison. Asterisks' indicate the position of the interrogated genes on the X chromosome. 3D FISH analysis confirming the interactions identified by 4C can be found in Supplemental Fig. S5A-C.

template. When analyzed, we found that *Pctk1* also escapes XCI (Fig. 4A). We used this gene and *Xist,* well known for its activity on the X_i, as new viewpoints for further querying the structure of chromatin that escapes XCI. Like *Jarid1C*, *Pctk1* and *Xist* are engaged in many long-range interactions across the entire inactive X chromosome. Most interactions were shared between the interrogated escaping genes (Fig. 4C), providing more evidence for the concept of escapees being capable of meeting each other in the nuclear space. Both *Jarid1C* and *Pctk1* showed many interchromosomal interactions when analyzed from the X_i, even more than when analyzed from the X_a chromosome (Fig. 3B and data not shown), confirming that these escapees preferentially locate at the periphery or outside of the nuclear territory occupied by the inactive X (Chaumeil et al. 2006; Clemson et al. 2006). RNA FISH experiments using probes against nascent transcripts have shown that most genes transcribe in bursts and are active only part of the time (Osborne et al. 2004; Chubb et al. 2006). This would suggest that interaction frequencies between discontinuously transcribed genes will be much higher when measured by nascent RNA-FISH compared to DNA FISH, if such contacts are dependent on ongoing transcription. We found no obvious differences when interaction frequencies between pairs of genes on the inactive X were measured by RNA or DNA FISH, even for genes that are active less than 65% of the time. This argues that the co-localization of escapee loci may not be dependent of ongoing transcription (Supplemental Fig. S5D-E). We further noticed that the regions contacted by *Jarid1C* and *Pctk1* on the X_i were not necessarily the same as those contacted on the X_a. In fact, spearman rank correlation analysis, comparing the X_i and X_a for the same locus (Supplemental Fig. S6) revealed that the *cis*-interaction profiles of escapees were as dissimilar as those of the non-escaping genes. Collectively, these data argue for a unique topology of the inactive X in which randomly folded inactive chromatin makes up the core, while escapees are positioned at the periphery of the X_i territory. The trans-interaction profiles of escapees on the X_a and X_i, on the other hand, showed a high percentage of overlap (Supplemental Table S1), arguing that the two X chromosomes in female cells do not adopt different positions relative to autosomes.

Deletion of Xist causes partial re-folding of the inactive X to a structure reminiscent of the active X chromosome without affecting its inactive state

The X chromosome is particularly suited for studying factors that dictate chromosome shape, as its conformation completely changes upon the expression of the long non-coding Xist transcript. Xist is crucial for initiation, but is not required for maintenance of XCI. In order to assess whether Xist RNA dictates X chromosome conformation, even after the initiation phase of XCI, we used a similar conditional knock-out strategy to ablate *Xist* in NPCs (Fig. 5A) as was published before (Csankovszki et al. 1999; Kohlmaier et al. 2004; Pullirsch et al. 2010). After inducing deletion of *Xist* and subsequent culturing of the $Xist^{KO}$ NPCs for two weeks (cell numbers increased >10 fold), analysis of DNA and RNA levels revealed that approximately 80% of the cells harbored a recombined *Xist* locus (Fig. 5B, C and Supplemental Fig. S7A). As reported previously, the knock-out of *Xist* showed a dramatic reduction in Ezh2 and H3K27me3 accumulation on the inactive X (Fig. 5D, E). This occurred without changing DNA methylation levels on the X_i, or re-activating expression of silenced genes (Fig. 5F, G).

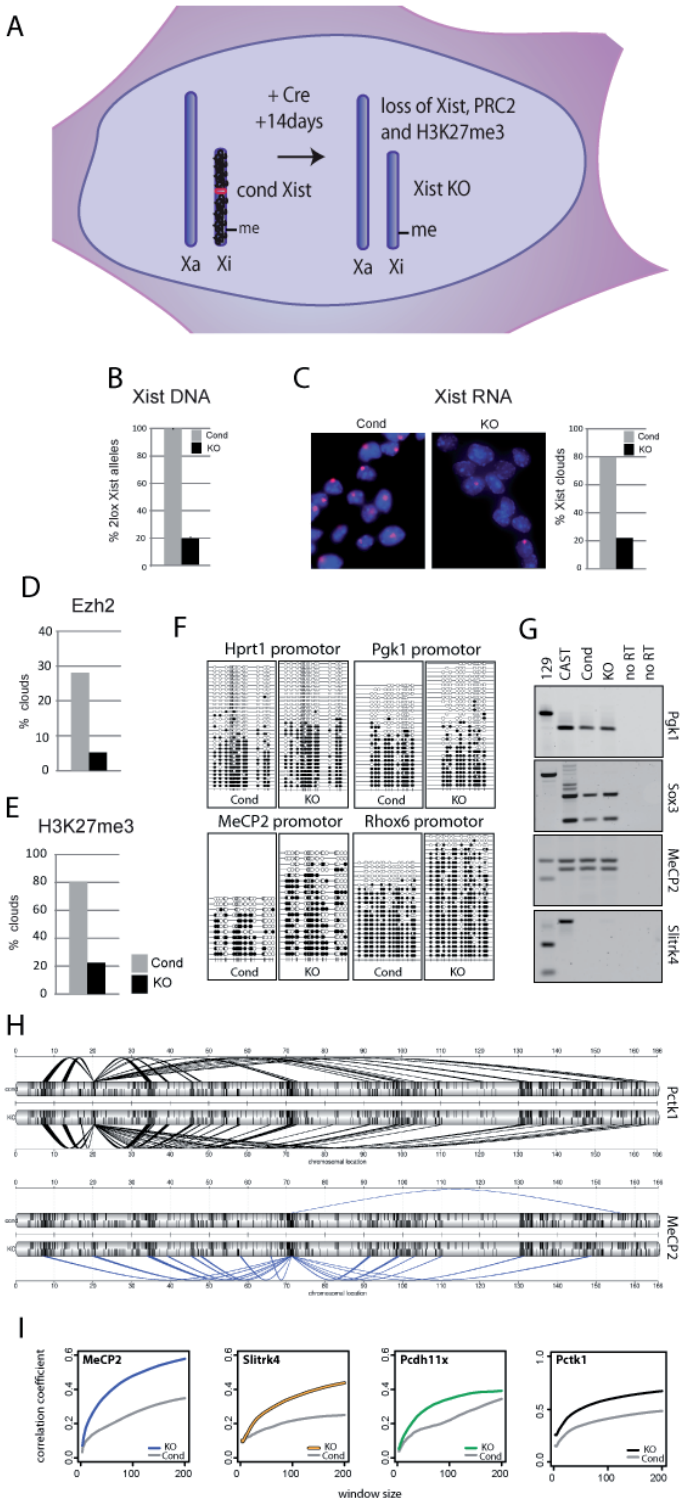


Figure 5: The inactive X chromosome partially refolds to the active X chromosome upon conditional deletion of *Xist*. **A**: Schematic representation of the *Xist* knock out strategy. **B**: Allele-specific qPCR result, measuring recombination efficiency. **C**: The quantification of the Xist RNA FISH applied on $Xist^{KO}$ and $Xist^{CON}$ NPC. A representative example of the FISH experiment is shown. **D** and **E**: quantification of Ezh2 and H3K27me3 clouds respectively, identified by IF applied on $Xist^{KO}$ NPC. **F**: Bisulfite sequencing result of X linked gene promoters in both the control NPC and $Xist^{KO}$ NPC. Open circles represent non methylated CpGs, while methylated CpGs are represented by the filled dots. **G**: Allele-specific expression analysis of four X linked genes in control NPC and $Xist^{KO}$ NPC. Allelic transcript contribution is visualized by separating digested RT-PCR products, using a RFLP recognizing restriction enzyme, on an agarose gel (see also Supplemental Fig. S7). **H**: Spiderplots depicting long-range interactions of $Pctk1^{Xi}$ (black) and $MeCP2^{Xi}$ (blue) identified in the control NPC (top) and $Xist^{KO}$ NPC (bottom). **I**: Spearman's rank correlation calculated comparing the 4C profiles identified in the control NPC and $Xist^{KO}$ NPC to the corresponding X_a profile of the indicated viewpoints (see also Supplemental Fig. S7).

To determine the consequences of Xist depletion on the 3D structure of the X chromosome, we performed allele-specific 4C, as before, on $Xist^{KO}$ and control NPCs. We analyzed DNA interaction profiles of the aforementioned escapees *Pctk1* and *Jarid1C*, as well as the silenced genes *Slitrk4*, *MeCP2* and *Pcdh11x* and *Sox3* (ChrX 58.1Mb). The latter was found as an interaction partner of active genes on the X_a, but is subject to X inactivation and appeared ignored by these and other genes on the X_i chromosome. Upon depletion of Xist, the two escapees were found to contact the same (but larger) regions, as well as additional ones, including the regions containing *MeCP2* and *Sox3*. This suggests they may re-engage in specific contacts after removal of *Xist* (Fig. 5H). Indeed, both *MeCP2* and to a lesser degree *Sox3*, as well as two other inactive genes, *Pcdh11x* and *Slitrk4*, were all found to partially regain their preference for specific genomic neighborhoods when *Xist* expression was lost from the inactive X chromosome. We noted that many of the re-gained interactions appeared to be similar to those found on the active X chromosome, and therefore asked whether depletion of Xist leads to refolding of the X_i into the ground-state configuration of the active X chromosome. A Spearman's rank correlation analysis revealed that indeed when *Xist* is no longer expressed the silenced genes on the X_i adopt interactions more reminiscent of those seen on the X_a (Fig. 5I and Supplemental Fig. S7B). This was the case not only for silenced genes, but also for one of the escapees, *Pctk1*. On the other hand, *Jarid1C* showed no such changes. Taken together this data demonstrates that depletion of Xist results in partially refolding of the X_i chromosome into a structure more reminiscent of the X_a.

Changes seen in overall chromosome topology have frequently been correlated to changes in gene activity. Indeed, this led to the idea that the act of transcription might be crucial for the nuclear positioning of active genes (Sexton et al. 2007; Cook 2010). Although drug-induced transcription inhibition experiments have usually failed to show appreciable changes in chromosome topology (Tumbar et al. 1999; Palstra et al. 2008; Muller et al. 2010), these experiments could not formally exclude that an initial act of gene transcription prior to drug treatment might be the force driving gene positioning. The conditional *Xist* knockout system used here is unique in that it is known to retain gene silencing, yet appears to affect gene positioning. To further exclude that genes re-activate in our system, we used an allele-specific transcription assay for 11 additional X-linked genes, 9 of which are silenced and 2 genes escaping XCI. Only one gene, *Atp7a*, showed slight de-repression, as previously reported (Zhang et al. 2007). All others, including *MeCP2*, *Sox3* and *Slitrk4* loci, showed an identical silent status on the X_i, with or without *Xist* expression (Supplemental Fig. S7C). Our data thus demonstrate that the detailed folding of the inactive X chromosome is very much reliant on Xist RNA, even in cells where XCI is complete. The role of Xist RNA on X_i chromosome conformation may either be direct or due to the downstream recruitment of PcG and/or macro-H2A. Our data demonstrate that this function of Xist RNA is independent of gene transcription and DNA methylation, neither of which is affected after deletion of *Xist*.

DISCUSSION

Chromosome topology and X inactivation

The three-dimensional conformation of the X chromosome and its position in the nucleus has been implicated in initiation and maintenance of X inactivation. In order to better understand the interplay between DNA topology and XCI, detailed structural maps of both the inactive and active X chromosome are needed. Here, we provide the first high resolution interaction maps of the active and inactive X chromosomes in somatic cells that have stably established X inactivation. FISH experiments previously showed that escapees tend to locate at the periphery of the Xist RNA territory. In agreement with this, using 4C, providing high resolution molecular interaction maps, we consistently find these genes to be engaged in many inter-chromosomal contacts. In contrast, silenced genes on the X_i which are believed to reside preferentially inside the inactive Xist RNA domain, are found engaged in very few inter-chromosomal interactions. Importantly, we also show that escapees tend to contact each other, including the highly active *Xist* locus (see below), as well as the active pseudo-autosomal region at the tip of the X chromosome. We were able to use the DNA interaction maps of escapees to identify a total of nineteen genes that escape XCI in NPCs (Supplemental Table S2). Previously, only a few escapees had been discovered in mice. The highest reported number of escaping genes came from a recent RNA-sequencing study that identified thirteen escapees in a cell line cultured from mouse embryonic kidney (Yang et al. 2010). We confirm that nine of these escapees interact with both *Jarid1c* and *Pctk1* in NPCs. Further allele-specific expression analysis of interacting genes led to the identification of another ten genes that are active on the X_i chromosome in NPCs (Supplemental Table S2).

Some of the escaping genes identified (*Usp9x*, *Pctk1* and *Fmr1*) can be classified as tissue-specific genes, showing that tissue-specific factors may impact on the mechanisms underlying escape from XCI in mice, as was found previously in human (Prothero et al. 2009). Escapees appear to be randomly distributed along the X chromosome, with some in gene deserts, and others in areas dense with silenced genes. This supports the idea that in mice, the ability to escape is a gene-intrinsic property (Li and Carrel 2008; Yang et al. 2010). The situation is slightly different in humans, where many more genes can escape XCI and many escapees are clustered on the short arm of the human X chromosome (Carrel and Willard 2005).

The fact that during the XCI maintenance phase the *Xist* locus interacts with escapees is interesting when considering the mechanisms of Xist RNA spreading. If diffusion through the nucleoplasm plays a role in the efficient spreading of Xist RNA across the X chromosome, it is striking that the regions spatially close to the source of Xist RNA production are not necessarily those that become silenced by Xist. It emphasizes that escapees must have powerful mechanisms to counteract Xist-induced repression. Escapees are randomly distributed across the entire X chromosome, implying that most will be flanked on the linear chromosome by sequences sensitive to repression by Xist. These sequences will automatically be dragged along when an escape gene loops towards the Xist locus. One could speculate that this 3D organization helps Xist spreading across the X chromosome at this stage of

development when X inactivation has long been established.

Our study reports for the first time that silenced genes fail to show preferred neighboring sequences in *cis*. This has not been reported before, as inactive regions on autosomes (Simonis et al. 2006; Lieberman-Aiden et al. 2009) and on the active X (this study) so far invariably were found to be engaged in specific contacts with other inactive regions elsewhere on their chromosomes. This novel type of organization may be a reflection of highly dynamic interactions inside the inactive X domain, but will also be found when a gene's spatial orientation to other chromosomal parts is relatively stable but different for each cell in the population analyzed. We speculate that the latter organization may be expected if chromosomal regions show little differences in their epigenetic landscape (see below). It will be interesting to determine whether this type of random organization of genes within the chromosome territory also occurs during other stages of differentiation and/or at other parts of the genome.

Transcription and the shape of the inactive X chromosome

Our study provides further evidence supporting the correlation between a gene's expression status and its exact position relative to other genomic regions in the nucleus (Fraser and Bickmore 2007). The outstanding question is now whether transcription and genome topology are causally related or not. Previous genome-wide DNA topology studies provided detailed evidence for the spatial segregation of active and inactive chromatin inside the cell nucleus (Simonis et al. 2006; Lieberman-Aiden et al. 2009). It was additionally suggested that functionally related genes preferentially come together in the nuclear space (Schoenfelder et al. 2009), although this was not immediately clear from other studies on the same genes (Simonis et al. 2006). Two extreme models may explain the link between transcription and chromosome conformation. One proposes that genes need to migrate to specific nuclear locations for their transcription (Chakalova and Fraser 2010). Another possibility is that chromosome structure and transcription are independent parameters that both follow some physical properties of chromatin. The structure of the inactive X chromosome, with escapees locating peripheral and silenced genes more inside the chromosome territory is compatible with both models for nuclear organization. Our observation that escapees tend to interact with each other, though, seems to be further evidence for the idea that active genes meet at dedicated transcription sites. However, the fact that with RNA and DNA FISH the same percentage of interaction between pairs of escapees is measured (Supplemental Fig. S5D) suggests that the loci that are not actively transcribing meet as frequently as the transcribing loci. For example, the escaping genes *Pctk1* and *Fmr1*, located on the X_i, are both active only 60% of the time (Supplemental Fig. S5E). Assuming that they independently control their bursts of transcription (Chubb et al. 2006), this means that only ~36% of the cells will have both alleles transcribed at the same time on the same chromosome. Although located 40Mb apart, they show an interaction on the X_i in 20% of the cells, as measured by DNA FISH. If this were accounted for exclusively by the actively transcribed alleles, RNA FISH would yield a much higher interaction frequency within the 36% of cells transcribing both genes. However this is not the case, with RNA and DNA FISH measurements in cells showing the two (active) alleles do not differ significantly (23% versus 20%).

The data therefore suggests that escapees co-localize in the nucleus irrespective of their transcriptional status, but perhaps due to chromatin features of these genes.

Experiments that try to address the causal relationship between nuclear positioning and transcription often manipulate the one factor and analyze its consequences for the other. Gene repositioning from the nuclear interior to the periphery, via induced anchoring to the nuclear lamina, is one such experiment. In one study this was found to have no effect on a reporter's ability to activate transcription (Kumaran and Spector 2008), in another study to cause gene silencing (Reddy et al. 2008) and in a third study to repress some genes while not affecting others (Finlan et al. 2008). A similar lack of coherent results is apparent from studies that use transcription inhibitors to study the impact of transcription on gene positioning. Some studies reported (minor) DNA topological changes upon transcription inhibition (Mahy et al. 2002b; Branco and Pombo 2006; Naughton et al. 2010), while others, including a 4C study (Palstra et al. 2008) failed to measure an effect of transcription on DNA structure (Tumbar et al. 1999; Muller et al. 2010). Inherently such studies may not be well suited to address this relationship though. The chemical and heat-shock treatments that block transcription also induce other types of cellular stress, making it difficult to unambiguously relate observed conformational changes to changes in transcription. Vice versa, not observing an effect does not exclude that an early act of transcription prior to (drug) treatment was responsible for setting up the 3D structure measured. Nuclear repositioning independent of changes in transcription has been observed before, but only for some individual loci and measured relative to nuclear landmarks like the chromosome territory or nuclear lamina (Morey et al. 2008; Peric-Hupkes et al. 2010). Our data provide high resolution molecular interaction maps that uncouple topological changes from transcription on a chromosome wide scale. Following depletion of Xist, we find that genes present on the inactive X chromosome remained transcriptionally silent, but that the inactive X chromosome shows profound changes in its shape, adopting a structure reminiscent of the active X chromosome. This observation suggests that Xist induced X chromosome topology does not play a role in maintenance of XCI. A recent study generated chromatin compaction profiles of the human active and inactive X chromosome. Only minor compaction differences between the two chromosomes were found at the 30 nanometer scale. However, the volume of the X_a, normally larger than that of the X_i, was demonstrated to become similar to the X_i volume after transcription inhibition (Naughton et al. 2010). Our detailed DNA interaction maps of the active and inactive X chromosome suggest that although similar in size, the two X chromosomes in their study are likely to be folded very differently. Collectively, our data shows that gene interactions and chromosome folding are at least partially independent of gene expression. A remaining possibility is that Xist depletion allows for re-loading of RNA polymerase II molecules onto inactive genes that cannot complete gene transcription; such paused polymerase molecules may then be responsible for the newly gained long-range DNA interactions. Alternatively or in parallel, other factors than polymerase may play an important role in shaping the inactive X chromosome.

Factors shaping the inactive X chromosome

Recently PcG proteins were found to mediate long-range intrachromosomal interactions between cognate regions tens of megabases apart in *Drosophila* (Bantignies et al. 2011). In addition, the polycomb group protein complex PRC1 has been shown to alter large-scale chromatin structures (Eskeland et al. 2010). The latter observations may be relevant for X chromosome folding as well. A scattered presence of PcG proteins along the inactive regions of the X_i chromosome may well account for the random organization of inactive chromatin that we observe inside the Barr body. After all, preferential interactions are expected to occur only when chromosomal regions differ sufficiently from each other in chromatin composition and associated factors. In addition to PcG proteins, macro-H2A and obviously the long-non-coding RNA molecule Xist itself may play a direct role in shaping the inactive X chromosome. The knockdown of such factors in NPCs turned out to be difficult also because their depletion, unlike that of Xist, causes genome-wide, rather than X-specific, epigenetic and expression changes. Further deciphering the factors that shape the X therefore awaits more sophisticated strategies. Collectively, our data show that that the inactive X chromosome in female cells adopts a unique three-dimensional structure that is dependent on Xist RNA and, at least partially, independent of transcription and DNA methylation.

METHODS

Cell culture

NPCs were generated according to Conti and collogues (Conti et al. 2005) with slight modifications. In brief three independent 129SVJ/CAST ES cell lines (Csankovszki et al. 1999; Luikenhuis et al. 2001; Jonkers et al. 2008) that remain XX upon differentiation were differentiated in N2B27 (StemCell Recourses) for 7 days, followed by the formation of neural spheres in N2B27 supplemented with EGF and FGF (10ng/ml). 3 day old spheres were allowed to attach to the culture dish to expand NPCs. After 2 passages cells were seeded in low density and colonies were picked and analyzed for NPC identity using IF (see below) and proper XCI. A single clone was selected from each cell line that exclusively had silenced the 129SVJ (2x) or the CAST (1x) X chromosome. Cre mediated *Xist* knock out was achieved by first differentiating 129SVJ-Xist2lox/CAST ES cells (Csankovszki et al. 1999) to NPCs. NPCs showing proper XCI were transfected with pCMV-Cre-puro using Amaxa® according to manufacturer's protocol, followed by transient puromycin selection for 2 days. $Xist^{KO}$ NPCs were cultured an additional 12 days with cell numbers increasing >10 fold before analysis to dilute out stable X_i markers.

4C analysis

To distinguish between the conformation of the X_a and X_i in female cells we designed an allele-specific 4C approach that is outlined in Figure 1A. Comparing the DNA sequence of *Mus musculus domesticus* and *Mus musculus castaneus*, employing a database of SNPs (Frazer et al. 2007) or after PCR amplification and sequencing, we confirmed/ identified restriction fragment length polymorphisms (RFLPs). Based on identified RFLPs we were able to design an allele-specific 4C strategy, when applied on SVJ129/CAST

cells. The initial steps of the 4C procedure, as published before [Simonis et al. 2006), remain unchanged. Cells are cross linked using formaldehyde, after which chromatin is digested and subsequently ligated under diluted conditions. After reversal of the cross links the DNA is purified and ready for the second restriction enzyme treatment. In order to design an allele-specific 4C procedure the viewpoint fragment is chosen such that a RFLP is located in between the primers used in the PCR amplification. As a consequence only the allele which is not digested by the second restriction enzyme can contribute to the 4C PCR product and is analyzed either by dedicated micro-array or high throughput sequencing. An example of the allele-specific formation of 4C PCR product is shown in Figure 1B, where Jarid1C amplifies only from the SVJ129 template with high efficiency. As a control the Hbb-b1 primers, which were not designed around a RFLP, were able to amplify similar amounts of PCR product from both templates. An alternative strategy, which allows the usage of a 'RFLP recognizing' restriction enzyme that is sensitive to CpG methylation is depicted in Supplemental Figure S1. Primer design and PCR product analysis were adapted to include the use of Illumina sequencing (see also Supplemental Figure S1 for detailed information). SNPs and primers used in the 4C analysis are available upon request. All data obtained for the X_i and some for the X_a (MeCP2) were verified in independent NPC clones, giving highly similar interaction profiles (data not shown). RFLPs in Pcdh11x allowed independent 4C analysis of both the X_i^{129} and X_i^{CAST} chromosome, which demonstrated that folding was highly similar between the genetically distinct X chromosomes (data not shown).

Data analysis

To identify interacting regions we set up a standardized 4C-seq data analysis workflow. The initial step in the 4C-seq analysis is the alignment of the sequencing reads to a reduced genome of sequences that flank HindIII sites (fragment ends), using custom perl scripts. Due to their ambiguous nature in reporting contacts, repetitive fragment ends were excluded from subsequent analysis. The reduced genome was based on mouse mm9. The data have been deposited in the Gene Expression Omnibus under accession nr. GSE29509. The proportional distribution of reads identified in the different experiments can be found in Supplemental Fig. S4A.

All statistical analysis was performed using the R programming language (R Development Core Team 2010). To avoid possible PCR artifacts we transformed the data to unique coverage (> 1 reads per fragment end is set to 1). Because the coverage declines as a function of the distance from the viewpoint (i.e. high coverage close to the viewpoint and low at larger distances), we normalized the coverage for the background coverage. To this end we calculated z-scores for a given window of fragment ends *i*, of size *w*, based on the relative unique coverage in the background window *W* ($p_{W,I} = cov_I/W$, where cov_I is the number of unique fragment ends covered in window *I*). We choose *I* such that I»i and i runs from ⌈w/2⌉ until N-⌊w/2⌋, where N is the number of fragment ends on the chromosome. Window i spans i-⌊w/2⌋ until i+⌊w/2⌋ for odd values of w and (i-w/2)+1 until i+w/2 for even values of w. In general I = i, except for values where I < W and I > N-W, there I = W or I = N-W, respectively. We calculated estimators for the mean and standard deviation, μ and σ, following the binomial distribution, for every window i given a window size w:

$$\mu_{w,i} = w \cdot p_W \,,\, \sigma_{w,i} = w \cdot p_{W,i} \cdot (1 - p_{W,i})$$

We use the relative unique coverage in window *w*, p_w, to calculate the z-score:

$$z_{w,i} = \frac{p_{w,i} - \mu_{w,i}}{\sigma_{w,i}}$$

To identify regions of non-random 4C signals (i.e. contacted regions) we used the false discovery rate. To this end we randomly permuted the dataset 100 times and determined the threshold z-score at which the false discovery rate was 0.01. For the *trans* interactions an FDR threshold of 0.01 was determined based on 100 random permutations of the data for every chromosome. A window size of 500 was used; windows that exceeded the threshold were scored as *trans* interactions.

The domainogram analysis was performed analogous to de Wit et al. [de Wit et al. 2008), but using a matrix of probability scores based on the matrix of z-scores (z_w, w = 2,3,…200). Probability scores were calculated based on the normal distribution. A $\log_{10}$ transformation of the probability score is used in the visualization.

Correlation analysis

Correlation analysis between different 4C experiments was performed by calculating the Spearman's coefficient of rank correlation (ρ) between the set of z-scores z_w (collection of all $z_{w,i}$), between two 4C experiments over a range of values for w. Supplemental Fig. S7A depicts a visual example.

Genomic annotation

Gene density scores were calculated based on the RefSeq annotation downloaded from the UCSC Table Browser [Karolchik et al 2004]. Repeat density analysis was based on LINE and SINE annotation from the Ensembl core v59 (Flicek et al. 2011). Neural progenitor expression data from Mikkelsen et al. (Mikkelsen et al. 2007) was used and can be downloaded under GEO accession GSE8024. Raw CEL files were normalized using RMA in the Bioconductor affy package (Bolstad et al. 2003) and were subsequently combined to one expression value. Probe locations were downloaded from the UCSC Table Browser, redundant probe locations were removed from the data.

Immunofluorescence

ES cells and NP cells cultured on coverslips were fixed in PBS containing 3% paraformaldehyde for 10 minutes and permeablized in PBS/0.4% Triton X-100 for 5 minutes on ice. Blocking and antibody hybridizations were performed for two hours at room temperature, using PBS/10% Fetal Calf Serum/0.05% tween-20. Antibodies used are as follows: for Oct4, ab19857 (Abcam) for Nestin, ab6142 (Abcam) for Gfap, ab4648 (Abcam) for Tubulin, ab7751 (Abcam) for EZH2, 612666 (BD Biosciences) for H3K27me3, ab6002 (Abcam) for donkey anti mouse IgG A594, 715-515-150 (Jackson) and for goat anti rabbit IgG FITC, 111-095-003 (Jackson). Dapi was used for DNA counterstaining. Images were collected using a Leica DM6000 B microscope equipped with a Leica DFC360 FX camera and Leica application suite 2.2.1 software.

RNA FISH and DNA FISH

RNA FISH and DNA FISH experiments were performed as described previously (Chaumeil et al. 2006), except cells were fixed before permeabilization. The following BAC clones were used as probes in the analysis. For *Pctk1*, RP23-362P12 for *Fmr*, RP24-183G11 for *Mecp2*, RP23-77L16 for ChrX: 87.2Mb, RP23-25P21 for *Xist*, CT7-399K20 for *Taf1*/*Ogt*, RP23-268G11 for *Chm*, RP23-118M16 for *Pcdh11x*, RP23-20G23 and for *Jarid1C*, RP24-148H21. For the generation

of probes, 300ng *Sau*3AI digested BAC DNA was fluorescently labeled with Cy5 dUTP (NEL 579001 EA, Perkin ELmer), Spectrum Orange dUTP (02N33-050, Enzo life sciences) or spectrum green dUTP (02N32-050, Enzo life sciences) using the BioPrime Array CGH Genomic Labelling System (Invitrogen) following manufacturers protocol. Specificity of the labeled probes was confirmed on metaphase spreads from mouse ES cells. 3D images were collected using a Leica DM6000 B microscope equipped with a 100X objective, Leica DFC360 FX camera, taking z-steps of 0.2 µm. Leica application suite 2.2.1 software was used both for image collection and deconvolution. 3D distance measurements were taken of 100 nuclei per data point using ImageJ software. Xist RNA co-staining allowed us to measure distances of both the X_a and X_i in the same nucleus. Measurements were only taken if signals from both chromosomes could be identified.

Allele-specific expression analysis

Allele-specific expression analysis based on RFLPs between X^{129} and X^{CAST} was performed as described before (Huynh and Lee 2003). In brief, RNA was isolated using TRIzol® following manufacturer's instructions. RNA was treated with DNase and converted into cDNA using random primers (Promega). PCR primers spanning an RFLP were used for the detection of gene transcripts. PCR products were digested with the appropriate restriction enzyme and separated by agarose gel electrophoreses. Images were captured using a Thyphoon9410 scanner (GE Healthcare) and analyzed using Image Quant software. SNPs and primers used in this analysis are available upon request.

qPCR

qPCR analysis was performed using MyIQ PCR machines and MyIQ software (BioRad) using standard sybergreen incorporation to detect PCR products. Primers used are available upon request.

Bisulfite sequencing

Bisulfite conversion was performed using the EZ DNA Methylation-Direct™ Kit (Zymo Research). Primers were designed using MethPrimer (Li and Dahiya 2002) and are available upon request. Correct PCR products were isolated from agarose gel, cloned into pGEMT-easy and transformed. 48 Colonies were picked and sequenced per condition. Sequencing data was analyzed using QUMA analysis tool (http://quma.cdb.riken.jp/), only including sequences showing >90% identity and C-T conversion.

ACKNOWLEDGEMENTS

The authors like to thank Eric Engelen for technical assistance and Patrick Wijchers for critical reading of the manuscript. This work was financially supported by grants from the Dutch Scientific Organization (NWO) to EdW (700.10.402, 'Veni') and to WdL (91204082 and 935170621), InteGeR FP7 Marie Curie ITN (PITN-GA-2007-214902) and a European Research Council Starting Grant (209700, '4C') to WdL.

SUPPLEMENTAL MATERIAL

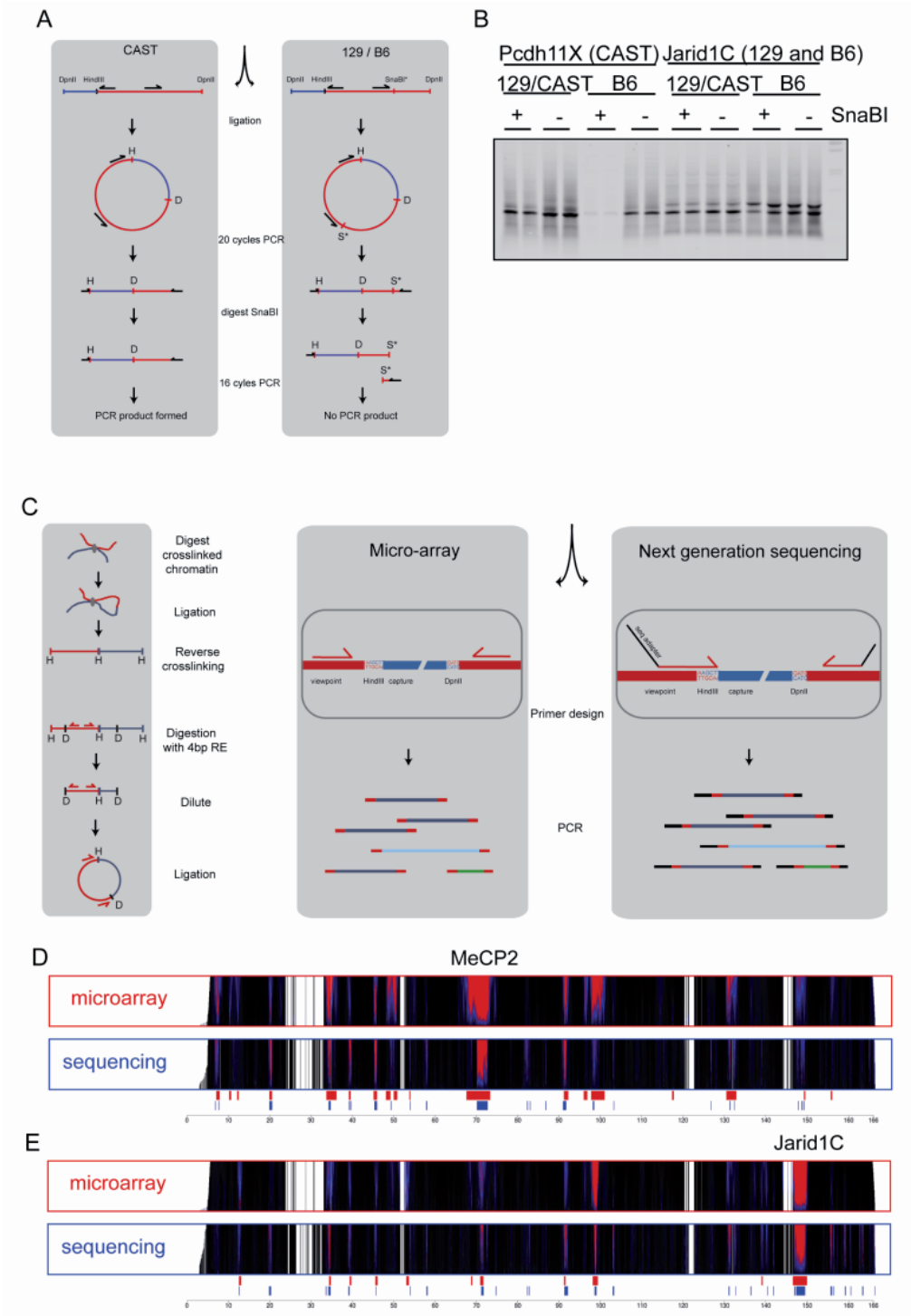
Supplemental Table S1: Active genes located in gene-dense areas share interactions identified on other chromosomes. Values indicated the percentage of overlapping region contacted by the different viewpoints comparing the different experiments. The first number in each cell indicates the percentage of interacting sequence shared between the 4C experiment on the left and the 4C experiment plotted on top. The second percentage represents the reciprocal analysis.

	MeCP2_Xa_trans	Pctk1_Xa_trans	Pctk1_Xi_trans	Jarid1C_Xa_trans	Jarid1C_Xi_trans
MeCP2_Xa_trans	100.0% / 100.0%	47.1% / 49.0%	66.9% / 34.7%	29.1% / 53.5%	28.6% / 31.7%
Pctk1_Xa_trans	49.0% / 47.1%	100.0% / 100.0%	67.5% / 33.7%	28.0% / 49.6%	19.9% / 21.2%
Pctk1_Xi_trans	34.7% / 66.9%	33.7% / 67.5%	100.0% / 100.0%	16.7% / 59.3%	25.0% / 53.3%
Jarid1C_Xa_trans	53.5% / 29.1%	49.6% / 28.0%	59.3% / 16.7%	100.0% / 100.0%	39.1% / 23.5%
Jarid1C_Xi_trans	31.7% / 28.6%	21.2% / 19.9%	53.3% / 25.0%	23.5% / 39.1%	100.0% / 100.0%

Supplemental Table S2: List of escaping genes. In total twenty X-linked genes, contacted by an escaping gene as measured by 4C, were either reported or confirmed to escape XCI. Escape from XCI was confirmed by allele-specific expression analysis, except for Taf1 and Ogt1. For these two genes, gene activity on the X_i was detected by RNA FISH in 82% of the cells, using a BAC probe (RP23-268G11) spanning both genes. Tissue specificity was determined based on data provided by http://biogps.gnf.org/.

gene ID	picked up by	gene position 1	gene position 2
ENSMUSG00000039231	Pctk1 and Jarid1C	7638297	7651886
ENSMUSG00000044148	Pctk1 and Jarid1C	12232005	12250794
ENSMUSG00000031010	Pctk1 and Jarid1C	12648624	12766815
ENSMUSG00000000787	Pctk1 and Jarid1C	12858096	12871178
ENSMUSG00000037369	Pctk1 and Jarid1C	17739701	17857062
ENSMUSG00000031065	Jarid1C	20265080	20277006
ENSMUSG00000016409	Pctk1 and Jarid1C	34666790	34690741
ENSMUSG00000036551	Pctk1 and Jarid1C	34690693	34708837
ENSMUSG00000085396	Pctk1	47908921	47988498
ENSMUSG00000000838	Pctk1 and Jarid1C	65931716	65971138
ENSMUSG00000031386	Pctk1 and Jarid1C	71188131	71211696
ENSMUSG00000031197	Pctk1	72759638	72780281
ENSMUSG00000035150	Pctk1 and Jarid1C	91434046	91458201
ENSMUSG00000031314	Pctk1 and Jarid1C	98728073	98797128
ENSMUSG00000034160	Pctk1 and Jarid1C	98835399	98879690
ENSMUSG00000086503	Pctk1 and Jarid1C	100655714	100678556
ENSMUSG00000031226	Pctk1	102275095	102312429
ENSMUSG00000025531	Pctk1 and Jarid1C	110154204	110299126
ENSMUSG00000025332	Pctk1	148667563	148709078
ENSMUSG00000035299	Pctk1 and Jarid1C	166123131	166428730

domain start	domain end	gene name	source	tissue specific
7162709	7878720	Suv39h1	confirmed	-
12107199	13056309	810030o07Rik	Yang et al, 2010	-
12107199	13056309	Usp9x	confirmed	+
12107199	13056309	Ddx3x	Yang et al, 2010	-
17551677	17930443	Utx	Yang et al, 2010	-
20082653	20518714	Pctk1	confirmed	+
34475296	35100569	Nkap	confirmed	-
34475296	35100569	Akap	confirmed	-
47642658	48182857	6720401G13Rik	Yang et al, 2010	
65726353	66151529	Fmr1	confirmed	+
71145255	71390304	Hcfc1	confirmed	+/-
72623098	73138391	Vbp1	confirmed	-
91264822	91574717	Eif2s3x	Yang et al, 2010 and confirmed	-
98498880	99512449	Taf1	confirmed by RNA FISH	-
98498880	99512449	Ogt		-
100601273	100990400	Xist	Yang et al, 2010 and confirmed	-
102141536	102407292	2610029G23Rik	Yang et al, 2010	-
110088894	110414742	Chm	confirmed	-
148481293	149025730	Jarid1C	Yang et al, 2010 and confirmed	-
166166237	166525117	Midl	Yang et al, 2010	-


A
CAST
129 / B6
DpnII HindIII DpnII
DpnII HindIII SnaBI* DpnII
ligation
20 cycles PCR
digest SnaBI
16 cyles PCR
PCR product formed
No PCR product
B
Pcdh11X (CAST)
Jarid1C (129 and B6)
129/CAST B6
129/CAST B6
+ - + - + - + - SnaBI
C
Digest crosslinked chromatin
Ligation
Reverse crosslinking
Digestion with 4bp RE
Dilute
Ligation
Micro-array
Next generation sequencing
viewpoint
HindIII
capture
DpnII
Primer design
PCR
D
MeCP2
microarray
sequencing
E
Jarid1C
microarray
sequencing

◁ **Figure S1: Additional information on an alternative allele-specific 4C strategy and 4C analysis by next generation sequencing. A:** To distinguish between the conformation of the X_a and X_i in female cells we designed an allele-specific 4C approach as outlined in Figure 1A, B. An alternative strategy, which allows the usage of a 'RFLP recognizing' restriction enzyme that is sensitive to CpG methylation is depicted here. In this strategy 4C product is first amplified using 20 cycles of PCR, after which it is treated with the RFLP enzyme to create the allele specificity. An example of the allele-specific formation of 4C PCR product is shown in **B,** where Pcdh11x primers only amplify in an allele-specific (here: CAST) manner upon *Sna*BI digestion. As a control the Jarid1C primers, which were not designed around a *Sna*BI restriction site, were able to amplify similar amounts of PCR product from both digested and undigested template. **C:** 4C sample preparation is similar between array and NGS analysis. The difference between these two options of analyzing 4C data mainly concerns primer design in two ways. One: using the Illumina platform the sample requires adaptor sequences that allow the DNA to bind to the flowcell. By using primers containing these adaptor sequences we limit sequencing to products formed in the PCR reaction, thus improving the signal to noise ratio. The second restriction in primer design concerns the limited read length of NGS on the Illumina platform. Although improving over time, experimental design is such that 36bp read length results in the generation of informative data. This is done by designing the *Hin*dIII –or read primer as short as possible (typically 18-20bp) and placing it on top of the *Hin*dIII restriction site. The first 18bp analyzed in the sequencing reaction are from the primer and are used as a barcode to identify captures from each experiment pooled in one sequencing lane. Typically we mix 10-15 experiments per lane, mostly yielding more than 1 million mappable interactions per viewpoint. Because the *Hin*dIII sequence in the primer is redundant for the capture it adds 6bp to last 18bp of the sequencing read resulting in a total of 24bp of capture sequence that is used for mapping purposes. A comparison between two experiments, PCR amplified from the same template but analyzed differently, resulted in the identification of highly similar regions, but with higher resolution (**D** and **E**) due to an increase in signal to noise ratio.

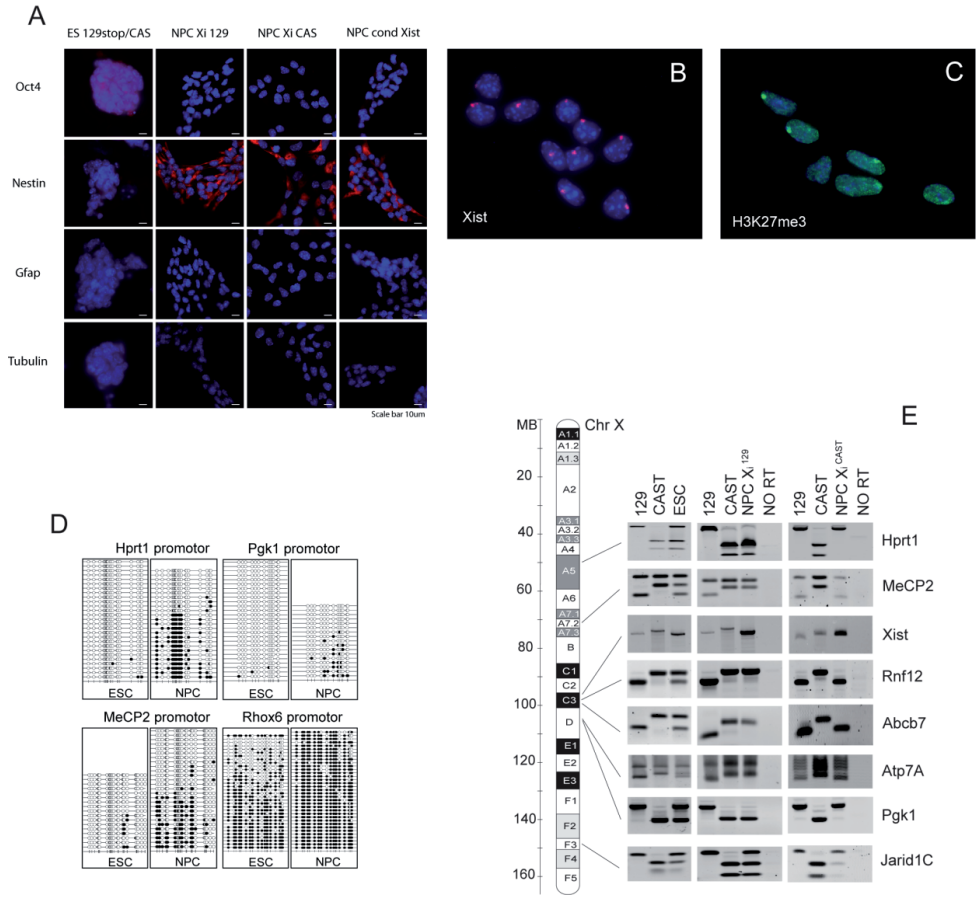


Figure S2: The X chromosome inactivation process is well represented in NPCs. Dependent on which X chromosome carries the RFLP, the 4C analysis is directed to either the 129SvJ or the CAST chromosome. To study both the X_a and X_i two cell lines were generated, one where 129SvJ was active (X_a^{129}) and CAST is silenced (X_i^{CAST}) and one with X_a^{CAST} X_i^{129}. **A:** NPC identity was confirmed by IF using antibodies against various proteins representing different stages of differentiation. Oct4, loss of embyonic stem cell (ESC) features; Nestin, neural lineage marker (typically >98% positive n=100); Gfap, astrocyte and Tubulin, neuronal marker. Scale bar represents 10µm. **B-E:** Verification of different marks of XCI in NPC. The formation of a Xist RNA cloud, an early mark of XCI, was detected by RNA FISH (**B**). PRC2 recruitment to the X_i was confirmed by an enrichment of Ezh2 (data not shown and Fig. 6D) and its mark, H3K27me3 by IF (**C**). Bisulphate sequencing confirmed the presence of DNA methylation as roughly 50% methylated alleles in NPC while in ES cells almost no DNA methylation could be detected (**D**). Rhox6 is XCI independently methylated and served as a control. Gene silencing was confirmed by allele-specific cDNA analysis (details in Methods) (**E**). Interrogated genes and cell types are indicated. 129 and CAST templates (first two lanes of each panel) and ESC served as references to determine the origin of the transcripts found in NPC Xi^{129} and NPC Xi^{CAST}.

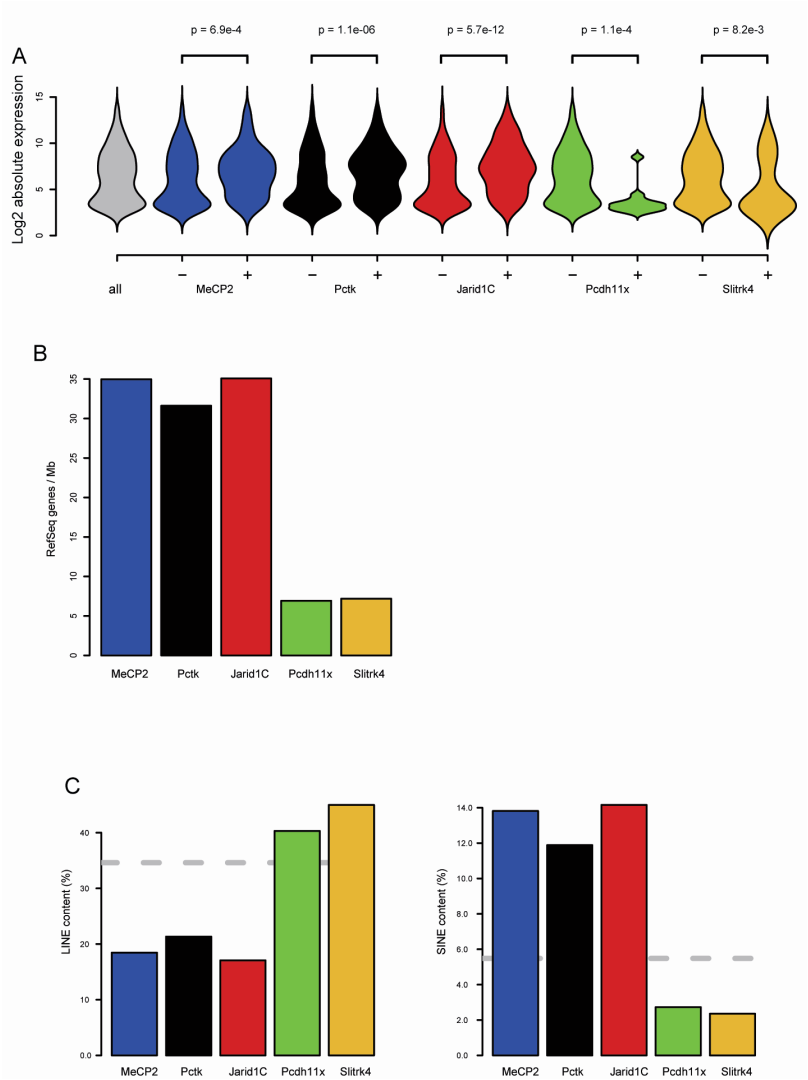


Figure S3: Characterization of X_a interacting regions. Active genes located in gene-dense regions spatially separate from inactive genes located in gene-poor areas on the active X chromosome. Preferred interacting regions of the different viewpoints on X_a are interrogated for gene activity (**A**). Violin plots depict the distribution of normalized absolute expression of all genes on the X chromosome (grey) and regions that were not contacted (-) and contacted (+) respectively by the different viewpoints indicated. The width of the bar relates to the number of data points present at that position. P-values were calculated using a Wilkinson test. Average gene density present in the *cis* contacted regions is plotted in **B**. C*is* interacting regions of the different viewpoints were also queried for LINE and SINE content (**C**). The dashed line represents the average content of the interrogated repeat elements on the X chromosome. On the X_i little interacting regions were identified due to the lack of clustered 4C-captured-fragment-ends.

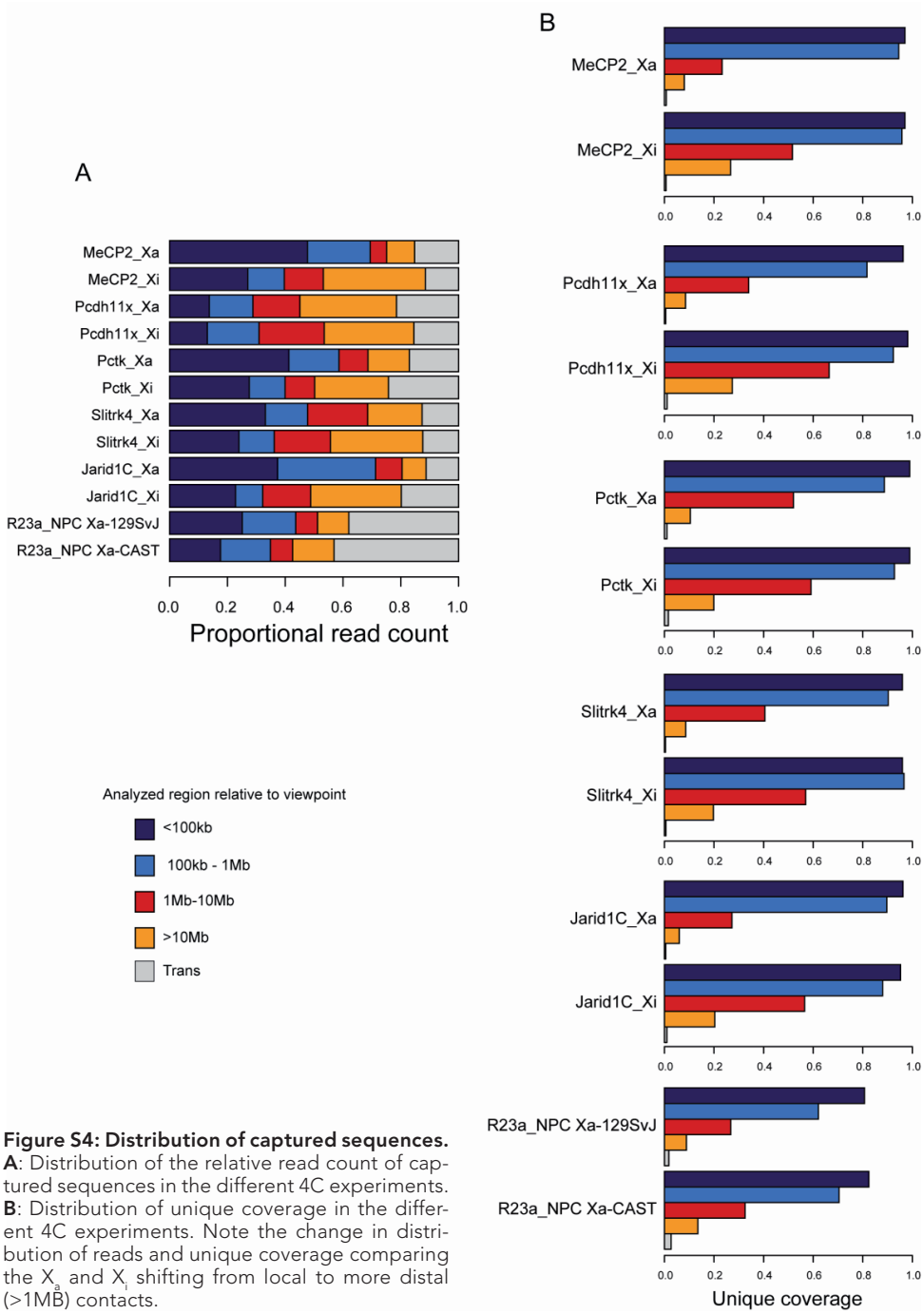
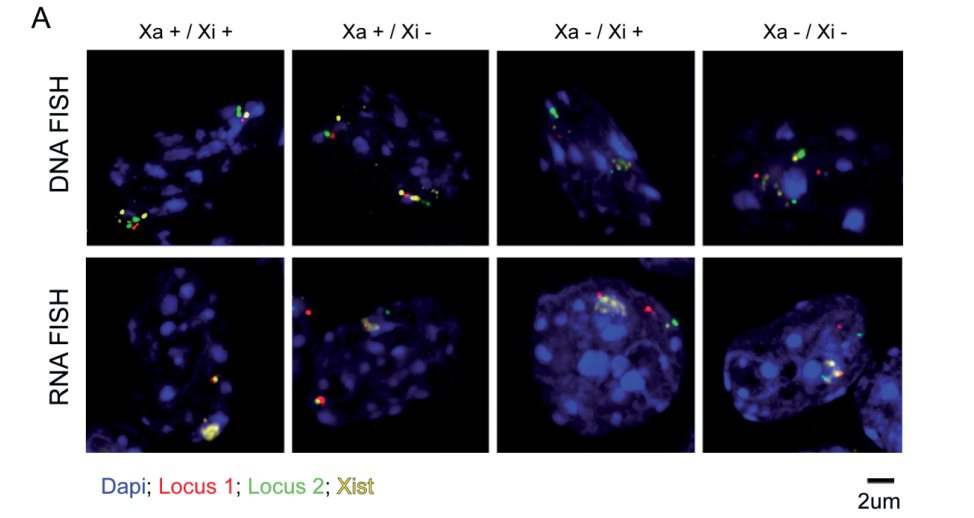
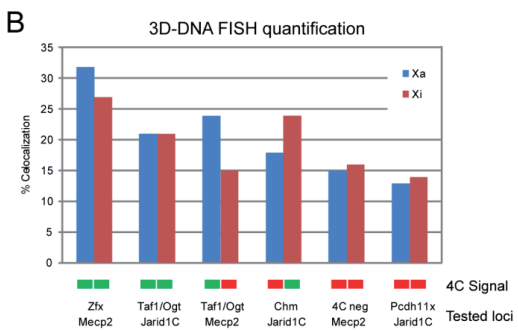
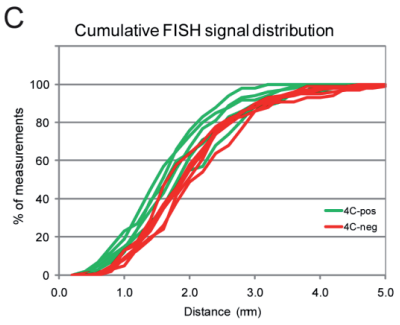
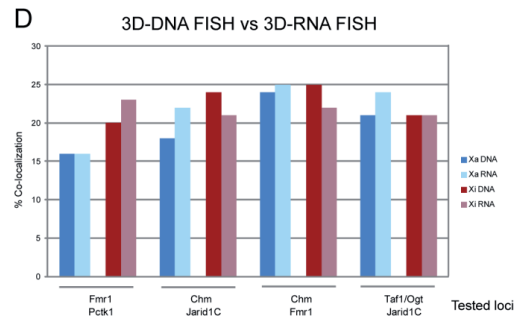


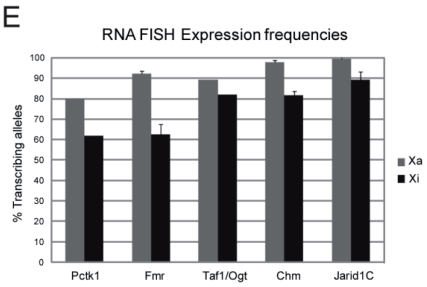
Figure S4: Distribution of captured sequences. **A**: Distribution of the relative read count of captured sequences in the different 4C experiments. **B**: Distribution of unique coverage in the different 4C experiments. Note the change in distribution of reads and unique coverage comparing the X_a and X_i shifting from local to more distal (>1MB) contacts.











Locus	BAC	Position
Pctk1	RP23-362P12	(20.2MB)
Fmr	RP24-183G11	(64.9MB)
MeCP2	RP23-77L16	(71.3MB)
4C neg	RP23-25P21	(87.2MB)
Zfx	RP23-269L6	(91.3MB)
Taf1/Ogt	RP23-268G11	(97.8MB)
Chm	RP23-118M16	(110.2MB)
Pcdh11x	RP23-20G23	(117.5MB)
Jarid1C	RP24-148H21	(148.7MB)

Figure S5: 4C interactions are confirmed by 3D-DNA FISH and comparing 3D-RNA FISH with 3D-DNA FISH suggest co-localization of genes is not mediated by active transcription. Co-localizing frequencies were determined by performing 3D distance measurements on 100 nuclei. Loci were scored co-localizing when the center of both signals was detected within 1µm apart. Only cells containing four signals with clear separation between spots

Continued on next page

▽

belonging to the X_a or X_i (indicated by Xist RNA staining; BAC probe: CT7-399K20) were analyzed. Co-localizing frequencies were calculated using the total number of measurements taken within each experiment. The BACs, representing the different loci used in the FISH experiments including their position on the X chromosome, are shown below panel E. **A:** Representative examples of DNA and RNA FISH experiments showing co-localization of loci on both X chromosomes, on the X_a, on the X_i or non co-localizing loci respectively. For discrimination between X_a and X_i a BAC probe spanning *Xist* was added to the experiment. Although in the DNA FISH the detection of the X_i resulted in a dotted appearance of the Xist RNA cloud, this did not affect its detection. **B:** DNA FISH co-localizing percentages are plotted for the different combinations of loci indicated. Below the graph the green and red bars indicate if corresponding loci were scored interacting or non-interacting by 4C. Co-localizing percentages for non interacting 4C sequences were found to be lower than 4C interacting sequences, although separated further apart on the linear chromosome (compare *Pcdh11x-Jarid1C* and *Taf1/Ogt-Jarid1C*; 4Cneg-*MeCP2* and *Zfx-MeCP2*). **C**: Depicting the cumulative FISH signal distribution of the experiments shown in (B) further supports this observation. **D**: If co-localization would only occur when two genes are actively transcribed, co-localizing percentages are expected to increase focusing only on cells actively transcribing those two genes. This is tested by comparing co-localizing frequencies found in DNA FISH (interrogation of loci, irrespective of gene activity) to frequencies found in RNA FISH (only actively transcribed loci will show a primary transcript signal). Importantly, interrogated genes were not active in all cells (**E**). Comparing RNA and DNA FISH co-localizing frequencies no difference could be observed, even not for genes expressed in as few as 60% of the cells.

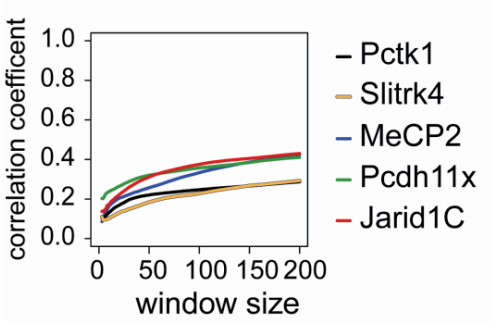


Figure S6: Spearman rank correlation analysis, comparing the X_i and X_a for the same locus (as indicated) revealed that the *cis*-interaction profiles of escapees were as dissimilar as those of the non-escaping genes.

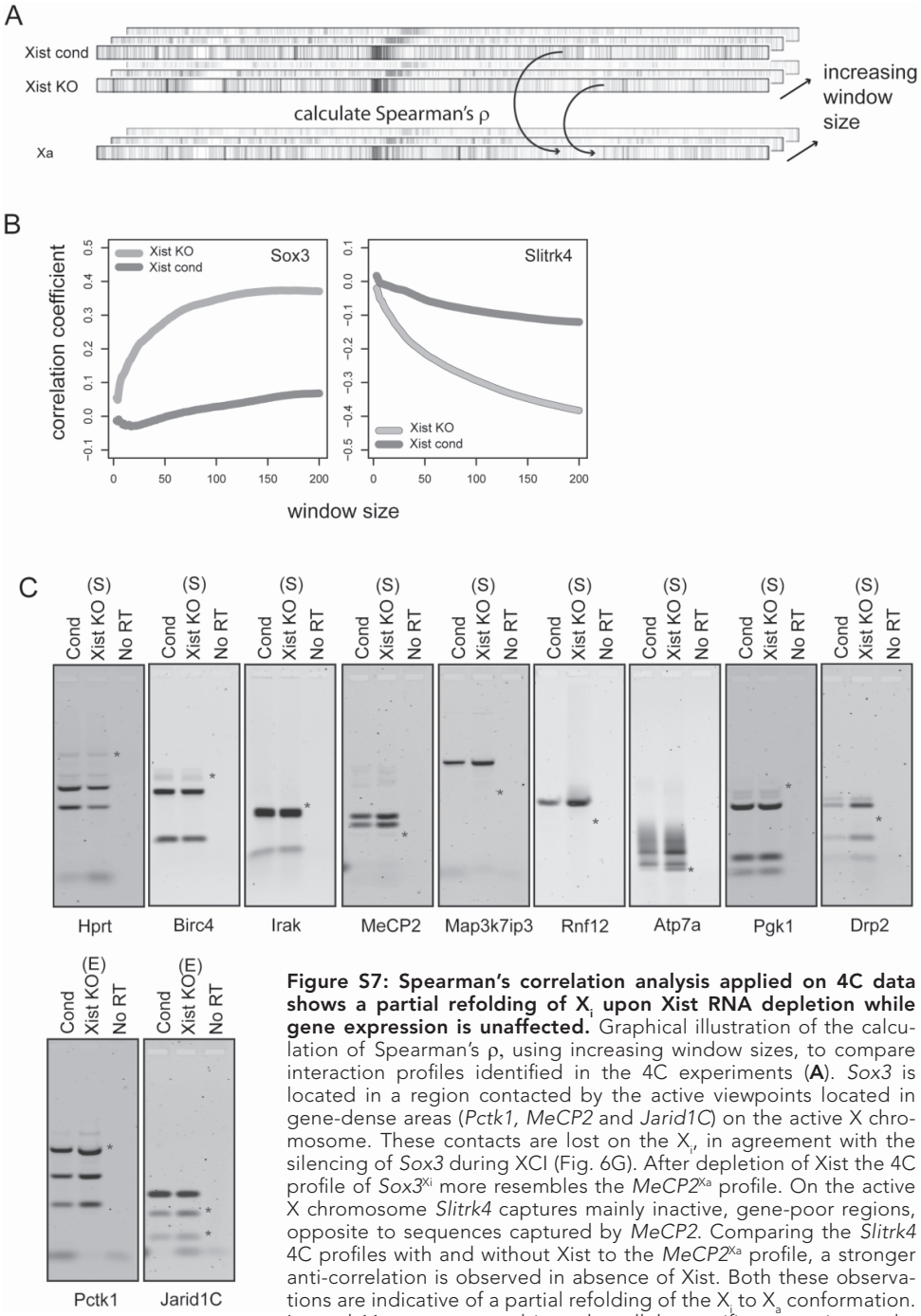


Figure S7: Spearman's correlation analysis applied on 4C data shows a partial refolding of X_i upon Xist RNA depletion while gene expression is unaffected. Graphical illustration of the calculation of Spearman's ρ, using increasing window sizes, to compare interaction profiles identified in the 4C experiments (**A**). *Sox3* is located in a region contacted by the active viewpoints located in gene-dense areas (*Pctk1, MeCP2* and *Jarid1C*) on the active X chromosome. These contacts are lost on the X_i, in agreement with the silencing of *Sox3* during XCI (Fig. 6G). After depletion of Xist the 4C profile of *Sox3*^{Xi} more resembles the *MeCP2*^{Xa} profile. On the active X chromosome *Slitrk4* captures mainly inactive, gene-poor regions, opposite to sequences captured by *MeCP2*. Comparing the *Slitrk4* 4C profiles with and without Xist to the *MeCP2*^{Xa} profile, a stronger anti-correlation is observed in absence of Xist. Both these observations are indicative of a partial refolding of the X_i to X_a conformation. In total 11 genes were subjected to allele-specific expression analysis, 9 silenced (S) and 2 escaping XCI (E). The position of X_i originating transcript signals, indicative of escape, is indicated by the asterisk. Of all genes tested only *Atp7a* was found to show de-repression, indicating most genes on X_i remain stably repressed upon *Xist* deletion.

SUMMARY

All genetic information required for the development and functioning of an organism is stored in the DNA which in eukaryotes is organized in chromosomes. Within these chromosomes genes are located, encoding the functional biomolecules that are responsible for executing cellular functions. A human body consists over 200 different cell types, all containing the same genetic information, but different in function and morphology. Gene regulatory programs are in place that drive the specification of different cell types during development. But also during the later stages of life correct gene expression is critical as misregulation can lead to diseases such as cancer. Therefore, mechanisms that regulate gene expression are topic of investigation.

Genes are regulated by various *cis*-regulatory elements, like promoter, enhancers and insulators. Some of these elements are located away from the genes they control raising the question how these elements function over distance. The three-dimensional conformation of chromatin in the nucleus appears to play a role in this. Over the past years, intense efforts were made to improve the methodologies for studying genome structure. New techniques have allowed obtaining unique insight into DNA topology, the factors that shape our genome and the impact structure has on function of the genome.

The development of new molecular technologies, especially that of Chromosome Conformation Capture (3C) and derivative methods, has greatly contributed to the current knowledge on the relation between chromosome folding and gene transcription. 3C allows the investigation of the spatial organization of individual loci at high resolution, but requires many controls and considerations that need to be taken into account before the results can be interpreted. A detailed description of 3C technology and the necessary controls required when applying this method to a mammalian model system have been provided in chapter 3. Chapter 4 describes the 4C methodology, which can be described as a high-throughput version of 3C (Simonis et al. 2006; Zhao et al. 2006). This technology allows the identification of the genomic environment of a locus of choice. Originally employing dedicated micro-array chips to identify co-localizing sequences, currently the technology more relies on novel next generation sequencing strategies. The application of sequencing is more cost effective and provides an unlimited range of detection, solving previous saturation issues. But there are more advantages; contrary to micro-array, sequencing analysis is not limited to the selection of probes present on the micro-array. Thereby the resolution of the technology is only limited by the mappability of analyzed DNA fragments. When analyzing small DNA fragments, this allows the detection of precise genomic contacts between cis-regulatory elements. Application of this technology, named highres-4C-seq, in future experiments will further expand our knowledge on gene regulatory networks and identify the regulatory elements involved.

In chapter 5 and 6 the murine β-globin locus was used as a model system. The β-globin locus encodes the β-subunits of the haemoglobin molecule that is responsible for oxygen transportation within red blood cells. Next to the globin genes this locus consist of several well characterized flanking *cis*-regulatory elements of which the locus control region (LCR) is most important for regulating globin transcription. Other sites flanking the locus involve those binding the protein CTCF, a factor known for its insulator capacity. These CTCF binding sites were

shown to cluster in erythroid progenitor cells (Palstra et al. 2003) and chapter 5 describes experiments investigating the role of CTCF in mediating these interactions. We showed that depletion of the CTCF protein and targeted mutation of the 3'HS1 CTCF binding site resulted in a reduction of long-range interaction between the affected CTCF sites. However, a functional role of these loops in the regulation of β-globin gene expression could not be detected at the time. In a follow-up study, we therefore investigated the effects of aberrant loop formation in more detail. In this study, presented in chapter 6, highres-4C-seq was applied on cells with the wild-type and cells with the mutant β-globin locus that lacked a CTCF binding site. We demonstrated that CTCF loops directs the spatial contacts of sequences within the loops to preferentially contact each other while showing reduced contact frequencies with sites outside the loop. Mutating the CTCF binding site and disrupting the loop was now found to cause preferential upregulation of the proximal embryonic gene by the LCR, at the expense of the more distal adult globin genes. Although preliminary, these results show that these CTCF mediated loops at later stages of development help directing the LCR to the distal adult globin genes. We predict that loop formation is a common mechanism by which contacts between regulatory elements are regulated. The position of these elements on the loops and their affinity for their targets will eventually determine the effect on gene expression. In agreement, besides the β-globin locus, also other developmentally regulated multi-gene loci were reported to be organized in spatial interaction domains, suggesting this type of organization can be more commonly found within the genome.

Little is known about the molecular details on how CTCF is able to mediate chromatin loops. CTCF could execute this function independent of other factors as it was shown to be able to homo-dimerize (Yusufzai et al. 2004). Another option includes the recruitment of co-factors, like cohesin that was shown to be recruited to the DNA via CTCF (Rubio et al. 2008). Cohesin, classically linked to sister chromatid cohesion during mitosis, was recently shown to play a role in mediating chromatin looping providing support for this hypothesis (Kagey et al. 2010).

Besides CTCF, SatBI (special AT-rich binding protein 1) has been implicated in DNA loop formation. Two studies performed at the Th2 and MHC class I locus in thymocytes, the cell type where SatBI is highest expressed, were able to show a role for SatBI in mediating long-range chromatin interactions that were required for correct gene expression (Cai et al. 2006; Kumar et al. 2007). The ability of SatBI to mediate chromatin looping also in other cell types was shown recently at the BCL2 locus, where the role of SatBI loop formation was investigated in Jurkat cells (Gong et al. 2011). It will be interesting to see if also other proteins are involved in dictating chromatin loops and reveal the mechanisms by which they are regulated.

In chapter 7 experiments are described investigating nuclear organization on a chromosome wide scale. Here the X-chromosome was selected as a model system to study the relationship between gene expression and DNA organization. In female mammalian cells one X-chromosome is silenced to achieve dosage compensation. Provided one can separate between the interactions of the active and inactive X-chromosome, this allows the investigation of the folding of two

identical chromosomes that are differentially treated within the same cell. By employing single nucleotide polymorphisms we developed such allele specific 4C method that was used to investigate the active and inactive X.

The X inactivation process starts by the upregulation of *Xist* on the future inactive X-chromosome upon differentiation of embryonic stem cells. Xist encodes a noncoding RNA that subsequently spreads across the X-chromosome *in cis* and induces silencing possibly via the recruitment of various silencing factors, like polycomb group proteins (PcG) and macro-H2A, which stably establish gene silencing. Once established, Xist RNA is no longer required for maintaining the silencing.

Studying the conformation of the X chromosomes we found on the active X a segregation of active-gene dense regions and inactive-gene poor regions, an organization that was previously identified also on other chromosomes. The inactive X however, showed an a-typical organization, as preferred contacts between inactive sequences were limited. This was not due to a lack of contacts, as many were detected, but due to their seemingly randomly distribution across the chromosome. An exception was found for the genes escaping the X inactivation process, as these were found to be able to contact each other. This ability was readily picked up by the 4C interaction profiles, allowing the identification of novel escapees in the process. Among the clustered escapees is the *Xist* gene, which is remarkable as this is the site where the noncoding Xist RNA is produced that is responsible for the silencing of the X-chromosome. Possibly the localization of escaping genes near *Xist* could facilitate the spreading of Xist RNA across the chromosome, a mechanism seen for the spreading of the dosage compensation complex (DCC) in *Drosophila* (Grimaud and Becker 2009).

The folding of the inactive X depends on the presence of the Xist RNA, because when depleted the inactive X folds into a structure reminiscent of the active X chromosome. This observation has multiple implications. First, as the re-folding of the inactive X occurred in the absence of re-activation of silenced genes, it shows the specific folding pattern on the inactive X is not required to maintain X chromosome silencing. Secondly, because the inactive X is not re-activated this result shows that active transcription is not required to mediate long-range contacts. The results therefore challenge a currently popular model that states that genes need to come together in specialized 'transcription factories' for their expression. Although we cannot exclude an increased affinity of the inactive X chromosome for the transcription machinery upon Xist depletion, we do show that chromosome topology is dictated also by factors other than the active transcription machinery. It will be interesting to determine what factors these are and how they implicate in chromosome folding.

REFERENCES

Ahituv, N., Zhu, Y., Visel, A., Holt, A., Afzal, V., Pennacchio, L.A., and Rubin, E.M. 2007. Deletion of ultraconserved elements yields viable mice. *PLoS Biol* **5**(9): e234.

Amano, T., Sagai, T., Tanabe, H., Mizushina, Y., Nakazawa, H., and Shiroishi, T. 2009. Chromosomal dynamics at the Shh locus: limb bud-specific differential regulation of competence and active transcription. *Dev Cell* **16**(1): 47-57.

Banerji, J., Rusconi, S., and Schaffner, W. 1981. Expression of a beta-globin gene is enhanced by remote SV40 DNA sequences. *Cell* **27**(2 Pt 1): 299-308.

Bantignies, F., Roure, V., Comet, I., Leblanc, B., Schuettengruber, B., Bonnet, J., Tixier, V., Mas, A., and Cavalli, G. 2011. Polycomb-dependent regulatory contacts between distant Hox loci in Drosophila. *Cell* **144**(2): 214-226.

Barakat, T.S. and Gribnau, J. 2010. X chromosome inactivation and embryonic stem cells. *Adv Exp Med Biol* **695**: 132-154.

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**(4): 823-837.

Bau, D., Sanyal, A., Lajoie, B.R., Capriotti, E., Byron, M., Lawrence, J.B., Dekker, J., and Marti-Renom, M.A. 2010. The three-dimensional folding of the alpha-globin gene domain reveals formation of chromatin globules. *Nat Struct Mol Biol* **18**(1): 107-114.

Bell, A.C. and Felsenfeld, G. 1999. Stopped at the border: boundaries and insulators. *Curr Opin Genet Dev* **9**(2): 191-198.

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

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

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**(4): 1395-1401.

Blanton, J., Gaszner, M., and Schedl, P. 2003. Protein:protein interactions and the pairing of boundary elements in vivo. *Genes Dev* **17**(5): 664-675.

Bolstad, B.M., Irizarry, R.A., Astrand, M., and Speed, T.P. 2003. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* **19**(2): 185-193.

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**(5): e157.

Bourgeois, C.A., Laquerriere, F., Hemon, D., Hubert, J., and Bouteille, M. 1985. New data on the in-situ position of the inactive X chromosome in the interphase nucleus of human fibroblasts. *Hum Genet* **69**(2): 122-129.

Bowman, G.D. 2010. Mechanisms of ATP-dependent nucleosome sliding. *Curr Opin Struct Biol* **20**(1): 73-81.

Boyle, A.P., Davis, S., Shulha, H.P., Meltzer, P., Margulies, E.H., Weng, Z., Furey, T.S., and Crawford, G.E. 2008. High-resolution mapping and characterization of open chromatin across the genome. *Cell* **132**(2): 311-322.

Branco, M.R. and Pombo, A. 2006. Intermingling of chromosome territories in interphase suggests role in translocations and transcription-dependent associations. *PLoS Biol* **4**(5): e138.

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**(6): 1083-1097.

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. *Mol Cell Biol* **23**(15): 5234-5244.

Butler, J.E. and Kadonaga, J.T. 2001. Enhancer-promoter specificity mediated by DPE or TATA core promoter motifs. *Genes Dev* **15**(19): 2515-2519.

Cai, H.N. and Shen, P. 2001. Effects of cis arrangement of chromatin insulators on enhancer-blocking activity. *Science* **291**(5503): 493-495.

Cai, S., Lee, C.C., and Kohwi-Shigematsu, T. 2006. SATB1 packages densely looped, transcriptionally active chromatin for coordinated expression of cytokine genes. *Nat Genet* **38**(11): 1278-1288.

Cajiao, I., Zhang, A., Yoo, E.J., Cooke, N.E., and Liebhaber, S.A. 2004. Bystander gene activation by a locus control region. *EMBO J* **23**(19): 3854-3863.

Carmo-Fonseca, M., Mendes-Soares, L., and Campos, I. 2000. To be or not to be in the nucleolus. *Nat Cell Biol* **2**(6): E107-112.

Caron, H., van Schaik, B., van der Mee, M., Baas, F., Riggins, G., van Sluis, P., Hermus, M.C., van Asperen, R., Boon, K., Voute, P.A. et al. 2001. The human transcriptome map: clustering of highly expressed genes in chromosomal domains. *Science* **291**(5507): 1289-1292.

Carotta, S., Pilat, S., Mairhofer, A., Schmidt, U., Dolznig, H., Steinlein, P., and Beug, H. 2004. Directed differentiation and mass cultivation of pure erythroid progenitors from mouse embryonic stem cells. *Blood* **104**(6): 1873-1880.

Carrel, L. and Willard, H.F. 2005. X-inactivation profile reveals extensive variability in X-linked gene expression in females. *Nature* **434**(7031): 400-404.

Carter, D., Chakalova, L., Osborne, C.S., Dai, Y.F., and Fraser, P. 2002. Long-range chromatin regulatory interactions in vivo. *Nat Genet* **32**(4): 623-626.

Chakalova, L. and Fraser, P. 2010. Organization of transcription. *Cold Spring Harb Perspect Biol* **2**(9): a000729.

Changolkar, L.N., Singh, G., Cui, K., Berletch, J.B., Zhao, K., Disteche, C.M., and Pehrson, J.R. 2010. Genome-wide distribution of macroH2A1 histone variants in mouse liver chromatin. *Mol Cell Biol* **30**(23): 5473-5483.

Chao, W., Huynh, K.D., Spencer, R.J., Davidow, L.S., and Lee, J.T. 2002. CTCF, a candidate trans-acting factor for X-inactivation choice. *Science* **295**(5553): 345-347.

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 Dev* **20**(16): 2223-2237.

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**(6): 1106-1117.

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**(20): 17870-17878.

Cho, D.H., Thienes, C.P., Mahoney, S.E., Analau, E., Filippova, G.N., and Tapscott, S.J. 2005. Antisense transcription and heterochromatin at the DM1 CTG repeats are constrained by CTCF. *Mol Cell* **20**(3): 483-489.

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**(6): 439-445.

Chubb, J.R., Trcek, T., Shenoy, S.M., and Singer, R.H. 2006. Transcriptional pulsing of a developmental gene. *Curr Biol* **16**(10): 1018-1025.

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**(3): 505-514.

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. *Proc Natl Acad Sci U S A* **103**(20): 7688-7693.

Comet, I., Savitskaya, E., Schuettengruber, B., Negre, N., Lavrov, S., Parshikov, A., Juge, F., Gracheva, E., Georgiev, P., and Cavalli, G. 2006. PRE-mediated bypass of two Su(Hw) insulators targets PcG proteins to a downstream promoter. *Dev Cell* **11**(1): 117-124.

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. *Proc Natl Acad Sci U S A* **108**(6): 2294-2299.

Conti, L., Pollard, S.M., Gorba, T., Reitano, E., Toselli, M., Biella, G., Sun, Y., Sanzone, S., Ying, Q.L., Cattaneo, E. et al. 2005. Niche-independent symmetrical self-renewal of a mammalian tissue stem cell. *PLoS Biol* **3**(9): e283.

Cook, P.R. 2010. A model for all genomes: the role of transcription factories. *J Mol Biol* **395**(1): 1-10.

Crawford, G.E., Holt, I.E., Mullikin, J.C., Tai, D., Blakesley, R., Bouffard, G., Young, A., Masiello, C., Green, E.D., Wolfsberg, T.G. et al. 2004. Identifying gene regulatory elements by genome-wide recovery of DNase hypersensitive sites. *Proc Natl Acad Sci U S A* **101**(4): 992-997.

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 Res* **9**(7): 541-567.

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. *Proc Natl Acad Sci U S A*.

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. *J Cell Biol* **145**(6): 1119-1131.

Csankovszki, G., Nagy, A., and Jaenisch, R. 2001. Synergism of Xist RNA, DNA methylation, and histone hypoacetylation in maintaining X chromosome inactivation. *J Cell Biol* **153**(4): 773-784.

Csankovszki, G., Panning, B., Bates, B., Pehrson, J.R., and Jaenisch, R. 1999. Conditional deletion of Xist disrupts histone macroH2A localization but not maintenance of X inactivation. *Nat Genet* **22**(4): 323-324.

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 Res* **19**(1): 24-32.

de Laat, W. and Grosveld, F. 2003. Spatial organization of gene expression: the active chromatin hub. *Chromosome Res* **11**(5): 447-459.

de Laat, W. and Grosveld, F. 2007. Inter-chromosomal gene regulation in the mammalian cell nucleus. *Curr Opin Genet Dev* **17**(5): 456-464.

de Villiers, J., Olson, L., Banerji, J., and Schaffner, W. 1983. Analysis of the transcriptional enhancer effect. *Cold Spring Harb Symp Quant Biol* **47 Pt 2**: 911-919.

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**(3): e1000045.

de Wit, E. and van Steensel, B. 2009. Chromatin domains in higher eukaryotes: insights from genome-wide mapping studies. *Chromosoma* **118**(1): 25-36.

Deaton, A.M. and Bird, A. 2011. CpG islands and the regulation of transcription. *Genes Dev* **25**(10): 1010-1022.

Defossez, P.A. and Gilson, E. 2002. The vertebrate protein CTCF functions as an insulator in Saccharomyces cerevisiae. *Nucleic Acids Res* **30**(23): 5136-5141.

Dekker, J., Rippe, K., Dekker, M., and Kleckner, N. 2002. Capturing chromosome conformation. *Science* **295**(5558): 1306-1311.

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. *Exp Cell Res* **252**(2): 363-375.

Dillon, N. and Sabbattini, P. 2000. Functional gene expression domains: defining the functional unit of eukaryotic gene regulation. *Bioessays* **22**(7): 657-665.

Dillon, N., Trimborn, T., Strouboulis, J., Fraser, P., and Grosveld, F. 1997. The effect of distance on long-range chromatin interactions. *Mol Cell* **1**(1): 131-139.

Disteche, C.M. 1995. Escape from X inactivation in human and mouse. *Trends Genet* **11**(1): 17-22.

Dolznig, H., Boulme, F., Stangl, K., Deiner, E.M., Mikulits, W., Beug, H., and Mullner, E.W. 2001. Establishment of normal, terminally differentiating mouse erythroid progenitors: molecular characterization by cDNA arrays. *FASEB J* **15**(8): 1442-1444.

Dorschner, M.O., Hawrylycz, M., Humbert, R., Wallace, J.C., Shafer, A., Kawamoto, J., Mack, J., Hall, R., Goldy, J., Sabo, P.J. et al. 2004. High-throughput localization of functional elements by quantitative chromatin profiling. *Nat Methods* **1**(3): 219-225.

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**(10): 1299-1309.

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**(20): 2485-2490.

Droge, P. and Muller-Hill, B. 2001. High local protein concentrations at promoters: strategies in prokaryotic and eukaryotic cells. *Bioessays* **23**(2): 179-183.

Duan, Z., Andronescu, M., Schutz, K., McIlwain, S., Kim, Y.J., Lee, C., Shendure, J., Fields, S., Blau, C.A., and Noble, W.S. 2010. A three-dimensional model of the yeast genome. *Nature* **465**(7296): 363-367.

Dunn, K.L., Zhao, H., and Davie, J.R. 2003. The insulator binding protein CTCF associates with the nuclear matrix. *Exp Cell Res* **288**(1): 218-223.

Epner, E., Reik, A., Cimbora, D., Telling, A., Bender, M.A., Fiering, S., Enver, T., Martin, D.I., Kennedy, M., Keller, G. et al. 1998. The beta-globin LCR is not necessary for an open chromatin structure or developmentally regulated transcription of the native mouse beta-globin locus. *Mol Cell* **2**(4): 447-455.

Eskeland, R., Leeb, M., Grimes, G.R., Kress, C., Boyle, S., Sproul, D., Gilbert, N., Fan, Y., Skoultchi, A.I., Wutz, A. et al. 2010. Ring1B compacts chromatin structure and represses gene expression independent of histone ubiquitination. *Mol Cell* **38**(3): 452-464.

Farrell, C.M., West, A.G., and Felsenfeld, G. 2002. Conserved CTCF insulator elements flank the mouse and human beta-globin loci. *Mol Cell Biol* **22**(11): 3820-3831.

Fedoriw, A.M., Stein, P., Svoboda, P., Schultz, R.M., and Bartolomei, M.S. 2004. Transgenic RNAi reveals essential function for CTCF in H19 gene imprinting. *Science* **303**(5655): 238-240.

Filippova, G.N., Cheng, M.K., Moore, J.M., Truong, J.P., Hu, Y.J., Nguyen, D.K., Tsuchiya, K.D., and Disteche, C.M. 2005. Boundaries between chromosomal domains of X inactivation and escape bind CTCF and lack CpG methylation during early development. *Dev Cell* **8**(1): 31-42.

Filippova, G.N., Thienes, C.P., Penn, B.H., Cho, D.H., Hu, Y.J., Moore, J.M., Klesert, T.R., Lobanenkov, V.V., and Tapscott, S.J. 2001. CTCF-binding sites flank CTG/CAG repeats and form a methylation-sensitive insulator at the DM1 locus. *Nat Genet* **28**(4): 335-343.

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**(3): e1000039.

Flicek, P., Amode, M.R., Barrell, D., Beal, K., Brent, S., Chen, Y., Clapham, P., Coates, G., Fairley, S., and Fitzgerald, S. 2011. Ensembl 2011. *Nucleic acids research* **39**(suppl 1): D800.

Follenzi, A., Sabatino, G., Lombardo, A., Boccaccio, C., and Naldini, L. 2002. Efficient gene delivery and targeted expression to hepatocytes in vivo by improved lentiviral vectors. *Hum Gene Ther* **13**(2): 243-260.

Forrester, W.C., Epner, E., Driscoll, M.C., Enver, T., Brice, M., Papayannopoulou, T., and Groudine, M. 1990. A deletion of the human beta-globin locus activation region causes a major alteration in chromatin structure and replication across the entire beta-globin locus. *Genes Dev* **4**(10): 1637-1649.

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

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**(7157): 1050-1053.

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

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**(7269): 58-64.

Fussner, E., Ching, R.W., and Bazett-Jones, D.P. 2011. Living without 30nm chromatin fibers. *Trends Biochem Sci* **36**(1): 1-6.

Gerasimova, T.I., Byrd, K., and Corces, V.G. 2000. A chromatin insulator determines the nuclear localization of DNA. *Mol Cell* **6**(5): 1025-1035.

Gerlich, D., Beaudouin, J., Kalbfuss, B., Daigle, N., Eils, R., and Ellenberg, J. 2003. Global chromosome positions are transmitted through mitosis in mammalian cells. *Cell* **112**(6): 751-764.

Gilbert, D.M. 2010. Evaluating genome-scale approaches to eukaryotic DNA replication. *Nat Rev Genet* **11**(10): 673-684.

Gilbert, D.M., Takebayashi, S.I., Ryba, T., Lu, J., Pope, B.D., Wilson, K.A., and Hiratani, I. 2010. Space and time in the nucleus: developmental control of replication timing and chromosome architecture. *Cold Spring Harb Symp Quant Biol* **75**: 143-153.

Giresi, P.G., Kim, J., McDaniell, R.M., Iyer, V.R., and Lieb, J.D. 2007. FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) isolates active regulatory elements from human chromatin. *Genome Res* **17**(6): 877-885.

Gong, F., Sun, L., Wang, Z., Shi, J., Li, W., Wang, S., Han, X., and Sun, Y. 2011. The BCL2 gene is regulated by a special AT-rich sequence binding protein 1-mediated long range chromosomal interaction between the promoter and the distal element located within the 3'-UTR. *Nucleic Acids Res*.

Grimaud, C. and Becker, P.B. 2009. The dosage compensation complex shapes the conformation of the X chromosome in Drosophila. *Genes Dev* **23**(21): 2490-2495.

Grimaud, C. and Becker, P.B. 2010. Form and function of dosage-compensated chromosomes--a chicken-and-egg relationship. *Bioessays* **32**(8): 709-717.

Grosveld, F., van Assendelft, G.B., Greaves, D.R., and Kollias, G. 1987. Position-independent, high-level expression of the human beta-globin gene in transgenic mice. *Cell* **51**(6): 975-985.

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**(7197): 948-951.

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**(7253): 410-413.

Hakim, O., John, S., Ling, J.Q., Biddie, S.C., Hoffman, A.R., and Hager, G.L. 2009. Glucocorticoid receptor activation of the Ciz1-Lcn2 locus by long range interactions. *J Biol Chem* **284**(10): 6048-6052.

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**(7): 630-638.

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**(8): 1387-1394.

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**(6785): 486-489.

Heintzman, N.D., Hon, G.C., Hawkins, R.D., Kheradpour, P., Stark, A., Harp, L.F., Ye, Z., Lee, L.K., Stuart, R.K., Ching, C.W. et al. 2009. Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature* **459**(7243): 108-112.

Heintzman, N.D., Stuart, R.K., Hon, G., Fu, Y., Ching, C.W., Hawkins, R.D., Barrera, L.O., Van Calcar, S., Qu, C., Ching, K.A. et al. 2007. Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat Genet* **39**(3): 311-318.

Heuchel, R., Matthias, P., and Schaffner, W. 1989. Two closely spaced promoters are equally activated by a remote enhancer: evidence against a scanning model for enhancer action. *Nucleic Acids Res* **17**(22): 8931-8947.

Hindorff, L.A., Junkins, H.A., Mehta, J.P., and Manolio, T.A. 2009. A catalog of published genome-wide association studies. *Available at: <wwwgenomegov/gwastudies>*.

Horike, S., Cai, S., Miyano, M., Cheng, J.F., and Kohwi-Shigematsu, T. 2005. Loss of silent-chromatin looping and impaired imprinting of DLX5 in Rett syndrome. *Nat Genet* **37**(1): 31-40.

Horn, P.J. and Peterson, C.L. 2002. Molecular biology. Chromatin higher order folding--wrapping up transcription. *Science* **297**(5588): 1824-1827.

Horowitz-Scherer, R.A. and Woodcock, C.L. 2006. Organization of interphase chromatin. *Chromosoma* **115**(1): 1-14.

Hou, C., Zhao, H., Tanimoto, K., and Dean, A. 2008. CTCF-dependent enhancer-blocking by alternative chromatin loop formation. *Proc Natl Acad Sci U S A* **105**(51): 20398-20403.

Huang da, W., Sherman, B.T., and Lempicki, R.A. 2009. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc* **4**(1): 44-57.

Huynh, K.D. and Lee, J.T. 2003. Inheritance of a pre-inactivated paternal X chromosome in early mouse embryos. *Nature* **426**(6968): 857-862.

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.

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**(6): 1523-1536.

Jackson, V. 1999. Formaldehyde cross-linking for studying nucleosomal dynamics. *Methods* **17**(2): 125-139.

Jones, P.A. and Takai, D. 2001. The role of DNA methylation in mammalian epigenetics. *Science* **293**(5532): 1068-1070.

Jonkers, I., Monkhorst, K., Rentmeester, E., Grootegoed, J.A., Grosveld, F., and Gribnau, J. 2008. Xist RNA is confined to the nuclear territory of the silenced X chromosome throughout the cell cycle. *Mol Cell Biol* **28**(18): 5583-5594.

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**(7314): 430-435.

Kellum, R. and Schedl, P. 1991. A position-effect assay for boundaries of higher order chromosomal domains. *Cell* **64**(5): 941-950.

Kellum, R. and Schedl, P. 1992. A group of scs elements function as domain boundaries in an enhancer-blocking assay. *Mol Cell Biol* **12**(5): 2424-2431.

Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringle, T.H., Zahler, A.M., and Haussler, D. 2002. The human genome browser at UCSC. *Genome Res* **12**(6): 996-1006.

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**(6): 1231-1245.

Kimura, H., Sugaya, K., and Cook, P.R. 2002. The transcription cycle of RNA polymerase II in living cells. *J Cell Biol* **159**(5): 777-782.

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**(1): 8-32.

Kohlmaier, A., Savarese, F., Lachner, M., Martens, J., Jenuwein, T., and Wutz, A. 2004. A chromosomal memory triggered by Xist regulates histone methylation in X inactivation. *PLoS Biol* **2**(7): E171.

Kouzarides, T. 2007. Chromatin modifications and their function. *Cell* **128**(4): 693-705.

Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D., Jones, S.J., and Marra, M.A. 2009. Circos: an information aesthetic for comparative genomics. *Genome Res* **19**(9): 1639-1645.

Kumar, P.P., Bischof, O., Purbey, P.K., Notani, D., Urlaub, H., Dejean, A., and Galande, S. 2007. Functional interaction between PML and SATB1 regulates chromatin-loop architecture and transcription of the MHC class I locus. *Nat Cell Biol* **9**(1): 45-56.

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**(1): 51-65.

Kunarso, G., Chia, N.Y., Jeyakani, J., Hwang, C., Lu, X., Chan, Y.S., Ng, H.H., and Bourque, G. 2010. Transposable elements have rewired the core regulatory network of human embryonic stem cells. *Nat Genet* **42**(7): 631-634.

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**(28): 10684-10689.

Leach, K.M., Vieira, K.F., Kang, S.H., Aslanian, A., Teichmann, M., Roeder, R.G., and Bungert, J. 2003. Characterization of the human beta-globin downstream promoter region. *Nucleic Acids Res* **31**(4): 1292-1301.

Lee, J.M. and Sonnhammer, E.L. 2003. Genomic gene clustering analysis of pathways in eukaryotes. *Genome Res* **13**(5): 875-882.

Lettice, L.A., Heaney, S.J., Purdie, L.A., Li, L., de Beer, P., Oostra, B.A., Goode, D., Elgar, G., Hill, R.E., and de Graaff, E. 2003. A long-range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly. *Hum Mol Genet* **12**(14): 1725-1735.

Li, H., Ruan, J., and Durbin, R. 2008. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res* **18**(11): 1851-1858.

Li, N. and Carrel, L. 2008. Escape from X chromosome inactivation is an intrinsic property of the Jarid1c locus. *Proc Natl Acad Sci U S A* **105**(44): 17055-17060.

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**(5950): 289-293.

Ling, J.Q., Li, T., Hu, J.F., Vu, T.H., Chen, H.L., Qiu, X.W., Cherry, A.M., and Hoffman, A.R. 2006. CTCF mediates interchromosomal colocalization between Igf2/H19 and Wsb1/Nf1. *Science* **312**(5771): 269-272.

Lister, R., Pelizzola, M., Kida, Y.S., Hawkins, R.D., Nery, J.R., Hon, G., Antosiewicz-Bourget, J., O'Malley, R., Castanon, R., Klugman, S. et al. 2011. Hotspots of aberrant epigenomic reprogramming in human induced pluripotent stem cells. *Nature* **471**(7336): 68-73.

Liu, Z. and Garrard, W.T. 2005. Long-range interactions between three transcriptional enhancers, active Vkappa gene promoters, and a 3' boundary sequence spanning 46 kilobases. *Mol Cell Biol* **25**(8): 3220-3231.

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**(51): 21771-21776.

Luger, K., Mader, A.W., Richmond, R.K., Sargent, D.F., and Richmond, T.J. 1997. Crystal structure of the nucleosome core particle at 2.8 A resolution. *Nature* **389**(6648): 251-260.

Luikenhuis, S., Wutz, A., and Jaenisch, R. 2001. Antisense transcription through the Xist locus mediates Tsix function in embryonic stem cells. *Mol Cell Biol* **21**(24): 8512-8520.

Mahy, N.L., Perry, P.E., and Bickmore, W.A. 2002a. Gene density and transcription influence the localization of chromatin outside of chromosome territories detectable by FISH. *J Cell Biol* **159**(5): 753-763.

Mahy, N.L., Perry, P.E., Gilchrist, S., Baldock, R.A., and Bickmore, W.A. 2002b. Spatial organization of active and inactive genes and noncoding DNA within chromosome territories. *J Cell Biol* **157**(4): 579-589.

Maksimenko, O., Golovnin, A., and Georgiev, P. 2008. Enhancer-promoter communication is regulated by insulator pairing in a Drosophila model bigenic locus. *Mol Cell Biol* **28**(17): 5469-5477.

Mallin, D.R., Myung, J.S., Patton, J.S., and Geyer, P.K. 1998. Polycomb group repression is blocked by the Drosophila suppressor of Hairy-wing [su(Hw)] insulator. *Genetics* **148**(1): 331-339.

Martin, S. and Pombo, A. 2003. Transcription factories: quantitative studies of nanostructures in the mammalian nucleus. *Chromosome Res* **11**(5): 461-470.

Mikkelsen, T.S., Ku, M., Jaffe, D.B., Issac, B., Lieberman, E., Giannoukos, G., Alvarez, P., Brockman, W., Kim, T.K., Koche, R.P. et al. 2007. Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature* **448**(7153): 553-560.

Misteli, T. 2001. The concept of self-organization in cellular architecture. *J Cell Biol* **155**(2): 181-185.

Misteli, T. 2004. Spatial positioning; a new dimension in genome function. *Cell* **119**(2): 153-156.

Misteli, T. 2007. Beyond the sequence: cellular organization of genome function. *Cell* **128**(4): 787-800.

Morey, C., Da Silva, N.R., Kmita, M., Duboule, D., and Bickmore, W.A. 2008. Ectopic nuclear reorganisation driven by a Hoxb1 transgene transposed into Hoxd. *J Cell Sci* **121**(Pt 5): 571-577.

Muller, I., Boyle, S., Singer, R.H., Bickmore, W.A., and Chubb, J.R. 2010. Stable morphology, but dynamic internal reorganisation, of interphase human chromosomes in living cells. *PLoS One* **5**(7): e11560.

Muravyova, E., Golovnin, A., Gracheva, E., Parshikov, A., Belenkaya, T., Pirrotta, V., and Georgiev, P. 2001. Loss of insulator activity by paired Su(Hw) chromatin insulators. *Science* **291**(5503): 495-498.

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**(11): e1000739.

Naughton, C., Sproul, D., Hamilton, C., and Gilbert, N. 2010. Analysis of active and inactive X chromosome architecture reveals the independent organization of 30 nm and large-scale chromatin structures. *Mol Cell* **40**(3): 397-409.

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**(3): e1000016.

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*.

Nunez, E., Fu, X.D., and Rosenfeld, M.G. 2009. Nuclear organization in the 3D space of the nucleus - cause or consequence? *Curr Opin Genet Dev* **19**(5): 424-436.

Ohtsuki, S., Levine, M., and Cai, H.N. 1998. Different core promoters possess distinct regulatory activities in the Drosophila embryo. *Genes Dev* **12**(4): 547-556.

Orlando, V., Strutt, H., and Paro, R. 1997. Analysis of chromatin structure by in vivo formaldehyde cross-linking. *Methods* **11**(2): 205-214.

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**(10): 1065-1071.

Oudet, P., Gross-Bellard, M., and Chambon, P. 1975. Electron microscopic and biochemical evidence that chromatin structure is a repeating unit. *Cell* **4**(4): 281-300.

Pal, C. and Hurst, L.D. 2003. Evidence for co-evolution of gene order and recombination rate. *Nat Genet* **33**(3): 392-395.

Palmer, D.K., O'Day, K., Trong, H.L., Charbonneau, H., and Margolis, R.L. 1991. Purification of the centromere-specific protein CENP-A and demonstration that it is a distinctive histone. *Proc Natl Acad Sci U S A* **88**(9): 3734-3738.

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**(2): e1661.

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**(2): 190-194.

Papantonis, A., Larkin, J.D., Wada, Y., Ohta, Y., Ihara, S., Kodama, T., and Cook, P.R. 2010. Active RNA polymerases: mobile or immobile molecular machines? *PLoS Biol* **8**(7): e1000419.

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**(3): 422-433.

Pennings, S., Meersseman, G., and Bradbury, E.M. 1994. Linker histones H1 and H5 prevent the mobility of positioned nucleosomes. *Proc Natl Acad Sci U S A* **91**(22): 10275-10279.

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**(4): 603-613.

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

Pickersgill, H., Kalverda, B., de Wit, E., Talhout, W., Fornerod, M., and van Steensel, B. 2006. Characterization of the Drosophila melanogaster genome at the nuclear lamina. *Nat Genet* **38**(9): 1005-1014.

Pombo, A., Jackson, D.A., Hollinshead, M., Wang, Z., Roeder, R.G., and Cook, P.R. 1999. Regional specialization in human nuclei: visualization of discrete sites of transcription by RNA polymerase III. *EMBO J* **18**(8): 2241-2253.

Prothero, K.E., Stahl, J.M., and Carrel, L. 2009. Dosage compensation and gene expression on the mammalian X chromosome: one plus one does not always equal two. *Chromosome Res* **17**(5): 637-648.

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

Pullirsch, D., Hartel, R., Kishimoto, H., Leeb, M., Steiner, G., and Wutz, A. 2010. The Trithorax group protein Ash2l and Saf-A are recruited to the inactive X chromosome at the onset of stable X inactivation. *Development* **137**(6): 935-943.

Rada-Iglesias, A., Bajpai, R., Swigut, T., Brugmann, S.A., Flynn, R.A., and Wysocka, J. 2011. A unique chromatin signature uncovers early developmental enhancers in humans. *Nature* **470**(7333): 279-283.

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. *Proc Natl Acad Sci U S A* **99**(10): 6883-6888.

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**(7184): 243-247.

Rippe, K. 2001. Making contacts on a nucleic acid polymer. *Trends Biochem Sci* **26**(12): 733-740.

Rippe, K., von Hippel, P.H., and Langowski, J. 1995. Action at a distance: DNA-looping and initiation of transcription. *Trends Biochem Sci* **20**(12): 500-506.

Robinson, J.T., Thorvaldsdottir, H., Winckler, W., Guttman, M., Lander, E.S., Getz, G., and Mesirov, J.P. 2011. Integrative genomics viewer. *Nat Biotechnol* **29**(1): 24-26.

Rodley, C.D., Bertels, F., Jones, B., and O'Sullivan, J.M. 2009. Global identification of yeast chromosome interactions using Genome conformation capture. *Fungal Genet Biol* **46**(11): 879-886.

Rozen, S. and Skaletsky, H. 2000. Primer3 on the WWW for general users and for biologist programmers. *Methods Mol Biol* **132**: 365-386.

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**(24): 8309-8314.

Ruf, S., Symmons, O., Uslu, V.V., Dolle, D., Hot, C., Ettwiller, L., and Spitz, F. 2011. Large-scale analysis of the regulatory architecture of the mouse genome with a transposon-associated sensor. *Nat Genet* **43**(4): 379-386.

Ryba, T., Hiratani, I., Lu, J., Itoh, M., Kulik, M., Zhang, J., Schulz, T.C., Robins, A.J., Dalton, S., and Gilbert, D.M. 2010. Evolutionarily conserved replication timing profiles predict long-range chromatin interactions and distinguish closely related cell types. *Genome Res* **20**(6): 761-770.

Sabatino, D.E., Cline, A.P., Gallagher, P.G., Garrett, L.J., Stamatoyannopoulos, G., Forget, B.G., and Bodine, D.M. 1998. Substitution of the human beta-spectrin promoter for the human agamma-globin promoter prevents silencing of a linked human beta-globin gene in transgenic mice. *Mol Cell Biol* **18**(11): 6634-6640.

Saitoh, N., Bell, A.C., Recillas-Targa, F., West, A.G., Simpson, M., Pikaart, M., and Felsenfeld, G. 2000. Structural and functional conservation at the boundaries of the chicken beta-globin domain. *EMBO J* **19**(10): 2315-2322.

Sanyal, A., Bau, D., Marti-Renom, M.A., and Dekker, J. 2011. Chromatin globules: a common motif of higher order chromosome structure? *Curr Opin Cell Biol*.

Saxonov, S., Berg, P., and Brutlag, D.L. 2006. A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. *Proc Natl Acad Sci U S A* **103**(5): 1412-1417.

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. 2009. Preferential associations between co-regulated genes reveal a transcriptional interactome in erythroid cells. *Nat Genet* **42**(1): 53-61.

Schubeler, D., Scalzo, D., Kooperberg, C., van Steensel, B., Delrow, J., and Groudine, M. 2002. Genome-wide DNA replication profile for Drosophila melanogaster: a link between transcription and replication timing. *Nat Genet* **32**(3): 438-442.

Segal, E., Fondufe-Mittendorf, Y., Chen, L., Thastrom, A., Field, Y., Moore, I.K., Wang, J.P., and Widom, J. 2006. A genomic code for nucleosome positioning. *Nature* **442**(7104): 772-778.

Segal, E. and Widom, J. 2009. What controls nucleosome positions? *Trends Genet* **25**(8): 335-343.

Senner, C.E. and Brockdorff, N. 2009. Xist gene regulation at the onset of X inactivation. *Curr Opin Genet Dev* **19**(2): 122-126.

Sexton, T., Umlauf, D., Kurukuti, S., and Fraser, P. 2007. The role of transcription factories in large-scale structure and dynamics of interphase chromatin. *Semin Cell Dev Biol* **18**(5): 691-697.

Sigrist, C.J. and Pirrotta, V. 1997. Chromatin insulator elements block the silencing of a target gene by the Drosophila polycomb response element (PRE) but allow trans interactions between PREs on different chromosomes. *Genetics* **147**(1): 209-221.

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**(11): 1348-1354.

Simonis, M., Kooren, J., and de Laat, W. 2007. An evaluation of 3C-based methods to capture DNA interactions. *Nat Methods* **4**(11): 895-901.

Smith, R.D., Seale, R.L., and Yu, J. 1983. Transcribed chromatin exhibits an altered nucleosomal spacing. *Proc Natl Acad Sci U S A* **80**(18): 5505-5509.

Solomon, M.J. and Varshavsky, A. 1985. Formaldehyde-mediated DNA-protein crosslinking: a probe for in vivo chromatin structures. *Proc Natl Acad Sci U S A* **82**(19): 6470-6474.

Spector, D.L. 2006. SnapShot: Cellular bodies. *Cell* **127**(5): 1071.

Spilianakis, C.G. and Flavell, R.A. 2004. Long-range intrachromosomal interactions in the T helper type 2 cytokine locus. *Nat Immunol* **5**(10): 1017-1027.

Spitz, F., Gonzalez, F., and Duboule, D. 2003. A global control region defines a chromosomal regulatory landscape containing the HoxD cluster. *Cell* **113**(3): 405-417.

Splinter, E., De Wit, E., Nora, E., Klous, P., Van der Werken, H., Zhu, Y., Kaaij, L.J.T., 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 and Development*.

Splinter, E., Grosveld, F., and de Laat, W. 2004. 3C technology: analyzing the spatial organization of genomic loci in vivo. *Methods Enzymol* **375**: 493-507.

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 Dev* **20**(17): 2349-2354.

Strouboulis, J., Dillon, N., and Grosveld, F. 1992. Developmental regulation of a complete 70-kb human beta-globin locus in transgenic mice. *Genes Dev* **6**(10): 1857-1864.

Tanizawa, H., Iwasaki, O., Tanaka, A., Capizzi, J.R., Wickramasinghe, P., Lee, M., Fu, Z., and Noma, K. 2010. Mapping of long-range associations throughout the fission yeast genome reveals global genome organization linked to transcriptional regulation. *Nucleic Acids Res* **38**(22): 8164-8177.

Team, R.D.C. 2010. R: A language and environment for statistical computing. *R Foundation for Statistical Computing*.

Tiwari, V.K., McGarvey, K.M., Licchesi, J.D., Ohm, J.E., Herman, J.G., Schubeler, D., and Baylin, S.B. 2008. PcG proteins, DNA methylation, and gene repression by chromatin looping. *PLoS Biol* **6**(12): 2911-2927.

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**(3): e1001343.

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**(6): 1453-1465.

Tremethick, D.J. 2007. Higher-order structures of chromatin: the elusive 30 nm fiber. *Cell* **128**(4): 651-654.

Tschopp, P., Tarchini, B., Spitz, F., Zakany, J., and Duboule, D. 2009. Uncoupling time and space in the collinear regulation of Hox genes. *PLoS Genet* **5**(3): e1000398.

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**(7): 1341-1354.

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**(3): 453-462.

Van der Ploeg, 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**(5748): 637-642.

van der Werken, H. In Prep.

van Steensel, B. and Dekker, J. 2010. Genomics tools for unraveling chromosome architecture. *Nat Biotechnol* **28**(10): 1089-1095.

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**(8): 2041-2051.

Versteeg, R., van Schaik, B.D., van Batenburg, M.F., Roos, M., Monajemi, R., Caron, H., Bussemaker, H.J., and van Kampen, A.H. 2003. The human transcriptome map reveals extremes in gene density, intron length, GC content, and repeat pattern for domains of highly and weakly expressed genes. *Genome Res* **13**(9): 1998-2004.

Visel, A., Blow, M.J., Li, Z., Zhang, T., Akiyama, J.A., Holt, A., Plajzer-Frick, I., Shoukry, M., Wright, C., Chen, F. et al. 2009a. ChIP-seq accurately predicts tissue-specific activity of enhancers. *Nature* **457**(7231): 854-858.

Visel, A., Prabhakar, S., Akiyama, J.A., Shoukry, M., Lewis, K.D., Holt, A., Plajzer-Frick, I., Afzal, V., Rubin, E.M., and Pennacchio, L.A. 2008. Ultraconservation identifies a small subset of extremely constrained developmental enhancers. *Nat Genet* **40**(2): 158-160.

Visel, A., Rubin, E.M., and Pennacchio, L.A. 2009b. Genomic views of distant-acting enhancers. *Nature* **461**(7261): 199-205.

Wai, A.W., Gillemans, N., Raguz-Bolognesi, S., Pruzina, S., Zafarana, G., Meijer, D., Philipsen, S., and Grosveld, F. 2003. HS5 of the human beta-globin locus control region: a developmental stage-specific border in erythroid cells. *EMBO J* **22**(17): 4489-4500.

Wall, L., deBoer, E., and Grosveld, F. 1988. The human beta-globin gene 3' enhancer contains multiple binding sites for an erythroid-specific protein. *Genes Dev* **2**(9): 1089-1100.

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*.

Wasylyk, B., Wasylyk, C., Augereau, P., and Chambon, P. 1983. The SV40 72 bp repeat preferentially potentiates transcription starting from proximal natural or substitute promoter elements. *Cell* **32**(2): 503-514.

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**(7180): 796-801.

West, A.G., Gaszner, M., and Felsenfeld, G. 2002. Insulators: many functions, many mechanisms. *Genes Dev* **16**(3): 271-288.

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

Wutz, A. and Jaenisch, R. 2000. A shift from reversible to irreversible X inactivation is triggered during ES cell differentiation. *Mol Cell* **5**(4): 695-705.

Wutz, A., Rasmussen, T.P., and Jaenisch, R. 2002. Chromosomal silencing and localization are mediated by different domains of Xist RNA. *Nat Genet* **30**(2): 167-174.

Xie, X., Mikkelsen, T.S., Gnirke, A., Lindblad-Toh, K., Kellis, M., and Lander, E.S. 2007. Systematic discovery of regulatory motifs in conserved regions of the human genome, including thousands of CTCF insulator sites. *Proc Natl Acad Sci U S A* **104**(17): 7145-7150.

Yang, F., Babak, T., Shendure, J., and Disteche, C.M. 2010. Global survey of escape from X inactivation by RNA-sequencing in mouse. *Genome Res* **20**(5): 614-622.

Yu, H., Kessler, J., and Shen, J. 2000. Heterogeneous populations of ES cells in the generation of a floxed Presenilin-1 allele. *Genesis* **26**(1): 5-8.

Yusufzai, T.M. and Felsenfeld, G. 2004. The 5'-HS4 chicken beta-globin insulator is a CTCF-dependent nuclear matrix-associated element. *Proc Natl Acad Sci U S A* **101**(23): 8620-8624.

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

Zhang, L.F., Huynh, K.D., and Lee, J.T. 2007. Perinucleolar targeting of the inactive X during S phase: evidence for a role in the maintenance of silencing. *Cell* **129**(4): 693-706.

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**(11): 1341-1347.

Zufferey, R., Nagy, D., Mandel, R.J., Naldini, L., and Trono, D. 1997. Multiply attenuated lentiviral vector achieves efficient gene delivery in vivo. *Nat Biotechnol* **15**(9): 871-875.

SAMENVATTING

Alle genetische informatie die nodig is voor de ontwikkeling en het functioneren van een organisme is opgeslagen in het DNA. Dit DNA is in eukaryoten georganiseerd in chromosomen. Binnen deze chromosomen bevinden zich de genen, welke coderen voor de functionele biomoleculen die verantwoordelijk zijn voor het uitvoeren van cellulaire functies. Een menselijk lichaam bestaat uit meer dan 200 verschillende celtypen, alle met dezelfde genetische informatie, maar verschillend van vorm en functie. Het reguleren van de expressie van genen resulteert in de specificatie van de verschillende celtypen tijdens de ontwikkeling. Maar ook tijdens de latere stadia van het leven is de expressie van de juiste genen kritisch, omdat misregulatie kan leiden tot ziektes zoals kanker. Om deze reden wordt er intensief onderzoek verricht naar de verschillende mechanismen waarmee expressie van genen wordt gereguleerd.

Het DNA dat in elke cel is opgeslagen heeft een lengte van ongeveer twee meter. Om het DNA in de celkern met een diameter van 10 micrometer te passen moet het opgevouwen en ingepakt worden. Belangrijk hierbij is dat het DNA ook weer uitgepakt moet kunnen worden zodra genen moet worden afgelezen. De manier waarop het DNA in de celkern gevouwen zit zou daarom heel belangrijk kunnen zijn voor de regulatie van genen. De afgelopen jaren zijn de methoden die het mogelijk maken om de vouwing van het genoom in de celkern te bestuderen sterk verbeterd, wat heeft geleid tot een vernieuwd inzicht in hoe het DNA in de celkern is gevouwen en de factoren die hierbij betrokken zijn.

De ontwikkeling van nieuwe moleculaire technologieën, voornamelijk die van '*Chromosome Conformation Capture*' (3C) en de daarvan afgeleide methoden, heeft sterk bijgedragen aan de huidige kennis over hoe vouwing van het DNA relateert aan de transcriptie van genen. 3C maakt het mogelijk onderzoek te verrichten naar de ruimtelijke organisatie van individuele loci met hoge resolutie, maar vereist veel controles voordat de resultaten kunnen worden geïnterpreteerd. Een gedetailleerde beschrijving van de 3C methode en de controles die nodig zijn bij het toepassen van deze methode op zoogdier cellen zijn beschreven in hoofdstuk 3. Hoofdstuk 4 beschrijft de 4C methode die kan worden omschreven als een high-throughput versie van 3C. Deze technologie maakt het mogelijk om de genomische omgeving van een locus naar keuze te identificeren. Oorspronkelijk werd er gebruik gemaakt van micro-arrays om deze co-localizerende sequenties te identificeren, momenteel wordt er voornamelijk gebruik gemaakt van de nieuwe '*next generation sequencing*' strategieën. De toepassing van sequencing is goedkoper en niet gebonden aan een detectie limiet, dat een verbetering is ten opzichte van micro-arrays. Maar er zijn meer voordelen; in tegenstelling tot micro-array is de sequentie-analyse niet beperkt tot de selectie van '*probes*' die aanwezig zijn op de micro-array. Te samen worden daarmee zowel de gevoeligheid als de resolutie van de 4C methode sterk verhoogd.

Een verdere verhoging van de resolutie wordt bereikt doormiddel van het analyseren van kleine DNA-fragmenten. Hierdoor wordt het mogelijk om precieze genomische contacten tussen cis-regulatoire elementen op te sporen. Toepassing van deze technologie, genaamd highres-4C-seq in toekomstige experimenten zal nieuwe informatie verschaffen over de regulatoire netwerken die de

expressie van genen beïnvloeden en over de elementen die hierbij betrokken zijn.

In hoofdstuk 5 en 6 wordt het muizen β-globine locus gebruikt als een modelsysteem. Het β-globine locus codeert voor de β-eiwitten van het hemoglobine molecuul dat verantwoordelijk is voor het transport van zuurstof in rode bloedcellen. Naast de globine genen bevat dit locus meerdere goed gekarakteriseerde cis-regulatoire elementen waarvan de Locus Control Region (LCR) het meest belangrijk is voor het reguleren van globine transcriptie. Andere sequenties aan weerszijden van de locus binden het eiwit CTCF. Voorheen is aangetoond dat deze CTCF bindingsplaatsen samenkomen in erytroïde voorlopercellen. Hoofdstuk 5 beschrijft experimenten die de rol van CTCF bij het tot stand brengen van deze interacties onderzoeken. We konden aantonen dat, door het verwijderen van het CTCF eiwit en het maken van een mutatie in de 3'HS1 CTCF bindingsplaats, dit resulteert in een vermindering van de lange-afstands interacties tussen de betreffende CTCF bindingsplaatsen. Echter een functioneel effect van deze interacties op de regulatie van β-globine gen expressie kon niet worden gedetecteerd. In een vervolg studie hebben we om die reden de effecten van de afwijkende interacties nader onderzocht. In deze studie, gepresenteerd in hoofdstuk 6, werd highres-4C-seq toegepast op cellen met het wild-type en cellen met het mutant β-globine locus, waaraan de CTCF bindingsplaats ontbrak. We hebben aangetoond dat CTCF lussen vormt in het DNA waardoor de ruimtelijke contacten van sequenties binnen deze lussen wordt beïnvloed. Dit lijdt ertoe dat stukken DNA binnen een lus de voorkeur geven om elkaar te contacteren en contacten buiten de lus verminderen. Naar aanleiding hiervan hebben we onderzocht of er als gevolg van de vertoorde ruimtelijke organizatie binnen het globine locus een preferentiële opregulatie van een van de globine genen gedetecteerd kon worden. Dit bleek inderdaad het geval te zijn; het proximale embryonale gen wordt preferentieel opgereguleerd ten koste van de meer distale volwassen globine genen. Hoewel nog niet onomstotelijk bewezen, toont dit resultaat aan dat de CTCF gemedieerde DNA-lussen helpen om de LCR naar de distale volwassen globine genen te dirigeren. Wij voorspellen dat de formatie van lussen een algemeen mechanisme is waarmee de contacten tussen regulatoire elementen wordt beïnvloed. Daarnaast zal de positie van deze elementen binnen een dergelijke lus en de affiniteit voor elkaar uiteindelijk het effect op gen-expressie bepalen. Deze hypothese wordt ondersteund door de observatie dat, naast het β-globine locus, ook andere gen loci zijn georganiseerd in DNA lussen.

In hoofdstuk 7 worden experimenten beschreven die de organisatie van een volledig chromosoom onderzoeken. Om de relatie tussen gen-expressie en DNA-organisatie op deze schaal te bestuderen was het X chromosoom geselecteerd als modelsysteem. In vrouwelijke zoogdiercellen worden de genen van een van de twee X chromosomen uitgezet om te compenseren voor het verschil in het aantal X chromosomen tussen vrouwelijke en mannelijke cellen. Mits er een onderscheid gemaakt kan worden tussen de interacties van het actieve en inactieve X chromosoom, maakt dit model systeem het mogelijk om de vouwing van twee identieke chromosomen te onderzoeken die totaal verschillend worden behandeld in dezelfde cel. Door gebruik te maken van kleine afwijkingen tussen de twee X chromosomen hebben we een dergelijk allel-specifieke 4C methode ontwikkeld welke voor dit doel is gebruikt. Het X-inactivatie proces begint met de

opregulatie van het gen *Xist* op het toekomstige inactieve X chromosoom zodra embryonale stamcellen differentiëren. *Xist* codeert voor RNA dat zich over het X chromosoom verspreidt om deze vervolgens uit te zetten doormiddel van het aantrekken van verschillende repressieve factoren, waaronder Polycomb Groep eiwitten (PcG) en macro-H2A. Deze repressieve factoren zorgen er op hun beurt voor dat de genen stabiel uitgeschakeld blijven. Zijn de genen eenmaal uitgezet, dan is het Xist RNA niet langer nodig om de genen uit te houden. Door het bestuderen van de vouwing van de beide X chromosomen zijn we erachter gekomen dat op het actieve X chromosoom een scheiding wordt gemaakt tussen actievegen dichte gebieden en inactieve-gen arme regio's, een organisatie die eerder ook werd gevonden op andere chromosomen. Het inactieve X chromosoom liet echter een atypische organisatie zien; op dit chromosoom werden vrijwel geen specifieke contacten tussen de inactieve sequenties gedetecteerd. Dit was niet te wijten aan een gebrek aan contacten, deze zijn in overvloed aanwezig, maar werd veroorzaakt door hun willekeurige verdeling over het chromosoom. Een uitzondering hierop werd gevonden voor de genen die weten te ontsnappen aan het X inactivatie proces, voor deze genen werd gevonden dat ze wel specifiek met elkaar in contact kunnen komen. Door dit vermogen werden deze genen gemakkelijk opgepikt met behulp van de 4C methode, waardoor het aantal bekende genen dat weet te ontsnappen aan het X-inactivatie proces vergroot is. Onder deze genen bevond zich ook het Xist gen, wat opmerkelijk is, aangezien juist dit gen de plaats is waar het Xist RNA wordt geproduceerd dat op zijn beurt verantwoordelijk is voor het uitzetten van het X-chromosoom. Mogelijk zou de lokalisatie van genen die ontsnappen aan het X-inactivatie proces bij het *Xist* gen het verspreiden van het Xist RNA over het chromosoom kunnen vergemakkelijken. Een vergelijkbaar mechanisme is eerder gevonden bij het verspreiden van het 'Dosage Compensation Complex' (DCC) in Drosophila (Grimaud en Becker 2009). Het unieke vouwing patroon van het inactieve X chromosoom hangt af van de aanwezigheid van het Xist RNA, want wanneer dit RNA werd verwijderd nam de inactieve X een structuur aan die deed denken aan de vouwing van het actieve X chromosoom. Deze observatie heeft meerdere consequenties. Als eerste, doordat de hervouwing van het inactieve X chromosoom plaats vond in afwezigheid van het heractiveren van de genen op dit chromosoom, toont dit aan dat de specifieke vouwing van het inactieve X chromosoom niet noodzakelijk is voor het handhaven van de inactiviteit. Om dezelfde reden toont dit resultaat aan dat actieve transcriptie niet vereist is om lange-afstands contacten op te zetten. Deze resultaten komen niet overeen met een op dit moment populair model, dat stelt dat genen samen moeten komen in gespecialiseerde 'transcriptie fabrieken' voor hun expressie. Hoewel we niet kunnen uitsluiten dat na het verwijderen van het Xist RNA het inactieve X-chromosoom een verhoogde affiniteit voor het transcriptie mechanisme heeft, tonen we aan dat chromosoom vouwing wordt bewerkstelligd door andere factoren dan actieve transcriptie. Het zal interessant zijn om te bepalen welke factoren dit zijn en hoe deze betrokken zijn bij chromosoom vouwing.

DANKWOORD

En dan tot slot het meest gelezen en belangrijkste stuk tekst van dit boekje: het dankwoord. Overigens volledig terecht het meest gelezen onderdeel, want zonder de hulp van deze en gene had dit boekje er heel anders uit gezien. Daarvoor wil ik iedereen hartelijk bedanken, met een aantal personen in het bijzonder.

Allereerst Wouter; 10 jaar hebben we het met elkaar uitgehouden. Ik herinner mij het nog als de dag van gister dat ik bij jou, als net gestarte groepsleider, kon starten als analist. Wat een avontuur is het sindsdien geweest. Dankzij jou heldere kijk op verwarrende zaken (lees: verprutste experimenten) en altijd systematische aanpak, gecombineerd met de juiste dosis wilde ideeën -transvectie in mammals?- heb ik mij nooit hoeven vervelen. Het was een voorrecht om met jou te kunnen samenwerken en heb veel respect voor de manier waarop je altijd voor iedereen tijd hebt –ook al heb je dat eigenlijk niet. En als volgend jaar Limburgs Mooiste weer in de planning staat dan houd ik mij aanbevolen!

&

Maar natuurlijk ook dank aan Frank, niet alleen voor de mogelijkheid om na mijn stage op de afdeling Celbiologie te blijven werken, maar zeker ook voor het vertrouwen in mijn kunnen als OIO. Ik blijf mij verbazen over jou mogelijkheid om na drie woorden door te dringen tot de essentie van de materie en er dan ook nog iets relevants over te kunnen vertellen. Hoewel onze contact frequentie (om nog maar even in de geest van dit boekje te blijven) na de verhuizing sterk is afgenomen wil ik je bedanken dat je er altijd was en dat je nu samen met Wouter mijn promotor bent. Ik kan mij geen betere combinatie voorstellen!

Wetenschap is een typische hobby waarin samenwerken vaak efficiënter is dan in je eentje zitten prutsen. Dit is zeker wat ik aan den lijve heb ondervonden bij de (inmiddels ook vaak vruchtbare) samenwerkingen die zijn opgezet. Helen en Niels bedankt voor jullie onmisbare bijdrage aan het CTCF artikel. Elphège and Edith, it was a pleasure collaborating with you on the X chromosome project, thanks for all your help. Measuring distances between FISH spots (almost) became my new hobby ☺. Dit laatste project had er trouwens nooit geweest als Joost niet had bijgestaan met raad&daad en de oh zo belangrijke X-goodies waarmee dit project echt van de grond kon komen. Om die reden waardeer ik het dan ook extra dat je plaats hebt willen nemen in mijn beoordelingscommissie en je op die manier betrokken kon blijven bij dit project, dank daarvoor.

Dan de oud groeps/lab-genoten: Bas en Robert-Jan, 3C-ers van het eerste uur, bedankt voor die onvergetelijke pipetteerzomer van 2002 waarin we, met de radio op volume 10 zeven dagen per week aan het werk waren om de concurrentie af te troeven. Maar natuurlijk ook Jurgen, Daan en Marieke, zelf inmiddels al (lang) gepromoveerd, wil ik bedanken voor jullie bijdrage en gezelligheid. Zo waren er de spelletjes avonden en de 'Cluster 15 bowling competitie', waar een ieder zijn (verborgen) fanatisme goed tot zijn recht kon laten komen. En natuurlijk de invoering van 'de Bassie' (infameuze prijs voor prutswerk), met de 'permanente Bassie' en 'Bassie devaluatie' als afgeleiden. Emilie, your positive and friendly attitude made you very welcome. I hope we meet again soon. Sanja and Petros, I still miss those days that we were sharing desk -and bench space. Petra, jij hebt je de afgelopen jaren helemaal onmisbaar gemaakt in het lab en was dan ook erg blij dat je mee bent gegaan naar Utrecht. Ik wens je veel succes met het voltooien van je master en straks je eigen promotie –maar dat zit wel goed. Ook Sjoerd en Harmen zijn meeverhuisd vanuit Rotterdam. Sjoerd, de laatste maanden hebben wij

nog even intensief samengewerkt en ik hoop dat we met deze eindspurt dit project nog tot een mooi einde kunnen brengen. Succes met jouw laatste loodjes. Harmen, jou vermogen om over echt elk onderwerp een gesprek te beginnen blijft mij verbazen. Met weemoed denk ik terug aan die tijd dat mijn bureau tussen die van jou en Elzo stond.. alleen de maan -en woensdagen zijn echt productief geweest ☺.

Sinds de verhuizing is de groep weer behoorlijk uitgebreid. Elzo, bedankt voor de fantastische samenwerking, het sparren en je mening over echt van alles. Ik ben dan ook erg blij dat jij mijn paranimf wilt zijn. Ook mijn andere paranimf wil ik bedanken; Patrick, bedankt voor jouw vrolijke noot, je goeie tips en hulp met de muizen. Jammer alleen van die kerstman, daarna ben jij voor mij nooit meer helemaal dezelfde geworden. Blaise it was great working with you, all the best in Cambridge. En dan zijn er natuurlijk de mede OIO's: Yun, Paula, Yuva en Britta. With you around there's never a dull moment (read: moment of complete silence), thanks for your enthusiasm and positive attitude.

Over de jaren zijn er ook studenten die gek genoeg zijn geweest om hun stage bij mij te voltooien. Yun (jij zou nu toch wel genoeg Nederlands moeten kunnen om dit te vertalen), Anita en Dylan, bedankt voor jullie bijdrage en het enthousiasme waarmee jullie je hebben ingezet, het was een plezier om met jullie samen te werken. Inmiddels zijn jullie zelf ook (bijna) gestart met je promotie traject en ik wens jullie daarbij alle succes van de wereld.

Maar ook buiten de labdeuren was altijd wel iemand te vinden voor een praatje, kop koffie of een biertje op zijn tijd. In het speciaal hiervoor dank aan de (oud)buren in Rotterdam (lab 706, 710, 902 en 1030), zonder wie menig (celkweek) uurtje een stuk langer had geduurd. Maureen bedankt voor de Wii avondjes en Alex, wanneer gaan we nu eindelijk eens fietsen? Maar laat ik ook Jeroen en Melle niet vergeten; zonder jullie had het bestellen van de restrictie enzymen zeker in de soep gelopen.

Ook in Utrecht was er altijd wel reden voor een praatje of om een biertje te pakken op de vrijdag-middag-borrel (sorry Romke, maar die frituur is daarbij echt onmisbaar). En er werd gefietst, de laatste tijd wat minder frequent, maar het was genieten met volle teugen. De harde fietskern - Frank, Bart, Sjoerd en Arnoud bedankt daarvoor.

En dan zijn er natuurlijk de computer mannen, P&O, de media keuken, de dierverzorgers en de mannen van het magazijn, bedankt voor al jullie hulp en ondersteuning. Een speciale vermelding verdient Stieneke; jou inzet voor het algemeen belang kan ik erg waarderen. En in de buitencategorie: Ira. De nauwkeurigheid waarmee jij de ultieme stress momenten binnen het OIO-bestaan weet op te sporen om deze vervolgens te vergroten is een kunst apart ;).

Mam en Pap, René & Uting, Bart & Alexandra en Niels & Sarina ik hoop dat we elkaar binnenkort gewoon weer thuis kunnen zien i.p.v. in het ziekenhuis. Wij hebben de laatste tijd onze portie toch wel gehad. Wat de toekomst ook mag zijn, volgend jaar als (gepensioneerde) opa en oma, dat klinkt toch een stuk gezelliger! :-p

En dan lest-best: Jessica. Bij deze een officieel -en gemeend- sorry voor het gebrek aan tijd waarmee ik het laatste (half) jaar te kampen heb gehad. Dat je ondanks dat mij toch bent blijven steunen en nu met mij op avontuur in het buitenland wilt gaan, betekent heel veel voor mij. Ik hou van je.

Erik

CORRICULUM VITAE

Erik Cornelis Splinter is geboren op 21 mei 1980 in Alphen a/d Rijn, Nederland. In 1997 behaalde hij zijn HAVO diploma aan de scholengemeenschap Prins Maurits te Middelharnis. In datzelfde jaar begon hij met zijn studie aan het Hoger Laboratorium Onderwijs aan de Hogeschool Rotterdam waaraan hij in 2001 is afgestudeerd. Gedurende deze opleiding heeft hij stage gelopen aan de afdeling Celbiologie van het ErasmusMC te Rotterdam onder supervisie van Rien van Haperen en dr. Rini de Krom en heeft meegewerkt aan het genereren en karakteriseren van transgene muis modellen voor aterosclerose. Na het afronden van zijn opleiding is hij begonnen als research-analist in de groep van dr. Wouter de Laat, waar hij heeft bijgedragen aan de toepassing van de Chromosome Conformation Capture techniek op zoogdiercellen met als doel de relatie tussen DNA vouwing en transcriptie te bestuderen. In 2007 heeft hij dit onderzoek voortgezet in zijn promotieonderzoek, dat vanaf september 2009, onder supervisie van Prof.dr. Frank Grosveld en Prof.dr. Wouter de Laat in het Hubrecht Instituut te Utrecht heeft plaatsgevonden. De resultaten van dit onderzoek zijn beschreven in dit proefschrift.

&

LIST OF PUBLICATIONS

The complex transcription regulatory landscape of our genome: control in three dimensions.
Splinter E and de Laat W
EMBO J. Accepted for publication. Review.

The inactive X chromosome adopts a unique three-dimensional conformation that is dependent on Xist RNA.
Splinter E, de Wit E, Nora EP, Klous P, van de Werken HJ, Zhu Y, Kaaij LJ, van IJcken W, Gribnau J, Heard E, de Laat W.
Genes Dev. 2011 Jul 1;25(13):1371-83.

An evolutionarily conserved three-dimensional structure in the vertebrate Irx clusters facilitates enhancer sharing and coregulation.
Tena JJ, Alonso ME, de la Calle-Mustienes E, **Splinter E**, de Laat W, Manzanares M, Gómez-Skarmeta JL.
Nat Commun. 2011;2:310.

Diverse gene reprogramming events occur in the same spatial clusters of distal regulatory elements.
Hakim O, Sung MH, Voss TC, **Splinter E**, John S, Sabo PJ, Thurman RE, Stamatoyannopoulos JA, de Laat W, Hager GL.
Genome Res. 2011 May;21(5):697-706.

Studying physical chromatin interactions in plants using Chromosome Conformation Capture (3C).
Louwers M, **Splinter E**, van Driel R, de Laat W, Stam M.
Nat Protoc. 2009;4(8):1216-29.

Transcription and chromatin organization of a housekeeping gene cluster containing an integrated beta-globin locus control region.
Noordermeer D, Branco MR, **Splinter E**, Klous P, van IJcken W, Swagemakers S, Koutsourakis M, van der Spek P, Pombo A, de Laat W.
PLoS Genet. 2008 Mar 7;4(3):e1000016.

Three-dimensional organization of gene expression in erythroid cells.
de Laat W, Klous P, Kooren J, Noordermeer D, Palstra RJ, Simonis M, **Splinter E**, Grosveld F.
Curr Top Dev Biol. 2008;82:117-39. Review.

Quantitative analysis of chromosome conformation capture assays (3C-qPCR).
Hagège H, Klous P, Braem C, **Splinter E**, Dekker J, Cathala G, de Laat W, Forné T.
Nat Protoc. 2007;2(7):1722-33.

Beta-globin active chromatin Hub formation in differentiating erythroid cells and in p45 NF-E2 knock-out mice.
Kooren J, Palstra RJ, Klous P, **Splinter E**, von Lindern M, Grosveld F, de Laat W.
J Biol Chem. 2007 Jun 1;282(22):16544-52.

Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4C).
Simonis M, Klous P, **Splinter E**, Moshkin Y, Willemsen R, de Wit E, van Steensel B, de Laat W.
Nat Genet. 2006 Nov;38(11):1348-54.

CTCF mediates long-range chromatin looping and local histone modification in the beta-globin locus.
Splinter E, Heath H, Kooren J, Palstra RJ, Klous P, Grosveld F, Galjart N, de Laat W.
Genes Dev. 2006 Sep 1;20(17):2349-54.

The active spatial organization of the beta-globin locus requires the transcription factor EKLF.
Drissen R, Palstra RJ, Gillemans N, **Splinter E**, Grosveld F, Philipsen S, de Laat W.
Genes Dev. 2004 Oct 15;18(20):2485-90.

3C technology: analyzing the spatial organization of genomic loci in vivo.
Splinter E, Grosveld F, de Laat W.
Methods Enzymol. 2004;375:493-507.

The beta-globin nuclear compartment in development and erythroid differentiation.
Palstra RJ, Tolhuis B, **Splinter E**, Nijmeijer R, Grosveld F, de Laat W.
Nat Genet. 2003 Oct;35(2):190-4.

Looping and interaction between hypersensitive sites in the active beta-globin locus.
Tolhuis B, Palstra RJ, **Splinter E**, Grosveld F, de Laat W.
Mol Cell. 2002 Dec;10(6):1453-65.