The role of elements binding CTCF and cohesin in directing tissue-specific enhancer activity

Lars Lee Perry Hanssen

Linacre College

A thesis submitted for the degree of

Doctor of Philosophy

of

University of Oxford

Wellcome Trust Chromosome and Developmental Biology
Weatherall Institute of Molecular Medicine
Medical Sciences Division
University of Oxford

Trinity Term 2016

Abstract

Lars Lee Perry Hanssen
Linacre College

Doctor of Philosophy
Trinity term 2016

Title: **The role of elements binding CTCF and cohesin in directing tissue-specific enhancer activity**

Distal enhancer elements regulate the tissue-specific expression of their target genes via the establishment of physical interactions with the gene promoter. In mice, a cluster of five enhancers, jointly classified as a super-enhancer, specifically upregulate α-globin gene expression during erythroid differentiation. Aside from the *Nprl3* gene, whose promoter is located inside this enhancer region, expression-levels of other genes within a short distance (<50kb) of the enhancer region are not affected by the activation of the enhancer in erythroid cells, despite being located within the same sub-TAD in erythroid cells. The CCCTC-binding factor (CTCF) is implicated in the organisation of chromosome topology through the formation of interactions between its binding sites in an orientation-dependent manner. In this thesis, I demonstrate that CTCF functions *in vivo* as a boundary to maintain α-globin enhancer-promoter specificity in erythroid cells

The study of the local chromatin architecture by next-generation Capture-C reveals that α-globin enhancer and promoter interactions are constrained to a compartment of roughly 70kb. The unidirectional interaction profiles of the α-globin enhancers are delimited by the interactions between two genomic domains flanking the α-globin cluster. Further investigation shows that each of these domains contains several CTCF binding sites orientated in tandem, such that CTCF binding orientation between domains is convergent. Although CTCF binding across the α-globin locus is identical between mouse embryonic stem (ES) cells and erythroid cells, interaction between these domains occurs only in erythroid cells suggesting it is dependent on the formation of tissue-specific α-globin enhancer-promoter interactions.

By generating a series of mouse models, deleting CTCF binding sites at the α-globin enhancers singly and in combination, I show that the deletion of two CTCF binding sites directly flanking the enhancer cluster results in a shift in interactions between flanking domains, away from the enhancer region. This leads to an expansion of enhancer interactions to include two genes directly upstream of the α-globin enhancers: *Rhbdf1* and *Mpg*. Despite the *Rhbdf1* gene being subject to polycomb group protein-mediated gene repression in erythroid cells, ablation of CTCF binding results in increased interactions between both the *Rhbdf1* and *Mpg* gene promoters and the α-globin enhancers and concurrent strong transcriptional upregulation of both genes. The *Rhbdf1* gene promoter acquires the active histone mark H3K4me3, but doesn't lose Polycomb Repressive Complex 2 (PRC2) mark H3K27me3 or binding of its catalytic component Ezh2. Despite the presence of this repressive mark, robust levels of *Rhbdf1* expression are detected at levels higher than those in ES cells where this gene is actively expressed under the influence of its own enhancer. I conclude that regulation of the direction of enhancer interactions by CTCF is required for the promoter specificity of enhancers and the maintenance of transcriptional states of nearby genes.

Acknowledgements

First and foremost, I would like to express my gratitude to my supervisors Doug Higgs and Ben Davies for giving me the opportunity to work in their respective labs and for guiding my work presented in this thesis over the last four years. I thank Jim Hughes for his advice with respect to the bio-informatics carried out in this thesis and Andrew Smith for the valuable insights into cell culture and genome engineering. I would also like acknowledge Shona Murphy and Chris Norbury for their advice and support in the capacity of my thesis committee. This work was made possible by generous funding from the Wellcome Trust and the Medical Research Council UK.

My work and this thesis were also influenced by current and previous members of the Higgs, Hughes, Gibbons and Buckle labs. I would like to thank James Davies, Andrew King, Mira Kassouf, Marieke Oudelaar, Danuta Jeziorska, Rosa Stolper, Martin Larke, Yavor Bozhilov, Anthony Cheong, Bryony Graham, Helena Ayub, David Clynes, Diu Nguyen, Deborah Hay, Hsiao Voon, Christian Babbs, Sachith Mettananda, and Jill Brown to name but a few for their help and support.

I would like to thank Marieke Oudelaar for our joint efforts in elucidating the topology of the locus, and the many useful discussions following from our work; Jackie Sharpe, Jackie Sloane-Stanley, and Sue Butler were all invaluable in helping me with the mouse work, both the maintenance of colonies and the analysis of haematological phenotypes; Chris Preece, without whose expert micro-injection skills this project would not have been possible; Alberto Cebrian-Serrano, Daniel Biggs, and Nicole Hortin for their help with cloning and the maintenance of mice and cells.

Last but certainly not least, I would like to thank my family and friends for being there for me throughout my DPhil; my parents, Geert and Ria Hanssen for their support and understanding; my sister, Malu, for coming to visit me in Oxford often and forgiving me my absent mindedness. Nico Kist and Jordan Mansell for the occasional coffee or beer and for helping me keep my sanity throughout.

List of abbreviations

3C	Chromosome conformation capture
4C	Circularised chromosome conformation capture
5C	Chromosome conformation capture carbon copy
ATAC-seq	Assay for Transposase-Accessible Chromatin with high throughput sequencing
BCB	Brilliant cresyl blue staining
bp	Base pair
BSA	Bovine serum albumin
Cas	CRISPR-associated
cDNA	Complementary DNA
ChIA-PET	Chromatin Interaction Analysis by Paired-End Tag Sequencing
ChIP	Chromatin immunoprecipitation
CK2	Caseine kinase 2
CRISPR	Clustered regularly interspaced short palindromic repeats
CT	Chromosome territory
CTCF	CCCTC binding factor
D29	Mouse with a deletion of the HS-29 CTCF binding sequence
D38	Mouse with a deletion of the HS-38 CTCF binding sequence
D3839	Mouse with a deletion of the HS-38 and HS-39 CTCF binding sequences
D39	Mouse with a deletion of the HS-39 CTCF binding sequence
DHS	DNaseI hypersensitive site
DMSO	Dimethyl sulfoxide
DNA	Deoxyribonuleic acid
DNMT1	DNA (Cytosine-5-)-Methyltransferase 1
dpc	days post coitum
DSB	Double strand break
DSG	Disuccinimidyl glutarate
EDTA	Ethylenediaminetetraacetic acid
Eed	Embryonic ectoderm development

Epha4	PH Receptor A4
ES	Embryonic stem
Ezh2	Enhancer of zeste homolog 2
FACS	Fluorescence-activated cell sorting
FBS	Foetal Bovine Serum
FDR	False discovery rate
FOXA1	Forkhead box A1
Gata1	GATA-binding factor 1
Gata2	GAT-binding factor 2
gRNA	Guide RNA
H	Histone
Hb	Haemoglobin
Hba-a1/2	α-globin genes
Hba-x	Embryonic ζ-globin gene
Hbb-b1	β-globin gene
Hbq1-a/b	θ-globin pseudo-genes
HCT	Haematocrit
HDR	Homology-directed repair
HDR	Homology-directed repair
HEPES	4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid
HGB	Haemoglobin count
HS	Hyper Sensitive
ICR	Imprinting control region
IDH	Isocitrate dehydrogenase
Igf2	Insulin-like growth factor 2
Il9r	Interleukin 9 receptor
K	Lysine
kb	Kilo base
Klf1	Krüppel-like factor 1
LAD	Lamina-associated domain
LIF	Leukemia inhibitory factor
MACS	Magnetic activated cell sorting
Mb	Mega base

MCH	Mean corpuscular haemoglobin
MCS	Multispecies conserved sequence
MCS-R	Multi-species conserved sequence regulatory
MCV	Mean corpuscular volume
MHC	Major histocompatibility complex
MLL	Mixed-lineage leukemia
Mpg	N-Methylpurine DNA Glycosylase
MRC	Medical Research Council
mRNA	messenger RNA
n.s.	non-significant
neo	Neomycin
Nf-e2	Nuclear factor erythroid 2
NHEJ	Non-homologous end joining
Nipbl	Nipped B-like protein
NLS	Nuclear localisation signal
NPC	Neural progenitor cell
Nprl3	Nitrogen Permease Regulator-Like 3
PARP1	Poly (ADP-ribose) polymerase I
PBS	Phosphate buffered saline
Pcdh	Protocadherin
PcG	Polycomb group proteins
PCR	Polymerase chain reaction
PH	Phenylhydrazine
PIC	Protease inhibitor complex
PLT	Platelet count
PolII	RNA Polymerase II
PRC1	Polycomb repressive complex 1
PRC2	Polycomb repressive complex 2
puro	Puromycin
Rbbp4	Retinoblastoma binding protein 4
RBC	Red blood cell count
RNA	Ribonucleic acid
RNA-seq	RNA sequencing
ROSE	Rank ordering of super enhancers

RPKM	Reads per kb per million mapped
RT	Reverse transcriptase
RT-qPCR	Real-time quantitative polymerase chain reaction
RVD	Repeat Variable Domain
SD	Standard deviation
SDS	Sodium dodecyl sulfate
SET1	Su(var)3-9, Enhancer of Zeste, Trithorax 1
sgRNA	Single guide RNA
Sh3pxd2b	SH3 and PX Domains 2B
Shh	Sonic Hedgehog
Smc1	Structural maintenance of chromosomes 1
Smc3	Structural maintenance of chromosomes 3
Snrnp25	Small Nuclear Ribonucleoprotein U11/U12 Subunit 25
ssODN	single-stranded oligodeoxynucleotide
SUZ12	Suppressor of zeste 12 protein homolog
TAD	Topologically-associated domain
TAF3	TATA-box binding protein associated factor 3
TALEN	Transcription activator-like effector nucleases
TF	Transcription factor
Tris	Tris(hydroxymethyl)aminomethane
TrxG	Trithorax group proteins
UK	United Kingdom
USF1	Upstream stimulatory factor 1
UV	Ultra-violet
WBC	White blood cell count
XIC	X-inactivation centre
YY1	Yin Yang 1

Table of contents

List of figures

List of tables

Chapter 1: Introduction

1.1 The organisation of the eukaryotic genome

1.1.1 Nuclear organisation and gene expression

The eukaryotic cell is faced with the challenge of packaging the genome several thousand-fold into the cell nucleus, while orchestrating the compaction and decompaction of chromatin for cellular processes such as gene transcription, DNA replication and DNA repair. The structure of interphase chromosomes has been studied since the late 19th and early 20th century, when Carl Rabl and Theodor Boveri first observed a non-random organisation into discrete regions of nuclear space which Boveri termed “chromosome territories” (Cremer and Cremer 2010). Since the existence of chromosome territories was confirmed in the 1980s by laser-UV-microirradiation experiments (Cremer *et al.* 1982), subsequent experiments have gradually established links between the nuclear positioning of chromosomes and gene density. Gene-dense, early-replicating chromosomes are located towards the nuclear interior, whereas late-replicating and gene-poor chromosomes are located closer to the nuclear border (Boyle *et al.* 2001; Tanabe *et al.* 2002). More recently, by mapping interactions of chromatin with the nuclear lamina, this correlation was also shown to hold true at the level of chromosome domains; gene-poor chromosome regions between 0.1 and 10 Mb in size, termed lamina associated domains (LADs), were found to interact more with the nuclear lamina (Guelen et al. 2008).

These observations have led to the proposal of a model in which the interior nuclear compartment is associated with active transcription, whereas inactive genes are positioned in the nuclear periphery (Lanctôt *et al.* 2007; Geyer, Vitalini and Wallrath 2011). Although this is consistent with the observations that some gene loci are repositioned towards the nuclear interior upon gene activation (Kosak et al. 2002; Hewitt et al. 2004; Williams 2006) and that genes located on LADs are expressed at lower levels and are enriched for repressive chromatin marks (Guelen et al. 2008), the discovery of several functional nuclear substructures suggest that the relationship between nuclear organisation and gene expression is likely to be more complex. The best characterised of these, the nucleolus (Pederson 2011) and Cajal bodies (Nizami, Deryusheva and Gall 2010), have been shown to be specifically involved in the processing of ribosomal RNA and nuclear RNA respectively. More generally, actively-transcribed genes have been proposed to co-localise into similar nuclear sub-compartments, termed “transcription factories” (Iborra et al. 1996; Osborne et al. 2004). Similarly, genes that are developmentally repressed by polycomb group proteins (PcG) have been shown to be recruited to Polycomb bodies in *Drosophila* (Ficz 2005; Grimaud *et al.* 2006; Boettiger *et al.* 2016).

These findings highlight that the nucleus is highly organised and suggest that nuclear architecture is tightly linked to genome function, but whether this is important for gene regulation or merely reflects self-association is not clear. Although the mechanisms by which genes or chromosome domains form and interact with nuclear sub-compartments are poorly understood, evidence suggests that the nuclear position of genomic loci is established on the basis of chromatin domains (Mahy

2002; Fraser and Bickmore 2007; Bickmore and van Steensel 2013), although individual genes have been shown to be able to reposition away from their chromosome territories upon gene activation (Chambeyron and Bickmore 2004; Ferrai et al. 2010). The repositioning of genes away from chromosome territories or the nuclear periphery upon gene activation is thought to be driven by the binding of transcription factors to gene regulatory elements such as enhancers (Lundgren *et al.* 2000; Noordermeer *et al.* 2008) .

The emergence of new genomics methods to study genome organisation has fuelled an increased understanding of chromosome domains and has allowed chromosome topology to be studied in detail at the level of individual genes via the measurement of DNA-DNA contacts.

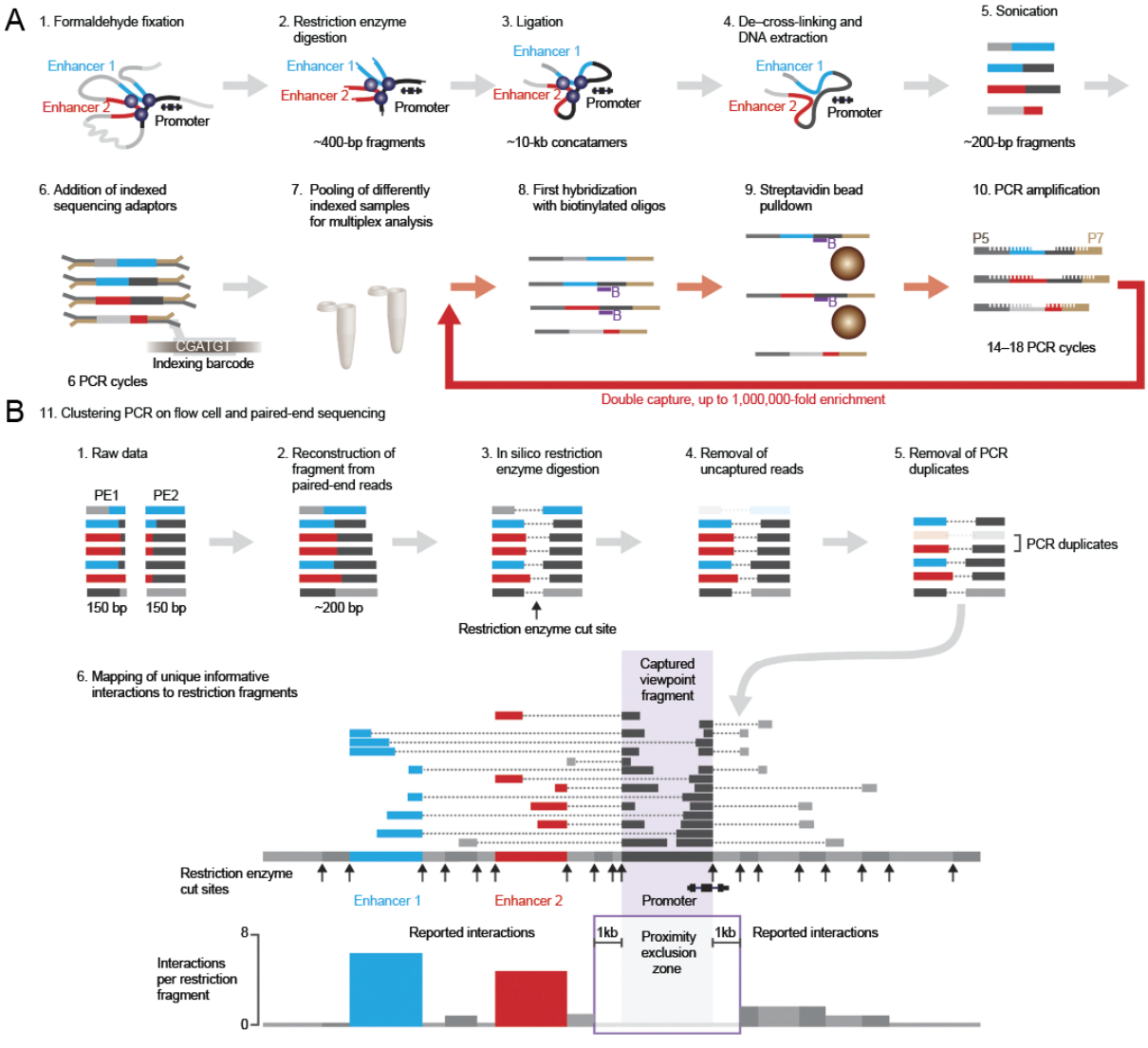


Figure 1.1 Schematic overview of the Capture-C protocol. **A.** Experimental workflow **B.** Bioinformatic analysis. Figure from Davies et al. 2015

1.1.2 Chromosome Conformation Capture (3C) technologies

The development of Chromosome Conformation Capture (3C) and derived technologies based on the same "proximity ligation principle", has enabled high-resolution analyses of physical contacts between genomic elements in nuclear space (Dekker et al. 2002). 3C relies on the fixation of the three-dimensional chromatin organisation by chemical crosslinking, after which chromatin is digested with a restriction enzyme. Digested DNA fragments that were in in close spatial proximity in the nucleus, but not necessarily contiguous in the linear genome, are preferentially re-ligated in the subsequent ligation reaction. The choice of restriction enzyme limits the resolution at which interactions can be detected to the size of the restriction fragments. Commonly used enzymes have six- or four-bp recognition sequences which results in average fragment sizes of 4096 and 256 bp respectively. While PCR amplification across specific ligated junctions was initially (i.e. in 3C) used to quantify the pairwise interaction frequency between two selected genomic loci (one vs one), this methodology was later combined with high-throughput sequencing to allow detection of interactions between one locus (or "viewpoint") and the rest of the genome (4C, one vs all), across a genomic region (5C, many vs many), and across the entire genome (Hic, all vs all). While the ability to generate whole-genome contact maps has greatly refined our understanding of chromosome topology, Hi-C currently still requires prohibitory amounts of sequencing for it to be used in the hi-resolution functional study of single human or mouse loci. While published Hi-C maps of mouse and human genomes have improved in resolution from about a megabase (Lieberman-Aiden et al. 2009) to 1-5 kb by extremely deep sequencing (Rao et al. 2014), most current Hi-C maps have a resolution of around 40kb (Dixon

et al. 2015, Fig 1.1). Thus, more targeted approaches such as 4C and 5C are still preferred for the high-resolution analysis of local chromatin contacts at single loci (Denker and de Laat 2016).

One such approach is Capture-C, which allows contact profiles of multiple sites to be generated in parallel (many vs all, Hughes *et al.* 2014; Davies *et al.* 2015). In the Capture-C protocol, 3C libraries are prepared with a four-cutter enzyme (generally DpnII), and, following ligation and de-crosslinking, fragmented by sonication. Following the ligation of sequencing adaptors, libraries are hybridised to biotinylated oligos designed against the fragments of interest (i.e. the Capture-C viewpoints) and purified by immobilisation to beads. While this allows the generation of genome-wide interaction profiles of many loci simultaneously, the interpretation of contacts between viewpoints studied in the same experiment is compromised as these are enriched with different efficiencies. Thus, to establish contacts between two genomic sites in *cis*, independent Capture-C experiments have to be performed. A major advantage of Capture-C is provided by the use of sonication to fragment the 3C template. As sonication creates random DNA fragment ends, each unique interaction is characterised by the fragmentation positions on either end of the ligated fragment. This allows independent ligation events (different fragment ends) to be discerned from PCR duplicates when used in combination with paired-end sequencing, making Capture-C a quantitative method to measure chromatin contact frequencies. The latest iteration of this protocol, next-generation Capture-C, further improves the enrichment of fragments of interest by adding a second round of hybridisation and this protocol is used in this thesis (Davies *et al.* 2015, Fig. 1.1).

Another method that attempts to enrich for interactions between sites of interest is ChIA-PET, which combines 3C technology with chromatin immunoprecipitation (ChIP-seq) by pulling down fragments bound by a protein of interest (Fullwood et al. 2009). While this potentially allows the identification of rare interactions between sites bound by the protein of interest, it is difficult to quantitatively interpret ChIA-PET data (Denker and de Laat 2016). As sites in close proximity in the linear genome may form ligation junctions, these proximity ligation events may be detected as *bona fide* loops due to the ChIP enrichment. In addition, the degree of enrichment by ChIP limits the number of interactions that can be detected from each site, making the analyses often based on scoring relatively few interactions.

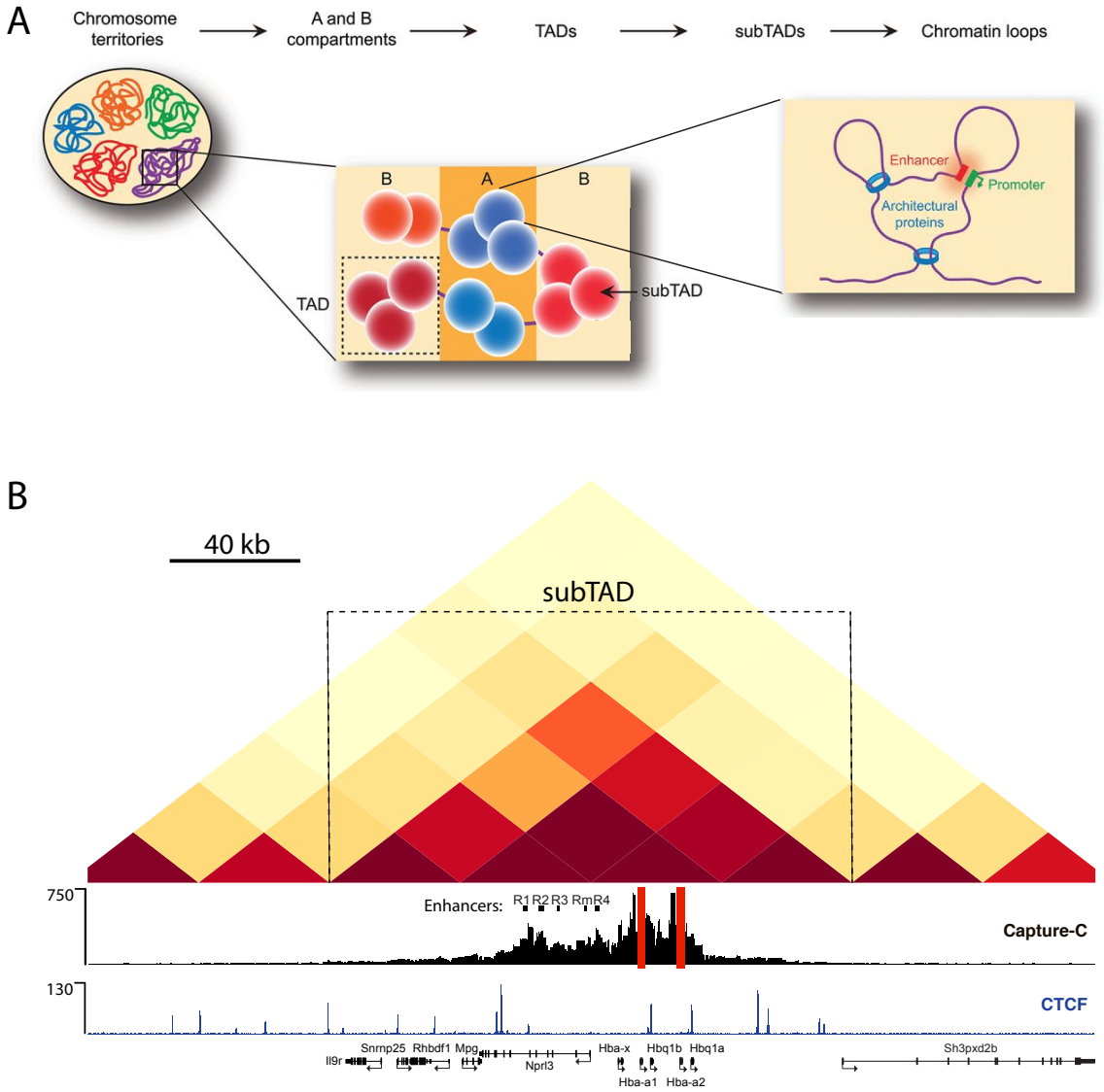


Figure 1.2 Different scales of nuclear organisation and interaction. A. Schematic overview of the hierarcHical organisation of chromatin at different scales. Chromosomes are organised into large interacting TADs that segregate into active A (blue) and inactive B (red) compartments. TADs are further subdivided into smaller domains of interaction termed sub-TADs or "contact domains" that are shaped by loops between enhancers and promoters and architectural proteins such as CTCF. Figure from Denker and De Laat 2016. **B.** Comparison between Hi-C and Capture-C interaction data. Hi-C data from erythroid cells at a resolution of 40 kb shows that the α-globin cluster is located in a sub-TAD of ~200kb. Capture-C in erythroid cells using the α-globin promoters as a viewpoint (red bars). Whereas Hi-C data barely resolves contacts between the α-globin enhancers and promoters, Capture-C data identifies interactions with the individual enhancers (annotated R1-4 and Rm). Notice that the sub-TAD structure is reflected in the Capture-C plot. Plot is annotated with gene annotations and CTCF ChIP-seq in erythroid cells. Capture-C data was taken from Davies et al. 2015 and unpublished Hic data was generated by James Davies and Job Dekker (unpublished).

1.1.3 Hierarchical chromatin interactions organise genome topology

From the initial discovery of interactions between the β-globin genes and enhancers by 3C (Tolhuis et al. 2002) to the latest high-resolution Hi-C map of the human genome (Rao et al. 2014), 3C derived technologies have been instrumental in mapping the interactions between DNA elements in detail. Moreover, these methods have provided support for the existence of chromosome territories and the separation of active and inactive chromatin in nuclear space (Simonis et al. 2006; Lieberman-Aiden et al. 2009; Rao et al. 2014). The analysis of chromosome-wide interactions by Hi-C revealed that mammalian chromatin falls into two genome-wide compartments, labelled A and B (Lieberman-Aiden et al. 2009). The A compartment contains regions of chromatin that are gene-dense, highly expressed and DNaseI accessible, whereas the B compartment consists of more inactive chromatin regions that are relatively gene-poor and lowly expressed. As higher-resolution Hi-C data has become available, a further subdivision into six compartments has been proposed (Rao et al. 2014); two associated with active chromatin (A1 and A2) and four with a more repressed chromatin state (B1-B4). These sub-compartments differ from each other in replication timing and chromatin modifications and their propensity to localise to nuclear organelles such as the lamina and the nucleolus.

However, perhaps the most important discovery made by Hi-C studies, is the observation that chromosomes are divided into structural domains termed topologically associated domains (TADs) that were observed both in mouse (Dixon et al. 2015) and *Drosophila* cells (Sexton et al. 2012). In mammals, TADs are

genomic regions roughly 1 Mb in size that are highly self-interacting and delimited by domain boundaries, across which contacts occur much less frequently. Again, the analysis of higher resolution mammalian interaction data, using either 5C (Phillips-Cremins et al. 2013) or Hi-C (Rao et al. 2014), has shown that domain structures analogous to TADs exist on a scale much smaller than TADs. These domains, termed sub-TADs or “contact domain”, vary in size between ~40 kb and 3 Mb, with a median size of 185 kb. This discovery illustrates the hierarchical, fractal, nature of chromatin architecture in which small chromatin loops are enclosed by contacts over progressively larger genomic distances. While this suggests that the distinction between TAD and sub-TAD structures may be artificial and mainly based on the resolution of the analysed data, it has been shown that TADs are more conserved between closely related species than sub-TAD structures (Rudan et al. 2015). In addition, while TADs appear to be largely invariant between cell types (Dixon et al. 2012), clear differences in chromatin topology can be observed between different cell types at this smaller scale (Phillips-Cremins *et al.* 2013; Jin *et al.* 2013; Rao *et al.* 2014), implicating that a functional hierarchy exists for these structural domains as well.

1.1.4 Linking genome topology and gene regulation

While the mechanisms by which structural genome organisation affects gene regulation are only partially understood, compelling evidence now confirms this relationship and has started to bridge the understanding of gene regulation at a local scale and the chromatin architecture within which it takes place. Ever since the discovery that gene regulation is dependent on distal regulatory elements, it was

suggested that these elements may physically interact with the genes that are under their control (Ptashne 1986). This was later confirmed by experiments at several model loci, including the α- (Vernimmen *et al.* 2007a) and β-globin genes (Wijgerde, Grosveld and Fraser 1995; Tolhuis *et al.* 2002), that identified interactions between these genes and several DNaseI hypersensitive sites located more than 20kb away from the gene promoters. These interactions were characterised in more detail in later studies (van de Werken *et al.* 2012; Hughes *et al.* 2014; Davies *et al.* 2015) and have been shown to occur specifically in cells where the genes are active.

More direct evidence for the functional importance of physical interactions between distal enhancers and gene promoters was more recently provided in an elegant study in which the β-globin enhancers were artificially tethered to the gene promoters in cells lacking Gata1 (Deng et al. 2012). In Gata1 mutant erythroid cells, no interaction is established between the enhancer region and the promoters and, consequently, the β-globin genes are expressed at low levels. Ldb1, which has a self-association domain required for the establishment of the enhancer-promoter interaction, is recruited to the enhancers but not the promoter in the absence of Gata1. By targeting a zinc-finger protein fused to Ldb1 (or its self-association domain) to the β-globin promoter, a strong interaction was established that resulted in the strong activation of β-globin. This demonstrated the requirement of physical interactions between distal enhancers and the genes they regulate. A follow up study also demonstrated that the artificial tethering of the enhancers to the developmentally silenced γ-globin gene resulted in the aberrant activation of this gene in adult primary erythroid cells (Deng et al. 2014). At the same time, a reduction of interactions between the adult β-globin gene and the enhancers resulted

in a reduction in transcription. Thus, these findings illustrate the functional relationship between gene expression and the establishment of physical contacts between enhancers and promoters. Additionally, they suggest that careful regulation of these interactions is required, and that aberrant interactions between genes and active regulatory elements could have dramatic effects on their transcriptional output.

A number of studies now provide insight into the interplay between these gene regulatory interactions and chromatin architecture. Cell-type specific changes in chromatin topology at a sub-TAD level has been linked to changes in gene expression patterns (Phillips-Cremins et al. 2013). Indeed, the latest high-resolution Hi-C datasets allow chromatin interactions between enhancers and promoters to be identified and show that these interactions correlate strongly with gene expression (Rao et al. 2014). Functionally, the importance of topological organisation is illustrated by a study of the boundary between TADs separating the non-coding transcripts of *Xist* and *Tsix* at the X-inactivation centre. The ratio of the expression between these two transcripts determines the activation state of the X-chromosome in *cis*. *Xist* and *Tsix* are regulated by antagonistic enhancer elements located in two adjacent TADs. The deletion of a boundary between *Xist/Tsix* increased contacts between the adjacent TADs and was accompanied by widespread transcriptional misregulation (Nora et al. 2012). A similar topology has been described at the mouse *HoxD* cluster, where the spatiotemporal activation of regulatory enhancers located in flanking TADs differentially activates genes in the cluster (Montavon et al. 2011; Andrey et al. 2013) Another study of the developmentally regulated mouse Sonic Hedgehog (*Shh*) locus also provided evidence that TADs may delimit functional gene-regulatory domains. This study showed that all functionally validated distal

regulatory elements are contained within the same TAD (Anderson et al. 2014). Conclusive evidence for this notion was provided by a study of the *Epha4* locus. Chromosomal rearrangements resulting in deletion or inversion of the TAD boundaries flanking the *Epha4* gene, resulted in abnormal limb development in human and mouse. Various disruptions of the TAD boundary were shown to result in ectopic interactions between the *Epha4* enhancers and the neighbouring *ihh*, *Wnt6*, and *Pax3* genes, suggesting that chromosome topology regulates gene expression by linking enhancers to their target genes (Lupiáñez et al. 2015). Chromosome domain architecture may even assist in establishing these interactions; the well-characterised *Shh* gene and limb bud enhancer co-localise with the boundaries of a tissue-invariant TAD and constitutively interact (Dixon et al. 2012). Similar instances of invariant loops between promoters and their regulatory enhancers have been described at other loci, including the HoxD cluster (Montavon *et al.* 2011; Jin *et al.* 2013).

A final, tentative link between chromosome domain architecture and genome function is provided by the observation that some TADs correlate with domains of chromatin modifications linked to transcriptional activation or repression (Dixon et al. 2012). Although this correlation has most clearly been observed in *Drosophila* (Sexton et al. 2012), similar findings have now been made in higher resolution studies of mammalian cells (Rao et al. 2014), suggesting that the initial discrepancy in findings between mammals and *Drosophila* may have been caused by *Drosophila*'s smaller genome size.

In summary, the organisation of chromosomes in the nucleus has been studied with a wide range of microscopy and sequencing based technologies that agree on the basic principles of genome organisation. These various experimental approaches have jointly confirmed the strong links between chromosome organisation and gene regulation, illustrating the importance of understanding the mechanisms that underlie chromosome organisation at every scale.

1.2 CTCF: regulator of genome architecture

Since the zinc-finger protein CCCTC-binding factor (CTCF) was first identified as a vertebrate insulator protein (Klenova *et al.* 1993; Chung, Bell and Felsenfeld 1997), it has been suggested that CTCF may contribute to the topological organisation of loci through the formation of interactions between its binding sites. The discovery that CTCF is enriched at the boundaries between TADs further implicated this factor in the regulation of chromosome domains (Dixon et al. 2012). A growing body of evidence now suggests that CTCF is a key factor in the establishment of mammalian genome architecture and this has begun to unravel the mechanisms by which it operates.

1.2.1 CTCF is a conserved 11 zinc-finger protein

The protein CTCF is conserved throughout bilatarians (i.e. animals with bilateral symmetry), with the notable exception of *C. elegans* and other derived nematodes, but is not present in other eukaryotes, including plants and yeast (Heger et al. 2012). The 11 zinc fingers of CTCF recognise a uniquely information-rich DNA sequence motif, that is similarly well-conserved and has been shown to be indistinguishable between cell types and several model organisms (Filippova *et al.* 1996; Schmidt *et al.* 2012; Nakahashi *et al.* 2013; Rudan *et al.* 2015). A detailed study of CTCF binding to the genome discovered that CTCF is anchored to a 20 bp core sequence at most (>80%) CTCF binding sites (Kim *et al.* 2007; Xie *et al.* 2007; Schmidt *et al.* 2010; Nakahashi *et al.* 2013). Zinc-fingers 4-7 have been proposed to recognise this core motif while non-specific association of the other zinc-fingers with DNA stabilises

binding (Nakahashi et al. 2013). In addition, shorter 10 bp motifs lying up- and downstream of the core binding site are found in a smaller subsets of sites and have been proposed to further anchor or destabilise CTCF binding respectively (Rhee and Pugh 2011; Schmidt *et al.* 2012; Nakahashi *et al.* 2013). While, in the case of the core and upstream motifs, both the presence as well as the similarity to the consensus motif increase CTCF binding and its residency time on chromatin. Conversely, the presence of a strong match to the downstream consensus sequence decreases CTCF binding. Notably, both the extended consensus motif (including up- and downstream sites) as well as the CTCF core motif are non-palindromic, allowing the orientation of CTCF binding to DNA to be determined. The residency time of CTCF on chromatin was found to be ~11 minutes as measured in FRAP experiments (Nakahashi *et al.* 2013). This is an order of magnitude longer than various other transcription factors (TF) which were found to have a residence time in the order of 1 minute (Boyle *et al.* 2011), suggesting CTCF binds chromatin with high affinity.

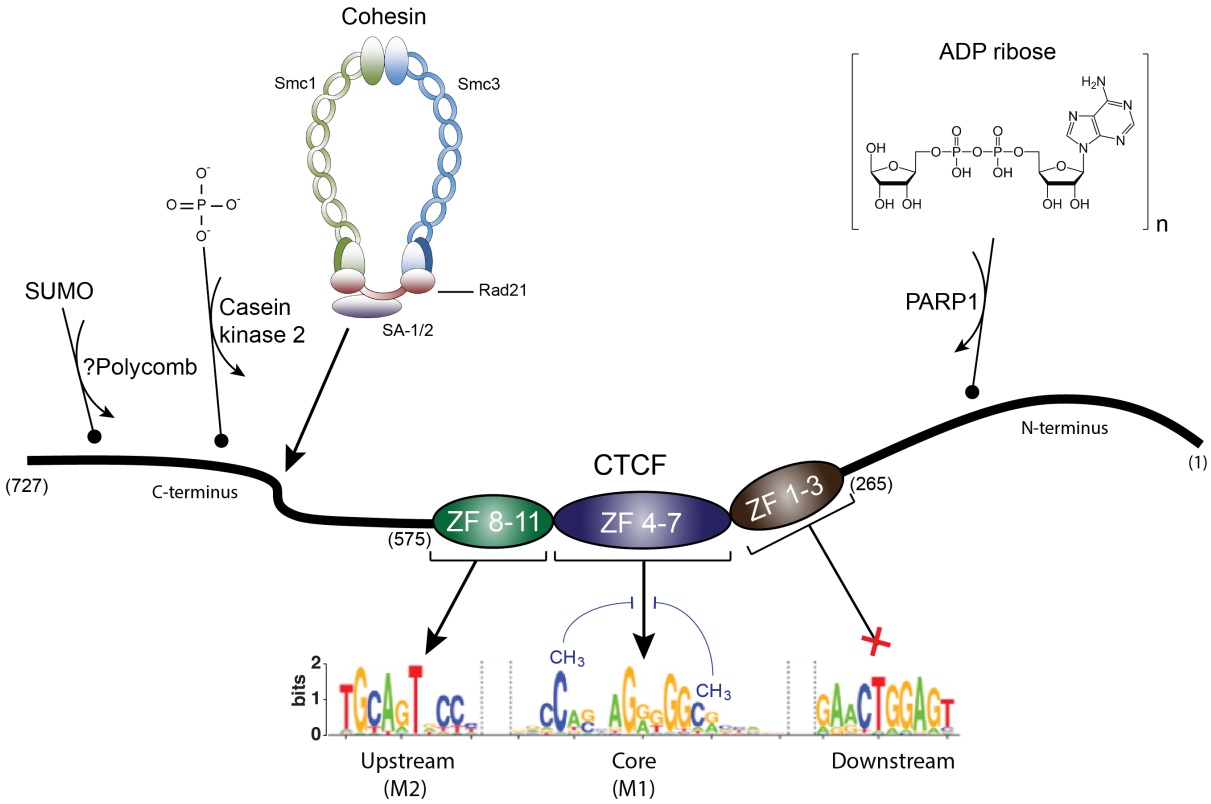


Figure 1.3 Overview of the CTCF protein and sequence binding motif. CTCF is a 727 amino acid 11-zinc-finger protein that is modified in several ways. The C-terminal domain is phosphorylated and SUMOylated, whereas the N-terminus is targeted for poly(ADP-ribosyl)ation. The Cohesin complex associates with the C-terminus through SA-1/2. Zinc-fingers 4-7 bind the core motif, while zinc-fingers 8-11 bind to the upstream motif. The presence of the downstream motif destabilises CTCF binding to DNA. Methylation of CpGs in the consensus sequence prevents CTCF binding. Figure modified from Nakahashi et al. 2013; Ong and Corces. 2014; Xiao et al. 2011; Williams and Somervaille, 2015. and Wang et al. 2012.

1.2.2 The distribution and regulation of CTCF binding

CTCF binds between 45,000 and 65,000 sites in the mammalian genome and binds preferentially to gene-dense regions (Kim *et al.* 2007; Wang *et al.* 2012). Roughly half of these sites are constitutively bound by CTCF regardless of tissue-type, whereas 30-60% of these sites are cell-type specific (Kim *et al.* 2007; Cuddapah *et al.* 2008; Wang *et al.* 2012). It has been found that well-conserved CTCF binding sites are less likely to vary in binding between tissues and contain stronger matches to the CTCF consensus motif (Schmidt *et al.* 2012; Rudan *et al.* 2015). While CTCF binding sites are mostly found in intergenic regions (~45%), some are found at promoters (~20%) and in the introns and exons of genes (~35%, Kim *et al.* 2007).

Tissue-specific CTCF binding sites have been shown to be subject to DNA methylation, which prevents the binding of CTCF to the sequence motif (Bell and Felsenfeld 2000; Hark et al. 2000; Engel et al. 2004). A comparison of bisulphite sequencing data between 19 human cell lines found that 41% of tissue-specific sites are linked to differential methylation, suggesting that a subset of CTCF sites is regulated in this way (Wang et al. 2012). This analysis also revealed that two CpG residues in the CTCF core motif were targeted for methylation. It is unclear whether methylation is a regulatory mechanism for CTCF binding or simply occurs in the absence of CTCF, as it has been suggested that CTCF binding could protect against DNA methylation (Stadler et al. 2011). It was suggested that CTCF is able to regulate the methylation state of DNA through association with poly(ADP-ribose) polymerase 1 (PARP1) and DNA (Cytosine-5)-methyltransferase 1 (DNMT1, Zampieri et al. 2011). Upon association with CTCF, PARP1 becomes active and

targets DNMT1 inactivation through poly(ADP-ribosyl)ation, preventing the local methylation of DNA.

1.2.3 Functions of CTCF

CTCF was originally identified as a transcriptional repressor of *Myc* (Lobanenkov et al. 1990). Subsequent studies showed that CTCF was able to bind insulator elements in vertebrates which had already been extensively described in *Drosophila* (Udvardy, Maine and Schedl 1985; Geyer and Corces 1992). Insulator elements were classified as such by their ability to block activating enhancer activity when placed in between gene promoters and distal *cis*-acting elements in transgenic reporter assays. At the chicken β-globin locus, one such study identified a 42 bp fragment with enhancer-blocking function that was shown to contain a binding site for CTCF (Chung, Bell and Felsenfeld 1997; Bell, West and Felsenfeld 1999). Consistent with these findings, a sequence containing the human or mouse imprinting control region (ICR) of the *Igf2/H19* locus was found to act as an insulator in transgenic assays in a CTCF-dependent manner. Insulator activity and CTCF binding to this element was abolished when the ICR was methylated, as is the case on the paternal allele, showing that CTCF is involved in the transcriptional regulation of imprinted genes (Bell and Felsenfeld 2000; Hark et al. 2000). These initial findings have inspired many follow up studies that identified a wide range of additional functions of CTCF.

1.2.3.1 CTCF as a chromatin barrier

The results of these early transgenic assays for insulator elements, were often interpreted to indicate that insulator proteins have the ability to separate genomic regions with active and repressive chromatin states (Kellum and Schedl 1991). Such barrier function has indeed been described at the *HMR* locus in yeast, where CTCF is absent, through the recruitment of the acetyltransferase SAGA (Oki and Kamakaka 2005). Several publications have suggested that the unique properties of CTCF binding may provide a mechanism by which it could serve a similar function in metazoans. CTCF carefully positions the nucleosomes around its binding site and is enriched for histone variants H2A.Z and H3.3 (Fu et al. 2008; Weth et al. 2014). Incorporation of H3.3 and the spacing of nucleosomes was linked to the removal of the repressive H3K27me3 modification and the opening of compact chromatin at the binding sites (Weth et al. 2014). Indeed, several studies have observed that domains of H3K27me3 at the *Hox* clusters contain gaps in the domains of these histone modifications at sites where CTCF is bound (Bowman *et al.* 2014; Narendra *et al.* 2015).

Genome-wide localisation studies of CTCF binding have provided some support for the notion that CTCF acts as a chromatin barrier. In a study of HeLa and human CD4+ T cells, a subset of H3K27me3 domains was found to be flanked by CTCF at at least on one of the domain borders (Cuddapah et al. 2008). The mapping of genome-wide CTCF interactions by ChIA-PET in mouse embryonic stem (ES) cells supported the existence of CTCF-flanked functional domains. Clustering analyses of the 1,480 identified CTCF-CTCF interactions with RNA polymerase II and several

histone modifications indicative of chromatin state identified a subset of loops that contained either active or repressed chromatin within a differentially marked chromatin environment outside the loop (Handoko et al. 2011). These correlative studies are complemented by a functional study at the mouse *Wnt4* locus, where knockdown of CTCF was shown to result in the spreading of histone modifications to neighbouring genes, affecting their expression (Essafi et al. 2011). Although these results are in support of a *bona fide* chromatin boundary function for CTCF, it has been argued that other mechanisms could explain these observations (Ong and Corces 2014). Indeed, the dissection of the chicken β-globin HS4 boundary element showed that, while enhancer blocking was dependent on CTCF, chromatin barrier activity was dependent on recruitment of USF1 (Gaszner and Felsenfeld 2006; Huang *et al.* 2007). This suggests that CTCF may cooperate with other factors to establish chromatin insulation.

1.2.3.2 CTCF as a multifunctional regulator of genome architecture

While CTCF has been extensively studied as an enhancer blocker in transgenic assays, the mechanism by which CTCF achieves this is unclear. The initial discovery that CTCF is able to establish chromatin loops between its binding sites at the β-globin (Splinter 2006a) and *Igf2/H19* loci (Kurukuti et al. 2006) has emerged as a key mechanism to explain CTCF function. Indeed, CTCF's ability to bring distal sequences in close physical proximity can explain, at least part, the many different roles that have been proposed for CTCF in gene regulation and nuclear organisation. The many functions assigned to CTCF binding include; the tethering of distal enhancers to promoters (Kuzmin et al. 2005; Sanyal et al. 2012), a role in

transcriptional pausing and splicing (Shukla *et al.* 2011; Paredes, Melgar and Sethupathy 2013; Laitem *et al.* 2015), the regulation of somatic recombination at the antigen receptor loci (Ribeiro de Almeida *et al.* 2012; Guo *et al.* 2012a; Seitan *et al.* 2012), the establishment of interactions between chromatin domains and the nuclear lamina (Guelen et al. 2008), and the regulation of gene expression at several complex loci such as the *Hox* gene clusters (Moon *et al.* 2005; Kim *et al.* 2007), the MHC class II locus (Majumder and Boss 2010), and the Protocadherins (Yagi et al. 2012). At the β-globin locus, deletion of a CTCF binding site flanking the locus was first shown to cause a partial loss of the chromatin topology of the locus in erythroid cells (Splinter 2006a). These findings were corroborated by the observation that a mutation in a CTCF binding site at the *Igf2/H19* locus, where the CTCF-bound maternal ICR normally prevents the expression of the *Igf2* gene. Mutation of CTCF binding sites in the ICR allowed the maternally-contributed *Igf2* gene promoter to establish an interaction with an enhancer cluster located on the other side of the ICR (Kurukuti et al. 2006). Similar observations of CTCF-mediated interactions were made at the MHC class II receptor locus (Majumder et al. 2008). These data, and observations linking CTCF-mediated loops to other proposed functions mentioned above (Guo *et al.* 2011; Seitan *et al.* 2012), have resulted in the suggestion that specific functional outcomes of CTCF binding may depend on the context within which CTCF-CTCF interactions are established, such as the nature of genomic elements that are brought together and the relative position of the CTCF binding sites (Ong and Corces 2014).

1.2.4 Binding partners as modulators for CTCF function

1.2.4.1 CTCF co-localises with cohesin

Another explanation of the diversity of CTCF's multiple functions may be its ability to interact with other proteins (Zlatanova and Caiafa 2009). The most ubiquitous and well-studied of these is the co-localisation of CTCF at many binding sites with the cohesin complex (Wendt et al. 2008; Rubio et al. 2008; Parelho et al. 2008; Stedman et al. 2008). Before being linked to CTCF, the ring-shaped Cohesin complex was mainly associated with DNA replication and sister chromatid cohesion during mitosis (Nasmyth and Haering 2009). Cohesin is composed of several "Structural Maintenance of Chromosome" family subunits, Smc1 and Smc3, and a kleisin family subunit, known as Rad21 in vertebrates (Scc1 in yeast). These are further bound by Scc3 which has three variants in mammals; SA-1, SA-2, and SA-3, and form a ring-shaped structure that is thought to encircle two DNA strands during mitosis.

The discovery that cohesin associates with chromatin in post-mitotic cells suggested that it may have a function outside of this classic role in mitosis. Between half and 80% of cohesin binding sites co-localise with CTCF and depletion of CTCF results in loss of cohesin at these sites. Conversely, CTCF binds to chromatin independently in the absence of cohesin, leading to the suggestion that CTCF anchors cohesin to its sites (Wendt et al. 2008; Rubio et al. 2008; Parelho et al. 2008). CTCF was shown to interact with cohesin directly through association of its C-terminal domain with SA-1 or SA-2 (Xiao, Wallace and Felsenfeld 2011). These observations and cohesin's known role in sister-chromatid cohesion, quickly led to the suggestion that cohesin

may be involved in the establishment or stabilisation of interactions between CTCF binding sites. Indeed, depletion of cohesin alters chromatin interactions and gene expression of the *Igf2/H19* locus (Nativio et al. 2009). Similarly, cohesin was shown to be important for the establishment of chromatin loops at the interferon gamma locus (Hadjur et al. 2009) and the β-globin locus (Chien et al. 2011).

While these studies show that cohesin is an important contributor to CTCF-mediated loop formation, it has been suggested that CTCF may be able to form DNA loops via cohesin-independent mechanisms, such as through homo-dimerisation (Yusufzai et al. 2004). Conversely, CTCF is not the only protein associated with cohesin on chromatin. Cohesin components and the cohesin-loading factor Nipbl associate with the Mediator complex at enhancers (Kagey et al. 2010), which has prompted the suggestion that cohesin may also be involved independently of CTCF in stabilising promoter-enhancer interactions (Schmidt et al. 2010; Faure et al. 2012).

1.2.4.2 Other binding partners of CTCF

CTCF has been shown to share binding sites with many other proteins, which have been suggested to modulate CTCF function in various ways (Wallace and Felsenfeld 2007; Lee and Iyer 2012). Most of these co-associations are linked to the ability of CTCF to function as an insulator. These include interactions with Kaiso (Defossez et al. 2005), the histone deacetylase SIN3A (Lutz et al. 2000), nucleophosmin (Yusufzai et al. 2004), and the Thyroid Hormone Receptor (Lutz et al. 2003). The DEAD-box RNA helicase, p68, and the associated non-coding RNA, SRA, also aid insulator function, potentially through stabilisation of the interaction between CTCF

and cohesin (Yao et al. 2010). Association of CTCF with the transcription factors FOXA1 and the oestrogen receptor has been linked to transcriptional activation (Ross-Innes, Brown and Carroll 2011). Moreover, the recruitment of transcription factor TAF3 to distal regulatory elements by CTCF in ES cells was shown to lead to TAF3-dependent loop formation and gene activation (Liu et al. 2011). A subset of highly conserved CTCF binding sites associated with transcriptional activation co-bind YY1 (Schwalie et al. 2013), which was also shown to be a required co-factor for CTCF function in X-inactivation (Donohoe et al. 2007). While these associations have been shown to be important for CTCF function at specific loci and in specific cell-types, only cohesin has so far been shown to co-localise to a large degree across the genome and different cell-types.

Finally, several CTCF binding partners have been shown to post-translationally modify CTCF and affect its function. Poly(ADP-ribosyl)ation of CTCF by PARP1 has been shown to be important for insulator function (Yu et al. 2004; Ong et al. 2013). CTCF is further modified by SUMOylation (MacPherson et al. 2009) and phosphorylated by caseine kinase 2 (CK2, El-Kady and Klenova 2005).

Thus, while association of CTCF with cohesin is likely important for CTCF function across cell-types, association with tissue-specific transcription factors and post-translational modifications, combined with the regulation of its binding to DNA, can be used to modulate CTCF function in a cell-type and locus-specific manner.

1.2.5 Interactions between convergent CTCF sites shape the genome

Since the role of CTCF in the establishment of chromatin contacts was first discovered, it has been hypothesised that CTCF may also have a role in the regulation of genome architecture on a larger scale. When Hi-C data were computationally intersected with CTCF binding data, a clear enrichment of CTCF at the boundaries of TADs was found (Dixon et al. 2015). While most (76%) TAD boundaries contained a CTCF binding site, only 15% of all CTCF binding events was located in boundary regions. This led to the suggestion that CTCF binding is not sufficient to establish a TAD and that it may have other functions at a sub-TAD scale. However, higher-resolution (4 kb) studies applying 5C to several loci in mouse ES and Neural Progenitor Cells (NPCs) identified a role for CTCF, in combination with cohesin component SMC1 and Mediator, in the establishment of sub-TAD interactions (Phillips-Cremins et al. 2013). In a different experimental approach, the depletion of CTCF was shown to reduce intra-domain contacts while increasing contacts between domains (Zuin et al. 2014). More recent Hi-C studies allow genome-wide interactions to be mapped at an even higher resolution (down to 1 kb), revealing a large number of sub-TAD structures termed “contact domains” (Rao et al. 2014). Furthermore, this resolution allowed individual loops to be identified, the majority of which (86%) are anchored by both CTCF and cohesin. Roughly 40% of these loops were shown to demarcate a contact domain (Rao et al. 2014).

As CTCF can only bind to its sequence motif in one direction (as shown in Fig. 1.3), two CTCF binding sites in the genome can have four possible orientations with respect to each other (shown in Fig. 3.2B). An intriguing observation made by Rao et

al. was that the loops detected between CTCF binding sites occurred predominantly (>90%) when these CTCF binding sites were configured in a convergent orientation, i.e. positioned with the N-terminal domain towards each other. Similar observations were made by an evolutionary comparison of Hi-C data between syntenic regions in four vertebrates. The most conserved CTCF sites had the highest affinity for CTCF and marked the boundaries of Hi-C domains. These strongly conserved sites had a predominantly convergent orientation with respect to each other (Rudan et al. 2015).

1.2.6 Functional studies of CTCF binding

Over the past year, a number of small-scale functional studies investigating the role of CTCF in genome organisation complemented these genome-wide correlative approaches. First, the importance of DNA methylation for the regulation of CTCF binding was recently illustrated; a mutation in the isocitrate dehydrogenase (IDH) gene common in certain human gliomas was shown to inhibit CpG demethylation pathways. The increased global levels of methylation resulted in loss of CTCF binding at susceptible sites and was shown to disrupt a TAD boundary. Loss of CTCF-mediated insulation allows the subsequent establishment of contacts between a distal enhancer (~900 kb away) and the promoter of the *PDGFRA* gene, which has a known function in glioma oncogenesis (Flavahan et al. 2015).

An analysis of the *Six* homeodomain locus in zebrafish, revealed that TAD boundaries were occupied by pairs of divergent CTCF binding sites (and thus convergent within the TAD). A deletion of one of these boundaries results in aberrant interactions between enhancers and promoters between TADs (Gómez-Marín et al.

2015). The precise removal of several CTCF binding sites at the *HoxA* cluster in mouse ES cells, results in the disruption of a boundary between topological domains and causes activation of polycomb-repressed *Hox* genes. Similarly, the removal of one of a pair of interacting CTCF binding sites flanking active genes within a TAD was shown to cause mis-regulation of nearby genes (Dowen et al. 2014). The TAD boundaries of previously-discussed functional studies at the XIC and the *Epha4* locus were both shown to be bound by CTCF (Nora et al. 2012; Lupiáñez et al. 2015).

The importance of CTCF convergence was addressed in a study of the human Protocadherin and β-globin loci, where the inversion of single CTCF binding sites results in a completely different pattern of 4C interactions. New contacts are formed according to the rule of convergence; with convergent CTCF binding sites on the other side of the inversion (Guo *et al.* 2015). Correspondingly, 4C loop interactions between CTCF sites were lost when one of the sites was inverted (de Wit et al. 2015) and deletion and inversion of individual CTCF binding sites in a cluster of two looped domains resulted in Hi-C contact changes that were consistent with the requirement of convergence for CTCF-CTCF interactions (Sanborn et al. 2015).

1.2.7 Proposed model for generating convergence in CTCF interactions

While these studies confirm the prevalence of interactions between convergent CTCF binding sites in genome organisation, it is important to note that not all CTCF-CTCF interactions abide by this rule. Although interactions between divergent sites are rare (2%), two ChIA-PET studies for CTCF identify a significant number (33%) of

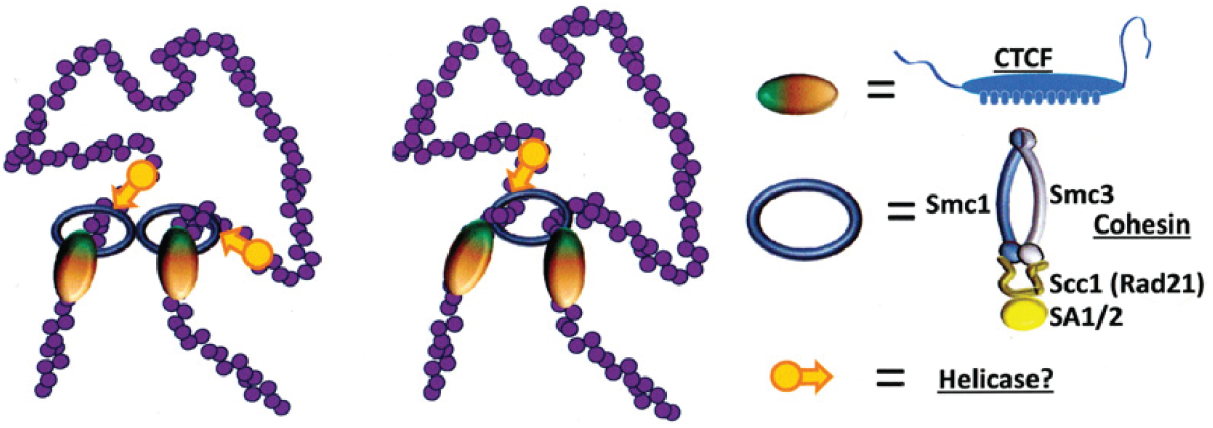


Figure 1.4 Proposed mechanism of loop extrusion for the formation of orientation-dependent CTCF loops. Loading of cohesin extrudes a loop that progressively grows as cohesin proceeds along the chromosome arms until a correctly orientated CTCF site is reached and anchors the complex. CTCF is pre-bound or co-migrates with cohesin. One or two cohesin molecules may be involved in the loop extrusion. As this process would require an energy source, a currently unknown helicase is thought to be involved in this process. Figure from Ghirlando and Felsenfeld 2016.

interactions between CTCF binding sites in a tandem orientation (i.e. facing the same way, Guo *et al.* 2015; Tang *et al.* 2015). Nonetheless, the preference for interactions between CTCF sites in specific orientations is incompatible with a simple model of loop formation through random collisions between the sites, as these interactions would not be able to distinguish between CTCF site orientations in 3-dimensional nuclear space.

Different variations of a loop extrusion model have been proposed to resolve this discrepancy (Nasmyth 2001; Ghirlando *et al.* 2012; Sanborn *et al.* 2015; Nichols and Corces 2015). This model proposes that the loading of the cohesin complex creates an incipient loop by forming a “handcuff” structure around the chromosome arm (Fig. 1.4). To explain the directionality of CTCF-CTCF interactions, it has been proposed that a processive enlargement of this loop along the chromosome then continues until a correctly orientated CTCF binding site is encountered and anchors the loop. As this model would essentially transform a 3-dimensional process into a linear (1-

dimensional) process, it would theoretically be able to recognise the orientation of neighbouring CTCF binding sites. It is unclear whether one or two cohesin molecules are involved in this mechanism. Computational models based on a loop extrusion model were able to better explain Hi-C contact map observations than previously proposed models to describe chromosome architecture (Sanborn et al. 2015).

In summary, large strides have been made in elucidating the molecular mechanism by which interactions between CTCF binding sites are formed. While it is clear that CTCF-mediated chromatin architecture has key roles in the regulation of gene expression and genome function, the interplay between architectural and other distal *cis*-regulatory elements is still only partially understood, especially on the scales at which gene regulation commonly occurs.

1.3 Gene regulation of the α-globin locus

The α-globin locus is a well-conserved regulatory domain which has been extensively studied in a variety of model organisms and humans. As the gene cluster is flanked by several binding sites for the CTCF protein, I used the α-globin locus as a model to investigate the requirement and function of CTCF for regulation of gene expression and local genome topology.

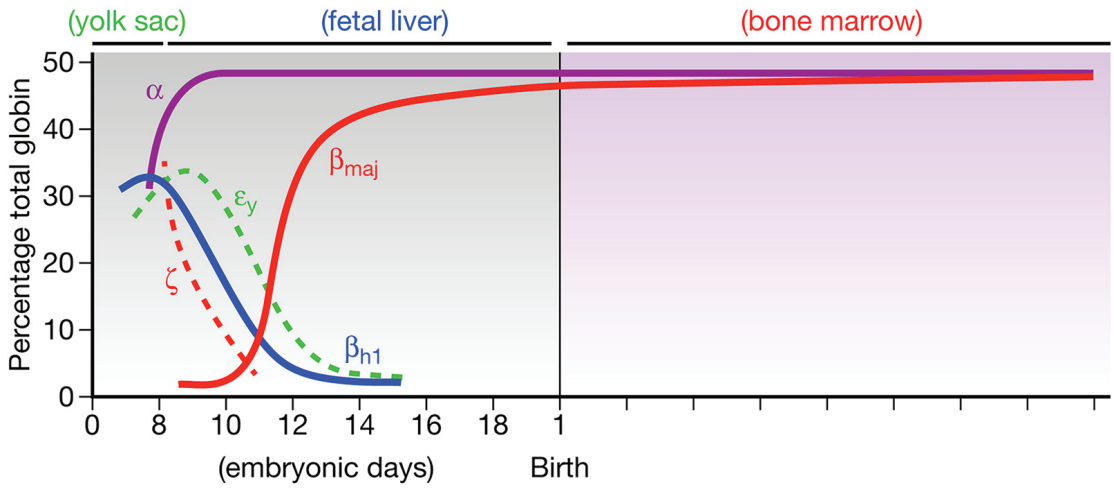


Figure 1.5 Developmental expression of the mouse globin genes. The timing and abundance of β-type and α-type globin genes during embryonic, foetal, and postnatal murine development. Figure from Blobel et al. 2015.

1.3.1 Haemoglobin and globin genes

Two α-globin and two β-globin chains combine in adult red blood cells to form a haemoglobin tetramer, which is responsible for the delivery of oxygen from the lungs to the peripheral tissues. As early embryonic and foetal development have different requirements due to more hypoxic conditions, different α- and β-type chains are expressed in a developmental-stage specific manner (Wells 1979; Blobel *et al.* 2015) (see Fig. 1.5). The mouse α-locus transcribes the embryonic ζ-globin gene only in early embryonic development (until ~E10.5) and starts expressing the two adult α-globin genes around the same time, expression of which continues into the adult stage. These two α-type chains combine with β-type chains (εγ, β_{H1}, β_{maj}) to produce different haemoglobin complexes in development. Stringent regulation of the expression of these proteins is essential, as deviation from an equimolar cellular concentration of α- and β-chains leads to the formation of insoluble homotetramers which is the cause of the genetic disorders α- and β- thalassemia (Cao and Galanello 2010; Galanello and Cao 2011; Higgs 2013)

1.3.2 Genetic structure of the α-globin locus in human and mouse

The study of mutations in patients with α-thalassemia and complementary functional studies in the mouse have identified important *cis*-regulatory elements for the regulation of the α-like globin genes. Both in mouse and human, the regulation of α-globin expression is under the control of four *cis*-regulatory elements that were identified as erythroid-specific DNaseI hypersensitive sites (DHS) located upstream of the α-globin genes (Fig. 1.6A). This basic genetic structure of the α-globin locus

has been shown to be remarkably conserved throughout evolution and lies within a region of synteny of between 135 and 155 kb In all mammalian species that were analysed (Hughes et al. 2005). Moreover, a 117 kb genomic fragment of the human region was shown to reproduce the appropriate developmental-stage-specific expression of the α-globin genes in a transgenic mouse model (Wallace *et al.* 2007, see Fig. 1.6A). The four regulatory elements are also conserved across mammals and, as such, have been termed multispecies conserved sequence (MCS) R1 to R4. Three of these enhancers (MCS-R1, MCS-R2, and MCS-R3) are located in the introns of the neighbouring housekeeping gene *Nprl3*, while MCS-R4 is positioned just upstream of the *Nprl3* promoter in both human and mouse. However, some differences also exist between these two species. The mouse contains an additional DHS in the first intron of *Nprl3* and 12kb upstream of the ζ-globin promoter that is called HS-12 or Rm, for mouse-specific regulatory element. This site is thought to be important in priming the α-globin locus for activation as it is one of the first DHSs to bind erythroid erythroid transcription factors (Anguita et al. 2004). In addition, the mouse α-globin cluster contains two θ-pseudogenes (arranged 5'-ζ-α1-5'θ-α2-3'θ-3') whereas the human locus contains a θ- and α^D-pseudogene (arranged 5'-ζ-α^D-α2-α1-θ-3'). Note, however, that the order of the functional α-like genes remains unchanged (5'-ζ-α-α-3', Hughes *et al.* 2005; Vernimmen 2014).

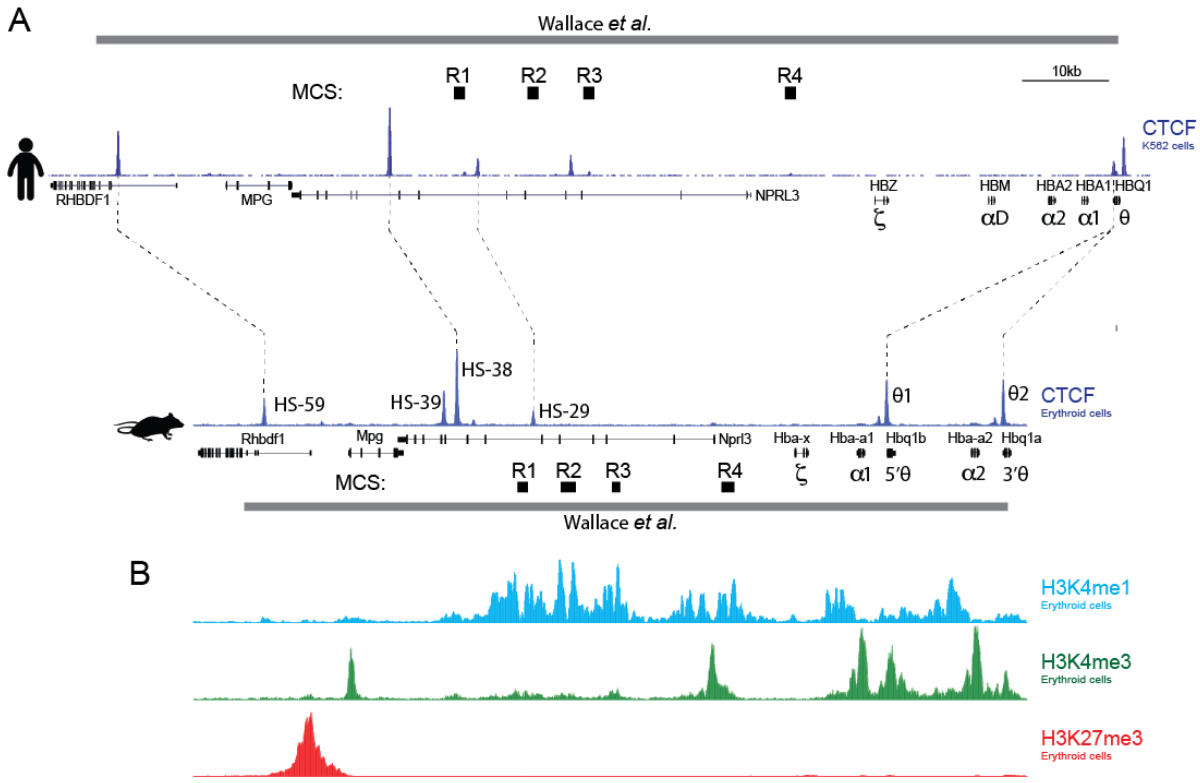


Figure 1.6 Overview of the mouse and human α-globin loci.. **A.** Overview of conservation of CTCF binding sites between the mouse and the human α-globin locus. Shown are annotated CTCF ChIP-seq tracks in human and mouse erythroid cells. Conservation of CTCF binding is indicated by dotted lines between human and mouse ChIP-seq peaks and based on phastCons conservation data (UCSC genome browser). Blocks marked Wallace et al. signify the region of synteny that was replaced in the humanised mouse model. The human and mouse α-globin loci are annotated with their gene names and conserved enhancer regions (labelled R1-4). **B.** Chromatin modifications at the mouse α-globin locus. Shown are ChIP-seq data for enhancer mark H3K4me1, promoter mark H3K4me3, and repressive mark H3K27me3. Data is from Kowalczyk et al. 2012.

The functional importance of the α-globin enhancers first became apparent in rare patients with α-thalassemia where upstream deletions were linked to the disease, leaving the globin genes and their promoters fully intact (Hatton *et al.* 1990). The further study of upstream deletions within the *NPRL3* gene resulting in the disease showed that the elements MCS-R1 and MCS-R2 are consistently removed in these cases (Higgs and Wood 2008). Subsequent characterisation of these elements in transgenic assays have shown that MCS-R2 (also known as HS-40) has the

strongest enhancer activity among the human regulatory elements (Higgs *et al.* 1990; Sharpe *et al.* 1992; Bernet *et al.* 1995).

In mice, the deletion of MCS-R2 has a much milder effect on α-globin expression than observed in humans (Anguita 2002). A recent comprehensive study of the mouse α-globin regulatory regions, deleting the mouse enhancers individually and in informative combinations, revealed that no single element is critical for α-globin expression in the mouse. Both deletion of R1 and R2 resulted in significant decreases in transcriptional output (~30% and ~50% respectively) and the deletion of these two elements together reduced transcription to about 10% of normal levels. By contrast, deletion of R3, R4, or the mouse specific regulatory element Rm only resulted in minor changes in α-globin expression *in vivo*, showing that R1 and R2 are the strongest functional enhancer elements during embryonic, foetal, and adult erythropoiesis in mouse (Hay *et al.* 2016).

1.3.3 Transcription factor binding at the α-globin locus

The binding of transcription factors to enhancers has a leading role in the initiation of gene expression in development and cell differentiation (Spitz and Furlong 2012). Enhancer regions typically contain clusters of transcription factor binding motifs that are thought to be bound cooperatively. Different combinations of transcription factors may then recruit transcriptional cofactors (such as p300) or components of multiprotein complexes involved in transcriptional activation (such as Mediator) in order to regulate transcription of the target gene, although the exact mechanisms by

which enhancer activation is achieved are still the topic of much debate (Spitz and Furlong 2012).

The developmental regulation of expression of the α- and β-globin genes is under control of a set of erythroid transcription factors that includes GATA-binding factor 1 (Gata1), Gata2, stem-cell leukemia factor (Scl, also known as Tal1), nuclear factor-erythroid 2 (Nfe2), and Krüppel-like factor 1 (Klf1). Initial recruitment of Gata2 and Gata1 is thought to drive α-globin enhancer activation, after which the other factors are recruited during erythroid differentiation (Anguita *et al.* 2004; De Gobbi *et al.* 2007). Not all transcription factors are bound at each enhancer; only MCS-R1 and MCS-R2 bind Nf-e2 in erythroid cells. The binding of these transcription factors to the α-globin enhancers is accompanied by the recruitment of RNA polymerase II and general transcription cofactors (Vernimmen *et al.* 2007b; 2009; Vernimmen 2014). This includes components of Mediator, a multi-protein complex that is ubiquitously involved in the regulation of RNA polymerase II transcription (Allen and Taatjes 2015). It was recently proposed that enhancer regions bound by high levels of this master-regulator form a special class of enhancers that are drive high levels of expression of key regulators of cell lineages (Whyte *et al.* 2013). These super-enhancers typically comprise a cluster of individual enhancers and have now been described in many cell-types (Whyte *et al.* 2013; Qian *et al.* 2014; Siersbæk *et al.* 2014). However, it is unclear whether super-enhancers represent a truly new class of element, the properties of which are more than the sum of the individual regulatory elements of varying strengths that make up these composite enhancers (Pott and Lieb 2015). Based on this definition, the α-globin enhancers may fit the definition of a super-enhancer, as they are multiple Mediator-occupied enhancer elements located

within a short genomic distance of each other that drive high levels of α-globin expression. Interestingly, it has been reported that super-enhancers are often flanked by interacting CTCF binding sites, suggesting their activity needs to be insulated by these boundary elements (Dowen *et al.* 2014).

1.3.4 Epigenetic regulation and histone modifications at the α-globin cluster

Different functional elements in the genome have been correlated to the presence of specific post-translational modifications of the N-terminal amino acid tails of histone proteins. In addition, different modifications have also been linked to states of transcriptional activation or repression (Bannister and Kouzarides 2011). Together with the DNA methylation, these modifications are thought to function as epigenetic signals that allow expression states to be stably established and maintained through differentiation and inherited across cell divisions (Riddihough and Zahn 2010; Allis and Jenuwein 2016).

The chromatin state of genes at the α-globin locus has been well characterised. The mouse α-globin cluster is located within a region of open chromatin and is flanked by several housekeeping genes which are expressed in all cell types. In erythroid cells, the promoters of these genes and the α-globin genes are marked by histone H3 lysine 4 trimethylation (H3K4me3) which is associated with active gene promoters (Fig. 1.6B, De Gobbi *et al.* 2007; Kowalczyk *et al.* 2012b). Similarly, the α-globin enhancers are marked by H3K4me1 and H3K27ac in erythroid cells, both marks associated with active enhancers (Bannister and Kouzarides 2011). Finally, the *Rhbdf1* gene is marked by histone H3 lysine 27 trimethylation (H3K27me3), a mark

associated with transcriptional repression via the establishment of facultative heterochromatin. In humans, the α-globin genes are repressed by H3K27me3 in non-erythroid cells (Garrick *et al.* 2008), but this is not the case in mouse *(Lynch et al.* 2011).

H3K27me3 is deposited by the polycomb repressive complex 2 (PRC2), one of two polycomb complexes involved in the establishment of polycomb repression (Margueron and Reinberg 2011). PRC2 is composed of catalytic subunit enhancer of zeste homolog 2 (Ezh2) and four additional core subunits: suppressor of zeste 12 protein homolog (Suz12), embryonic ectoderm development (Eed), and retinoblastoma binding protein 4 (Rbbp4) (Margueron and Reinberg 2011; Klose et al. 2013). This repressor complex stabilises its own binding in a positive feedback loop as Ezh2 is able to deposit H3K27me3 which is in turn bound by Eed. PRC2 functions together with polycomb repressive complex 1 (PRC1) to induce gene repression and both of these complexes are thought to be recruited to unmethylated CG-rich sequences in the absence of features of active transcription such as RNA polymerase II (Blackledge *et al.* 2015).

The repressive activity of polycomb group proteins is opposed in development by the trithorax group proteins that are responsible for the deposition of the active H3K4me3 at gene promoters. In mice, SET1A/B and MLL histone methyltransferases deposit this mark in conjunction with gene activation (Schuettengruber et al. 2011). While H3K4me3 and H3K27me3 have antagonistic roles in regulating gene expression, bivalent domains that are enriched for both histone modifications are observed at the promoters of mammalian genes. It has

been proposed that bivalent domains poise genes for activation in development, allowing rapid transcriptional upregulation when required whilst maintaining repression in the absence of differentiation signals (Voigt *et al.* 2013).

1.3.5 Conservation of CTCF binding between human and mouse α-globin locus

The mouse α-globin cluster, including the flanking housekeeping genes, is entirely situated within an erythroid TAD (unpublished, see Fig. 1.2). Despite this local chromatin organisation, genes within this TAD are marked by different epigenetic modifications and vary dramatically in their transcriptional regulation in erythroid cells, suggesting that these genes are insulated from the activity of the α-globin enhancer.

CTCF binds at several locations within the erythroid TAD in mice and I hypothesise that these intra-TAD CTCF binding sites may be responsible for the separation of transcriptional programmes of genes located within the cluster. Both humans and mice have a conserved CTCF binding site located upstream of the α-globin enhancers (Fig. 1.6A). This site, called HS-38 in the mouse, separates the enhancers from the promoters of the directly adjacent *Mpg* and *Rhbdf1* genes. In mice, a species-specific second site (HS-39) is located ~1.5kb upstream of HS-38. Similarly, the downstream end of the gene cluster is also demarcated by CTCF binding sites in both mice and humans. While the 5' and 3' ends of the human θ-globin gene are each flanked by a CTCF binding site, a duplication of this gene in mice has resulted in a duplication of the conserved 5' θ site (named θ1 and θ2). The θ1 CTCF site is located between the second α-globin gene (Hba-a2) and the α-

globin enhancers in mice. Interestingly, the first α-globin gene (Hba-a1) interacts more strongly with the enhancers than Hba-a2, suggesting this site may interfere with these interactions (Davies *et al.* 2015). Two other CTCF binding sites present in the mouse α-globin cluster, named HS-29 and HS-59, are both conserved between mouse and human. HS-29 is located in between the R1 and R2 enhancers and was observed to be constitutively DNaseI hypersensitive and marked by histone acetylation throughout erythroid differentiation, suggesting this site may have a role in the establishment of α-globin enhancer activity in differentiation (Anguita et al. 2004).

Thus, the α-globin cluster contains several CTCF binding sites that separate differentially expressed and epigenetically marked genes and *cis*-regulatory elements. To establish whether these insulator elements are involved in the functional separation of local genes and enhancer activity, I set out to target CTCF binding sites at the α-globin locus *in vivo*.

1.4 A revolution in genome-editing

Over the past few years, several advances were made in genome editing technologies that has enabled biologists to manipulate virtually any sequence both in cell culture systems and *in vivo* in a wide range of organisms. Genome editing technology relies on the ability to target engineered nucleases to specific genome sequences and induce double strand breaks (DSBs). Subsequently, breaks will either be repaired via the error-prone non-homologous end-joining (NHEJ) repair pathway or via homology-directed repair (HDR). Whereas NHEJ repair results in small *indel* mutations at the target site, repair via HDR can be used to introduce specific sequence changes by introducing a DNA template containing the desired mutation (Ran *et al.* 2013b; Hsu, Lander and Zhang 2014; Sander and Joung 2014).

The first such tools to become available were zinc-finger nucleases (ZFNs) which linked several Cys_2-His_2 zinc-finger domains to a *FokI* cleavage domain. However, limited targeting density and variable targeting efficiencies prevented ZFNs from becoming widely adopted (Kim and Kim 2014). The development of transcription activator like effector nucleases (TALENs) resolved these issues. TALENs are constructed by combining the *FokI* cleavage domain with DNA-binding domains that each recognise a single base pair in the DNA sequence allowing almost any sequence to be targeted. The only restriction to targeting is that TALEN binding sites have to start with a T. Although the cloning of TALENs was originally challenging due to the repetitive nature of the DNA binding domains, the development of a novel molecular cloning protocol (‘Golden Gate’ cloning) in which sequences of multiple DNA-binding domains can be combined in a single reaction has enabled the rapid

assembly of TALE arrays (Cermak et al. 2011). Finally, the emergence of the RNA-guided nuclease clustered regularly interspaced short palindromic repeat (CRISPR)-CRISPR-associated (Cas) system has further revolutionised genome-editing. The most commonly used variant of this system allows *S. Pyogenes* Cas9 nuclease to be targeted to genomic sites by a single-guide RNA that contains a sequence complementary to the target site (Cong *et al.* 2013; Ran *et al.* 2013a; 2013b). The CRISPR-Cas9 system has been demonstrated to be an efficient tool for HDR mutagenesis in both mouse embryonic stem (ES) cells and mouse embryos (Wang *et al.* 2013; Yang *et al.* 2013). Thus, these novel genome-editing tools can be used to introduce directed mutations *in vivo*, allowing carefully orchestrated changes in *cis*-regulatory elements such as CTCF to be introduced and studied in development.

1.5 Summary and aims

Although CTCF is implicated in the regulation of both genome topology and gene expression, the mechanism by which these are linked is not clear. Given that CTCF is able to mediate interactions between its binding sites and the growing evidence for the importance of genome topology in gene regulation, I hypothesise that CTCF contributes to the regulation of gene expression through its ability to restrict and establish local chromatin contacts. In this thesis, I aim to elucidate the relationship between CTCF-mediated local chromatin topology and gene regulation at the α-globin locus.

The α-globin cluster and neighbouring housekeeping genes are all located within an erythroid cell specific sub-TAD. The close proximity of non-erythroid housekeeping and repressed genes to the α-globin enhancers within this self-interacting domain suggests that CTCF binding sites within the sub-TAD may further restrict interaction within this domain. To investigate intra-TAD interactions at higher resolution and investigate how enhancer-activation impacts on local chromatin topology, I analyse interactions from regulatory elements across the α-globin locus by Capture-C in ES cells and primary erythroid cells. The resulting interaction profiles are interpreted in the context of local CTCF binding and orientation.

To investigate how intra-TAD CTCF binding impacts on this local chromatin structure, I employ genome editing technologies to delete CTCF binding sequences at the α-globin cluster. Mutations are verified to result in a loss of CTCF and primary erythroid cells of mice homozygous for the CTCF mutation are analysed by Capture-

C to investigate changes in local chromatin interactions upon loss of CTCF binding sites. I next analyse expression of local local genes in various CTCF binding site mutants and aim to link observed changes in local chromatin architecture to changes in gene expression. Finally, to assess whether CTCF insulation is required for the epigenetic regulation of local genes, I compare the chromatin state between wild-type and CTCF-mutant erythroid cells.

The aim of the performed work is to elucidate the interplay between local genome topology and function by relating changes in CTCF-mediated interactions to gene expression and epigenetic marks.

Chapter 2: Materials and Methods

2.1 Cell culture and cell selection methods

2.1.1 Culture of mouse embryonic stem cells

Mouse ES cells were routinely cultured on *neo*^{R}-*puro*^{R} cassette containing DR-4 feeders (Tucker et al. 1997) in knockout DMEM supplemented with 10% foetal calf serum (FCS), 2 mM Glutamine, 1 mM Non-Essential Amino Acids, 1000 U/ml Leukemia Inhibitory Factor (LIF, Millipore), antibiotics (100 U/ml penicillin and 100 µg/ml streptomycin), and 0.1 mM β-Mercaptoethanol at 37°C with 5% CO_2 incubation. When required, cells were frozen in media supplemented with 20% FBS and 10% dimethyl sulfoxide (DMSO).

2.1.2 Transfection of mouse embryonic stem cells

Transfections were performed using the Neon electroporation system (Invitrogen), according to the manufacturer's instructions. Trypsinised cells were resuspended in buffer R and 10^6 cells were used per transfection. A transfection protocol optimized for mouse ES cells was used (3 pulses of 1400V/10ms). A total of 5 µg plasmid DNA and 2 µl of ssODN reconstituted at a concentration of 100 µM were used per transfection. Cells were plated on a 10cm dish at single cell density and selected for the presence of puromycin or neomycin resistance cassettes from 24h to 72h post transfection. Colonies were picked into 96-well plates and screened for mutations.

Positive clones were thawed and karyotyped by Daniella Moralli (Chromosome Dynamics Core Facility, Wellcome Trust Centre for Human Genetics) before blastocyst injection.

2.1.3 Extraction and screening of genomic DNA from 96-well plates

Colony-picked clonal cell populations were plated in individual gelatinised wells of a 96-well plate. When a majority of wells were confluent, cells were lysed in 0.1% SDS and proteinase K at 37°C overnight. Genomic DNA was precipitated by addition of 10 µl 8M LiCl and 100 µl of isopropanol, washed with 200 µl 70% ethanol, and resuspended in 0.1 TE (1 mM Tris-HCl, pH 7.4; 0.1 mM EDTA, pH 8). PCR reactions specific to CRISPR-Cas9 targeted sites were performed and approximately 500ng of product was used directly for mutation screening by restriction digestion. After incubation for 2 hours at 37°C, the digestions were run on a 2% agarose gel. Positive clones were sent for Sanger sequencing.

2.1.4 Isolation and selection of ter119+ cells

Mature primary erythroid cells were obtained from phenylhydrazine-treated mice (Spivak, Toretti and Dickerman 1973). Spleens of phenylhydrazine-treated mice were mechanically disrupted to single cell suspension in RPMI media (Thermo Fischer Scientific) supplemented with 10% foetal bovine serum (FBS, Gibco). Cells were kept on ice and washed in cold PBS (Gibco) supplemented with 2% bovine serum albumin (BSA). To isolate late-stage erythroid cells, cells from a single spleen were resuspended in 5 mL of cold PBS/2% BSA and stained with 120 µL PE anti-

ter119 antibody (Ly-76, BD Biosciences) at 4°C for 15 minutes (Kina et al. 2000). After washing stained cells in PBS/0.5% BSA, cells were resuspended in 1.6 mL of PBS/0.5% BSA and 400 µL of anti-PE magnetic beads (Miltenyi Biotec) and incubated for 20 minutes at 4°C. Ter119 positive cells were isolated via auto-magnetic-activated cell sorting (autoMACS, Miltenyi Biotec) and processed for downstream applications. Purity of the isolated erythroid cells was routinely verified by FACS.

2.2 Mouse methods and husbandry

2.2.1 Mouse maintenance

C57BL/6 and CD1 mice were sourced from MRC Harwell/Charles River Laboratories. Mice were housed in individually ventilated cages and all husbandry and procedures occurred under Home Office Project License approval, in accordance with European Union Directive 2010/63/EU and the UK Animals (Scientific Procedures) Act, 1986.

2.2.2 Phenylhydrazine treatment of mice

Phenylhydrazine treatment was carried out by Jacqueline Sharpe at the molecular haematology unit (WIMM). Mice between 6-9 months old were given 3 doses of freshly neutralised phenylhydrazine (40 mg/g body weight) by intraperitoneal injection 12h apart and were harvested on day 5 following the start of treatment

(Spivak, Toretti and Dickerman 1973). Phenylhydrazine treatment resulted in haemolytic anaemia and erythroid expansion in the spleen so that >80% of cells are late-stage erythroid lineage (defined as CD71+ter119+).

2.3 Genome editing

2.3.1 Preparation of CRISPR-Cas9 expression constructs

Guide RNA sequences designed to target CTCF binding sequences were cloned into pX330-U6-Chimeric_BB-CBh-hSpCas9 (Addgene plasmid #42230, pX330) or pX335-U6-Chimeric_BB-CBh-hSpCas9n(D10) (Addgene plasmid #42335, pX335) vectors as described previously (Cong et al. 2013). pX330 and pX335 were modified to contain a puromycin and neomycin selection cassette respectively. Complementary DNA oligos containing the 20nt guide sequence (10 μM, Table 2.1) were treated with T4 PNK (New England Biolabs) at 37°C for 30 minutes and subsequently incubated at 95°C for 5 minutes and annealed by gradual cooling to 25°C. To clone the gRNA sequence into the pX330 or pX335 backbone vector, gel-purified *BbsI* digested vector (50 ng) was mixed with a 1:200 dilution of annealed gRNA oligo duplex and ligated (New England Biolabs Quick Ligase and 10X Quick Ligation buffer) in a final volume of 11 μL for 10 minutes at room temperature. Ligated vectors were transformed into competent DH5α *E. coli* bacteria. Bacteria were plated on ampicillin-containing LB Agar plates and colonies were picked and grown overnight at 37°C. Plasmid DNA was purified from bacterial cultures using the QIAprep miniprep kit (QIAgen) and screened by Sanger sequencing.

Table 2.1 gRNA sequences of targeted CTCF binding sites at the α-globin locus

ID	Target	Vector	Guide sequence
1	HS-38	pX330	GGTAGGCCTCTGCTACCCTC
2	HS-39	pX330/335	GAATGGCGCCCCCAGTGGCC
3	HS-39	pX335	CGCCATTAAAAGGTCCTGCT
4	HS-29	pX330	TCCCTCCAAATTGGTCCACT
5	θ1	pX330	TGGAACGATGCAGCGCCCCC
6	θ2	pX335	TGAAACACAAGAGGCCGCCA
7	θ2	pX335	GACATCTTTGAGCTCAGCCA
8	HS44	pX330	GAAAGCCAGTGGCGCCACCT
9	HS44	pX330	CCCTGCAGGCCACTATAAGT
10	HS48	pX330	TCCAAGGTCCTCAAGCAGAC
11	HS48	pX330	CGACGAGCACCCCCGTGTGG

2.3.2 Preparation of TALEN expression constructs

For TALEN construction, a 500bp sequence centred around the HS-38 CTCF consensus sequence was submitted to the TALEN Targeter (https://tale-nt.cac.cornell.edu/node/add/talen) using NN for G recognition (Golden Gate TALEN and TAL Effector Kit 1.0, Addgene). Two TALEN pairs with a differential spacer region that targeted the HS-38 CTCF binding sequence were selected and constructed via the Golden Gate assembly method (Cermak et al. 2011). TALEN-AF targeted the sequence 5'-TCCTGGGTAGGCCTCT-3' with the RVD array HD-HD-

NG-NN-NN-NN-NG-NI-NN-NN-HD-HD-NG-HD-NG and TALEN-AR targeted the sequence 5'-GAGTCCCACGTATCGT-3' on the reverse strand with the RVD array NN-HD-NG-NI-NG-NN-HD-NI-HD-HD-HD-NG-NN-NI-NN. Similarly, a second TALEN pair (TALEN-B) was designed, with the forward TALEN targeting the sequence 5'-TGAGGTCCTGGGTAGG-3' with the RVD array NN-NI-NN-NN-NG-HD-HD-NG-NN-NN-NN-NG-NI-NN-NN and the antisense sequence 5'-TCCCACGTATCGTGAT-3' with the RVD array NI-NN-NG-NN-HD-NG-NI-NG-NN-HD-NI-HD-HD-HD-NG. The vectors pC-Goldytalen (38143, addgene) and RCIscript-Goldytalen (38142, addgene) were used as target vectors in the final step of the Golden Gate cloning protocol.

2.3.3 Preparation and injection of TALEN mRNA

TALEN micro-injections were performed as previously described (Davies *et al.* 2013). The mMessage mMachine T7 Kit (Life Technologies) was used to generate mRNA from 1 µg of linearised RCIscript-Goldytalen constructs, according to the manufacturer's instructions. After purification of the mRNAs with the MEGAclear kit (Life Technologies), mRNAs were eluted in 2x50 µl of pre-warmed elution buffer. Microinjection and embryo transfer were carried out by Ben Davies and Chris Preece at the Transgenics Core (Wellcome Trust Centre for Human Genetics). Purified mRNA was diluted to 5 ng/µL in 1 mM Tris/HCl pH 7.5/0.1 mM EDTA and was microinjected into the cytoplasm of fertilised oocytes that were prepared from superovulated plugged females at 0.5 dpc. After injection, oocytes were immediately transferred to pseudopregnant CD1 foster mothers at 0.5 dpc.

2.3.4 Preparation and injection of sgRNA

DNA templates for use in *in vitro* transcription were generated from CRISPR-Cas9 expression constructs by PCR (described in section 2.3.1). The forward, gRNA-specific primer was modified with a 5' extension that contained a T7 polymerase binding site, and used to amplify the gRNA with a reverse primer binding downstream of the mature gRNA sequence (gRNA-R) (see table 2.2). The MEGAshortscript™ T7 Transcription Kit (Thermo Scientific) was used for *in vitro* transcription of the gRNAs. *In vitro* transcribed RNAs were purified with the MEGAclear Kit (Thermo Scientific) and diluted in 10 mM Tris-HCl pH7.5, 0.1 mM EDTA pH8.0 before microinjection. Female mice homozygous or heterozygous for the CAG-Cas9 transgene insertion were superovulated and mated with C57BL/6 or DEL38 studs. Oocytes were prepared for microinjection from plugged females and 20 ng/µl of gRNA was injected into the pronucleus. Depending on the experiment, ssODN templates for HDR (Eurogentec) were added at a final concentration of 20 ng/µl (see Table 2.3). For the single mutation of HS-39, D10A Cas9 protein (PNA Bio) was injected with two sgRNAs at a concentration of 40 ng/µl into C57BL/6 oocytes. The microinjected zygotes were immediately transferred to pseudopregnant CD1 foster mothers.

Table 2.2 Forward and reverse primers used in the generation of the DNA template for *in vitro* transcription of gRNAs

ID	Target	Primer sequence
2	**HS-39**	AGTCCATTAATACGACTCACTATAgGAATGGCGCCCCCAGTGGCC
3	**HS-39 (D10A)**	AGTCCATTAATACGACTCACTATAgGCGCCATTAAAAGGTCCTGCT
4	**HS-29**	AGTCCATTAATACGACTCACTATAgGTCCCTCCAAATTGGTCCACT
5	**θ1**	AGTCCATTAATACGACTCACTATAgGTGAAACACAAGAGGCCGCCA
8	**HS44**	AGTCCATTAATACGACTCACTATAgGAAAGCCAGTGGCGCCACCT
9	**HS44**	AGTCCATTAATACGACTCACTATAgGCCCTGCAGGCCACTATAAGT
10	**HS48**	AGTCCATTAATACGACTCACTATAgGTCCAAGGTCCTCAAGCAGAC
11	**HS48**	AGTCCATTAATACGACTCACTATAgGCGACGAGCACCCCCGTGTGG
	gRNA-R	AAAAGCACCGACTCGGTGCC

Table 2.3 Sequence of ssODNs used for micro-injection

ID	Target	ssODN sequence (5' -3')
2	HS-39	TAGTGAGATCTGGCCTCATGGATTCAAAGCCACTGAGGCCTG G**AGTACT**CGCCATTCGCCATTAAAAGGTCCTGCTGGGCTTTT CTAGCTCCAGATTCCAGATTTTTGGCAGCCACTGGTACTTAC AGACACACA
4	HS-29	TGTCATCTGCCAGGCACAGCTCAGGGCTTGAGGCCTCCAAGT G**CAGCTG**GACCAATTTGGAGGGACACAGGAATTTGAGCTTTT GGTGAAAAGGCATTCAAGGTCCTACTAGCCAGATACCCTGTT TGTTAGTGG
5	**θ1**	CGCTCTGCCCGCTGGCTGAGCTCAAAGACGTCCTGAAACACA AGAGGCGGATCCCGCTGCATCGTTCCAGGATGCCTAGGTGTT CACAGATTCTGGTTCAGCTTTGAGCCCTCTGTTTCCCTGGGC TCCCCCTCC
6	**θ2**	CTTGCGCTCTGCCCCCTGGCTGAGCTCAAAGATGTCCTGAAA CACAAGAGGCGGATCCCGCTGCATCGTTCCAGGATGCCTAGG TGTTCACAGATTCTGGTTCAGCTTTGAGCCCTCTGTTTCCCT GGGCTCCCT

2.3.5 Genotyping

Ear biopsies from mice were lysed in lysis buffer (10 mM Tris-HCl pH 8.0, 50 mM KCl, 0.45% NP40, 0.45% TWEEN 20) and proteinase K overnight at 55°C. The lysate was directly used as template for PCR reactions except in the case of a long range PCR covering both θ1 and θ2 (primer ID #5), in which case genomic DNA was first purified. Target genes were amplified with the primers listed in Table 2.4. When genotyping founder mice with heterozygous mutation or mosaic mutations, PCR reactions with ambiguous mixed sequencing traces were cloned into pCR2.1-TOPO (Thermo Scientific) and multiple plasmids were sequenced to reliably identify individual alleles.

Different strategies were used for the routine genotyping of mice, depending on the mutation. Small deletions in HS-38 and HS-39 in D3839 and D38 mice were detected by resolving wild-type and mutant alleles on a 3% agarose gel (primer IDs #8 and #9). The D39 mutant was genotyped by PCR (primer ID #7) followed by restriction digestion with ScaI. Similarly, the D29 mutant was digested with PvuII following PCR (primer ID #6). The θ2 mouse was genotyped by PCR (primer ID #4) followed by BamHI restriction digestion. The 100bp deletion of HS44 was resolved on a 2% agarose gel following PCR (primer ID #1). Screening for HS48 and θ1 mutations is performed with the primers in Table 2.3. PCR products are run on an agarose gel and sent for Sanger sequencing to screen for mutations. Cloning of the PCR products is performed where required as described above.

Table 2.4 PCR primers used for the genotyping of mice

ID	Target	Primer	Sequence
1	HS44	Forward	AAGCAAGCACTTCCCATCCA
		Reverse	TGAGGTGCCTTCTAGTCCCA
2	HS48	Forward	GCCCCACAAAACTGTACCCT
		Reverse	TGTGTCATAAGGAAGCCAGGG
3	θ1	Forward	GGGCTCCTGACACTCAGTCC
		Reverse	GGCATGGGCTTTAACCTGGC
4	θ2	Forward	TTCCGAAGGACTCGGGAAGC
		Reverse	ATGCACAGAGGCAATGCAGC
5	θ1 and θ2	Forward	ATGCATCTTTTTCGCCGCTGAAGTCTC
		Reverse	GGTACGGGCTCCTGACACTCAGTCCT
6	HS-29	Reverse	AGCCCACAACTTCCTGTCTT
		Reverse	GCTGGCCTGGAACTAACTCA
7	HS-39	Forward	ACAGCAACCATCTGGGTGAG
		Reverse	TGCTGGTGTCTGTGGACAAG
8	HS-39	Forward	GCACGATATGGGCAGGCAGC
		Reverse	GGAGCTAGAAAAGCCCAGCAG
9	HS-38	Forward	GAGAAGGCTGGCCTTTGA
		Reverse	CAGGACCCAGGGAATGAA

2.4 Chromatin Immunoprecipitation

2.4.1 Chromatin immunoprecipitation

Chromatin immunoprecipitation (ChIP) was performed on purified ter119-positive primary erythroid cells (1×10^7 cells/ChIP) using the ChIP Assay Kit (Cat. No. 17-295, Millipore). Briefly, for ChIP of Cohesin component Rad21, cells were subjected to dual cross-linking with 2 mM disuccinimidyl glutarate (DSG, Thermo Fischer Scientific) for 50 min and 1% (v/v) formaldehyde for 10 min, whereas a single 10 min 1% formaldehyde fixation was used for all other antibodies. The cross-linking reaction was quenched by addition of a final concentration of 0.125M glycine and cells were washed twice in cold PBS with cOmplete™ Proteinase Inhibitor Cocktail (PIC, Roche). Cells were lysed in sodium dodecyl sulfate (SDS) lysis buffer with PIC for 10 min. Fixed chromatin samples were fragmented using the Bioruptor sonicator (Diagenode) for 15 min at 4°C to obtain an average fragment size between 200 and 500bp. Sonication of dual cross-linked chromatin samples was improved by increasing sonication time to 20 min. Sonicated chromatin samples were diluted 1:10 in ChIP Dilution buffer containing PIC to 2 mL diluted chromatin per ChIP or no antibody control. Before addition of the antibody, 100 µL was taken from each 2 mL of sonicated, diluted chromatin to be used as 5% input. After pre-clearing the fragmented chromatin with Protein A Agarose, immunoprecipitation was performed at 4°C overnight. An overview of all antibodies and conditions used is included in table 2.5. Protein A agarose beads were added and incubated for 60 min at 4°C after which the beads were sequentially washed with Low Salt Buffer, High Salt Buffer, LiCl Buffer, and TE Buffer. Finally, bound chromatin fragments were eluted from the beads in Bicarbonate/SDS elution buffer (1% SDS, 0.1M $NaHCO_3$), crosslinks were

reversed by incubating samples at 65°C for 5 h, and samples were Proteinase K treated (Roche, 20 mg/ml) at 55°C for 1 h. Bound DNA fragments were purified by phenol-chloroform extraction and ethanol precipitation and analysed by qPCR or sequencing.

Table 2.5 Antibodies used for ChIP

Target	Manufacturer	Product code
CTCF	Millipore	07-729
Rad21	Abcam	ab992
H3K4me3	Abcam	ab8580
H3K27me3	Cell signaling	#9733
Ezh2	Cell signalling	#5246
Pol II	Santa Cruz	sc-56767

2.4.2 ChIP analysis by quantitative PCR (ChIP-qPCR)

Immunoprecipitated and input DNA fragments were resuspended in 150 µL water after ethanol precipitation. The amount of genomic DNA co-precipitated with antibody was quantified relative to input using SYBR green PCR master mix (Applied Biosystems) and custom designed primers (Table 2.6). Standard curves of dilution series of input chromatin material (5%, 1%, 0.2%, and 0.04% input) were used for relative quantification. All primer sets had efficiencies between 80% and 120% and contained a single peak in melt curve analysis. Enrichment of ChIP at genomic loci was expressed as a percentage of input or as a fraction of a positive control.

Table 2.6 Primers used for ChIP-qPCR analysis

Target	Primer	Sequence
Rhbdf1 K27	Forward	TGTTTACACAGACACGTGCG
	Reverse	TGCTGGGAAAGAAGTAGGGG
Ubtd2 K27	Forward	CGACTCGCACTTCTGGGTTA
	Reverse	GGGCTCCGCACTTTTAAGGT
control	Forward	AGCACACCCAGGGTTCTCTA
(Nprl3)	Reverse	CAGAGCTCCCAGACAACCAG

2.4.3 ChIP analysis by sequencing (ChIP-seq)

Immunoprecipitated and input DNA fragments were resuspended in 20 µL water after ethanol precipitation. For Rad21 in ter119+ cells, the reconstituted DNA libraries were sent to the Oxford Genomics Centre (Wellcome Trust Centre for Human Genetics, Oxford) for library prep and sequencing on the Illumina HiSeq 2000 platform. For all other experiments, the concentration of precipitated DNA fragments was determined by Qubit dsDNA HS assay (Thermo Fischer) and indexed sequencing libraries for the Illumina platform were prepared using NEBNext Ultra™ II DNA library prep kit (New England Biolabs) following the manufacturer's instructions. No size selection was performed after adaptor ligation. PCR amplification was performed for 7-11 cycles depending on the quatity of input DNA using NEBNext Multiplex Oligos (New England Biolabs). Libraries (4 nM) were sequenced on the Illumina NextSeq platform according to the manufacturer's instructions using High-Output v2 75 cycle kits (Illumina).

2.5 Capture-C

2.5.1 Preparation of 3C libraries

Chromosome conformation capture (3C) libraries were prepared as previously described with minor modifications (Davies *et al.* 2015). Briefly, 2 × 10^7 ter119+ cells isolated from wild-type or CTCF site-mutant C57BL/6 mouse spleens were fixed in RPMI with 10% FCS to which formaldehyde was added to a final concentration of 2% (v/v). The cross-linking reaction was quenched after a 10 min incubation at room temperature through addition of 1 mL 2M glycine. Fixed cells were washed twice in cold PBS, resuspended in 5 mL cold lysis buffer (10 mM Tris-HCl pH 8, 10 mM NaCl, 0.2% IGEPAL CA630 (Sigma), 1x PIC), incubated for 20 min on ice, and centrifuged (5 min, 500g, 4°C). The pellet containing cell nuclei was snap-frozen and stored for up to several months at -80°C or immediately resuspended in 1 mL of water, homogenised on ice (45 strokes in a Dounce homogeniser), and resuspended in 650 mL water. The cross-linked chromatin was subsequently split across three parallel tubes (containing 200 µL each) and digested by addition of 80 µL DpnII 10X restriction buffer, 10 µL of 20% SDS (v/v), and 414 µL water, to which 66 µL Triton X-100 was added after a 1 h incubation at 37°C. After a further 1 h incubation, three aliquots of 500 U DpnII enzyme were added several hours apart over a total incubation time of 16-24 hours (1,400 rpm shaking, 37 °C). To control for non-specific degradation, the remaining 50 µL cross-linked nuclei was treated identically except for the addition of DpnII enzyme.

To assess the digestion efficiency, DNA was extracted from undigested control (200 µL) and pooled material containing 100 µL from each digestion reaction. Controls were proteinase K and RNase treated before phenol-chloroform extraction followed by ethanol precipitation.

The remaining digested material was incubated at 65°C for 20 min to heat-inactivate remaining DpnII and cooled on ice. To ligate chromatin fragments in close physical proximity, each of the three digested samples was diluted with 500 µL water after which 133 µL 10X ligation buffer and 8 µL high-concentration T4 DNA ligase (Thermo Scientific, 30 U/µL) was added before incubating samples overnight (1,400 rpm shaking, 16°C). Ligated DNA was reverse cross-linked and proteinase K treated (Thermo scientific, >600 U/mL) overnight at 65°C. Ligation reactions were pooled and treated with RNase (30 µL, Roche) at 37°C for 30 min and subsequently phenol-chloroform extracted and ethanol precipitated. For the ethanol precipitation, the 4 mL sample was diluted with 7 mL water and precipitated with 1.5 mL of 2M sodium acetate and 35 mL 100% ethanol, to improve DTT removal from the sample. Ligated 3C-libraries and digestion controls were run on a 1% agarose (w/v) gel to determine digestion and ligation efficiency.

2.5.2 Addition of sequencing adaptors and indices

Indexed sequence libraries were generated as described in Davies *et al.* 2015. Briefly, two 5 µg aliquots of a 3C-library were sonicated using a Covaris S220 Focussed ultra-sonicator (six cycles of 60s, duty cycle: 10%, intensity: 5, cyles per burst: 200) and processed separately to retain maximum library complexity. Illumina

TruSeq adaptors and indices were added to libraries using NEBNext reagents (New England Biolabs), following the manufacturer's instructions with some exceptions. DNA clean-up steps were performed with Ampure XP beads (Beckman Coulter Genomics) at a 1.8:1 ratio to minimize loss of material. The Herculase II PCR kit (Agilent) was used to add TruSeq indices in a 7-cycle PCR reaction.

2.5.3 Oligonucleotide capture

Capture of fragments containing the viewpoint of interest were performed with Nimblegen SeqCap EZ reagents (Roche) using custom biotinylated DNA oligonucleotides, as described in Davies *et al.* 2015. Between 1-2 µg of up to six differently indexed libraries were pooled with 5 µg of mouse Cot-1 DNA (Invitrogen) and SeqCap EZ HE blocking oligos (Roche) to be processed in parallel. For each indexed library, 1000 pmol of universal TS HE blocking oligo and 1,000 pmol of HS blocking oligo matching the library index, were added. This allowed simultaneous capture of fragments of interest from two separate biological triplicate experiments. The mixture was split between three tubes and dried completely using a vacuum centrifuge (55°C). Dried pellets containing the DNA libraries were each resuspended in 15 µL of SeqCap EZ hybridization buffer and 6 µL SeqCap EZ hybridization component A and subsequently denatured at 95°C for 10 min (SeqCap EZ hybridization and wash kit). The denatured library mixes were then each added to 9 µL of biotinylated capture oligonucleotide library preheated to 47°C, and allowed to hybridize in the thermocycler for 64-72 h. For each experiment, the oligonucleotide library contained an equimolar mix of oligonucleotides targeting fragments of interest

(2.9 µM total oligonucleotide concentration). An overview of all oligonucleotides used is shown in Supplementary table 1.

After the 64-72 h hybridization reaction, hybridized biotinylated oligos were bound to 100 µL M270 streptavidin beads per library according to the manufacturer's instructions (SeqCap EZ hybridization and wash kit). Streptavidin beads were washed twice in 200 µL of bead wash buffer (per 100 100 µL beads) and the hybridization reaction was added to the beads immediately after the final wash, mixed thoroughly, and incubated at 47°C for 45 min with regular vortexing to allow the captured material to bind the beads. Magnetic beads were subsequently washed once with 100 µL pre-warmed wash buffer I (47°C, vortexed), twice with 200 µL pre-warmed stringent wash buffer (47°C, incubated 5 min), once with 200 µL wash buffer I (vortexed for 2 min), once with 200 µL wash buffer II (vortexed for 1 min), and once with 200 µL wash buffer III (vortexed for 30 s). After removal of the wash buffer III, beads were resuspended in 40 µL water and stored at -20°C or immediately split over two PCR reactions to amplify captured material without eluting material from the beads (KAPA master mix, SeqCap EZ accessory kit v2). Amplified captured material was recovered with an Ampure XP bead (Beckman Coulter Genomics) clean-up and quantified using the D1000 Tapestation (Agilent) and the Qubit dsDNA BR assay kit (Invitrogen).

As described in (Davies *et al.* 2015), a second round of capture was performed following the same protocol described above. A minimum of 500 ng of material from the first capture was used as input for the second capture and hybridization time was decreased to 24 h. After the second capture, amplified captured libraries were

diluted to 4nM after quantification by D1000 Tapestation (Agilent) and the Qubit dsDNA BR assay kit (Invitrogen) and sequenced on the Illumina MiSeq platform according to the manufacturer's instructions (300 bp, v2 chemistry kits).

2.6 RNA isolation and gene expression analysis

2.6.1 RNA isolation

Isolation of total RNA was performed by lysing cells in TRI reagent (Sigma). For ter119+ cells isolated from PH-treated mouse spleens, 1×10^7 cells were lysed in 1 mL of TRI reagent. Mouse ES cells (E14TG2a) were lysed in 1 mL TRI reagent in a single well of a 6-well plate when cells reached 70% confluency. For long-term storage, lysed cells in TRI reagent were snap-frozen and kept at -80°C. To isolate the RNA from lysed cells, 200 µL chloroform was added and incubated at room temperature after vigorous vortexing. Samples were subsequently spun for 15 min (12,000 rpm, 4°C) after which the aqueous phase was added to a tube containing 500 µL isopropanol and incubated for 5 min at room temperature to allow precipitation of RNA. After spinning samples for 10 min (12,000 rpm, 4°C), the pellet containing RNA was washed with 70% ethanol and dried at room temperature for 8 min. RNA was resuspended in 44 µL of DEPC-treated water and incubated at 55°C for 10 min. Finally, to remove genomic DNA from RNA samples, samples were treated with TURBO™ DNase with the DNA-free™ DNA removal kit (Ambion) according to the manufacturer's instructions. DNase-treated RNA samples were stored at -80°C.

2.6.2 RT-qPCR expression analysis

To assess relative changes in gene expression by qPCR, 1 µg of total RNA was used for cDNA synthesis using the Superscript III first-strand synthesis SuperMix (Invitrogen), according to the manufacturer's instructions. Briefly, 1 µg of RNA and 2 µL of Reverse Transcriptase (RT) enzyme mix were added to 10 µL 2X RT reaction mix and made up to a total volume of 20 µL. The reaction mixture was sequentially incubated for 10 min at 25°C, for 30 min at 50°C, 5 min at 85°C, and cooled down to 4°C. Samples were incubated for a final 20 min at 37°C after the addition of 1 µL of *E. coli* RNase H to remove RNA following cDNA synthesis. For every RNA sample, a second control reaction was run without the addition of the RT enzyme. The 20 µL reaction was diluted with 180 µL water and used for qPCR with SYBR green PCR mastermix at a 5X dilution. The $\Delta\Delta C_t$ method was used for relative quantitation of RNA abundance using the primers in table 2.7.

2.6.3 RNA-sequencing

For RNA-seq libraries, rRNA and globin mRNA species were removed using the Globin-Zero Gold kit (Illumina) with 5 µg of total RNA according to the manufacturer's instructions. Globin-Zero depleted samples were purified using Ampure XP beads (Beckman Coulter Genomics) and eluted in 50 µL DEPC-treated water.

To further enrich for mRNA, poly(A)+ were isolated using the NEBNext Poly(A) mRNA magnetic isolation module (New England Biolabs) and combined with the

NEBNext Ultra™ directional RNA library prep kit (New England Biolabs) according to the manufacturer's instructions. Fragmentation of mRNA was achieved by incubating samples at 95°C for 12 min. To achieve maximal strand specificity, Actinomycin D was added (5 µL of 0.1 µg/µL) to the first strand cDNA synthesis reaction. Poly(A)+ libraries (4 nM) were sequenced on the Illumina NextSeq platform using a NextSeq High Output v2 150 cycle sequencing kit, according to the manufacturer's instructions.

Table 2.7 Primers used for RT-qPCR gene expression analysis

Target	Primer	Sequence
Hba-a1/2	Forward	CTGGGGAAGACAAAAGCAAC
	Reverse	GCCGTGGCTTACATCAAAGT
Hbq1-a/b	Forward	GACCTGCCTGCTTCTCTGTC
	Reverse	GATAGTGCCTGGCGAGAGTC
Hbb-b1	Forward	ATGGCCTGAATCACTTGGAC
	Reverse	ACGATCATATTGCCCAGGAG
Rn18s	Forward	GTAACCCGTTGAACCCCATT
	Reverse	CCATCCAATCGGTAGTAGCG
Nprl3	Forward	GCTCTTCAGGTACCCCTTCC
	Reverse	ATGTTCGCCAGTGTTGTTGA
Mpg	Reverse	CTTCTCCAGCCCAGAGGAC
	Reverse	ATGCCTCAGTCTCCACAATG
Snrnp25	Forward	GAGGTAATGCCTGTGGTCGT
	Reverse	GGTCAGATGGTATGTCCGCC
Rhbdf1	Forward	ATCCTAGTGCCCCAGACCTT
	Reverse	CGGCCACTTGGTGATATCTT

2.7 ATAC-seq

ATAC-seq was employed to identify regions of open chromatin. This assay relies on the action of the mutated and hyperactivated transposase *Tn5*, which efficiently cuts exposed DNA found within regions of open chromatin such as active enhancers, promoters, and CTCF binding sites. After the cut, the transposase ligates adapter sequences into the site of the cut that allow these genomic regions to be sequenced by next-generation sequencing as previously described (Buenrostro *et al.* 2013).

ATAC-seq was performed on ter119+ cells isolated from PH-treated mouse spleens and ES cells as previously described (Buenrostro et al. 2013). Cells (65,000) were lysed in 50 µL of cold lysis buffer (10 mM Tris-HCl, pH 7.4, 10 mM NaCl, 3 mM MgCl2, 0.1% IGEPAL CA-630) to isolate nuclei before transposition with *Tn5* transposase (Nextera, Illumina) for 30 min at 37°C. After purification of DNA using a MinElute kit (Qiagen), libraries were amplified and indexed using NEBNext 2X Mastermix (New England Biolabs) and custom primers as described (Buenrostro et al. 2013). Library quality was assessed by D1000 Tapestation (Agilent) and libraries were quantified using the universal library quantification kit (KAPA Biosystems). ATAC-seq libraries (4 nM) were sequenced on the Illumina Nextseq platform using a NextSeq High Output v2 75 cycle sequencing kit, according to the manufacturer's instructions. ATAC-seq data in ES cells was generated by Damien Downes (Hughes lab, WIMM), following the above protocol.

2.8 Bioinformatics

The following software packages were used and are referred to throughout the methods section:

- Bedtools 2.25.0 (Quinlan and Hall 2010)
- Bowtie 1.1.2 (Langmead et al. 2009)
- ROSE tool (Whyte et al. 2013)
- Samtools 0.1.19 (Li et al. 2009)
- MACS 2.0.10 (Zhang *et al.* 2008)
- UCSC browser (Kent et al. 2002)
- UCSCtools 1.0
- R 3.2.1 (R Core Team 2014)
- STAR 2.4.2a (Dobin et al. 2013)
- SRAtoolkit 2.3.4-2 (Sequence Read Archive Submissions Staff 2011)
- Subread 1.4.5-p1 (Liao, Smyth and Shi 2014)
- Deeptools 2.2.2 (Ramirez et al. 2014)
- FLASH 1.2.8 (Magoc and Salzberg 2011)
- FASTQC 0.11.4 (http://www.bioinformatics.babraham.ac.uk/projects/fastqc)
- Trim Galore 0.3.1 (http://www.bioinformatics.babraham.ac.uk/projects/trim_galore)
- MEME 4.9.1_1 (Bailey et al. 2006)
- DESeq2 R package 1.8.2 (Love, Huber and Anders 2014)
- DiffBind R package 1.14.6 (Stark and Brown 2011)
- DpnII2E and CCanalyser (Davies *et al.* 2015), https://github.com/telenius/captureC/releases

All other data analysis was done using R, Bioconductor, and Perl scripts.

2.8.1 Mapping and visualisation of ChIP-seq and ATAC-seq data

Quality of sequence files was assessed by FASTQC. All ChIP-seq and ATAC-seq datasets were aligned to the mm9 mouse genome build using Bowtie with the following parameters: -m 2, -k 1, --best. Unmapped reads were then trimmed in Trim Galore (--length 10, --qualFilter 20) after which a second round of Bowtie mapping was performed. Reads that still failed to map were treated with FLASH (-m 9 –x 0.125) to merge paired-end reads that originated from short DNA fragments and mapped again using Bowtie. Duplicate reads were removed from aligned data using samtools (rmdup). For visualization, original DNA fragments are restored from paired-end reads using Bedtools (bedpe) and read coverage was counted using Bedtools (genomecov). Fragments were normalized as fragments per kilobase per million mapped (FPKM) using Deeptools bamCoverage (--normalizeUsingRPKM) and visualized in the UCSC genome browser. Published datasets were downloaded from the NCBI Gene Expression Omnibus (GEO) database and converted to FASTQ files using the SRAtoolkit. FASTQ files were processed as described above. An overview of all publicly available datasets used is shown in table 2.8.

Table 2.8 Publicly available datasets used in this thesis

Full references refered to in the table are (Kowalczyk *et al.* 2012b), (Hosseini et al. 2013), (Davies *et al.* 2015), (Consortium et al. 2012), and (Stadler et al. 2011).

Cell type	Data type	GEO submission codes	Publication
Ter119+ cells	H3K4me1 ChIP-seq	GSM689846	Kowalczyk et al.
Ter119+ cells	DNaseI	GSM1199553	Hosseini et al.
Ter119+ cells	α-prom Capture-C	GSM1659660, GSM1659661	Davies et al.
ES cells	H3K4me3 ChIP-seq	GSM769008	ENCODE consortium
ES cells	H3K4me1 ChIP-seq	GSM747542	Stadler et al.
ES cells	H3K27me3 ChIP-seq	GSM747539, GSM747540, GSM747541	
ES cells	CTCF ChIP-seq	GSM747534, GSM747535, GSM747536	

2.8.2 Mapping and visualisation of RNA-seq data

Quality of sequence files was assessed by FASTQC. All RNA-seq datasets were aligned to the mm9 mouse genome build using STAR with the following parameters: --outFilterType BySJout --outFilterMultimapNmax 20 --alignSJoverhangMin 8 --alignSJDBoverhangMin 1 --outFilterMismatchNmax 999. Deeptools bamCoverage (--filterRNAstrand, --normalizeUsingRPKM) was used to calculate normalized (RPKM) and strand-specific read coverage which was visualized in the UCSC genome browser. Published datasets were downloaded from the NCBI Gene Expression Omnibus (GEO) database and converted to FASTQ files using the SRAtoolkit. FASTQ files were processed as described above.

2.8.3 Identifying regions of ChIP-seq enrichment

The MACS (Model based analysis of ChIP-seq) peak finding algorithm was used to identify regions of ChIP-seq enrichment over background in an unbiased manner. The MACS2 callpeak function was used on biological duplicate ChIP-seq data of CTCF (-q 10^{-5}) and H3K4me3 (-q 10^{-3}). For H3K27me3, the MACS2 callpeak function was used with the --broad-cutoff option (--broad-cutoff 0.05) on biological duplicate ChIP-seq data.

2.8.4 Analysis of differential ChIP-seq enrichment

To identify regions that were differentially enriched between wild-type and CTCF binding site mutant mice, the R package DiffBind was used. Two biological duplicate datasets and independent peak calls of CTCF, H3K4me3 and H3K27me3 were used to identify differentially enriched regions with a false discovery rate (FDR) of 0.05. Differential analysis within the DiffBind package was performed with DESeq2.

2.8.5 Classification of enhancer elements

Enhancer and promoter annotations were generated by Jelena Telenius (Computational Biology Research Group, WIMM) as described (Hay et al. 2016). Briefly, 15,849 peak-called DNaseI hypersensitive sites (DHSs) were sub-classified as either putative enhancer or promoter based on the ratio between normalized H3K4me1 and H3K4me3 signal within each peak call. DHSs with a signal below a

treashold (less than 10 reads per million mapped) for both H3K4me1 and H3K4me3 were excluded. Regions with a H3K4me3/H3K4me1 ratio <1 were annotated as enhancers, excluding those located within 250bp within annotated transcription start sites (RefSeq).

Super-enhancers were called using the ROSE tool as described (Whyte et al. 2013; Hay et al. 2016). Briefly, individual enhancers within 12.5kb of each other were stitched together to form a single large enhancer domain. Stitched enhancer domains were then ranked for input-corrected Med1 ChIP-seq enrichment. Med1 signal was normalized to a maximum value of 1.0 and visualized after sorting the enhancers for signal. Enhancer domains were called as super-enhancers if Med1 occupancy was above a threshold placed at the point where the tangent to the graph ranking enhancers was equal to 1.

2.8.6 Analysis of differentially expressed genes

Mapped RNA-seq reads were assigned to genes using Subread featureCounts using RefSeq gene annotation. Normalised differential gene expression between biological triplicate data from wild type and CTCF binding site mutant mice was calculated with the DESeq2 R package.

2.8.7 Analysis of Capture-C data

Analysis of Capture-C data was performed as previously described (Davies *et al.* 2015). First, adaptor sequences were removed with Trim Galore (--length 10, --

qualFilter 20) and reads were reconstructed with FLASH (-m 9 –x 0.125). Next, reads were *in silico* digested at DpnII restriction sites using the DpnII2E.pl script and aligned to the mm9 genome using Bowtie (-p 1, -m 2, --best, --strata). Aligned DpnII-digested read fragments were classified as reporter fragments, if they mapped to regions other than the capture fragment (i.e. the viewpoint) or a 1 kb proximity-exclusion zone on either side of the capture fragment. A stringent filter requiring both start and stop coordinates of reads to be unique was used to remove all PCR duplicates. Reporter fragments were called as unique interactions if (i) the sequence read was unique and contained (ii) a single capture fragment next to (iii) one or more reporter fragments.

To determine differential interactions between erythroid (ter119+) and ES cells, and between wild-type and CTCF binding site mutant mice, read counts were quantified per DpnII-fragment and normalised to the total number of detected interactions to give the number of interactions per 100,000 total interactions. Data was exported as a BigBed file and visualized in the UCSC genome browser. Statistically significant changes were calculated on non-normalized counts per restriction fragment from biological triplicates using the DESeq2 R package.

For visualisation, of differences in Capture-C profiles, normalised interactions in ES cells or erythroid cells of CTCF binding site mutants were subtracted from wild-type interactions to generate a differential Capture-C track. For plotting of multiple interaction profiles simultaneously, Capture-C interactions were binned in 250bp bins and a sliding 5 kb window was used. The mean of three biological replicates and standard deviation were plotted in R.

2.8.8 *de novo* CTCF motif analysis in Ter119+ cells

Motif analysis was performed similar to previously described (Nakahashi et al. 2013). A set of 2000 peaks was randomly selected from the set of CTCF peaks that were called in ter119+ cells. Peak sequences were retrieved and used for *de novo* motif discovery using the MEME suite with the following command: meme -revcomp –dna – nmotifs 1 –w 20 –mod zoops –fa peaks.fa –bfile flanks.bg. A background model was constructed from regions directly flanking the peaks using fasta-get-markov (-m 0 flanks.fa). The motif with the highest score matched the previously published consensus CTCF core binding motif. Significant matches ($p < 10^{-3}$) for the CTCF core motif within all peak regions were identified using fimo (-motif 1 -bgfile flanks.bg core.meme.txt peaks.fa). If multiple core motifs were detected within the same peak region, only the best match was retained.

To identify previously described motifs up- and downstream of the core motif (Nakahashi et al. 2013), 20 bp sequences of up- and downstream flanks were extracted in a strand specific manner so that the CTCF core motif was always on the top strand. *De novo* motif discovery was performed again on sets of 6000 randomly selected up- and downstream flanks using meme (-nmotifs 5 -minw 5 -maxw 10 -minsites 100 -mod zoops –bfile flanks.bg). The top upstream motif (U) and downstream motif (D) were 10 bp long and resembled those previously identified (Nakahashi et al. 2013). Fimo was used to identify all occurances of U and D motifs in the CTCF peak sequences with a P-value threshold of 10^{-2}. Analysis of spacing between the core and flanking motifs revealed a preferential spacing for both up- and downstream motifs (spamo –eps –dumpseqs –inc 1 –bgfile flanks.bg core.meme.txt

up/down.meme.txt). Significant upstream or downstream motif matches were added to CTCF peak annotation only if the motifs were present at the preferred spacing (5-6 bp for U, 4-6 for D). Motif logos and heatmaps were created using the R seqLogo package and image() function respectively.

2.8.9 DNaseI footprint analysis

DNaseI footprints and meta-plots at CTCF binding sites were generated using a custom perl script based around Samtools using previously published C57BL/6 DNaseI-seq data (Hosseini et al. 2013). DNaseI-seq cuts were counted as the 5' end of the first read and the 3' end of the second read of DNA fragments. For meta-plots of CTCF peaks with different combinations of core and auxiliary motifs, average cut-counts relative to the start of the CTCF core sequence were calculated for each category.

2.8.10 Data visualizations and graphs

Graphical visualizations were created with the UCSC genome browser, the R plot function, and Microsoft Excel. Aesthetic modifications only were made with Adobe Illustrator.

2.9 Analysis of mouse haematological phenotypes

Haematological analysis of CTCF binding site mutant mouse models was carried out by Jacqueline Sharpe at the Molecular Haematology Unit (MHU, WIMM). Mice were backcrossed on a C57BL/6 genetic background. Reticulocyte preparations were generated using BCB staining, and counts were made by two independent assessors blinded to genotype.

Chapter 3: Interacting clusters of convergent CTCF binding sites flank the α-globin gene cluster

3.1 Introduction

The mouse α-globin gene cluster is located in a gene-dense region, within an erythroid sub-TAD containing 10 genes within a 50kb range of its enhancer region. The cluster contains several broadly-expressed housekeeping genes (*Snrnp25, Mpg*, and *Nprl3*) and transcriptionally repressed genes (*Rhbdf1*, *Il9r*, and the embryonic stage ζ-globin) with roles in a wide variety of cellular processes including; cell signalling (*Rhbdf1* (Christova et al. 2013), *Il9r* (Knoops and Renauld 2009)), splicing (*Snrnp25* (Will 2004)), DNA repair (*Mpg* (Hedglin and O'Brien 2008)), and metabolism (*Nprl3* (Kowalczyk *et al.* 2012a)). The ability to activate and repress these genes independently is essential for the regulation of these cellular processes. Thus, despite the close proximity of genes flanking the α-globin cluster, their transcriptional states appear to be shielded from transcriptional upregulation by the strong α-globin enhancer in erythroid cells.

In humans, α-globin expression is regulated by four enhancers located upstream of the α-globin genes. Full or partial deletions of these four regulatory elements have long been known to cause α-thalassemia by decreasing α-globin transcription (Higgs and Wood 2008). The genetic structure of the α-globin locus is conserved in mice (Hughes et al. 2005), where transcription of α-globin is controlled by a block of five

distal enhancers located 15 to 40kb away from the nearest α-globin promoter. Four out of the five α-globin enhancers are located in the introns of the neighbouring gene *Nprl3*. In erythroid cells, these enhancer elements specifically interact with the α-globin promoters driving high levels of transcription (Davies *et al.* 2015). Careful *in vivo* study of enhancer element deletions, both singly and in informative combinations, shows that the R1 and R2 enhancers have the largest contributions to the transcriptional output of the α-globin genes. Their combined loss results in a >90% reduction in transcript levels, whereas individual deletions of R1 and R2 result in reductions of ~35% and ~50% (Hay et al. 2016). While the functional study of these *cis*-regulatory elements in both patients and mouse models has identified the distal elements that are important for cell type-specific expression of the α-globin genes, the mechanism by which they specifically interact with and subsequently upregulate their target genes are still poorly understood.

Composite enhancer regions consisting of multiple individual DNaseI hypersensitive sites have been shown to drive exceptionally high levels of their target gene expression (Whyte et al. 2013). These enhancers, termed super-enhancers, are characterised by high levels of binding by the Mediator complex. In ES cells, super-enhancers and the genes they regulate have been shown to be insulated from surrounding chromatin by interactions between flanking binding sites of CTCF and Cohesin (Dowen et al. 2014). Deletion of these sites was shown to result in upregulation of flanking genes. These observations fit with longstanding observations that CTCF is able to act as an insulator from local enhancers through formation of interactions between its binding sites (Splinter 2006a). Accordingly,

CTCF is enriched at the boundaries of regions of interacting chromatin termed topologically-associated domains (TADs) (Dixon et al. 2012).

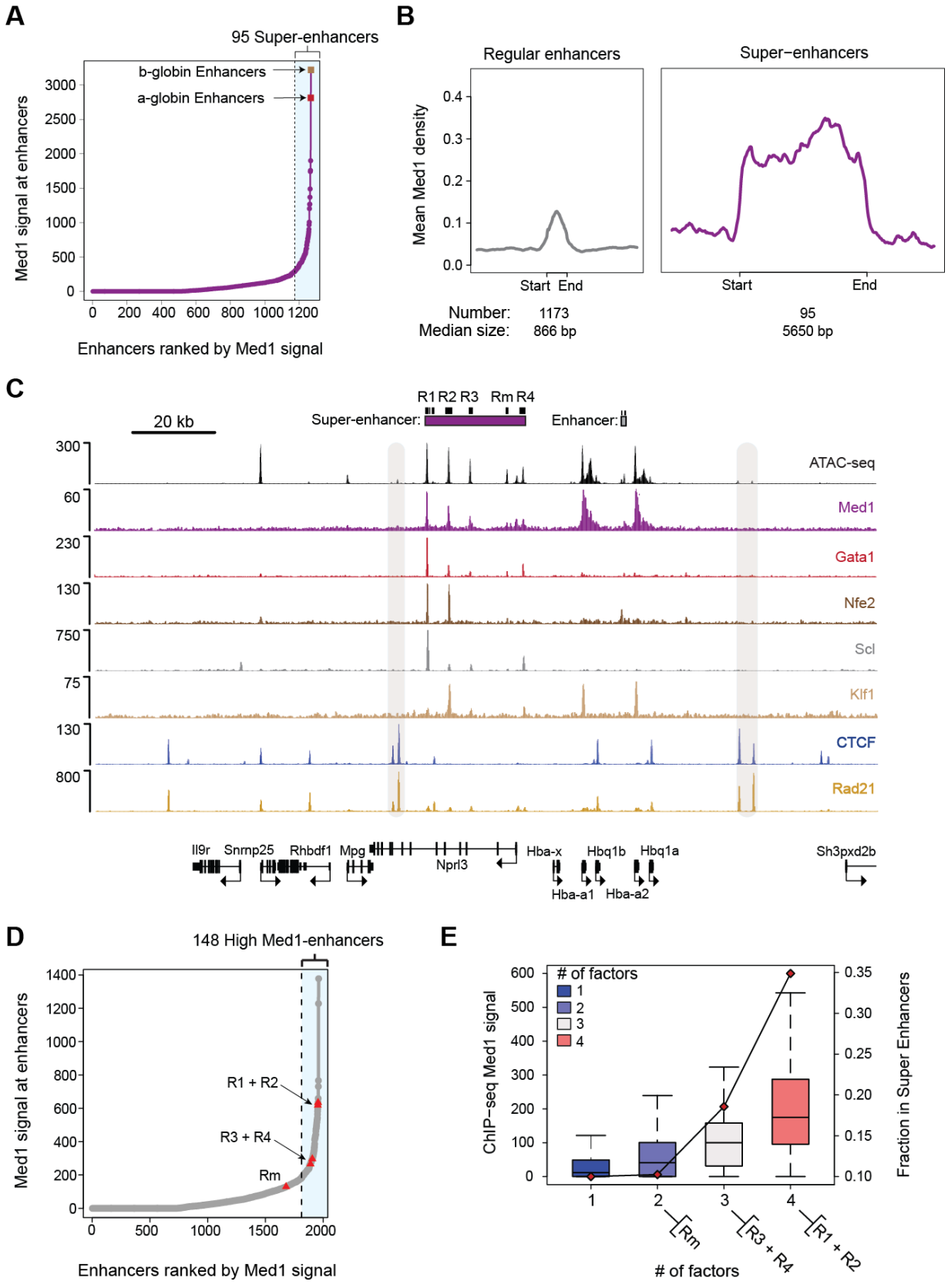
A recent advance in understanding the mechanism by which interactions between CTCF binding sites shape the genome was made with the discovery that the orientation of CTCF binding is important for the establishment of CTCF interactions (de Wit *et al.* 2015; Sanborn *et al.* 2015; Guo *et al.* 2015). These studies show that interactions between CTCF sites are preferentially formed if binding occurs in a convergent manner (Fig. 3.2B). In studies of genome-wide CTCF interactions by Hi-C and ChIA-PET, the majority of detected CTCF interactions occurred between convergent sites (see Fig. 3.2B). However, a large number of interactions between tandem sites was also detected, suggesting that only divergent CTCF pairs are certain to be devoid of reciprocal interactions (Rao *et al.* 2014; Tang *et al.* 2015).

In this chapter, I have described the distribution and orientation of CTCF binding sites at the α-globin gene cluster in primary mouse erythroid cells. I examine chromatin interactions of CTCF binding sites and the α-globin enhancers in embryonic stem (ES) cells and primary erythroid cells. Furthermore, I propose a mechanism by which CTCF contributes to enhancer specificity in erythroid cells through the establishment of a local chromatin architecture.

3.2 Results

3.2.1 Five α-globin regulatory elements form an erythroid super-enhancer

The need for the insulation of strong enhancer clusters occupied by high levels of the Mediator complex, termed super-enhancers, has previously been established (Dowen et al. 2014). To investigate whether the α-globin enhancer cluster classifies as an erythroid super-enhancer, all *cis*-regulatory elements were identified by characterising DNaseI-hypersensitive sites (DHSs) in erythroid cells (ter119+) isolated from C57BL/6 mice. This DNaseI data was re-analysed from the original publication in Hosseini et al. 2013. Previously published H3K4me3 and H3K4me1 ChIP-seq data was used to sub-classify regulatory elements as enhancers or promoters (data published in Kowalczyk *et al.* 2012b). Elements were classified as enhancers if they contained a high ratio of H3K4me1 to H3K4me3 and were not located within 250bp of transcription start sites. This resulted in a list of 1,963 putative enhancers in erythroid cells.



See next page for figure legend

Figure 3.1 Analysis of super-enhancers in mouse erythroid cells. A. Ranked distribution of input-subtracted mediator ChIP-seq signal (total reads) for all stitched enhancers (n = 1,268). In total, 95 enhancers, including the α- and β-globin enhancers, contain high levels of mediator and are classified as super-enhancers. **B.** Average distribution of normalised (input-subtracted reads per bp per million mapped) Med1 ChIP-seq signal across 1173 regular enhancers and 95 super-enhancers and a region of 3kb up- and downstream. The distance between start and end was scaled to the median size of regular or super-enhancers.ES **C.** Normalised (RPKM) ATAC-seq and ChIP-seq data for Med1, Gata1, Nfe2, Scl, Klf1, CTCF, and Rad21 across the α-globin locus in primary erythroid cells. Pairs of CTCF and Rad21 bound sites are highlighted in brown. **D.** Ranked distribution of input-subtracted mediator ChIP-seq signal (total reads) for all identified enhancers without stitching (n = 1,963). In total, 148 enhancers are classified as high-Med1 enhancers. The α-globin enhancers are shown as red triangles. **E.** The fraction of individual enhancers that are constituents of a super-enhancer and the average Med1 ChIP-seq signal (input-subtracted total reads) at individual enhancers as a function of the number of transcription factors bound to the individual enhancer (Number of enhancers in each group: 1 TF: 570, 2 TF: 459, 3 TF: 237, and 4 TF: 129).5

To identify super-enhancers, this list of erythroid enhancer elements was used as input for the ROSE tool (Whyte et al. 2013). In short, this algorithm merges individual enhancer elements located within 12.5kb of each other into a single genomic region. The stitched and remaining individual enhancer regions are then ranked for the occupancy of Mediator (Med1, data generated by Mira Kassouf (Hay et al. 2016)). As was previously observed in other cell types (Whyte et al. 2013), a small number of merged erythroid enhancers (95) are bound by exceptionally high levels of Med1. By definition, enhancer regions are classified as super-enhancers if levels of Med1 occupancy are above a cut-off value determined by the point at which the slope of the Med1 distribution plot is equal to 1 (Fig. 3.1A).

Consistent with previous studies (Whyte et al. 2013; Dowen et al. 2014), the median size of erythroid super-enhancers is one order of magnitude above that of regular enhancers (Fig. 3.1B). The previously identified regulatory regions of the α-globin and β-globin genes are the two erythroid super-enhancers marked by the highest levels of Mediator binding, consistent with these genes being very highly expressed

in erythroid cells (Fig. 3.1A). The mouse α-globin super-enhancer spans 24kb, covers most of the adjacent *Nprl3* gene, and encompasses all five individual α-globin enhancers (Fig. 3.1C). Consistent with previous observations that Mediator recruits Cohesin to enhancers and promoters (Kagey et al. 2010), low levels of binding of Cohesin component Rad21 are observed at super-enhancer constituents (Fig. 3.1C).

I next evaluated transcription factor binding by re-analysing previously published ChIP-seq data for four key erythroid transcription factors (published in Kassouf *et al.* 2010; Kowalczyk *et al.* 2012b; Hay *et al.* 2016). The individual enhancer elements that constitute the α-globin super-enhancer are all bound by different combinations of erythroid transcription factors and have varying levels of Med1 occupancy (Fig. 3.1C). When individual erythroid enhancers were ranked for the level of Med1 binding without stitching, Med1 occupancy was found to vary considerably between α-globin super-enhancer constituents (Fig. 3.1D). Using the same cut-off as described above, the four conserved α-globin regulatory elements (R1-R4) still fell into a class of high-Med1 enhancers whereas the mouse-specific enhancer Rm did not. Consistent with the observation that 'hotspots' of transcription factor binding are defining components of adipogenic super-enhancers (Siersbæk et al. 2014), the number of transcription factors bound by individual enhancer elements correlated to the level of Med1 occupancy and to the likelihood that enhancer elements were a part of a super-enhancer (Fig. 3.1E). Within the α-globin super-enhancer, the R1 and R2 constituents bind the greatest number of transcription factors and are occupied by high levels of Med1.

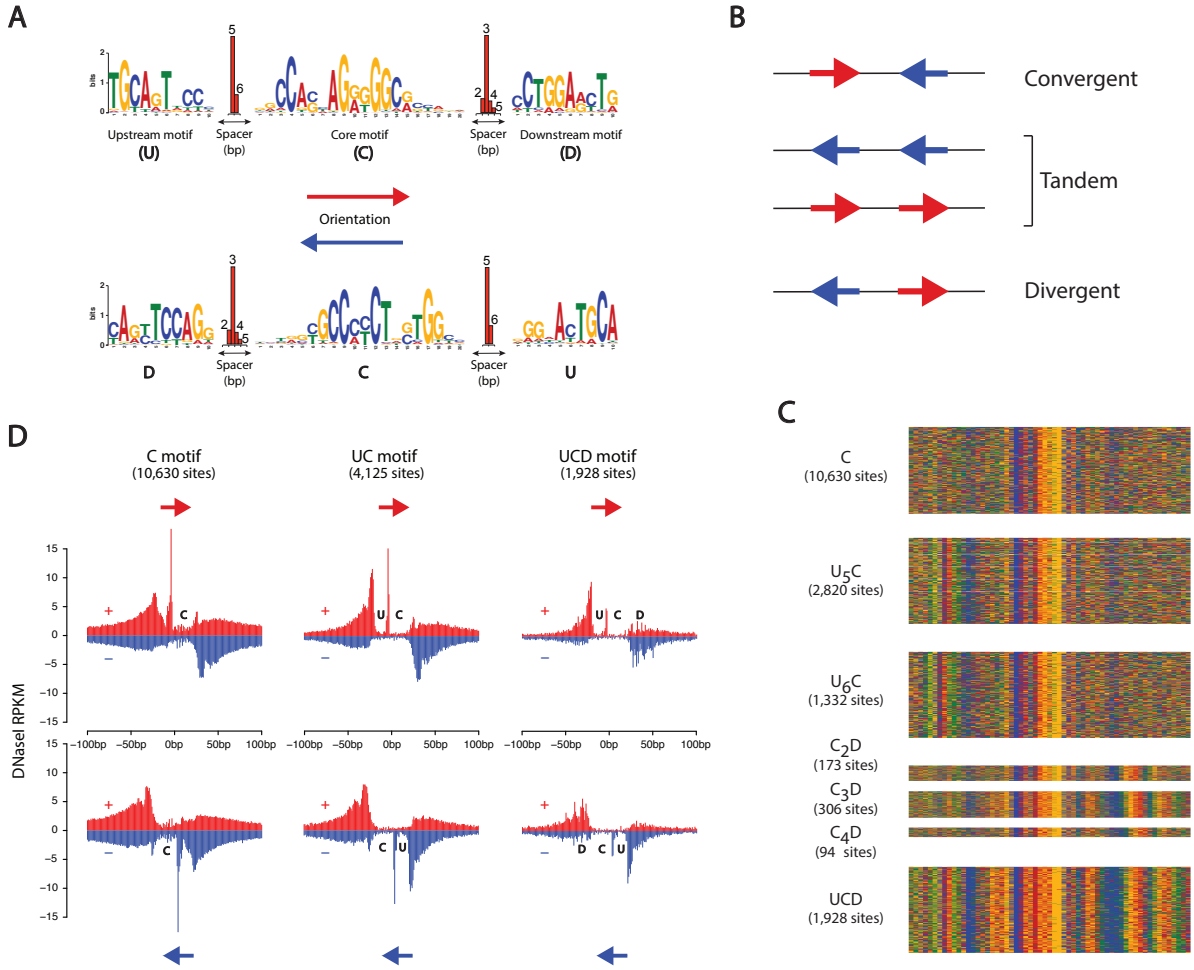


Figure 3.2 *De novo* analysis of the CTCF binding motif in mouse erythroid cells. A. CTCF binding motif resulting from MEME motif analysis of CTCF ChIP-seq in erythroid cells in forward and reverse orientation. A core, upstream, and downstream sequence element are identified with preferential spacing to the core motif (red bar histograms) **B**. Possible orientations of two CTCF motifs in the genome with respect to each other. **C.** Heatmap representation of a 60 bp sequence around all CTCF binding motifs grouped by the presence and spacing of up- and downstream motifs. The total number of sites in each heatmap is shown in parentheses. **D.** Plots of the average DNaseI footprints of C, UC, and UCD motif containing CTCF binding sites if forward orientation (top panel) and reverse orientation (lower panel). Upper (red, +) and lower (blue, -) strand specific DNaseI cleavage signals are shown.

This corresponds with the observations that deletion of one or both of these elements has strong effects on α-globin gene expression, whereas deletion of R3, R4, or both in combination, does not (Hay et al. 2016). Thus, both with respect to their transcription factor binding profiles and function, R1 and R2 are the strongest constituents of the α-globin super-enhancer.

3.2.2 Analysis of CTCF orientation at the α-globin locus reveals two flanking clusters of convergent CTCF binding

Having established that the α-globin enhancer region classifies as a super-enhancer located in the centre of a dense gene-cluster, I next investigated whether CTCF binding sites separate the enhancer from surrounding genes as has previously been reported (Dowen et al. 2014). The study of genome-wide binding of CTCF in erythroid cells revealed that the α-globin genes and enhancer-region are flanked by several CTCF binding sites (Fig. 3.1C). The majority of CTCF binding sites at the α-globin cluster are co-bound by Cohesin (Fig. 3.1C). This includes two pairs of CTCF bound sites directly flanking the gene cluster that are marked by high levels of Cohesin component Rad21, suggesting they may be involved in insulation of the super-enhancer cluster.

A large number of recent studies shows a preference for interactions between CTCF-binding sites in a convergent orientation (Guo *et al.* 2015; Sanborn *et al.* 2015; de Wit *et al.* 2015). To determine the orientation of CTCF binding sites at the α-globin cluster, I performed MEME *de novo* motif discovery on genome-wide CTCF ChIP-seq peaks in erythroid cells as previously described (Nakahashi et al. 2013).

This analysis revealed a 20 bp core (C) CTCF binding motif that matched those previously reported and could be identified in 98% (17,284) of the binding sites (Fig. 3.2A, C) (Kim *et al.* 2007; Schmidt *et al.* 2010; Rhee and Pugh 2011). Further analysis of the up- and downstream flanks of the core motif also discovered the previously identified 10bp upstream (U) and downstream (D) motifs (Nakahashi et al. 2013). The CTCF core binding sequence was found associated with different combinations of the flanking up- and downstream motif depending on the binding site (Fig. 3.2C).

The upstream motif was found to be present in 35% (6080) of CTCF peaks and occurred at a preferential distance of 5 or 6 bp from the core motif, whereas the downstream motif was only present in 14% (1501) of binding sites (Fig. 3.2A, C). Although the erythroid downstream motif resembled part of the sequence (CTGGA) of the motif previously identified in B-cells, analysis in erythroid cells suggests this motif occurs with a preferential spacing of 2-4 bp instead of 6-8 bp as previously suggested. A closer look at the motif analysis in both cell types revealed that two instances of the downstream sequence (CTGGA) are present downstream of the central core sequence, the start of the more distal occurrence of the motif corresponding to the spacing previously described (Nakahashi et al. 2013). The more proximal motif is more strongly identified in the analysis of erythroid CTCF binding sites.

The identified CTCF binding sequence is information-rich and non-palindromic, allowing the orientation of CTCF binding to be established with high confidence (Fig. 3.2A). The orientation of CTCF binding sites was determined genome-wide by identification of the most significant match to the consensus motif below each CTCF

binding peak (ChIP-seq). To verify that the prediction of binding orientation was correct, genome-wide CTCF motif predictions were compared with DNaseI footprinting data (DNaseI data was analysed from Hosseini et al. 2013). Genome-wide motifs were grouped by orientation (+ or – strand) and class (C, UC, UCD) and average, strand-specific DNaseI cut-counts were plotted relative to the start of the sequence motif (Fig. 3.2D). As was shown previously (Nakahashi et al. 2013), the upstream and core motifs provide marked protection against DNaseI digestion whereas the downstream motif is less well-protected and further characterised by three to four smaller footprints.

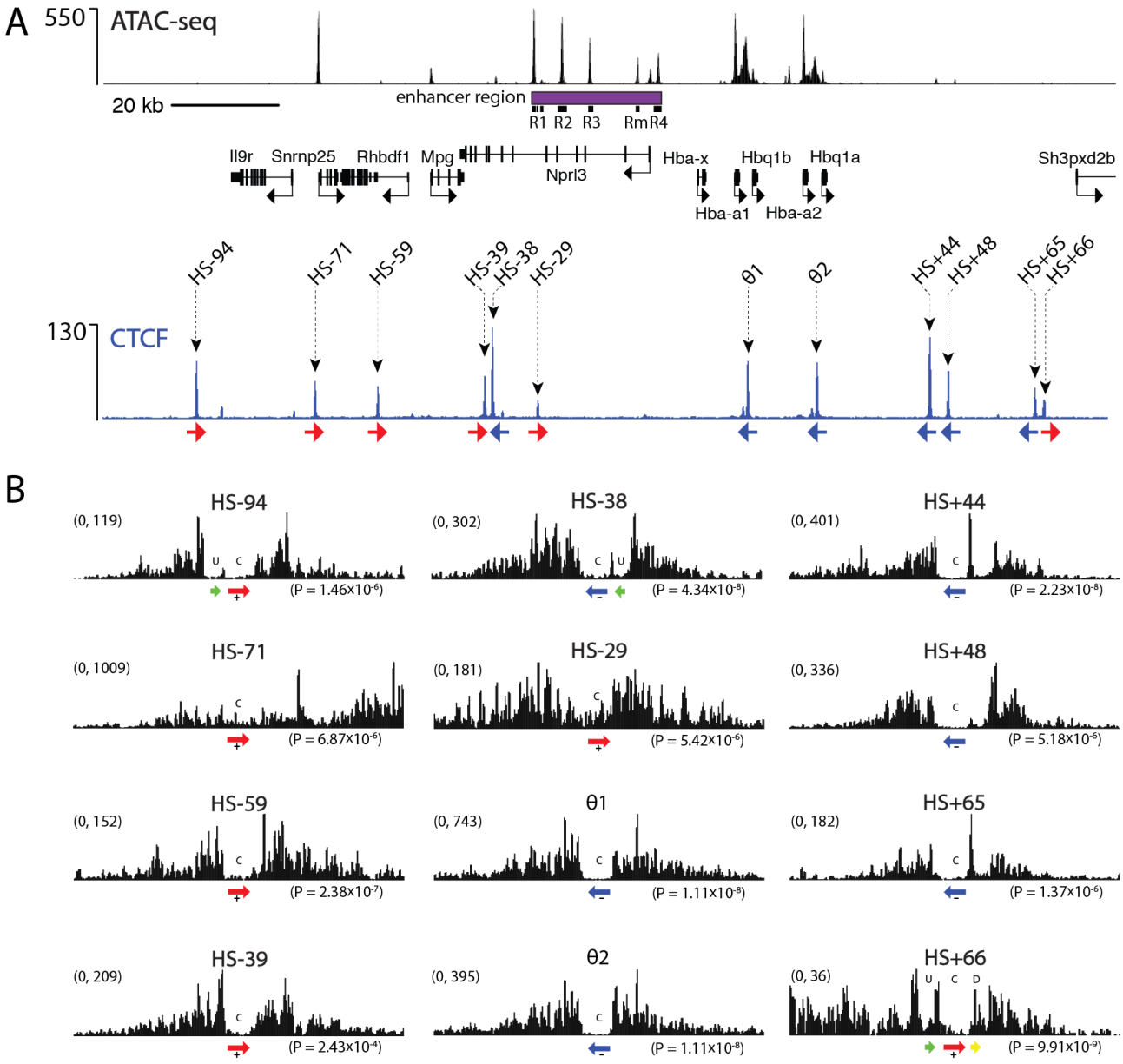


Figure 3.3 CTCF binding orientation at the α-globin locus. A. Normalised CTCF ChIP-seq (RPKM) and ATAC-seq (RPKM) in erythroid cells, merged across two biological duplicates. Annotated with the CTCF site names and orientation. Gene annotation is Refseq. **B**. DNaseI footprints and top CTCF binding motif hit for each of the CTCF binding sites at the α-globin locus. Motif P-values are shown. Orientation is indicated by the colour of the core motif; forward (red) or reverse (blue). Upstream motif is shown in green and downstream motif is shown in yellow.

This results in an asymmetrical footprint that is dependent on the orientation of CTCF binding in the presence of either motif, which can clearly be observed in groups containing the flanking motifs (Fig. 3.2D). Moreover, a clear strand-specific DNaseI footprint is observed even when only the core binding motif is present; a strand-specific peak of DNaseI cuts directly upstream of the C motif can be observed in meta-footprint plots (Fig. 3.2D). This validates that the analysis of the CTCF sequence motif is able to reliably distinguish between both orientations of CTCF binding.

At the α-globin cluster, analysis of CTCF binding orientation revealed a striking pattern of CTCF binding orientation (Fig. 3.3A). Each of the flanking genomic regions contains several CTCF bound sites orientated in tandem, such that two clusters of convergent CTCF binding sites are created with the α-globin enhancer region located at the focal point of these clusters. A notable exception to this organisation is the HS-38 CTCF binding site, which faces away from the enhancer region. To validate that genome-wide CTCF motif predictions at the α-globin cluster were correct, I verified that the predicted CTCF motifs matched DNaseI footprints created by CTCF binding at these sites (Fig. 3.3B).

3.2.3 Clusters of convergent CTCF binding sites delimit the α-globin chromatin compartment in erythroid cells

To investigate how the organisation of CTCF binding to flanking, convergent clusters shaped local chromatin interactions in erythroid cells, I performed Capture-C from viewpoints across α-globin cluster. It has previously been shown that the interactions of the α-globin promoters are contained within an upstream, tissue-specific 'compartment' of roughly 70kb (Davies *et al.* 2015). To examine whether the interactions of α-globin regulatory elements is constrained in a similar manner, Capture-C was performed using the enhancers as viewpoints.

As expected, the α-globin enhancers interact strongly and specifically with the α-globin promoters in erythroid cells (Fig. 3.4). The interaction profile of the strongest and most distal enhancers, R1 and R2, is remarkably unidirectional: these enhancer elements only interact with chromatin downstream of their genomic position while showing minimal interactions with the upstream gene-rich region. Reciprocally, the α-globin promoters interact specifically with the genomic region directly upstream containing the enhancer region (α-globin promoter data re-analysed from Davies *et al.* 2015). While the centrally located enhancer elements R3, R4 and Rm have more bidirectional interaction profiles, their interactions are still delimited by boundaries directly upstream of R1 and downstream of the α-globin genes. The ~70kb compartment of reciprocally interacting chromatin defined by these boundaries contains the globin genes (α-, θ-, and ζ- globin) and their regulatory elements, but excludes most flanking genes with the exception of *Nprl3*.

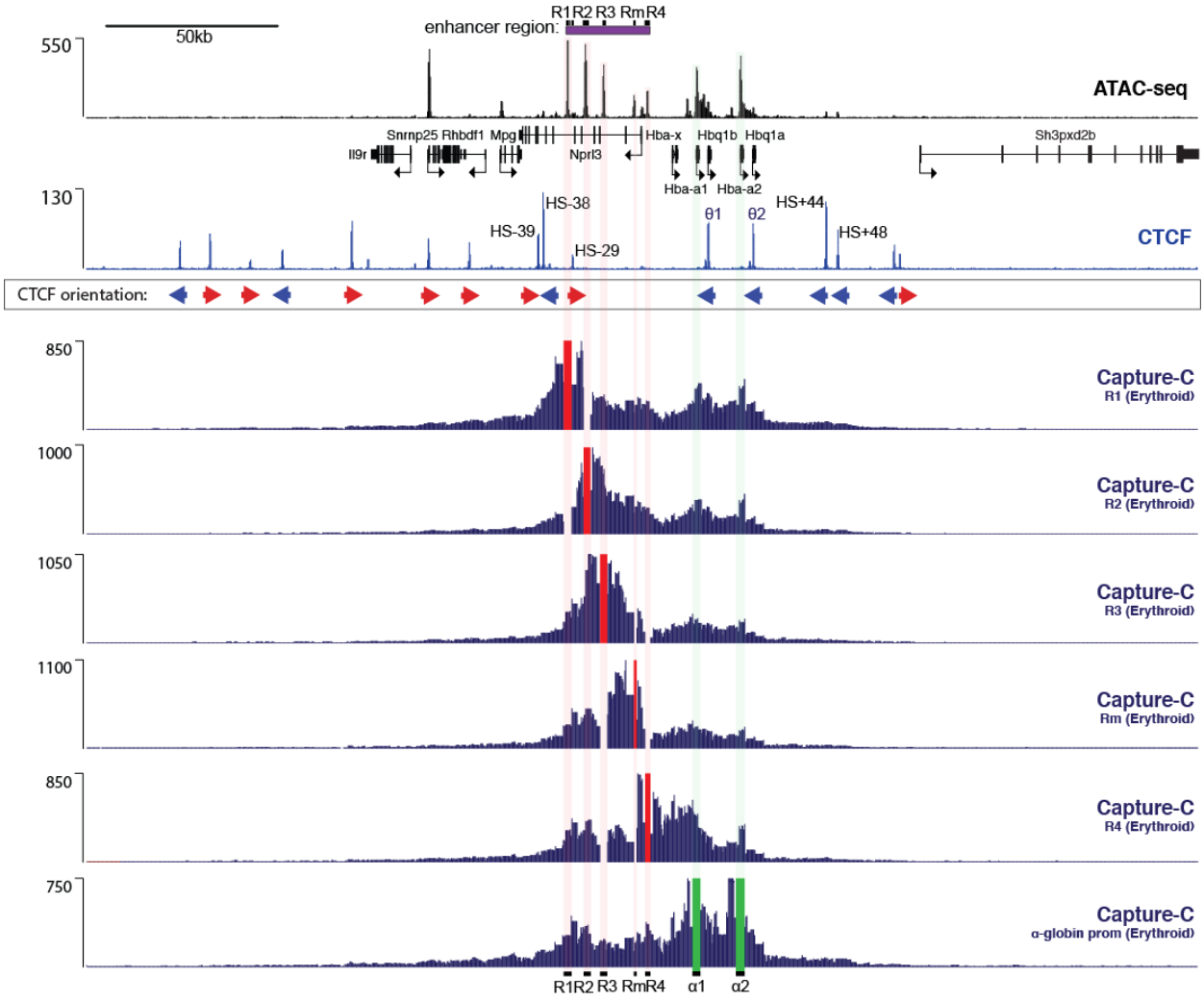


Figure 3.4 Interactions of the α-globin regulatory elements in erythroid cells. Panels show normalised Capture-C data for the indicated viewpoints in erythroid cells merged across three biological replicates. The mean number of meaningful interactions per restriction fragment, scaled to a total of 100,000 interactions genome-wide are shown. Annotated with forward (red) and reverse (blue) CTCF binding orientation. Red vertical bars indicate the position of the enhancer viewpoints, whereas green vertical bars indicate the position of the α-globin promoters. Also shown are normalised CTCF ChIP-seq (RPKM) and ATAC-seq (RPKM) for erythroid cells, merged across two biological duplicates. Gene annotation is Refseq. α-globin promoter Capture-C data is from Davies et al. 2015

The edges of the α-globin compartment roughly coincide with the start of the flanking clusters of CTCF binding sites. To investigate whether interactions between these flanking regions delimit the α-globin compartment, Capture-C was performed on CTCF binding sites on both sides of the compartment (Fig. 3.5). Strikingly, CTCF binding sites located outside the α-globin compartment do not interact with the enhancer region, but interact instead with the clusters of convergently orientated CTCF binding sites on the opposite flank of the gene cluster. The CTCF binding sites HS-38 and HS-39 interact with a genomic region containing the θ1/2 and HS44/48 sites, whereas reciprocal interactions with the upstream cluster of convergent CTCF binding sites are observed from the HS44 and HS48 sites. The θ1 and θ2 CTCF binding sites, located at the θ-globin promoters, are situated on the boundary of the α-globin compartment and interact with both the enhancer region and upstream CTCF sites. Moreover, the θ2 CTCF site, which is located more towards the edge of the central compartment than θ1, interacts less with the enhancers and more with the upstream region of chromatin showing the transition between these two interaction domains.

Not only CTCF binding sites follow this interaction pattern; interaction profiles of the promoters of flanking genes *Rhbdf1* and *Mpg* are similarly devoid of interactions with the α-globin enhancer region. These genes instead also interact with the downstream flank of the α-globin gene cluster on the other side of the α-globin compartment.

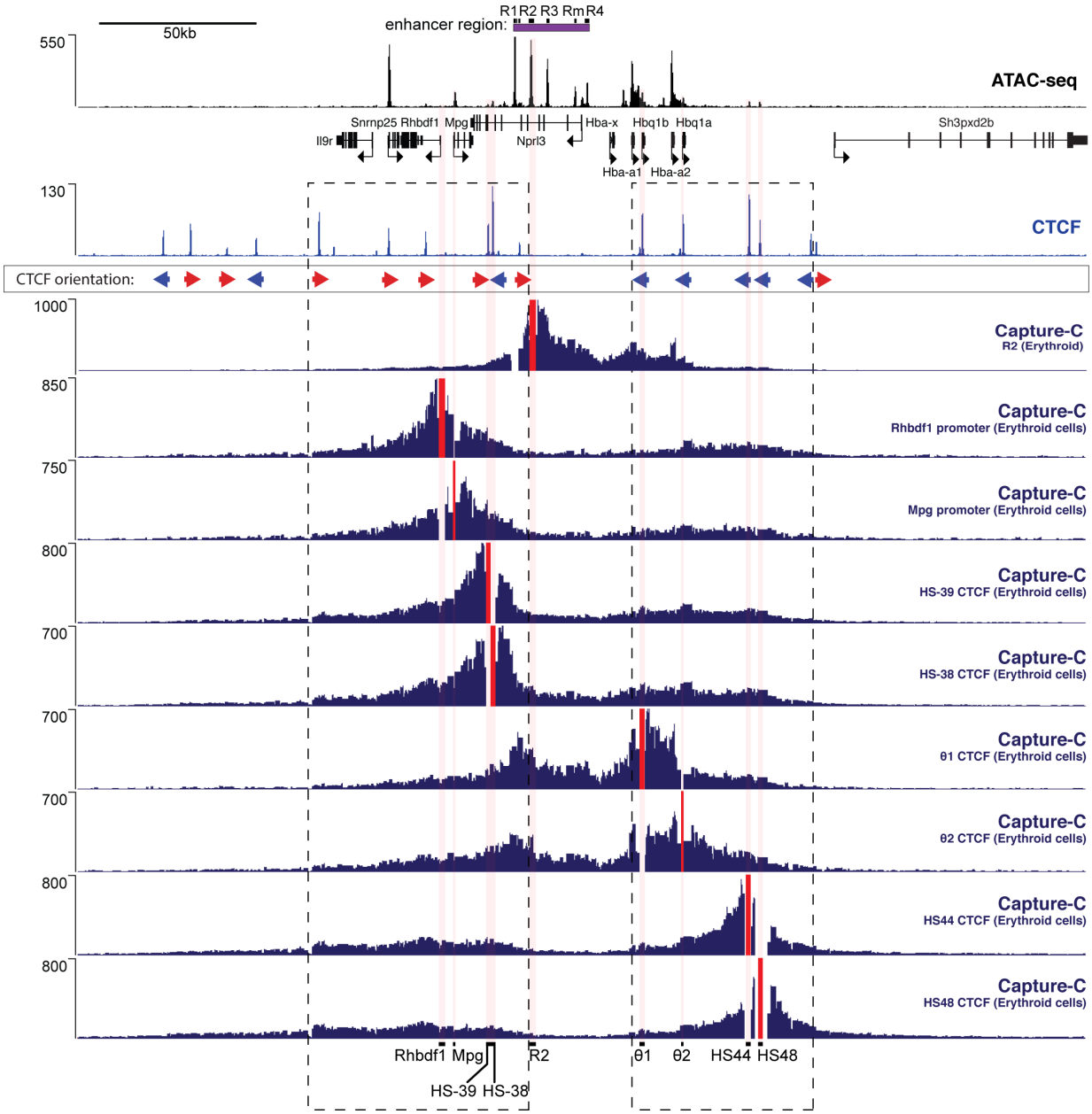


Figure 3.5 Interactions of viewpoints flanking the α-globin cluster in erythroid cells. Panels show normalised Capture-C data for the indicated viewpoints in primary erythroid cells isolated from mouse spleens and merged across three biological replicates. The mean number of meaningful interactions per restriction fragment, scaled to a total of 100,000 interactions genome-wide are shown. Annotated with forward (red) and reverse (blue) CTCF binding orientation. Boxes indicate clusters of convergent CTCF binding sites. Red vertical bars indicate the position of the viewpoint. Also shown are normalised CTCF ChIP-seq (RPKM) and ATAC-seq (RPKM) for erythroid cells, merged across two biological duplicates. Gene annotation is Refseq.

3.2.4 Interactions between flanking domains are induced upon α-globin enhancer activation

Active regulatory elements form strong interactions with the promoters of their target genes (Fig. 3.4). To understand how CTCF-mediated chromatin interactions are affected by α-globin enhancer activation, I compared interactions of regulatory elements between mouse erythroid and embryonic stem (ES) cells. In ES cells, the α-globin genes are not expressed and transcription of flanking genes *Snrnp25* and *Mpg* occur at similar levels in both cell types independent of α-globin enhancer activity (Fig. 3.6A, B). Whereas the *Rhbdf1* gene is actively transcribed under the influence of its own enhancer in ES cells, this gene is repressed in erythroid cells. The *Il9r* gene is not expressed in either cell type. Notably, the *Nprl3* gene, whose promoter is located inside the α-globin compartment, is the only gene in the cluster that is upregulated in erythroid cells apart from the globin genes (Fig. 3.6B).

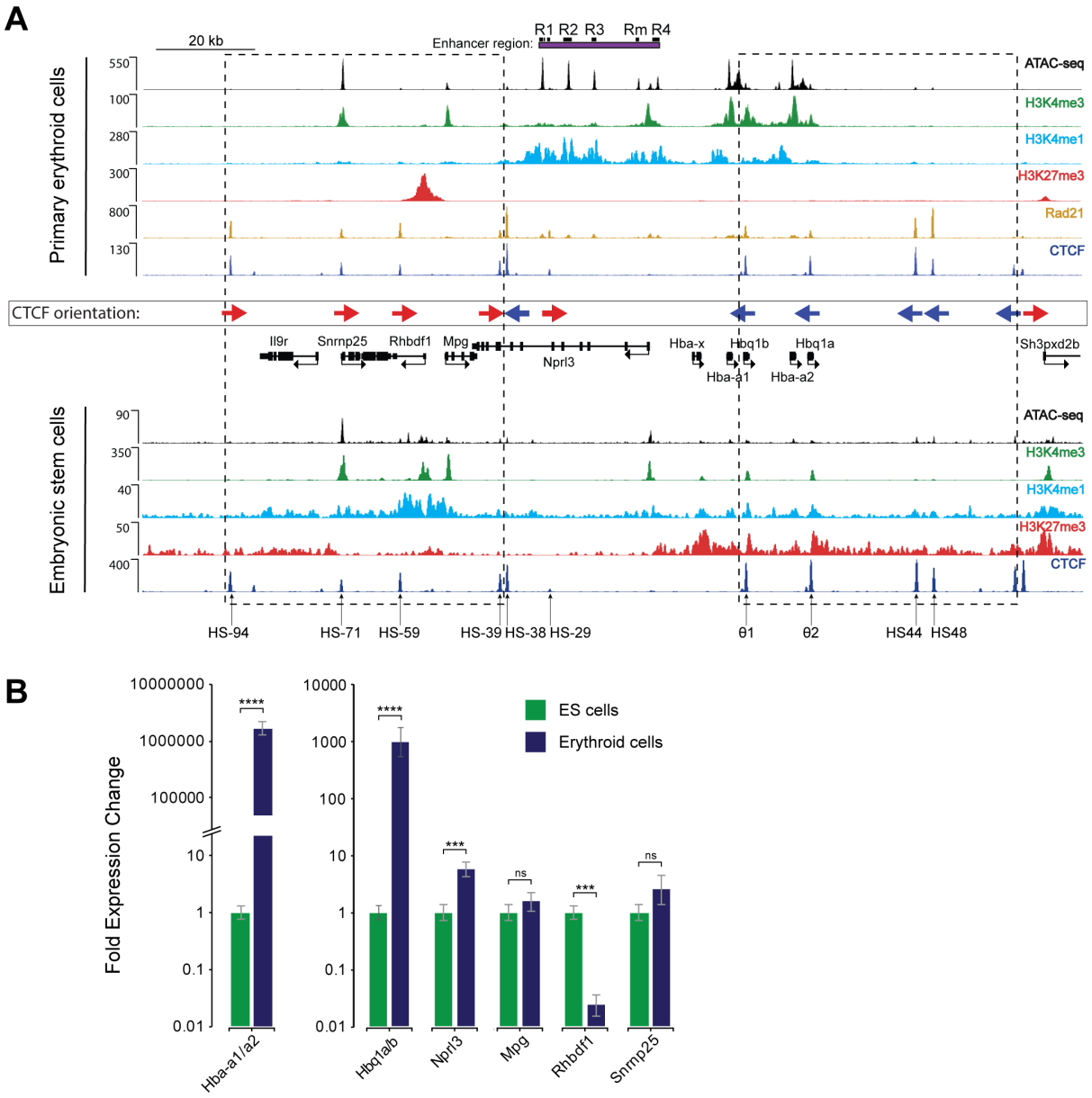


Figure 3.6 Comparison of chromatin state and gene expression between erythroid and ES cells. A. Normalised (RPKM) ChIP-seq read-densities at the α-globin locus in erythroid cells and ES cells. Read density is the average of two biological replicates, except for erythroid H3K4me1 (one replicate). Annotated with forward (red) and reverse (blue) CTCF binding orientation. Boxes indicate clusters of convergent CTCF binding sites. Gene annotation is Refseq. Erythroid H3K4me1 is from. ES cell ChIP-seq data is from Stadler *et al.* 2011 and Consortium *et al.* 2012. ES cell ATAC-seq was generated by Damien Downes (WIMM) **B**. Relative expression of erythroid versus ES cells. Measured by real-time qPCR and representing 3 biological replicates. P-values are obtained via a student t-test. ns>0.05, ***<0.001, ****<0.0001.

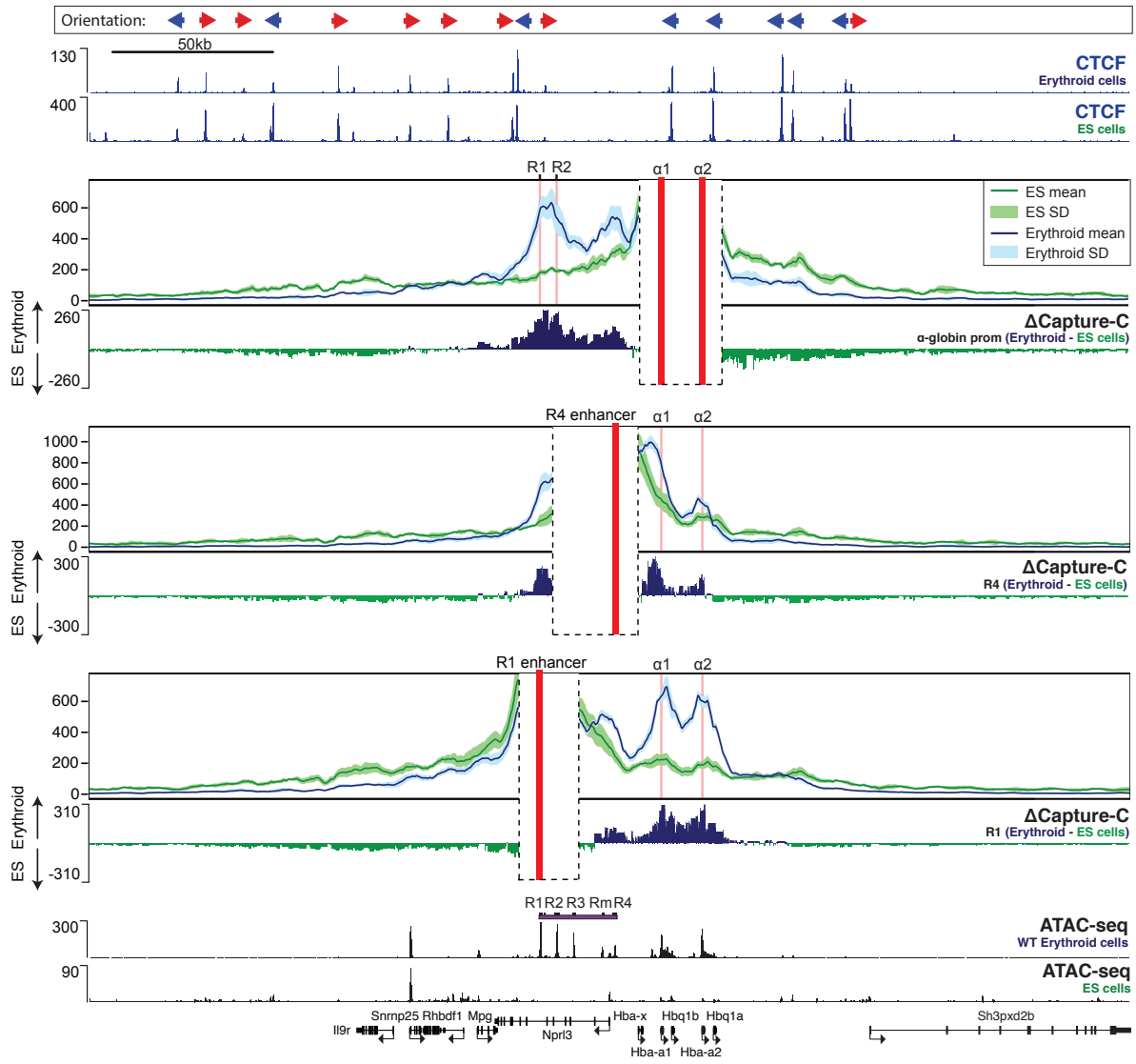


Figure 3.7 Differential interactions of α-globin regulatory elements between ES and erythroid cells. Panels show overlaid, normalised Capture-C data for the α-globin promoter, R4 enhancer, and R1 enhancer viewpoints in ES and erythroid cells merged across three biological replicates. The mean number of interactions, scaled to a total of 100,000 interactions genome-wide plus and minus one standard deviation (SD) of sliding 5kb windows are visualised. Differential tracks (ΔCapture-C) show a subtraction (WT - ES) of the mean number of meaningful interactions per restriction fragment, scaled to a total of 100,000 interactions genome-wide. Red vertical bars indicate the position of the viewpoint. Also shown are normalised CTCF ChIP-seq (RPKM) and ATAC-seq (RPKM) for both ES and erythroid cells, all merged across two biological duplicates. Gene annotation is Refseq. α-globin promoter Capture-C data is from Davies et al. 2015

To investigate the chromatin state at these genes in ES cells; I analysed ChIP-seq data for CTCF, H3K4me3, H3K4me1, and H3K27me3 in both primary erythroid and ES cells (ES cell data from Stadler et al. 2011 and Consortium *et al.* 2012. Erythroid H3K4me1 is from Kowalczyk *et al.* 2012b. All other erythroid ChIP-seq data was generated by me). Chromatin at the gene promoters reflects their transcriptional states (Fig. 3.6A); active promoters are marked by accessible chromatin and the presence of H3K4me3, whereas transcriptionally repressed genes lack these marks. Moreover, the inactive *Rhbdf1* gene is marked by high levels of H3K27me3 which is a repressive epigenetic mark deposited at regions of facultative heterochromatin by the Polycomb Repressive Complex 2 (PRC2) (Simon and Kingston 2013a). The enhancer mark H3K4me1 marks the active *Rhbdf1* enhancer in ES cells and the α-globin enhancer region in erythroid cells.

As expected, a comparison of normalised interactions between ES cells and erythroid cells shows that interactions between α-globin regulatory elements are erythroid-specific and not present in ES cells (Fig. 3.7). Instead, broader and more diffuse interaction profiles covering the gene cluster are observed from the α-globin promoter and enhancer viewpoints. Furthermore, the strong directionality that is observed in R1 and α-globin promoter interaction profiles in erythroid cells is not present in ES cells. Rather, more symmetrical interaction profiles are observed indicating the absence of specific regulatory interactions of these DNA elements in this cell type.

To examine whether these gene specific interactions between regulatory elements affect interactions between flanking CTCF binding sites, interaction profiles of two

flanking CTCF sites were compared between ES and erythroid cells. Despite near-identical CTCF binding profiles across the α-globin cluster in both cell types, a striking decrease in interactions between the domains flanking the α-globin cluster is observed in ES cells compared to erythroid cells (Fig. 3.8A). Instead, HS44 interacts more strongly with proximal CTCF binding sites in a tandem orientation such as θ1 and θ2. Similarly, the promoters of the *Mpg* and *Rhbdf1* genes do not interact with the genomic region downstream of the α-globin cluster in ES cells, showing that this local chromatin topology is specific to erythroid cells (Fig. 3.8B). It is possible that interactions between convergent CTCF binding sites flanking the α-globin cluster are dependent on the strong interactions between the active α-globin enhancer and the gene promoters that are only established in erythroid cells.

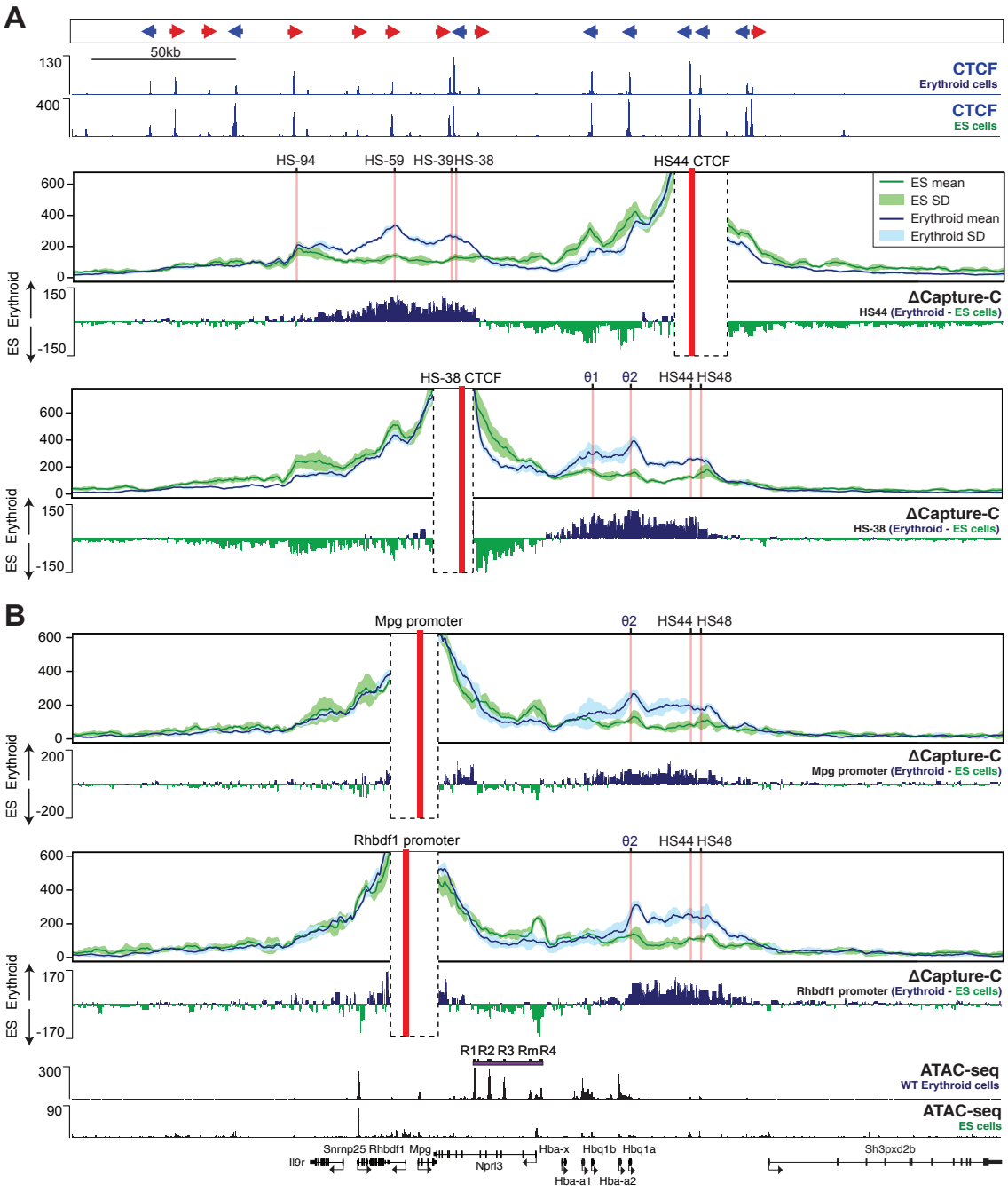


Figure 3.8 Differential interactions of CTCF binding sites flanking the α-globin cluster between ES and erythroid cells. Panels show overlaid, normalised Capture-C data for the indicated viewpoints in ES and erythroid cells merged across three biological replicates. The mean number of interactions, scaled to a total of 100,000 interactions genome-wide plus and minus one standard deviation (SD) of sliding 5kb windows are visualised. Differential tracks (ΔCapture-C) show a subtraction (WT - ES) of the mean number of meaningful interactions per restriction fragment, scaled to a total of 100,000 interactions genome-wide. Red vertical bars indicate the position of the viewpoint. Also shown are normalised CTCF ChIP-seq (RPKM) and ATAC-seq (RPKM) for both ES and erythroid cells, all merged across two biological duplicates. Gene annotation is Refseq. **A.** HS44 and HS-38 enhancer viewpoints. **B.** Mpg and Rhbdf1 promoter viewpoints.

3.3 Discussion

A detailed analysis of CTCF binding sites and their interactions at the α-globin locus, reveals two clusters of CTCF and cohesin bound sites that flank the α-globin genes and super-enhancer. Moreover, the CTCF sites comprising these clusters are orientated such that CTCF binding sites in one flanking cluster are convergent with CTCF sites located in the other. The analysis of local chromatin interactions by Capture-C reveals that interactions between these flanking CTCF clusters delimit the interactions of the α-globin genes and enhancer elements.

Indeed, it was recently shown that strong enhancer clusters and the genes they regulate are flanked by interacting CTCF and Cohesin bound sites in ES cells (Dowen et al. 2014). However, this study investigated only interactions between cohesin-bound sites by ChIA-PET, identifying only punctate interactions between CTCF binding sites, enhancers and promoters. The high-resolution interactions observed by Capture-C at the α-globin locus are consistent with these observations, but in addition reveal the full topology of a super-enhancer locus in erythroid cells. The α-globin regulatory elements do not interact with genomic positions outside of the compartment delimited by the CTCF clusters, suggesting that the interactions between these clusters may physically constrain interactions within the central ~70kb α-globin compartment. The interaction profiles of flanking genes *Rhbdf1* and *Mpg* are similar to those of nearby CTCF sites HS-38 and HS-39, suggesting that these too, may be constrained by CTCF-CTCF interactions across the α-globin cluster. Thus, this local chromatin topology prevents flanking genes from interacting with the α-globin enhancer region, identifying a potential mechanism for the insulation of

these non-erythroid genes from α-globin enhancer activity. In support of this hypothesis, expression of the *Nprl3* gene, whose promoter is located inside the α-globin compartment, is upregulated more than 6-fold in erythroid cells whereas expression of genes directly flanking the α-globin compartment does not change.

Other examples of gene clusters where CTCF binding sites follow a highly organised organisation into clusters with the same binding orientation have been recently described. The mouse protocadherin (*Pcdh*) gene cluster has been found to contain clusters of CTCF binding sites that are paired between the enhancers and *Pcdh* promoters (Guo *et al.* 2012b; 2015). Similarly, two clusters of CTCF binding sites in divergent orientations separate the rostral and caudal genes of the human *HoxD* cluster (Guo *et al.* 2015). This is consistent with studies showing that these genes are differentially activated by interacting with enhancers in flanking genomic regions in mice (Andrey et al. 2013), showing that such domains are likely of great importance for the proper regulation of enhancer-promoter specificity.

While it is well-established that interactions between gene enhancers and promoters are largely confined to the cell types in which the genes are active (Davies *et al.* 2015; Whalen, Truty and Pollard 2016), a comparison of Capture-C data between erythroid and ES cells reveals that CTCF-CTCF interactions between flanking clusters are also stronger in erythroid cells despite similar levels of CTCF binding to these sites in both cell types. It is possible that the strong, erythroid-specific interactions between the α-globin enhancers and gene promoters assist in stabilising the interaction between flanking CTCF clusters. Alternatively, additional factors could be recruited to these CTCF clusters in erythroid cells. As cohesin is thought to be

required for interactions between CTCF sites , an attractive model would be that the additional recruitment of the cohesin complex to the active α-globin enhancer leads to more cohesin binding at these CTCF clusters in erythroid cells (Kagey et al. 2010). Cohesin is loaded onto chromatin by Nipbl, which co-localises with Mediator binding, but is not present at CTCF binding sites. Enhancer-loaded cohesin could then stabilise flanking CTCF-CTCF interactions via a loop-extrusion mechanism, as was recently proposed (Nasmyth 2001; Sanborn *et al.* 2015). A quantitative comparison of cohesin binding would reveal whether more cohesin is present at enhancers and flanking CTCF sites in erythroid cells (Hu et al. 2015).

Contrary to previous reports (Nakahashi et al. 2013; Rao et al. 2014), most CTCF binding sites identified by ChIP-seq in primary erythroid cells contained the core CTCF binding motif. Additionally, a smaller number of genome-wide CTCF binding sites (17,460) were identified by ChIP-seq in erythroid cells than those previously reported in other cell types (48,156 in B cells (Nakahashi et al. 2013), average of 55,000 in 19 diverse cell-types (Wang et al. 2012)). As binding sites with a close match to the CTCF consensus motif bind CTCF more strongly (Schmidt et al. 2012), this suggests that only strong CTCF binding sites are present in definitive erythroid cells.

Although the results presented in this chapter strongly suggest that the observed interactions between genomic regions flanking the α-globin cluster are mediated by CTCF, functional analysis of these CTCF binding sites via genome-editing is required to confirm whether their presence is required for the maintenance of this local chromatin topology.

Chapter 4: Deletion of a CTCF boundary results in an expansion of the α-globin compartment

4.1 Introduction

CTCF has been the subject of intense study since it was first discovered and characterised as a vertebrate transcription factor (Lobanenkov et al. 1990; Klenova et al. 1993). Enhancer-blocking reporter assays with different fragments from the β-globin (Chung, Bell and Felsenfeld 1997; Bell, West and Felsenfeld 1999) and *Igf2/H19* loci (Bell and Felsenfeld 2000; Hark et al. 2000) loci first identified CTCF binding sites as a key component of vertebrate insulators. Early evidence that this insulator function may arise from the ability to form interactions between its binding sites (Splinter 2006b) corresponds with the findings of recent high-resolution Hi-C studies, identifying a large number of smaller topological domains (termed “contact domains”) that are enclosed by CTCF and cohesin-anchored interactions (Rao et al. 2014).

Although the depletion of CTCF has been shown to result in partial loss of boundaries between TADs (Zuin et al. 2014), many attempts at functional study of CTCF in its natural chromatin environment by knockout or knockdown have been hampered by the essential role of CTCF in development and the regulation of apoptosis (Docquier *et al.* 2005; Wan *et al.* 2008; Gomes and Espinosa 2010; Soshnikova *et al.* 2010; Moore *et al.* 2012). Even when CTCF could be efficiently depleted, a core set of highly conserved CTCF binding sites was found to be

resistant to knockdown (Schmidt et al. 2012). While a small number of studies have used gene targeting approaches to remove individual CTCF binding elements (Engel 2006; Splinter 2006b; Gombert and Krumm 2009; Spencer *et al.* 2011; Volpi *et al.* 2012), the recent emergence of novel genome-editing technologies has vastly improved the ease and efficiency of targeted CTCF binding disruption at loci of interest. These technical advances have enabled the detailed functional analysis of these regulatory elements in shaping local chromatin topology and gene regulation.

Several genome-editing tools allowing the *in vivo* targeting of mammalian model organisms have been developed in rapid succession. These technologies rely on the creation of a targeted double-strand break (DSB) at the genomic site of interest. Repair of these DSBs via the non-homologous end joining (NHEJ) pathway results in the creation of small insertions or deletions (indels) which can be employed to create mutation that result in loss of gene function or disruption of non-coding regulatory elements. If a repair template is provided, desired specific mutations can be obtained via homology directed repair (HDR). Zinc-finger nucleases (ZFNs) are the first class of such programmable nucleases, but suffer from poor targeting density (~one site per 100bp), cytotoxicity, and variable targeting efficiency (Kim and Kim 2014). The second class, transcription activator-like effector nucleases (TALENs), eliminates most of these problems, improving upon both targeting efficiency and the availability of target sites (Cermak et al. 2011). However, both ZFNs and TALENs rely on combining arrays of DNA-binding domains to tether a *FokI* endonuclease to specific genomic sites, construction of which is challenging and time-consuming. The emergence of the RNA-guided CRISPR-Cas system, in which the Cas9 nuclease is recruited to target DNA by a complementary RNA strand, markedly improved the

ease of design and construction. In addition, this method allows multiplexed targeting of genomic sites while maintaining targeting efficiency and specificity (Ran *et al.* 2013b).

Various studies employing CRISPR-Cas9 genome-editing to probe CTCF function have confirmed the importance of CTCF in the maintenance of local chromosome organisation. Deletion of CTCF binding sites within the *Hox* clusters in mouse ES cells results in increased interactions between active and repressed chromatin domains, with concurrent activation of *Hox* genes located in the silenced domain (Narendra et al. 2015). The potential phenotypic consequences of such disruptions in domain boundaries are made clear in experiments deleting and inverting a CTCF-associated TAD boundary near the limb enhancers that normally control the mouse *Epha4* gene. These rearrangements result in ectopic enhancer contacts and altered gene expression patterns, leading to limb malformations (Lupiáñez et al. 2015). Similar studies explored the importance of the orientation of CTCF binding sites in domain organisation. Inversion of CTCF binding sites at the human *Pcdh* (protocadherin) and β-globin locus leads to a shift in 4C interactions conforming to the change in binding orientation. The discovery that CTCF binding sites preferentially interact when convergently orientated was further explored in a Hi-C study in which deletions and inversions of CTCF binding sites result in the loss and gain of interactions conforming to this rule (Sanborn et al. 2015). Similarly, a loss of looping is observed at several loci in ES cells when inversion of CTCF binding results in a loss of convergence between sites (de Wit et al. 2015).

In this chapter, I describe the use of a combination of TALEN and CRISPR-Cas9 mutagenesis to investigate the role of enhancer-proximal CTCF binding sites in the establishment of chromatin topology at the mouse α-globin locus. In erythroid cells, domains flanking the α-globin cluster interact with each other and delimit interactions of the α-globin genes and enhancer region that span a central chromatin compartment of roughly 70 kb (see **Chapter 3** and (Davies *et al.* 2015)). To test whether CTCF binding sites in these domains are responsible for the establishment of this chromatin topology, the three CTCF binding sites closest to the α-globin enhancer region and demarcating the upstream edge of the α-globin compartment were deleted in mice, singly and in an informative combination. After confirmation of deletion using ChIP-seq for CTCF, I analysed chromatin interactions in the HS-38 and HS-39 double deletion using Capture C. Chromatin interactions in the HS-29 deleted model were also analysed using the same method. Given time limitations, Capture C of the HS-38 and HS-39 single deletions, was not performed.

4.2 Results

4.2.1 Deletion of three CTCF binding sites at the α-globin enhancer region

A cluster of three CTCF binding sites marking the upstream edge of the erythroid-specific α-globin compartment was targeted for mutagenesis *in vivo*. Two sites, HS-38 and HS-39, are located less than 1.5 kb apart just upstream of the R1 enhancer while the third, less prominent site, HS-29, is located between the R1 and R2 enhancers (Fig. 4.1A). Due to the location of these CTCF binding sites both close to the α-globin enhancers and exons of the *Nprl3* gene, precise disruption of CTCF binding was required at each of the sites. Nuclease activity was targeted to the best match to the CTCF binding motif below the CTCF ChIP-seq peak. Before targeting, CTCF binding to these sequences was confirmed by DNaseI footprinting analysis (Fig. 4.1A).

Individual CTCF binding sites were sequentially targeted for mutagenesis. A pair of TALENs flanking the HS-38 core binding sequence were designed and constructed as previously described (Cermak *et al.* 2011; Davies *et al.* 2013). Microinjection of TALEN mRNA into mouse oocytes (C57BL/6J) resulted in mice with a 19 bp deletion that removed most of the HS-38 CTCF core binding motif (D38, Fig. 4.1B).

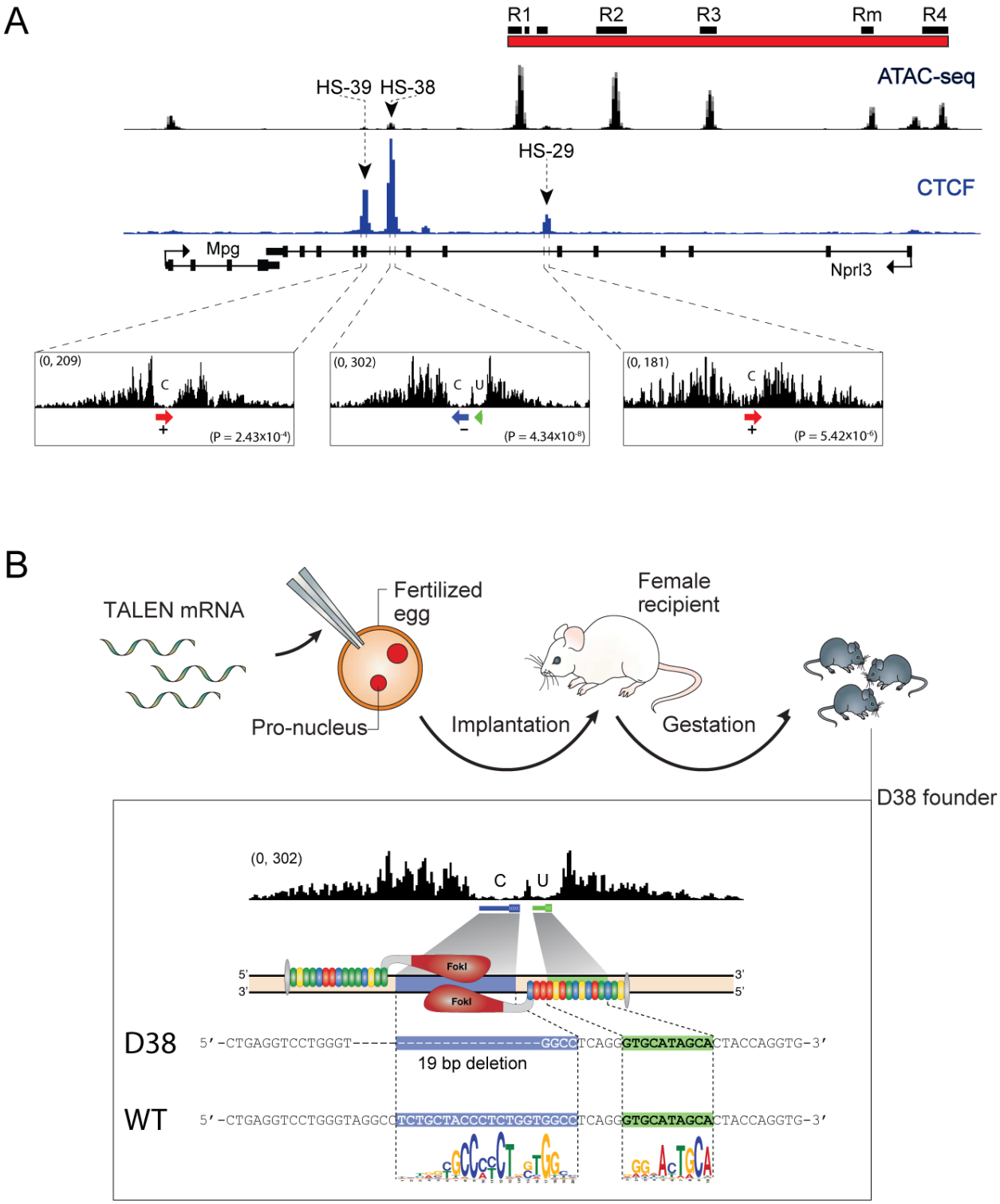


Figure 4.1 Overview of targeted CTCF sites and the deletion of HS-38 in mice. A. Annotated CTCF ChIP-seq and ATAC-seq at the three targeted CTCF binding sites, DNaseI footprints and the position of the CTCF motif are shown for each site. **B**. Schematic overview of the production method and TALEN induced mutation at HS-38 (D38). DNaseI footprints and the position of the CTCF motif (red = forward, blue = reverse) are shown for each targeted site.

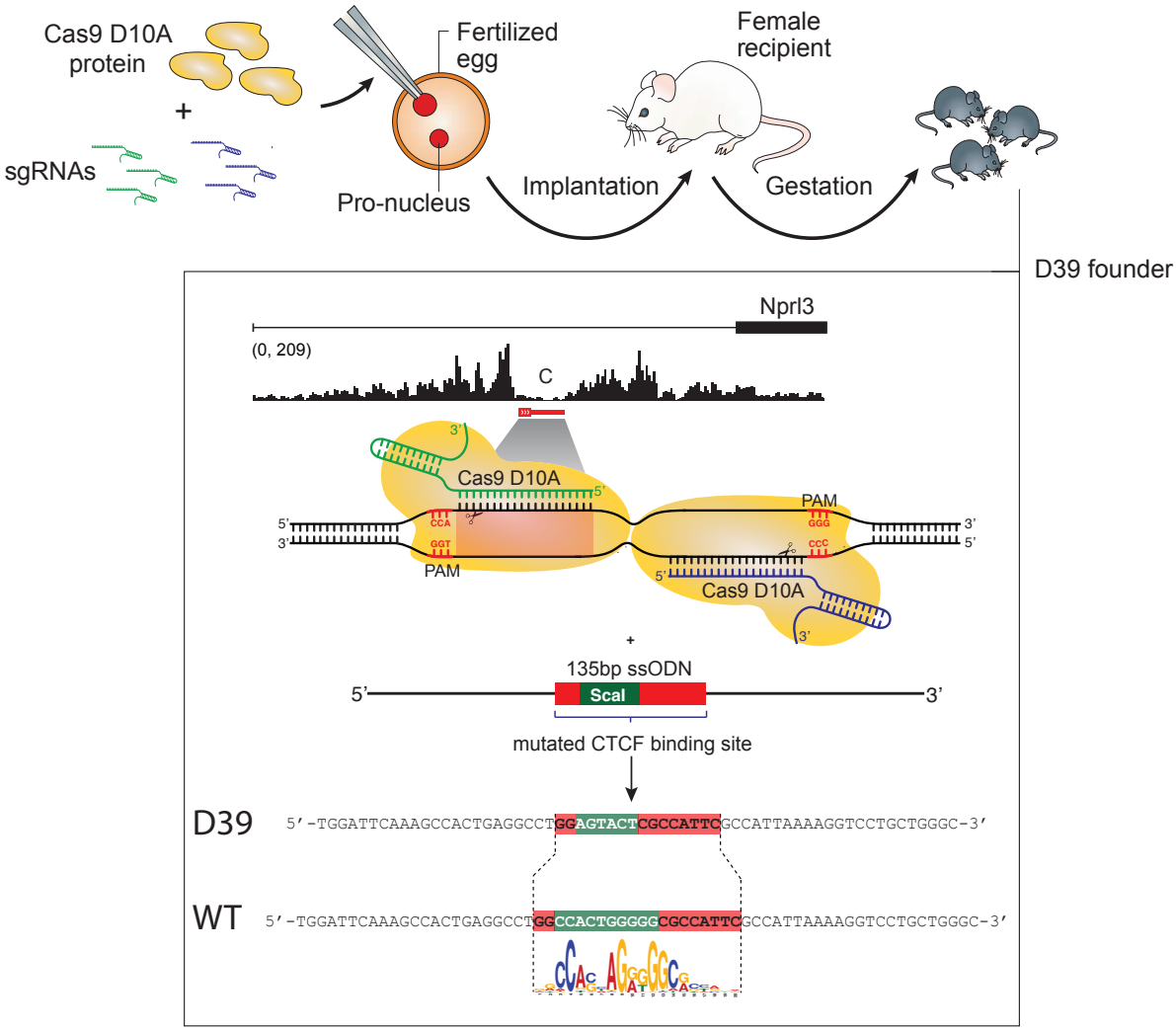


Figure 4.2 The deletion of HS-39 in mice. Schematic overview of the production method and induced mutation at HS-39 (D39). The DNaseI footprint and the position of the CTCF motif (red = forward) is shown for the HS-39 site. The green block indicates the position of the ScaI restriction motif in D39 and the deleted segment of the CTCF binding sequence in the WT.

Next, the HS-39 site was targeted singly and in combination with the deletion of HS-38 by CRISPR-Cas9 mutagenesis. For the single deletion of HS-39, Cas9 D10A nickase protein and two sgRNAs directed against the CTCF core binding sequence were injected into mouse oocytes (C57BL/6J) by pronuclear injection (Sung et al. 2014). To control the size of the sequence disruption and to allow easy screening of mice, a single-stranded oligodeoxynucleotide (ssODN, 135 bp) in which a 10 bp CTCF core binding fragment was replaced with a ScaI restriction recognition site was added (Wang et al. 2013). This resulted in mutation of the CTCF core sequence by HDR (D39, Fig. 4.2). To create a mouse lacking CTCF binding at both HS-38 and HS-39, a D38 male was crossed with a female in which a CAG-NLS-Cas9 overexpression cassette is integrated into the *Gt(ROSA26)Sor* locus (Fig. 4.3). In these mice, constitutive overexpression of Cas9 results in the genetic delivery of Cas9 to oocytes requiring only the injection of the sgRNA (*submitted for publication*, Cebrian-Serrano et al.), similar to previous reports (Zhang *et al.* 2016). Injection of the sgRNA against HS-39 created a mouse containing a 26 bp deletion that included the CTCF core binding sequence and was in *cis* with the previously generated D38 mutation, resulting in an allele with mutations in both the HS-38 and HS-39 core binding sequences (D3839).

Finally, a mouse carrying a single mutation for the HS-29 binding site was created by microinjection of sgRNA and a ssODN into oocytes derived from female mice homozygous for the Cas9 transgene (Fig. 4.4). This resulted in the replacement of a 10 bp sequence of the CTCF core motif with a PvuII restriction site (D29). Thus, four mouse models were created in total to study the function of CTCF binding at the α-globin enhancers.

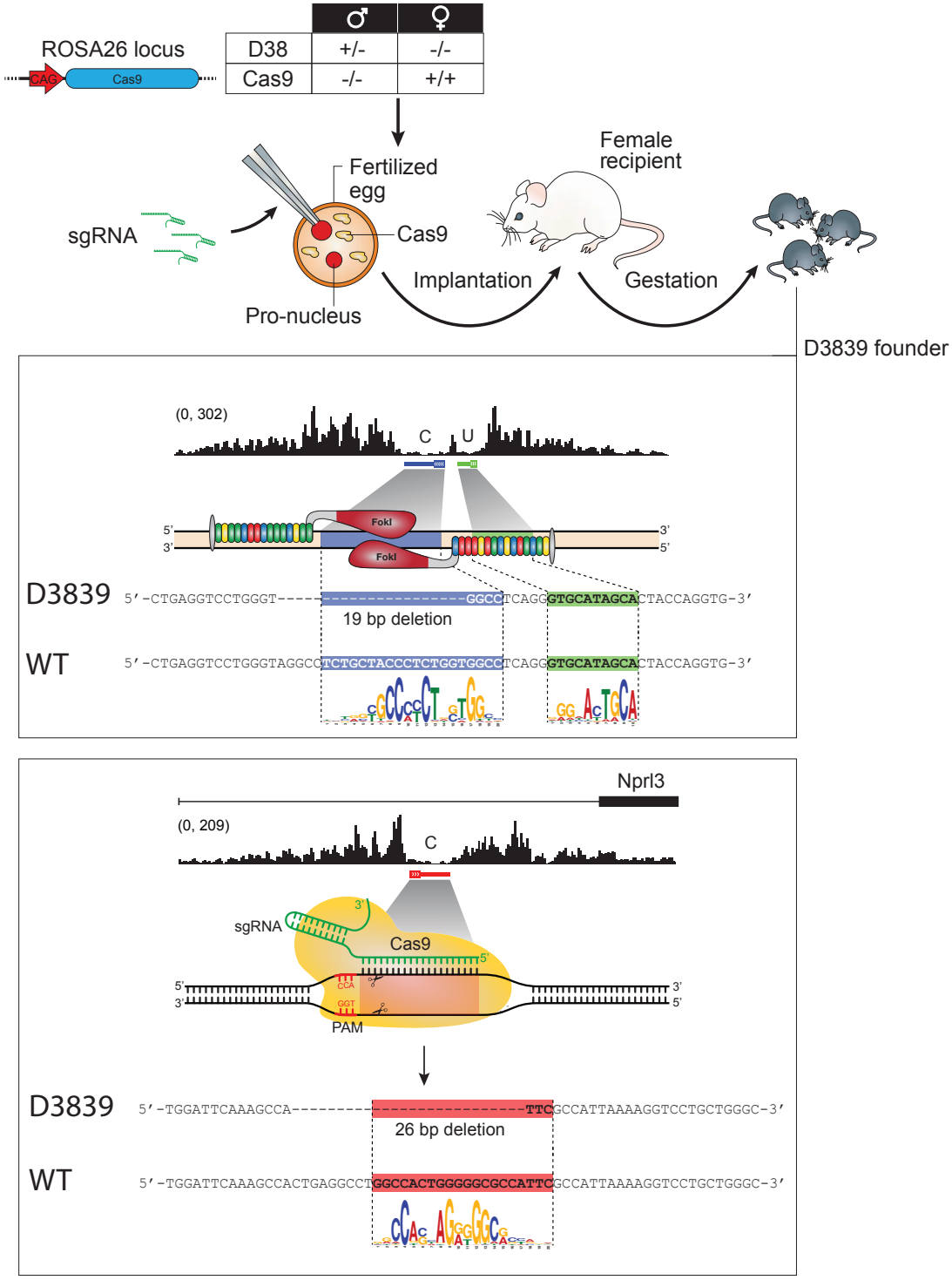


Figure 4.3 The combined deletion of HS-38 and HS-39 in mice. Schematic overview of the production method and induced mutations at HS-38 (top panel) and HS-39 (bottom panel) CTCF binding sites in the D3839 double mutant. DNaseI footprints and the position of the CTCF motif (red = forward, blue = reverse) is shown for the HS-38 and HS-39 sites.

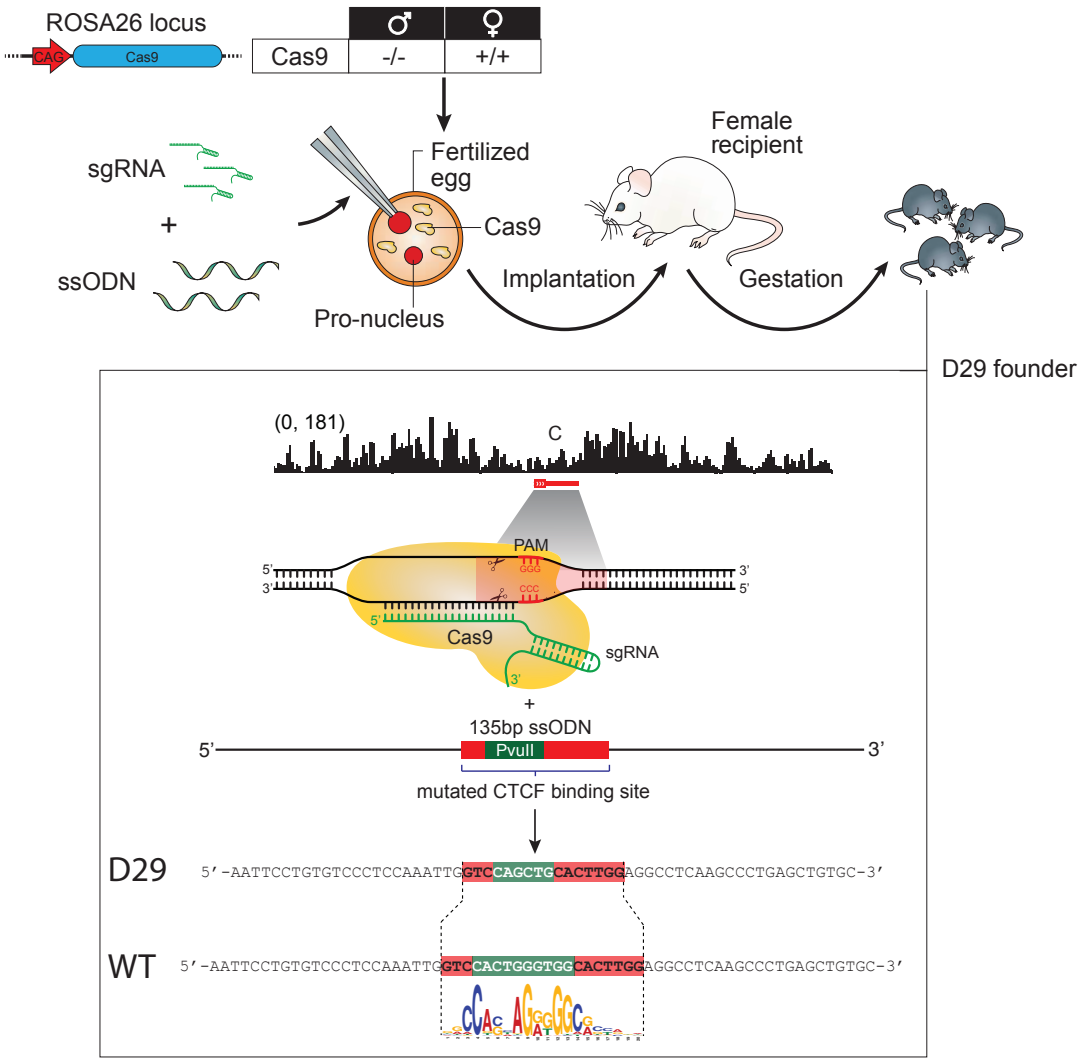


Figure 4.4 The deletion of HS-29 in mice. Schematic overview of the production method and induced mutation at HS-29 (D29). The DNaseI footprint and the position of the CTCF motif (red = forward) is shown for the HS-29 site. The green block indicates the position of the PvuII restriction site in the D29 sequence and deleted segment of the CTCF binding site in the WT sequence.

4.2.2 Mutations in the CTCF core motif result in the abrogation of CTCF binding

To assess whether these mutations result in a loss of CTCF binding at these genomic sites, I performed and analysed ChIP-seq for CTCF in erythroid cells isolated from each mouse model (Fig. 4.5). A complete abrogation of CTCF binding is observed at each of the targeted sites. It was previously reported that deletion of individual CTCF binding elements can result in either loss of CTCF binding (Narendra et al. 2015) or compensation in CTCF binding (Yang *et al.* 2015) at CTCF sites near the deletion. However, none of the CTCF deletion mice showed significant alterations in CTCF binding at other CTCF binding sites throughout the α-globin locus in erythroid cells (Fig. 4.5).

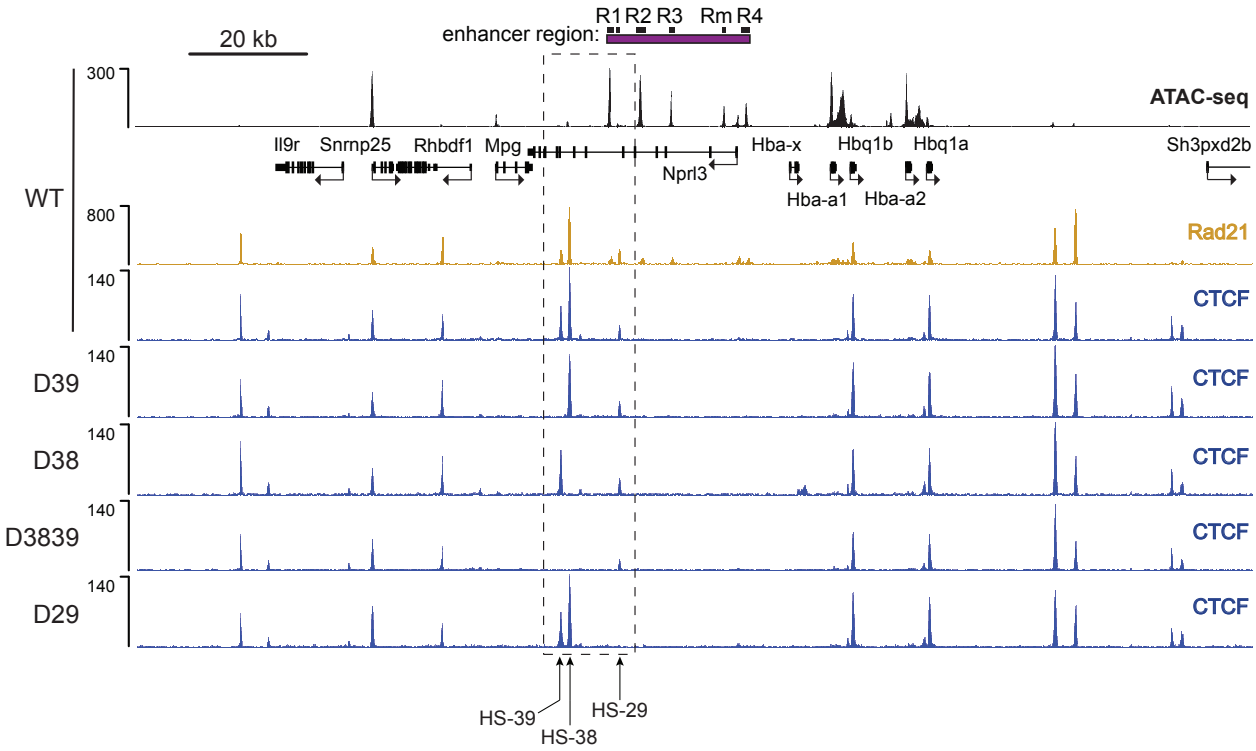


Figure 4.5 Overview of CTCF binding in CTCF deletion mouse models. Normalised CTCF ChIP-seq reads (RPKM) of two biological replicates along the α-globin locus for each of the generated CTCF binding site mutants. Also shown are normalized Rad21 ChIP-seq (RPKM) and ATAC-seq (RPKM) from wild-type erythroid cells.

4.2.3 Combined deletion of HS-38 and HS-39 results in loss of directionality and specificity of α-globin enhancer interactions

To investigate whether the removal of the enhancer-flanking HS-38 and HS-39 CTCF binding sites results in changes in local genome topology, I initially performed Capture-C from viewpoints across the α-globin locus in wild-type (WT) and D3839 erythroid cells. Capture-C data in WT erythroid cells shows strong interactions between the HS-38 and HS-39 sites and a cluster of CTCF binding sites on the other side of the α-globin locus (see **Chapter 3**). To examine whether these interactions are lost upon removal of these sites in the D3839 mutant, interaction profiles centered on the downstream, flanking HS44 and HS48 CTCF binding sites were generated in D3839 mutant erythroid cells and compared to WT (Fig 4.6). The mutation of HS-38 and HS-39 causes a local loss of interactions between HS44 and HS48 and the deleted CTCF sites. Although interactions between flanking clusters of CTCF sites are still maintained in D3839 erythroid cells, the start of this interacting chromatin region moves away from HS-38 and HS-39 towards the upstream HS-59 CTCF site. Furthermore, HS44 and HS48 interact more strongly with the HS-94 CTCF site in the D3839 mutant. Thus, deletion of HS-38 and HS-39 results in a shift of interactions across the α-globin compartment towards upstream, more distal CTCF binding sites.

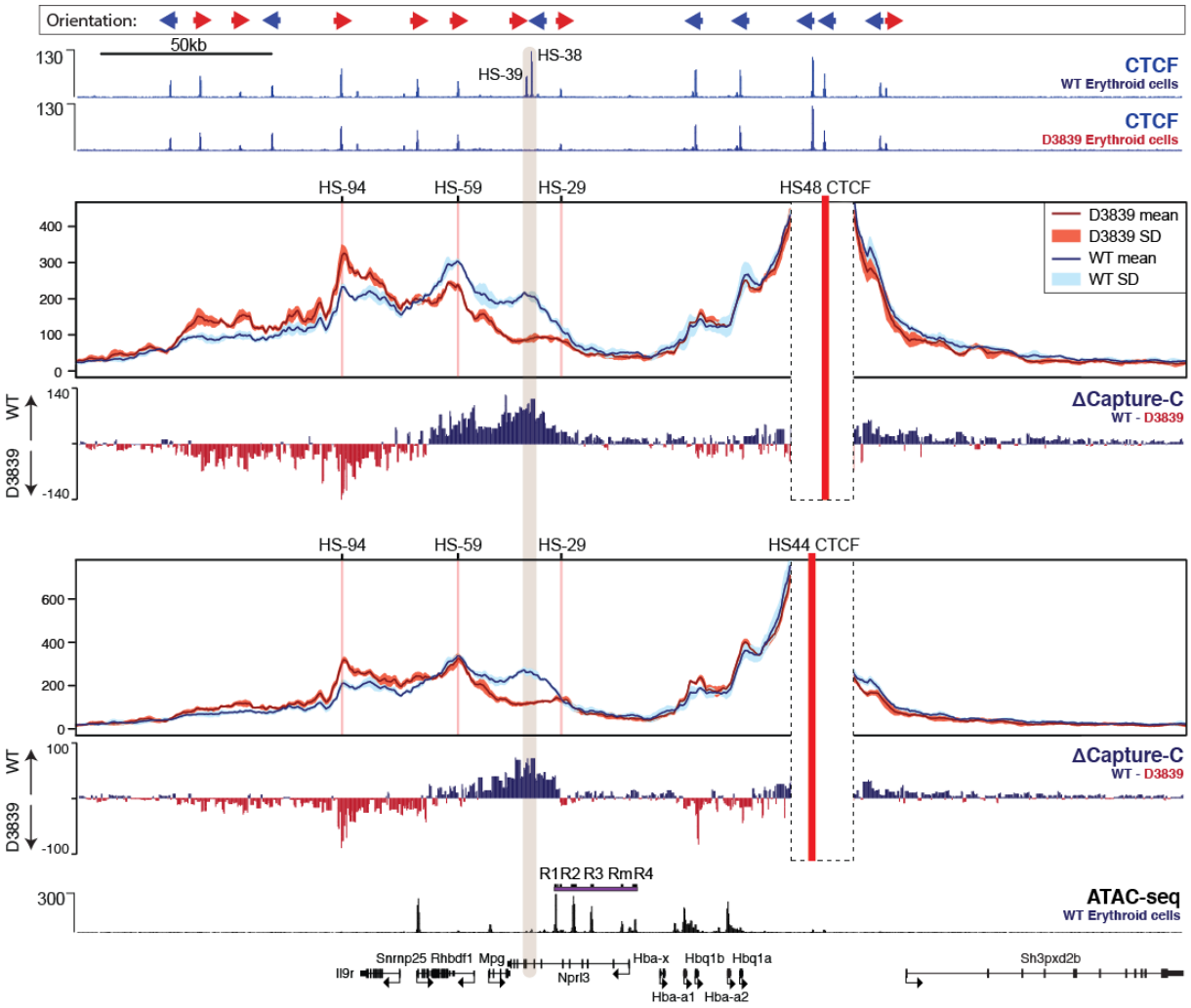
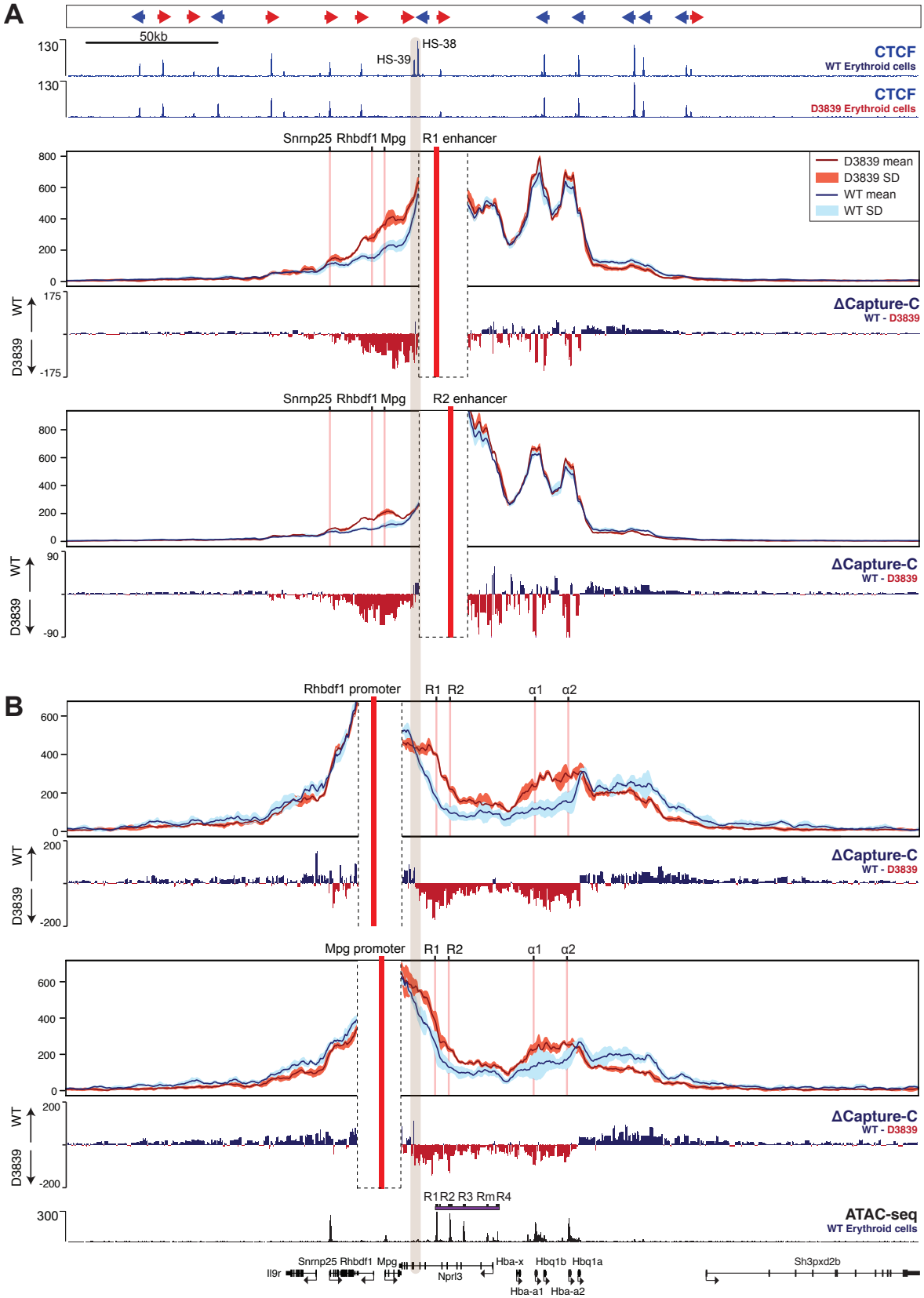


Figure 4.6 Differential interactions of HS44 and HS48 CTCF binding sites between wild-type and D3839 erythroid cells. Panels show overlaid, normalised Capture-C data for the HS44 and HS48 CTCF site viewpoints in wild-type and D3839 erythroid cells merged across three biological replicates. The mean number of interactions, scaled to a total of 100,000 interactions genome-wide plus and minus one standard deviation (SD) of sliding 5kb windows are visualised. Differential tracks (ΔCapture-C) show a subtraction (WT – D3839) of the mean number of meaningful interactions per restriction fragment, scaled to a total of 100,000 interactions genome-wide. Red vertical bars indicate the position of the viewpoint. Mutated CTCF sites are indicated with a transparent grey vertical bar. Also shown are normalised CTCF ChIP-seq (RPKM) for both wild-type and D3839 cells and ATAC-seq (RPKM) from wild-type erythroid cells, all merged across two biological duplicates. Gene annotation is Refseq.

To see how the loss of HS-38 and HS-39 affected the interactions of the enhancers, Capture-C interaction profiles using the R1 and R2 enhancers as viewpoints were compared between WT and D3839 mutants. While interactions between the enhancers (R1 and R2) and the α-globin promoters (α1 and α2) appeared unchanged, ablation of CTCF binding at the HS-38/-39 sites results in increased interactions between the enhancers and a region of chromatin, directly upstream, containing the *Mpg* and *Rhbdf1* genes (Fig. 4.7A). Particularly the most distal enhancer, R1, shows a strong increase in interactions with the upstream region of chromatin. These observations are further confirmed by interaction profiles obtained using the *Rhbdf1* and *Mpg* promoters as viewpoints, which show reciprocal interactions with the R1 and R2 enhancers and the α-globin genes (Fig. 4.7B). While these two promoters interact more with chromatin inside the α-globin compartment in the D3839 mutant, decreases in interactions with flanking genomic regions, especially the region downstream of the α-globin genes, are observed. Thus, the abrogation of CTCF binding in the D3839 mutant is associated with a partial loss in directionality of α-globin enhancer interactions resulting in an entirely new set of contacts between the α-globin enhancers and the upstream *Rhbdf1* and *Mpg* genes.



See next page for figure legend

Figure 4.7 Differential interactions of α-globin enhancers and flanking genes between wild-type and D3839 erythroid cells. Panels show overlaid, normalised Capture-C data for the indicated viewpoints in wild-type and D3839 erythroid cells merged across three biological replicates. The mean number of interactions, scaled to a total of 100,000 interactions genome-wide plus and minus one standard deviation (SD) of sliding 5kb windows are visualised. Differential tracks (ΔCapture-C) show a subtraction (WT – D3839) of the mean number of meaningful interactions per restriction fragment, scaled to a total of 100,000 interactions genome-wide. Red vertical bars indicate the position of the viewpoint. Mutated CTCF sites are indicated with a transparent grey vertical bar. Also shown are normalised CTCF ChIP-seq (RPKM) for both wild-type and D3839 cells and ATAC-seq (RPKM) from wild-type erythroid cells, all merged across two biological duplicates. Gene annotation is Refseq. **A.** R1 and R2 enhancer viewpoints. **B.** Rhbdf1 and Mpg promoter viewpoints. Figure is on the previous page.

Together, these results suggest that the interactions between the downstream HS44/48 and the upstream HS-38/-39 CTCF binding sites constrain the interactions of flanking genes and the α-globin enhancer in WT erythroid cells. Loss of this interaction in the D3839 mutant, results in an expansion of α-globin compartment, such that it is now delimited by the upstream HS-59 CTCF site. This then permits interactions between the non-erythroid genes *Mpg* and *Rhbdf1* and the α-globin enhancer.

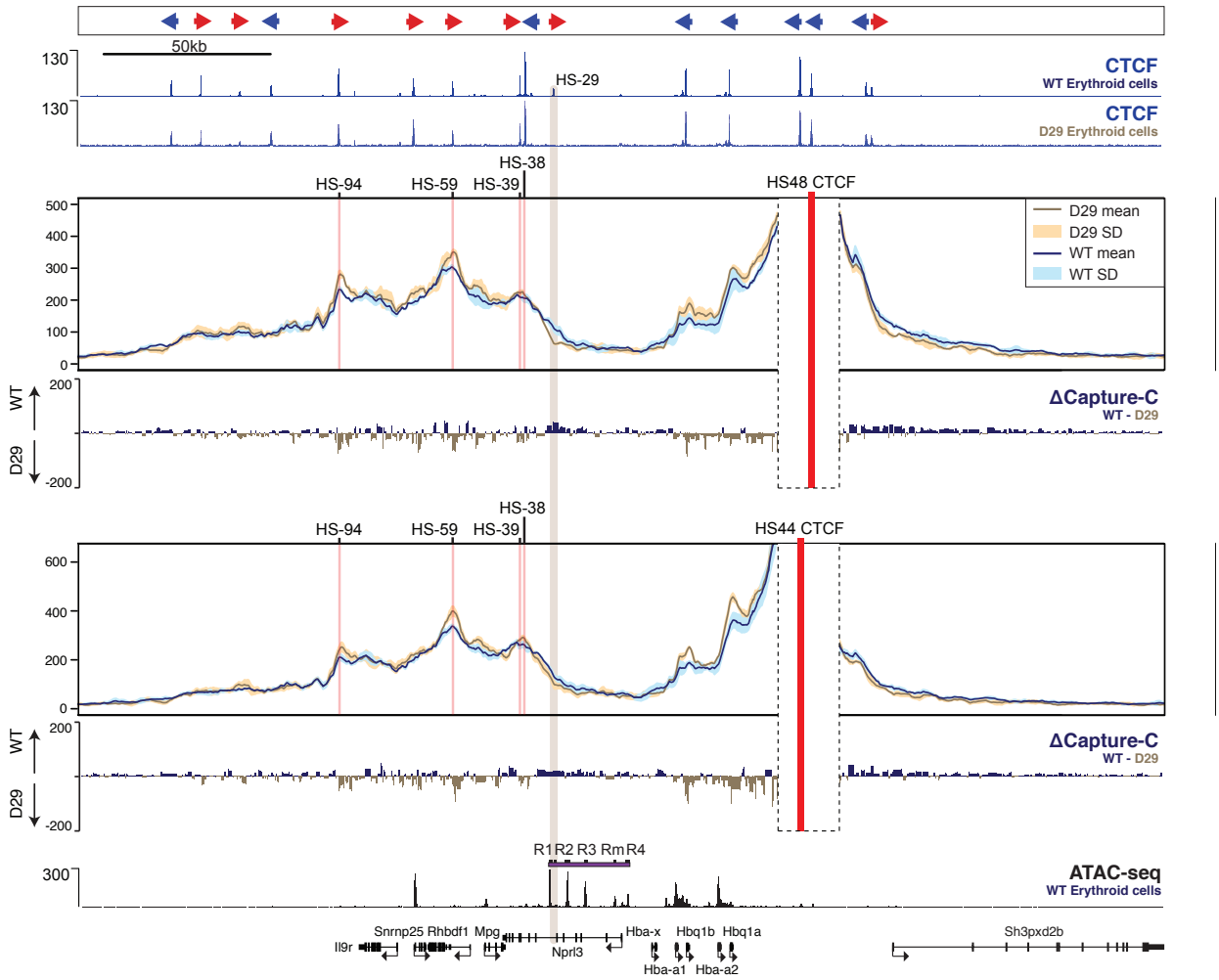


Figure 4.8 Differential interactions of HS44 and HS48 CTCF binding sites between wild-type and D29 erythroid cells. Panels show overlaid, normalised Capture-C data for the HS44 and HS48 CTCF site viewpoints in wild-type and D29 erythroid cells merged across three biological replicates. The mean number of interactions, scaled to a total of 100,000 interactions genome-wide plus and minus one standard deviation (SD) of sliding 5kb windows are visualised. Differential tracks (ΔCapture-C) show a subtraction (WT – D29) of the mean number of meaningful interactions per restriction fragment, scaled to a total of 100,000 interactions genome-wide. Red vertical bars indicate the position of the position of the viewpoint. Mutated CTCF sites are indicated with a transparent grey vertical bar. Also shown are normalised CTCF ChIP-seq (RPKM) for both wild-type and D29 cells and ATAC-seq (RPKM) from wild-type erythroid cells, all merged across two biological duplicates. Gene annotation is Refseq.

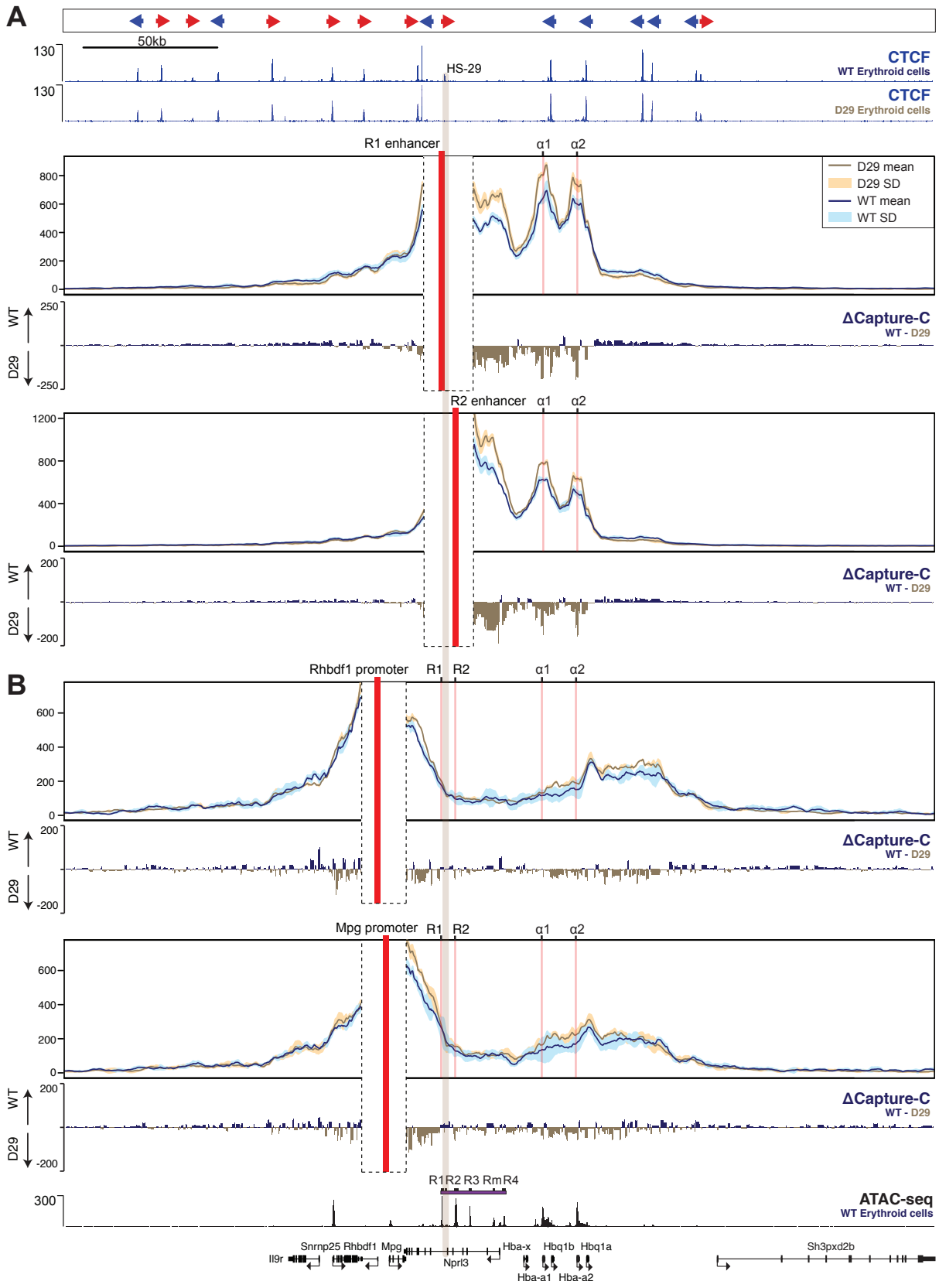
4.2.4 Deletion of HS-29 does not result in aberrant interactions

The HS-29 site is located between the R1 and R2 enhancers and is the first in cluster of CTCF binding sites upstream of the α-globin genes. It is located at the start of the genomic region interacting with the downstream HS44/48 sites. To test whether removal of HS-29 resulted in changes in these interactions, Capture-C was again performed using HS44 and HS48 as viewpoints (Fig. 4.8). No gross changes in interaction profiles generated for several viewpoints were observed in the absence of HS-29. Deletion of CTCF binding at HS-29 results in a subtle, local loss of interactions between HS44/48 and HS-29 in combination with small increases in interactions with other proximal CTCF binding sites.

Interestingly, the α-globin enhancers show increased contacts with chromatin inside the α-globin compartment in D29 mice, including interactions with the α-globin promoters (Fig. 4.9A). These enhanced interactions inside the α-globin compartment are accompanied by diffuse decreases in interactions with flanking genomic regions, especially the region directly downstream of the α-globin genes containing the HS44/48 CTCF binding sites. These data suggest that the subtle loss of interactions between HS-29 and flanking CTCF sites in the D29 mutant may lead to the consolidation of interactions inside the central α-globin compartment, including those between the α-globin enhancers and promoters.

Finally, interaction profiles of the flanking *Rhbdf1* and *Mpg* promoters were compared between D29 and WT to investigate whether, similar to the D3839 mutant, removal of HS-29 results in increased contacts between the α-globin enhancers and these genes (Fig. 4.9B). No such changes in interactions were observed in the D29 mutant, in which the interaction profiles of the two flanking gene promoters were remarkably similar to WT.

Figure 4.9 Differential interactions of α-globin enhancers and flanking genes between wild-type and D29 erythroid cells. Panels show overlaid, normalised Capture-C data for the HS44 and HS48 CTCF site viewpoints in wild-type and D29 erythroid cells merged across three biological replicates. The mean number of interactions, scaled to a total of 100,000 interactions genome-wide plus and minus one standard deviation (SD) of sliding 5kb windows are visualised. Differential tracks (ΔCapture-C) show a subtraction (WT – D29) of the mean number of meaningful interactions per restriction fragment, scaled to a total of 100,000 interactions genome-wide. Red vertical bars indicate the position of the position of the viewpoint. Mutated CTCF sites are indicated with a transparent grey vertical bar. Also shown are normalised CTCF ChIP-seq (RPKM) for both wild-type and D29 cells and ATAC-seq (RPKM) from wild-type erythroid cells, all merged across two biological duplicates. Gene annotation is Refseq. **A.** R1 and R2 enhancer viewpoints. **B.** Rhbdf1 and Mpg promoter viewpoints. Figure is on the next page.



See previous page for figure legend

4.3 Discussion

By employing TALEN- and CRISPR-Cas9-mediated targeting of CTCF binding sequences, I generated four mouse models with mutations in either one (D29, D38, D39) or two (D3839) CTCF binding sites located upstream the α-globin enhancers. The analysis of local chromatin interactions in two of these mutants (D29 and D3839), revealed that the two CTCF binding sites directly upstream of the α-globin enhancer region constrain the interactions of R1 and R2, the strongest of the α-globin enhancer elements (see **Chapter 3**, section 3.2.1). Although the strong interactions between the enhancers and the α-globin promoters are not affected in D3839 mutants, directionality of enhancer interactions is partially lost in erythroid cells of these mice. Moreover, changes in interactions from flanking Capture-C viewpoints show that a genomic region containing the *Mpg* and *Rhbdf1* genes shifts in the local chromatin topology to be included in the α-globin compartment in D3839 mice.

These results are important as they suggest that CTCF insulator activity may rely on a mechanism by which CTCF topologically shields nearby genes from strong, tissue-specific enhancer activity. In this model, interactions between flanking CTCF clusters do not only create a loop across the α-globin locus, but also constrain the interactions of flanking genes via the establishment of this interaction. These data are consistent with the observation that removal of CTCF boundaries of a ~2.5 Mb TAD can lead to ectopic enhancer interactions (Lupiáñez et al. 2015); however, are here observed within a sub-TAD structure. This illustrates the hierarchical nature of

topological organisation and the importance of the use of high-resolution, quantitative 3C-based methods to resolve topological changes on this scale.

Although the deleted CTCF binding sites (HS-38 and HS-39) are orientated in a convergent orientation to each other, it was shown that two convergently orientated sites in close proximity to each other interact with surrounding sites similarly to a pair of CTCF binding sites in a divergent orientation (Sanborn et al. 2015). It was proposed that such pairs of oppositely orientated CTCF binding sites create effective TAD boundaries by forming interactions with CTCF binding sites in both directions (Gómez-Marín et al. 2015). Although the HS-38/39 CTCF pair does not form a TAD boundary, both CTCF sites show bidirectional interaction profiles interacting with up- and downstream CTCF binding sites (see **Chapter 3**, section 3.2.3). While the observed interactions between flanking CTCF clusters suggest that the local chromatin topology orchestrated by these interactions is responsible for delimiting the α-globin compartment, the ability of the HS-38/39 pair to establish bidirectional interactions may also underlie some of its ability to restrict enhancer contacts..

Evidence suggests that CTCF binding sites located near enhancers may function to tether these distal elements to their promoters at some loci (Yagi *et al.* 2012; Ong and Corces 2014). It is conceivable that CTCF binding sites at the α-globin cluster could serve such a function; the two CTCF binding sites at the θ-globin promoters are in a convergent orientation with HS-29, which is located in between the R1 and R2 enhancers. However, neither of the CTCF deletions studied here (D29 and D3839) show a decrease in enhancer interactions. Rather, deletion of HS-29 results in increased interactions between the enhancers and their target promoters in

erythroid cells. Although it is possible that CTCF helps in the initial establishment of the enhancer-promoter interaction during erythroid differentiation, the interaction data from CTCF mutants presented here is not compatible for a role of CTCF in the establishment of the interaction between the α-globin enhancers and promoters.

While the single deletion of HS-29 did not result in an expansion of enhancer contacts, I noticed that the R1 enhancer saw a greater loss in directionality of its interactions than R2 in the D3839 mutants. As HS-29 is positioned between these two elements, it may still partially restrain R2 interactions in the absence of HS-38 and HS-39. Although a mutant of all three sites was not created, this predicts the α-globin R2 enhancer would further expand its upstream interactions in the erythroid cells of such a mouse.

In conclusion, this work shows that the *in vivo* deletion of a CTCF boundary upstream the α-globin locus results in partial loss in the directionality of enhancer interactions. High-resolution study of local interactions reveals the topological changes that occur within the sub-TAD that contains the α-globin cluster. The genomic region subject to increased enhancer contacts contains three non-erythroid genes. Based on the proposed model for CTCF-dependent insulation, these genes would have been predicted to be upregulated in D3839 erythroid cells.

Chapter 5: Topological shielding of promoters from tissue-specific enhancers by CTCF is required for maintenance of transcriptional states

5.1 Introduction

Interphase mammalian chromosomes are folded into compartmentalised structures termed TADs which are separated by boundaries that are enriched for CTCF/cohesin binding sites. These boundaries have been shown to separate domains of “active” and “repressed” chromatin states (Dixon et al. 2012), especially when studied at a sub-TAD resolution (Sexton *et al.* 2012; Rao *et al.* 2014). Further evidence for the involvement of CTCF in the establishment of these domains is provided by a study of genome-wide CTCF-CTCF interactions by ChIA-PET, which found that interacting CTCF sites enclosed domains of active or repressed chromatin (Handoko et al. 2011).

Two well-characterised histone modifications that define their local chromatin environment are H3K27me3 and H3K4me3. Both are strongly linked to the regulation of gene expression; H3K4me3 is found at the promoters of active genes, while H3K27me3 is thought to mediate developmental repression of genes via the establishment of facultative heterochromatin (Simon and Kingston 2013b). In *Drosophila*, where Polycomb group proteins were first discovered, it was first shown that H3K4me3 and H3K27me3 are deposited on chromatin by antagonistic protein complexes (Geisler and Paro 2015). H3K27me3 is deposited by Ezh2, the catalytic

component of the Polycomb Repressive Complex 2 (PRC2) (Margueron and Reinberg 2011). While the antagonistic roles of H3K4me3 and H3K27me3 are conserved in mammals, the situation is complicated by the existence of bivalent domains at which both histone modifications are present (Voigt, Tee and Reinberg 2013). Bivalently marked genes have been suggested to be poised for activation upon development, and are generally repressed or expressed at low levels.

While genomic regions enriched for these histone modifications have been shown to reside in different sub-compartments, as defined by their Hi-C interaction pattern, the median size of these genomic intervals was ~300 kb (Rao et al. 2014). However, domains marked by H3K4me3 and H3K27me3 often occur interspersed with each other at much shorter genomic distances, potentially requiring boundaries that operate on a smaller genomic scale.

The mechanisms by which boundaries prevent the spread of chromatin state between adjacent regions are not clear. CTCF has been suggested to have specific chromatin barrier activity, preventing the spread of repressive chromatin marks (Cuddapah et al. 2008), although this is disputed (Phillips-Cremins et al. 2013; Ong and Corces 2014). CTCF spaces the nucleosomes around its binding site and incorporates the H3.3 histone variant, which was proposed to prevent the spread of H3K27me3 (Fu *et al.* 2008; Weth *et al.* 2014). Indeed, discontinuities of H3K27me3 deposition at the *Hox* gene clusters are demarcated by CTCF binding sites (Bowman *et al.* 2014; Narendra *et al.* 2015). However, when CTCF binding sites separating the rostral and caudal domains at the mouse *HoxA* cluster were deleted, an expansion of active chromatin (H3K4me3) was observed, resulting in reductions of H3K27me3

and transcriptional upregulation (Narendra et al. 2015). The *HoxA* gene cluster is located at the boundary between adjacent TADs (Andrey et al. 2013). Removal of CTCF binding sites resulted in increased interactions between the edges of the two topological domains, suggesting that CTCFs ability to separate active from repressive chromatin may stem from its ability to orchestrate local chromatin interactions. Similarly, the removal of one of a pair of interacting CTCF binding sites flanking the *Tcfap2e* gene and enhancers resulted in the modest upregulation of a neighbouring polycomb-repressed gene (Dowen et al. 2014).

Located within a sub-TAD of ~200kb, the *Rhbdf1* gene flanks the α-globin enhancer region and is marked by high levels of Polycomb repression in erythroid cells (Kowalczyk *et al.* 2012b). The deletion of two CTCF binding sites directly upstream the enhancer creates several ectopic chromatin contacts, including increased interactions between the α-globin enhancers and several non-erythroid genes, including *Rhbdf1* (see **Chapter 4**). In order to determine the function of CTCF in the separation of local chromatin and gene expression states, I analysed the effect of deleting CTCF binding sites in four mouse models with deletions in CTCF binding sites near the α-globin enhancer (see **Chapter 4**). For each of these mice, gene expression and chromatin states are analysed and interpreted in the context of changes in the local topology of the locus.

5.2 Results

5.2.1 Removal of a boundary upstream the α-globin enhancer results in loss of enhancer-promoter specificity

To assess the effect of deletion of CTCF binding in the D3839 mutant on gene expression in an unbiased manner, RNA-seq was compared between wild-type and D3839 erythroid cells (Fig. 5.1A). Whereas no change in expression of the Nprl3 gene, whose promoter is situated inside the α-globin enhancer region, is detected, three genes immediately upstream of the deleted CTCF binding sites are strongly upregulated in D3839 erythroid cells. The location of these genes corresponds to the area of expanded enhancer contacts observed in these mutants (see **Chapter 4**) and included *Mpg*, *Rhbdf1,* and the gene delimiting this expansion, *Snrnp25*. By contrast, expression of the *Il9r* gene was not increased in the D3839 mutant, consistent with the presence of two remaining CTCF binding sites between its promoter and the α-globin enhancers and lack of increased enhancer contacts with its promoter (see **Chapter 4**). Similarly, no change in transcription of the Polycomb-repressed *Sh3pxd2b* gene, located downstream of the α-globin cluster, was detected (Fig. 5.1A). As all globin transcripts were depleted from the sequenced RNA library due to their high-abundance, globin expression was quantitatively assessed by real-time PCR (Fig. 5.1B). No significant changes in α- or θ- globin expression were detected in the absence of the HS-38 and HS-39 CTCF binding sites, consistent with the absence of changes in enhancer interactions with the globin promoters in the D3839 mutant.

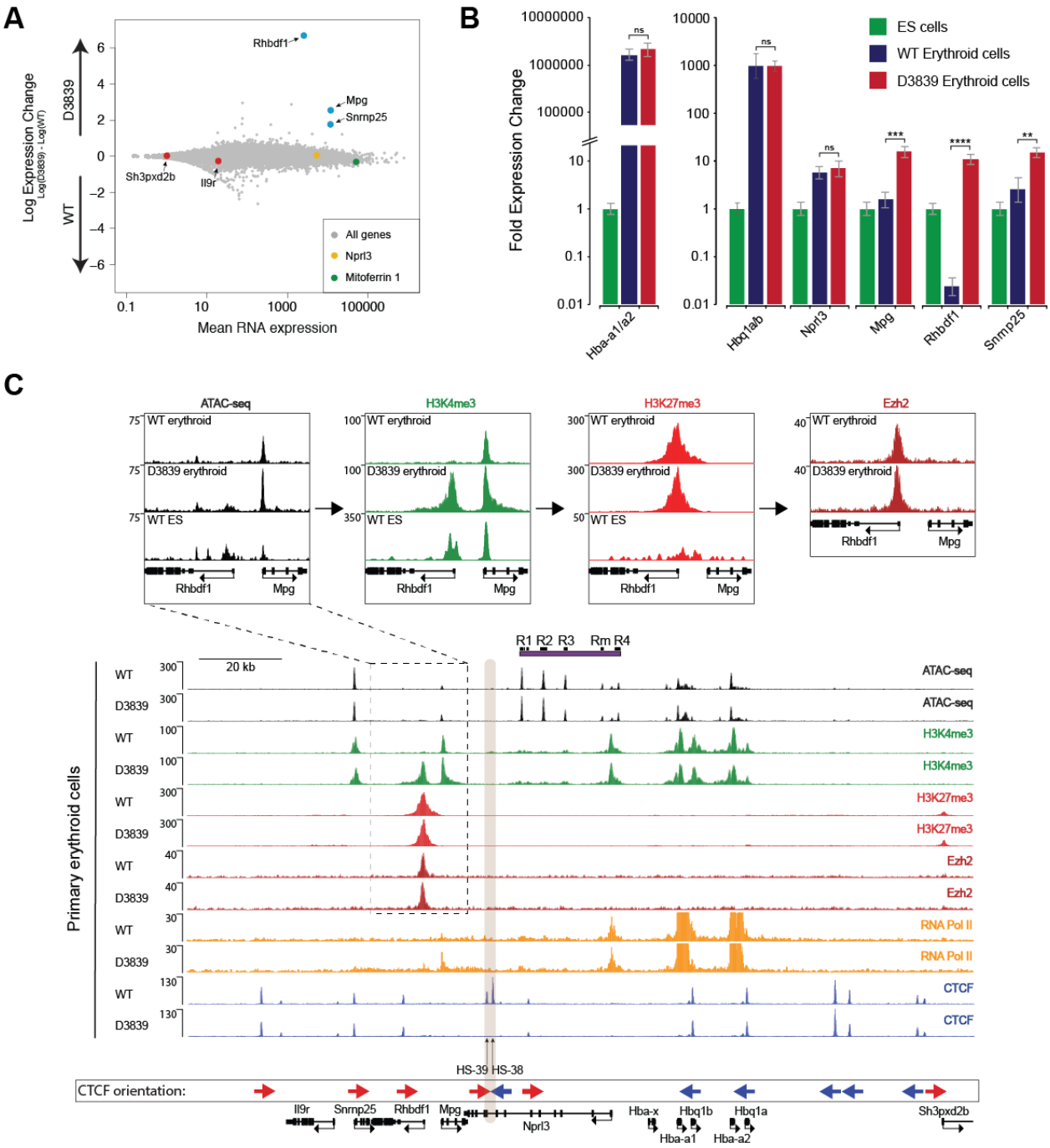


Figure 5.1 Effects of combined deletion of HS-38 and HS-39 on local gene expression and chromatin state. A. RNA-seq MA plot of WT versus D3839 erythroid cells. Data represent three biological replicate experiments. Mean RNA abundance is plotted on the X axis and enrichment is plotted on the y-axis. Mitoferrin 1 is a highly expressed erythroid control gene. *Snrnp25*: P = 1.46e-46, *Rhbdf1*: P<9.99e-99, *Mpg*: P=6.71e-64 **B**. Relative gene expression in WT and D3839 erythroid cells versus ES cells. Measured by real-time qPCR and representing 3 biological replicates. P-values are obtained via a student t-test. ns>0.05, *<0.05, **<0.01, ***<0.001, ****<0.0001. For significance of changes between ES and WT erythroid cells, see Fig. 3.6. **C**. Normalised (RPKM) ChIP-seq read-densities at the α-globin locus. Read density is the average of two biological replicates.

To determine the effect of the HS-38/39 deletion on α-globin gene expression, I examined the haematological indices in wild type (WT), heterozygous (het) and homozygous (hom) mice (haematological phenotyping of mice was carried out by Jacqueline Sharpe, MHU). No significant differences in the main red cell indices that might have been affected by changes in alpha globin expression, including the levels of Hb, RBC, MCV, MCH or the proportion of reticulocytes were found. Together with the similar alpha/beta globin RNA ratios in these mice, these parameters suggest that there are no significant changes in globin expression in the peripheral blood of these mice (Supplementary table 2 and 3).

In combination, these results suggest that more promiscuous enhancer interactions in the absence of upstream CTCF sites result in a loss of enhancer selectivity.

5.2.2 The CTCF HS-38/39 boundary is required for the maintenance of epigenetically controlled transcriptional states

As a strong transcriptional upregulation of the Polycomb-repressed *Rhbdf1* gene was observed in D3839 erythroid cells, I next investigated whether these expression changes were accompanied by changes in chromatin marks. In wild-type erythroid cells, the *Rhbdf1* gene is transcriptionally silent and is marked by high levels of H3K27me3 and an absence of active H3K4me3, consistent with its repressed transcriptional state (Fig 5.1C). As expected, upon gene activation in D3839 erythroid cells, high levels of H3K4me3 are deposited at the *Rhbdf1* promoter and the promoter-proximal part of the gene. Increased levels of H3K4me3 are also observed at the *Mpg* gene and, to a lesser extent, *Snrnp25* (Fig. 5.2A).

Transcriptional upregulation in D3839 erythroid cells is further correlated to the additional recruitment of RNA Polymerase II and increased chromatin accessibility by ATAC-seq at the promoters of *Mpg* and *Rhbdf1* (Fig 5.1C). Surprisingly, no significant decrease in the levels of H3K27me3 was detected at the *Rhbdf1* promoter region indicating that PRC2 activity at this gene is retained in D3839 cells (Fig. 5.1C and Fig. 5.2B). To verify that PRC2 is still present and to exclude the possibility that H3K27me3 was retained after the expulsion of PRC2 upon α-globin enhancer activation, the presence of the catalytic PRC2 component Ezh2 in D3839 erythroid cells was confirmed (Fig. 5.1C). These observations were also confirmed by real-time PCR (Fig 5.2C).

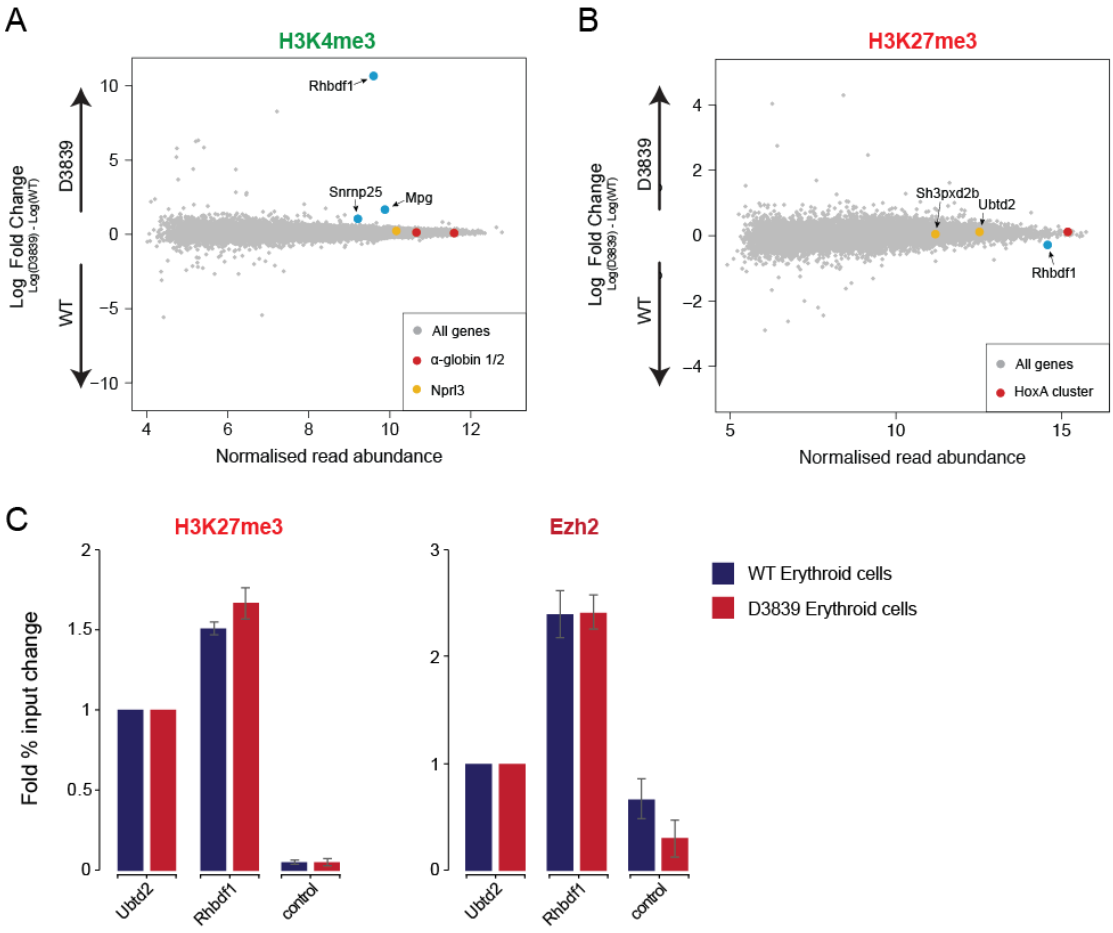


Figure 5.2 Effects of the combined deletion of HS-38 and HS-39 on histone modifications. A. H3K4me3 ChIP-seq MA plot of WT versus D3839 erythroid cells. Data represent two biological replicate experiments. Mean read abundance is plotted on the X axis and enrichment is plotted on the y-axis. *Snrnp25*: FDR<0.1, *Rhbdf1*: FDR<0.05, *Mpg*: FDR<0.05 **B**. H3K27me3 ChIP-seq MA plot of WT versus D3839 erythroid cells. Data represent two biological replicate experiments. Mean read abundance is plotted on the X axis and enrichment is plotted on the y-axis. *Ubtd2* and *Sh3pxd2b* are Polycomb repressed genes directly downstream of the α-globin cluster. *Rhbdf1*: FDR=0.19. **C**. ChIP-qPCR for H3K27me3 and Ezh2 in wild-type and D3839 erythroid cells. The control is an amplicon within the *Nprl3* gene. *Ubtd2* is a polycomb repressed gene within 200kb downstream of the α-globin cluster and is used as a positive control. No significant changes are detected by student t-test on biological triplicates.

The combined presence of H3K4me3 and H3K27me3 at bivalent promoters is thought to dictate low levels of gene expression (Voigt, Tee and Reinberg 2013). I investigated how levels of *Rhbdf1* transcription in D3839 erythroid cells compared to those in ES cells where *Rhbdf1* is actively transcribed in the absence of Polycomb repression (see **Chapter 3** and Fig. 5.1C). Surprisingly, D3839 erythroid cells express *Rhbdf1* at a level 10-fold that observed in ES cells, showing that the gene is robustly activated in these cells (Fig 5.1B). Similar to ES cells, chromatin at the *Rhbdf1* promoter is accessible in D3839 cells, whereas this accessible chromatin region is not observed in wild-type erythroid cells (Fig. 5.1C). Thus, the removal of the HS-38/39 CTCF boundary results in loss of transcriptional repression of the neighbouring Polycomb-occupied *Rhbdf1* gene.

5.2.3 Individual HS-38 and HS-39 CTCF binding sites retain partial boundary capacity

Plainly, the removal of both HS-38 and HS-39 sites causes a change in the functional interactions between the α-globin enhancers and other genes within the TAD. It has been suggested that divergent pairs of CTCF form effective boundaries because of their ability to establish interactions in both orientations (Gómez-Marín et al. 2015). However, this notion has not been addressed functionally. I analysed the expression of the genes that were affected by the loss of both HS-38 and HS-39 in the D38 and D39 single site mutants (Fig. 5.3A). As expected, neither α-globin nor *Nprl3* expression are affected by the removal of these sites. In contrast, several significant changes in the expression of the upstream *Mpg*, *Rhbdf1*, and *Snrnp25* are detected. The deletion of HS-38 has stronger effects on the transcription of these

genes than the removal of HS-39, albeit not as strong as the removal of both sites; the *Rhbdf1* gene is upregulated 20-fold in the D38 mutant, whereas a ~500-fold increase in expression is observed in the D3839 mutant.

Thus, although the deletion of HS-39 (D39) does not have a strong effect on the expression of the three upstream genes, its single presence (in D38 mice) is sufficient to retain significant boundary activity compared to the double D3839 mutant. Conversely, the single presence of HS-38 (in D39 mice) is sufficient to retain the majority of the boundary element's function. These observations and the fact that only HS-38 is conserved in human erythroid cells suggest that the presence of HS-38 is sufficient to provide adequate enhancer insulation.

Changes in transcription in single CTCF site mutant cells are again reflected in the levels of H3K4me3 at the gene promoters; increases in H3K4me3 are seen at the *Rhbdf1* and *Mpg* promoters in the D38 and D39 mutants (Fig 5.3B), although levels were much lower than those observed in the D3839 mice. No strong changes in chromatin accessibility are observed in the single D38 and D39 mutants, consistent with the smaller increases in gene expression compared to the D3839 mutant where these changes are already subtle (Fig. 5.3B and Fig. 5.1C).

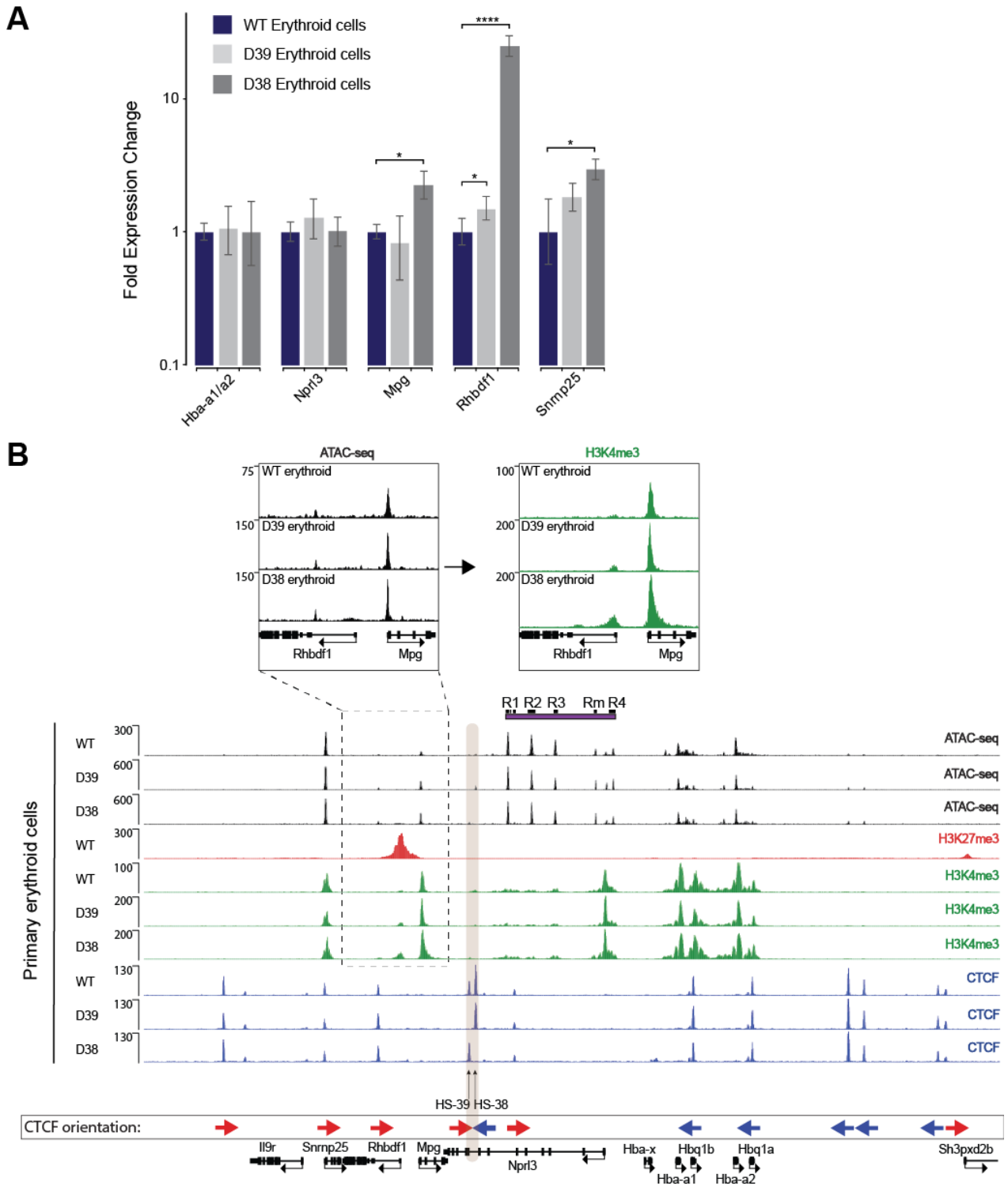


Figure 5.3 Effects of the individual deletion of HS-38 or HS-39 on local gene expression and chromatin state. A. Relative expression of WT and D3839 erythroid cells versus ES cells. Measured by real-time qPCR and representing 3 biological replicates. P-values are obtained via a student t-test. *<0.05, ****<0.0001. **B**. Normalised (RPKM) ChIP-seq read-densities at the α-globin locus. Read density is the average of two biological replicates.

5.2.4 Single deletion of HS-29 has minor effects on local gene regulation

Although the analysis of chromatin interactions in erythroid cells of D29 mice did not identify any gross changes in local genome architecture, I investigated whether the deletion of HS-29 affected the expression of local genes through some other mechanism. A comparison of RNA-seq data between wild-type and D29 erythroid cells identified no significant changes in the expression of local genes (Fig. 5.4A). However, a modest, but significant, increase in *Rhbdf1* is detected by real-time PCR, suggesting that HS-29 may also have a small role in contributing to the directionality of the enhancer (Fig. 5.4B).

The removal of D29 resulted in intensified interactions between the R1/R2 enhancers and the globin genes (see **Chapter 3**). To test whether the increase in these enhancer-promoter contacts resulted in the upregulation of the globin genes, α- and θ- globin gene expression was compared between D29 and wild-type erythroid cells by real-time PCR (Fig. 5.4B). However, no significant increase in the expression of either gene is detected.

Finally, I analysed chromatin accessibility and deposition of H3K4me3 in the D29 mice (Fig. 5.4C). In agreement with the absence of any changes in local gene expression, no difference in chromatin state is observed within the α-globin gene cluster.

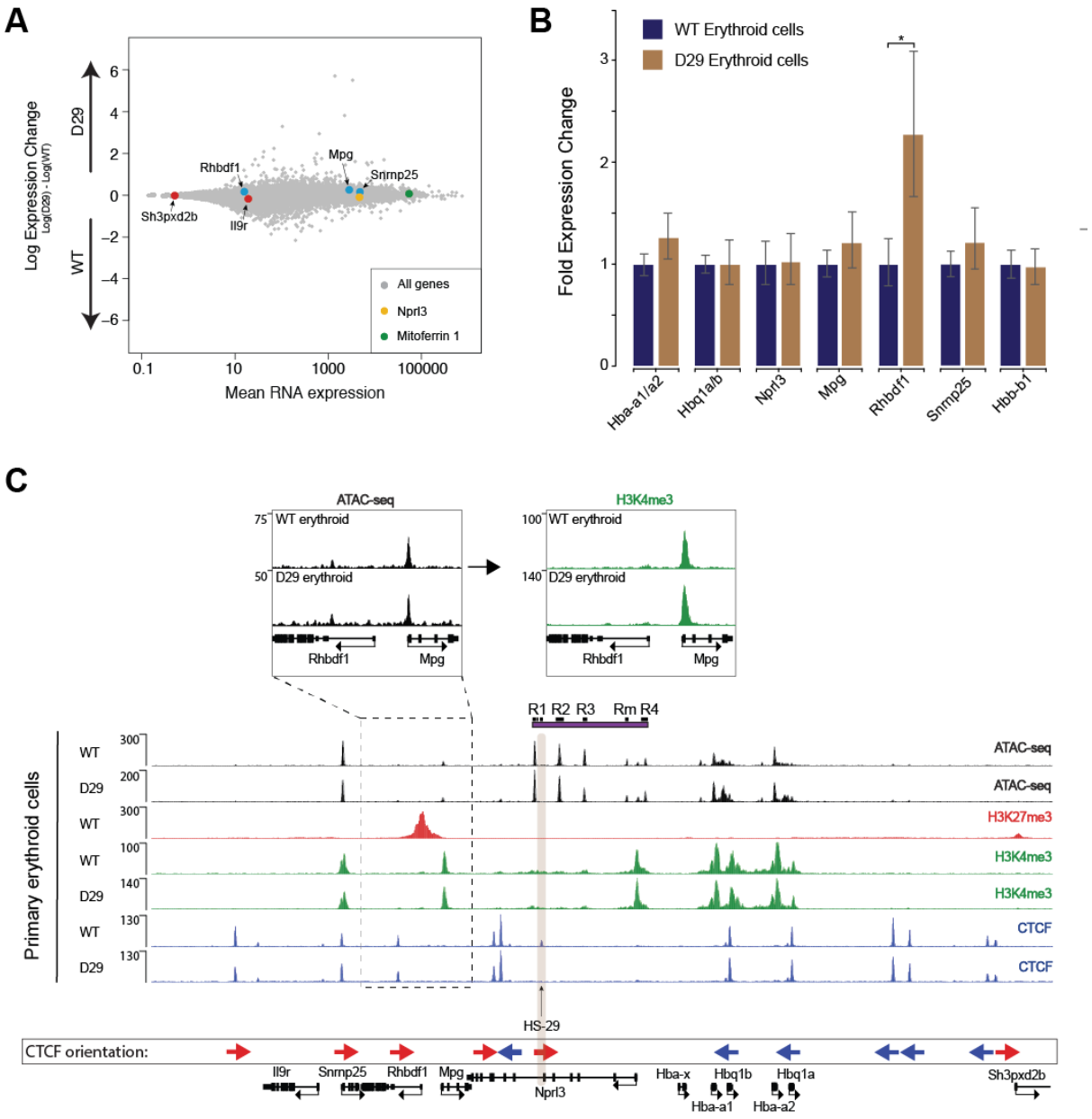


Figure 5.4 Effects of the deletion of HS-29 on local gene expression and chromatin state. A. RNA-seq MA plot of WT versus D29 erythroid cells. Data represent three biological replicate experiments. Mean RNA abundance is plotted on the X axis and enrichment is plotted on the y-axis. Mitoferrin 1 is a highly expressed erythroid control gene. No significant changes in local gene expression were detected **B**. Relative expression of WT vs D29 erythroid cells. Measured by real-time qPCR and representing three biological replicates. P-values are obtained via a student t-test. *<0.05. **C**. Normalised (RPKM) ChIP-seq read-densities at the α-globin locus. Read density is the average of two biological replicates.

5.2 Discussion and Future Work

The results in this chapter indicate that, in wild-type mice, the activity of the α-globin enhancers is constrained and directed by CTCF to result in the specific upregulation of the α-globin genes. In the D3839 mutant, full removal of a CTCF boundary upstream of the α-globin enhancers results in the upregulation of the upstream *Mpg*, *Rhbdf1*, and *Mpg* genes, demonstrating a loss of enhancer specificity. These transcriptional changes are reflected in the chromatin state as increased levels of the active H3K4me3 are present at the promoters of the three upregulated genes.

Contrary to previous studies making similar observations (Dowen *et al.* 2014; Lupiáñez *et al.* 2015), a mechanistic explanation for these observations is provided by the changes in local topology caused by CTCF binding site deletions (see **Chapter 4**). The abrogation of CTCF binding in the D3839 mutant results in a loss of the looping interaction between the downstream HS44/48 and the deleted HS-38/-39 sites. *Mpg* and *Rhbdf1,* the two genes located closest to the deleted sites, are now unconstrained by this CTCF-mediated interaction. This results in an expansion of the α-globin compartment and increased interactions with the R1 and R2 enhancers. *The Snrnp25* gene promoter is flanked by a CTCF binding site (HS-59) which marks the upstream end of these expanded enhancer interactions and of the observed changes in gene expression.

These findings are consistent with a model in which CTCF binding within TADs regulates gene expression by creating a local chromosome architecture that directs enhancer interactions to act specifically on their target genes. Furthermore, it is

unlikely that CTCF serves as an insulator by directly preventing the spread of active or repressive chromatin at the α-globin locus, as no CTCF sites are present between the adjacent active (H3K4me3) promoter of the housekeeping gene *Mpg* and inactive Polycomb-repressed (H3K27me3) *Rhbdf1* promoter. Thus, the management of enhancer interactions is likely one of the key functions of topological chromatin organisation within TADs.

Surprisingly, the strong upregulation of the Polycomb-repressed gene *Rhbdf1* in the D3839 mutant does not result in a loss of H3K27me3 or PRC2 component Ezh2. This suggests that Polycomb repression, while not lost, is overcome by the strong gene-activating influence of the α-globin enhancers upon increased interaction. While the existence of bivalency of mammalian promoters is now well-established, this chromatin state is generally associated with gene repression or low levels of gene expression (Voigt, Tee and Reinberg 2013). However, levels of *Rhbdf1* in D3839 erythroid cells exceed those in ES cells where the gene is active. Although it is possible that *Rhbdf1* activation is limited to a small subpopulation of cells where Polycomb repression is lost, the observed large effect on gene expression and H3K4me3 acquisition suggest that *Rhbdf1* becomes bivalent in D3839 erythroid cells. To confirm that *Rhbdf1* is upregulated under the influence of the α-globin enhancers in every cell, I attempted to use the Fluidigm Biomark HD™ to measure single-cell expression of genes within the α-globin cluster. However, this analysis was not conclusive due to the detection limit of this assay.

While the single deletions of HS-38 or HS-39 result in a partial loss of boundary activity, significant insulation from the enhancers is retained in the absence of either

CTCF site. These data suggest that CTCF insulation provided by multiple binding sites is not additive; the gene expression changes upon removal of either single CTCF site are smaller than half the effect size observed in the D3839 mutant. However, these experiments don't address the question of whether insulator strength is dependent on the number of CTCF binding sites, the divergent orientation of those sites, or a combination of both (Gómez-Marín et al. 2015). This question could be addressed at the α-globin cluster by creating mice with a single inversion of HS-38 or HS-39. If the ability to interact with other CTCF binding sites in both orientations is a major determinant of boundary activity, the inversion of HS-38 or HS-39 should result in similar gene expression changes as observed in the D38 and D39 mutants upon CTCF binding site deletion. Finally, HS-38 is conserved in human and its binding sequence is highly similar to the CTCF consensus binding sequence. The observation that conserved CTCF sites are enriched at domain borders and strongly insulated loci across 4 different mammalian genomes corresponds with the observation that HS-38 has the strongest boundary activity of the two sites (Rudan et al. 2015).

The limited effect of HS-29 deletion on gene expression is consistent with the minor changes to local topology observed in this mutant (see **Chapter 4**). In the light of the observed increase in α-globin enhancer-promoter interactions in the D29 mutant, it is intriguing that α-globin gene expression shows a non-significant increase in these mice, whereas β-globin does not. As erythroid cells contain a large pool of accumulated steady-state α-globin mRNA, small changes in transcriptional output may be difficult to detect (Hay et al. 2016). Thus, the analysis of labelled nascent

RNA via the Nanostring nCounter gene expression system may be able to resolve this potential increased expression (Geiss et al. 2008).

To further increase our understanding of how CTCF binding regulates local chromatin topology and, ultimately, gene expression, I have created mice with deletions in CTCF binding sites downstream of the α-globin locus. Similar to the mice analysed in this thesis, mice lacking the θ2, HS44, and a combination of HS44 and HS48 were created but could not be analysed due to time constrains. As current Capture-C interaction data from upstream sites suggest that HS44 and HS48 are a key anchor point for interactions between up- and downstream domains, the analysis of HS44/48 mutant mice may shed further light on the link between genome topology and gene regulation. Finally, ES cells containing ectopic insertions of CTCF in different orientations in between the α-globin enhancers and genes were also constructed with this aim. For an overview of this work, see appendix 2**.**

In conclusion, these findings show that CTCF-mediated organisation of chromatin architecture within a sub-TAD structure is required to ensure the promoter specificity and directionality of enhancers. The maintenance of this organisation is required in order to topologically shield housekeeping and repressed genes from the influence of tissue-specific enhancers, as was suggested in a recent correlative, bioinformatic study (Oti et al. 2016). The mechanism described here shows how variation in CTCF binding due to mutation or methylation of the sequence motif can lead to variation in gene expression and disease (Flavahan *et al.* 2015; Hnisz *et al.* 2016).

Chapter 6: General discussion and conclusions

This thesis explores the regulation of gene expression in the context of local chromatin topology by the *in vivo* dissection of enhancer-proximal CTCF and cohesin binding at the α-globin locus. By analysing interactions of *cis*-regulatory elements between different cell-types and in CTCF binding site mutants, the results in this thesis have revealed an important role of CTCF in constraining the interactions of tissue-specific enhancer elements. Interactions between flanking clusters of CTCF binding sites in a convergent orientation are strongly induced in erythroid cells, where the α-globin enhancers are active, and delimit interactions of the α-globin genes and regulatory elements. CTCF binding is responsible for shaping this topology, as deletion of two CTCF sites (D3839) flanking the α-globin enhancers results in a local loss of contacts with the CTCF cluster flanking the α-globin genes. The concomitant increase of interactions between the enhancers and flanking genes demonstrates that CTCF normally constrains these interactions. This management of enhancer contacts links local chromosome architecture to gene expression, as the expanded area of enhancer contacts in mice lacking the enhancer-flanking CTCF sites corresponds to the position of three strongly upregulated genes. Importantly, the strongly activated *Rhbdf1* gene is occupied by the repressive PRC2 complex, illustrating that insulation from enhancer activity is required for the maintenance of epigenetic repression. The deletion of the constituents of the CTCF boundary (HS-38 and HS-39) revealed that individual CTCF binding sites are able to retain partial boundary activity. When only the conserved HS-38 site is retained upon HS-39 deletion, changes in the transcription of upstream genes are barely detectable.

Similarly, the deletion of a less prominent CTCF site in between the R1 and R2 enhancers resulted in a minor increase of *Rhbdf1* expression. A 2D representation of the changes in interaction observed upon loss of HS-38 and HS-39 is shown in Figure 6.1.

These findings provide a mechanism for CTCF-mediated insulation of enhancer activity by linking the interactions of architectural elements and enhancers to local gene expression. This work expands on our understanding of CTCF and enhancer biology in several ways, which have been outlined below.

6.1 CTCF constrains enhancer activity by topologically shielding flanking genes

The notion that CTCF is able to block interactions between gene enhancers and promoters stems from transgenic assays in which insulators are placed in between enhancers and the gene they regulate (Chung, Bell and Felsenfeld 1997; Bell, West and Felsenfeld 1999; Bell and Felsenfeld 2000). However, evidence for this role in its natural genomic context has only recently emerged. In a study of super-enhancers and their target genes in ES cells, it was shown that removal of flanking CTCF and cohesin bound sites resulted in the upregulation of neighbouring genes. Although the authors speculated that this might be due to increased enhancer contacts with these genes, this was not shown (Dowen et al. 2014). The deletion of a genomic region containing a CTCF-bound TAD boundary was shown to result in ectopic enhancer interactions and misregulation of flanking genes. However, the role of CTCF was not explicitly studied, allowing for the possibility that other factors involved in the establishment of TAD boundaries are responsible for the insulation of enhancer activity (Lupiáñez et al. 2015).

The study of the interactions of regulatory elements within and flanking the α-globin compartment reveals a chromatin structure in which enhancer interactions are unidirectionally orientated towards the target α-globin genes. I show that the deletion of two CTCF sites flanking the enhancers results in a partial loss of the directionality in enhancer interactions, resulting in increased interactions with the upstream neighbouring genes and concomitant upregulation of expression. Moreover, this domain of expanded enhancer interactions corresponds to the genomic region that

loses interactions with CTCF sites downstream of the α-globin genes. These data provide strong evidence for a role of CTCF in ensuring enhancer-promoter specificity by topologically constraining enhancer interactions. Flanking housekeeping genes are similarly constrained by the interactions of CTCF clusters interspersed between these genes. Thus, CTCF directs α-globin enhancer activity by orchestrating local chromatin architecture.

These data are consistent with the enhancer-blocking function of CTCF originally observed in transgenic assays and provide mechanistic insight into how this function is utilised in a natural chromatin environment *in vivo*. In addition, they expand on observations that tethered interactions between the erythroid β-globin enhancers and the developmentally silenced γ-globin promoter result in activation of the gene (Deng et al. 2012). While in agreement with the observation that enhancer-promoter contacts are directly linked to the expression state of the gene, the results presented here suggest that this is not limited to enhancers and promoters that are active in similar lineages, as interactions with the erythroid α-globin enhancer result in the activation of three adjacent non-erythroid genes. Interestingly, the deletion of HS-38 and HS-39 does not result in decreased interactions between the α-globin genes and enhancers, despite an expansion of enhancer interactions with the upstream region. Correspondingly, no difference in expression of α-globin or θ-globin is observed in the D3839 mutant. This suggests that the function of these CTCF sites is not to tether the enhancers to their target promoter, as has been observed at other loci (Liu *et al.* 2011; Monahan *et al.* 2012; Golan-Mashiach *et al.* 2012; Guo *et al.* 2012b), but purely to ensure their specificity as described above.

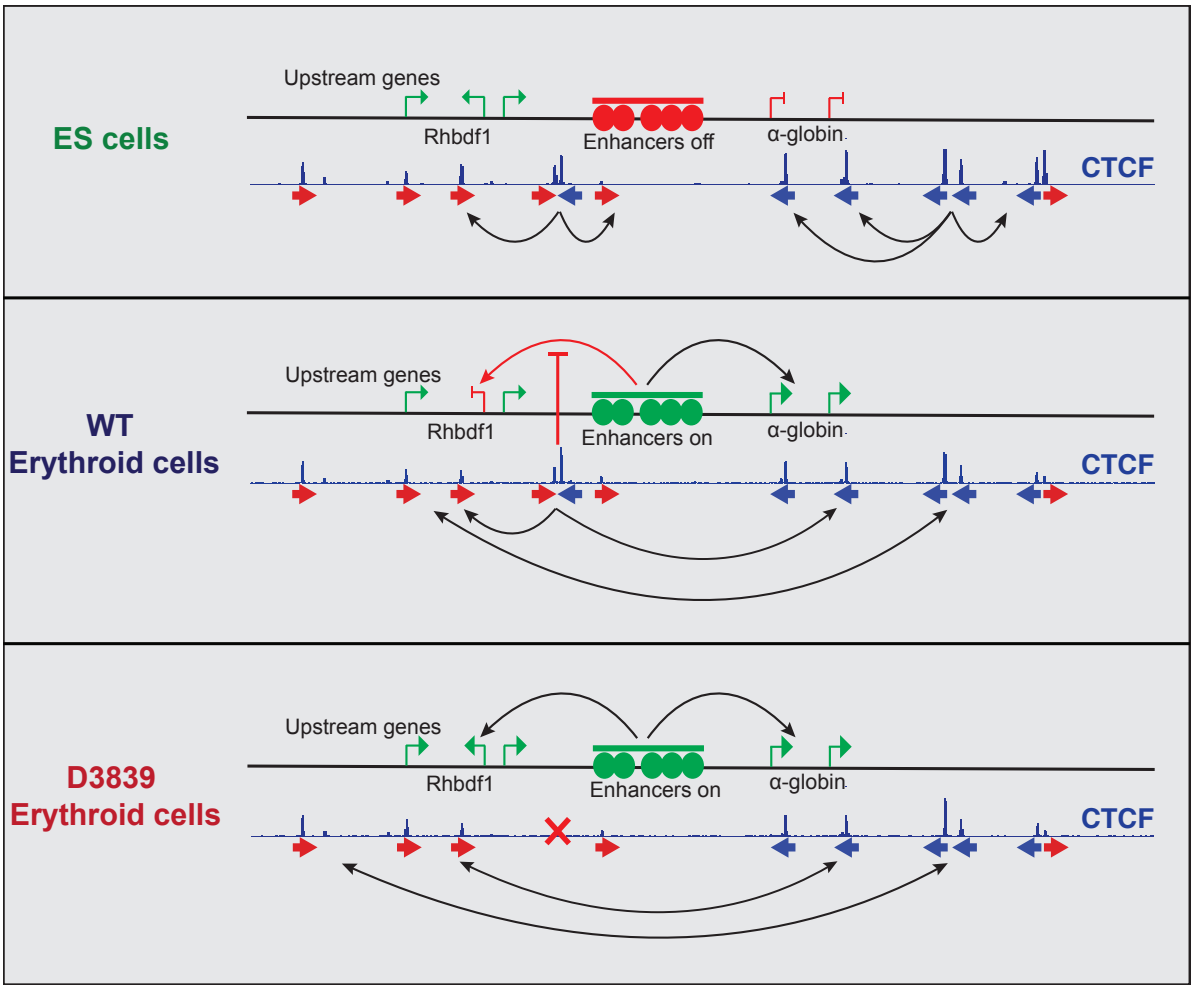


Figure 6.1 CTCF boundary activity prevents enhancer promiscuity in erythroid cells. In ES cells, the inactive α-globin enhancer and CTCF binding sites predominantly interact with chromatin in close proximity. Upon enhancer activation in erythroid cells, longer range interactions are established between the α-globin promoters and enhancers. Interactions between clusters of flanking CTCF sites prevent contacts between the α-globin enhancers and upstream genes from forming. Upon deletion of HS-38 and HS-39 CTCF binding sites (D3839), CTCF interactions shift away from this boundary towards more distally located upstream sites, allowing bidirectional α-globin enhancer interactions and upregulation of upstream genes.

However, it is possible that contacts between remaining CTCF binding sites are sufficient to maintain a favourable local chromatin topology for the formation of enhancer-promoter interactions. This notion is supported by the remaining interactions between HS44 and HS48 with HS-59 and other upstream sites. A comprehensive removal of all CTCF binding sites in the α-globin sub-TAD is required to exclude the possibility that CTCF binding is required for the efficient establishment of enhancer-promoter interactions at the α-globin cluster.

6.2 Topological shielding from the enhancers is required for the maintenance of epigenetic repression

Polycomb repression of genes in development is thought to be brought about via the establishment of a facultative repressed chromatin state, inhibiting transcription until Polycomb is evicted upon gene activation. It is currently unclear what molecular mechanisms trigger Polycomb eviction, but two previous studies have suggested that the loss of insulation from active chromatin marks or enhancer activity may result in a partial loss of polycomb repression. Deletion of CTCF sites on the boundary of active and repressed domains of genes at the *HoxA* cluster resulted in the upregulation of previously repressed *Hox* genes combined with a ~50% reduction in levels of the H3K27me3 repressive mark. However, it was not clear from these experiments whether this activation was brought about by a spread of active chromatin state or enhancer activity (Narendra et al. 2015). A second study suggested the latter may be true, as the removal of a CTCF binding site flanking a super-enhancer in ES cells resulted in the mild (~2.5 fold) upregulation of the neighbouring Polycomb-repressed *Tcfap2e* gene. This observation was not further investigated, making it unclear whether this upregulation was paired to increased enhancer interactions and loss of Polycomb (Dowen et al. 2014).

The data presented in this thesis expands on these observations to show that insulation from enhancer activity by CTCF is required for the maintenance of Polycomb mediated gene repression. In the absence of CTCF, *Rhbdf1* interacts with the α-globin enhancers and is concomitantly subject to strong transcriptional activation, suggesting that increased enhancer-interactions drive the transcriptional

upregulation. While experiments presented here do not formally exclude the possibility that activation occurs via the spread of an active chromatin state, the presence of the active, H3K4me3-marked *Mpg* promoter adjacent to the H3K27me3 marked domain repressing *Rhbdf1* suggests that this mechanism is unlikely to explain these observations. Interestingly, in contrast with the observations made at the *HoxA* cluster (Narendra et al. 2015), the strong upregulation of *Rhbdf1* does not lead to a significant loss of H3K27me3 or PRC2 component Ezh2 despite the acquisition of high levels of H3K4me3.

Although the analysis of single-cell expression is required to confirm that acquisition of H3K4me3 does not occur within a subpopulation of D3839 erythroid cells in which H3K27me3 is lost, these results, if confirmed, have potential implications for the model of Polycomb-mediated gene repression. Opposing repressive Polycomb group (PcG) and activating Trithorax group (TrxG) activities act as antagonistic forces in gene regulation. In the prevailing “responsive model”, the sampling of unmethylated CpG islands by these protein complexes is thought to result in low levels of PcG deposited H3K27me3 and TrxG deposited H3K4me3 at these sites. The absence of features of active transcription, such as RNA polymerase II or activating transcription factors, would result in the establishment of a stable Polycomb chromatin domain to prevent low-level stochastic reactivation of the repressed gene. By contrast, the presence of active transcription would decrease the residency time or catalytic activity of Polycomb complexes (Klose *et al.* 2013; Blackledge *et al.* 2015). Such a chromatin-encoded bi-stable switch between a Polycomb repressed and active transcriptional state is not consistent with the strong transcriptional activation in the presence of H3K27me3 and Ezh2 that is observed at

the *Rhbdf1* gene in the D3839 mutant. In these cells, the acquisition of high levels of H3K4me3 and RNA polymerase at the *Rhbdf1* promoter does not result in a significant decrease in H3K27me3 or Ezh2. In the described responsive model in which Polycomb acts to stabilise gene repression in the absence of gene transcription, the recruitment of active H3K4me3 and RNA polymerase II in combination with with the robust transcriptional upregulation of the gene should destabilise Polycomb repression. Thus, observations made at the *Rhbdf1* gene in the D3839 mutant are in support of the classic “instructive model”, in which Polycomb repression is targeted to defined promoters at which PcG-deposited histone modifications inhibit transcription via the establishment of a facultative heterochromatic state (Voigt, Tee and Reinberg 2013; Klose *et al.* 2013; Blackledge, Rose and Klose 2015). To test whether the continued presence of Polycomb acts as an brake on *Rhbdf1* transcription in D3839 cells, PRC2’s catalytic component Ezh2 may be pharmacologically inhibited in these cells.

6.3 Flanking CTCF clusters establish tissue-specific interactions

While genomic domain structures originally described as TADs, ranging between 1-5 Mb, were described to be largely invariant between different tissue types, more recent higher resolution studies have shown that smaller topological domains (described as sub-TADs or “contact domains”) differ considerably between cell-types (Jin et al. 2013; Phillips-Cremins et al. 2013; Rao et al. 2014). While the formation of cell-type specific sub-TADs has been linked to cell-type-specific gene expression patterns (Phillips-Cremins et al. 2013), it is not clear how these structures are formed. The comparison of the interactions between *cis*-regulatory elements at the α-globin locus between erythroid and ES cells presented in this thesis, reveals that both interactions between the α-globin enhancers and their target promoters as well as interactions between flanking CTCF clusters are strongly induced in erythroid cells. The edges of the interactions between these CTCF clusters correspond to the edges of the erythroid sub-TAD observed by Hi-C which contains the α-globin gene cluster (see Fig. 1.1), suggesting that the combination of enhancer and CTCF mediated interactions may be responsible for the establishment of this cell-type specific topological domain structure.

While the deletions of CTCF binding sites at the α-globin locus in mice presented here show that interactions between flanking domains are dependent on CTCF binding, it is not clear how these interactions are specifically induced in erythroid cells as binding of CTCF within the erythroid sub-TAD does not change between ES and erythroid cells. As cohesin has been suggested to be required for the formation of interactions between CTCF binding sites (Rudan et al. 2015), it is possible that

higher levels of cohesin are recruited to these CTCF sites upon enhancer activation and subsequently stabilise CTCF-mediated interactions. A current gap in the understanding of the molecular mechanisms by which cohesin is positioned at CTCF binding sites is the way the cohesin complex is loaded onto DNA, although this is thought to involve the cohesin loading factor NIPBL (Ocampo-Hafalla and Uhlmann 2011). The observation that NIPBL binding does not co-localise with cohesin found at CTCF binding sites suggests that cohesin is likely not loaded at these sites, but merely anchored there by CTCF. Instead, Nipbl was found to bind to a subset of enhancers and promoters that were bound by the mediator complex (Kagey et al. 2010). As the α-globin enhancers and promoters are bound by Mediator (Med1) and cohesin-component Rad21 specifically in erythroid cells, these findings provide the fascinating possibility that increased loading of the cohesin complex at the α-globin cluster in erythroid cells does not only mediate the formation of enhancer-promoter interactions, but also stabilises interactions between flanking CTCF clusters via a loop extrusion mechanism.

Alternatively, these CTCF-CTCF interactions may be stabilised by other unknown binding partners or post-translational modifications. Poly-(ADP-ribosyl)ation has been shown to influence CTCF's ability to form interactions and act as an insulator in *Drosophila* (Ong et al. 2013). Such modifications may in turn also regulate the ability of CTCF to partner with cohesin. The protein kinase Caseine Kinase 2 (CK2) phosphorylates residues close to the binding interface between CTCF and cohesin-component SA-1/2 (Xiao, Wallace and Felsenfeld 2011). The specific mutation of serine residues to completely prevent phosphorylation of CTCF has been shown to strengthen its ability to act as a transcriptional repressor at the *c-myc* locus, raising

the intriguing possibility that phosphorylation of CTCF is a regulatory mechanism for the interaction with cohesin (Klenova *et al.* 2001; El-Kady and Klenova 2005). The C-terminal domain of CTCF is further modified by SUMOylation which may alternatively be involved in modulating CTCF function. In contrast to phosphorylation, the SUMOylation of CTCF increased its ability to repress *c-myc*, suggesting these modifications may be employed to fine-tune CTCF function on chromatin (MacPherson et al. 2009). While these post-translational modifications may alter CTCF function by regulating its interaction with cohesin, it is also possible that they modulate CTCF function directly or by affecting its association with other binding partners.

Thus, induced interactions between CTCF clusters flanking the α-globin cluster define the topology of the α-globin cluster in erythroid cells and provide insight into the establishment of cell-type specific topological domains (sub-TADs).

6.4 Loop extrusion as a mechanism for the establishment of α-globin cluster chromatin topology

The model proposed for the topology of the α-globin cluster based on Capture-C data from viewpoints across the locus (see Fig. 6.1 for graphical representation) relies on the ability of CTCF to restrict α-globin enhancer and promoter interactions past CTCF-bound insulators. The recently proposed mechanism of loop-extrusion provides an explanation for how such insulation may occur.

In the context of the loop-extrusion mechanism, the formation of 3D interactions is essentially transformed into a 1D (linear) process that can consequently be controlled by insulator proteins. The linearity by which this mechanism operates has the ability to explain the preference for interactions between convergent CTCF binding sites, the ability of CTCF to reliably insulate *cis*-elements in close proximity, and the robust formation of interactions over long distances (~1 Mb), all of which would be unlikely to occur if interactions between *cis*-elements were established through spontaneous formation of 3D interactions (Sanborn *et al.* 2015; Fudenberg *et al.* 2016; Dekker and Mirny 2016). In the latter model, the establishment of a loop between CTCF binding sites flanking a gene and its enhancers would not necessarily prevent interactions between these elements and genes outside of this loop, as the enhancer could fold towards these genes in 3D nuclear space. However, if functional interactions between all *cis*-elements are indeed established via extrusion mechanisms, CTCF could prevent loop-extrusion, and thus interactions, past its binding sites (Nichols and Corces 2015; Dekker and Mirny 2016; Ghirlando and Felsenfeld 2016). While the cohesin complex is implicated in loop-extrusion and

stabilisation of interactions between CTCF sites, other molecules capable of tracking, such as RNA polymerase II, have been proposed to potentially play a role in establishing enhancer-promoter contacts via this mechanism (Dekker and Mirny 2016).

While the loop extrusion model is mostly consistent with my observations in this thesis, a number of apparent discrepancies highlight the gaps of knowledge that still exist. If cohesin binding is indeed stabilised specifically at CTCF binding sites, punctate “looping” interactions between CTCF binding sites would be expected to be observed by Capture-C. However, CTCF binding sites within genomic regions flanking the α-globin locus appear to interact with each other across a domain, being only slightly enriched at CTCF binding sites. As clear looping interactions are observed over longer distances in Hi-C data (Rao et al. 2014; Sanborn et al. 2015), this may simply reflect the resolution of 3C-based assays and the high density of CTCF binding sites within the α-globin cluster. However, the more defined interactions between the α-globin enhancers and promoters suggest that Capture-C is able to resolve interactions at the theoretical restriction fragment resolution. Perhaps CTCF binding sites merely act as “speed bumps”, temporarily slowing down the progression of loop-extrusion locally. The observed broad interactions across a genomic region could then be a reflection of the different positions of cohesin in the erythroid cell population.

Although my results agree with the notion that CTCF binding sites preferentially interact when in a convergent orientation towards each other, interactions between CTCF sites orientated in tandem are also observed over short genomic distances.

For example, HS44 and HS48 both interact with the CTCF binding sites at the θ-globin promoters that are positioned in the same orientation. Such tandem interactions have also been observed in CTCF ChIA-PET data, where ~33% of CTCF-CTCF interactions was found to be established between CTCF sites in tandem. These interactions were weaker on average than interactions between convergent CTCF binding sites and occurred over shorter distances (Tang *et al.* 2015; Guo *et al.* 2015). To explain these observations within the loop-extrusion model, one has to assume that CTCF can stabilise cohesin at lower efficiency in the reverse orientation. The processive extrusion mechanism would then be efficiently halted by two convergent CTCF binding sites, more rarely by two CTCF sites in tandem, and even less likely by two divergent CTCF sites (Ghirlando et al. 2012). However, in a static looping model, this does not explain why these interactions are observed to be weaker (Ghirlando et al. 2012). Again, a dynamic "speed bump" model for CTCF, in which the extruding cohesin complex is slowed down more when it passes CTCF sites orientated toward the complex, and less when orientated away from the complex, could explain the 3C-based observations on cell populations.

A final question that will have to be answered if the regulation of chromatin topology through loop extrusion is to be fully understood, is where the cohesin complex is loaded onto chromatin. The presence of cohesin-loading factor NIPBL at a subset of enhancers and promoters makes these *cis*-regulatory elements prime candidates for cohesin loading sites. However, not all promoters or enhancers are occupied by NIPBL and cohesin, suggesting cohesin may be loaded at other genomic sites as well. While a minimal loop-extrusion model is able to explain Hi-C contact maps without information of cohesin loading or CTCF binding (all that is given is the size of

loops framed), the specification of cohesin loading sites may further refine this model, allowing more accurate predictions of 3D genomic structure based on only *cis*-elements.

6.5 Variation in CTCF binding and variation in gene expression

The results presented in this thesis show that the disruption of intra-TAD CTCF binding sites can disrupt the regulation of proximal genes, overcoming mechanisms of epigenetic repression of transcription. Single nucleotide polymorphisms of the CTCF sequence motif can be sufficient to abrogate CTCF binding (Nakahashi et al. 2013) and the methylation state of the CTCF sequence motif has similarly been shown to regulate CTCF binding to DNA (Wang et al. 2012). My results show that abrogation of CTCF binding can have severe consequences for the regulation of transcription of local genes. Therefore, I predict that disruption of CTCF binding through naturally occurring genetic variation can result in variation in gene expression with potential consequences for disease susceptibility. Genome-wide association studies have identified that genetic susceptibility variants are mostly located outside of protein-coding regions (Hindorff *et al.* 2009; Ricaño-Ponce and Wijmenga 2013). A significant proportion of these variants may be located in CTCF binding sequences. Indeed, a recent study that focused on lung cancer susceptibility identified three single nucleotide polymorphisms (SNPs) below CTCF binding sites that were associated with increased risk of lung cancer (Dai et al. 2015). The analysis of CTCF binding between 51 lymphoblastoid cell lines revealed hundreds of differentially bound CTCF binding sites that could be linked to genetic variation, suggesting that genetic effects on CTCF are common within the human population and providing a database of SNPs that should be prioritised in studying the role of CTCF binding in disease risk (Ding et al. 2014).

Variation in CTCF binding has similarly been implicated in disease by several other studies of cancer cells. Acute lymphoblastic leukaemia cells were found to contain recurrent microdeletions that removed CTCF boundaries and resulted in the activation of proto-oncogenes. The removal of these boundaries in wild-type cells was sufficient for the transcriptional activation of proto-oncogenes at three of the investigated loci. Hepatocellular carcinoma and oesophageal adenocarcinoma cells were also found to contain recurrent somatic mutations of CTCF boundaries flanking proto-oncogenes, suggesting this is a common mechanism in oncogenesis (Hnisz et al. 2016). In a comprehensive study of whole-genome sequencing and ChIP-exo data of 213 colorectal cancer samples showed that point mutations frequently occurred at CTCF binding sites. The identification of CTCF-binding site mutations in publicly available datasets of several other cancers led the authors to propose that CTCF binding sites are major mutational hotspots in the noncoding genome of cancer cells (Katainen et al. 2015). Interestingly, point mutations close to the CTCF core sequence identified in both studies preferentially identified mutations in A·T base pairs. It is not clear what causes this mutation pattern, although it is of note that mutations occur outside of motif nucleotides within the strong DNaseI footprint, suggesting the high-affinity and stable binding of CTCF to its sequence motif may protect it from mutations.

Evidence also supports the notion that methylation of the CTCF sequence motif may be involved in disease. Cancer cells are often hyper-methylated, raising the possibility that some of these methylation events may disrupt CTCF binding (Jones and Baylin 2007). Indeed, a comparison of DNA methylation and CTCF binding between 19 different cell lines revealed that CTCF relocated away from DNA

methylated sites in immortalised cells (Wang et al. 2012). The interplay between the regulation of CTCF binding and DNA methylation has been described as a mechanism for the repression of several oncogenes and tumour suppressor genes (Witcher and Emerson 2009; Lai et al. 2010; Soto-Reyes and Recillas-Targa 2010). The most striking example of a functional study connecting loss of CTCF insulation by hyper-methylation to carcinogenesis is an investigation by Flavahan and colleagues, in which hyper-methylation caused by a gain-of-function mutation in *IDH* results in a global loss of CTCF binding at methylated sites in glioma cells. Loss of CTCF at the *PDGFRA* locus resulted in aberrant enhancer interactions resulting in upregulation of this glioma oncogene (Flavahan et al. 2015). Thus, both mutations and variation in methylation of the CTCF sequence motif have been shown to result in aberrant gene expression patterns. My results provide a detailed mechanism by which these changes in gene expression may occur and urge the careful consideration of CTCF as important functional elements in GWAS studies.

6.6 Conclusion

In summary, the results presented in this thesis reveal a mechanistic link between local CTCF-mediated genome architecture and the regulation of gene transcription via the restriction of enhancer contacts. These observations provide a significant contribution to the understanding of the interplay between genome topology, epigenetic regulation of gene expression, and *cis*-regulatory elements.

References

Allen BL, Taatjes DJ. The Mediator complex: a central integrator of transcription. *Nature Reviews Molecular Cell Biology* 2015;**16**:155–66.

Allis CD, Jenuwein T. The molecular hallmarks of epigenetic control. *Nat Rev Genet* 2016;**17**:487–500.

Anderson E, Devenney PS, Hill RE *et al.* Mapping the Shh long-range regulatory domain. *Development* 2014;**141**:3934–43.

Andrey G, Montavon T, Mascrez B *et al.* A switch between topological domains underlies HoxD genes collinearity in mouse limbs. *Science* 2013;**340**:1234167.

Anguita E, Hughes J, Heyworth C *et al.* Globin gene activation during haemopoiesis is driven by protein complexes nucleated by GATA-1 and GATA-2. *EMBO J* 2004;**23**:2841–52.

Anguita E. Deletion of the mouse alpha α-globin regulatory element (HS -26) has an unexpectedly mild phenotype. *Blood* 2002;**100**:3450–6.

Bailey TL, Williams N, Misleh C *et al.* MEME: discovering and analyzing DNA and protein sequence motifs. *Nucleic Acids Research* 2006;**34**:W369–73.

Bannister AJ, Kouzarides T. Regulation of chromatin by histone modifications. *Cell Res* 2011;**21**:381–95.

Bell AC, Felsenfeld G. Methylation of a CTCF-dependent boundary controls imprinted expression of the Igf2 gene. *Nature* 2000;**405**:482–5.

Bell AC, West AG, Felsenfeld G. The protein CTCF is required for the enhancer blocking activity of vertebrate insulators. *Cell* 1999;**98**:387–96.

Bernet A, Sabatier S, Picketts DJ *et al.* Targeted Inactivation of the Major Positive Regulatory Element (Hs-40) of the Human Alpha-Globin Gene Locus. *Blood* 1995;**86**:1202–11.

Bickmore WA, van Steensel B. Genome Architecture: Domain Organization of Interphase Chromosomes. *Cell* 2013;**152**:1270–84.

Blackledge NP, Rose NR, Klose RJ. Targeting Polycomb systems to regulate gene expression: modifications to a complex story. *Nature Reviews Molecular Cell Biology* 2015;**16**:643–9.

Blobel GA, Bodine D, Brand M *et al.* An international effort to cure a global health problem: A report on the 19th Hemoglobin Switching Conference. *Experimental Hematology* 2015;**43**:821–37.

Boettiger AN, Bintu B, Moffitt JR *et al.* Super-resolution imaging reveals distinct chromatin folding for different epigenetic states. *Nature* 2016;**529**:418–22.

Bowman SK, Deaton AM, Domingues H *et al.* H3K27 modifications define segmental regulatory domains in the Drosophila bithorax complex. *eLife* 2014;**3**:16847–13.

Boyle AP, Song L, Lee BK *et al.* High-resolution genome-wide in vivo footprinting of diverse transcription factors in human cells. *Genome Research* 2011;**21**:456–64.

Boyle S, Gilchrist S, Bridger JM *et al.* The spatial organization of human chromosomes within the nuclei of normal and emerin-mutant cells. *Human Molecular Genetics* 2001;**10**:211–9.

Buenrostro JD, Giresi PG, Zaba LC *et al.* Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat Meth* 2013;**10**:1213–8.

Cao A, Galanello R. Beta-thalassemia. *Genet Med* 2010;**12**:61–76.

Cermak T, Doyle EL, Christian M *et al.* Efficient design and assembly of custom TALEN and other TAL effector-based constructs for DNA targeting. *Nucleic Acids Research* 2011;**39**:e82–2.

Chambeyron S, Bickmore WA. Chromatin decondensation and nuclear reorganization of the HoxB locus upon induction of transcription. *Genes Dev.* 2004;**18**:1119–30.

Chien R, Zeng W, Kawauchi S *et al.* Cohesin Mediates Chromatin Interactions That Regulate Mammalian β-globin Expression. *Journal of Biological Chemistry* 2011;**286**:17870–8.

Christova Y, Adrain C, Bambrough P *et al.* Mammalian iRhoms have distinct physiological functions including an essential role in TACE regulation. *EMBO Rep* 2013;**14**:884–90.

Chung JH, Bell AC, Felsenfeld G. Characterization of the chicken beta-globin insulator. *Proc Natl Acad Sci USA* 1997;**94**:575–80.

Cong L, Ran FA, Cox D *et al.* Multiplex Genome Engineering Using CRISPR/Cas Systems. *Science* 2013;**339**:819–23.

Consortium TEP, Consortium TEP, data analysis coordination OC *et al.* An integrated encyclopedia of DNA elements in the human genome. *Nature* 2012;**488**:57–74.

Cremer T, Cremer C, Schneider T *et al.* Analysis of chromosome positions in the interphase nucleus of Chinese hamster cells by laser-UV-microirradiation experiments. *Hum Genet* 1982;**62**:201–9.

Cremer T, Cremer M. Chromosome Territories. *Cold Spring Harbor Perspectives in Biology* 2010;**2**:a003889–9.

Cuddapah S, Jothi R, Schones DE *et al.* Global analysis of the insulator binding protein CTCF in chromatin barrier regions reveals demarcation of active and repressive domains. *Genome Research* 2008;**19**:24–32.

Dai J, Zhu M, Wang C *et al.* Systematical analyses of variants in CTCF-binding sites identified a novel lung cancer susceptibility locus among Chinese population. *Sci Rep* 2015;**5**:7833–6.

Davies B, Davies G, Preece C *et al.* Site Specific Mutation of the Zic2 Locus by Microinjection of TALEN mRNA in Mouse CD1, C3H and C57BL/6J Oocytes. Schmidt EE (ed.). *PLoS ONE* 2013;**8**:e60216–7.

Davies JOJ, Telenius JM, McGowan SJ *et al.* Multiplexed analysis of chromosome

conformation at vastly improved sensitivity. *Nat Meth* 2016:**13**:74–80.

De Gobbi M, Anguita E, Hughes J *et al.* Tissue-specific histone modification and transcription factor binding in alpha globin gene expression. *Blood* 2007;**110**:4503–10.

de Wit E, Vos ESM, Holwerda SJB *et al.* CTCF Binding Polarity Determines Chromatin Looping. *Molecular Cell* 2015;**60**:676–84.

Defossez PA, Kelly KF, Filion G *et al.* The human enhancer blocker CTC-binding factor interacts with the transcription factor Kaiso. *J Biol Chem* 2005;**280**:43017–23.

Dekker J, Mirny L. The 3D Genome as Moderator of Chromosomal Communication. *Cell* 2016;**164**:1110–21.

Dekker J, Rippe K, Dekker M *et al.* Capturing chromosome conformation. *Science* 2002;**295**:1306–11.

Deng W, Lee J, Wang H *et al.* Controlling Long-Range Genomic Interactions at a Native Locus by Targeted Tethering of a Looping Factor. *Cell* 2012;**149**:1233–44.

Deng W, Rupon JW, Krivega I *et al.* Reactivation of Developmentally Silenced Globin Genes by Forced Chromatin Looping. *Cell* 2014;**158**:849–60.

Denker A, de Laat W. The second decade of 3C technologies: detailed insights into nuclear organization. *Genes Dev.* 2016;**30**:1357–82.

Ding Z, Ni Y, Timmer SW *et al.* Quantitative Genetics of CTCF Binding Reveal Local Sequence Effects and Different Modes of X-Chromosome Association. Gibson G (ed.). *PLoS Genetics* 2014;**10**:e1004798.

Dixon JR, Jung I, Selvaraj S *et al.* Chromatin architecture reorganization during stem cell differentiation. *Nature* 2015;**518**:331–6.

Dixon JR, Selvaraj S, Yue F *et al.* Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* 2012;**485**:376–80.

Dobin A, Davis CA, Schlesinger F *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 2013;**29**:15–21.

Docquier F, Farrar D, D'Arcy V *et al.* Heightened expression of CTCF in breast cancer cells is associated with resistance to apoptosis. *Cancer Res* 2005;**65**:5112–22.

Donohoe ME, Zhang L-F, Xu N *et al.* Identification of a Ctcf Cofactor, Yy1, for the X Chromosome Binary Switch. *Molecular Cell* 2007;**25**:43–56.

Dowen JM, Fan ZP, Hnisz D *et al.* Control of Cell Identity Genes Occurs in Insulated Neighborhoods in Mammalian Chromosomes. *Cell* 2014;**159**:374–87.

El-Kady A, Klenova E. Regulation of the transcription factor, CTCF, by phosphorylation with protein kinase CK2. *FEBS Letters* 2005;**579**:1424–34.

Engel N, West AG, Felsenfeld G *et al.* Antagonism between DNA hypermethylation and enhancer-blocking activity at the H19 DMD is uncovered by CpG mutations. *Nat Genet* 2004;**36**:883–8.

Engel N. CTCF binding sites promote transcription initiation and prevent DNA methylation on the maternal allele at the imprinted H19/Igf2 locus. *Human Molecular Genetics* 2006;**15**:2945–54.

Essafi A, Webb A, Berry RL *et al.* A Wt1-Controlled Chromatin Switching Mechanism Underpins Tissue-Specific Wnt4 Activation and Repression. *Developmental Cell* 2011;**21**:559–74.

Faure AJ, Schmidt D, Watt S *et al.* Cohesin regulates tissue-specific expression by stabilizing highly occupied cis-regulatory modules. *Genome Research* 2012;**22**:2163–75.

Ferrai C, de Castro IJ, Lavitas L *et al.* Gene Positioning. *Cold Spring Harbor Perspectives in Biology* 2010;**2**:a000588–8.

Ficz G. Polycomb group protein complexes exchange rapidly in living Drosophila. *Development* 2005;**132**:3963–76.

Filippova GN, Fagerlie S, Klenova EM *et al.* An exceptionally conserved transcriptional repressor, CTCF, employs different combinations of zinc fingers to bind diverged promoter sequences of avian and mammalian c-myc oncogenes. *Molecular and Cellular Biology* 1996;**16**:2802–13.

Flavahan WA, Drier Y, Liau BB *et al.* Insulator dysfunction and oncogene activation in IDH mutant gliomas. *Nature* 2015;**529**:74–80.

Fraser P, Bickmore W. Nuclear organization of the genome and the potential for gene regulation. *Nature* 2007;**447**:413–7.

Fu Y, Sinha M, Peterson CL *et al.* The Insulator Binding Protein CTCF Positions 20 Nucleosomes around Its Binding Sites across the Human Genome. van Steensel B (ed.). *PLoS Genetics* 2008;**4**:e1000138–13.

Fudenberg G, Imakaev M, Lu C *et al.* Formation of Chromosomal Domains by Loop Extrusion. *Cell Reports* 2016;**15**:2038–49.

Fullwood MJ, Liu MH, Pan YF *et al.* An oestrogen-receptor-α-bound human chromatin interactome. *Nature* 2009;**461**:58–64.

Galanello R, Cao A. Alpha-thalassemia. *Genet Med* 2011;**13**:83–8.

Garrick D, De Gobbi M, Samara V *et al.* The role of the polycomb complex in silencing alpha-globin gene expression in nonerythroid cells. *Blood* 2008;**112**:3889–99.

Gaszner M, Felsenfeld G. Insulators: exploiting transcriptional and epigenetic mechanisms. *Nat Rev Genet* 2006;**7**:703–13.

Geisler SJ, Paro R. Trithorax and Polycomb group-dependent regulation: a tale of opposing activities. *Development* 2015;**142**:2876–87.

Geiss GK, Bumgarner RE, Birditt B *et al.* Direct multiplexed measurement of gene expression with color-coded probe pairs. *Nature Biotechnology* 2008;**26**:317–25.

Geyer PK, Corces VG. DNA position-specific repression of transcription by a Drosophila zinc finger protein. *Genes Dev.* 1992;**6**:1865–73.

Geyer PK, Vitalini MW, Wallrath LL. Nuclear organization: taking a position on gene expression. *Current Opinion in Cell Biology* 2011;**23**:354–9.

Ghirlando R, Felsenfeld G. CTCF: making the right connections. *Genes Dev.* 2016;**30**:881–91.

Ghirlando R, Giles K, Gowher H *et al.* Chromatin domains, insulators, and the regulation of gene expression. *BBA - Gene Regulatory Mechanisms* 2012;**1819**:644–51.

Golan-Mashiach M, Grunspan M, Emmanuel R *et al.* Identification of CTCF as a master regulator of the clustered protocadherin genes. *Nucleic Acids Research* 2012;**40**:3378–91.

Gombert WM, Krumm A. Targeted Deletion of Multiple CTCF-Binding Elements in the Human C-MYC Gene Reveals a Requirement for CTCF in C-MYC Expression. Blagosklonny MV (ed.). *PLoS ONE* 2009;**4**:e6109.

Gomes NP, Espinosa JM. Gene-specific repression of the p53 target gene PUMA via intragenic CTCF-Cohesin binding. *Genes Dev.* 2010;**24**:1022–34.

Gómez-Marín C, Tena JJ, Acemel RD *et al.* Evolutionary comparison reveals that diverging CTCF sites are signatures of ancestral topological associating domains borders. *Proc Natl Acad Sci USA* 2015;**112**:7542–7.

Grimaud C, Bantignies F, Pal-Bhadra M *et al.* RNAi Components Are Required for Nuclear Clustering of Polycomb Group Response Elements. *Cell* 2006;**124**:957–71.

Guelen L, Pagie L, Brasset E *et al.* Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions. *Nature* 2008;**453**:948–51.

Guo C, Yoon HS, Franklin A *et al.* CTCF-binding elements mediate control of V(D)J recombination. *Nature* 2011;**477**:424–30.

Guo Y, Monahan K, Wu H *et al.* CTCF/cohesin-mediated DNA looping is required for protocadherin α promoter choice. *Proc Natl Acad Sci USA* 2012b;**109**:21081–6.

Guo Y, Xu Q, Canzio D *et al.* CRISPR Inversion of CTCF Sites Alters Genome Topology and Enhancer/Promoter Function. *Cell* 2015;**162**:900–10.

Hadjur S, Williams LM, Ryan NK *et al.* Cohesins form chromosomal cis-interactions at the developmentally regulated IFNG locus. *Nature* 2009;**460**:410–3.

Handoko L, Xu H, Li G *et al.* CTCF-mediated functional chromatin interactome in pluripotent cells. *Nat Genet* 2011;**43**:630–8.

Hark AT, Schoenherr CJ, Katz DJ *et al.* CTCF mediates methylation-sensitive enhancer-blocking activity at the H19/Igf2 locus. *Nature* 2000;**405**:486–9.

Hatton C, Wilkie A, Drysdale HC *et al.* Alpha-Thalassemia Caused by a Large (62 Kb) Deletion Upstream of the Human Alpha-Globin Gene-Cluster. *Blood* 1990;**76**:221–7.

Hay D, Hughes JR, Babbs C *et al.* Genetic dissection of the α-globin super-enhancer in vivo. *Nat Genet* 2016;**48**:895–903.

Hedglin M, O'Brien PJ. Human Alkyladenine DNA Glycosylase Employs a Processive Search for DNA Damage †. *Biochemistry* 2008;**47**:11434–45.

Heger P, Marin B, Bartkuhn M *et al.* The chromatin insulator CTCF and the emergence of metazoan diversity. *Proc Natl Acad Sci USA* 2012;**109**:17507–12.

Hewitt SL, High FA, Reiner SL *et al.* Nuclear repositioning marks the selective exclusion of lineage-inappropriate transcription factor loci during T helper cell differentiation. *Eur J Immunol* 2004;**34**:3604–13.

Higgs DR, Wood WG, Jarman AP *et al.* A Major Positive Regulatory Region Located Far Upstream of the Human Alpha-Globin Gene Locus. *Genes Dev.* 1990;**4**:1588–601.

Higgs DR, Wood WG. Long-range regulation of alpha globin gene expression during erythropoiesis. *Curr Opin Hematol* 2008;**15**:176–83.

Higgs DR. The Molecular Basis of alpha-Thalassemia. *Cold Spring Harbor Perspectives in Medicine* 2013;**3**:a011718–8.

Hindorff LA, Sethupathy P, Junkins HA *et al.* Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc Natl Acad Sci USA* 2009;**106**:9362–7.

Hnisz D, Weintraub AS, Day DS *et al.* Activation of proto-oncogenes by disruption of chromosome neighborhoods. *Science* 2016;**351**:1454–8.

Hosseini M, Goodstadt L, Hughes JR *et al.* Causes and Consequences of Chromatin Variation between Inbred Mice. de Villena FPM (ed.). *PLoS Genetics* 2013;**9**:e1003570.

Hsu PD, Lander ES, Zhang F. Development and Applications of CRISPR-Cas9 for Genome Engineering. *Cell* 2014;**157**:1262–78.

Hu B, Petela N, Kurze A *et al.* Biological chromodynamics: a general method for measuring protein occupancy across the genome by calibrating ChIP-seq. *Nucleic Acids Research* 2015;**43**:e132.

Huang S, Li X, Yusufzai TM *et al.* USF1 Recruits Histone Modification Complexes and Is Critical for Maintenance of a Chromatin Barrier. *Molecular and Cellular Biology* 2007;**27**:7991–8002.

Hughes JR, Cheng J-F, Ventress N *et al.* Annotation of cis-regulatory elements by identification, subclassification, and functional assessment of multispecies conserved sequences. *Proc Natl Acad Sci USA* 2005;**102**:9830–5.

Hughes JR, Roberts N, McGowan S *et al.* Analysis of hundreds of *cis*-regulatory landscapes at high resolution in a single, high-throughput experiment. *Nat Genetics* 2014;**46**:205–12.

Iborra FJ, Pombo A, Jackson DA *et al.* Active RNA polymerases are localized within discrete transcription "factories' in human nuclei. *Journal of Cell Science* 1996;**109 (Pt 6)**:1427–36.

Jin F, Li Y, Dixon JR *et al.* A high-resolution map of the three-dimensional chromatin interactome in human cells. *Nature* 2013;**503**:290–4.

Jones PA, Baylin SB. The Epigenomics of Cancer. *Cell* 2007;**128**:683–92.

Kagey MH, Newman JJ, Bilodeau S *et al.* Mediator and cohesin connect gene expression and chromatin architecture. *Nature* 2010;**467**:430–5.

Kassouf MT, Hughes JR, Taylor S *et al.* Genome-wide identification of TAL1's functional targets: Insights into its mechanisms of action in primary erythroid cells. *Genome Research* 2010;**20**:1064–83.

Katainen R, Dave K, Pitkänen E *et al.* CTCF/cohesin-binding sites are frequently mutated in cancer. *Nat Genet* 2015;**47**:818–21.

Kellum R, Schedl P. A Position-Effect Assay for Boundaries of Higher-Order Chromosomal Domains. *Cell* 1991;**64**:941–50.

Kent WJ, Sugnet CW, Furey TS *et al.* The human genome browser at UCSC. *Genome Research* 2002;**12**:996–1006.

Kim H, Kim J-S. A guide to genome engineering with programmable nucleases. *Nat Rev Genet* 2014;**15**:321–34.

Kim TH, Abdullaev ZK, Smith AD *et al.* Analysis of the Vertebrate Insulator Protein CTCF-Binding Sites in the Human Genome. *Cell* 2007;**128**:1231–45.

Kina T, Ikuta K, Takayama E *et al.* The monoclonal antibody TER-119 recognizes a molecule associated with glycophorin A and specifically marks the late stages of murine erythroid lineage. *Br J Haematol* 2000;**109**:280–7.

Klenova EM, Chernukhin IV, El-Kady A *et al.* Functional Phosphorylation Sites in the C-Terminal Region of the Multivalent Multifunctional Transcriptional Factor CTCF. *Molecular and Cellular Biology* 2001;**21**:2221–34.

Klenova EM, Nicolas RH, Paterson HF *et al.* Ctcf, a Conserved Nuclear Factor Required for Optimal Transcriptional Activity of the Chicken C-Myc Gene, Is an 11-Zn-Finger Protein Differentially Expressed in Multiple Forms. *Molecular and Cellular Biology* 1993;**13**:7612–24.

Klose RJ, Cooper S, Farcas AM *et al.* Chromatin sampling--an emerging perspective on targeting polycomb repressor proteins. Reik W (ed.). *PLoS Genetics* 2013;**9**:e1003717.

Knoops L, Renauld J-C. IL-9 and its Receptor: From Signal Transduction to Tumorigenesis. *Growth Factors* 2009;**22**:207–15.

Kosak ST, Skok JA, Medina KL *et al.* Subnuclear compartmentalization of immunoglobulin loci during lymphocyte development. *Science* 2002;**296**:158–62.

Kowalczyk MS, Hughes JR, Babbs C *et al.* Nprl3 is required for normal development of the cardiovascular system. *Mamm Genome* 2012a;**23**:404–15.

Kowalczyk MS, Hughes JR, Garrick D *et al.* Intragenic Enhancers Act as Alternative Promoters. *Molecular Cell* 2012b;**45**:447–58.

Kurukuti S, Tiwari VK, Tavoosidana G *et al.* 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 USA* 2006;**103**:10684–9.

Kuzmin I, Geil L, Gibson L *et al.* Transcriptional Regulator CTCF Controls Human Interleukin 1 Receptor-associated Kinase 2 Promoter. *J Mol Biol* 2005;**346**:411–22.

Lai AY, Fatemi M, Dhasarathy A *et al.* DNA methylation prevents CTCF-mediated silencing of the oncogene BCL6in B cell lymphomas. *J Exp Med* 2010;**207**:1939–50.

Laitem C, Zaborowska J, Tellier M *et al.* CTCF regulates NELF, DSIF and P-TEFb recruitment during transcription. *Transcription* 2015;**6**:79–90.

Lanctôt C, Cheutin T, Cremer M *et al.* Dynamic genome architecture in the nuclear space: regulation of gene expression in three dimensions. *Nat Rev Genet* 2007;**8**:104–15.

Langmead B, Trapnell C, Pop M *et al.* Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* 2009;**10**, DOI: 10.1186/gb-2009-10-3-r25.

Lee BK, Iyer VR. Genome-wide Studies of CCCTC-binding Factor (CTCF) and Cohesin Provide Insight into Chromatin Structure and Regulation. *Journal of Biological Chemistry* 2012;**287**:30906–13.

Li H, Handsaker B, Wysoker A *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 2009;**25**:2078–9.

Liao Y, Smyth GK, Shi W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* 2014;**30**:923–30.

Lieberman-Aiden E, van Berkum NL, Williams L *et al.* Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome. *Science* 2009;**326**:289–93.

Liu Z, Scannell DR, Eisen MB *et al.* Control of Embryonic Stem Cell Lineage Commitment by Core Promoter Factor, TAF3. *Cell* 2011;**146**:720–31.

Lobanenkov VV, Nicolas RH, Adler VV *et al.* A novel sequence-specific DNA binding protein which interacts with three regularly spaced direct repeats of the CCCTC-motif in the 5'-flanking sequence of the chicken c-myc gene. *Oncogene* 1990;**5**:1743–53.

Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* 2014;**15**:31–21.

Lundgren M, Chow CM, Sabbattini P *et al.* Transcription factor dosage affects changes in higher order chromatin structure associated with activation of a heterochromatic gene. *Cell* 2000;**103**:733–43.

Lupiáñez DG, Kraft K, Heinrich V *et al.* Disruptions of Topological Chromatin Domains Cause Pathogenic Rewiring of Gene-Enhancer Interactions. *Cell* 2015;**161**:1012–25.

Lutz M, Burke LJ, Barreto G *et al.* Transcriptional repression by the insulator protein CTCF involves histone deacetylases. *Nucleic Acids Research* 2000;**28**:1707–13.

Lutz M, Burke LJ, LeFevre P *et al.* Thyroid hormone-regulated enhancer blocking: cooperation of CTCF and thyroid hormone receptor. *EMBO J* 2003;**22**:1579–87.

Lynch MD, Smith AJH, De Gobbi M *et al.* An interspecies analysis reveals a key role for unmethylated CpG dinucleotides in vertebrate Polycomb complex recruitment. *EMBO J* 2011;**31**:317–29.

MacPherson MJ, Beatty LG, Zhou W *et al.* The CTCF Insulator Protein Is Posttranslationally Modified by SUMO. *Molecular and Cellular Biology* 2009;**29**:714–25.

Magoc T, Salzberg SL. FLASH: fast length adjustment of short reads to improve genome assemblies. *Bioinformatics* 2011;**27**:2957–63.

Mahy NL. Gene density and transcription influence the localization of chromatin outside of chromosome territories detectable by FISH. *The Journal of Cell Biology* 2002;**159**:753–63.

Majumder P, Boss JM. CTCF Controls Expression and Chromatin Architecture of the Human Major Histocompatibility Complex Class II Locus. *Molecular and Cellular Biology* 2010;**30**:4211–23.

Majumder P, Gomez JA, Chadwick BP *et al.* The insulator factor CTCF controls MHC class II gene expression and is required for the formation of long-distance chromatin interactions. *J Exp Med* 2008;**205**:785–98.

Margueron R, Reinberg D. The Polycomb complex PRC2 and its mark in life. *Nature* 2011;**469**:343–9.

Monahan K, Rudnick ND, Kehayova PD *et al.* Role of CCCTC binding factor (CTCF) and cohesin in the generation of single-cell diversity of protocadherin-α gene expression. *Proc Natl Acad Sci USA* 2012;**109**:9125–30.

Montavon T, Soshnikova N, Mascrez B *et al.* A Regulatory Archipelago Controls Hox Genes Transcription in Digits. *Cell* 2011;**147**:1132–45.

Moon H, Filippova G, Loukinov D *et al.* CTCF is conserved from Drosophila to humans and confers enhancer blocking of the Fab-8 insulator. *EMBO Rep* 2005;**6**:165–70.

Moore JM, Rabaia NA, Smith LE *et al.* Loss of Maternal CTCF Is Associated with Peri-Implantation Lethality of Ctcf Null Embryos. Wu Q (ed.). *PLoS ONE* 2012;**7**:e34915–10.

Nakahashi H, Kwon K-RK, Resch W *et al.* A Genome-wide Map of CTCF

Multivalency Redefines the CTCF Code. *Cell Reports* 2013;**3**:1678–89.

Narendra V, Rocha PP, An D *et al.* CTCF establishes discrete functional chromatin domains at the Hox clusters during differentiation. *Science* 2015;**347**:1017–21.

Nasmyth K, Haering CH. Cohesin: Its Roles and Mechanisms. *Annu Rev Genet* 2009;**43**:525–58.

Nasmyth K. Disseminating the genome: joining, resolving, and separating sister chromatids during mitosis and meiosis. *Annu Rev Genet* 2001;**35**:673–745.

Nativio R, Wendt KS, Ito Y *et al.* Cohesin Is Required for Higher-Order Chromatin Conformation at the Imprinted IGF2-H19 Locus. Bickmore WA (ed.). *PLoS Genetics* 2009;**5**:e1000739–15.

Nichols MH, Corces VG. A CTCF Code for 3D Genome Architecture. *Cell* 2015;**162**:703–5.

Nizami Z, Deryusheva S, Gall JG. The Cajal Body and Histone Locus Body. *Cold Spring Harbor Perspectives in Biology* 2010;**2**:a000653–3.

Noordermeer D, Branco MR, Splinter E *et al.* Transcription and Chromatin Organization of a Housekeeping Gene Cluster Containing an Integrated β-Globin Locus Control Region. Lee JT (ed.). *PLoS Genetics* 2008;**4**:e1000016–13.

Nora EP, Lajoie BR, Schulz EG *et al.* Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature* 2012;**485**:381–5.

Ocampo-Hafalla MT, Uhlmann F. Cohesin loading and sliding. *Journal of Cell Science* 2011;**124**:685–91.

Oki M, Kamakaka RT. Barrier Function at HMR. *Mol. Cell* 2005;**19**:707–16.

Ong C-T, Corces VG. CTCF: an architectural protein bridging genome topology and function. *Nat Rev Genet* 2014;**15**:234–46.

Ong C-T, Van Bortle K, Ramos E *et al.* Poly(ADP-ribosyl)ation Regulates Insulator Function and IntrachromosomalInteractions in Drosophila. *Cell* 2013;**155**:148–59.

Osborne CS, Chakalova L, Brown KE *et al.* Active genes dynamically colocalize to shared sites of ongoing transcription. *Nat Genet* 2004;**36**:1065–71.

Oti M, Falck J, Huynen MA *et al.* CTCF-mediated chromatin loops enclose inducible gene regulatory domains. *BMC Genomics* 2016;**17:**252.

Paredes SH, Melgar MF, Sethupathy P. Promoter-proximal CCCTC-factor binding is associated with an increase in the transcriptional pausing index. *Bioinformatics* 2013;**29**:1485–7.

Parelho V, Hadjur S, Spivakov M *et al.* Cohesins Functionally Associate with CTCF on Mammalian Chromosome Arms. *Cell* 2008;**132**:422–33.

Pederson T. The Nucleolus. *Cold Spring Harbor Perspectives in Biology* 2011;**3**:a000638–8.

Phillips-Cremins JE, Sauria MEG, Sanyal A *et al.* Architectural Protein Subclasses Shape 3D Organization of Genomes during Lineage Commitment. *Cell* 2013;**153**:1281–95.

Pott S, Lieb JD. What are super-enhancers? *Nat Genet* 2015;**47**:8–12.

Ptashne M. Gene regulation by proteins acting nearby and at a distance. *Nature* 1986;**322**:697–701.

Qian J, Wang Q, Dose M *et al.* B Cell Super-Enhancers and Regulatory Clusters Recruit AID Tumorigenic Activity. *Cell* 2014;**159**:1524–37.

Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 2010;**26**:841–2.

R Core Team. R: A Language and Environment for Statistical Computing. 2014:1–3501.

Ramirez F, Dundar F, Diehl S *et al.* deepTools: a flexible platform for exploring deep-sequencing data. *Nucleic Acids Research* 2014;**42**:W187–91.

Ran FA, Hsu PD, Lin C-Y *et al.* Double Nicking by RNA-Guided CRISPR Cas9 for Enhanced Genome Editing Specificity. *Cell* 2013a;**154**:1380–9.

Ran FA, Hsu PD, Wright J *et al.* Genome engineering using the CRISPR-Cas9 system. *Nat Protoc* 2013b;**8**:2281–308.

Rao SSP, Huntley MH, Durand NC *et al.* A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping. *Cell* 2014;**159**:1665–80.

Rhee HS, Pugh BF. Comprehensive Genome-wide Protein-DNA Interactions Detected at Single-Nucleotide Resolution. *Cell* 2011;**147**:1408–19.

Ribeiro de Almeida C, Stadhouders R, Thongjuea S *et al.* DNA-binding factor CTCF and long-range gene interactions in V(D)J recombination and oncogene activation. *Blood* 2012;**119**:6209–18.

Ricaño-Ponce I, Wijmenga C. Mapping of Immune-Mediated Disease Genes. *Annu Rev Genom Hum Genet* 2013;**14**:325–53.

Riddihough G, Zahn LM. What Is Epigenetics? *Science* 2010;**330**:611–1.

Ross-Innes CS, Brown GD, Carroll JS. A co-ordinated interaction between CTCF and ER in breast cancer cells. *BMC Genomics* 2011;**12**:593.

Rubio ED, Reiss DJ, Welcsh PL *et al.* CTCF physically links cohesin to chromatin. *Proc Natl Acad Sci USA* 2008;**105**:8309–14.

Rudan MV, Barrington C, Henderson S *et al.* Comparative Hi-C Reveals that CTCF Underlies Evolution of Chromosomal Domain Architecture. *Cell Reports* 2015;**10**:1297–309.

Sanborn AL, Rao SSP, Huang S-C *et al.* Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes. *Proc Natl Acad Sci USA* 2015:201518552.

Sander JD, Joung JK. CRISPR-Cas systems for editing, regulating and targeting genomes. *Nature Biotechnology* 2014;**32**:347–55.

Sanyal A, Lajoie BR, Jain G *et al.* The long-range interaction landscape of gene promoters. *Nature* 2012;**489**:109–13.

Schmidt D, Schwalie PC, Ross-Innes CS *et al.* A CTCF-independent role for cohesin in tissue-specific transcription. *Genome Research* 2010;**20**:578–88.

Schmidt D, Schwalie PC, Wilson MD *et al.* Waves of Retrotransposon Expansion Remodel Genome Organization and CTCF Binding in Multiple Mammalian Lineages. *Cell* 2012;**148**:335–48.

Schuettengruber B, Martinez A-M, Iovino N *et al.* Trithorax group proteins: switching genes on and keeping them active. *Nature Reviews Molecular Cell Biology* 2011;**12**:799–814.

Schwalie PC, Ward MC, Cain CE *et al.* Co-binding by YY1 identifies the transcriptionally active, highly conserved set of CTCF-bound regions in primate genomes. *Genome Biol* 2013;**14**:R148.

Seitan VC, Hao B, Tachibana-Konwalski K *et al.* A role for cohesin in T-cell-receptor rearrangement and thymocyte differentiation. *Nature* 2012;**476**:467–71.

Sequence Read Archive Submissions Staff. Using the SRA Toolkit to convert .sra files into other formats. *SRA Knowledge Base Internet* 2011:1–2.

Sexton T, Yaffe E, Kenigsberg E *et al.* Three-Dimensional Folding and Functional Organization Principles of the Drosophila Genome. *Cell* 2012;**148**:458–72.

Sharpe JA, Chan-Thomas PS, Lida J *et al.* Analysis of the human alpha globin upstream regulatory element (HS-40) in transgenic mice. *EMBO J* 1992;**11**:4565–72.

Shukla S, Kavak E, Gregory M *et al.* CTCF-promoted RNA polymerase II pausing links DNA methylation to splicing. *Nature* 2011;**479**:74–9.

Siersbæk R, Rabiee A, Nielsen R *et al.* Transcription Factor Cooperativity in Early Adipogenic Hotspots and Super-Enhancers. *Cell Reports* 2014;**7**:1443–55.

Simon JA, Kingston RE. Occupying Chromatin: Polycomb Mechanisms for Getting to Genomic Targets, Stopping Transcriptional Traffic, and Staying Put. *Molecular Cell* 2013a;**49**:808–24.

Simon JA, Kingston RE. Occupying Chromatin: Polycomb Mechanisms for Getting to Genomic Targets, Stopping Transcriptional Traffic, and Staying Put. 2013b;**49**:808–24.

Simonis M, Klous P, Splinter E *et al.* Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture–on-chip (4C). *Nat Genet* 2006;**38**:1348–54.

Soshnikova N, Montavon T, Leleu M *et al.* Functional Analysis of CTCF During Mammalian Limb Development. *Developmental Cell* 2010;**19**:819–30.

Soto-Reyes E, Recillas-Targa F. Epigenetic regulation of the human p53 gene promoter by the CTCF transcription factor in transformed cell lines. *Oncogene* 2010;**29**:2217–27.

Spencer RJ, del Rosario BC, Pinter SF *et al.* A Boundary Element Between Tsix and Xist Binds the Chromatin Insulator Ctcf and Contributes to Initiation of X-Chromosome Inactivation. *Genetics* 2011;**189**:441–54.

Spitz F, Furlong EEM. Transcription factors: from enhancer binding to developmental control. *Nat Rev Genet* 2012;**13**:613–26.

Spivak JL, Toretti D, Dickerman HW. Effect of phenylhydrazine-induced hemolytic anemia on nuclear RNA polymerase activity of the mouse spleen. *Blood* 1973;**42**:257–66.

Splinter E. CTCF mediates long-range chromatin looping and local histone modification in the beta-globin locus. 2006a;**20**:2349–54.

Splinter E. CTCF mediates long-range chromatin looping and local histone modification in the beta-globin locus. 2006b;**20**:2349–54.

Stadler MB, Murr R, Burger L *et al.* DNA-binding factors shape the mouse methylome at distal regulatory regions. *Nature* 2011;**480:**490–495.

Stark R, Brown G. DiffBind: Differential binding analysis of ChIP-Seq peak data. 2011:1–29.

Stedman W, Kang H, Lin S *et al.* Cohesins localize with CTCF at the KSHV latency control region and at cellular c-myc and H19/Igf2 insulators. *EMBO J* 2008;**27**:654–66.

Sung YH, Kim JM, Kim H-T *et al.* Highly efficient gene knockout in mice and zebrafish with RNA-guided endonucleases. *Genome Research* 2014;**24**:125–31.

Tanabe H, Müller S, Neusser M *et al.* Evolutionary conservation of chromosome territory arrangements in cell nuclei from higher primates. *Proc Natl Acad Sci USA* 2002;**99**:4424–9.

Tang Z, Luo OJ, Li X *et al.* CTCF-Mediated Human 3D Genome Architecture Reveals Chromatin Topology for Transcription. *Cell* 2015;**163**:1611–27.

Tolhuis B, Palstra RJ, Splinter E *et al.* Looping and interaction between hypersensitive sites in the active beta-globin locus. *Molecular Cell* 2002;**10**:1453–65.

Tucker KL, Wang Y, Dausman J *et al.* A transgenic mouse strain expressing four drug-selectable marker genes. *Nucleic Acids Research* 1997;**25**:3745–6.

Udvardy A, Maine E, Schedl P. The 87a7 Chromomere - Identification of Novel Chromatin Structures Flanking the Heat-Shock Locus That May Define the Boundaries of Higher-Order Domains. *J Mol Biol* 1985;**185**:341–58.

van de Werken HJG, Landan G, Holwerda SJB *et al.* Robust 4C-seq data analysis to screen for regulatory DNA interactions. *Nat Meth* 2012;**9**:969–72.

Vernimmen D, De Gobbi M, Sloane-Stanley JA *et al.* Long-range chromosomal interactions regulate the timing of the transition between poised and active gene expression. *EMBO J* 2007a;**26**:2041–51.

Vernimmen D, De Gobbi M, Sloane-Stanley JA *et al.* Long-range chromosomal interactions regulate the timing of the transition between poised and active gene expression. *EMBO J* 2007b;**26**:2041–51.

Vernimmen D, Marques-Kranc F, Sharpe JA *et al.* Chromosome looping at the human alpha-globin locus is mediated via the major upstream regulatory element

(HS-40). *Blood* 2009;**114**:4253–60.

Vernimmen D. Uncovering Enhancer Functions Using the α-Globin Locus. Rada-Iglesias A (ed.). *PLoS Genetics* 2014;**10**:e1004668–11.

Voigt P, Tee WW, Reinberg D. A double take on bivalent promoters. 2013;**27**:1318–38.

Volpi SA, Verma-Gaur J, Hassan R *et al.* Germline Deletion of Igh 3' Regulatory Region Elements hs 5, 6, 7 (hs5-7) Affects B Cell-Specific Regulation, Rearrangement, and Insulation of the Igh Locus. *The Journal of Immunology* 2012;**188**:2556–66.

Wallace HAC, Marques-Kranc F, Richardson M *et al.* Manipulating the Mouse Genome to Engineer Precise Functional Syntenic Replacements with Human Sequence. *Cell* 2007;**128**:197–209.

Wallace JA, Felsenfeld G. We gather together: insulators and genome organization. *Current Opinion in Genetics & Development* 2007;**17**:400–7.

Wan LB, Pan H, Hannenhalli S *et al.* Maternal depletion of CTCF reveals multiple functions during oocyte and preimplantation embryo development. *Development* 2008;**135**:2729–38.

Wang H, Maurano MT, Qu H *et al.* Widespread plasticity in CTCF occupancy linked to DNA methylation. *Genome Research* 2012;**22**:1680–8.

Wang H, Yang H, Shivalila CS *et al.* One-Step Generation of Mice Carrying Mutations in Multiple Genes by CRISPR/Cas-Mediated Genome Engineering. *Cell* 2013;**153**:910–8.

Wells RMG. Hemoglobin-Oxygen Affinity in Developing Embryonic Erythroid-Cells of the Mouse. *Journal of Comparative Physiology* 1979;**129**:333–8.

Wendt KS, Yoshida K, Itoh T *et al.* Cohesin mediates transcriptional insulation by CCCTC-binding factor. *Nature* 2008;**451**:796–801.

Weth O, Paprotka C, Gunther K *et al.* CTCF induces histone variant incorporation,

erases the H3K27me3 histone mark and opens chromatin. *Nucleic Acids Research* 2014;**42**:11941–51.

Whalen S, Truty RM, Pollard KS. Enhancer–promoter interactions are encoded by complex genomic signatures on looping chromatin. *Nat Genet* 2016;**48**:488–96.

Whyte WA, Orlando DA, Hnisz D *et al.* Master Transcription Factors and Mediator Establish Super-Enhancers at Key Cell Identity Genes. *Cell* 2013;**153**:307–19.

Wijgerde M, Grosveld F, Fraser P. Transcription complex stability and chromatin dynamics in vivo. *Nature* 1995;**377**:209–13.

Will CL. The human 18S U11/U12 snRNP contains a set of novel proteins not found in the U2-dependent spliceosome. *RNA* 2004;**10**:929–41.

Williams RRE. Neural induction promotes large-scale chromatin reorganisation of the Mash1 locus. *Journal of Cell Science* 2006;**119**:132–40.

Witcher M, Emerson BM. Epigenetic Silencing of the p16INK4a Tumor Suppressor Is Associated with Loss of CTCF Binding and a Chromatin Boundary. *Molecular Cell* 2009;**34**:271–84.

Xiao T, Wallace J, Felsenfeld G. Specific Sites in the C Terminus of CTCF Interact with the SA2 Subunit of the Cohesin Complex and Are Required for Cohesin-Dependent Insulation Activity. *Molecular and Cellular Biology* 2011;**31**:2174–83.

Xie X, Mikkelsen TS, Gnirke A *et al.* Systematic discovery of regulatory motifs in conserved regions of the human genome, including thousands of CTCF insulator sites. *Proc Natl Acad Sci USA* 2007;**104**:7145–50.

Yagi THETYYNGT, Tarusawa E, Yoshimura Y *et al.* CTCF Is Required for Neural Development and Stochastic Expression of Clustered Pcdh Genes in Neurons. *Cell Reports* 2012;**2**:345–57.

Yang H, Wang H, Shivalila CS *et al.* One-Step Generation of Mice Carrying Reporter and Conditional Alleles by CRISPR/Cas-Mediated Genome Engineering. *Cell* 2013;**154**:1370–9.

Yang R, Kerschner JL, Gosalia N *et al.* Differential contribution of cis-regulatory elements to higher order chromatin structure and expression of the CFTRlocus. *Nucleic Acids Research* 2015:gkv1358.

Yao H, Brick K, Evrard Y *et al.* Mediation of CTCF transcriptional insulation by DEAD-box RNA-binding protein p68 and steroid receptor RNA activator SRA. *Genes Dev* 2010;**24**:2543–55.

Yu W, Ginjala V, Pant V *et al.* Poly(ADP-ribosyl)ation regulates CTCF-dependent chromatin insulation. *Nat Genet* 2004;**36**:1105–10.

Yusufzai TM, Tagami H, Nakatani Y *et al.* CTCF tethers an insulator to subnuclear sites, suggesting shared insulator mechanisms across species. *Molecular Cell* 2004;**13**:291–8.

Zampieri M, Guastafierro T, Calabrese R *et al.* ADP-ribose polymers localized on Ctcf–Parp1–Dnmt1 complex prevent methylation of Ctcf target sites. *Biochem J* 2011;**441**:645–52.

Zhang L, Zhou J, Han J *et al.* Generation of an Oocyte-Specific Cas9 Transgenic Mouse for Genome Editing. Sun Q-Y (ed.). *PLoS ONE* 2016;**11**:e0154364–10.

Zhang Y, Liu T, Meyer CA *et al.* Model-based Analysis of ChIP-Seq (MACS). *Genome Biol* 2008;**9**, DOI: 10.1186/gb-2008-9-9-r137.

Zlatanova J, Caiafa P. CTCF and its protein partners: divide and rule? *Journal of Cell Science* 2009;**122**:1275–84.

Zuin J, Dixon JR, van der Reijden MIJA *et al.* Cohesin and CTCF differentially affect chromatin architecture and gene expression in human cells. *Proc Natl Acad Sci USA* 2014;**111**:996–1001.

Appendix

Appendix 1 – Supplementary tables

Supplementary table 1 – List of biotinylated oligonucleotides used in Capture-C experiments. Page 1

Name	Oligo sequence (5' - 3')	5' modification
HS44 F1	GATCTGCACCAGACATTCCTTCACAAGCACGGGTTAGAAGCCTAAG CTGGGAACAGCCAGTCCTCCCCTACCCACACTGCCCTGCCCTGTGG GAGAGGGCACTCGTCTTGGCTTAACTCT	Biotin [Btn}
HS44 R1	TGGGACTAGAAGGCACCTCACTGCCCTTGGGTAGCTTCTTGAGTGC CAGGGGAGATTCTTATACCACGTAAGGACAAAGATGTGACAAAGCT GAACTGAAGCCATGAGCAGTCCTGGATC	Biotin [Btn}
HS48 F1	GATCACACTGTCTGCTTGGCCTTCAGGTTTCAAACACAGGTCAGGG TGAATTCAGAGTCTAGATTCTCACAACAGCAGTCATGTCCTTACAA CCCACAAAATAACGTCTAGCTGTGCAAA	Biotin [Btn}
HS48 R1	TCATTATCGTCTCCTTGCTTAATTGACCATTCTATTTTATGGTGTC AACTTCTGAGAATTTCTAGGAAAATACAAAAAGGGACCACTCCAGG GAGCTGGAGGGTAAGACTGCATAAGATC	Biotin [Btn}
HS-38 F1	GATCCCAGTCAAGACCTAGACAAGGAAATGGAAGAGACATCAAAGC TTTATCAAAGGACTGTGACAGTGAAGTCAGGCACAGTCCTAGCACA GATTCCTTCAAACAGCCTCTTAGTAAAG	Biotin [Btn}
HS-38 R1	CTGCCACATGTCTTCCGCCACAAGAATGAGGGTATCTTGATGACTC CAGGGCTGGGTTGGATAAAAGGGGAGAAGGAGCAGAACAAAAGGGA CCCATGTGGAGCGTACCTCACTTGGATC	Biotin [Btn}
HS-39 F1	GATCTGGCCTCATGGATTCAAAGCCACTGAGGCCTGGCCACTGGGG GCGCCATTCGCCATTAAAAGGTCCTGCTGGGCTTTTCTAGCTCCAG ATTCCAGATTTTTGGCAGCCACTGGTAC	Biotin [Btn}
HS-39 R1	CACTGGCATTGGGAGACATGACATAGACGTTGTTCTCACACAGTGG GTAGATGATGACAGCCTTGCCCCAGTATACCAGGTGAGCTGCCAAC TGGAAGACCTGCAGACCAAGGAGAGATC	Biotin [Btn}

Supplementary table 1 – List of biotinylated oligonucleotides used in Capture-C experiments. Page 2

Name	Oligo sequence (5' - 3')	5' modification
Theta2 F1	GATCTACCTGACCTGAATTTACAGGCAAATGTGAGCAGGCATGTTT GTGCTGTGAACCAAAGCCAGGTACTCTGGGAGAGCAGTTAGTACTC TCAACTGCTGAGCCATTTCTCCATCCCC	Biotin [Btn}
Theta2 R1	CATCCTGGAACGATGCAGCGCCCCCTGGCGGCCTCTTGTGTTTCAG GACATCTTTGAGCTCAGCCAGGGGGCAGAGCGCAAGGCTCAGTTCA TTGAAGATGGCTCGGTCCCAGGATGATC	Biotin [Btn}
Theta1 R1	CATCCTGGAACGATGCAGCGCCCCCTGGCGGCCTCTTGTGTTTCAG GACGTCTTTGAGCTCAGCCAGCGGGCAGAGCGCAAGGCTCAGTTCA CTGAAGATGGCTCGGTCCCAGGATGATC	Biotin [Btn}
R3 F1	GATCAGCTGGGGTTGGCATTGCCCTCAGGAGCCATTCACTGAGGCT CACAGGCCTCTCCTTACTGCACTGCAGTTTGGGAACTGTACCTCTC AACATCTTTGACCCCAGTCCCTCCACAA	Biotin [Btn}
R3 R1	GGGAACCTGGATAAGTGAGTTGTCTGCAGGGAGGAGAGCCCTGGGC CTGGTGTTGCCTCTTATCTAAGAGACCAGGGACTGAGGGCCTCCAC TGTAGACATACAGGCACTGTTTCTGATC	Biotin [Btn}
R4 F1	GATCCCAGTCTACAGAAGTCCACTGTACTCCCCTCTGAGGCTGCCT TACCCAATACTGACAGCACTGGATAGGAGAACCCAGGACTAACTAA CCTGGGATGCATCACCTCAACTCCCAAG	Biotin [Btn}
R4 R1	GAGATGACCCTTCATCACACTGGGTCGGAAGTTGAGTCTAAAGTCA TCAGAAAGGTTTAATTGGTGAAACTGAATGGACGAGATTAATTTGC CTCCCATGGTCACCTTTCTTGGGGGATC	Biotin [Btn}
Rm F1	GATCTCTGTGAGTTCTAGACCAGCGTAGTCTAACTCCAGGCAGCCA GGGACACATAATGGAACCCTGTCTTAAGTAAATAAGTTCATTTTAA AATGAGACAATAAATGTTCTAAAGAAAA	Biotin [Btn}

Supplementary table 1 – List of biotinylated oligonucleotides used in Capture-C experiments. Page 3

Name	Oligo sequence (5' - 3')	5' modification
R1 F1	GATCCCCAGCCTATGTCTTTCCCTCTGAACATTTTGTACTAAGCTG GAGCCTCTTTACAAAGTAAAAAAGCTCTGGCATGGAGACACTGAAC ACTGCCATTAGGAGTCCCCCGAGTATCA	Biotin [Btn}
R1 R1	CTGTTAATTCAGAGAATCACTTCTTCACACCACAAGGAGTTTATAT CCTCATTCCTCCTGCCATTCCTTATCTCTTTAACTGTGGGTGAGCT GTGTTATGTTATAAAGTGAAGGGAGATC	Biotin [Btn}
R2 F1	GATCTATGGAGATGCTTGAACGAGCAGATAACTAAGCCAAGCATGA CTCAGAGTTTCTAGAGGCCACTAGGACTGCTGAGTAATACTTGGGG GTACAGAGTCAGAAAGGAAAGGACAAAT	Biotin [Btn}
R2 R1	CTGAGAAAGCTGCCACCTAGAATGAGGCAGAGTTTAATGGGAATGC TAACAAAAAGTGTATTTACAGCTAGGTATGGTAGCTTACCCCTATA ATGCCATCACTTGGGAGGTAGAAGGATC	Biotin [Btn}
Mpg F1	GATCTAGAGAACCTGGGTTCCAATCCTACTGGTGAAAGGTTTCTAC ACTCGCCCCGCCCTGTCCCGCCCCTAGAGTCCTCGCTCCGCCTCTT CTGCAGATTCCTTCCACCCCCGCGCTAC	Biotin [Btn}
Mpg R1	TGGTAGTGCGCGCCCGGGCAGAGGAGCCCTAAAACCGGTGTCCGTG ACCCTGCTCCCCGACACCGAGCAGCCTCCATTTCTTGGACGAGCCC GCCGCCCAGGGAATGCCAGAGCAGGATC	Biotin [Btn}
Rhbdf1 F1	GATCGGCGGACACGCGCCAGAACGCAAGCTCCGCACTCGCCCGCCC GGAACCCCGCGCAAAGACCCCGGCCGCCACTCACCGCCGTAGGGTC CTCAGAAGGGTCTGCACCGGCTGCTGCC	Biotin [Btn}
Rhbdf1 R1	CAAGAGGGTGGCGTGGGAAGGGACCTAAAAGGCAGATTGAACGCCC CTTCCGAAGGCACGCACTGGCACATCACTGAGGGTCCCGAAGCTCT CACTTGTCCTGTGCACACGCGCCTGATC	Biotin [Btn}

Supplementary table 2 – Summary of the haematology of D3839 mutant mice

WBC = white blood cell count, RBC = red blood cell count, HGB = haemoglobin count, HCT = haematocrit, MCV = mean corpuscular volume, MCH = mean corpuscular haemoglobin, PLT = platelet count, Retics % = percentage reticulocytes, Spleen/body % = spleen weight as a percentage of body weight.

	n=	WBC	RBC	HGB	HCT
WT	4	11.2±6.3	9.9±0.5	15.2±0.34	48.0±2.9
Het	16	9.6±3.0	9.6±0.6	14.4±0.9	47.4±4.5
Hom	6	7.5±2.8	9.6±0.5	14.3±0.6	45.9±2.2

	MCV	MCH	PLT	Retics %	spleen/body %
WT	49±1.7	15.3±0.53	454±24	1.5±0.72	0.3±0.04
Het	50±3.3	15.0±0.4	480±26	1.5±0.6	0.3±0.1
Hom	48±1.4	14.9±0.4	487±97	1.6±0.7	0.3±0.1

Supplementary table 3 – Complete overview of the haematology of D3839 mutant mice

WBC = white blood cell count, RBC = red blood cell count, HGB = haemoglobin count, HCT = haematocrit, MCV = mean corpuscular volume, MCH = mean corpuscular haemoglobin, PLT = platelet count, Retics % = percentage reticulocytes, Spleen/body % = spleen weight as a percentage of body weight.

Genotype	Sex	Age	WBC	RBC	HGB	HCT	MCV	MCH	PLT	Retics %	%spleen/body	mouse weight	spleen weight
wt	F	8 mons	5.6	9.31	15.0	43.8	47	16.1	445	2.0	0.35	26.3	0.091
wt	F	8 mons	9.9	9.83	15.0	49.7	51	15.2	464	0.6	0.32	33.5	0.108
wt	F	8 mons	9.1	10.4	15.7	50.1	48	15.1	481	2.1	0.27	28.9	0.079
wt	M	6 mons	20.2	10.1	15.1	48.5	48	14.9	426	1.1	0.26	31	0.081
			11.2±6.3	**9.9±0.5**	**15.2±0.34**	**48.0±2.9**	**49±1.7**	**15.3±0.53**	**454±24**	**1.5±0.72**	**0.3±0.04**	**29.9**	**0.1**
het	F	6 mons	10.4	10.26	15.6	49.7	48	15.2	407	2.6	0.39	25	0.097
het	M	6 mons	5.9	9.52	14.4	45.0	47	15.1	546	2.1	0.24	32	0.077
het	M	6 mons	11.8	9.95	14.4	46.6	47	14.4	355	1.6	0.27	30	0.081
het	M	6 mons	10.1	9.54	13.6	44.1	46	14.2	741	1.2	0.23	27	0.061
het	F	6 mons	17.1	9.91	15.1	51.9	52	15.2	282	2.1	0.05	25	0.012
het	M	4 mons	10.2	9.83	14.9	46.4	47	15.1	574	1.8	0.32	29	0.092
het	F	4 mons	7.3	9.95	15.3	48.9	49	15.4	431	1.2	0.36	25	0.09
het	M	3 mons	10.4	9.99	15.1	47.6	48	15.1	417	1.4	0.26	28	0.074
het	M	3 mons	8.6	9.85	15.1	53.6	54	15.3	500	0.6	0.24	28	0.068
het	F	3 mons	4.0	7.77	12.0	37.6	48	15.4	393	1.0	0.32	25	0.081
het	F	3 mons	9.3	9.56	14.7	50.5	53	15.4	395	1.5	0.36	23.5	0.085
het	M	8 mons	6.7	8.95	13.5	41.7	47	15	690	1.8	0.24	31.1	0.076
het	F	8 mons	10.5	9.92	15.1	54.3	55	15.2	363	0.6	0.22	34.2	0.074
het	F	8 mons	7.5	9.84	14.1	47.7	48	14.4	609	1.8	0.27	25.0	0.068
het	F	8 mons	12.4	9.03	13.6	42.5	47	15.1	494	1.9	0.40	29.4	0.118
het	F	8 mons	11.0	8.94	13.5	50.0	56	15.1	486	not countable	0.24	31.2	0.075
			9.6±3.0	**9.6±0.6**	**14.4±0.9**	**47.4±4.5**	**50±3.3**	**15.0±0.4**	**480±26**	**1.5±0.6**	**0.3±0.1**	**28.0**	**0.1**
hom	F	6 mons	5.1	10.12	14.8	47.6	47	14.6	358	2.5	0.37	24	0.089
hom	M	6 mons	10.5	9.9	14.3	45.4	46	14.5	541	1.2	0.18	33	0.06
hom	M	4 mons	4.6	8.72	13.3	42.0	48	15.3	573	1.1	0.28	29.6	0.083
hom	M	3 mons	11.2	9.86	14.6	46.8	47	14.8	468	2.3	0.33	26	0.086
hom	F	3 mons	7.1	9.55	14.0	48.0	50	14.6	391	0.7	0.33	24	0.078
hom	F	8 mons	6.6	9.59	14.6	45.8	48	15.3	588	2.0	0.25	31.8	0.079
			7.5±2.8	**9.6±0.5**	**14.3±0.6**	**45.9±2.2**	**48±1.4**	**14.9±0.4**	**487±97**	**1.6±0.7**	**0.3±0.1**	**28.1**	**0.1**

Appendix 2 – Generation of mouse models for the study of CTCF function

In this thesis, I have described the molecular phenotype of several mouse models containing deletions of CTCF sites at the α-globin locus. While these mice were successfully generated and analysed, various attempts to generate CTCF binding site mutants at different sites within the α-globin locus failed initially.

I first set out to target CTCF binding sites at the mouse α-globin cluster in ES cells heterozygous for the previously published humanised α-globin locus (see Fig. 1.6) (Wallace *et al.* 2007). The ability to specifically target CTCF binding sites on a single allele enabled the rapid successive deletion of different combinations of CTCF binding sites in mouse ES cells, including the mutations analysed in this thesis and, in addition, the combined mutation of a total of four CTCF sites at the cluster; HS-38, HS-39, θ1.and θ2. These cells were all blastocyst injected and consistently gave rise to high-percentage chimaeras (40-95%). Unfortunately, none of these chimaeras gave rise to transgenic offspring.

To avoid similar delays in future experiments, the experimental strategy was changed to sequential direct *in vivo* mutagenesis of CTCF binding sites by micro-injection of TALENs and CRISPR-Cas9. While this allowed quicker generation of founder mice carrying individual mutations of CTCF binding sites, it became more difficult to verify that sequential mutations had occurred in *cis* without first breeding mutant mice to homozygosity.

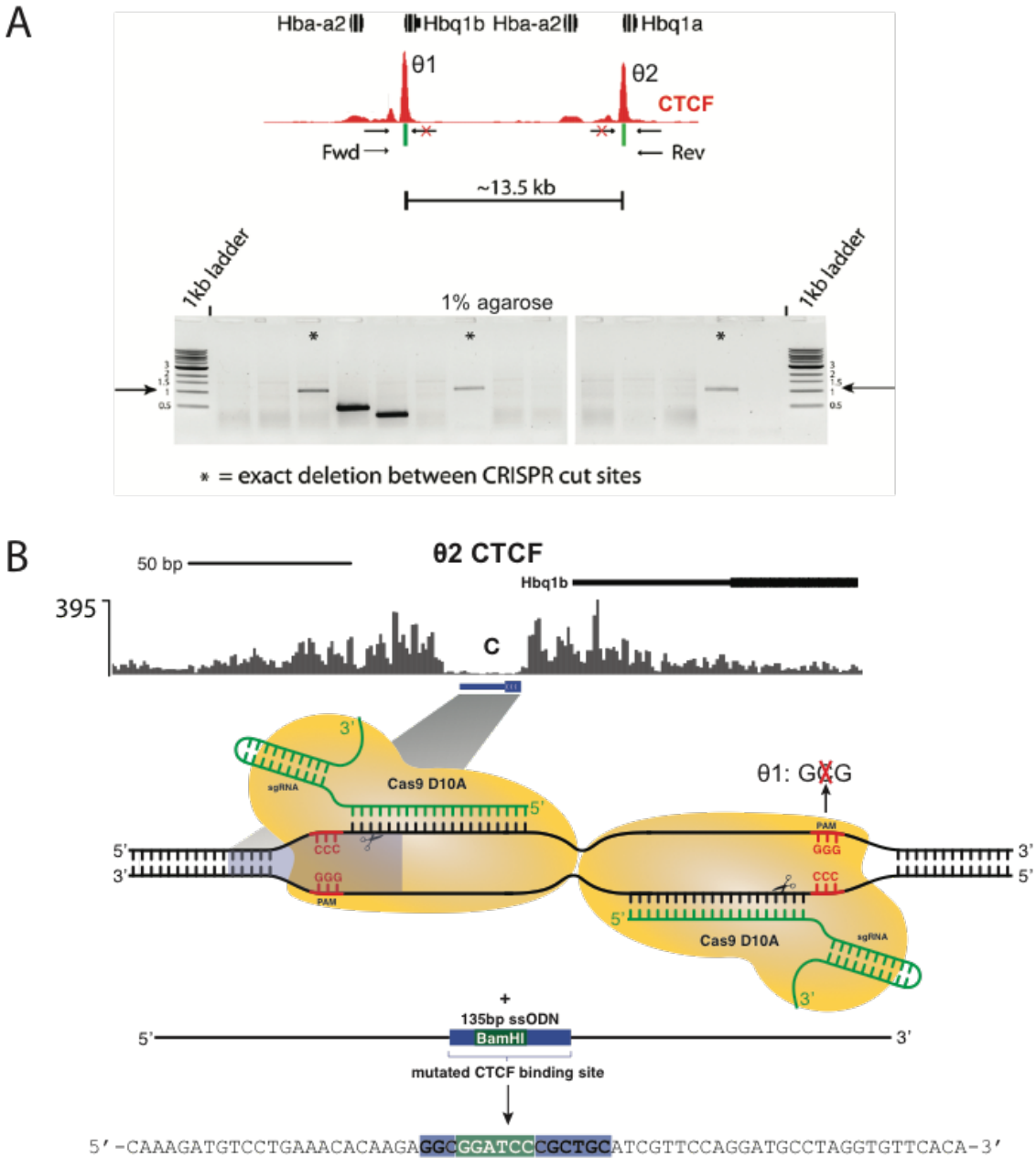


Figure A1. Mutagenesis of the θ2 CTCF binding site A. Simultaneous targeting of the homologous θ1 and θ2 CTCF binding sites results in the frequent deletion of the 13.5 kb sequence in between the binding sites. **B.** Double-nickase D10A Cas9 strategy for the specific targeting of θ2. A single nucleotide change between θ1 and θ2 destroys a PAM sequence ~30 bp away from the CTCF binding site at θ1. The CTCF core motif is shown in blue (reverse orientation) and annotated with DNaseI footprint data and gene annotation (Refseq).

For the mice analysed in this thesis, this posed no problem as the CTCF binding sites in the only double mutant (D3839) are located only 1.5 kb apart and could be analysed by cloning and sequencing of a PCR spanning both sites. Mice carrying mutations for these enhancer-proximal CTCF binding sites were generated and analysed, as has been described in this thesis.

To investigate the function of CTCF binding sites at the θ-globin promoters and the downstream HS44 and HS48 sites, I devised strategies to target these sites in pairs of two. The θ-globin promoters are highly homologous in their sequence. The use of a gRNA that targeted both sites simultaneously resulted in the frequent deletion of the 13.5 kb genomic region between θ1 and θ2 (Fig. A1). The existence of a single nucleotide difference between the θ1 and θ2 promoters that disrupted a potential CRISPR-Cas9 PAM sequence at the θ1 promoter made it possible to target the θ2 site specifically which was done in mouse ES cells (Fig. A1). This resulted in the successful and specific disruption of the θ2 CTCF binding site via HDR, also disrupting the CRISPR-Cas9 binding site so that the θ1 CTCF site could be targeted next. These embryonic stem cells were used in blastocyst injections and went germline successfully. The θ2 CTCF deletion mice are now used in combination with genetically delivered Cas9 and a micro-injected sgRNA against the remaining θ1 site to create a double θ1θ2 mutant. The θ2 single mutant is currently being bred for analysis and may yield interesting results, as α-globin enhancer interactions appear to be restricted by this site at the downstream end of the locus.

Similarly, I planned to target HS44 and HS48 sequentially to create a double HS44/48 mutant mouse. Genes and CTCF binding sites upstream of the α-globin

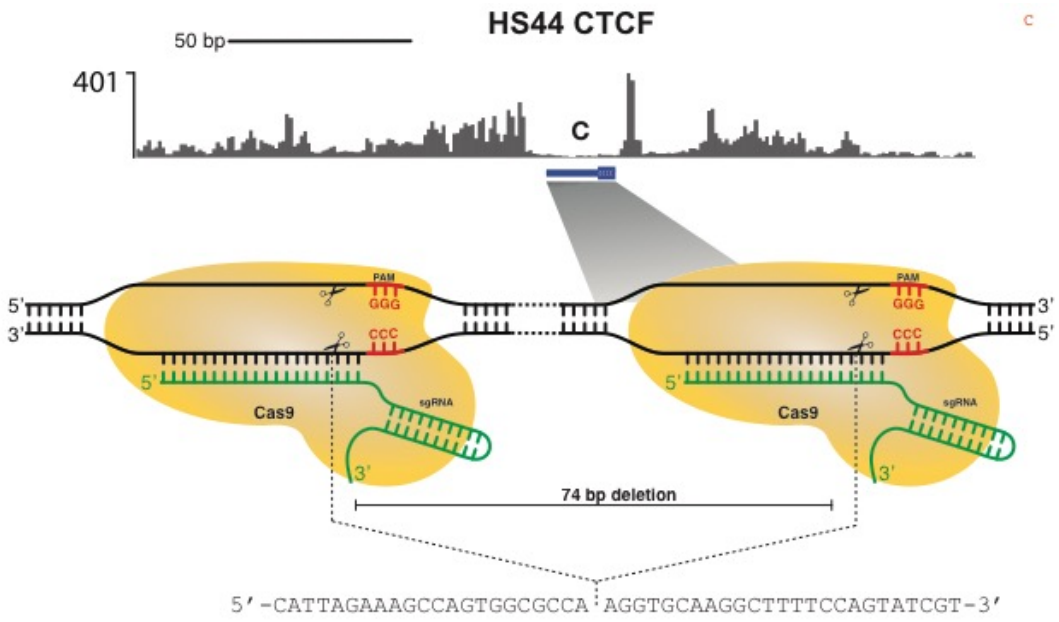


Figure A2. Deletion of the HS44 CTCF binding site Two Cas9 nucleases are simultaneously targeted to sites flanking the HS44 CTCF binding site, creating a 74 bp deletion containing the CTCF binding motif. The CTCF core motif is shown in blue (indicating reverse orientation) and annotated with DNaseI footprint data.

locus all interact with these downstream sites that delimit the α-globin sub-TAD, suggesting they may have an important architectural function. As the HS44 and HS48 regulatory elements are located away from genes and regulatory elements, I designed sgRNAs to target Cas9 to two sites flanking the HS44 CTCF binding site in ES cells, resulting in the deletion of a 74 bp sequence (Fig. A2). Again, these cells were used to create mice that successfully transmitted the mutant allele. HS44 CTCF site mutants were then crossed with the transgenic females that constitutively express Cas9. Oocytes were injected with two sgRNAs targeted to regions flanking HS48. While this experiment did not result in the complete deletion of HS48, a mouse containing a small deletion in HS48 that potentially disrupts CTCF binding was found. This mouse contained mutations in HS44 and HS48 in *cis*, making it a

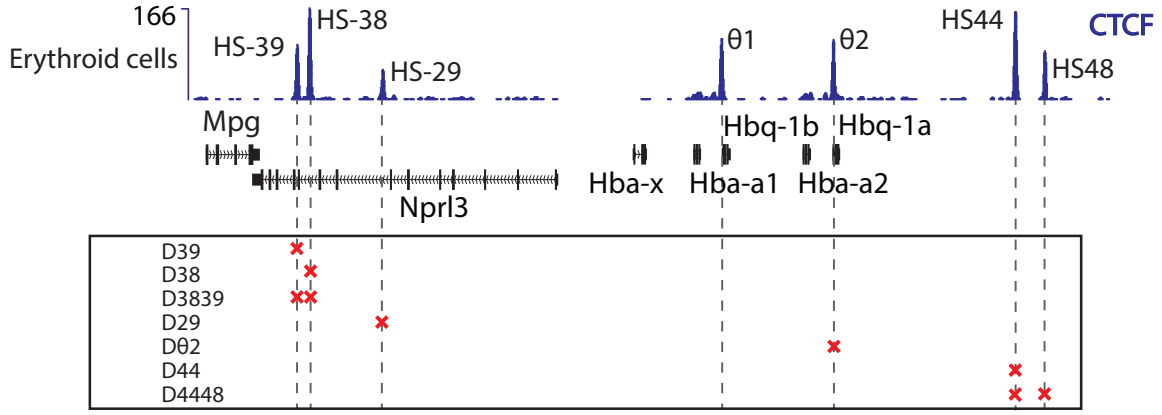


Figure A3. Overview of α-globin CTCF binding site mutants Red crosses indicate CTCF binding site mutations. Shown is CTCF ChIP-seq and Refseq genes.

potential double HS44/48 mutant. However, the loss of CTCF binding at HS48 will have to be confirmed. In summary, I have created seven CTCF binding site mutant mice, of which only those containing deletions of three upstream sites in close proximity to the enhancer have currently been analysed (Fig. A3). The analysis of downstream CTCF deletions will likely further expand our understanding of the mechanisms by which CTCF-mediated interactions are involved in gene regulation.

Finally, in addition to the systematic deletion of CTCF binding sites, I have also created several ectopic insertions of a CTCF binding sequence in between the *Nprl3* promoter and the ζ-globin gene (Fig. A4). Double nickase D10A Cas9 pairs were designed against two genomic sites at a distance of 1.5 and 5 kb from the ζ-globin gene promoter. Plasmids encoding the D10A Cas9 nickase and sgRNAs were transfected into ES cells together with a 130bp ssODN designed to introduce a 30 bp sequence that contained the HS-38 CTCF binding sequence. The introduction of this sequence was sufficient to acquire high levels of CTCF binding in ES cells at both

sites (Fig. A4). Unfortunately, although several chimaeras were produced, these ES cells did not go germline and could not be differentiated down the erythroid lineage.

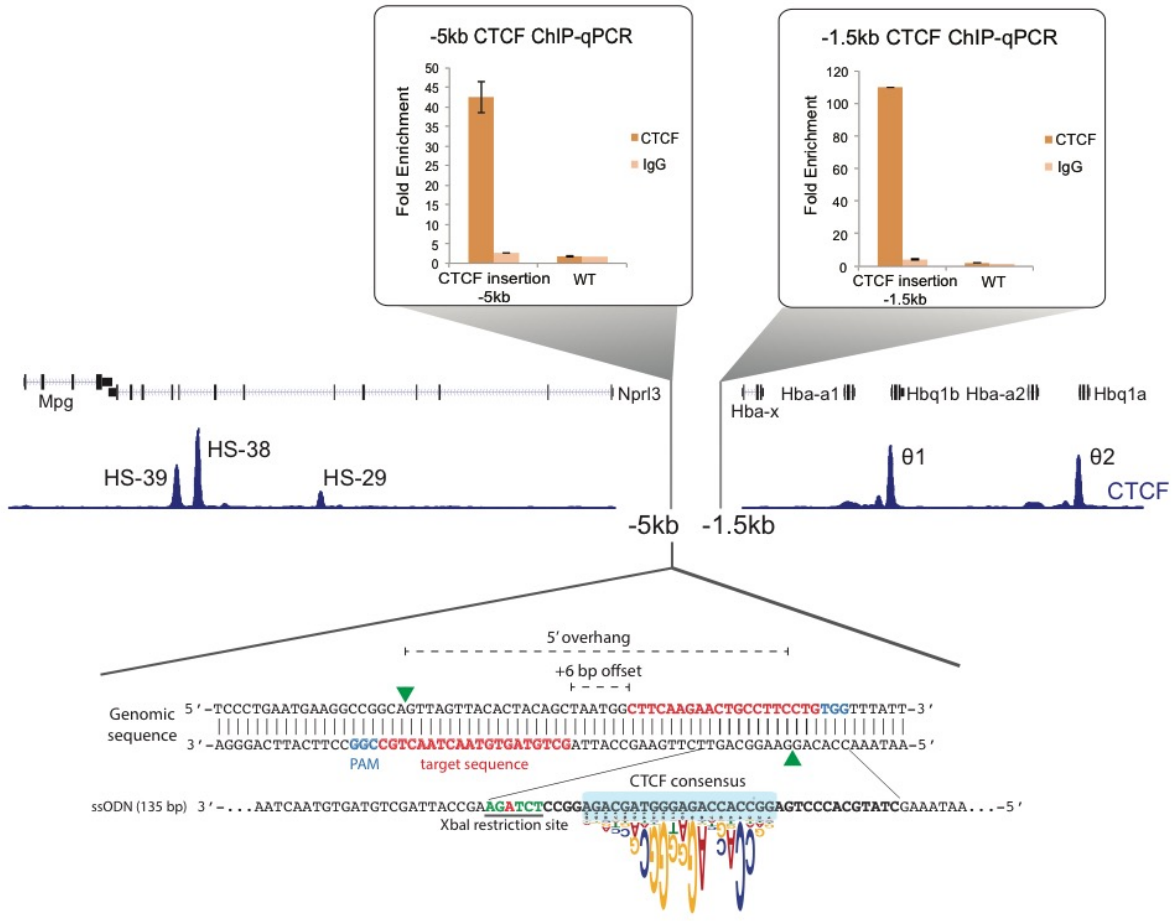


Figure A4. Insertion of an ectopic CTCF binding site at the α-globin locus The introduction of a 30 bp sequence containing the HS-38 CTCF binding sequence at two locations in between the α-globin enhancer and promoter results in high levels of CTCF binding to these genomic sites in ES cells.