Investigating the mechanisms and elements responsible for regulatory domain formation during gene regulation

A thesis submitted for the degree of Doctor of Philosophy of the University of Oxford



Caroline L. Harrold
Jesus College

Wellcome Doctoral Training Programme in Genomic Medicine and Statistics
MRC Weatherall Institute of Molecular Medicine
University of Oxford

Trinity Term 2019

To my partner, Tom,
for always providing me with food.

Abstract

The genome is a highly organised and complex structure; it is separated into distinct chromatin domains whose structures influence transcriptional regulation. Regulatory domains contain the promoters of active genes and their *cis*-regulatory elements, such as enhancers, and are flanked by boundary elements which bind CTCF. Processes that initiate and maintain this level of genome organisation are not well understood. In this thesis, I have investigated the role of promoters, enhancers, and CTCF-bound elements in generating and defining regulatory domains. To do this, I have exploited the well-studied regulatory domain and elements of the mouse α-globin locus.

In mouse erythroblasts, the five active enhancer elements and two active promoters of the α-globin genes interact and form a tissue-specific regulatory domain. These active elements have been well-characterised and are clearly involved in formation of the domain. In addition to the tissue-specific elements, there are numerous CTCF binding sites throughout the α-globin locus which may contribute to the structure of the regulatory domain. I have characterised a series of mouse models that harbour mutations of CTCF binding sites within the regulatory domain (θ1 & θ2) and downstream of the active promoters (HS+44 & HS+48) to understand their role in gene regulation and domain formation. Deletion of the downstream CTCF sites (Δ44-48) provides support for the loop extrusion model of domain formation. Deletion of CTCF sites within the active domain (Δθ1θ2 & Δθ2) suggests that promoters may act as domain boundaries, in addition to CTCF sites.

Genetic manipulations of the α-globin locus have helped us understand which elements are necessary for domain formation, but as the mouse α-globin locus is highly conserved and contains numerous elements that could be contributing to domain formation, it is hard to determine which elements are sufficient for domain formation. Therefore, I have identified "neutral" loci on chromosome X, devoid of genes and functional elements, and by genome editing, have generated artificial loci containing the mouse α-globin locus elements in various permutations in male mouse embryonic stem (ES) cells.

Introduction of the strongest α-globin enhancer, R2, into one such locus on chromosome X causes local changes in chromatin accessibility in erythroid cells. These changes in accessibility are accompanied by changes in chromatin modifications in which the accessible regions are marked with the active histone modifications H3K27ac and H3K4me1. Additionally, the R2 insertion gains another active histone modification H3K4me3 and produces long, unidirectional non-polyadenylated (~75 kb) and polyadenylated (~4 kb) transcripts. Finally, I found that the activity of the inserted R2 element is sufficient to form a regulatory domain structure similar to that observed at the α-globin locus. These results indicate that strong enhancer elements may be the key drivers in the formation of regulatory domains and provide insights into enhancer evolution and the mechanisms by which they regulate transcription of their target promoters.

Acknowledgements

I would like to express my gratitude and thanks to my supervisors Prof Doug Higgs and Prof Jim Hughes, for giving me the opportunity to work in their groups and for all of their advice, encouragement, and kindness over the last three years. Their supervision has allowed me to grow as a scientist: learning when to take risks, learning when to be cautious, and learning when to ask for help.

I would like to thank all members of the Higgs and Hughes groups for their help and expertise and the many valuable discussions. My 'PhD family' have provided a fun and stimulating research environment. I would like to thank the great post-docs in the groups: Dr Damien Downes and Dr Mira Kassouf for their support at the bench; Dr Marieke Oudelaar for discussions about the structure of the α-globin locus; and in particular, Dr Jelena Telenius and Dr Robert Beagrie for being so incredibly patient when teaching me bioinformatic analyses. I feel lucky to have made amazing friends throughout this experience who I could share all of the ups and downs with. I would like to thank Ron Schwessinger for his computational help and making me laugh with his dark humour. I value the numerous times I went to the 'downstairs lab' and sought the support and comfort of Lea Nussbaum and Steph Hill. I would like to thank Rosa Stolper for teaching me about the world of CTCF and answering my many questions about what it was like to learn a second language. Matt Gosden provided technical and emotional support from day one and was always there to tell me to keep going (especially when I was being dramatic); a true friend. Finally, I would like to thank Hele Francis who really has been my other half in our double act throughout this whole experience, and I hope Doug refers to us forever as 'The Terrible Twins'.

I would like to acknowledge the collaborators who have worked on the projects involved in this thesis. I am grateful to Dr Benjamin Davies and Dr Lars Hanssen for the generation of mouse models. My thanks go to the members of the Transgenics Facility (both past and present) for their invaluable help with the mouse work: Jackie Sharpe, Jackie Sloane-Stanley, Sue Butler, and Hannah Robinson. I would like to thank Philip Hublitz for the generation of reagents used for genome editing; Kevin Clark for performing numerous FACS sorts; and Jon Kerry for help with designing Capture-C probes. Lastly, the advice provided by Veronica Buckle and Tudor Fulga on my work during our thesis committee meetings has been immensely helpful.

Finally, I would like to thank my friends and my family. To the WhatsApp group 'Amazing People' for providing encouragement in the form of memes and quirky chat. To my parents, Tracey and David Harrold, my brother, Stephen, and my sister, Annabel, for the love and support they have always shown me throughout all my endeavours. Last but not least, my partner, Tom, who has been by my side for the last seven years and who has done more for me than can be put into words.

Table of Contents

ABSTRACT **3**

ACKNOWLEDGEMENTS **4**

TABLE OF CONTENTS **5**

LIST OF FIGURES **7**

LIST OF TABLES **9**

LIST OF ABBREVIATIONS **10**

CHAPTER 1: INTRODUCTION **12**

SPATIO-TEMPORAL REGULATION OF GENE EXPRESSION 12
THE GENOME IS A HIGHLY-ORGANISED THREE-DIMENSIONAL STRUCTURE 16
PROPOSED MECHANISMS OF CHROMOSOME ORGANISATION 25
THE WELL-CHARACTERISED, TISSUE-SPECIFIC REGULATORY LANDSCAPE OF THE MOUSE α-GLOBIN LOCUS 30
AIMS AND THESIS OUTLINE 37

CHAPTER 2: DISSECTING THE ROLE OF CTCF BOUNDARY ELEMENTS AT THE MOUSE α-GLOBIN LOCUS **39**

INTRODUCTION 39
RESULTS 41
Deletion of CTCF binding sites situated within and downstream of the mouse α-globin regulatory domain *41*
Deletion of downstream CTCF binding sites resulted in an expansion of the α-globin regulatory domain *46*
Deletion of CTCF binding sites within the α-globin regulatory domain revealed they do not regulate differential interactions and expression of the α-globin genes *50*
A revised model for promoter competition based on multi-way chromatin analysis of an extended regulatory domain *64*
DISCUSSION 69
DATA AVAILABILITY 72
ACKNOWLEDGEMENTS 72

CHAPTER 3: THE DESIGN AND CONSTRUCTION OF ARTIFICIAL LOCI **73**

INTRODUCTION 73
RESULTS 78
Identification of neutral regions of the mouse genome *78*
Selection of regions that aided targeting strategy and analysis of NGS-based assays *84*
Targeting strategy used to introduce the basic regulatory elements of the mouse α-globin locus *90*
DISCUSSION 95
ACKNOWLEDGEMENTS 97

CHAPTER 4: ANALYSING THE MOUSE α-GLOBIN LOCUS ELEMENTS IN ARTIFICIAL LOCI OUTSIDE OF THEIR NATIVE CHROMOSOMAL CONTEXT **98**

INTRODUCTION 98
RESULTS 99
The R2 enhancer showed variable local chromatin accessibility in artificial loci *99*
The inserted R2 enhancer appears as a promoter-like regulatory element in erythroid cells *108*

R2 is sufficient to form a regulatory domain at Artificial Locus 2 in erythroid cells *112*
Insertion of the α-globin promoter into the artificial loci does not lead to expression of a fluorescent reporter gene *114*
DISCUSSION 118
ACKNOWLEDGEMENTS 121

CHAPTER 5: DISCUSSION 122

CIS-REGULATORY ELEMENTS PLAY CONTEXT-DEPENDENT INTERCHANGEABLE ROLES WITHIN REGULATORY DOMAINS 123
INSIGHTS INTO ENHANCER EVOLUTION AND THE MECHANISMS FOR REGULATING GENE EXPRESSION 128
MOLECULAR MECHANISMS UNDERLYING CHROMOSOME ORGANISATION 130
CONCLUSION 137

CHAPTER 6: MATERIALS AND METHODS 138

CELLS 138
Isolation of erythroid cells derived from adult mouse spleen *138*
Culture of mouse embryonic stem cells *138*
Differentiation of mouse embryonic stem cells into embryoid bodies and isolation of erythroid cells *139*
GENOME EDITING AND GENOTYPING 140
Generation of mouse models *140*
Genotyping of mouse models *142*
Generation of mouse ES cell lines harbouring artificial loci *143*
Extraction of genomic DNA from 96-well plates *148*
Genotyping of mouse ES cell lines harbouring artificial loci *149*
ATAC-SEQ 151
CHIP-SEQ 152
Chromatin immunoprecipitation *152*
ChIP enrichment by qPCR *153*
Library preparation for Illumina sequencing *155*
CAPTURE-C 155
3C library preparation *156*
Capture-C library preparation *157*
Oligonucleotide capture *158*
TRI-C 162
RNA ISOLATION AND GENE EXPRESSION ANALYSIS 163
RNA isolation *163*
RT-qPCR expression analysis *163*
Poly(A)+ RNA-seq *164*
Poly(A)- RNA-seq *165*
DATA ANALYSIS 166
Mapping and visualisation of ATAC-seq and ChIP-seq data *166*
Peak calling of ChIP-seq data *169*
CTCF motif analysis *169*
Analysis of Capture-C data *169*
Analysis of Tri-C data *170*
Mapping and visualisation of RNA-seq data and variant analysis *170*

REFERENCES 172

List of Figures

Chapter 1

Figure 1.1: Layers of genome organisation. 18
Figure 1.2: Chromosome conformation capture techniques. 20
Figure 1.3: Formation of self-interacting domains by stable chromatin loops. 26
Figure 1.4: Formation of self-interacting domains by a dynamic loop extrusion mechanism. 28
Figure 1.5: Formation of self-interacting domains by compartmentalisation based on chromatin state. 29
Figure 1.6: The mouse α-globin locus. 32
Figure 1.7: Transcription factor and RNA polymerase II occupancy at the active mouse α-globin locus. 34
Figure 1.8: Chromatin interactions profiles at the mouse α-globin locus assayed by NG Capture-C 36

Chapter 2

Figure 2.1: CTCF binding sites at the α-globin locus. 43
Figure 2.2: Mutated sequences of CTCF binding sites at the α-globin locus 44
Figure 2.3: The presence of a large deletion in Δθ2 ATAC-seq. 45
Figure 2.4: Effects of the combined deletion of HS+44 and HS+48 on local chromatin accessibility, CTCF binding, genome topology, and gene expression. 50
Figure 2.5: Chromatin accessibility, CTCF binding, and quality of NG Capture-C profile in Δθ1 erythroid cells. 53
Figure 2.6: Effects of the deletion of θ2 on local chromatin accessibility, CTCF binding, and genome topology. 54
Figure 2.7: Effects of the combined deletion of θ1 and θ2 on local chromatin accessibility, CTCF binding, and genome topology. 55
Figure 2.8: Hba-α1/2-specific interaction profiles in Δθ2 erythroid cells. 58
Figure 2.9: Hba-α1/2-specific interaction profiles in Δθ1θ2 erythroid cells. 60
Figure 2.10: Comparison of Hba-α1/2-specific interaction profiles in Δθ2 and Δθ1θ2 erythroid cells. 62
Figure 2.11: Relative expression of Hba-α1/2 mRNA in WT, Δθ1, Δθ2, and Δθ1θ2 erythroid cells. 63
Figure 2.12: Effects of the combined deletion of HS-38 and HS-39 on multi-way chromatin interactions at the α-globin locus from the viewpoint of R2 in erythroid cells. 66
Figure 2.13: Effects of the combined deletion of HS-38 and HS-39 on multi-way chromatin interactions at the α-globin locus from the viewpoint of Mpg in erythroid cells. 68

Chapter 3

Figure 3.1: Overview of the design and targeting strategy for artificial loci. 77
Figure 3.2: Genome-wide chromatin accessibility and CTCF binding in erythroid cells. ..80
Figure 3.3: CTCF stitch blocks and genomic distances between CTCF peak calls. 82
Figure 3.4: A neutral region of chromatin on chrX. 83
Figure 3.5: Capture-C interaction profiles of AL1 from the viewpoints of LP1-5. 87
Figure 3.6: Capture-C interaction profiles of AL2 from the viewpoints of LP1-6. 89
Figure 3.7: Selected region of the R2 enhancer inserted into artificial loci. 91
Figure 3.8: Alignment of WT R2 and mutated R2 sequences inserted into artificial loci. .92
Figure 3.9: Selected region of the α-globin promoter inserted into artificial loci. 94
Figure 3.10: Selected sequence of the HS-38 CTCF binding site. 95

Chapter 4

Figure 4.1: Chromatin accessibility in ES cells after insertion of the R2 enhancer in AL1. 101
Figure 4.2: Chromatin accessibility in ES cells after insertion of the R2 enhancer in AL2. 101
Figure 4.3: Chromatin accessibility in erythroid cells after insertion of the R2 enhancer in AL1. 103
Figure 4.4: ATAC-seq reads originating from R2 in erythroid cells. 104
Figure 4.5: Chromatin accessibility in erythroid cells after insertion of the R2 enhancer in AL2. 106
Figure 4.6: Chromatin accessibility in erythroid cells after insertion of the mutated R2 enhancer in AL2. 107
Figure 4.7: Chromatin marks in erythroid cells after insertion of the R2 enhancer in AL2. 109
Figure 4.8: Comparison of chromatin marks in erythroid cells between the endogenous R2 on chromosome 11 and the inserted R2 in AL2. 110
Figure 4.9: Transcriptional changes in erythroid cells after insertion of the R2 enhancer in AL2. 111
Figure 4.10: Local genome topology in erythroid cells after insertion of the R2 enhancer in AL2. 113
Figure 4.11: Analysis of mVenus expression by FACS and reverse transcription qPCR in embryoid bodies containing an insertion of the enhancer R2 and the α-globin promoter in artificial loci. 117

Chapter 5

Figure 5.1: A summary of proposed proteins and mechanisms involved in the formation of self-interacting chromatin structures. 135

List of Tables

Table 6.1: gRNA target sequences of targeted CTCF binding sites at the mouse α-globin locus 141
Table 6.2: ssODN sequences used for disruption of CTCF binding sites at the mouse α-globin locus 141
Table 6.3: PCR primers used for the genotyping of mice 143
Table 6.4: gRNA target sequences of targeted LPs at artificial loci 144
Table 6.5: Donor sequences used for insertion of mouse α-globin locus elements into artificial loci 145
Table 6.6: PCR primers used for the genotyping of ES cell lines harbouring artificial loci 149
Table 6.7: Antibodies and conditions used for ChIP 153
Table 6.8: Primers used for ChIP-qPCR analysis 154
Table 6.9: Primers used for the quantification of digestion efficiency of 3C libraries 157
Table 6.10: Custom biotinylated DNA oligonucleotides used in Capture-C experiments 159
Table 6.11: Custom biotinylated DNA oligonucleotides used in Tri-C experiments 162
Table 6.12: Primers used for RT-qPCR gene expression analysis 164
Table 6.13: Taqman probes used for RT-qPCR gene expression analysis 164
Table 6.14: Publicly available datasets used in this thesis 167

List of Abbreviations

3C	Chromosome conformation capture
AL1	Artificial locus 1
AL1R2	Artificial locus 1 with inserted R2 sequence
AL2	Artificial locus 2
AL2R2	Artificial locus 2 with inserted R2 sequence
ATAC-seq	Assay for transposase-accessible chromatin with high-throughput sequencing
bp	Base pairs
BSA	Bovine serum albumin
Cas9	CRISPR-associated protein 9
cDNA	Complementary deoxyribonucleic acid
ChIP-seq	Chromatin immunoprecipitation with high-throughput sequencing
Chr11-R2	Endogenous R2 enhancer element on chromosome 11
CRISPR	Clustered regularly interspaced short palindromic repeats
CTCF	CCCTC-binding factor
D3839	Mouse with deletions of the HS-38 and HS-39 CTCF binding sites
dCas9	Dead endonuclease Cas9
DNA	Deoxyribonucleic acid
DNase I	Deoxyribonuclease I
DMSO	Dimethyl sulfoxide
EDTA	Ethylenediaminetetraacetic acid
eGFP	Enhanced green fluorescent protein
ENCODE	Encyclopaedia of DNA elements
eRNA	Enhancer RNA
ES	Embryonic stem
FACS	Fluorescence-activated cell sorting
FBS	Fetal bovine serum
Gata1	GATA-binding factor 1
GEO	Gene Expression Omnibus
GMEM	Glasgow’s minimal essential medium
H3K4me1	Histone 3 lysine 4 mono-methylation
H3K4me3	Histone 3 lysine 4 tri-methylation
H3K27ac	Histone 3 lysine 27 acetylation
H3K27me3	Histone 3 lysine 27 tri-methylation
H3K9me3	Histone 3 lysine 9 tri-methylation
HDR	Homology-directed repair
HPRT	Hypoxanthine-guanine phosphoribosyltransferase
IMDM	Iscove's modified dulbecco's medium
kb	Kilobase pairs
Klf1	Krüppel-like factor 1
LADs	Lamina-associated domains
LCR	Locus control region
LIF	Leukemia inhibitory factor
lncRNA	Long noncoding RNA
LP	Landing pad

MACS	Magnetic-activated cell sorting
Mb	Megabase pairs
MCS	Multispecies conserved sequences
mg	Milligram
ml	Millilitre
mM	Millimolar
mRNA	Messenger RNA
MTG	α-Monothioglycerol
mVenus	Monomeric variant of Venus yellow fluorescent protein
Nfe2	Nuclear factor erythroid 2
NG Capture-C	Next Generation Capture-C
NGS	Next-generation sequencing
PAM	Protospacer adjacent motif
PBS	Phosphate-buffered saline
PCR	Polymerase chain reaction
PFHM-II	Protein-free hybridoma medium
PIC	Pre-initiation complex
PolII	RNA polymerase II
qPCR	Quantitative polymerase chain reaction
RIN	RNA integrity number
RNA	Ribonucleic acid
RNA-seq	RNA sequencing
RPKM	Reads per kilobase per million
RPM	Reads per million
RPMI	Roswell Park Memorial Institute
rRNA	Ribosomal ribonucleic acid
RT	Reverse transcriptase
scaRNA-seq	Short capped RNA sequencing
Scl (Tal1)	Stem cell leukaemia factor
SDS	Sodium dodecyl sulfate
Seq	Sequencing
sgRNA	Single guide RNA
SMC	Structural maintenance of chromosomes
SNP	Single nucleotide polymorphism
ssODN	Single-stranded oligodeoxynucleotide
TAD	Topologically associating domain
Tris	Tris(hydroxymethyl)aminomethane
TSS	Transcriptional start site
WCHG	Wellcome Centre for Human Genetics
WIMM	Weatherall Institute of Molecular Medicine
WT	Wild-type
Δθ1	Mouse with a deletion of the θ1 CTCF binding site
Δθ2	Mouse with a deletion of the θ2 CTCF binding site
Δθ1θ2	Mouse with deletions of the θ1 and θ2 CTCF binding sites
Δ44-48	Mouse with deletions of the HS+44 and HS+48 CTCF binding sites
°C	Degrees Celsius
µg	Microgram
µl	Microlitre

Chapter 1: Introduction

Spatio-temporal regulation of gene expression

Complex multi-cellular organisms can be made up of trillions of cells of hundreds of distinct types and subtypes that have a wide variety of functions. The genome provides the code for the specification, development, and maintenance of these cells. The human genome contains ~22,000 protein-coding genes (Pertea & Salzberg, 2010) which is a similar number to that of the much simpler organism, the *Caenorhabditis elegans* (Chen et al., 2005). This suggests that gene number alone does not encode complexity. The *C. elegans* genome is made up of 10^8 base pairs, however the human genome contains 3 x 10^9 base pairs, of which only 1.2% is protein-coding genes (Lander et al., 2001). The remaining non-coding DNA has been characterised to reveal a vast number (~2-3 x 10^6) of genomic elements with predicted regulatory functions (ENCODE Project Consortium, 2012). This is consistent with the hypothesis that one way in which complexity can arise is from the different combinations of genes being switched on and off in different cell types and tissues across differentiation and development. Precise spatio-temporal regulation of gene expression is therefore essential for the development and maintenance of complex multi-cellular organisms.

Cis-regulatory elements are the non-coding genomic elements that have been implicated in the control of gene expression. The exact abundance of these elements is not clear; however, it's predicted that they vastly outnumber protein-coding genes in mammalian genomes (Lindblad-Toh et al., 2011). Sequence conservation analysis of 29 mammalian genomes revealed ~3.6 million evolutionarily constrained elements that encompassed ~4.2% of the human genome. The majority of constrained elements (~2.5 million) were found within intronic and intergenic regions (Lindblad-Toh et al., 2011). In addition, the maintenance of

functional *cis*-regulatory elements does not necessarily require strict sequence conservation (Ludwig et al., 2000) and the above may be an underestimate of the true value. Through chromatin profiling, the ENCODE Project Consortium (2012) assigned biochemical functions for ~80% of the human genome, in particular outside of the protein-coding regions, suggesting a range of functions for *cis*-regulatory elements. Nevertheless, the fundamental *cis*-regulatory elements of the genome include promoters, enhancers, and boundary elements.

Promoters are DNA sequences located directly upstream of the 5' transcriptional start site (TSS) of protein-coding genes (Maxam et al., 1977). They contain binding motifs for general transcription factors which nucleate the formation of the pre-initiation complex (PIC) which recruits RNA polymerase II and initiates transcription (Sainsbury et al., 2015). Unlike promoters, enhancers do not always lie adjacent to the genes that they regulate (Sanyal et al., 2012). In some cases, they are located more than one megabase pairs (Mb) away from their target gene (Lettice et al., 2003). Enhancers are defined as short regions of DNA that can upregulate (enhance) transcription of their target gene, independent of their relative position, distance, or orientation to the gene promoter (Banerji et al., 1981). These elements typically contain multiple binding sites for tissue- and developmental stage-specific transcription factors which can impact gene activity by initiating the opening of chromatin via recruitment of remodelling complexes to generate nucleosome-free regions. They can also recruit histone modifying complexes and co-activator proteins including the Mediator complex which, in mammals, is a 33-subunit complex that has been implicated in most stages of RNA polymerase II activation. Co-activator proteins are thought to promote the formation of the PIC to initiate transcription and/or overcome RNA polymerase II pausing (Heinz et al., 2015; Long et al., 2016; Vermunt et al., 2019). Developmental genes can be

controlled by multiple enhancers along the linear chromosome. Clusters of enhancer elements have been identified multiple times in the literature with varying names: locus control region (LCR), stretch-, spread-, or super-enhancers. These definitions are based on size, histone modification, and level of occupancy by nuclear factors such as tissue-specific transcription factors and Mediator (Heinz et al., 2015; Pott & Lieb, 2015; Vermunt et al., 2019; Whyte et al., 2013). Some methods used to identify enhancers rely on the concept that they are promiscuous and capable of controlling the expression of commonly used reporter genes with minimal promoters (Picard & Schaffner, 1983). However, some enhancers display a high-degree of specificity with the promoters of their target genes *in vivo*. One way to constrain enhancer function and promote effective communication between enhancers and promoters comes from enhancer-blocking insulators or boundary elements (Kellum & Schedl, 1991; West et al., 2002) which bind architectural proteins that contribute to the formation of specific enhancer-promoter complexes (Bickmore & van Steensel, 2013; Mirny et al., 2019; Vermunt et al., 2019).

An important insight for the identification of potential *cis*-regulatory elements is that they are associated with specific chromatin signatures. Promoters, enhancers, and boundary elements are typically located in regions with increased chromatin accessibility in the cells in which they are active/bound. The DNase I assay was originally used to identify these elements as sensitivity to DNase I is observed in chromatin which is depleted in nucleosomes, and bound by *trans*-acting factors (Parslow & Granner, 1983; Varshavsky et al., 1978). A more current and preferred method used to identify chromatin accessibility couples next-generation sequencing (NGS) with a hyperactive mutant Tn5 transposase which inserts sequencing adaptors into accessible regions of the genome (Buenrostro et al., 2013).

Having identified regions of open chromatin, the underlying regulatory elements can be classified via their chromatin signatures. Active promoters are associated with Histone 3 lysine 27 acetylation (H3K27ac) and Histone 3 lysine 4 tri-methylation (H3K4me3), and the presence of actively transcribing RNA polymerase II. Active enhancers are also marked with H3K27ac and are associated with actively transcribing RNA polymerase II, but usually display enrichment of Histone 3 lysine 4 mono-methylation (H3K4me1) and low levels of H3K4me3. Primed enhancer-like regions are marked with H3K4me1 but lack histone acetylation, and enhancers marked additionally by Histone 3 lysine 27 tri-methylation (H3K27me3, a repressive mark) are considered to be poised (Bannister & Kouzarides, 2011; Creyghton et al., 2010; Heintzman et al., 2007; Heinz et al., 2015; Rada-Iglesias et al., 2011). Boundary elements are bound by architectural proteins, such as the highly conserved zinc-finger protein CCCTC-binding factor (CTCF) (Chung et al., 1997), which appears to play a structural role in mammalian gene regulation (reviewed in Merkenschlager & Nora, 2016; Ong & Corces, 2014).

Accurate regulation of gene expression relies on robust and specific communication between *cis*-regulatory elements and physical proximity can mediate this communication. Enhancer-promoter interactions have been observed at multiple loci, including at the α- and β-globin loci in erythroid cells (Carter et al., 2002; Tolhuis et al., 2002; Vernimmen et al., 2007). More direct evidence for the functional importance of physical interactions between *cis*-regulatory elements came from tethering experiments in which proximity between *cis*-regulatory elements was forced in living cells (Deng et al., 2012; Nolis et al., 2009). In erythroid cells lacking Gata1, no interaction between the LCR and β-globin was established and β-globin was expressed at low levels. Zinc-finger-mediated tethering of Ldb1 (a transcription co-factor and candidate looping factor) to the β-globin promoter led to

recruitment of the LCR and activation of the β-globin gene in erythroid cells lacking Gata1 (Deng et al., 2012). A follow up study used this approach to rewire the LCR to the developmentally silenced γ-globin gene which resulted in the aberrant activation of this gene in adult erythroid cells, and the simultaneous reduction of interactions and transcription of the adult β-globin gene (Deng et al., 2014). Additionally, dCas9-mediated tethering of the transcription factor YY1 to the *Etv4* promoter in mouse ES cells increased interaction frequency between the promoter and its enhancer and resulted in increased transcription (Weintraub et al., 2017). These studies indicate that the structural organisation of the genome within the cell nucleus plays a critical role in establishing precise, tissue-specific patterns of gene expression.

The genome is a highly-organised three-dimensional structure

DNA within the nucleus of a eukaryotic cell is a highly-organised and complex three-dimensional structure (Figure 1.1). Individual interphase chromosomes occupy their own distinct territories within the nucleus, in which they are folded in such a way that they more frequently interact with themselves in *cis* rather than with other chromosomes in *trans* (T. Cremer & Cremer, 2001; T. Cremer et al., 1982; 2006). The radial positioning of chromosome domains within the nucleus correlates with gene content and activity. Gene poor regions and repressed genes are closer to the nuclear periphery and are often found in close contact with the lamina forming lamina-associated domains (LADS); whereas, gene-rich and active regions tend to localise more centrally (Bickmore & van Steensel, 2013; Boyle et al., 2001; Guelen et al., 2008; van Steensel & Belmont, 2017). Transcriptional activation often results in re-localisation of genes towards the nuclear interior. At the β-globin locus, activation of the LCR drove positioning of the locus away from

heterochromatin (Francastel et al., 1999) and promoted movement of the locus to the nuclear interior during erythroid cell maturation, which was accompanied by increased gene activity (Ragoczy et al., 2006). In haematopoietic progenitors and pro-T cells, the inactive Ig heavy (H) and Igκ loci were preferentially positioned at the nuclear periphery but were centrally located in the nuclei of pro-B cells, in which these loci are transcribed (Kosak et al., 2002). *Mash1* is an important pro-neural regulator gene. As mouse embryonic stem (ES) cells became committed to the neural lineage, the *Mash1* locus was transcriptionally activated and relocated towards the interior of the nucleus, away from the nuclear periphery (Williams et al., 2006). In addition, upon activation, loci can be extruded from their chromosome territories (Chambeyron & Bickmore, 2004; Ragoczy et al., 2003). More generally, actively transcribing genes have been proposed to co-localise into similar nuclear sub-compartments called 'transcription factories' (Jackson et al., 1993; Osborne et al., 2004; Ragoczy et al., 2006). Together, these findings suggest a link between transcriptional activity and nuclear positioning, but the driving forces behind genes/chromosome domains localising away from the nuclear periphery upon gene activation remain unclear. Ectopic integration of the β-globin LCR caused the new integration site to move away from the chromosome territory and resulted in activation of some of the surrounding genes (Noordermeer et al., 2008). This indicates that *cis*-regulatory elements can influence nuclear localisation, possibly through the binding of transcription factors. In ES cells, targeted chromatin de-condensation at a chosen gene in the absence of transcription activation was sufficient to mobilise the locus towards the nuclear interior (Therizols et al., 2014). This indicates that chromatin state and not transcription may be a driving force in nuclear repositioning.

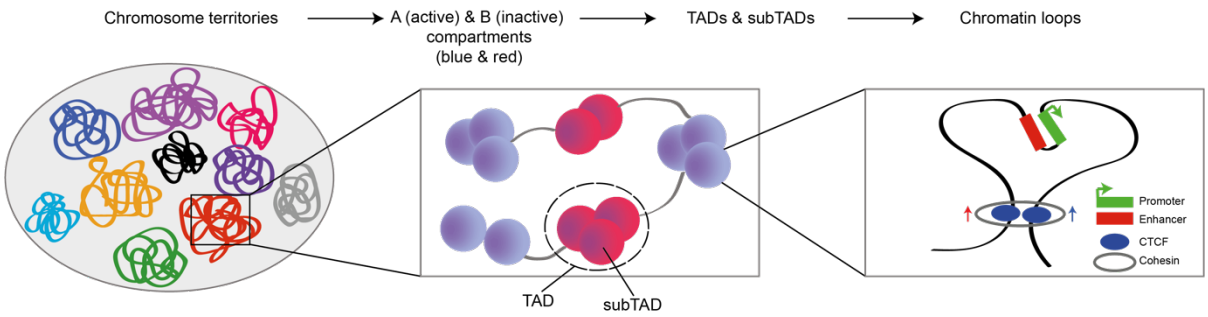


Figure 1.1: Layers of genome organisation. *Schematic overview of the organisation of chromatin at different scales. Interphase chromosomes occupy distinct territories within the cell nucleus. Within individual chromosomes, compartments are formed by increased inter-domain interactions between domains with similar transcriptional activity and epigenetic signatures known as A (blue) and B (red) compartments. The A compartment is predominantly formed of transcriptionally active chromatin and the B compartment represents a more repressed chromatin environment and is associated with inactivity. Chromatin regions within these compartments form domains of preferential interactions, often referred to as TADs. TADs are further subdivided into smaller domains of interaction termed sub-TADs. Within self-interacting domains, chromatin loops are formed between genes and/or cis-regulatory elements. Figure adapted from Denker & de Laat, 2016.*

With the emergence of Chromosome Conformation Capture (3C: Cullen et al., 1993; Dekker, 2002) technology and its high-throughput derivatives (4C, 5C, Capture-C, and Hi-C, reviewed in Davies et al., 2017 and Denker & de Laat, 2016), it has been possible to study DNA folding at high resolution and in a more systematic manner to reveal that chromosomes display multiple layers of genome organisation which guide communication between genes and their *cis*-regulatory elements. The above 3C technologies employ stabilisation of chromosomal contacts via chemical cross-linking (usually formaldehyde) and fragmentation of the cross-linked chromatin. Usually, fragmentation is achieved by digestion using a restriction enzyme. The most commonly used restriction enzymes target either six or four base pairs (bp) recognition sequences, resulting in the genome being cut, on average, every 4096 bp and 256 bp respectively. The fragmented chromatin is ligated back together, and

the resulting ligation junctions of chimeric fragments are detected and quantified (Figure 1.2). The underlying concept is that the closer genomic regions are positioned in nuclear space, the more likely they are to be cross-linked and ligated to each other. These methods provide independent and more detailed evidence for the existence of chromosome territories, their limited capacity to intermingle, and the spatial separation of active and inactive chromatin in the nucleus (Lieberman-Aiden et al., 2009; Simonis et al., 2006).

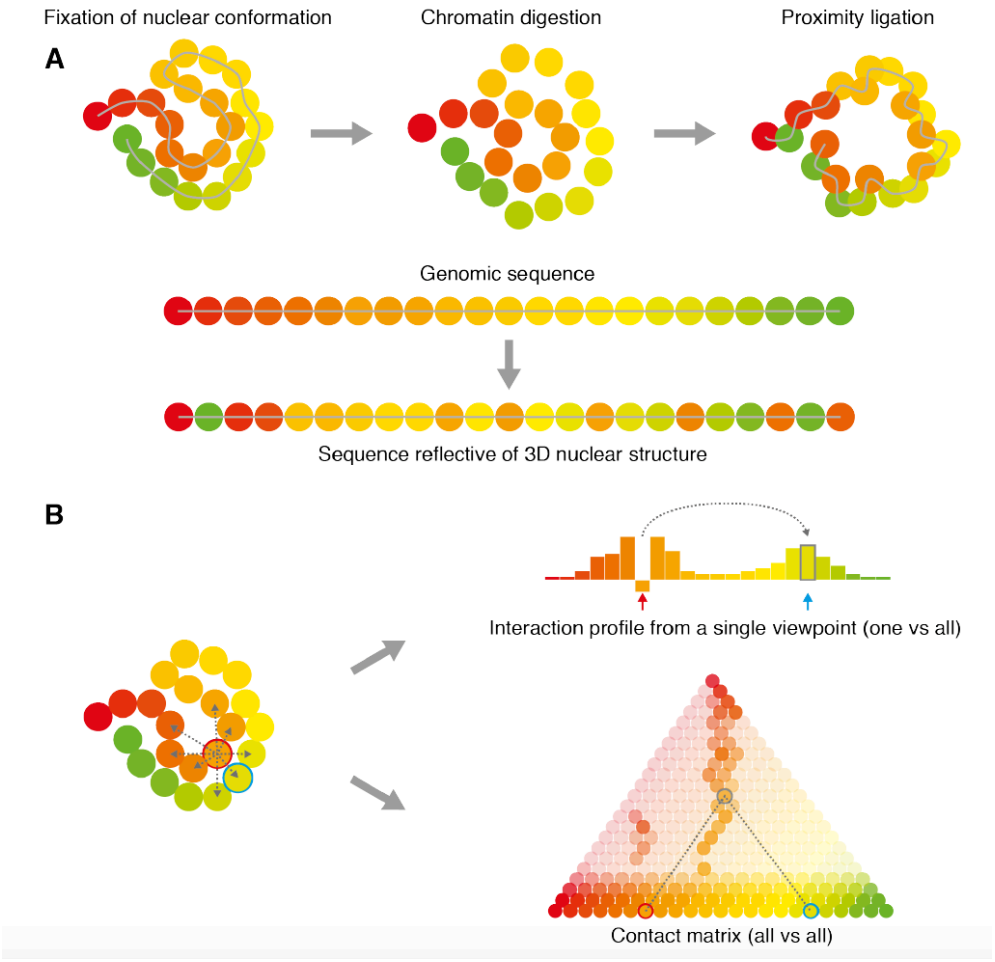


Figure 1.2: Chromosome conformation capture techniques. A: *3C techniques rely on cross-linking of DNA sequences that are close in nuclear proximity, followed by fragmentation with restriction enzymes and* in situ *proximity ligation. This results in large DNA concatemers in which the order of the ligated fragments reflects the three-dimensional structure of the chromatin in the cell nucleus. Quantification of the newly formed ligation junctions in 3C libraries provides information about the frequencies of interactions between DNA elements.* ***B:*** *Several techniques based on the 3C principle are used to study different aspects of chromatin architecture. 'One vs all' techniques can identify all interactions with a DNA fragment of interest at high resolution, which are usually represented in genome-wide interaction profiles from the viewpoint of interest. 'All vs all' techniques can analyse structural interactions across the entire genome at moderate resolution, which are typically shown in contact matrices in which the interaction frequency is represented by the colour intensity at the intersection of the interacting DNA elements. Figure adapted from Davies et al., 2017.*

Based on the analysis of data obtained from the above 3C methods, it has been shown that chromosomes are organised into sub-chromosomal domains. At the megabase scale, chromosomes form large, functionally distinct domains, referred to as A and B compartments (Lieberman-Aiden et al., 2009) (Figure 1.1). The A compartment is formed of blocks of predominantly gene-rich, transcriptionally active, and accessible chromatin which associate with other regions of active chromatin within the interior of the nucleus. The B compartment represents a more repressed environment with fewer genes and low levels of expression. Chromatin associated with B compartments is associated with repressive histone marks. The B compartments are located closer to the nuclear periphery and include LADs, which have also been associated with gene repression (Gibcus & Dekker, 2013; Peric-Hupkes et al., 2010).

Within such compartments, 3C methods have shown that chromosomes are subdivided into smaller structural domains at the scale of hundreds of kilobases (kb) to a few Mb known as topologically associating domains (TADs; Dixon et al., 2012; Nora et al., 2012; Sexton et al., 2012) (Figure 1.1). TADs have been detected with high-resolution imaging as well as genomic methods that are independent of proximity ligation, suggesting that they are not reflective of technical biases (Beagrie et al., 2017; Nir et al., 2018; Quinodoz et al., 2018; Szabo et al., 2018; H. Wang et al., 2016). TADs represent chromosomal units within which chromatin regions preferentially interact with each other, whereas interactions between regions located in adjacent TADs, separated by intervening boundaries are relatively depleted. It is now widely believed that TADs not only form structural entities but also serve as the functional units of chromosomes. TADs correlate with chromatin signatures, gene activity, and replication timing (Dileep et al., 2015; Dixon et al., 2012; Nora et al., 2012). A TAD forms a framework within which the correct enhancers and promoters find each other

(Shen et al., 2012). Transposition of a regulatory sensor into >1000 integration sites of the mouse genome revealed that enhancers function within large regulatory domains that correlate with TADs (Symmons et al., 2014). When TADs were first identified, they were reported to be stable between different cell types (Dixon et al., 2012; Nora et al., 2012). While some conservation was observed, 35% and 50% of the TADs differed between mouse ES cells and cortex or lung fibroblasts (Dixon et al., 2012). In addition, analysis of Hi-C data at 40 kb resolution in 21 human cell lines and primary tissues found that nearly 10% of all TAD boundaries were present in only one cell type and that <40% were conserved between all examined cell types (Schmitt et al., 2016). Genomic regions of which the TAD structure was found to differ between cell types were highly enriched for differentially expressed genes (Dixon et al., 2012). These findings indicate that conserved TADs reflect areas of the genome with comparable activity in the analysed cell types, but that regions in which gene activity was altered were structurally reorganised. Therefore, some TADs may represent tissue-specific structures that are closely related to genome function (Bonev et al., 2017).

High resolution 3C data have revealed smaller, self-interacting domains that are nested within larger TADs and have been referred to as sub-TADs (Phillips-Cremins et al., 2013), contact domains (Rao et al., 2014), insulated neighbourhoods (Dowen et al., 2014), and frequently interacting regions (Schmitt et al., 2016) (Figure 1.1). These domains vary in size between ~40 kb and 3 Mb, with a median size of 185 kb (Rao et al., 2014). Tissue-specific interactions between genes and their *cis*-regulatory elements are thought to occur within these domains and are often described as chromatin loops (Bonev & Cavalli, 2016) (Figure 1.1). The α- and β-globin loci provide good examples of tissue-specific chromatin loops. At both loci, the cluster of enhancers interact strongly with their target promoters in erythroid

cells, where the genes are active, but show little or no interaction in cells where the genes are inactive (Davies et al., 2016; Palstra et al., 2003). These interactions have been proposed to form an 'active chromatin hub' (Tolhuis et al., 2002). Recent, preliminary evidence suggests that transcriptional activation takes place in nuclear condensates, which contain a high concentration of transcription factors, co-factors and components of the basal transcription machinery recruited by enhancer elements (Boehning et al., 2018; Boija et al., 2018; Cho et al., 2018b; Chong et al., 2018; Sabari et al., 2018). Such condensates imply that multiple *cis*-regulatory elements interact and function together in a hub-like complex. These hub-like structures have recently been identified using chromosome conformation assays at both of these loci (Allahyar et al., 2018; Oudelaar et al., 2018).

A common feature of all self-interacting domains (compartments/TADs/sub-TADs/chromatin loops) is that their borders are enriched for boundary elements that among many other proteins commonly bind CTCF (Dixon et al., 2012; Nora et al., 2012; Phillips-Cremins et al., 2013; Rao et al., 2014). CTCF is an 11 zinc-finger, sequence-specific DNA binding protein. CTCF is conserved in most bilaterian metazoan (Heger et al., 2012; Klenova et al., 1993), broadly expressed, and essential for both cellular function (Burcin et al., 1997; Fedoriw et al., 2004) and embryonic development (Wan et al., 2008). As the DNA recognition sequence of CTCF is not palindromic, the binding orientation to chromatin can be determined as forward or reverse. It has been shown that CTCF plays an important role in the formation of domains and their boundaries, and of interest the CTCF binding motifs at these sites are usually found in convergent orientations (de Wit et al., 2015; Gómez-Marín et al., 2015; Guo et al., 2015; Rao et al., 2014; Sanborn et al., 2015; Vietri Rudan et al., 2015).

In addition to CTCF binding, the cohesin complex is often enriched at the boundaries of self-interacting domains. CTCF and cohesin co-occupy tens of thousands of binding sites across the genome (Hadjur et al., 2009; Parelho et al., 2008; Rubio et al., 2008; Stedman et al., 2008; Wendt et al., 2008). Cohesin is a ring-like multiprotein complex that provides cohesion between sister chromatids during mitosis and is involved in DNA replication (Nasmyth & Haering, 2009). Cohesin is composed of a heterodimer of structural maintenance of chromosomes (SMC) proteins, Smc1 and Smc3, and further bound by a kleisin family subunit known as Rad21 (Scc1 in yeast) and SA1 or SA2 (Scc3 in yeast) to form a structure large enough to enclose two chromatin fibres (Nasmyth & Haering, 2009). Although many cohesin sites often coincide with the binding of CTCF, others often map to active promoters and enhancers, where it co-localises with its loading factor Nipbl, Mediator components, and tissue-specific transcription factors (Faure et al., 2012; Kagey et al., 2010; Schmidt et al., 2010). These findings have prompted the suggestion that cohesin may also act independently of CTCF in stabilising enhancer-promoter interactions.

CTCF and cohesin are both thought to play a role in organising the genome into self-interacting domains. The functional relevance of genome organisation in the regulation of gene expression has been demonstrated by perturbation of these proteins both at individual loci and globally. Various studies have shown that perturbation of CTCF- (and cohesin-) associated domain boundaries can lead to changes in chromatin structure that allow enhancers to interact with genes in neighbouring domains, resulting in mis-regulated gene expression (Dowen et al., 2014; Franke et al., 2016; Gómez-Marín et al., 2015; Hanssen et al., 2017; H. K. Lee et al., 2017; Lupiáñez et al., 2015; Narendra et al., 2015; 2016; Nora et al., 2012; Symmons et al., 2016). Interestingly, experiments in which such proteins were depleted globally reported limited transcriptional changes, even though TAD structures

were markedly perturbed (Hyle et al., 2019; Nora et al., 2017; Rao et al., 2017; Schwarzer et al., 2017; Wutz et al., 2017). However, removal of these proteins does lead to changes in expression of a subset of genes. The majority of the downregulated genes are predicted to be regulated by distal *cis*-regulatory elements. Therefore, without the structure imposed by TADs, the distal elements no longer interact with their correct target promoter, and in some cases, cause upregulation of neighbouring genes and/or intergenic regions located in their vicinity (Schwarzer et al., 2017). It is therefore believed that self-interacting domains delimit chromatin regions in which gene regulatory mechanisms take place and they can be thought of as functional regulatory domains. It still remains unclear how regulatory domains are formed and how they contribute to specific enhancer-promoter interactions.

Proposed mechanisms of chromosome organisation

In Hi-C contact matrices obtained from bulk populations of cells, the boundaries of regulatory domains often form punctate interactions (also known as 'peaks'; Rao et al., 2014) which are thought to reflect the presence of chromatin loops, with the peaks representing the anchor points of the chromatin loop. Such loop domains have been interpreted as relatively stable structures, formed by affinity between CTCF proteins located at the anchor points (Rao et al., 2014) (Figure 1.3). Reports that CTCF can form dimers/multimers *in vitro* and in two-hybrid assays provide some support for this mechanism (Pant et al., 2004; Yusufzai et al., 2004). However, single-cell Hi-C studies (Nagano et al., 2013; Stevens et al., 2017), simulations based on polymer models (Giorgetti et al., 2014), and a recent microscopy-based study (Bintu et al., 2018) suggested that regulatory domain structures in single cells are highly diverse. In addition, this model falls short in explaining why not all CTCF sites interact with each other, and in particular, why

interactions at boundaries are predominantly between convergent pairs of CTCF binding sites (Rao et al., 2014; Vietri Rudan et al., 2015).

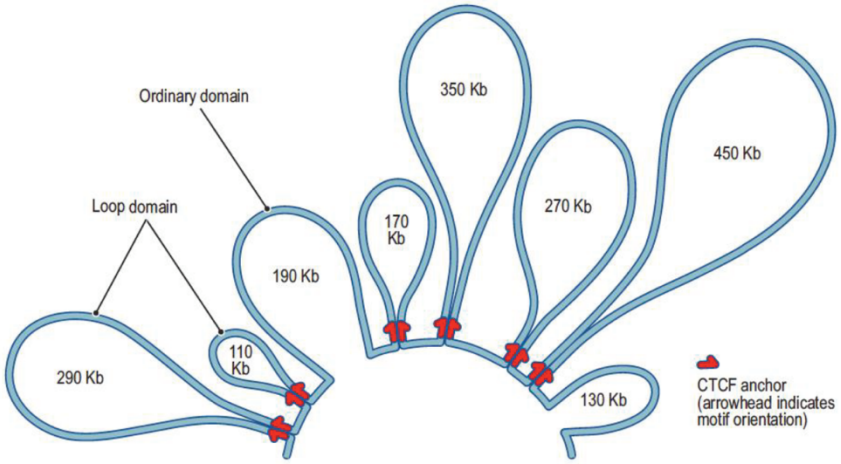


Figure 1.3: Formation of self-interacting domains by stable chromatin loops. *Interpretation of the structure of a 2.1 Mb region on chromosome 20 in GM12878 cells. The structure is represented by eight domains, ranging in size from 110 kb to 450 kb, in which 95% of the chromatin in the region is contained. Six of the eight domains are demarcated by loops between convergent CTCF binding sites located at the domain boundaries. Figure adapted from Rao et al., 2014.*

A more dynamic mechanism proposed to contribute to regulatory domain formation is the loop extrusion model (Fudenberg et al., 2016; Sanborn et al., 2015) (Figure 1.4), which was first proposed to organise mitotic chromosomes (Alipour & Marko, 2012; Nasmyth, 2001). Loop extrusion is thought to be an active process in which a loop extruding factor, with two DNA-binding domains, associates with chromatin and travels along the fibre in opposite directions. This creates a progressively larger intervening loop until the loop extruding factor is stalled at boundary elements, such as convergently bound CTCF sites, which defines the limits of the regulatory domain (Fudenberg et al., 2016; Sanborn et al., 2015). Due to its co-localisation with CTCF binding sites, its ATPase activity, and its ability to interact with two

double-stranded DNA helices, cohesin has been proposed as a most likely loop extruding factor during interphase in the mammalian genome (Abdennur et al., 2018). This model of regulatory domain formation recapitulates many features of interphase chromosome organisation: the formation of non-overlapping dynamic chromatin domains that are primarily self-interacting, insulation between *cis*-regulatory elements located in adjacent domains, enrichment of architectural proteins at the boundaries of domains, and preferential convergent orientation of CTCF binding sites at the boundaries. The molecular mechanism of loop extrusion by SMC protein complexes is being actively investigated (reviewed in Nichols & Corces, 2018), but one open question is whether loop extrusion proceeds in both directions from the site of association as predicted by the models. Loop extrusion has been demonstrated to occur *in vitro* by the yeast condensin complex in single-molecule experiments (Ganji et al., 2018), which showed it reeling in DNA only on one side. The process of loop extrusion has not been directly observed *in vivo*. Other open questions include if and how loop extrusion could contribute to the formation of tissue-specific regulatory domains, how and where loop extruding factors are recruited along the genome, the mechanism that allows CTCF to impede extrusion in spite of the relatively rapid turnover of CTCF on DNA (Hansen et al., 2017), and interactions of extrusion with transcription (Brandão et al., 2019).

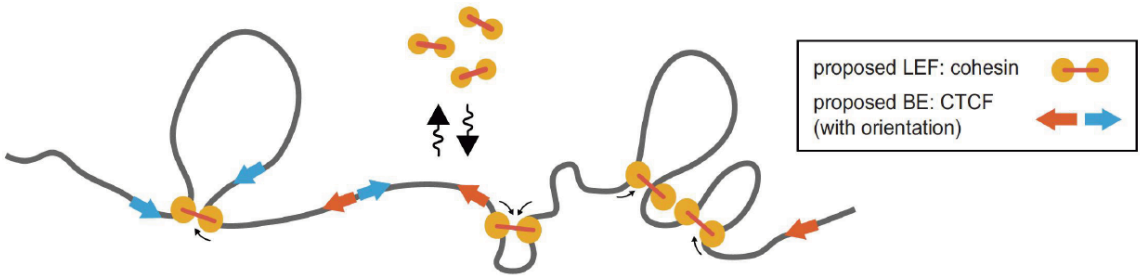


Figure 1.4: Formation of self-interacting domains by a dynamic loop extrusion mechanism. *Schematic of the hypothetical process of loop extrusion, in which self-interacting domains result from* cis*-acting loop extruding factors (LEFs; likely cohesin), which form progressively larger loops, until stalled at boundary elements (BEs), such as CTCF-bound elements. Figure adapted from Fudenberg et al., 2016.*

Another model proposed to organise the genome relies on CTCF/cohesin-independent segregation of the genome into compartments defined by chromatin state (Lieberman-Aiden et al., 2009; Mirny et al., 2019) (Figure 1.5). Support for this model originates from the observation that active and inactive (A/B) compartments of the genome are unaffected by global CTCF depletion (Hyle et al., 2019; Nora et al., 2017), and removal of chromatin-associated cohesin led to enhanced compartmentalisation patterns in Hi-C contact matrices which more accurately reflected transcriptional activity and chromatin state (Rao et al., 2017; Schwarzer et al., 2017; Wutz et al., 2017). This spatial separation of different chromatin states is similar to microphase separation in polymers with two or more types of monomers, where weak attractions between monomers of the same type is sufficient to drive phase separation (known as block copolymers) (Leibler, 1980). Therefore, it has been suggested that attractions among regions with similar epigenetic signatures could establish the segregation of active and inactive compartments due to phase separation (Haddad et al., 2017; Jost et al., 2014; Nuebler et al., 2018). Additional support for this model comes from

experiments performed on cells with naturally or artificially inverted nuclei and polymer simulations (Falk et al., 2018). The re-localisation of heterochromatin to the centre of the nucleus and euchromatin to the nuclear periphery found that affinity between heterochromatic regions was essential for establishing compartmentalisation by phase separation, and that heterochromatin-lamina interactions and euchromatic interactions were dispensable for compartmental segregation. However, this model alone does not explain smaller scale chromosome structures such as regulatory domains which contain specific enhancer-promoter interactions.

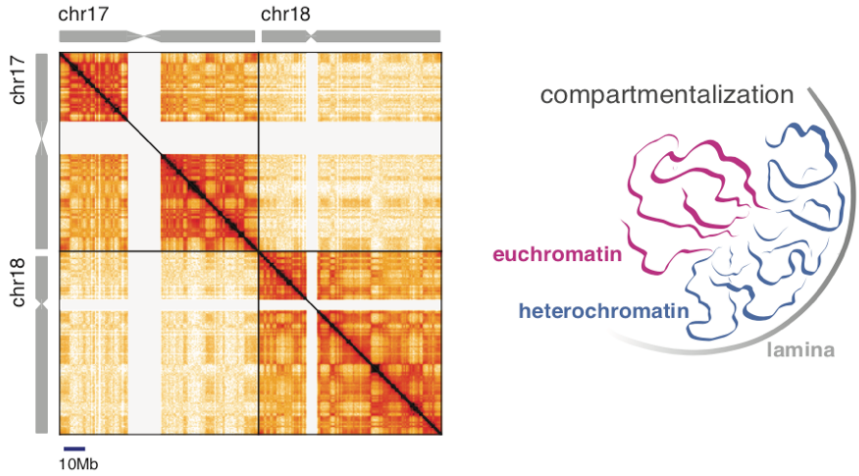


Figure 1.5: Formation of self-interacting domains by compartmentalisation based on chromatin state. *Hi-C contact matrices for chromosomes 17 and 18 in HFFc6 human fibroblast cells (left) show the characteristic checkerboard pattern of compartmentalisation with alternating zones of contact enrichment and depletion spanning the intrachromosomal and interchromosomal areas of the genome-wide Hi-C map. Schematic interpretation (right) of genomic intervals of different type exhibiting differential affinities that cause them to segregate spatially in the nucleus. At the coarsest level, this can be broken down into two compartmental types: A (active; magenta) and B (inactive; blue) chromatin. Figure adapted from Mirny et al., 2019.*

Finally, both the process of transcription and affinity between specific genomic elements have been suggested to play roles in genome organisation. Actively transcribing RNA polymerase II could contribute to chromatin architecture either by direct polymerase-mediated extrusion (Lee et al., 2015) or by influencing the distribution of cohesin along chromosomes (Busslinger et al., 2017; Heinz et al., 2018). Affinity between specific genomic elements is detected in Hi-C contact matrices as an enrichment in contact frequency over the baseline predicted by compartmentalisation. Enrichments have been observed for highly transcribed, exon-rich genes thought to be co-localising at splicing factories (Bonev et al., 2017; Y. Chen et al., 2018; Quinodoz et al., 2018) and for Polycomb-silenced loci that are typically located within euchromatin (Kundu et al., 2017). This affinity-based mechanism therefore has been proposed to be mediated by specific protein-protein interactions at these elements.

The well-characterised, tissue-specific regulatory landscape of the mouse α-globin locus

Even though genome-wide studies of nuclear organisation and its role in gene regulation have been important for uncovering the general principles and potential mechanisms involved (Bonev & Cavalli, 2016), many discoveries regarding the interplay between genome structure and gene expression have been made by analysing individual loci in detail (de Laat & Duboule, 2013). The α-globin genes, which encode the α-subunits of haemoglobin, have long been used as a model locus for studying regulation of gene expression and its associated mechanisms (Higgs & Wood, 2008). Naturally occurring mutations which downregulated α-globin expression and caused α-thalassemia despite intact sequences of the genes themselves, led to the identification of the distal *cis*-regulatory

elements involved in controlling α-globin expression (Hatton et al., 1990; Kioussis et al., 1983). The α-globin genes provided one of the first few examples of long-range regulation of mammalian gene expression (Higgs et al., 1990) and highlighted the importance of distal regulatory elements for controlling gene activity. As a result, the *cis*-regulatory elements of the α-globin locus have been well-characterised. The specific expression pattern of the α-globin genes also makes the locus a convenient experimental model system in which to ask questions about gene regulation. The α-globin genes are expressed in red blood cells, which are easily obtained from patients and model organisms, and their activity is upregulated throughout erythroid differentiation in a predictable manner. Importantly, the product of the α-globin genes is only functionally important at terminal differentiation of erythroid cells. This is unlike other model loci such as the developmental *Hox* genes. Changes in α-globin expression do not affect cell fate decisions, development, or differentiation, which could confound the interpretation of experimental manipulations.

The mouse α-globin locus comprises the duplicated adult α-globin genes *Hba-α1* and *Hba-α2*, as well as the embryonic ζ-globin gene *Hba-x*, and two θ-globin pseudogenes *Hbq1a* and *Hbq1b* (Figure 1.6). This gene cluster resides in a highly conserved region which has a defined segment of synteny of between ~135-155 kb and has been maintained throughout 500 million years of evolution (Flint et al., 2001; Hughes et al., 2005). This region contains four highly conserved enhancer elements (R1-4) (Hughes et al., 2005) and one murine-specific enhancer element (Rm), all of which are erythroid-specific. R1-3 and Rm are located in the introns of the neighbouring housekeeping gene *Nprl3*. R4 lies just downstream of the *Nprl3* promoter in an intergenic region. In erythroid cells, the α-globin genes are active and highly expressed, and their *cis*-regulatory elements have characteristic chromatin marks which are not present in ES cells in which the genes are silent (Figure 1.6).

The active promoters and enhancers are marked by high levels of H3K4me3 and H3K4me1 respectively, and the histone modification H3K27ac, which is associated with active transcription, is found at all of these active elements. The locus is flanked by multiple convergent CTCF binding sites that are bound in both erythroid and ES cells, and subunits of the cohesin complex co-localise at these sites (Figure 1.6).

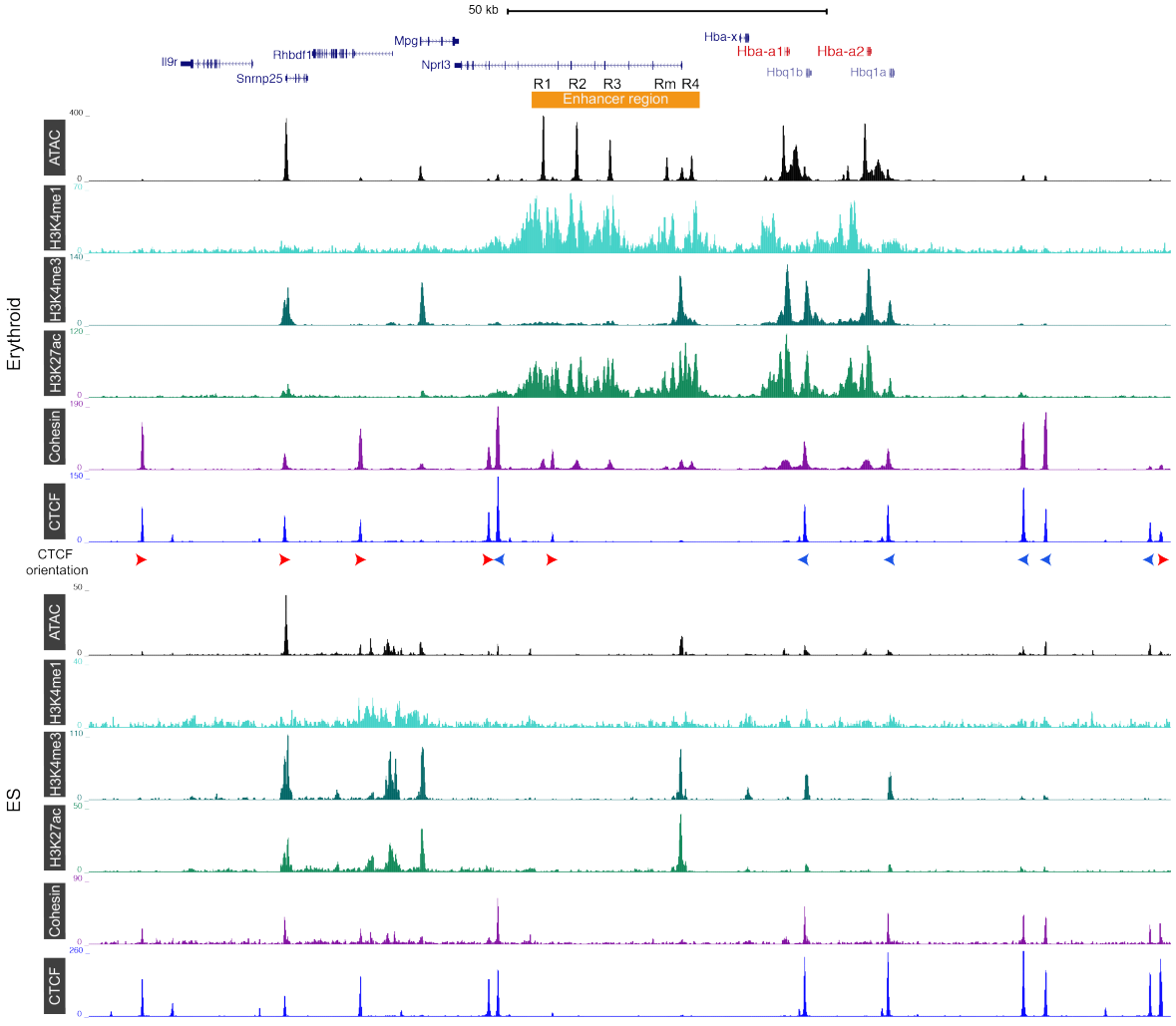


Figure 1.6: The mouse α-globin locus. *Overview of the mouse α-globin locus on chromosome 11, including characteristic histone modifications and occupancy of transcription factors in primary erythroid (top) cells isolated from wild-type (WT) mice and ES (bottom) cells. From top to bottom, tracks show profiles for ATAC-seq and ChIP-seq for H3K4me1, H3K4me3, H3K27ac, cohesin (Rad21 subunit in erythroid cells; Smc1 subunit in ES cells), and CTCF, with arrows indicating the orientation of CTCF motifs under peaks (red for forward; blue for reverse). Profiles show*

normalised (RPKM) and averaged data from differing numbers of biological replicates: three replicates for erythroid ATAC, H3K4me3 and CTCF (Hanssen et al., 2017), two replicates for erythroid Rad21 (Hanssen et al., 2017), one replicate for erythroid H3K4me1 and H3K27ac (Kowalczyk et al., 2012), three replicates for ES ATAC (Simon et al., 2017) and CTCF (Stadler et al., 2011), two replicates for ES H3K27ac (Wamstad et al., 2012) and Smc1 (Kagey et al., 2010), and on replicate for ES H3K4me1 (Stadler et al., 2011) and H3K4me3 (ENCODE). The five enhancer elements are labelled R1-4 and Rm and the enhancer region has been highlighted with an orange bar. Gene annotation from Refseq with the α-globin genes highlighted in red.

The activity of the *cis*-regulatory elements of the α-globin locus is controlled by several erythroid-specific transcription factors. These include GATA-binding factor 1 (Gata1), Stem cell leukaemia factor (Scl, also known as Tal1), Nuclear factor erythroid 2 (Nfe2), and Krüppel-like factor 1 (Klf1) (De Gobbi et al., 2007). It has been shown that binding of these transcription factors to the α-globin enhancers is accompanied by binding of the general transcriptional co-factor Mediator (Hay et al., 2016) and recruitment of RNA polymerase II (Vernimmen et al., 2007) (Figure 1.7). The enhancers have been further characterised *in vivo* by studying the effects of engineered deletions of these elements. This revealed that the enhancers behave in an additive manner to upregulate the expression of α-globin but differ markedly in their individual contribution to transcriptional output (Hay et al., 2016).

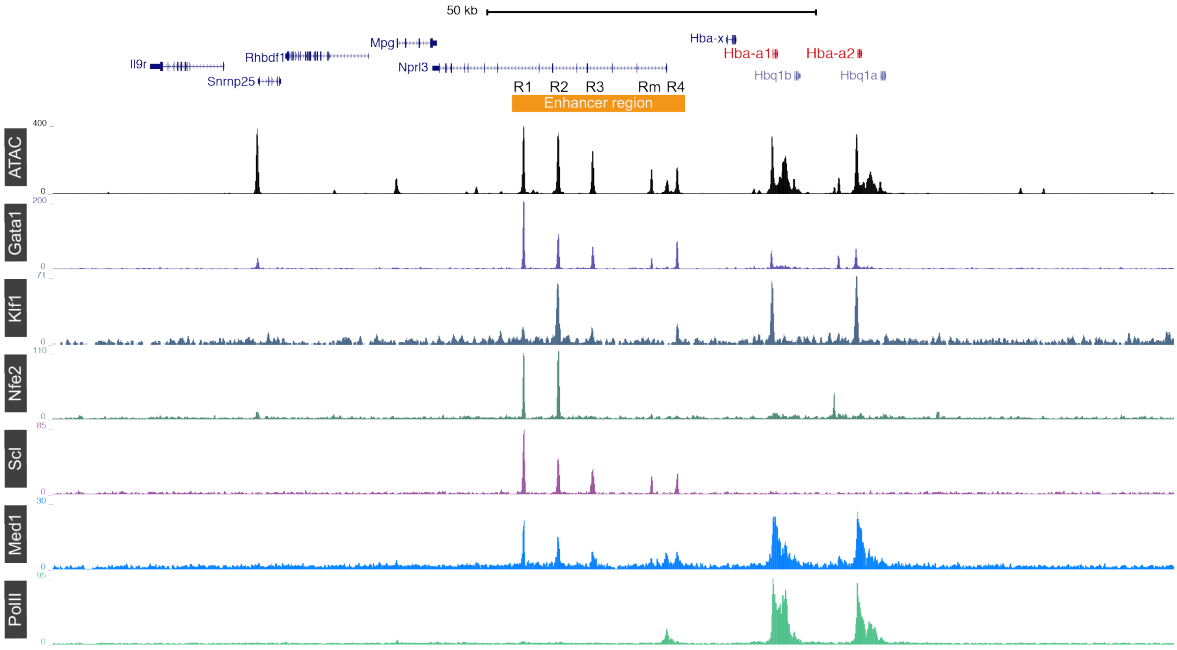


Figure 1.7: Transcription factor and RNA polymerase II occupancy at the active mouse α-globin locus. *From top to bottom, tracks show profiles for ATAC-seq and ChIP-seq in primary erythroid cells isolated from WT mice for erythroid transcription factors Gata1, Klf1, Nfe2, Scl, transcriptional co-activator Mediator (Med1 subunit), and RNA polymerase II (PolII) for the α-globin locus on chromosome 11. Profiles show normalised (RPKM) and averaged data from differing numbers of biological replicates: three replicates for ATAC (Hanssen et al., 2017) and Scl (Pimkin et al., 2014), two replicates for Gata1 (Martin Larke, WIMM; unpublished) and PolII (Hanssen et al., 2017), one replicate for Klf1 (Tallack et al., 2010), Nfe2 (Hughes et al., 2014), and Med1 (Hay et al., 2016). The five enhancer elements are labelled R1-4 and Rm and the enhancer region has been highlighted with an orange bar. Gene annotation from Refseq with the α-globin genes highlighted in red.*

In addition to the extensive functional annotation of the *cis*-regulatory elements of the mouse α-globin locus, the structural organisation of the locus has been well-characterised using 3C-based methods. NG Capture-C (hereafter referred to as Capture-C) is a 3C technique that can generate genome-wide interaction profiles from thousands of viewpoints across multiple 3C libraries in a single assay (Davies et al., 2016; Hughes et al., 2014). It is a 'many vs all' approach in that all interactions with the viewpoints of interest in the genome are analysed

and is one of the most sensitive and highest resolution assays currently available for mammalian genomes (Davies et al., 2017). We have used this technique to thoroughly characterise the regulatory landscape of the α-globin locus by capturing from the viewpoints of the *cis*-regulatory elements (Figure 1.8) (Brown et al., 2018; Davies et al., 2016; Hanssen et al., 2017; Oudelaar & Harrold et al., 2019). In erythroid cells, the promoters of *Hba-α1/2* and their five enhancer elements form a ~70 kb tissue-specific regulatory domain which is primarily self-interacting and excludes the flanking regions. The multiple convergent CTCF binding sites around the locus mark the boundaries of the regulatory domain; they form diffuse interactions with each other and avoid interacting with the active elements within the domain. Further characterisation of the regulatory landscape of this locus has been completed using another 'many vs all' 3C method: Tri-C. Tri-C can detect multi-way chromatin interactions at single alleles from specific viewpoints to determine if multiple *cis*-regulatory elements are interacting simultaneously (Oudelaar et al., 2018). From viewpoints within the mouse α-globin regulatory domain, a higher-order hub-like structure was observed which involved simultaneous interactions between the multiple enhancers and promoters in erythroid cells. Tri-C performed from the viewpoint of HS-39 (one of the upstream flanking CTCF sites) in erythroid cells showed multi-way interactions preferentially formed with the region located downstream of the α-globin regulatory domain containing many CTCF binding sites, which is consistent with the pair-wise interaction data described above. The observed diffuse interactions across the entire region suggest that CTCF binding elements do not form higher-order hub-like structures in which multiple CTCF-binding sites interact simultaneously, but instead participate in dynamic interactions. These observations demonstrate that the α-globin locus provides a good model for the study of regulatory domains and their role in gene regulation; however, it still remains unclear

exactly how this domain forms, especially in a tissue-specific manner, and whether the *cis*-regulatory elements play a role in domain formation.

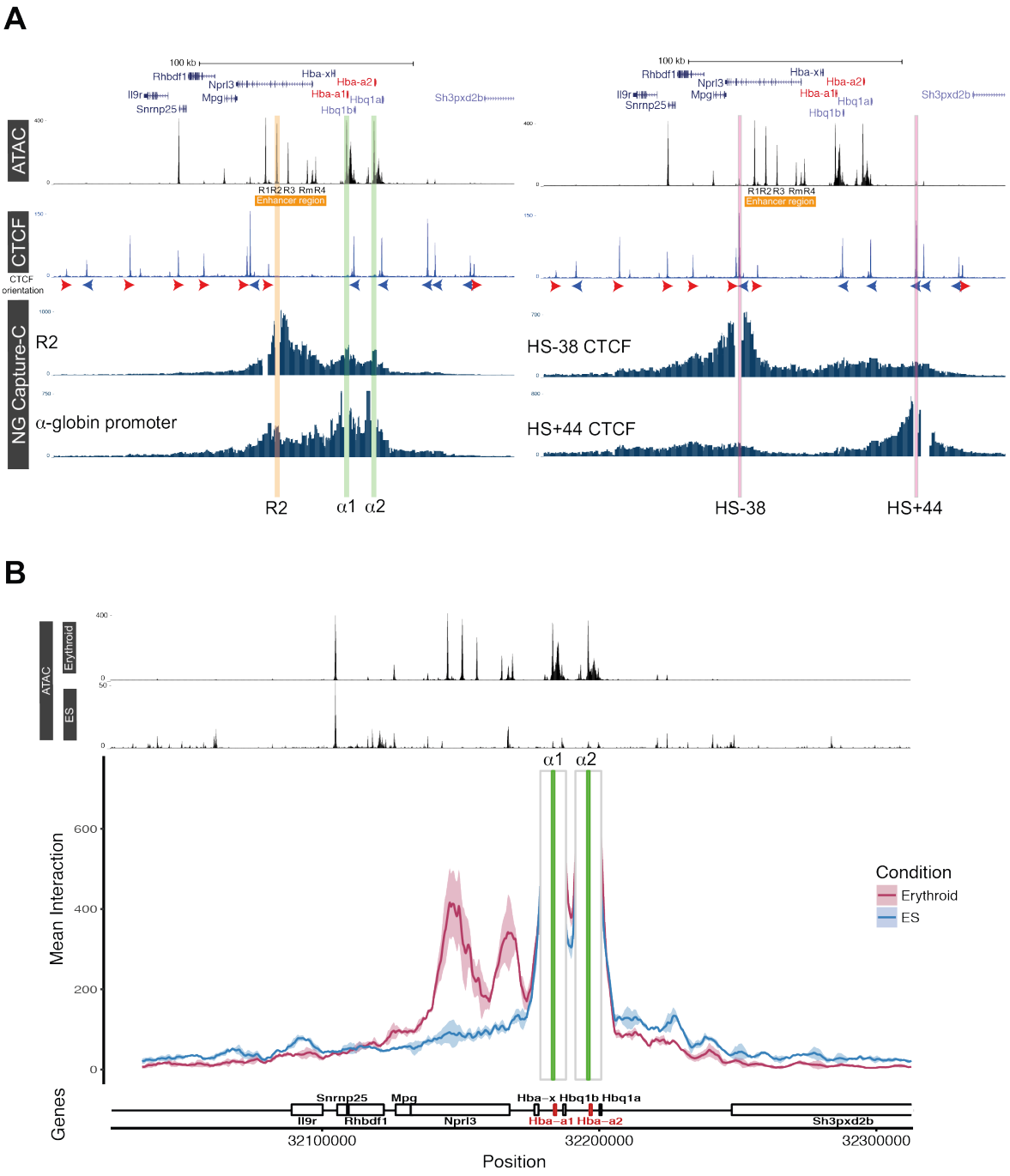


Figure 1.8: Chromatin interactions profiles at the mouse α-globin locus assayed by NG Capture-C. A: *Interaction profiles from viewpoints within (left) and outside (right) of the α-globin regulatory domain on chromosome 11 in primary erythroid cells isolated from WT mice. Top tracks show profiles for ATAC-seq and CTCF ChIP-seq. Profiles show normalised (RPKM) and averaged*

data from three biological replicates (Hanssen et al., 2017). The orientation of CTCF motifs is shown under peaks by red (forward) and blue (reverse) arrows and the five enhancer elements are labelled R1-4 and Rm and the enhancer region has been highlighted with an orange bar. Left: NG Capture-C interaction profiles of the α-globin locus from the viewpoints of R2 (orange) and the α-globin promoters (green), with a 1 kb exclusion zone around the viewpoint, in WT primary erythroid cells. Right: NG Capture-C interaction profiles of the α-globin locus from the viewpoints of the flanking CTCF binding sites (pink) HS-38 (upstream) and HS+44 (downstream), with a 1 kb exclusion zone around the viewpoint, in WT primary erythroid cells The profiles represent normalised and averaged unique interactions from three biological replicates. NG Capture-C data from (Brown et al., 2018; Davies et al., 2016; Hanssen et al., 2017; Oudelaar & Harrold et al., 2019). Gene annotation from Refseq with the α-globin genes highlighted in red. ***B:*** *Interaction profiles from the viewpoints of the α-globin promoters (green) on chromosome 11 in primary erythroid cells isolated from WT mice and ES cells showing that the α-globin regulatory domain (shown in A) is tissue-specific. Top tracks show profiles for ATAC-seq in primary erythroid cells isolated from WT mice and ES cells. Profiles show normalised (RPKM) and averaged data from three biological replicates (Hanssen et al., 2017; Simon et al., 2017). NG Capture-C interaction profiles of the α-globin locus from the viewpoints of the α-globin promoters (green), with 1 kb exclusion zone around the viewpoint, in erythroid (red) and ES (blue) cells. The profiles represent normalised and averaged unique interactions from three biological replicates, with the halo representing the standard deviation of a sliding 3 kb window. NG Capture-C data generated by Damien Downes, WIMM (unpublished). This method of displaying Capture-C interaction profiles is preferred to that shown in A when comparing different tissue-types or WT vs mutant. Genes and genomic position below interaction profiles with α-globin genes highlighted in red.*

Aims and thesis outline

Processes that create, regulate, and maintain regulatory domains are not well understood, therefore this thesis aims to address these questions by building on previous studies at the α-globin locus to investigate the roles *cis*-regulatory elements (CTCF-bound elements and active enhancers and promoters) play in generating and defining regulatory domains.

To this aim, I will present the characterisation of mouse models harbouring genetic perturbations of CTCF binding sites around the α-globin locus to assess their impact on the locus' structural conformation and gene regulation. Next, I will describe how I devised a strategy to determine which, if any, of the α-globin locus *cis*-regulatory elements are sufficient for the formation of a regulatory domain, by inserting them into neutral regions of the mouse genome to construct artificial loci. Finally, I will present the characterisation of various permutations of artificial loci to assess whether it was possible to build a regulatory domain from the bottom-up, and to determine the roles each element plays alone and in combination in the formation of functional regulatory domains.

Chapter 2: Dissecting the role of CTCF boundary elements at the mouse α-globin locus

Introduction

The zinc-finger transcription factor CTCF is thought to play an important role in the organisation of the genome across different scales. Enhancer-blocking reporter assays first identified CTCF binding sites as key components of vertebrate insulators (Bell et al., 1999; Chung et al., 1997). Other studies have since described the roles of CTCF to include acting as boundaries to chromatin modifications (Cuddapah et al., 2009; Handoko et al., 2011; Weth et al., 2014) and facilitating interactions between regulatory elements (Hirayama et al., 2012; Liu et al., 2011). TAD boundaries are enriched for CTCF binding motifs (Dixon et al., 2012) and high-resolution Hi-C data have shown CTCF binding at the base of chromatin loops (Rao et al., 2014). CTCF binding motifs at these sites predominantly (>90%) occurred in a convergent orientation (Gómez-Marín et al., 2015; Rao et al., 2014; Sanborn et al., 2015; Vietri Rudan et al., 2015), and convergent CTCF binding sites contribute to specific chromatin structures within TADs (de Wit et al., 2015; Guo et al., 2015).

The precise roles of CTCF in the formation of regulatory domains remain unclear, however previous studies have shown that perturbating specific CTCF binding sites resulted in altered chromatin interactions/structure and mis-regulation of gene expression (Dowen et al., 2014; Flavahan et al., 2016; Hanssen et al., 2017; Hnisz et al., 2016; Narendra et al., 2015; 2016). In addition, studies exploring the orientation of CTCF binding sites in the organisation of chromatin domains have found that inversions of CTCF binding sites alter chromatin

interactions (de Wit et al., 2015; Guo et al., 2015; Lupiáñez et al., 2015; Sanborn et al., 2015). Depletion of CTCF has been used to assess the effects of global loss of CTCF on chromatin architecture. Such studies found that CTCF depletion resulted in a complete loss of TADs across the genome but did not lead to widespread mis-regulation of gene expression (Hyle et al., 2019; Nora et al., 2017; Wutz et al., 2017). Moreover, the removal of CTCF did not alter the relationship between contact frequency and linear genomic distance between loci throughout the genome, indicating that there was a similar level of chromatin packaging despite the loss of CTCF.

As described in Chapter 1, the mouse α-globin locus provides a good model to study the structure and function of a mammalian regulatory domain. The α-globin genes and their five enhancer elements form a ~70 kb compartmentalised, tissue-specific regulatory domain in erythroid cells, which is flanked by multiple, largely convergent CTCF binding sites (Introduction: Figure 1.8). This domain is primarily self-interacting and was found to form a higher-order hub-like structure, characterised by strong simultaneous interactions between the multiple enhancers and the α-globin promoters which excluded the flanking regions of chromatin bound by CTCF (Davies et al., 2016; Oudelaar et al., 2018). The flanking CTCF binding sites form diffuse interactions with each other in erythroid cells and appear to define the boundaries of the α-globin regulatory domain. The 5' functional boundary of the domain has been experimentally determined by mutation of the CTCF binding sites HS-38 and HS-39 (Hanssen et al., 2017). When these sites were deleted, there was an expansion of the α-globin regulatory domain to include the upstream *Mpg*, *Rhbdf1*, and *Snrnp25* genes which resulted in their upregulation in erythroid cells. This boundary delimits enhancer interactions and therefore contributes to tissue-specific regulation of gene expression. However, removal of the CTCF boundary did not affect the interactions between the enhancers and the α-globin

genes, indicating that CTCF may not directly mediate these interactions, but rather constrains them within the chromatin domain.

In this chapter, I have investigated the role of CTCF binding sites downstream and within the α-globin regulatory domain to further our understanding of the role of such elements in the formation of chromatin domains. I analysed primary erythroid cells isolated from mice harbouring mutations of four CTCF binding sites (θ1, θ2, HS+44, and HS+48; Figure 2.1) individually and in informative combinations to map the key boundary elements at the 3' region of the domain. In addition, to address the question of whether CTCF plays a role in mediating enhancer-promoter interactions in the context of a chromatin domain, as previously suggested (Hirayama et al., 2012; Liu et al., 2011), I have re-analysed a mouse model harbouring deletions of the HS-38 and HS-39 CTCF boundaries (D3839; Hanssen et al., 2017).

Results

Deletion of CTCF binding sites situated within and downstream of the mouse α-globin regulatory domain

Using designs and reagents from Lars L. P. Hanssen (WIMM) and in collaboration with Benjamin Davies (WCHG), we targeted a total of four CTCF binding sites within the α-globin locus: HS+44 and HS+48, which are situated downstream of the erythroid-specific α-globin regulatory domain, and θ1 and θ2, which are situated within the regulatory domain. The CTCF binding site HS+44 is located ~25 kb downstream of the *Hba-α2* promoter and the HS+48 site is a further ~4 kb downstream of HS+44 (Figure 2.1). The θ1 and θ2 CTCF

binding sites are associated with the θ-globin promoters which are positioned downstream of *Hba-α1/2,* ~3.5 kb away from the α-globin promoters (Figure 2.1).

HS+44/+48 were selected for targeting because, in erythroid cells, they display diffuse interactions with the two CTCF-bound sites (HS-38 and -39) which lie upstream of the locus. It appears that these CTCF sites avoid all possible interactions with the active regulatory elements within the α-globin domain, but rather delimit the interactions of the active elements with promoters lying outside of the domain (Introduction: Figure 1.8). The upstream CTCF-bound sites denote the 5' functional boundary of the regulatory domain (Hanssen et al., 2017), therefore we hypothesised that HS+44/+48 sites may have similar architectural and functional roles to those of the upstream sites.

The θ sites were selected for targeting because of their location around the α-globin genes. In human, the 5' α-globin gene (*HBA2*) located proximal to the enhancers, is the more active of the two α-globin genes, producing ~70% of the total mRNA (Liebhaber & Kan, 1981; Liebhaber et al., 1985; Orkin & Goff, 1981). We wanted to investigate whether this was the case for mouse and whether the θ sites might play a role in the differential expression of the duplicated α-globin genes. Mouse models were generated using CRISPR-Cas9 genome editing to target the CTCF binding motif at each site, individually and in informative combinations.

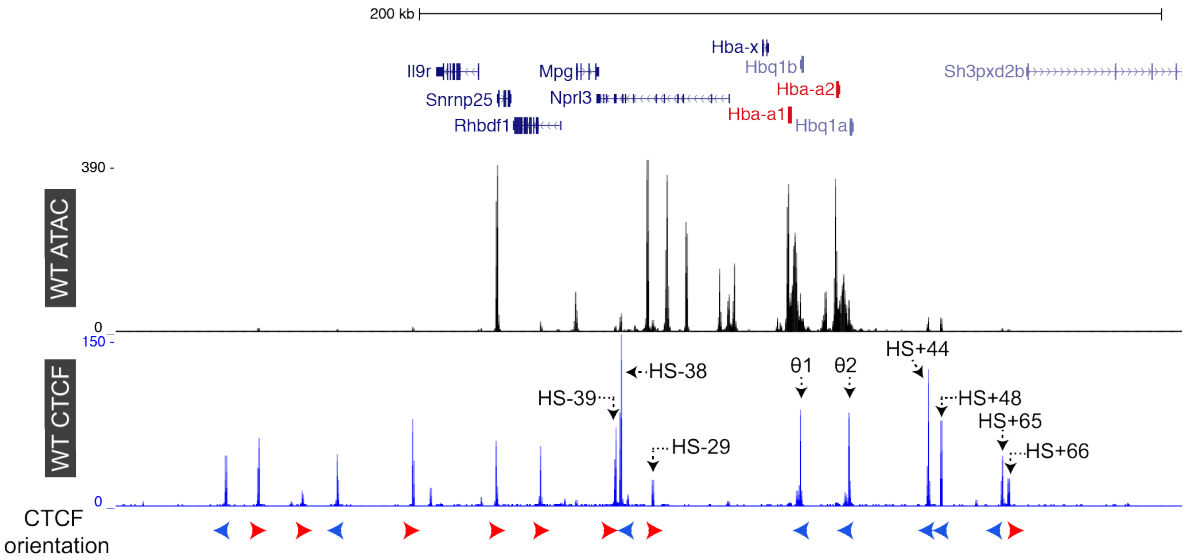


Figure 2.1: CTCF binding sites at the α-globin locus. *Tracks show ATAC-seq and CTCF ChIP-seq profiles in primary erythroid cells isolated from WT mice (Ter119+) for the α-globin locus on chromosome 11. Profiles show normalised (RPKM) and averaged data from three biological replicates (Hanssen et al., 2017). CTCF binding sites in the α-globin locus are labelled with dotted arrows. The orientation of CTCF motifs is shown under peaks by red (forward) and blue (reverse) arrows. Gene annotation from Refseq with the α-globin genes highlighted in red.*

The first model to be generated sequentially mutated the HS+44 site and then the HS+48 site to produce a double knock-out: Δ44-48. As the HS+44 and HS+48 sites are located away from genes and other *cis*-regulatory elements, Hanssen devised a strategy to target each site with two single guide RNAs (sgRNAs) flanking the CTCF binding motif (Materials and Methods: Table 6.1). This produced two separate deletions: targeting of the HS+44 site resulted in a 77 bp deletion, and targeting of the HS+48 site resulted in a 6 bp deletion (Figure 2.2A).

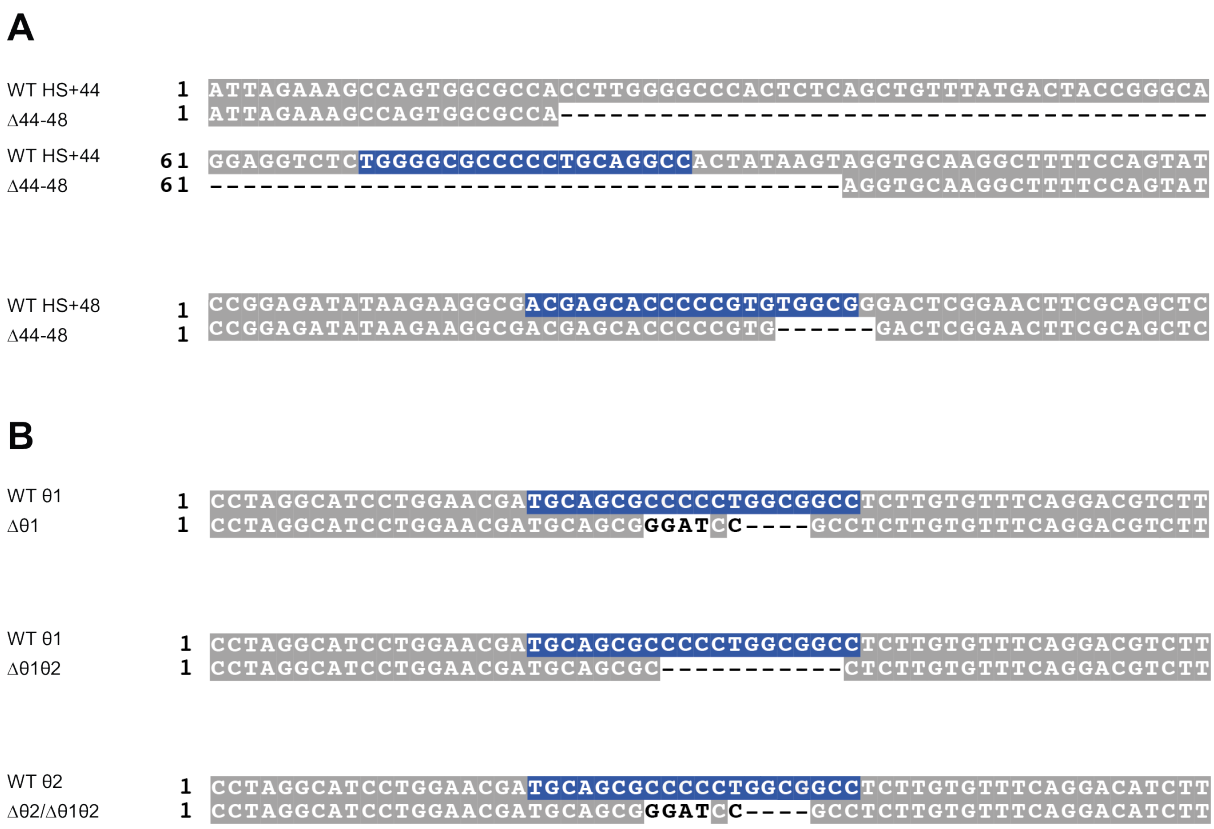


Figure 2.2: Mutated sequences of CTCF binding sites at the α-globin locus. A: *Alignments show WT sequences of HS+44 and HS+48 CTCF binding sites, with the 20 bp core binding motif highlighted in blue (reverse orientation), and deleted sequences of HS+44 and HS+48 CTCF binding sites in Δ44-48 mice.* ***B:*** *Alignments show WT sequences of θ1 and θ2 CTCF binding sites, with the 20 bp core binding motif highlighted in blue (reverse orientation), and mutated sequences of θ1 CTCF binding site in Δθ1 and Δθ1θ2 mice, and mutated sequence of θ2 CTCF binding in Δθ2 and Δθ1θ2 mice.*

The CTCF binding sites at the θ-globin promoters are highly homologous as they are located within a duplicated region of the α-globin locus. Previous attempts at mutating these CTCF binding sites by Hanssen using a sgRNA that targeted both sites simultaneously resulted in the frequent deletion of the 13.5 kb genomic region between θ1 and θ2. Hanssen utilised the presence of a single nucleotide polymorphism (SNP) between the θ1 and θ2 promoters that

disrupted a potential protospacer adjacent motif (PAM) sequence at θ1. This made it possible to mutate the θ2 site with one sgRNA that was complementary to both sites and one sgRNA that was specific to θ2 (Materials and Methods: Table 6.1), and a single-stranded oligodeoxynucleotide (ssODN) donor (Materials and Methods: Table 6.2) that caused homology-directed repair (HDR) of the induced break. However, when I isolated erythroid cells (Ter119+) from homozygous mice (Δθ2) and performed ATAC-seq, this revealed a large deletion between the θ sites that was not detected in screening and genotyping (Figure 2.3). This deletion most likely included one of the α-globin genes, however due to alignment issues resulting from short reads (40 bp paired-end) over the duplicated region, it was not possible to conclusively determine this.

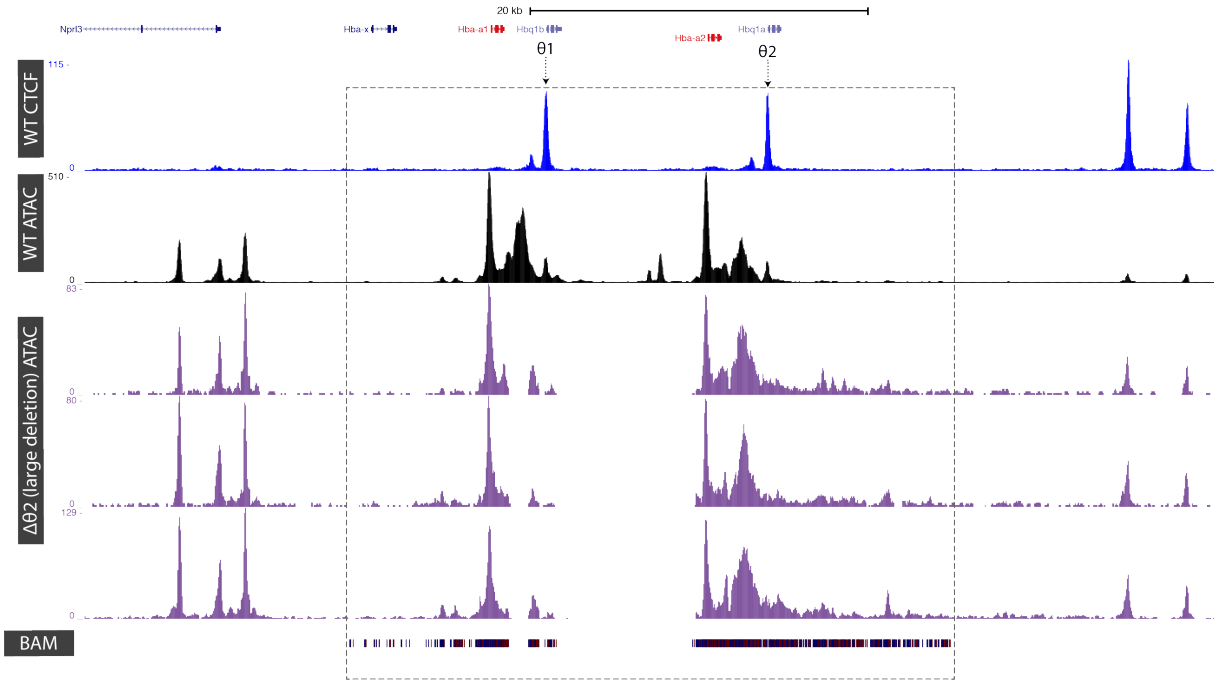


Figure 2.3: The presence of a large deletion in Δθ2 ATAC-seq. *Top tracks show ATAC-seq and CTCF ChIP-seq profiles in primary erythroid cells isolated from WT mice (Ter119+) for the α-globin locus on chromosome 11. WT profiles show normalised (RPKM) and averaged data from three biological replicates (Hanssen et al., 2017). Bottom tracks show ATAC-seq in primary erythroid cells isolated from Δθ2 mice (Ter119+) from three biological replicates. A section of the BAM file of the ATAC-seq in Δθ2 erythroid cells is highlighted by the dashed box and shows no*

reads over the large deletion. Gene annotation from Refseq with the α-globin genes highlighted in red.

Due to mosaicism in the founders of the Δθ2 model, however, Davies was able to generate a double θ1 and θ2 mutant (Δθ1θ2) using a sgRNA targeted to θ1 that was no longer complementary to the mutated θ2 site (Materials and Methods: 6.1). The Δθ1θ2 founder mice exhibited a combination of alleles: the large deletion present in the original Δθ2 model, an 11 bp deletion at the θ1 CTCF binding site, and the intended HDR mutation at the θ2 CTCF binding site. Therefore, not only was the Δθ1θ2 model generated, but through breeding, Davies was able to isolate the HDR θ2 mutation to obtain the Δθ2 model (Figure 2.2B).

In collaboration with Davies, we generated the final model which disrupted the CTCF binding site of θ1 (Δθ1). We used the same sgRNA used to disrupt the site in Δθ1θ2 but in combination with a ssODN donor to promote HDR at this site (Methods and Materials: Table 6.2). We were able to genotype the mutated θ1 sequence (Figure 2.2B) and the wild-type (WT) θ2 sequence showing this method of targeting was successful.

Deletion of downstream CTCF binding sites resulted in an expansion of the α-globin regulatory domain

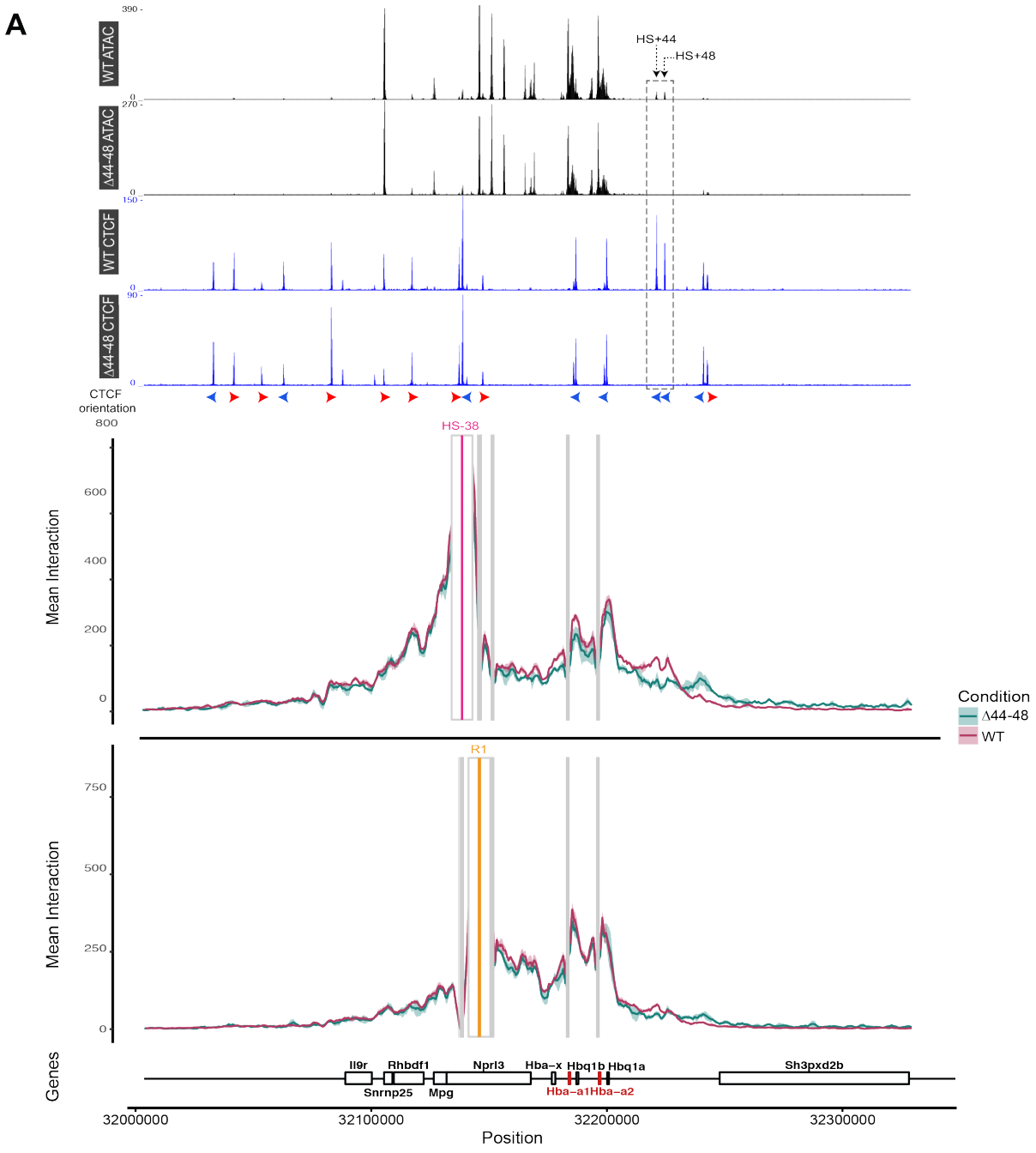
To assess whether the mutations in the HS+44 and HS+48 sites resulted in a loss of CTCF binding, I performed ChIP-seq for CTCF in primary erythroid cells (Ter119+) isolated from the spleens of homozygous Δ44-48 mice (Figure 2.4A). A complete loss of CTCF binding was observed at each of the targeted sites. In addition, I performed ATAC-seq in primary erythroid cells (Ter119+) from homozygous Δ44-48 mice to address whether the genome

editing had altered other tissue-specific regulatory elements in the locus. Chromatin accessibility around the α-globin locus remained unaltered when compared to ATAC-seq performed on WT erythroid cells (Ter119+), other than the loss of peaks at the deleted HS+44 and HS+48 sites (Figure 2.4A).

To investigate whether the removal of the HS+44 and HS+48 CTCF binding sites resulted in changes in local genome topology, I performed Capture-C from viewpoints across the α-globin locus in WT and homozygous Δ44-48 primary erythroid cells (Ter119+). Capture-C data in WT erythroid cells showed diffuse interactions between HS+44/+48 and the two CTCF-bound sites upstream of the α-globin locus (HS-38 and HS-39; Introduction: Figure 1.8) which denote the 5' boundary of the regulatory domain (Hanssen et al., 2017). To see if these interactions were altered in Δ44-48 mice, I analysed Capture-C profiles from the viewpoint of HS-38. When the two CTCF binding sites were absent, the diffuse interactions over the downstream HS+44/+48 sites had shifted to the next pair of downstream CTCF binding sites (HS+65 and HS+66), resulting in an expansion of the regulatory domain (Figure 2.4A). Moreover, analysis from the viewpoint of the R1 enhancer showed that this shift was accompanied by slightly increased interactions between R1 and the region of chromatin downstream of the deleted HS+44/+48 sites, even though this region does not contain any regulatory elements (Figure 2.4A).

The two genes directly downstream of the α-globin locus (*Sh3pxd2b* and *Ubtd2*) are not expressed in WT primary erythroid cells. To assess whether the deletion of the HS+44/+48 sites had an effect on the expression of these downstream genes and α-globin, I performed reverse transcription qPCR analysis. There was no detectable difference in expression of

Sh3pxd2b, Ubtd2 and *Hba-a1/2* in Δ44-48 erythroid cells when compared to WT erythroid cells (Figure 2.4B).


A
WT ATAC
Δ44-48 ATAC
WT CTCF
Δ44-48 CTCF
HS+44
HS+48
CTCF orientation
HS-38
R1
Mean Interaction
Condition
Δ44-48
WT
Genes
Il9r
Rhbdf1
Nprl3
Hba-x
Hbq1b
Hbq1a
Sh3pxd2b
Snrnp25
Mpg
Hba-a1
Hba-a2
Position
32000000
32100000
32200000
32300000

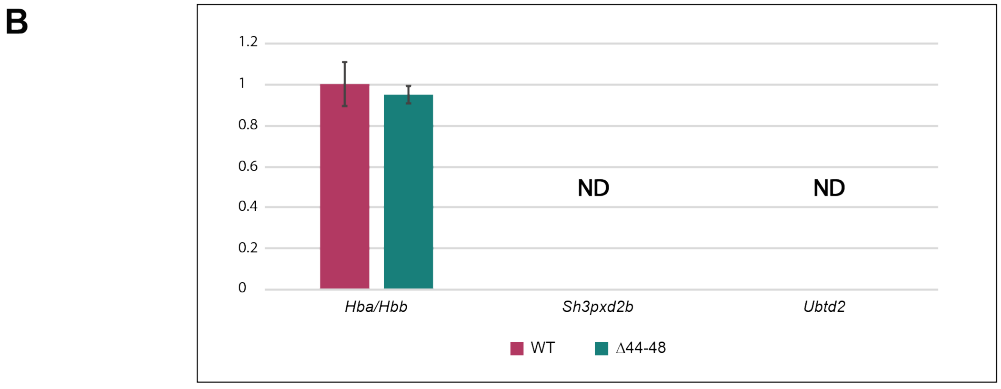

B
ND
ND
Hba/Hbb
Sh3pxd2b
Ubtd2
WT
Δ44-48

Figure 2.4: Effects of the combined deletion of HS+44 and HS+48 on local chromatin accessibility, CTCF binding, genome topology, and gene expression. A: *Top tracks show profiles for ATAC-seq and CTCF ChIP-seq in primary erythroid cells isolated from WT (Hanssen et al., 2017) and Δ44-48 mice (Ter119+) for the α-globin locus on chromosome 11. Profiles show normalised (RPKM) and averaged data from three biological replicates. The orientation of CTCF motifs is shown under peaks by red (forward) and blue (reverse) arrows. NG Capture-C interaction profiles of the α-globin locus from the viewpoints of HS-38 (pink) and R1 (orange), with a 1 kb exclusion zone around the viewpoint, in WT (red) and Δ44-48 (green) Ter119+ primary erythroid cells. The profiles represent normalised and averaged unique interactions from three biological replicates, with the halo representing the standard deviation of a sliding 3 kb window. Grey bars denote other capture points included in this experiment. Genes and genomic position below interaction profiles with α-globin genes highlighted in red.* ***B:*** *Reverse transcription qPCR expression analysis of α- and β-globin mRNA ratio, Sh3pxd2b mRNA, and Ubtd2 mRNA in WT (red) and Δ44-48 (green) Ter119+ erythroid cells, normalised to 18S RNA. Mean and standard deviation of three biological replicates shown. Data normalised to WT. (Figure on previous page.)*

Deletion of CTCF binding sites within the α-globin regulatory domain revealed they do not regulate differential interactions and expression of the α-globin genes

As discussed, the 5' α-globin gene (*HBA2*) located proximal to the enhancers, is the more active of the two α-globin genes in human. To determine if this holds true for mouse, I performed Poly(A)+ RNA-seq on primary erythroid cells (Ter119+) isolated from WT mice (Figure 2.11). The presence of a single SNP in the third exon of *Hba-α1* and *Hba-α2* allowed for variant calling analysis of the reads originating from α-globin transcripts; *Hba-α1* accounted for ~66% of the total α-globin mRNA and *Hba-α2* accounted for ~34%. In addition to this preferential expression of *Hba-α1*, previous work within the group showed that the *Hba-α1* promoter preferentially interacted with the enhancers, particularly the proximal Rm and R4 elements, whereas the *Hba-α2* promoter interacted much more strongly with the chromatin between the two genes (Davies et al., 2016). Interestingly, the promoters of *Hba-α1* and *Hba-α2* had similar interactions with the more distal enhancer R1. Therefore,

we hypothesised that the θ CTCF binding sites (in particular, θ1 as it is situated in between the α-globin genes) could be regulating the differential interactions and expression of the α-globin genes.

To assess whether the HDR mutations in Δθ1 and Δθ2 mice resulted in a specific loss of CTCF binding at θ1 and θ2 respectively, I performed CTCF ChIP-seq in primary erythroid cells (Ter119+) isolated from the spleens of homozygous Δθ1 and Δθ2 mice using long paired-end reads (150 bp compared to 40 bp). Due to a small number of SNPs in the θ1 and θ2 regions, the longer read lengths allowed for specific mapping of these highly similar sequences. A loss of binding was observed at the θ1 site in Δθ1 erythroid cells, whilst binding at the θ2 site was preserved (Figure 2.5). The reciprocal was observed in Δθ2 erythroid cells; a loss of binding at the θ2 site whilst the θ1 site was still bound by CTCF (Figure 2.6). The mutations in the θ1 and θ2 binding sites in Δθ1θ2 mice (11 bp deletion and HDR mutation, respectively) were predicted to cause loss of CTCF binding at both sites, therefore I performed the standard experimental procedure for CTCF ChIP-seq in Δθ1θ2 primary erythroid cells (Ter119+) (40 bp paired-end read sequencing), which confirmed the loss of binding (Figure 2.7). In all three models (Δθ1, Δθ2, and Δθ1θ2), I observed no changes in chromatin accessibility around the α-globin locus in erythroid cells when compared to WT erythroid cells (Figure 2.5, Figure 2.6, and Figure 2.7).

To investigate changes in local genome topology in the mice harbouring mutations at the θ CTCF binding sites, I performed Capture-C from the viewpoints of the α-globin promoters in WT, Δθ1, Δθ2, and Δθ1θ2 primary erythroid cells (Ter119+). Capture-C data in WT erythroid cells showed a pronounced reduction in interactions immediately downstream of *Hba-α2* from any viewpoint inside the regulatory domain (Introduction: Figure 1.8),

showing that this is the 3' edge of the regulatory domain. The Capture-C profiles generated in Δθ2 and Δθ1θ2 erythroid cells showed this 3' edge was unaffected and the overall regulatory domain interaction profiles remained relatively unchanged compared to WT (Figure 2.6 and Figure 2.7), despite the loss of CTCF binding at θ2 and both θ sites, respectively. Unfortunately, the Capture-C profile generated in Δθ1 erythroid cells was of poor quality and displayed low complexity, therefore I could not determine any differences between the mutant and WT (Figure 2.5). Preliminary analysis suggests that this model may harbour SNPs in the DpnII sites around the α-globin locus originating from the genetic backgrounds of the mice used to generate the model. The presence of SNPs in DpnII sites impacts the digestion of 3C libraries and drastically alters the complexity of interaction profiles. Due to time constraints, I did not continue analysis of the Capture-C data generated from the Δθ1 model.

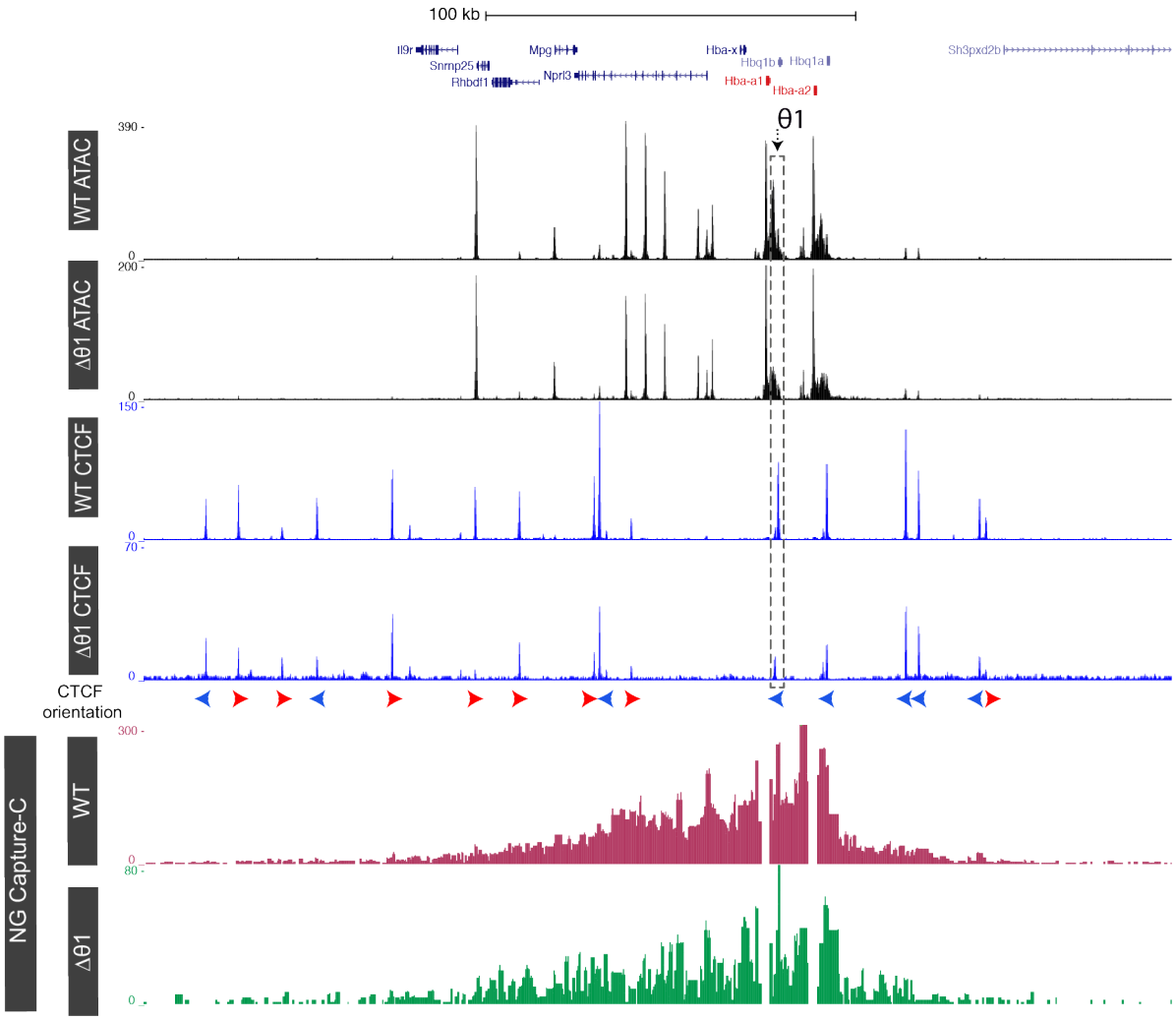


Figure 2.5: Chromatin accessibility, CTCF binding, and quality of NG Capture-C profile in Δθ1 erythroid cells. *Top tracks show profiles for ATAC-seq and CTCF ChIP-seq in primary erythroid cells isolated from WT (Hanssen et al., 2017) and Δθ1 mice (Ter119+) for the α-globin locus on chromosome 11. Profiles show normalised (RPKM) and averaged data from three biological replicates. The orientation of CTCF motifs is shown under peaks by red (forward) and blue (reverse) arrows. NG Capture-C interaction profiles of the α-globin locus from the viewpoints of the α-globin promoters, with a 1 kb exclusion zone around the viewpoints, in WT (red) and Δθ1 (green) Ter119+ primary erythroid cells. The profiles show one representative biological replicate to highlight the low complexity library in Δθ1 erythroid cells. Gene annotation from Refseq with the α-globin genes highlighted in red.*

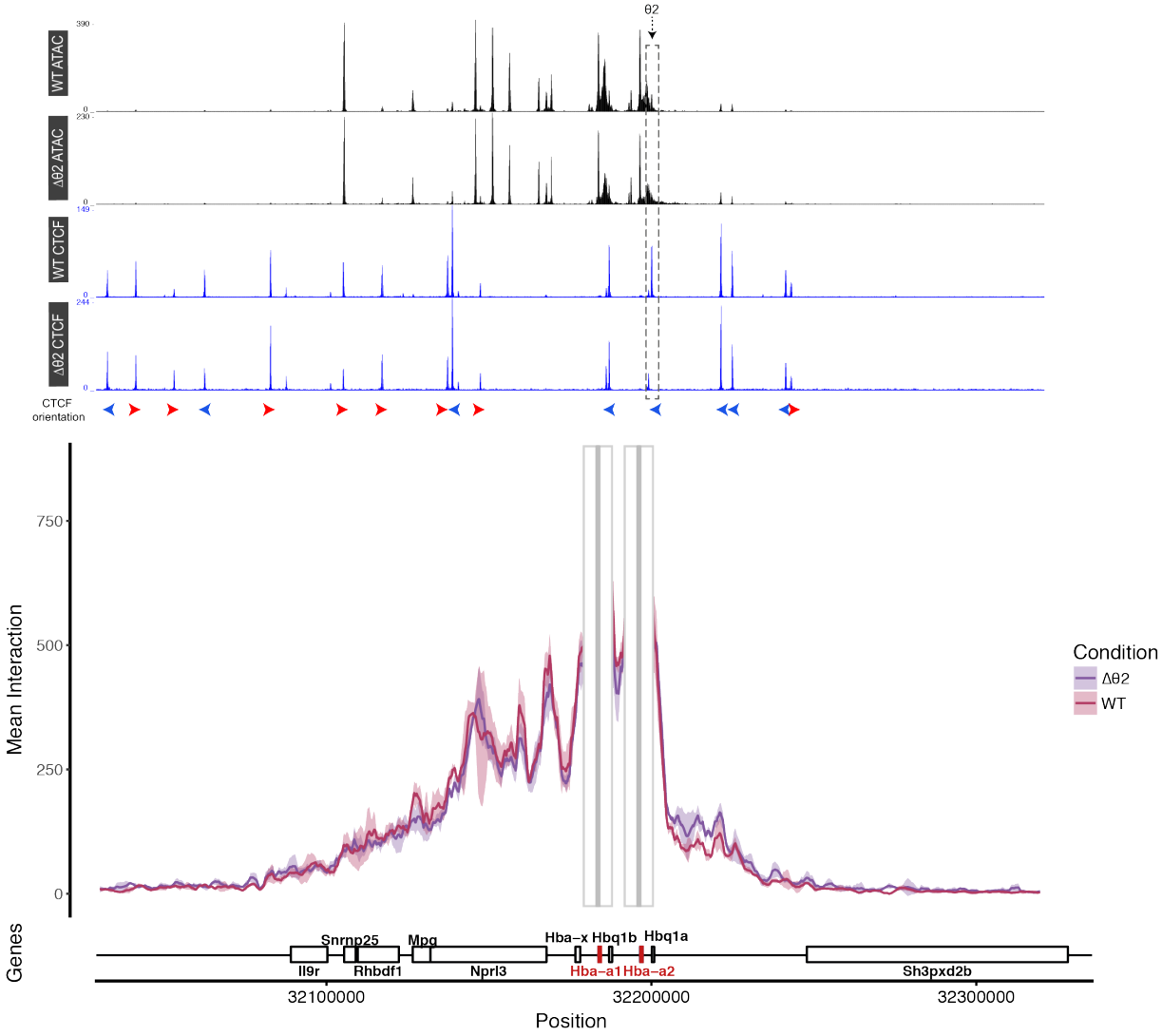


Figure 2.6: Effects of the deletion of θ2 on local chromatin accessibility, CTCF binding, and genome topology. *Top tracks show profiles for ATAC-seq and CTCF ChIP-seq in primary erythroid cells isolated from WT (Hanssen et al., 2017) and Δθ2 mice (Ter119+) for the α-globin locus on chromosome 11. Profiles show normalised (RPKM) and averaged data from three biological replicates. The orientation of CTCF motifs is shown under peaks by red (forward) and blue (reverse) arrows. NG Capture-C interaction profiles of the α-globin locus from the viewpoints of the α-globin promoters (grey), with a 1 kb exclusion zone around the viewpoints, in WT (red) and Δθ2 (purple) Ter119+ primary erythroid cells. The profiles represent normalised and averaged unique interactions from three biological replicates, with the halo representing the standard deviation of a sliding 3 kb window. Genes and genomic position below interaction profiles with α-globin genes highlighted in red.*

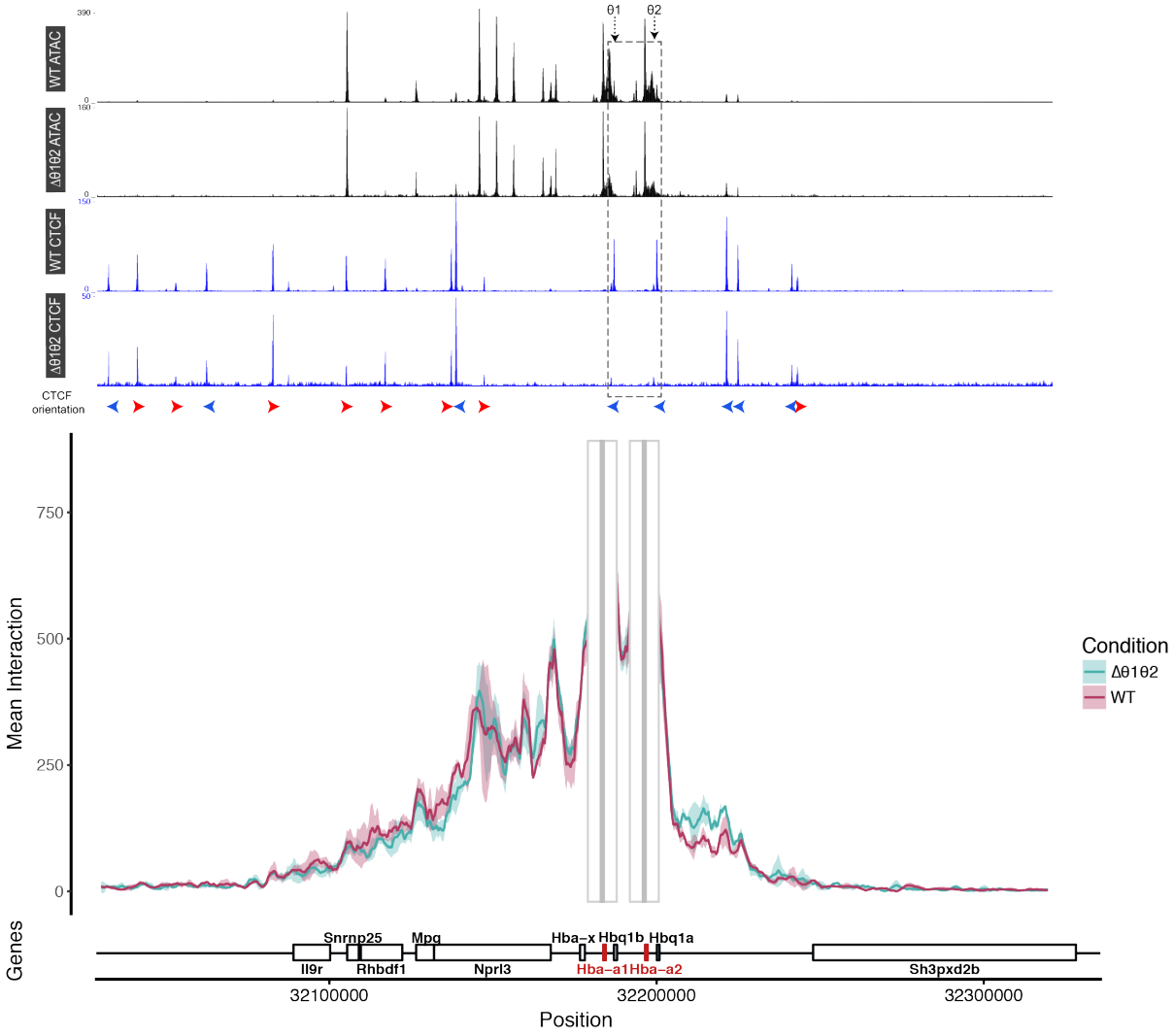


Figure 2.7: Effects of the combined deletion of θ1 and θ2 on local chromatin accessibility, CTCF binding, and genome topology. *Top tracks show profiles for ATAC-seq and CTCF ChIP-seq in primary erythroid cells isolated from WT (Hanssen et al., 2017) and Δθ1θ2 mice (Ter119+) for the α-globin locus on chromosome 11. Profiles show normalised (RPKM) and averaged data from three biological replicates. The orientation of CTCF motifs is shown under peaks by red (forward) and blue (reverse) arrows. NG Capture-C interaction profiles of the α-globin locus from the viewpoints of the α-globin promoters (grey), with a 1 kb exclusion zone around the viewpoints, in WT (red) and Δθ1θ2 (teal) Ter119+ primary erythroid cells. The profiles represent normalised and averaged unique interactions from three biological replicates, with the halo representing the standard deviation of a sliding 3 kb window. Genes and genomic position below interaction profiles with α-globin genes highlighted in red.*

The deep signal provided by Capture-C can be used to generate allele-specific interaction profiles when there is a differing SNP(s) between the two alleles in the capture viewpoint. The SNP(s) is sequenced, and all reporter reads can be assigned to the correct allele. This type of analysis can also be applied to non-allelic SNPs. The paralogous genes *Hba-α1* and *Hba-α2* are almost identical, differing at only a few positions, one of which lies near the 5' promoter capture viewpoint. From the Capture-C data analysed from the viewpoints of the α-globin promoters in WT, Δθ2, and Δθ1θ2 primary erythroid cells, I generated separate interactions profiles for the *Hba-α1* and *Hba-α2* promoters. When erythroid cells from Δθ2 and Δθ1θ2 were compared to WT, the same differential interaction profiles of the α-globin promoters that were previously reported (Davies et al., 2016) were observed (Figure 2.8 and Figure 2.9). To investigate whether the loss of CTCF binding between the α-globin promoters (at θ1) altered the differential interactions, I generated comparisons of the *Hba-α1/2* specific interaction profiles between Δθ2 and Δθ1θ2 erythroid cells (Figure 2.10). The only difference between the two models was the mutation at the θ1 site in Δθ1θ2 mice. There were no observable differences between the interaction profiles despite the loss of CTCF binding between the α-globin genes.

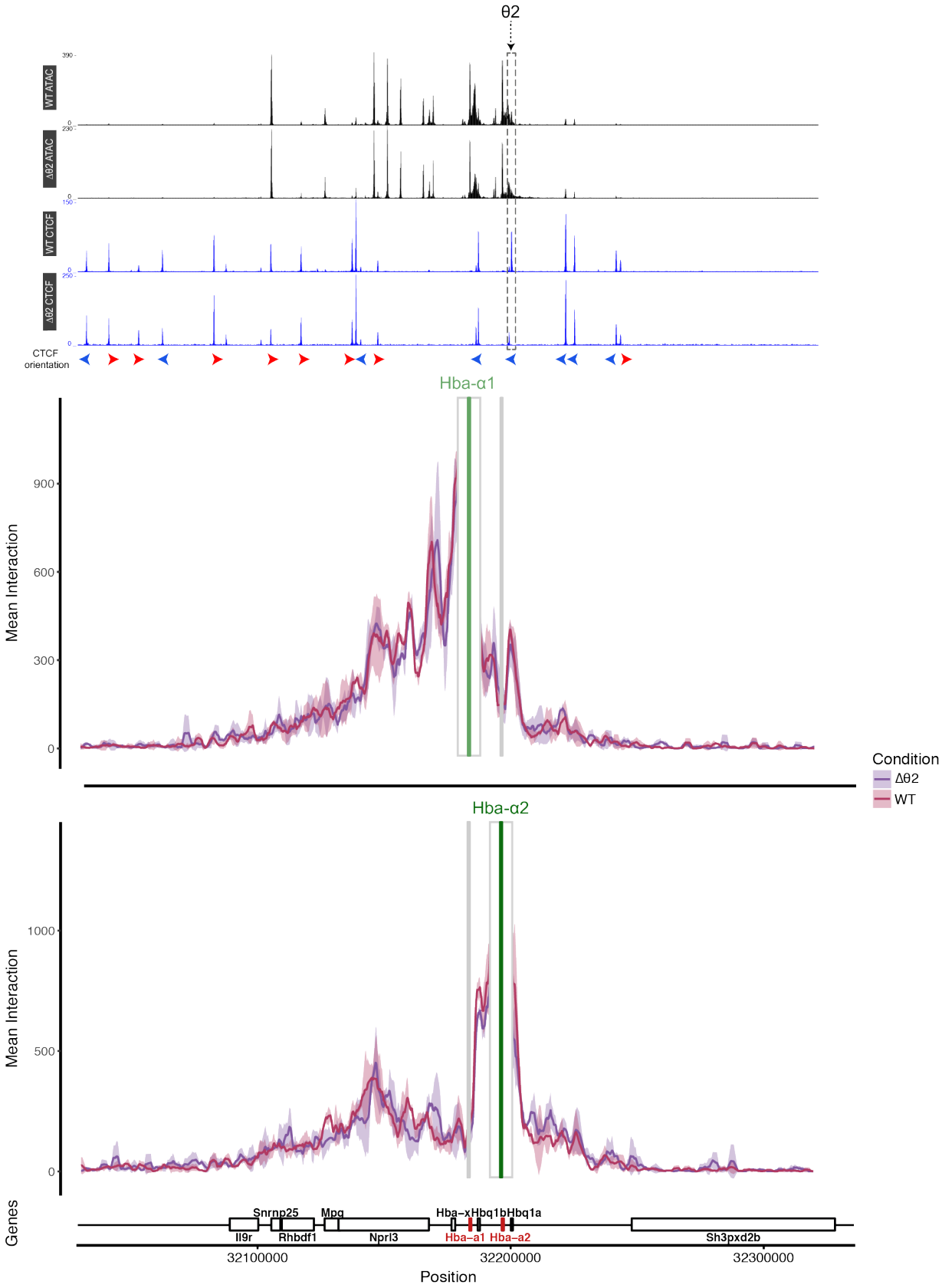

θ2
WT ATAC
Δθ2 ATAC
WT CTCF
Δθ2 CTCF
CTCF orientation
Hba-α1
Mean Interaction
900
600
300
0
Condition
Δθ2
WT
Hba-α2
1000
500
0
Genes
Snrnp25
Mpg
Hba-x
Hbq1b
Hbq1a
Il9r
Rhbdf1
Nprl3
Hba-a1
Hba-a2
Sh3pxd2b
32100000
32200000
32300000
Position

Figure 2.8: Hba-α1/2-specific interaction profiles in Δθ2 erythroid cells. *Top tracks show profiles for ATAC-seq and CTCF ChIP-seq in primary erythroid cells isolated from WT (Hanssen et al., 2017) and Δθ2 mice (Ter119+) for the α-globin locus on chromosome 11. Profiles show normalised (RPKM) and averaged data from three biological replicates. The orientation of CTCF motifs is shown under peaks by red (forward) and blue (reverse) arrows. NG Capture-C interaction profiles of the α-globin locus from the viewpoints of the α-globin promoters (Hba-α1 in light green; Hba-α2 in dark green), with a 1 kb exclusion zone around the viewpoints, in WT (red) and Δθ2 (purple) Ter119+ primary erythroid cells. The profiles represent normalised and averaged unique interactions from three biological replicates, with the halo representing the standard deviation of a sliding 3 kb window. Genes and genomic position below interaction profiles with α-globin genes highlighted in red. (Figure on previous page.)*

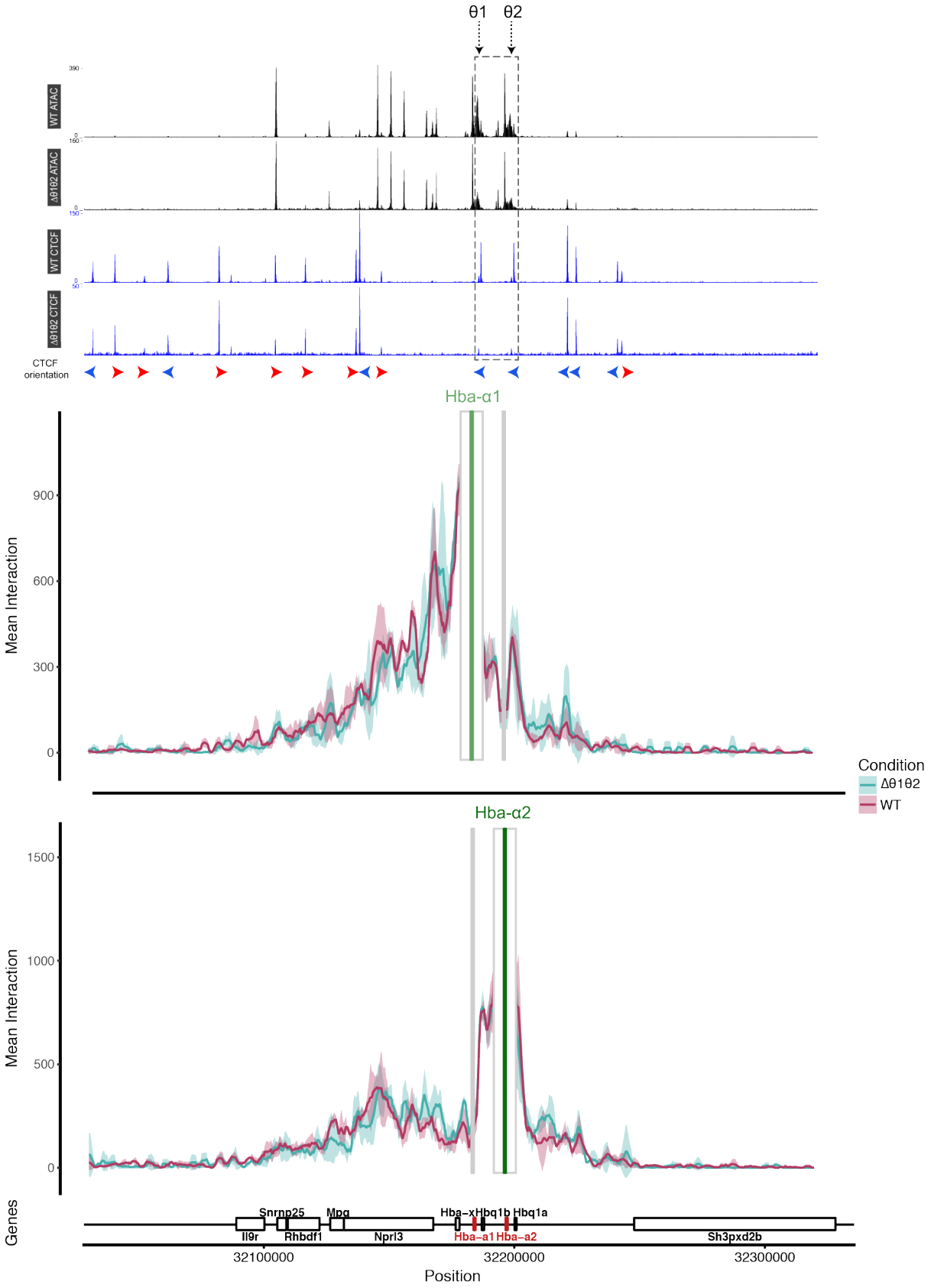

θ1
θ2
WT ATAC
Δθ1θ2 ATAC
WT CTCF
Δθ1θ2 CTCF
CTCF orientation
Hba-α1
Mean Interaction
900
600
300
0
Condition
Δθ1θ2
WT
Hba-α2
1500
1000
500
0
Genes
Snrnp25
Mpg
Hba-x
Hbq1b
Hbq1a
Il9r
Rhbdf1
Nprl3
Hba-a1
Hba-a2
Sh3pxd2b
32100000
32200000
32300000
Position

Figure 2.9: Hba-α1/2-specific interaction profiles in Δθ1θ2 erythroid cells. *Top tracks show profiles for ATAC-seq and CTCF ChIP-seq in primary erythroid cells isolated from WT (Hanssen et al., 2017) and Δθ1θ2 mice (Ter119+) for the α-globin locus on chromosome 11. Profiles show normalised (RPKM) and averaged data from three biological replicates. The orientation of CTCF motifs is shown under peaks by red (forward) and blue (reverse) arrows. NG Capture-C interaction profiles of the α-globin locus from the viewpoints of the α-globin promoters (Hba-α1 in light green; Hba-α2 in dark green), with a 1 kb exclusion zone around the viewpoints, in WT (red) and Δθ1θ2 (teal) Ter119+ primary erythroid cells. The profiles represent normalised and averaged unique interactions from three biological replicates, with the halo representing the standard deviation of a sliding 3 kb window. Genes and genomic position below interaction profiles with α-globin genes highlighted in red. (Figure on previous page.)*

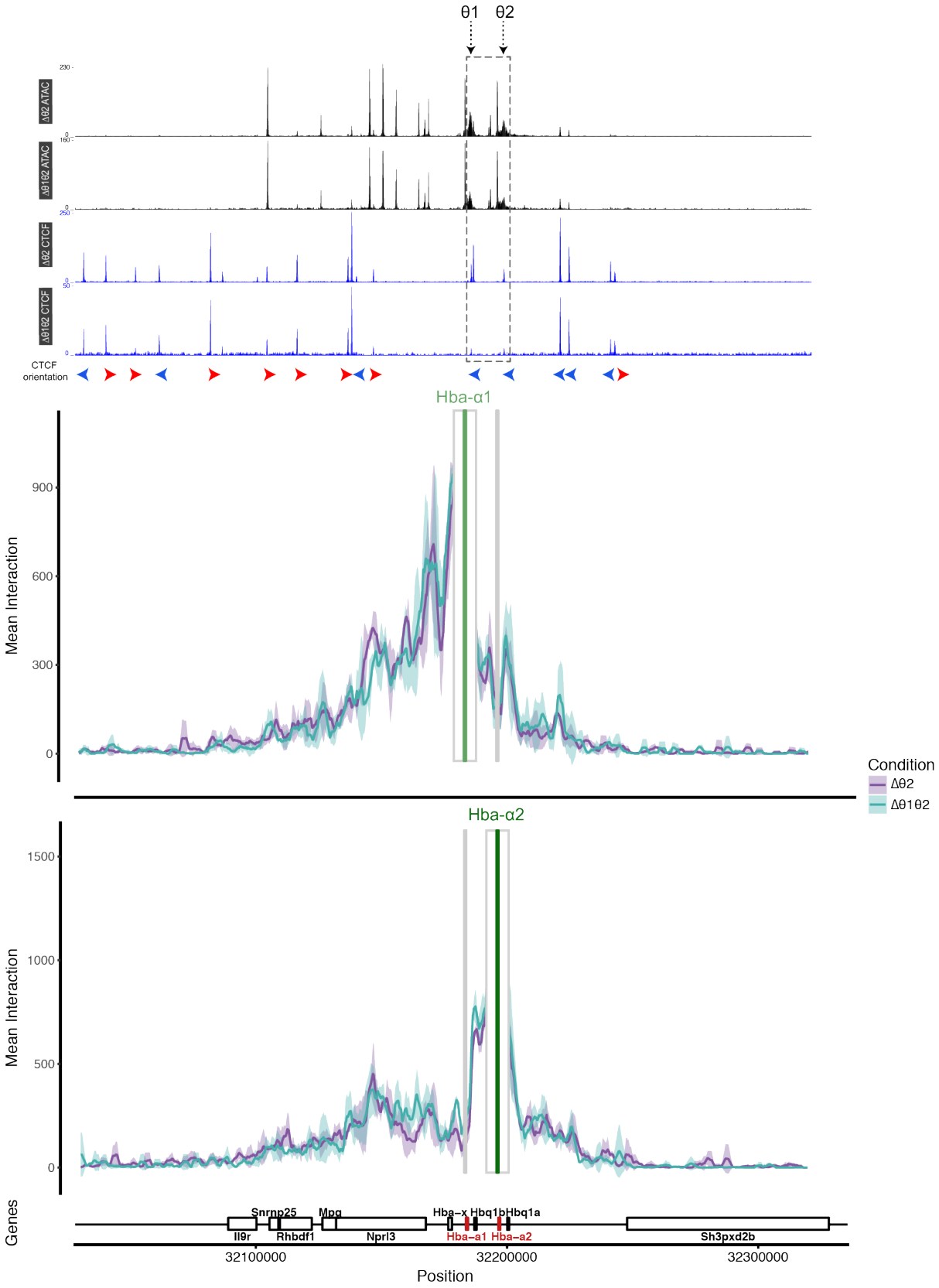

θ1
θ2
Δθ2 ATAC
Δθ1θ2 ATAC
Δθ2 CTCF
Δθ1θ2 CTCF
CTCF orientation
Hba-α1
Mean Interaction
900
600
300
0
Condition
Δθ2
Δθ1θ2
Hba-α2
1500
1000
500
0
Genes
Snrnp25
Mpg
Hba-x
Hbq1b
Hbq1a
Il9r
Rhbdf1
Nprl3
Hba-a1
Hba-a2
Sh3pxd2b
32100000
32200000
32300000
Position

Figure 2.10: Comparison of Hba-α1/2-specific interaction profiles in Δθ2 and Δθ1θ2 erythroid cells. *Top tracks show profiles for ATAC-seq and CTCF ChIP-seq in primary erythroid cells isolated from Δθ2 and Δθ1θ2 mice (Ter119+) for the α-globin locus on chromosome 11. Profiles show normalised (RPKM) and averaged data from three biological replicates. The orientation of CTCF motifs is shown under peaks by red (forward) and blue (reverse) arrows. NG Capture-C interaction profiles of the α-globin locus from the viewpoints of the α-globin promoters (Hba-α1 in light green; Hba-α2 in dark green), with a 1 kb exclusion zone around the viewpoints, in Δθ2 (purple) and Δθ1θ2 (teal) Ter119+ primary erythroid cells. The profiles represent normalised and averaged unique interactions from three biological replicates, with the halo representing the standard deviation of a sliding 3 kb window. Genes and genomic position below interaction profiles with α-globin genes highlighted in red. (Figure on previous page.)*

To interrogate the second part of our hypothesis, I performed Poly(A)+ RNA-seq on primary erythroid cells (Ter119+) isolated from Δθ1, Δθ2, and Δθ1θ2 mice and executed variant calling analysis on the α-globin transcripts. In all three models, the ratio of transcripts produced by *Hba-α1* and *Hba-α2* were similar to that of WT (Figure 2.11), showing that the loss of CTCF binding between and downstream of the α-globin genes did not affect the preferential expression of *Hba-α1*. In addition, there were no significant differences in gene expression profiles around the α-globin locus in the three models compared to WT (data not shown).

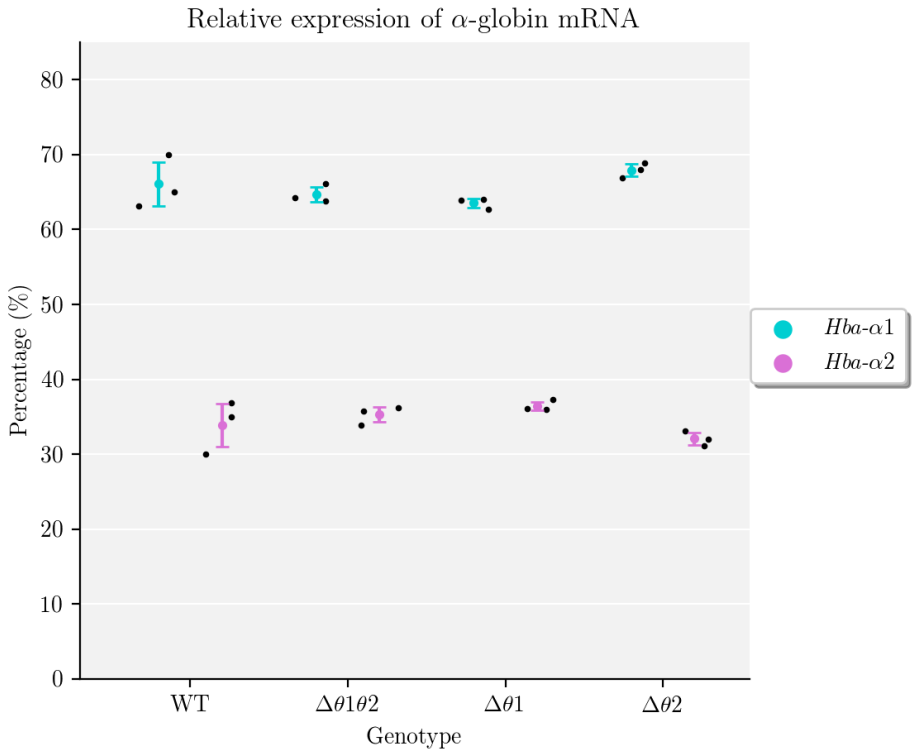


Figure 2.11: Relative expression of Hba-α1/2 mRNA in WT, Δθ1, Δθ2, and Δθ1θ2 erythroid cells. *Variant calling analysis performed on Poly(A)+RNA-seq data from biological triplicates of WT, Δθ1, Δθ2, and Δθ1θ2 erythroid cells (Ter119+) revealed percentage of reads originating from transcripts of Hba-α1 (teal) or Hba-α2 (pink). Mean and standard deviation for each model shown, and each point represents a biological replicate.*

A revised model for promoter competition based on multi-way chromatin analysis of an extended regulatory domain

Previous work in the group found that the α-globin regulatory domain consists of a higher-order structure in which the multiple enhancers interact simultaneously within hub-like complexes in erythroid cells (Oudelaar et al., 2018). The same was found to be true at the β-globin locus in mouse fetal liver cells where the multiple enhancers of the locus control region (LCR) form a hub-like complex (Allahyar et al., 2018). These results imply that multiple enhancers within a regulatory domain/TAD can interact together to act on their target promoter(s). However, it is not clear how specificity between multiple enhancer elements and different promoters within a single regulatory domain/TAD is regulated. Enhancers often exert different effects on what appear to be equally accessible genes within individual TADs. It has been proposed that enhancer-driven transcription from different promoters within a TAD is dependent on distance, orientation or affinity of the enhancers with respect to the specific promoters (Furlong & Levine, 2018; Long et al., 2016). Previous studies have suggested that enhancers may only interact with one accessible promoter at a time. This has led to a model in which the pattern of gene expression within a TAD containing multiple genes is determined by competition between promoters for limited access to shared enhancers. Based on this model, it has been proposed that co-expression of multiple genes regulated by shared enhancers in a single TAD results from rapidly alternating interactions of these genes with the enhancers in a "flip-flop" mechanism (Bartman et al., 2016; Wijgerde et al., 1995). Therefore, we investigated whether enhancer elements and multiple different promoters could interact simultaneously to form higher-order regulatory domains and address questions about promoter competition within regulatory domains by using the previously generated D3839 model (Hanssen et al., 2017).

D3839 erythroid cells show an expansion of the α-globin regulatory domain to include the upstream *Mpg*, *Rhbdf1*, and *Snrnp25* genes which results in their upregulation in erythroid cells (Hanssen et al., 2017). This extended domain enables us to address the mechanism by which the α-globin enhancers interact with multiple accessible gene promoters in a single regulatory domain. To do this, I performed Tri-C from viewpoints in the α-globin locus in WT and D3839 primary erythroid cells (Ter119+). Tri-C allows for the analysis of multi-way chromatin interactions at single alleles. Computational analyses were performed by A. Marieke Oudelaar (WIMM) and multi-way interactions are displayed in contact matrices in which we exclude the viewpoint of interest and plot the frequencies with which two elements interact simultaneously with this viewpoint in a single allele. Essentially this provides a two-dimensional representation of a three-dimensional phenomenon. Preferential, simultaneous interactions are visible as enrichments at the intersections between these elements, whereas mutually exclusive contacts between elements appear as depletions in the matrix.

From the viewpoint of R2, we observed strong, simultaneous R2 interactions with the α-globin promoter and all other α-globin enhancer elements in WT cells which is consistent with previous findings. These interactions were not decreased in the D3839 cells, as would be expected if there was competition between these elements (Figure 2.12). Rather, there was a trend for increased interactions contributing to the α-globin hub in the D3839 cells, although this was not statistically significant (Figure 2.12: green). In addition to the multi-way contacts between the α-globin enhancers and promoters, in D3839 we observed simultaneous interactions between the α-globin enhancer elements and the *Mpg* and *Rhbdf1* promoters (Figure 2.12: purple). Interestingly, the R2 contact matrix also showed simultaneous contacts with both the α-globin promoters and the *Mpg* and *Rhbdf1* promoters

(Figure 2.12: grey). This indicates that all promoters in the extended D3839 regulatory domain interact together in a single hub.

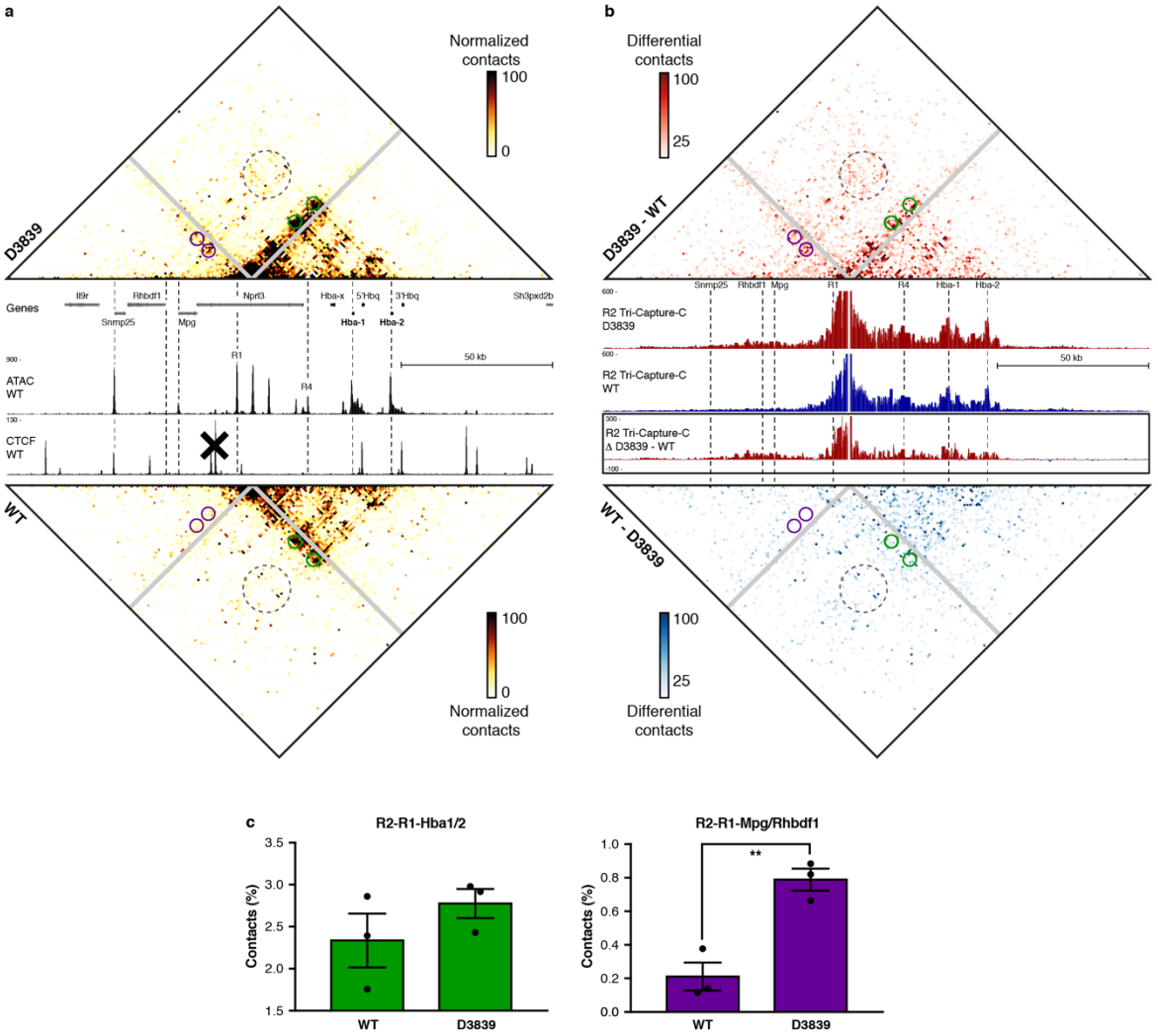


Figure 2.12: Effects of the combined deletion of HS-38 and HS-39 on multi-way chromatin interactions at the α-globin locus from the viewpoint of R2 in erythroid cells. ***A:*** *Tri-C contact matrices showing multi-way chromatin interactions with R2 in D3839 (top) and WT (bottom) erythroid cells (Ter119+). Matrices represent mean numbers of normalized, unique contact counts at 1 kb resolution in 3 replicates with proximity contacts around the R2 viewpoint excluded (grey diagonal). Gene annotation, open chromatin (ATAC) and CTCF occupancy in WT erythroid cells are shown in the middle.* ***B:*** *Tri-C contact matrices showing differential multi-way chromatin interactions with R2 between D3839 and WT erythroid cells (top) and vice versa (bottom). Pair-wise interaction profiles derived from the Tri-C data from the R2 viewpoint (R2 Tri-Capture-C) are shown in the middle (D3839 in red, WT in blue), with a differential profile in the bottom panel.* ***C:***

Quantification of multi-way contacts between R2, R1, and the α-globin promoters (R2-R1-Hba1/2, green, P = 0.29)) and R2, R1, and the Mpg and Rhbdf1 promoters (R2-R1-Mpg/Rhbdf1, purple, P = 0.0055). Quantified contacts are highlighted with corresponding colours in the matrices above. Numbers represent the proportion of these 3-way contacts relative to the total in the matrix and are averages of 3 replicates, with individual data points overlaid as dot plots and the standard error of the mean denoted by the error bar. Figure from Oudelaar & Harrold et al., 2019.

To allow more extensive examination of the simultaneous interactions occurring when the upstream genes interact with the α-globin enhancers in D3839 cells, we generated a Tri-C contact matrix from the viewpoint of the *Mpg* promoter (Figure 2.13). Comparison of multi-way *Mpg* interactions in D3839 and WT cells revealed a strong increase in interactions downstream of *Mpg* after removal of the CTCF boundary. These interactions were strongest proximal to the *Mpg* promoter and reduced in strength beyond the R1 enhancer, which is located close to the HS-29 CTCF binding site. We also observed a clear increase in more distal downstream multi-way interactions, predominantly with the regions containing the α-globin promoters and enhancers. In a dynamic flip-flop model, the *Mpg* promoter would structurally compete with the α-globin promoters for dynamic interactions with the α-globin enhancers. Such mutually exclusive interactions would be reflected as a depletion of the corresponding multi-way interactions in the Tri-C matrix. However, we found preferential interactions between these elements. For example, when *Mpg* interacted with R1, it preferentially interacted with the α-globin promoters (Figure 2.13: green). Similarly, we found enrichment of multi-way interactions between *Mpg*, R1 and the *Rhbdf1* promoter (Figure 2.13: purple). This shows that *Mpg* preferentially interacts with the α-globin enhancers in a complex that contains multiple enhancer elements and promoters. We also found that multi-way interactions between the three promoters were enriched (Figure 2.13:

orange), which further confirmed that there was no structural competition between active promoters for contact with the enhancers within the extended regulatory domain.

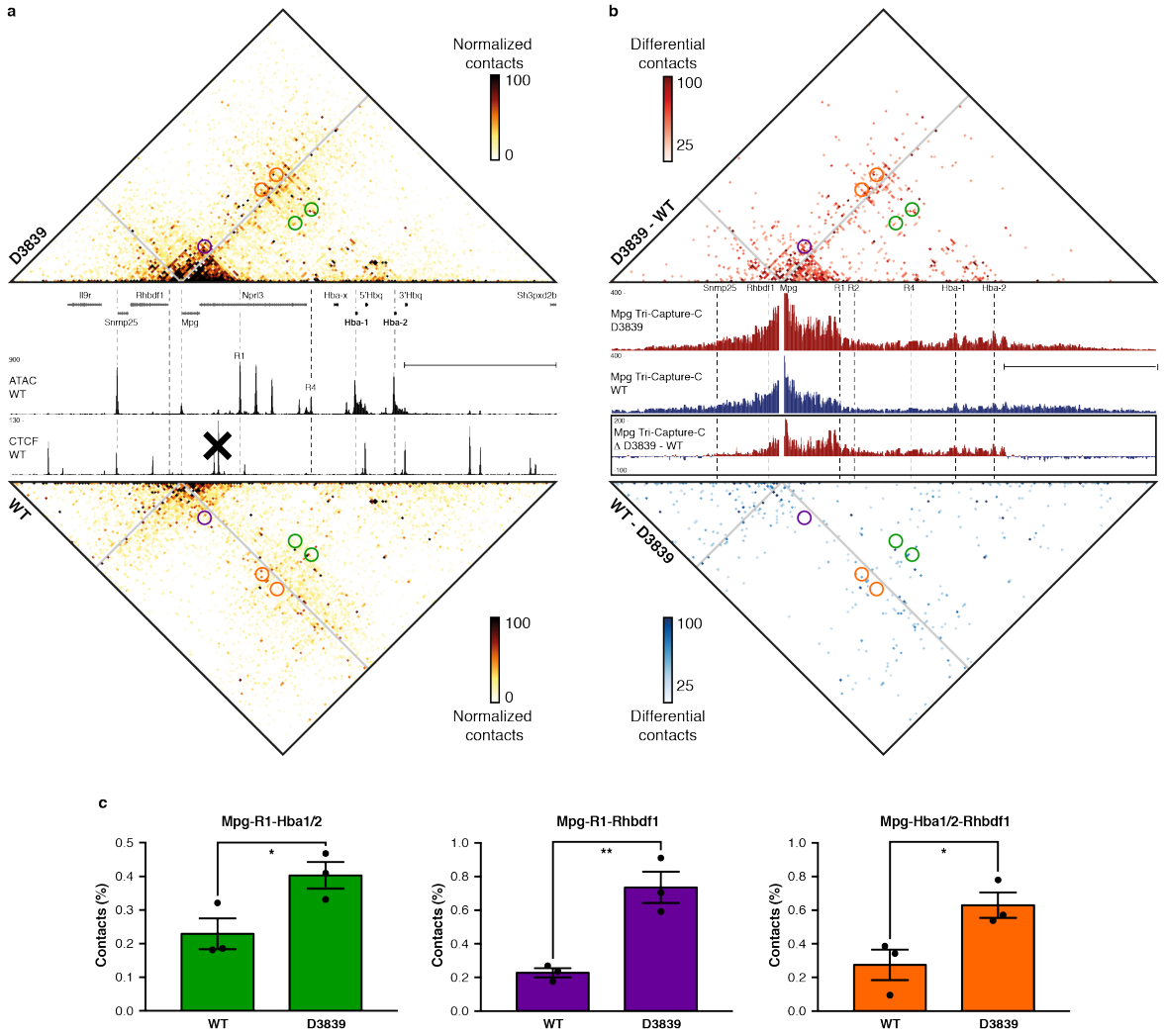


Figure 2.13: Effects of the combined deletion of HS-38 and HS-39 on multi-way chromatin interactions at the α-globin locus from the viewpoint of Mpg in erythroid cells. A: *Tri-C contact matrices showing multi-way chromatin interactions with Mpg in D3839 (top) and WT (bottom) erythroid cells (Ter119+). Matrices represent mean numbers of normalized, unique contact counts at 1 kb resolution in 3 replicates with proximity contacts around the Mpg viewpoint excluded (grey diagonal). Gene annotation, open chromatin (ATAC) and CTCF occupancy in WT erythroid cells are shown in the middle.* ***B:*** *Tri-C contact matrices showing differential multi-way chromatin interactions with Mpg between D3839 and WT erythroid cells (top) and vice versa (bottom). Pair-wise interaction profiles derived from the Tri-C data from the Mpg viewpoint (Mpg Tri-Capture-C) are shown in the middle (D3839 in red, WT in blue), with a differential profile in the bottom panel.*

C: *Quantification of multi-way contacts between Mpg, R1 and the α-globin promoters (Mpg-R1-Hba1/2, green, P = 0.046); Mpg, R1 and the Rhbdf1 promoter (Mpg-R1-Rhbdf1, purple, P = 0.0064); and Mpg, the α-globin promoters and the Rhbdf1 promoter (Mpg-Hba1/2-Rhbdf1, orange, P = 0.040). Quantified contacts are highlighted with corresponding colours in the matrices above. Numbers represent the proportion of these 3-way contacts relative to the total in the matrix and are averages of 3 replicates, with individual data points overlaid as dot plots and the standard error of the mean denoted by the error bar. Figure from Oudelaar & Harrold et al., 2019.*

Discussion

In this chapter, I have described how using CRISPR-Cas9 mediated targeting of CTCF binding motifs we have generated mouse models harbouring mutations in four CTCF binding sites at the α-globin locus (θ1, θ2, HS+44, and HS+48) individually and in informative combinations. Mutations in the CTCF binding sites led to a loss of CTCF binding at all targeted sites. Despite the loss of CTCF binding, analysis of local genome topology in all models showed similar overall interaction profiles at the α-globin locus, with some modest differences, when compared to WT. Additionally, expression analysis from all models showed no significant differences compared to WT. Finally, I have shown that analysis of multi-way chromatin interactions in the previously generated D3839 model reveals that multiple enhancers and promoters within a shared TAD can all interact simultaneously to form a single regulatory hub structure.

The chromatin interaction profiles in Δ44-48 erythroid cells showed an expansion of the α-globin regulatory domain which supports the loop extrusion model for domain formation described in Chapter 1 (Fudenberg et al., 2016). The increased interactions observed between the R1 enhancer and the region of chromatin downstream of the deleted CTCF binding sites suggests that this expansion could occur even without the formation of new

interactions between *cis*-regulatory elements. Therefore, a model in which a loop extruding factor, such as cohesin, is loaded onto chromatin within the α-globin domain and extrudes a loop until it is stalled by boundary elements, such as the next downstream pair of CTCF binding sites (HS+65 and HS+66), could explain these increased interactions. A domain could still form without CTCF binding at HS+44 and HS+48, therefore suggesting that these sites play a role in delimiting active interactions rather than in forming the regulatory domain. However, as the region contains many CTCF binding sites, it is possible that there is functional redundancy; CTCF-mediated structures could be playing a role in facilitating interactions between *cis*-regulatory elements and regulatory domain formation. Only removal of all the CTCF binding sites in and around the α-globin locus could answer this question.

My investigation into the function of the CTCF binding sites at θ1/2 showed that deletion of sites within the active α-globin regulatory domain had no effect on the differential interactions of *Hba-α1/2* or the preferential expression of *Hba-α1*. This suggests that *Hba-α1* could be behaving as a partial boundary through promoter competition, in that *Hba-α2* has reduced enhancer interactions and expression relative to *Hba-α1*. Reported examples of promoter competition have often described situations where an active promoter located between an enhancer and another, more distal promoter causes reduced activity of the distal promoter (Bartman et al., 2016; Cho et al., 2018a; De Gobbi et al., 2006; Wijgerde et al., 1995). It is possible that the proximal, highly transcribed gene forms a barrier to loop extrusion (Brandão et al., 2019), due to accumulation of large amounts of transcriptional machinery and regulatory factors. This could reduce interactions between the more distal promoter and the enhancers and thus decrease expression of the distal gene. Importantly, the 3' edge of the α-globin regulatory domain (immediately downstream of *Hba-α2*) was

preserved in the absence of CTCF binding at θ2. This suggests that the promoter of *Hba-α2* is the 3' boundary of the domain rather than the CTCF site playing this role. Our data provide evidence that actively transcribed genes may also behave as regulatory domain boundaries, in addition to CTCF binding sites, and this is in agreement with studies that found some TAD boundaries are enriched for the promoters of actively transcribed genes, such as housekeeping genes (Dixon et al., 2012).

The analysis of multi-way chromatin interactions in an engineered extended α-globin regulatory domain revealed that multiple gene promoters (globin and non-globin) can simultaneously interact with the five clustered α-globin enhancers at a single allele to form a hub-like regulatory domain. These results demonstrate that the previously reported "flip-flop" model of promoter competition, in which individual gene promoters interact with enhancers in a mutually exclusive manner, is not universally true, and that in this context (where the newly incorporated promoters lie upstream of the enhancers and their target promoters) there is no intrinsic competition between gene promoters for physical access to shared enhancers within a single regulatory domain. This model is supported by recent live imaging experiments in *Drosophila* which showed coordinated bursting of two genes regulated by a single shared enhancer (Fukaya et al., 2016). These findings in combination with the above data (from the deletions of the θ1/2 CTCF binding sites) suggest another underlying mechanism of promoter competition which is not based on mutual exclusivity of enhancer-promoter interactions, but a structural boundary which reduces access of distal promoters to a cooperative hub.

Data availability

The Tri-C data presented in this chapter are published in the following article:

Oudelaar, A. M. *, Harrold, C. L. *, Hanssen, L. L. P., Telenius, J. M., Higgs, D. R., & Hughes, J. R. (2019). A revised model for promoter competition based on multi-way chromatin interactions at the α-globin locus. *Nature Communications*, 10(1), 5412–8. http://doi.org/10.1038/s41467-019-13404-x

* authors contributed equally

Acknowledgements

This was a collaborative project between the WIMM and the WCHG; we worked closely with Benjamin Davies (WCHG) to generate the mouse models described in this chapter. Jacqueline A. Sharpe (WIMM) maintained the mouse colonies and provided the analysis of haematological phenotypes. Matthew E. Gosden (WIMM) helped isolate and process primary erythroid cells obtained from the spleens of phenylhydrazine-treated mice, and with performing ATAC-seq experiments. Rosa Stolper (WIMM) helped perform CTCF ChIP-seq experiments. A. Marieke Oudelaar (WIMM) performed the computational analyses for the Tri-C experiments.

Chapter 3: The Design and Construction of Artificial Loci

Introduction

As described in Chapter 2, deletions of CTCF binding sites within and downstream of the mouse α-globin locus do not cause the collapse of the regulatory domain in erythroid cells. This is in agreement with previous work performed in the group where mice harbouring small deletions in the binding sequences of the two upstream CTCF sites (HS38 and HS39) showed an expansion of the α-globin regulatory domain in erythroid cells which included the immediate upstream genes (*Snrnp25, Mpg and Rhbdf1*) (Hanssen et al., 2017). This suggests that HS-38 and HS-39 CTCF binding sites behave as a boundary that delimits the regulatory domain, but a somewhat enlarged domain still forms in the absence of these elements. Additionally, the mouse α-globin regulatory domain is tissue-specific whereas CTCF binding around the α-globin locus is generally tissue-invariant, and similar patterns are observed between tissues in which the locus is either active, and forming a domain (in erythroid cells), or inactive and not forming a domain (e.g. in ES cells) (Introduction: Figure 1.6). This suggests that CTCF binding alone is not sufficient to form a regulatory domain and such domains may be formed by the combined action of other tissue-specific processes or elements, such as transcriptional activity, the loading of cohesin, or interactions between active enhancers and promoters by bridging transcription factors.

The mouse α-globin locus is a highly conserved and complex locus containing multiple *cis*-regulatory elements. It has five known erythroid-specific enhancer elements, both erythroid and non-erythroid promoters/genes, and multiple CTCF binding sites. Comparison of the α-globin locus and its surrounding sequence in human, mouse, chicken, and pufferfish has identified a highly conserved region which has a defined segment of synteny of between

~135-155 kb (Flint et al., 2001). In another, more extensive comparison of the α-globin locus across 22 divergent species, this region of synteny was shown to be maintained throughout 500 million years of evolution (Hughes et al., 2005), including being found in jawless fish (Philipsen & Hardison, 2018). This well-defined chromosomal unit is predicted to contain all of the *cis*-acting sequences required to fully regulate α-globin gene expression from its natural chromosomal environment. Bioinformatic approaches that had no prior knowledge of previously characterised *cis*-regulatory elements identified 24 multispecies conserved sequences (MCS) in noncoding regions across the α-globin locus which correspond to both known and presumed *cis*-acting regulatory elements (Hughes et al., 2005). Moreover, when a 117 kb genomic fragment of the human α-globin region replaced the equivalent 85 kb region in a transgenic mouse model, it was shown to reproduce the appropriate tissue- and developmental stage-specific expression of the α-globin genes (Wallace et al., 2007). The high levels of conservation around the mouse α-globin locus suggest the presence of *cis*-acting elements that could contribute to the regulation of α-globin expression, potentially through the formation of a regulatory domain.

Previous work within the group has focused on characterising the functional role of the mouse α-globin elements, individually and in combination, in the native locus. Individual knockouts of the five enhancer elements in mice caused very little change in the regulatory domains in erythroid cells when compared to WT (Hay et al., 2016). However, a double knockout out of the two strongest enhancers (R1 and R2) caused markedly reduced interactions between the promoters and the regions around the R1 and R2 enhancers. This could suggest that when there is only one enhancer element missing, there is enough functional redundancy among the remaining enhancers for a regulatory domain to form, but in the absence of strong enhancers, full regulatory domain contacts cannot be established.

Therefore, strong enhancers could be playing a role in the formation of a regulatory domain. Furthermore, the formation of the α-globin regulatory domain happens early in terminal erythroid differentiation (Brown et al., 2018). Despite the low/absent expression of α-globin in early erythroid cells, the *cis*-regulatory elements are bound by erythroid-specific transcription factors. This indicates that tissue-specific regulatory domains may depend on the presence of activated lineage-specific *cis*-regulatory elements for their formation, which subsequently contribute to a permissive environment for gene activation. Additionally, in the absence of the *Hba-α1/2* promoters, the regulatory domain was still present in erythroid cells and looked comparable to that of WT (Brown et al., 2018), suggesting that erythroid promoters may not be needed for α-globin domain formation. However, it must be noted that this model contained a PGK-neo selection cassette within the locus; this heterologous promoter within the regulatory domain may have contributed to domain formation.

Genetic manipulations of the α-globin locus have helped us understand which *cis*-regulatory elements are necessary for domain formation, but the mouse α-globin locus is highly conserved and contains numerous tissue-specific elements that could also be contributing to domain formation, so it has proven difficult to explore which elements are sufficient for domain formation at this locus. Therefore, I aimed to identify the minimal requirements needed for the formation of an erythroid-specific regulatory domain, by constructing various permutations of artificial loci using the mouse α-globin elements (Figure 3.1). Artificial loci were generated in neutral regions of the mouse genome to determine if the tissue-specific α-globin elements were sufficient for domain formation outside of their native highly-conserved chromosomal context. ES cell lines (E14-derived) containing the artificial loci were used in an *in vitro* haematopoietic differentiation system which produces primitive-like erythroid cells (Keller, 1995; Francis et al., in preparation).

In this chapter, I describe how I defined and identified neutral regions of the mouse genome in both ES and erythroid cells (Ter119+), and how I selected candidate regions on chromosome X to develop a targeting strategy appropriate for methods of data analysis. I have also described the targeting strategy, and selection and definition of mouse α-globin elements used in the artificial loci.

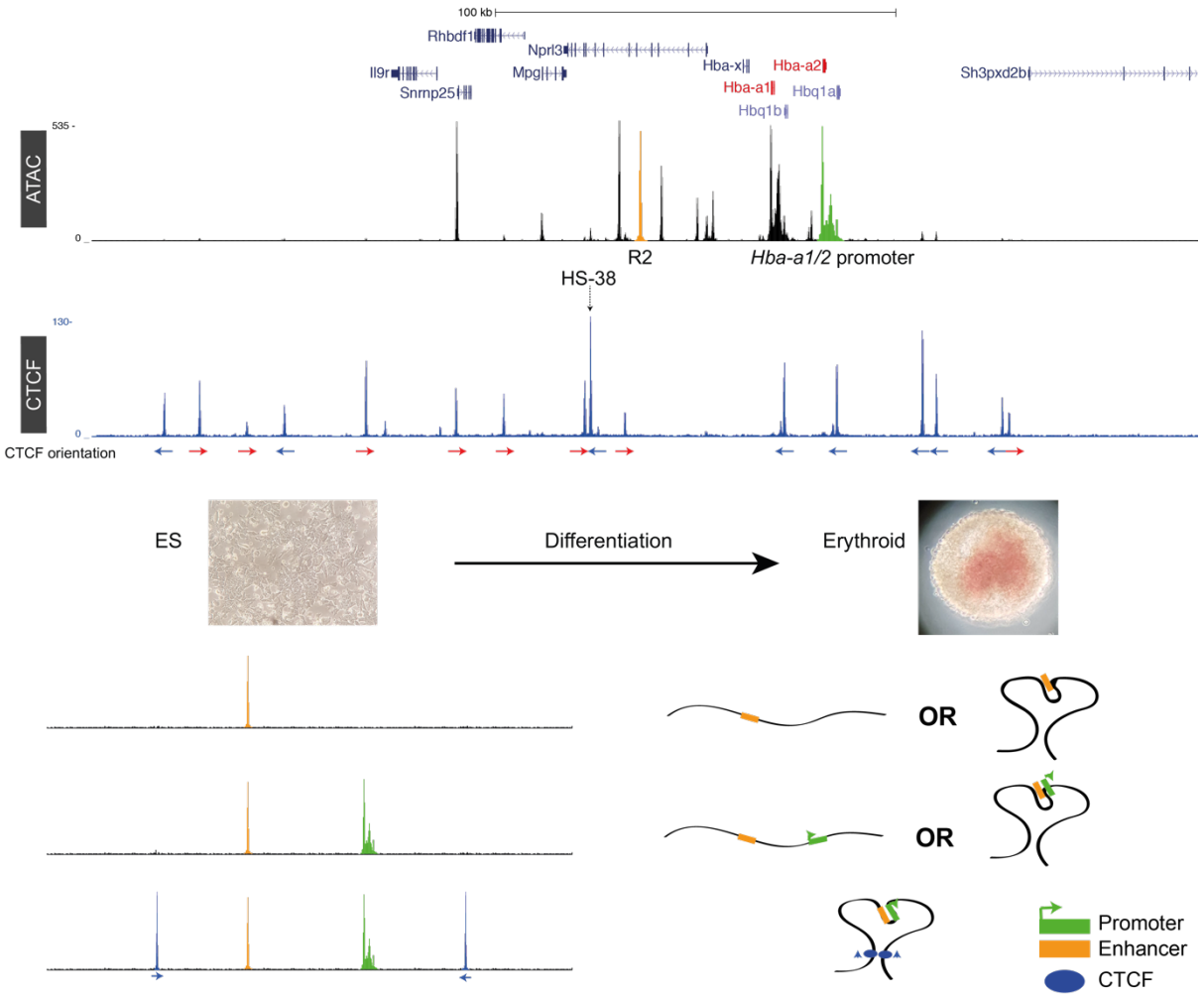


Figure 3.1: Overview of the design and targeting strategy for artificial loci. *Schematic representation showing the introduction of the basic regulatory elements of the mouse α-globin locus (R2, Hba-α1/2 promoter, HS-38) into neutral chromatin regions in ES cells. ES cell lines then used in an in vitro haematopoietic differentiation system to produce erythroid cells from embryoid bodies. Predictions of regulatory domain formation shown in bottom right. Top tracks show the mouse α-globin locus on chromosome 11: gene annotation from Refseq with the α-globin genes highlighted in red, ATAC-seq and CTCF ChIP-seq profiles in primary erythroid cells isolated from WT mice (Ter119+). Profiles show normalised (RPKM) and averaged data from three biological replicates (Hanssen et al., 2017). The ATAC-seq peak for the enhancer element R2 is coloured orange, the peak for the Hba-α1/2 promoter is coloured green, and the CTCF ChIP-seq peak for HS-38 is identified by the dotted arrow. Due to time constraints of this thesis project, the bottom model was not generated.*

Results

Identification of neutral regions of the mouse genome

To investigate chromatin accessibility and CTCF binding genome-wide, we performed ATAC-seq and CTCF ChIP-seq on primary erythroid cells derived from WT mice (Hanssen et al., 2017; Figure 3.2A). Visualisation of the data over large regions (>20 Mb) revealed clusters of accessible chromatin peaks and CTCF binding peaks, but also showed these clusters to be interspaced with suspiciously large regions of chromatin that showed low/no accessibility or CTCF binding. The regions were also relatively devoid of genes and so could represent optimal regions for genome engineering. However, a major concern of the experimental design is that large parts of the genome are known to be highly repressive (van Steensel & Belmont, 2017) and so even if I successfully engineered *cis*-regulatory elements into these regions they might not function. I decided to test this using the unique binding properties of CTCF which, unlike typical transcription factors, requires only its own binding site to associate with chromatin. Therefore, to determine whether these large gaps in the binding profiles for CTCF were due to a repressive environment or due to a lack of CTCF binding motifs, I performed motif analysis to identify significant matches to the CTCF core motif across the genome and intersected these sites with CTCF peak calls from ChIP-seq data. The CTCF core binding motif is present throughout the entire genome, including the regions described above, and the majority of CTCF motifs are not bound by CTCF in erythroid cells (Figure 3.2), showing that these regions were indeed refractory to binding and so likely lying within regions of repressive chromatin.

Using publicly available data, I determined that regions lacking chromatin accessibility and CTCF binding were generally associated with the nuclear lamina (Figure 3.2A); a transcriptionally repressive environment (Akhtar et al., 2013; Finlan et al., 2008; Reddy et

al., 2008). Chromatin that interacts with the lamina form lamina-associated domains (LADs), which are enriched for the repressive chromatin modification H3K9me2 (Ugarte et al., 2015). This suggests that the epigenetic landscape of LADs can impact chromatin accessibility and CTCF binding. LADs are relatively consistent between different cell types and across differentiation, and only show local alterations in contacts around genes that are important for cellular identity (Figure 3.A; Peric-Hupkes et al., 2010). Furthermore, in collaboration with Robert A. Beagrie (WIMM), I performed compartmentalisation analysis on published Hi-C data obtained from ES cells (Selvaraj et al., 2013), which determined the A and B compartments of the genome. This showed that LADs generally reside in the B compartment of the genome and confirmed that they are associated with inactive chromatin, and are, therefore, not neutral and unsuitable for genome engineering (Figure 3.2A).

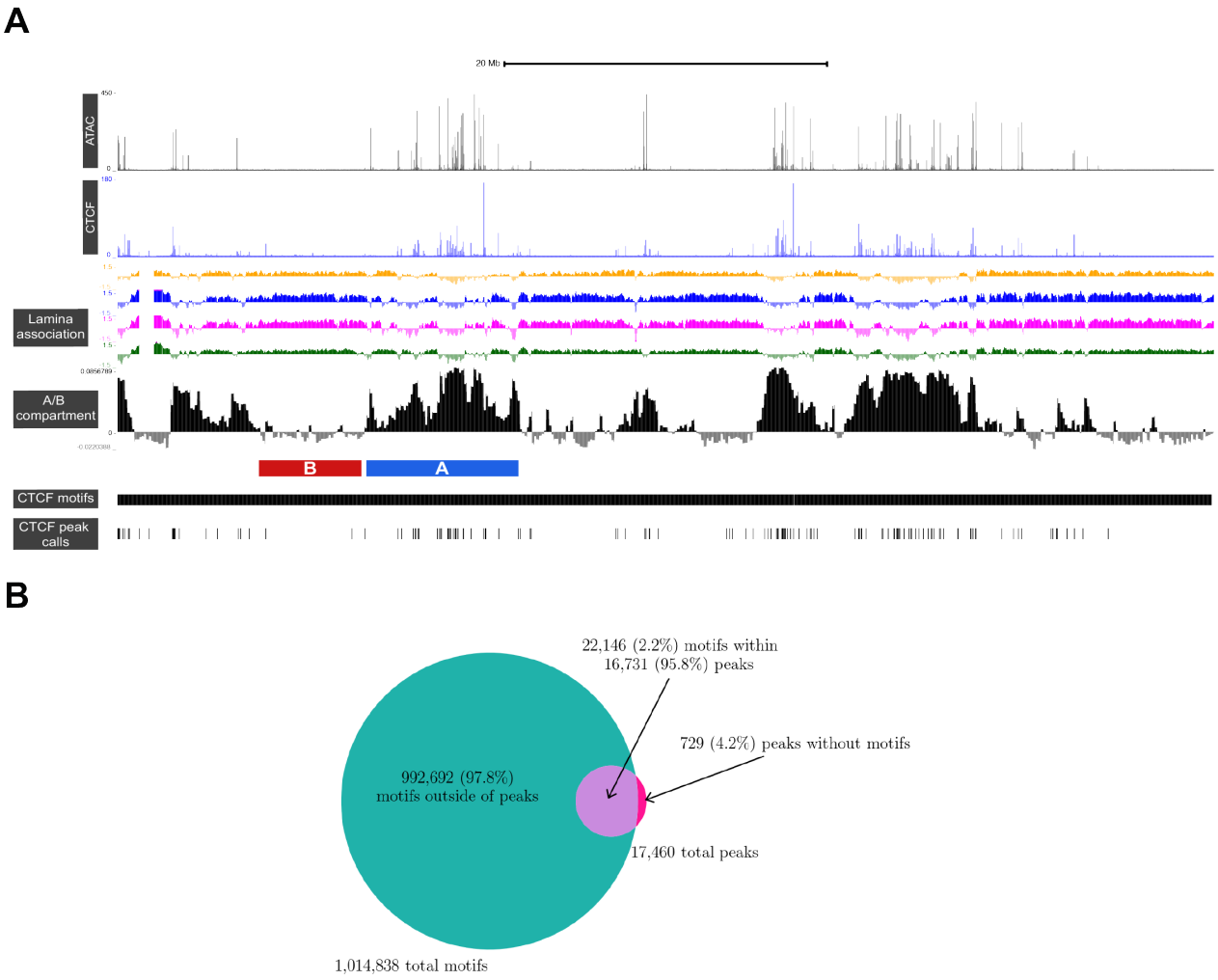


Figure 3.2: Genome-wide chromatin accessibility and CTCF binding in erythroid cells. A: *Top tracks show profiles for ATAC-seq and ChIP-seq in primary erythroid cells isolated from WT mice (Ter119+) for a region on chromosome X. Profiles show normalised (RPKM) and averaged data from three biological replicates (Hanssen et al., 2017). Middle tracks show profiles of DamID of LaminB1 data in four different mouse cell types: ES cells (orange), neural progenitor cells (blue), astrocytes (pink), and embryonic fibroblasts (green) (Peric-Hupkes et al., 2010). These profiles represent regions of chromatin that are associated with the lamina. Bottom tack shows the profile of compartmentalisation analysis of Hi-C data from ES cells (Selvaraj et al., 2013): positive values represent chromatin in the A compartment and negative values represent chromatin in the B compartment (an example region of chromatin residing in the A and B compartment are highlighted with blue and red bars respectively); matches to the CTCF core binding motif; and the CTCF peak calls from CTCF ChIP-seq in erythroid cells.* ***B:*** *A Venn diagram showing the intersection between CTCF motif analysis across the mouse genome and CTCF peak calls from CTCF ChIP-seq data in WT primary erythroid cells (Ter119+).*

To avoid generating artificial loci in the B compartment, I used a bioinformatic approach to identify chromatin regions residing in the A compartment, in which it was more probable that the elements of the artificial loci would become accessible and active. As the presence/absence of CTCF binding clusters behaved as a reasonable proxy for defining A/B compartments respectively, I analysed the CTCF ChIP-seq data obtained from primary erythroid cells to determine genomic regions which either did or did not have the ability to bind CTCF. I called peaks from CTCF ChIP-seq data and 'stitched' these peaks together into genomic blocks representing CTCF binding. As LADs are megabases in length, I reasoned that a gap of substantially less than a Mb between CTCF-bound regions would represent a true gap in binding rather than a repressive domain. Computationally, I analysed the CTCF ChIP-seq data such that if the distance between the peaks was 500 kb or less they were joined to form a genomic block; if the distance was greater than 500 kb then the genomic block ended at the last CTCF peak and the next genomic block started at the next peak. In addition to identifying genomic blocks of CTCF binding, this script simultaneously determined the genomic distance between CTCF peaks. This identified candidate chromatin regions that had sufficient genomic distance between CTCF-bound sites and likely within a block of CTCF-binding/residing in the A compartment, in which to target the elements from the α-globin locus (>200 kb). An example of genomic blocks representing CTCF binding and genomic distances between CTCF peaks can be seen in Figure 3.3.

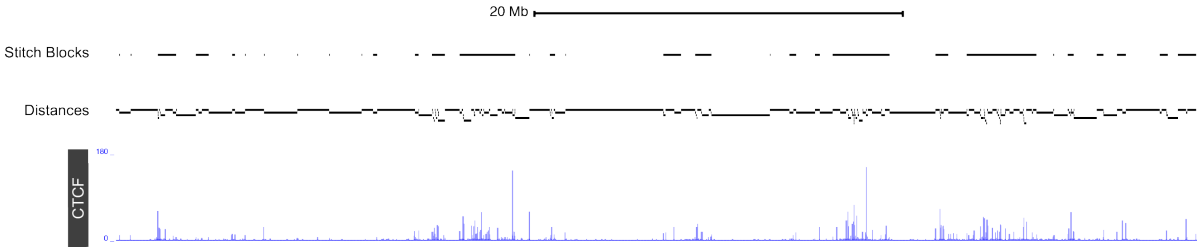


Figure 3.3: CTCF stitch blocks and genomic distances between CTCF peak calls. *Tracks show genomic blocks that represent CTCF binding, stitched together from the CTCF peak calls identified in the CTCF ChIP-seq profile in primary erythroid cells isolated from WT mice (Ter119+), and the genomic distances between the peak calls for a region on chromosome X. CTCF ChIP-seq profile shows normalised (RPKM) and averaged data from three biological replicates (Hanssen et al., 2017).*

To further determine if the candidate regions identified above were neutral rather than part of a repressive LAD, I analysed previously published data obtained from primary erythroid (Ter119+) and ES cells to assess genome organisation and epigenetic landscapes (Figure 3.4). Regions were classified as neutral if they were not enriched for the transcriptionally repressive chromatin marks generally enriched in LADs (H3K9me3 and H3K27me3), and had no pre-existing functionality as judged by accessible chromatin signals as determined by ATAC-seq and ChIP-seq signals for histone modifications associated with active enhancers and promoters (H3K4me1 and H3K4me3 respectively). Neutral regions were also chosen to be devoid of annotated genes and had no underlying TAD structure as determined by Hi-C (TADs appear as triangles of signal in Hi-C matrices; data not shown).

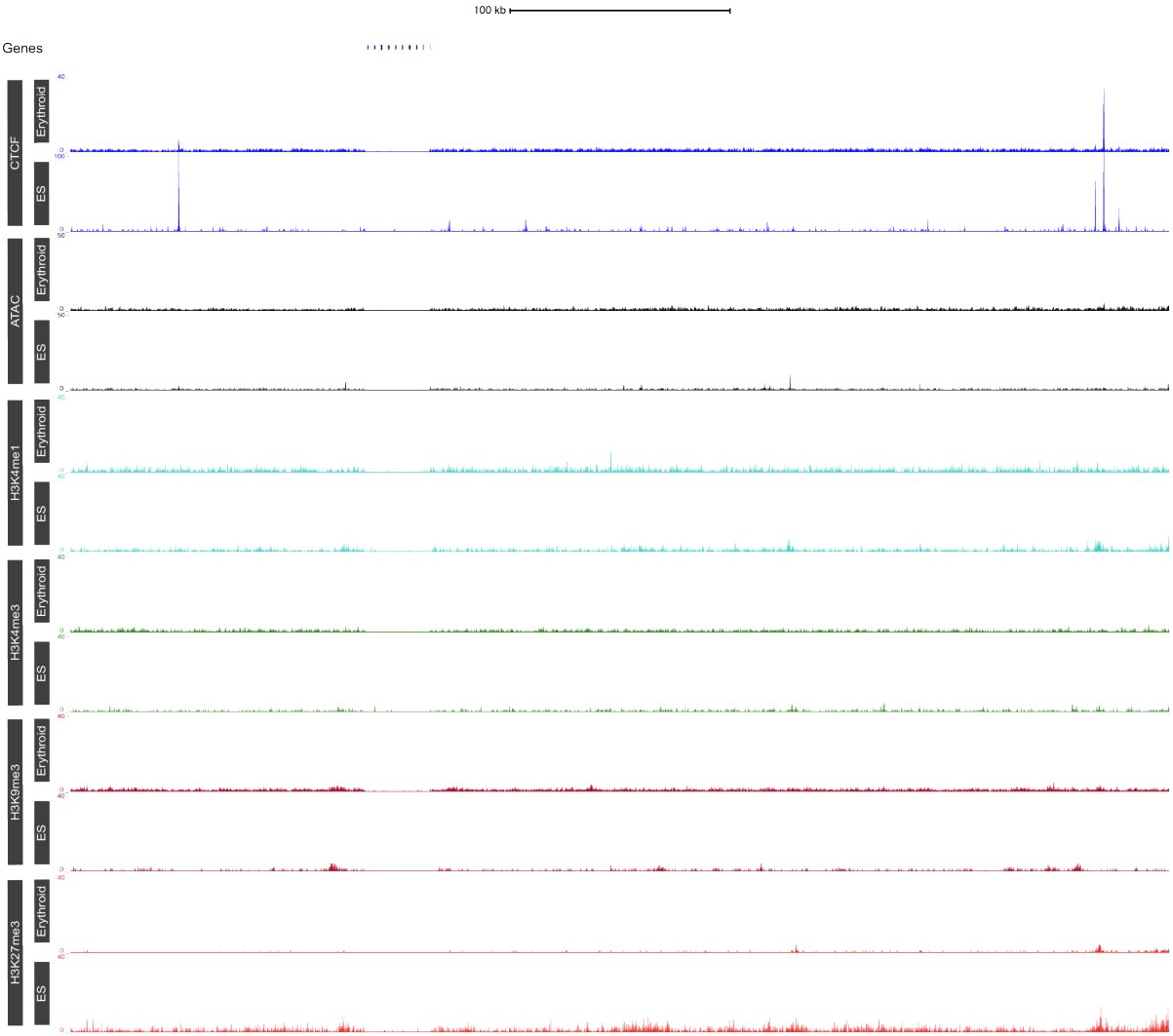


Figure 3.4: A neutral region of chromatin on chrX. *Tracks show profiles for ATAC-seq and ChIP-seq in primary erythroid cells isolated from WT mice (Ter119+) and ES cells for a region on chromosome X. Profiles show normalised (RPKM) and averaged data from differing numbers of biological replicates: three replicates for erythroid ATAC, CTCF, H3K4me3 and H3K27me3 (Hanssen et al., 2017), one replicate for erythroid H3K4me1 (Kowalczyk et al., 2012) and H3K9me3 (ENCODE), three replicates for ES ATAC (Simon et al., 2017) and CTCF (Stadler et al., 2011), one replicate for ES H3K4me1 (Stadler et al., 2011) and H3K4me3 (ENCODE), two replicates for ES H3K9me3 (Bulut-Karslioglu et al., 2014), and three replicates for ES H3K27me3 (Stadler et al., 2011). Gene annotation from Refseq.*

Selection of regions that aided targeting strategy and analysis of NGS-based assays

In order to aid the targeting strategy and subsequent data analysis of NGS-based assays, I identified candidate regions that satisfied the above criteria on chromosome X. Any genome editing accomplished on chromosome X in male E14 ES cells would cause all mutations at the artificial loci to be in *cis* and would remove the need to isolate homozygously edited clones, therefore increasing editing efficiency. Additionally, any changes observed in the NGS-based assays, in particular Capture-C data, of the artificial loci when compared to WT can be directly credited to the targeted region as there is no signal originating from a homologous chromosome.

After manual curation of candidate regions on chromosome X, there were only two regions that satisfied all above criteria and represented potentially neutral chromatin in both ES and erythroid cells (Ter119+). Artificial locus 1 (AL1; mm9 coordinates chrX:10,402,215-10,497,675) is ~96 kb in length and artificial locus 2 (AL2; mm9 coordinates chrX:10,935,857-11,046,248) is ~110 kb in length. To ensure both loci had no complex underlying chromosome structure other than the normal proximity signal expected of the chromatin fibre, I performed Capture-C across the loci in WT ES cells and WT CD71+ erythroid cells isolated from cultures of embryoid bodies (Figure 3.5 and Figure 3.6). I identified DpnII fragments between 1-3 kb in length that spanned across the loci to act as viewpoints of interest. I was able to design suitable capture oligonucleotides to five viewpoints across AL1 and six viewpoints across AL2. This provided much higher resolution interaction data than Hi-C and confirmed the absence of any structure in the two cell types. The DpnII viewpoint fragments were also used as 'landing pads' (LPs) in which to target the α-globin locus elements. This allowed me to assay local chromatin structure in the loci both before and after editing with the same set of capture oligonucleotides.

Hereafter, the nomenclature I will use to refer to the five LPs in AL1 is LP1-5 and for the six LPs in AL2 is LP1-6, with LP1 in both loci corresponding to the most 5' LP and LP5/6 corresponding to the most 3' LP in AL1/AL2 respectively.

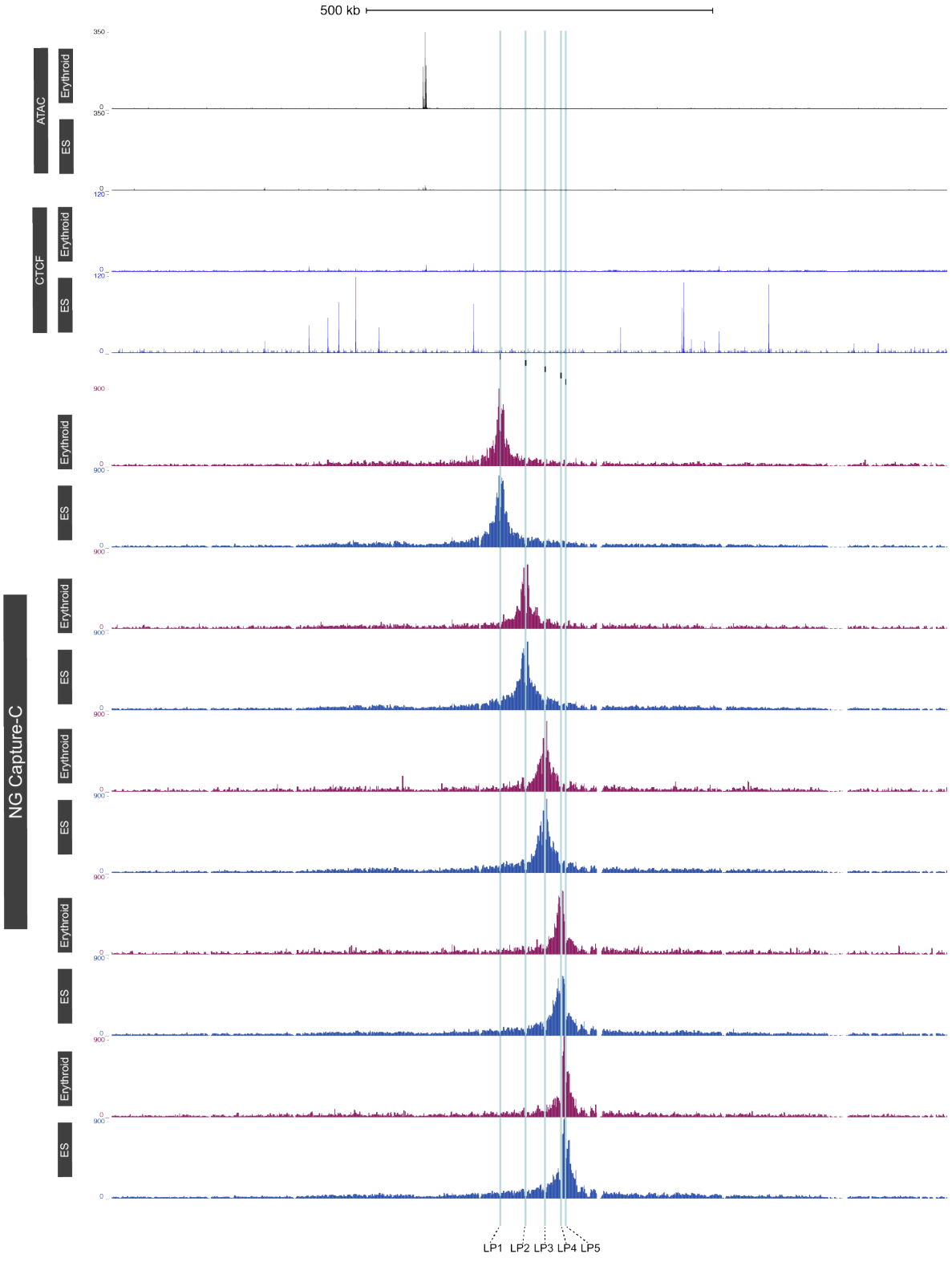

500 kb
ATAC
Erythroid
ES
CTCF
Erythroid
ES
NG Capture-C
Erythroid
ES
Erythroid
ES
Erythroid
ES
Erythroid
ES
Erythroid
ES
LP1
LP2
LP3
LP4
LP5

Figure 3.5: Capture-C interaction profiles of AL1 from the viewpoints of LP1-5. *Top tracks show profiles for ATAC-seq and CTCF ChIP-seq in primary erythroid cells isolated from WT mice (Ter119+) and ES cells for AL1 on chromosome X. Profiles show normalised (RPKM) and averaged data from three biological replicates for erythroid ATAC and CTCF (Hanssen et al., 2017) and three replicates for ES ATAC (Simon et al., 2017) and CTCF (Stadler et al., 2011). Bottom tracks show NG Capture-C interaction profiles of AL1 from the viewpoints of LP1-5 (highlighted by green vertical bars), with a 1 kb exclusion zone around each viewpoint, in WT CD71+ erythroid cells isolated from day7 embryoid bodies (purple) and WT ES cells (blue). The profiles represent normalised and averaged unique interactions from three biological replicates. (Figure on previous page.)*

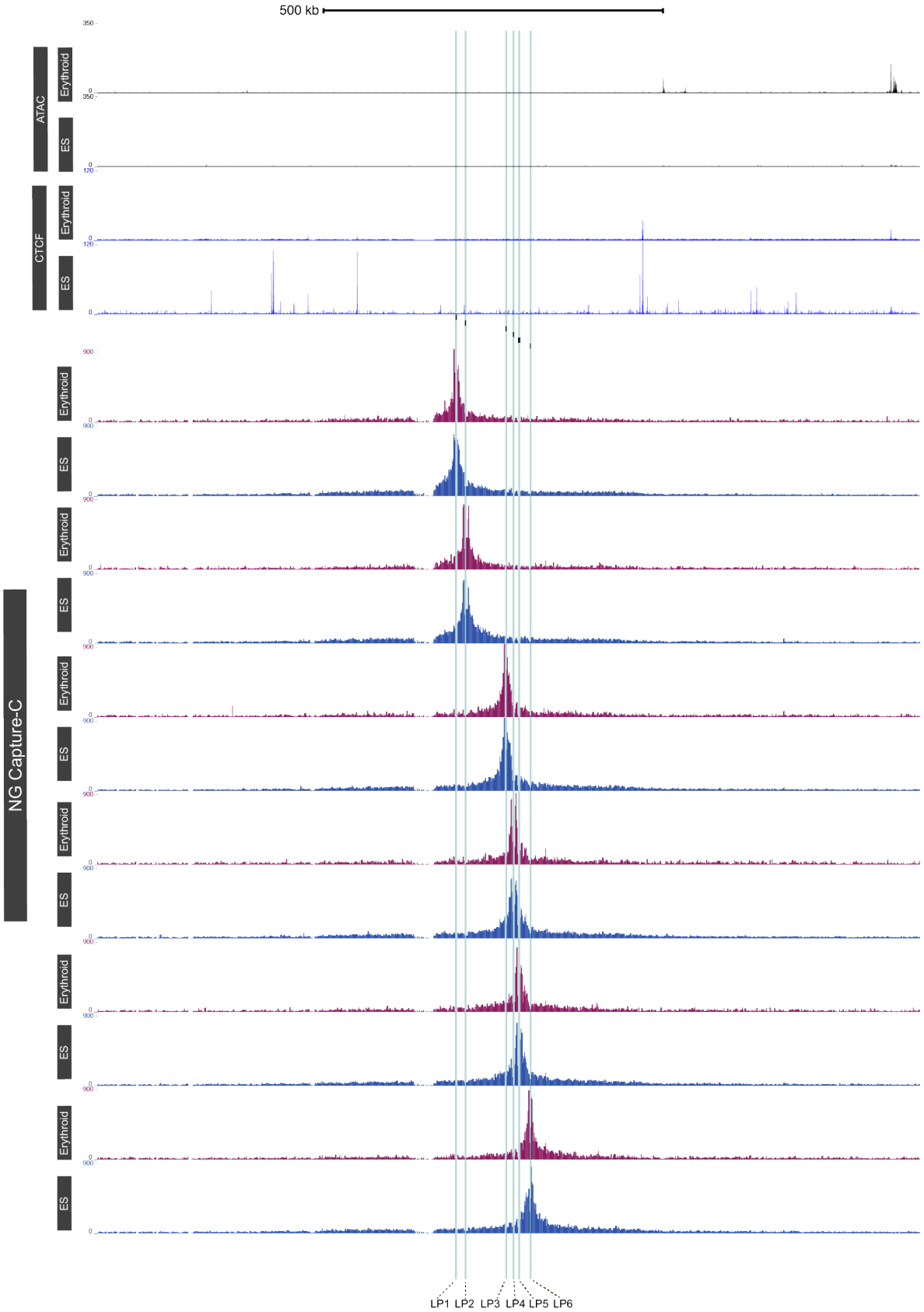

500 kb
ATAC
Erythroid
ES
CTCF
Erythroid
ES
NG Capture-C
Erythroid
ES
Erythroid
ES
Erythroid
ES
Erythroid
ES
Erythroid
ES
Erythroid
ES
LP1 LP2 LP3 LP4 LP5 LP6

Figure 3.6: Capture-C interaction profiles of AL2 from the viewpoints of LP1-6. *Top tracks show profiles for ATAC-seq and CTCF ChIP-seq in primary erythroid cells isolated from WT mice (Ter119+) and ES cells for AL1 on chromosome X. Profiles show normalised (RPKM) and averaged data from three biological replicates for erythroid ATAC and CTCF (Hanssen et al., 2017) and three replicates for ES ATAC (Simon et al., 2017) and CTCF (Stadler et al., 2011). Bottom tracks show NG Capture-C interaction profiles of AL2 from the viewpoints of LP1-6 (highlighted by green vertical bars), with a 1 kb exclusion zone around each viewpoint, in WT CD71+ erythroid cells isolated from day7 embryoid bodies (purple) and WT ES cells (blue). The profiles represent normalised and averaged unique interactions from three biological replicates. (Figure on previous page.)*

Targeting strategy used to introduce the basic regulatory elements of the mouse α-globin locus

I used CRISPR-Cas9 mediated homology-directed repair (HDR) to introduce the basic regulatory elements of the α-globin locus into the two loci on chromosome X in ES cells. In collaboration with Philip Hublitz (Genome Engineering Service, WIMM), I designed three sgRNAs for each LP. Genome Engineering Services cloned the sgRNAs into pX458 (Ran et al., 2013), a plasmid encoding Cas9 from *Streptococcus pyogenes* and enhanced green fluorescent protein (eGFP), and provided functional validation of 33 sgRNAs by performing Surveyor nuclease assay on DNA obtained from transfected B16F10 cells. The sgRNA that provided the highest targeting efficiency for each LP in the two loci was used for targeting (Materials and Methods: Table 6.4).

The first element that was inserted into the artificial loci on chromosome X was the mouse α-globin enhancer R2 (Figure 3.7). R2 is the strongest enhancer of the α-globin promoters; the deletion of R2 in primary erythroid cells caused a ~50% reduction in steady-state α-globin RNA, whereas individual deletions of the other four enhancers (R1, R3, Rm, and R4) caused smaller or no reductions (Hay et al., 2016). I designed donor plasmids with a 325 bp sequence predicted to contain the R2 enhancer and homology arms of ~500 bp for LP2 in both AL1 and AL2 (Materials and Methods: Table 6.5). This 325 bp sequence lies directly under the accessible chromatin peak observed in ATAC-seq and DNase-seq performed in WT primary erythroid cells. It includes the MCS identified for R2 (Hughes et al., 2005), and it contains all of the six known erythroid-specific transcription factor binding sites. Additionally, this sequence encompasses the footprints observed in DNase I hypersensitivity data in WT erythroid cells (unpublished Maria Suciu; Schwessinger & Suciu et al., 2017), which indicate the exact locations of transcription factor occupancy on

chromatin. Therefore, I predicted the insertion of this sequence was sufficient to encode the activity of the R2 enhancer on chromosome X.

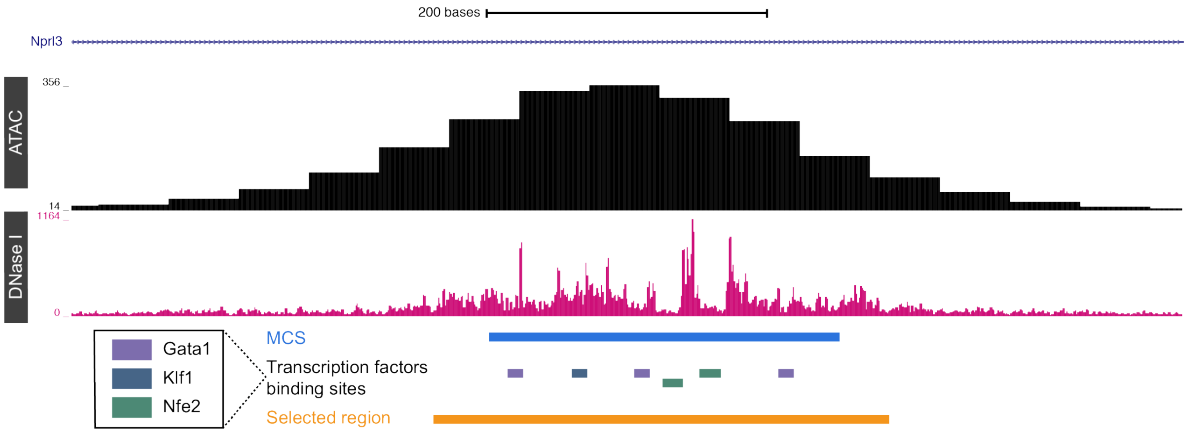


Figure 3.7: Selected region of the R2 enhancer inserted into artificial loci. *Tracks show profiles for ATAC-seq and DNase I footprinting in primary erythroid cells isolated from WT mice (Ter119+) on chromosome 11. ATAC-seq profile shows normalised (RPKM) and averaged data from three biological replicates (Hanssen et al., 2017). The coloured bars represent the MCS (blue), the six known transcription factor binding sites in R2 (for Gata1, Klf1, and Nfe2), and the 325 bp region selected to be inserted into artificial loci (orange). Gene annotation from Refseq and shows R2 is within an intron of Nprl3.*

In parallel to inserting the R2 sequence on chromosome X, I targeted as a control a mutated version of R2 that was predicted not to have any enhancer activity into the same LPs (Figure 3.8). I designed donors that were identical to those described above, apart from multiple base-pair mutations in the six known transcription factor binding sites and one predicted binding site (Klf1) (Materials and Methods: Table 6.5). Each mutation was a transversion mutation which maintained the overall GC content of the sequence (e.g. A ↔ T and G ↔ C). The base-pair mutations in each binding site ranged from two to five substitutions; a total of 23 substitutions across the 325 bp sequence. Therefore, I predicted this would stop

the erythroid-specific transcription factors from being able to bind and stop the mutated R2 from becoming accessible and functional. This would allow me to investigate whether an almost identical but non-functional R2 sequence was sufficient for regulatory domain formation.

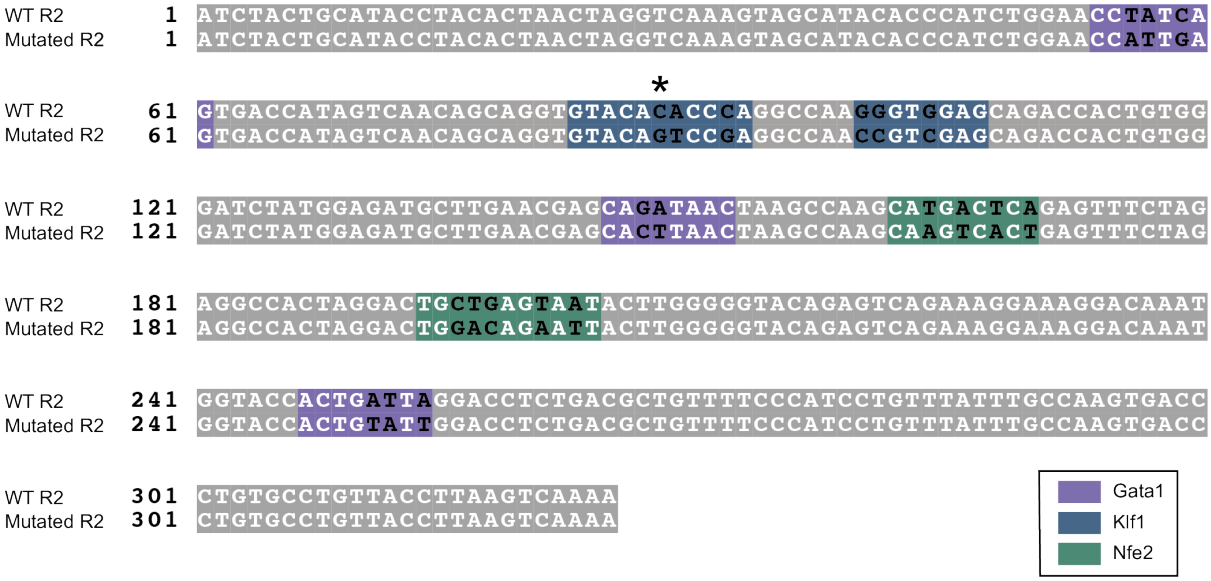


Figure 3.8: Alignment of WT R2 and mutated R2 sequences inserted into artificial loci. *Base-pair mutations in transcription factor binding sequences coloured in black. Gata1 binding sites are highlighted in purple, Klf1 binding sites are highlighted in blue, and Nfe2 binding sites are highlighted in green. The additional predicted Klf1 binding site is marked with the asterisk.*

The next element to be inserted on chromosome X was the mouse α-globin promoter (Figure 3.9), in combination with the R2 insertion. In order to test the minimal requirements needed for an erythroid-specific regulatory domain to form, I did not want to insert the whole *Hba-α1/2* gene, but rather just the promoter with a non-erythroid reporter gene. This would allow me to investigate whether an erythroid-specific enhancer and promoter pair are

necessary and sufficient for regulatory domain formation. Therefore, I designed donor plasmids with a 267 bp sequence predicted to contain the α-globin promoter, a 717 bp sequence encoding mVenus (a monomeric variant of Venus yellow fluorescent protein; Kremers et al., 2006), a 75 bp sequence encoding a triple nuclear localisation sequence, a 102 bp sequence of the α-globin 3' untranslated region, and homology arms of ~500-1000 bp for LP4 in AL1 and LPs 4 and 5 for AL2 (Materials and Methods: Table 6.5).

I ascertained the predicted α-globin promoter sequence using data from ATAC-seq, DNase I hypersensitivity, and a novel method developed within the group by Martin Larke: short capped RNA-seq (scaRNA-seq; Larke et al., in preparation) (Figure 3.9), which identifies transcription start sites with high resolution. As with the predicted R2 sequence, the 267 bp predicted promoter sequence lies directly under the accessible chromatin peak and encompasses the DNase I footprints observed in WT primary erythroid cells. It also includes the MCS identified for the α-globin promoter (Hughes et al., 2005). scaRNA-seq identifies sites of transcriptional initiation and promoter proximal regions where RNA ends accumulate. The scaRNA-seq data in WT primary erythroid cells revealed bidirectional transcriptional initiation sites at the α-globin promoter ~200 bp apart. Previous work has established that bidirectional initiation typically occurs within a window of 90-120 bp, suggesting that the DNA sequence within this window is sufficient for initiation to occur (Core et al., 2014). Therefore, I selected a slightly larger region which contained the bidirectional initiation sites as a conservative minimal α-globin promoter.

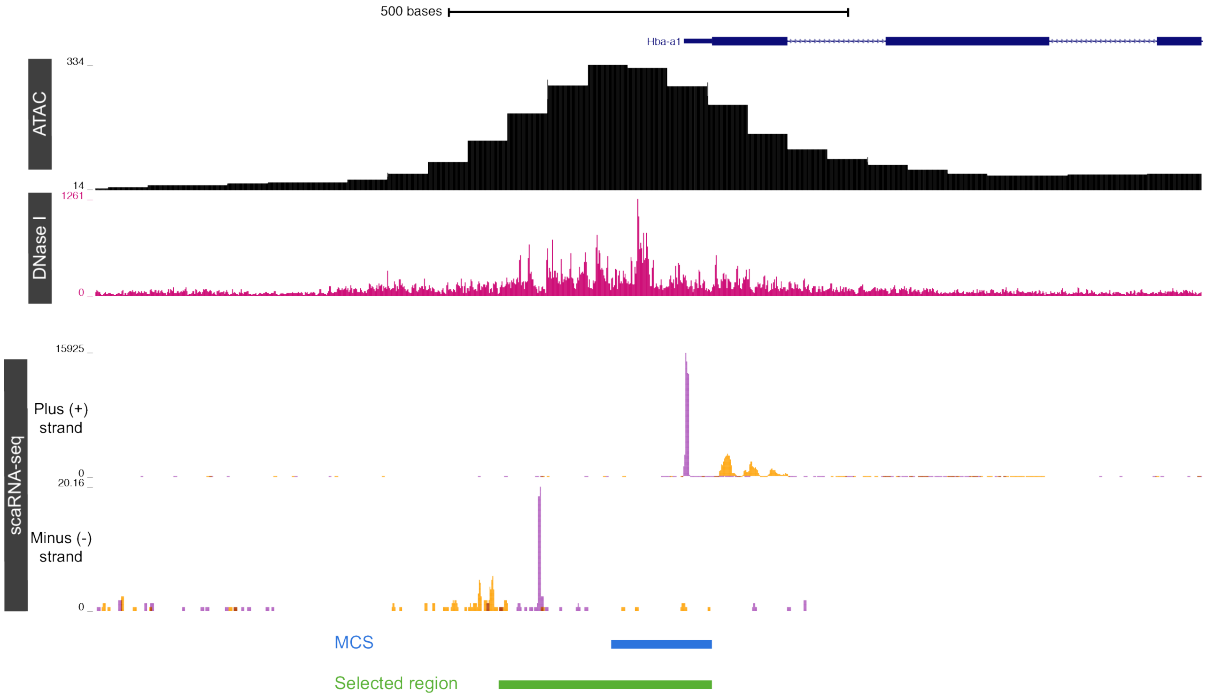


Figure 3.9: Selected region of the α-globin promoter inserted into artificial loci. *Tracks show profiles for ATAC-seq, DNase I footprinting and scaRNA-seq in primary erythroid cells isolated from WT mice (Ter119+) on chromosome 11. ATAC-seq profile shows normalised (RPKM) and averaged data from three biological replicates (Hanssen et al., 2017). scaRNA-seq profiles show transcriptional initiation sites (purple) and RNA ends (orange) for both plus and minus strands. The coloured bars represent the MCS (blue) and the 267 bp region selected to be inserted into artificial loci (green). Gene annotation from Refseq. It should be noted that this sequence is identical for Hba-α1/2.*

Finally, I designed constructs to insert the HS-38 CTCF binding site into the artificial loci on chromosome X in combination with R2 and the α-globin promoter. However, due to time constraints I was not able to perform the genome editing in addition to the analysis of the insertional models already generated as part of this project. I would have targeted the most 3' LP of each locus with the sequence in its native orientation (reverse) and the most 5' LP of each locus with the sequence in an inverted orientation (forward) (Figure 3.1). The 50 bp sequence containing the HS-38 binding site predicted to be sufficient for CTCF binding can

be seen in Figure 3.10. This 50 bp sequence contains the previously identified core and upstream binding motifs of CTCF (Nakahashi et al., 2013) and binding at these motifs has been observed in DNase I footprinting analysis of HS-38 in WT primary erythroid cells (Hanssen et al., 2017).

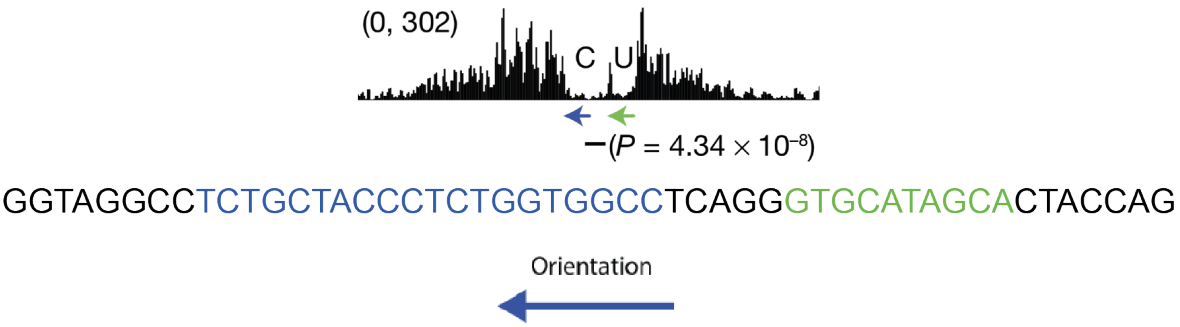


Figure 3.10: Selected sequence of the HS-38 CTCF binding site. *DNase I footprint for HS-38 and the top CTCF binding motif for that site (Hanssen et al., 2017). Motif P-value shown. Reverse core motif highlighted in blue and upstream motif highlighted in green on selected 50 bp sequence.*

Discussion

To determine which, if any, of the known elements of the mouse α-globin locus are sufficient to form a regulatory domain, I have designed a targeting strategy to generate artificial loci containing various permutations of the α-globin locus elements outside of their native chromosomal context. To ensure maximum likelihood of element accessibility and activation upon erythroid-specific differentiation, and to avoid perturbation of normal gene expression, artificial loci were designed in regions of chromatin that I defined as neutral: residing in the A compartment, no CTCF-bound elements, no open chromatin signals, no signals associated with active/repressive histone modifications, devoid of genes, and having no complex underlying genome organisation. Using bioinformatic approaches, I identified two loci on chromosome X: AL1 and AL2. Using Capture-C, I confirmed they had the

expected chromosome structure of neutral chromatin both in ES cells and erythroid cells isolated from embryoid bodies (CD71+). Using a multi-omics approach, I have defined the sequences predicted to encode the activity of the mouse enhancer R2, the mouse α-globin promoter, and for the CTCF binding site HS-38, and outlined the targeting strategy used to generate artificial loci.

At the time of identifying regions of neutral chromatin, we had not fully characterised CD71+ erythroid cells isolated from embryoid bodies (Francis et al., in preparation). Therefore, the identification of neutral regions was performed on available data, both published and unpublished, that originated from different strains of ES cells and definitive Ter119+ primary erythroid cells isolated from either phenylhydrazine-treated mice or from fetal livers (Material & Methods: Table 6.14) However, we subsequently found that AL2 is not completely neutral in CD71+ erythroid cells isolated from embryoid bodies, as small peaks could be observed in the ChIP-seq of H3K4me1. But, these were not large enough to be detected by peak calling algorithms and since all other data was consistent with that observed in Ter119+ erythroid cells, I defined AL2 as 'nearly neutral' and continued with generation and analysis of the model.

Even though insertion of the mouse α-globin locus elements on chromosome X would result in cell lines harbouring multiple copies of the elements, and therefore make NGS-data analysis challenging, I decided against using human α-globin locus elements which would have computationally facilitated distinguishing signal from the native and artificial loci. This was because a transgenic mouse model harbouring a humanised α-globin locus had only 40% of α-globin expression compared to that of the endogenous mouse α-globin genes, and had abnormal red blood cells (Wallace et al., 2007). It may be that the structure or

recognition sequences of the key transcription factors may have altered during evolution of the two species such that binding or stability of mouse transcription factors on human sequences are suboptimal. Therefore, I wanted to insert α-globin elements that had the highest likelihood of achieving maximum activity through the most optimal binding of mouse transcription factors to mouse sequences in a foreign chromosomal context.

I have provided a bioinformatic workflow that identified likely neutral regions of the mouse genome. This workflow provides a framework that can be applied to multiple genomes and can be developed further to be used to identify chromatin associated with other chromatin signatures, such as active and repressive, to make it more widely applicable.

Acknowledgements

Philip Hublitz (Genome Engineering Service, WIMM) cloned and validated the sgRNAs described in this chapter. Robert A. Beagrie (WIMM) helped with various computational analyses throughout this project.

Chapter 4: Analysing the mouse α-globin locus elements in artificial loci outside of their native chromosomal context

Introduction

As described in Chapter 3, I devised a strategy to investigate the minimal requirements needed to form an erythroid-specific regulatory domain. In this chapter, I describe the insertion of mouse α-globin elements into the two loci on chromosome X (AL1 and AL2) in mouse ES cells using HDR via CRISPR-Cas9 genome editing. The CRISPR-Cas9 system is an efficient tool for HDR mutagenesis in mouse ES cells (Wang et al., 2015). My work confirmed this as, combined with targeting a single copy chromosome, I routinely achieved relatively high efficiency of HDR yielding the desired insertion events. All results described in this chapter have been obtained from multiple independently targeted clones, ranging from three to eleven clones per model.

ES cell lines containing the various permutations of the artificial loci were used in an *in vitro* differentiation system to produce embryoid bodies that contain cells of the haematopoietic, endothelial, muscle, and neuronal lineages (Keller, 1995). The lineage-specific differentiation programmes within embryoid bodies are similar to those found in the embryo, and therefore the culture system produces primitive (embryonic) erythroid cell populations. These erythroid cells are somewhat different from definitive erythroid cells (Chapter 3). However, the α-globin elements used to generate artificial loci have been shown to be fully active in these primitive cells. To fully characterise the experimental system, we have assayed the transcriptional programme, chromatin landscape, and cellular morphology of erythroid cells isolated from embryoid bodies (day seven; CD71+), confirming that they are

similar to the primitive erythroid cells in mouse embryonic blood and faithfully recapitulate changes in gene expression seen in full mouse models carrying mutations of key *cis*-regulatory elements (Francis et al., in preparation). We found that the α-globin enhancers and promoters are accessible, have active chromatin marks, and interact with each other when assayed by Capture-C. The embryonic ζ-globin gene is also active in these cells. I was therefore confident that this culture system would be appropriate to investigate the ability of the α-globin elements to form a regulatory domain in AL1/AL2.

In this chapter, I describe the effects of inserting the α-globin elements into AL1/AL2 on chromatin accessibility, histone modifications, transcription, and local genome topology.

Results

The R2 enhancer showed variable local chromatin accessibility in artificial loci

A 325 bp sequence predicted to encode the α-globin enhancer R2 was successfully inserted into LP2 of AL1 and AL2 in mouse ES cells using CRISPR-Cas9-mediated HDR (hereby referred to as AL1R2 and AL2R2, respectively). To investigate whether the inserted R2 became accessible and/or caused changes in local chromatin in ES cells, I performed ATAC-seq on AL1R2, AL2R2, and WT ES cells. AL1R2 and AL2R2 cells harboured an extra copy of the R2 sequence, as the endogenous R2 on chromosome 11 was intact (hereby referred to as chr11-R2), resulting in a total of three copies of R2 in these cells. As discussed later, the fact that there were three copies of R2 in the genome required specialised genomic analysis. However, in ES cells there was no accessible chromatin signal at either the endogenous or the insertion locations, so I was able to conclude that neither chr11-R2 nor AL1R2/AL2R2 were accessible in ES cells without needing to distinguish between reads originating from

different genomic locations. This was expected given that R2 is a known erythroid-specific enhancer. Additionally, there were no differences in local chromatin accessibility in both AL1R2 and AL2R2 ES cells compared to WT (Figure 4.1 and Figure 4.2). Thus, I subsequently differentiated AL1R2, AL2R2, and WT ES cells to form embryoid bodies, isolated the erythroid cells at day seven of differentiation (CD71+), and performed ATAC-seq to determine whether the inserted R2 became accessible in erythroid cells.

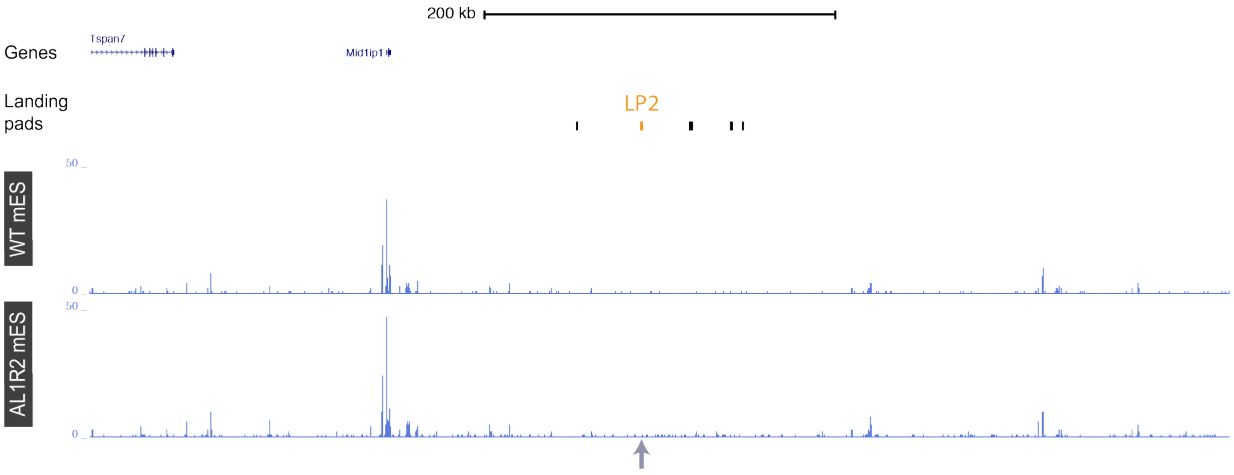


Figure 4.1: Chromatin accessibility in ES cells after insertion of the R2 enhancer in AL1. *Tracks show profiles for ATAC-seq in WT and AL1R2 ES cells for AL1. Profiles show normalised (RPKM) and averaged data from three biological replicates for WT and four independently targeted clones for AL1R2. Landing pads are shown as horizontal bars and LP2 which harbours the R2 insertion is highlighted in orange. The grey arrow denotes the position of the inserted R2 sequence in the ATAC-seq profile. Gene annotation from Refseq.*

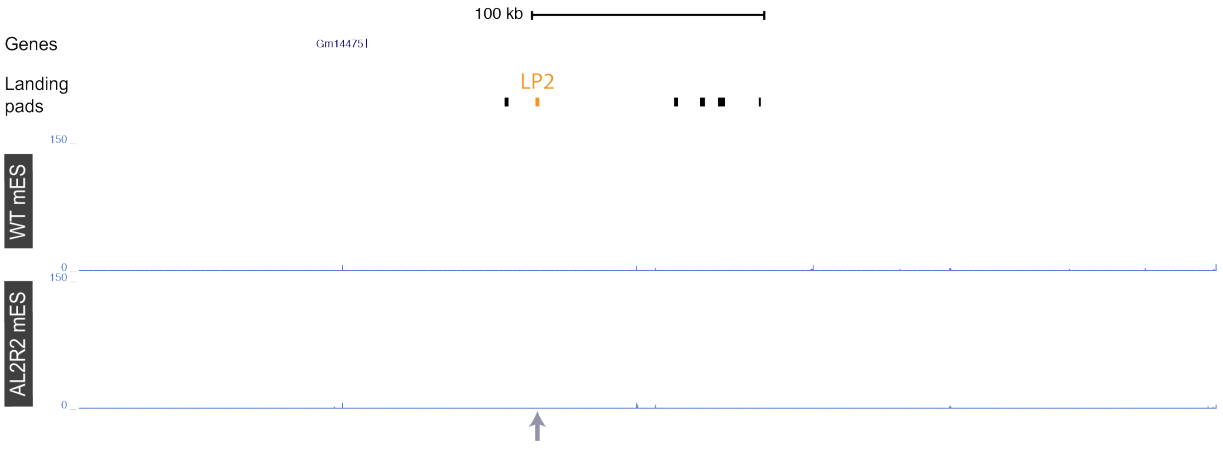


Figure 4.2: Chromatin accessibility in ES cells after insertion of the R2 enhancer in AL2. *Tracks show profiles for ATAC-seq in WT and AL2R2 ES cells for AL2. Profiles show normalised (RPKM) and averaged data from three biological replicates for WT and three independently targeted clones for AL2R2. Landing pads are shown as horizontal bars and LP2 which harbours the R2 insertion is highlighted in orange. The grey arrow denotes the position of the inserted R2 sequence in the ATAC-seq profile. Gene annotation from Refseq.*

I aligned the ATAC-seq from the AL1R2 erythroid cells to a custom reference genome which contained the inserted R2 sequence at AL1. As chr11-R2 is accessible in erythroid cells, alignment of the reads to the custom reference genome had to be performed to allow for multi-mapping to align data which matched both chr11-R2 and the inserted R2 (Figure 4.3A); reads that originated directly from the 325 bp duplicated sequence were randomly mapped to one of the two genomic locations harbouring R2 resulting in peaks at both locations. To investigate whether the accessible chromatin peak over the inserted R2 was a real signal, I isolated the paired-end reads that spanned the junctions of the insertion and compared the profile to the junction reads of chr11-R2. This revealed that two out of four AL1R2 clones had a similar level of chromatin accessibility around the junctions of the R2 insertion compared to the junctions of chr11-R2 in erythroid cells, whereas the remaining two AL1R2 clones showed a reduced level of chromatin accessibility around the junctions of the insertion compared to the junctions of chr11-R2 (Figure 4.3B). In addition, I isolated all ATAC-seq reads that overlapped with the R2 sequence in both WT erythroid cells and in AL1R2 erythroid cells and normalised the read counts to reads per million. The two clones that displayed higher chromatin accessibility at the junctions of the inserted R2 had a higher total count of R2 reads than WT, whereas the two clones that had lower chromatin accessibility at the junctions of the inserted R2 had total counts of R2 reads similar to that of WT (Figure 4.4). Therefore, the 325 bp core R2 sequence was capable of generating accessible chromatin at this locus in erythroid cells, but the accessibility was not consistent between clones despite them harbouring the same artificial sequence. This could imply that additional elements in the native α-globin locus are required for fully robust chromatin-opening activity of R2, or that the predicted neutral region of AL1 could be somewhat repressive leading to sporadic opening of the chromatin over the inserted R2.

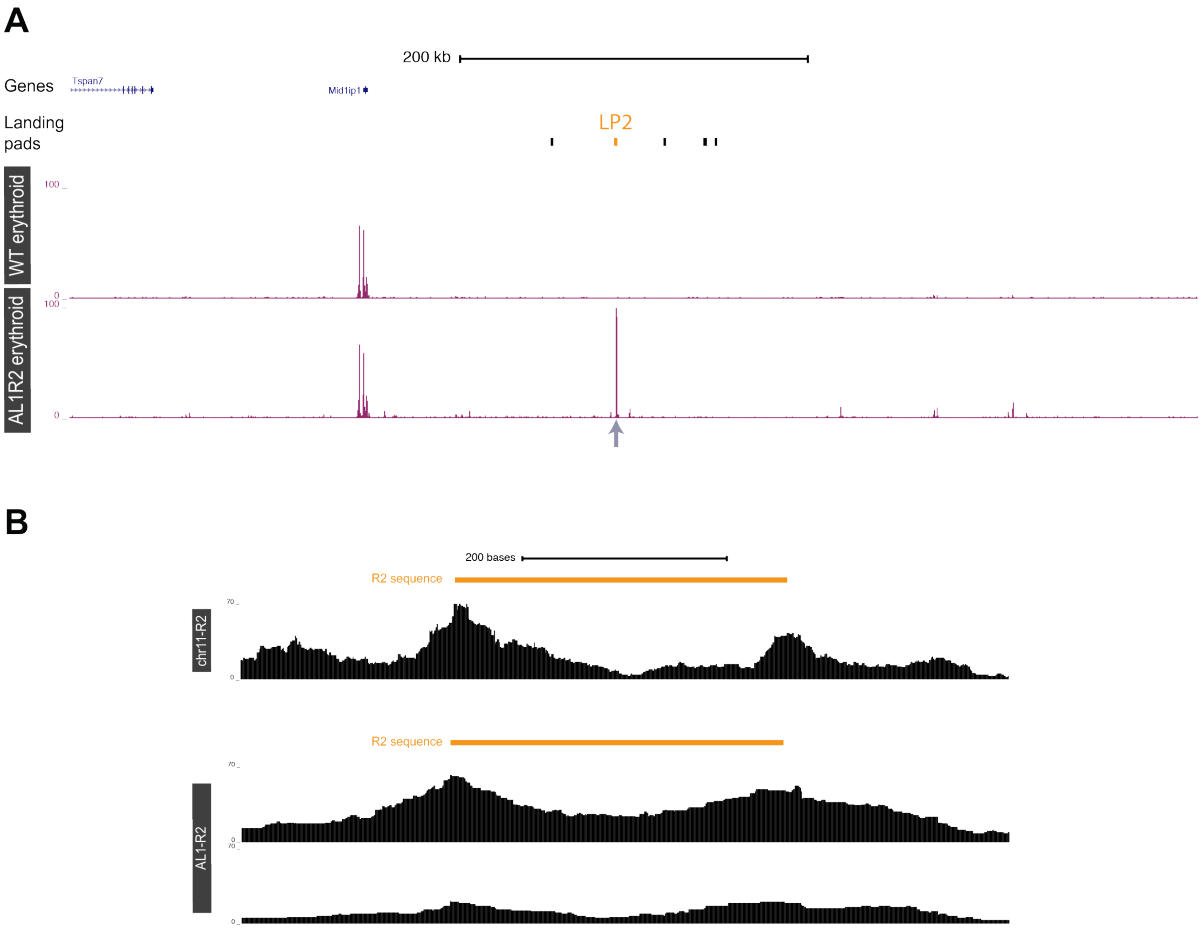


Figure 4.3: Chromatin accessibility in erythroid cells after insertion of the R2 enhancer in AL1. ***A:*** *Tracks show profiles for ATAC-seq in WT and AL1R2 erythroid cells (CD71+) for AL1. Profiles show normalised (RPKM) and averaged data from three biological replicates for WT and four independently targeted clones for AL1R2. Landing pads are shown as horizontal bars and LP2 which harbours the R2 insertion is highlighted in orange. The grey arrow denotes the position of the inserted R2 sequence in the ATAC-seq profile. Gene annotation from Refseq.* ***B:*** *Profiles show paired-end reads from ATAC-seq in AL1R2 erythroid cells that spanned the junctions of R2 in chr11 and AL1. Paired-end reads that originated from the 325 bp sequence of R2 (orange bar) were discarded. Profiles from representative clones; two clones appeared to have a similar level of chromatin accessibility to chr11-R2 junctions, two clones appeared to have a reduced level of chromatin accessibility.*

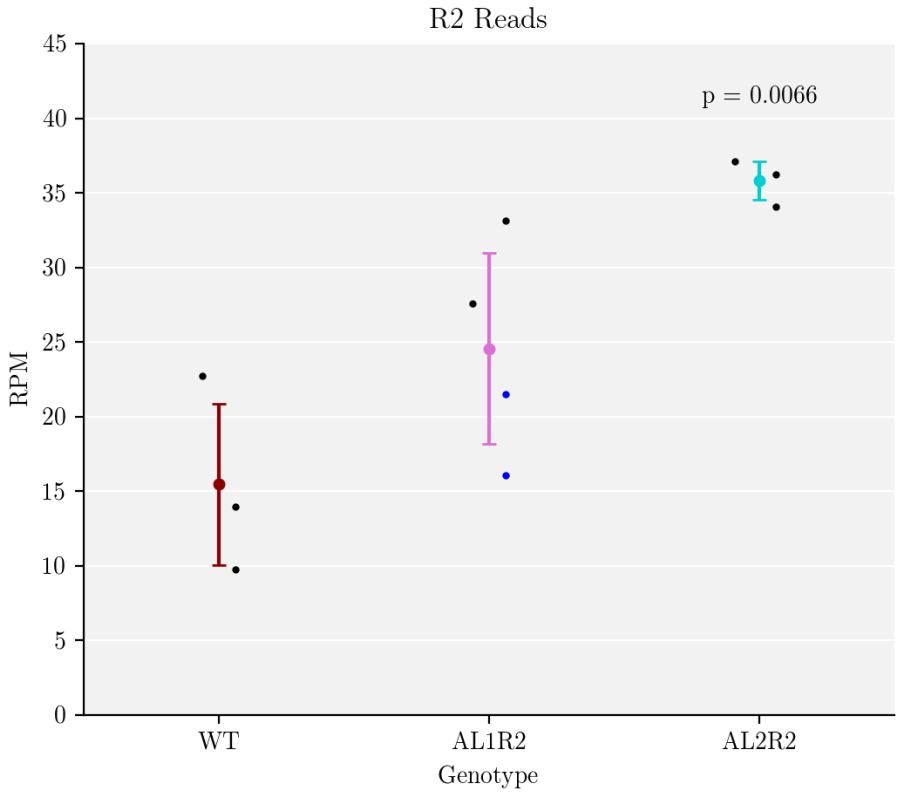


Figure 4.4: ATAC-seq reads originating from R2 in erythroid cells. *Reads that overlapped with the 325 bp R2 sequence normalised to reads per million (RPM) for ATAC-seq performed on WT, AL1R2, and AL2R2 erythroid cells. Mean and standard deviation for each model shown, and each point represents a biological replicate/independently targeted clone. The two clones that displayed lower chromatin accessibility around the junctions of the R2 insertion in AL1 are highlighted in blue. Statistical analysis shows the mean R2 reads from ATAC-seq in AL2R2 erythroid cells significantly differ from that of WT, P-value shown (two-tailed student's t-test).*

I performed the same ATAC-seq analyses described above on AL2R2 erythroid cells and in this case was able to show robust chromatin accessibility of the inserted R2 in AL2 in all clones. All three independently targeted clones showed similar levels of chromatin accessibility around the junctions of the inserted R2 sequence to the junctions of chr11-R2 (Figure 4.5B), and they all had a higher total count of R2 reads than WT (Figure 4.4). In addition, the activation of R2 in AL2 caused local changes in accessibility in the vicinity of the insertion on chromosome X, with the generation of novel ATAC-seq peaks up to ~75 kb downstream (Figure 4.5A). These results show that not only was the chromatin accessibility of the inserted R2 enhancer more stable in AL2 than AL1, but also had the ability to exert changes on nearby chromatin, which suggests that R2 was active as a *cis*-regulatory element at this locus in addition to being an open chromatin site. Therefore, I continued to characterise AL2R2 erythroid cells as the inserted R2 was more likely to recapitulate the activity of chr11-R2 in erythroid cells.

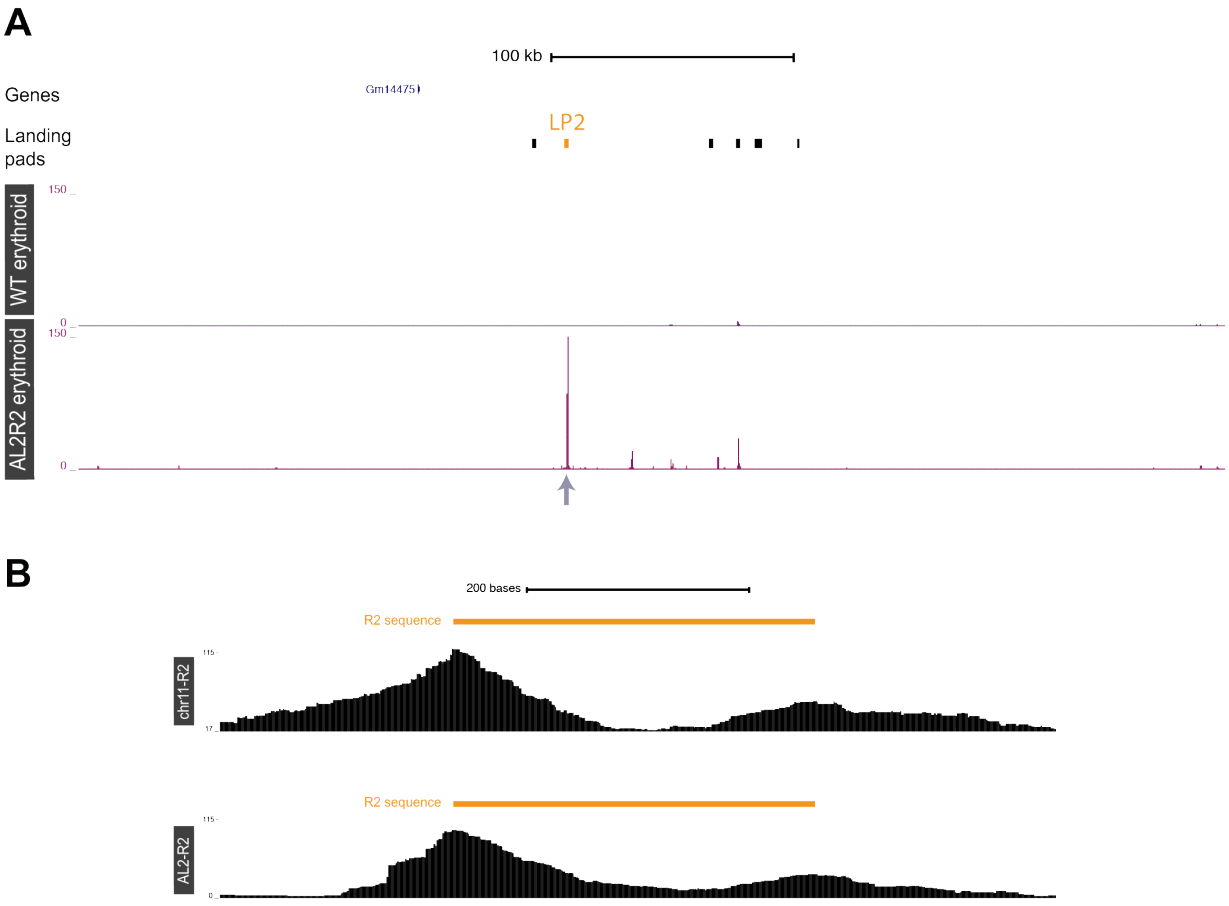


Figure 4.5: Chromatin accessibility in erythroid cells after insertion of the R2 enhancer in AL2. ***A:*** *Tracks show profiles for ATAC-seq in WT and AL2R2 erythroid cells (CD71+) for AL2. Profiles show normalised (RPKM) and averaged data from three biological replicates for WT and three independently targeted clones for AL2R2. Landing pads are shown as horizontal bars and LP2 which harbours the R2 insertion is highlighted in orange. The grey arrow denotes the position of the inserted R2 sequence in the ATAC-seq profile. Gene annotation from Refseq.* ***B:*** *Profiles show paired-end reads from ATAC-seq in AL2R2 erythroid cells that spanned the junctions of R2 in chr11 and AL2. Paired-end reads that originated from the 325 bp sequence of R2 (orange bar) were discarded. Profiles from representative clones.*

Importantly, the inserted sequence of the control, mutated R2 into LP2 of AL2 did not become accessible in erythroid cells when assayed by ATAC-seq and did not cause any local changes in chromatin accessibility around the locus (Figure 4.6). I aligned the ATAC-seq reads to a custom reference genome which contained the mutated R2 sequence in AL2. There were a sufficient number of base pair polymorphisms in the mutated R2 to distinguish reads from chr11-R2. Thus, suggesting that the above observations in AL2R2 erythroid cells were caused by the activity of the inserted R2 and not from the insertion of DNA sequence itself, as the mutated R2 sequence was predicted not to have enhancer activity.

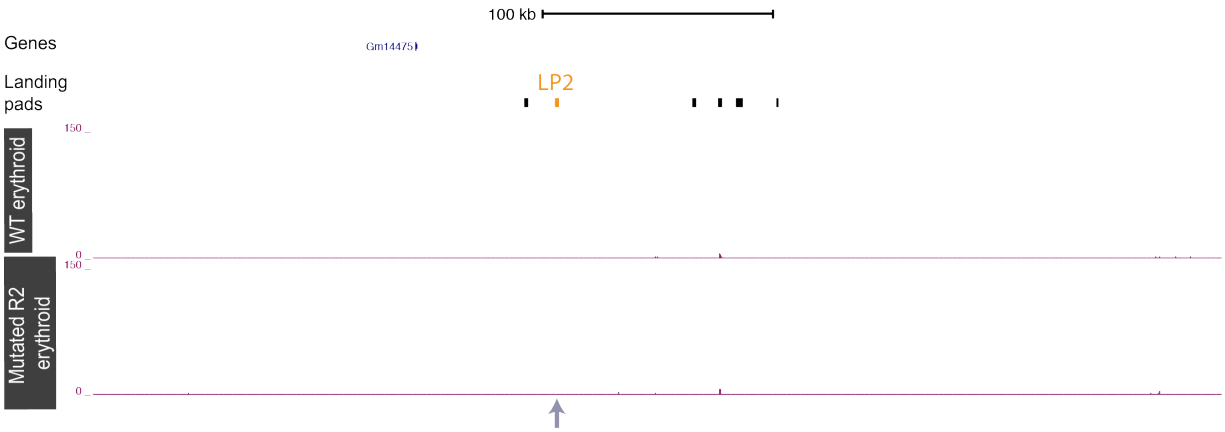


Figure 4.6: Chromatin accessibility in erythroid cells after insertion of the mutated R2 enhancer in AL2. *Tracks show profiles for ATAC-seq in WT erythroid cells (CD71+) and erythroid cells harbouring an insertion of a mutated version of R2 (CD71+) for AL2. Profiles show normalised (RPKM) and averaged data from three biological replicates for WT and three independently targeted clones for the mutated R2 insertion. Landing pads are shown as horizontal bars and LP2 which harbours the mutated R2 insertion is highlighted in orange. The grey arrow denotes the position of the inserted sequence in the ATAC-seq profile. Gene annotation from Refseq.*

The inserted R2 enhancer appears as a promoter-like regulatory element in erythroid cells

To investigate whether the changes in chromatin accessibility were associated with changes in histone modifications, I performed H3K27ac, H3K4me1, and H3K4me3 ChIP-seq in AL2R2 and WT erythroid cells. H3K27ac is associated with active chromatin and is a mark for active transcription, H3K4me1 is associated with active/primed enhancers though can also be found at active promoters, and H3K4me3 is associated with active promoters. The histones around the inserted R2 sequence and the novel ATAC-seq peaks in AL2R2 erythroid cells were marked with H3K27ac and H3K4me1, however only the histones around the inserted R2 sequence were marked with H3K4me3 (Figure 4.7). This was different to the levels of histone modifications found at chr11-R2; comparison of the chromatin signatures of the inserted R2 with chr11-R2 in AL2R2 erythroid cells showed that chr11-R2 had higher levels of H3K27ac and H3K4me1 than the inserted R2, but much lower levels of H3K4me3 (Figure 4.8). Therefore, the inserted R2 enhancer had the chromatin signature of an active promoter in AL2R2 erythroid cells, whereas the newly formed accessible regions of chromatin lying downstream of the inserted R2 had the signatures of enhancers. These observations were confirmed by performing chromatin segmentation analysis on these datasets from AL2R2 erythroid cells, which identified the inserted R2 as a weak-medium strength promoter and the downstream accessible regions as weak enhancers (hereby referred to as enhancer-like elements). However, it must be noted that H3K4me1 peaks can be observed in WT embryoid-body-derived erythroid cells at the sites of the novel enhancer-like elements in AL2R2 erythroid cells, even though no ATAC-seq/H3K27ac peaks were present in WT erythroid cells. Hence these regions may be primed enhancers in WT erythroid cells that become active in AL2R2 erythroid cells following the upstream insertion of the R2 sequence.

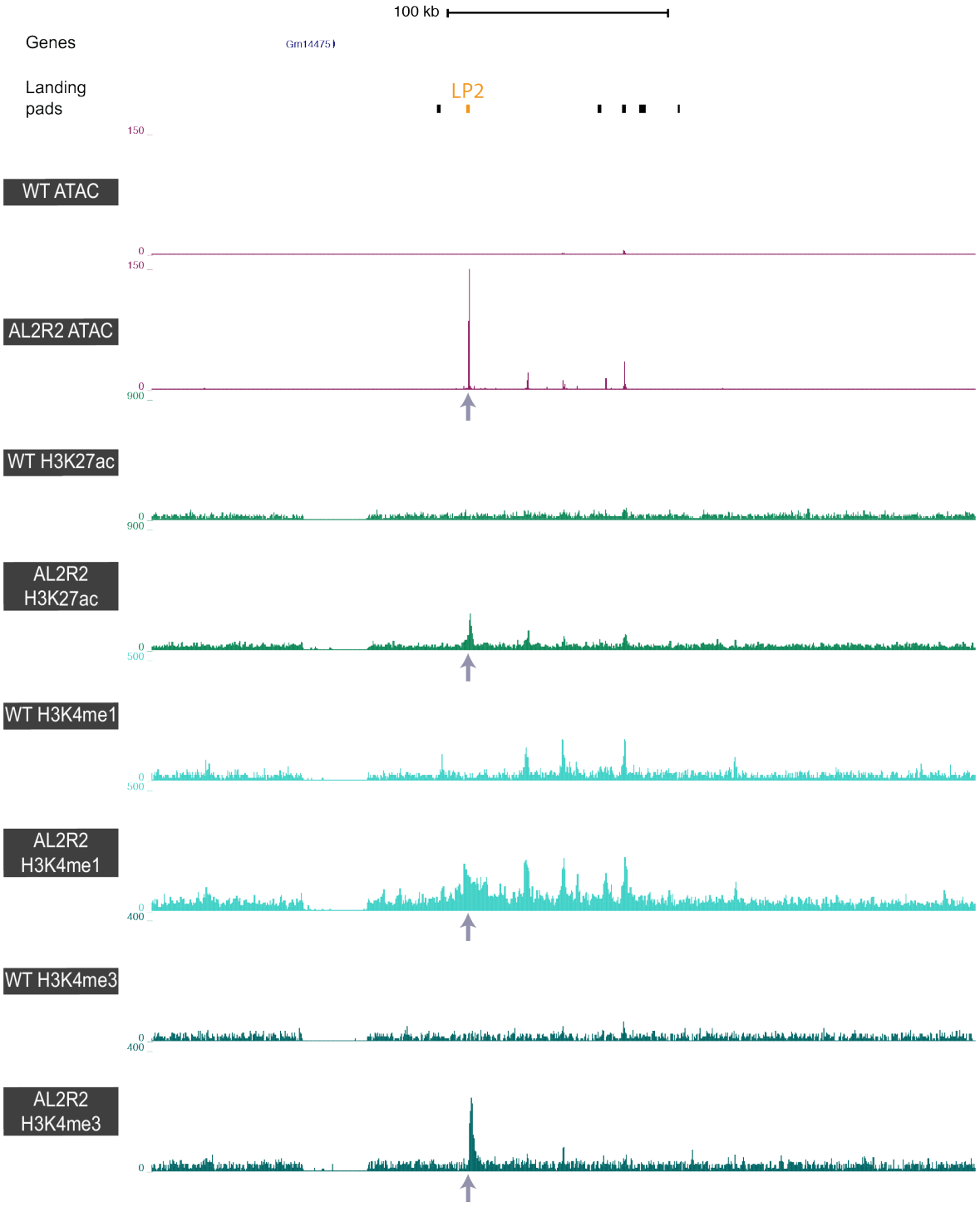


Figure 4.7: Chromatin marks in erythroid cells after insertion of the R2 enhancer in AL2. *Tracks show profiles for ATAC-seq and ChIP-seq for H3K27ac, H3K4me1, and H3K4me3 in WT and AL2R2 erythroid cells (CD71+) for AL2. Profiles show normalised (RPKM) and averaged data from two/three biological replicates for WT (ChIP-seq/ATAC-seq data respectively) and three independently targeted clones for AL2R2. Landing pads are shown as horizontal bars and LP2 which harbours the R2 insertion is highlighted in orange. The grey arrows denote the position of the inserted R2 sequence in the mutant profiles. Gene annotation from Refseq.*

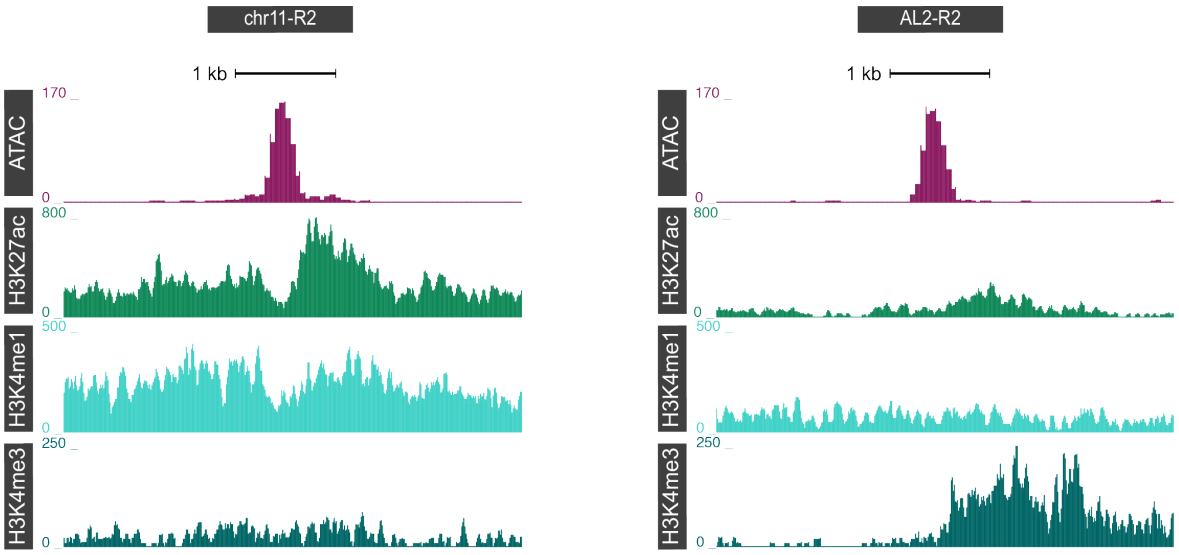


Figure 4.8: Comparison of chromatin marks in erythroid cells between the endogenous R2 on chromosome 11 and the inserted R2 in AL2. *Profiles show ATAC-seq and ChIP-seq for H3K27ac, H3K4me1, and H3K4me3 in AL2R2 erythroid cells (CD71+) for the endogenous R2 on chromosome 11 (left; chr11-R2) and the inserted R2 in AL2 (right; AL2-R2). Profiles show normalised (RPKM) and averaged data from three independently targeted clones for AL2R2 erythroid cells.*

As the inserted R2 enhancer in AL2R2 erythroid cells had an active promoter signature, I wanted to assess whether it caused any transcriptional effects, so I performed strand-specific poly(A)- and poly(A)+ RNA-seq on AL2R2 and WT erythroid cells to detect the production of new enhancer RNAs (eRNAs) and long noncoding RNAs (lncRNAs). The inserted R2 sequence in AL2 was transcribed to generate a unidirectional, ~ 75 kb transcript visible in the poly(A)- data that included the regions containing the novel enhancer-like elements (Figure 4.9). In addition, a unidirectional, ~4 kb transcript transcribed from the inserted R2 was visible in the poly(A)+ data (Figure 4.9). These data suggest that the inserted R2 in AL2R2 erythroid cells acts as a promoter-like element which is transcribed to produce both non-polyadenylated and polyadenylated unidirectional transcripts.

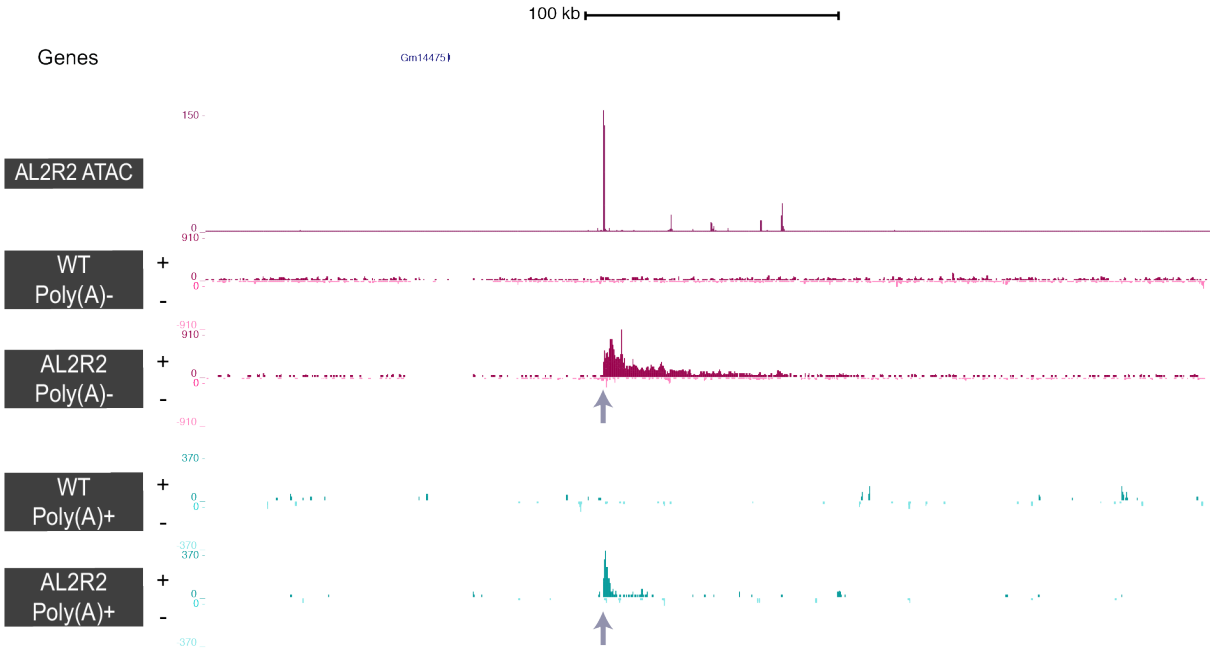


Figure 4.9: Transcriptional changes in erythroid cells after insertion of the R2 enhancer in AL2. *Tracks show profiles for ATAC-seq in AL2R2 erythroid cells (CD71+), and strand-specific poly(A)- and poly(A)+ RNA-seq in WT and AL2R2 erythroid cells (CD71+) for AL2. Profiles show normalised (RPKM) and averaged data from two/three biological replicates for WT and two/three independently targeted clones for AL2R2 for poly(A)+/poly(A)- RNA-seq data respectively. In all profiles, data originating from the plus strand is labelled with + and data from the minus strand is labelled with -. The grey arrows denote the position of the inserted R2 sequence in the mutant profiles. Gene annotation from Refseq.*

R2 is sufficient to form a regulatory domain at Artificial Locus 2 in erythroid cells

I performed Capture-C from the viewpoint of LP2 in AL2R2 and WT erythroid cells to investigate whether the activity of the inserted R2 altered chromatin interactions around AL2 (Figure 4.10). WT erythroid cells showed a normal distribution of interactions around the viewpoint, consistent with the expected proximity signal of a chromatin fibre that has no underlying regulatory structure. AL2R2 erythroid cells, however, showed enriched interactions between LP2, which contained the inserted R2, and the regions downstream spanning the novel enhancer-like elements. These interactions only reduced to the level observed in WT beyond the most 3' novel enhancer-like element located ~75 kb away. Additionally, there was a slight enrichment in interactions upstream of LP2 in AL2R2 erythroid cells, however these did not extend as far as those observed downstream. This pattern of interactions encompassed the region of active transcription and were not specifically focused at the novel enhancer-like elements, but rather extended over the entire domain, similar to that observed at the α-globin locus in erythroid cells (Introduction: Figure 1.8). This suggests that this interaction pattern arises from a scanning mechanism such as loop extrusion. Overall, this indicates that the activity of the inserted R2 enhancer element, in AL2R2 erythroid cells, was sufficient to form a regulatory domain at AL2 in the absence of other *cis*-regulatory elements, such as promoters and CTCF-bound sites.

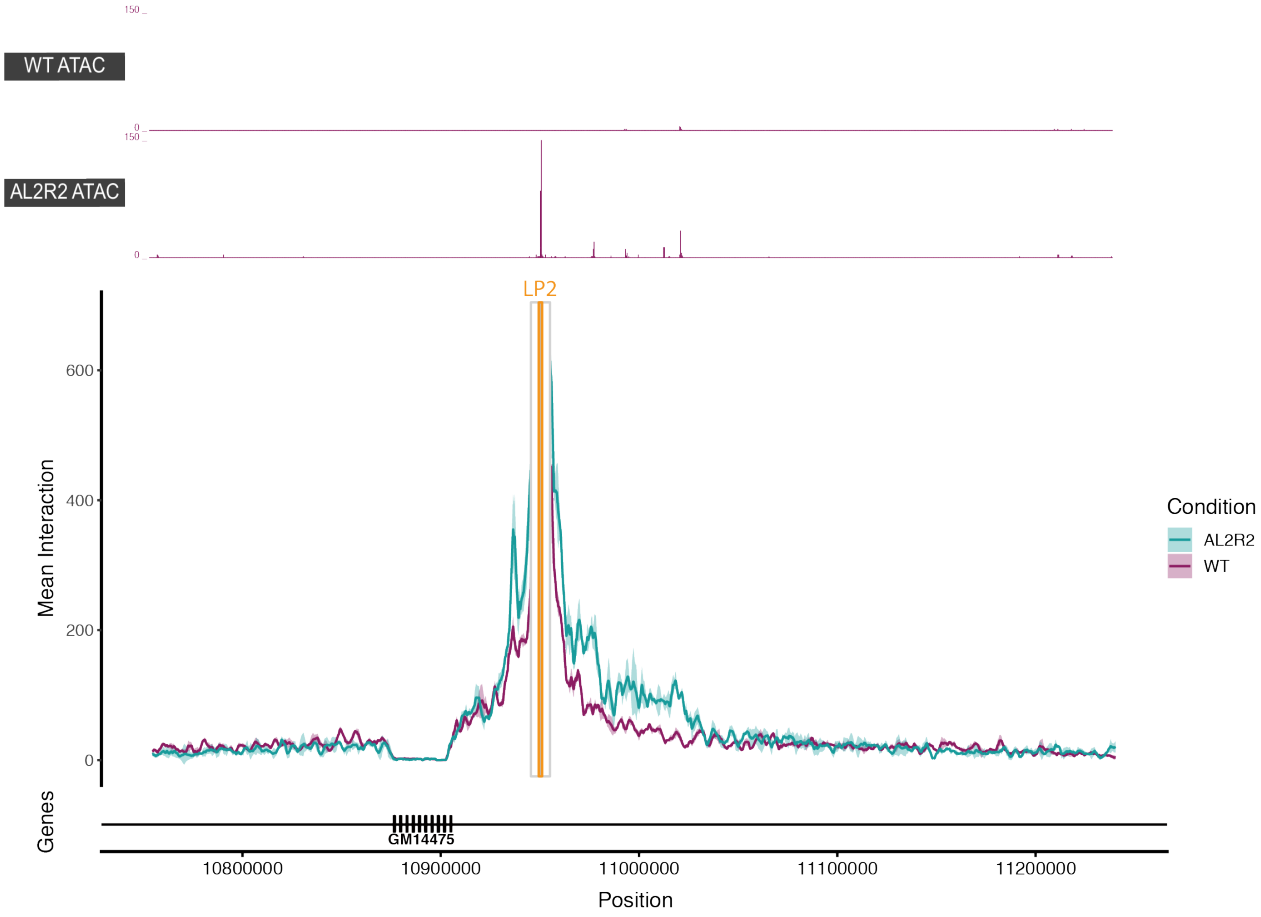


Figure 4.10: Local genome topology in erythroid cells after insertion of the R2 enhancer in AL2. *Tracks show profiles for ATAC-seq in WT and AL2R2 erythroid cells (CD71+) for AL2. Profiles show normalised (RPKM) and averaged data from three biological replicates for WT and three independently targeted clones for the mutated R2 insertion. NG Capture-C interaction profiles of AL2 from the viewpoint of LP2 (orange), with a 1 kb exclusion zone around the viewpoint, in WT (purple) and AL2R2 (cyan) CD71+ erythroid cells. The profiles represent normalised and averaged unique interactions from three biological replicates/independently targeted clones, with the halo representing the standard deviation of a sliding 3 kb window. Genes and genomic position below interaction profiles.*

Insertion of the α-globin promoter into the artificial loci does not lead to expression of a fluorescent reporter gene

A 1173 bp sequence encoding the predicted α-globin promoter and an mVenus fluorescent reporter was successfully inserted into LP4 of AL1R2 ES cells using CRISPR-Cas9 HDR; LP4 is ~53 kb downstream of the inserted R2 in LP2 (Chapter 3: Figure 3.5). Even though the chromatin accessibility/activity of the inserted R2 in AL1R2 erythroid cells was variable, I wanted to investigate whether it was sufficient to drive the production of mVenus RNA and/or protein. In particular, I was interested in whether the presence of an active promoter would provide a more stable chromatin environment in which the activity of the inserted R2 was more robust, and whether this would help in the formation of a regulatory domain. Therefore, I differentiated the AL1R2 ES cells containing the α-globin promoter-mVenus sequence to form embryoid bodies and performed Fluorescence-Activated Cell Sorting (FACS) to assess mVenus protein levels and reverse transcription qPCR analysis to assess mVenus RNA levels. Neither mVenus RNA nor protein were detected in these cells (Figure 4.11).

The same α-globin promoter-mVenus sequence was inserted into either LP4 or LP5 of AL2R2 ES cells using CRISPR-Cas9 HDR; LP4 and LP5 are ~73 and ~81 kb downstream of the inserted R2 in LP2 respectively (Chapter 3: Figure 3.6). Even though the inserted R2 was sufficient to form a regulatory domain at AL2, I wanted to investigate if the erythroid-specific enhancer and promoter pair could produce mVenus RNA/protein at this locus. Moreover, I wanted to investigate whether inserting the α-globin promoter-mVenus sequence both within and outside the regulatory domain had an impact on expression; LP5 lies just outside the domain observed in AL2R2 erythroid cells, whereas LP4 lies just within the 3' edge. In addition, one of the novel enhancer-like elements observed in AL2R2

erythroid cells is within LP4 and I thought that inserting the α-globin promoter-mVenus sequence close to accessible chromatin may increase the likelihood of expression of the insertion. However, FACS and reverse transcription qPCR analysis of AL2R2 embryoid bodies containing the inserted α-globin promoter-mVenus sequence in either LP4 or LP5 again revealed no mVenus RNA or protein in these cells (Figure 4.11). This suggests that the inserted R2 was not behaving as an enhancer to the inserted promoter in any of these permutations of artificial loci; possibly due to the genomic distance between the elements or that R2 cannot act as an enhancer alone in these situations.

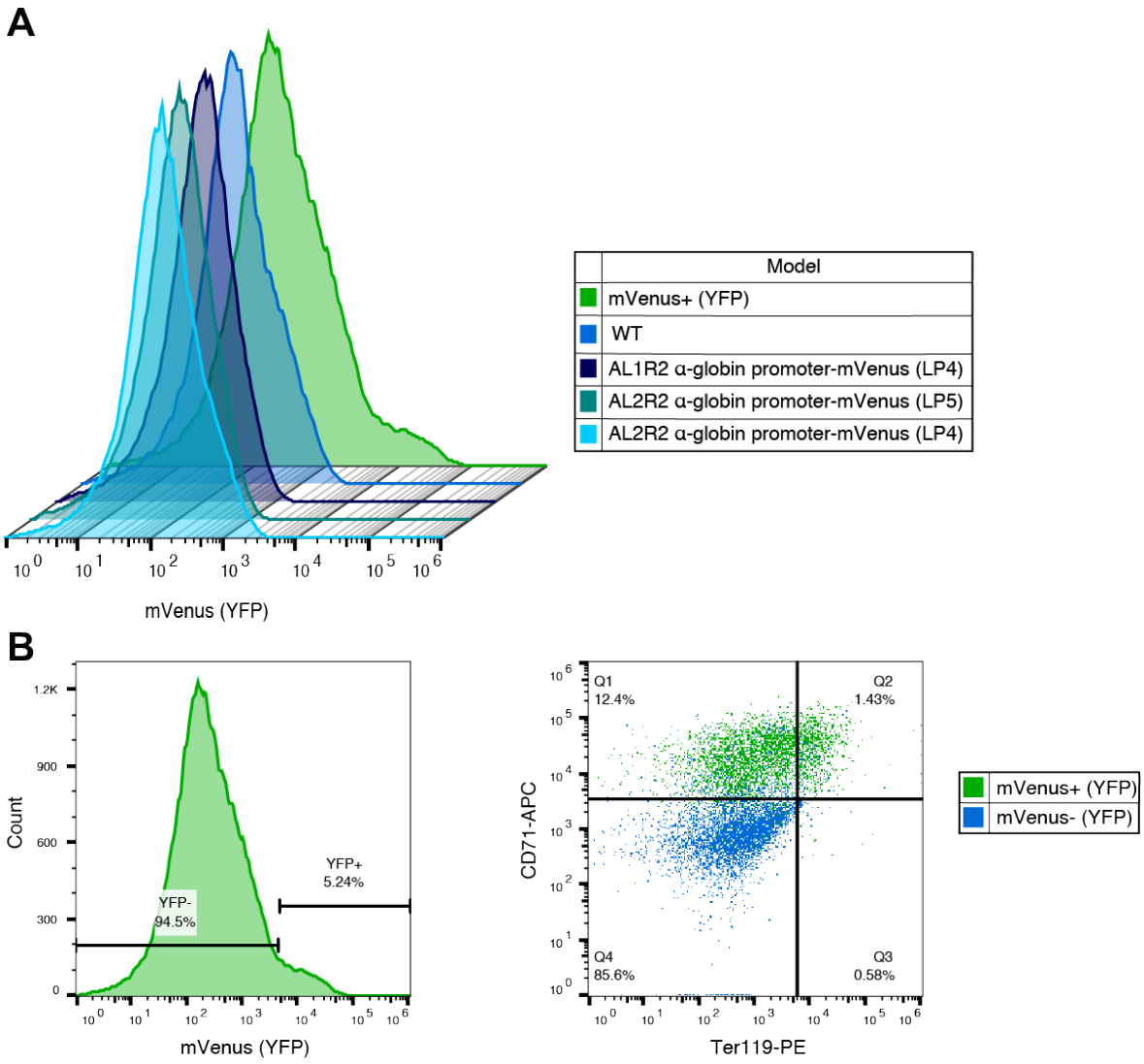

A
Model
mVenus+ (YFP)
WT
AL1R2 α-globin promoter-mVenus (LP4)
AL2R2 α-globin promoter-mVenus (LP5)
AL2R2 α-globin promoter-mVenus (LP4)
mVenus (YFP)
B
Count
YFP-
94.5%
YFP+
5.24%
mVenus (YFP)
Q1
12.4%
Q2
1.43%
Q4
85.6%
Q3
0.58%
CD71-APC
Ter119-PE
mVenus+ (YFP)
mVenus- (YFP)
C
mVenus RNA
WT
mVenus+ (YFP)
AL1R2 (LP4)
AL2R2 (LP5)
AL2R2 (LP4)

Figure 4.11: Analysis of mVenus expression by FACS and reverse transcription qPCR in embryoid bodies containing an insertion of the enhancer R2 and the α-globin promoter in artificial loci. A: *FACS analysis of mVenus protein levels in WT embryoid bodies, embryoid bodies in which Hba-α1 had been tagged with mVenus (positive control; provided by Helena S. Francis, WIMM), and AL1R2 and AL2R2 embryoid bodies with an insertion of the α-globin promoter-mVenus sequence.* ***B:*** *FACS analysis of mVenus positive control embryoid bodies. The cells that were positive for mVenus were also positive for erythroid markers (CD71 and Ter119).* ***C:*** *Reverse transcription qPCR analysis to detect mVenus RNA in WT embryoid bodies, mVenus positive control embryoid bodies, and AL1R2 and AL2R2 embryoid bodies with an insertion of the α-globin promoter-mVenus sequence. Mean and standard deviation of mVenus RNA shown. Data normalised to 18S RNA. Mean fluorescence and expression for multiple replicates shown: three biological replicates for WT, three clones for mVenus positive control, eight clones for AL1R2 (LP4), six clones for AL2R2 (LP5), and eleven clones for AL2R2 (LP4). (Figure on previous page.)*

Discussion

To investigate the mechanisms governing the formation of tissue-specific regulatory domains, I have inserted predicted *cis*-regulatory elements of the mouse α-globin locus into two neutral loci within the mouse genome. Using an *in vitro* haematopoietic differentiation system, I assessed the tissue-specific activity of these elements and their impact on local histone modifications, transcription, and chromosome conformation outside of their native, highly-conserved chromosomal context. Insertion of the R2 enhancer displayed tissue-specific chromatin accessibility which was variable between the two loci in erythroid cells, with more stable accessibility observed in AL2R2 erythroid cells. Furthermore, the inserted R2 in AL2R2 erythroid cells caused local changes in accessibility of the native chromatin in AL2 which resulted in novel ATAC-seq peaks. These peaks appeared as active enhancer-like elements based on chromatin signatures in that there were associated with the histone modifications H3K27ac and H3K4me1, whereas the inserted R2 had the chromatin marks of an active promoter (H3K27ac, H3K4me1, and H3K4me3). The inserted R2 was transcribed to produce both non-polyadenylated and polyadenylated unidirectional transcripts (lncRNAs), therefore in this locus R2 was behaving as a promoter-like element. Additionally, the activity of the inserted R2 in erythroid cells was sufficient to form a regulatory domain, in which enriched interactions between the inserted R2 and the enhancer-like elements encompassed the region of active transcription. Interestingly, in both loci, the combination of R2 and the α-globin promoter did not lead to expression of a fluorescent reporter from the α-globin promoter in erythroid cells.

The results of the inserted R2 in AL2R2 erythroid cells suggests that strong enhancers have the ability to recruit both transcription activation factors and regulatory domain formation factors in a tissue-specific manner outside of their native chromosomal context. The ~75 kb

non-polyadenylated and ~4 kb polyadenylated lncRNAs produced from the inserted R2 and the chromatin interactions spanning the region of active transcription could be the result of these inherent functions of the R2 enhancer in combination with the chromatin context of AL2. In the WT locus, the α-globin enhancers are positioned within the introns of *Nprl3* which is expressed in erythroid cells. The α-globin enhancers are transcribed to produce bidirectional eRNAs that are ~1 kb in length (Kowalczyk et al., 2012; Larke et al., in preparation). Upon deletion of the *Nprl3* promoter, the enhancers were revealed to be able to produce polyadenylated and spliced lncRNAs (Kowalczyk et al., 2012). The observation that a tissue-specific α-globin regulatory domain still formed in erythroid cells lacking the α-globin promoters also complements the idea that the enhancers are key drivers of regulatory domain formation (Brown et al., 2018). Therefore, this provides two mechanisms by which enhancers can regulate transcription of their target promoters: by bringing them into close proximity via regulatory domain formation and stimulating their transcriptional activation.

The novel enhancer-like elements observed in AL2R2 erythroid cells are reminiscent of the cluster of enhancers found at the α-globin locus. The five α-globin enhancers meet the criteria of a super-enhancer and was found to be the second highest scoring super-enhancer in erythroid cells (Hay et al., 2016). However, the individual elements of the super-enhancer are not functionally equivalent (Hay et al., 2016); deletion of R1 or R2 had a significant impact on the synthesis of α-globin RNA in erythroid cells, deletion of R4 had a minor impact, and deletion of R3 or Rm had no detectable effect. In addition, R3, Rm and R4 did not display significant enhancer activity in a well-established mouse transgenic system. This implies that R3, Rm and R4 are either very weak enhancers or not enhancers at all, despite their chromatin marks. Instead, these enhancers may be similar to the enhancer-like elements

I observed in AL2R2 erythroid cells in that they could be the result of the activation of a strong enhancer nearby. However, it may be that once these 'secondary' enhancer-like elements have been activated by a strong enhancer, they play a stabilising role in transcription and/or regulatory domain formation possibly through the binding of tissue-specific transcription factors. These results can provide insights into super-enhancer evolution; why enhancers appear in clusters and why there are ~100 fold more enhancers than genes in the genome (Lindblad-Toh et al., 2011).

The lack of mVenus expression from the α-globin promoter inserted in combination with R2 in AL1R2 and AL2R2 erythroid cells was an interesting and somewhat unexpected result. Mice have been generated in which four out of the five enhancer elements have been deleted in *cis* at the α-globin locus, and the only remaining enhancer is R2 (Helena S. Francis, WIMM; unpublished). Heterozygotes for the R2-only model were viable and able to breed normally. However, cross-breeding of heterozygotes gave reduced litter sizes and no homozygotes for the R2-only allele. Analysis of the α-globin:β-globin mRNA ratio in the embryonic blood (E10.5) of homozygous R2-only embryos showed 15% of the α-globin mRNA level found in WT embryonic blood. Even though homozygote R2-only mice were not viable there was still detectable α-globin mRNA. Therefore, we predicted that there would be low but readily detectable expression of the fluorescent reporter in the above models. The variable activity of R2 in AL1R2 erythroid cells could explain the lack of expression from the α-globin promoter, however this does not account for the observations in AL2R2 erythroid cells. The distance between the inserted R2 and the inserted α-globin promoter in AL1/2 was larger than the distance between R2 and the promoter of *Hba-α1* in the WT locus, which could be a contributing factor (~33 kb between the elements in the WT locus, ~53 kb between the elements in AL1, and ~73 kb or ~81 kb between the elements in

AL2). Therefore, future experiments could include inserting the α-globin promoter-mVenus sequence closer to the inserted R2, and not necessarily within an LP, in order to recapitulate the WT locus as faithfully as possible. The WT locus has two α-globin promoters whereas there was only one in the artificial loci, so there could be more recruitment of transcriptional machinery within the regulatory domain resulting in successful production of transcripts in the WT locus. Additionally, it may be that the neutral chromatin of the artificial loci does not create as permissive an environment as the chromatin at the WT locus. It may be that the sequence selected to contain the minimal α-globin promoter was not sufficient and further testing of the minimal promoter sequence is required. It would be interesting to assess whether the inserted α-globin promoter in these artificial loci becomes as accessible as the promoter in its natural environment by performing ATAC-seq in erythroid cells. Finally, it may be that the new transcripts originating from the inserted R2 in AL2R2 erythroid cells could be interfering with the inserted promoter and preventing transcription. Possibly, the transcription of the inserted R2 may prevent the appropriate interactions between the enhancer and the inserted promoter from forming, however this needs further investigation.

Acknowledgements

Helena S. Francis (WIMM) provided advice for all stages of ES cell culture, transfection, and differentiation, and helped perform the histone modification ChIP-seq experiments in WT CD71+ erythroid cells. Kevin Clark (FACS Facility, WIMM) performed the FACS analysis to isolate successfully transfected ES cells (eGFP+). Ron Schwessinger performed the chromatin segmentation analysis. Matthew E. Gosden (WIMM) performed the ATAC-seq in WT CD71+ erythroid cells and helped perform the Capture-C experiment. Jelena Telenius (WIMM) helped perform computational analysis for the Capture-C experiment.

Chapter 5: Discussion

Previous studies investigating the mechanisms and elements involved in the formation of regulatory domains have mostly focused on perturbing existing loci, either by disrupting individual *cis*-regulatory elements at specific loci (Bender et al., 2012; Brown et al., 2018; Dowen et al., 2014; Franke et al., 2016; Gómez-Marín et al., 2015; Hanssen et al., 2017; Hay et al., 2016; Lee et al., 2017; Lupiáñez et al., 2015; Narendra et al., 2015; 2016; Nora et al., 2012; Symmons et al., 2016) or by abolishing the architectural proteins which bind them (Hyle et al., 2019; Nora et al., 2017; Rao et al., 2017; Schwarzer et al., 2017; Wutz et al., 2017). A potential caveat in such studies is that unknown elements and mechanisms may contribute to the formation of regulatory domains and such combinatorial effects might be difficult to dissect.

While using conventional approaches at the well-characterised mouse α-globin locus, I have also developed a different strategy to investigate how a regulatory domain may be established and maintained. Having defined the fundamental *cis*-regulatory elements which contribute to the mouse α-globin regulatory domain, I have inserted these into neutral regions of chromatin to generate artificial loci, and I have thereby started to determine which elements are sufficient to build a tissue-specific regulatory domain. This allows me to investigate the elements and mechanisms required to form a regulatory domain away from the potentially confounding effects within an established regulatory landscape.

Cis-regulatory elements play context-dependent interchangeable roles within regulatory domains

My investigation into the function of CTCF binding sites downstream of the mouse α-globin regulatory domain found that a domain still forms in the absence of CTCF binding at the sites HS+44 and HS+48. However, in the absence of these elements the domain expands to include chromatin downstream of the deleted CTCF binding sites. This suggests that these sites behave as a boundary; delimiting active interactions rather than playing a role in forming the regulatory domain. This agrees with previous studies which have found that CTCF binding sites behave as functional boundary elements at the borders of regulatory domains (Dowen et al., 2014; Flavahan et al., 2016; Hanssen et al., 2017; Hnisz et al., 2016; Narendra et al., 2015; 2016). Interestingly, my analyses of mice harbouring deletions of CTCF binding sites within the α-globin domain found that the α-globin genes themselves also behave as boundary elements.

In erythroid cells, the promoter of the α-globin gene closest to the enhancers (*Hba-α1*) shows preferential interactions with the enhancers compared to the promoter of the more distal α-globin gene (*Hba-α2*) (Davies et al., 2016). In addition to having differential interactions with the enhancers, I found that the mouse α-globin genes are differentially expressed; *Hba-α1* accounts for ~66% of the total α-globin mRNA in erythroid cells and *Hba-α2* accounts for ~34%. This indicates that the preferential expression of *Hba-α1* could be due to its preferential interactions with the enhancer elements. As the θ CTCF binding sites are located within the α-globin domain, we thought that these CTCF binding sites could be regulating the differential interactions and expression of the α-globin genes (in particular, θ1 as it is situated in between the α-globin genes). However, both the preferential interactions and expression of *Hba-α1* were maintained in erythroid cells lacking CTCF

binding at these sites, suggesting that *Hba-α1* could be behaving as a partial boundary through promoter competition, in that *Hba-α2* has reduced enhancer interactions and expression relative to *Hba-α1*. Importantly, the 3' edge of the α-globin regulatory domain (immediately downstream of *Hba-α2*) was preserved in the absence of CTCF binding just downstream of the active genes. This suggests that the distal α-globin gene (*Hba-α2*) is the 3' boundary of the domain rather than the CTCF site playing this role. Previous studies have indicated that promoters of actively transcribed genes might behave as boundary elements (Brandão et al., 2019; Dixon et al., 2012), however, our data provide direct evidence for this.

Further support for promoters behaving as boundary elements comes from a naturally occurring mutation in the human α-globin locus. A SNP which creates a new promoter between the enhancers and the α-globin genes is transcriptionally active and appears to downregulate expression of the α-globin genes in *cis* by outcompeting and "stealing (enhancer) activity" away from the promoters (De Gobbi et al., 2006). Work within the group has found that this new promoter sequence interacts with the α-globin enhancers and causes a reduction in interactions between the enhancers and the more distal α-globin genes, contributing to their downregulation (Bozhilov et al., in preparation).

An additional set of relevant observations, is that when there are more than two α-globin genes present within the locus (αα, ααα, αααα etc), the additional genes lying further downstream appear to be expressed at ever decreasing levels (Vestri et al., 1994). This suggests that the promoters of the more proximal α-globin genes act as boundaries between the enhancers and the more distal α-globin genes. This configuration might reduce the effect of the enhancers on the expression of the more distal genes via reduced enhancer-promoter interactions. This suggests that a promoter may have multiple roles in shaping the structure

of a regulatory domain, one of which is to act as a structural boundary element which may reduce the access of more distal promoters to enhancers.

Of interest, multi-way chromatin interaction analysis of erythroid cells isolated from the D3839 model (Hanssen et al., 2017), showed that the interactions between the enhancers and the α-globin promoters do not decrease upon incorporation of the three upstream genes into the regulatory domain. By contrast they become highly upregulated under the influence of the α-globin enhancers. We found that the multiple gene promoters (globin and non-globin) simultaneously interact with the five clustered α-globin enhancers at a single allele to form a hub-like regulatory domain. This suggests that the context of the promoter determines whether or not it acts as a boundary element. In the case of the D3839 model, the newly incorporated promoters lie upstream of the enhancers and not between the enhancers and the downstream α-globin genes. This suggests that active promoters can behave as boundary elements when they are situated between an enhancer and another more distal target gene.

In addition to showing that promoters can behave as boundary elements, I also found that enhancers can behave as tissue-specific promoters when positioned outside of their native chromosomal context. When the strongest enhancer element (R2) from the α-globin locus (Hay et al., 2016) was inserted into a nearly neutral chromatin environment at AL2, it appeared to be more promoter-like in its chromatin signature compared to that of the endogenous R2 on chromosome 11. Additionally, the inserted R2 produced both non-polyadenylated and polyadenylated unidirectional transcripts (~75 kb and ~4 kb respectively) that were considerably longer than the bi-directional eRNAs produced by the

endogenous R2 enhancer element (~1 kb; Kowalczyk et al., 2012; Larke et al., in preparation).

There is growing evidence that enhancers also have some inherent promoter capacity. *In vitro* studies using massively parallel cell culture-based reporter assays to test promoter activity across thousands of small DNA elements (Nguyen et al., 2016) or across the entire human genome (van Arensbergen et al., 2017), suggested that many enhancers can act as weak autonomous promoters, although at levels ~10-fold lower than that of annotated gene promoters. An *in vivo* study in *Drosophila* found that the balance between enhancer and promoter activity was correlated with the levels and directionality of eRNA transcription originating from the *cis*-regulatory element and that this was most likely to be an inherent sequence property of the elements themselves (Mikhaylichenko et al., 2018). Promoters with high levels of unidirectional transcription functioned mainly as orientation-dependent promoters (with little or no enhancer function) in the orientation in which they were transcribed. Whereas *cis*-regulatory elements (either enhancers or promoters) with high levels of bidirectional transcription functioned as both enhancers and promoters in an orientation-independent manner, and more distal, intergenic enhancer elements with low levels of bidirectional transcription mainly functioned as enhancers, with a subset having weak orientation-independent promoter activity (Mikhaylichenko et al., 2018). This suggests there is a continuum of enhancer and promoter activity of *cis*-regulatory elements which matches the continuum of their endogenous transcription.

Upon deletion of the *Nprl3* promoter at the endogenous mouse α-globin locus, the intragenic α-globin enhancers were shown to act as alternative promoters, regulating unidirectional transcription in the direction of the host gene *Nprl3* (minus (-) strand) to produce lncRNAs

that were abundant, stable (polyadenylated), and spliced (Kowalczyk et al., 2012). Conversely, genome-wide analysis of intergenic enhancers in mouse and *Drosophila* found no evidence that they produced stable, long transcripts like those described above (De Santa et al., 2010; Kim et al., 2010; Mikhaylichenko et al., 2018), suggesting that RNA polymerase II elongation may be fundamentally different at intergenic and intragenic enhancers. However, whether this is due to sequence-dependent differences or context-dependent differences is not clear. Therefore, this indicates that intragenic enhancers, rather than intergenic enhancers, may be more likely to display promoter-like behaviours. Interestingly, the inserted R2 in AL2 produced a long polyadenylated transcript which was transcribed in the opposite direction (plus (+) strand) to that of the endogenous R2. This is an unusual observation as the normally intragenic-located enhancer displayed promoter-like behaviour when inserted in an intergenic location. It may be that the context of the nearly neutral chromatin of AL2 allowed the inserted R2 enhancer to change behaviour to become more promoter-like. Possibly, the novel enhancer-like elements in AL2R2 erythroid cells generated a permissive chromatin environment that contributed to the histone modifications at the inserted R2, and to the production of unidirectional transcripts in the plus (+) strand orientation. Alternatively, it may be that something inherent to the R2 enhancer sequence contributed to this behaviour. However, insertion of the same R2 enhancer sequence into AL1 led to inconsistent chromatin accessibility of the element in erythroid cells, suggesting that the context of AL1 was not permissive for the promoter-like behaviour observed in AL2.

To investigate in further detail this relationship between the R2 enhancer and the chromatin context in which it resides, we plan to insert the R2 sequence in its native orientation and reverse orientation in different LPs in AL2. It would be interesting to see if insertion of the

R2 sequence downstream of the 'primed' enhancer elements in AL2 (e.g. LP5) or insertion of its reverse complement sequence in LP2 causes a change in the direction of both transcriptional activation and regulatory domain formation. Additionally, we are collaborating with Jef Boeke (Institute for Systems Genetics, NYU Langone Health), whose research group specialises in the high-throughput synthesis of large DNA constructs, in order to generate multiple permutations of AL2 to address this question and others about the potential mechanisms of regulatory domain formation in a much more systematic manner.

Insights into enhancer evolution and the mechanisms for regulating gene expression

Activation of the inserted R2 enhancer element in AL2 led to multiple interesting observations providing insights into enhancer evolution and the potential mechanisms by which they regulate transcription. AL2R2 erythroid cells displayed local changes in chromatin accessibility around the inserted sequence on chromosome X, resulting in novel ATAC-seq peaks up to ~75 kb downstream of the insertion. These novel ATAC-seq peaks are associated with histone modifications characteristic of active enhancers (H3K27ac and H3K4me1) and were defined as weak enhancers by chromatin segmentation analysis. As discussed in Chapter 4, these enhancer-like elements appear similar to the enhancers R3, Rm, and R4 at the endogenous α-globin locus and may play a role in creating a permissive environment which results in stabilising transcription and/or the formation of a regulatory domain. Support for this idea comes from unpublished results generated within the group. It was predicted that mice harbouring deletions of four out of the five enhancer elements, so that only R2 remained and was driving the expression of α-globin, would produce ~50% of the α-globin mRNA level found in WT animals. This was due to the observation that the

regulatory effect of the α-globin enhancer cluster was additive and depended mainly on two elements, R1 and R2, with very little/no contribution from the other three elements to enhancer activity (Hay et al., 2016). Therefore, the R2-only model described above was expected to display roughly the same enhancer activity as mice with a deletion of the R1 element. However, analysis of α-globin mRNA from the R2-only model showed ~15% of the level found in WT animals (Helena S. Francis, WIMM; unpublished). Therefore, despite the R3, Rm, and R4 elements appearing to have very weak or no enhancer activity (Hay et al., 2016), they caused a greater reduction in α-globin mRNA when deleted in combination with R1 than expected. This suggests they could be contributing to α-globin expression indirectly by altering the local chromatin environment to make it more permissive for the processes mentioned above.

It has been suggested that transcriptional activation takes place in nuclear condensates which contain a high concentration of transcription factors, co-factors, and components of the basal transcriptional machinery recruited by enhancer elements (Boehning et al., 2018; Boija et al., 2018; Cho et al., 2018b; Chong et al., 2018; Sabari et al., 2018). These condensates have been suggested to form by liquid-liquid phase separation of proteins with disordered regions (Hnisz et al., 2017). Clusters of multiple enhancers would recruit high concentrations of such proteins and condensates that contain Mediator have been found at clustered enhancers (Sabari et al., 2018). It may be that condensates are a reflection of regulatory domains in which multiple *cis*-regulatory elements interact simultaneously and function together in a hub-like complex or *vice versa*. Therefore, my results from AL2R2 erythroid cells could provide insights into why enhancers appear in clusters. There may be parts of the genome that contain 'primed' enhancer elements but only become active when in the close vicinity of a strong active enhancer element in the correct tissue type. AL2 may be one such area of

the genome; in erythroid cells, insertion of the R2 enhancer may have caused a sufficiently high local concentration of factors through binding at R2 and the other enhancer-like elements which led to transcriptional activation and regulatory domain formation. This indicates that strong enhancers, such as R2, (possibly in combination with other weaker enhancer elements) may have at least two fundamental, closely intertwined mechanisms for regulating expression of their target promoters: by bringing them into close proximity by recruiting factors that result in regulatory domain formation and stimulating their transcriptional activation by recruiting activators and transcriptional machinery.

Molecular mechanisms underlying chromosome organisation

The interaction profile of the regulatory domain at AL2 in AL2R2 erythroid cells, from the viewpoint of the inserted R2, showed interactions extending over the entirety of the actively transcribed region that were not specifically focused at the novel enhancer-like elements. This suggests that this structure was produced by a scanning mechanism such as the proposed loop extrusion model for domain formation described in Chapter 1 (Fudenberg et al., 2016). In addition, my analyses in the mouse model with engineered deletions of the CTCF binding sites downstream of the α-globin locus also provides support for the loop extrusion model. Removal of these elements resulted in an expansion of the α-globin regulatory domain to include the region of chromatin downstream of the deleted CTCF binding sites. These results therefore support a model of regulatory domain formation in which extruding factors are recruited to chromatin near active *cis*-regulatory elements and translocate along chromatin to bring together progressively more distant loci along a chromosome, thereby extruding dynamically expanding loops of chromatin. Extruding factors are probabilistically halted at genomic loci that are barriers to extrusion, such as

CTCF-bound elements or promoters of actively transcribed genes, which form transient domain boundaries.

As described in Chapter 1, loop extrusion could explain many observed aspects of chromosome organisation and indeed, simulations of loop extrusion with directional and partially permeable barriers (Fudenberg et al., 2016; 2017; Sanborn et al., 2015) can reproduce a broad range of features seen in Hi-C maps. These include stripes along TAD edges which form when an extruding factor reaches a single barrier in one direction and continues to track the chromatin fibre in the other direction (Vian et al., 2018), connecting this site to downstream loci, and peaks at TAD corners between convergently oriented CTCF-bound elements (Rao et al., 2014). However, both the strength and the complexity of interactions between enhancers and the promoters of their target genes, such as those observed at the α-globin locus, cannot be explained by a loop extrusion mechanism alone. Chromatin domains that are believed to be formed by loop extrusion, such as TADs, are characterised by a mere ~two-fold increase in interactions within the domain compared to interactions between neighbouring domains (Dekker & Mirny, 2016); this cannot account for the strongly enriched interactions between enhancers and promoters within these domains. It is likely that the two-fold difference in interaction frequencies resulting from loop extrusion is further strengthened by other mechanisms that stabilise specific interactions between *cis*-regulatory elements. One such mechanism could involve protein-protein interactions between transcription factors bound at enhancers and the PIC at the promoter. This idea is supported by our Tri-C analyses of the α-globin locus in erythroid cells isolated from D3839 mice which revealed that multiple gene promoters (globin and non-globin) can simultaneously interact with multiple α-globin enhancer elements at a single allele to form a hub-like regulatory domain. Interestingly, there was a trend for

increased interactions contributing to the α-globin hub in the D3839 cells compared to WT cells. This suggests that the inclusion of the three upstream genes into the regulatory hub in D3839 cells may have increased the local concentration of factors which may have, in turn, strengthened interactions mediated by multi-protein complexes. Further support for this model comes from the suggestion that multi-molecular assemblies formed through phase separation of interacting proteins bound at *cis*-regulatory elements may strengthen interactions between enhancers and promoters (Hnisz et al., 2017).

My observations at the α-globin locus and AL2 are supported by genome-wide analyses of chromatin architecture upon perturbations of the loop extrusion machinery using systems that result in rapid, acute depletion of target proteins, such as the auxin-inducible degron (AID) system (Nishimura et al., 2009). Studies that used the AID system to deplete CTCF showed a loss of TADs across the genome (Hyle et al., 2019; Nora et al., 2017; Wutz et al., 2017). However, the removal of CTCF did not alter the genome-wide relationship between contact frequency and linear genomic distances between loci, indicating a similar level of chromatin packaging upon loss of CTCF. Depletion of chromatin-associated cohesin, either by deleting its loading factor *Nipbl* (Schwarzer et al., 2017) or by direct AID-mediated depletion (Rao et al., 2017; Wutz et al., 2017), resulted in a similar loss of TAD structure, but also affected the rate of contact frequency decay with genomic distance. Therefore, these results suggest that both of these factors are required for the formation of TADs in a model of loop extrusion by cohesin delimited by CTCF-bound sites. As CTCF only plays an instructive role for the extrusion activity, loss of CTCF results in perturbed domain structures but does not alter contact frequency decay. Whereas without cohesin, the process of extrusion can no longer occur resulting in both the loss of TADs and an altered relationship between contact frequency and genomic distance. This is supported by a study

that used microscopy to generate distance maps for extensive regions of chromatin before and after cohesin depletion. Upon cohesin depletion, intra-TAD distances increased resulting in decompaction of TADs which is consistent with the loss of extrusion (Bintu et al., 2018). Interestingly, segregation of chromosomes into active (A) and inactive (B) compartments, however, was unaffected by the loss of CTCF (Hyle et al., 2019; Nora et al., 2017; Schwarzer et al., 2017), but the compartmentalisation pattern became stronger upon the loss of chromatin-associated cohesin and, after the disappearance of TADs, revealed a finer compartment structure with shorter compartmental domains (Rao et al., 2017; Schwarzer et al., 2017; Wutz et al., 2017). This finer structure more accurately reflected transcriptional activity and chromatin state. This suggests that chromosome architecture is the result of at least two independent mechanisms: CTCF/cohesin-mediated formation of self-interacting domains by loop extrusion and CTCF/cohesin-independent segregation of the genome into compartments defined by chromatin state and transcriptional activity. Consistent with this model, an increase in cohesin activity, resulting from the deletion or depletion of the cohesin release factor WAPL, increased the strength and size of TADs and reduced compartmentalisation (Haarhuis et al., 2017; Wutz et al., 2017).

Evidence for a mechanism of loop extrusion underlying the CTCF/cohesin-dependent mode of chromosome folding is accumulating (Fudenberg et al., 2017), but the molecular mechanisms contributing to the formation of A and B compartments remain unknown. As described in Chapter 1, it is thought that separation of different chromatin states into compartments may be driven by phase separation. Simulations have demonstrated that a combination of loop extrusion and compartmental segregation by phase separation can faithfully reproduce Hi-C data (Nuebler et al., 2018), thus supporting this hypothesis. In addition, these simulations show that the active process of loop extrusion may antagonise

the proper phase separation of a block copolymer which suggests that loop extrusion weakens compartmentalisation by locally mixing compartmental domains of a different type, having a particularly strong impact on short (<100 Kb) euchromatic or heterochromatin regions (Nuebler et al., 2018). This suggests that even though compartments were originally described as megabase-scale chromatin structures, compartmental separation may be able to take place at much smaller scales than initially appreciated. Further support for this originates from the finer compartmental structures observed without the overlapping TADs upon cohesin loss (Rao et al., 2017; Schwarzer et al., 2017; Wutz et al., 2017). Therefore, this segregation mechanism which brings together genomic regions of similar epigenetic signatures might not only contribute to the formation of large-scale structures defined by their chromatin state (such as A and B compartments) but may also stabilise enhancer-promoter interactions within regulatory domains at a much smaller scale. The observation that cohesin-independent interactions are highly enriched for 'super-enhancers' supports the idea that this segregation can take place at the level of individual *cis*-regulatory elements (Rao et al., 2017). Compartmentalisation across different scales could contribute to the formation and/or maintenance of interactions between *cis*-regulatory elements based on their specific chromatin signatures and may be why regulatory domains have a hub-like structure (Allahyar et al., 2018; Oudelaar et al., 2018). This would also help explain why the loss of CTCF or chromatin-associated cohesin did not lead to widespread mis-regulation of gene expression (Hyle et al., 2019; Nora et al., 2017; Rao et al., 2017; Schwarzer et al., 2017; Wutz et al., 2017), but genes that were affected were most likely regulated by the structure imposed by TADs, as described in Chapter 1, which shows that loop extrusion plays an important role in guiding communication between *cis*-regulatory elements over long distances. A summary of the potential proteins and mechanisms that control genome organisation are presented in Figure 5.1.

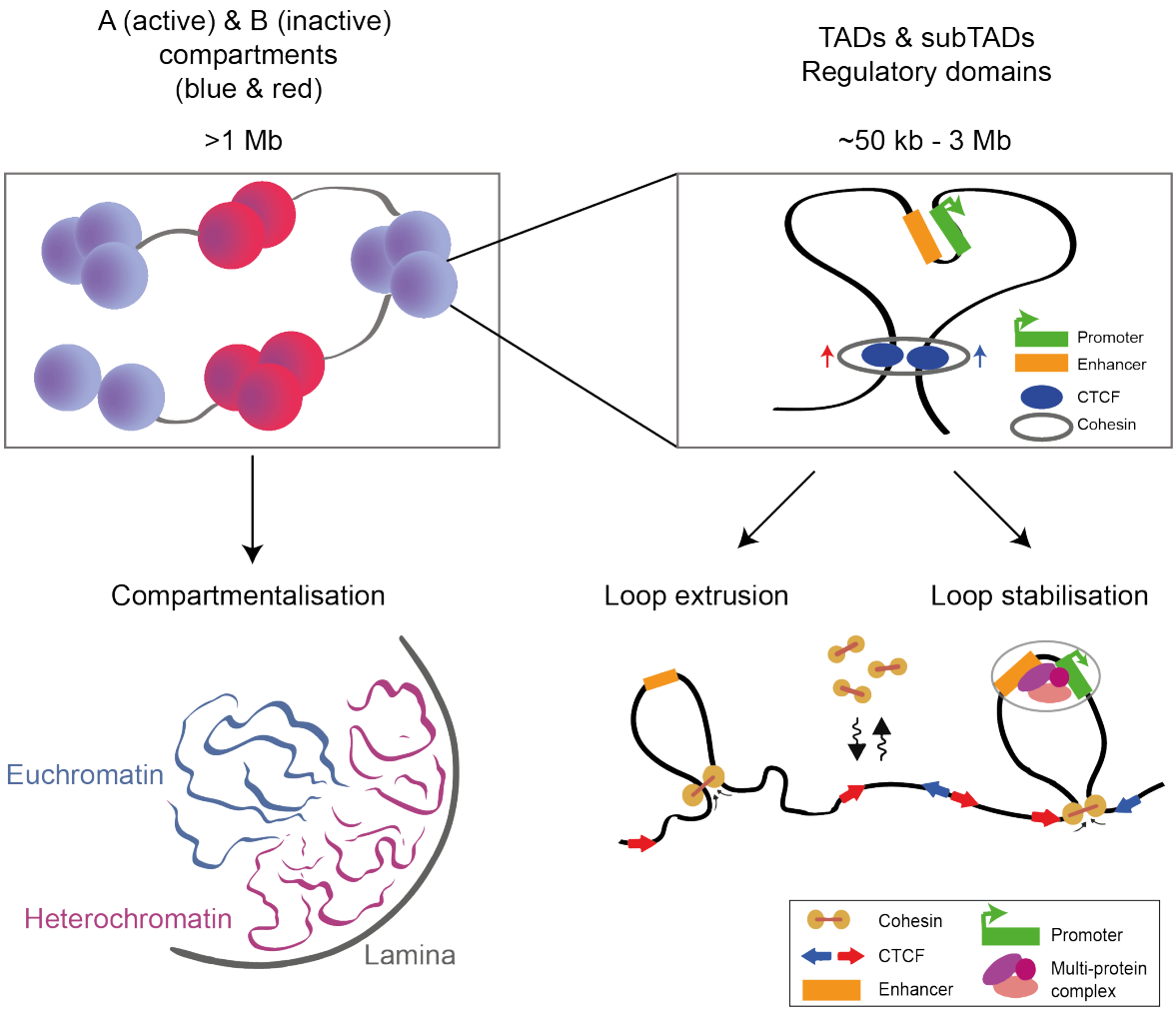


Figure 5.1: A summary of proposed proteins and mechanisms involved in the formation of self-interacting chromatin structures. *The separation of different chromatin states into megabase-scale A (active euchromatin; blue) and B (inactive heterochromatin; red) compartments is thought to be driven by weak attractions between chromatin regions with similar epigenetic signatures, resulting in phase separation of compartments. In contrast, smaller self-interacting domains, such as TADs and regulatory domains, are thought to form through an active process of loop extrusion. My data are supportive of a model in which loop extruding factors, such as cohesin, are recruited to active* cis*-regulatory elements, such as enhancers, to form these structures. Complex higher-order structures, in which multiple enhancers and promoters interact, are formed within these domains by distinct tissue-specific mechanisms, possibly involving the formation of phase-separated protein assemblies of transcriptional machinery, such as transcription factories or sub-nuclear condensates (grey circle).*

As the insertion of the R2 enhancer element into AL2 was sufficient to form a tissue-specific regulatory domain that most likely resulted from loop extrusion, this suggests that extrusion factors, potentially cohesin, are recruited to active *cis*-regulatory elements such as enhancers (Figure 5.1). Support for this hypothesis comes from a recent study assessing cohesin binding and chromatin interactions upon WAPL depletion (Liu et al., 2019). This study found that regions that lost cohesin binding and enhancer-promoter interactions are associated with lineage-specific genes in mouse ES cells. These regulatory regions are enriched for active enhancer marks and lineage-specific transcription factor binding, but largely devoid of CTCF binding sites. This suggests that cohesin is recruited at lineage-specific enhancers and extrudes chromatin to form regulatory domains containing tissue-specific enhancer-promoter interactions. However, in the case of the AL2 regulatory domain in AL2R2 erythroid cells, I am yet to complete experiments investigating what the potential loop extruding factor is at this locus. As the interaction patterns of this regulatory domain span the entire region that is actively transcribed, it would be interesting to look at both RNA polymerase II occupancy and cohesin binding at this locus. In collaboration with Nigel Roberts, we are tagging Rad21, a cohesin subunit, with a Strep-tag II in AL2R2 and WT ES cells to enable us to perform quantitative ChIP-seq of this protein to investigate cohesin loading dynamics at AL2. We are also interested to see if there is an additive/synergistic effect of inserting the second strongest enhancer element, R1, in combination with R2 at AL2 on both transcriptional activation and regulatory domain formation and aim to perform the same experiments performed in this thesis in a model harbouring both insertions.

Conclusion

The in-depth analysis of deletions of CTCF binding sites both within and downstream of the mouse α-globin regulatory domain, and the generation of artificial loci using the well-characterised α-globin elements outside of their native chromosomal context contribute to our fundamental understanding of the formation of regulatory domains, providing insights into the rules guiding three-dimensional gene regulation.

Chapter 6: Materials and Methods

Cells

Isolation of erythroid cells derived from adult mouse spleen

Primary Ter119+ erythroid cells were obtained from the spleens of adult mice treated with phenylhydrazine. Phenylhydrazine treatment causes haemolytic anaemia which results in an expansion of erythroid cells in the spleen (>80% of the cells are CD71+/Ter119+) (Spivak, Toretti, & Dickerman, 1973). Phenylhydrazine treatment was carried out by Jacqueline A. Sharpe (WIMM). Mice were treated with three doses of phenylhydrazine (40 mg/g body weight) given 12 hours apart and were sacrificed on day five after the first dose. Spleens were mechanically dissociated into single cells suspensions in cold phosphate-buffered saline (PBS; Gibco: 10010023)/10% fetal bovine serum (FBS; Gibco: 10270106) and passed through a 70 μm filter to remove clumps. Cells were washed with cold PBS/10% FBS and resuspended in 10 μl of cold PBS/10% FBS per 10^6 cells and stained with a 1/100 dilution of anti-Ter119-PE antibody (Miltenyi Biotec: 130-102-336) at 4 °C for 20 minutes. Stained cells were washed with cold PBS/10% FBS and resuspended in 8 μl of cold PBS/0.5% BSA/2 mM EDTA and 2 μl of anti-PE MACS microbeads (Miltenyi Biotec: 130-048-801) per 10^6 cells and incubated at 4 °C for 15 minutes. Ter119+ cells were positively selected via MACS lineage selection columns (Miltenyi Biotec: 130-042-401) and processed for downstream applications. Purity of the isolated erythroid cells was routinely verified by Fluorescence-Activated Cell Sorting (FACS).

Culture of mouse embryonic stem cells

Wild type mouse ES cells were derived from the 129/Ola mouse strain and gave rise to the male HPRT-deficient ES-E14TG2a cell line (J. Nichols, Evans, & Smith, 1990; Smith,

1991). The mouse ES cell line used in this thesis is ES-E14TG2a.IV; a subclone of the ES-E14TG2a line that is cultured feeder-free (provided by Austin Smith, Wellcome/MRC Cambridge Stem Cell Institute). Mouse ES cells were routinely cultured on 0.1% gelatin (Sigma-Aldrich: G1393) in Glasgow's Minimal Essential Medium (GMEM; Gibco: 21710025) supplemented with 10% FBS, 1 mM sodium pyruvate (Gibco: 11360039), 1X Minimum Essential Medium Non-Essential Amino Acids (Gibco: 11140035), 2 mM L-Glutamine (Gibco: 25030024), 1000 U/ml LIF, 0.1 mM 2-Mercaptoethanol (Gibco: 31350010), and 100 U/ml Penicillin-Streptomycin (Gibco: 15140122); and were incubated at 37 °C with 5% CO_2. When required, cells were frozen in media/20% FBS/10% dimethyl sulfoxide (DMSO) and temporarily stored at -80 °C before being moved to long-term storage in liquid nitrogen (~-160 °C). Mouse ES cells were trypsinised and washed with media before being harvested for downstream applications.

Differentiation of mouse embryonic stem cells into embryoid bodies and isolation of erythroid cells

Mouse ES cells were differentiated using an *in vitro* haematopoietic differentiation system to form embryoid bodies, as previously described (Keller, 1995). In brief, ES cells were cultured on 0.1% gelatin in Iscove's Modified Dulbecco's Medium (IMDM; Gibco: 31980022) with 1.5 x 10^{-4} M α-monothioglycerol (MTG; Sigma-Aldrich: M6145) and 100 U/ml Penicillin-Streptomycin (referred to hereafter as IMDM/MTG/PenStrep), supplemented with 15% heat-inactivated FBS and 1000 U/ml LIF for 24 to 48 hours at 37 °C with 5% CO_2. Cells were then trypsinised, washed and resuspended in IMDM/MTG/PenStrep/10% heat-inactivated FBS at a concentration of 1 x 10^6 cells/ml. To avoid cell adherence during differentiation, cells were then plated in bacterial petri dishes (~20,000 cells/dish) and cultured in IMDM/MTG/PenStrep supplemented with 15% heat-

inactivated FBS, 5% Protein-Free Hybridoma Medium (PFHM-II; Gibco: 12040077), 2 mM L-Glutamine, 300 µg/ml transferrin (Roche, Sigma-Aldrich: 10652202001), 50 µg/ml L-Ascorbic acid (Sigma-Aldrich: A4544), and 3 x 10^{-4} M MTG; and were incubated at 37 °C with 5% CO_2 for 7 days.

Embryoid bodies were collected after seven days of differentiation and CD71+ erythroid cells were isolated. Embryoid bodies were collected, washed with cold PBS, and trypsinised. The resulting single cells were resuspended in 500 µl of PBS/10% FBS per 10^7 cells and stained with a 1/200 dilution of anti-CD71-FITC antibody (eBioscience: 11-0711-85) at 4 °C for 20 minutes. Stained cells were washed with cold PBS/10% FBS and resuspended in 90 µl of cold PBS/0.5% BSA/2 mM EDTA and 10 µl of anti-FITC MACS microbeads (Miltenyi Biotec: 130-048-701) per 10^7 cells and incubated at 4 °C for 15 minutes. CD71+ cells were positively selected via MACS lineage selection columns and processed for downstream applications. Purity of the isolated erythroid cells was routinely verified by FACS.

Genome editing and genotyping

Generation of mouse models

Mouse models harbouring mutations of CTCF binding sites around the mouse α-globin locus were generated by Lars L. P. Hanssen and Benjamin Davies (WCHG), using CRISPR-Cas9 mediated genome editing. Hanssen and Davies designed the targeting strategy which involved both targeting mouse ES cells which were then used in blastocyst injections and direct microinjections of oocytes, to obtain the desired models. They prepared the CRISPR-Cas9 expression constructs for targeting of the mouse ES cells, prepared ssODN donors, and prepared the sgRNA for direct microinjection of oocytes. Full methods have been described

in “The role of elements binding CTCF and cohesin in directing tissue-specific enhancer activity”, L. L. P. Hanssen 2016. The 20 nucleotide guide sequences used to direct the Cas9 protein to the correct CTCF binding sites can be seen in Table 6.1 and the ssODN donor sequences in Table 6.2. Founder mice harbouring the desired mutations were crossed with C57BL/6 mice to obtain heterozygotes by Benjamin Davies (WCHG) before being transferred to Jacqueline A. Sharpe (WIMM) to be bred to homozygosity.

Table 6.1: gRNA target sequences of targeted CTCF binding sites at the mouse α-globin locus

Target	Guide Sequence (5’-3’)
θ1	TGGAACGATGCAGCGCCCCC
θ2	TGAAACACAAGAGGCCGCCA
θ2	GACATCTTTGAGCTCAGCCA
HS+44	GAAAGCCAGTGGCGCCACCT
HS+44	CCCTGCAGGCCACTATAAGT
HS+48	TCCAAGGTCCTCAAGCAGAC
HS+48	CGACGAGCACCCCCGTGTGG

Table 6.2: ssODN sequences used for disruption of CTCF binding sites at the mouse α-globin locus

Target	ssODN Sequence (5'-3')
θ1	CGCTCTGCCCGCTGGCTGAGCTCAAAGACGTCCTGAAACACAA GAGGCGGATCCCGCTGCATCGTTCCAGGATGCCTAGGTGTTCA CAGATTCTGGTTCAGCTTTGAGCCCTCTGTTTCCCTGGGCTCCC CCTCC
θ2	CTTGCGCTCTGCCCCCTGGCTGAGCTCAAAGATGTCCTGAAAC ACAAGAGGCGGATCCCGCTGCATCGTTCCAGGATGCCTAGGTG TTCACAGATTCTGGTTCAGCTTTGAGCCCTCTGTTTCCCTGGGC TCCCT

Genotyping of mouse models

Ear biopsies from mice were lysed in DIRECTPCR-EAR PEQGOLD (peqlab, VWR: VIAG402-E) and proteinase K (ThermoFisher: EO0491) overnight at 55 °C, followed by a 5-minute incubation at 95 °C. The lysate was directly used as template for PCR reactions except in the case of a long-range PCR covering the α-globin genes, θ1, and θ2 (primer ID no.5), in which case genomic DNA was first purified. Target regions were amplified with the primers listed in Table 6.3 using IMMOLASE DNA polymerase (Bioline: BIO-21047).

Different strategies were used for routine genotyping of mice, depending on the mutation. The deletion of HS+44 in Δ44-48 mice was detected by resolving WT and mutant alleles on a 2% agarose gel following PCR (primer ID no. 1). Screening for the small deletion of HS+48 in Δ44-48 mice was completed by PCR products (primer ID no. 2) being run on an agarose gel and sent for Sanger sequencing. The θ1 mutation in Δθ1 mice was detected by BamHI digestion of the PCR product (primer ID no. 3). To detect the small deletion of θ1 in Δθ1θ2 mice, two sets of PCR primers were used: primer ID no. 4 only amplifies the WT sequence without the deletion and primer ID no. 5 only amplifies the sequence containing the deletion. The θ2 mutation in both Δθ2 and Δθ1θ2 mice was genotyped by digestion of the PCR product (primer ID no. 6) with BamHI.

Table 6.3: PCR primers used for the genotyping of mice

ID	Target	Primer	Sequence (5'-3')
1	HS+44	Forward	AAGCAAGCACTTCCCATCCA
		Reverse	TGAGGTGCCTTCTAGTCCCA
2	HS+48	Forward	GCCCCACAAAACTGTACCCT
		Reverse	TGTGTCATAAGGAAGCCAGGG
3	θ1	Forward	GGTACGGGCTCCTGACACTC
		Reverse	TGGAGGGGAAAGCCACGAAG
4	θ1 (WT)	Forward	CCCTGGCGGCCTCTTG
		Reverse	GTCCAGGACAACAATGCAGC
5	θ1 (DEL)	Forward	TGTGGAAGACAGAAGAGGGC
		Reverse	CTGAAACACAAGAGGCGCTG
6	θ2	Forward	TTCCGAAGGACTCGGGAAGC
		Reverse	ATGCACAGAGGCAATGCAGC
7	Long range (α-globin genes, θ1 & θ2)	Forward	AGTCCTTCCCAGGACTTTCCA
		Reverse	AGACATGCATCTTTTTCGCCG

Generation of mouse ES cell lines harbouring artificial loci

The introduction of the basic *cis*-regulatory elements of the mouse α-globin locus into the two neutral loci selected on chromosome X was achieved using CRISPR-Cas9 mediated HDR in mouse ES cells. In collaboration with Philip Hublitz (Genome Engineering Service, WIMM), three sgRNAs for each LP were designed. Genome Engineering Services cloned the sgRNAs into pX458 (Ran et al., 2013), a plasmid encoding Cas9 from *Streptococcus pyogenes* and enhanced green fluorescent protein (eGFP), and provided functional validation of 33 sgRNAs by performing Surveyor nuclease assay on DNA obtained from transfected B16F10 cells. The sgRNA that provided the highest targeting efficiency for each

LP in the two loci was used for targeting (Table 6.4). HDR templates containing the sequence predicted to encode the *cis*-regulatory element to be inserted and sufficient homology arms for each LP (Table 6.5) were synthesised and cloned by GeneArt (Invitrogen, ThermoFisher) into double-stranded plasmid vectors (pMX). sgRNA/Cas9 plasmid DNA and donor plasmid DNA were prepared for transfection using QIAprep Spin Miniprep Kit (Qiagen: 27106) according to manufacturer's instructions. Mouse ES cells (~1 x 10^6) were transfected using Lipofectamine LTX Reagent with PLUS Reagent (Invitrogen, ThermoFisher: 15338100) which were performed in 6-well format by plating freshly trypsinised cells in 2 ml media supplemented with the transfection mix prepared following manufacturer's instructions: 5 µl LTX reagent, a total of 2.5 µg purified plasmid DNA (1:3 ratio of sgRNA/Cas9 plasmid DNA:donor plasmid DNA, 0.625 µg and 1.875 µg respectively), 2 µl PLUS reagent, and 250 µl Opti-MEM (Gibco: 31985062). After 24 hours incubation at 37 °C, transfected cells were given fresh media. Transfected cells were selected based on the presence of eGFP by FACS after 48 hours of incubation at 37 °C. eGFP+ cells were plated at low densities (~2,000-6,000 cells/10 cm dish) to allow for isolation of colonies originating from single cells. After seven days of growth, with the cells receiving fresh media every two days, individual colonies were picked into 96-well plates, grown for an additional seven days receiving fresh media every two days, and screened for mutants.

Table 6.4: gRNA target sequences of targeted LPs at artificial loci

Target		**Guide Sequence (5'-3')**
AL1	LP1	GGTGTGTAAGGCATCACATTC
	LP2	GGTTGGTGCTTCCGGAGACA
	LP3	GGCAACATAAATCCTTGTGAG
	LP4	GGGACAGTTTTGCCAGACGT
	LP5	GGCACTGCCAGTACGTTCAT

AL2	LP1	GGCATACCAGTGGTACCCCCT
	LP2	GGAGAGTAGTGGCCCAACTCT
	LP3	GGCCAACTCGACCGCTTA
	LP4	GGACCCATCGTTCTTCCCCGC
	LP5	GGTAGTGCCTTTAGGGTGTTA
	LP6	GGCACTCAGGACGGGTAGAT

Table 6.5: Donor sequences used for insertion of mouse α-globin locus elements into artificial loci

Target		**Donor Sequence (including homology arms) (5'-3')**
AL1	LP2 (R2 sequence)	CGTATTTCTGTGTCTCCATAATAAAGAGGATATTTCTGAGGCTCTAAACTCTATTTAC GTGTCTCAGTTGGGAAACACTCTTTATTTACTTGGGGCTAAGGTCACTGCTTGTCTGT GTTTATTGATTTCAAAGACCATCTTAGTCTCTAGAGCCTCTTTCTTGGCTAAGGCCCC TAATATGGGTACCAGAAATGAAGTTATGGCCCTGTGTTTCATGTTCTTCTTGTCTAGA ACTTGGACTTTTTTGTTTTTGTCTTTCATTTCTCACTGAGGTTTAGGTTGTTTTGGGAT GACAGGAGAGGTCATGGTTTATTTTGACTTTTTACATGTTGTAGCTGGAGGCTGTTAT TTAGTTCTTAATGTTTTAAAAACAGGTTTAGAGTCCTTGGATTAACTGTCCTTTGCTT ACTTATGCTTATTAATTGGAGACGTGCATAGTCAAGTATTTTACCTCAAATATTACAT CAGAAACAGACAATGGGAAAGTAGTTGGTGCTTCCGGAGATCTACTGCATACCTACA CTAACTAGGTCAAAGTAGCATACACCCATCTGGAACCTATCAGTGACCATAGTCAAC AGCAGGTGTACACACCCAGGCCAAGGGTGGAGCAGACCACTGTGGGATCTATGGAG ATGCTTGAACGAGCAGATAACTAAGCCAAGCATGACTCAGAGTTTCTAGAGGCCACT AGGACTGCTGAGTAATACTTGGGGGTACAGAGTCAGAAAGGAAAGGACAAATGGTA CCACTGATTAGGACCTCTGACGCTGTTTTCCCATCCTGTTTATTTGCCAAGTGACCCT GTGCCTGTTACCTTAAGTCAAAAATGACTTCACTATATATCATCACATAAAATAGTCT ATATTCACATTGATGAACTTGTTCATGTTATTTCTCTAATAGTATATATATATGACTG CTACTATATTATCTTTTCAAATCCCTTGATATTCTATATCAGCCAGTTGAGCTATGAG AGCTCACACAAGTGAGGATTTGGGGAGGGGATTCTGTGTGGCAATGGAGTGGAGAA CTGTGGAAGCTTCGGGAATGCTATGCCTTCACTGGATTTGTGTTATAGTGGTCATATT TGCTCTGCAAACTAGATATACCCCAATACCTACAAACTAAAACTGAGTGGGAGGCTT TCTTTGAATAGAAATAGGTATAATTGGAACTTGGTGGAAATTATTCTGGAAAAGTCC TTAAGAGACTCGTCATGGCCTCCACAGAACATAGTGAAGCTGCCATTGAAGTCCTCA GGCTTTCAGCAAACCTGGGTGGAAGGTTTTAAACTAGTAAGCAGGGAAGAAGTTTA AGTTCATG
	LP4 (α-globin promoter-mVenus sequence)	CATCTATGACTCCTGGAACAACCACACTTGCGTGGGAGGGTTGAGCCGGATCTTTTC CACCTTTTGCCACTCAGAAAGTAGCGCAGATCCCTGTTGGATCATCATCATCTCCAG AGAGATGGCCTTATATTAGGGTCAAGTAGTAGCTAGGCTTCTACACACCAAGCAGTA ATAAAAAGAACCTGGACCTTGGAATCTAATCCCAGCTATTGTATGGGAGGCATGTGT CCTAGCTCTGTTTTTTTTTTTTTTTCATTTATAACATAGATACTTGTAGTGCGTGTTAC TCCATTTTCTATTACTATAACAAAATACCCAAGGCTGAGCCTTTTTTTTTAATTCATAT ATTTGGAGTTCTAAGAGTATAGCACTGACTTCTGCCTAGCTTCATAGTAGATGGCCTC ACATATATTGTAGTATGTGTGCAAGGGTATATACTGCAAGACAGGAACAAAAAAAG AGCACGTTTTTGTCTTTTATAACAACCCATTCTCAGGAAAACTATCCCAAAGTCCCAT GAGTGCTACATTAATGTCTTCAAAGGTCAAGGATTCCTGTGACTTGCTACTTTCTGTT TGCCCCCACCTCTTAAAGAGTCTGCCGCCTCAGTCCCAGTATAATGGAGAATAAACT TCTAGCTCACAAAACTTTAGGAGACAAGCTACATTTAACATATAGCACAACACACAG TGCACTGGGCTTTCTGAATGAATAAGGAAAAATATACAGGTAAAAGGTTTTTTTTTTT TTTTTTTTTTTAATGAATCTTTCCCTCTACAGCCAACTCCATTTTACTGCTGAGGAAAA CTGAGGCCAGGTGAATTAAAGTTATTCTTTCAAGTCTATAGGCTACTTTACTGTTACA ACTCAGGCGAAAAACCTCAATGAACTTTACTAATACAAAGTCCAAAGAAATTATGGT CTGCCTGCTGATAAATGTTGAATAATTGGCTAAGTGCTTGGAGGGAAGATGGTGAAG AGAAAGGTCTGATTTTTTTGTCTTAACATCACTTTATATTCCCATACAACTGTAAGAA AGAAGGTCTTTTGAACCATCTCCCCATTTCCCTCCAGTGACTAGTCTGCCATCACTGG AGACGTAGTAAGGGGTAAGAAGTGTGTCCGGGCAACTGATAAGGATTCCCTGCACC

		CAGGGGAAGCACAACCCAGCCCCAGAATCTCAGGGGCCCTAACAAGTTTTACTGGG TAGAGCAAGCACAAACCAGCCAATGAGAACTGCTCCAAGGGCGTGTCCACCCTGCC TGGAGGACACGCCCTTGGAGGGCATATAAGTGCTACTTGCTGCAGGTCCAAGACACT TCTGATTCTGACAGACTCAGGAAGAAACCATGGTGAGCAAGGGCGAGGAGCTGTTC ACCGGGGTGGTGCCCATCCTGGTCGAGCTGGACGGCGACGTAAACGGCCACAAGTT CAGCGTGTCCGGCGAGGGCGAGGGCGATGCCACCTACGGCAAGCTGACCCTGAAGC TGATCTGCACCACCGGCAAGCTGCCCGTGCCCTGGCCCACCCTCGTGACCACCCTGG GCTACGGCCTGCAGTGCTTCGCCCGCTACCCCGACCACATGAAGCAGCACGACTTCT TCAAGTCCGCCATGCCCGAAGGCTACGTCCAGGAGCGCACCATCTTCTTCAAGGACG ACGGCAACTACAAGACCCGCGCCGAGGTGAAGTTCGAGGGCGACACCCTGGTGAAC CGCATCGAGCTGAAGGGCATCGACTTCAAGGAGGACGGCAACATCCTGGGGCACAA GCTGGAGTACAACTACAACAGCCACAACGTCTATATCACCGCCGACAAGCAGAAGA ACGGCATCAAGGCCAACTTCAAGATCCGCCACAACATCGAGGACGGCGGCGTGCAG CTCGCCGACCACTACCAGCAGAACACCCCCATCGGCGACGGCCCCGTGCTGCTGCCC GACAACCACTACCTGAGCTACCAGTCCAAGCTGAGCAAAGACCCCAACGAGAAGCG CGATCACATGGTCCTGCTGGAGTTCGTGACCGCCGCCGGGATCACTCTCGGCATGGA CGAGCTGTACAAGGATCCAAAAAAGAAGAGAAAGGTAGATCCAAAAAAGAAGAGA AAGGTAGATCCAAAAAAGAAGAGAAAGGTATAAGCTGCCTTCTGCGGGGCTTGCCT TCTGGCCATGCCCTTCTTCTCTCCCTTGCACCTGTACCTCTTGGTCTTTGAATAAAGCC TGAGTAGGAAGAAGCCTGCACAATTGAGGCAGGATATGGATAGGATAGAATCAAGC AATCTTTGTGTTTTCCTCGTTTGTCTTTGTTTAGTTTCATGTCATTTCATTGCATGTATC CCCGGGCCACTATCACAATCAAGACACAGAACAGTTCCATTAGCACTGCTGAATAAA GGACTGCTCTACTGCCATCTCCTTATCTTGTCCTGTCCTTAAGAAGCTTACAGTCCAG TCTCTAATAGGCTCTCTTCTTCTAAAACTGCACTATTTCAAAAAACCATCATCATATA AAGCAACACAGTCCCATGCAGTTTATATTGTATTTTGGATACAATAATATTCACCCCA CACTACCCCCCCCATCTCTTTTCACTGCCTGTAGACCTCTTCGTCTCCAAAAATACAT CTTCTTTCTTTCATATCTTTATAATAATAATGCACTGAGTTTAATTGGGATTTCTTGCA TGTGTGTGGGTTTGGCTGTGGAATAATTTTATTGGCTCATGGGAAAACCACTGAAAG TAAAAACAAGTCCCCTTTCCCAGAGGCTGTTAACTGCCAATATAGTCCCCTCAGCTG AGGCAGAGCCTTGGGCGCCTCTCTTGCATCAGTGCTAGAATGTCCACAAGCTAGCTC TTGTTCAGATCTTGTGCAGGTAAACACAGCTATTGTGAATTCAGAGTGCATGGCC
AL2	LP2 (R2 sequence)	CCCACAATGGCATGGCATTGCTGAATGTATCCCAGTTTTTTGGCTTTTGATGCCTATA TTGTCAGGGTCATATATCCCCAACCACTACTGAGAAAAACTTCCTTGGGTCAGTGTT GGATAGCAGCAGCTCCCTTATGCTTACTTGGAGTGGTTCTAAGATTTTGTGTTTTACA TTAGGTATTTTATCTATTTGCAGTTGGTTTGGGATGTTGTGTGAGACAAGGGTCCAGT TTCTTTCTTCTGCATGTACACATCTAGTCTTCAAACAGAGTTCAATGAAGAAATTGTC AATTTTCTATTGTATATACTTTTGCCATATTCATGCTGGAAATGGCTTTCTTTGCTTTT CCCAAGGTAATCTGCATTTAATGTGGATGAGAGGATGCTGGCCATAACCCCCCCCCC CTTCCATCCCTGCACTGGAAAATATTCTGGGCATCTGACAGTCTATAAGACCATGCTC TTCCCAGGTAGTTATCTTCTTATGACTGAGCACAGTTGGTTGATCTACTGCATACCTA CACTAACTAGGTCAAAGTAGCATACACCCATCTGGAACCTATCAGTGACCATAGTCA ACAGCAGGTGTACACACCCAGGCCAAGGGTGGAGCAGACCACTGTGGGATCTATGG AGATGCTTGAACGAGCAGATAACTAAGCCAAGCATGACTCAGAGTTTCTAGAGGCC ACTAGGACTGCTGAGTAATACTTGGGGGTACAGAGTCAGAAAGGAAAGGACAAATG GTACCACTGATTAGGACCTCTGACGCTGTTTTCCCATCCTGTTTATTTGCCAAGTGAC CCTGTGCCTGTTACCTTAAGTCAAAAACTTATGACTTTTCTGATGGGCAGTTTTTCCC CTGGTCTCCACTGGCCTGAAGAGACGTTTTCAAAGTGATGCTTTTCAGCTGAGGTGT ATTCCATCCTTCCCTCTGCTTCCACAGGTCTTGGATGGACAACCTCTGTTTGTTCCATT TATCTTTTTGCAGGTTTCTAGCCCCATCTAGGTGTCTGCTTCCAGGAGGACAAGAACT GGCATACATAACCATAAAGTCTCAACAACGCAGTAGGTGTTTTCCAGATATCCTCCC CTGAGTTCTTTCCAGCTGAATGAGGGAGCCACTGAGGTCTGTGCCTTTGCTTACCTCC CCAACTACCATGCAGCAGCCTTGACAACAGCACTTCTGTTCTGCTATGCAAATTTGTT TTCTCTCTCTACTGAATTCCCTGAAGGTAAGAACAGGAATTATCTTATCTCTAAACCA GGTATTGGGCAATTCAGGTGGATCATGCATGATACTCTGGGTCACACATCCCCC
	LP2 (mutated R2 sequence)	CCCACAATGGCATGGCATTGCTGAATGTATCCCAGTTTTTTGGCTTTTGATGCCTATA TTGTCAGGGTCATATATCCCCAACCACTACTGAGAAAAACTTCCTTGGGTCAGTGTT GGATAGCAGCAGCTCCCTTATGCTTACTTGGAGTGGTTCTAAGATTTTGTGTTTTACA TTAGGTATTTTATCTATTTGCAGTTGGTTTGGGATGTTGTGTGAGACAAGGGTCCAGT TTCTTTCTTCTGCATGTACACATCTAGTCTTCAAACAGAGTTCAATGAAGAAATTGTC AATTTTCTATTGTATATACTTTTGCCATATTCATGCTGGAAATGGCTTTCTTTGCTTTT CCCAAGGTAATCTGCATTTAATGTGGATGAGAGGATGCTGGCCATAACCCCCCCCCC CTTCCATCCCTGCACTGGAAAATATTCTGGGCATCTGACAGTCTATAAGACCATGCTC TTCCCAGGTAGTTATCTTCTTATGACTGAGCACAGTTGGTTGATCTACTGCATACCTA CACTAACTAGGTCAAAGTAGCATACACCCATCTGGAACCATTGAGTGACCATAGTCA ACAGCAGGTGTACAGTCCGAGGCCAACCGTCGAGCAGACCACTGTGGGATCTATGG

		AGATGCTTGAACGAGCACTTAACTAAGCCAAGCAAGTCACTGAGTTTCTAGAGGCCA CTAGGACTGGACAGAATTACTTGGGGGTACAGAGTCAGAAAGGAAAGGACAAATGG TACCACTGTATTGGACCTCTGACGCTGTTTTCCCATCCTGTTTATTTGCCAAGTGACC CTGTGCCTGTTACCTTAAGTCAAAAACTTATGACTTTTCTGATGGGCAGTTTTTCCCC TGGTCTCCACTGGCCTGAAGAGACGTTTTCAAAGTGATGCTTTTCAGCTGAGGTGTAT TCCATCCTTCCCTCTGCTTCCACAGGTCTTGGATGGACAACCTCTGTTTGTTCCATTTA TCTTTTTGCAGGTTTCTAGCCCCATCTAGGTGTCTGCTTCCAGGAGGACAAGAACTGG CATACATAACCATAAAGTCTCAACAACGCAGTAGGTGTTTTCCAGATATCCTCCCCT GAGTTCTTTCCAGCTGAATGAGGGAGCCACTGAGGTCTGTGCCTTTGCTTACCTCCCC AACTACCATGCAGCAGCCTTGACAACAGCACTTCTGTTCTGCTATGCAAATTTGTTTT CTCTCTCTACTGAATTCCCTGAAGGTAAGAACAGGAATTATCTTATCTCTAAACCAG GTATTGGGCAATTCAGGTGGATCATGCATGATACTCTGGGTCACACATCCCCC
	LP4 (α-globin promoter- mVenus sequence)	GTGAAAAGGACTGAAGACCTCGGCTTACATAACTGCAAGGAGCATATTGCAAGTTA GCAGAATTCACACTCACCCTAGGGAAGGGGCCAGAAGTTTTGTTTAAGGTCCTGAAG GGAACGAAAGGGGAGGATAATGAAAGAGGGAATGGAGCTATTGCGTGCCACACAC AGCTTGTACATACATACTACTGGCTTGAGAGAGAAGAAAGAGGAATTGTCTCCATTC ACAGAATATTCCTTAGAGAATCCATCCCAGGCATCATTTTCTCCAACAGACGTTTCCC ACCTGCGCACATGTACCTGTGTGTCTGCTTGCACATGCACGAGCAAAGTCATATTTA CACAGAGTCACACTCACACGGGTGTGTACATACAGAAGCATATGTGTCACCACGGG GACTGAACTGCTTGGCTTACATTACACTATCAAAGGACACAACAGTGTTGGTAGGGA AGCCACAGACCTACCCTGGTGGCTTTATCTGCAGCCATTCCTTGTGGAGACTAGTCTG CCATCACTGGAGACGTAGTAAGGGGTAAGAAGTGTGTCCGGGCAACTGATAAGGAT TCCCTGCACCCAGGGGAAGCACAACCCAGCCCCAGAATCTCAGGGGCCCTAACAAG TTTTACTGGGTAGAGCAAGCACAAACCAGCCAATGAGAACTGCTCCAAGGGCGTGTC CACCCTGCCTGGAGGACACGCCCTTGGAGGGCATATAAGTGCTACTTGCTGCAGGTC CAAGACACTTCTGATTCTGACAGACTCAGGAAGAAACCATGGTGAGCAAGGGCGAG GAGCTGTTCACCGGGGTGGTGCCCATCCTGGTCGAGCTGGACGGCGACGTAAACGGC CACAAGTTCAGCGTGTCCGGCGAGGGCGAGGGCGATGCCACCTACGGCAAGCTGAC CCTGAAGCTGATCTGCACCACCGGCAAGCTGCCCGTGCCCTGGCCCACCCTCGTGAC CACCCTGGGCTACGGCCTGCAGTGCTTCGCCCGCTACCCCGACCACATGAAGCAGCA CGACTTCTTCAAGTCCGCCATGCCCGAAGGCTACGTCCAGGAGCGCACCATCTTCTT CAAGGACGACGGCAACTACAAGACCCGCGCCGAGGTGAAGTTCGAGGGCGACACCC TGGTGAACCGCATCGAGCTGAAGGGCATCGACTTCAAGGAGGACGGCAACATCCTG GGGCACAAGCTGGAGTACAACTACAACAGCCACAACGTCTATATCACCGCCGACAA GCAGAAGAACGGCATCAAGGCCAACTTCAAGATCCGCCACAACATCGAGGACGGCG GCGTGCAGCTCGCCGACCACTACCAGCAGAACACCCCCATCGGCGACGGCCCCGTGC TGCTGCCCGACAACCACTACCTGAGCTACCAGTCCAAGCTGAGCAAAGACCCCAACG AGAAGCGCGATCACATGGTCCTGCTGGAGTTCGTGACCGCCGCCGGGATCACTCTCG GCATGGACGAGCTGTACAAGGATCCAAAAAAGAAGAGAAAGGTAGATCCAAAAAA GAAGAGAAAGGTAGATCCAAAAAAGAAGAGAAAGGTATAAGCTGCCTTCTGCGGGG CTTGCCTTCTGGCCATGCCCTTCTTCTCTCCCTTGCACCTGTACCTCTTGGTCTTTGAA TAAAGCCTGAGTAGGAAGAAGCCTGCACAATTGCGGGCCTGAGCAGAAGTTGTCTG CAGATGTCAGTGTTAATGCCTTGATAAACAGGGAGCTTGCTGGAGTGCAGCCTTCAG CTCCCGGAGCTGTGGGTAATGACAGGCCACTGGGGGAGCCAAAGACAGTTGCCACT GTTCCTGCCTGTCCTCGGAGCTTCTTAAACAGCAGTATACATTTAAAAGCTTGAATAT GCTTTCTTGGGAATCATTGTGTCTTAGAAATTCTGAGAATGCCAGCAAATTTCACAGC TCCCTGGCCAGTGACCTTCTACAGGCCAGGGAAGAGCAGCCTTTCTGCCTTCCAGGC CACAGCTTCACCTTGCCTCCCCCTGCTGTGGAGCCACTGGGGGGGGGGGGGTGTTTG AAGAAGCTTCAGGGCTAATTTACAGTTGACTGGCAGCTGTGGGAGAAAGAGCTTAG AAGAGACTCTTTAGACTTCCAAGGGAAGAATTTTCTAGATGCTCTAAACTGACCCAG AAGCCACAACCTCTTTTTTTC
	LP5 (α-globin promoter- mVenus seqeunce)	CTAGTAAGGAAAAGAGATTCCCCCTAAGAGCACATTGCCTTCTTCAGGGACACAGTG ACCAGACATGAAGAGCTACTTAAGCCCAGAACAGTTTTCTGTGTCACTGTTCACACA TTATGGACAGCCATATGTGCAAATTTCCATTTCTATTTAGCAGTGGGAGGCCATGAT ATGGTTTAGATGTGTGCCCACCCCCACCCCCGCACTTCCTCACAAAGTTCACTTCAAA CAAGCTTAGTTAAGTCAAGGTGAAGGTGATCTGGGATGCTGTGAACCTTTAAGAATT ACAGATTAGTGGGAAGTTTTTGAGTCATTGAAGGTTGTCTTAGGAAGTGAGACCCTA GAACTTCACTGGCTTTCTGATTTGTCCCATGATCTAGTCCTCACACATTTATGAGATG GTGTCACCAAGAGCACCTTCACAAGAAGACCAATGCGATACCATTGAGCCTCACAGT GAGCCAAACAAAAGTCTTTATAAAATTAGCCTGCTTGGCCTACTTTACTGTAGCAAC AACAAACAGGTTAATGTAGGTGGGGCCCTAGAATCTGCCTTTCCAACACACTGCTGT GTGAGGCAGTTGCCTCTTCCTCTGAGACTTCATTTTGAGACTAGTCTGCCATCACTGG AGACGTAGTAAGGGGTAAGAAGTGTGTCCGGGCAACTGATAAGGATTCCCTGCACC CAGGGGAAGCACAACCCAGCCCCAGAATCTCAGGGGCCCTAACAAGTTTTACTGGG TAGAGCAAGCACAAACCAGCCAATGAGAACTGCTCCAAGGGCGTGTCCACCCTGCC

		TGGAGGACACGCCCTTGGAGGGCATATAAGTGCTACTTGCTGCAGGTCCAAGACACT TCTGATTCTGACAGACTCAGGAAGAAACCATGGTGAGCAAGGGCGAGGAGCTGTTC ACCGGGGTGGTGCCCATCCTGGTCGAGCTGGACGGCGACGTAAACGGCCACAAGTT CAGCGTGTCCGGCGAGGGCGAGGGCGATGCCACCTACGGCAAGCTGACCCTGAAGC TGATCTGCACCACCGGCAAGCTGCCCGTGCCCTGGCCCACCCTCGTGACCACCCTGG GCTACGGCCTGCAGTGCTTCGCCCGCTACCCCGACCACATGAAGCAGCACGACTTCT TCAAGTCCGCCATGCCCGAAGGCTACGTCCAGGAGCGCACCATCTTCTTCAAGGACG ACGGCAACTACAAGACCCGCGCCGAGGTGAAGTTCGAGGGCGACACCCTGGTGAAC CGCATCGAGCTGAAGGGCATCGACTTCAAGGAGGACGGCAACATCCTGGGGCACAA GCTGGAGTACAACTACAACAGCCACAACGTCTATATCACCGCCGACAAGCAGAAGA ACGGCATCAAGGCCAACTTCAAGATCCGCCACAACATCGAGGACGGCGGCGTGCAG CTCGCCGACCACTACCAGCAGAACACCCCCATCGGCGACGGCCCCGTGCTGCTGCCC GACAACCACTACCTGAGCTACCAGTCCAAGCTGAGCAAAGACCCCAACGAGAAGCG CGATCACATGGTCCTGCTGGAGTTCGTGACCGCCGCCGGGATCACTCTCGGCATGGA CGAGCTGTACAAGGATCCAAAAAAGAAGAGAAAGGTAGATCCAAAAAAGAAGAGA AAGGTAGATCCAAAAAAGAAGAGAAAGGTATAAGCTGCCTTCTGCGGGGCTTGCCT TCTGGCCATGCCCTTCTTCTCTCCCTTGCACCTGTACCTCTTGGTCTTTGAATAAAGCC TGAGTAGGAAGAAGCCTGCACAATTGAACTGAAGTGCATCTCCATGAATCAGCCATG ACAACCACAGAAGGTACGTCAAATTGTCCTATAAACTACCAAGGTTCTTTGGCTAAG TACTCAGATGGAGAACCTGAGTGGGCAGGACTATGATACTGAAGAAGAACACAGAC AAAACAAGTCCTATCTCTTTGGACCATGGAGAGGTATGACAGAAGGACCAGGACCT GATGGGAAGGCTGCATGGGAGGAAAGGGTCCATATCTGTGTGTGTCTGCTTACTGCT GTTCTGCCACCTGGGTTTCAGGAGGTCCCAAGCATACAGATGACGAGGAGCATCTGA ACCAGGATTCCAGGAGGCTGAAGCTACGTTCATCCTCAGCAGCAAGAGTTGTTTTGT TTTGTTTGTTTGTTTTGTTTTGTTTTTGTTTTAAGTGTCACATTCCTCCATCAGAAGTGT CGTAGCGCCCTGAGTACAGGCCTCTTGCTTCTCTGTTCTTCTGACTGTTTGCAGCAGT CTGAGGGAGAAGGGCTAACAAGAGGCCACACTGAGAGCAGAAGCAATGAGCACAG GTTCTGTCCAGAACAATCTGGCAGCAGTAGGCTGATGATAAAAAGCTCATTGGCTGG ATGTGGTCATAGACATCAAAGTCACAGTTACTTCCTGGGGTGGATATTAATGGGAGA GAAGCATGAGTAAGCCTTCTGCGTGTGTTAAGGGAGCAGTGGTTGTATAGGTATGTA CTATATAAAAAGCTACCAAGCTATGCACTTAAGCTTCCTGAGCTTGGACCACACAGT TGGCCTAGCTCAATTAAAACTGGAAAGAAATTTAAATGAAGCCAAATGAGAGGAGT GAGTAGGCAAGCCAATCCCCAGGCCCTCCCAGAGCATTGTCCAGAAGGCTATTAAAT TCCATGTGAATGCCATCTGCCTGGGGGCTACAAAAGGAGACTATCTCTGCTCGGACC AACCTAGCTAGATAACTCGAATCTAGAGCCACTGCCCACGTGATGATTAAGCAGCTG AGATGTGTGCAAAGATGTCCTGTAATTTTCCAGCCACATAAACTGTTTAGGGG

Extraction of genomic DNA from 96-well plates

When the majority of wells were confluent, clonal cell populations in 96-well plates were lysed in lysis buffer (EDTA 5 mM pH 8.0, Tris 100 mM, SDS 0.2%, NaCl 200 mM, pH 8) and treated with proteinase K at 37 °C overnight. Genomic DNA was precipitated by addition of cold isopropanol and incubated at room temperature for 20 minutes on a shaking platform (75 rpm). After centrifugation, the supernatant was removed, and the DNA pellets were washed twice with cold 70% ethanol, DNA pellets were left to dry completely and resuspended fully in water by incubating at 65 °C for 30 minutes on a shaking heating block (1400 rpm). Genomic DNA was used as the template for PCR reactions described below

using the primers listed in Table 6.6 and DreamTaq Green PCR Master Mix (2X) (ThermoFisher: K1081).

Genotyping of mouse ES cell lines harbouring artificial loci

Genomic DNA extracted from 96-well plates described above was used in PCR reactions designed to detect the presence of insertion events by HDR at the artificial loci to identify potential clones that had been correctly targeted. Following PCR (primer ID nos. 1, 4, 7, 8, 11, 14), the products were run on a 2% agarose gel and potentially positive clones were identified by the presence of a band. Potentially positive clones were screened further using primers designed both outside of the homology arms (primer ID nos. 2, 5, 9, 12, 15) to ensure the insertion was in the correct location and within the homology arms (primer ID nos. 3, 6, 10, 13, 16) to detect the presence of mixed clones (with contamination from WT resulting in a smaller PCR product). The PCR products were resolved on a 2% agarose gel, and the products from within the homology arms were also sent for Sanger Sequencing to confirm the presence of the intact inserted sequence.

Table 6.6: PCR primers used for the genotyping of ES cell lines harbouring artificial loci

ID	**Target**	**Primer**	**Sequence (5'-3')**
1	AL1 (insertion of R2 in LP2)	Forward	AGCAGACCACTGTGGGATCTAT
		Reverse	GTTGGCCCCCTTTTAGGAGC
2	AL1 (insertion of R2 in LP2)	Forward	TCTAGGCCCTGGACCAGTTT
		Reverse	CTGGCCAGGTACCACTGATG
3	AL1 (insertion of R2 in LP2)	Forward	TGGGATGACAGGAGAGGTCA

		Reverse	TCCCGAAGCTTCCACAGTTC
4	AL2 (insertion of R2 in LP2)	Forward	GAACGAGCAGATAACTAAGCCA
		Reverse	GGCCTTGTACATGTTACAGAAGAG
5	AL2 (insertion of R2 in LP2)	Forward	TGTTCACCTCATGTTTTCCATGT
		Reverse	TGATGGCCTTTGCTGTTGTTG
6	AL2 (insertion of R2 in LP2)	Forward	TTCTGGGCATCTGACAGTCT
		Reverse	AACGTCTCTTCAGGCCAGTG
7	AL2 (insertion of mutated R2 in LP2)	Forward	CGAGCACTTAACTAAGCCAAGC
		Reverse	CCACCAGTAAGTTGCATTTCCA
8	AL1 (insertion of α-globin promoter-mVenus in LP4)	Forward	TGGTGGGCCAAATCTACTCTC
		Reverse	TGGGTGCAGGGAATCCTTATC
9	AL1 (insertion of α-globin promoter-mVenus in LP4)	Forward	CCAGCAGAACACCCCCATC
		Reverse	AGCTAAGGTGGAAGAGACACCA
10	AL1 (insertion of α-globin promoter-mVenus in LP4)	Forward	CAACTCAGGCGAAAAACCTCA
		Reverse	GAGATGGCAGTAGAGCAGTCC
11	AL2 (insertion of α-globin promoter-mVenus in LP4)	Forward	GGACGAAGGTTGAGAACCAGAT
		Reverse	TGGGTGCAGGGAATCCTTATC
12	AL2 (insertion of α-globin promoter-mVenus in LP4)	Forward	GACAACCACTACCTGAGCTAC
		Reverse	GTCTTTTGGCAATGGAAGGCTTA
13	AL2 (insertion of α-globin promoter-mVenus in LP4)	Forward	GAAGGGAACGAAAGGGGAGG
		Reverse	CAGTGGCAACTGTCTTTGGC

14	AL2 (insertion of α-globin promoter-mVenus in LP5)	Forward	CTGGAAGTGGTATTTTTGTGCCA AG
		Reverse	GAGCAGTTCTCATTGGCTGGTT
15	AL2 (insertion of α-globin promoter-mVenus in LP5)	Forward	GCCACAACATCGAGGACGG
		Reverse	GCAGGTTGCTACATGCCCA
16	AL2 (insertion of α-globin promoter-mVenus in LP5)	Forward	GGTGATCTGGGATGCTGTGA
		Reverse	CTCCCTCAGACTGCTGCAAA

ATAC-seq

To identify regions of accessible chromatin, ATAC-seq was performed on ter119+ cells isolated from phenylhydrazine-treated mouse spleens, CD71+ cells isolated from cultures of embryoid bodies, and mouse ES cells as previously described (Buenrostro et al., 2013). In brief, 75,000 cells were lysed in 50 µl of cold lysis buffer (10mM Tris-HCl pH 7.4, 10 mM NaCl, 3 mM $MgCl_2$, 0.1% IGEPAL CA-630) and the nuclei were isolated before transposition with *Tn5* transposase (Nextera, Illumina: FC-121-1030) for 30 minutes at 37 °C. DNA was purified using a MinElute kit (Qiagen: 28004) and libraries were amplified and indexed using NEBNext High-Fidelity 2X Mastermix (New England Biolabs: M0541) and custom primers. Amplified libraries were purified using a QIAquick PCR Purification Kit (Qiagen: 28104) and eluted in 23 µl of elution buffer. Library profiles were assessed by visualisation on a D1000 tape on a Tapestation (Agilent Technologies) and quantified using a universal library quantification kit (KAPA Biosystems, Roche: 07960140001). ATAC-seq libraries were sequenced on the Illumina Nextseq platform using a 75-cycle paired-end kit (NextSeq 500/550 High Output Kit v2.5: 20024906).

ChIP-seq

Chromatin immunoprecipitation

Chromatin immunoprecipitation (ChIP) was performed on 3-10 x 10^6 ter119+ and CD71+ cells (isolated from phenylhydrazine-treated mouse spleens and from cultures of embryoid bodies, respectively) using a Chromatin Immunoprecipitation (ChIP) Assay Kit (Millipore: 17-295) according to the manufacturer's instructions. Briefly, cells were crosslinked in Roswell Park Memorial Institute (RPMI) 1640 Medium (Gibco: 21875034) supplemented with 10% FBS using 1% formaldehyde for 10 minutes. The cross-linking reaction was quenched by addition of a final concentration of 125 mM cold glycine. Cells were washed with cold PBS and, if required, fixed pellets could be stored long-term at -80 °C. Fixed cells were lysed in sodium dodecyl sulfate (SDS) Lysis Buffer with cOmplete Proteinase Inhibitor Cocktail (Roche, Sigma-Aldrich: 11873580001). Fixed chromatin samples were fragmented using the Bioruptor Pico sonicator for a total sonication time of 3 minutes (6 cycles) and 4 minutes (8 cycles) at 4°C, for CD71+ cells and Ter119+ cells, respectively. This resulted in an average fragment size between 200 and 400 bp. Sonicated chromatin samples were diluted 1:10 in ChIP Dilution Buffer containing proteinase inhibitor cocktail, resulting in 2 ml of diluted chromatin per ChIP/no antibody control. Before the addition of antibody, 100 µl was taken from each sample to be used as 5% input. Table 6.7 contains an overview of all antibodies and conditions used. Immunoprecipitation was performed overnight at 4 °C on a rotator. Protein A Agarose Beads were added and incubated for 60 minutes at 4 °C after which the beads were sequentially washed with Low Salt Immune Complex Wash Buffer, High Salt Immune Complex Wash Buffer, LiCl Immune Complex Wash Buffer, and twice with TE Buffer. Bound chromatin fragments were twice eluted in elution buffer (1% SDS, 0.1M $NaHCO_3$), de-crosslinked at 65 °C for 4 hours, and treated with proteinase K (ThermoFisher: EO0491) at 45 °C for 60 minutes. DNA fragments were extracted using

Phenol-Chloroform-Isoamyl Alcohol Mixture (Sigma-Aldrich: 77617) with Phase Lock Gel Light Tubes (VWR: 733-2477), purified by ethanol precipitation, and resuspended in 28 µl water.

Table 6.7: Antibodies and conditions used for ChIP

ChIP type	Supplier	Catalogue Number	Lot Number	Monoclonal/ Polyclonal	Volume used per ChIP
H3K4me1	Abcam	ab8895	GR3206285-1	Polyclonal	2 µl
H3K4me3	Abcam	ab8580	GR3190162-1	Polyclonal	5 µl
H3K27ac	Abcam	ab4729	GB3205523-1	Polyclonal	1 µl
CTCF	Merck Millipore	07-729	2836926	Polyclonal	10 µl

ChIP enrichment by qPCR

To assess for enrichment in the ChIP samples, the amount of DNA was quantified relative to the input using qPCR with Fast SYBR Green Master Mix (Applied Biosystems, ThermoFisher: 4385612) and custom primers (Table 6.8). Standard curves of a dilution series of input chromatin material (5%, 1%, 0.2%, 0.04%) were used for relative quantification. Primers for known positive and negative regions of protein binding/histone modification were used to assess enrichment in the ChIP sample. This determined both percentage of input recovery and fold enrichment over a negative control of the ChIP sample. Only samples with sufficient enrichment were used in library preparation and sequencing.

Table 6.8: Primers used for ChIP-qPCR analysis

ID	Target	Primer	Sequence (5'-3')
1	H3K4me1 positive (R2)	Forward	GTACAGAGTCAGAAAGGAAAGGACAA
		Reverse	GGCAAATAAACAGGATGGGAAA
2	H3K4me1 control (Chr15 intergenic region)	Forward	GTAAGATGCCATTGTCCTGACC
		Reverse	TGTGTTCTCCTCCCTTCCTT
3	H3K4me3 positive (*Hba* promoter)	Forward	CTAGCTTCCCCACCACCAAG
		Reverse	CTTCTTGCCGTGACCCTTGA
4	H3K4me3 control (Chr15 intergenic region)	Forward	GTAAGATGCCATTGTCCTGACC
		Reverse	TGTGTTCTCCTCCCTTCCTT
5	H3K27ac positive (*Hba* promoter)	Forward	CTAGCTTCCCCACCACCAAG
		Reverse	CTTCTTGCCGTGACCCTTGA
6	H3K27ac control (Chr15 intergenic region)	Forward	GTAAGATGCCATTGTCCTGACC
		Reverse	TGTGTTCTCCTCCCTTCCTT
7	CTCF positive (HS-38 site)	Forward	TGAGAAGGCTGGCCTTTGAG
		Reverse	CCCAGGGAATGAATGCCAGT
8	CTCF control (*Nprl3* gene body)	Forward	CAGAGCTCCCAGACAACCAG
		Reverse	AGCACACCCAGGGTTCTCTA

Library preparation for Illumina sequencing

Library preparation of immunoprecipitated DNA fragments was performed using NEBNext Ultra II DNA Library Prep Kit for Illumina (New England Biolabs: E7645) according to the manufacturer's instructions. For CTCF, two ChIP technical replicates per biological replicate were combined for library preparation, whereas a single technical replicate per biological replicate was used for all other antibodies. Before library preparation, the concentration of each sample was determined by Qubit dsDNA High Sensitivity Assay (Invitrogen, ThermoFisher: Q32854) to ensure the correct adaptor concentration and number of amplification cycles were used. Library profiles were assessed by visualisation on a D1000 tape on a Tapestation (Agilent Technologies) and quantified using a universal library quantification kit (KAPA Biosystems, Roche: 07960140001). ChIP-seq libraries were sequenced on the Illumina Nextseq platform using a 75-cycle paired-end kit (NextSeq 500/550 High Output Kit v2.5: 20024906).

Capture-C

Capture-C experiments were performed on ter119+ cells isolated from phenylhydrazine-treated mouse spleens, CD71+ cells isolated from cultures of embryoid bodies, and mouse ES cells according to the Next-Generation Capture-C protocol which has been previously described (Davies et al., 2016). There are three stages to Capture-C: 3C library preparation, Capture-C library preparation, and oligonucleotide capture, which enriches for restriction fragments of interest.

3C library preparation

Aliquots of 5-20 x 10^6 cells were crosslinked in RPMI Medium supplemented with 10% FBS using 2% formaldehyde for 10 minutes. The reaction was quenched by addition of a final concentration of 125 mM cold glycine. Fixed cells were washed in cold PBS, resuspended in 5 ml cold lysis buffer (10 mM Tris pH 8, 10 mM NaCl, 0.2% IGEPAL CA-630, 1X proteinase inhibitor cocktail, water), incubated on ice for 20 minutes, and centrifuged. Pelleted cells were snap frozen and stored at -80 °C for long-term storage. Fixed cells were resuspended in 1X restriction buffer and Dounce homogenised on ice (65-75 strokes), centrifuged, and resuspended in 650 µl of 1X restriction buffer. Cells were split across three parallel tubes (200 µl per tube) and treated with SDS (incubated 1 hour at 37 °C) and Triton X-100 (incubated a further 1 hour at 37 °C) before digestion with DpnII. Three aliquots of 500 U DpnII restriction enzyme (New England Biolabs: R0543M) were added to each tube (a total of 4500 U per 3C library) several hours apart over a total incubation time of 16-24 hours on a shaking thermomixer at 37 °C (500 rpm; 30 seconds on/ 30 seconds off). The remaining 50 µl of fixed cells was used as a non-digested control and was treated identically except for the addition of DpnII enzyme.

The digestion reaction was heat-inactivated at 65 °C for 20 minutes and cooled on ice; a non-ligated control was taken and processed in parallel. Each of the three digested chromatin samples were diluted with 500 µl water and 134 µl 10X Ligation buffer and ligated with 240 U T4 DNA ligase (ThermoFisher: EL0013) (a total of 720 U per sample) overnight on a shaking thermomixer at 16 °C (500 rpm; 30 seconds on/ 30 seconds off). Ligated chromatin was de-crosslinked and proteinase K (ThermoFisher: EO0491) treated overnight at 65 °C. The three ligation reactions were pooled and treated with RNase (Roche: 11119915001) at 37 °C for 30 minutes. DNA was extracted using Phenol-Chloroform-Isoamyl Alcohol

Mixture and chloroform, purified by ethanol precipitation in a large volume to improve DTT removal, resuspended in water, and stored at -20 °C.

To assess the quality of 3C libraries (by assessing unspecific degradation, digestion efficiency, and ligation efficiency), the non-digested control, the non-ligated control, and the 3C library were visualised by gel electrophoresis (1% agarose gel) or by a Genomic tape on a Tapestation (Agilent Technologies), depending on DNA concentration. The digestion efficiency was determined by qPCR with Fast SYBR Green Master Mix (Applied Biosystems, ThermoFisher: 4385612) and primers designed across a DpnII digestion site (DpnII digestion control; primer ID no. 1) and across a region without a DpnII digestion site (genomic control; primer ID no. 2) (Table 6.9). Libraries were quantified using a Qubit dsDNA Broad-Range Assay kit (Invitrogen, ThermoFisher: Q32853).

Table 6.9: Primers used for the quantification of digestion efficiency of 3C libraries

ID	Target	Primer	Sequence (5'-3')
1	DpnII digestion control	Forward	GGAGAAAGAAGGCTGGTGTTAT
		Reverse	TATCTGAGTTGGACAGCATTGG
2	Genomic control	Forward	TTATCTTGCATTTGCCAACTCG
		Reverse	TGGGTTTCCCTGATTCTGAAA

Capture-C library preparation

Aliquots of 4 µg of 3C library were sonicated to ~200 bp fragments using a Covaris S220 Focused Ultrasonicator (settings: one cycle of 360s; duty cycle: 10%; intensity: 5; cycles per burst: 200; set mode: frequency sweeping). The recovered material (~3 µg) was split over two parallel reactions and Illumina TruSeq adaptors were added using the NEBNext

Ultra II DNA Library Prep Kit for Illumina (New England Biolabs: E7645) according to the manufacturer's instructions. Adaptor-ligated libraries were amplified and indexed (TruSeq indices) using Herculase II Fusion Enzyme with dNTPs Combo (Agilent Technologies: 600677) in a 7-cycle PCR reaction, following the manufacturer's instructions. The two parallel reactions were barcoded with the same index to generate maximum library complexity. All DNA purification steps were performed using Agencourt AMPure XP beads (Beckman Coulter: A63881) at a 1:1.8 sample:bead ratio to minimise loss of smaller fragments. Sonication efficiency and library profiles were assessed by visualisation on a D1000 tape on a Tapestation (Agilent Technologies).

Oligonucleotide capture

Restriction fragments containing the viewpoint of interest were enriched by oligonucleotide capture using NimbleGen SeqCap EZ Hybridization and Wash Kit (Roche: 05634261001), NimbleGen SeqCap EZ Accessory Kit v2 (Roche: 07145594001), and custom biotinylated DNA oligonucleotides (Table 6.10). Differently indexed libraries were pooled and hybridised to multiple capture oligonucleotides simultaneously, allowing for multiplexing of both samples and viewpoints. Biotinylated DNA oligonucleotides were reconstituted to a concentration of 100 μM and stored at -20 °C. The oligonucleotides were pooled in equimolar quantities and diluted such that each individual oligonucleotide was at a concentration of 2.9 nM within the pool. The experimental procedure outlined below was scaled up to the number of multiplexed libraries and performed in a single reaction.

Table 6.10: Custom biotinylated DNA oligonucleotides used in Capture-C experiments

Viewpoint	5' or 3' Capture DNA oligonucleotides	Sequence (5'-3')
HS-38 CTCF	5'	[Btn]GATCCCAGTCAAGACCTAGACAAGGAAATGGAAGAGACATCAAAGCTTTATCAAAGGACTGTGACAGTGAAGTCAGGCACAGTCCTAGCACAGATTCCTTCAAACAGCCTCTTAGTAAAG
	3'	[Btn]CTGCCACATGTCTTCCGCCACAAGAATGAGGGTATCTTGATGACTCCAGGGCTGGGTTGGATAAAAGGGGAGAAGGAGCAGAACAAAAGGGACCCATGTGGAGCGTACCTCACTTGGATC
R1	5'	[Btn]GATCCCCAGCCTATGTCTTTCCCTCTGAACATTTTGTACTAAGCTGGAGCCTCTTTACAAAGTAAAAAAGCTCTGGCATGGAGACACTGAACACTGCCATTAGGAGTCCCCCGAGTATCA
	3'	[Btn]CTGTTAATTCAGAGAATCACTTCTTCACACCACAAGGAGTTTATATCCTCATTCCTCCTGCCATTCCTTATCTCTTTAACTGTGGGTGAGCTGTGTTATGTTATAAAGTGAAGGGAGATC
Hba-α1	5'	[Btn]GATCCATGGTAGCACAGGGCAGCTAAGATGCAAGTCTGAAGGAGGAGTCTGGCGAGCTGCTCCTGCAGTTCCCTGGACCCAGAAGGATGAGCTAGCAGATTCACTTGAGCCAAAGGATTC
	3'	[Btn]AAACCATGGTGCTCTCTGGGGAAGACAAAAGCAACATCAAGGCTGCCTGGGGGAAGATTGGTGGCCATGGTGCTGAATATGGAGCTGAAGCCCTGGAAAGGTGAGAACAGGACCTTGATC
Hba-α2	5'	[Btn]GATCCATGGTAGCACAGGGCAGCTAAGATGCAAGTCTGAAGGAGGAGTCTGGCGAGCTGCTCCTGCAGTTCCCTGGACCCAGAAAGATGAGCTAGCAGATTCACTTGAGCCAAAGGATTC
	3'	[Btn]AAACCATGGTGCTCTCTGGGGAAGACAAAAGCAACATCAAGGCTGCCTGGGGGAAGATTGGTGGCCATGGTGCTGAATATGGAGCTGAAGCCCTGGAAAGGTGAGAACAGGACCTTGATC
AL1 LP1	5'	[Btn]GATCTCTAGAAATGTTCTGTATCCCAGTCTGTTGGAGAATAAGTGCAAGCATTTAAACAACTTTGATTTG
	3'	[Btn]CTTGAGAATTATGAAAATTGAATTAACCATGTGCTAGATACAATGTCAAATCTATGCAAATGAAGGGATC
AL1 LP2	5'	[Btn]GATCTTAACTTTTAGGTATCTCATTCATTTCCCTTTTTGTCTCCAGGTTTCTTTTTCCTCTGATGCTAAT
	3'	[Btn]TACTTCTTCAAACAACTGGAACAGGACAGGTACGAGGAATGAAGAGGAAAGTTCTGGAGGTCAGAGGATC
AL1 LP3	5'	[Btn]GATCAGAATTAGAAGTCTGTCAAACAATAACCATAGCATGTCCCTGAGTAAAATAATCCTTCAACCTTAG
	3'	[Btn]TTCTTCTCCATGCCCCCACAATCTTTATTAGCCTCCTCCACGTTAAACTCAATTTGCTATTTTTTGGATC
AL1 LP4	5'	[Btn]GATCATCATCATCTCCAGAGAGATGGCCTTATATTAGGGTCAAGTAGTAGCTAGGCTTCTACACACCAAG
	3'	[Btn]TGAGGCAGAGCCTTGGGCGCCTCTCTTGCATCAGTGCTAGAATGTCCACAAGCTAGCTCTTGTTCAGATC
AL1 LP5	5'	[Btn]GATCATAGTCCCAAAGGCTTCTCCATCCTACTGCAGAGGTACCTGCTCACTCCTGTTCAGTGCTGTTCTA

	3’	[Btn]GTAAATGTAGCCCAGTTGCTAGAGGACTTGCTTAGCATGCTCAGAGCTCTAGTTTCAAGTTTCCAGGATC
AL2 LP1	5’	[Btn]GATCTATATGGGATGAGATATTAGTTGGGGTTCTCCAGAGACAGGGAGCCAATAAGATGTATGTTGTTGG
	3’	[Btn]CCCAGCCAAGAGAATGGTGTCATTCACAACTGGCTGGGTTTTCTTGTATCAATTAAGAGTCAAGATGATC
AL2 LP2	5’	[Btn]GATCATGGAATATTTTTTCTTTTGGACTTGCTTCACTCAAATAAAAAAACTGTTCACCTCATGTTTTCCA
	3’	[Btn]TGAATTCCCTGAAGGTAAGAACAGGAATTATCTTATCTCTAAACCAGGTATTGGGCAATTCAGGTGGATC
AL2 LP3	5’	[Btn]GATCATTAAAGGAGACAGGAGGAAAGGATACCATCCCCGCAGCAGGGAAATGGACTGCAACAGAGCATCC
	3’	[Btn]ATGAGGCTTTGTCCTAAGCGGCTAAGTGTCCCCGGTGTTCCTAGCAGGAATACAGGCTGGCTCATGGATC
AL2 LP4	5’	[Btn]GATCCAAAATGCTCTGCTGAACTCAAGATAGTTCATTTCACTAAAGGCCGTGCTGAGGAGCAGGGGCAAA
	3’	[Btn]GCTACCATGTAAGCCCCAGGACCAGCTATGTCTTGCTCCTGGAGAAGGCGGTCATCATGGCTACCTGATC
AL2 LP5	5’	[Btn]GATCTAGTCCTCACACATTTATGAGATGGTGTCACCAAGAGCACCTTCACAAGAAGACCAATGCGATACC
	3’	[Btn]TACTTTTCAAAGAAACTAAAAATTCTCACTACCTGCATTCATCATCCAAAATTCTCTAGAGAAACAGATC
AL2 LP6	5’	[Btn]GATCCAAATGTTCTTAAGGTCCTGAAGTAGTTTTTAAAAATAGACTTCCGTGAAGAAAGAACTCTGGTTG
	3’	[Btn]TATTTTAAAAGAAGAAAAGGAATAGACATTTGTGTAGGTCATGTTGCTGGAGTCTAGAAAGCCTAAGATC

Aliquots of 1-2 µg of differently indexed libraries were mixed with 5 µg Mouse Cot-1 DNA (Invitrogen, ThermoFisher: 18440016) and 1 nmol of universal TS HE blocking oligonucleotides per library, and 1 nmol of the index-specific blocking oligonucleotide for each library (NimbleGen SeqCap HE-Oligo Kit A, NimbleGen SeqCap HE-Oligo Kit B; Roche: 06777287001, 06777317001 respectively). The mixture was dried completely in a vacuum centrifuge (55 °C) and resuspended in 7.5 µl of NimbleGen Hybridization Buffer and 3 µl of NimbleGen Hybridisation Component A per library. The reaction was denatured at 95 °C for 10 minutes. During this time, 4.5 µl of the pooled biotinylated capture oligonucleotides (resulting in 13 fmol of each individual oligonucleotide) per sample library were heated to 47 °C in a PCR thermocycler. The denatured library mix was added to the

pre-heated oligonucleotides and hybridisation was performed at 47 °C in the thermocycler for 64-72 hours.

Enrichment for DNA fragments hybridised to the biotinylated oligonucleotides was performed by adding 100 µl of washed M-270 Streptavidin beads (Invitrogen, ThermoFisher: 65306) per library and incubating at 47 °C for 45 minutes on a shaking thermomixer to allow binding of the beads. The magnetic beads were subsequently washed with NimbleGen SeqCap EZ Wash Buffers, according to manufacturer's instructions. The washed beads were resuspended in 40 µl of water per captured library and either stored at -20 °C or split across two PCR reactions and the captured DNA was PCR amplified directly off the beads using KAPA HiFi HotStart ReadyMix and Post-LM-PCR-Oligo-nucleotides 1&2 in a 10-cycle PCR reaction. The amplified captured material was purified using Agencourt AMPure XP beads (Beckman Coulter: A63881).

To increase enrichment, a second round of oligonucleotide capture was performed following the same procedure, using all of the recovered material in a hybridisation reaction of 20-24 hours and treating it as one library. The enriched Capture-C library profile was assessed by visualisation on a D1000 tape on a Tapestation (Agilent Technologies) and quantified using a universal library quantification kit (KAPA Biosystems, Roche: 07960140001). The Capture-C library was sequenced on the Illumina Nextseq platform using a 300-cycle paired-end kit (NextSeq 500/550 Mid Output Kit v2.5: 20024905). At least 1 x 10^6 150 bp paired-end reads were generated per viewpoint, per sample in multiplexed libraries to produce sufficient sequencing depth.

Tri-C

Tri-C experiments were performed on ter119+ cells isolated from phenylhydrazine-treated mouse spleens according to the Tri-C protocol which has been previously described (Oudelaar et al., 2018). The three experimental stages of Tri-C are the same as Capture-C described above with some alterations. Tri-C requires the viewpoint of interest to reside on a small restriction fragment to allow the detection of multi-way chromatin contacts. Therefore, the NlaIII restriction enzyme (New England Biolabs: R0125L) was used for digestion of 3C libraries, instead of DpnII, using three aliquots of 200 U NlaIII per tube (a total of 1800 U per 3C library). To be able to capture multiple restriction fragments in individual reads, the subsequent sonication step of the 3C library was optimised to generate DNA fragments of ~450 bp, using a Covaris S220 Focussed Ultra-Sonicator with the following settings: one cycle of 55 s; duty cycle: 10%; intensity: 4; cycles per burst: 200; set mode: frequency sweeping. DNA purification and size selection were performed after sonication using Agencourt AMPure XP beads (Beckman Coulter: A63881) at a 1:0.7 sample:bead ratio to select for larger sonication fragments. Subsequent ligation of adaptors, indexing, multiplexing of libraries, double oligonucleotide capture to enrich for the viewpoint fragments of interest, and sequencing were performed as described above, using the biotinylated DNA oligonucleotides in Table 6.11, which were designed to target the middle of the restriction fragments.

Table 6.11: Custom biotinylated DNA oligonucleotides used in Tri-C experiments

Viewpoint	**Sequence (5'-3')**
R2	[Btn]GGTCAAAGTAGCATACACCCATCTGGAACCTATCAGTGACCATAGTCAACAGCAGGTGTACA CACCCAGGCCAAGGGTGGAGCAGACCACTGTGGGATCTATGGAGATGCTTGAACGAGC
Mpg	[Btn]TCCGGTGGCCTGGCCTGTGCTGGCGGCGACTAGATGCCCGCGCGCGGTGGTAGTGCGCGCCC GGGCAGAGGAGCCCTAAAACCGGTGTCCGTGACCCTGCTCCCCGACACCGAGCAGCCT

RNA isolation and gene expression analysis

RNA isolation

Total RNA was isolated from 1-5 x 10^6 ter119+ and CD71+ cells (isolated from phenylhydrazine-treated mouse spleens and from cultures of embryoid bodies, respectively) lysed in TRI reagent (Sigma-Aldrich: T9424) using a Direct-zol RNA MiniPrep kit (Zymo Research: R2050). DNase I treatment was performed on the column as recommended in the manufacturer's instructions but with an increased incubation of 30 minutes at room temperature (rather than 15 minutes). RNA was eluted in DNase/RNase-Free Water and degradation was assessed using RNA Screentape Analysis (Tapestation, Agilent Technologies) which determined the RNA integrity number (RIN) score of the sample. RNA was quantified either by Qubit RNA Broad-Range Assay (Invitrogen, ThermoFisher: Q10210) or by NanoDrop (ThermoFisher). RNA samples were stored at -80 °C.

RT-qPCR expression analysis

To assess relative changes in gene expression by qPCR, cDNA was synthesised from 1 µg of total RNA using SuperScript III First-Strand Synthesis SuperMix for qRT-PCR (Invitrogen, ThermoFisher: 11752-050) according to the manufacturer's instructions. For every RNA sample, a control reaction was also run without the addition of the Reverse Transcriptase (RT) enzyme mix to account for any incomplete genomic DNA digestion in the sample. The resulting 20 µl reaction was diluted 1:6 (an addition of 100 µl water) and could be stored at -20 °C. The diluted cDNA was used in qPCR with either Fast SYBR Green Master Mix (Applied Biosystems, ThermoFisher: 4385612) using the primers in Table 6.12, or with TaqMan Universal PCR Master Mix (Applied Biosystems, ThermoFisher: 4304437) using the TaqMan probes in Table 6.13. The ΔΔCt method was used for relative quantification of RNA abundance for both master mixes.

Table 6.12: Primers used for RT-qPCR gene expression analysis

ID	Target	Primer	Sequence (5'-3')
1	mVenus	Forward	CAAAGACCCCAACGAGAAGC
		Reverse	CTTGTACAGCTCGTCCATGC
2	Rn18s	Forward	GTAACCCGTTGAACCCCATT
		Reverse	CCATCCAATCGGTAGTAGCG

Table 6.13: Taqman probes used for RT-qPCR gene expression analysis

ID	Target	Taqman Probe Code
1	Hba-α1/Hba-α2	Mm00845395_s1
2	Hbb-b1	Mm01611268_g1
3	Sh3pxd2b	Mm00616672_m1
4	Ubtd2	Mm00612868_m1
5	Rn18s	Mm04277571_s1

Poly(A)+ RNA-seq

To assess global changes in mRNA, 1-2 µg of total RNA was depleted of rRNA and globin mRNA using the Globin-Zero Gold rRNA Removal Kit (Illumina: GZG1224) according to the manufacturer's instructions. Globin-Zero depleted samples were purified by ethanol precipitation overnight at -80 °C, before being spun at 12,000 g for 1 hour, washed twice with freshly prepared 70% ethanol, and resuspended in 50 µl of RNase-free water. RNA samples were checked to ensure sufficient depletion and lack of degradation using RNA Screentape Analysis (Tapestation, Agilent Technologies).

To enrich for mRNA, poly(A)+ RNA was isolated, strand-specific cDNA was synthesised, and the resulting libraries prepared for Illumina sequencing using the NEBNext Poly(A) mRNA Magnetic Isolation Module (New England Biolabs: E7490) and the NEBNext Ultra II Directional RNA Library Prep Kit for Illumina (New England Biolabs: E7760) following the manufacturer's instructions. The poly(A)- RNA fraction was retained by storing the supernatant and all subsequent washes of the magnetic Oligo dT Beads bound with mRNA (steps 1.2.16 - 1.2.34; ~800 µl in total) at -80 °C. Purification of the double-stranded cDNA, the adaptor ligation reaction, and the PCR reaction of adaptor ligated DNA were all performed using Agencourt AMPure XP beads (Beckman Coulter: A63881). Library profiles were assessed by visualisation on a D1000 tape on a Tapestation (Agilent Technologies) and quantified using a universal library quantification kit (KAPA Biosystems, Roche: 07960140001). Poly(A)+ RNA-seq libraries were sequenced on the Illumina Nextseq platform using a 75-cycle paired-end kit (NextSeq 500/550 High Output Kit v2.5: 20024906).

Poly(A)- RNA-seq

The poly(A)- RNA fraction (isolated from samples as described above) was purified using Agencourt RNAClean XP beads (Beckman Coulter: A63987) and eluted in 50 µl of RNase-free water. An additional round of poly(A)+ selection using the magnetic Oligo dT Beads from the NEBNext Poly(A) mRNA Magnetic Isolation Module was performed to deplete the sample further of any contaminating poly(A)+ RNA. The resulting supernatant (step 1.2.16, ~100 µl) was purified using Agencourt RNAClean XP beads and eluted in 11.5 µl of the First Strand Synthesis Reaction Buffer and Random Primer Mix (2X) prepared earlier from the NEBNext Ultra II Directional RNA Library Prep Kit for Illumina. Fragmentation of RNA, strand-specific cDNA synthesis, and library preparation for Illumina sequencing

were all performed using the NEBNext Ultra II Directional RNA Library Prep Kit for Illumina according to the manufacturer's instructions (step 1.2.37 onwards). Purification of the double-stranded cDNA, the adaptor ligation reaction, and the PCR reaction of adaptor ligated DNA were all performed using Agencourt AMPure XP beads. Library profiles were assessed by visualisation on a D1000 tape on a Tapestation (Agilent Technologies) and quantified using a universal library quantification kit (KAPA Biosystems, Roche: 07960140001). Poly(A)- RNA-seq libraries were sequenced on the Illumina Nextseq platform using a 75-cycle paired-end kit (NextSeq 500/550 High Output Kit v2.5: 20024906).

Data Analysis

The following languages and software packages were used throughout data analyses and visualisation: Python, R, ucsctools (Kent et al., 2002), bedtools (Quinlan & Hall, 2010), samtools (Li et al., 2009), Bowtie1 (Langmead, Trapnell, Pop, & Salzberg, 2009), STAR (Dobin et al., 2013), deeptools (Ramirez et al., 2016), SRAtoolkit (Leinonen, Sugawara, Shumway, International Nucleotide Sequence Database Collaboration, 2011), MACS2 (Y. Zhang et al., 2008), The MEME Suite (Bailey et al., 2009) and the UCSC genome browser (Kent et al., 2002).

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

Both in-house generated and publicly-available ATAC-seq and ChIP-seq datasets were analysed using an in-house pipeline developed by Jelena Telenius, WIMM (Telenius & Hughes, 2018). Using the default settings of the pipeline, the reads underwent quality control, were aligned to the mouse mm9 reference genome, and visualisation files were generated. Full documentation can be found at:

http://userweb.molbiol.ox.ac.uk/public/telenius/NGseqBasicManual/external/. Briefly, the pipeline checked the quality of the FASTQ files using FASTQC and used Trim Galore when needed to trim adaptor sequences from the reads. Mapping was performed by using Bowtie1 (the -m 2 option was used to allow for multi-mapping in the duplicated regions in the globin loci), and samtools was used to filter out PCR duplicates and ploidy regions. Finally, ucsctools was used to generate visualisation files of the alignments, so that DNA fragments could be viewed in the UCSC genome browser. Publicly-available datasets were downloaded from the NCBI Gene Expression Omnibus (GEO) (Table 6.14) and converted into FASTQ file format using the SRAtoolkit.

Table 6.14: Publicly available datasets used in this thesis

Cell type (mouse)	**Data Type**	**GEO project submission code**	**Publication**
Ter119+	ATAC-seq; CTCF ChIP-seq H3K4me3 ChIP-seq H3K27me3 ChIP-seq Rad21 ChIP-Seq RNA polymerase II ChIP-seq	GSE97871	Hanssen et al., 2017
Ter119+	H3K4me1 ChIP-seq H3K27ac ChIP-seq	GSE27921	Kowalczyk et al., 2012
Ter119+	H3K9me3 ChIP-seq	GSE36028	ENCODE Consortium
Ter119+	Scl (Tal1) ChIP-seq	GSE51337	Pimkin et al., 2014
Ter119+	Nfe2 ChIP-seq	GSE47492	Hughes et al., 2014

Ter119+	Med1 ChIP-seq	GSE78835	Hay et al., 2016
Erythroid	DNase-seq	GSE86393	Schwessinger & Suciu et al., 2017
Erythroid	Klf1 ChIP-seq	GSE20478	Tallack et al., 2010
Ter119+	Gata1 ChIP-seq scaRNA-seq	N/A	Martin Larke, WIMM, unpublished
ES	ATAC-seq	GSE94250	Simon et al., 2017
ES	CTCF ChIP-seq H3K4me1 ChIP-seq H3K27me3 ChIP-seq	GSE30206	Stadler et al., 2011
ES	H3K27ac ChIP-seq	GSE47950	Wamstad et al., 2012
ES	H3K4me3 ChIP-seq	GSE31039	ENCODE Consortium
ES	H3K9me3 ChIP-seq	GSE57092	Bulut-Karslioglu et al., 2014
ES	Smc1 ChIP-seq	GSE22557	Kagey et al., 2010

ATAC-seq and ChIP-seq data were normalised and averaged across biological replicates to generate a single track that could be visualised in the UCSC genome browser. The replicates were normalised to Reads Per Kilobase per Million (RPKM) mapped reads using deeptools bamCoverage with the sorted alignment file generated from the in-house pipeline as input. For ATAC-seq samples, mitochondrial DNA was excluded from the normalisation. A union bedgraph of the biological replicates was generated using bedtools unionbedg and awk was used to generate a single bedgraph file representing the mean of the replicates. To visualise

averaged DNA fragments in the UCSC genome browser, ucsctools bedGraphToBigWig was used to convert the bedgraph file to bigwig format.

Peak calling of ChIP-seq data

The MACS (Model-based Analysis for 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 triplicate ChIP-seq of CTCF with a -q value of 10^{-5}.

CTCF motif analysis

To identify all significant CTCF core binding motifs throughout the mouse genome, the FIMO (Find Individual Motif Occurrences) algorithm belonging to The MEME Suite was used with a --thresh value of 10^{-4} and a --max-stored-scores value of 10^{9}. Bedtools intersect was used to intersect these motifs with CTCF peak calls from ChIP-seq data.

Analysis of Capture-C data

Capture-C data were analysed as previously described (Davies et al., 2016) using an in-house pipeline developed by Jelena Telenius. Using the default settings of the pipeline, the reads underwent quality control, were aligned to the mouse mm9 reference genome, capture-containing reads were identified, interaction fragments were reported, and visualisation files were generated. Full documentation can be found at: http://userweb.molbiol.ox.ac.uk/public/telenius/CCseqBasicManual/prePrintManuals/CCseqBasic/index.html. To generate comparisons of Capture-C interaction profiles, the pipeline outputs were used in scripts developed by Damien Downes, WIMM (https://github.com/Hughes-Genome-Group/CaptureCompare). Briefly, unique interactions

were normalised to *cis* interactions, averaged (n = 3), and the difference between the means were calculated. For visualisation, means and standard deviations were binned (150 bp) and smoothed with a sliding window (3 kb) and plotted using ggPlot in R.

Analysis of Tri-C data

Tri-C data were analysed as previously described (Oudelaar et al., 2018) using a modified version of the in-house pipeline described above (Jelena Telenius) and custom scripts. Full scripts can be found at: https://github.com/Hughes-Genome-Group/CCseqBasicS and https://github.com/oudelaar/TriC.

Mapping and visualisation of RNA-seq data and variant analysis

Reads originating from RNA-seq experiments were aligned to the mouse mm9 genome using STAR with the --outFilterMultimapNmax value set to 2. Alignment files were filtered for proper pairs, sorted, and PCR duplicates were removed using samtools. Using the sorted alignment file as input, biological replicates were normalised to Reads Per Kilobase per Million (RPKM) mapped reads in a strand-specific manner using deeptools bamCoverage (performed twice; once with --filterRNAstrand forward and again with --filterRNAstrand reverse to generate a file for each strand). Files for biological replicates were merged and converted into bigwig file format for visualisation in the UCSC genome browser.

To perform variant calling analysis on RNA-seq reads originating from α-globin transcripts an in-house variant-caller tool developed by Jelena Telenius was used, which was based on the samtools1 version of mpileup to count variants. Each sample was aligned twice to the mouse mm9 genome: once with *Hba-α1* masked and once with *Hba-α2* masked. The resulting alignments were used as inputs for the variant-caller which counted mismatches at

Hba-α1/2. Full documentation for the variant-caller tool can be found here: http://userweb.molbiol.ox.ac.uk/public/telenius/variantApp/variantApp_JTelenius_GPL3_2019.pdf.

References

Abdennur, N., Schwarzer, W., Pekowska, A., Shaltiel, I. A., Huber, W., Haering, C. H., et al. (2018). Condensin II inactivation in interphase does not affect chromatin folding or gene expression. *bioRxiv*, 437459.

Akhtar, W., de Jong, J., Pindyurin, A. V., Pagie, L., Meuleman, W., de Ridder, J., et al. (2013). Chromatin position effects assayed by thousands of reporters integrated in parallel. *Cell*, *154*(4), 914–927. http://doi.org/10.1016/j.cell.2013.07.018

Alipour, E., & Marko, J. F. (2012). Self-organization of domain structures by DNA-loop-extruding enzymes. *Nucleic Acids Research*, *40*(22), 11202–11212. http://doi.org/10.1093/nar/gks925

Allahyar, A., Vermeulen, C., Bouwman, B. A. M., Krijger, P. H. L., Verstegen, M. J. A. M., Geeven, G., et al. (2018). Enhancer hubs and loop collisions identified from single-allele topologies. *Nature Genetics*, *50*(8), 1151–1160. http://doi.org/10.1038/s41588-018-0161-5

Bailey, T. L., Boden, M., Buske, F. A., Frith, M., Grant, C. E., Clementi, L., et al. (2009). MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Research*, *37*(Web Server issue), W202–8. http://doi.org/10.1093/nar/gkp335

Banerji, J., Rusconi, S., & Schaffner, W. (1981). Expression of a β-globin gene is enhanced by remote SV40 DNA sequences. *Cell*, *27*(2), 299–308. http://doi.org/10.1016/0092-8674(81)90413-X

Bannister, A. J., & Kouzarides, T. (2011). Regulation of chromatin by histone modifications. *Cell Research*, *21*(3), 381–395. http://doi.org/10.1038/cr.2011.22

Bartman, C. R., Hsu, S. C., Hsiung, C. C.-S., Raj, A., & Blobel, G. A. (2016). Enhancer Regulation of Transcriptional Bursting Parameters Revealed by Forced Chromatin Looping. *Molecular Cell*, *62*(2), 237–247. http://doi.org/10.1016/j.molcel.2016.03.007

Beagrie, R. A., Scialdone, A., Schueler, M., Kraemer, D. C. A., Chotalia, M., Xie, S. Q., et al. (2017). Complex multi-enhancer contacts captured by genome architecture mapping. *Nature*, *543*(7646), 519–524. http://doi.org/10.1038/nature21411

Bell, A. C., West, A. G., & Felsenfeld, G. (1999). The Protein CTCF Is Required for the Enhancer Blocking Activity of Vertebrate Insulators. *Cell*, *98*(3), 387–396. http://doi.org/10.1016/S0092-8674(00)81967-4

Bender, M. A., Ragoczy, T., Lee, J., Byron, R., Telling, A., Dean, A., & Groudine, M. (2012). The hypersensitive sites of the murine β-globin locus control region act independently to affect nuclear localization and transcriptional elongation. *Blood*, *119*(16), 3820–3827. http://doi.org/10.1182/blood-2011-09-380485

Bickmore, W. A., & van Steensel, B. (2013). Genome architecture: domain organization of interphase chromosomes. *Cell*, *152*(6), 1270–1284. http://doi.org/10.1016/j.cell.2013.02.001

Bintu, B., Mateo, L. J., Su, J.-H., Sinnott-Armstrong, N. A., Parker, M., Kinrot, S., et al. (2018). Super-resolution chromatin tracing reveals domains and cooperative interactions in single cells. *Science (New York, N.Y.)*, *362*(6413), eaau1783. http://doi.org/10.1126/science.aau1783

Boehning, M., Dugast-Darzacq, C., Rankovic, M., Hansen, A. S., Yu, T., Marie-Nelly, H., et al. (2018). RNA polymerase II clustering through carboxy-terminal domain phase separation. *Nature Structural & Molecular Biology*, *25*(9), 833–840. http://doi.org/10.1038/s41594-018-0112-y

Boija, A., Klein, I. A., Sabari, B. R., Dall'Agnese, A., Coffey, E. L., Zamudio, A. V., et al. (2018). Transcription Factors Activate Genes through the Phase-Separation Capacity of Their Activation Domains. *Cell*, *175*(7), 1842–1855.e16. http://doi.org/10.1016/j.cell.2018.10.042

Bonev, B., & Cavalli, G. (2016). Organization and function of the 3D genome. *Nature Reviews. Genetics*, *17*(11), 661–678. http://doi.org/10.1038/nrg.2016.112

Bonev, B., Mendelson Cohen, N., Szabo, Q., Fritsch, L., Papadopoulos, G. L., Lubling, Y., et al. (2017). Multiscale 3D Genome Rewiring during Mouse Neural Development. *Cell*, *171*(3), 557–572.e24. http://doi.org/10.1016/j.cell.2017.09.043

Boyle, S., Gilchrist, S., Bridger, J. M., Mahy, N. L., Ellis, J. A., & Bickmore, W. A. (2001). The spatial organization of human chromosomes within the nuclei of normal and emerin-mutant cells. *Human Molecular Genetics*, *10*(3), 211–219. http://doi.org/10.1093/hmg/10.3.211

Brandão, H. B., Wang, X., Paul, P., van den Berg, A. A., Rudner, D. Z., & Mirny, L. A. (2019). RNA polymerases as moving barriers to condensin loop extrusion. *bioRxiv*, 604280. http://doi.org/10.1101/604280

Brown, J. M., Roberts, N. A., Graham, B., Waithe, D., Lagerholm, C., Telenius, J. M., et al. (2018). A tissue-specific self-interacting chromatin domain forms independently of enhancer-promoter interactions. *Nature Communications*, *9*(1), 3849. http://doi.org/10.1038/s41467-018-06248-4

Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y., & Greenleaf, W. J. (2013). Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nature Methods*, *10*(12), 1213–1218. http://doi.org/10.1038/nmeth.2688

Bulut-Karslioglu, A., La Rosa-Velázquez, De, I. A., Ramirez, F., Barenboim, M., Onishi-Seebacher, M., Arand, J., et al. (2014). Suv39h-dependent H3K9me3 marks intact retrotransposons and silences LINE elements in mouse embryonic stem cells. *Molecular Cell*, *55*(2), 277–290. http://doi.org/10.1016/j.molcel.2014.05.029

Burcin, M., Arnold, R., Lutz, M., Kaiser, B., Runge, D., Lottspeich, F., et al. (1997). Negative protein 1, which is required for function of the chicken lysozyme gene silencer in conjunction with hormone receptors, is identical to the multivalent zinc finger repressor CTCF. *Molecular and Cellular Biology*, *17*(3), 1281–1288. http://doi.org/10.1128/mcb.17.3.1281

Busslinger, G. A., Stocsits, R. R., van der Lelij, P., Axelsson, E., Tedeschi, A., Galjart, N., & Peters, J.-M. (2017). Cohesin is positioned in mammalian genomes by transcription, CTCF and Wapl. *Nature*, *460*, 410. http://doi.org/10.1038/nature22063

Carter, D., Chakalova, L., Osborne, C. S., Dai, Y.-F., & Fraser, P. (2002). Long-range chromatin regulatory interactions in vivo. *Nature Genetics*, *32*(4), 623–626. http://doi.org/10.1038/ng1051

Chambeyron, S., & Bickmore, W. A. (2004). Chromatin decondensation and nuclear reorganization of the HoxB locus upon induction of transcription. *Genes & Development*, *18*(10), 1119–1130. http://doi.org/10.1101/gad.292104

Chen, N., Harris, T. W., Antoshechkin, I., Bastiani, C., Bieri, T., Blasiar, D., et al. (2005). WormBase: a comprehensive data resource for Caenorhabditis biology and genomics. *Nucleic Acids Research*, *33*(Database issue), D383–9. http://doi.org/10.1093/nar/gki066

Chen, Y., Zhang, Y., Wang, Y., Zhang, L., Brinkman, E. K., Adam, S. A., et al. (2018). Mapping 3D genome organization relative to nuclear compartments using TSA-Seq as a cytological ruler. *The Journal of Cell Biology*, *217*(11), 4025.

Cho, S. W., Xu, J., Sun, R., Mumbach, M. R., Carter, A. C., Chen, Y. G., et al. (2018a). Promoter of lncRNA Gene PVT1 Is a Tumor-Suppressor DNA Boundary Element. *Cell*, *173*(6), 1398–1412.e22. http://doi.org/10.1016/j.cell.2018.03.068

Cho, W.-K., Spille, J.-H., Hecht, M., Lee, C., Li, C., Grube, V., & Cisse, I. I. (2018b). Mediator and RNA polymerase II clusters associate in transcription-dependent condensates. *Science (New York, N.Y.)*, *361*(6400), 412–415. http://doi.org/10.1126/science.aar4199

Chong, S., Dugast-Darzacq, C., Liu, Z., Dong, P., Dailey, G. M., Cattoglio, C., et al. (2018). Imaging dynamic and selective low-complexity domain interactions that control gene transcription. *Science (New York, N.Y.)*, *361*(6400), eaar2555. http://doi.org/10.1126/science.aar2555

Chung, J. H., Bell, A. C., & Felsenfeld, G. (1997). Characterization of the chicken beta-globin insulator. *Proceedings of the National Academy of Sciences*, *94*(2), 575–580. http://doi.org/10.1073/pnas.94.2.575

Core, L. J., Martins, A. L., Danko, C. G., Waters, C. T., Siepel, A., & Lis, J. T. (2014). Analysis of nascent RNA identifies a unified architecture of initiation regions at mammalian promoters and enhancers. *Nature Genetics*, *46*(12), 1311–1320. http://doi.org/10.1038/ng.3142

Cremer, T., & Cremer, C. (2001). Chromosome territories, nuclear architecture and gene regulation in mammalian cells. *Nature Reviews. Genetics*, *2*(4), 292–301. http://doi.org/10.1038/35066075

Cremer, T., Cremer, C., Schneider, T., Baumann, H., Hens, L., & Kirsch-Volders, M. (1982). Analysis of chromosome positions in the interphase nucleus of Chinese hamster cells by laser-UV-microirradiation experiments. *Human Genetics*, *62*(3), 201–209. http://doi.org/10.1007/BF00333519

Cremer, T., Cremer, M., Dietzel, S., Müller, S., Solovei, I., & Fakan, S. (2006). Chromosome territories – a functional nuclear landscape. *Nucleus and Gene Expression*, *18*(3), 307–316.

Creyghton, M. P., Cheng, A. W., Welstead, G. G., Kooistra, T., Carey, B. W., Steine, E. J., et al. (2010). Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(50), 21931–21936. http://doi.org/10.1073/pnas.1016071107

Cuddapah, S., Jothi, R., Schones, D. E., Roh, T.-Y., Cui, K., & Zhao, K. (2009). Global analysis of the insulator binding protein CTCF in chromatin barrier regions reveals demarcation of active and repressive domains. *Genome Research*, *19*(1), 24–32. http://doi.org/10.1101/gr.082800.108

Cullen, K. E., Kladde, M. P., & Seyfred, M. A. (1993). Interaction between transcription regulatory regions of prolactin chromatin. *Science*, *261*(5118), 203–206. http://doi.org/10.1126/science.8327891

Davies, J. O. J., Oudelaar, A. M., Higgs, D. R., & Hughes, J. R. (2017). How best to identify chromosomal interactions: a comparison of approaches. *Nature Methods*, *14*(2), 125–134. http://doi.org/10.1038/nmeth.4146

Davies, J. O. J., Telenius, J. M., McGowan, S. J., Roberts, N. A., Taylor, S., Higgs, D. R., & Hughes, J. R. (2016). Multiplexed analysis of chromosome conformation at vastly improved sensitivity. *Nature Methods*, *13*(1), 74–80. http://doi.org/10.1038/nmeth.3664

De Gobbi, M., Anguita, E., Hughes, J., Sloane-Stanley, J. A., Sharpe, J. A., Koch, C. M., et al. (2007). Tissue-specific histone modification and transcription factor binding in alpha globin gene expression. *Blood*, *110*(13), 4503–4510. http://doi.org/10.1182/blood-2007-06-097964

De Gobbi, M., Viprakasit, V., Hughes, J. R., Fisher, C., Buckle, V. J., Ayyub, H., et al. (2006). A regulatory SNP causes a human genetic disease by creating a new transcriptional promoter. *Science (New York, N.Y.)*, *312*(5777), 1215–1217. http://doi.org/10.1126/science.1126431

de Laat, W., & Duboule, D. (2013). Topology of mammalian developmental enhancers and their regulatory landscapes. *Nature*, *502*(7472), 499–506. http://doi.org/10.1038/nature12753

De Santa, F., Barozzi, I., Mietton, F., Ghisletti, S., Polletti, S., Tusi, B. K., et al. (2010). A large fraction of extragenic RNA pol II transcription sites overlap enhancers. *PLoS Biology*, *8*(5), e1000384. http://doi.org/10.1371/journal.pbio.1000384

de Wit, E., Vos, E. S. M., Holwerda, S. J. B., Valdes-Quezada, C., Verstegen, M. J. A. M., Teunissen, H., et al. (2015). CTCF Binding Polarity Determines Chromatin Looping. *Molecular Cell*, *60*(4), 676–684. http://doi.org/10.1016/j.molcel.2015.09.023

Dekker, J. (2002). Capturing Chromosome Conformation. *Science (New York, N.Y.)*, *295*(5558), 1306–1311. http://doi.org/10.1126/science.1067799

Dekker, J., & Mirny, L. (2016). The 3D Genome as Moderator of Chromosomal Communication. *Cell*, *164*(6), 1110–1121. http://doi.org/10.1016/j.cell.2016.02.007

Deng, W., Lee, J., Wang, H., Miller, J., Reik, A., Gregory, P. D., et al. (2012). Controlling long-range genomic interactions at a native locus by targeted tethering of a looping factor. *Cell*, *149*(6), 1233–1244. http://doi.org/10.1016/j.cell.2012.03.051

Deng, W., Rupon, J. W., Krivega, I., Breda, L., Motta, I., Jahn, K. S., et al. (2014). Reactivation of developmentally silenced globin genes by forced chromatin looping. *Cell*, *158*(4), 849–860. http://doi.org/10.1016/j.cell.2014.05.050

Denker, A., & de Laat, W. (2016). The second decade of 3C technologies: detailed insights into nuclear organization. *Genes & Development*, *30*(12), 1357–1382. http://doi.org/10.1101/gad.281964.116

Dileep, V., Ay, F., Sima, J., Vera, D. L., Noble, W. S., & Gilbert, D. M. (2015). Topologically associating domains and their long-range contacts are established during early G1 coincident with the establishment of the replication-timing program. *Genome Research*, *25*(8), 1104–1113. http://doi.org/10.1101/gr.183699.114

Dixon, J. R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., et al. (2012). Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, *485*(7398), 376–380. http://doi.org/10.1038/nature11082

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., et al. (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics (Oxford, England)*, *29*(1), 15–21. http://doi.org/10.1093/bioinformatics/bts635

Dowen, J. M., Fan, Z. P., Hnisz, D., Ren, G., Abraham, B. J., Zhang, L. N., et al. (2014). Control of cell identity genes occurs in insulated neighborhoods in mammalian chromosomes. *Cell*, *159*(2), 374–387. http://doi.org/10.1016/j.cell.2014.09.030

ENCODE Project Consortium. (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature*, *489*(7414), 57–74. http://doi.org/10.1038/nature11247

Falk, M., Feodorova, Y., Naumova, N., Imakaev, M., Lajoie, B. R., Leonhardt, H., et al. (2018). Heterochromatin drives organization of conventional and inverted nuclei. *bioRxiv*, 244038.

Faure, A. J., Schmidt, D., Watt, S., Schwalie, P. C., Wilson, M. D., Xu, H., et al. (2012). Cohesin regulates tissue-specific expression by stabilizing highly occupied cis-regulatory modules. *Genome Research*, *22*(11), 2163–2175. http://doi.org/10.1101/gr.136507.111

Fedoriw, A. M., Stein, P., Svoboda, P., Schultz, R. M., & Bartolomei, M. S. (2004). Transgenic RNAi reveals essential function for CTCF in H19 gene imprinting. *Science (New York, N.Y.)*, *303*(5655), 238–240. http://doi.org/10.1126/science.1090934

Finlan, L. E., Sproul, D., Thomson, I., Boyle, S., Kerr, E., Perry, P., et al. (2008). Recruitment to the nuclear periphery can alter expression of genes in human cells. *PLoS Genetics*, *4*(3), e1000039. http://doi.org/10.1371/journal.pgen.1000039

Flavahan, W. A., Drier, Y., Liau, B. B., Gillespie, S. M., Venteicher, A. S., Stemmer-Rachamimov, A. O., et al. (2016). Insulator dysfunction and oncogene activation in IDH mutant gliomas. *Nature*, *529*(7584), 110–114. http://doi.org/10.1038/nature16490

Flint, J., Tufarelli, C., Peden, J., Clark, K., Daniels, R. J., Hardison, R., et al. (2001). Comparative genome analysis delimits a chromosomal domain and identifies key regulatory elements in the alpha globin cluster. *Human Molecular Genetics*, *10*(4), 371–382.

Francastel, C., Walters, M. C., Groudine, M., & Martin, D. I. (1999). A functional enhancer suppresses silencing of a transgene and prevents its localization close to centrometric heterochromatin. *Cell*, *99*(3), 259–269. http://doi.org/10.1016/s0092-8674(00)81657-8

Franke, M., Ibrahim, D. M., Andrey, G., Schwarzer, W., Heinrich, V., Schöpflin, R., et al. (2016). Formation of new chromatin domains determines pathogenicity of genomic duplications. *Nature*, *538*(7624), 265–269. http://doi.org/10.1038/nature19800

Fudenberg, G., Abdennur, N., Imakaev, M., Goloborodko, A., & Mirny, L. A. (2017). Emerging Evidence of Chromosome Folding by Loop Extrusion. *Cold Spring Harbor Symposia on Quantitative Biology*, *82*, 45–55. http://doi.org/10.1101/sqb.2017.82.034710

Fudenberg, G., Imakaev, M., Lu, C., Goloborodko, A., Abdennur, N., & Mirny, L. A. (2016). Formation of Chromosomal Domains by Loop Extrusion. *Cell Reports*, *15*(9), 2038–2049. http://doi.org/10.1016/j.celrep.2016.04.085

Fukaya, T., Lim, B., & Levine, M. (2016). Enhancer Control of Transcriptional Bursting. *Cell*, *166*(2), 358–368. http://doi.org/10.1016/j.cell.2016.05.025

Furlong, E. E. M., & Levine, M. (2018). Developmental enhancers and chromosome topology. *Science (New York, N.Y.)*, *361*(6409), 1341–1345. http://doi.org/10.1126/science.aau0320

Ganji, M., Shaltiel, I. A., Bisht, S., Kim, E., Kalichava, A., Haering, C. H., & Dekker, C. (2018). Real-time imaging of DNA loop extrusion by condensin. *Science (New York, N.Y.)*, *360*(6384), 102–105. http://doi.org/10.1126/science.aar7831

Gibcus, J. H., & Dekker, J. (2013). The hierarchy of the 3D genome. *Molecular Cell*, *49*(5), 773–782. http://doi.org/10.1016/j.molcel.2013.02.011

Giorgetti, L., Galupa, R., Nora, E. P., Piolot, T., Lam, F., Dekker, J., et al. (2014). Predictive polymer modeling reveals coupled fluctuations in chromosome conformation and transcription. *Cell*, *157*(4), 950–963. http://doi.org/10.1016/j.cell.2014.03.025

Gómez-Marín, C., Tena, J. J., Acemel, R. D., López-Mayorga, M., Naranjo, S., la Calle-Mustienes, de, E., et al. (2015). Evolutionary comparison reveals that diverging CTCF sites are signatures of ancestral topological associating domains borders. *Proceedings of the National Academy of Sciences of the United States of America*, *112*(24), 7542–7547. http://doi.org/10.1073/pnas.1505463112

Guelen, L., Pagie, L., Brasset, E., Meuleman, W., Faza, M. B., Talhout, W., et al. (2008). Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions. *Nature*, *453*(7197), 948–951. http://doi.org/10.1038/nature06947

Guo, Y., Xu, Q., Canzio, D., Shou, J., Li, J., Gorkin, D. U., et al. (2015). CRISPR Inversion of CTCF Sites Alters Genome Topology and Enhancer/Promoter Function. *Cell*, *162*(4), 900–910. http://doi.org/10.1016/j.cell.2015.07.038

Haarhuis, J. H. I., van der Weide, R. H., Blomen, V. A., Yáñez-Cuna, J. O., Amendola, M., van Ruiten, M. S., et al. (2017). The Cohesin Release Factor WAPL Restricts Chromatin Loop Extension. *Cell*, *169*(4), 693–707.e14. http://doi.org/10.1016/j.cell.2017.04.013

Haddad, N., Jost, D., & Vaillant, C. (2017). Perspectives: using polymer modeling to understand the formation and function of nuclear compartments. *Chromosome Research : an International Journal on the Molecular, Supramolecular and Evolutionary Aspects of Chromosome Biology*, *25*(1), 35–50. http://doi.org/10.1007/s10577-016-9548-2

Hadjur, S., Williams, L. M., Ryan, N. K., Cobb, B. S., Sexton, T., Fraser, P., et al. (2009). Cohesins form chromosomal cis-interactions at the developmentally regulated IFNG locus. *Nature*, *460*(7253), 410–413. http://doi.org/10.1038/nature08079

Handoko, L., Xu, H., Li, G., Ngan, C. Y., Chew, E., Schnapp, M., et al. (2011). CTCF-mediated functional chromatin interactome in pluripotent cells. *Nature Genetics*, *43*(7), 630–638. http://doi.org/10.1038/ng.857

Hansen, A. S., Pustova, I., Cattoglio, C., Tjian, R., & Darzacq, X. (2017). CTCF and cohesin regulate chromatin loop stability with distinct dynamics. *eLife*, *6*, 2848. http://doi.org/10.7554/eLife.25776

Hanssen, L. L. P., Kassouf, M. T., Oudelaar, A. M., Biggs, D., Preece, C., Downes, D. J., et al. (2017). Tissue-specific CTCF-cohesin-mediated chromatin architecture delimits enhancer interactions and function in vivo. *Nature Cell Biology*, *19*(8), 952–961. http://doi.org/10.1038/ncb3573

Hatton, C. S., Wilkie, A. O., Drysdale, H. C., Wood, W. G., Vickers, M. A., Sharpe, J., et al. (1990). Alpha-thalassemia caused by a large (62 kb) deletion upstream of the human alpha globin gene cluster. *Blood*, *76*(1), 221.

Hay, D., Hughes, J. R., Babbs, C., Davies, J. O. J., Graham, B. J., Hanssen, L. L. P., et al. (2016). Genetic dissection of the α-globin super-enhancer in vivo. *Nature Genetics*, *48*(8), 895–903. http://doi.org/10.1038/ng.3605

Heger, P., Marin, B., Bartkuhn, M., Schierenberg, E., & Wiehe, T. (2012). The chromatin insulator CTCF and the emergence of metazoan diversity. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(43), 17507–17512. http://doi.org/10.1073/pnas.1111941109

Heintzman, N. D., Stuart, R. K., Hon, G., Fu, Y., Ching, C. W., Hawkins, R. D., et al. (2007). Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nature Genetics*, *39*(3), 311–318. http://doi.org/10.1038/ng1966

Heinz, S., Romanoski, C. E., Benner, C., & Glass, C. K. (2015). The selection and function of cell type-specific enhancers. *Nature Reviews. Molecular Cell Biology*, *16*(3), 144–154. http://doi.org/10.1038/nrm3949

Heinz, S., Texari, L., Hayes, M. G. B., Urbanowski, M., Chang, M. W., Givarkes, N., et al. (2018). Transcription Elongation Can Affect Genome 3D Structure. *Cell*, *174*(6), 1522–1536.e22. http://doi.org/10.1016/j.cell.2018.07.047

Higgs, D. R., & Wood, W. G. (2008). Long-range regulation of alpha globin gene expression during erythropoiesis. *Current Opinion in Hematology*, *15*(3), 176–183. http://doi.org/10.1097/MOH.0b013e3282f734c4

Higgs, D. R., Wood, W. G., Jarman, A. P., Sharpe, J., Lida, J., Pretorius, I. M., & Ayyub, H. (1990). A major positive regulatory region located far upstream of the human alpha-globin gene locus. *Genes & Development*, *4*(9), 1588–1601. http://doi.org/10.1101/gad.4.9.1588

Hirayama, T., Tarusawa, E., Yoshimura, Y., Galjart, N., & Yagi, T. (2012). CTCF is required for neural development and stochastic expression of clustered Pcdh genes in neurons. *Cell Reports*, *2*(2), 345–357. http://doi.org/10.1016/j.celrep.2012.06.014

Hnisz, D., Shrinivas, K., Young, R. A., Chakraborty, A. K., & Sharp, P. A. (2017). A Phase Separation Model for Transcriptional Control. *Cell*, *169*(1), 13–23. http://doi.org/10.1016/j.cell.2017.02.007

Hnisz, D., Weintraub, A. S., Day, D. S., Valton, A.-L., Bak, R. O., Li, C. H., et al. (2016). Activation of proto-oncogenes by disruption of chromosome neighborhoods. *Science (New York, N.Y.)*, *351*(6280), 1454–1458. http://doi.org/10.1126/science.aad9024

Hughes, J. R., Cheng, J.-F., Ventress, N., Prabhakar, S., Clark, K., Anguita, E., et al. (2005). Annotation of cis-regulatory elements by identification, subclassification, and functional assessment of multispecies conserved sequences. *Proceedings of the National Academy of Sciences*, *102*(28), 9830–9835. http://doi.org/10.1073/pnas.0503401102

Hughes, J. R., Roberts, N., McGowan, S., Hay, D., Giannoulatou, E., Lynch, M., et al. (2014). Analysis of hundreds of cis-regulatory landscapes at high resolution in a single,

high-throughput experiment. *Nature Genetics*, *46*(2), 205–212. http://doi.org/10.1038/ng.2871

Hyle, J., Zhang, Y., Wright, S., Xu, B., Shao, Y., Easton, J., et al. (2019). Acute depletion of CTCF directly affects MYC regulation through loss of enhancer-promoter looping. *Nucleic Acids Research*, *47*(13), 6699–6713. http://doi.org/10.1093/nar/gkz462

Jackson, D. A., Hassan, A. B., Errington, R. J., & Cook, P. R. (1993). Visualization of focal sites of transcription within human nuclei. *The EMBO Journal*, *12*(3), 1059–1065.

Jost, D., Carrivain, P., Cavalli, G., & Vaillant, C. (2014). Modeling epigenome folding: formation and dynamics of topologically associated chromatin domains. *Nucleic Acids Research*, *42*(15), 9553–9561. http://doi.org/10.1093/nar/gku698

Kagey, M. H., Newman, J. J., Bilodeau, S., Zhan, Y., Orlando, D. A., van Berkum, N. L., et al. (2010). Mediator and cohesin connect gene expression and chromatin architecture. *Nature*, *467*(7314), 430–435. http://doi.org/10.1038/nature09380

Keller, G. M. (1995). In vitro differentiation of embryonic stem cells. *Current Opinion in Cell Biology*, *7*(6), 862–869.

Kellum, R., & Schedl, P. (1991). A position-effect assay for boundaries of higher order chromosomal domains. *Cell*, *64*(5), 941–950. http://doi.org/10.1016/0092-8674(91)90318-s

Kent, W. J., Sugnet, C. W., Furey, T. S., Roskin, K. M., Pringle, T. H., Zahler, A. M., & Haussler, D. (2002). The human genome browser at UCSC. *Genome Research*, *12*(6), 996–1006. http://doi.org/10.1101/gr.229102

Kim, T.-K., Hemberg, M., Gray, J. M., Costa, A. M., Bear, D. M., Wu, J., et al. (2010). Widespread transcription at neuronal activity-regulated enhancers. *Nature*, *465*(7295), 182–187. http://doi.org/10.1038/nature09033

Kioussis, D., Vanin, E., deLange, T., Flavell, R. A., & Grosveld, F. G. (1983). Beta-globin gene inactivation by DNA translocation in gamma beta-thalassaemia. *Nature*, *306*(5944), 662–666. http://doi.org/10.1038/306662a0

Klenova, E. M., Nicolas, R. H., Paterson, H. F., Carne, A. F., Heath, C. M., Goodwin, G. H., et al. (1993). CTCF, a conserved nuclear factor required for optimal transcriptional activity of the chicken c-myc gene, is an 11-Zn-finger protein differentially expressed in multiple forms. *Molecular and Cellular Biology*, *13*(12), 7612–7624. http://doi.org/10.1128/mcb.13.12.7612

Kosak, S. T., Skok, J. A., Medina, K. L., Riblet, R., Le Beau, M. M., Fisher, A. G., & Singh, H. (2002). Subnuclear compartmentalization of immunoglobulin loci during lymphocyte development. *Science (New York, N.Y.)*, *296*(5565), 158–162. http://doi.org/10.1126/science.1068768

Kowalczyk, M. S., Hughes, J. R., Garrick, D., Lynch, M. D., Sharpe, J. A., Sloane-Stanley, J. A., et al. (2012). Intragenic enhancers act as alternative promoters. *Molecular Cell*, *45*(4), 447–458. http://doi.org/10.1016/j.molcel.2011.12.021

Kremers, G.-J., Goedhart, J., van Munster, E. B., & Gadella, T. W. J. (2006). Cyan and yellow super fluorescent proteins with improved brightness, protein folding, and FRET Förster radius. *Biochemistry*, *45*(21), 6570–6580. http://doi.org/10.1021/bi0516273

Kundu, S., Ji, F., Sunwoo, H., Jain, G., Lee, J. T., Sadreyev, R. I., et al. (2017). Polycomb Repressive Complex 1 Generates Discrete Compacted Domains that Change during Differentiation. *Molecular Cell*, *65*(3), 432–446.e5. http://doi.org/10.1016/j.molcel.2017.01.009

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., et al. (2001). Initial sequencing and analysis of the human genome. *Nature*, *409*(6822), 860–921. http://doi.org/10.1038/35057062

Langmead, B., Trapnell, C., Pop, M., & Salzberg, S. L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology*, *10*(3), R25–10. http://doi.org/10.1186/gb-2009-10-3-r25

Lee, H. K., Willi, M., Wang, C., Yang, C. M., Smith, H. E., Liu, C., & Hennighausen, L. (2017). Functional assessment of CTCF sites at cytokine-sensing mammary enhancers using CRISPR/Cas9 gene editing in mice. *Nucleic Acids Research*, 1–13. http://doi.org/10.1093/nar/gkx185

Lee, K., Hsiung, C. C.-S., Huang, P., Raj, A., & Blobel, G. A. (2015). Dynamic enhancer–gene body contacts during transcription elongation. *Genes & Development*, *29*(19), 1992–1997. http://doi.org/10.1101/gad.255265.114

Leibler, L. (1980). Theory of Microphase Separation in Block Copolymers. *Macromolecules*, *13*(6), 1602–1617. http://doi.org/10.1021/ma60078a047

Leinonen, R., Sugawara, H., Shumway, M., International Nucleotide Sequence Database Collaboration. (2011). The sequence read archive. *Nucleic Acids Research*, *39*(Database issue), D19–21. http://doi.org/10.1093/nar/gkq1019

Lettice, L. A., Heaney, S. J. H., Purdie, L. A., Li, L., de Beer, P., Oostra, B. A., et al. (2003). A long-range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly. *Human Molecular Genetics*, *12*(14), 1725–1735. http://doi.org/10.1093/hmg/ddg180

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., et al. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics (Oxford, England)*, *25*(16), 2078–2079. http://doi.org/10.1093/bioinformatics/btp352

Lieberman-Aiden, E., van Berkum, N. L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., et al. (2009). Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome. *Science (New York, N.Y.)*, *326*(5950), 289–293. http://doi.org/10.1126/science.1181369

Liebhaber, S. A., & Kan, Y. W. (1981). Differentiation of the mRNA transcripts originating from the alpha 1- and alpha 2-globin loci in normals and alpha-thalassemics. *The Journal of Clinical Investigation*, *68*(2), 439–446. http://doi.org/10.1172/jci110273

Liebhaber, S. A., Cash, F. E., & Main, D. M. (1985). Compensatory increase in alpha 1-globin gene expression in individuals heterozygous for the alpha-thalassemia-2 deletion. *The Journal of Clinical Investigation*, *76*(3), 1057–1064. http://doi.org/10.1172/JCI112058

Lindblad-Toh, K., Garber, M., Zuk, O., Lin, M. F., Parker, B. J., Washietl, S., et al. (2011). A high-resolution map of human evolutionary constraint using 29 mammals. *Nature*, *478*(7370), 476–482. http://doi.org/10.1038/nature10530

Liu, N. Q., Maresca, M., van den Brand, T., Braccioli, L., Schijns, M. M. G. A., Teunissen, H., et al. (2019). WAPL maintains dynamic cohesin to preserve lineage specific distal gene regulation. *bioRxiv*, 731141.

Liu, Z., Scannell, D. R., Eisen, M. B., & Tjian, R. (2011). Control of embryonic stem cell lineage commitment by core promoter factor, TAF3. *Cell*, *146*(5), 720–731. http://doi.org/10.1016/j.cell.2011.08.005

Long, H. K., Prescott, S. L., & Wysocka, J. (2016). Ever-Changing Landscapes: Transcriptional Enhancers in Development and Evolution. *Cell*, *167*(5), 1170–1187. http://doi.org/10.1016/j.cell.2016.09.018

Ludwig, M. Z., Bergman, C., Patel, N. H., & Kreitman, M. (2000). Evidence for stabilizing selection in a eukaryotic enhancer element. *Nature*, *403*(6769), 564–567. http://doi.org/10.1038/35000615

Lupiáñez, D. G., Kraft, K., Heinrich, V., Krawitz, P., Brancati, F., Klopocki, E., et al. (2015). Disruptions of Topological Chromatin Domains Cause Pathogenic Rewiring of Gene-Enhancer Interactions. *Cell*, *161*(5), 1012–1025. http://doi.org/10.1016/j.cell.2015.04.004

MAXAM, A. M., TIZARD, R., SKRYABIN, K. G., & GILBERT, W. (1977). Promoter region for yeast 5S ribosomal RNA. *Nature*, *267*(5612), 643–645.

Merkenschlager, M., & Nora, E. P. (2016). CTCF and Cohesin in Genome Folding and Transcriptional Gene Regulation. *Annual Review of Genomics and Human Genetics*, *17*(1), 17–43. http://doi.org/10.1146/annurev-genom-083115-022339

Mikhaylichenko, O., Bondarenko, V., Harnett, D., Schor, I. E., Males, M., Viales, R. R., & Furlong, E. E. M. (2018). The degree of enhancer or promoter activity is reflected by the levels and directionality of eRNA transcription. *Genes & Development*, *32*(1), 42–57. http://doi.org/10.1101/gad.308619.117

Mirny, L. A., Imakaev, M., & Abdennur, N. (2019). Two major mechanisms of chromosome organization. *Current Opinion in Cell Biology*, *58*, 142–152. http://doi.org/10.1016/j.ceb.2019.05.001

Nagano, T., Lubling, Y., Stevens, T. J., Schoenfelder, S., Yaffe, E., Dean, W., et al. (2013). Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. *Nature*, *502*(7469), 59–64. http://doi.org/10.1038/nature12593

Nakahashi, H., Kwon, K.-R. K., Resch, W., Vian, L., Dose, M., Stavreva, D., et al. (2013). A genome-wide map of CTCF multivalency redefines the CTCF code. *Cell Reports*, *3*(5), 1678–1689. http://doi.org/10.1016/j.celrep.2013.04.024

Narendra, V., Bulajić, M., Dekker, J., Mazzoni, E. O., & Reinberg, D. (2016). CTCF-mediated topological boundaries during development foster appropriate gene regulation. *Genes & Development*, *30*(24), 2657–2662. http://doi.org/10.1101/gad.288324.116

Narendra, V., Rocha, P. P., An, D., Raviram, R., Skok, J. A., Mazzoni, E. O., & Reinberg, D. (2015). CTCF establishes discrete functional chromatin domains at the Hox clusters during differentiation. *Science (New York, N.Y.)*, *347*(6225), 1017–1021. http://doi.org/10.1126/science.1262088

Nasmyth, K. (2001). Disseminating the genome: joining, resolving, and separating sister chromatids during mitosis and meiosis. *Annual Review of Genetics*, *35*(1), 673–745. http://doi.org/10.1146/annurev.genet.35.102401.091334

Nasmyth, K., & Haering, C. H. (2009). Cohesin: its roles and mechanisms. *Annual Review of Genetics*, *43*(1), 525–558. http://doi.org/10.1146/annurev-genet-102108-134233

Nguyen, T. A., Jones, R. D., Snavely, A. R., Pfenning, A. R., Kirchner, R., Hemberg, M., & Gray, J. M. (2016). High-throughput functional comparison of promoter and enhancer activities. *Genome Research*, *26*(8), 1023–1033. http://doi.org/10.1101/gr.204834.116

Nichols, J., Evans, E. P., & Smith, A. G. (1990). Establishment of germ-line-competent embryonic stem (ES) cells using differentiation inhibiting activity. *Development (Cambridge, England)*, *110*(4), 1341–1348.

Nichols, M. H., & Corces, V. G. (2018). A tethered-inchworm model of SMC DNA translocation. *Nature Structural & Molecular Biology*, *25*(10), 906–910. http://doi.org/10.1038/s41594-018-0135-4

Nir, G., Farabella, I., Pérez Estrada, C., Ebeling, C. G., Beliveau, B. J., Sasaki, H. M., et al. (2018). Walking along chromosomes with super-resolution imaging, contact maps, and integrative modeling. *PLoS Genetics*, *14*(12), e1007872. http://doi.org/10.1371/journal.pgen.1007872

Nishimura, K., Fukagawa, T., Takisawa, H., Kakimoto, T., & Kanemaki, M. (2009). An auxin-based degron system for the rapid depletion of proteins in nonplant cells. *Nature Methods*, *6*(12), 917–922. http://doi.org/10.1038/nmeth.1401

Nolis, I. K., McKay, D. J., Mantouvalou, E., Lomvardas, S., Merika, M., & Thanos, D. (2009). Transcription factors mediate long-range enhancer-promoter interactions. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(48), 20222–20227. http://doi.org/10.1073/pnas.0902454106

Noordermeer, D., Branco, M. R., Splinter, E., Klous, P., van Ijcken, W., Swagemakers, S., et al. (2008). Transcription and chromatin organization of a housekeeping gene cluster containing an integrated beta-globin locus control region. *PLoS Genetics*, *4*(3), e1000016. http://doi.org/10.1371/journal.pgen.1000016

Nora, E. P., Goloborodko, A., Valton, A.-L., Gibcus, J. H., Uebersohn, A., Abdennur, N., et al. (2017). Targeted Degradation of CTCF Decouples Local Insulation of Chromosome Domains from Genomic Compartmentalization. *Cell*, *169*(5), 930–944.e22. http://doi.org/10.1016/j.cell.2017.05.004

Nora, E. P., Lajoie, B. R., Schulz, E. G., Giorgetti, L., Okamoto, I., Servant, N., et al. (2012). Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature*, *485*(7398), 381–385. http://doi.org/10.1038/nature11049

Nuebler, J., Fudenberg, G., Imakaev, M., Abdennur, N., & Mirny, L. A. (2018). Chromatin organization by an interplay of loop extrusion and compartmental segregation. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(29), E6697–E6706. http://doi.org/10.1073/pnas.1717730115

Ong, C.-T., & Corces, V. G. (2014). CTCF: an architectural protein bridging genome topology and function. *Nature Reviews. Genetics*, *15*(4), 234–246. http://doi.org/10.1038/nrg3663

Orkin, S. H., & Goff, S. C. (1981). The duplicated human alpha-globin genes: their relative expression as measured by RNA analysis. *Cell*, *24*(2), 345–351. http://doi.org/10.1016/0092-8674(81)90324-x

Osborne, C. S., Chakalova, L., Brown, K. E., Carter, D., Horton, A., Debrand, E., et al. (2004). Active genes dynamically colocalize to shared sites of ongoing transcription. *Nature Genetics*, *36*(10), 1065–1071. http://doi.org/10.1038/ng1423

Oudelaar, A. M., Davies, J. O. J., Hanssen, L. L. P., Telenius, J. M., Schwessinger, R., Liu, Y., et al. (2018). Single-allele chromatin interactions identify regulatory hubs in dynamic compartmentalized domains. *Nature Genetics*, *50*(12), 1744–1751. http://doi.org/10.1038/s41588-018-0253-2

Oudelaar, A. M., Harrold, C. L., Hanssen, L. L. P., Telenius, J. M., Higgs, D. R., & Hughes, J. R. (2019). A revised model for promoter competition based on multi-way chromatin interactions at the α-globin locus. *Nature Communications*, 10(1), 5412–8. http://doi.org/10.1038/s41467-019-13404-x

Palstra, R. J., Tolhuis, B., Splinter, E., Nijmeijer, R., Grosveld, F., & de Laat, W. (2003). The beta-globin nuclear compartment in development and erythroid differentiation. *Nature Genetics*, *35*(2), 190–194. http://doi.org/10.1038/ng1244

Pant, V., Kurukuti, S., Pugacheva, E., Shamsuddin, S., Mariano, P., Renkawitz, R., et al. (2004). Mutation of a single CTCF target site within the H19 imprinting control region leads to loss of Igf2 imprinting and complex patterns of de novo methylation upon maternal inheritance. *Molecular and Cellular Biology*, *24*(8), 3497–3504. http://doi.org/10.1128/mcb.24.8.3497-3504.2004

Parelho, V., Hadjur, S., Spivakov, M., Leleu, M., Sauer, S., Gregson, H. C., et al. (2008). Cohesins functionally associate with CTCF on mammalian chromosome arms. *Cell*, *132*(3), 422–433. http://doi.org/10.1016/j.cell.2008.01.011

Parslow, T. G., & Granner, D. K. (1983). Structure of a nuclease-sensitive region inside the immunoglobin kappa gene: evidence for a role in gene regulation. *Nucleic Acids Research*, *11*(14), 4775–4792. http://doi.org/10.1093/nar/11.14.4775

Peric-Hupkes, D., Meuleman, W., Pagie, L., Bruggeman, S. W. M., Solovei, I., Brugman, W., et al. (2010). Molecular maps of the reorganization of genome-nuclear lamina interactions during differentiation. *Molecular Cell*, *38*(4), 603–613. http://doi.org/10.1016/j.molcel.2010.03.016

Pertea, M., & Salzberg, S. L. (2010). Between a chicken and a grape: estimating the number of human genes. *Genome Biology*, *11*(5), 206.

Philipsen, S., & Hardison, R. C. (2018). Evolution of hemoglobin loci and their regulatory elements. Blood Cells, Molecules & Diseases, 70, 2–12. http://doi.org/10.1016/j.bcmd.2017.08.001

Phillips-Cremins, J. E., Sauria, M. E. G., Sanyal, A., Gerasimova, T. I., Lajoie, B. R., Bell, J. S. K., et al. (2013). Architectural protein subclasses shape 3D organization of genomes during lineage commitment. *Cell*, *153*(6), 1281–1295. http://doi.org/10.1016/j.cell.2013.04.053

Picard, D., & Schaffner, W. (1983). Correct transcription of a cloned mouse immunoglobulin gene in vivo. *Proceedings of the National Academy of Sciences of the United States of America*, *80*(2), 417–421.

Pimkin, M., Kossenkov, A. V., Mishra, T., Morrissey, C. S., Wu, W., Keller, C. A., et al. (2014). Divergent functions of hematopoietic transcription factors in lineage priming and differentiation during erythro-megakaryopoiesis. *Genome Research*, *24*(12), 1932–1944. http://doi.org/10.1101/gr.164178.113

Pott, S., & Lieb, J. D. (2015). What are super-enhancers? *Nature Genetics*, *47*(1), 8–12. http://doi.org/10.1038/ng.3167

Quinlan, A. R., & Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics (Oxford, England)*, *26*(6), 841–842. http://doi.org/10.1093/bioinformatics/btq033

Quinodoz, S. A., Ollikainen, N., Tabak, B., Palla, A., Schmidt, J. M., Detmar, E., et al. (2018). Higher-Order Inter-chromosomal Hubs Shape 3D Genome Organization in the Nucleus. *Cell*, *174*(3), 744–757.e24. http://doi.org/10.1016/j.cell.2018.05.024

Rada-Iglesias, A., Bajpai, R., Swigut, T., Brugmann, S. A., Flynn, R. A., & Wysocka, J. (2011). A unique chromatin signature uncovers early developmental enhancers in humans. *Nature*, *470*(7333), 279–283. http://doi.org/10.1038/nature09692

Ragoczy, T., Bender, M. A., Telling, A., Byron, R., & Groudine, M. (2006). The locus control region is required for association of the murine beta-globin locus with engaged transcription factories during erythroid maturation. *Genes & Development*, *20*(11), 1447–1457. http://doi.org/10.1101/gad.1419506

Ragoczy, T., Telling, A., Sawado, T., Groudine, M., & Kosak, S. T. (2003). A genetic analysis of chromosome territory looping: diverse roles for distal regulatory elements. *Chromosome Research : an International Journal on the Molecular, Supramolecular and Evolutionary Aspects of Chromosome Biology*, *11*(5), 513–525.

Ramirez, F., Ryan, D. P., Grüning, B., Bhardwaj, V., Kilpert, F., Richter, A. S., et al. (2016). deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Research*, *44*(W1), W160–5. http://doi.org/10.1093/nar/gkw257

Ran, F. A., Hsu, P. D., Wright, J., Agarwala, V., Scott, D. A., & Zhang, F. (2013). Genome engineering using the CRISPR-Cas9 system. *Nature Protocols*, *8*(11), 2281–2308. http://doi.org/10.1038/nprot.2013.143

Rao, S. S. P., Huang, S.-C., Glenn St Hilaire, B., Engreitz, J. M., Perez, E. M., Kieffer-Kwon, K.-R., et al. (2017). Cohesin Loss Eliminates All Loop Domains. *Cell*, *171*(2), 305–320.e24. http://doi.org/10.1016/j.cell.2017.09.026

Rao, S. S. P., Huntley, M. H., Durand, N. C., Stamenova, E. K., Bochkov, I. D., Robinson, J. T., et al. (2014). A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell*, *159*(7), 1665–1680. http://doi.org/10.1016/j.cell.2014.11.021

Reddy, K. L., Zullo, J. M., Bertolino, E., & Singh, H. (2008). Transcriptional repression mediated by repositioning of genes to the nuclear lamina. *Nature*, *452*(7184), 243–247. http://doi.org/10.1038/nature06727

Rubio, E. D., Reiss, D. J., Welcsh, P. L., Disteche, C. M., Filippova, G. N., Baliga, N. S., et al. (2008). CTCF physically links cohesin to chromatin. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(24), 8309–8314. http://doi.org/10.1073/pnas.0801273105

Sabari, B. R., Dall'Agnese, A., Boija, A., Klein, I. A., Coffey, E. L., Shrinivas, K., et al. (2018). Coactivator condensation at super-enhancers links phase separation and gene control. *Science (New York, N.Y.)*, *361*(6400), eaar3958. http://doi.org/10.1126/science.aar3958

Sainsbury, S., Bernecky, C., & Cramer, P. (2015). Structural basis of transcription initiation by RNA polymerase II. *Nature Reviews. Molecular Cell Biology*, *16*(3), 129–143. http://doi.org/10.1038/nrm3952

Sanborn, A. L., Rao, S. S. P., Huang, S.-C., Durand, N. C., Huntley, M. H., Jewett, A. I., et al. (2015). Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes. *Proceedings of the National Academy of Sciences of the United States of America*, *112*(47), E6456–65. http://doi.org/10.1073/pnas.1518552112

Sanyal, A., Lajoie, B. R., Jain, G., & Dekker, J. (2012). The long-range interaction landscape of gene promoters. *Nature*, *489*(7414), 109–113. http://doi.org/10.1038/nature11279

Schmidt, D., Schwalie, P. C., Ross-Innes, C. S., Hurtado, A., Brown, G. D., Carroll, J. S., et al. (2010). A CTCF-independent role for cohesin in tissue-specific transcription. *Genome Research*, *20*(5), 578–588. http://doi.org/10.1101/gr.100479.109

Schmitt, A. D., Hu, M., Jung, I., Xu, Z., Qiu, Y., Tan, C. L., et al. (2016). A Compendium of Chromatin Contact Maps Reveals Spatially Active Regions in the Human Genome. *Cell Reports*, *17*(8), 2042–2059. http://doi.org/10.1016/j.celrep.2016.10.061

Schwarzer, W., Abdennur, N., Goloborodko, A., Pekowska, A., Fudenberg, G., Loe-Mie, Y., et al. (2017). Two independent modes of chromatin organization revealed by cohesin removal. *Nature*, *551*(7678), 51–56. http://doi.org/10.1038/nature24281

Schwessinger, R., Suciu, M. C., McGowan, S. J., Telenius, J., Taylor, S., Higgs, D. R., & Hughes, J. R. (2017). Sasquatch: predicting the impact of regulatory SNPs on transcription factor binding from cell- and tissue-specific DNase footprints. *Genome Research*, *27*(10), 1730–1742. http://doi.org/10.1101/gr.220202.117

Selvaraj, S., R Dixon, J., Bansal, V., & Ren, B. (2013). Whole-genome haplotype reconstruction using proximity-ligation and shotgun sequencing. *Nature Biotechnology*, *31*(12), 1111–1118. http://doi.org/10.1038/nbt.2728

Sexton, T., Yaffe, E., Kenigsberg, E., Bantignies, F., Leblanc, B., Hoichman, M., et al. (2012). Three-dimensional folding and functional organization principles of the Drosophila genome. *Cell*, *148*(3), 458–472. http://doi.org/10.1016/j.cell.2012.01.010

Shen, Y., Yue, F., McCleary, D. F., Ye, Z., Edsall, L., Kuan, S., et al. (2012). A map of the cis-regulatory sequences in the mouse genome. *Nature*, *488*(7409), 116–120. http://doi.org/10.1038/nature11243

Simon, C. S., Downes, D. J., Gosden, M. E., Telenius, J., Higgs, D. R., Hughes, J. R., et al. (2017). Functional characterisation of cis-regulatory elements governing dynamic Eomes expression in the early mouse embryo. *Development (Cambridge, England)*, *144*(7), 1249–1260. http://doi.org/10.1242/dev.147322

Simonis, M., Klous, P., Splinter, E., Moshkin, Y., Willemsen, R., de Wit, E., et al. (2006). Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4C). *Nature Genetics*, *38*(11), 1348–1354. http://doi.org/10.1038/ng1896

Smith, A. G. (1991). Culture and differentiation of embryonic stem cells. *Journal of Tissue Culture Methods*, *13*(2), 89–94.

Spivak, J. L., Toretti, D., & Dickerman, H. W. (1973). Effect of phenylhydrazine-induced hemolytic anemia on nuclear RNA polymerase activity of the mouse spleen. *Blood*, *42*(2), 257–266.

Stadler, M. B., Murr, R., Burger, L., Ivanek, R., Lienert, F., Schöler, A., et al. (2011). DNA-binding factors shape the mouse methylome at distal regulatory regions. *Nature*, *480*(7378), 490–495. http://doi.org/10.1038/nature10716

Stedman, W., Kang, H., Lin, S., Kissil, J. L., Bartolomei, M. S., & Lieberman, P. M. (2008). Cohesins localize with CTCF at the KSHV latency control region and at cellular c-myc and H19/Igf2 insulators. *The EMBO Journal*, *27*(4), 654–666. http://doi.org/10.1038/emboj.2008.1

Stevens, T. J., Lando, D., Basu, S., Atkinson, L. P., Cao, Y., Lee, S. F., et al. (2017). 3D structures of individual mammalian genomes studied by single-cell Hi-C. *Nature*, *589*(7648), 2931–64. http://doi.org/10.1038/nature21429

Symmons, O., Pan, L., Remeseiro, S., Aktas, T., Klein, F., Huber, W., & Spitz, F. (2016). The Shh Topological Domain Facilitates the Action of Remote Enhancers by Reducing the Effects of Genomic Distances. *Developmental Cell*, *39*(5), 529–543. http://doi.org/10.1016/j.devcel.2016.10.015

Symmons, O., Uslu, V. V., Tsujimura, T., Ruf, S., Nassari, S., Schwarzer, W., et al. (2014). Functional and topological characteristics of mammalian regulatory domains. *Genome Research*, *24*(3), 390–400. http://doi.org/10.1101/gr.163519.113

Szabo, Q., Jost, D., Chang, J.-M., Cattoni, D. I., Papadopoulos, G. L., Bonev, B., et al. (2018). TADs are 3D structural units of higher-order chromosome organization in Drosophila. *Science Advances*, *4*(2), eaar8082. http://doi.org/10.1126/sciadv.aar8082

Tallack, M. R., Whitington, T., Yuen, W. S., Wainwright, E. N., Keys, J. R., Gardiner, B. B., et al. (2010). A global role for KLF1 in erythropoiesis revealed by ChIP-seq in primary erythroid cells. *Genome Research*, *20*(8), 1052–1063. http://doi.org/10.1101/gr.106575.110

Telenius, J., & Hughes, J. R. (2018). NGseqBasic - a single-command UNIX tool for ATAC-seq, DNaseI-seq, Cut-and-Run, and ChIP-seq data mapping, high-resolution visualisation, and quality control. *bioRxiv*, 393413.

Therizols, P., Illingworth, R. S., Courilleau, C., Boyle, S., Wood, A. J., & Bickmore, W. A. (2014). Chromatin decondensation is sufficient to alter nuclear organization in embryonic stem cells. *Science (New York, N.Y.)*, *346*(6214), 1238–1242. http://doi.org/10.1126/science.1259587

Tolhuis, B., Palstra, R. J., Splinter, E., Grosveld, F., & de Laat, W. (2002). Looping and interaction between hypersensitive sites in the active beta-globin locus. *Molecular Cell*, *10*(6), 1453–1465. http://doi.org/10.1016/s1097-2765(02)00781-5

Ugarte, F., Sousae, R., Cinquin, B., Martin, E. W., Krietsch, J., Sanchez, G., et al. (2015). Progressive Chromatin Condensation and H3K9 Methylation Regulate the Differentiation of Embryonic and Hematopoietic Stem Cells. *Stem Cell Reports*, *5*(5), 728–740. http://doi.org/10.1016/j.stemcr.2015.09.009

van Arensbergen, J., FitzPatrick, V. D., de Haas, M., Pagie, L., Sluimer, J., Bussemaker, H. J., & van Steensel, B. (2017). Genome-wide mapping of autonomous promoter activity in human cells. *Nature Biotechnology*, *35*(2), 145–153. http://doi.org/10.1038/nbt.3754

van Steensel, B., & Belmont, A. S. (2017). Lamina-Associated Domains: Links with Chromosome Architecture, Heterochromatin, and Gene Repression. *Cell*, *169*(5), 780–791. http://doi.org/10.1016/j.cell.2017.04.022

Varshavsky, A. J., Sundin, O. H., & Bohn, M. J. (1978). SV40 viral minichromosome: preferential exposure of the origin of replication as probed by restriction endonucleases. *Nucleic Acids Research*, *5*(10), 3469–3477. http://doi.org/10.1093/nar/5.10.3469

Vermunt, M. W., Zhang, D., & Blobel, G. A. (2019). The interdependence of gene-regulatory elements and the 3D genome. *The Journal of Cell Biology*, *218*(1), 12–26. http://doi.org/10.1083/jcb.201809040

Vernimmen, D., De Gobbi, M., Sloane-Stanley, J. A., Wood, W. G., & Higgs, D. R. (2007). Long-range chromosomal interactions regulate the timing of the transition between poised and active gene expression. *The EMBO Journal*, *26*(8), 2041–2051. http://doi.org/10.1038/sj.emboj.7601654

Vestri, R., Pieragostini, E., & Ristaldi, M. S. (1994). Expression gradient in sheep alpha alpha and alpha alpha alpha globin gene haplotypes: mRNA levels. *Blood*, *83*(8), 2317–2322.

Vian, L., Pekowska, A., Rao, S. S. P., Kieffer-Kwon, K.-R., Jung, S., Baranello, L., et al. (2018). The Energetics and Physiological Impact of Cohesin Extrusion. *Cell*, *173*(5), 1165–1178.e20. http://doi.org/10.1016/j.cell.2018.03.072

Vietri Rudan, M., Barrington, C., Henderson, S., Ernst, C., Odom, D. T., Tanay, A., & Hadjur, S. (2015). Comparative Hi-C reveals that CTCF underlies evolution of chromosomal domain architecture. *Cell Reports*, *10*(8), 1297–1309. http://doi.org/10.1016/j.celrep.2015.02.004

Wallace, H. A. C., Marques-Kranc, F., Richardson, M., Luna-Crespo, F., Sharpe, J. A., Hughes, J., et al. (2007). Manipulating the mouse genome to engineer precise functional syntenic replacements with human sequence. *Cell*, *128*(1), 197–209. http://doi.org/10.1016/j.cell.2006.11.044

Wamstad, J. A., Alexander, J. M., Truty, R. M., Shrikumar, A., Li, F., Eilertson, K. E., et al. (2012). Dynamic and coordinated epigenetic regulation of developmental transitions in the cardiac lineage. *Cell*, *151*(1), 206–220. http://doi.org/10.1016/j.cell.2012.07.035

Wan, L.-B., Pan, H., Hannenhalli, S., Cheng, Y., Ma, J., Fedoriw, A., et al. (2008). Maternal depletion of CTCF reveals multiple functions during oocyte and preimplantation embryo development. *Development (Cambridge, England)*, *135*(16), 2729–2738. http://doi.org/10.1242/dev.024539

Wang, B., Li, K., Wang, A., Reiser, M., Saunders, T., Lockey, R. F., & Wang, J.-W. (2015). Highly efficient CRISPR/HDR-mediated knock-in for mouse embryonic stem cells and zygotes. *BioTechniques*, *59*(4), 201–2– 204– 206–8. http://doi.org/10.2144/000114339

Wang, H., La Russa, M., & Qi, L. S. (2016). CRISPR/Cas9 in Genome Editing and Beyond. *Annual Review of Biochemistry*, *85*(1), 227–264. http://doi.org/10.1146/annurev-biochem-060815-014607

Weintraub, A. S., Li, C. H., Zamudio, A. V., Sigova, A. A., Hannett, N. M., Day, D. S., et al. (2017). YY1 Is a Structural Regulator of Enhancer-Promoter Loops. *Cell*, *171*(7), 1573–1588.e28. http://doi.org/10.1016/j.cell.2017.11.008

Wendt, K. S., Yoshida, K., Itoh, T., Bando, M., Koch, B., Schirghuber, E., et al. (2008). Cohesin mediates transcriptional insulation by CCCTC-binding factor. *Nature*, *451*(7180), 796–801. http://doi.org/10.1038/nature06634

West, A. G., Gaszner, M., & Felsenfeld, G. (2002). Insulators: many functions, many mechanisms. *Genes & Development*, *16*(3), 271–288. http://doi.org/10.1101/gad.954702

Weth, O., Paprotka, C., Günther, K., Schulte, A., Baierl, M., Leers, J., et al. (2014). CTCF induces histone variant incorporation, erases the H3K27me3 histone mark and opens chromatin. *Nucleic Acids Research*, *42*(19), 11941–11951. http://doi.org/10.1093/nar/gku937

Whyte, W. A., Orlando, D. A., Hnisz, D., Abraham, B. J., Lin, C. Y., Kagey, M. H., et al. (2013). Master transcription factors and mediator establish super-enhancers at key cell identity genes. *Cell*, *153*(2), 307–319. http://doi.org/10.1016/j.cell.2013.03.035

Wijgerde, M., Grosveld, F., & Fraser, P. (1995). Transcription complex stability and chromatin dynamics in vivo. *Nature*, *377*(6546), 209–213. http://doi.org/10.1038/377209a0

Williams, R. R. E., Azuara, V., Perry, P., Sauer, S., Dvorkina, M., Jørgensen, H., et al. (2006). Neural induction promotes large-scale chromatin reorganisation of the Mash1 locus. *Journal of Cell Science*, *119*(Pt 1), 132–140. http://doi.org/10.1242/jcs.02727

Wutz, G., Várnai, C., Nagasaka, K., Cisneros, D. A., Stocsits, R. R., Tang, W., et al. (2017). Topologically associating domains and chromatin loops depend on cohesin and are regulated by CTCF, WAPL, and PDS5 proteins. *The EMBO Journal*, *36*(24), 3573–3599. http://doi.org/10.15252/embj.201798004

Yusufzai, T. M., Tagami, H., Nakatani, Y., & Felsenfeld, G. (2004). CTCF tethers an insulator to subnuclear sites, suggesting shared insulator mechanisms across species. *Molecular Cell*, *13*(2), 291–298. http://doi.org/10.1016/s1097-2765(04)00029-2

Zhang, Y., Liu, T., Meyer, C. A., Eeckhoute, J., Johnson, D. S., Bernstein, B. E., et al. (2008). Model-based analysis of ChIP-Seq (MACS). *Genome Biology*, *9*(9), R137–9. http://doi.org/10.1186/gb-2008-9-9-r137