The role of *cis*-acting elements in regulating 3D structure and gene expression

A thesis submitted for the degree of Doctor of Philosophy



Rosa J. Stolper

New College

Wellcome Doctoral Training Programme in Chromosome and Developmental Biology

Weatherall Institute of Molecular Medicine

University of Oxford

Hilary Term 2022

Abstract

Precise spatial-temporal gene expression is essential for normal development and differentiation, and, as such, must be tightly controlled. *Cis*-acting regulatory elements (enhancers, promoters and boundary elements) play a key role in gene regulation, either by enhancing transcription directly or by altering the 3D chromatin conformation to modify interactions between enhancers and promoters. Neither the contributions of individual elements to these processes, nor the complex interplay between them is fully understood. In this thesis, I have specifically investigated the ability of an ectopic CTCF binding site to shape the 3D genome and perturb gene expression. Furthermore, I have studied which aspects of gene activation and transcription depend on the enhancers using a model of the alpha-globin locus from which all enhancers have been removed.

The mouse alpha-globin locus provides a well-characterised model of mammalian gene expression. Five enhancer elements regulate the expression of the alpha-globin genes, which lie in a ~65 kb sub-TAD flanked by largely convergent CTCF binding sites. First, I have studied the effect of restricting the 3D interactions of the enhancers by placing an insulator element between the alpha-globin enhancers and genes. This single CTCF binding site significantly reduces enhancer-promoter interactions and strongly decreases the expression of the alpha-globin genes. I created two CTCF insertion models, with the site either oriented to predominantly block loop extrusion from the direction of the enhancers or from the direction of the genes. Gene expression is most strongly affected in the model that is predominantly blocking cohesin coming from the direction of the enhancers. In this orientation, there is a prominent accumulation of cohesin at the insertion site, which is absent in the other orientation.

I have also characterized a model in which all five alpha-globin enhancer elements have been deleted. This had a much stronger effect on gene expression than restricting enhancer interactions and sheds light on the role that the enhancers play in activating the locus: for example, the binding of transcription factors and co-factors was almost completely lost from the alpha-globin promoters. This all enhancer knockout model provides a “blank” slate that will be valuable resource for re-building the enhancers and establishing the precise role of each element and their combinations in the future.

Acknowledgments

First, I would like to thank my supervisors Doug Higgs and Rob Beagrie. Doug, thank you for giving me this opportunity and for your enthusiasm and kind encouragement throughout my time in the lab. Rob, thank you for our weekly meetings which helped me make sense of my rambling thoughts and for your practical support throughout this project. Also, thank you for occasionally reminding me that I was planning to do too much.

My time in Oxford started with a collaborative project between the Higgs lab in the WIMM and the Davies lab in the WCHG, conceived by Lars Hanssen. I want to thank him for introducing me to Oxford student life and to the world of CTCF. I would also like to thank Ben Davies and everyone in the Davies lab for their kind help when I undertook my first CRISPR/Cas9 editing experiments.

The Higgs lab is full of wonderful people who have greatly influenced my DPhil and thesis, and without whose support I would not have been able to do it. Ed, Joe, Susie, Felice, Siyu, Chris, Mohsin, Lucy, Danuta, Marieke, Martin, Jelena, Yavor, Jill, Caroline, Aude – thank you for your help with troubleshooting experiments, our interesting scientific and non-scientific discussion and for all the tea breaks that went on for way too long. We are also all indebted to Liz Rose for her patient help and support in organising any practicalities.

Without the supportive environment in the WIMM this project would not have been possible. I would particularly like to thank Jackie Sharpe, Jackie Sloane-Stanley, Sue Butler and Hannah Robinson of the Transgenic Facility for all their help with mouse work. I would also like to thank Kevin Clark for his help with FACS.

For helpful discussions and experimental support, I would like to thank Damien Downes, Mira Kassouf, Jim Hughes and James Davies. I would like to acknowledge my thesis committee, Liz Robertson and Tom Milne, for their advice. This work was generously funded by Wellcome for which I am grateful.

As the cliché goes, the most important part of any journey is the friends you make along the way, and nothing could be more true for my DPhil. To my lovely lab mates turned friends – Caz, Hele, Lea, Matt, Ron, Steph – thank you for your support during our time in the lab together and all the love and fun times outside of it.

I am glad that I have people in my life who do not know what a failed ChIP is, so they can remind me there is more to life than lab work. I want to thank my friends from the Netherlands – Judith, Sarah, Layla, Susette, Ghaada, Emma, Tessa, Esmee – for coming to visit me in Oxford, and for making time to meet up whenever I came to the Netherlands, even if I was often disorganised and only messaging you when I was already on the train there. To my parents and brother, thank you for all your love and support, and for your very genuine attempts to understand my research – I will try to explain it better again in the future (but I do not think this thesis will be helpful in that regard). Finally to Louis, thank you for being such an amazing human being, for supporting me in everything and for bringing me numerous cups of tea while writing this thesis. I feel so grateful that you are in my life.

Table of Contents

ABSTRACT **2**

ACKNOWLEDGMENTS **3**

TABLE OF CONTENTS **4**

LIST OF FIGURES **7**

LIST OF TABLES **8**

LIST OF ABBREVIATIONS **9**

CHAPTER 1: INTRODUCTION **10**

REGULATION OF TRANSCRIPTION 10
ENHANCER ELEMENTS 12
Mechanisms of enhancer activation *12*
3D GENOME ORGANISATION 16
LOOP EXTRUSION 19
CTCF BINDING SITES 22
THE ALPHA-GLOBIN LOCUS AS A MODEL FOR GENE EXPRESSION REGULATION 24
AIMS AND OUTLINE OF THIS THESIS 27

CHAPTER 2: MATERIALS AND METHODS **29**

CELL CULTURE 29
Foetal liver expansion *29*
Mouse ES cells *29*
Mouse ES cell differentiation *29*
Doxycycline induction of dCas9-p300 ..... *30*
ISOLATION OF ERYTHROID CELLS 30
Selection of CD71+ cells after embryoid body differentiation *30*
Selection of Ter119+ cells after foetal liver culture *30*
Selection of Ter119+ cells from phenylhydrazine-treated spleen *31*
E10.5 embryonic blood *31*
A note on replicates *31*
GENOME EDITING 32
Cloning of CTCF insertion HDR plasmid and associated gRNA plasmids *32*
Cloning for the CTCF reporter assay HDR plasmids *33*
Cloning of inducible dCas9-p300 plasmid *33*
Generation of CTCF insertion mouse ES cells *37*
Generation of CTCF insertion mouse models and animal procedure *38*
Generation of the CTCF reporter ES cells *38*
Generation of all enhancer KO and wild type hemizygous mouse ES cells *39*
Generation of dCas9-p300 expressing AEKO mouse ES cells *40*
GENE EXPRESSION ANALYSIS 41
RNA extraction *41*
qPCR analysis *42*

RNA sequencing ... *42*
NEXT-GENERATION SEQUENCING ASSAYS ... 43
ATAC-seq ... *43*
ChIP-seq ... *43*
Chipmentation ... *44*
Capture-C ... *45*
COMPUTATION ANALYSIS ... 46
Mapping and visualisation of ATAC-seq, ChIP-seq and ChIPmentation samples ... *46*
Analysis of Capture-C data ... *47*
Analysis of RNA-seq data ... *47*
Read count analysis and peak calling ... *47*
PREVIOUSLY PUBLISHED AND PUBLICLY AVAILABLE DATASETS USED IN THIS THESIS ... 48

CHAPTER 3: INSERTION OF A CTCF BINDING SITE REDUCES ENHANCER-PROMOTER INTERACTIONS IN AN ORIENTATION-DEPENDENT MANNER ... 49

INTRODUCTION ... 49
Previous characterisation of CTCF binding sites at the alpha-globin locus ... *49*
Not all CTCF binding sites are functional as an insulator ... *54*
RESULTS ... 55
Introducing a CTCF site between the alpha-globin genes and their enhancer region ... *56*
The newly inserted CTCF binding site recruits CTCF ... *58*
The new CTCF binding site changes 3D genome organisation ... *59*
Rad21 accumulates at the ectopic CTCF site in the Reverse model ... *64*
DISCUSSION ... 67
ACKNOWLEDGEMENTS ... 69

CHAPTER 4: INSERTION OF A SINGLE CTCF BINDING SITE IS SUFFICIENT TO DISRUPT ENHANCER FUNCTION ... 70

INTRODUCTION ... 70
RESULTS ... 71
Expression of alpha-globin is reduced ... *72*
Expression changes of non-erythroid genes at the alpha-globin locus ... *75*
The alpha-globin enhancers and promoters remain accessible ... *76*
Development of a reporter assay ... *77*
DISCUSSION ... 81
ACKNOWLEDGEMENTS ... 85

CHAPTER 5: STUDYING TRANSCRIPTIONAL ACTIVATION AT AN ENHANCERLESS ALPHA-GLOBIN LOCUS ... 86

INTRODUCTION ... 86
RESULTS ... 90
Characterisation of histone modifications in the AEKO model ... *92*
Recruitment of transcription factors in the AEKO model ... *94*
Using dCas9 to rescue alpha-globin expression in the AEKO model ... *97*
DISCUSSION ... 102
ACKNOWLEDGEMENTS ... 105

CHAPTER 6: DISCUSSION 106

A single CTCF binding site can reduce enhancer-promoter interactions 107
CTCF binding sites have different insulating potential 107
Studying aspects of enhancer regulation at the alpha-globin locus 109
Loading of cohesin and loop extrusion dynamics 112
Rhbdf1 expression 116
CONCLUSIONS AND FURTHER DIRECTIONS 117

REFERENCES 119

List of Figures

Figure 1.1: Potential mechanisms by which enhancers regulate transcription 14
Figure 1.2: Schematic overview of loop extrusion for domain formation and gene expression regulation 20
Figure 1.3: The mouse alpha-globin locus 26
Figure 3.1: A knockout study reveals the upstream CTCF binding sites function to restrict the alpha-globin enhancers 51
Figure 3.2: Characterisation of downstream CTCF binding sites 53
Figure 3.3: DNase I hypersensitivity of the HS-38 CTCF binding site 56
Figure 3.4: Strategy for the introduction of the CTCF binding site at the alpha-globin locus and subsequent genetic manipulations 58
Figure 3.5: CTCF ChIP-seq confirms recruitment of CTCF to the newly introduced binding site .. 59
Figure 3.6: Capture-C analysis using two enhancer elements as a viewpoint 60
Figure 3.7: Capture-C analysis using two upstream CTCF binding sites as a viewpoint 62
Figure 3.8: Capture-C analysis using two downstream CTCF sites as a viewpoint 63
Figure 3.9: Comparison between Capture-C interaction plots from Forward and Reverse model64
Figure 3.10: Rad21 ChIP-shows accumulation of cohesin at the ectopic site in the Reverse model 66
Figure 4.1: Expression of alpha-globin in various erythroid tissues in the Forward and Reverse model 73
Figure 4.2: Expression of zeta-globin in various erythroid tissues in the Forward and Reverse model 74
Figure 4.3: Expression of non-erythroid genes at the alpha-globin locus in the Forward and Reverse model 75
Figure 4.4: ATAC-seq tracks in the Forward and Reverse model 77
Figure 4.5: Strategy for high-throughput screening to assess CTCF sites' insulating potential .. 79
Figure 4.6: Additional CTCF binding sites analysed using the alpha-globin locus landing pad 84
Figure 5.1: Mouse ES cell lines used in this chapter 89
Figure 5.2: Differentially expressed genes in AEKO vs WT CD71+ EB-derived erythroid cells..... 91
Figure 5.3: ATAC-seq analysis of hemizygous WT and AEKO erythroid cells as well as mouse ES cells 92
Figure 5.4: Characterisation of histone modifications in the AEKO model 94
Figure 5.5: Characterisation of promoter-associated factors in the AEKO model 95
Figure 5.6: Rad21 ChIP-seq in the AEKO model 97
Figure 5.7: dCas9-p300 strategy to rescue alpha-globin expression in the AEKO model 100
Figure 6.1: CTCF, Rad21, NIPBL and WAPL ChIP-seq tracks 116

List of Tables

Table 2.1: Guide RNA sequences used for inserting the CTCF binding site 32
Table 2.2: Primers used for cloning HDR plasmid 32
Table 2.3: Synthetic fragment used in cloning the CTCF insertion HDR plasmid 33
Table 2.4: DNA fragments and primers used for the CTCF reporter assay cloning 33
Table 2.5: DNA sequences ordered from Thermo Fisher for cloning the dCas9-p300 plasmid 35
Table 2.6: primer sequences used for genotyping PCRs 38
Table 2.7: Guide RNAs used to create allele and R2 deletions 40
Table 2.8: Variants detected in the WT synthetic BAC compared to design (Francis 2019) 40
Table 2.9: Variants detected in the R2-only synthetic BAC compared to design (Francis 2019) 40
Table 2.10: Primers used to verify integration of the inducible dCas9-p300 plasmid 41
Table 2.11: Primers used for RT-qPCR expression analysis 42
Table 2.12: Antibodies used for ChIP 44
Table 2.13: Antibodies used for Chipmentation 45
Table 2.14: Primer used for the quantification of the digestion efficiency of 3C libraries 46
Table 2.15: Biotinylated oligos used as Capture-C probes 46

List of Abbreviations

3C	Chromosome conformation capture
3D	Three-dimensional
AEKO	All enhancer knockout model
ATAC-seq	Assay for transposase-accessible chromatin followed by sequencing
BAC	Bacterial artificial chromosome
bp	Base pair
BSA	Bovine serum albumin
Cas9	CRISPR-associated protein 9
ChIP-seq	Chromatin immunoprecipitation followed by sequencing
CRISPR	Clustered regularly interspaced short palindromic repeats
CTCF	CCCTC-binding factor
dCas9	Dead Cas9
DNA	Deoxyribonucleic acid
DNase I	Deoxyribonuclease I
E10.5/E12.5	Embryonic day 10.5/12.5
EB	Embryonic body
Epo	Erythropoietin
ES	Embryonic stem
FACS	Fluorescence-activated cell sorting
FBS	Foetal bovine serum
Flp	Flippase
Frt	FLP recombinase target
Fwd	CTCF insertion model – Forward orientation
GFP	Green fluorescent protein
HDR	Homology directed repair
HS	Hypersensitive site
Kb	Kilobase pair
KO	Knockout
LADs	Lamina-associated domain
MACS	Magnetic-activated cell sorting
Mb	Megabase pair
mRNA	Messenger RNA
PBS	Phosphate-buffered saline
PCR	Polymerase chain reaction
PIC	Pre-initiation complex
Pol II	RNA polymerase II
qPCR	Quantitative PCR
Rev	CTCF insertion model – Reverse orientation
RMGR	Recombinase-mediated genomic replacement
RPKM	Reads per kb per million mapped reads
SCF	Stem cell factor
SDS	Sodium dodecyl sulfate
sgRNA	Single guide RNA
TAD	Topologically-associated Domain
TBP	TATA-binding protein
TSS	Transcription start site
WT	Wild type
YFP	Yellow fluorescent protein
ZF	Zinc finger

Chapter 1: Introduction

One of the central questions of molecular biology is how a single cell can give rise to a whole organism with a wide range of cell types. As almost every cell contains the same DNA encoding ~20,000 protein-coding genes (Pertea and Salzberg 2010), differential gene expression is key to cell type identity. Better understanding of the regulation of gene expression is therefore crucial to understanding differentiation and development. Gene expression starts with transcription, and this forms a major regulatory stage.

Regulation of transcription

Protein-coding genes are expressed via transcription by RNA polymerase II (Pol II), which produces messenger RNA (mRNA). This is a tightly controlled process, which starts with the recruitment of Pol II and general transcription factors necessary for transcription. These assemble to form the pre-initiation complex (PIC) at the promoter (Roeder 1996; Thomas and Chiang 2006; Sainsbury et al. 2015)). Transcription then occurs in four recognised stages: initiation, pausing, elongation and termination (Venkatesh and Workman 2015).

All the stages of the transcription cycle as well as PIC recruitment can be regulated to affect the amount of transcript produced. Non-coding *cis*-acting DNA elements play an important role in this regulation, affecting gene expression on the same chromosome (cis: “on this side”). These elements, including promoters, enhancers and insulator elements, can lie close (bp) to the gene they regulate or far away (Mb) on the linear genome. Many of these elements function in a cell-type specific fashion. Generally, genomic elements that are functional in a specific cell type are free of nucleosomes, enabling their DNA to sequence to be recognised and bound by other factors. Therefore, open chromatin is a key characteristic of regulatory elements, and can be assayed by DNase I foot printing (Varshavsky et al. 1978; Parslow and Granner 1983; Boyle et al. 2008) or ATAC-seq (Buenrostro et al. 2013).

Proximal gene regulation occurs via promoters. The core promoter is located directly (~50 bp) adjacent to the transcription start site (TSS) and is sufficient for basal transcription. It functions as the site where the PIC is assembled (Smale and Kadonaga 2003; Juven-Gershon et al. 2008; Baumann et al. 2010). Core promoter sequences can contain a diverse set of binding motifs (Baumann et al. 2010). The first of such motifs to be discovered was the TATA-box, which is only found at a relatively small subset (3.5% - 28%) of core promoters (Breathnach and Chambon 1981; Suzuki et al. 2001; Vo Ngoc et al. 2017). It is recognised by TBP, a component of the PIC, and has a role in PIC positioning. Presence of a TATA-box also increases PIC stability at promoters (Krebs et al. 2017), although TBP can also bind to promoters without a TATA-box (Blair et al. 2012; Haberle et al. 2019). Another common feature of mammalian core promoters is CG-rich stretches of DNA, called CpG islands. CpG island promoters make up 71% of mouse promoters and are thought to lead to more disperse transcription initiation with multiple TSSs (Bajic et al. 2006; Baumann et al. 2010).

Sequences directly upstream of the core promoter but in the same open chromatin site can also function to modulate gene expression. These promoter-proximal elements are bound by transcription factors, which are often tissue-specific (Calhoun et al. 2002; Zabidi et al. 2015). Generally, the promoter refers to the combination of the core promoter as well as the proximal transcription factor binding sites (Andersson and Sandelin 2020). Promoter-proximal elements can play an essential role in enabling gene regulation by distal regulatory elements (Wen et al. 1993; Calhoun et al. 2002; Schoneveld et al. 2005) and promoter-proximal CTCF sites can function in the same way (Kubo et al. 2021). However, promoters can also drive expression independent from enhancer elements. For instance, the integration of a number of promoters in different genomic regions showed that the ability to drive transcription in both repressive and active chromatin environments is partly encoded in the promoter sequence (Leemans et al. 2019). Finally, some promoter sequences have intrinsic properties that drive high levels of expression at most genomic sites and most tissue types (Norrman et al. 2010).

Enhancer elements

In addition to gene-proximal elements, gene expression is also regulated by elements further away from the TSS. These enhancer elements regulate expression in a position and orientation-independent manner and can be far away from their target genes (> 1 Mb), which they regulate in *cis* (Banerji et al. 1981; Moreau et al. 1981; Lettice et al. 2003). Enhancers are an important feature of gene regulation; it has been estimated that there are tens to hundreds of thousands of enhancers in the human genome, many more than the total number of genes (~20,000) (Pennacchio et al. 2013; Andersson et al. 2014).

In essence, enhancer elements are regions of DNA that bind multiple transcription factors via transcription factor motifs present in their sequence (Spitz and Furlong 2012; Zabidi and Stark 2016). Typically, they are a few hundred base pairs in length, and they bind both cell type-specific and general transcription factors. In turn, these transcription factors recruit co-activators that regulate transcription from core promoters (Zabidi and Stark 2016). Importantly, many genes are regulated by several enhancer elements jointly. These elements can have different functions (Thomas et al. 2021), affect different cell populations during differentiation (Sacilotto et al. 2016), may be redundant (Frankel et al. 2010; Osterwalder et al. 2018), or act synergistically (Whyte et al. 2013). Clusters of enhancers that drive high levels of expression have been variously called locus control regions (Levings and Bungert 2002), stretch enhancers (Parker et al. 2013) or super-enhancers (Whyte et al. 2013), with slight variations of definition. Although the super-enhancer concept has proved useful in revealing which enhancers play an important role in cell fate, whether super-enhancers represent a separate class of elements or simply a cluster of enhancer elements remains debated (Blobel et al. 2021; Grosveld et al. 2021).

Mechanisms of enhancer activation

Many mechanisms have been proposed by which enhancers can activate transcription from a gene promoter (Beagrie and Pombo 2016; Zabidi and Stark 2016; Grosveld et al. 2021), and these can work on different points of the transcription cycle. First, it is possible that enhancers act

early in the transcription cycle to create a permissive environment for PIC formation at the promoters. Changing the local chromatin structure around a promoter can involve the creation of a nucleosome-free region by chromatin remodellers, the deposition of active histone modifications such as H3K4me3 and H3K27ac, as well as the inclusion of non-canonical histones, all of which can aid PIC formation and transcription initiation (Figure 1.1.a; (Chen et al. 2013; Venkatesh and Workman 2015)). Some co-activators, such as CBP and p300 (Brown et al. 2000; Hsia et al. 2010), interact with transcription factors at enhancers and are known to have chromatin remodelling and modifying activity. Support for this type of regulation also comes from studies that artificially target p300 to promoters, which has been shown to upregulate transcription (Cheng et al. 2013; Hilton et al. 2015).

Furthermore, enhancers could also create a permissive environment by removing repressive chromatin marks (Figure 1.1a). Many genes are repressed by Polycomb complexes and these could be evicted by enhancer activity (Beagrie and Pombo 2016). This potential mechanism was observed at the humanised alpha-globin locus, in which the mouse alpha-globin locus has been replaced by the human alpha-globin locus in the mouse. Polycomb is usually present at the human alpha-globin locus before gene activation, but disappears during differentiation (Garrick et al. 2008). When one of the main enhancers was deleted, however, H3K27me3 and Polycomb group proteins remained at the alpha-globin promoters even in erythroid cells (Vernimmen et al. 2011), showing that this enhancer normally plays a role in evicting Polycomb.

Enhancers could also play a role in the recruitment or stabilisation of the PIC (Figure 1.1b). Mediator is an important co-activator that interacts with a large number of transcription factors, is essential for the formation of the PIC (Allen and Taatjes 2015) and stabilises several general transcription factors that are part of the PIC (Conaway and Conaway 2011). Mediator directly interacts with Pol II and could therefore also play a role in recruiting Pol II to the promoter.

It is also possible that enhancer-associated transcription factors recruit co-activators with enzymatic activities which activate components of the transcription machinery via post-translational modifications (Figure 1.1c). Pol II phosphorylation, for instance, is essential for initiation and is dependent on Mediator (Boeing et al. 2010; Allen and Taatjes 2015; Robinson et al. 2016). After initiation, Pol II stalls close to the promoter and requires pause release factors to commence efficient elongation (Gariglio et al. 1981; Adelman and Lis 2012). This step has been proposed as one that can be regulated by distal enhancers (Lin et al. 2013; Liu et al. 2013).

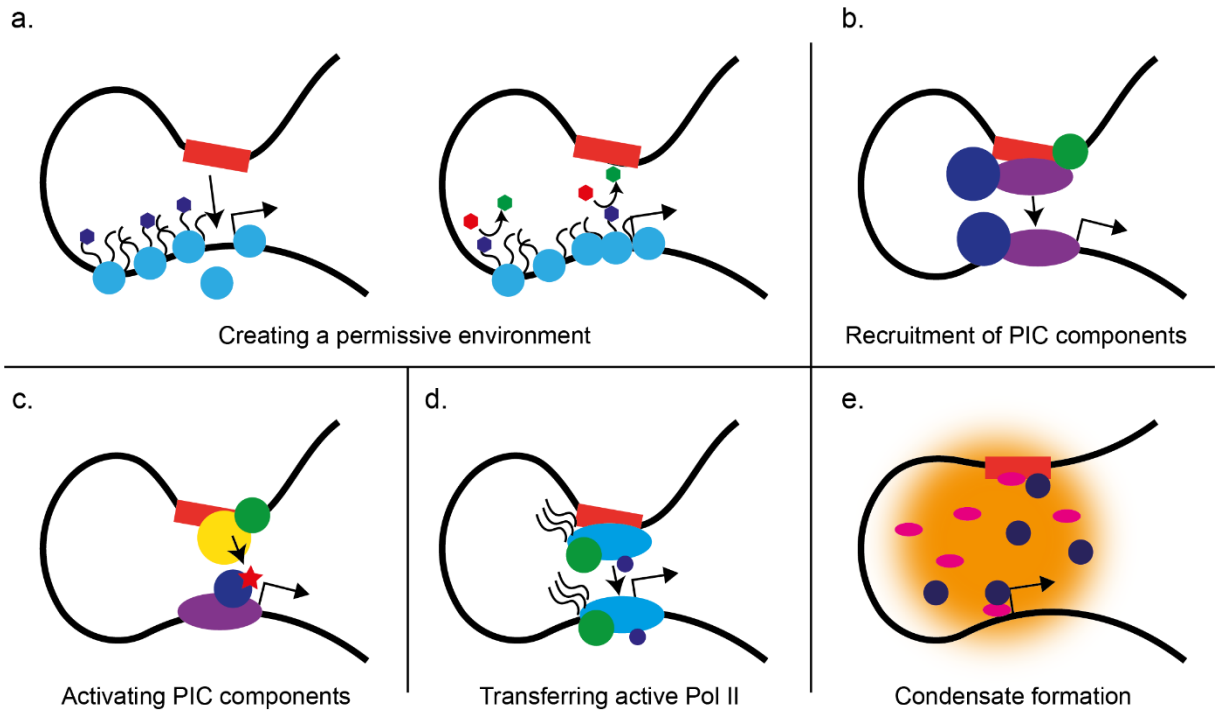


Figure 1.1: Potential mechanisms by which enhancers regulate transcription
a) Enhancers could create a permissive environment at the promoter by evicting nucleosomes and changing histone modifications. b) Enhancers could aid in the recruitment or stabilisation of PIC components at the enhancer. c) Enhancers could activate PIC components or release them to allow elongation. d) Transcription-competent RNA Pol II could transfer directly from the enhancers to the promoters. e) Enhancers could create biomolecular condensates with a high local concentration of transcription machinery.

Pol II and PIC components are present at enhancers as well as promoters, and in some cases their presence at enhancers precedes their presence at promoters in differentiation (Johnson et al. 2001; Vieira et al. 2004). Pol II at the enhancers is transcription competent and forms

bidirectional enhancer RNAs (De Santa et al. 2010; Kim et al. 2010). These enhancer RNAs could have a function in gene regulation (Li et al. 2013; Melo et al. 2013). Also, some have suggested that the transcription competent Pol II could be transferred directly onto the promoter, and this is another potential enhancer regulatory mechanism (Figure 1.1d; (Vieira et al. 2004)).

Finally, it is possible that the enhancers do not directly interact with a promoter or regulate steps of the transcription cycle, but change the nuclear environment in which the gene is located (Figure 1.1e). Certain nuclear compartments could provide an environment with a sufficiently high concentration of transcription machinery for efficient transcription to occur. Enhancers would then activate gene expression via forced relocation of the promoter to a site where active transcription occurs, such as transcription factories (Jackson et al. 1993; Wansink et al. 1993). Alternatively, it is possible that instead of relocating to specific nuclear compartment, enhancers create such an environment with a high concentration of transcription machinery, perhaps through condensate formation or phase separation (Hnisz et al. 2017).

Most enhancers bind many different transcription factors and co-activators, and it is therefore likely that enhancers combine several mechanisms to affect gene expression at their target genes, especially as these mechanisms are not mutually exclusive (Beagrie and Pombo 2016; Grosveld et al. 2021). It has also been observed that enhancers which bind a variety of transcription factors are better able to increase expression of their target genes than those with multiple binding sites for the same transcription factor (Smith et al. 2013). Furthermore, in many cases genes are regulated by several enhancer elements, and it is possible that these affect gene expression via different mechanisms (Thomas et al. 2021).

It has long been observed that enhancers do not increase the expression of all genes around them, but rather regulate the expression of specific genes (Li and Noll 1994; Merli et al. 1996). The specificity of certain enhancers to regulate specific genes is at least partly driven by intrinsic compatibilities (Zabidi et al. 2015; Martinez-Ara et al. 2021). Communication between

enhancers and promoters has long been thought to be mediated by physical proximity (Spitz and Furlong 2012; Pennacchio et al. 2013). Physical proximity could mean direct interactions between proteins bound at the promoter and enhancer, and there is evidence for such protein-mediated interactions (Hua et al. 2021). However, it could also entail more generally that the enhancer and the promoter are in a similar nuclear compartment, such as is postulated by gene regulation via biomolecular condensate formation (Figure 1.1e). Recently the requirement for close proximity between enhancers and promoters has been questioned (Alexander et al. 2019; Benabdallah et al. 2019). However, the 3D conformation of the genome at least plays a role in restricting enhancer-promoter interactions, as well as potentially directing them. Therefore, the three-dimensional organisation of the genome provides another important layer of genome regulation in addition to enhancers and promoters.

Finally, it is important to note that enhancers and promoters might not be distinct classes of elements per se, as they share many molecular characteristics (Andersson and Sandelin 2020). As mentioned earlier, enhancers are capable of initiating local transcription in the form of enhancer RNAs (De Santa et al. 2010; Kim et al. 2010; Andersson et al. 2014; Van Arensbergen et al. 2017) and promoters have been observed to act as enhancers (Dao et al. 2017; Diao et al. 2017). Certain histone modifications have been traditionally associated with either enhancers or promoters: H3K4me3 primarily marks promoters and its levels correlate with transcriptional output (Core et al. 2014), while H3K27ac and H3K4me1 are primarily associated with enhancer regions (Creyghton et al. 2010), although these are found at promoters as well. The most fundamental distinction is the stability of the post-initiation transcript, with promoters producing a long, stable transcript in one direction (Core et al. 2014).

3D genome organisation

Chromatin within the nuclei of interphase cells appears to have a distinct and highly organised 3D conformation. Proximity ligation assays, often called chromatin conformation capture or 3C-based assays, have been instrumental in understanding the 3D conformation of the genome (De

Wit and De Laat 2012; Davies et al. 2017). These methods focus mainly, but not exclusively (Ramani et al. 2017), on the 3D conformation in a population of cells, and most observations therefore reveal population averages. These 3C methods rely on crosslinking of chromatin in the nucleus, followed by digestion and a ligation step in which DNA fragments that are crosslinked together (and thus close in the 3D genome) are more likely to be ligated (Cullen et al. 1993; Dekker et al. 2002). These ligation junctions function as a read out of how often two fragments are found in close proximity. Initially the ligation junctions were quantified only between two fragments using qPCR (Dekker et al. 2002), but subsequent methods identify all the interactions of a single DNA fragment of interest (the so called viewpoint) using sequencing. These methods include 4C (Simonis et al. 2006; Zhao et al. 2006) and Capture-C (Hughes et al. 2014; Davies et al. 2016). Genome-wide analysis of all chromatin interactions became possible with the development of Hi-C (Lieberman-Aiden et al. 2009). Orthogonal methods have been in good agreement with 3C-based methods while omitting essential steps from the 3C protocol, such as crosslinking (Brant et al. 2016), ligation (Beagrie et al. 2017; Quinodoz et al. 2018) or both crosslinking and ligation (Redolfi et al. 2019). Most recently, 3C-based methods have been improved by using micrococcal nuclease for the digestion step (Hsieh et al. 2015; Hua et al. 2021). Finally, many of the observations have also been corroborated by imaging studies (Nora et al. 2012; Bintu et al. 2018).

Hi-C data revealed that chromatin spatially segregates to form compartments (Lieberman-Aiden et al. 2009). On a chromosomal level, these compartments associate with other compartments with similar chromatin marks, forming a plaid pattern on Hi-C matrices. Compartments can be classified as regions of active chromatin, called the A compartment, and of repressive chromatin, called the B compartment. The A compartment is mainly composed of euchromatin, which is more loosely packed, gene dense, and contains actively transcribed genes and active histone modifications (Gilbert et al. 2004). The B compartment contains more densely packed heterochromatin and histone modification associated with gene repression. It also strongly

overlaps with lamina-associated domains (LADs), regions of the genome that are associated with the nuclear lamina in which most genes are only very lowly expressed (Van Steensel and Belmont 2017).

Another layer of 3D chromatin conformation comes from topologically associated domains (TADs), originally defined as regions of genome with increased interactions within the domain rather than with sequences outside of the domain (Dixon et al. 2012; Nora et al. 2012). This is similar to the definition of compartments, and terminology has been inconsistent and not always very distinct (Mirny et al. 2019; Beagan and Phillips-Cremins 2020; Oudelaar and Higgs 2021). However, several studies have shown that disrupting TAD structure does not affect compartmentalisation into A and B domains. It has therefore been proposed that at least two layers 3D chromatin organisation exists, TADs and compartments, and they are distinguished by the mechanisms that drive their formation (Schwarzer et al. 2017; Mirny et al. 2019).

Compartments are characterised by the clustering of chromatin with similar structure. It is thought that the formation of biomolecular condensates or phase separation could drive interactions between chromatin domains with similar epigenetic states, leading to A and B compartments (Beagan and Phillips-Cremins 2020; Oudelaar and Higgs 2021). The heterochromatin protein 1 could play a role in the condensation of heterochromatin in a similar manner (Hildebrand and Dekker 2020).

Structurally, one important difference is that TADs do not tend to associate with other TADs over longer distances in the way compartments do (Mirny et al. 2019). Moreover, boundaries flanking TADs are often in close proximity to one another across the TAD, leading to a characteristic dot on Hi-C matrices. These are thought to represent persistent interactions present in many cells and are often thought of as a loop domain (Beagan and Phillips-Cremins 2020). CTCF binding sites are an important component of TAD boundaries, and sites of active transcription are enriched at TAD boundaries as well (Dixon et al. 2012; Ulianov et al. 2016).

Other genomic elements, such as lncRNAs (Barutcu et al. 2018), have also been proposed as contributing to TAD boundaries. The disruption of TAD boundaries often leads to misexpression, revealing the functional role of TAD boundaries in restriction enhancer interactions (Nora et al. 2012; Lupianez et al. 2015; Tsujimura et al. 2015). Many TADs are conserved between different cell types (Barrington et al. 2019).

Within TADs lie smaller interaction domains which have been called sub-TADs. These are also often bound by CTCF at their boundaries, but are more likely to be cell type specific and their boundaries are weaker (Phillips-Cremins et al. 2013; Beagan and Phillips-Cremins 2020).

Loop extrusion

Loop extrusion has been proposed as the mechanism that drives TAD formation (Sanborn et al. 2015; Fudenberg et al. 2016). The central proposition of loop extrusion is that cohesin, or another loop extrusion factor, slides along the chromatin fibre, thereby extruding the fibre in a loop that continuously increases in size (Figure 1.2a). This process continues until the loop extrusion complex reaches a boundary element which blocks further translocation, or the loop extrusion complex dissociates from the chromatin fibre. It is likely that loop extrusion is a highly dynamic process in which loops are constantly formed and falling apart.

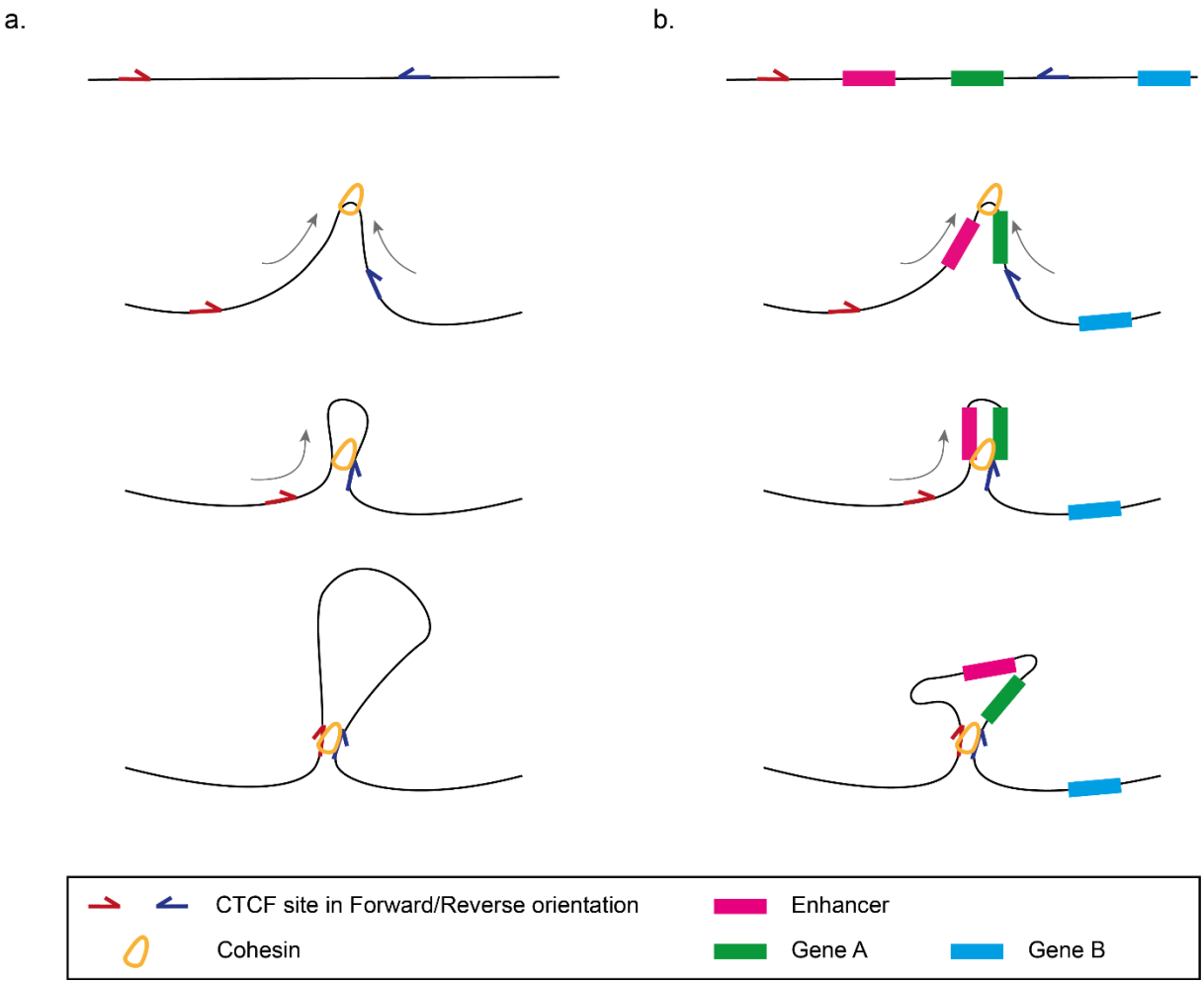


Figure 1.2: Schematic overview of loop extrusion for domain formation and gene expression regulation

a) Loop extrusion has been proposed as a dynamic process important for the establishment of TADs. One or multiple cohesin complexes associate topologically or non-topologically with the chromatin fibre and extrude a loop that continues to increase in size until the loop extrusion complex reaches a boundary element, usually CTCF. b) In addition to TAD formation, loop extrusion could also contribute to the establishment of enhancer-promoter interactions. The enhancer and gene A are brought into close proximity while loop extrusion occurs. However, gene B is insulated from the enhancer function by a CTCF site, which blocks loop extrusion.

The multiprotein complex cohesin has generally been proposed as the main loop extruding factor in interphase chromosomes (Sanborn et al. 2015; Fudenberg et al. 2016). The core cohesin complex consists of two Smc proteins, Smc1 and Smc3, as well as Rad21 (also known as Scc1 or Mcd1) and Scc3 (also known as SA1 or SA2). Smc1, Smc3 and Rad21 form a ring that mediates sister chromatid cohesion during mitosis and is physically capable of enclosing a DNA strand (Nasmyth and Haering 2009). In addition to its role in sister chromatid cohesion it has long been

known that cohesin plays a role in gene regulation (Parelho et al. 2008), and this can now be understood in the context of loop extrusion.

Cohesin has been observed to translocate along the DNA fibre and extrude loops *in vitro* (Kim et al. 2019), as has condensin (Ganji et al. 2018). Extrusion by cohesin depends *in vitro* on the presence of NIPBL-Mau2 and ATP, and occurs symmetrically while cohesin travels along the DNA at the base of the loop. These experiments also showed that extrusion likely occurs via a single cohesin complex and that this complex might not be topologically associated with the DNA, rather than enclosing the DNA fibre (Kim et al. 2019). Depletion of cohesin causes a loss of TADs and chromatin decompaction. However, A and B compartments are not affected by cohesin depletion and might even become more pronounced (Rao et al. 2017; Wutz et al. 2017; Bintu et al. 2018). The effect of removing Cohesin on gene expression is unexpectedly mild, with affected genes often located in the vicinity of super-enhancers (Rao et al. 2017).

Cohesin is loaded onto chromatin by NIPBL (Nasmyth and Haering 2009). NIPBL is present at active elements, marking these as potential loading sites for cohesin (Barrington et al. 2019), although it is not clear if it is exclusively loaded there. Stripe domains are a feature of TADs arising in Hi-C data, in which one boundary seems to interact with the entire TAD (Vian et al. 2018; Barrington et al. 2019). This might be caused by specific loading of cohesin close to the boundary forming the stripe, as this could force unidirectional loop extrusion. Interestingly, NIPBL as well as active enhancers and active histone modifications were often seen close to the stripe border. Deletion of NIPBL leads to changes in the chromatin that are very similar to cohesin depletion: a loss of TAD structures while the A and B compartments are preserved or even reinforced (Schwarzer et al. 2017). The NIPBL deletion caused a change in expression in ~1000 genes and a loss of active histone marks at promoters, but not at enhancers. Interestingly, there was an increase in exogenic expression, indicating that regulation by enhancers is

misdirected. It is important to note that this study looked at the conditional deletion of NIPBL using a timepoint 10 days after Cre-mediated deletion was induced (Schwarzer et al. 2017).

Cohesin is released from DNA by WAPL, and depletion of WAPL leads to the clustering of cohesin into axial structures, termed "vermicelli" DNA (Tedeschi et al. 2013). Furthermore, WAPL depletion leads to larger TADs and weaker compartments (Haarhuis et al. 2017; Wutz et al. 2017). This further indicates that compartments are formed in a loop extrusion independent manner, and that loop extrusion might compete with the mechanism for compartment formation in shaping the 3D genome. In the absence of WAPL, interactions between non-convergent CTCF sites increased, indicating that CTCF binding sites are more easily skipped in the absence of WAPL. This effect is even stronger when the PDS5 proteins, WAPL's binding partners, are also depleted (Wutz et al. 2017). When CTCF and WAPL are both depleted, cohesin accumulates at the 3' end of actively transcribed genes (Busslinger et al. 2017).

Loop extrusion could also explain how enhancer-promoter interactions are regulated. During loop extrusion, enhancers and promoters might be brought into close proximity by cohesin which could lead to protein-protein interactions between factors bound at either element (Figure 1.2b). Interactions could be prevented if there is a boundary element between an enhancer and a gene. This mechanism fits well with the observed effect of boundary elements; however, it is not consistent with the generally mild effects on gene expression seen when loop extrusion components are depleted.

CTCF binding sites

The final component of loop extrusion is CTCF, which is the main barrier to loop extrusion as exemplified by its large number of binding sites at TAD and sub-TAD boundaries (Dixon et al. 2012). CTCF is widely expressed and its function as an insulator was first discovered in reporter assays, where it prevents interactions between enhancers and promoters (Chung et al. 1997; Bell et al. 1999). Often, CTCF binding sites are found to be in close physical proximity across

TADs, as they are drawn together in the process of loop extrusion (Figure 1.2). Many CTCF binding sites are also bound by cohesin (Parelho et al. 2008), which matches its function in stalling the protein complex.

CTCF has 11 zinc fingers with which it can bind to a complex, non-palindromic motif. This high number of zinc fingers leads to a wide variety of binding site sequences, with different binding site strengths associated with different elements within the motif. About 80% of CTCF binding sites contain the 20-bp core motif which is bound by zinc fingers (ZF) 4-7. A subset of binding sites has a 10 bp upstream motif (about 13%), which increases CTCF binding stability. About 8% of binding sites have a downstream motif, often associated with both the core and upstream motif (5%). The role of the downstream motif is less clear, and sites with the downstream motif have lower levels of CTCF binding compared to the core motif alone (Nakahashi et al. 2013).

Interestingly, the vast majority CTCF sites that interact across TADs are in convergent orientation (Rao et al. 2014), indicating that CTCF can block loop extrusion from a specific direction. Several studies have also shown that when CTCF sites are inverted they lose the interactions with interacting CTCF sites (De Wit et al. 2015; Guo et al. 2015; Sanborn et al. 2015). An N-terminal segment of CTCF forms a specific interaction with the Rad21-Scc3 subunits of cohesin, and thereby stabilises it on the chromatin (Li et al. 2020c; Pugacheva et al. 2020). Furthermore, the crystal structure shows that WAPL binds to the same groove as CTCF. Therefore, it seems likely that CTCF binding prevents cohesin from being released by WAPL (Li et al. 2020c), in that manner also stabilising TAD structure.

Depletion of CTCF leads to a loss of TAD structure, but has only a mild effect on gene expression (Nora et al. 2017; Wutz et al. 2017). Interestingly, CTCF is also necessary for the expression of certain genes, confirming it can both block and establish enhancer-promoter interactions (Beagan et al. 2017; Nora et al. 2017). However, as with cohesin depletion, the A/B compartments remain unchanged, highlighting that these are formed by a mechanism other

than loop extrusion (Nora et al. 2017; Wutz et al. 2017). Most CTCF depletion studies have reported several persistent CTCF binding sites that remain bound after degradation of cytoplasmic CTCF (Nora et al. 2017; Khoury et al. 2020). In photobleaching experiments, CTCF has a considerably longer recovery time (~11 minutes) than most transcription factors, indicating a longer residency time on DNA (Nakahashi et al. 2013). This could mean that persistent sites in depletion experiments are sites where CTCF dissociates from the DNA infrequently, and the observation that there are differences between binding sites could reflect binding site strength.

In addition to CTCF, sites of active transcription might function as boundaries to loop extrusion (Dixon et al. 2012; Fudenberg et al. 2016; Ulianov et al. 2016; Harrold et al. 2020). Furthermore, chromatin interactions could also be established or maintained independently of cohesin and CTCF, for instance by proteins such as YY1 and Mediator (Beagan et al. 2017)

The alpha-globin locus as a model for gene expression regulation

Genome-wide studies have improved our understanding of gene regulation and genome organisation considerably, but studies focussing on a single locus also play an important part in unravelling these processes (De Laat and Duboule 2013). The mouse alpha-globin locus is very well characterised and has been used extensively to study genome regulation and organisation (Oudelaar et al. 2021). It is especially well suited for studies affecting gene expression levels, as the process of erythropoiesis is largely unaffected by changes in alpha-globin expression. This ensures that the cell lineage, differentiation, and maturation remain largely unchanged even in genetic models that have a substantial effect on alpha-globin expression (Hay et al. 2016; Oudelaar et al. 2021).

The mouse alpha-globin locus lies in a 65 kb sub-TAD which is part of a 165 kb TAD that is present in most cell types (Figure 1.3a; (Oudelaar et al. 2020)). There are five globin genes at the alpha-globin locus: two, almost identical, adult alpha-globin genes (*Hba-a1* and *Hba-a2*), the

embryonic zeta-globin (*Hba-x*), and two theta-globin genes (*Hbq-1* and *Hbq-2*) of unknown function (Figure 1.3b). Zeta-globin is repressed in definitive erythroid cells, but expressed in primitive tissue, where it is marked as an open chromatin region in ATAC-seq.

The expression of these genes in mouse is regulated by five enhancer elements (R1, R2, R3, Rm and R4), four of which are located within introns of the constitutively expressed *Nprl3* gene while R4 lies just downstream of the *Nprl3* promoter. Together these elements qualify as a super-enhancer (Hay et al. 2016). The strongest enhancer elements are R1 and R2, while Rm is a mouse-specific element. The enhancer elements are bound by various erythroid-specific transcription factors, such as Gata1, Klf-1, and Nf-e2 (Figure 1.3c). The alpha-globin sub-TAD is flanked by largely convergently orientated CTCF binding sites (Figure 1.3d), which bind CTCF in a tissue-invariant manner (Oudelaar et al. 2021).

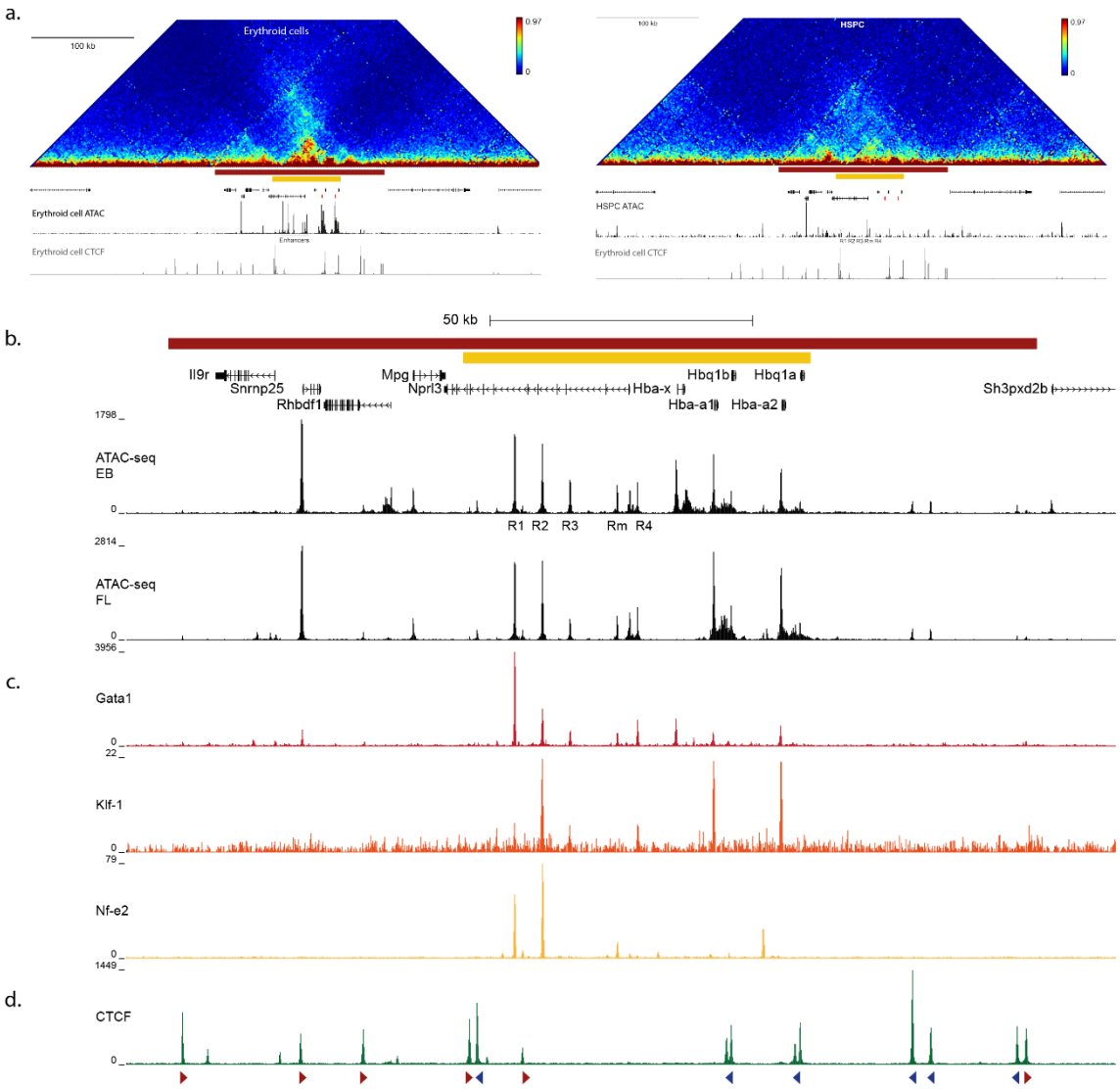


Figure 1.3: The mouse alpha-globin locus

a) Tiled-C contact matrices of 500 kb spanning the mouse alpha-globin locus, in erythroid cells (left) and haematopoietic stem and progenitor cells (HSPCs; right), showing the larger 165 kb TAD (dark red) and the 65 kb erythroid-specific sub-TAD (yellow). Data previously published (Oudelaar et al. 2020). b) Functional elements are marked by peaks in the ATAC-seq profile, and are very similar in primitive-like erythroid cells (embryoid body (EB); previously published in Francis et al. (Francis et al. 2022)), and definitive erythroid cells (foetal liver (FL)). Enhancer elements are marked under the EB ATAC-seq track. c) RPKM-normalised ChIP-seq tracks for erythroid transcription factor: Gata1 (n=3, EB), Klf-1 (spleen, unpublished), Nf-e2 (n=2, EB). d) RPKM-normalised CTCF ChIP-seq tracks (n=2) in EB-derived erythroid cells. CTCF site orientation is marked under each peak.

The activation of the alpha-globin locus, as well as the beta-globin locus, during differentiation along the erythroid lineage has been extensively characterised, and these erythroid loci have long been used as model loci for gene expression (Bulger et al. 2002; Oudelaar et al. 2021). Erythropoiesis is well suited to studying gene regulation, not only because manipulation of

globin gene expression does not interfere with differentiation, but also because erythroid cells can be easily obtained and purified in large numbers. Primary erythroid cells can be obtained from the mouse spleen following phenylhydrazine treatment, which expands the proportion of erythroblasts in the spleen to over 70% (Spivak et al. 1973). In this thesis, Ter119+ cells were purified from spleen tissue to obtain a pure population of definitive erythroid cells. Foetal livers are another source of primary definitive erythroid cells, which can be expanded in culture and purified in the same way as spleen-derived erythroblasts to supply a similar population of cells. Certain mutations that severely impair alpha-globin expression lead to embryonic lethality, however embryos generally survive long enough to assess erythroid cells from embryonic day 12.5 (E12.5) foetal livers (Hay et al. 2016).

Obtaining erythroid cells from mice has the benefit of providing large numbers of cells (in the order of 10^8-10^9), however, the time and costs associated with testing many different genetic models are limiting. Alternatively, mouse ES cells are easily manipulated and differentiated *in vitro* into embryoid bodies (EBs) that provide a mixed population of cells from which erythroid cells can be isolated. In contrast to the Ter119+ definitive erythroid cells from spleen and foetal liver, CD71+ erythroblasts isolated from EBs are primitive-like erythroid cells (Francis et al. 2022). These express adult alpha-globin as the definitive erythroblasts do, but in addition also express embryonic zeta-globin (Figure 1.3a). Primary primitive erythroid cells can be isolated in small numbers from E10.5 embryonic blood.

Aims and outline of this thesis

Understanding the individual contribution of *cis*-regulatory elements to gene expression regulation, as well as their interplay, is an ongoing process in molecular biology. Much has been accomplished in characterising the role of enhancers, promoters and insulators at the alpha-globin locus, however recent advances in genome editing have allowed us to isolate specific elements and investigate their contribution. In this thesis, I first investigate the most effective CTCF binding site at the locus. By placing it at a different site within the same locus I find that

this single CTCF site is sufficient to change the 3D chromatin interactions of the alpha-globin locus. The inserted sequence disrupts the enhancer-promoter interactions and leads to substantially decreased alpha-globin expression. Further developing this model, I have tested other CTCF sites and established a general model for assaying the functional role of any chosen CTCF binding site. Finally, I characterised a model of the alpha-globin locus in which all enhancer elements have been deleted, to determine the extent to which the enhancers affect the structure and function of the locus.

Chapter 2: Materials and Methods

Cell culture

Foetal liver expansion

The *in vitro* mouse foetal liver culturing system is based on previous protocols (von Lindern et al. 2001; Chen et al. 2009). Briefly, E12.5 mouse foetal liver were dissected and cultured in StemPro medium (Invitrogen) supplemented with Epo (1 U/mL; Janssen, PL 00242/029), SCF (50 ng/mL; Peprotech, 250-03), dexamethasone (1 µM; Hameln, DEXA3.3) and 1× L-Glutamine (Invitrogen) for 6–8 days to expand the erythroid progenitor population. During culture the erythroblast cell count was maintained at 1 x 10^6 cells/ml or under. If necessary cells were frozen at day 4 and after thawing expanded for a further 2-4 days.

Mouse ES cells

E14Tg2a mouse ES cells (Nichols et al. 1990; Smith 1991) were maintained in KnockOut™ DMEM or Glasgow's Minimal Essential Medium supplemented with 10% FBS, 1 mM sodium pyruvate, 2 mM L-Glutamine, 1x non-essential amino acids (all Gibco), beta-mercaptoethanol, and leukemia inhibiting factors. During gene targeting, the media was supplemented with penicillin and streptomycin (Gibco). Cells were usually grown on 0.1% (v/v) gelatin (Sigma-Aldrich: G1393), except during gene targeting for the CTCF insertion mouse model and prior to blastocyst injection, when they were grown on mitotically inactivated mouse embryonic fibroblasts (MEFs).

Mouse ES cell differentiation

An in vitro haematopoietic differentiation protocol (Francis et al. 2022) was used to differentiate mouse ES cells into embryoid bodies (EBs), based on a previously published protocol (Keller 1995). In brief, proliferating ES cells were cultured in IMDM-ES media (Iscove's Modified Dulbecco Medium (IMDM) (Gibco) supplemented with 15% FBS, 100 U/ml penicillin/streptomycin, 1000 U/ml LIF, and 1.5 x 10^{-4} M monothioglycerol (MTG; Sigma-Aldrich: M6145)) for up to 48 hours prior to primary plating. Subsequently, cells were plated in 10 ml

differentiation media on bacterial petri dishes (15000-50000 cells/plate) and cultured for 7 days. Differentiation medium consisted of IMDM (Gibco) supplemented with 15% heat-inactivated FBS (Gibco, heat inactivated for 1 hour at 56 °C), 2 mM L-Glutamine (Gibco), 300 µg/ml transferrin (Roche), 50 µg/ml ascorbic acid (Sigma-Aldrich), 3 x 10^{-4} M MTG and 5% Protein-Free Hybridoma Media II (Gibco).

Doxycycline induction of dCas9-p300

Doxycycline (Abcam, ab141091) was added to cell culture media at concentrations ranging from 0.1-10 µg/ml for 24-48 hours, either to undifferentiated ES cells or EBs.

Isolation of erythroid cells

Selection of CD71+ cells after embryoid body differentiation

After 7 days of differentiation, EBs were disaggregated using 0.25% trypsin (Gibco) and stained with 0.5 µg FITC Rat anti-mouse CD71 antibody (BD Pharmingen or eBioscience) per 10^6 cells for 20 minutes at 4 °C, followed washing and by incubation with anti-FITC MACS microbeads (Miltenyi Biotec) for 15 minutes at 4 °C. Stained cells were then purified using MACS lineage selection columns (Miltenyi Biotec) and immediately processed for various assays. Purity of isolated erythroid population was verified by fluorescent-activated cell sorting (FACS; Attune Nxt Flow Cytometer (Thermo Fisher)).

Selection of Ter119+ cells after foetal liver culture

After 6 to 8 days of expansion, cells were stained with 1 µg anti-mouse Ter119-PE antibody (BD Pharmingen) per 10^6 cells for 20 minutes at 4 °C. Cells were washed with PBS + 10% FCS and resuspended in 135 µl cold PBE buffer (PBS with 0.5%BSA and 2 mM EDTA) with 15 µl Ter119-PE microbeads (Miltenyi Biotec) per 10^7 cells, and incubated for 15 minutes at 4 °C. Erythroid cells were then isolated using MACS lineage selection columns (Miltenyi Biotec) and processed for various assays. Purity of isolated populations was assessed as described above.

Selection of Ter119+ cells from phenylhydrazine-treated spleen

Primary erythroid cells were obtained from the spleen of adult mice treated with phenylhydrazine, which results in an expansion of erythroid cells in the spleen (Spivak et al. 1973). Mice were treated with phenylhydrazine by Jacqueline Sharpe or Jacqueline Sloane-Stanley (WIMM). Treatment consisted of 3 doses (40 mg/g body weight) given 12 hours apart, and mice were sacrificed 5 days after the first dose. Spleens were mechanically disrupted in cold PBS + 10% FCS to produce a single cell suspension. Cells were resuspended in 1 ml cold PBS + 10% FCS and stained with 6 µg anti-mouse Ter119-PE antibody (Miltenyi Biotec) per 10^8 cells for 20 minutes at 4 °C. After washing with PBS + 10% FCS, cells were resuspended in 8 µl cold PBE buffer (PBS with 0.5%BSA and 2 mM EDTA) and 2 µl anti-PE microbeads (Miltenyi Biotec) and incubated for 15 minutes at 4 °C. Ter119+ cells were selected with MACS lineage specification columns (Miltenyi Biotec), typically using one column for up to 5 x 10^8 cells, and processed for various assays. Purity of isolated populations was assessed as for foetal liver and EB-derived cells.

E10.5 embryonic blood

Pregnant mice were dissected on embryonic day 10.5 (E10.5) after observation of vaginal plug. Embryos were individually collected and erythroid cells were used directly without selection.

A note on replicates

Replicates for foetal liver, spleen and E10.5 erythroid cells are from separate animals. For EB-derived cells, three independently-edited clones were used for the Forward, Reverse and AEKO model. Any experiments with more than three replicates consist of the three independent clones over multiple EB differentiations. For wild-type EB samples, replicates consist of independent EB differentiations.

Genome editing

Cloning of CTCF insertion HDR plasmid and associated gRNA plasmids

Mouse ES cells were modified using double nickase D10A Cas9 pairs (Cong et al. 2013). Two plasmids encoding D10A Cas9 protein and either sgRNA were generated by cloning the targeting oligonucleotides (see Table 2.1) into pX335-U6-Chimeric_BB-CBh-hSpCas9n(D10) (Addgene plasmid no. 42335, pX335). pX335 had been modified to contain a puromycin selection cassette. The homology directed repair (HDR) plasmid was prepared by amplifying 1.4 and 1.2 kb homology arms from genomic DNA (primers Table 2.2) using extension-overlap PCR to merge the two fragments and introduce a SacI binding site. This fragment was then cloned into the pJET vector (CloneJET PCR cloning kit, Thermo Fisher). A synthetic DNA fragment (Genewiz, Table 2.3) containing the 83 bp CTCF binding site, flanked by heterospecific lox sites (loxP and lox2272, (Lee and Saito 1998)), two FRT sites and flanking rox sites was cloned between the homology arms using SacI. A neomycin selection cassette was cloned between the FRT sites via blunt end ligation and the sequence of the final plasmid sequence was confirmed by Sanger sequencing.

Table 2.1: Guide RNA sequences used for inserting the CTCF binding site

Guide	Name	Target sequence
1	pDN1	GCTGTAGTGTAACTAACTGC
2	p1	GCTTCAAGAACTGCCTTCCTG

Table 2.2: Primers used for cloning HDR plasmid

Homology arm	F/R	Sequence
Left homology arm	F	TCCTTGGCTATATATTGCACTTGA
Left homology arm	R	TGCGAGAGCTCTACACCTTCATTCAGGGACATTGC
Right homology arm	F	TGTAGAGCTCTCGCAACAATTAAAAGAGAATAACCGCTGT
Right homology arm	R	ATCCCTTGGAGTTCAGAGTTACAG

Table 2.3: Synthetic fragment used in cloning the CTCF insertion HDR plasmid

	Sequence
Fragment containing CTCF binding site	CGCACGACGAGCTCGTGGCTAACTTTAAATAATGCCAATTATTTAAAGTTAAAACGCATAGC TCGCTCAATGGAAGTTCCTATTCCGAAGTTCCTATTCTCTAGAAAGTATAGGAACTTCGACT CGGTACCGAGCCGAATTCTTTCAGAAGTTCCTATTCCGAAGTTCCTATTCTCTAGAAAGTAT AGGAACTTCACTTCTCAATATGTGAGCACTATAACTTCGTATAATGTATGCTATACGAAGTT ATGGTCATTGGCTCACGGAACTTATAACTTCGTATAAAGTATCCTATACGAAGTTATGCAAA CTGAGGTCCTGGGTAGGCCTCTGCTACCCTCTGGTGGCCTCAGGGTGCATAGCACTACCAG GTGAGGATACTGGCATTATAACTTCGTATAGCATACATTATACGAAGTTATGAGATCCTGTC GTCCTGGGTTATAACTTCGTATAGGATACTTTATACGAAGTTATTATGAATTTCGTATTTGCA GCTAACTTTAAATAATGCCAATTATTTAAAGTTACTGCAGAGCTCTCGGTCCTGG

Cloning for the CTCF reporter assay HDR plasmids

For sequences used for the control DNA fragment and HS-38 binding site DNA fragment see Table 2.4. The HS-38/H-39 fragment was amplified from the C57BL/6J genome (for primers see Table 2.4). These fragments were cloned into a vector with the same homology arms as the CTCF insertion HDR vector. In addition, the vector contained a hygromycin and ganciclovir (tyrosine-kinase) selection cassette flanked by lox sites.

Table 2.4: DNA fragments and primers used for the CTCF reporter assay cloning

Oligonucleotide	**F/R**	**Sequence**
HS38 Reverse	F	AAACAATGCCAGTATCCTCACCTGGTAGTGCTATGCACCCTGAGGCCACCAG AGGGTAGCAGAGGCCTACCCAGGACCTCAGTTTGC
HS38 Reverse	R	TCCGGCAAACTGAGGTCCTGGGTAGGCCTCTGCTACCCTCTGGTGGCCTCAG GGTGCATAGCACTACCAGGTGAGGATACTGGCATT
Control sequence	F	AAACACAATAAATAATTAAAATGTAATGAAAATATGGCTAATTAATCTTTACA TTCTGTCTTGGCCAAGATGCAGAGAAGCGACAGT
Control sequence	R	TCCGACTGTCGCTTCTCTGCATCTTGGCCAAGACAGAATGTAAAGATTAATTA GCCATATTTTCATTACATTTTAATTATTTATTGT
HS38-39 primer	F	CGTCTCGAAACGGCCTCATGGATTCAAAGCCAC
HS38-39 primer	R	CGTCTCATCCGGAATGAATGCCAGTATCCTCACCTG
HS38-39 reverse primer	F	CGTCTCGAAACGAATGAATGCCAGTATCCTCACCTG
HS38-39 reverse primer	R	CGTCTCATCCGGGCCTCATGGATTCAAAGCCAC

Cloning of inducible dCas9-p300 plasmid

All restriction enzymes used here were from New England Biolabs (NEB) and were used following the manufacturer's instructions. First, the inducible Caspex expression plasmid (Addgene plasmid #97421) was digested with ClaI and Acc65I resulting in 8.3 and 5.1 kb fragments. The

fragments were run on an agarose gel and the 8.3 kb fragment was excised and extracted (QIAquick gel extraction kit, Qiagen). This 8.3 kb fragment was assembled with the T2A repair oligo (Table 2.5) using NEBuilder HiFi DNA Assembly Cloning Kit (NEB) following manufacturer's instructions, and DH5α-derived 5-alpha competent *E. Coli* cells (NEB) were transformed with the product. Bacteria were plated on ampicillin-containing LB agar plates, incubated overnight and single colonies were picked for expansion in LB media. Plasmid DNA was isolated using a QIAquick Spin Miniprep kit (Qiagen) and the plasmid sequence was confirmed using Sanger sequencing (PB-T2A-GFP-EF1alpha-Tet-puro plasmid).

Next, the PGK-TetOn-mCherry fragment (Table 2.5), which was ordered as part of a plasmid, was digested with ScaI, AsiSI and BstXI and the 2.5 kb fragment was excised from an agarose gel and extracted using a Zymoclean Gel DNA Recovery Kit (Zymo Research). The PB-T2A-GFP-EF1α-Tet-Puro plasmid produced in the previous step was digested with AsiSI and BstXI and phosphatase treated with Quick CIP (NEB) following the manufacturer's instruction. The restricted and phosphatase-treated PB-T2A-GFP-EF1α-Tet-Puro fragments were purified using a QIAquick PCR Purification kit (Qiagen) and ligated with the 2.5 kb fragment from the PGK-TetOn-mCherry plasmid using T4 ligase (NEB). The ligation product was transformed into 5-alpha competent *E. coli* cells and plasmid DNA was screened as described above. This resulted in the PB-T2A-GFP-PGK-TetOn-mCherry plasmid.

Then, dCas9-p300 core plasmid (Addgene plasmid # 61357) was isolated using the Miraprep protocol (Pronobis et al. 2016) in combination with the QIAquick Spin Miniprep kit (Qiagen) and digested with ClaI, BsiWI and AvrII, resulting in 6.1 kb, 4.3 kb and 1.1 kb fragments. The 6.1 kb was purified by agarose gel purification. This fragment was then ligated (T4 ligase, NEB) with the fragment resulting from the digest of PB-T2A-GFP-PGK-TetOn-mCherry with ClaI and Acc65I, which was treated with quick CIP (NEB). The ligation product was transformed into DH10B-derived 10-beta competent *E. Coli* cells (NEB) and plasmids from resulting colonies were

screened by digest and Sanger sequencing. This resulted in the PB-dCas9-p300-T2A-GFP-PGK-TetOn-mCherry plasmid.

Finally, the alpha-promoter gRNA array (Table 2.5) was synthesised by GeneART and the pMA cloning vector was digested with AvrII and AflII. The resulting 3 kb fragment was purified using agarose gel purification and ligated (T4, NEB) with fragment resulting from the digest of the PB-dCas9-p300-T2A-GFP-PGK-TetOn-mCherry plasmid with AvrII and AflII. The ligation product was transformed into 10-beta competent *E. coli* bacteria. Plasmid DNA was purified as described above, and the resulting plasmids were screened by digest for the correct ligation. The sequence was then verified using sanger sequencing.

Table 2.5: DNA sequences ordered from Thermo Fisher for cloning the dCas9-p300 plasmid

Oligo	**Sequence**
T2A repair oligo	GACGGTGATTATAAAGATCATGACATCGATTACAAGGATGACGATGACAAGATGGCCGC CCAAGAAGGTACCGCGCGCGAGACGTCGAGGGCAGAGGAAGTCTTCTAACATGCGGTG ACGTGGAGGAGAATCCCGGCCCTCGTACCGTGAGCAAGGGCGAGGAGCTGTTCACCGG G
PGK-TetOn-mCherry	CCCCCAAGTTCGGGTGAAGGCCCAGGGCTCGCAGCCAACGTCGGGGCGGCAGGCCCTG CCATAGCAGATCTGCGAAATCGGATCCAGCTGAATCTAAGTGGATCTGCGATCGCTCCG GTGCCCGTCAGTGGGTAGGGGAGGCGCTTTTCCCAAGGCAGTCTGGAGCATGCGCTTTA GCAGCCCCGCTGGGCACTTGGCGCTACACAAGTGGCCTCTGGCCTCGCACACATTCCACA TCCACCGGTAGGCGCCAACCGGCTCCGTTCTTTGGTGGCCCCTTCGCGCCACCTTCTACTC CTCCCCTAGTCAGGAAGTTCCCCCCCGCCCCGCAGCTCGCGTCGTGCAGGACGTGACAAA TGGAAGTAGCACGTCTCACTAGTCTCGTGCAGATGGACAGCACCGCTGAGCAATGGAAG CGGGTAGGCCTTTGGGGCAGCGGCCAATAGCAGCTTTGCTCCTTCGCTTTCTGGGCTCAG AGGCTGGGAAGGGGTGGGTCCGGGGGCGGGCTCAGGGGCGGGCTCAGGGGCGGGGC GGGCGCCCGAAGGTCCTCCGGTGGCCCGGCATTCTGCACGCTTCAAAAGCGCACGTCTG CCGCGCTGTTCTCCTCTTCCTCATCTCCGGGCCTTTCGTAAGCTGTGACCGGCGCCTACTC TAGCCTGCAGGAGCCGCCATGTCTAGACTGGACAAGAGCAAAGTCATAAACTCTGCTCT GGAATTACTCAATGGAGTCGGTATCGAAGGCCTGACGACAAGGAAACTCGCTCAAAAGC TGGGAGTTGAGCAGCCTACCCTGTACTGGCACGTGAAGAACAAGCGGGCCCTGCTCGAT GCCCTGCCAATCGAGATGCTGGACAGGCATCATACCCACTCCTGCCCCCTGGAAGGCGA GTCATGGCAAGACTTTCTGCGGAACAACGCCAAGTCATACCGCTGTGCTCTCCTCTCACA TCGCGACGGGGCTAAAGTGCATCTCGGCACCCGCCCAACAGAGAAACAGTACGAAACCC TGGAAAATCAGCTCGCGTTCCTGTGTCAGCAAGGCTTCTCCCTGGAGAACGCACTGTACG CTCTGTCCGCCGTGGGCCACTTTACACTGGGCTGCGTATTGGAGGAACAGGAGCATCAA GTAGCAAAAGAGGAAAGAGAGACACCTACCACCGATTCTATGCCCCCACTTCTGAAACA AGCAATTGAGCTGTTCGACCGGCAGGGAGCCGAACCTGCCTTCCTTTTCGGCCTGGAACT AATCATATGTGGCCTGGAGAAACAGCTAAAGTGCGAAAGCGGCGGGCCGACCGACGCC CTTGACGATTTTGACTTAGACATGCTCCCAGCCGATGCCCTTGACGACTTTGACCTTGATA TGCTGCCTGCTGACGCTCTTGACGATTTTGACCTTGACATGCTCCCCGGGGAGGGCAGAG GAAGTCTTCTAACATGCGGTGACGTGGAGGAGAATCCCGGCCCTATGGTGAGCAAGGG CGAGGAGGATAACATGGCCATCATCAAGGAGTTCATGCGCTTCAAGGTGCACATGGAG

	GGCTCCGTGAACGGCCACGAGTTCGAGATCGAGGGCGAGGGCGAGGGCCGCCCCTACG AGGGCACCCAGACCGCCAAGCTGAAGGTGACCAAGGGTGGCCCCCTGCCCTTCGCCTGG GACATCCTGTCCCCTCAGTTCATGTACGGCTCCAAGGCCTACGTGAAGCACCCCGCCGAC ATCCCCGACTACTTGAAGCTGTCCTTCCCCGAGGGCTTCAAGTGGGAGCGCGTGATGAA CTTCGAGGACGGCGGCGTGGTGACCGTGACCCAGGACTCCTCTCTGCAGGACGGCGAGT TCATCTACAAGGTGAAGCTGCGCGGCACCAACTTCCCCTCCGACGGCCCCGTAATGCAGA AGAAGACAATGGGCTGGGAGGCCTCCTCCGAGCGGATGTACCCCGAGGACGGCGCCCT GAAGGGCGAGATCAAGCAGAGGCTGAAGCTGAAGGACGGCGGCCACTACGACGCTGA GGTCAAGACCACCTACAAGGCCAAGAAGCCCGTGCAGCTGCCCGGCGCCTACAACGTCA ACATCAAGTTGGACATCACCTCCCACAACGAGGACTACACCATCGTGGAACAGTACGAA CGCGCCGAGGGCCGCCACTCCACCGGCGGCATGGACGAGCTGTACAAGTAACAACCTCT GGATTACAAAATTTGTGGGTAAACTTGTTTATTGCAGCTTATAATGGTTACAAATAAAGC AATAGCATCACAAATTTCACAAATAAAGCATTTTTTTCACTGCATTCTAGTTGTGGTTTGT CCAAACTCATCAATGTATCTTATCATGTCTGGAATTGACTCAAATGATGTCAATTAGTCTA TCAGAAGCTCATCTGGTCTCCCTTCCGGGGGACAAGACATCCCTGTTTAATATTTAAACA GCAGTGTTCCCAAACTGGGTTCTTATATCCCTTGCTCTGGTCAACCAGGTTCTTAAGGCAG GGTTTCCTGTCCTCACAGGAACGAAGTCCCTAAAGAAACAGTGGCAGCCAGGTTTAGCC CCGGACCTAGGATTGACTGGATTCCTTTTTTAGGGCCCATTGGTATGGCTTTTTCCCCGTA TCCCCCCAGGTGTCTGCAGGCTCAAAGAGCAGCGAGAAGC
Alpha-promoter gRNA array	CACGCATCTGGAATAAGGAAGTGCCATTCCGCCTGACCTCTTAAGTCGGGATCCGCTGG AGCGGCCGCTTTCCCATGATTCCTTCATATTTGCATATACGATACAAGGCTGTTAGAGAG ATAATTAGAATTAATTTGACTGTAAACACAAAGATATTAGTACAAAATACGTGACGTAGA AAGTAATAATTTCTTGGGTAGTTTGCAGTTTTAAAATTATGTTTTAAAATGGACTATCATA TGCTTACCGTAACTTGAAAGTATTTCGATTTCTTGGCTTTATATATCTTGTGGAAAGGACG AAACACCATATCCAGGGGCTGGGGCTGGTTTTAGAGCTAGAAATAGCAAGTTAAAATAA GGCTAGTCCGTTATCAACTTGAAAAAGTGGCACCGAGTCGGTGCTTTTTTGAATGGATCC GCTGGAGCGGCCGCTTTCCCATGATTCCTTCATATTTGCATATACGATACAAGGCTGTTA GAGAGATAATTAGAATTAATTTGACTGTAAACACAAAGATATTAGTACAAAATACGTGAC GTAGAAAGTAATAATTTCTTGGGTAGTTTGCAGTTTTAAAATTATGTTTTAAAATGGACTA TCATATGCTTACCGTAACTTGAAAGTATTTCGATTTCTTGGCTTTATATATCTTGTGGAAA GGACGAAACACCAGACCTGCTCACATAAAAAGGTTTTAGAGCTAGAAATAGCAAGTTAA AATAAGGCTAGTCCGTTATCAACTTGAAAAAGTGGCACCGAGTCGGTGCTTTTTTGAATG GATCCGCTGGAGCGGCCGCTTTCCCATGATTCCTTCATATTTGCATATACGATACAAGGCT GTTAGAGAGATAATTAGAATTAATTTGACTGTAAACACAAAGATATTAGTACAAAATACG TGACGTAGAAAGTAATAATTTCTTGGGTAGTTTGCAGTTTTAAAATTATGTTTTAAAATGG ACTATCATATGCTTACCGTAACTTGAAAGTATTTCGATTTCTTGGCTTTATATATCTTGTGG AAAGGACGAAACACCAGCATTTCCTTATTGATCCAGTTTTAGAGCTAGAAATAGCAAGTT AAAATAAGGCTAGTCCGTTATCAACTTGAAAAAGTGGCACCGAGTCGGTGCTTTTTTGAA TGGATCCGCTGGAGCGGCCGCTTTCCCATGATTCCTTCATATTTGCATATACGATACAAG GCTGTTAGAGAGATAATTAGAATTAATTTGACTGTAAACACAAAGATATTAGTACAAAAT ACGTGACGTAGAAAGTAATAATTTCTTGGGTAGTTTGCAGTTTTAAAATTATGTTTTAAAA TGGACTATCATATGCTTACCGTAACTTGAAAGTATTTCGATTTCTTGGCTTTATATATCTTG TGGAAAGGACGAAACACCACTATAAGAGCAAGAGGTAGGTTTTAGAGCTAGAAATAGC AAGTTAAAATAAGGCTAGTCCGTTATCAACTTGAAAAAGTGGCACCGAGTCGGTGCTTTT TTGAATGGATCCGCTGGAGCGGCCGCTTTCCCATGATTCCTTCATATTTGCATATACGATA CAAGGCTGTTAGAGAGATAATTAGAATTAATTTGACTGTAAACACAAAGATATTAGTACA AAATACGTGACGTAGAAAGTAATAATTTCTTGGGTAGTTTGCAGTTTTAAAATTATGTTTT AAAATGGACTATCATATGCTTACCGTAACTTGAAAGTATTTCGATTTCTTGGCTTTATATA TCTTGTGGAAAGGACGAAACACCAAGGGGGGGGATCAGTGATGGTTTTAGAGCTAGAA ATAGCAAGTTAAAATAAGGCTAGTCCGTTATCAACTTGAAAAAGTGGCACCGAGTCGGT GCTTTTTTGAATGGATCCGCTGGAGCGGCCGCTTTCCCATGATTCCTTCATATTTGCATAT ACGATACAAGGCTGTTAGAGAGATAATTAGAATTAATTTGACTGTAAACACAAAGATATT

AGTACAAAATACGTGACGTAGAAAGTAATAATTTCTTGGGTAGTTTGCAGTTTTAAAATT
ATGTTTTAAAATGGACTATCATATGCTTACCGTAACTTGAAAGTATTTCGATTTCTTGGCT
TTATATATCTTGTGGAAAGGACGAAACACCGAATAAATTGCTCAAGGACAGTTTTAGAGC
TAGAAATAGCAAGTTAAAATAAGGCTAGTCCGTTATCAACTTGAAAAAGTGGCACCGAG
TCGGTGCTTTTTTGAATGGATCCGCTGGAGCGGCCGCTTTCCCATGATTCCTTCATATTTG
CATATACGATACAAGGCTGTTAGAGAGATAATTAGAATTAATTTGACTGTAAACACAAAG
ATATTAGTACAAAATACGTGACGTAGAAAGTAATAATTTCTTGGGTAGTTTGCAGTTTTA
AAATTATGTTTTAAAATGGACTATCATATGCTTACCGTAACTTGAAAGTATTTCGATTTCTT
GGCTTTATATATCTTGTGGAAAGGACGAAACACCAGCAGCAAAAAGCTTGTCCGGTTTTA
GAGCTAGAAATAGCAAGTTAAAATAAGGCTAGTCCGTTATCAACTTGAAAAAGTGGCAC
CGAGTCGGTGCTTTTTTGAATGGATCCGCTGGAGCGGCCGCTTTCCCATGATTCCTTCAT
ATTTGCATATACGATACAAGGCTGTTAGAGAGATAATTAGAATTAATTTGACTGTAAACA
CAAAGATATTAGTACAAAATACGTGACGTAGAAAGTAATAATTTCTTGGGTAGTTTGCAG
TTTTAAAATTATGTTTTAAAATGGACTATCATATGCTTACCGTAACTTGAAAGTATTTCGA
TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCACCTTTCTGAAGCTGCATTAG
TTTTAGAGCTAGAAATAGCAAGTTAAAATAAGGCTAGTCCGTTATCAACTTGAAAAAGTG
GCACCGAGTCGGTGCTTTTTTGAATGGATCCCCTAGGAGGCTAGGTGGAGGCTCAGTGA
TGATAAGTCTGCG

Generation of CTCF insertion mouse ES cells

ES cell electroporation was performed using the Neon transfection system (Invitrogen) using 3 x 1400V for 10 ms with 2.5 µg of the two sgRNA nickase-Cas9 plasmids and 5 µg HDR plasmid. The transfected ES cells were plated at low density and selected with 600 ng/ml puromycin for two days, followed by 350 ng/ml G418 selection for 5 days. Individual clones were selected, expanded and initially screened for correct integration using homology-arm spanning PCRs (Immolase DNA polymerase, Bioline, supplemented with Q solution, Qiagen; for screening primers see Table 2.6). Genotypes of positive clones were confirmed by Sanger sequencing. Three clones with the correct insertion were transiently transfected with either 5 µg pFlpO (Addgene #13793) to remove the neomycin selection cassette or both 5 µg pFlpO (Addgene #13792) and 5 µg of pCrePac (Taniguchi et al. 1998) to simultaneously remove the selection cassette and invert the CTCF binding site. This provided clones with the forward and reverse orientated CTCF site and three clones were selected for each genotype.

Table 2.6: primer sequences used for genotyping PCRs

Screen ID	Target	F/R	Sequence
1	CTCF side insert	F	GAGTGGCTCTGAGGGTTGAG
1	CTCF side insert	R	GATCCTGTCGTCCTGGGTTA
2	Neo side insert	F	TCGCGAGTCGAAGTTCCTAT
2	Neo side insert	R	CCAGAGGATCTGGGTTCAAA
3	Across insert	F	AGTTTGTGGTTGGGAGAGGAGG
3	Across insert	R	GTGCACACGGGGAAGAAGTTTG

Generation of CTCF insertion mouse models and animal procedure

ES cells with the original insertion (as described above), prior to Flp and Cre recombinase treatment, were used for blastocyst injection. After germline transmission, these were crossed with a Flp-expressing mouse line (B6.Cg-Tg(ACTFLPe)9205Dym/J; Jackson Laboratory stock #005703), and following positive genotyping, with a Cre-expressing line (Mao et al. 1999) to produce both a forward and reverse CTCF insertion strains.

The mutant mouse strains were backcrossed and maintained on a C57BL/6J background. Protocols were approved through the Oxford University Local Ethical Review process. Experimental procedures were in accordance with the European Union Directive 2010/63/EU and/or the UK Animals (Scientific Procedures) Act (1986), and reviewed by the clinical medicine Animal Welfare and Ethical Review Body (AWERB). Experimental procedures were conducted under project licences PPL 30/2966 and PPL 30/3339.

Blastocyst injection, initial breeding and genotyping steps were performed by Ben Davies (Wellcome Centre for Human Genetics, Oxford). Subsequent crosses with Flp- and Cre-expressing mice as well as colony maintenance were performed by Jacqueline Sharpe, Jacqueline Sloane-Stainley, Sue Butler and Hannah Robinson (Transgenic core, WIMM).

Generation of the CTCF reporter ES cells

CTCF reporter ES cells were targeted by Muhammad Hanifi. Mouse ES cells with a YFP-tag at the *Hba-a1*, as previously reported (Francis et al. 2022), were transfected with 1.25 µg sgRNA nickase-Cas9 plasmid and 1.25 µg HDR plasmid, using Lipofectamine LTX (Thermo Fisher) according to the manufacturer's instructions. After 24 hours, the cells were then incubated with

1 µg/ml puromycin for 48 hours, followed by 7 days of incubation with 120 µg/ml hygromycin. Subsequently, cells were transfected with 2.5 µg Cre-recombinase plasmid to delete the selection cassette. Cells were then incubated for 5 days before addition of ganciclovir at 3 µM for 7 days.

Generation of all enhancer KO and wild type hemizygous mouse ES cells

The all enhancer KO and wild type hemizygous mouse ES cell clones were generated by Helena Francis in collaboration with Jef Boeke (New York University, USA), following a recombinase-mediated genomic replacement (RMGR) strategy (Wallace et al. 2007). A RMGR-ready mouse ES cell line, initially created from WT E14 mESCs, has two lox sites 85 kb apart (85,145 bp; chr11:32,115,389-32,200,533 (mm9)), flanking the alpha-globin. These sites facilitate the integration of bacterial artificial chromosomes (BACs) at the alpha-globin locus, making it possible to replace many elements in one recombination event. A further modification of the RMGR-ready cells deleted 117 kb of the allele without the lox sites which removed the entire alpha-globin-associated self-interacting domain and resulting in cell lines that were hemizygous for the alpha-globin locus (guide RNAs in Table 2.7).

Initially, two BAC sequences were integrated into the RMGR-ready cells: a BAC with R1, R3, Rm and R4 enhancer sequences deleted, leaving only R2, and a BAC with a wild type sequence. There are a small number of mutations in the BACs compared to the design (Table 2.8 and Table 2.9), which occurred during synthesis. However, none of them are predicted to impact functional elements at the locus. In total, there are three clones with the wild type BAC integrated correctly, called the wild-type hemizygous lines. In the R2-only clones, the R2 enhancer sequence was then deleted using CRISPR/Cas9 (Table 2.7), resulting in four independent all enhancer KO clones. The sequence of the R2-only clones was confirmed with linked-read library preparation (10X Genomics) followed by Illumina sequencing.

Full methods have been described in “Functional dissection of a single enhancer at the mouse α-globin locus” (Francis 2019).

Table 2.7: Guide RNAs used to create allele and R2 deletions

Target	Location in mm10	Sequence
Allele deletion left		ACAAGATGCTCTCCGCTGT
Allele deletion right		AGGACAAATCCCGAAGGTC
R2 deletion left		TACCTCCAAGGTTTTGCTC
R2 deletion right		GCCGTGACACTTCATGCTCA

Table 2.8: Variants detected in the WT synthetic BAC compared to design (Francis 2019)

Position in BAC design	Type of variant	Reference	Variant	Description
5679	Substitution	C	T	Rhbdf1 exon 2 - synonymous
27421	Substitution	C	T	Nprl3 intron
31550	Substitution	C	T	Nprl3 intron
35798	Substitution	C	T	Nprl3 intron
39735	Substitution	C	T	Nprl3 exon 5 – synonymous
45789	Substitution	T	C	R3 enhancer
68981	Substitution	T	C	Intergenic
74052	Substitution	G	A	Intergenic

Table 2.9: Variants detected in the R2-only synthetic BAC compared to design (Francis 2019)

Position in BAC design	Type of variant	Reference	Variant	Description
6145	Substitution	C	T	Rhbdf1 intron
44425	Deletion	30 bp	-	Nprl3 intron
46911	Substitution	G	A	Nprl3 intron
58887	Substitution	G	A	Intergenic

Generation of dCas9-p300 expressing AEKO mouse ES cells

1 x 10^6 cells of a AEKO clone was transfected with 2.5 µg pCAGS-PBase (kindly provided by Austin Smith) and 2.5 µg inducible dCas9-p300 plasmid using Lipofectamine LTX Reagent with PLUS reagent (Thermo Fisher) following manufacturer's instructions. After 24 hours, the media containing the Lipofectamine was exchanged for fresh media and after 48 hours the single cells mCherry positive cells were sorted into individual wells using FACS. The cells were then grown in conditioned media for a week and passaged onto replicate plates, which were then used to screen for mCherry expression by FACS again. This resulted in 4 clones positive for mCherry expression. Integration of the plasmid was further verified by PCR using the primers in Table 2.10.

Table 2.10: Primers used to verify integration of the inducible dCas9-p300 plasmid

Target	F/R	Sequence
N-terminus dCas9	F	CCGTACCACTTCCTACCCTC
N-terminus dCas9	R	CGATTCTTCCGACGTGTATACC
mCherry	F	CCCCGTAATGCAGAAGAAGA
mCherry	R	GAAATTTGTGATGCTATTGC
sgRNA array	F	GGGGGGGATCAGTGATGG
	R	CTGTCCTTGAGCAATTTATTCGGTG

Gene expression analysis

RNA extraction

Total RNA was extracted from 5 x 10^5 – 5 x 10^6 cells. Cells were lysed in TRI Reagent (Sigma-Aldrich) according to manufacturer's instruction and immediately stored at -80 °C. Total RNA was isolated either via chloroform extraction or using a Direct-zol MiniPrep or MicroPrep kit (Zymo Research).

For extraction using chloroform, chloroform was added to the sample in TRI reagent in a ratio 1:5 (v/v) and the sample was mixed vigorously by vortexing, followed by room temperature incubation. Then the sample was centrifuged (15000 RPM, 4 °C) for 10 minutes and the aqueous layer was transferred and precipitated with isopropanol and glycoblue (Thermo Fisher) at room temperature for a few minutes. The samples were centrifuged (15000 RPM, 4 °C, 10 minutes) and the pellet containing the RNA was washed with 70% ethanol, air dried and resuspended in DNAse/RNase-free water. Samples were then treated with DNase using the TURBO DNA-free kit (Invitrogen) according to manufacturer's instruction. Briefly, samples were treated with DNase enzyme for 30 minutes followed by inactivation for 5 minutes, after which the samples were centrifuged and the supernatant was used for cDNA synthesis after quality control steps.

RNA isolated using the Direct-zol kits was performed according to manufacturer's instructions, but the DNase I treatment was extended to 30 minutes at room temperature (rather than 15 minutes).

The quality of all isolated RNA was assessed using RNA Analysis Screentape (Tapestation, Agilent Technologies) and only samples with high (>7.5) RNA integrity number (RIN) scores were used

in subsequent steps. RNA was stored at -80 °C.

cDNA synthesis

RNA concentration was determined using a NanoDrop (Thermo Fisher) and an equal amount of RNA was used for cDNA synthesis where samples were processed simultaneously. Up to 1 µg RNA was used for cDNA synthesis using the Superscript III Reverse Transcriptase kit (Invitrogen) following manufacturer's instructions. For each sample, a control without reverse transcriptase enzyme was included to account for any incomplete genomic DNA digestion.

qPCR analysis

cDNA was diluted 5x before qPCR analysis to determine the relative changes in gene expression. qPCR analysis was performed in triplicate using Fast SYBR green master mix (Applied Biosystems) using primers listed in Table 2.11. The ΔΔCt-method was used for relative quantification of RNA abundance.

Table 2.11: Primers used for RT-qPCR expression analysis

Primer target	F/R	Sequence
Hba-a1/2	F	CTGGGGAAGACAAAAGCAAC
	R	GCCGTGGCTTACATCAAAGT
Hba-x	F	ATGCGGTTAAGAGCATCGAC
	R	GGGACAGGAGCTTGAAGTTG
Hbb-b1	F	ATGGCCTGAATCACTTGGAC
	R	ACGATCATATTGCCCAGGAG
Hbb-y	F	TGGCCTGTGGAGTAAGGTCAA
	R	GAAGCAGAGGACAAGTTCCCA
Nprl3	F	GCTCTTCAGGTACCCCTTCC
	R	ATGTTCGCCAGTGTTGTTGA
Rhbdf1	F	ATCCTAGTGCCCCAGACCTT
	R	CGGCCACTTGGTGATATCTT
Mpg	F	CTTCTCCAGCCCAGAGGAC
	R	ATGCCTCAGTCTCCACAATG
RN18s	F	GTAACCCGTTGAACCCCATT
	R	CCATCCAATCGGTAGTAGCG

RNA sequencing

RNA was isolated using a Direct-zol MiniPrep Kit as described above. RNA was prepared for sequencing using NEBNext Single Cell/Low Input RNA Library Prep Kit for Illumina (NEB) following manufacturer's instructions, with 200 ng RNA of each sample. Libraries were

sequenced on the Illumina NextSeq platform using a 75-cycle paired end kit (NextSeq 500/550 High Output Kit).

Next-generation sequencing assays

ATAC-seq

Assay for transposase-accessible chromatin (ATAC)-seq was performed as previously described (Buenrostro et al. 2013). Briefly, 60 000 – 75 000 cells were lysed in cold lysis buffer (10 mM Tris HCl, 10 mM NaCl, 3.4 mM $MgCl_2$, 0.1% IGEPAL CA-630) to isolate nuclei. Pelleted nuclei were incubated in transposase reaction mix (1 X TD reaction buffer, 2.5 µl Tn5 transposase (Illumina Tagment DNA Enzyme and Buffer kit)) for 30 minutes at 37 °C. Tagmented DNA was purified using a MinElute PCR purification kit (Qiagen) and amplified and indexed with NEBNext High-Fidelity 2X PCR Master Mix (NEB) and custom primers (Buenrostro et al. 2013). Amplified libraries were purified with QIAquick PCR Purification kit (Qiagen) and library quality was assessed by D1000 Tapestation (Agilent Technologies). Libraries were sequenced on the Illumina NextSeq platform using a 75-cycle paired end kit (NextSeq 500/550 High Output Kit).

ChIP-seq

A Chromatin Immunoprecipitation (ChIP) Assay Kit (Merck Millipore 17-295) was used for all ChIP experiments following manufacturer's instructions. First, chromatin of 3 – 10 x 10^7 isolated erythroid cells was fixed in 1% formaldehyde for 10 minutes. For Rad21 ChIP specifically, 10 x 10^7 cells were fixed with 2 mM disuccinimidyl glutarate (DSG; Thermo Fisher) for 50 minutes, followed by 10 minutes of fixation with 1% formaldehyde. The crosslinking reaction was quenched with a final concentration of 125 mM glycine for 5 minutes. After cell lysis, chromatin was sonicated using the Bioruptor (Diagenode), Bioruptor pico (Diagenode) or ME220 Focused-ultrasonicator (Covaris) to produce fragments of 200-500 bp. Sonicated chromatin was incubated with antibodies (Table 2.12) at 4 °C on a rotator overnight. Protein A agarose beads were added and incubated with the chromatin for 60 minutes at 4 °C with rotation, and beads were sequentially washed and eluted following manufacturer's instructions. Good quality

libraries, as determined by qPCR enrichment, were prepared for sequencing with an Ultra II Library Prep Kit (NEB) and sequenced on the Illumina NextSeq platform using a 75-cycle paired end kit (NextSeq 500/550 High Output Kit).

Table 2.12: Antibodies used for ChIP

Protein target	Product code	Polyclonal/Monoclonal	Volume per 10^7 cells
CTCF	Merck 07-729	Polyclonal	10 µl
Rad21	Abcam ab992	Polyclonal	10 µl
H3K27me3	CST 9733 C36B11	Monoclonal	5 µl
Pol2 (N20)	Santa Cruz sc-899	Polyclonal	10 µl
Gata1	Abcam ab11852	Polyclonal	10 µl
Nf-e2	Santa Cruz sc-22827	Polyclonal	2 µl
TBP	Santa Cruz sc-204	Polyclonal	10 µl

Chipmentation

Chipmentation combines chromatin immunoprecipitation with Tn5 transposase to prepare libraries for sequencing by treating immunoprecipitated chromatin with Tn5 transposase while the chromatin is on bound to magnetic beads (Schmidl et al. 2015). The Chipmentation protocol was based on Schmidl et al. (Schmidl et al. 2015). Cells were fixed with 1% formaldehyde for 10 minutes at room temperature, after which the reaction was stopped by incubation with 125 mM glycine for 5 minutes. Samples were centrifuged (5 minutes, 1000 RPM, 4 °C), washed with PBS and snap frozen before storage at -80 °C. Cells were sonicated in lysis buffer (50mM Tris, 10 mM EDTA, 0.5% SDS) using ME220 Focused-ultrasonicator (Covaris) to produce fragments of 200-500 bp. Triton-X was added to the samples to a final concentration of 1% to neutralise SDS in the lysis buffer. In parallel, 5 µl of both Protein A and Protein G dynabeads (Invitrogen) per sample/antibody combination were incubated with the appropriate antibody (see Table 2.13), in PBS with 0.5% BSA and proteinase inhibitors (cOmplete Proteinase Inhibitor Cocktail, Roche) for 2-4 hours at 4 °C. Sonicated chromatin corresponding to 150 000 cells was then added to antibody-incubated dynabeads and incubated in the cold room overnight with rotation.

The next day, the beads were washed 3 times with RIPA buffer (10 mM Tris-HCl, 1 mM EDTA, 0.5 mM EGTA, 1% Triton-X, 0.1% SDS, 0.1% Na Deoxycholate, 140 mM NaCl, 1x Proteinase Inhibitor Cocktail), 2 times with TE buffer (Sigma Aldrich) and 2 times with 10mM Tris. All washes

were performed at 4 °C and in the cold room. The beads were then resuspended in 30 µl pre-warmed tagmentation buffer and 1 µl Tn5 (Illumina Tagment DNA Enzyme and Buffer kit) and incubated for 10 minutes at 37 °C. The tagmentation reaction was inactivated by a wash step with RIPA and 10mM Tris.

Libraries were prepared using NEBNext High-Fidelity 2X PCR Master Mix (NEB) and custom primers, followed by purification by Ampure XP beads (Beckman Coulter) and sequencing on the Illumina NextSeq platform using a 75-cycle paired end kit (NextSeq 500/550 High Output Kit).

Table 2.13: Antibodies used for Chipmentation

Protein target	Product code	Polyclonal/Monoclonal	Volume per reaction
H3K4me3	Millipore 07-473	Polyclonal	1 µl
H3K4me1	Abcam ab8895	Polyclonal	2 µl
H3K27ac	Abcam ab4729	Polyclonal	3 µl

Capture-C

Capture-C was performed as previously described, either following the next-generation Capture-C protocol (Davies et al. 2016) or the NuCapture-C protocol (Downes et al. 2020). Briefly, a total of 3 - 10 x 10^6 EB-derived CD71+ erythroid cells were fixed with 2% formaldehyde for 10 minutes. The crosslinking reaction was quenched by the addition of 125 mM glycine and the cells were washed with PBS, followed by a 20 minute incubation in 5 ml lysis buffer (10 mM Tris, 10 mM NaCl, 0.2% Igepal CA-630, 1x proteinase inhibitor cocktail). Cells were then snap-frozen and stored at -80 °C. Libraries were produced following the protocols as published using DpnII digestion. Library quality was assessed by qPCR using digested and undigested controls (for primers see Table 2.14) and agarose gel electrophoresis. Libraries of sufficient quality were indexed using NEBNext Ultra II DNA Library Prep Kit for Illumina (New England Biolabs: E7645) according to the manufacturer's instructions. Capture enrichment was performed using 120 bp-long biotinylated oligos designed to the 5' and 3' end of DpnII fragments listed in Table 2.15.

Table 2.14: Primer used for the quantification of the digestion efficiency of 3C libraries

Target	F/R	Sequence
Digested control	F	GGAGAAAGAAGGCTGGTGTTAT
	R	TATCTGAGTTGGACAGCATTGG
Genomic control	F	TTATCTTGCATTTGCCAACTCG
	R	TGGGTTTCCCTGATTCTGAAA

Table 2.15: Biotinylated oligos used as Capture-C probes

DpnII fragment	Position in mm9
HS-39	chr11:32138079-32139000
HS-38	chr11:32137176-32137426
R1	chr11:32145277-32146568
R2	chr11:32151060-32151883
HS+44	chr11:32220918-32221720
HS+48	chr11:32224323-32227298

Computation analysis

Mapping and visualisation of ATAC-seq, ChIP-seq and ChIPmentation samples

Next-generation sequencing reads were trimmed using trimgalore (version 0.3.1, https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/) with stringency 2 and aligned with bowtie2 (version 2.1.0, (Langmead and Salzberg 2012)) to a custom genome based on mm9 for the CTCF insertion models or mm10 for models used in the AEKO study. Samtools (version 1.3, (Li et al. 2009)) was used to remove PCR duplicates and bedgraphs were created using bamCoverage (part of deeptools version 2.2.2, (Ramírez et al. 2016)), with normalisation set to RPKM. Replicates were merged with unionbedg (bedtools version 2.25.0, (Quinlan and Hall 2010)) and bigwig tracks were created with bedGraphToBigWig (deeptools). To compare bigwig files from different genotypes in the CTCF insertion project, tracks were shifted to align over genes and other genomic features.

Some samples in the AEKO chapter (ATAC-seq and Pol II ChIP) were analysed using an in-house pipeline developed by Jelena Telenius (Telenius et al. 2018).

Finally, data was visualised using the UCSC genome browser (Kent et al. 2002).

Analysis of Capture-C data

Data was analysed as previously described (Davies et al. 2016) using scripts available at https://github.com/Hughes-Genome-Group/CCseqBasicS. Data visualisation and statistical analysis was performed using CaptureSee (Telenius et al. 2020).

Analysis of RNA-seq data

Trimgalore (version 0.3.1, https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/) was used to trim reads. Reads were then aligned to mm10 using STAR (2.4.2a, (Dobin et al. 2013)). Read counts over genes were generated using featureCounts of the R package subread (version 1.6.2). Differential expression analysis was performed using edgeR (Robinson et al. 2009) and adjusted p-value was calculated with p.adjust, method "BH".

Read count analysis and peak calling

Reads were counted over the ectopic CTCF site (chr11:32,171,303-32,172,502) for every sample (aligned to CTCFins custom genome). For normalisation, peaks were called with Lanceotron (Hentges et al. 2021), using model wide-and-deep_jan-2021, on bigWig files generated using bamCoverage (part of deeptools 3.4.3, (Ramírez et al. 2016)) with extendReads (default), binsize 1 and normalisation set to RPKM, after aligning to mm9. Peaks with a peak score between 0.8 and 1 were used for further analysis. The peak call files were merged with replicates in the same cell type and a peak call list was created by merging the peaks and only preserving the peaks that were present in at least half of the replicates (4 replicates for EB Rad21 ChIP, 6 replicates for spleen Rad21 ChIP). Next, bedtools multicov (bedtools v2.29.2 (Quinlan and Hall 2010)) was used to count reads over the peak call list for every sample (aligned to mm9) and the sum of those reads was used to normalise read counts over the ectopic CTCF site and other sites of interest.

AEKO ATAC-seq tracks were normalised in a similar fashion. First peaks were called using macs2 (Zhang et al. 2008) using -q 0.1 as a cut off. Then peaks present in more than one replicate were used to create a peak call list. Bedtools multicov (Quinlan and Hall 2010) was used to count the

number of reads over peaks for every sample, and this was then used to normalise the ATAC-seq tracks. The Ensembl annotation of promoters was used to calculate reads over promoters using featureCounts.

Previously published and publicly available datasets used in this thesis

Dataset	**Tissue**	**GEO number if available**	
ATAC-seq	Mouse ES cells (E14TG2a)	GSE184435	(Francis et al. 2022)
Rad21 ChIP-seq	Mouse ES cells (E14TG2a)	GSE144336	(Hua et al. 2021)
CTCF ChIP	CD71+ EB	GSE184435	(Francis et al. 2022)
ATAC-seq	CD71+ EB	GSE184435	(Francis et al. 2022)
DNase-seq	Ter119+ Spleen	GSE86393	(Schwessinger et al. 2017)
KLF1 ChIP-seq	Ter119+ Spleen		(Hay et al. 2016)
NIPBL ChIP-seq	Ter119+ Spleen	GSE144336	(Hua et al. 2021)
WAPL ChIP-seq	Quiescent mouse embryonic fibroblasts	GSE76303	(Busslinger et al. 2017)
WAPL ChIP-seq	Mouse ES cells (E14TG2a)	GSE135180	(Liu et al. 2021)

Chapter 3: Insertion of a CTCF binding site reduces enhancer-promoter interactions in an orientation-dependent manner

Introduction

Previous characterisation of CTCF binding sites at the alpha-globin locus

The mouse alpha-globin locus is flanked by several CTCF binding site that interact with each other across the locus, but infrequently interact with the enhancers and alpha-globin genes in between (Hanssen et al. 2017). Motif analysis shows that the upstream binding sites (HS-94, HS-71, HS-59, HS-39 and HS-29; see Figure 3.1a) are in forward orientation. This means they are positioned to block cohesin and thereby loop extrusion coming from the alpha-globin locus. Conversely, the downstream CTCF binding sites (θ1, θ2, HS+44, HS+48 and HS+65) are in reverse orientation, also oriented to block loop extrusion complexes originating from the alpha-globin locus. The CTCF sites are thus in convergent orientation across the locus, a configuration which is often seen across insulated interaction domains (Rao et al. 2014). The exceptions are HS+66, which is at the edge of the interaction domain and positioned facing outward, and more notably HS-38. The reverse orientation of the HS-38 binding site, towards the upstream genes, means it is primarily oriented to block loop extrusion and interactions from upstream of the alpha-globin.

Previous work has used knockout mouse models to interrogate the function of several CTCF binding sites at the alpha-globin locus. The knockout of the upstream HS-39, HS-38 and HS-29 sites, as well as the combined knockout of HS-39 and HS-38, did not affect the expression of alpha-globin (Figure 3.1) or the 3C interaction profile of the alpha-globin enhancers with the alpha-globin genes in definitive erythroid cells (Hanssen et al. 2017), showing that these CTCF binding sites are not necessary for the promoter-enhancer interactions between the alpha-globin genes and enhancers. However, in all of these mouse models, expression of one or more upstream genes increased, revealing that these CTCF binding sites play an important role in insulating the alpha-globin enhancers from nearby non-erythroid genes. The effect was

strongest in the double knockout (HS-38 and HS-39), with a substantial increase in expression of *Mpg*, *Rhbdf1* and *Snrnp25* (Figure 3.1c). These three genes were also significantly upregulated in the HS-38 knockout, albeit to a lower extent (Figure 3.1d). The knockout of HS-39 or HS-29 had only a mild effect on gene expression, with some upregulation of *Rhbdf1* (Figure 3.1d and e). These knockout models thus revealed that the CTCF binding sites at the 5' of the locus restrict interactions between the alpha-globin enhancers and the flanking genes, in a redundant fashion. It also shows that these sites are not all functionally equivalent, with HS-38 having the greatest influence on gene expression.

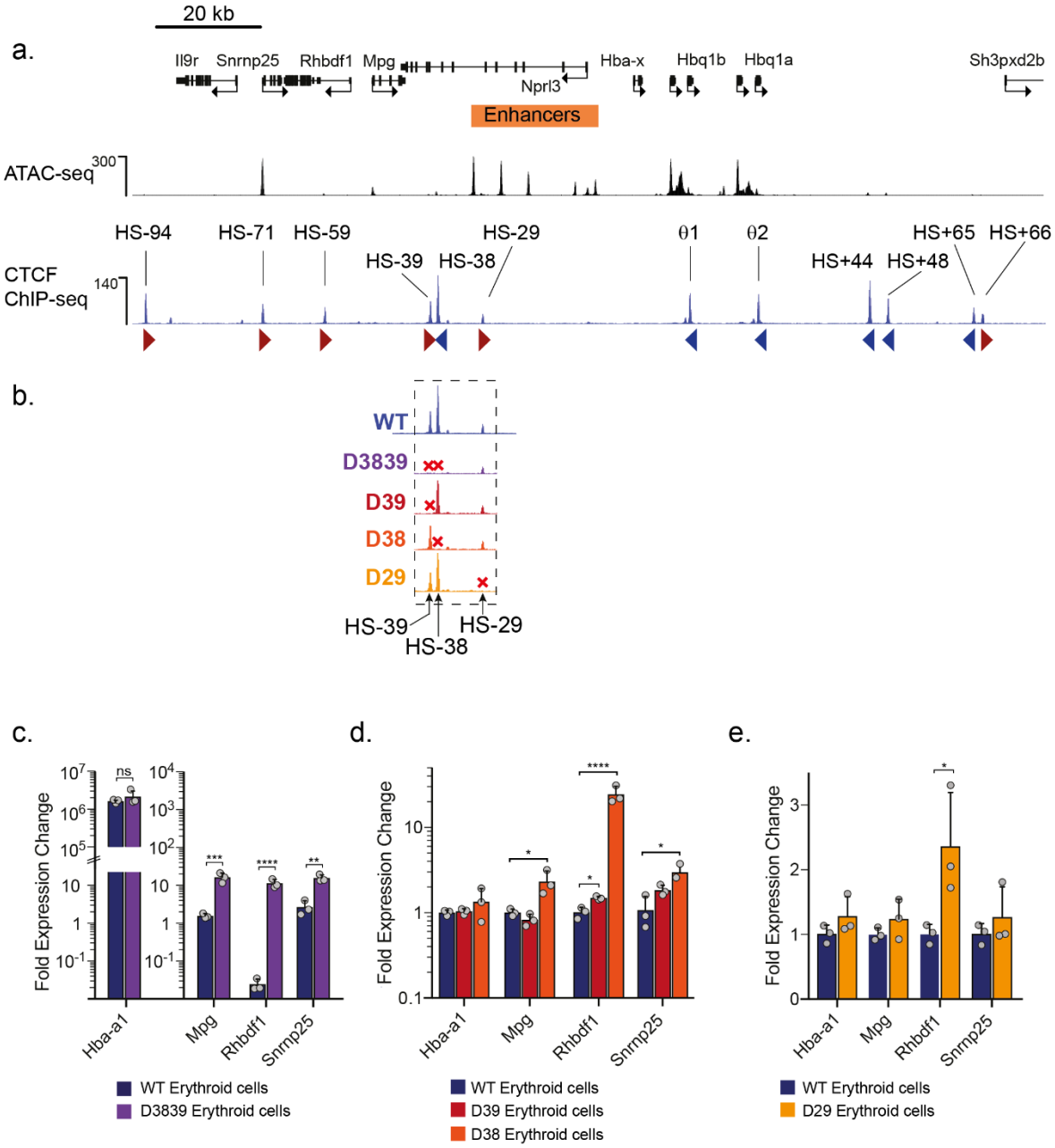


Figure 3.1: A knockout study reveals the upstream CTCF binding sites function to restrict the alpha-globin enhancers

*a) Overview of the alpha-globin locus, showing ATAC-seq and CTCF ChIP-seq tracks in Ter119+ spleen-derived erythroid cells. The orientation of the CTFC binding sites is indicated by forward arrows (red) and reverse arrows (blue). b) CTCF ChIP-seq tracks showing the CTCF knockout models created by Hanssen et al. to study the role of three CTCF sites at the 5' end of the sub-TAD. c) Fold change relative to ES cell expression in WT and D3839 erythroid cells as measured by qPCR. Data represent n=3 independent experiments in which 3 animals were analysed in total. Bars represent the mean and the error bars represent the standard deviation, while grey dots represent individual data points. d) Relative gene expression changes in D38 and D39 compared to WT, data represented as in c). e) Relative gene expression changes in D29 compared to WT, data represented as in c). P values were obtained using an unpaired, two-tailed Student's t-test. NS, P>0.05; *P<0.05; **P<0.01; ***P<0.001; **** P<0.0001. Figure adapted from Hanssen et al. (Hanssen et al. 2017).*

The effect of deleting the downstream CTCF binding sites has also been studied (Susannah Holliman, unpublished; (Harrold et al. 2020)). The two theta-globin associated CTCF sites were deleted either individually or in combination, and the HS+44 and HS+48 further downstream were deleted together (Figure 3.2a). A model in which all of these four downstream CTCF sites were deleted was also generated and called the Quad deletion model. Capture-C analysis revealed changes in the 3D conformation in accordance with loop extrusion, as enhancer interactions with regions downstream of the deleted CTCF sites increased. In wild-type erythroid cells, *Hba-a1* is more highly expressed than *Hba*-a2, and this differential expression has been speculated to be regulated by the theta-globin CTCF sites. However, RNA-seq analysis revealed that the expression ratio of the two alpha-globin genes remained the same when the theta-globin CTCF sites are deleted (Figure 3.2b). In the deletion of the HS+44 and HS+48, there was no change in alpha-globin expression, nor were the downstream genes (*Sh3pxd2b* and *Ubtd2*) upregulated (Figure 3.2c). This was also the case when all four CTCF sites were deleted (Figure 3.2d). In summary, the deletion of any of the downstream CTCF sites does not lead to changes in gene expression.

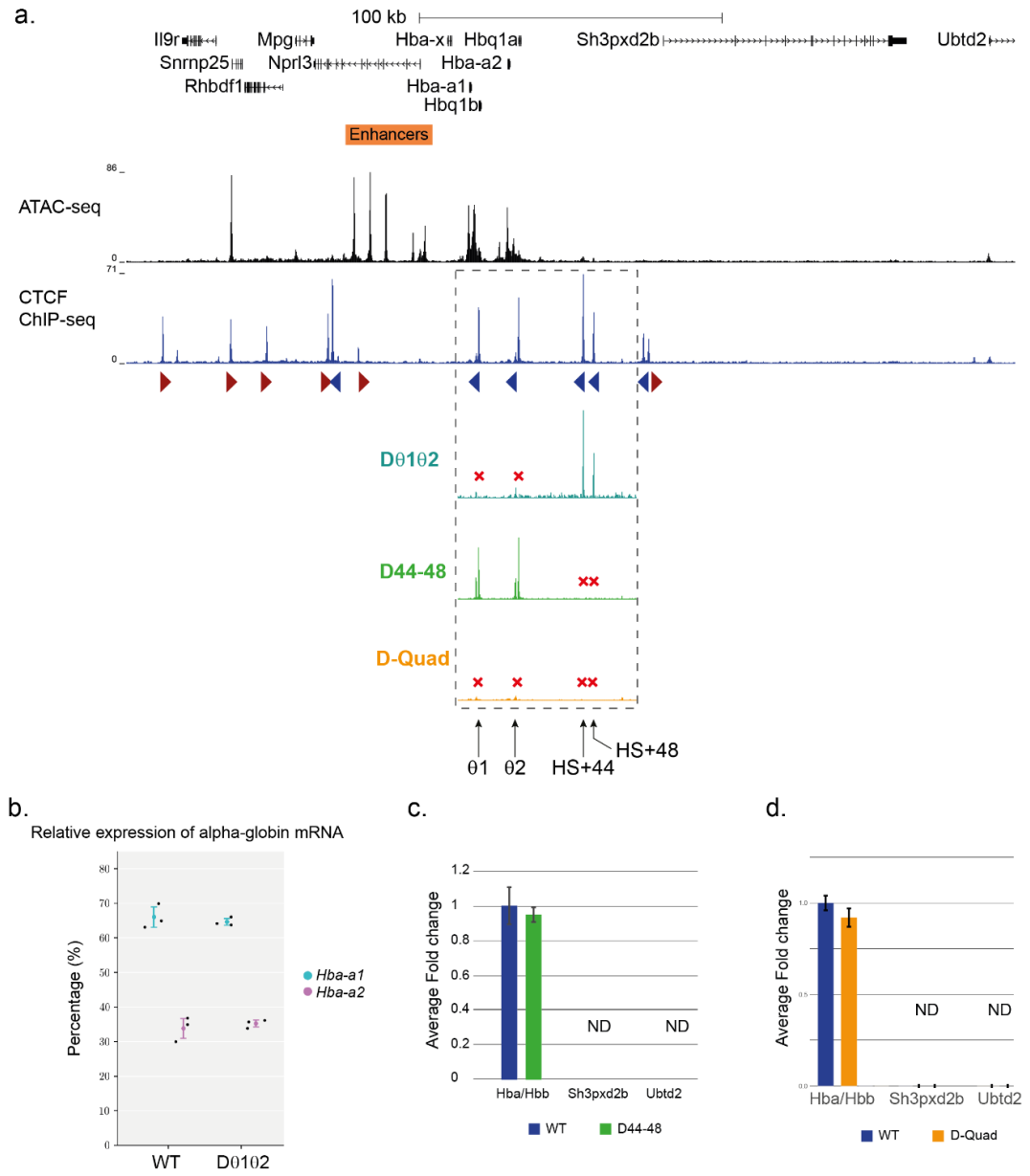


Figure 3.2: Characterisation of downstream CTCF binding sites

a) Overview of the alpha-globin locus, showing ATAC-seq and CTCF ChIP-seq tracks in spleen-derived erythroid cells, and the some of the models created to study the downstream CTCF binding sites. The theta-globin-associated CTCF sites deletion (Dθ1θ2) and the HS+44/HS+48 deletion (D44-48) were characterised by Caroline Harrold and the deletion of all four CTCF binding sites (D-Quad) was analysed by Susannah Holliman. b) Poly(A)+ RNA-seq analysis in the theta-deletion model shows that the higher expression of Hba-a1 remains after the deletion of the CTCF site between the two alpha-globin genes. The presence of a single SNP in the third exon allows for the reads to be analysed by variant calling and attributed to either Hba-a1 or Hba-a2 (Harrold et al. 2020). c) Expression of alpha-globin normalised to beta-globin, normalised to wild type, and of Sh3pxd2b and Ubtd2 normalised to RN18s in wild-type and D44-48 erythroid cells. Mean and standard deviation of three biological replicates (Harrold et al. 2020). d) Expression of alpha-globin normalised to beta-globin, normalised to wild type, and of Sh3pxd2b and Ubtd2 normalised to RN18s in wild-type and D-Quad erythroid cells (Susannah Holliman, unpublished).

Not all CTCF binding sites are functional as an insulator

The dissection of CTCF binding sites at the alpha-globin locus reveals that not all binding sites have the same effect on the surrounding chromatin environment: some are much more effective as insulators than others. Indeed, it has been observed that not all CTCF binding sites are equal in several studies (Barutcu et al. 2018; Huang et al. 2021). Some of the observed variety is likely due to binding site sequence, as the sites have diverse binding motifs, and the presence of the upstream motif in addition to the core motif is known to increase binding strength and insulation (Nakahashi et al. 2013; Jia et al. 2020). Other factors are likely to affect a CTCF binding site's ability to function as an insulator as well, and it has for instance been suggested that the presence of multiple CTCF binding sites is beneficial or even necessary for their insulator function (Jia et al. 2020; Huang et al. 2021). However, much remains unclear about the interplay between CTCF binding site sequence, its local chromatin environment and its position relative to other genomic elements.

Furthermore, CTCF's insulator function is derived from its ability to block loop extrusion, and thus loop extrusion dynamics are likely to influence its insulator function. Since CTCF predominantly but not exclusively blocks cohesin coming from one specific direction, the site as well as the level of cohesin loading are likely to influence how CTCF shapes 3D chromatin interactions. The interplay between CTCF binding and loop extrusion might explain why chromatin interactions can be very different at loci in different stages of differentiation, while the binding of CTCF is invariant, such as at the alpha-globin locus (Hua et al. 2021).

Ectopic insertions of CTCF binding sites have helped answer some of the questions with regard to CTCF binding site function. However, many of these insertion studies have inserted large fragments (i.e. 1 to 4 kb (Hou et al. 2008; Huang et al. 2021)) and therefore cannot differentiate the effect of the CTCF binding site itself from the contribution of local sequence context. Furthermore, many studies have not fully exploited the potential of ectopic CTCF insertions in better understanding cohesin loading dynamics, by only inserting the CTCF site in a single

orientation (Hou et al. 2008). Finally, many studies have concluded that a single CTCF site is not sufficient, or have mainly focussed on multiple CTCF sites (Jia et al. 2020; Huang et al. 2021), which makes it complicated to draw conclusion about contribution of a single CTCF site and its sequence contribution.

In this chapter, I describe an insertion of HS-38, the CTCF binding site within the alpha-globin locus with the highest insulating potential (Hanssen et al. 2017; Harrold et al. 2020). It is inserted between the alpha-globin genes and enhancers in both orientations. I then show that CTCF is effectively recruited to the ectopic CTCF site in both the forward and the reverse orientation. Furthermore, Capture-C analysis shows that the 3D conformation of the locus is altered upon introduction of this new site, and the change in 3D conformation depends on the orientation of the new binding site. Finally, Rad21 ChIP-seq revealed that cohesin is present at the ectopic CTCF site only when it is inserted in the reverse orientation.

Results

Regions of the genome that are active are often devoid of nucleosomes and thus more susceptible to digestion by DNase I, and are therefore marked by DNase I accessibility assays (Galas and Schmitz 1978; Wu et al. 1979; Parslow and Granner 1983; Neph et al. 2012). High sequence depth DNase-seq data reveals footprints at sites protected by stable protein binding, such as binding of CTCF (Schwessinger et al. 2017). At the HS-38 CTCF binding site, DNase-seq data shows two footprints (Figure 3.3), one for the 20 bp core motif and a smaller one for the 10 bp upstream motif (Nakahashi et al. 2013). There is a 5 bp spacer motif between these two, making the total length of the motif 35 bp. The CTCF binding site's orientation is reverse, and can be confidently determined because of the presence of the upstream motif.

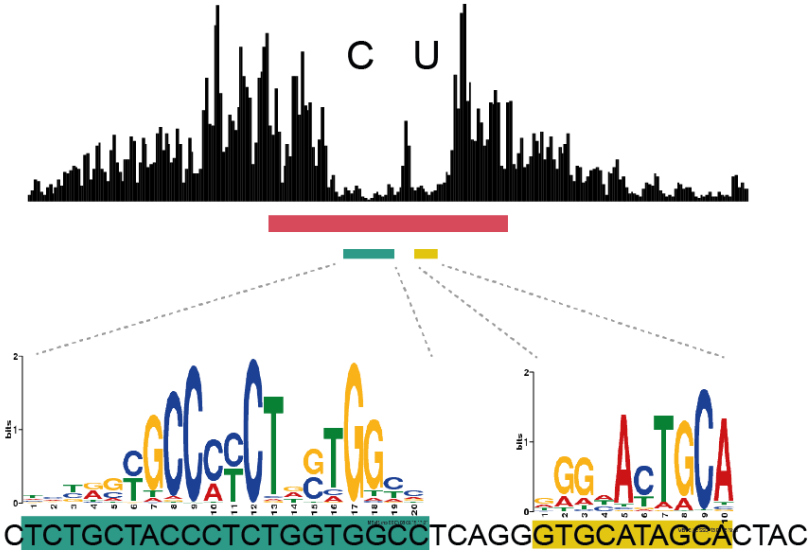


Figure 3.3: DNase I hypersensitivity of the HS-38 CTCF binding site
DNase I footprint at the HS-38 CTCF binding site and a comparison of the DNA sequence with the CTCF consensus motif (green) and upstream motif (yellow). Figure adapted from Hanssen et al. (Hanssen et al. 2017). The sequence selected to be inserted ectopically is shown in red.

Introducing a CTCF site between the alpha-globin genes and their enhancer region

To test if the HS-38 boundary sequence is functional outside of its native context, and if it can interfere with the strong enhancer-promoter interactions at the alpha-globin locus, I aimed to introduce it between the alpha-globin genes and their enhancers. Taking into account the design of guide RNAs and the position with regards to other genomic elements, I decided to insert the CTCF binding site 5 kb from the embryonic zeta-globin promoter and 3 kb from the R4 enhancer element (Figure 3.4a). This places the CTCF site at the centre of the sub-TAD (Figure 3.4a, yellow bar), where the orientation of the insertion will determine if it is primarily able to block cohesin tracking from the direction of the enhancers or from the alpha-globin genes. In this position, the newly inserted copy of the HS-38 CTCF site is flanked by the usual convergently oriented CTCF sites surrounding the alpha-globin cluster, and is therefore poised to be drawn together with either the 5' boundary or the 3' boundary of the locus by loop extrusion.

As a CTCF binding site introduced at this position might function in both forward and reverse orientation, I designed a construct that allowed for inversion of the inserted sequence using

heterospecific lox sites (Figure 3.4b; (Lee and Saito 1998)). I also included 24 bp of native sequences on either side to ensure that these flanking lox sites do not interfere with the 35 bp CTCF binding site. I thus selected an 83 bp fragment to be introduced (Figure 3.3 and 3.4b). Finally, the construct also contained a neomycin resistance cassette to aid selection of cells with the insertion. This was subsequently removed by transfection with Flp recombinase.

This construct was used for homology-directed repair in combination with a nickase Cas9 strategy with two guide RNAs to decrease the probability of off-target editing (Ran et al. 2013). I generated two mouse ES cell lines: the "Forward" model with the binding site orientation convergent with downstream CTCF binding sites and orientated towards the alpha-globin genes, and the "Reverse" model in its native orientation, convergent with upstream CTCF sites and oriented towards the enhancer cluster (see Figure 3.4b). I selected three independently-generated clones of each genotype to characterise. We also used these ES cells to generate two mouse lines bearing the inserted CTCF sites.

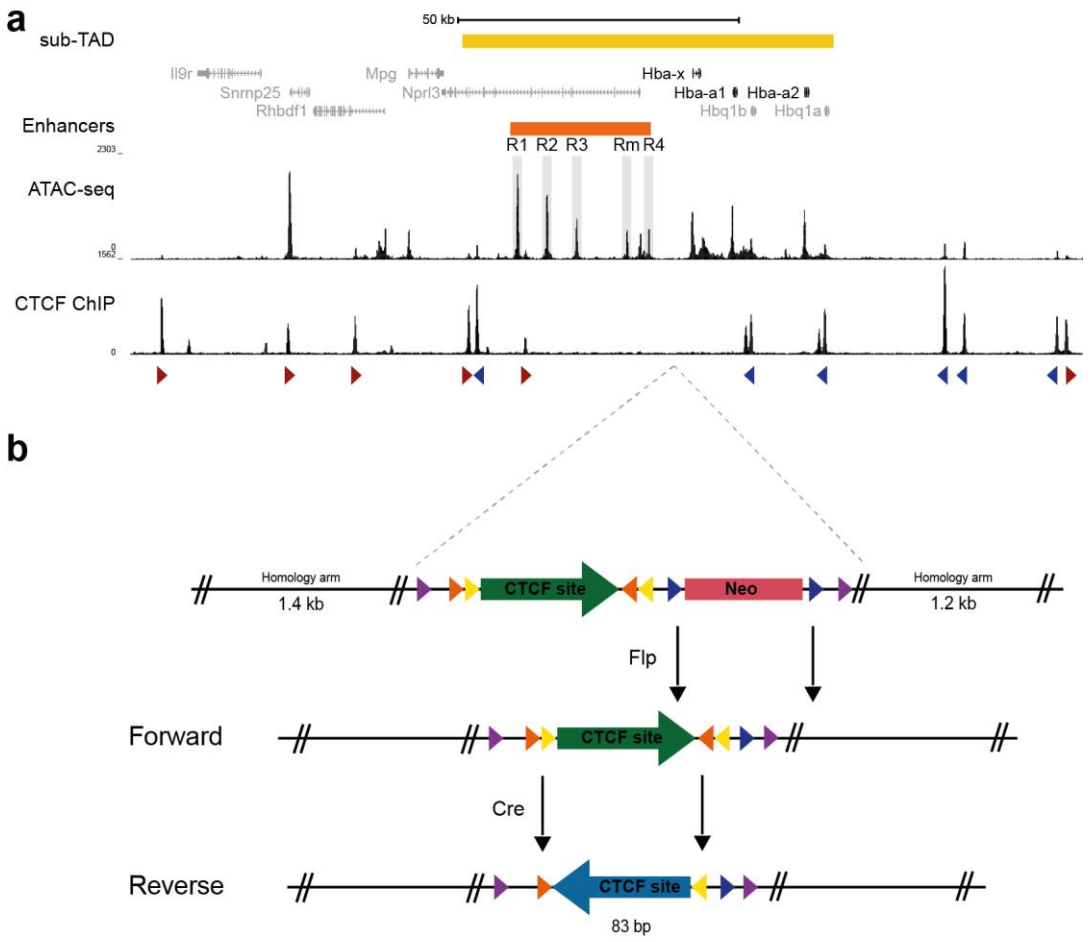


Figure 3.4: Strategy for the introduction of the CTCF binding site at the alpha-globin locus and subsequent genetic manipulations

a) The CTCF site was introduced in between the alpha-globin enhancers and genes, 5 kb upstream of the zeta-globin promoter. ATAC-seq and CTCF ChIP track from EB-derived erythroid cells. b) The construct used for the creation of the Forward and Reverse CTCF insertion model. An 83 bp CTCF site (green for forward orientation, blue for reverse orientation) is flanked by lox2272 (red triangles) and loxP (yellow triangles). A neomycin resistance cassette (Neo, red) is flanked by Frt sites (blue triangles). The whole construct can be removed by flanking rox sites (purple triangles). After successfully introducing the fragment, transfection with Flp removed the neomycin resistance cassette and created the Forward model, and following transfection with Cre the CTCF binding site orientation was reversed to create the Reverse model.

The newly inserted CTCF binding site recruits CTCF

The recruitment of CTCF to the new binding site was assessed in both the Forward and Reverse model using *in vivo* derived definitive erythroid foetal liver cells and *in vitro* derived primitive-like erythroid cells from day 7 EBs (Francis et al. 2022). ChIP-seq for CTCF revealed that the newly introduced site recruited high levels of CTCF in both cell types, and in both orientations of the CTCF binding site (Figure 3.5).

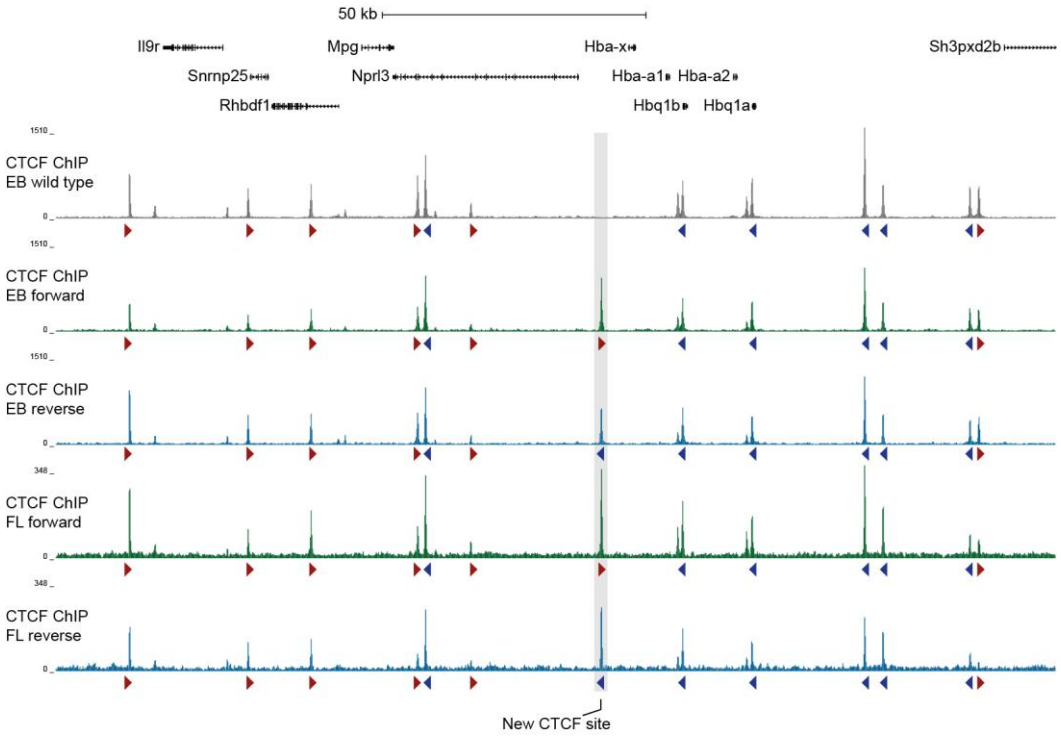


Figure 3.5: CTCF ChIP-seq confirms recruitment of CTCF to the newly introduced binding site
Normalised (RPKM) CTCF ChIP-seq tracks in WT (grey), Forward (green) and Reverse (blue) day 7 EB-derived erythroid cells (EB, n=2 for each) and Forward and Reverse foetal liver-derived erythroid cells (FL, n=3 for each). CTCF binding site orientation is indicated by forward arrows (red) and reverse arrows (blue). The insertion site is highlighted in grey.

The new CTCF binding site changes 3D genome organisation

To test if the new CTCF binding site changes the 3D organisation at the alpha-globin locus, I performed Capture-C in erythroid cells isolated from day 7 EBs. Capture-C can reveal high-resolution chromatin interactions from viewpoints of interest at a locus (Hughes et al. 2014; Davies et al. 2016). First, I examined the interactions from the two main enhancer elements, R1 and R2. In wild-type cells, both enhancer elements are in close proximity to the alpha-globin genes (Figure 3.6, grey line). In both the Forward and Reverse models, there is a reduction of interactions between enhancers and the alpha-globin genes, as well as the zeta-globin gene. When comparing the interaction profiles from R2 in the Forward and Reverse model with each other (Figure 3.9a), the interactions are more reduced in the Reverse model than the Forward model, particularly over the adult alpha-globin genes.

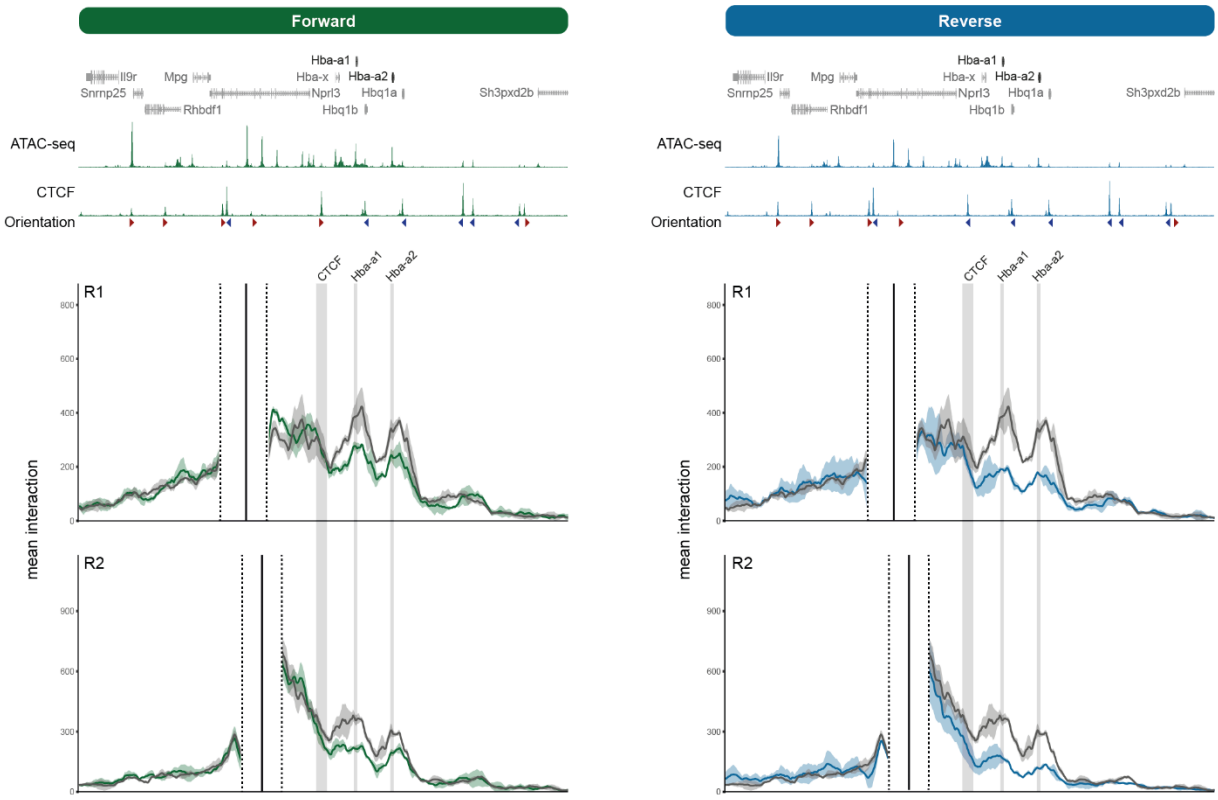


Figure 3.6: Capture-C analysis using two enhancer elements as a viewpoint

Top tracks show ATAC-seq and CTCF ChIP-seq in the Forward or Reverse EB model. The middle panels show capture profiles from the viewpoint of R1 and the lower panels show the viewpoint R2. The profile shows normalised and averaged interactions with a halo representing one standard deviation from biological replicates, for WT (grey, n=3), Forward (Fwd, green, n=3) and Reverse (Rev, blue, n=2) EB-derived erythroid cells. The position of the alpha-globin genes and the ectopic CTCF site is indicated in grey. The exclusion zones around capture probes are marked by dashed black lines and the position of the viewpoint by a solid black line.

In 3C-based methods, CTCF binding sites are often observed to form specific interactions with each other. I therefore selected downstream and upstream CTCF binding sites to further analyse changes in chromatin interactions. HS-38 and HS-39 are two CTCF binding sites upstream of the alpha-globin enhancers, located 38 and 39.5 kb respectively from the zeta-globin transcription start site (TSS). The binding sites are in convergent orientation, so that HS-39 is oriented towards the enhancers and genes and poised to mainly block loop extrusion from that direction, whereas HS-38 is in the reverse orientation, pointed away from the alpha-globin enhancers. Both sites show very similar interaction profiles in Capture-C, and, given that they are only separated by

1.5 kb, this could reflect physical constraints of the DNA fibre, the limited resolution of Capture-C, or both.

HS-38 and HS-39 both interact with the theta-globin-associated CTCF site at the 3' end of the alpha-globin genes (θ2), and to a lesser extent with the reverse oriented CTCF sites (HS+44 and HS+48) further downstream (Figure 3.7). Interactions with the enhancer region are very infrequent for both CTCF binding sites. In the Forward model, HS-38's interactions with the theta-globin CTCF sites (θ1 and θ2) are slightly reduced, while the interactions of HS-39 with the theta-globin CTCF sites are more strongly reduced (Figure 3.7). In the Reverse model, the interactions between both CTCF sites and the theta-globin CTCF sites are even more reduced, and there is also a reduction in interactions with the HS+44/HS+48 CTCF sites further downstream. Strikingly, a new interaction forms between the inserted CTCF binding site and both HS-38 and HS-39 only in the reverse orientation.

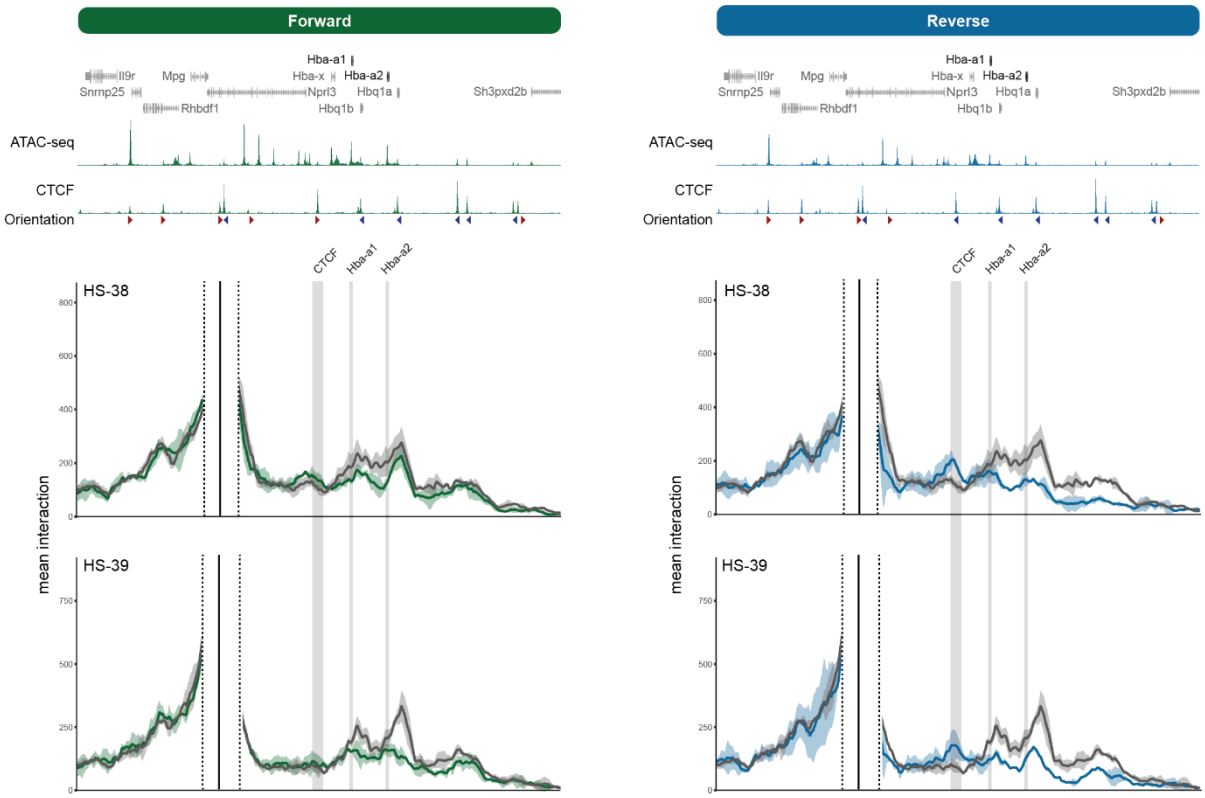


Figure 3.7: Capture-C analysis using two upstream CTCF binding sites as a viewpoint
The top tracks show ATAC-seq and CTCF ChIP-seq in the Forward or Reverse EB model. The middle panels show capture profiles from the viewpoint of HS-38 and the lower panels show the viewpoint HS-39. The profile shows normalised and averaged interactions with a halo representing one standard deviation from biological replicates, for WT (grey, n=3), Forward (Fwd, green, n=3) and Reverse (Rev, blue, n=2) EB-derived erythroid cells. The position of the alpha-globin genes and the ectopic CTCF site is indicated in grey. The exclusion zones around capture probes are marked by dashed black lines and the position of the viewpoint by a solid black line.

In wild-type cells, the HS+44 and HS+48 CTCF binding sites, which are about 44 and 48 kb downstream of the zeta-globin TSS respectively, frequently interact with the CTCF binding sites upstream of the sub-TAD (Figure 3.8). As they are in the reverse orientation, they are poised to block cohesin coming from the upstream genes and enhancer region. In addition to the interactions with CTCF sites across the locus, they also form interactions with the theta-globin-associated CTCF binding sites. In the Forward model, these interactions with the theta-globin sites are greatly increased. Furthermore, a new interaction is formed with the inserted forward CTCF site. Interactions with sites across the locus remain largely intact. In the Reverse model, interactions with the theta-globin CTCF sites remains the same except for a small decrease in

interactions between HS+44 and θ1. However, the interactions between both HS+44 and HS+48 and the 5' CTCF sites, upstream of the enhancer region, are very strongly reduced.

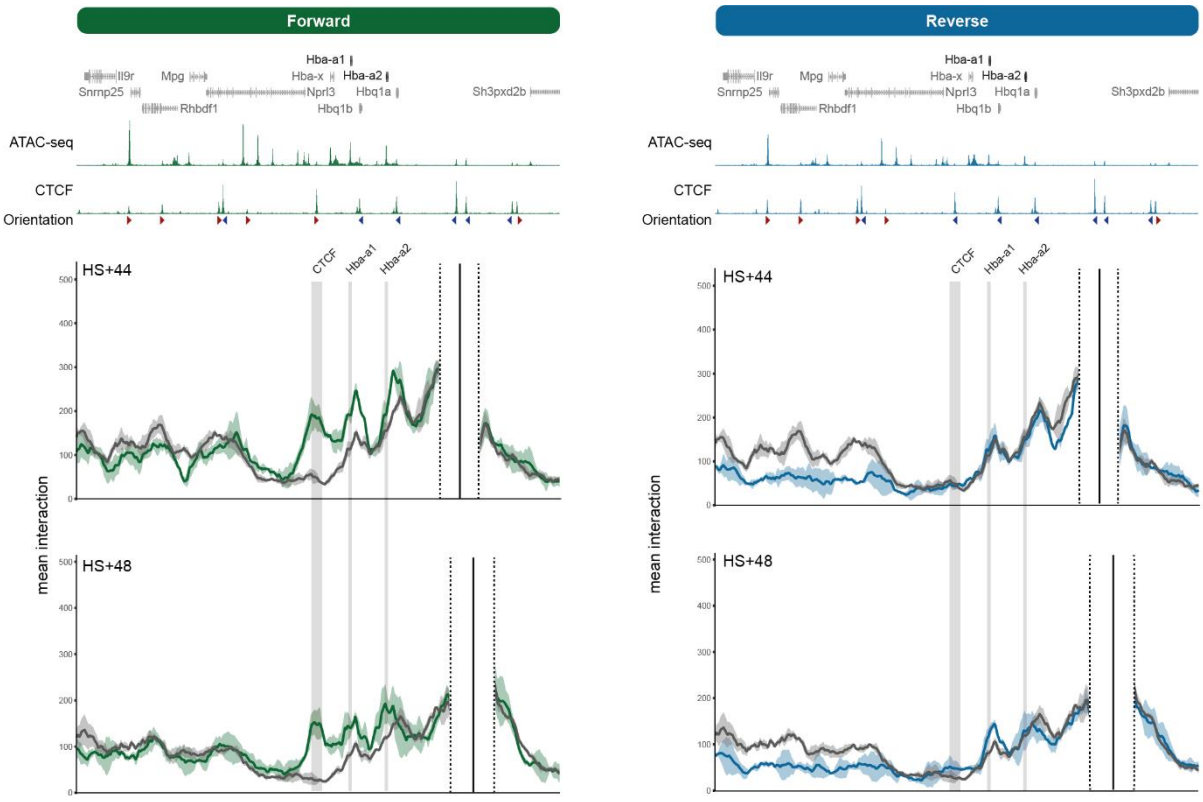


Figure 3.8: Capture-C analysis using two downstream CTCF sites as a viewpoint
Top tracks show ATAC-seq and CTCF ChIP-seq in the Forward or Reverse EB model. Middle panels show capture profiles from the HS+44 viewpoint and lower panels show the HS+48 viewpoint. The profile shows normalised and averaged interactions with a halo representing one standard deviation from biological replicates, for WT (grey, n=3), Forward (Fwd, green, n=3) and Reverse (Rev, blue, n=2) EB-derived erythroid cells. The position of the alpha-globin genes and the ectopic CTCF site is indicated in grey. The exclusion zones around capture probes are marked by dashed black lines and the position of the viewpoint by a solid black line.

When comparing the interaction profile of the HS+44 viewpoint between the Forward and Reverse model (Figure 3.9), the interaction with the new CTCF site is clearly specific to the Forward model, as is the increase in interactions with the theta-globin-associated CTCF sites. However, the decrease in interactions with the CTCF sites upstream of the enhancers is unique to the Reverse model.

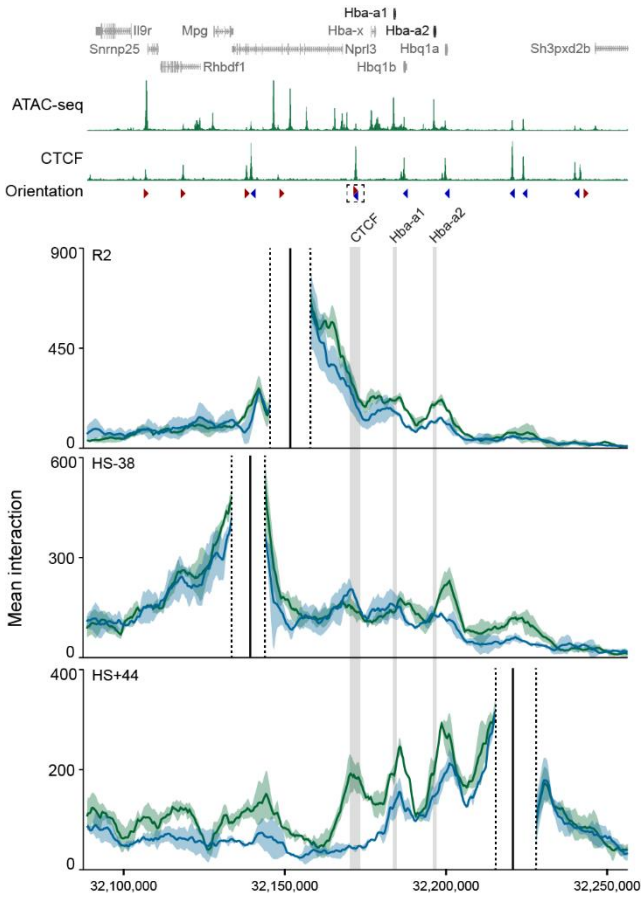


Figure 3.9: Comparison between Capture-C interaction plots from Forward and Reverse model

The top tracks show ATAC-seq and CTCF ChIP-seq in the Forward model. The three lower panels show the comparison of interaction plots in the Forward and Reverse model: The top panel shows the enhancer R2 viewpoint, the middle panel shows the HS-38 CTCF site and the bottom panel the HS+44 CTCF site. The profile shows normalised and averaged interactions with a halo representing one standard deviation from biological replicates, Forward (Fwd, green, n=3) and Reverse (Rev, blue, n=2) EB-derived erythroid cells. The position of the alpha-globin genes and the ectopic CTCF site is indicated in grey. The exclusion zones around capture probes are marked by dashed black lines and the position of the viewpoint by a solid black line.

Rad21 accumulates at the ectopic CTCF site in the Reverse model

There are striking changes in the 3D conformation of the alpha-globin locus upon the introduction of an ectopic CTCF binding site. However, the precise nature of these changes is highly dependent on the orientation. To better understand the mechanism behind this orientation-dependence, I performed ChIP for Rad21, which is a component of the loop extrusion complex cohesin. Usually, cohesin is found to colocalise with CTCF sites, which is the case at the alpha-globin locus as well (Figure 3.10a). However, cohesin is not present at every

CTCF binding site nor are the levels of cohesin equal or directly proportional to the CTCF occupancy. Furthermore, cohesin is also found at other genomic features such as promoters and enhancers (Figure 3.10a). In the Forward model, there is not much cohesin present at the new CTCF site, but in the Reverse model there is a clear cohesin peak in both the EB-derived and spleen-derived erythroid cells. Furthermore, there appears to be a new peak over the *Rhbdf1* promoter in the Reverse spleen cells, and the peaks over the HS+44/HS+48 CTCF site appear reduced. Unfortunately, I only managed to obtain one replicate per genotype and tissue type, and Rad21 ChIPs are variable, so this limits how confidently I can draw these conclusions.

I quantified the reads in the Rad21 ChIP over the newly inserted site (Figure 3.10b), which showed a strong increase in read coverage the Reverse orientation and a small increase in the Forward model. However, due to a lack of replicates, it is not clear if these differences are significant. When I performed the same analysis on the CTCF ChIP, I did not see a difference between the Forward and Reverse model (Figure 3.10c).

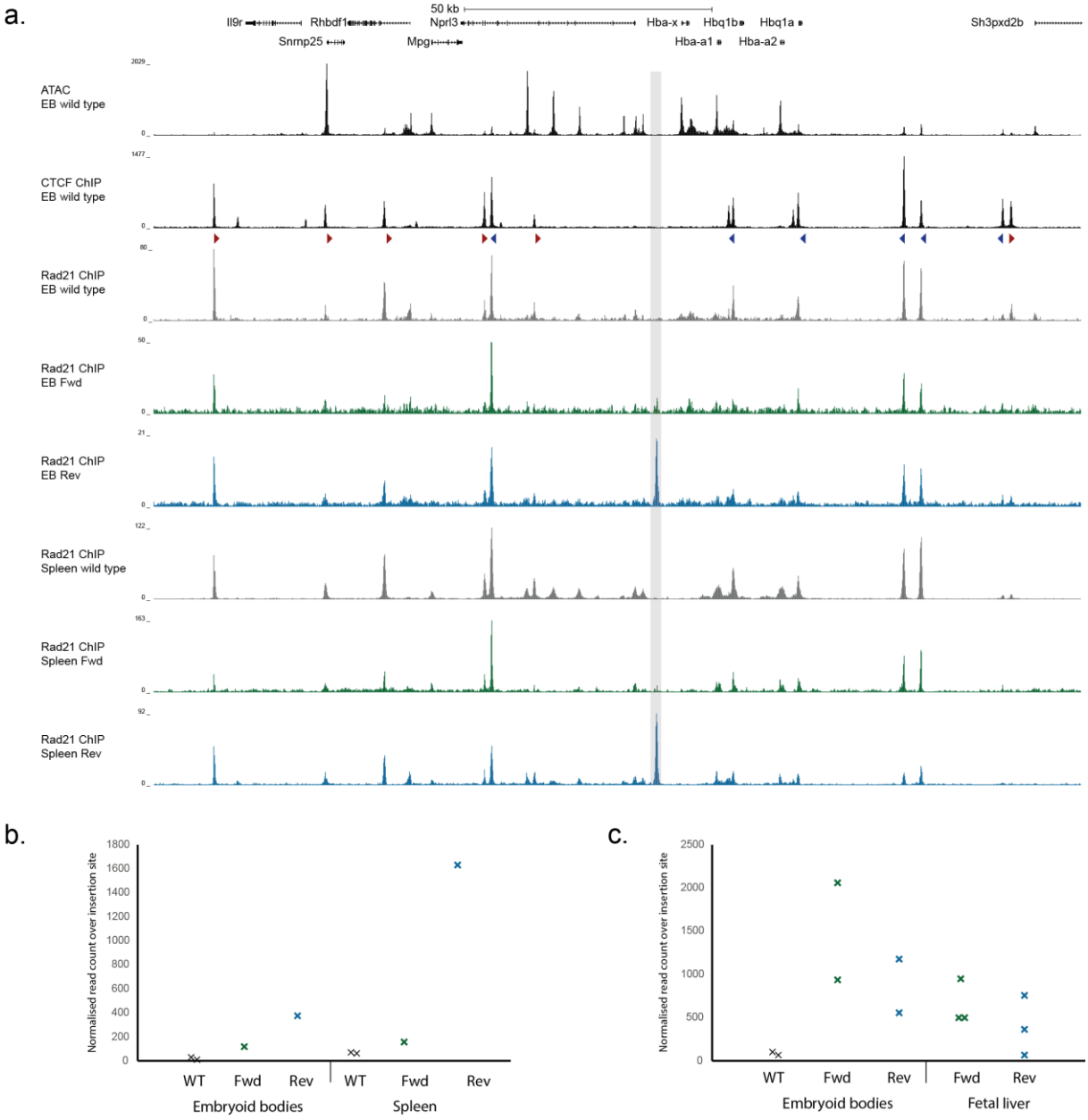


Figure 3.10: Rad21 ChIP-shows accumulation of cohesin at the ectopic site in the Reverse model

a) Top two tracks: ATAC-seq and CTCF ChIP-seq tracks in wild-type EB-derived erythroid cells. RPKM-normalised Rad21 ChIP-seq tracks in WT (grey), Forward (green) and Reverse (blue) in EB-derived erythroid cells (n=1 for all), and in spleen-derived erythroid cells, also WT (grey, n=2), Forward (green, n=1) and Reverse (blue, n=1). The site where the ectopic CTCF site is inserted is highlighted in grey. b) Read counts from Rad21 ChIP-seq over the inserted CTCF site, normalised to total reads over called peaks for each sample. c) Read counts from CTCF ChIP-seq over the inserted CTCF site, normalised to total reads over called peaks for each sample.

Discussion

In this chapter, I have described how I created a mouse ES cell model with an ectopic CTCF binding site at the alpha-globin locus in both orientations, which I named the Forward and Reverse models. I differentiated these ES cells *in vitro* to analyse the effect of the ectopic binding site in EB-derived erythroid cells. Additionally, we used these cells to generate mouse lines bearing the same insertions and examined the effect of the insertion *in vivo* in foetal liver-derived erythroid cells. The single CTCF binding site recruited CTCF equally effectively in both the Forward and Reverse model. This shows that the binding site orientation does not affect CTCF recruitment at this site.

Capture-C revealed that the 3D conformation at the locus was significantly altered in both models, but in different ways, showing that the direction of the CTCF binding site plays a relevant role in establishing the 3D conformation of the locus. This agrees with many other reports (De Wit et al. 2015; Guo et al. 2015; Huang et al. 2021), although others have reported that insulator activity does not depend on orientation (Jia et al. 2020). In the Forward model, the new CTCF site forms strong interactions with the downstream HS+44/HS+48 sites, and these sites also increase their interactions with the nearby theta-globin-associated CTCF sites. Furthermore, when analysed from either the HS-39/HS-38 or the HS+44/HS+48 viewpoints, interactions across the locus are decreased in the Forward model compared to WT, but are not completely absent.

In contrast, in the Reverse model, chromatin interactions across the locus were reduced to a far greater extent. This strong reduction was observed even though only a small new interaction between the upstream CTCF site and the ectopic CTCF site is formed. Enhancer-promoter interactions decrease in both models, but this change was more pronounced in the Reverse model.

The Rad21 ChIP indicates that a substantial amount of cohesin accumulates at the ectopic binding site only in the Reverse model, whereas in the Forward model only low levels of cohesin

are present at the new site. This is surprising, as both models recruit equal levels of CTCF. One explanation for the unequal accumulation of cohesin at the ectopic CTCF site between the two models could be that the ability to stall extruding cohesin complexes depends on the binding site sequence. However, since the binding site sequence and levels of CTCF are so similar, this does not seem like a likely explanation. It seems more plausible that the sites encounter different levels of translocating cohesin, as they are poised to block cohesin coming from different directions.

The dynamics of cohesin loading onto the chromatin fibre remain insufficiently understood (Zuin et al. 2014). It has been proposed that cohesin is primarily loaded at enhancer or super-enhancer sites (Vos et al. 2021), or at both enhancers and promoters (Hua et al. 2021). Enhancers as primary sites of cohesin loading fits with the observation that the ectopic site accumulates more cohesin in the reverse orientation, as it is oriented to block cohesin coming from the enhancer region. Cohesin loading at enhancer regions also fits with the observed difference in 3D conformation in undifferentiated compared to differentiated cells (Hua et al. 2021). CTCF binding at the alpha-globin locus is almost invariant between ES cells and erythroid cells. Levels of cohesin do not differ greatly either, although there are new peaks present at the enhancers and promoters at the active locus. However, the interactions between CTCF sites change drastically between the active and the inactive locus (Hua et al. 2021). Finally, NIPBL, a factor which has been proposed to have a function in loading cohesin onto the chromatin, is found to bind at the alpha-globin enhancers (Hua et al. 2021). However, it is also found at the alpha-globin promoters which are also only active in erythroid cells.

Cohesin loading at the enhancer is also consistent with the persistence of interactions between the HS-38/-39 and HS+44/+48 in the Forward model, as cohesin coming from the enhancers would not be blocked by the forward-oriented ectopic site. However, in the Forward model there are also new interactions between the ectopic binding site and binding sites further

downstream. Therefore, either cohesin loading occurs at a lower level downstream of the inserted CTCF site, potentially at the active promoters, or these interactions are set up by an alternative mechanism. Although studies have shown that CTCF can self-dimerise (Bonchuk et al. 2020), this is not thought to play an important role in interactions between CTCF sites. For example, cohesin depletion studies have shown that chromatin interactions between CTCF sites are lost in the absence of cohesin (Rao et al. 2017). Phase separation may be another driver of chromatin interactions, however it is more likely to generally set up active domains rather than orchestrate interactions between specific CTCF binding sites, especially not in an orientation-dependent manner. Therefore, there is no clear alternative mechanism by which the interactions between the ectopic CTCF site in the Forward model and the downstream CTCF sites are established.

Finally, having shown that the ectopic CTCF site changes the 3D conformation of the locus and decreases enhancer-promoter interactions, it remains important to verify whether the ectopic site acts as a functional boundary, in the sense that it influences the gene expression. In previous studies at the alpha-globin locus, genetic manipulation of the CTCF binding sites frequently changed the 3D conformation, however this did not always result in changes in gene expression.

Acknowledgements

The generation of the mouse model was a collaboration with Benjamin Davies at the Wellcome Centre for Human Genetics (WCHG). Chris Preece (WCHG) and Daniel Biggs (WCHG) performed blastocyst injections and initial genotyping. Transgenic mouse strains were maintained, crossed with Flp and Cre-expressing strains, and genotyped by Jacqueline A Sharpe and Jacqueline A Sloane-Stanley of the Transgenic Facility (WIMM). Foetal livers were dissected and expanded by Sue Butler (WIMM). Caroline Harrold helped with genotyping of the CTCF insertion mice. Damien Downes generated one replicate of the Capture-C libraries.

Chapter 4: Insertion of a single CTCF binding site is sufficient to disrupt enhancer function

Introduction

Enhancers can regulate the expression of their target genes over large genomic distances, yet 3C-based methods have shown that enhancers and promoters far apart on the linear genome are often found in close physical proximity in the 3D genome (Rao et al. 2014; Davies et al. 2016; Furlong and Levine 2018). This has also been observed by imaging methods (Brown et al. 2018; Li et al. 2020a). Thus, physical proximity has long been thought to be an important component of transcriptional regulation by enhancers.

The cohesin complex has been proposed as an important factor in establishing or maintaining enhancer-promoter interactions via loop extrusion (Fudenberg et al. 2016). However, several studies in which cohesin function has been perturbed observed only a small effect on gene expression (Busslinger et al. 2017; Rao et al. 2017; Schwarzer et al. 2017). It is therefore unclear how important loop extrusion is for enhancer-promoter interactions and it has been suggested that at least a subset of short-range interactions do not depend on cohesin (Thiecke et al. 2020), potentially relying on transcription factors and coactivators to establish enhancer-promoter interactions in addition to or instead of cohesin. The mediator complex (Kagey et al. 2010), YY1 (Beagan et al. 2017) and Lbd1 in erythroid cells (Krivega and Dean 2017) have been proposed to function as a structural bridge between enhancers and promoters.

Moreover, recent studies have found examples in which enhancers are not in close physical proximity to their target promoters when regulating their expression (Alexander et al. 2019; Benabdallah et al. 2019). A role for polyADP-ribosylation was suggested in the observed increase in enhancer-promoter distance at the Sonic hedgehog locus upon activation of this gene (Benabdallah et al. 2019). It is unclear if this concerns a subset of enhancers that regulate their

genes via an alternative mechanism. It is also possible that there is only a brief, transient interaction necessary for gene activation, which could be missed by these studies (Dean et al. 2021). Alternatively, regulation by biomolecular condensate formation, in which components of transcriptional machinery are present in high concentrations at the site of active transcription, would not require close proximity (Hnisz et al. 2017). This could also explain why cohesin depletion is not as detrimental as might be expected.

In summary, although proximity of enhancers to their target promoters has long been thought to be central to their activity, it is possible that it is not required at all loci. Despite this, CTCF is able to interfere with gene expression in an orientation and position-dependent manner, which is difficult to reconcile in the absence of loop extrusion or some similar mechanism involving tracking along the chromatin fibre.

In the CTCF insertion models, described in Chapter 3, alpha-globin enhancer-promoter interactions have been disrupted while all other regulatory elements are preserved. Therefore, these models can contribute to our understanding of the interplay between enhancer-promoter interactions and gene expression. In this chapter, I analyse the expression of alpha-globin and find that a single CTCF site is effective at significantly reducing the expression of alpha-globin. I also demonstrate that this approach when combined with a YFP-tagged alpha-globin gene can be used to screen DNA fragments for their potential to interfere with enhancer-promoter interactions.

Results

As described in Chapter 3, the insertion of an ectopic CTCF site between the alpha-globin enhancers and their cognate promoters led to changes in chromatin interactions at the alpha-globin locus. New interactions between the ectopic CTCF site and existing CTCF sites were formed, and the enhancer-promoter interactions were decreased, most notably in the Reverse

model in which the CTCF site is orientated towards the alpha-globin enhancers. Next, I wanted to assess what effect this has on gene expression at the alpha-globin locus.

Expression of alpha-globin is reduced

To assess the effect of the ectopic CTCF binding site, I first determined levels of alpha-globin in erythroid cells from EBs. The expression of alpha-globin is 44% lower in the Forward model compared to wild type, and 86% lower in the Reverse model (Figure 4.1a). I next determined levels of alpha-globin expression in E10.5 primitive embryonic erythroid cells, as these are the primary-tissue equivalents of EB-derived erythroid cells. In this tissue, alpha-globin expression is not significantly altered in the Forward model, but in the Reverse model expression is reduced by 68% (Figure 4.1b).

Next, I assessed alpha-globin expression in definitive erythroid cells, namely in foetal liver-derived and spleen-derived erythroid cells. In foetal liver-derived erythroid cells, the expression of alpha-globin is reduced by 25% in the Forward model, and by 38% in the Reverse model (see Figure 4.1c). In the spleen-derived cells, expression of alpha-globin is 18% and 41% lower in the Forward and Reverse model respectively (Figure 4.1d).

In summary, insertion of a CTCF site reduced alpha-globin expression in almost all tissue types that were analysed, and the reverse orientation of the inserted element had a consistently larger effect. Furthermore, the expression of alpha-globin was usually more strongly affected in primitive erythroid cells, which express embryonic zeta-globin as well as alpha-globin.

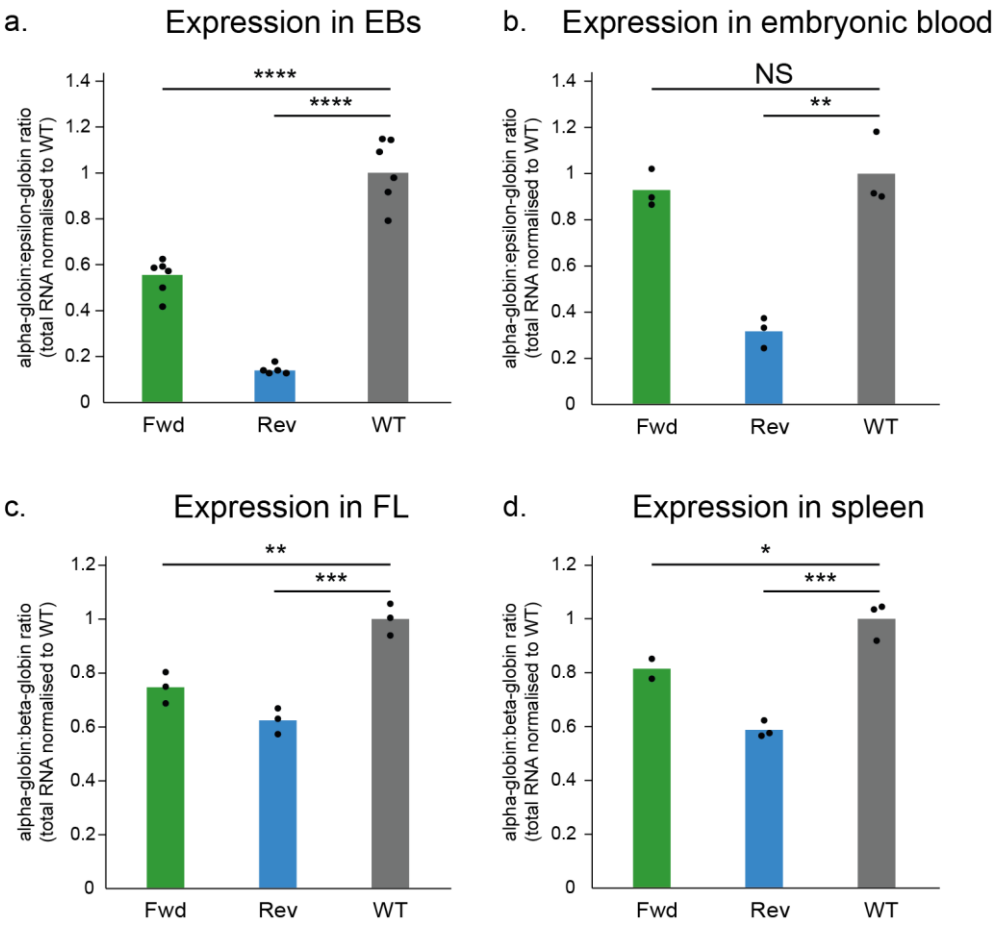


Figure 4.1: Expression of alpha-globin in various erythroid tissues in the Forward and Reverse model

*a) Expression of alpha-globin in EB-derived CD71+ erythroid cells, normalised to the expression of embryonic beta-like globin, epsilon-globin (Hbb-y; n=5 for Rev, n=6 for Fwd and WT). Bar plot shows the mean ratio with individual data points marked by black dots. b) Expression of alpha-globin in E10.5 embryonic blood normalised to epsilon-globin (n=3 for all). c) Expression of alpha-globin normalised to adult beta-globin in cultured E12.5 foetal liver-derived Ter119+ erythroid cells (n=3 for all). d) Expression of alpha-globin normalised to beta-globin in spleen-derived Ter119+ erythroid cells (n=2 for Fwd, n=3 for Rev and WT). P-values were obtained using an unpaired two-tailed Student's t-test. Not significant (NS) p > 0.05, * p <0.05, ** p < 0.01, *** p <0.001, **** p <0.0001.*

Primitive and definitive erythroid cells are similar and both express high levels of alpha-globin (*Hba-a1/2*), but differ in the expression of a few other genes at the alpha-globin locus. Most notably, zeta-globin (*Hba-x*) is expressed in primitive cells but is fully silenced in definitive cells. The expression of zeta-globin is less dependent on the alpha-globin super-enhancer, and is less affected by the knockout of both of the R1 and R2 enhancer elements (King et al. 2021). In addition to the enhancers, its expression is regulated by its promoter and it is actively silenced

in definitive erythroid tissue (King et al. 2021). Consistent with this model of zeta-globin regulation, the CTCF site insertion does not dramatically change zeta-globin expression in either EB-derived erythroid cells, E10.5 embryonic blood, foetal liver-derived erythroid cells or spleen-derived erythroid cells (Figure 4.2). Zeta-globin expression is very low in definitive erythroid cells and remains silenced in both the Forward and the Reverse model (Figure 4.2c and d).

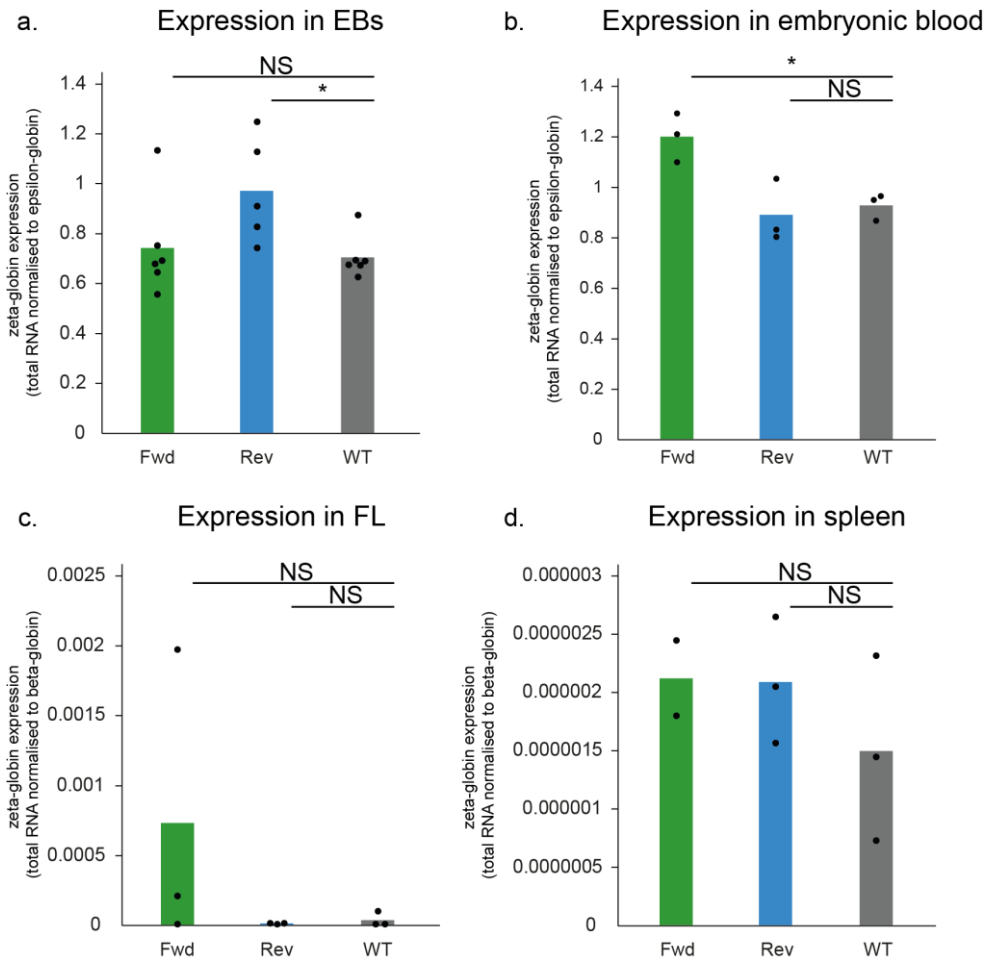


Figure 4.2: Expression of zeta-globin in various erythroid tissues in the Forward and Reverse model

*a) Expression of the embryonic zeta-globin (Hba-x) in EB-derived CD71+ erythroid cells, normalised to the embryonic beta-like globin, epsilon-globin (Hbb-y; n=5 for Rev, n=6 for Fwd and WT). Bar plot shows the mean expression with individual data points marked by black dots. b) Expression of zeta-globin in E10.5 embryonic blood normalised to epsilon-globin (n=3 for all). c) Expression of zeta-globin normalised to beta-globin in cultured E12.5 foetal liver Ter119+ erythroid cells (n=3 for all). Note the lower y-axis values compared to a) and b). d) Expression of zeta-globin normalised to beta-globin in spleen-derived Ter119+ erythroid cells (n=2 for Fwd, n=3 for Rev and WT). Note the lower y-axis values compared to a) and b). P-values were obtained using an unpaired two-tailed Student's t-test. Not significant (NS) $p > 0.05$, * $p < 0.05$.*

Expression changes of non-erythroid genes at the alpha-globin locus

In addition to the globin genes, there are several other genes close to alpha-globin enhancers. *Nprl3* is part of the sub-TAD and its promoter lies between the alpha-globin enhancers and genes. It is more highly expressed in red blood cells (Hanssen et al. 2017). *Mpg* and *Rhbdf1* were shown to be shielded from enhancer function by the HS-38 CTCF site in definitive erythroid cells (Hanssen et al. 2017). Thus, as these genes display varying levels of sensitivity to the alpha-globin enhancers, I wanted to probe their expression in the Forward and Reverse models. Importantly, alpha-globin is very highly expressed in erythroid cells, at a similar level to 18S ribosomal subunit (RN18S). The non-erythroid genes are expressed at a much lower level (1000-2000 fold lower).

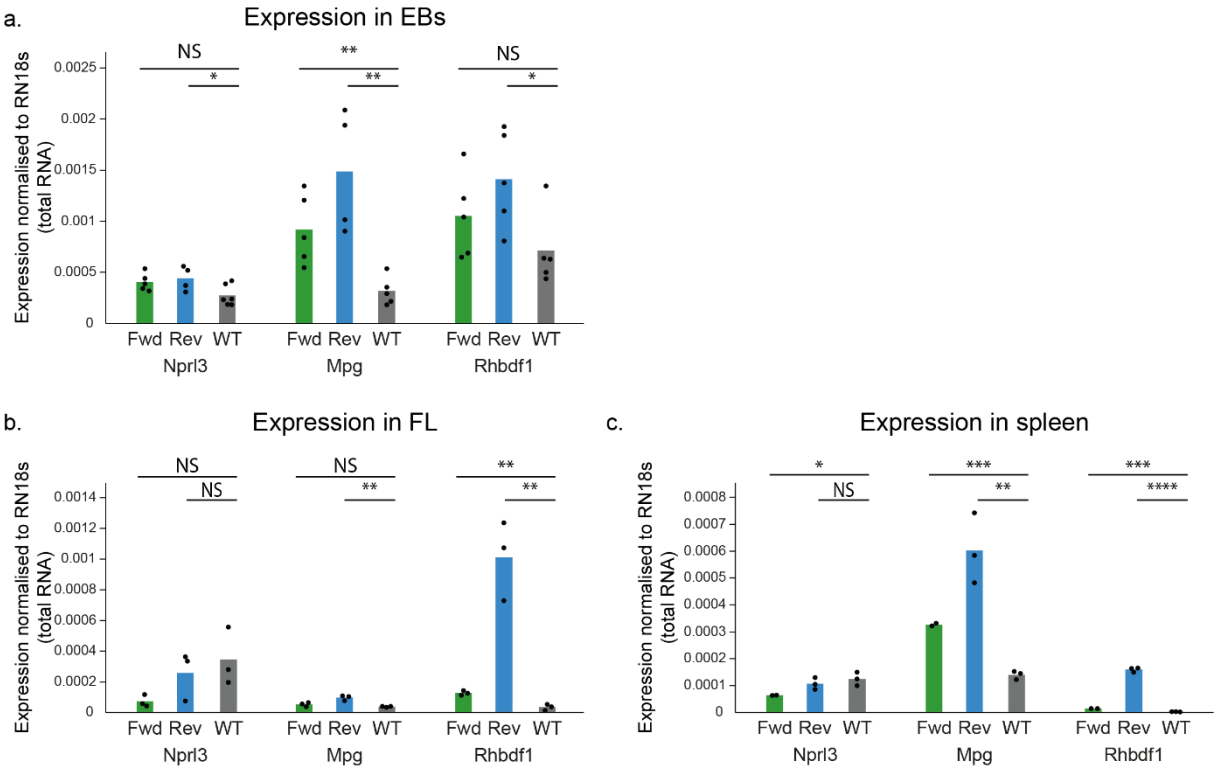


Figure 4.3: Expression of non-erythroid genes at the alpha-globin locus in the Forward and Reverse model

*a) Expression of Nprl3 (Fwd n=5, Rev n=4, WT n= 6), Mpg (Fwd and WT n=5, Rev n=4), and Rhbdf1 (n=5 for all) in EB-derived CD71+ erythroid cells, normalised to RN18s. Bar plot shows the mean expression with individual data points marked by black dots. b) Expression of Nprl3, Mpg and Rhbdf1 in cultured E12.5 foetal liver Ter119+ erythroid cells (n=3 for all), normalised to RN18s. c) Expression of Nprl3, Mpg and Rhbdf1 in spleen-derived Ter119+ erythroid cells (n=2 for Fwd, n=3 for Rev and WT), normalised to RN18s. P-values were obtained using an unpaired two-tailed Student's t-test. Not significant (NS) p > 0.05, * p <0.05, ** p < 0.01, *** p <0.001, **** p <0.0001.*

Nprl3 is expressed is expressed both in primitive and definitive erythroid cells, and its expression is not strongly affected by the ectopic CTCF site (Figure 4.3). *Mpg* is also expressed in both primitive and definitive erythroid cells, although its expression is lower in foetal liver-derived cells. Interestingly, its expression increases in EB and spleen-derived erythroid cells, most strongly in the Reverse model. The effect is not as pronounced in the foetal liver-derived cells, although the expression does increase (Figure 4.3). Finally, the expression of *Rhbdf1* is repressed by Polycomb group complexes in definitive erythroid cells, such as those derived from foetal liver and spleen (Hanssen et al. 2017). In primitive cells however, the *Rhbdf1* is expressed. The expression of *Rhbdf1* increases somewhat in the EB-derived erythroid cells. However, the effect on *Rhbdf1* expression is strongest in definitive cells, and particularly in the Reverse model. In the Reverse spleen cells, the expression is upregulated 60-fold, ten times higher than in the Forward model. The upregulation is not as high as in the foetal liver cells, but is significant in both models, with the strongest effect again in the Reverse model (Figure 4.3).

The alpha-globin enhancers and promoters remain accessible

There are substantial changes in gene expression in the CTCF insertion models, but this is not likely to be caused by changes in the enhancers themselves. To verify this, I performed ATAC-seq, which marks functional genomic elements, in both the Forward and Reverse model in primitive-like (EB) and definitive erythroid cells (foetal liver). In general, the ATAC-seq profile is not altered between the CTCF insertion models and wild-type erythroid cells, showing that the integrity of the locus was not affected and all functional elements are still present(Figure 4.4). Upon the introduction of the ectopic CTCF site in the forward or reverse orientation, a small new ATAC-seq peak is visible at the insertion site, comparable to other CTCF sites. Furthermore, the open chromatin region over the alpha-globin gene body is reduced in most of the models compared to wild-type cells, especially in the Reverse model in EBs, in accordance with changes

in gene expression. Finally, in the foetal liver cells the *Rhbdf1* promoter has become hypersensitive, matching the increase in expression I observed (Figure 4.3).

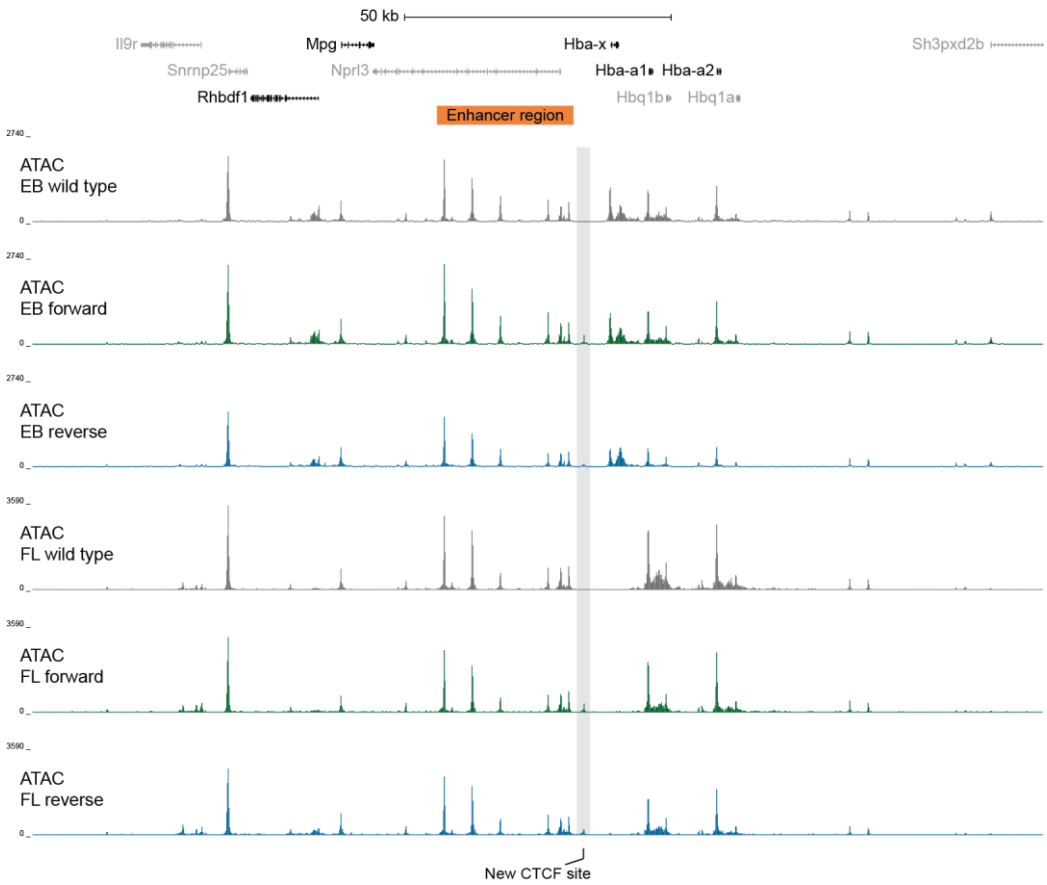


Figure 4.4: ATAC-seq tracks in the Forward and Reverse model

RPKM-normalised ATAC-seq tracks in the WT (grey), Forward (green) and Reverse (blue) erythroid cells either from day 7 EBs or E12.5 foetal liver. Three replicates were merged for each track. WT EB tracks has been published previously (Francis et al. 2022).

Development of a reporter assay

As alpha-globin expression in EBs is highly sensitive to the presence of a single CTCF binding site, we wondered if this locus might be well suited to study the insulating capacity of different CTCF binding sites. There have been some CTCF binding sites reporter assays developed at different loci, however these only observed substantial disruption of reporter gene expression when multiple CTCF binding sites were inserted in tandem (Huang et al. 2021) or studied just the chromatin interactions rather than the effect on gene expression (Jia et al. 2020). Based on the change in expression observed in the Reverse model, we expected that the alpha-globin locus

to provide a more sensitive screen, with the site of the ectopic insertion functioning as a "landing pad" for different CTCF sequences. Importantly, as we mainly introduce short sequences (<100 bp) and always at the same integration site, this screen will allow us to separate insulator function intrinsic to the binding site sequence from external factors.

To allow efficient screening of many clones, we decided to use a mouse ES cell line in which the first alpha-globin gene (*Hba-a1*) has been tagged with YFP. This hemizygous cell line has been previously characterised and published (Francis et al. 2022), and provides the potential for high-throughput screening of alpha-globin expression by fluorescence-activated cell sorting (FACS; Figure 4.5a). Using the same guides as for the creation of the Forward and Reverse model, many different sites can be inserted to test their ability to reduce enhancer promoter interactions and gene expression. Initially, five different insertions were tested: a control fragment, consisting of a "neutral" DNA sequence from the alpha-globin locus that is neither enriched in any ChIP-seq datasets nor contains known transcription factor motifs; the HS-38 binding site, in forward and reverse orientation; and the combination of HS-38 and HS-39, as they occur in the genome, in forward and reverse orientation (Figure 4.5b). The control fragment and the HS-38 in forward and reverse orientation are 83 bp, the same length as used for the Forward and Reverse model. The insertion of both HS-38 and HS-39 is a much larger fragment at 1.5 kb, since this is the distance between the two sites in the genome. It thus provides a good test for the landing pad to see if it can work with larger sequences as well. In addition, previous work has shown that the combination of these two CTCF sites is stronger than the HS-38 site by itself (Hanssen et al. 2017).

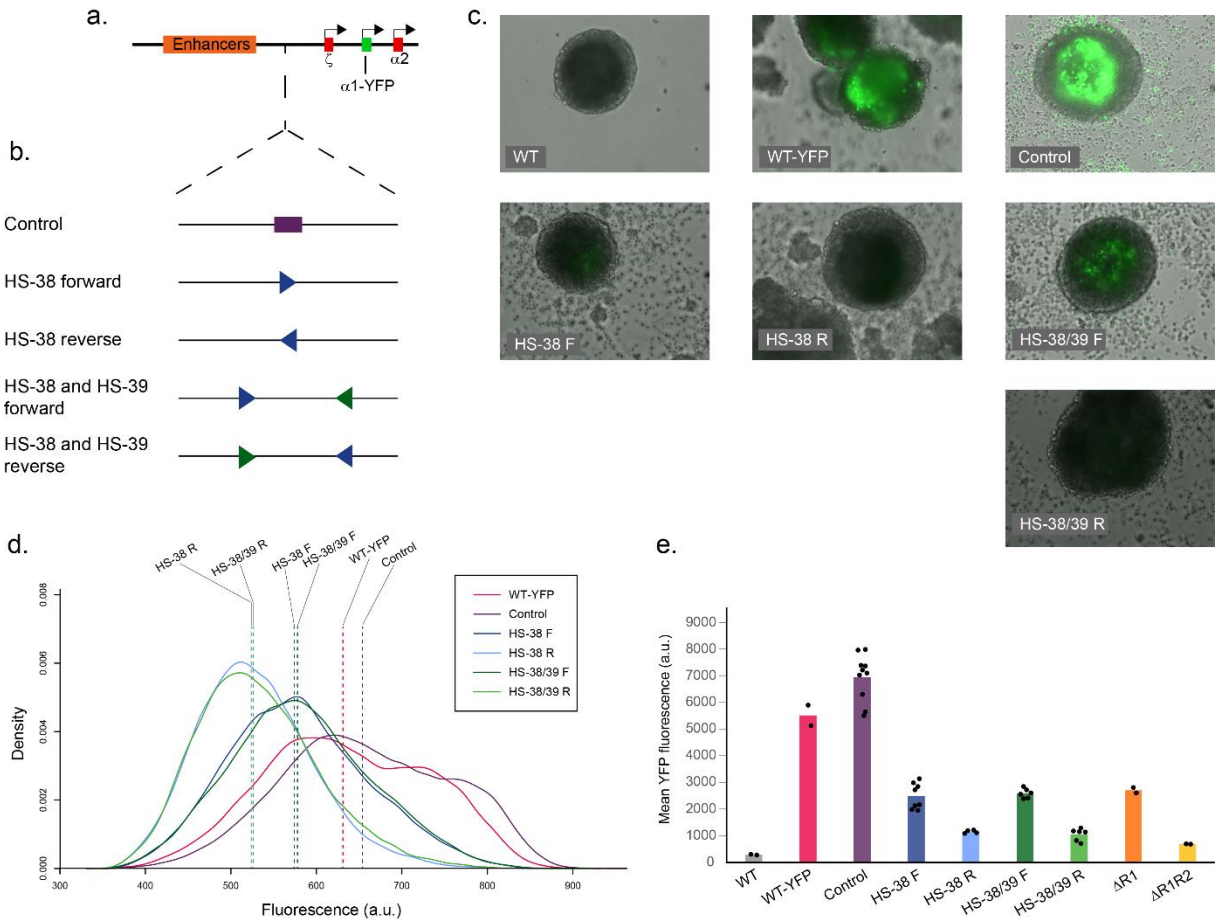


Figure 4.5: Strategy for high-throughput screening to assess CTCF sites' insulating potential

a) Modified alpha-globin locus used for the screening of CTCF sites, which can be inserted between the enhancers and the genes, while the first alpha-globin gene (Hba-a1) is tagged with YFP, as previously published (Francis et al. 2022). b) Initial sites used in the proof of principle screen. The control sequence is an 83-bp fragment with no predicted boundary function. The HS-38 forward and reverse contain the same 83-bp binding site as the Forward and Reverse model. The combined HS-38 and HS-39 site is a larger fragment (1.5 kb) amplified from genomic DNA. c) Brightfield images of day 7 embryoid bodies with several genotypes overlaid with YFP fluorescence signal. d) Flow cytometry histograms showing YFP fluorescence from the models shown in b (average of two experiments, using 5 clones for control, 4 clones for HS-38 F, 2 clones for HS-38 R, 3 clones for both HS-38/39 F and HS-38/39 R). Dotted line depicts the median. e) Median fluorescence of different clones with CTCF insertions from two experiments, replicates described in d). In addition to the CTCF insertion models, I differentiated one clone of several controls for both experiments. These controls were WT without YFP, WT with YFP tag, R1 deletion (ΔR1), and R1 and R2 double deletion (ΔR1R2), and have been previously reported (Francis 2019). Bars represent mean values and black dots median fluorescence of individual clones. P-values were obtained using an unpaired two-tailed Student's t-test: WT-YFP vs Control not significant (p= 0.051), all other genotypes vs Control p <0.0001

When ES cells with the alpha-globin-YFP allele are differentiated using the EB differentiation protocol, the subset of cells that differentiates along the erythroid lineages starts expressing YFP. This is also the case in the cells with the CTCF insertion, and varying levels of YFP were

observed in the different genotypes (Figure 4.5c). To quantify the YFP levels, the EBs were disaggregated and analysed using FACS. The distributions of fluorescence (reflecting alpha-globin expression) of the various clones overlap, but can be distinguished easily by their medians (Figure 4.5d; (Francis 2019)). The median fluorescence of the control "neutral" insertion is somewhat higher than in the wild-type YFP cells (Figure 4.5e), but this is not significant (p= 0.0516). The median fluorescence is significantly lower when a CTCF site is present between the enhancers and the genes. The fluorescence follows a similar pattern as previously observed when analysing alpha-globin RNA expression in the Forward and Reverse model, with a stronger decrease in expression when the CTCF site is oriented towards the enhancers. Surprisingly, the median fluorescence was not affected by the addition of the HS-39 site and the intervening sequence, with the median expression of HS-38 forward and HS-38/HS-39 forward almost identical. In both these models the HS-38 site is directed towards the genes, while the HS-39 site in the HS-38/HS-39 model is in the reverse orientation. The HS-38 reverse model also showed very similar levels of fluorescence to the HS-38/HS-39 reverse model, indicating that HS-39 does not contribute additional insulation when combined with HS-38.

Finally, as a comparison, I also differentiated two additional cell lines in parallel: one with the enhancer element R1 deleted, and one with both R1 and R2 deleted. These have been previously characterised (Francis et al. 2022) and the deletion of these elements reduces expression of alpha-globin by 40% for the single deletion and 90% for the combination. The reduction observed in the CTCF deletions, expected at about 45% for the Forward and 85% for the Reverse, is in a similar range. Indeed, the median fluorescence levels seem in agreement with the expected expression levels when comparing the different genetic backgrounds, with the Forward being similar to the R1 deletion, and the Reverse similar to the R1 and R2 double deletion (Figure 4.5e).

Discussion

In Chapter 3 I introduced two models in which an ectopic CTCF binding site is placed between the alpha-globin genes and its enhancers. I showed that the introduction of the new CTCF site changed chromatin interactions in an orientation-dependent manner, and that cohesin specifically accumulates at the CTCF site in the Reverse model. Here, I investigated if the ectopic site is a functional boundary by insulating enhancer function in addition to changing chromatin interactions.

The introduction of the HS-38 CTCF binding site generates a functional insulator that decreases alpha-globin expression in almost all tissues and models assessed. The boundary effect is the strongest in the Reverse model in all erythroid tissues analysed. This corresponds with the observed effect in chromatin interactions between the enhancers and the promoters: these were also more strongly reduced in the Reverse model compared to the Forward model (see Chapter 3). Thus, despite recent reports that promoter-enhancer proximity and gene expression are not always linked (Alexander et al. 2019; Benabdallah et al. 2019), the well-established correlation does seem to hold for the alpha-globin locus.

In the Reverse model, the CTCF site is oriented to block loop extrusion complexes translocating from the enhancers. In Chapter 3 I observed that cohesin, the main loop extrusion factor, accumulates at the reverse-oriented site specifically. Combined with the more severe loss of expression in the Reverse model, it is possible that the disruption of loop extrusion is the mechanism by which alpha-globin expression is reduced in the Reverse model. At least, some mechanism involving tracking along the chromatin fibre seems likely, because mechanisms invoking diffusion cannot easily explain the observation that the structural and functional effects of the inserted element on the enhancer-promoter interaction depends on the position and orientation of the its sequence (Beagan and Phillips-Cremins 2020).

I also observed that the expression of alpha-globin was generally more reduced in primitive erythroid cells. It is possible that alpha-globin expression in primitive tissue is more sensitive to perturbation than in definitive erythroid cells, as the interactions between the adult alpha-globin genes and the enhancers are weaker than in definitive cells (King et al. 2021). Additionally, it is also possible that expression of zeta-globin in primitive cells could sensitise the adult alpha-globin gene expression to the effect of the CTCF site insertion, either by enhancer competition or via its promoter acting as a boundary element . Indeed, it has been suggested that sites of active gene expression can contribute to genomic boundaries (Dixon et al. 2012; Harrold et al. 2020; Bozhilov et al. 2021).

Surprisingly, I also observed a change in the expression of some non-erythroid genes upstream of the CTCF binding sites that have previously been identified as the boundary of the sub-TAD (Hanssen et al. 2017), namely *Rhbdf1* and *Mpg* expression in definitive erythroid cells. The increase in *Rhbdf1* expression in the Reverse model is similar to the increase observed in the knockout of the HS-38 CTCF site (Hanssen et al. 2017). Clearly, although *Mpg* and *Rhbdf1* are expressed in mouse ES cells (Hanssen et al. 2017) and therefore are not solely dependent on the alpha-globin enhancers for their expression, they are susceptible to regulation by these enhancers and their expression goes up when the CTCF sites in between are deleted. It is possible that this regulation occurs via an indirect mechanism: the presence of a strong enhancer could for instance alter the chromatin accessibility of the TAD. It is also possible that this occurs via direct interactions, which is the case in the knockout of the boundary CTCF sites, where interactions between the enhancers and *Rhbdf1* and *Mpg* promoters increase (Hanssen et al. 2017). However, I do not have chromatin interaction data in definitive cells and in primitive cells the increase in expression is only very moderate, so I cannot conclude how the strong upregulation occurs in this model.

Finally, the ATAC-seq profiles show that all enhancer elements remain accessible in the CTCF insertion models. The effect on alpha-globin expression is therefore due to the restriction of chromatin interactions only. In future, I could test if this affects the binding of transcription factors at the promoters by ChIP for these proteins.

As expression of alpha-globin is strongly affected by the introduction of a single CTCF binding site in the EB-derived erythroid cells, we reasoned that this could provide an opportunity to test several CTCF binding sequences for their insulating potential. Using the original insertion site as a "landing pad" for other sequences, while keeping the surrounding chromatin unaltered, could give an interesting insight into how binding site sequence relates to the insulating potential. I showed that a system in which one alpha-globin gene is tagged with YFP can provide efficient output of insulating potential using FACS. Insertion of a neutral control sequence did not affect YFP fluorescence levels, and we were also able to integrate a larger 1.5kb fragment and assess its insulation potential. Unexpectedly, the insertion of the combined HS-38 and HS-39 site did not increase the insulation compared to the insertion of just HS-38, although their combined deletions in the wild-type position showed that they are a stronger boundary together. This lack of additional insulation when a weaker CTCF site is combined with a strong CTCF site placed in convergent orientation has been noted by others (Huang et al. 2021), but is surprising when compared to the knockout data from the original locus (Hanssen et al. 2017).

Recent work in the Higgs lab has analysed more CTCF binding sites using the assay described in Figure 4.5. Interestingly, these showed a range of decreased YFP expression (Figure 4.6), emphasising the dynamic range of the assay and its ability to distinguish CTCF binding sites with greater or lesser insulation potential. Previous characterisation of the CTCF binding sites at the alpha-globin locus showed that HS-29 is not a strong insulator compared to HS-38 (Figure 3.1, (Hanssen et al. 2017)). This difference is again observed in the CTCF insulator assay, with HS-29 decreasing YFP expression to a lesser extent than HS-38 (Figure 4.6a). We next assessed the

insulator potential of a CTCF site from the beta-globin locus, the 3'HS1 CTCF site. This CTCF binding site was capable of reducing YFP levels prominently (Figure 4.6b), as was the 5'HS4 CTCF site from the chicken beta-globin locus, which is one of the first insulator sequences described (Chung et al. 1997). We next analysed some well-characterised CTCF binding sites, such as from the HoxA (Narendra et al. 2015), Igf2-H19 (Bell and Felsenfeld 2000; Hark et al. 2000) and Sox9 (Despang et al. 2019) loci. These CTCF sites all reduced YFP levels considerably (Figure 4.6c), whereas the insertion of a less prominent CTCF site from the Epha4-Pax3 TAD boundary (Anania et al. 2021) had a more modest effect on YFP levels. Interestingly, the insertion of the two Sox9 CTCF binding site showed a substantial reduction in YFP level in our reporter assay, but in another study that used a similar approach based on the Sox2 locus these sites only showed very subtle or no boundary activity (Huang et al. 2021). This suggests that our reporter assay is more sensitive in quantifying the activities of boundary elements.

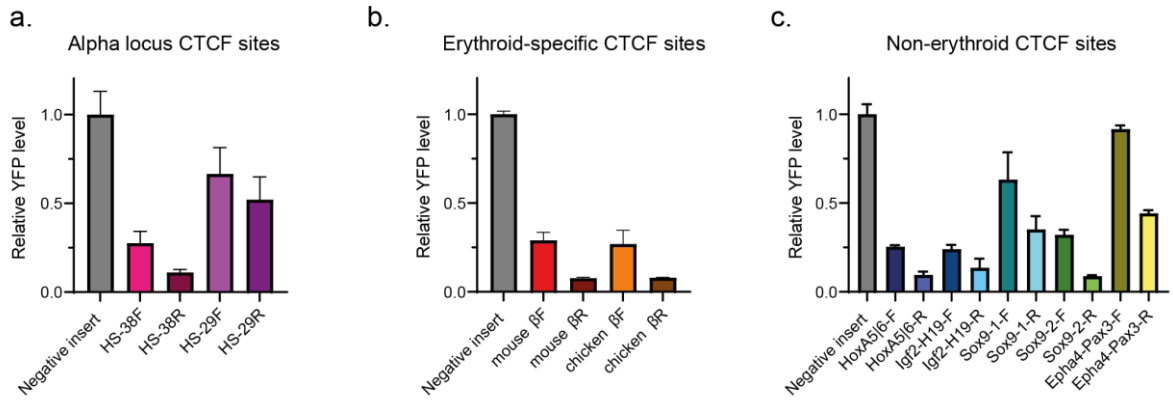


Figure 4.6: Additional CTCF binding sites analysed using the alpha-globin locus landing pad
YFP levels normalised to the neutral control insert (Negative insert) in CTCF site in forward and reverse orientation. Either sequences taken from the alpha-globin locus (a), beta-globin locus (b) or elsewhere in the genome (c). Replicates for a) Negative insert n=2, HS-38 forward and reverse n=4, HS-29 forward and reverse n=4. b) Negative insert n=4, mouse beta-globin forward n=4, mouse beta-globin reverse n=3, chicken beta-globin forward n=3, chicken beta-globin reverse n=4. c) Negative insert n=8. For HoxA5|6 forward and reverse, and Sox9-1 and Sox9-2 reverse n=3. For Igf2-H19 reverse, Sox9-1 and Sox9-2 forward, and Epha4-Pax3 forward and reverse n=4. For Igf2-H19 forward n=5. Data acquired and analysed by Felice Tsang.

Importantly, all of the sites tested showed a stronger insulation when inserted in the reverse orientation. This shows that the orientation effect is not specific to the HS-38 site that was tested

in the Forward and Reverse models, but more likely a general feature of this locus and the position of the inserted CTCF sites relative to the enhancers and promoters. Future work will focus on trying to understand the wide range of insulating capacity CTCF sites have and how individual sequences contribute to this.

Acknowledgements

E10.5 embryonic blood was collected by Mira Kassouf. ATAC-seq in wild-type EBs has been published previously, as was the mouse ES cell line with the YFP-tagged alpha-globin allele both in wild type and with the enhancer mutations (Francis et al. 2022). Targeting of the control and CTCF site insertion in the YFP-tagged ES cells was done by Muhammad Hanifi and genotyping of these clones was performed by Mira Kassouf and me. I carried out the EB differentiations and data acquisition for Figure 4.5, and the data was analysed together with Helena Francis. Data in Figure 4.6 was acquired and analysed by Felice Tsang.

Chapter 5: Studying transcriptional activation at an enhancerless alpha-globin locus

Introduction

Tissue-specific transcription factors play a key role in regulating the expression of genes in specific tissues (Spitz and Furlong 2012; Neal et al. 2021). However, it has been possible, to some extent, to bypass tissue-specific transcription factors and activate transcription in tissues where genes are usually not expressed by targeting catalytically dead Cas9 (dCas9) fused to transcriptional co-activators to *cis*-regulatory elements (Cheng et al. 2013; Maeder et al. 2013; Mali et al. 2013; Hilton et al. 2015; Li et al. 2020b). Interestingly, it was even possible to activate gamma-globin expression distally by targeting dCas9-p300 to the beta-globin enhancer region in HEK293T cells (Hilton et al. 2015). Gamma-globin expression went up by 269 fold, while beta-globin was not transcriptionally activated, and the expression of epsilon and delta-globin increased modestly (~15.5 and 7.5 respectively). These experiments provide important insights into some aspects of how enhancers might regulate gene expression.

In addition to epigenetic manipulation, enhancer knockout studies have also been instrumental in understanding how enhancers or specific enhancer elements regulate gene expression (Hay et al. 2016; Thomas et al. 2021; Vos et al. 2021). By definition, when enhancers are deleted, expression of their target genes is reduced. Examining how the deletion of an enhancer affects the transcription cycle enables unravelling of the mechanisms by which enhancers normally regulate gene expression. For example, if enhancers function by creating a permissive environment, an enhancer knockout might be reflected by a change in histone modifications or open chromatin at the promoters. If recruiting PIC factors is the main role of the enhancers, an enhancer knockout would result in lower levels of PIC formation and initiation of transcription. If it affected pause release or elongation, we might expect to see a build-up of short transcripts

at the 5’ end of the gene. How much of the gene expression remains depends on the basal activities of the promoters and any remaining enhancer elements.

It has been previously suggested that enhancers may play an important role in pause release (Lin et al. 2013; Liu et al. 2013). In particular, it was suggested the beta-globin enhancers affect beta-globin transcription via pause release (Sawado et al. 2003). However, detailed analysis of a model in which the main alpha-globin enhancers are deleted (ΔR1R2), showed that this is not the case at the alpha-globin locus. In this model, alpha-globin expression is reduced by 90%, but there was a uniform loss of nascent RNA and no build-up of paused Pol II at the promoters (Larke et al. 2021). It is therefore more likely that the enhancers play a role in an early step of transcription, such as recruiting Pol II and/or other PIC components. Genome-wide analysis also suggested that enhancers act primarily to recruit the PIC (Larke et al. 2021). In the ΔR1R2 model, the alpha-globin promoters retain their open chromatin peak, as assayed by ATAC-seq, and some Pol II is still recruited to the promoters. The same pattern is observed for the beta-globin genes when their two main enhancers are deleted (Larke et al. 2021). It is possible that this is mediated by the presence of the remaining enhancer elements, or the basal transcriptional activity of the promoters. There are also several erythroid-specific transcription factor binding sites at the promoters that could play a role in gene regulation.

Previous work at the alpha-globin locus has shown that the combined deletion of the R1 and R2 enhancer elements (ΔR1R2) reduces the expression of nascent alpha-globin by 90% (Hay et al. 2016). Individual deletions of the other enhancer elements (R3, Rm and R4), have little or no effect on alpha-globin expression. However, the presence of just R1 and R2, without the other enhancer elements, only provides 43% of wild-type expression rather than the expected 90% expression (Francis 2019). This indicates that the R3, Rm and R4 elements do play a role in regulation and that the ΔR1R2 model is does not represent an enhancerless alpha-globin locus.

To gain a more detailed understanding of how the entire cluster of the alpha-globin enhancers, acting in concert, activate gene expression, and the independent role of the alpha-globin promoters therein, I characterised the all enhancer knockout (AEKO) model, in which all five enhancer elements are deleted. This ES cell model was created by Helena Francis using a recombinase-mediated genomic replacement (RMGR) strategy to exchange one allele of the alpha-globin locus with an engineered bacterial artificial chromosome (BAC; (Francis 2019)). The recipient mouse ES cells have heterospecific lox sites on one allele to facilitate the exchange. On the other allele they have a 117 kb deletion that removes a region larger than the sub-TAD, including *Rhbdf1* and the downstream HS+44/HS+48 sites (Figure 5.1a). This method was used to create two hemizygous cell lines: one in which the remaining alpha-globin locus was replaced by a wild-type alpha-globin locus (creating WT hemizygous cells, Figure 5.1b) and one in which it was replaced by an allele in which all enhancers except for R2 were deleted. The R2 element was subsequently deleted by CRISPR/Cas9 editing to create the hemizygous AEKO model (Figure 5.1b). The creation of these models has previously been described in detail (Francis 2019).

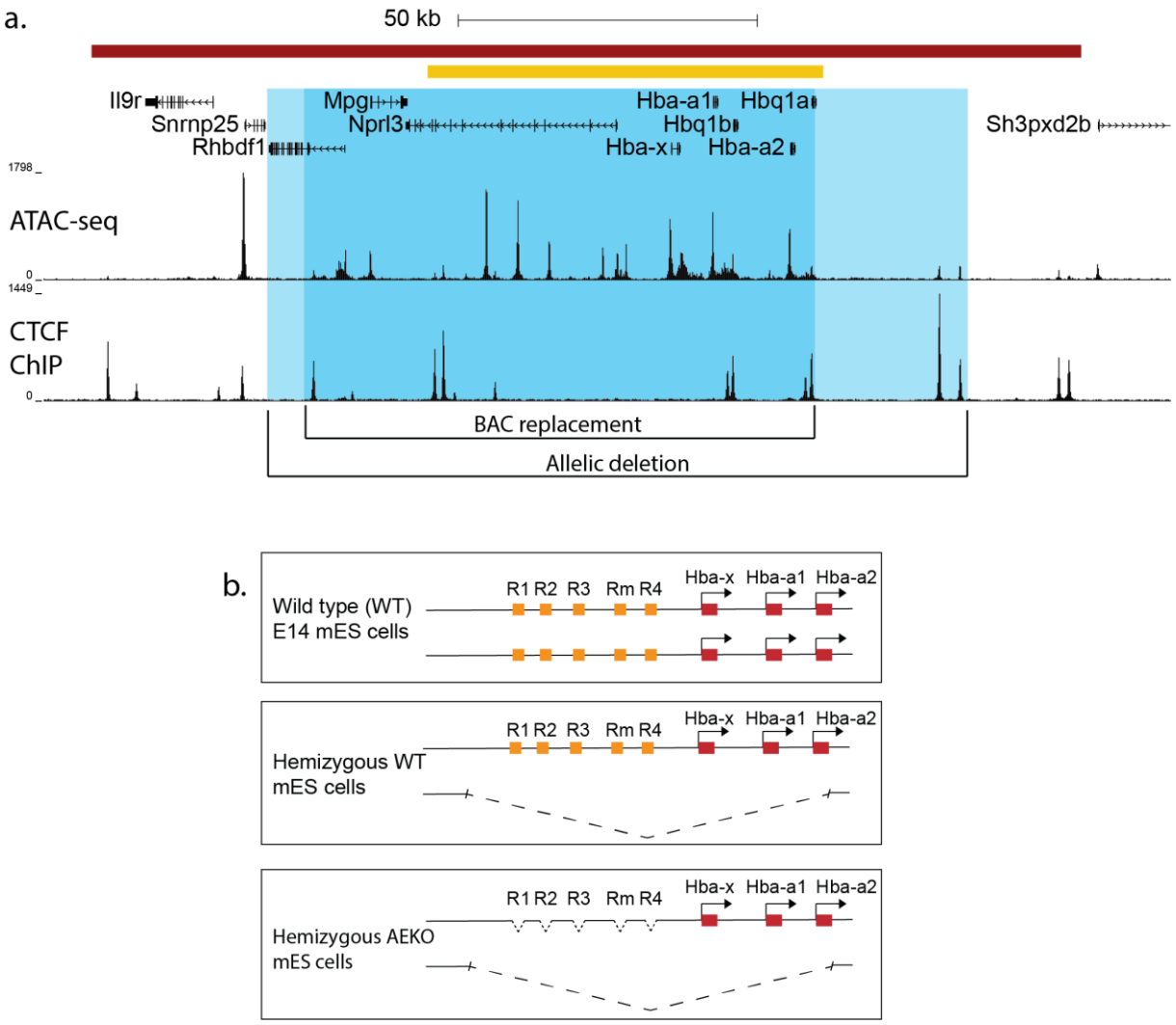


Figure 5.1: Mouse ES cell lines used in this chapter

a) The BAC-recipient mouse ES cell line has heterospecific lox sites that allow for the replacement of a large region (86 kb), highlighted in dark blue (mm9 chr11:32115389-32200533). A large (117 kb) removes the entire sub-TAD and flanking sequences from one allele (highlighted in light blue (mm9 chr11:32109272-32226122)), making the cells hemizygous for the alpha-globin locus. TAD is indicated in dark red and sub-TAD in yellow. b) Schematic depiction of the three mouse ES cell lines used in this chapter. The deletions removing the enhancers range from 682 - 1146 bp.

In the previous chapters, I showed how the introduction of a single CTCF binding site between the alpha-globin enhancers and genes was able to disrupt enhancer-promoter interactions. When the CTCF binding site was oriented towards the enhancers, it reduced alpha-globin expression by 86% or about 2.8 fold. Here, I analyse a model in which all five alpha-globin enhancer elements have been deleted, to contrast absence of the enhancers with reduced access of the enhancers to the promoters. Alpha-globin expression is almost completely lost in

the AEKO and Pol II is not recruited. Analysis of several histone modifications and transcription factors shows the enhancers play a pivotal role in activating the locus.

Results

First, I employed total RNA-seq to assess gene expression in the AEKO model, using EB-derived erythroid cells. These cells are primitive-like erythroid cells, and usually express high levels of embryonic zeta-globin as well as alpha-globin, and also express *Rhbdf1*. As I compare wild type erythroid cells with the AEKO model that is hemizygous for the alpha-globin locus (Figure 5.1), it is important to note that some genes (*Rhbdf1, Mpg, Nprl3,* and all of the globin genes at the alpha-globin locus) are expected to have only half their usual expression, because there is only one copy of these genes in the AEKO cells. I therefore have adjusted the fold change accordingly, to account for the copy number of every gene. Alpha-globin expression is severely reduced (~750 fold) in the AEKO model (Figure 5.2). Zeta-globin expression is not as strongly affected as alpha-globin expression, but is still ~150 fold lower. Expression of the two theta-globin pseudogenes is also reduced, with *Hbq1b* most strongly affected (~650 fold lower). Surprisingly, expression of upstream *Rhbdf1* is also strongly reduced by the deletion of the alpha-globin enhancers, as its expression is about 60 fold lower. Other non-erythroid genes the locus (*Nprl3*, *Snrnp25*, *Mpg*) also have a lower expression, but by a more modest ~2-5 fold.

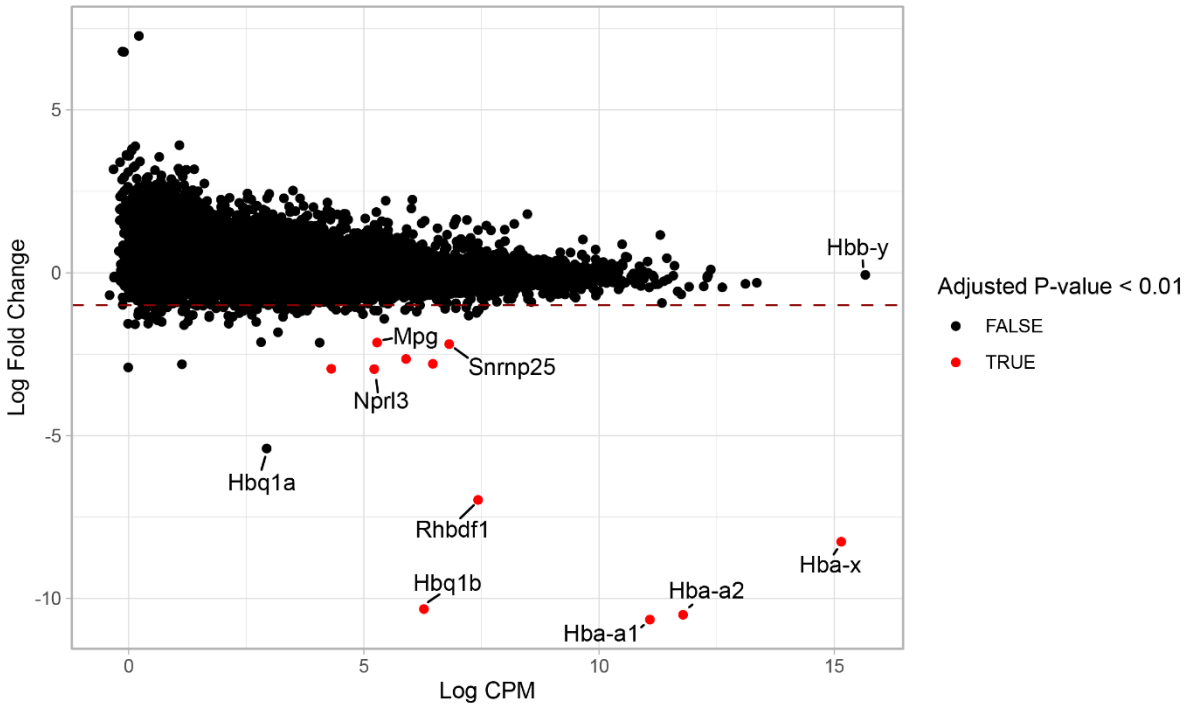


Figure 5.2: Differentially expressed genes in AEKO vs WT CD71+ EB-derived erythroid cells
MA plot of RNA-seq data in AEKO vs WT erythroid cells (n=2). Log counts per million reads mapped is plotted on the x-axis and log fold change is plotted on the y-axis. Log fold change of -1, which is expected for genes that are hemizygous in the AEKO model, is indicated by a red dotted line.

Next, to assess the functional elements at the alpha-globin locus, I analysed ATAC-seq data previously generated by Helena Francis (Francis 2019). The peaks over the enhancer elements are no longer present (Figure 5.3a). The peaks over the alpha-globin and zeta-globin gene bodies are also absent, but very small peaks over the alpha-globin promoters remain (Figure 5.3b). Small peaks coinciding with the theta-globin promoters also remain. These theta-globin associated peaks are also present in mouse ES cells and coincide with tissue-invariant CTCF binding sites. The zeta-globin promoter appears more strongly affected than the alpha-globin promoters. This is confirmed by a comparison of read counts over promoters genome wide in the AEKO and hemizygous wild-type cells (Figure 5.3c). It also shows that embryonic beta-globin-like genes (*Hbb-y* and *Hbb-bs*) remain fully hypersensitive, as do the *Nprl3* and *Rhbdf1* promoters.

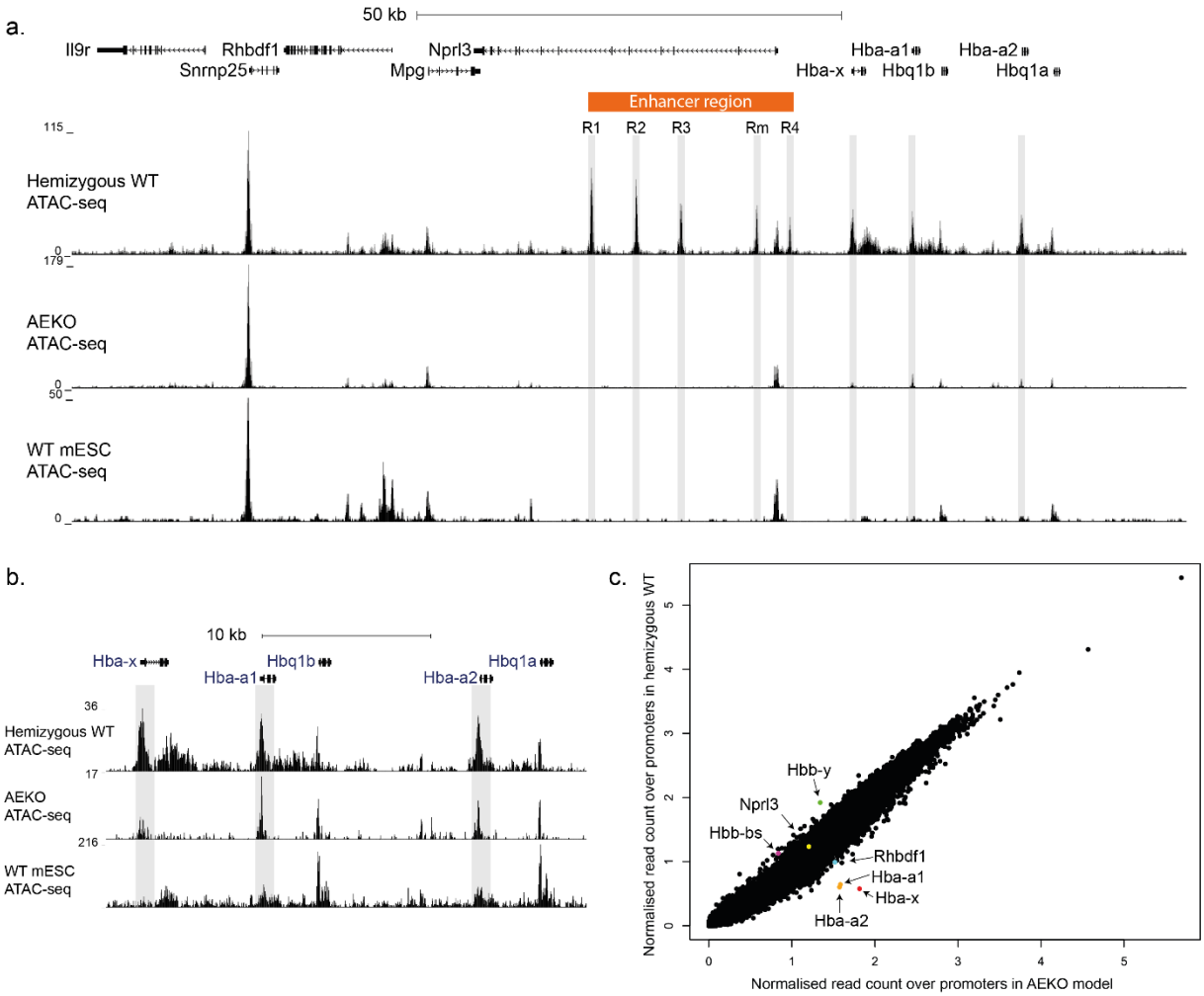


Figure 5.3: ATAC-seq analysis of hemizygous WT and AEKO erythroid cells as well as mouse ES cells

a) ATAC-seq tracks for hemizygous WT (n=3) and AEKO (n=4) CD71+ EB-derived erythroid cells normalised to 10 million reads over all called peaks (genome wide) for both samples. RPKM-normalised ATAC-seq profile in mouse ES cells (n=3). b) Same tracks as in a), but focussed on the alpha-globin genes. c) ATAC-seq read counts over all promoter regions (genome wide) in the hemizygous WT and AEKO models. Promoters highlighted: Hba-a1 and Hba-a2 in orange, Hba-x in red, Hbb-bs in pink, Hbb-y in green, Nprl3 in yellow, and Rhbdf1 in blue. Data acquired by Helena Francis (Francis 2019).

Characterisation of histone modifications in the AEKO model

To better understand chromatin changes in the AEKO model, I investigated histone modifications by ChIPmentation and ChIP. H3K4me3 is usually found at active or poised promoters (Core et al. 2014), and is present at erythroid genes (*Hba-a1, Hba-a2, Hba-x* and *Hbq1b)* as well as some non-erythroid genes (*Snrnp25, Rhbdf1, Mpg* and *Nprl3*) at the wild-type alpha-globin locus in erythroid cells (Figure 5.4). In the AEKO model, H3K4me3 is lost from the

alpha-globin promoters as well as the *Hbq1b* promoter. A small peak remains at the zeta-globin promoter. H3K4me3 is also lost from the *Rhbdf1* promoter, consistent with the observed change in expression.

Chromatin around active enhancers, and to a lesser extent promoters, is marked by H3K27ac. It is usually present at the alpha-globin enhancer region and at the alpha-globin genes (Figure 5.4). When the enhancer elements are deleted, the H3K27ac levels are strongly reduced over the entire locus. Small residual peaks of H3K27ac can be observed over the *Nprl3* and *Hba-x* promoters, consistent with their continued expression in the absence of the alpha-globin enhancers.

H3K27me3 is a histone modification generally associated with Polycomb-mediated repression (Azuara et al. 2006). It is present at *Sh3pxd2b*, a gene downstream of the alpha-globin locus that is not expressed in EB-derived erythroid cells. Furthermore, in definitive spleen-derived erythroid cells, *Rhbdf1* is repressed and marked by H3K27me3 (Hanssen et al. 2017), although this mark is absent in primitive erythroid cells where *Rhbdf1* is expressed. In the AEKO model, there is a large H3K27me3 peak at the *Rhbdf1* promoter, but no other new peaks arise.

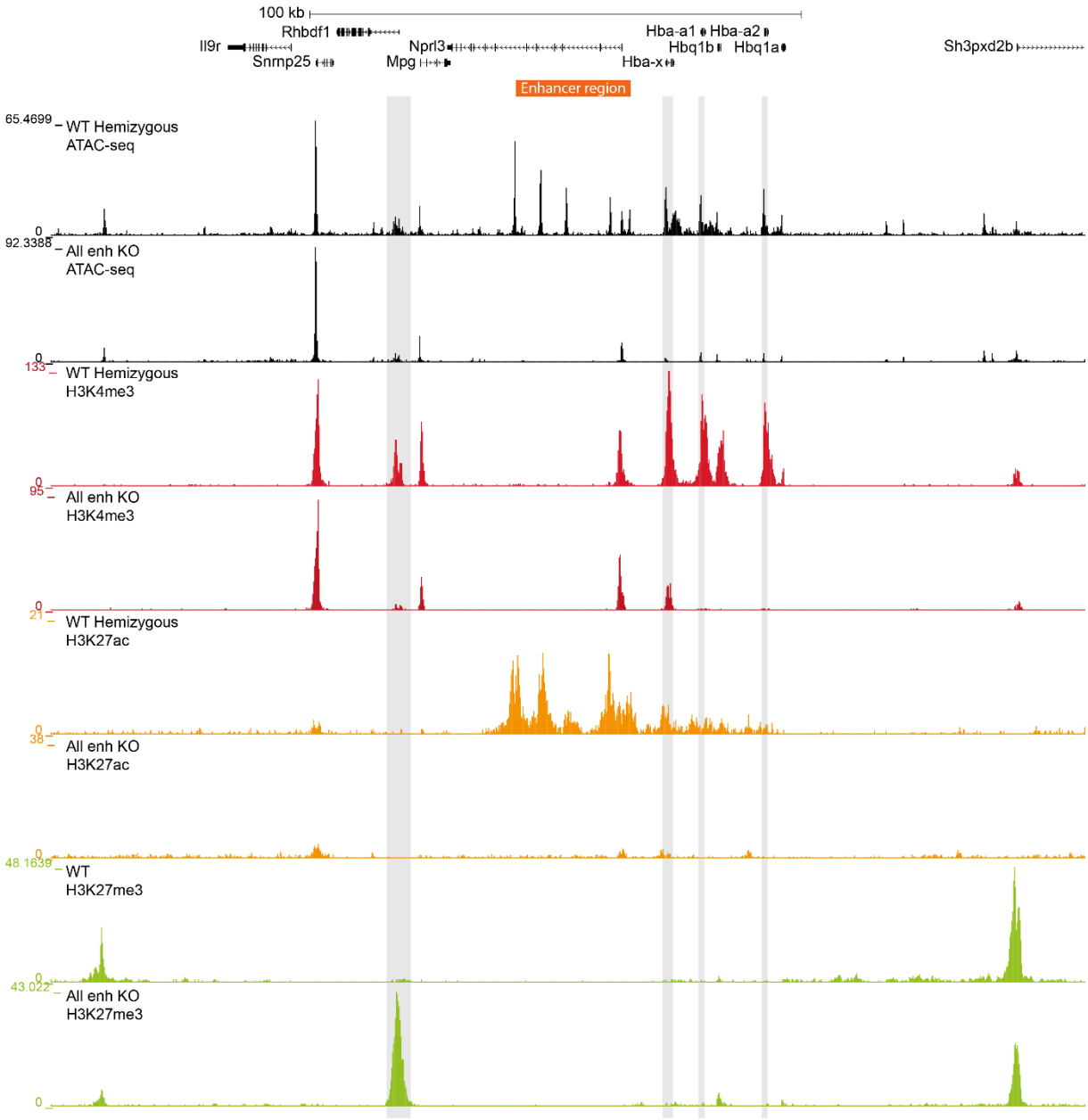


Figure 5.4: Characterisation of histone modifications in the AEKO model

Top: ATAC-seq tracks in hemizygous WT and AEKO EB-derived erythroid cells. Middle: Chipmentation tracks for H3K4me3 and H3K27ac histone modifications in hemizygous WT and AEKO EB-derived erythroid cells (n=2). Bottom: ChIP-seq track for H3K27me3 histone modification in WT and hemizygous AEKO EB-derived erythroid cells (n=2).

Recruitment of transcription factors in the AEKO model

Next, I investigated some of the proteins and protein complexes that are usually present at the alpha-globin promoters. For instance, there is a Gata1 binding site just upstream of the core promoter of alpha-globin and zeta-globin. General transcription factor TBP is also present at the alpha-globin and zeta-globin promoters in wild-type erythroid cells. Surprisingly, both Gata1 and

TBP are absent from the promoters in the AEKO cells. This is unexpected, as there is still an open chromatin site at the promoters, even though it is reduced, and the absence of Gata1 and TBP therefore suggest that these transcription factors require more than an accessible binding motif. Finally I looked at RNA polymerase II, and this is lost from the alpha-globin and zeta-globin promoters as well.

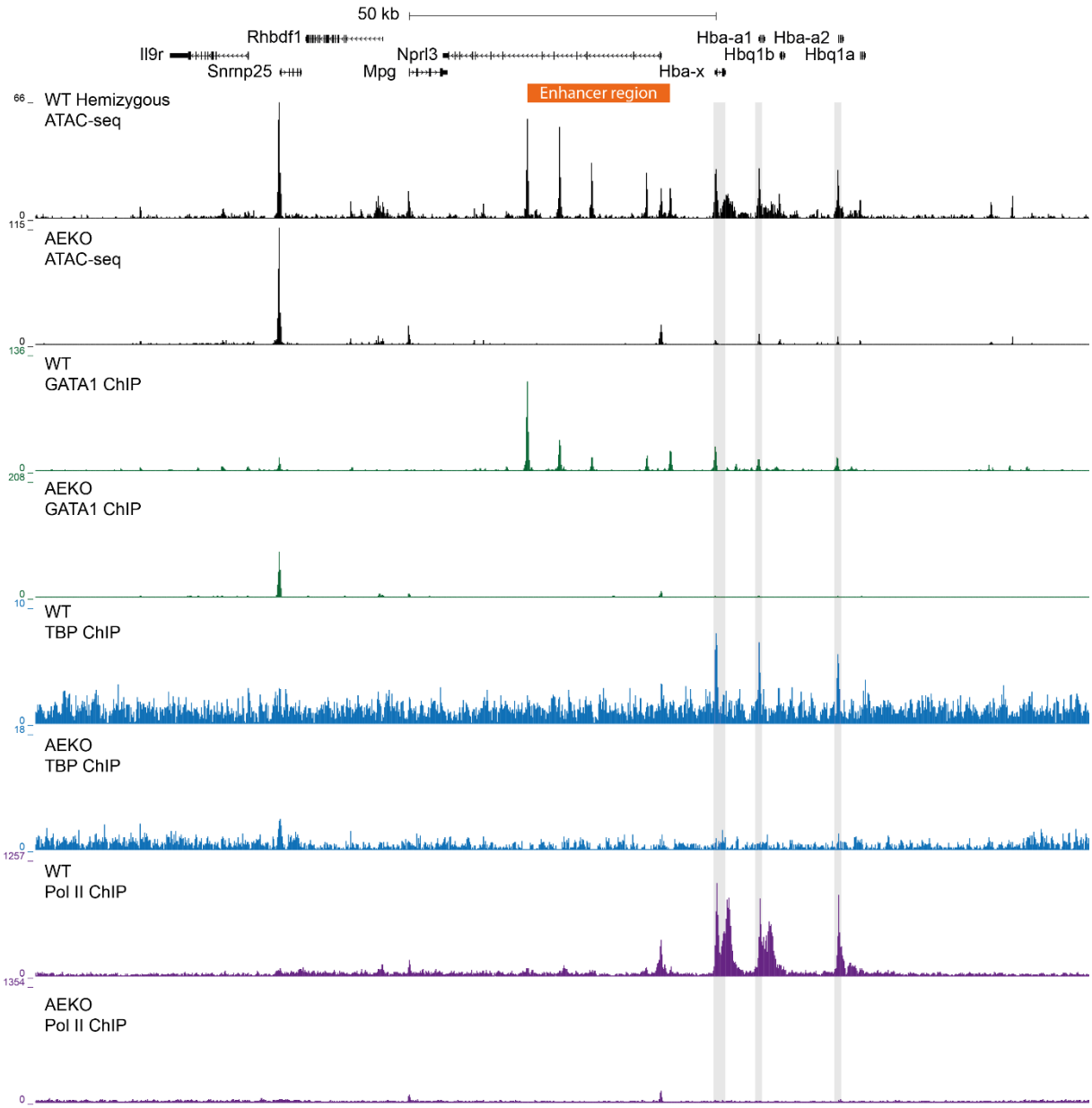


Figure 5.5: Characterisation of promoter-associated factors in the AEKO model
Top: ATAC-seq tracks in hemizygous WT and AEKO EB-derived erythroid cells. Bottom: ChIP-seq track for Gata1 in WT (n=3) and AEKO (n=2) EB-derived erythroid cells, and for TATA-binding protein (TBP, n=1) and Pol II (n=2) in WT and AEKO EB-derived erythroid cells.

In addition to affecting transcription, enhancer clusters are also reported to play a role in 3D chromatin structure, by acting as recruitment sites for cohesin loading (Hua et al. 2021; Vos et al. 2021). To investigate if the alpha-globin enhancers play such a role, I performed ChIP-seq for Rad21, a cohesin subunit. Cohesin, the main loop extrusion complex, is often found to colocalise with CTCF binding sites. In addition to CTCF binding sites, it is also present at active enhancer elements and highly expressed genes. In wild-type erythroid cells, cohesin is present at the CTCF binding sites, as well as the alpha-globin enhancers and genes but at a lower level (Figure 5.6). In the AEKO model, cohesin is lost from the enhancer region and there seems to be a lower overall level of cohesin. This latter effect is more difficult to quantify, particular as I only have one replicate in both wild-type and AEKO erythroid cells. Cohesin is however still present at the CTCF sites. Importantly, in Figure 5.6 I compare WT EB-derived cells with AEKO cells that are hemizygous for the alpha-globin locus. As such, there is only one allele and the enrichment is lower at the CTCF sites accordingly, but not much more than expected. Comparing the Rad21 recruitment in the AEKO model to mouse ES cells shows that the profiles are very similar, and therefore the AEKO model seems to reflect the inactive locus in ES cells.

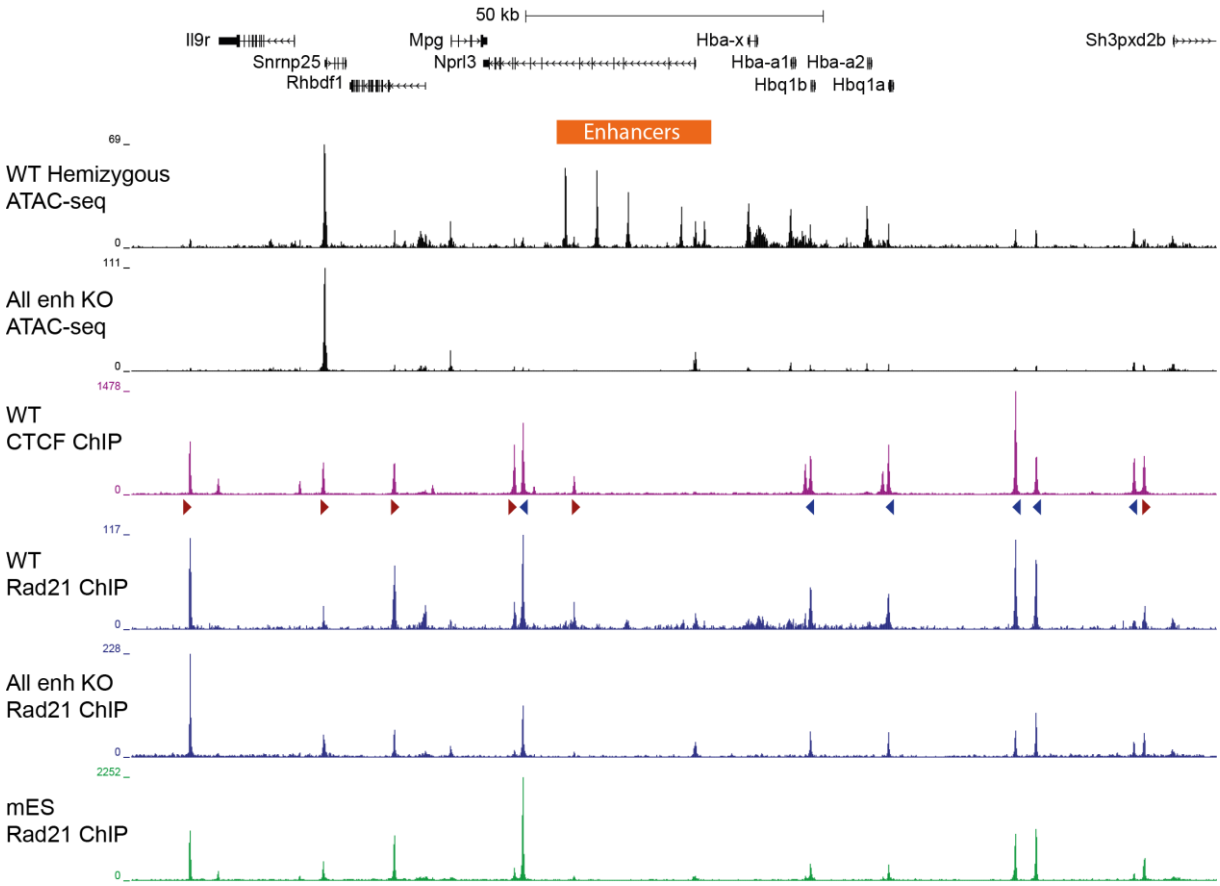


Figure 5.6: Rad21 ChIP-seq in the AEKO model

ATAC-seq tracks in hemizygous WT and AEKO EB-derived erythroid cells. CTCF ChIP-seq track in EB-derived cells (n=2, (Francis et al. 2022)). Rad21 ChIP-seq in EB-derived WT (n=1, unpublished, generated by Caroline Harrold) and AEKO (n=1) erythroid cells. Rad21 ChIP-seq in mouse ES (mES) cells (n=1, (Hua et al. 2021)).

Using dCas9 to rescue alpha-globin expression in the AEKO model

The characterisation of the AEKO model so far has shown that most activity is lost from the locus upon the deletion of the enhancers. Small open chromatin sites over the promoters remain, but binding of Gata1 and TBP is lost. Presumably, the wide range of transcription factors and associated co-activators normally bound at the enhancer elements normally play a role in recruiting these proteins to the promoters. However, it remains unclear what each individual transcription factor or co-activator contributes to enhancer function. The AEKO model has great potential as a blank canvas to ask this question, especially as all *trans*-acting factors normally present at the promoter are still expressed but the *cis*-regulatory elements are absent; Gata1 is still expressed and therefore able to bind the promoter if the chromatin environment allows it.

To gain a better understanding of what transcription factors and co-activators are central to the enhancers' function, I decided to employ a dCas9 activation strategy. Via dCas9-mediated recruitment, transcription factors or co-activators can be brought to different regions in the locus, such as the alpha-globin promoters or the locations of the deleted enhancers (Figure 5.7a). Many different factors could be fused to dCas9: erythroid transcription factors like Gata1 and Nf-e2, or co-activators such as p300 or mediator, as well as the cohesin loading factor NIPBL. As a proof of principle of the dCas9 activation strategy, I decided to start by recruiting p300 to the alpha-globin promoters, reasoning that strong activation at the promoters might provide the maximum recovery of gene expression possible.

The AEKO model requires EB differentiation to produce erythroid cells, which takes seven days. Expressing dCas9-p300 constitutively during the differentiation protocol risks causing changes to the alpha-globin locus chromatin, as it would be activated for multiple days in cells in which the locus is not normally active. It would then be difficult to distinguish if difference between the wild-type and AEKO model are due to prolonged activation of the locus by dCas9-p300, or due to its activity in the absence of the enhancers. Furthermore, overexpressing p300 could negatively affect cell viability and differentiation, as p300 acetylation targets many different proteins (Weinert et al. 2018). I therefore decided to design a system in which dCas9-p300 expression is inducible, by utilising the Tet-On system to control dCas9-p300 expression by doxycycline (Das et al. 2016). I modified a plasmid previously used to express dCas9 fused to APEX2 (Myers et al. 2018), which contains of the doxycycline-sensitive transactivator rtTA. The rtTA is expressed in frame with a T2A self-cleaving peptide followed by mCherry, which can be used to monitor rtTA expression. A TRE-3G promoter (which binds rtTA in an inducible manner) drives dCas9-p300 expression, which can be similarly monitored by an in-frame T2A-GFP cassette. After introducing this construct into mouse ES cells, the cells and resulting EB-derived cells should constitutively express mCherry and be inducible for dCas9-p300 and GFP expression.

This will allow for selection of erythroid cells expressing dCas9-p300 from the bulk EB population (Figure 5.7b).

Recruiting multiple dCas9-p300 fusion proteins is beneficial for transcription (Cheng et al. 2013; Maeder et al. 2013; Mali et al. 2013), and therefore I decided to design multiple sgRNAs to target the alpha-globin promoters (Figure 5.7c). As the promoter sequence of the two alpha-globin sequences is very similar, four sgRNAs bind to both promoters. I added two sgRNAs that bind uniquely to either of the two promoters outside of the homologous region, for a total of six sgRNAs for each alpha-globin gene. I designed the sgRNAs to bind outside of the region of the core promoter and transcription factor binding sites (> 370 bp from TSS) and spaced them 165-270 bp apart, to avoid disrupting nucleosome spacing.

The sgRNAs were expressed using a CARGO array (Gu et al. 2018), and together with the Tet-On system described above this formed a plasmid to recruit dCas9-p300 to the alpha-globin promoters (Figure 5.7d). As the total sequence to be integrated was large (>14 kb), I opted to use the piggyBac transposase system (Wang et al. 2008) which is especially suited for this purpose.

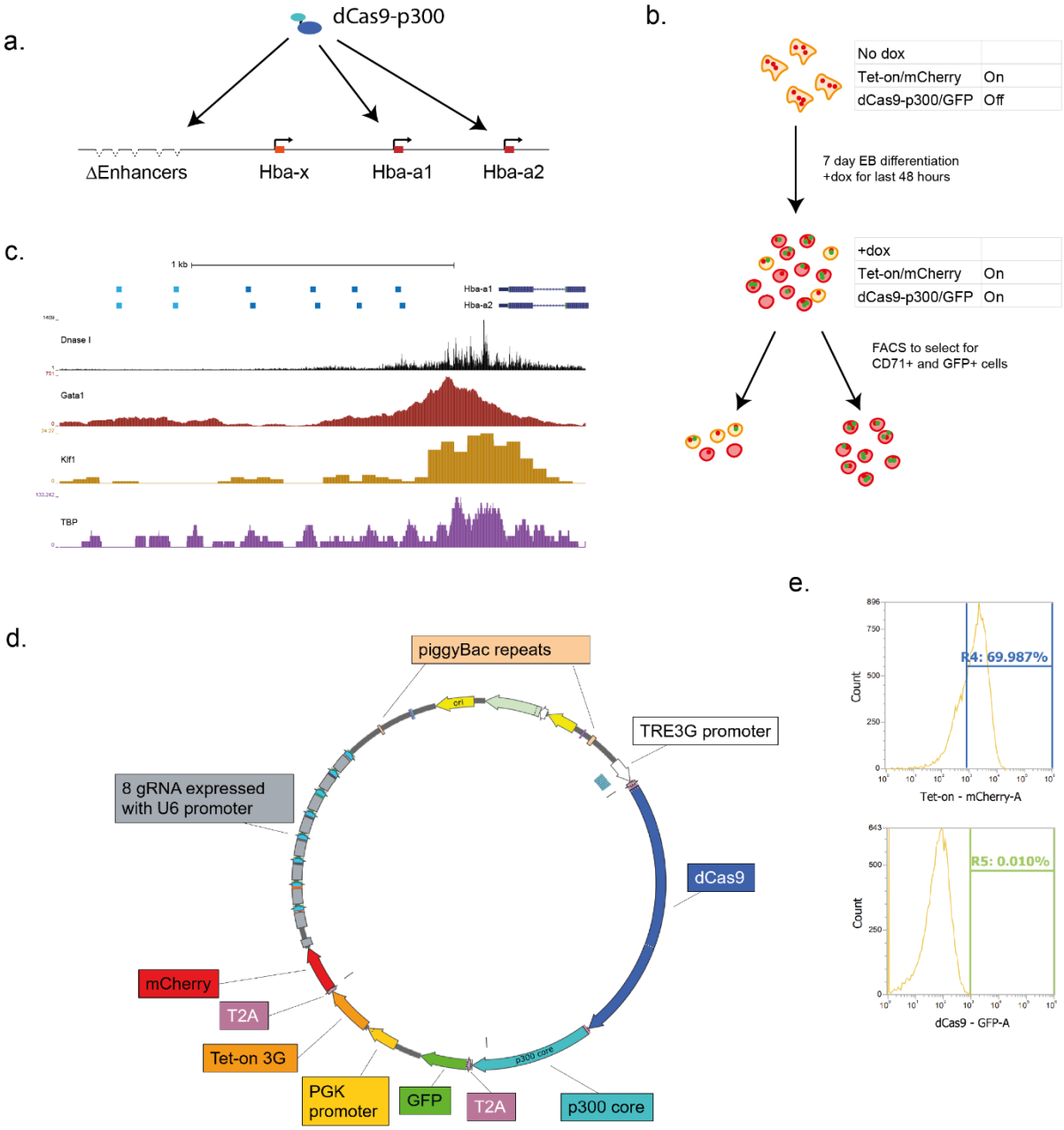


Figure 5.7: dCas9-p300 strategy to rescue alpha-globin expression in the AEKO model

a) A range of transcription factors and co-activators can be fused to dCas9 and be recruited to different sites, to test if that specific factor is sufficient to activate expression from the alpha-globin promoters. b) The Tet-On system can be employed to control expression of dCas9-p300, so that it is only expressed during the last 48 hours of EB differentiation. By co-expressing mCherry with the Tet-On transactivator and GFP with the dCas9-p300 fusion protein, the expression of all components of the system can be assessed using FACS, and only cells that differentiate into erythroid cells (CD71+) and expresses dCas9-p300 can be selected by FACS. c) The eight sgRNA are designed to target the dCas9-p300 to the promoters of both alpha-globin genes. Four are in to the homologous region and bind both promoters, while two sgRNAs further upstream are unique to either Hba-a1 or Hba-a2. DNase-seq track in Ter119+ spleen has been previously published (Schwessinger et al. 2017), Gata1 and TBP ChIP-seq track as in Figure 5.5, and a Klf1 ChIP-seq track in spleen (n=1). Sequencing tracks are aligned to Hba-a1. d) Plasmid designed to deliver the dCas9-p300 expression under doxycycline control, with the sgRNA array.

PiggyBac repeats allow for the integration by piggybac transposase. e) Histograms showing the expression of mCherry and GFP as acquired by FACS in day 7 EB-derived cells, after 48 hours of doxycycline induction (10 µg/ml).

I transfected AEKO mouse ES cells with this plasmid and a piggyBac transposase plasmid and, using FACS to screen for mCherry expression, I selected four AEKO+dCas9-p300 clones. I subsequently attempted to induce dCas9-p300 expression, which I monitored via GFP expression. However, even after trying different timepoints of induction (24 and 48 hours), different doxycycline concentrations (0.1-12 mg/µl), and inducing both undifferentiated ES cells and EBs, I did not observe any GFP expression (Figure 5.7e).

Importantly, mCherry expression was very variable, especially in EB-derived cells. The PGK promoter, which controls the Tet-On and mCherry expression, might cause variable expression levels in differentiation, which has been noted in other projects (Ed Tunnacliffe, unpublished). However, I did not observe GFP expression even in cells with high levels of mCherry expression (Figure 5.7e).

The lack of GFP expression could be explained in two ways: the doxycycline induction is not efficient, or the construct has a mutation or is only partially integrated. As I initially confirmed the presence of the cassette by mCherry expression, I further verified the presence of various parts of the construct in the clones by PCR. Primers targeting the N-terminus of dCas9, GFP, mCherry, and the sgRNA cassette all returned DNA fragments of expected sizes (data not shown), except for a mutation in the sgRNA cassette in one clone. Partial integration of the construct therefore seems unlikely. However, it is possible that there is a mutation in the sequence. I used primer walking to verify most of the plasmid's sequence by Sanger sequencing, and there are only 16 out of 7318 bp of the dCas9:T2A:GFP open reading frame that require re-sequencing due to poor or ambiguous traces. A mutation in these could possibly lead to a premature stop codon which would prevent GFP expression. As for doxycycline induction, I have

tried many timepoints and concentrations, but it is possible this requires additional optimisation.

Discussion

Upon deletion of all five alpha-globin enhancer elements, almost all alpha-globin expression is lost. Small ATAC-seq peaks over the alpha-globin promoters remain, however the promoters are no longer marked by H3K4me3. Set1 complexes are responsible for the majority of trimethylation at H3K4, however, loss of this methylation only affects gene expression in small subset of cells. It is therefore unclear how instructive this modification is in transcriptional regulation or if it merely reflects transcriptional activity (Howe et al. 2017). Active enhancers are usually marked by H3K27ac, for which histone acetyl transferases such as p300 and CBP are responsible (Calo and Wysocka 2013). These histone acetyl transferases are recruited to enhancers via interactions with transcription factors, such as Gata1 (Blobel et al. 1998). It is therefore not surprising that in the absence of the enhancers H3K27ac is lost from the locus, with only a small peak within the zeta-globin gene remaining.

The promoters are also no longer bound by Gata1, Pol II or TBP. The loss of these proteins at the promoter is unexpected, because there is still an open chromatin site. It suggests that these transcription factors need more than an accessible motif at their usual binding site to be recruited and this is usually provided by the enhancers. It is also in contrast with the ΔR1R2 model, where some Pol II is still recruited to the promoters (Larke et al. 2021), and shows again that although R3, Rm and R4 are not classical enhancers like R1 and R2, they nevertheless contribute to enhancer function. Furthermore, these data suggest that the enhancers are functional in the very early stages of the transcription cycle, by recruiting or stabilising PIC factors at the promoter. And although there is still an open chromatin site at the enhancers, it is reduced, so the enhancers also play a role in creating a permissive environment.

Zeta-globin expression is only 60% reduced in the ΔR1R2 model, compared to 90% for alpha-globin (King et al. 2021). Similarly, although zeta-globin expression is very strongly reduced in the AEKO model, it is not as strongly affected as alpha-globin. It is therefore likely that the zeta-globin promoter plays a more important role in gene regulation than the alpha-globin promoters do. However, the chromatin accessibility is more strongly affected than for alpha-globin and binding of Gata1 is lost as well. A smaller H3K4me3 peak does remain.

Surprisingly, the expression of *Rhbdf1*, a non-erythroid gene upstream of the alpha-globin sub-TAD, is strongly affected in the AEKO model. In the AEKO model, its promoter is marked by the repressive H3K27me3 mark, which is also the case in definitive erythroid cells (Hanssen et al. 2017). When the CTCF binding sites between *Rhbdf1* and the enhancers are deleted, however, the *Rhbdf1* promoter becomes marked with both H3K27me3 and H3K4me3. Some Polycomb complex remains, but *Rhbdf1* is expressed in this model (Hanssen et al. 2017).

Since *Rhbdf1* is also expressed in non-erythroid tissues where the alpha-globin enhancers are not active, it is unlikely that the alpha-globin enhancers are the sole regulators of its expression. However, based on the data from the AEKO model, the alpha-globin enhancers do affect its expression in primitive EB-derived erythroid cells. The enhancers are capable of evicting Polycomb and responsible for creating a permissive environment at the *Rhbdf1* promoters. This could result either via a direct mechanism, in which the enhancers contact the promoter to affect chromatin changes, or via an indirect mechanism, in which the local concentration of transcriptional machinery is increased by the enhancers and also affects *Rhbdf1* expression.

Thus, the alpha-globin enhancers do seem to be capable of Polycomb eviction. This has been previously noted, in work on a humanised alpha-globin locus which showed that upon the deletion of MCS-R2, the human equivalent to the mouse R2 element, H3K27me3 and Polycomb proteins accumulated at the alpha-globin promoters (Vernimmen et al. 2011). I did not observe H3K27me3 at the alpha-globin promoters in the AEKO model. Polycomb group complexes are

present at the Human alpha-globin promoters in non-erythroid tissue (Garrick et al. 2008). The difference between mouse and human could be because the CpG islands, present at the human promoters, have eroded in mouse (Hughes et al. 2005).

It has been speculated that super-enhancers play a role in the tissue-specific recruitment of cohesin (Vos et al. 2021). NIPBL, which plays a role in cohesin loading, is present at the alpha-globin enhancers in erythroid cells (Hua et al. 2021). Furthermore, there are small Rad21 ChIP-seq peaks at the enhancer elements. These Rad21 peaks are lost in the AEKO model, but there is no great change in cohesin binding at the surrounding CTCF sites. Others have noted a small decrease of Rad21 at CTCF sites around a super-enhancer when that super-enhancer is deleted (Vos et al. 2021). As I only have one replicate, it is possible that the loss of Rad21 binding was too small to be picked up. However, as cohesin is present at the CTCF binding sites in mouse ES cells as well, it is also possible that there are other sites of cohesin loading in addition to the enhancer elements. These could be further away, as there are interactions across the tissue-invariant TAD that bring flanking CTCF binding sites together and are therefore likely mediated by cohesin (Oudelaar et al. 2020). In this light, it will be informative to investigate the 3D chromatin conformation in the absence of the enhancers using a 3C-based method. I have attempted Tiled-C (Oudelaar et al. 2020) in the AEKO model, but it was not successful using the EB-derived erythroid cells. It will be explored again in a mouse model of AEKO which is currently being analysed. A future Capture-C experiment could reveal if the 3D conformation is similar to mouse ES cells or if there is some tissue specificity even in the absence of the enhancers.

The AEKO model provides a unique opportunity to test mechanisms of enhancer function using dCas9 fusion proteins, because only the *cis*-regulatory elements have been removed while all the *trans*-acting factors remain. For instance, in HEK293T cells gamma-globin transcription could be activated from its enhancer region using dCas9-p300 (Hilton et al. 2015) and maybe this could be replicated in the AEKO model. The expression of gamma-globin increased by 269-fold, which

is a considerable increase, but it is more difficult to assess expression increases when the gene was not expressed in the tissue before. If we can recover alpha-globin expression in a similar manner in the AEKO model, it would be easier to compare the expression to wild-type erythroid cells.

Considering that R1 and R2, the main enhancers, are further from the TSS than the other enhancer elements, it could show how the distance from the TSS affects the transcriptional activation potential. Furthermore, it would be interesting to see if the binding of Gata1 or other factors at the promoter can be recovered. Finally, it would be informative to see if *Rhbdf1* recovers its expression when the alpha-globin genes are rescued by the recruitment of dCas9-p300 to their promoters. If this is the case, it would mean that an indirect regulatory mechanism, which does not involve the enhancers per se, but rather a site of exceptionally high transcription is responsible for the activation of Rhbdf1 in EB-derived erythroid cells.

Acknowledgements

The AEKO and WT hemizygous model were generated by Helena Francis in a collaboration with Jef Boeke (NYU Langone Medical Center). Kevin Clark (FACS facility, WIMM) performed single cell sorting of ES cells transfected with piggyBac and inducible dCas9-p300. Nicole Jackson (WIMM) provided the dCas9-p300 template plasmid.

Chapter 6: Discussion

Genome-wide studies that describe genomic features such as open chromatin, histone modifications and transcription factor binding have provided great insight into gene regulation (Dunham et al. 2012; Neph et al. 2012; Thurman et al. 2012). To better understand the mechanism behind what is observed in genome-wide descriptive studies, many recent studies have perturbed proteins and protein complexes involved in gene regulation, particularly those involved in the regulation of the 3D structure of the genome (Busslinger et al. 2017; Haarhuis et al. 2017; Nora et al. 2017; Wutz et al. 2017). The effects of these perturbations have been studied genome-wide, however, this approach has limitations. By perturbing *trans*-acting factors that play an important role in correct gene expression, it is likely that the expression of a substantial number of genes will be affected and this can confound interpretation of the original perturbation. Furthermore, it is possible but difficult to mutate several factors in combination, as this disrupts genome-wide expression and cell viability further. However, this can be valuable to really understand the interplay between different factors regulating gene expression. Finally, changes in *trans*-acting factors can also cause changes in cell type, for instance ES cells can start differentiating along inappropriate lineages, which further complicates interpreting the effect of a specific perturbation.

It is therefore important to complement genome-wide studies with studies focussing on a single, appropriate, well-characterised locus, in which contributions of different *cis*-acting elements can be carefully dissected leading to better understanding of their interplay and individual contributions. This approach has brought valuable insight in the past, which has shown that careful and thorough characterisation can allow for enhanced interpretation of the resulting phenotype (Lupianez et al. 2015; Hay et al. 2016; Hanssen et al. 2017; Vos et al. 2021). When establishing a model for understanding the roles of *cis*-acting elements, it is equally important to investigate a gene which has no downstream targets. From this point of view the globin genes

are ideal. It is possible to mutate these genes with no downstream effect and no matter what the mutation, cells proceed along the erythroid lineage and differentiate into mature red cells.

A single CTCF binding site can reduce enhancer-promoter interactions

In this thesis, I have used several models in which *cis*-acting elements have been genetically modified to study their contribution to gene expression at the alpha-globin locus. First, I asked if a single CTCF binding site could effectively perturb enhancer-promoter interactions at this locus, by inserting it between the alpha-globin genes and their enhancers. I initially studied a CTCF site (HS-38) which was previously shown to have the strongest insulator function in the alpha cluster (Hanssen et al. 2017). This inserted HS-38 CTCF site alters the 3D conformation of the locus in an orientation-dependent manner and reduces alpha-globin expression, demonstrating that enhancer-promoter interactions are essential for the alpha-globin enhancers' function. The strongest effect on expression was observed when the CTCF site is oriented to block loop extrusion coming from the direction of the alpha-globin enhancers.

As the expression of alpha-globin was strongly affected in these CTCF insertion models, we decided to exploit this by developing an assay to test the insulating potential of any chosen CTCF binding site. By tagging one of the alpha-globin genes with YFP, we created a system in which CTCF sites can be tested in a high-throughput fashion. I showed that this strategy is viable and indeed reports different insulator strengths for different CTCF sites.

CTCF binding sites have different insulating potential

The advantage of this assay lies in keeping the chromatin environment constant and only altering a very small sequence (83 bp), thereby separating sequence contributions from other factors affecting insulator potential. Part of the varying insulator potential that is observed in native CTCF site deletions might be explained by the genomic position of a CTCF binding site, as illustrated at the alpha-globin locus and elsewhere (Hanssen et al. 2017; Harrold et al. 2020; Zhang et al. 2020). For example, it is possible that HS-38 is recognised as a strong insulator, not only because of its sequence but also because *Rhbdf1* and *Mpg* are sensitive to the alpha-globin

enhancers. Deletion of the downstream CTCF sites (θ1, θ2, HS+44 and H+48) led to a shift of enhancer chromatin interactions past the deleted CTCF sites to the next CTCF sites, in accordance with loop extrusion (Susannah Holliman, unpublished data; (Harrold et al. 2020)). However, it did not lead to an upregulation of downstream genes (*Sh3pxd2b* and *Ubtd2*). This could be because some intrinsic properties of these CTCF binding sites make them genuinely weaker insulators. However, it is also possible that these genes (*Sh3pxd2b* and *Ubtd2*) are not sensitive to the alpha-globin enhancers, as in some instances enhancers have been shown to preferentially activate some promoters over others (Zabidi et al. 2015; Martinez-Ara et al. 2021). It is also possible that there are elements present at this site that do not bind CTCF but do function as a boundary, such as the alpha-globin genes (Harrold et al. 2020).

In summary, there are several reasons that these CTCF sites are not considered strong insulators and some of these depend on their genomic position rather than binding site characteristics. In summary, for a CTCF binding site to be considered an insulator, it requires the presence of a suitable enhancer-promoter pair and be placed at a position where it encounters loop extruding cohesin. These conditions for a CTCF binding site to be recognised as an insulator are all met in the CTCF binding site assay developed here.

In addition to its genomic position relative to other genomic elements, the binding of CTCF itself is regulated by the local chromatin context. DNA methylation may prevent CTCF from binding (Hark et al. 2000) and CTCF binding also requires nucleosomal spacing induced by a chromatin remodeller SNF2H (Wiechens et al. 2016; Barisic et al. 2019). Furthermore, there are also factors that actively compete with CTCF for its binding sites (Pugacheva et al. 2015; Kaaij et al. 2019; Nishana et al. 2020). The CTCF insertion assay developed here allows us to keep all these extrinsic factors constant, thereby allowing direct comparison of sequence-intrinsic insulator function between different CTCF binding sites.

CTCF binding sites have widely different ChIP-seq intensity, which is a proxy for binding site strength or percentage of cells in which that particular site is bound. Binding site strength is largely determined by the sequence motif (Nakahashi et al. 2013), and is also likely to affect insulator potential. Generally, the binding site strength of a CTCF site increases when its sequence closely resembles the core motif. The presence of the upstream motif also greatly increases the binding site strength (Nakahashi et al. 2013). Furthermore, several experimental observations correlate with CTCF binding sites that are strong boundaries, such as conservation across species and tissue types (Schmidt et al. 2012) and persistent binding after auxin or RNA interference treatment (Schmidt et al. 2012; Khoury et al. 2020; Luan et al. 2021). However, it is not clear if these just select binding sites with exceptional stable binding, which could already explain increased insulator function, or if there are any additional features in these binding site sequences. It has also been suggested that CTCF ChIP-seq peaks with multiple CTCF motifs could contribute to binding site strength and that this might also contribute to insulator function (Chang et al. 2021).

In the future, using the CTCF reporter assay, we (together with Felice Tsang) plan to investigate the sequence contribution to insulator potential. We are asking if insulator potential is just based on binding site strength or if there are any additional features in the binding site motif that specifically contributes to insulator potential. For instance, it would be informative if a binding site with a strong core motif is a better insulator than a binding site with an upstream motif but a weaker core motif. It has been speculated in the past that the diverse sequences that are recognised by CTCF could account for its many different functions (Nakahashi et al. 2013).

Studying aspects of enhancer regulation at the alpha-globin locus

The CTCF insertion model can also provide insight into enhancer function. The importance of enhancer-promoter proximity for gene regulation by enhancers has recently been debated (Alexander et al. 2019; Benabdallah et al. 2019), even though it was previously regarded as a

fundamental requirement (Spitz and Furlong 2012; Pennacchio et al. 2013). Enhancer-promoter interactions are reduced in both the HS-38 Forward and Reverse model, but more strongly in the Reverse model. This corresponds to changes in alpha-globin expression, indicating that enhancer-promoter interactions are important for enhancer function at the alpha-globin locus. However, some enhancer-promoter interactions still exist, so to better understand how important they are, future experiments will include the introduction of multiple CTCF binding sites (e.g. 2-4 copies of HS-38). Previous studies have shown that this would likely reduce these interactions further (Chang et al. 2021; Huang et al. 2021). If this in turn leads to even lower alpha-globin expression, it could give an indication of how much the regulation of gene expression relies on the close proximity between enhancers and promoters.

The effect on alpha-globin expression was mild in the CTCF insertion models compared to the model where all enhancer elements were deleted. In the Reverse CTCF insertion model the expression is reduced by 2.8-fold, whereas the expression in the AEKO model is 750-fold lower. This is probably due to various reasons: as discussed previously, the single CTCF binding site does not restrict enhancer-promoter interactions completely. However, it is possible that enhancer-promoter interactions cannot be completely abolished, as CTCF sites are not likely to form absolute boundaries. Importantly, even if any tracking from the enhancers to the promoters is prevented, there might still be mechanisms by which the enhancers can regulate expression that are independent of tracking (e.g. diffusion or long range effects on chromatin accessibility). This can be studied further in the multiple CTCF insertion model by investigating elements of gene regulation in parallel with the AEKO model, such as the recruitment of transcription factors to the promoters.

Characterisation of the AEKO model has uncovered that the enhancers play a pivotal role in activating the alpha-globin locus. In their absence, binding of Gata1 and TBP is lost from the promoter, even though the promoter regions retain a small ATAC-seq peak indicating some

accessibility. H3K4me3, normally found at the promoters, is lost as well. Although it is not known if Gata1 binding at the promoter is necessary for expression, it is nonetheless remarkable that it should be lost when its binding site still accessible. It is also in contrast to earlier studies of the human alpha-globin locus, where some transcription factors were still bound to the promoter in absence of the enhancers (Vernimmen et al. 2007). TBP is known to be important for transcription, and is lost from the promoter as well. Together, this indicates that the enhancers are functional very early stages of transcriptional activation, as the binding of several factors at the promoter depends on their presence. Potentially, they function in altering the chromatin environment such that Gata1 can bind to the promoters.

Further experiments might uncover what is binding at the small open chromatin region which remains at the alpha-globin promoter in the absence of the enhancers, and could give insight in protein recruitment to the promoters in absence of the enhancers. There are many factors that could be investigated by ChIP, such as Klf1 which is known to bind at the enhancer regions. However, ChIP is limited by the work involved in probing many factors as well as the availability and quality of antibodies. Single locus proteomics, in which proteins associated with a locus are purified and analysed by mass spectrometry, has the advantage of being unbiassed and probing many proteins simultaneously. There have been some promising methods published for single locus proteomics (Gao et al. 2018; Myers et al. 2018), however these have not worked in our hands.

It would also be interesting to study the cell distribution of the activity I observe, such as the open chromatin site or the basal expression levels. It is possible that the promoters are briefly accessible in many cells, or if there are a few cells that have a more stably activated locus. In future, we can therefore perform single cell RNA-seq or RNA FISH to see if the expression, especially zeta-globin expression, is present at a low level in many cells or if there are a few highly expressing cells.

The AEKO also provides a blank slate in which all *trans*-acting factors are present but the *cis*-regulatory elements are removed. To better understand what the enhancers bring to the locus to activate transcription, I propose to use a dCas9-mediated recruitment strategy, in which different transcription factors or co-activators are recruited to different sites at the alpha-globin locus. It could be used to test the hierarchy of recruitment (i.e. which factors are able to recruit others) to the enhancers and whether this recruitment is necessary for gene activation. Also, as I observed that cohesin is no longer present at the promoters, it would be interesting to see if cohesin recruitment is recovered when the locus is artificially activated in this way.

Although it has been observed that enhancers and promoters can have overlapping activities (Andersson and Sandelin 2020), the alpha-globin enhancers and promoters appear to act differently. The enhancers do express enhancer RNAs (Kowalczyk et al. 2012), however there is no clear H3K4me3 peak in EB Chipmentation data. Furthermore, the promoters seem to have very little activity in the absence of the enhancers. In contrast, previous work has shown that the enhancers still recruit Pol II in the absence of the promoters, although it is reduced (Vernimmen et al. 2007). It has previously been observed that promoters of genes active in few tissue types have weaker independent promoter activity (Van Arensbergen et al. 2017). It was also observed that if more enhancer elements were present near (<50 kb) a promoter, the weaker the promoter tended to be. These two observations fit the alpha-globin promoters well and can therefore at least partly account for the lack of independent promoter activity I observed, although it has to be noted that 5 enhancer elements is on the lower half of the range of enhancer elements observed in this study.

Loading of cohesin and loop extrusion dynamics

Loop extrusion dynamics, such as cohesin loading, extrusion speed, and cohesin dissociation, are likely to have a strong influence on gene expression and chromatin 3D structure, and are still insufficiently understood. Interestingly, in the Forward and Reverse CTCF insertion models, I observed a difference in cohesin accumulation at the ectopic CTCF site, with cohesin

accumulating specifically at the insertion site only when it is oriented towards the enhancers. This is likely to be functionally relevant, as the expression of alpha-globin is also more affected in the Reverse model.

The different levels of cohesin in Forward and Reverse models make cohesin-mediated loop extrusion a plausible mechanism behind the different expression levels in the Forward and Reverse model. CTCF is known to exert its function by blocking cohesin, forming a specific interaction to stabilise cohesin on the DNA (Li et al. 2020c; Pugacheva et al. 2020). However, it is conceivable that another mechanism causes the changes in 3D chromatin structure and gene expression. This mechanism would likely involve tracking along the chromatin fibre, otherwise the importance of binding site directionality would be difficult to explain. There could be some other proteins however involved such as those in the Mediator complex or Ldb1. It is therefore important to verify the central role cohesin plays in our current understanding of the regulation of gene expression at the alpha-globin locus, which could be achieved by depleting cohesin. If cohesin depletion reduces alpha-globin expression in wild-type erythroid cells it is an important indication that the enhancers depend on cohesin for their action. Moreover, it would be interesting to see if the different effects on gene expression in the Forward and Reverse model is lost in the absence of cohesin. If this is the case, then the expression loss between wild-type and CTCF insertion cells would be expected to be the same, unless CTCF has an additional mechanism by which it affects enhancer function.

The difference in cohesin accumulation between the Forward and Reverse model is unlikely to be based on the inserted sequence, as the sequences are the same except for an extra pair of lox sites in the Forward model. Moreover, in the CTCF insertion assay, where we have tested several binding site sequences, we consistently observed the same trend: the orientation of CTCF sites towards the enhancers produces a stronger insulator than when they are orientated towards the genes. It is most likely that this difference reflects cohesin dynamics at the locus,

either by cohesin being loaded at specific sites or preferentially extruding in a specific direction. NIPBL is responsible for cohesin loading, and it binds to the alpha-globin enhancers as well as the genes (Figure 6.1). If extrusion by cohesin is bi-directional, as has been observed *in vitro* (Kim et al. 2019), then the forward-oriented site will preferentially block cohesin loaded at the genes, whereas in the reverse orientation cohesin from the enhancers would be preferentially blocked. From the distribution of NIPBL, it seems likely that both sites encounter high levels of cohesin. The directionality of loop extrusion could therefore underlie the different cohesin levels at the inserted CTCF site in the Forward and Reverse models.

Cohesin loading via NIPBL has been proposed to change depending on chromatin activity. It could explain why 3D chromatin structure varies widely in different cell types while CTCF binding is less variable (Hua et al. 2021). Several observations fit with this idea: cohesin sites that do not overlap with CTCF sites are often cell type specific (Pugacheva et al. 2020), and NIPBL often colocalises with active promoters and enhancers. It has also been observed that when a super-enhancer was deleted, cohesin levels at the surrounding CTCF sites decreased, although the effect was moderate (Vos et al. 2021). I therefore expected to see an effect on cohesin levels in the AEKO model. However, with the caveat that I only have one replicate of the Rad21 ChIP, there was no clear decrease in cohesin at the surrounding CTCF sites. There was no cohesin at the enhancer region and at the alpha-globin promoters, but the general ChIP-seq profile was similar to that of mouse ES cells. This could mean that the cohesin levels we observe at CTCF binding sites depend more on properties of those particular binding sites, such as binding stability, than on the levels of cohesin they encounter. If this is the case, the cohesin observed at the CTCF sites in the AEKO might have been loaded far away (>100 kb), and the existence of relatively long-range interactions indicates this could be the case. However, this is not consistent with the CTCF insertion models, as then cohesin would be expected in equal levels at both the forward and reverse orientated site.

The fact that CTCF binding sites are largely invariant at the alpha-globin locus, while 3D conformation changes drastically during erythropoiesis, has previously been explained by loop extrusion dynamics (Hua et al. 2021). High resolution chromatin interaction data, acquired using micrococcal nuclease in combination with Capture-C, shows that proximity of CTCF binding sites to other CTCF binding sites alters during differentiation. Presumably this is because cohesin dynamics have changed.

Finally, cohesin unloading from the chromatin fibre by WAPL is another layer of loop extrusion dynamics. This has remained largely unexplored in our system, however, previously published data in mouse embryonic fibroblasts shows that, in that tissue, WAPL is associated with a subset of CTCF binding site (Figure 6.1; (Busslinger et al. 2017)). Another WAPL ChIP-seq dataset, from mouse ES cells, also shows enrichment of WAPL at mainly the same CTCF sites (Liu et al. 2021). In this dataset, there are some additional WAPL peaks at promoters. Possibly these CTCF binding sites with WAPL binding are more effective boundaries, as cohesin that dissociates from these CTCF binding site would be have a higher change of being unloaded from the chromatin fibre, as opposed to continuing loop extrusion past the boundary. It is important to note that these ChIP data is not erythroid specific, and WAPL expression levels fluctuate during haematopoietic differentiation (Hill et al. 2020), and therefore binding patterns could be different in erythroid cells. There are also differences between the two tissue types that these ChIPs have been acquired in. Performing a WAPL ChIP-seq in erythroid tissue would therefore contribute to better understanding of loop extrusion dynamics at the alpha-globin locus. Furthermore, a WAPL ChIP in the Forward and Reverse model would clarify if the binding of WAPL is encoded in the binding site sequence of CTCF, the sequence directly flanking the CTCF binding site or is imposed by the wider chromosomal context.

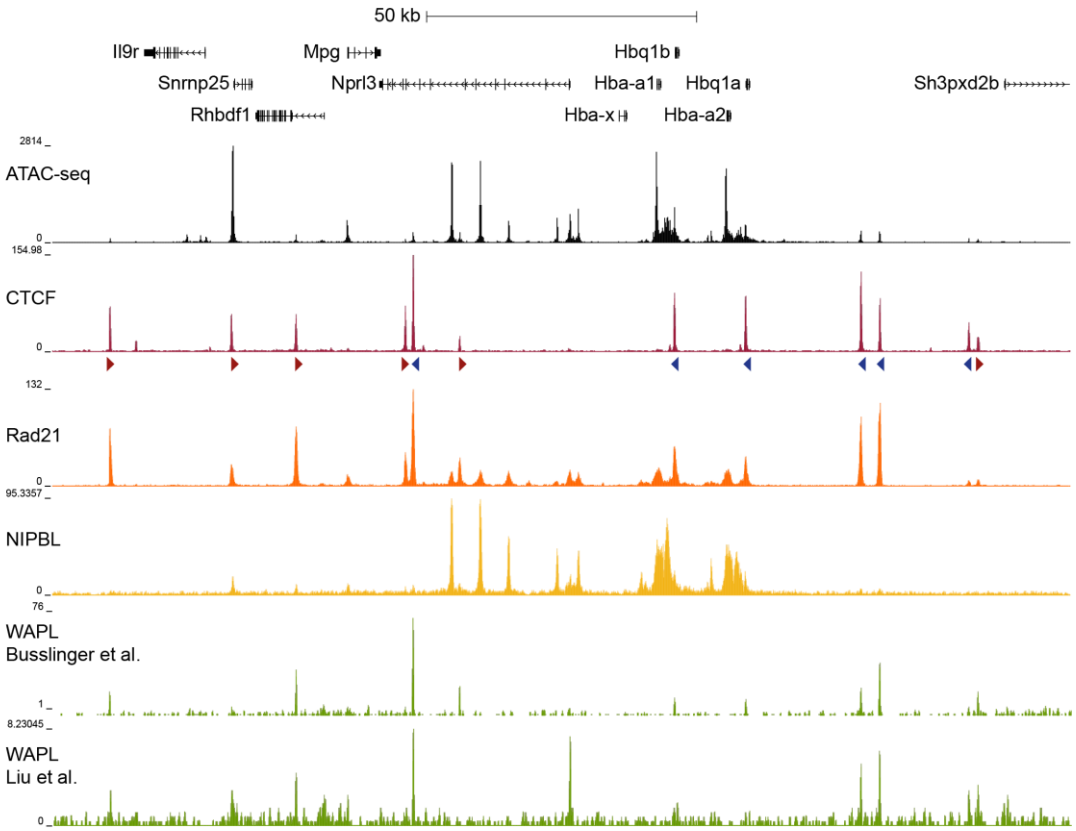


Figure 6.1: CTCF, Rad21, NIPBL and WAPL ChIP-seq tracks
ATAC-seq track in WT foetal liver-derived erythroid cells, CTCF ChIP-seq track in spleen-derived erythroid cells, Rad21 ChIP-seq track in spleen-derived erythroid cells and NIPBL ChIP-seq track in spleen-derived erythroid cells. CTCF, Rad21 and NIPBL ChIP as previously published (Hua et al. 2021). WAPL ChIP-seq track in primary quiescent mouse embryonic fibroblast, previously published by the Jan-Michael Peters lab (Busslinger et al. 2017). WAPL ChIP-seq track in mouse ES cells, previously published by the lab of Elzo de Wit (Liu et al. 2021).

Rhbdf1 expression

Surprisingly, in addition to expression changes at the alpha-globin locus, I also observed changes in non-erythroid genes flanking the alpha-globin locus. The expression of *Rhbdf1* is most strongly affected. *Rhbdf1* is expressed in primitive erythroid cells, such as those derived from day 7 EBs, but is repressed via a Polycomb-mediated mechanism in definitive erythroid cells, such as those derived from spleen and foetal liver (Hanssen et al. 2017). It is also expressed in some non-erythroid cell types, such as mouse ES cells.

In previously characterised deletions of the CTCF binding sites that form the 5' boundary of the sub-TAD, *Rhbdf1* retains Polycomb complexes at its promoter but its expression goes up by 600-

fold. This demonstrates that the 5' CTCF sites usually prevent the enhancers from affecting *Rhbdf1* expression (Hanssen et al. 2017). A previous study showed that interactions between the alpha-globin promoters, enhancers and *Rhbdf1* promoter increase when the 5' sub-TAD boundary is deleted. This did not affect interactions between the alpha-globin promoters and enhancers, indicating that there is no competition between these promoters for access to the enhancers (Oudelaar et al. 2019). In the Reverse CTCF insertion model, I observed an increase in *Rhbdf1* expression with the 5' sub-TAD boundary intact. The upregulation was not as profound as observed when two of the CTCF sites were deleted, but was still considerable at 60-fold in spleen-derived erythroid cells. It is not clear by which mechanism this occurs, but it seems that when the enhancers are prevented from interacting with the alpha-globin promoters by the CTCF insertion, the 5' CTCF sites become a less robust boundary.

The AEKO model was analysed in EB-derived cells, which resemble primitive erythroid cells and thus express *Rhbdf1*. Surprisingly, in the absence of the enhancers, *Rhbdf1* becomes repressed and its promoter marked by H3K27me3. The expression of *Rhbdf1* is dependent on the enhancers in this cell type, and these prevent Polycomb-mediated repression. In that way, the 5' CTCF sites seem to play opposing roles in primitive and definitive tissue and are not an absolute boundary in primitive cells. It could be that binding site properties underlie this different behaviour in different erythroid cell types. However, the main CTCF site of this boundary, HS-38, behaves as a stronger boundary in primitive erythroid cells when ectopically inserted. This is the opposite behaviour from its native site, so it is not likely caused by the binding site sequence.

Conclusions and further directions

I have shown that the insertion of a single CTCF binding site can alter the 3D chromatin structure at the alpha-globin locus in an orientation-dependent manner, and that this reduces alpha-globin expression considerably. Furthermore, I have characterised a model in which all alpha-

globin enhancer elements have been deleted, and shown the enhancers play a pivotal role in gene activation.

To test if cohesin is responsible for the different effect of the Forward and Reverse CTCF insertion model, I propose a cohesin depletion experiment in these models and wild-type erythroid cells. WAPL, which is responsible for the unloading of cohesin, has not been investigated yet in the context of the alpha-globin locus. Performing WAPL ChIP-seq will provide better understanding of loop extrusion dynamics at the alpha-globin locus, and can indicate if its presence at CTCF binding sites is encoded in the genome. Finally, I have shown that the CTCF insertion site can be used for a CTCF binding site assay and this will provide insight into how binding site sequence contributes to insulator potential.

Another aim is to understand enhancer function better. To test how much this relies on chromatin interactions between the enhancers and promoters, I propose a to study models in which multiple CTCF binding sites are inserted, which is likely to restrict interactions further. Finally, employing a dCas9-p300 strategy will provide insights into which factors are capable of recovering transcription and other aspects of gene regulation in the all enhancer knockout model, such as transcription factor binding at the promoters.

References

Adelman K, Lis JT. 2012. Promoter-proximal pausing of RNA polymerase II: emerging roles in metazoans. *Nature reviews Genetics* **13**: 720-731.

Alexander JM, Guan J, Li B, Maliskova L, Song M, Shen Y, Huang B, Lomvardas S, Weiner OD. 2019. Live-cell imaging reveals enhancer-dependent Sox2 transcription in the absence of enhancer proximity. *Elife* **8**.

Allen BL, Taatjes DJ. 2015. The Mediator complex: a central integrator of transcription. *Nat Rev Mol Cell Biol* **16**: 155-166.

Anania C, Acemel RD, Jedamzick J, Bolondi A, Cova G, Brieske N, Kühn R, Wittler L, Real FM, Lupiáñez DG. 2021. In vivo dissection of a clustered-CTCF domain boundary reveals developmental principles of regulatory insulation. *Biorxiv*.

Andersson R, Gebhard C, Miguel-Escalada I, Hoof I, Bornholdt J, Boyd M, Chen Y, Zhao X, Schmidl C, Suzuki T et al. 2014. An atlas of active enhancers across human cell types and tissues. *Nature* **507**: 455-461.

Andersson R, Sandelin A. 2020. Determinants of enhancer and promoter activities of regulatory elements. *Nat Rev Genet* **21**: 71-87.

Azuara V, Perry P, Sauer S, Spivakov M, Jørgensen HF, John RM, Gouti M, Casanova M, Warnes G, Merkenschlager M et al. 2006. Chromatin signatures of pluripotent cell lines. *Nat Cell Biol* **8**: 532-538.

Bajic VB, Tan SL, Christoffels A, Schönbach C, Lipovich L, Yang L, Hofmann O, Kruger A, Hide W, Kai C et al. 2006. Mice and men: their promoter properties. *PLoS genetics* **2**: e54-e54.

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

Barisic D, Stadler MB, Iurlaro M, Schübeler D. 2019. Mammalian ISWI and SWI/SNF selectively mediate binding of distinct transcription factors. *Nature* **569**: 136-140.

Barrington C, Georgopoulou D, Pezic D, Varsally W, Herrero J, Hadjur S. 2019. Enhancer accessibility and CTCF occupancy underlie asymmetric TAD architecture and cell type specific genome topology. *Nature Communications* **10**: 2908.

Barutcu AR, Maass PG, Lewandowski JP, Weiner CL, Rinn JL. 2018. A TAD boundary is preserved upon deletion of the CTCF-rich Firre locus. *Nature Communications* **9**: 1444.

Baumann M, Pontiller J, Ernst W. 2010. Structure and basal transcription complex of RNA polymerase II core promoters in the mammalian genome: an overview. *Mol Biotechnol* **45**: 241-247.

Beagan JA, Duong MT, Titus KR, Zhou L, Cao Z, Ma J, Lachanski CV, Gillis DR, Phillips-Cremins JE. 2017. YY1 and CTCF orchestrate a 3D chromatin looping switch during early neural lineage commitment. *Genome Research* **27**: 1139-1152.

Beagan JA, Phillips-Cremins JE. 2020. On the existence and functionality of topologically associating domains. *Nat Genet* **52**: 8-16.

Beagrie RA, Pombo A. 2016. Gene activation by metazoan enhancers: Diverse mechanisms stimulate distinct steps of transcription. *Bioessays* **38**: 881-893.

Beagrie RA, Scialdone A, Schueler M, Kraemer DC, Chotalia M, Xie SQ, Barbieri M, de Santiago I, Lavitas LM, Branco MR et al. 2017. Complex multi-enhancer contacts captured by genome architecture mapping. *Nature* **543**: 519-524.

Bell AC, Felsenfeld G. 2000. Methylation of a CTCF-dependent boundary controls imprinted expression of the Igf2 gene. *Nature* **405**: 482-485.

Bell AC, West AG, Felsenfeld G. 1999. The protein CTCF is required for the enhancer blocking activity of vertebrate insulators. *Cell* **98**: 387-396.

Benabdallah NS, Williamson I, Illingworth RS, Kane L, Boyle S, Sengupta D, Grimes GR, Therizols P, Bickmore WA. 2019. Decreased Enhancer-Promoter Proximity Accompanying Enhancer Activation. *Molecular cell* **76**: 473-484 e477.

Bintu B, Mateo LJ, Su JH, Sinnott-Armstrong NA, Parker M, Kinrot S, Yamaya K, Boettiger AN, Zhuang X. 2018. Super-resolution chromatin tracing reveals domains and cooperative interactions in single cells. *Science* **362**.

Blair RH, Goodrich JA, Kugel JF. 2012. Single-molecule fluorescence resonance energy transfer shows uniformity in TATA binding protein-induced DNA bending and heterogeneity in bending kinetics. *Biochemistry* **51**: 7444-7455.

Blobel GA, Higgs DR, Mitchell JA, Notani D, Young RA. 2021. Testing the super-enhancer concept. *Nat Rev Genet* **22**: 749-755.

Blobel GA, Nakajima T, Eckner R, Montminy M, Orkin SH. 1998. CREB-binding protein cooperates with transcription factor GATA-1 and is required for erythroid differentiation. *Proceedings of the National Academy of Sciences of the United States of America* **95**: 2061-2066.

Boeing S, Rigault C, Heidemann M, Eick D, Meisterernst M. 2010. RNA Polymerase II C-terminal Heptarepeat Domain Ser-7 Phosphorylation Is Established in a Mediator-dependent Fashion. *Journal of Biological Chemistry* **285**: 188-196.

Bonchuk A, Kamalyan S, Mariasina S, Boyko K, Popov V, Maksimenko O, Georgiev P. 2020. N-terminal domain of the architectural protein CTCF has similar structural organization and ability to self-association in bilaterian organisms. *Sci Rep* **10**: 2677.

Boyle AP, Davis S, Shulha HP, Meltzer P, Margulies EH, Weng Z, Furey TS, Crawford GE. 2008. High-resolution mapping and characterization of open chromatin across the genome. *Cell* **132**: 311-322.

Bozhilov YK, Downes DJ, Telenius J, Marieke Oudelaar A, Olivier EN, Mountford JC, Hughes JR, Gibbons RJ, Higgs DR. 2021. A gain-of-function single nucleotide variant creates a new promoter which acts as an orientation-dependent enhancer-blocker. *Nat Commun* **12**: 3806.

Brant L, Georgomanolis T, Nikolic M, Brackley CA, Kolovos P, van Ijcken W, Grosveld FG, Marenduzzo D, Papantonis A. 2016. Exploiting native forces to capture chromosome conformation in mammalian cell nuclei. *Molecular systems biology* **12**: 891.

Breathnach R, Chambon P. 1981. Organization and expression of eucaryotic split genes coding for proteins. *Annu Rev Biochem* **50**: 349-383.

Brown CE, Lechner T, Howe L, Workman JL. 2000. The many HATs of transcription coactivators. *Trends in Biochemical Sciences* **25**: 15-19.

Brown JM, Roberts NA, Graham BJ, Waithe D, Lagerholm C, Telenius JM, De Ornellas S, Oudelaar AM, Scott C, Szczerbal I et al. 2018. A tissue-specific self-interacting chromatin domain forms independently of enhancer-promoter interactions. *Nature Communications* **9**: 3849.

Buenrostro JD, Giresi PG, Zaba LC, Chang HY, Greenleaf WJ. 2013. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nature Methods* **10**: 1213-1218.

Bulger M, Sawado T, Schübeler D, Groudine M. 2002. ChIPs of the β-globin locus: unraveling gene regulation within an active domain. *Current Opinion in Genetics & Development* **12**: 170-177.

Busslinger GA, Stocsits RR, 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* **544**: 503-507.

Calhoun VC, Stathopoulos A, Levine M. 2002. Promoter-proximal tethering elements regulate enhancer-promoter specificity in the Drosophila Antennapedia complex. *Proc Natl Acad Sci U S A* **99**: 9243-9247.

Calo E, Wysocka J. 2013. Modification of enhancer chromatin: what, how, and why? *Molecular cell* **49**: 825-837.

Chang L-H, Ghosh S, Papale A, Miranda M, Piras V, Degrouard J, Poncelet M, Lecouvreur N, Bloyer S, Leforestier A et al. 2021. A complex CTCF binding code defines TAD boundary structure and function. *Biorxiv*.

Chen B, Gilbert Luke A, Cimini Beth A, Schnitzbauer J, Zhang W, Li G-W, Park J, Blackburn Elizabeth H, Weissman Jonathan S, Qi Lei S et al. 2013. Dynamic Imaging of Genomic Loci in Living Human Cells by an Optimized CRISPR/Cas System. *Cell* **155**: 1479-1491.

Chen K, Liu J, Heck S, Chasis JA, An X, Mohandas N. 2009. Resolving the distinct stages in erythroid differentiation based on dynamic changes in membrane protein expression during erythropoiesis. *Proc Natl Acad Sci U S A* **106**: 17413-17418.

Cheng AW, Wang H, Yang H, Shi L, Katz Y, Theunissen TW, Rangarajan S, Shivalila CS, Dadon DB, Jaenisch R. 2013. Multiplexed activation of endogenous genes by CRISPR-on, an RNA-guided transcriptional activator system. *Cell Research* **23**: 1163-1171.

Chung JH, Bell AC, Felsenfeld G. 1997. Characterization of the chicken beta-globin insulator. *Proc Natl Acad Sci U S A* **94**: 575-580.

Conaway RC, Conaway JW. 2011. Origins and activity of the Mediator complex. *Seminars in cell & developmental biology* **22**: 729-734.

Cong L, Ran FA, Cox D, Lin S, Barretto R, Habib N, Hsu PD, Wu X, Jiang W, Marraffini LA et al. 2013. Multiplex Genome Engineering Using CRISPR/Cas Systems. *Science* **339**: 819-823.

Core LJ, Martins AL, Danko CG, Waters CT, Siepel A, Lis JT. 2014. Analysis of nascent RNA identifies a unified architecture of initiation regions at mammalian promoters and enhancers. *Nature Genetics* **46**: 1311-1320.

Creyghton MP, Cheng AW, Welstead GG, Kooistra T, Carey BW, Steine EJ, Hanna J, Lodato MA, Frampton GM, Sharp PA et al. 2010. Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proc Natl Acad Sci U S A* **107**: 21931-21936.

Cullen KE, Kladde MP, Seyfred MA. 1993. Interaction between transcription regulatory regions of prolactin chromatin. *Science* **261**: 203-206.

Dao LTM, Galindo-Albarrán AO, Castro-Mondragon JA, Andrieu-Soler C, Medina-Rivera A, Souaid C, Charbonnier G, Griffon A, Vanhille L, Stephen T et al. 2017. Genome-wide characterization of mammalian promoters with distal enhancer functions. *Nature Genetics* **49**: 1073-1081.

Das AT, Tenenbaum L, Berkhout B. 2016. Tet-On Systems For Doxycycline-inducible Gene Expression. *Curr Gene Ther* **16**: 156-167.

Davies JO, Oudelaar AM, Higgs DR, Hughes JR. 2017. How best to identify chromosomal interactions: a comparison of approaches. *Nature methods* **14**: 125-134.

Davies JOJ, Telenius JM, McGowan SJ, Roberts NA, Taylor S, Higgs DR, Hughes JR. 2016. Multiplexed analysis of chromosome conformation at vastly improved sensitivity. *Nature methods* **13**: 74-80.

De Laat W, Duboule D. 2013. Topology of mammalian developmental enhancers and their regulatory landscapes. *Nature* **502**: 499-506.

De Santa F, Barozzi I, Mietton F, Ghisletti S, Polletti S, Tusi BK, Muller H, Ragoussis J, Wei C-L, Natoli G. 2010. A large fraction of extragenic RNA pol II transcription sites overlap enhancers. *PLoS biology* **8**: e1000384-e1000384.

De Wit E, De Laat W. 2012. A decade of 3C technologies: insights into nuclear organization. *Genes & Development* **26**: 11-24.

De Wit E, Vos ES, Holwerda SJ, Valdes-Quezada C, Verstegen MJ, Teunissen H, Splinter E, Wijchers PJ, Krijger PH, de Laat W. 2015. CTCF Binding Polarity Determines Chromatin Looping. *Molecular cell* **60**: 676-684.

Dean A, Larson DR, Sartorelli V. 2021. Enhancers, gene regulation, and genome organization. *Genes & Development* **35**: 427-432.

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

Despang A, Schöpflin R, Franke M, Ali S, Jerković I, Paliou C, Chan W-L, Timmermann B, Wittler L, Vingron M et al. 2019. Functional dissection of the Sox9–Kcnj2 locus identifies nonessential and instructive roles of TAD architecture. *Nature Genetics* **51**: 1263-1271.

Diao Y, Fang R, Li B, Meng Z, Yu J, Qiu Y, Lin KC, Huang H, Liu T, Marina RJ et al. 2017. A tiling-deletion-based genetic screen for cis-regulatory element identification in mammalian cells. *Nature Methods* **14**: 629-635.

Dixon JR, Selvaraj S, Yue F, Kim A, Li Y, Shen Y, Hu M, Liu JS, Ren B. 2012. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485**: 376-380.

Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. 2013. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**: 15-21.

Downes DJ, Gosden ME, Telenius J, Carpenter SJ, Nussbaum L, De Ornellas S, Sergeant M, Eijsbouts CQ, Schwessinger R, Kerry J et al. 2020. Targeted high-resolution chromosome conformation capture at genome-wide scale. *biorxiv*: 2020.2003.2002.953745.

Dunham I Kundaje A Aldred SF Collins PJ Davis CA Doyle F Epstein CB Frietze S Harrow J Kaul R et al. 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**: 57-74.

Francis HS. 2019. Functional dissection of a single enhancer at the mouse α-globin locus. in *Weatherall Institute of Molecular Medicine*, p. 251. University of Oxford, Oxford.

Francis HS, Harold CL, Beagrie RA, King AJ, Gosden ME, Blayney JW, Jeziorska DM, Babbs C, Higgs DR, Kassouf MT. 2022. Scalable in vitro production of defined mouse erythroblasts. *PloS one* **17**: e0261950.

Frankel N, Davis GK, Vargas D, Wang S, Payre F, Stern DL. 2010. Phenotypic robustness conferred by apparently redundant transcriptional enhancers. *Nature* **466**: 490-493.

Fudenberg G, Imakaev M, Lu C, Goloborodko A, Abdennur N, Mirny Leonid A. 2016. Formation of Chromosomal Domains by Loop Extrusion. *Cell Reports* **15**: 2038-2049.

Furlong EEM, Levine M. 2018. Developmental enhancers and chromosome topology. *Science* **361**: 1341-1345.

Galas DJ, Schmitz A. 1978. DNAse footprinting: a simple method for the detection of protein-DNA binding specificity. *Nucleic acids research* **5**: 3157-3170.

Ganji M, Shaltiel IA, Bisht S, Kim E, Kalichava A, Haering CH, Dekker C. 2018. Real-time imaging of DNA loop extrusion by condensin. *Science* **360**: 102-105.

Gao XD, Tu L-C, Mir A, Rodriguez T, Ding Y, Leszyk J, Dekker J, Shaffer SA, Zhu LJ, Wolfe SA et al. 2018. C-BERST: defining subnuclear proteomic landscapes at genomic elements with dCas9–APEX2. *Nature Methods* **15**: 433-436.

Gariglio P, Bellard M, Chambon P. 1981. Clustering of RNA polymerase B molecules in the 5' moiety of the adult β-globin gene of hen erythrocytes. *Nucleic Acids Research* **9**: 2589-2598.

Garrick D, De Gobbi M, Samara V, Rugless M, Holland M, Ayyub H, Lower K, Sloane-Stanley J, Gray N, Koch C et al. 2008. The role of the polycomb complex in silencing alpha-globin gene expression in nonerythroid cells. *Blood* **112**: 3889-3899.

Gilbert N, Boyle S, Fiegler H, Woodfine K, Carter NP, Bickmore WA. 2004. Chromatin Architecture of the Human Genome: Gene-Rich Domains Are Enriched in Open Chromatin Fibers. *Cell* **118**: 555-566.

Grosveld F, van Staalduinen J, Stadhouders R. 2021. Transcriptional Regulation by (Super)Enhancers: From Discovery to Mechanisms. *Annual review of genomics and human genetics* **22**: 127-146.

Gu B, Swigut T, Spencley A, Bauer MR, Chung M, Meyer T, Wysocka J. 2018. Transcription-coupled changes in nuclear mobility of mammalian cis-regulatory elements. *Science* **359**: 1050-1055.

Guo Y, Xu Q, Canzio D, Shou J, Li J, Gorkin David U, Jung I, Wu H, Zhai Y, Tang Y et al. 2015. CRISPR Inversion of CTCF Sites Alters Genome Topology and Enhancer/Promoter Function. *Cell* **162**: 900-910.

Haarhuis JHI, van der Weide RH, Blomen VA, Yanez-Cuna JO, Amendola M, van Ruiten MS, Krijger PHL, Teunissen H, Medema RH, van Steensel B et al. 2017. The Cohesin Release Factor WAPL Restricts Chromatin Loop Extension. *Cell* **169**: 693-707 e614.

Haberle V, Arnold CD, Pagani M, Rath M, Schernhuber K, Stark A. 2019. Transcriptional cofactors display specificity for distinct types of core promoters. *Nature* **570**: 122-126.

Hanssen LLP, Kassouf MT, Oudelaar AM, Biggs D, Preece C, Downes DJ, Gosden M, Sharpe JA, Sloane-Stanley JA, Hughes JR et al. 2017. Tissue-specific CTCF–cohesin-mediated chromatin architecture delimits enhancer interactions and function in vivo. *Nature Cell Biology* **19**: 952-961.

Hark AT, Schoenherr CJ, Katz DJ, Ingram RS, Levorse JM, Tilghman SM. 2000. CTCF mediates methylation-sensitive enhancer-blocking activity at the H19/Igf2 locus. *Nature* **405**: 486-489.

Harrold CL, Gosden ME, Hanssen LLP, Stolper RJ, Downes DJ, Telenius JM, Biggs D, Preece C, Alghadban S, Sharpe JA et al. 2020. A functional overlap between actively transcribed genes and chromatin boundary elements. *bioRxiv*: 2020.2007.2001.182089.

Hay D, Hughes JR, Babbs C, Davies JOJ, Graham BJ, Hanssen L, Kassouf MT, Marieke Oudelaar AM, Sharpe JA, Suciu MC et al. 2016. Genetic dissection of the α-globin super-enhancer in vivo. *Nature genetics* **48**: 895-903.

Hentges LD, Sergeant MJ, Downes DJ, Hughes JR, Taylor S. 2021. LanceOtron: a deep learning peak caller for ATAC-seq, ChIP-seq, and DNase-seq. *biorxiv*.

Hildebrand EM, Dekker J. 2020. Mechanisms and Functions of Chromosome Compartmentalization. *Trends Biochem Sci* **45**: 385-396.

Hill L, Ebert A, Jaritz M, Wutz G, Nagasaka K, Tagoh H, Kostanova-Poliakova D, Schindler K, Sun Q, Bönelt P et al. 2020. Wapl repression by Pax5 promotes V gene recombination by Igh loop extrusion. *Nature* **584**: 142-147.

Hilton IB, D'Ippolito AM, Vockley CM, Thakore PI, Crawford GE, Reddy TE, Gersbach CA. 2015. Epigenome editing by a CRISPR-Cas9-based acetyltransferase activates genes from promoters and enhancers. *Nature biotechnology* **33**: 510-517.

Hnisz D, Shrinivas K, Young RA, Chakraborty AK, Sharp PA. 2017. A Phase Separation Model for Transcriptional Control. *Cell* **169**: 13-23.

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

Howe FS, Fischl H, Murray SC, Mellor J. 2017. Is H3K4me3 instructive for transcription activation? *Bioessays* **39**: 1-12.

Hsia EY, Goodson ML, Zou JX, Privalsky ML, Chen H-W. 2010. Nuclear receptor coregulators as a new paradigm for therapeutic targeting. *Adv Drug Deliv Rev* **62**: 1227-1237.

Hsieh TH, Weiner A, Lajoie B, Dekker J, Friedman N, Rando OJ. 2015. Mapping Nucleosome Resolution Chromosome Folding in Yeast by Micro-C. *Cell* **162**: 108-119.

Hua P, Badat M, Hanssen LLP, Hentges LD, Crump N, Downes DJ, Jeziorska DM, Oudelaar AM, Schwessinger R, Taylor S et al. 2021. Defining genome architecture at base-pair resolution. *Nature* **595**: 125-129.

Huang H, Zhu Q, Jussila A, Han Y, Bintu B, Kern C, Conte M, Zhang Y, Bianco S, Chiariello AM et al. 2021. CTCF mediates dosage- and sequence-context-dependent transcriptional insulation by forming local chromatin domains. *Nat Genet*.

Hughes JR, Cheng J-F, Ventress N, Prabhakar S, Clark K, Anguita E, De Gobbi M, de Jong P, Rubin E, Higgs DR. 2005. Annotation of cis-regulatory elements by identification, subclassification, and functional assessment of multispecies conserved sequences. *Proceedings of the National Academy of Sciences of the United States of America* **102**: 9830-9835.

Hughes JR, Roberts N, McGowan S, Hay D, Giannoulatou E, Lynch M, De Gobbi M, Taylor S, Gibbons R, Higgs DR. 2014. Analysis of hundreds of cis-regulatory landscapes at high resolution in a single, high-throughput experiment. *Nature Genetics* **46**: 205-212.

Jackson DA, Hassan AB, Errington RJ, Cook PR. 1993. Visualization of focal sites of transcription within human nuclei. *The EMBO journal* **12**: 1059-1065.

Jia Z, Li J, Ge X, Wu Y, Guo Y, Wu Q. 2020. Tandem CTCF sites function as insulators to balance spatial chromatin contacts and topological enhancer-promoter selection. *Genome biology* **21**: 75.

Johnson KD, Christensen HM, Zhao B, Bresnick EH. 2001. Distinct Mechanisms Control RNA Polymerase II Recruitment to a Tissue-Specific Locus Control Region and a Downstream Promoter. *Molecular cell* **8**: 465-471.

Juven-Gershon T, Hsu J-Y, Theisen JW, Kadonaga JT. 2008. The RNA polymerase II core promoter - the gateway to transcription. *Current opinion in cell biology* **20**: 253-259.

Kaaij LJT, Mohn F, van der Weide RH, de Wit E, Bühler M. 2019. The ChAHP Complex Counteracts Chromatin Looping at CTCF Sites that Emerged from SINE Expansions in Mouse. *Cell* **178**: 1437-1451.e1414.

Kagey MH, Newman JJ, Bilodeau S, Zhan Y, Orlando DA, van Berkum NL, Ebmeier CC, Goossens J, Rahl PB, Levine SS et al. 2010. Mediator and cohesin connect gene expression and chromatin architecture. *Nature* **467**: 430-435.

Keller GM. 1995. In vitro differentiation of embryonic stem cells. *Current opinion in cell biology* **7**: 862-869.

Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler, David. 2002. The Human Genome Browser at UCSC. *Genome Research* **12**: 996-1006.

Khoury A, Achinger-Kawecka J, Bert SA, Smith GC, French HJ, Luu PL, Peters TJ, Du Q, Parry AJ, Valdes-Mora F et al. 2020. Constitutively bound CTCF sites maintain 3D chromatin architecture and long-range epigenetically regulated domains. *Nat Commun* **11**: 54.

Kim TK, Hemberg M, Gray JM, Costa AM, Bear DM, Wu J, Harmin DA, Laptewicz M, Barbara-Haley K, Kuersten S et al. 2010. Widespread transcription at neuronal activity-regulated enhancers. *Nature* **465**: 182-187.

Kim Y, Shi Z, Zhang H, Finkelstein IJ, Yu H. 2019. Human cohesin compacts DNA by loop extrusion. *Science* **366**: 1345-1349.

King AJ, Songdej D, Downes DJ, Beagrie RA, Liu S, Buckley M, Hua P, Suciu MC, Marieke Oudelaar A, Hanssen LLP et al. 2021. Reactivation of a developmentally silenced embryonic globin gene. *Nat Commun* **12**: 4439.

Kowalczyk MS, Hughes JR, Garrick D, Lynch MD, Sharpe JA, Sloane-Stanley JA, McGowan SJ, De Gobbi M, Hosseini M, Vernimmen D et al. 2012. Intragenic enhancers act as alternative promoters. *Molecular cell* **45**: 447-458.

Krebs AR, Imanci D, Hoerner L, Gaidatzis D, Burger L, Schubeler D. 2017. Genome-wide Single-Molecule Footprinting Reveals High RNA Polymerase II Turnover at Paused Promoters. *Molecular cell* **67**: 411-422 e414.

Krivega I, Dean A. 2017. LDB1-mediated enhancer looping can be established independent of mediator and cohesin. *Nucleic Acids Res* **45**: 8255-8268.

Kubo N, Ishii H, Xiong X, Bianco S, Meitinger F, Hu R, Hocker JD, Conte M, Gorkin D, Yu M et al. 2021. Promoter-proximal CTCF binding promotes distal enhancer-dependent gene activation. *Nature structural & molecular biology*.

Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. *Nature Methods* **9**: 357-359.

Larke MSC, Schwessinger R, Nojima T, Telenius J, Beagrie RA, Downes DJ, Oudelaar AM, Truch J, Graham B, Bender MA et al. 2021. Enhancers predominantly regulate gene expression during differentiation via transcription initiation. *Molecular cell*.

Lee G, Saito I. 1998. Role of nucleotide sequences of loxP spacer region in Cre-mediated recombination. *Gene* **216**: 55-65.

Leemans C, van der Zwalm MCH, Brueckner L, Comoglio F, van Schaik T, Pagie L, van Arensbergen J, van Steensel B. 2019. Promoter-Intrinsic and Local Chromatin Features Determine Gene Repression in LADs. *Cell* **177**: 852-864 e814.

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

Levings PP, Bungert J. 2002. The human β-globin locus control region. *European Journal of Biochemistry* **269**: 1589-1599.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics (Oxford, England)* **25**: 2078-2079.

Li J, Hsu A, Hua Y, Wang G, Cheng L, Ochiai H, Yamamoto T, Pertsinidis A. 2020a. Single-gene imaging links genome topology, promoter-enhancer communication and transcription control. *Nature structural & molecular biology* **27**: 1032-1040.

Li K, Liu Y, Cao H, Zhang Y, Gu Z, Liu X, Yu A, Kaphle P, Dickerson KE, Ni M et al. 2020b. Interrogation of enhancer function by enhancer-targeting CRISPR epigenetic editing. *Nat Commun* **11**: 485.

Li W, Notani D, Ma Q, Tanasa B, Nunez E, Chen AY, Merkurjev D, Zhang J, Ohgi K, Song X et al. 2013. Functional roles of enhancer RNAs for oestrogen-dependent transcriptional activation. *Nature* **498**: 516-520.

Li X, Noll M. 1994. Compatibility between enhancers and promoters determines the transcriptional specificity of gooseberry and gooseberry neuro in the Drosophila embryo. *The EMBO journal* **13**: 400-406.

Li Y, Haarhuis JHI, Sedeño Cacciatore Á, Oldenkamp R, van Ruiten MS, Willems L, Teunissen H, Muir KW, de Wit E, Rowland BD et al. 2020c. The structural basis for cohesin–CTCF-anchored loops. *Nature* **578**: 472-476.

Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, Amit I, Lajoie BR, Sabo PJ, Dorschner MO et al. 2009. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**: 289-293.

Lin C, Garruss Alexander S, Luo Z, Guo F, Shilatifard A. 2013. The RNA Pol II Elongation Factor Ell3 Marks Enhancers in ES Cells and Primes Future Gene Activation. *Cell* **152**: 144-156.

Liu NQ, Maresca M, van den Brand T, Braccioli L, Schijns M, Teunissen H, Bruneau BG, Nora EP, de Wit E. 2021. WAPL maintains a cohesin loading cycle to preserve cell-type-specific distal gene regulation. *Nat Genet* **53**: 100-109.

Liu W, Ma Q, Wong K, Li W, Ohgi K, Zhang J, Aggarwal A, Rosenfeld MG. 2013. Brd4 and JMJD6-associated anti-pause enhancers in regulation of transcriptional pause release. *Cell* **155**: 1581-1595.

Luan J, Xiang G, Gomez-Garcia PA, Tome JM, Zhang Z, Vermunt MW, Zhang H, Huang A, Keller CA, Giardine BM et al. 2021. Distinct properties and functions of CTCF revealed by a rapidly inducible degron system. *Cell Reports* **34**: 108783.

Lupianez DG, Kraft K, Heinrich V, Krawitz P, Brancati F, Klopocki E, Horn D, Kayserili H, Opitz JM, Laxova R et al. 2015. Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. *Cell* **161**: 1012-1025.

Maeder ML, Linder SJ, Cascio VM, Fu Y, Ho QH, Joung JK. 2013. CRISPR RNA-guided activation of endogenous human genes. *Nature methods* **10**: 977-979.

Mali P, Aach J, Stranges PB, Esvelt KM, Moosburner M, Kosuri S, Yang L, Church GM. 2013. CAS9 transcriptional activators for target specificity screening and paired nickases for cooperative genome engineering. *Nature biotechnology* **31**: 833-838.

Mao X, Fujiwara Y, Orkin SH. 1999. Improved reporter strain for monitoring Cre recombinase-mediated DNA excisions in mice. *Proceedings of the National Academy of Sciences* **96**: 5037-5042.

Martinez-Ara M, Comoglio F, van Arensbergen J, van Steensel B. 2021. Systematic analysis of intrinsic enhancer-promoter compatibility in the mouse genome. *Biorxiv*.

Melo Carlos A, Drost J, Wijchers Patrick J, van de Werken H, de Wit E, Vrielink Joachim AFO, Elkon R, Melo Sónia A, Léveillé N, Kalluri R et al. 2013. eRNAs Are Required for p53-Dependent Enhancer Activity and Gene Transcription. *Molecular cell* **49**: 524-535.

Merli C, Bergstrom DE, Cygan JA, Blackman RK. 1996. Promoter specificity mediates the independent regulation of neighboring genes. *Genes & Development* **10**: 1260-1270.

Mirny LA, Imakaev M, Abdennur N. 2019. Two major mechanisms of chromosome organization. *Current Opinion in Genetics & Development* **58**: 142-152.

Moreau P, Hen R, Wasylyk B, Everett R, Gaub MP, Chambon P. 1981. The SV40 72 base repair repeat has a striking effect on gene expression both in SV40 and other chimeric recombinants. *Nucleic Acids Res* **9**: 6047-6068.

Myers SA, Wright J, Peckner R, Kalish BT, Zhang F, Carr SA. 2018. Discovery of proteins associated with a predefined genomic locus via dCas9–APEX-mediated proximity labeling. *Nature Methods* **15**: 437-439.

Nakahashi H, Kwon K-Rim K, Resch W, Vian L, Dose M, Stavreva D, Hakim O, Pruett N, Nelson S, Yamane A et al. 2013. A Genome-wide Map of CTCF Multivalency Redefines the CTCF Code. *Cell Reports* **3**: 1678-1689.

Narendra V, Rocha PP, An D, Raviram R, Skok JA, Mazzoni EO, Reinberg D. 2015. CTCF establishes discrete functional chromatin domains at the Hox clusters during differentiation. *Science (New York, NY)* **347**: 1017-1021.

Nasmyth K, Haering CH. 2009. Cohesin: its roles and mechanisms. *Annu Rev Genet* **43**: 525-558.

Neal A, Nornes S, Louphrasitthiphol P, Sacilotto N, Preston MD, Fleisinger L, Payne S, De Val S. 2021. ETS factors are required but not sufficient for specific patterns of enhancer activity in different endothelial subtypes. *Dev Biol* **473**: 1-14.

Neph S, Vierstra J, Stergachis AB, Reynolds AP, Haugen E, Vernot B, Thurman RE, John S, Sandstrom R, Johnson AK et al. 2012. An expansive human regulatory lexicon encoded in transcription factor footprints. *Nature* **489**: 83-90.

Nichols J, Evans EP, Smith AG. 1990. Establishment of germ-line-competent embryonic stem (ES) cells using differentiation inhibiting activity. *Development* **110**: 1341–1348.

Nishana M, Ha C, Rodriguez-Hernaez J, Ranjbaran A, Chio E, Nora EP, Badri SB, Kloetgen A, Bruneau BG, Tsirigos A et al. 2020. Defining the relative and combined contribution of CTCF and CTCFL to genomic regulation. *Genome biology* **21**: 108.

Nora EP, Goloborodko A, Valton AL, Gibcus JH, Uebersohn A, Abdennur N, Dekker J, Mirny LA, Bruneau BG. 2017. Targeted Degradation of CTCF Decouples Local Insulation of Chromosome Domains from Genomic Compartmentalization. *Cell* **169**: 930-944.e922.

Nora EP, Lajoie BR, Schulz EG, Giorgetti L, Okamoto I, Servant N, Piolot T, van Berkum NL, Meisig J, Sedat J et al. 2012. Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature* **485**: 381-385.

Norrman K, Fischer Y, Bonnamy B, Wolfhagen Sand F, Ravassard P, Semb H. 2010. Quantitative comparison of constitutive promoters in human ES cells. *PloS one* **5**: e12413.

Osterwalder M, Barozzi I, Tissières V, Fukuda-Yuzawa Y, Mannion BJ, Afzal SY, Lee EA, Zhu Y, Plajzer-Frick I, Pickle CS et al. 2018. Enhancer redundancy provides phenotypic robustness in mammalian development. *Nature* **554**: 239-243.

Oudelaar AM, Beagrie RA, Gosden M, de Ornellas S, Georgiades E, Kerry J, Hidalgo D, Carrelha J, Shivalingam A, El-Sagheer AH et al. 2020. Dynamics of the 4D genome during in vivo lineage specification and differentiation. *Nat Commun* **11**: 2722.

Oudelaar AM, Beagrie RA, Kassouf MT, Higgs DR. 2021. The mouse alpha-globin cluster: a paradigm for studying genome regulation and organization. *Current Opinion in Genetics & Development* **67**: 18-24.

Oudelaar AM, Harrold CL, Hanssen LLP, Telenius JM, Higgs DR, Hughes JR. 2019. A revised model for promoter competition based on multi-way chromatin interactions at the α-globin locus. *Nature Communications* **10**: 5412.

Oudelaar AM, Higgs DR. 2021. The relationship between genome structure and function. *Nature Reviews Genetics* **22**: 154-168.

Parelho V, Hadjur S, Spivakov M, Leleu M, Sauer S, Gregson HC, Jarmuz A, Canzonetta C, Webster Z, Nesterova T et al. 2008. Cohesins functionally associate with CTCF on mammalian chromosome arms. *Cell* **132**: 422-433.

Parker SCJ, Stitzel ML, Taylor DL, Orozco JM, Erdos MR, Akiyama JA, Bueren KLv, Chines PS, Narisu N, Black BL et al. 2013. Chromatin stretch enhancer states drive cell-specific gene regulation and harbor human disease risk variants. *Proceedings of the National Academy of Sciences* **110**: 17921-17926.

Parslow TG, Granner DK. 1983. Structure of a nuclease-sensitive region inside the immunoglobin kappa gene: evidence for a role in gene regulation. *Nucleic Acids Res* **11**: 4775-4792.

Pennacchio LA, Bickmore W, Dean A, Nobrega MA, Bejerano G. 2013. Enhancers: five essential questions. *Nat Rev Genet* **14**: 288-295.

Pertea M, Salzberg SL. 2010. Between a chicken and a grape: estimating the number of human genes. *Genome biology* **11**: 206.

Phillips-Cremins JE, Sauria ME, Sanyal A, Gerasimova TI, Lajoie BR, Bell JS, Ong CT, Hookway TA, Guo C, Sun Y et al. 2013. Architectural protein subclasses shape 3D organization of genomes during lineage commitment. *Cell* **153**: 1281-1295.

Pronobis MI, Deuitch N, Peifer M. 2016. The Miraprep: A Protocol that Uses a Miniprep Kit and Provides Maxiprep Yields. *PloS one* **11**: e0160509.

Pugacheva EM, Kubo N, Loukinov D, Tajmul M, Kang S, Kovalchuk AL, Strunnikov AV, Zentner GE, Ren B, Lobanenkov VV. 2020. CTCF mediates chromatin looping via N-terminal domain-dependent cohesin retention. *Proceedings of the National Academy of Sciences* **117**: 2020-2031.

Pugacheva EM, Rivero-Hinojosa S, Espinoza CA, Méndez-Catalá CF, Kang S, Suzuki T, Kosaka-Suzuki N, Robinson S, Nagarajan V, Ye Z et al. 2015. Comparative analyses of CTCF and BORIS occupancies uncover two distinct classes of CTCF binding genomic regions. *Genome biology* **16**: 161.

Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics (Oxford, England)* **26**: 841-842.

Quinodoz SA, Ollikainen N, Tabak B, Palla A, Schmidt JM, Detmar E, Lai MM, Shishkin AA, Bhat P, Takei Y et al. 2018. Higher-Order Inter-chromosomal Hubs Shape 3D Genome Organization in the Nucleus. *Cell* **174**: 744-757.e724.

Ramani V, Deng X, Qiu R, Gunderson KL, Steemers FJ, Disteche CM, Noble WS, Duan Z, Shendure J. 2017. Massively multiplex single-cell Hi-C. *Nature Methods* **14**: 263-266.

Ramírez F, Ryan DP, Grüning B, Bhardwaj V, Kilpert F, Richter AS, Heyne S, Dündar F, Manke T. 2016. deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Res* **44**: W160-165.

Ran FA, Hsu Patrick D, Lin C-Y, Gootenberg Jonathan S, Konermann S, Trevino AE, Scott David A, Inoue A, Matoba S, Zhang Y et al. 2013. Double Nicking by RNA-Guided CRISPR Cas9 for Enhanced Genome Editing Specificity. *Cell* **154**: 1380-1389.

Rao SS, Huntley MH, Durand NC, Stamenova EK, Bochkov ID, Robinson JT, Sanborn AL, Machol I, Omer AD, Lander ES et al. 2014. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159**: 1665-1680.

Rao SSP, Huang S-C, Glenn St Hilaire B, Engreitz JM, Perez EM, Kieffer-Kwon K-R, Sanborn AL, Johnstone SE, Bascom GD, Bochkov ID et al. 2017. Cohesin Loss Eliminates All Loop Domains. *Cell* **171**: 305-320.e324.

Redolfi J, Zhan Y, Valdes-Quezada C, Kryzhanovska M, Guerreiro I, Iesmantavicius V, Pollex T, Grand RS, Mulugeta E, Kind J et al. 2019. DamC reveals principles of chromatin folding in vivo without crosslinking and ligation. *Nature structural & molecular biology* **26**: 471-480.

Robinson MD, McCarthy DJ, Smyth GK. 2009. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics (Oxford, England)* **26**: 139-140.

Robinson PJ, Trnka MJ, Bushnell DA, Davis RE, Mattei PJ, Burlingame AL, Kornberg RD. 2016. Structure of a Complete Mediator-RNA Polymerase II Pre-Initiation Complex. *Cell* **166**: 1411-1422 e1416.

Roeder RG. 1996. The role of general initiation factors in transcription by RNA polymerase II. *Trends in Biochemical Sciences* **21**: 327-335.

Sacilotto N, Chouliaras KM, Nikitenko LL, Lu YW, Fritzsche M, Wallace MD, Nornes S, Garcia-Moreno F, Payne S, Bridges E et al. 2016. MEF2 transcription factors are key regulators of sprouting angiogenesis. *Genes & Development* **30**: 2297-2309.

Sainsbury S, Bernecky C, Cramer P. 2015. Structural basis of transcription initiation by RNA polymerase II. *Nat Rev Mol Cell Biol* **16**: 129-143.

Sanborn AL, Rao SSP, Huang S-C, Durand NC, Huntley MH, Jewett AI, Bochkov ID, Chinnappan D, Cutkosky A, Li J 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* **112**: E6456-E6465.

Sawado T, Halow J, Bender MA, Groudine M. 2003. The beta -globin locus control region (LCR) functions primarily by enhancing the transition from transcription initiation to elongation. *Genes & Development* **17**: 1009-1018.

Schmidl C, Rendeiro AF, Sheffield NC, Bock C. 2015. ChIPmentation: fast, robust, low-input ChIP-seq for histones and transcription factors. *Nature methods* **12**: 963-965.

Schmidt D, Schwalie PC, Wilson MD, Ballester B, Goncalves A, Kutter C, Brown GD, Marshall A, Flicek P, Odom DT. 2012. Waves of retrotransposon expansion remodel genome organization and CTCF binding in multiple mammalian lineages. *Cell* **148**: 335-348.

Schoneveld OJ, Gaemers IC, Hoogenkamp M, Lamers WH. 2005. The role of proximal-enhancer elements in the glucocorticoid regulation of carbamoylphosphate synthetase gene transcription from the upstream response unit. *Biochimie* **87**: 1033-1040.

Schwarzer W, Abdennur N, Goloborodko A, Pekowska A, Fudenberg G, Loe-Mie Y, Fonseca NA, Huber W, Haering CH, Mirny L. 2017. Two independent modes of chromatin organization revealed by cohesin removal. *Nature* **551**: 51-56.

Schwessinger R, Suciu MC, McGowan SJ, Telenius J, Taylor S, Higgs DR, Hughes JR. 2017. Sasquatch: predicting the impact of regulatory SNPs on transcription factor binding from cell- and tissue-specific DNase footprints. *Genome research* **27**: 1730-1742.

Simonis M, Klous P, Splinter E, Moshkin Y, Willemsen R, de Wit E, van Steensel B, de Laat W. 2006. Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture–on-chip (4C). *Nature Genetics* **38**: 1348-1354.

Smale ST, Kadonaga JT. 2003. The RNA polymerase II core promoter. *Annu Rev Biochem* **72**: 449-479.

Smith AG. 1991. Culture and differentiation of embryonic stem cells. *Journal of tissue culture methods* **13**: 89-94.

Smith RP, Taher L, Patwardhan RP, Kim MJ, Inoue F, Shendure J, Ovcharenko I, Ahituv N. 2013. Massively parallel decoding of mammalian regulatory sequences supports a flexible organizational model. *Nature Genetics* **45**: 1021-1028.

Spitz F, Furlong EE. 2012. Transcription factors: from enhancer binding to developmental control. *Nat Rev Genet* **13**: 613-626.

Spivak JL, Toretti D, Dickerman HW. 1973. Effect of Phenylhydrazine-induced Hemolytic Anemia on Nuclear RNA Polymerase Activity of the Mouse Spleen. *Blood* **42**: 257-266.

Suzuki Y, Tsunoda T, Sese J, Taira H, Mizushima-Sugano J, Hata H, Ota T, Isogai T, Tanaka T, Nakamura Y et al. 2001. Identification and characterization of the potential promoter regions of 1031 kinds of human genes. *Genome Research* **11**: 677-684.

Taniguchi M, Sanbo M, Watanabe S, Naruse I, Mishina M, Yagi T. 1998. Efficient production of Cre-mediated site-directed recombinants through the utilization of the puromycin resistance gene, pac: A transient gene-integration marker for ES cells. *Nucleic Acids Research* **26**: 679-680.

Tedeschi A, Wutz G, Huet S, Jaritz M, Wuensche A, Schirghuber E, Davidson IF, Tang W, Cisneros DA, Bhaskara V et al. 2013. Wapl is an essential regulator of chromatin structure and chromosome segregation. *Nature* **501**: 564-568.

Telenius J, Consortium TW, Hughes JR. 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.

Telenius JM, Downes DJ, Sergeant M, Oudelaar AM, McGowan S, Kerry J, Hanssen LLP, Schwessinger R, Eijsbouts CQ, Davies JOJ et al. 2020. CaptureCompendium: a comprehensive toolkit for 3C analysis. *biorxiv*: 2020.2002.2017.952572.

Thiecke MJ, Wutz G, Muhar M, Tang W, Bevan S, Malysheva V, Stocsits R, Neumann T, Zuber J, Fraser P et al. 2020. Cohesin-Dependent and -Independent Mechanisms Mediate Chromosomal Contacts between Promoters and Enhancers. *Cell Reports* **32**: 107929.

Thomas HF, Kotova E, Jayaram S, Pilz A, Romeike M, Lackner A, Penz T, Bock C, Leeb M, Halbritter F et al. 2021. Temporal dissection of an enhancer cluster reveals distinct temporal and functional contributions of individual elements. *Molecular cell* **81**: 969-982 e913.

Thomas MC, Chiang CM. 2006. The general transcription machinery and general cofactors. *Crit Rev Biochem Mol Biol* **41**: 105-178.

Thurman RE, Rynes E, Humbert R, Vierstra J, Maurano MT, Haugen E, Sheffield NC, Stergachis AB, Wang H, Vernot B et al. 2012. The accessible chromatin landscape of the human genome. *Nature* **489**: 75-82.

Tsujimura T, Klein FA, Langenfeld K, Glaser J, Huber W, Spitz F. 2015. A Discrete Transition Zone Organizes the Topological and Regulatory Autonomy of the Adjacent Tfap2c and Bmp7 Genes. *PLoS genetics* **11**: e1004897.

Ulianov SV, Khrameeva EE, Gavrilov AA, Flyamer IM, Kos P, Mikhaleva EA, Penin AA, Logacheva MD, Imakaev MV, Chertovich A et al. 2016. Active chromatin and transcription play a key role in chromosome partitioning into topologically associating domains. *Genome Research* **26**: 70-84.

Van Arensbergen J, FitzPatrick VD, de Haas M, Pagie L, Sluimer J, Bussemaker HJ, van Steensel B. 2017. Genome-wide mapping of autonomous promoter activity in human cells. *Nature biotechnology* **35**: 145-153.

Van Steensel B, Belmont AS. 2017. Lamina-Associated Domains: Links with Chromosome Architecture, Heterochromatin, and Gene Repression. *Cell* **169**: 780-791.

Varshavsky AJ, Sundin OH, Bohn MJ. 1978. SV40 viral minichromosome: preferential exposure of the origin of replication as probed by restriction endonucleases. *Nucleic acids research* **5**: 3469-3477.

Venkatesh S, Workman JL. 2015. Histone exchange, chromatin structure and the regulation of transcription. *Nat Rev Mol Cell Biol* **16**: 178-189.

Vernimmen D, De Gobbi M, Sloane-Stanley JA, Wood WG, Higgs DR. 2007. Long-range chromosomal interactions regulate the timing of the transition between poised and active gene expression. *The EMBO Journal* **26**: 2041–2051.

Vernimmen D, Lynch MD, De Gobbi M, Garrick D, Sharpe JA, Sloane-Stanley JA, Smith AJ, Higgs DR. 2011. Polycomb eviction as a new distant enhancer function. *Genes & Development* **25**: 1583-1588.

Vian L, Pekowska A, Rao SSP, Kieffer-Kwon KR, Jung S, Baranello L, Huang SC, El Khattabi L, Dose M, Pruett N et al. 2018. The Energetics and Physiological Impact of Cohesin Extrusion. *Cell* **173**: 1165-1178 e1120.

Vieira KF, Levings PP, Hill MA, Crusselle VJ, Kang S-HL, Engel JD, Bungert J. 2004. Recruitment of Transcription Complexes to the β-Globin Gene Locus in Vivo and in Vitro*. *Journal of Biological Chemistry* **279**: 50350-50357.

Vo Ngoc L, Wang Y-L, Kassavetis GA, Kadonaga JT. 2017. The punctilious RNA polymerase II core promoter. *Genes & Development* **31**: 1289-1301.

von Lindern M, Deiner EM, Dolznig H, Parren-Van Amelsvoort M, Hayman MJ, Mullner EW, Beug H. 2001. Leukemic transformation of normal murine erythroid progenitors: v- and c-ErbB act through signaling pathways activated by the EpoR and c-Kit in stress erythropoiesis. *Oncogene* **20**: 3651-3664.

Vos ESM, Valdes-Quezada C, Huang Y, Allahyar A, Verstegen M, Felder AK, van der Vegt F, Uijttewaal ECH, Krijger PHL, de Laat W. 2021. Interplay between CTCF boundaries and a super enhancer controls cohesin extrusion trajectories and gene expression. *Molecular cell*.

Wallace HA, Marques-Kranc F, Richardson M, Luna-Crespo F, Sharpe JA, Hughes J, Wood WG, Higgs DR, Smith AJ. 2007. Manipulating the mouse genome to engineer precise functional syntenic replacements with human sequence. *Cell* **128**: 197-209.

Wang W, Lin C, Lu D, Ning Z, Cox T, Melvin D, Wang X, Bradley A, Liu P. 2008. Chromosomal transposition of *PiggyBac* in mouse embryonic stem cells. *Proceedings of the National Academy of Sciences* **105**: 9290-9295.

Wansink DG, Schul W, van der Kraan I, van Steensel B, van Driel R, de Jong L. 1993. Fluorescent labeling of nascent RNA reveals transcription by RNA polymerase II in domains scattered throughout the nucleus. *J Cell Biol* **122**: 283-293.

Weinert BT, Narita T, Satpathy S, Srinivasan B, Hansen BK, Schölz C, Hamilton WB, Zucconi BE, Wang WW, Liu WR et al. 2018. Time-Resolved Analysis Reveals Rapid Dynamics and Broad Scope of the CBP/p300 Acetylome. *Cell* **174**: 231-244.e212.

Wen P, Carwford N, Locker J. 1993. A promoter-linked coupling region required for stimulation of α-fetoprotein transcription by distant enhancers. *Nucleic acids research* **21**: 1911-1918.

Whyte Warren A, Orlando David A, Hnisz D, Abraham Brian J, Lin Charles Y, Kagey Michael H, Rahl Peter B, Lee Tong I, Young Richard A. 2013. Master Transcription Factors and Mediator Establish Super-Enhancers at Key Cell Identity Genes. *Cell* **153**: 307-319.

Wiechens N, Singh V, Gkikopoulos T, Schofield P, Rocha S, Owen-Hughes T. 2016. The Chromatin Remodelling Enzymes SNF2H and SNF2L Position Nucleosomes adjacent to CTCF and Other Transcription Factors. *PLoS genetics* **12**: e1005940.

Wu C, M. Bingham P, Livak KJ, Holmgren R, Elgin SCR. 1979. The chromatin structure of specific genes: I. Evidence for higher order domains of defined DNA sequence. *Cell* **16**: 797-806.

Wutz G, Várnai C, Nagasaka K, Cisneros DA, Stocsits RR, Tang W, Schoenfelder S, Jessberger G, Muhar M, Hossain MJ. 2017. Topologically associating domains and chromatin loops depend on cohesin and are regulated by CTCF, WAPL, and PDS5 proteins. *The EMBO journal* **36**: 3573-3599.

Zabidi MA, Arnold CD, Schernhuber K, Pagani M, Rath M, Frank O, Stark A. 2015. Enhancer-core-promoter specificity separates developmental and housekeeping gene regulation. *Nature* **518**: 556-559.

Zabidi MA, Stark A. 2016. Regulatory Enhancer-Core-Promoter Communication via Transcription Factors and Cofactors. *Trends Genet* **32**: 801-814.

Zhang D, Huang P, Sharma M, Keller CA, Giardine B, Zhang H, Gilgenast TG, Phillips-Cremins JE, Hardison RC, Blobel GA. 2020. Alteration of genome folding via contact domain boundary insertion. *Nature Genetics* **52**: 1076-1087.

Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, Nusbaum C, Myers RM, Brown M, Li W et al. 2008. Model-based Analysis of ChIP-Seq (MACS). *Genome biology* **9**: R137.

Zhao Z, Tavoosidana G, Sjölinder M, Göndör A, Mariano P, Wang S, Kanduri C, Lezcano M, Singh Sandhu K, Singh U et al. 2006. Circular chromosome conformation capture (4C) uncovers extensive networks of epigenetically regulated intra- and interchromosomal interactions. *Nature Genetics* **38**: 1341-1347.

Zuin J, Franke V, van Ijcken WF, van der Sloot A, Krantz ID, van der Reijden MI, Nakato R, Lenhard B, Wendt KS. 2014. A cohesin-independent role for NIPBL at promoters provides insights in CdLS. *PLoS genetics* **10**: e1004153.