Investigating Factors Contributing to Enhancer-Promoter Interaction

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

Lucy Joy Cornell

St Edmund Hall

Wellcome Trust Doctoral Training Programme in Chromosome and Developmental Biology

Weatherall Institute of Molecular Medicine

University of Oxford

Michaelmas 2023

Abstract

Enhancer-promoter interactions are drivers of tissue-specific gene expression and there are several mechanisms which contribute to their establishment and maintenance. These include accessibility and recruitment of transcription factors and co-factors to enhancer and promoter elements within the context of chromatin; compatibility of enhancer-promoter pairs; and the extent to which interactions are constrained by the 3D structure of the genome. Enhancer promoter interactions usually occur within topologically associated domains (TADs); large regions (200-2000kb) of interacting chromatin delimited by binding sites for the insulating factor CTCF. Whilst we know that all these mechanisms play a role, it remains unclear how each contributes. This thesis aims to address less well understood mechanisms of how enhancer-promoter communication is fine tuned in the context of the murine α-globin locus. I report on a functional assay which highlights how actively transcribing genes can themselves act as insulator elements, similar in strength to a known strong CTCF site (HS-38). This assay also revealed the unexpected importance of gene orientation relative to their enhancers, such that maximum enhancer-induced activation is achieved when the genes are in their native orientation, whilst in the reverse orientation transcription is impeded. Finally, I also find enhancer-promoter communication, in the context of the α-globin locus, is partially independent from CTCF sites throughout the locus; removal of most sites individually or in combination has no effect on alpha globin expression. However, when all 9 CTCF sites across the locus are deleted, α-globin transcription is reduced. These results highlight the relationship between transcription and cohesin-mediated contacts and overall, the nuances of enhancer-promoter interactions which together achieve the optimised level of gene expression.

Acknowledgments

Firstly, I would like to give my deepest gratitude to Doug Higgs and Mira Kassouf for the opportunity to study under their guidance, for their advice and kindness. It has been an incredible experience and I have learnt so much under your supervision and your continued enthusiasm for research. I must give special thanks to Mira for her mentorship, her tireless support and all our discussions. This work was generously funded by the Wellcome Trust, for which I am incredibly grateful.

Thank you to all the members in the Higgs group who made working in the lab such a joy. It was an honour to have such a wonderful team around me to share their technical expertise and discussions about research but equally to share Wordle scores, the morning crosswords, and impromptu treks to Pret. I give special thanks to Felice for training me on the differentiation system and always providing thoughtful discussion about our experiments. Thank you to Rosa for teaching me how to do my first ever ChIP and sequencing run, to Emily for sharing her brilliant protocol for cohesin ChIPs, and to Joe, Damien, and Rob for sharing their knowledge of bioinformatics. Thank you to Chris Babbs for invaluable discussions about computing, sequencing, and the opportunity to meet and spend time with one of the sweetest dogs I've ever met, Freddy. Thank you to Noortje who was my very first supervisee and who gave me such joy to mentor and train. Also, I want to give my thanks to past members of the group, particularly Helena Francis, Caz Harrold, and Marieke Oudelaar, who I didn't have the chance to work with directly, but who left such an incredible framework of research and science for me to develop my studies upon.

The work in this thesis would not have been possible without the incredible core facilities at the WIMM; particularly the Flow Cytometry Facility, Sequencing facility and Philip Hublitz and the Genome Engineering Facility. Equally, my work on the synthetic locus was only made possible with our collaboration with Professor Jef Boeke and Brendan Camellato, who took on feat of synthesising the locus from scratch.

Away from the lab, I thank the Cardigang, namely, Lizzy, Aleks and Giannhs, for all our hijinks, all the quotable moments and our shared Oxford DPhil journeys. I just can't imagine doing this without you all and so happy that Teddy Hall brought us together.

Last, but definitely not least, thank you to Mum and Dad. For always been there through the ups and downs and been my anchor throughout my studies. Thank you for your patience and your unwavering support, you've done more for me than I can ever thank you enough for.

To Mum and Dad

Abbreviations

3C	Chromosome conformation capture
AID	Auxin-inducible degron
APH	Acetylphenylhydrazine
ATAC	Assay for transposase-accessible chromatin
aINV	Inverted Hba-a1
BAC	Bacterial artificial chromosome
bp	base pair
CD71	Transferrin receptor-1
ChIP	Chromatin immunoprecipitation
CRISPR	Clustered regularly interspaced short palindromic repeats
CTCF	CCCTC-binding factor
DHS	DNase I hypersensitive site
E14	E14-TG2a.IV (mESC clone)
EB	Embryoid body
FACS	Fluorescence-activated cell sorting
FCS	Fetal calf serum
FISH	Fluorescence in situ hybridization
Flp	Flippase
frt	Flippase recognition target
Gata1	GATA-binding factor 1
GF	Gene forward orientation
GR	Gene reverse orientation
HAT	Hypoxanthine-aminopterin-thymidine
HDR	Homology-directed repair
Hprt	Hypoxanthine phosphoribosyltransferase
HS	Hypersensitive site
HT	Hypoxanthine-thymidine
LAD	Lamina-associated domain
LCR	Locus control region
MCS	Multispecies conserved sequence
mESC	Mouse embryonic stem cell
MCC	Micro-Capture-C
MPRA	Massively parallel reporter assay
NGS	Next-generation sequencing
NLS	Nuclear localisation signal
PBS	Phosphate buffered saline
PCR	Polymerase chain reaction
PGF	Promoter and Gene forward orientation
PGR	Promoter and Gene reverse orientation
PF	Promoter forward orientation
PR	Promoter reverse orientation
PolII / RNAPII	RNA polymerase II
PolyA	Polyadenylated
RMGR	Recombinase-Mediated Genomic Replacement
RPKM	Reads per kilobase per million mapped reads
rpm	Rotations per minute
RT-qPCR	Real-time PCR
TAD	Topologically Associated Domain
TF	Transcription factor
tk	Thymidine kinase
UTR	Untranslated region
WT	Wild type
YFP	Yellow fluorescent protein

TABLE OF CONTENTS

TABLE OF FIGURES

TABLE OF TABLES

Chapter 1: Introduction

1.1 Regulatory units of the mammalian genome

Each nucleated cell contains the same genome sequence, which encodes all the information necessary to give rise to a multicellular organism comprised of meticulously organised cells, each with specialised function and purpose which can also adapt to environmental stresses. One of the most prominent questions is how a single blueprint genome can give rise to such a diverse range of cells. Of course, each cell type expresses a distinct combination of protein-coding genes which then create specialised cells, however, the structural genes in the human genome are encoded in only 1.2% of the genome (Frith *et al*., 2005). Remarkably, the number of protein-coding genes in the human genome (~22,000) is almost the same as that of a simple nematode, *Caenorhabditis elegans* (~20,000) which comprise ~25% of the genome, indicating that the complexity of an organism does not arise from the number of genes, rather is driven by further variation in when and where the genes are expressed and the various combinations in their expression. These aspects of gene expression creating further complexity are now thought to be determine by the "non-coding" proportion of the genome.

Each gene is regulated by its promoter; a region directly upstream (5') of its transcription start site (TSS). Promoters transduce regulatory signals which converge on the gene, integrating signals into transcription initiation events via the recruitment of general transcription factors, the pre-initiation complex (PIC) and initiation of transcription via RNA Polymerase II (RNAPII). Mammalian promoters are often transcribed bi-directionally; however, elongation is often limited to transcription of the associated gene, with transcription in the opposite direction giving rise to unstable antisense transcripts which are rapidly degraded (Core *et al*., 2008; Preker *et al*., 2008; Seila *et al*., 2008). Tissue-specific genes, ubiquitously expressed housekeeping genes and developmental genes tend to have distinct characteristics and sequence motifs (Haberle and Stark, 2018). Generally, both developmental and housekeeping gene promoters often do not contain TATA box motifs but are associated with high densities of CpG di-nucleotides forming unmethylated CG islands (CGIs); by contrast tissue-specific gene promoters often to contain TATA elements but are not associated with CGIs (Haberle and Stark, 2018). In different cellular contexts, there are distinct repertoires of transcription factors, which bind to promoter proximal sequences, in concert with general transcription factors, to modulate expression.

Promoters can also be activated by enhancers, which are cis-regulatory elements rich in binding sites for tissue-specific and developmental stage-specific transcription factors; therefore, conferring spacio-temporal activation signals to target genes. In addition, enhancers can also recruit RNAPII, co-activators and Mediator complexes to further regulate gene expression (Kagey *et al*., 2010). Enhancer activation upon cofactor and transcription factor binding is thought to initiate a series of mechanistic events leading to increased gene activation including but not limited to; nucleosome repositioning and eviction of repressive Polycomb complexes at activated promoters, recruitment of RNAPII, pause release of RNAPII and nuclear repositioning (reviewed in Long *et al*., 2016). Classically, enhancers are defined as acting on their cognate promoters independently of genomic distance and orientation (Banerji *et al*., 1983). Recent developments have disputed such definition, as enhancer function does appear to be inversely correlated with distance (Rinzema *et al*., 2022; Zuin *et al*., 2022) and clusters of enhancers may act in an orientation dependant manner (preprint Kassouf *et al*., 2022). This disparity is likely caused by the traditional validation of enhancers in episomal reporter assays (Banerji *et al*., 1983) or minimal reporter systems (Pennacchio *et al*., 2006) which do not fully recapitulate the enhancer's native genomic context. Enhancers can be close to or found at large distances from their target genes; in the case of the mammalian sonic hedgehog (shh) gene, a cognate enhancer element (ZRS) is located more than 1Mb away (Lettice, 2003). In addition, enhancers show remarkable selectivity to their target genes, this is exemplified by 'bystander' genes, which can be found proximal to enhancers but remain unresponsive to them (Kikuta *et al*., 2007). One mechanism by which this is achieved may be active silencing and in-accessibility of intervening 'bystander' genes; for example, the embryonic *Hba-x* gene lies proximal to strong enhancers but is silenced during differentiation and associated with extensive epigenetic modifications preventing its expression (King *et al*., 2021). Another mechanism by which genes can be overlooked by enhancers may be due to inherent specificity between enhancer promoter pairs. Development of massively parallel reporter assays (MPRA) and testing of multiple combinations of enhancer promoter pairs, indicate that sequence composition of elements may confer such specificity between enhancers and promoters in Drosophila (Arnold *et al*., 2017; Zabidi *et al*., 2015) and to a more limited degree in vertebrates (Bergman *et al*., 2022; Martinez-Ara *et al*., 2022). This may be facilitated by biochemical compatibility of transcription factors and their associated co-factors bound to each element (Haberle *et al*., 2019).

Given the large linear distances that separate enhancer and promoters, it is rational that for the elements to communicate they must at some point be in physical proximity and conceptually this can be achieved by DNA looping. Many mechanisms have been proposed to explain looping: passive diffusion; tracking, with a protein translocating from enhancers to target promoters; linking or

bridging, by oligomerisation of transcription factors between the enhancer and promoter; and relocation, with active regions migrating to sub-nuclear transcription factories (reviewed in Furlong and Levine, 2018). Looping may be aided shared occupation of potential 'tethering' factors at enhancers and promoters, such as Mediator, YY1 (Hnisz *et al*., 2016) and LDB1 (Deng *et al*., 2012). Enhancer proximal non-promoter CGIs, were also proposed to act as tethering elements, bringing their associated enhancers into proximity with developmental promoters containing CGIs (Pachano *et al*., 2021). After establishment of contact, a sub-nuclear compartment may be formed, often referred to as a 'transcriptional hub', which contains a high concentration of transcription factors and co-factors which enable transcription. Different mechanisms may be more important at different stages of enhancer-promoter contact, such as during establishment or maintenance of contact. Indeed, how enhancer-promoter pairs are drawn into proximity has been the subject of intensive study in recent years.

1.2 Structural organization of the genome plays a role in the regulation of gene expression

To fit into a cell, the genome (~2 metres of DNA) must be compacted and organised. DNA is wrapped around octameric histone proteins, providing the first layer of compaction; approximately 200 bps fit around a histone octamer to form nucleosomes. Whilst post-translational modifications to histones can affect the accessibility of the underlying DNA, further organisation is required to facilitate the precise and selective activation or repression of genes. One way in which this is achieved is by radial positioning, such that gene-poor regions or densely packed silenced heterochromatin are found at the nuclear periphery, sometimes forming lamina-associated domains (LADS) (Peric-Hupkes *et al*., 2010) whilst gene-rich or active euchromatin is found more centrally. In addition, individual chromosomes form discrete territories in the nucleus, preferentially interacting *in cis* rather than *in trans,* as shown by chromosomal DNA fluorescence *in situ* hybridization (FISH) (Cremer and Cremer, 2001).

1.2.1 Compartments

This organisation was recapitulated in early genome-wide chromosome conformation capture (Hi-C) studies, such that blocks of predominantly active chromatin colocalise with other active regions to form A-compartments, whilst regions of transcriptionally inactive chromatin coalesce forming B-compartments (Lieberman-Aiden *et al*., 2009). These A and B compartments are mutually exclusive and are found at large mega-base scales and visualised in genome-wide Hi-C matrices as checkerboard

patterns. This was the first instance of use of genome-wide implementation of chromosome conformation capture (3C) (Dekker *et al*., 2002). Briefly in 3C, chromatin in close spatial proximity is crosslinked together, DNA is then fragmented (for example by using restriction enzymes DpnII or NcoI) but remain held in place by cross links. *In situ* proximity ligation then connects spatially proximal DNA ends forming concatemers. The frequency of chromosomal contacts can be calculated from the number of newly formed ligation junctions in the resulting chimeric DNA concatemers, as the closer two genomic regions are in 3D space, the more likely they are to be ligated. The presence of A/B compartments each with shared epigenetic signatures, lead to development of the hypothesis that microphase separation may give rise to these compartments; however, it is difficult to determine if this occurs *in vivo* and whether is a cause or consequence of chromatin regulation (reviewed in Hildebrand and Dekker, 2020).

1.2.2 TADs and sub-topologies

Within the A/B compartments, higher resolution Hi-C analysis revealed smaller domains of the scale of 100s to 1000s of kilobases, which are described as topologically associating domains (TADs) (Dixon *et al*., 2012; Nora *et al*., 2012; Sexton *et al*., 2012). These domains were initially reported to be of a median size of ∼880 kb, but improvement to kilobase resolution suggested a smaller average size of 185 kb (range 40 kb–3 Mb) (Rao *et al*., 2014). Such structures have also been observed by orthogonal techniques not using proximity ligation, including high-resolution (FISH) imaging (Bintu *et al*., 2018; Wang *et al*., 2016), genome architecture mapping (GAM) (Beagrie *et al*., 2017), proximity methylation (DamC) (Redolfi *et al*., 2019) and proximity tagging (SPRITE) (Quinodoz *et al*., 2018) further confirming spatial partitioning of genomes into domains corresponding to TADs. TADs are contiguous regions of chromatin which show preferential intra-domain interactions and lower inter-domain interactions with neighbouring domains in *cis*. Many lines of evidence support a functional role of TADs in gene regulation. Random transposition of a reporter into the mouse genome revealed enhancer activity was limited to large regulatory regions corresponding to TADs (Symmons *et al*., 2014). When analysing Hi-C from 21 different tissues, some TADs were consistently present in different cell types, however a small number of TADs were specific to single cell types and enriched for differentially expressed tissue-specific genes and super enhancers (Schmitt *et al*., 2016). Indeed, during neural development and differentiation of mESCs into cortical neurons (CNs), regions that underwent architectural reorganisation were associated with altered gene expression (Bonev *et al*., 2017). In addition, disruption of TADs, such as by structural rearrangements, are associated with gene mis-regulation, which can give rise to developmental malformations (Cova *et al*., 2023; De Bruijn *et al*., 2020; Franke

et al., 2016; Lupiáñez *et al*., 2015; Symmons *et al*., 2016) and development of cancer (Flavahan *et al*., 2016; Hnisz *et al*., 2016). From this, TADs are understood to act as functional regulatory units, providing topological restriction to tissue-specific enhancer-promoter interactions.

Within mammalian TADs, smaller self-interacting nested structures were also identified; referred to as sub topologies (sub-TADs) (Phillips-Cremins *et al*., 2013), contact domains (Rao *et al*., 2014), insulated neighbourhoods (Dowen *et al*., 2014) and frequently interacting regions (FIREs) (Schmitt *et al*., 2016). This inconsistent nomenclature and differences across experimental setups confuse classification of these domains. However, there are some characteristics which can help toward distinguishing TADs and sub-TADs from one another. For example, megabase-scale TADs are largely cell-type invariant, whereas sub-TADs are more likely to exhibit cell-type-specific reconfiguration; this is indeed the case in the α-globin locus (Oudelaar *et al*., 2020a). TADS and sub-TADS may be governed by different architectural proteins, but this remains unclear (Beagan and Phillips-Cremins, 2020).

1.2.3 Microcompartments

Recent technical developments have improved the resolution of 3C methods still further, allowing finer-scale interrogation of the genome architecture. Micro-C, and its variants, use micrococcal nuclease (MNase) in the place of 4-6bp restriction enzymes, to fragment chromatin to the level of individual nucleosomes (~200 bp) (Hsieh *et al*., 2020, 2015; Krietenstein *et al*., 2020). Micro-C has improved signal-to-noise ratio compared to Hi-C, allowing visualisation of fine-scale structures, including promoter-promoter and enhancer-promoter contacts. Combining this approach with region enrichment, Micro Capture-C (MCC), reveals base-pair resolution of interactions to the extent that footprints of proteins bound at interacting *cis*-regulatory elements can be visualised (Aljahani *et al*., 2022; Hua *et al*., 2021).

Whilst TADs and chromatin loops are apparent in data population-averaged chromatin-interaction maps, single cell techniques have highlighted that chromosome structure varies markedly from cell to cell (Nagano *et al*., 2013; Stevens *et al*., 2017). Despite the variability that may arise from the difficult task of extracting Hi-C from single cells, upon the ensemble of data from multiple cells, TADs do appear from the noise, even if they are not evident structures in individual cells. This is further corroborated by high-resolution single cell imaging, which also show high levels of stochasticity and heterogeneity in chromatin conformations on a cell-to-cell basis (Bintu *et al*., 2018; Finn *et al*., 2019; Gabriele *et al*.,

2022). Together these findings demonstrate that TADs are not static structures and instead are statistical constructs which arise from variations in the probability of pair-wise interactions.

1.3 Loop extrusion as a mechanism of genome organization

1.3.1 Roles of CTCF and Cohesin

A common characteristic of TADs is that the borders of domains are demarcated by CCCTC-binding factor (CTCF) and can be seen in Hi-C contact maps as 'corner peaks' (Dixon *et al*., 2012; Nora *et al*., 2012; Phillips-Cremins *et al*., 2013; Rao *et al*., 2014). CTCF is a small (~82 kDa) zinc finger transcription factor which binds a non-palindromic consensus sequence; hence there is inherent orientation of CTCF binding. Interestingly, the CTCF binding sites at domain boundaries are often found to be convergently orientated, and it was shown that convergent CTCF sites are functionally significant in the insultation and formation of domain boundaries (de Wit *et al*., 2015; Guo *et al*., 2015; Rao *et al*., 2014; Sanborn *et al*., 2015; Vietri Rudan *et al*., 2015). The importance of CTCF in gene regulation is displayed by many instances where disruptions of CTCF sites at domain boundaries led to merger of neighbouring domains and mis-regulated gene expression (Dowen *et al*., 2014; Hanssen *et al*., 2017; Lee *et al*., 2017; Lupiáñez *et al*., 2015; Narendra *et al*., 2015; Symmons *et al*., 2016; Vos *et al*., 2021). This is in line with CTCF's recognised function as an insulator, so called because CTCF acts as a local barrier to *cis*-acting elements, insulating promoters from distal enhancers (Bell *et al*., 1999; Wendt and Peters, 2009). CTCF has also been described as a chromatin barrier, as its presence limits the spread of chromatin modifications such as H3K4me3 and H3K27me3 associated with active and inactive chromatin respectively, however this may be a secondary effect of limiting enhancer activity (Hanssen *et al*., 2017; Narendra *et al*., 2015).

CTCF bound sites are often co-occupied by the protein complex cohesin (Parelho *et al*., 2008; Rubio *et al*., 2008; Stedman *et al*., 2008; Wendt *et al*., 2008). Cohesin is a tripartite ring-complex comprised of structural maintenance of chromosomes (SMC) family proteins Smc1/3 and Rad21; this further associates with Scc which comes in 3 variants in mammals (SA-1, SA-2, and SA-3). Complex assembly leads to Smc1/Smc3 heterodimerisation, which then form a composite ABC-like ATPase domain (Haering *et al*., 2002). Cohesin was primarily associated with DNA replication and sister chromatid cohesion and is considered to encircle chromosomal helices (Nasmyth and Haering, 2009). It is thought to be loaded onto chromatin by Nipbl-Mau2 complexes (Ciosk *et al*., 2000) and released by Wapl and Pds5 (Kueng *et al*., 2006), presumably by opening of the cohesin ring. Co-localisation of cohesin with

CTCF was found to be dependent on CTCF, however CTCF could associate with chromatin independently of cohesin, suggesting that CTCF may anchor cohesin to chromatin (Parelho *et al*., 2008; Rubio *et al*., 2008; Wendt *et al*., 2008). CTCF and cohesin were identified at >86% of domain boundaries in mammalian Hi-C, indicating a key role of these two proteins in TAD formation (Rao *et al*., 2014).

1.3.2 Loop extrusion

The role of both CTCF and cohesin in building chromatin architecture is highlighted in acute depletion studies coupled with Hi-C. When cohesin is removed via an auxin-inducible degron (AID) chromatin domain structures are lost globally (Rao *et al*., 2017) whilst degron depletion of CTCF results in loss of TAD insulation and creating merged TADs (Nora *et al*., 2017). However, in this situation, A and B compartments remain intact (Nora *et al*., 2017). In addition, upon knockout or knockdown of the cohesin release factor Wapl, cohesin residency time increases and chromatin loops are extended, indicating that Wapl normally acts to restrict domain formation by off-loading cohesin and thereby limiting its translocation (Haarhuis *et al*., 2017; Wutz *et al*., 2017).

It has been proposed that TADs are formed, at least in part, by loop extrusion (Fudenberg *et al*., 2016; Sanborn *et al*., 2015). In this mechanism, a DNA-extruding factor (i.e. cohesin) is loaded onto chromatin, proceeding to translocate uni- or bi-directionally and extruding DNA into a loop as it progresses. Translocation continues until halted by boundary elements, such as convergent CTCF sites, at which point it is stalled, and a loop domain is formed and eventually off-loaded by Wapl. This mechanism can explain the curious observation of convergent CTCF sites at domain boundaries, as it is the N-terminal domain of the CTCF protein that interacts with the translocating Rad21 subunit of cohesin causing stalling of cohesin translocation long enough for stabilisation, whilst stabilisation may not be achieved with weaker stalling at the C-terminus (**Figure 1.1**) (Li *et al*., 2020; Pugacheva *et al*., 2020).

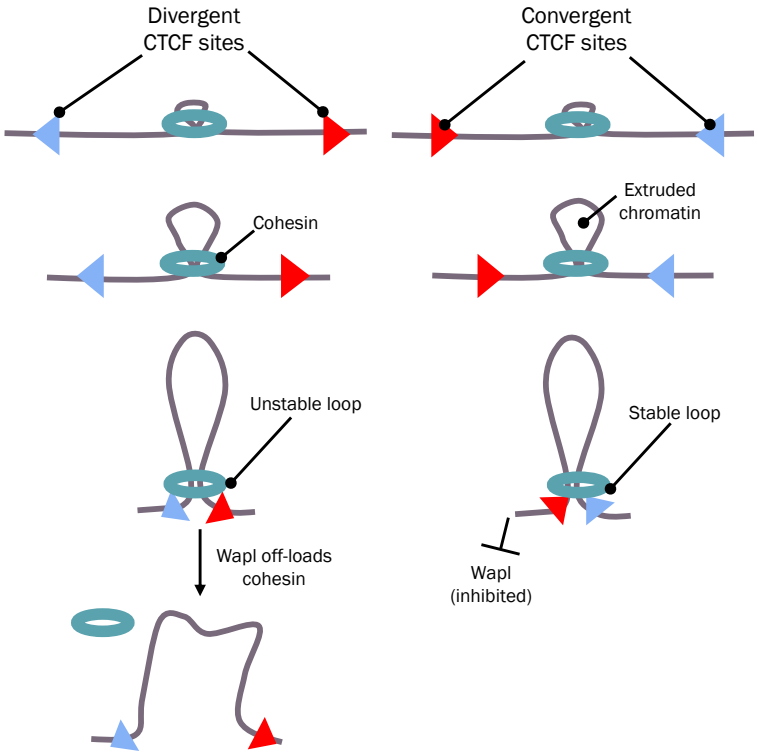

Figure 1.1: Proposed mechanism of preference for convergently orientated CTCF at TAD domain boundaries. *Cohesin is loaded onto chromatin, likely by a combination of Nipbl:Mau2 activity, and proceeds to pull chromatin into an extruded loop. Cohesin continues to translocate until it encounters CTCF sites. When CTCF sites are in divergent orientation, cohesin may be stalled but not stabilised and is promptly released from chromatin by Wapl. When CTCF sites are convergent, cohesin translocation is stalled and the N-terminal domain of CTCF may form a stabilising interaction with cohesin, preventing off-loading by Wapl. Figure adapted from (Davidson and Peters, 2021)*

Loop extrusion also explains the transient nature of TADs and subTADs as it is an inherently dynamic process. Single molecule tracking *in vitro* provide direct evidence of cohesin mediated DNA extrusion *in vitro,* using yeast cohesin (Ganji *et al*., 2018) and mammalian cohesin (Davidson *et al*., 2019; Kim *et al*., 2019) however it is yet to be observed directly *in vivo.* These studies as well as (Rhodes *et al*., 2017) suggest that Nipbl acts as a processivity factor to cohesin extrusion and so regions enriched for Nipbl may not simply be indicative of cohesin loading. Indeed, whether cohesin is pervasively loaded genome-wide or if there are sites of active loading remains an open question and is difficult to determine *in vivo* due to its dynamic nature; however, there are indirect observations that support the suggestion that cohesin may be loaded active enhancer elements. Firstly, cohesin co-localises with cell-type specific transcription factors and Mediator at promoters and enhancers (Faure *et al*., 2012; Kagey *et al*., 2010). In addition, cohesin occupancy at enhancers is reduced with degron-induced

depletion of Mediator (Ramasamy *et al*., 2023). Furthermore, if cohesin is indeed loaded at enhancer elements then it would be expected that disruption of enhancer elements would lead to a reduction in cohesin occupancy; this was observed when only one of the 5 enhancer elements is present in the α-globin locus cohesin accumulation across the locus is markedly reduced (preprint Georgiades *et al*., 2023). Similarly, this was also reported in a separate study, with reduced cohesin accumulation observed at flanking CTCF sites (Vos *et al*., 2021). Insertion of enhancer elements into 'neutral' regions of chromatin also leads to cohesin recruitment (Georgiades *et al*., 2023; Rinzema *et al*., 2022). Finally, predictions from polymer simulations predicted that targeted loading of cohesin and loop-extrusion would result in perpendicular 'jets' from loading sites in Hi-C maps, which were recently observed in quiescent murine thymocytes and associated with accessible chromatin, Nipbl and H3K27ac (Guo *et al*., 2022). However, it remains unclear whether those chromatin signatures are sufficient or causative of targeted cohesin loading.

It also unclear whether loop extrusion is a bi-directional or asymmetric process, however genomic context may dictate the resultant trajectory of cohesin translocation. For example, the phenomena of architectural stripes in Hi-C contact matrices can be interpreted as a forced asymmetric extrusion; cohesin loaded in the vicinity of a correctly orientated CTCF site is immediately blocked, anchoring one side of the loop. Extrusion then continues to reel in the rest of the domain and so presenting the anchor to the entire TAD (Barrington *et al*., 2019; Vian *et al*., 2018). In some cases, the anchor site is proximal to lineage-specific enhancers and super enhancers, thus the tethered enhancers can be considered to track along chromatin, permitting linear scanning for cognate promoters.

1.4 Transcriptional regulation: informed or informing genome architecture?

Despite these insights, the functional role of genome architecture and loop extrusion in gene regulation remains unclear, as there are numerous reports with surprisingly minor changes in gene expression upon the disruption of domain structures. Depletion of CTCF, cohesin, Wapl and Nipbl, whilst associated with massively perturbed global architecture, display very minor changes in transcription following depletion, with A/B compartmentalisation appearing to be more prominent, (Nora *et al*., 2017; Rao *et al*., 2017; Schwarzer *et al*., 2017; Wutz *et al*., 2017). The relatively small number of loci with significantly changed expression upon depletion of cohesin, corresponded to enhancer regulated tissue-specific genes. This led to the hypothesis that there may be two forces contributing to genome architecture; the first is loop extrusion (cohesin dependant) and the second is the molecular compartmentalisation (cohesin-independent). Molecular compartmentalisation may

arise from spatial separation of different chromatin states, such that regions with similar epigenetic signatures tend to attract and separate from other signatures (Nuebler *et al*., 2018). Loop extrusion appears to act as an antagonistic force to compartmentalisation, which may arise from more passive mechanisms, such as phase-separation, but this is still largely unclear. The extent to which each of these forces contributes to regulation remains an important line of study and may be dependent on genomic and cellular contexts.

Whilst global interactions are lost, the resolution of Hi-C prevented assessment of how individual enhancer-promoter interactions are affected. With the advent of Micro-C, it became possible to determine the role of cohesin in determining the fine-scale enhancer–promoter/promoter–promoter interactions, overlooked in Hi-C. 'Microcompartments' containing such interactions were reportedly resistant to acute depletion of cohesin (Hsieh *et al*., 2020). Even with the generation of region enrichments of Micro-C, deepening the mapping of 3D structures, uncertainty of the contribution of loop extrusion remains. Depending on the locus, some enhancer-promoter interactions are weakened, and some are unaffected upon acute depletion of cohesin in mESCs (Aljahani *et al*., 2022; Goel *et al*., 2023). This suggests that cohesin may not be important for the establishment and/or the maintenance of some enhancer-promoter interactions. However, these studies were performed in steady state conditions, and so the relatively minor changes in gene expression may be due to the presence of preformed stable enhancer-promoter contacts. Other reports suggest cohesin is required for long range enhancer-promoter interactions or inducible gene expression during differentiation (Calderon *et al*., 2022; Cuartero *et al*., 2018; Kane *et al*., 2022; Rinzema *et al*., 2022). In line with this, cohesin depletion during *in vitro* erythroid differentiation into the erythroid lineage caused a severe effect on α-globin expression only when depleted early in differentiation (preprint Stolper *et al*., 2023). Therefore, loop extrusion may be required during differentiation, to set up correct enhancer-promoter contacts. Once formed these become largely independent of extrusion and molecular affinities may hold the regions together, with extrusion providing robustness to the interaction.

An equally debated question is whether transcription can have a reciprocal influence on chromosomal architecture. Whilst CTCF and cohesin are observed at TAD borders, active genes were also observed at TAD boarders (Dixon *et al*., 2012; Rao *et al*., 2014). Later it was shown that this was not limited to housekeeping genes as the formation of a border was concomitant with activation of neuron-specific genes over different stages of differentiation into neuronal cells (Bonev *et al*., 2017). In addition, RNAPII is required to re-establish genome architecture following of after mitosis (Zhang *et al*., 2021). These observations suggested that transcription may indeed have a role in regulating genome

architecture. Whilst there was little change in Hi-C architecture upon depletion of RNAPs (Jiang *et al*., 2020; Valton *et al*., 2022) recently, degron depletion of RNAPII used in combination with Micro-C observed rewiring of fine-structure contacts from promoters to CTCF sites, which could be interpreted as RNAPII acting as a blockade to loop extrusion (S. Zhang *et al*., 2023). Chemical inhibition of RNAPII results in reduction in the intensity of enhancer–promoter and promoter–promoter stripes in Micro-C (Barshad *et al*., 2023; Hsieh *et al*., 2020). In addition, RNAPII and transcription can lead to relocation of cohesin (Busslinger *et al*., 2017) and can disrupt existing spatial organisation (Heinz *et al*., 2018). Therefore, it seems that actively transcribing RNAPII can also contribute to chromatin architecture, but it is unclear how transcription interacts with loop extrusion mechanistically.

1.5 The mouse α-globin locus as a model for cis-regulatory element function

Although the study of nuclear organization and its role in gene regulation has been crucial for uncovering general principles and potential mechanisms, a substantial number of discoveries concerning the interplay between genome structure and gene expression have emerged from detailed analyses of individual loci. The mouse alpha (α) globin locus has long been used as a model to study mechanisms of gene regulation. *Cis* regulatory elements are largely conserved in position and sequence throughout evolution (Hughes *et al*., 2005; Philipsen and Hardison, 2018). Due to its extensive *cis*-regulatory element characterisation and the ease of obtaining erythroid cells from patients and model organisms, the α-globin cluster remains an ideal model in which to investigate the role of mammalian *cis*-regulatory elements in transcriptional regulation (Oudelaar *et al*., 2021) (**Figure 1.2**). Furthermore, as α-globin is a product of terminal erythroid differentiation, experimental manipulations of the locus do not affect cell fate decisions, development, or differentiation, which could confound the interpretations.

The genes are distributed across the locus in the order of their developmental expression pattern starting with the zeta (ζ) globin gene (*Hba-x*) which encodes and expresses embryonic globin in primitive erythroid cells but is silenced in definitive erythroid cells (Palis, 2014). These are followed by the duplicated α-globin genes (*Hba-a1, Hba-a2*) which encode the alpha subunit of haemoglobin expressed in the embryo (primitive erythrocytes) and becomes the dominant α-globin in adult (definitive) erythroid cells. The α-globin genes are linked to the theta (θ) globin genes (*Hbq-1, Hbq-2*) which are of unknown function (Albitar *et al*., 1989). All of the genes within the α-globin locus are regulated exclusively by a cluster of 5 distinct enhancers R1, R2, R3, Rm and R4, which are not all functionally equivalent (Hay *et al*., 2016). These enhancers bind many transcription factors known to

regulate erythropoiesis including GATA1, NF-E2, KLF1 and TAL1) and also recruit high levels of Mediator, H3K4me1 and H3K27ac, fulfilling the requirements to be classified as a super enhancer (Hay *et al*., 2016). Of the elements, R2 was identified as the strongest element, binding the majority of transcription factors and alone contributing 50 % activation of the mouse α-globin genes (Hay *et al*., 2016) and 80 % of human activity (Badat *et al*., 2021). Further to this, R1 and R2 appear to act as classical enhancers, whilst R3, Rm and R4 do not appear to have inherent enhancer activity but are required for full activation potential; these elements have therefore been referred to as "facilitator" elements (Blayney *et al*., 2022 in press). The locus contains largely convergent CTCF sites which flank the super enhancer and promoters (Hanssen *et al*., 2017). Finally, the entire locus is contained within a 165 kb TAD domain, and over the course of erythropoiesis a smaller 65kb erythroid specific subTAD is formed, encapsulating the enhancer elements and α-globin genes (Hughes *et al*., 2014; Oudelaar *et al*., 2020) (**Figure 1.2**). The TAD has been visualised by high resolution microscopy (Brown *et al*., 2018). Interactions across the locus in red blood cells were recently determined to a single base-pair resolution using Micro-C with viewpoint enrichment (MCC) (Hua *et al*., 2021). Of note, in definitive erythrocytes, the enhancer elements each make punctate contacts with *Hba-a1/2* promoters, which footprints recapitulate transcription factor binding sites. The silenced *Hba-x* gene is packaged within its own 10kb compartment, that does not interact with the enhancers in definitive erythroid cells (King *et al*., 2021).

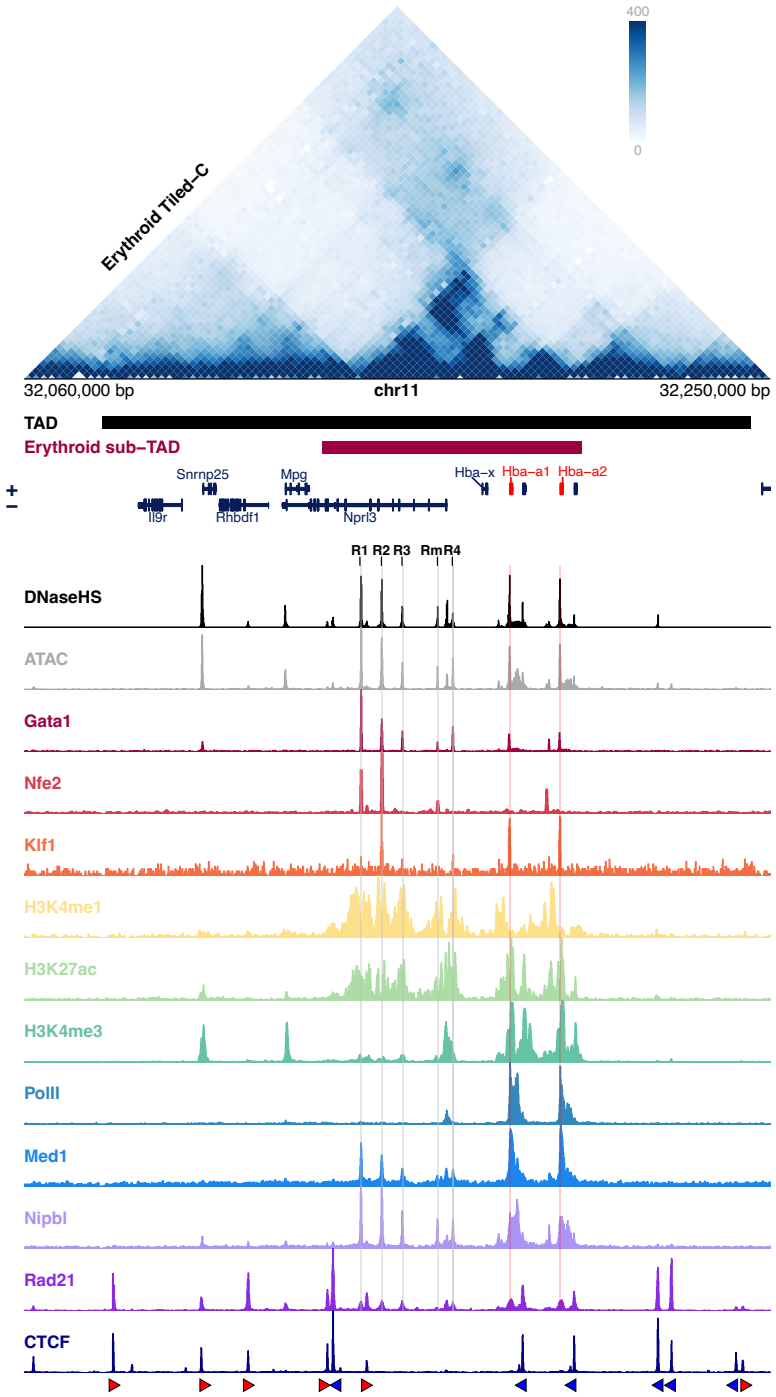

Figure 1.2: Structure of the murine α-globin locus TADs and underlying elements.

Tiled-C contact matrix displaying fine-structure of the α-globin locus at 2kb resolution, the black and red bars under the matrix represent the pre-existing 165 kb TAD (chr11:32,080,000-32,245,000) and the erythroid-specific 65 kb sub-TAD (chr11:32,136,000-32,202,000), respectively. Replotted from (Oudelaar et al., 2020a). Tracks below display DNA accessibility (DNase Hypersensitivty, ATAC-seq), ChIP-seq of transcription factors (Gata1, Nfe2, Klf1), ChIP-seq of histone modifications (H3K4me1, H3K27ac, H3K4me3), transcription machinery (Med1 and PolII) and ChIP-seq of architecture proteins (Rad21/Cohesin, CTCF). Enhancer elements (R1, R2, R3, Rm, R4) are highlighted in grey and orientation of CTCF binding sites are indicated by blue or red arrows. All data presented here has been reanalysed from published datasets from primary mature erythroid cells: DnaseHS (Schwessinger et al., 2017), ATAC, PolII & Rad21 (Hanssen et al., 2017); Nfe2, Gata1 & CTCF (Hughes et al., 2014); Klf1 (Tallack et al., 2010); Med1 (Hay et al., 2016); H3K4me1, H3K27ac & H3K4me3 (Kowalczyk et al., 2012).

Another powerful aspect of using the α-globin locus as a model system is that α-globin are terminal differentiation products, such that they do not affect the expression profile of other genes. Their regulation can also be observed over the course of differentiation during erythropoiesis. Furthermore, we have recently developed an *in vitro* differentiation protocol, which allows for generation of erythrocytes from mESCs (Francis *et al*., 2022), reducing the need to generate in vivo models and hence increasing the throughput of characterisation of genetic manipulations. These erythrocytes have very similar transcription factor binding, accessibility (**Figure 1.3**) and expression profiles to primitive erythrocytes (Francis *et al*., 2022).

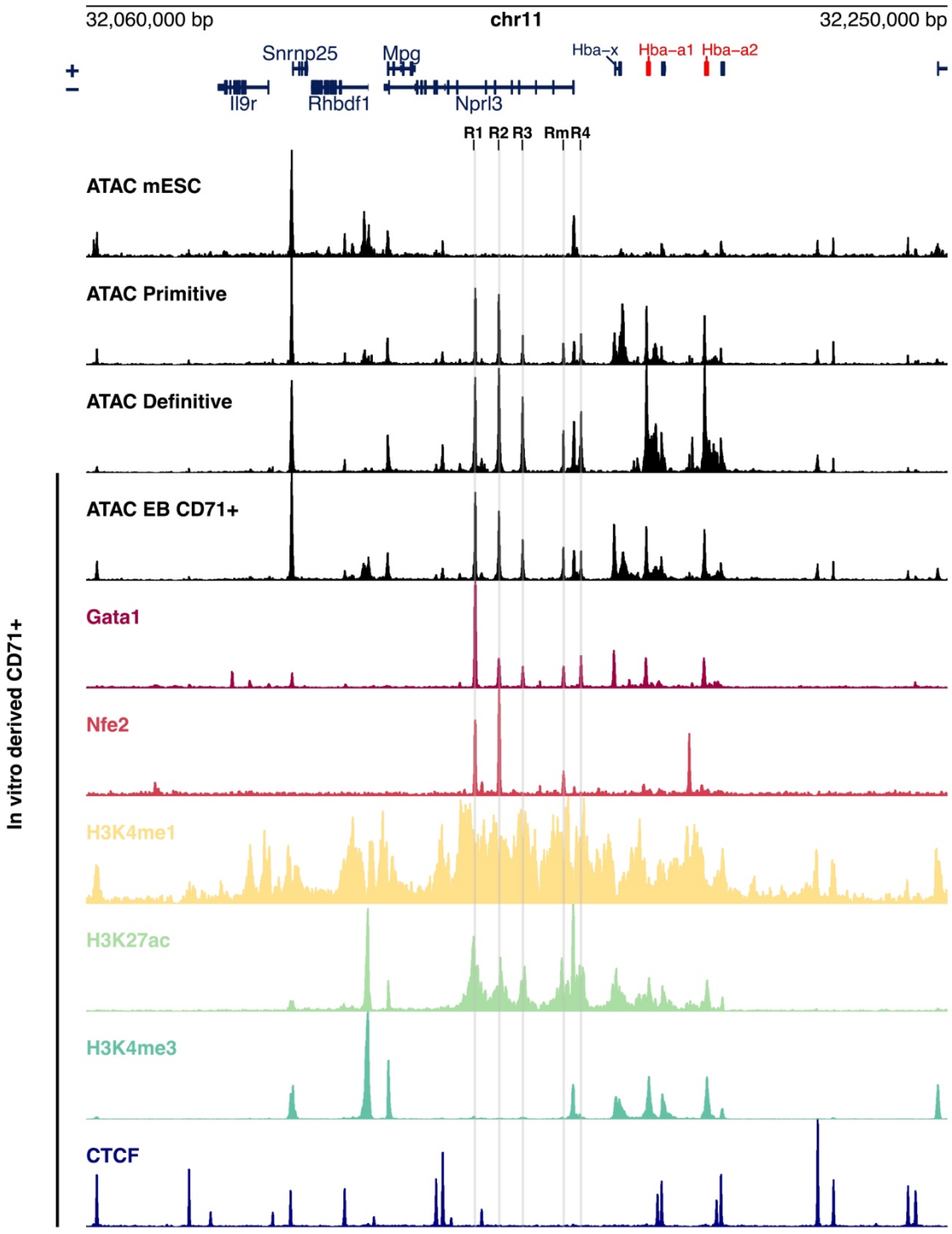

Figure 1.3: *In vitro derived erythrocytes molecularly resemble primitive erythrocytes.*

Accessibility across the α-globin locus as determined by ATAC-seq in embryonic stem cells (mESCs), primitive erythrocytes (E10.5), definitive erythrocytes from APH treated spleens and in vitro derived erythrocytes (EB CD71+). Below shows recruitment of various transcription factors, histone modifications and CTCF in WT CD71+ population of cells derived from in vitro embryoid body differentiation. Note that hba-x is accessible and associated with active chromatin signatures in EB derived cells. Data was previously published in (Francis et al., 2022) and primitive ATAC published in (King et al., 2021).

1.6 Aims and thesis outline

The interplay between cohesin mediated-loop extrusion and gene regulation remains poorly understood, therefore in this thesis I aim to investigate how transcription and CTCF elements interact with cohesin to coordinate formation of the erythroid specific subTAD and optimum gene expression. Firstly, I characterise how transcribing units can act as insulator-like elements, when placed between a cognate enhancer-gene pair. Secondly, I generate a gene inversion to determine how the direction of transcription and relative orientation of elements may affect enhancer driven activation. Finally, I present the generation and characterisation of a genome-integrated synthetic locus, lacking all CTCF sites to determine their combined impact on gene regulation. My study highlights underappreciated aspects which contribute to the optimisation of gene expression whilst also presenting a link between cohesin localisation and transcription.

Chapter 2: Materials and Methods

2.1 mESC genetic engineering

2.1.1 CRISPR/Cas9 editing

Boundary assay

Fragments of the α-globin gene were synthesised by TWIST Bioscience; these sequences were flanked by BsaI sites to facilitate Golden Gate Assembly. The HDR (Homology Directed Repair) template plasmid (pROSA-TV2, created by Prof. Ben Davies and colleagues) contained ~1.3 kb homology arms to either side of the landing site and a [Enh_{CMV}:$P_{β\text{-}actin}$:*HygromysinR*:ΔTK] selection cassette, flanked by homotypic loxP sites to facilitate CRE-recombinase mediated excision following integration into the genome (**Figure 2.1**). Synthesised fragments were cloned into the HDR plasmid by Golden Gate Assembly (NEB) and transformed into competent Top10 *E. coli.* Plasmids were miniprepped following colony screening and overnight growth in LB under antibiotic selection. The resulting HDR:insert vectors were co-transfected, using Lipofectamine (Plus/LTX), with 2 sgRNA:Cas9 expressing plasmids (pX355-pDN1 and pX355-p1) (**Table 2.1**) into parental reporter mESCs (clone G11). The parental reporter cells were generated previously (Francis Thesis, 2019): they were edited to be hemizygous for the α-globin locus, such that one allele had a 86kb deletion spanning the entire locus, and on the remaining allele, 2A-peptide:*mVenus*:3xSV40-NLS was incorporated at the end of the 3rd exon of *Hba-a1,* whilst the distal *Hba-a2* remained unedited (**Figure 2.1**). The emission and absorption spectrums of YFP are compatible with the absorbance spectrum for the haemoglobin protein and a strong nuclear localisation signal (3xSV40 NLS) was used to reduce the chance of signal quenching from cytoplasmic haemoglobin (Papadopoulos *et al*., 2012; Zijlstra *et al*., 1991). Following 24-hour recovery following transfection, cells were placed under Puromycin (1.5 μg/ml) selection for 48 hours, followed by Hygromycin (250 μg/ml) selection for 7 days. Cells were then transiently transfected with a CRE-recombinase expressing plasmid (Pcaggs-Cre-IRES*puro*) to remove the selection cassette and after 24 hours recovery, single cell sorted into 96 well plates with filter sterilsed 1:1 conditioned:fresh media. Single cell colonies were allowed to grow for up to 10 days, at which point they were split for -80 storage, expansion and gDNA extraction. 3 sets of PCRs were used for screening (**Table 2.4**).

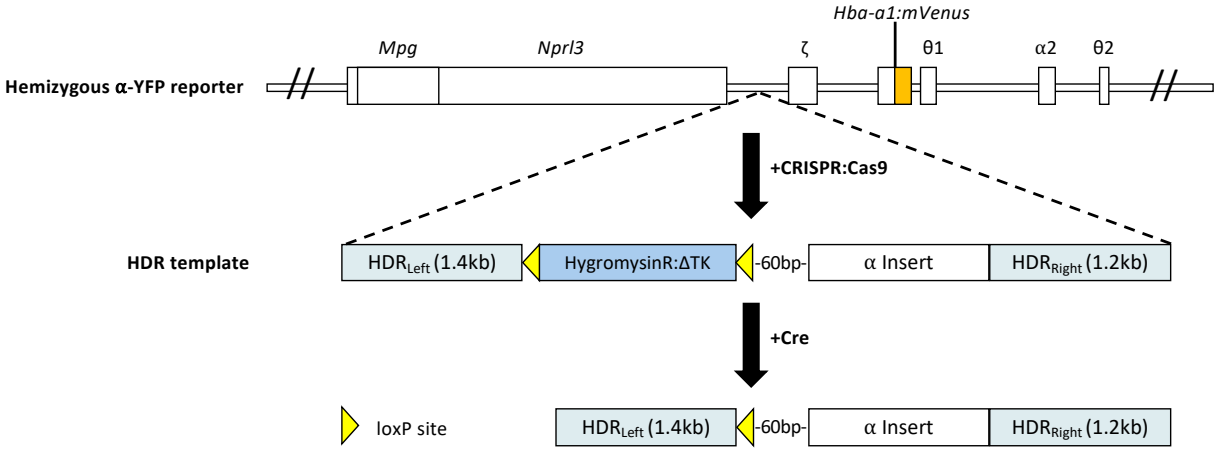

Figure 2.1: Strategy of inserting α-globin fragments into the boundary assay landing site.

The landing site is located between the enhancers and Hba-x gene (chr:32171587-32172056 mm9); ~2.6 kb from the R4 enhancer element. Hba-a1 has been edited so that the C-Terminus of the resulting α-globin protein is fused with 2A-peptide:mVenus:3xSV40-NLS. On the other allele (not shown) there is an 86kb deletion spanning the entire α-globin locus, such that it can be considered hemizygous for the locus. The inserts are introduced to the locus using CRISPR:Cas9 mediated HDR; the HDR template contains 2 arms homologous for either side of the insert site and a HygromycinR:ΔTK selection cassette flanked by homotypic loxP sites, which is removed upon Cre mediated recombination.

Table 2.1: Guide sequences used in landing pad insertions		
5' for landing site	PDN1	GCTGTAGTGTAACTAACTGC
3' for landing site	P1	GCTTCAAGAACTGCCTTCCTG

α-globin inversion

The targeting strategies for the inversion were designed, prepared, and tested by Dr Philip Hublitz and team in the Genome Engineering Core Facility at the Weatherall Institute of Molecular Medicine. Guide sequences for Cas9 were designed using CRISPOR and BreakingCas algorithms; guides with the fewest off-target effects were selected for further validation. The top 3 guides of each targeting region were cloned into pX458 derived plasmids and transfected into tetraploid B16f10 cells for ON-target Surveyor assays; the best performing guides were chosen to be used in targeting.

The parental line was derived previously (Francis, 2019) from RMGR-competent mESCs, derived from E14-TG2a.IV (E14) mESCs, in which one allele had the 86kb deletion across the α-globin locus and the other allele was exchanged for a WT BAC-derived α-globin locus, termed 'WT hemizygous. This cell line was used due to the ease of targeting a hemizygous allele. Due to the high homology around the

two copies of the α-globin genes (*Hba-a1/*2), the first step in the strategy required deletion of the theta pseudogenes and *Hba-a2*, to allow for direct targeting of only one copy of the gene Δ[*Hbq*-1a/Hba-a2/Hbq-1b] (**Figure 2.2**). Two sgRNA:Cas9 plasmids were generated to target the 5' and 3' ends of the target region (pX458-DH031-1 (Addgene 48138, eGFP), and pX458-R-DH031-4 (Addgene 110164, mRuby, respectively). The resulting deleted region was 16.1kb. A HDR vector with ~600 bp homology either side of the expected cut sites were synthesized by GeneArt (Invitrogen) into a pMA-RQ vector; the breakpoint was checked for to make sure no novel accessibility sites were generated using the sasquatch tool (Schwessinger *et al*., 2017). The sgRNA:Cas9 and HDR template plasmids were co-transfected by lipofectamine into WT hemizygous mESCs (clone C3), recovered for 24 hours then single cell sorted with gating for double positive GFP/RFP into 96 well plates with filter sterilsed 1:1 conditioned:fresh media. Following PCR screening and expansion the targeting to invert the remaining *Hba-a1* was performed.

In the second stage of targeting, two sgRNA:Cas9 plasmids were generated to target the 5' and 3' around *Hba-a1* (pX458-DH031-8 (Addgene 48138, eGFP), and pX458-R-DH031-10 (Addgene 110164, mRuby) respectively (**Table 2.2**)) (**Figure 2.2**). The target sites were located ~1.3 kb upstream of the TSS and ~900 bp downstream of the end of the 3'UTR, therefore the inverted region should include all the requisite sequence for native expression. Due to a reportedly high incidence of inversion upon induction of double stranded breaks, the inversion of the gene was allowed to happen by the default repair pathway of Non-Homologous End Joining (NHEJ) (Blayney *et al*., 2020). A clone from the first stage of targeting with one copy of *Hba-a1* (clone 7C, α1-only) was transfected with pX458-DH031-8 and pX458-R-DH031-10 targeting plasmids (**Table 2.2**), recovered for 24 hours then single cell sorted as in stage one. A series of orientation dependent genotyping PCRs were performed to confirm the inversion (α-INV) (**Table 2.5**). Sanger sequencing also confirmed the sequence underlying the inversion breakpoints and once again checked using the sasquatch tool with no *de novo* accessibility sites detected.

Table 2.2: Guide sequences used to create the *Hba-a1* inversion		
5' for Δ[*Hbq-1a*/*Hba-a2*/*Hbq-1b*]	DH031-1	GTTGGATGGTTACGGGCAAGTGG
3' for Δ[*Hbq-1a*/*Hba-a2*/*Hbq-1b*]	DH031-4	TGCCACTGTACGTTTATGACTGG
5' for inversion	DH031-8	TGAGACAGTTCCAAAATACGTGG
3' for inversion	DH031-10	ATTAAAAACTCTACTGACCGAGG

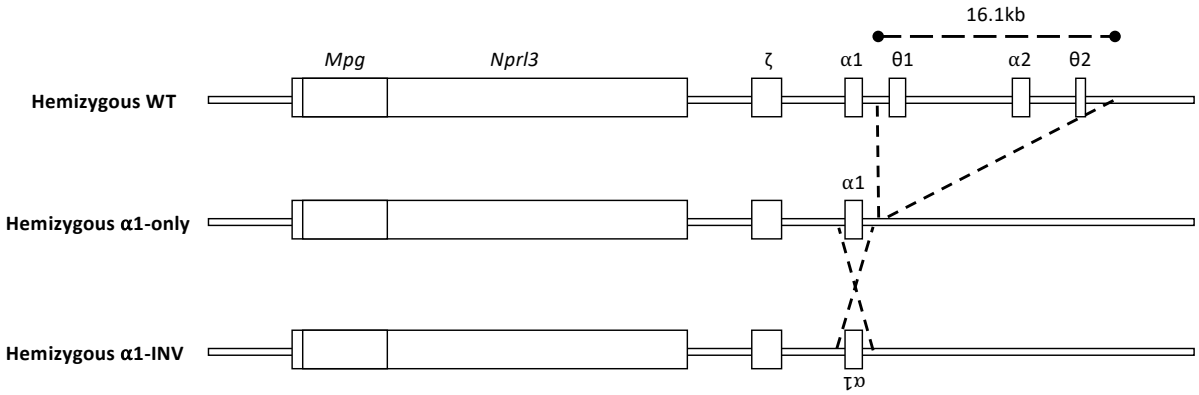

Figure 2.2: Strategy of inverting the α-globin gene.

The strategy began with a cell line hemizygous for the α-globin locus as on the other allele (not shown) there is an 86kb deletion spanning the entire α-globin locus. The α1-only model was generated by introducing a 16.1kb deletion encompassing Hbq-1a, Hba-a2 and Hbq1b using CRISPR:Cas9 mediated HDR. The remaining copy of Hba-a1 was then inverted using CRISPR:Cas9 mediated NHEJ.

2.1.2 Synthetic Recombinase Mediated Genetic Replacement (RMGR)

A synthetic Bacterial Artificial Chromosome (BAC) comprised of a variant α-globin locus with all the internal CTCF sites deleted (ΔCTCF), was generated using a version of the eSwAP-In method (Mitchell *et al*., 2021) by our collaborator Brendan Camellato (New York). Briefly, the synthetic locus was generated by iterative homologous recombination steps in *Saccharomyces cerevisiae* of PCR amplicons that tiled the entire locus. The integrity of the BAC was sequence-verified using short-read sequencing. BACs containing the synthetic ΔCTCF mouse α-globin locus were received as transformed bacteria samples and grown up overnight in LB under kanamycin (20 µg/ml) selection. Cells from 400 ml of culture were lysed using Qiagen P1/2/3 solutions and was filtered through a 70 µm cell strainer. BAC DNA was extracted by caesium chloride (CsCl) density gradient with ethidium bromide staining (EtBr) (**Figure 2.3**). Following separation, BAC DNA was collected and washed with TE-enriched butanol then precipitated in ethanol overnight at -80^{o}C. After dissolving the DNA pellet into sterile water overnight at 4 ^{o}C, the BAC was quantified by dsDNA Qubit and its integrity was checked using restriction digests before transfection.

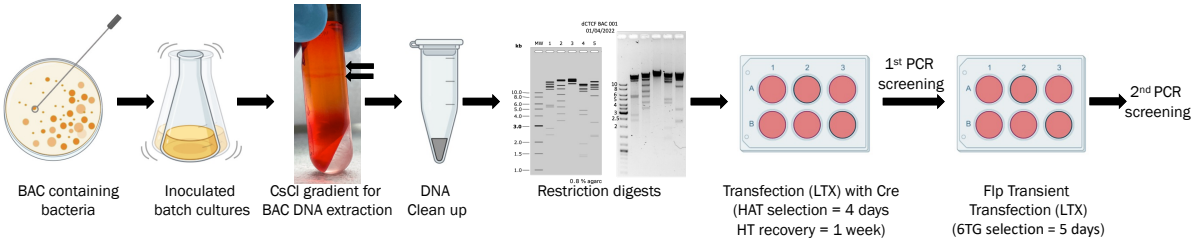

Figure 2.3: Series of steps in generating cells with synthetic Recombinase Mediated Genetic Replacement (RMGR)

Previously, an RMGR-receptive cell line was generated with the machinery in place to allow for BAC integration (Wallace *et al*., 2007). In addition, this cell line was edited further to delete the α-globin locus of non-receptive allele, creating an RMGR-ready hemizygous parental (Francis, 2019). The 86kb region encapsulating the RMGR-ready α-globin locus is flanked by heterotypic lox sites; at the 5' end of the locus a partial HPRT-$^{\Delta 3'}$ selection cassette was introduced, and at the 3' end is a Puromycin resistance/Thymidine Kinase (PuroR/TK) cassette (**Figure 2.4**). Integration into the RMGR-ready locus is facilitated by the inserted sequence having complementary loxP/lox511 sites and partial HPRT-$^{\Delta 5'}$ to complement the HPRT-$^{\Delta 3'}$ in the locus. Upon correct integration the PuroR:TK cassette is removed. 2.5 µg of ΔCTCF BAC was co-transfected with 0.25 µg of CRE-recombinase expressing plasmid (Pcaggs-Cre-IRESpuro) (Smith *et al*., 2002) using lipofectamine LTX/Plus reagents. After 24 hours of recovery, cells were placed under HAT selection for 6 days, followed by 5 days recovery in HT supplemented media. The few cells survived selection were genotyped by PCR over the RMGR breakpoints and then expanded in culture. The complemented *Hprt* gene selection cassette is flanked by homotypic frt sites, therefore was removed with transient transfection with a flippase (Flp) expressing plasmid. After 24 hours of recovery the cells were plated at clonal density and left to recover for another 24 hours. Cells were then placed under 6-ThioGuianine (6-TG) selection (10 µM) for 4 days. Cells were grown for a further 5 days before colony picking. Cells were allowed to grow and expand for 10 days after which they were genotyped by PCR (**Table 2.6**).

Table 2.3: CTCF binding sites deleted in the ΔCTCF model

CBS deletions in ΔCTCF	**Sequence**
HS-59	AATAAATGACCACGAGGTGGCGCCAA
Rhbdf_CTCF	ATGCCAGAGGGGGCATCAGA
HS-39	TGGCCACTGGGGGCGCCATT
HS-38	TCTGCTACCCTCTGGTGGCC
HS-29	GTCCACTGGGTGGCACTTGG
θ1/θ2 minor	AACCAGAAGAGGGCATCAGA
θ1/θ2 major	TGCAGCGCCCCCTGGCGGCCT

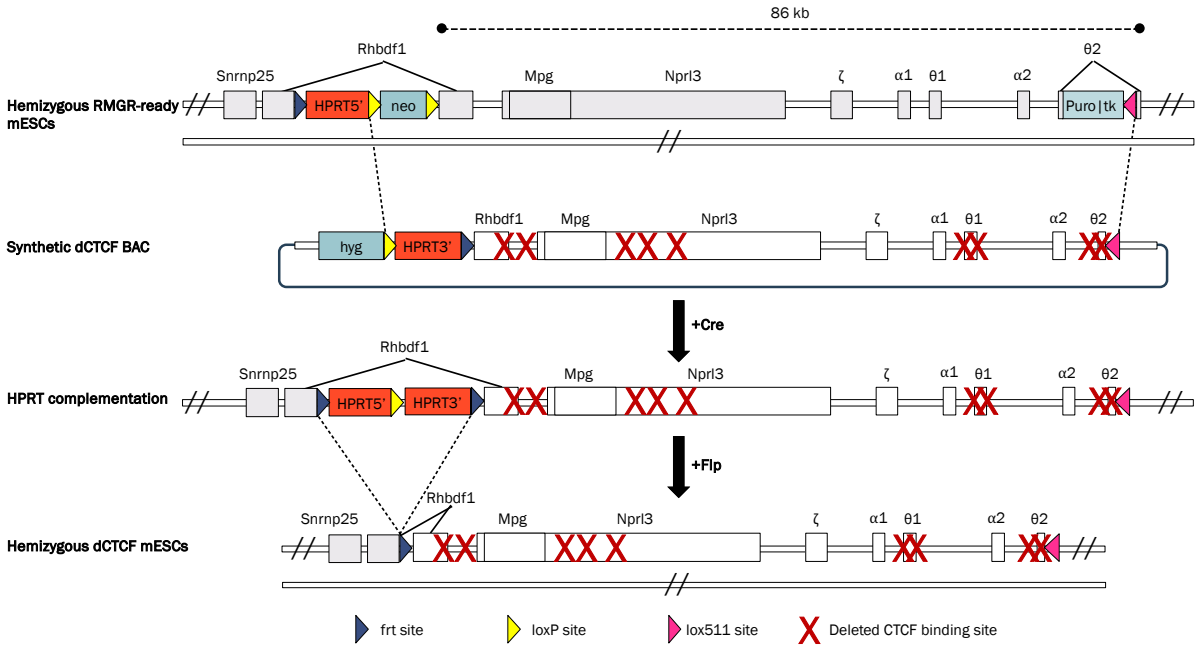

Figure 2.4: Schematic of the genome engineering approach followed to create the CTCF null model.

Parental hemizygous RMGR-ready cells, with Hprt-Δ3'|neomycinR cassette and PuromycinR|tk cassette integrated into the 5' and 3' regions of the α-globin locus (grey). A synthetic locus (white) with all CTCF sites deleted was created in a BAC, with complementary HygromycinR|Hprt-Δ5' cassette. The synthetic locus integrates into the RMGR-ready genome with the use of loxP/lox511:CRE-recombinase activity. The Hprt is removed by Frt:Flippase recombination.

2.1.3 Screening methods

Genomic DNA (gDNA) was extracted by firstly lysing cells in lysis buffer (50 mM Tris, pH 8.0, 1 mM EDTA, pH 8.0, 0.5% Tween 20) supplemented with 60 µg/ml proteinase K and incubating for 24 hours at 37 °C in 96-well plates. Proteinase K was inactivated with 10-minute incubation at 95°C. gDNA was then precipitated with cold isopropanol, washed with 70% ethanol, and resuspended in water and used for genotyping PCRs using the primers listed below.

Table 2.4: Genotyping primers and conditions for landing pad inserts

Primer 1	Primer 2	Polymerase	T_m	Description
U2 = CCTGCAATCTCAACCCTCAGGAATT	D2 = GGCAGAAAATTTAACCAGCAGTATTGTACC	DreamTaq	65°C	Outer right to left homology arm
U1 = TTTTTCGGGGCTTATACTCTCTCTT	2.2R = CAAACATCCTGGGAGAAGGA	DreamTaq	59 °C	Outer left to forward gene body
2.2F = TCCAATATGGACCTGGCACT	D2 = GGCAGAAAATTTAACCAGCAGTATTGTACC	DreamTaq	59 °C	Outer right to forward gene body

2.2R = CAAACATCCTGGGAGAAGGA	D2 = GGCAGAAAATTTAACCAGCAGTATTGTACC	DreamTaq	59 °C	Outer right to reverse gene body
LC1 = GATGGGCGCTGCTCAGTTTGGT	U1 = TTTTTCGGGGCTTATACTCTCTCTT	DreamTaq	62°C	Outer left to reverse promoter
LC2 = ACCAAACTGAGCAGCGCCCATC	U1 = TTTTTCGGGGCTTATACTCTCTCTT	DreamTaq	62°C	Outer left to forward promoter

Table 2.5: **Genotyping primers and conditions for α-globin inversion**

Primer 1	Primer 2	Polymerase	T_m	Description
DH_031_07 = GGCCTCGAACTCAGAAATCTGCC	DH_031_32 = CTTGAATTCAACCCATGGTTCCC	Platinum supermix	60 °C	Product upon 16.1kb deletion
Hba-a1-2_R = GCCGTGGCTTACATCAAAGT	DH_031_32 = CTTGAATTCAACCCATGGTTCCC	Platinum supermix	62.5 °C	Outer right to reverse gene body
DH_031_16 = AAGAGCCAGATAAAGGATGAGCC	DH_031_23 = TGATGATGTCACCCTCCAGTCAG	Platinum supermix	65 °C	Long PCR across the deleted and inverted region

Table 2.6: Genotyping primers and conditions for BAC integration (ΔCTCF)

Primer 1	Primer 2	Polymerase	T_m	Description
Tile56_Fw_1 = GTGGGACTCTGTAGGGACCA	Tile57_Rv_1 = GGGAGGGTTGAGGAGAGAAA	Platinum supermix	60 °C	5' breakpoint – checks for the completed HPRT gene
HFR2_3 = ACTGGGGCTGAATCACATGA	HFR2_4 = ACTCTTTCCCCTTGACCCAG	Platinum supermix	60 °C	5' breakpoint – checks for the completed HPRT gene
Tile31_Fw_2 = GCCCAACCCTGTTTTCCTAT	Tile35_Rv_1 = ATACACCGAGGCATGACACA	Platinum supermix	63 °C	3' breakpoint – checks for removal of the PuroR:TK casette
3422 = CTGTATTGAGCTCTGTCGACATAA CTTCGTATAATGTATACT	3424 = AGCAGCATCACACATCAGG CACACATGTAC	Platinum supermix	64 °C	3' breakpoint – checks for removal of the PuroR:TK casette
Tile51_Fw_1 = GGAGCTGGAGAAAGAGCAGA	Tile57_Rv_1 = GGGAGGGTTGAGGAGAGAAA	Platinum supermix	60 oC	Checks for removal of HPRT following FLP transfection
HS-38F = TGAGAAGGCTGGCCTTTGAG	HS-38R = CCCAGGGAATGAATGCCAGT	Platinum supermix	54°C	PCR across the HS38 CTCF binding site
Theta_CTCF_F2 = AAAATGAAAGGCTTGGAGAGG	Theta_CTCF_R1 = GGCTGAGCTCAAAGATGTCC	Platinum supermix	50°C	PCR across the theta CTCF binding sites
HS-39F = ACAGCAACCATCTGGGTGAG	HS-39R = TGCTGGTGTCTGTGGACAAG	Platinum supermix	65°C	PCR across the HS39 CTCF binding site
HS29R = GCTGGCCTGGAACTAACTCA	HS-29.5R = CAAAACTGCCACATCCTGCC	Platinum supermix	65°C	PCR across the HS29 CTCF binding site

PCR products were purified either by extraction from agarose gels or directly from PCR reactions. Purified products of potentially positive clones were Sanger sequenced by the Sequencing Facility in the Weatherall Institute of Molecular Medicine and aligned against expected sequences using Snapgene.

2.2 Cell culture and isolation

2.2.1 mES cell culture

Mouse embryonic stem cells derived from E14-TG2a.IV (E14) were maintained on gelatinised plates in GMEM media supplemented with 10% FCS, Amino Acids, Sodium Pyruvate, L-Glutamine, β-Mercaptoethanol and LIF as previously described in (Nichols *et al*., 1990; Smith, 1991). Cells were incubated at 37 ^{o}C with 5% CO_2 and trypsinised and passaged very 2 days. Cell suspensions were centrifuged at 1000 rpm at room temperature unless otherwise specified.

2.2.2 Isolation of in vitro-derived erythroid cells

Procedure as performed as previously described in (Francis *et al*., 2022). CD71+ cells derived from this protocol closely resemble primitive erythroblasts. Base media for differentiation was composed of IMDM + 1x Glutamax, 1x Pen/Strep and 140 µM MTG. Gibco FCS was heat inactivated (ΔFCS) at 56^{o}C for 1 hour. mESCs were grown for 48 hours in Adaptation media (15% ΔFCS, 1000 U/ml LIF in Base media). Cells were then de-adhered using 0.05% Trypsin and resuspended to single cell suspension in Resuspension media (10% ΔFCS in Base media, no LIF). Cells were seeded into Differentiation media (Base media supplemented with 15% ΔFCS, 5% PFHM II, 1x L-Glutamine, 1% Transferrin (Merck 10652202001), 1% Ascorbic Acid and additional MTG). Seeding density was between 10,000 to 20,000 cells per 10 cm plate, in 10-15 ml Differentiation media. Plates were incubated at 37^{o}C (5% CO_2) for 7 days with daily shaking to agitate the cells to discourage cell adherence. During this time, cells would grow into spherical embryoid bodies (EBs) with red/pink erythroid cells at the centre (**Figure 2.5a**). Before harvesting, EBs were imaged to check for consistent shape and colour. EBs were collected, washed once in PBS then trypsinised in 0.25% Trypsin at 37 ^{o}C for 5 mins with shaking until fully disaggregated. Cells were quenched with FCS and FACS buffer (10% FCS in PBS).

Table 2.7: Antibodies used for erythrocyte characterisation

Marker	Flour	Clone	Manufacturer
CD71	FITC	R17217	eBioscience, 11-0711-85
CD71	APC	R17217	eBioscience, 17-0711-80
Ter-119	PE	Ter-119	BD Biosciences, 553673
c-kit	APC	2B8	eBioscience, 17-1171-83

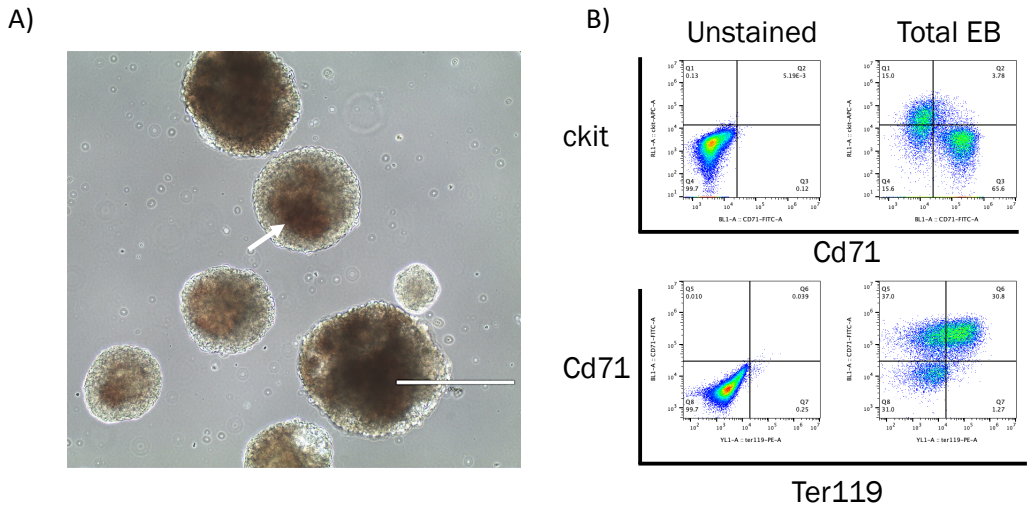

Figure 2.5: Embryoid bodies at day 7 of in vitro differentiation.

A) Arrow indicates the pocket of erythroid cells within the EB (embryoid bodies). Scale bar displays 400 µm. B) FACS plots of disaggregated embryoid bodies (EBs) stained, or unstained, for erythrocyte differentiation stage markers: ckit, CD71 and Ter119.

For erythrocyte enrichment by CD71 selection, disaggregated EB cells were strained through a 70 µm cell strainer, then resuspended in 3-5 ml FACS buffer with 2 µl FITC-conjugated α-CD71 antibody (eBioscience, R17217, cat. 11-0711-85, 0.5 mg/ml) per $1x10^7$ cells. Cells were left to roll at 4 °C for 20 mins, quenched with FACS buffer then resuspended in 200 µl PBE (2 mM EDTA ph8, 0.5 % BSA in PBS) with 20 µl α-FITC beads (Miltenyi 130-048-701) per $1x10^7$ cells and left to roll at 4 °C for 15 mins. Stained cells were washed then resuspended in cold PBE then loaded, via a 30 µm strainer, into an equilibrated LS column (Miltenyi 130-042-401) on a MACS magnet. Magnetised columns were washed through 3 times with cold PBE (negative fraction). The CD71 positive fraction was collected by releasing the column from the magnet, adding PBE and applying pressure with a plunger. The cells were resuspended in FACS buffer for counting, flow cytometry and preparation for other assays. For cells with no internal YFP (Yellow Fluorescent Protein), immunophenotype was checked by staining 200,000 cell aliquots with Ter-119-PE (1:100), CD71-FITC (1:200), c-kit-APC (1:100) and Hoechst (1:10,000) (**Table 2.7**). Cells were stained at 4 °C for 30 mins, washed in FACS buffer then analysed, with compensation, on an Attune NxT machine (**Figure 2.5b**). FACS data was analysed using FlowJo software and gating was kept consistent across replicate experiments.

2.3 Measurement of gene expression

2.3.1 RNA extraction and RT-qPCR

On the day of harvesting, between 250-500,000 cells were aliquoted in FACS buffer, centrifuged at 3500 rpm for 5 mins and pellets were resuspended in TRI-reagent, then stored at -80 °C. Prior to extraction, surfaces and utensils were cleaned with RNaseZap. RNA was extracted using the Zymo RNA MicroPrep kit, as per manufacturers' instructions with an extension of on column DNaseI treatment to 45 mins. RNA was quantified by HS RNA qubit and analysed using RNA Tapestation to check for overall quality and if there was degradation (**Figure 2.6**) (Samples were allowed through if RINe >=8 (**Figure 2.7**)). For samples within the same experiment, the same mass of RNA from each sample was used as input into cDNA synthesis allowing accurate comparison across samples. cDNA synthesis was done using the Superscript III Super mix as per manufacturer's instruction. For all experiments, "No Reverse Transcriptase" controls were included to check if there were signals from undigested genomic DNA. RT-qPCR was performed using TaqMan Universal Master Mix and TaqMan probes (**Table 2.8**); all plates were run on the same QuantStudio machine on the same day to reduce technical variation. The ΔΔCt method was used quantify the expression firstly relative to *RPS18,* then relative to the mean of β-like globins (*Hbb-bs, Hbb-bt, Hbb-bh1, Hbb-bh2, Hbb-y)* unless otherwise specified. Normalising against the β-like globin genes helps to correct for differences arising from heterogeneity across cell lines in *in vitro* differentiation.

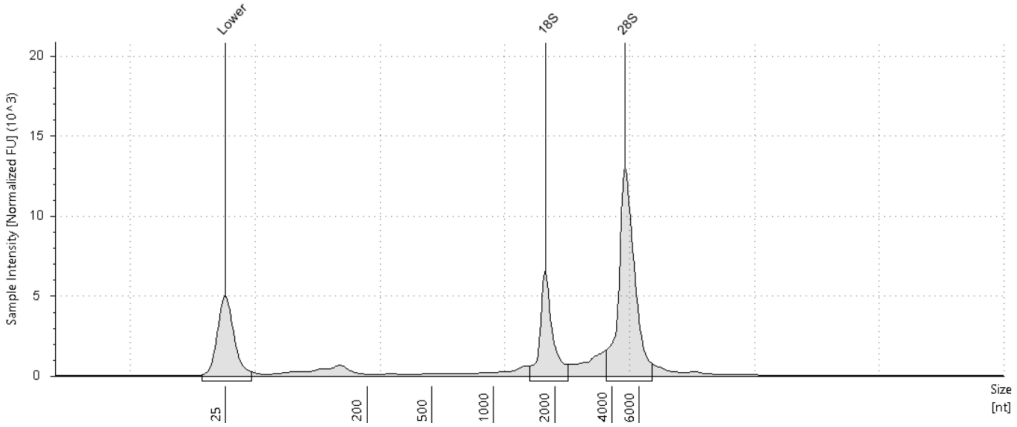

Figure 2.6: Representative tapestation trace of RNA used as inputs into rt-qPCR.

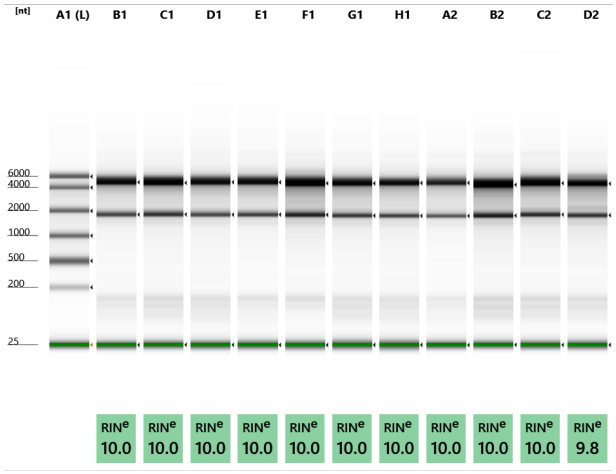

Figure 2.7: Tapestation traces and RINe scores of RNA used as inputs into rt-qPCR

Table 2.8: Taqman Probes used for rt-qPCR			
Gene	**Assay ID**	**Gene**	**Assay ID**
Hba-a1/2	Mm02580841_g1	Il9r	Mm00434313_m1
Hba-x	Mm00439255_m1	Mpg	Mm00447872_m1
Hbb-y	Mm00433936_g1	Nprl3	Mm01193449_m1
Hbb-bh1	Mm00433932_g1	Rhbdf1	Mm00711711_m1
Hbb-bs/t	Mm01611268_g1	Snrnp25	Mm00547218_m1
RPS18	Mm02601777_g1		

2.3.2 YFP tag design and clonal readout

As mentioned previously, the parental reporter line used for the landing pad was engineered so that the enhancer proximal *Hba-a1* was edited to have 2A-peptide:mVenus:3xSV40-NLS incorporated before the stop codon. The 2A-peptide facilitates self-cleaving from the α-globin peptide, and nuclear localisation signal repeats (3xSV40-NLS) determine localisation of the Venus (YFP) protein. The native *Hba-a1* 3' untranslated region (3'UTR) is maintained. Embryoid bodies were disaggregated by trypsinisation into single cell suspension as in 1.2.2. Cells were suspended in FACs buffer (10% FCS in PBS) and stained with α-CD71-APC (1:8000, eBioscience, 17-0711-80) (**Table 2.7**) at 4^{o}C for 30 mins, washed in FACS buffer and live/dead stained with Hoechst (1:10,000). Cells were analysed, with compensation, on an Attune NxT machine. The gating strategy comprised of gating for live single cells then stringent gating was used for CD71+/YFP+ cells (**Figure 2.8**). Median YFP florescence was

measured in FlowJo, plotted and statistics calculated in Prism (version 10.0.3 for MacOS). Density plots were generated in R using ggplot and ggridges packages.

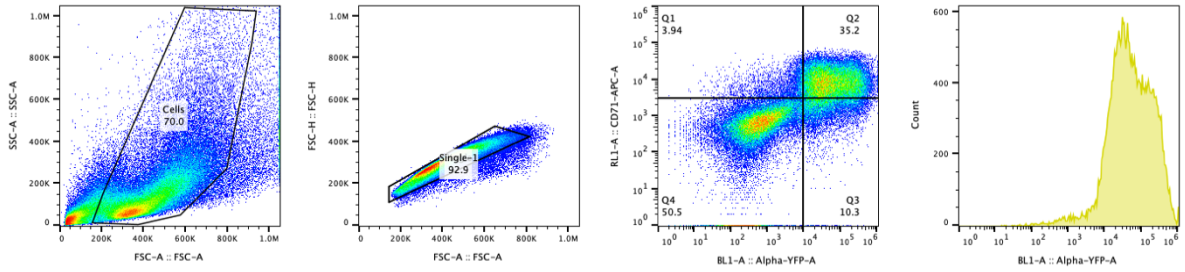

Figure 2.8: Representative FACS gating for measurement of YFP fluorescence as a proxy for α-globin expression.

2.4 Next Generation Sequencing assays

All Next Generation Sequencing was performed in-house by the WIMM sequencing facility on an Illumina NextSeq machine.

2.4.1 ATAC-seq

Assay for Transposase Accessible Chromatin (ATAC-seq) was performed as described previously (Buenrostro *et al*., 2015; Hay *et al*., 2016). $1.5x10^5$ cells were pelleted at 500 g for 10 min at 4°C then quickly resuspended in 50 µl cold ATAC lysis buffer (Tris HCl pH 7.5 10mM, NaCl 10mM, $MgCl_2$ 2 mM, Igepal CA-630 0.1%) and centrifuged 500 g for 10 min at 4°C. After removing the supernatant, the nuclear pellet was resuspended in 50 µl transposition mix (1x Tagment DNA buffer, TDE1 Tagment DNA Enzyme) from the Illumina Tagment DNA kit (20034197) and then incubated at 37 °C for 30 mins. The TDE1 Enzyme from this kit is pre-loaded with Illumina adaptor and integrates the adaptor upon transposition and fragmentation. Tagmented DNA was purified using the Qiagen MinElute Kit (28004). For sequencing, custom indices were added to fragments using NEBNext High-Fidelity 2X PCR Master Mix (M0541) and purified using a Qiagen PCR clean-up kit (28104). Libraries were checked using HS D1000 screen tape on a Tapestation machine; nucleosomal patterning is expected from tagmentation generated libraries (**Figure 2.9**). Libraries were quantified using the KAPA quantification kit and pooled to final concentrations of 4nM. Libraries were sequenced paired end using NextSeq 500/550 High Output Kit v2.5 (75 Cycles) kits (20024906) and sequenced on an Illumina NextSeq machine. Each library was allocated approximately 30-40 million reads.

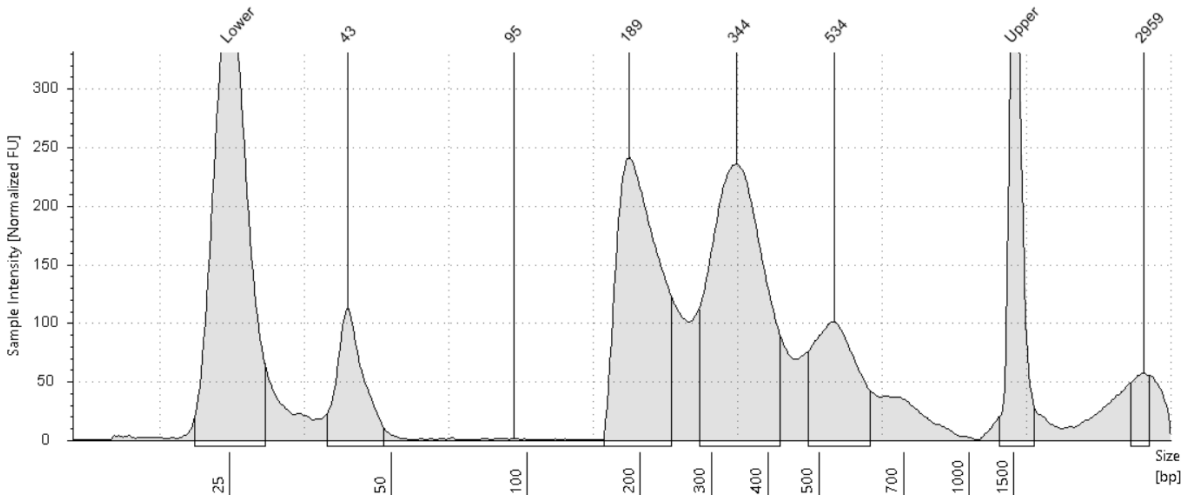

Figure 2.9: Representative tapestation trace of an ATAC library.

Nucleosome patterning can be seen as peaks at 189, 344 and 534 bp.

2.4.2 RNA-seq

Differentiations were performed in triplicate and for each sample $2x10^5$ cells were collected in TriReagent and stored at -80°C prior to extraction. Prior to extraction, surfaces and utensils were cleaned with RNaseZap. RNA/DNA was extracted by phase separation by addition of 1-Bromo-3-Chloropropane and centrifuging in PhaseLock tubes at 12,000 g for 5 mins at 4°C. RNA/DNA was precipitated using GlycoBlue and propan-2-ol and washed in cold 75% ethanol. Pellets were air dried then RNA/DNA was resuspended in DEPC-treated water and checked for quality on HS RNA Tapestation and quantified using HS RNA Qubit reagents. To remove genomic DNA contamination, samples were incubated with Turbo DNase 25 °C for 60 min and then precipitated at -80°C overnight with ethanol, GlycoBlue and 90mM sodium acetate. RNA was again tested on Tapestation and quantified by HS RNA Qubit. rRNA depletion was performed using the NEBNext rRNA Depletion Kit (E6310). PolyA Plus/Minus selection was then performed using the NEBNext Poly(A) mRNA Magnetic Isolation Module (E7490). RNA libraries were then generated using the NEBNext Ultra II Directional RNA Library Prep Kit (E7760) and indexed with NEBNext Multiplex Oligos for Illumina (E6609S). Libraries were checked by Tapestation using HS D1000 screen tape and quantified using a combination of KAPA quantification and ds DNA Qubit quantification. With the aim to capture important SNPs, approximately 100 bp apart, within one read libraries were sequenced paired end using NextSeq 500/550 Mid Output Kit v2.5 (300 Cycles) (20024905), allocating ~5 million reads per sample.

2.4.3 ChIP-seq

For CTCF ChIP-seq, $1x10^6$ CD71+ cells were fixed with 1% formaldehyde for 10 minutes at room temperature then quenched in cold glycine (125mM), washed once in PBS and snap frozen. Chromatin

immunopreciptiation (ChIP) was performed using the Millipore ChIP Assay Kit as described by (Hanssen *et al*., 2017). Fixed cells were lysed in lysis buffer and sonicated to ~200-500 bp on a Bioruptor Pico sonicator (5 cycles, 30 sec on/off, 4°C). Fragmented chromatin was incubated with 3 µl of polyclonal α-CTCF (Millipore, 07-729) and rotating overnight at 4°C. Chromatin was immunoprecipitated following incubation with Protein-A conjugated agarose beads, following the kit instructions for washes. DNA was purified by phenol-chloroform-isoamylalcohol extraction and enrichment:input was tested by qPCR using Fast SYBR green reagents and qPCR primers listed in **Table 2.9**.

For Rad21 ChIP-seq, 1-5x10^6 CD71+ cells were fixed with DGS (2 µM) for 30 minutes at room temperature, spun down and resuspended in 1% formaldehyde for 30 minutes at room temperature, rinsed in PBS and snap frozen prior to precipitation. Cells were lysed in SDS lysis buffer (1% SDS, 10 mM EDTA, 50 mM Tris-HCl ph8, 1x PIC) and sonicated to ~200-500 bp on a Covaris sonicator with the following settings: 600secs; 75 power; 25% Duty Factor; 1000 cycles per burst. Chromatin was diluted in dilution buffer (0.01% SDS, 1.10% Triton X-100, 1.2 mM EDTA, 16.7 mM TrisHCl ph8, 167 mM NaCl, 1x PIC). A 1:1 mix of proteinA:proteinG conjugated Dynabeads were added to chromatin prior to immunoprecipitation to reduce non-specific binding. Cleared chromatin was incubated with 4.4 ug polyclonal α-Rad21 (ab154769) with rotating overnight at 4°C. Chromatin was then incubated with rinsed A:G dynabeads for 5 hours, to bind to the antibody. Beads were then washed 3x with RIPA buffer (50 mM Hepes-KOH, 500 mM LiCl, 1 mM EDTA, 1% IGEPAL CA-630, 0.7% Na-Deoxycholate) and chromatin was eluted from beads using elution buffer (1% SDS, 10 mM EDTA, 50 mM TrisHCl ph8) and de-crosslinked at 65°C overnight. Chromatin was treated with 0.5 µg RNase (60 minutes at 37°C) followed by 20 µg Proteinase K (60 minutes at 37°C). Finally, DNA was purified using the ChIP DNA clean and concentration kit (Zymo), quantified by Qubit and profiled with Tapestation (**Figure 2.10**). Enrichment of immunoprecipitated DNA over input was determined by qPCR using Fast SYBR green reagents and qPCR primers listed in **Table 2.9**.

Libraries for ChIP-seq were generated using the NEBNext Ultra II DNA library prep kit as per manufacturer's instructions; 10 cycles were used during indexing. Libraries were sequenced paired end on an Illumina NextSeq platform using 500/550 High Output Kit Kit v2.5 (75 Cycles) allocating ~40 million reads per sample.

Table 2.9: Primers for ChIP qPCR		
Primer ID	**Sequence**	**Description**
Beta_5'HS2(CTCF)_F1	GCCCTCAGGTTGTCAACTAAAGC	Positive control for ChIP
Beta_5'HS2(CTCF)_R1	CGGAAATCAGCGGAACACTTC	Positive control for ChIP
Chr15_1F	GTAAGATGCCATTGTCCTGACC	Negative control for ChIP
Chr15_1R	TGTGTTCTCCTCCCTTCCTT	Negative control for ChIP

Table 2.10: Antibodies used for ChIP		
Antigen	**Product number**	**Antibody type**
Anti-Rad21	ab154769	Rabbit polyclonal
Anti-CTCF	07-729	Rabbit polyclonal

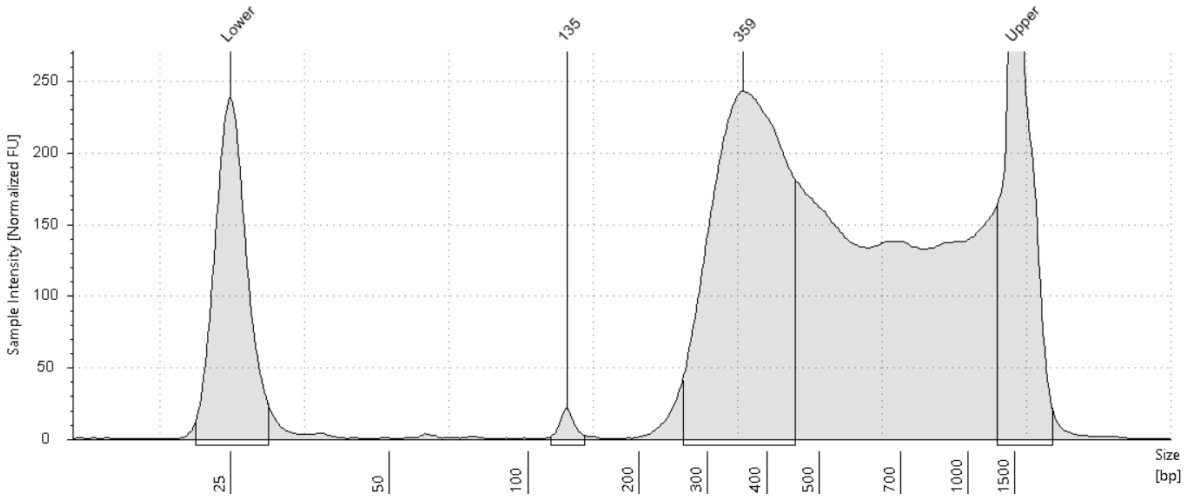

Figure 2.10: Tapestation trace of a ChIP library. Following indexing and amplification.

2.4.4 Tiled-C

TiledC was performed as first described in (Oudelaar *et al*., 2020). The protocol is a variation on the chromatin confirmation capture (3C) (Dekker *et al*., 2002); wherein oligonucleotides for region(s) of interest are used to enrich for interactions from 3C libraries (Davies *et al*., 2016; Downes *et al*., 2022; Hughes *et al*., 2014). Between 2-5x10^6 *in vitro* derived CD71+ erythrocytes were fixed with 2% formaldehyde for 10 minutes at room temperature, quenched with cold glycine (125 mM), washed once in PBS and resuspended in cold lysis buffer (10 mM Tris pH8, 10 mM NaCl, 0.2% Igepal CA-630, cOmplete Protease Inhibitor Cocktail) incubated on ice for 40 minutes with regular pipetting to disaggregate the cell pellets. This lysis buffer is formulated to be mild and aid in permeabilization of the cell and nuclear membranes, whilst minimally disrupting the nuclear structure. Cells were snap frozen in lysis buffer and stored at -80^{o}C. Cells were thawed and Triton X 100 added to a final

concentration of 1.7% to quench SDS in the buffer. An aliquot was taken here to act as an undigested control (C1). DpnII enzyme was added three times to cells over the course of 28 hours and incubated at 37°C with intermittent shaking (500rpm 30 sec on/30 sec off) to digest crosslinked chromatin. A second aliquot was taken here as a digested control (C2). DpnII digested chromatin was then treated with T4 ligase and incubated at 16°C with intermittent shaking (500rpm 30 sec on/30 sec off) for 22 hours. Throughout digestion and ligation, intact cells/nuclei were visible when observed under a microscope. Supernatant was removed from the cell pellets (and controls) and resuspended in TE with proteinase K and incubated overnight at 65°C. DNA was then incubated for 1 hour at 37°C with RNase (7.5 mU), purified by phenol-chloroform isoamylalcohol extraction and precipitated overnight at -80°C. Digested/ligated DNA was dissolved in TE buffer overnight at 4°C. Controls and ligated DNA were checked by electrophoresis in 1% agarose gel (**Figure 2.11**). A clear shift was observed in the ligated DNA samples (D+L), however undigested controls (C1) suggested there may have been degradation.

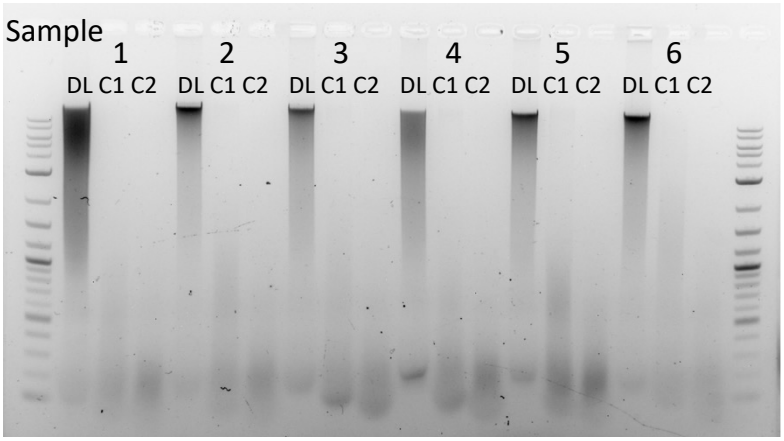

Figure 2.11: Quality control gels of DNA used as inputs in TiledC.

DL = digested and ligated DNA, C1 = undigested control, C2 = DpnII digested control (unligated).

Due to evidence of successful ligation, DNA was then taken through for library preparation and capture. 3 µg (3 x 1 µg) of Digested/ligated DNA was sonicated to ~200 bp fragments using a Covaris ME220 (settings: Duration: 130s, Peak Power: 70, Duty % Factor: 20, Cycle/Burst: 1000, Average power: 14), purified with Ampure XP beads and size distribution checked by Tapestation (**Figure 2.12**).

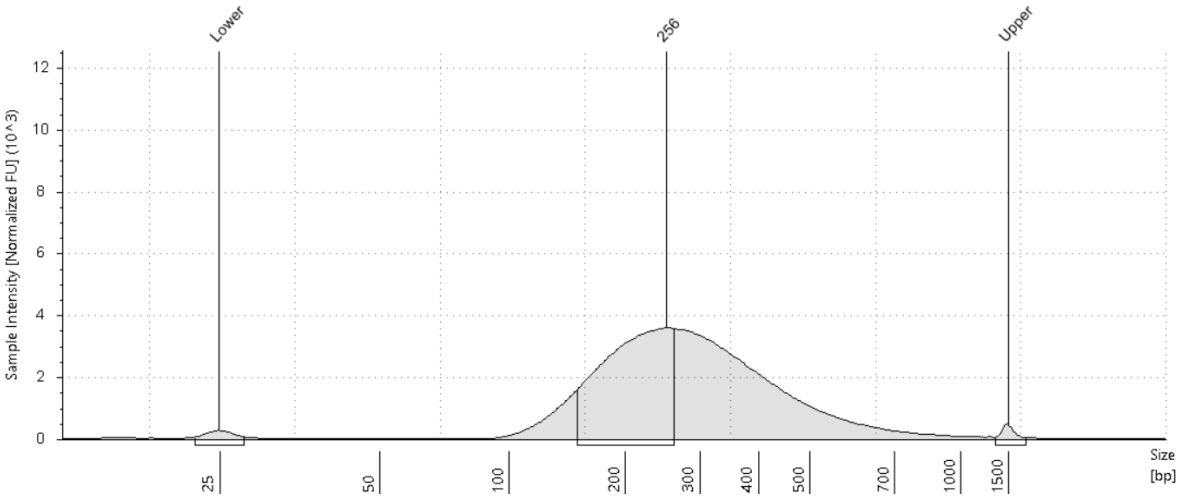

Figure 2.12: Representative Tapestation profile of sonicated DNA used in TiledC library preparation.

Libraries with unique indexes for each sample were made using an input of 1 µg of sonicate and the NEBNext Ultra II Directional RNA Library Prep Kit (E7760). 10 cycles of PCR amplifications were used. Capture was performed using biotinylated ds-oligonucleotides (TWIST Bioscience), designed to tile 3 Mb and bind to every DpnII fragment across the α-globin locus and used previously in (Oudelaar *et al*., 2020b). Probe hybridisation was carried out on a total of 2 µg of 8 multiplexed libraries (250 ng each) using the TWIST bioscience blocking and hybridization reagents and incubated overnight at 70 °C. Hybridised material was bound and pulled from solution using Streptavidin dynabeads. A final round of PCR amplification was performed using the KAPA HiFi HotStart ReadyMix (9 cycles) resulting in a final library (24 ng) was sequenced paired end on an Illumina NextSeq platform using 500/550 High Output Kit Kit v2.5 (75 Cycles) allocating ~40 million reads per sample.

2.5 Bioinformatic analysis

2.5.1 Custom genome sequences

To better analyse NGS data from the models produced in this thesis, several custom genome sequences were generated using an in-house pipeline. Firstly, the mm10 whole genome fasta sequences was edited to delete or insert desired sequences utilising the getfasta function in the BedTools suite (Quinlan and Hall, 2010). Secondly, the gene annotation file (.GTF) was also edited to

match the shifts in coordinates generated by the fasta edits. Thirdly, custom index files were generated using bowtie2-build (Langmead and Salzberg, 2012) or STAR --runMode genomeGenerate (Dobin *et al*., 2013) for use with their respective aligners. For initial visualisation, the custom genome references were uploaded to the UCSC genome browser (Kent *et al*., 2002) via a trackhub (Raney *et al*., 2014) using the faToTwoBit function of ucsctools. Custom genome references and annotations were also imported for use in the plotgardener suite (Kramer *et al*., 2022) using Bioconductor packages BSgenomeForge (Pagès, 2017) and makeTxDbFromGFF function of GenomicFeatures (Lawrence *et al*., 2013).

2.5.2 ATAC-seq and ChIP-seq

ATAC-seq and ChIP-seq FASTQ files were analysed using a sequential pipeline. Reads were aligned using Bowtie2 (Langmead and Salzberg, 2012) to the relevant mouse reference genome with multimapping set (-k 2). The resulting sam files were then processed and converted to bam format using SAMtools (Li *et al*., 2009); reads were filtered for mapped reads (view -F4), sorted, PCR duplicates removed (rmdup) and the resulting bam file was indexed. RPKM normalised BigWigs were generated using the bamCoverage function of deeptools with --smoothLength 100 and –binsize 20 (Ramírez *et al*., 2016). Bigwigs from multiple replicates were averaged using wiggletools (Zerbino *et al*., 2014) and converted back into Bigwig format using wigToBigWig function of ucsctools (Kent *et al*., 2002).

2.5.3 RNA-seq

PolyA selected RNA-seq FASTQ files were analysed using a sequential pipeline. Reads were aligned using STAR (Dobin *et al*., 2013) to a custom genome with a single copy of *Hba-a1* in its native position and *Venus* gene sequence appended as an independent chromosome. The resulting sam files were filtered for properpairs (-f3 -F4), sorted, PCR duplicates removed and indexed, resulting in indexed bam files. RPKM normalised and strand specific BigWigs were generated using the bamCoverage function of deeptools with --smoothLength 100 and --filterRNAstrand (Ramírez *et al*., 2016). The bigwig files from each replicate were merged by averaging in Wiggletools (Zerbino *et al*., 2014) and converted back into Bigwig format using wigToBigWig (Kent *et al*., 2002). Read coverage over each gene and over regions flanking the landing site were calculated using Rsubread featurecounts (Liao *et al*., 2014) and RPKM normalised using edgeR (Chen *et al*., 2016). For variant specific quantification, reads mapping over the positions of each of the SNPs were isolated from bam files using the mpileup function of BCFtools with the --annotate flag active to report high quality allele depth (Li, 2011a, 2011b). The results were filtered to only count bases with a minimum base quality of 30 (--min-BQ

30). Base identity and allele depth (AD) at the SNP position was extracted from the resulting vcf file, allowing quantification of insert and endogenous gene expression.

2.5.4 Tiled-C

FASTQ files from NGS sequencing of TiledC libraries were analysed by 2 methods. Data was first analysed using the HiCPro pipeline (Servant *et al*., 2015) using the coordinates of the regions covered by the tiled probes (mm9 chr11:29902951-33226736) and merging replicates using the -s merge_persample function. Briefly, reads are aligned by Bowtie; alignments were then assigned to DpnII restriction fragments. Paired interactions and contact maps are generated from valid pairs covering two different restriction fragments, whilst PCR duplicates and invalid ligation products are discarded. An Iterative Correction and Eigenvector decomposition (ICE) method is used to normalise data and correct for biases arising from GC content. Libraries generated here comprised of approximately 60-110 million valid interaction pairs of which approximately 30% were cis contacts, and 2-10% trans contacts. This unfortunately was not sufficient for 2 kb resolution as reported previously in (Oudelaar *et al*., 2020b). Data was analysed by a second pipeline which is in development, CapCruncher (v0.3.6) (https://github.com/sims-lab/CapCruncher), to generate virtual capture profiles from the TiledC data. In this case the R2 enhancer was used as a viewpoint (mm9, chr11:32151063-32151879). The pipeline consists of similar stages of analysis and generates averaged normalised interaction profiles from combined replicates as bigwigs.

2.5.5 Visualisation

ChIP-seq heatmaps were generated using the computeMatrix and plotHeatmap functions of deepTools (Ramírez *et al*., 2016). All bigwig tracks and interaction heatmaps in this thesis were generated using the plotgardener suite (Kramer *et al*., 2022). Other plots were generated using GraphPad Prism (version 10.0.3 for MacOS), utilising in-built statistical analysis.

Chapter 3: Transcription of an ectopic α-globin gene within a landing pad reveals insulator activity

3.1 Introduction

Metazoan chromosomes are folded and organised into TADs with domain borders highly enriched for the insulator protein CTCF (Dixon *et al*., 2012). Such CTCF bound flanking domains are predominantly found in convergent orientation (Rao *et al*., 2014); this is indeed the case for the α-globin locus (Hanssen *et al*., 2017). CTCF sites at the 5' end of the α-globin locus were shown to form the 5' boundary of a self-interacting erythroid compartment containing the enhancer elements and α-globin genes (sub-TAD). Upon combined deletion of the HS-38/39 CTCF sites, the sub-TAD spread to encompass the non-erythroid 5' genes and led to their inappropriately enhancer-induced activation (Hanssen *et al*., 2017). In sharp contrast, when 3' CTCF sites were deleted in pairs, HS+44/48 and θ1/θ1, there was little change in domain structure and equally no change in expression of either the 3'gene or the α-globin genes (Harrold *et al*., 2020). We have since characterised a further model, wherein all 4 3' CTCF sites were deleted in combination (HS44/HS48/θ1/θ1) (Susannah Holliman, unpublished). Again, there was no significant change in domain structure or gene expression suggesting that the 3' CTCF sites are not essential for limiting the 3' interactions of the sub-TAD. Capture-C profiles generated from the 3'CTCF site deletion models all displayed a common characteristic; enhancer interactions cut off sharply at the α-globin genes.

In addition to this, the enhancer proximal *Hba-a1* gene is expressed preferentially to the more distal *Hba-a2* with total α-globin transcripts comprised of 70% *Hba-a1* : 30% *Hba-a2;* this is reflected in SNP-specific interaction profiling, with more frequent enhancer interactions observed with *Hba-a1* relative to *Hba-a2* interactions (Davies *et al*., 2016). This of course could be due to the presence of the intervening CTCF sites between the two genes θ1/θ2, however upon deletion of these sites (both individually and in combination) preferential expression and enhancer interaction with the proximal *Hba-a1* was still observed (Harrold *et al*., 2020). This raised the hypothesis that the α-globin genes may themselves be forming the sub-TAD boundary, with *Hba-a1* also potentially acting as a partial boundary to *Hba-a2.* However, this remained to be functionally tested.

3.2 Results

3.2.1 Boundary Activity Assay design

The CTCF sites bounding the α-globin sub-TAD are inequivalent in their insulation strength, reflecting the general state of CTCF sites genome-wide. Our group set out to investigate what underlies this functional heterogeneity. We harnessed the α-globin sub-TAD structure to design a boundary activity assay whereby sequences to be tested for their boundary activity are inserted between the α-globin enhancer cluster and the α-globin genes (Tsang *et al*., 2023 BioRxiv and in review).

Boundary assay

A boundary activity reporter cell line has been designed and generated in mESCs to evaluate CTCF binding sequences in an otherwise controlled context at the α-globin locus (Tsang *et al*., 2023 BioRxiv and in review). The key features of the design are: 1) A single insertion site of query sequences is located between the globin genes and the enhancer elements (chr:32171587-32172056 mm9); 2) *Hba-a1* is tagged with YFP to allow readout of α-globin expression by FACS; 3) the cell line is hemizygous for the α-globin locus to allow for increased targeting efficiency and single allele genomics; finally, 4) engineered mESCs are differentiated into erythroid cells via *in vitro* differentiation protocol (Francis *et al*., 2022) allowing read out in the appropriate cell context. By testing a range of sequences with known boundary activity, we demonstrated that the expression of *Hba-a1*:*Venus* can be used as a proxy read-out for boundary strength; for example, insertion of a strong boundary results in significant reduction in YFP fluorescence .

Insert design – Promoter

As I wanted to test whether the α-globin genes can act as insulator-like elements, I synthesized six DNA fragments covering regions of the α-globin gene for insertion into the landing site (**Figure 3.1**): 1) the α-globin promoter, 2) the α-globin gene body, 3) the α-globin promoter and gene body, and all three in the reverse orientation.

The promoter sequence inserted was selected by using a combination of DNaseI footprinting, chromatin accessibility profiles, and mammalian sequence conservation analyses. The transcription initiation site at the α-globin gene was also confirmed using scaRNA-seq performed on primary erythrocytes, which allows for precise designation of transcription initiation sites and promoter-proximal pausing downstream of initiation (Larke *et al*., 2021). The resulting sequence included 373 bp from the transcription start site, encompassing the 5'UTR and much of the region accessible in

ATAC-seq in primary erythroid cells (chr11:32183340-32183712 mm9) (**Figure 3.1**). This region also contains a multiple conserved sequence (MCS) at the promoter (Hughes *et al*., 2005), general transcription factor binding sites (CAAT and TATA boxes) and erythroid-specific transcription factor binding sites, Gata1 and Klf1. With this design, I predict that I have captured all elements in the native α-globin promoter to allow transcription initiation from the inserts. To discriminate sequence reads between the native and inserted promoter sequences, a single TSS proximal SNP (-9T>A) was included in the promoter sequence (**Figure 3.1**).

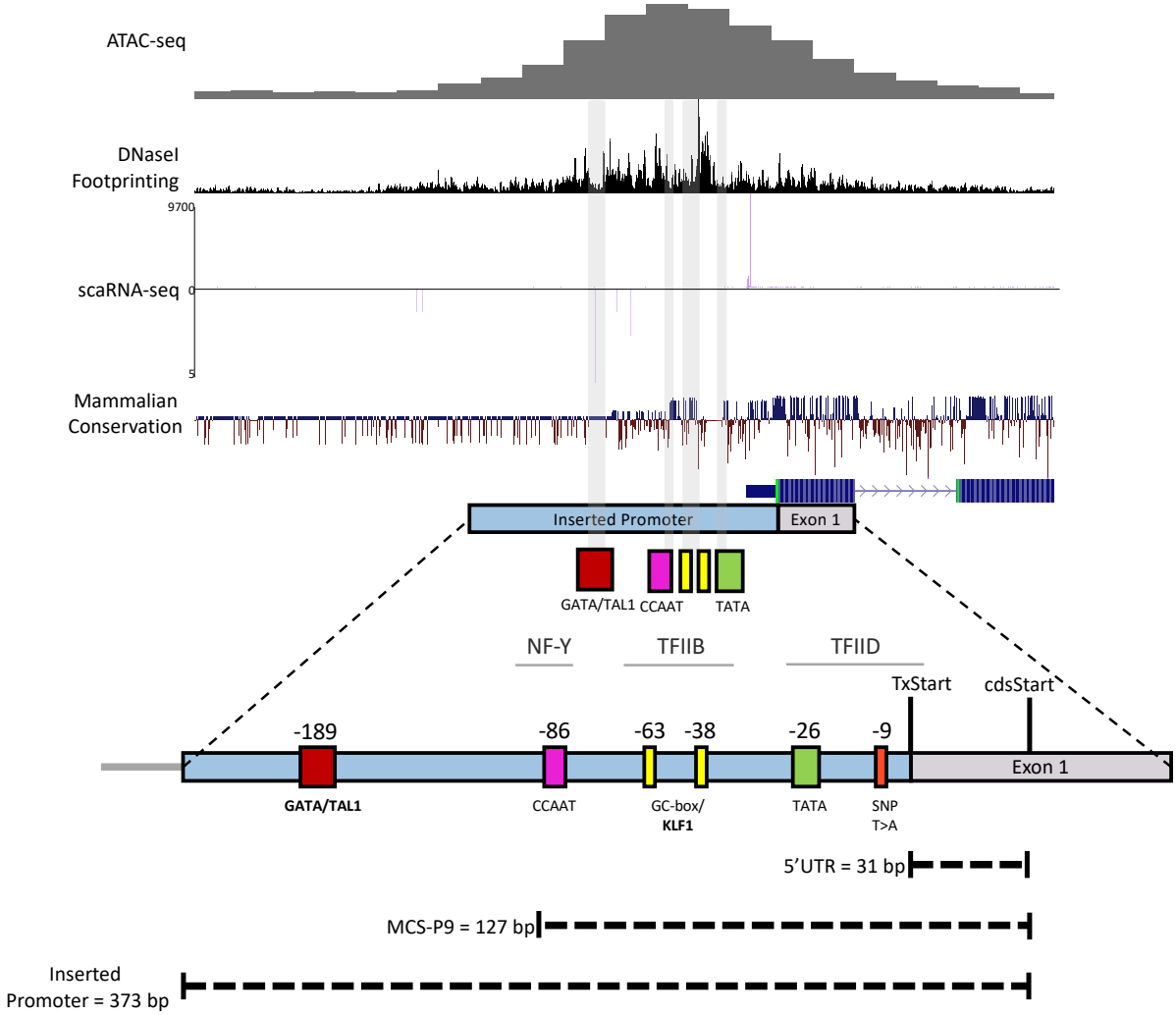

Figure 3.1: Defining the α-globin promoter and gene inserts.

ATAC-seq, DNaseI footprinting and scaRNA-seq tracks from primary mouse erythroid cells. Mammalian conservation available from the UCSC browser is also displayed. The region selected for insert designs is shown in blue (Inserted Promoter) and is aligned to scale against UCSC genome browser tracks with the highlighted locations of transcription factor binding sites as annotated (GATA/TAL1, CAAT, KLF1, TATA, NF-Y, TFIIB, TFIID). B) Schematic shows the positions of the highlighted transcription factor binding sites and the dashed lines specify the sequences corresponding to the 5'UTR bounded by the transcription start site (Txstart) and the coding start site (cdsStart), the multi-conserved sequence (MCS-P9) and the Inserted Promoter sequence.

Insert design – Gene body

The α-globin gene body inserts lack the native promoter sequence to abolish transcription initiation. These were included to control for any unexpected insulator activity present in the underlying sequence. The gene body was defined such that it starts at +37 bp from the TSS to exclude the promoter MCS, the 5'UTR and the 2 first codons of the coding sequence and ending at the annotated 3'UTR which includes a polyA addition signal (AATAAA) (**Figure 3.2b**).

The sequences were then combined for fragments containing both promoter and gene body (**Figure 3.2a, b**). Upon insertion of an ectopic copy of the α-globin gene into the landing site, we would need a strategy to discriminate amongst the resulting 3 nearly identical copies of the α-globin gene in the edited cells; the inserted α-globin sequence, the native *Hba-a1:Venus* and *Hba-a2* (**Figure 3.2a**). I designed an insert-specific-SNP within exon 3 (c.C312>T) to allow insert-specific reads to be identified (**Figure 3.2b**). *Hba-a1* and *Hba-a2* can also be distinguished from one another bioinformatically as there is a naturally occurring SNP in exon 3: c.C381 in *Hba-a1* and c.T381 in *Hba-a1* (**Figure 3.2b**). The SNPs are silent mutations and do not lead to any change in protein sequence.

We have previously reported that the orientation of a putative transcribing promoter impacted on it's on the expression of native target genes (Bozhilov *et al*., 2021). We, therefore, tested the designed Promoter-Gene inserts in both their native and reverse orientation (**Figure 3.2a**).

Custom genome reference

Due to the similarities in sequence between inserted α-globin fragments and endogenous genes, it was difficult to visualise NGS data (ATAC, ChIP and RNA-seq) against a standard genome reference. With a standard reference, reads would map to all copies of the gene, leading to many mapping artefacts. To help interpret NGS data, I therefore created a custom genome with only one copy of *Hba-a1* in its native position; allowing all α-globin reads to map to the same gene reference (*Hba-a1*) independently of their source. To help in aligning partly mapping reads, unmapped reads from the first round of alignments were trimmed and realigned. When interpreting the data presented later in this chapter, as there is no insert sequence in the reference genome at the landing site, there will be no signal at the site itself, but flanking regions will be informative in visualising the activity of the inserted sequence.

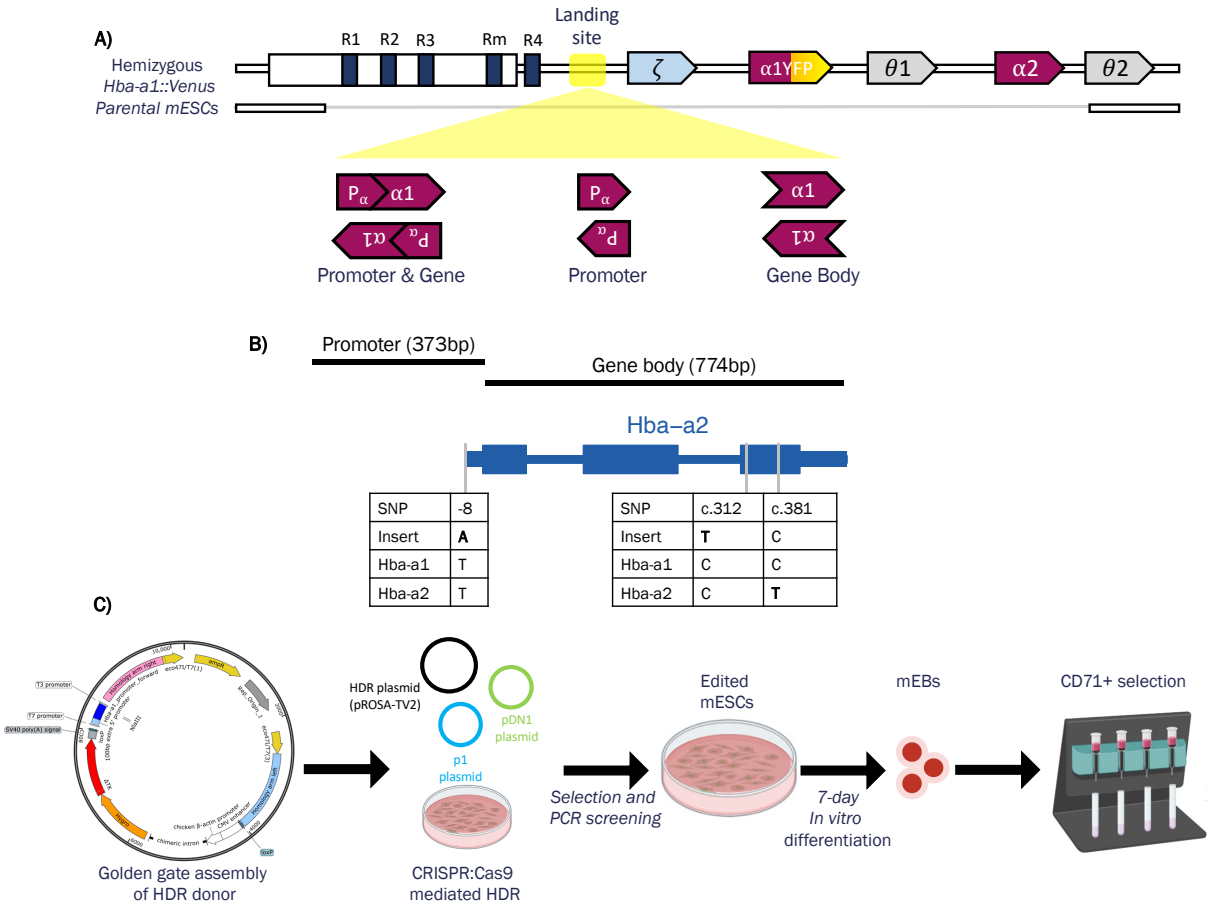

Figure 3.2: Landing pad design and experimental strategy.

A) *Schematic of the design of the boundary landing pad with the inserted promoter (Pα)-gene (α1) fragments used in this chapter; the insertion site is located between the endogenous globin genes (zeta, α1-YFP, α2) and the enhancer elements (R1, R2, R3, Rm, R4), ~2.6 kb from the R4 element.* ***B)*** *Detailed view of the inserted sequences (Promoter 373bp corresponding to Pα, Gene body 774bp corresponding to α1 used in the assay with SNP positions and α-globin variant specific bases displayed.* ***C)*** *Strategy used to engineer the mESC starting with the assembly of the homology driven repair (HDR) donor vector using the Golden Gate strategy followed by the mESC co-transfection of the CRISPR:Cas9 guide RNA plasmids (p1 and pDN1) and the HDR vector, then the screening and in vitro differentiation of the correctly targeted mESCs into erythroid cells marked by the erythroid-specific marker CD71.*

3.2.2 Testing the α-globin gene in the Boundary Assay

The landing site is located between the enhancer elements and the native globin genes, therefore an insulating sequence introduced at this site would intercept any enhancer: promoter interactions driven by cohesin-mediated loop extrusion and reduce the expression of the globin genes, as measured in this assay by the YFP reporter. To test the insulator activity of the α-globin genes, I inserted a copy of the α-globin gene with its promoter (as defined in section 3.2.1) in its native

orientation (PGF) into the boundary reporter landing site (**Figure 3.2**). To test for the boundary function, the correctly targeted mESCs were *in vitro* differentiated into embryoid bodies (EBs) containing erythroid cells identified using the erythroid-specific surface marker CD71 (transferrin receptor-1) (**Figure 3.2c**) (Dong *et al*., 2011). The α-YFP fluorescence measured in the CD71+/YFP+ cell population revealed a significant decrease in α-YFP compared to that in the parental reporter erythroid cells, indicating the ectopic gene is interfering with the expression of the downstream endogenous α-globin genes (**Figure 3.3**). The effect the PGF had on the α-YFP fluorescence was comparable to that observed when strong CTCF site (HS38) was tested in this assay (**Figure 3.3**). Therefore, the ectopic copy of the α-globin gene is exerting a strong insulator-like activity. To address the effect of gene orientation on boundary activity, I inserted the same sequence in the reverse orientation (PGR). The boundary effect of the reversed sequence (PGR) was comparable to PGF indicating gene orientation has no effect on its insulation activity (**Figure 3.3**).

The insulation observed following insertion of the α-globin gene (PGF, PGR) may be due to several factors including the transcription itself (transcription initiation, progressive transcription) or intrinsic properties of the gene-body sequence. To disentangle these factors, I tested the boundary effect of the α-globin promoter sequence only, in both orientations (Promoter Forward, PF, and Promoter Reverse, PR) and the α-globin gene sequence without its promoter in both orientations (Gene Forward, GF, and Gene Reverse, GR). Erythroid cells derived from clones with the insertion of the gene body alone (GF, GR) expressed high α-YFP fluorescence, comparable to that of the parental reporter cells, indicating the gene sequence alone does not have any insulator-like properties (**Figure 3.3**). Erythroid cells derived from clones with the insertion of the promoter sequence alone, in either orientation (PF, PR), expressed lower α-YFP compared to those derived from the inserted gene body sequence alone, indicative of insulator activity, albeit weaker than the promoter and gene inserted together (PGF, PGR) (**Figure 3.3**). As observed in the PGF and PGR models, the orientation of the promoter didn't lead to a significant difference in α-YFP. This trend was also apparent when plotting histograms of erythroid-specific YFP fluorescence (**Figure 3.4c**).

To summarise, the insulation effect was proportionate to the transcriptional potential of the inserted sequences; the greatest effect on αYF*P* reporter expression was observed with the insertion of the promoter and gene body (PGF, PGR), followed by the promoter only inserts (PF, PR), with the gene body alone (GF, GR) having negligible effect.

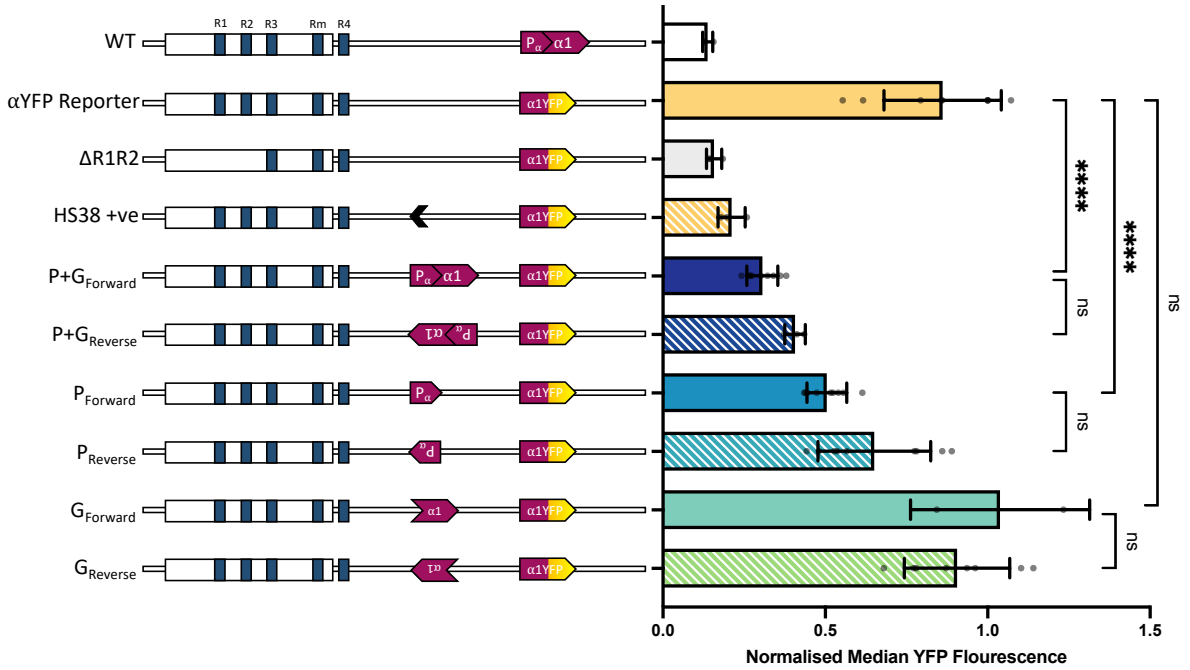

Figure 3.3: Insulation strength of α-globin gene fragments in a boundary reporter assay.

***Left**: Schematics of the models: WT = E14 unmodified mESCs; αYFP reporter = parental reporter with no insert (n=3); ΔR1R2 = mESC with deleted R1 and R2 enhancers and with αYFP; HS38 = positive control / mESC with a strong CTCF site HS-38 inserted at the landing pad; PGF = Promoter and Gene body in the native Forward orientation (n=3); PGR= Promoter and Gene body in the Reverse orientation (n=1); PF = Promoter in native Forward orientation (n=3); PR = Promoter in Reverse orientation (n=3); GF = Gene body only, in the Forward orientation (n=1); GR = Gene body only, in the Reverse orientation (n=3). **Right**: Median YFP fluorescence in the CD71+ fraction of cells measured by FACS and normalised to the parental αYFP reporter. Data acquired from clones (numbers indicated above) across 3 independent in vitro differentiations. Error bars display standard deviation. Statistical significance was calculated by performing an ANOVA with Tukey's multiple comparisons test ns = P > 0.05, **** = P ≤ 0.0001.*

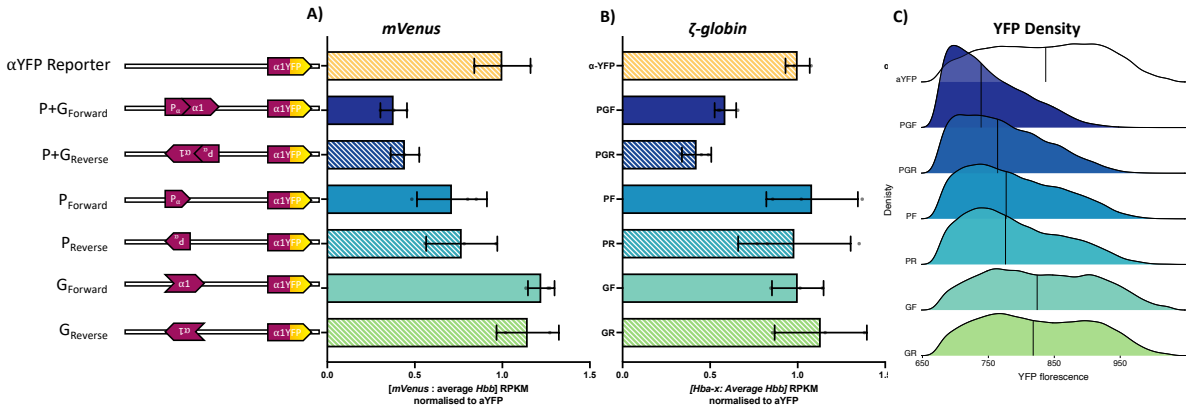

Figure 3.4: Insulation strength of α-globin gene fragments in a boundary reporter assay.

A) Virtual qPCR of mVenus expression, extracted from PolyA+ RNA-seq, normalised to RPKM then normalised to average expression of Beta-like globins. B) Virtual qPCR of Hba-x, analysed as in A. For A and B, n=3 independent differentiations of a representative clone per model, error bars display standard deviation. C) Density plots of YFP fluorescence in the CD71+ fraction of cells measured by FACS in representative clones of each model; median is marked by a central black line. Key as in ***Figure 3.3****.*

To check the integrity of the α-globin locus I performed ATAC-seq on *in vitro* derived CD71+ erythroid cells derived from the engineered mESC models (**Figure 3.5**). This confirmed that all α-globin cis-regulatory elements were accessible, and that chromatin opened at the landing site with the insertion of the fragments which displayed insulator activity (PGF, PGR, PF and PR). The ATAC-seq signal was highest at the PGF insert and absent at the landing site when only the gene body was inserted (GF and GR). This is consistent with the expected transcriptional output of each of the inserts and their insulator strengths.

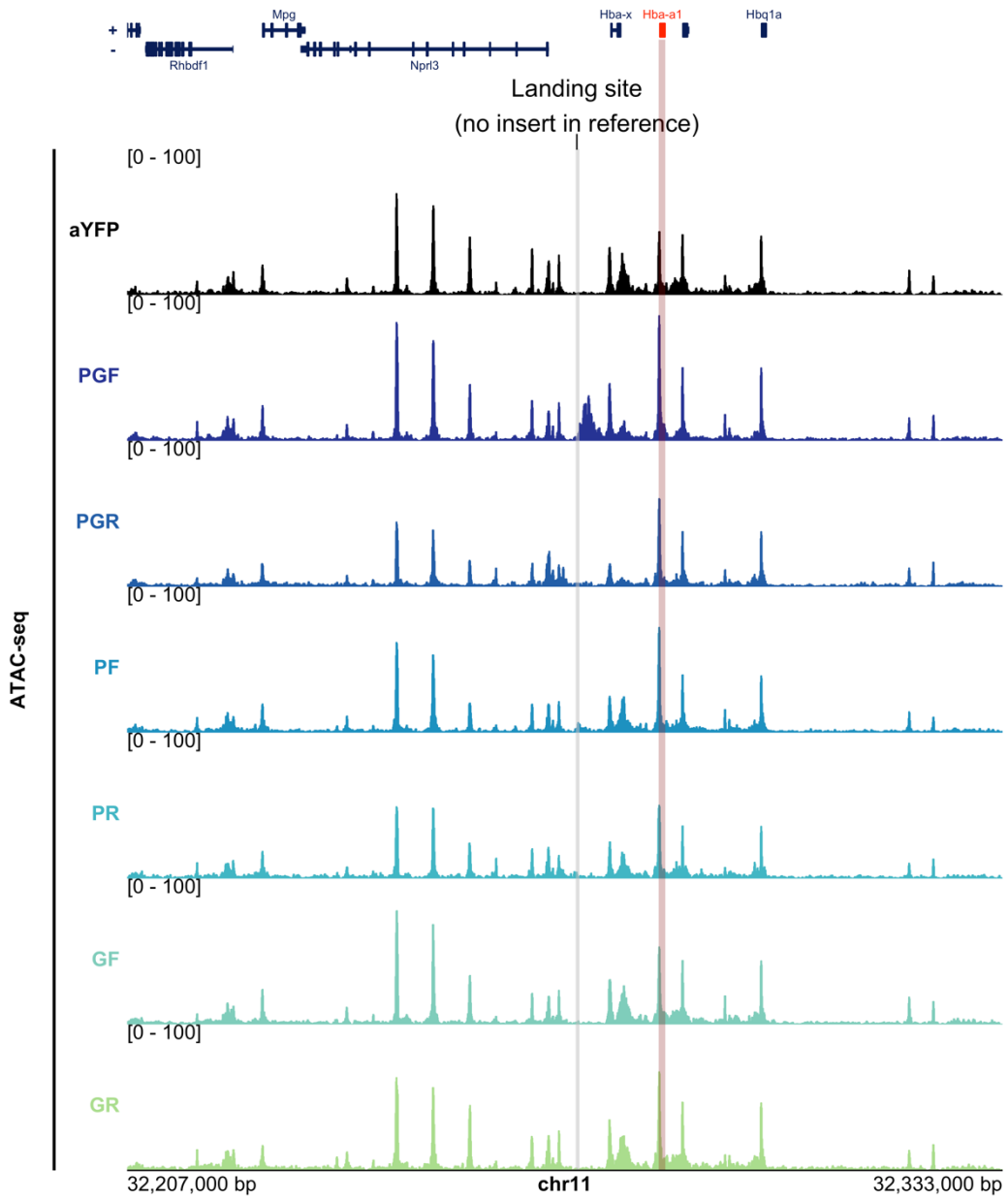

Figure 3.5: Accessibility across the α-globin locus in CD71+ derived from edited reporter cells.

RPKM normalised ATAC-seq performed on CD71+ erythroid cells derived from each of the indicated engineered models and aligned to a custom genome with only one copy of Hba-a1 in its native position, based on the mm10 reference. Note that the reference does not have insert sequences in the reference (see page 48 Custom genome reference). The landing site is highlighted in grey and the Hba-a1 reference in red. Bigwigs are visualised as mean of data from n=2 independent differentiations. Key as in ***Figure 3.3****.*

3.2.3 Transcription of fragments is correlated with insulation strength

Insulator strength appears to reflect the transcriptional potential of the inserts, with the PGF and PGR inserts displaying the strongest effect over promoter only inserts. To verify the transcriptional output from the inserted fragments, I performed PolyA selected RNA-seq from *in vitro* derived CD71+ cells from each of the models. RT-qPCR and digital droplet PCR (ddPCR) were initially considered as methods for assaying expression however due to the sequence similarity of the inserts and endogenous α-globin genes, designing probes with high enough specificity for each variant/SNP were not feasible, even with the incorporation of locked nucleic acids (LNAs) which massively increase the specificity of probes by increasing melting temperatures of probe:DNA hybrids (Ugozzoli *et al*., 2004). RNA-seq allows quantification of expression from each α-globin gene by bioinformatic discrimination of reads using SNPs (**Figure 3.2**). In addition, the ability to prepare PolyA selected libraries gives further information about the transcription happening at the inserts. The RNA-seq libraries were generated with a strand specific kit, therefore reads can be segregated according to the DNA strand they originate from.

Aligning stranded reads from PolyA+ RNA-seq displayed the expected mature RNA signals across the locus, such as high-level transcription of *hba-x* and *hba-a* genes from the forward strand (>8x10^6 RPKM) and *Rhbdf1* from the reverse strand (**Figure 3.6, 3.7**). At the insertion site, PolyA+ RNA-seq confirmed there was polyadenylated transcription originating from both the PGF, PGR inserts, and the PF insert, whilst the reverse promoter and gene body only inserts (PR, GF, and GR) had background level of transcription, similar to the equivalent position with no insert in the parental reporter. To determine if the α-YFP change in fluorescence was associated by a reduction in transcription, I performed a virtual qPCR, using Featurecounts and EdgeR to extract RPKM normalised reads mapping to *mVenus* and the β-like globins (*Hbb-bs, Hbb-bt, Hbb-bh1, Hbb-bh2, Hbb-y*) from PolyA+ RNA-seq (**Figure 3.4a**). Whilst the CD71+ population of cells derived from EB's is largely homogenous (Francis *et al*., 2022), some cell lines can drift in differentiation potential following targeting, therefore normalising against the β-like globin genes helps to correct for differences arising from heterogeneity in the *in vitro* differentiation across cell lines. Virtual qPCR confirmed *mVenus* transcription correlated well with α-YFP fluorescence in all models (**Figure 3.4a**). In addition, as the inserts are located between the enhancers and ζ-globin gene, I expected that ζ-globin could also be affected by enhancer-insulation by the inserts; virtual qPCR confirmed that ζ-globin transcription was reduced in both PGF and PGR models, confirming that the effect was not limited to only the α-globin genes (**Figure 3.4b**).

To better quantify the transcription from the promoter-only inserts, I also performed PolyA- RNA-seq to capture any non-polyadenylated transcripts that could be initiated from the promoter-only inserts. As a check, PolyA- RNA-seq confirmed the presence of non-polyadenylated RNAs across the *Nprl3* gene body, consistent with the presence of eRNA as previously reported (Blayney *et al*., 2022, in press; Kowalczyk *et al*., 2012) (**Figure 3.8, 3.9**). When examining the landing site, I observed transcription from all inserts containing a promoter (PGF, PGR, PF, and PR) and only background level from the gene body only inserts (GF, GR) (**Figure 3.8, 3.9**). Transcription into the flanking regions is congruent with the expected direction of transcription from the inserts; such that natively orientated inserts (Forward, PGF, PF) have transcription from the forward DNA strand travelling downstream (**Figure 3.6, 3.7**) and the reverse inserts (PGR, PR) transcribing from the reverse strand (**Figure 3.8, 3.9**).

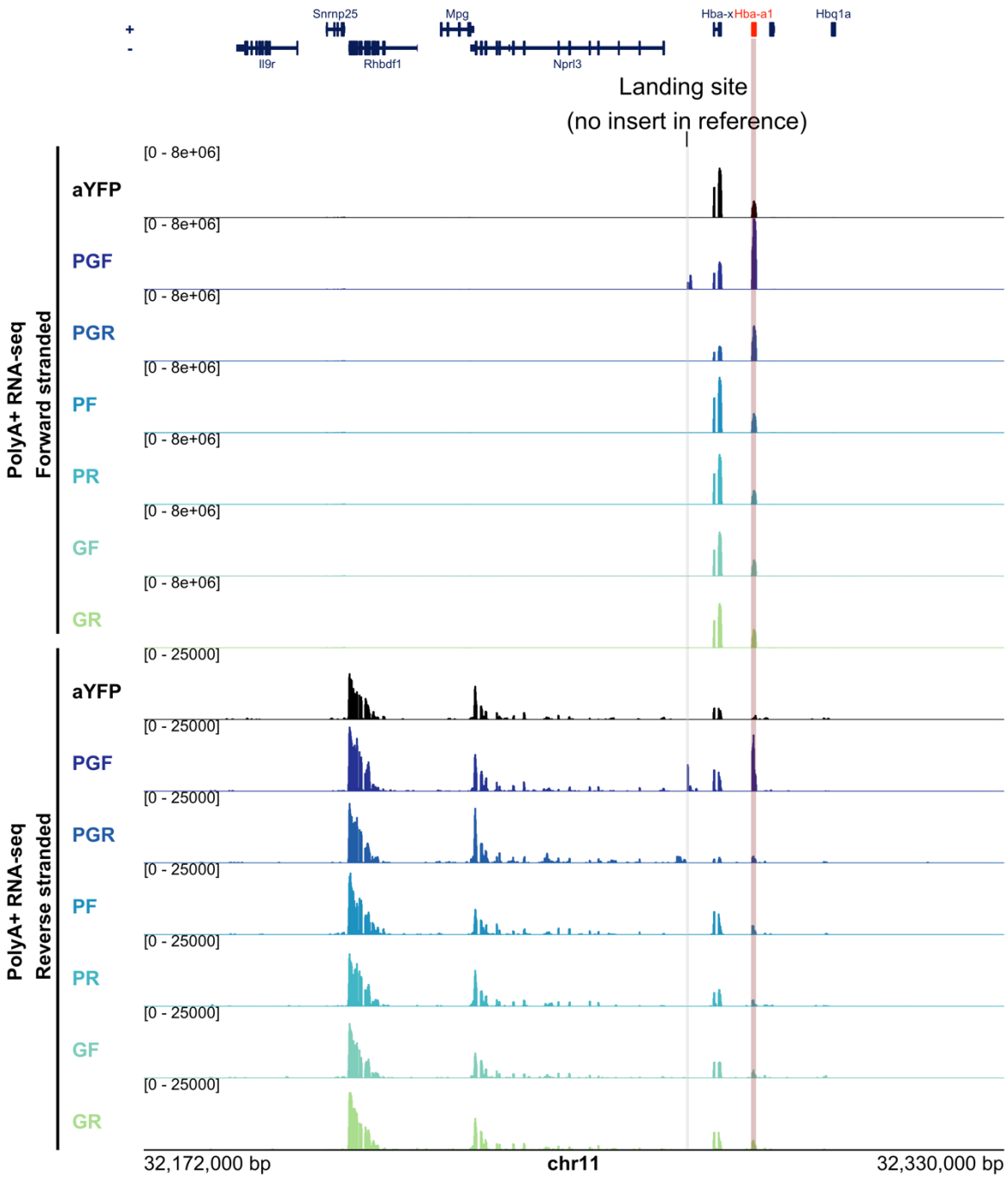

Figure 3.6: PolyA+ RNA-seq across the α-globin locus in CD71+ erythroid cells derived from edited reporter cells.

RPKM normalised PolyA+ RNA-seq derived from CD71+ erythroid cells from each of the indicated models, aligned to a custom genome with only one copy of Hba-a1 in its native position, based on the mm10 reference (see page 48 Custom genome reference). The landing site is highlighted in grey and the Hba-a1 reference in red. The top panel displays reads mapping to the forward strand whilst the bottom panel shows reads mapping to the reverse strand. Bigwigs are visualised as mean of data from n=3 independent differentiations of one representative clone per genotype. Key as in ***Figure 3.3****.*

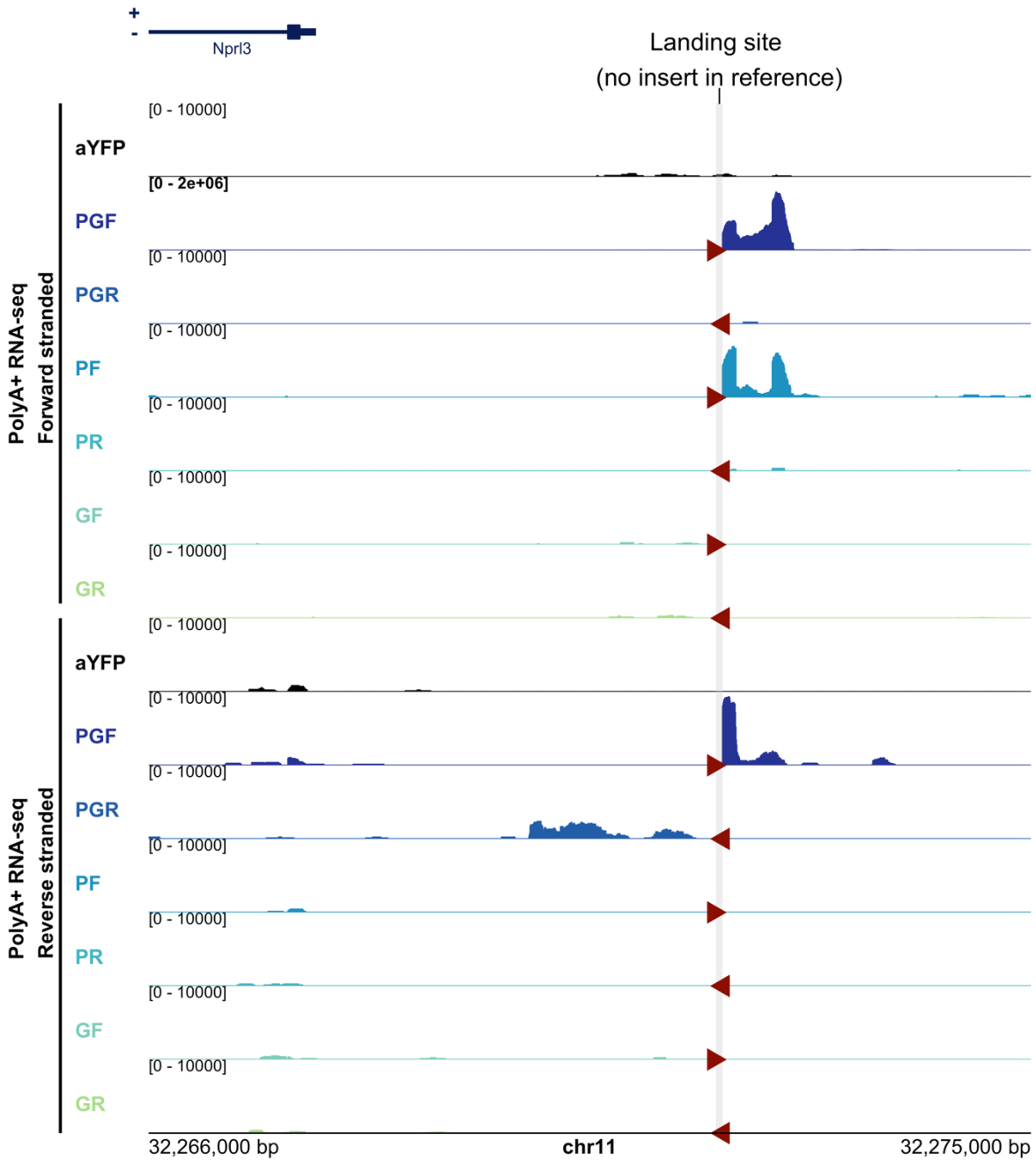

Figure 3.7: PolyA+ RNA-seq across a zoomed in region at the landing site at α-globin locus in CD71+ erythroid cells derived from edited reporter cells.

*As in **Figure 3.6**, focusing on region around the landing site (custom chr11: 32266000-32275000) (see page 48 Custom genome reference). The landing site is highlighted in grey, and the direction of the insert indicated by a red arrow. Note, PGF is scaled differently to the rest of the models, due to its high levels of transcription. Key as in **Figure 3.3**.*

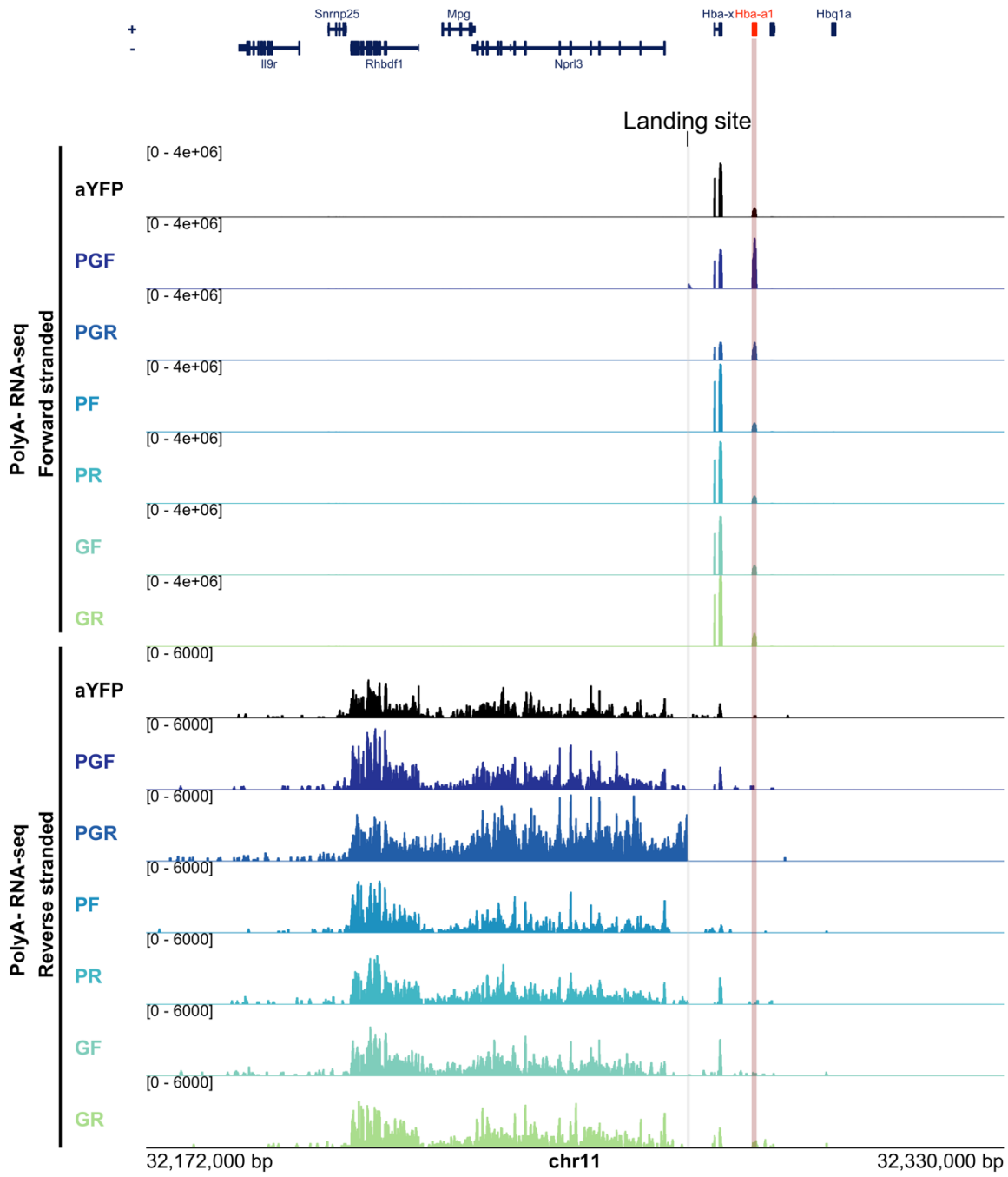

Figure 3.8: PolyA- RNA-seq across the α-globin locus in CD71+ erythroid cells derived from edited reporter cells. *RPKM normalised PolyA- RNA-seq derived from CD71+ erythroid cells from each of the indicated models, aligned to a custom genome with only one copy of Hba-a1 in its native position, based on the mm10 reference (see page 48 Custom genome reference). The landing site is highlighted in grey and the Hba-a1 reference in red. The top panel displays reads mapping to the forward strand whilst the bottom panel shows reads mapping to the reverse strand. Bigwigs are visualised as mean of data from n=3 independent differentiations of a representative clone per genotype. Key as in* ***Figure 3.3****.*

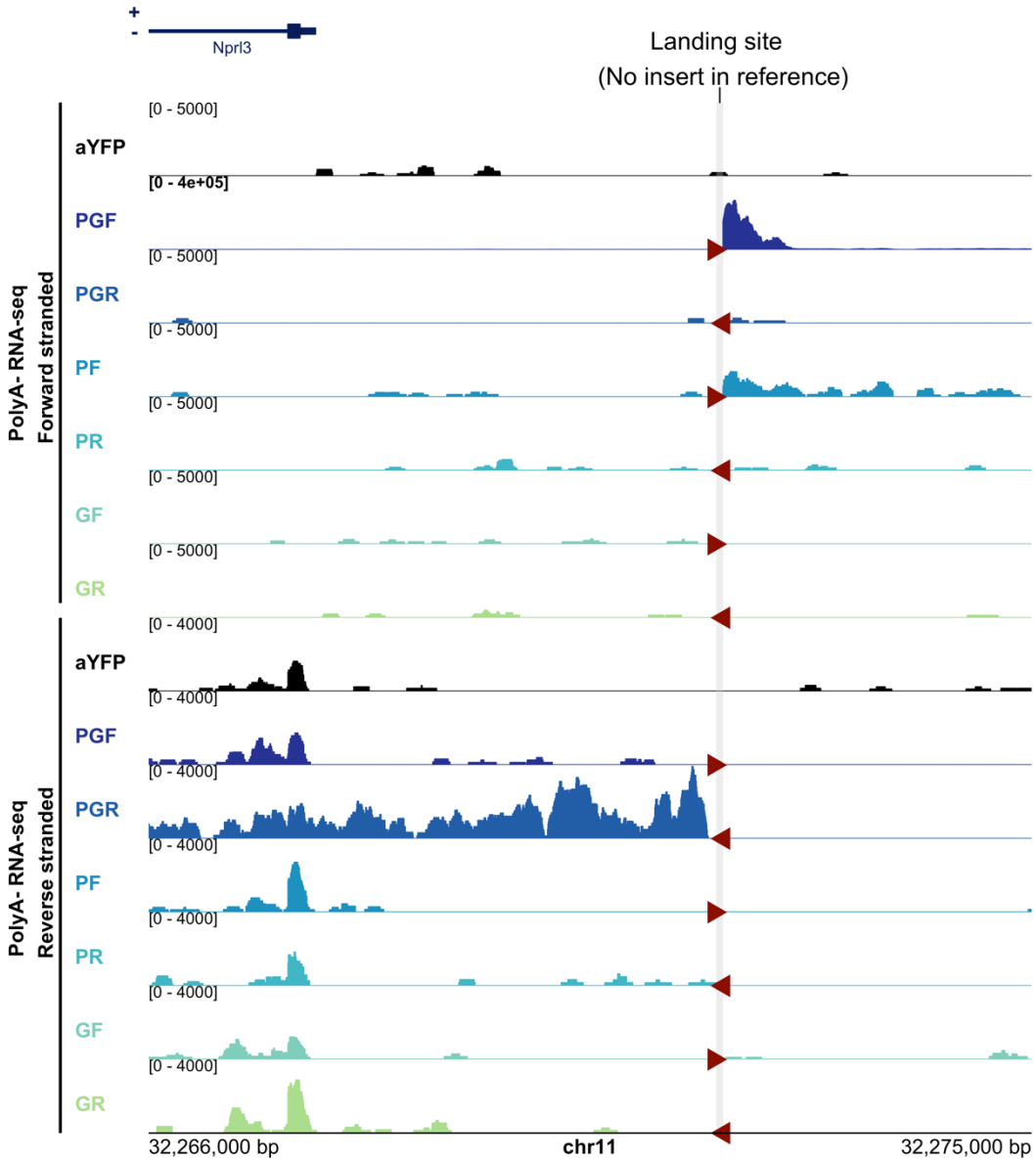

Figure 3.9: Zoomed region PolyA- RNA-seq across the α-globin locus in CD71+ derived from edited reporter cells.

*As in **Figure 3.8**, focusing on region around the landing site (custom chr11: 32266000-32275000) (see page 48 Custom genome reference). The landing site is highlighted in grey, and the direction of the insert indicated by a red arrow. Note, PGF is scaled differently to the rest of the models, due to its high levels of transcription. Key as in **Figure 3.3**.*

To better quantify the expression from the inserts I developed an analysis pipeline to count the number of reads from three sequences: the inserted copy, *Hba-a1::Venus* and *Hba-a2,* using the SNPs unique to each variant. In this case, analysis was performed on PolyA+ RNA-seq. Quantification of *Hba-a1/2* reads were consistent with the pattern of observed α-YFP fluorescence measured by FACS (**Figure 3.10a and Figure 3.3**). Using the exonic SNP to quantify expression from the inserted α-globin genes (PGF and PGR) revealed the inserted sequences were more highly expressed than the endogenous genes (**Figure 3.10a**). This could be due to the inserted sequences being closer to the enhancers than in their native position (~2.6 kb compared to ~14 kb); it is well-documented that enhancer activity is distance dependant and stronger when the enhancer is brought closer to its target promoter (Rinzema *et al*., 2022; Zuin *et al*., 2022). Interestingly, transcription of the natively orientated gene (PGF) was significantly higher than that of the sequence in reverse orientation (PGR) (**Figure 3.10a**). This observation will be the topic of a later chapter and will be briefly covered in the rest of this chapter.

Where no reads containing the promoter proximal SNP (-8) were captured, I quantified the transcription of the regions flanking the insertion sites using Featurecounts on the PolyA- RNA-seq data, normalised using RPKM. When applied to the inserts lacking the exonic SNP, i.e. PF, PR, GF and GR, the analysis followed by bigwig visualisation show transcription in all inserts with a promoter consistent with the direction of the insert; such that the forward inserts transcribed into the right flank, and the reverse inserts into the left flank (**Figure 3.10b**). This quantification also highlights the correlation between transcription level and the insulator strength of the inserts (**Figure 3.10b and Figure 3.3**).

A reduction in endogenous *Hba-a1/2* gene expression in this boundary assay is considered as a measure of insulator strength. However, the boundary effect observed following insertion of the α-globin gene can be interpreted by two mechanistic models. The first, is promoter competition whereby the ectopic promoter (that of the insert) may compete with the native α-globin promoters for the enhancers. Alternatively, the reduction on native gene expression could be caused by a blockade to the linear progression of loop extrusion machinery by the transcriptional machinery of the insert and hence form a partial physical boundary which hinders contact between the enhancers and their cognate promoters. These models were addressed previously when analysing *de novo* transcription from a pathogenic SNV that created a new promoter which intercepted the normal alpha globin enhancer-promoter interactions causing alpha thalassemia; transposition of this novel SNV promoter outside of the sub-TAD partly restored expression of the native α-globin genes, disfavouring

a competition model between the SNV and the native promoters for enhancer activity (Bozhilov *et al*., 2021). Therefore, to get a better understanding of the mechanisms underlying my new observations I performed further molecular characterization.

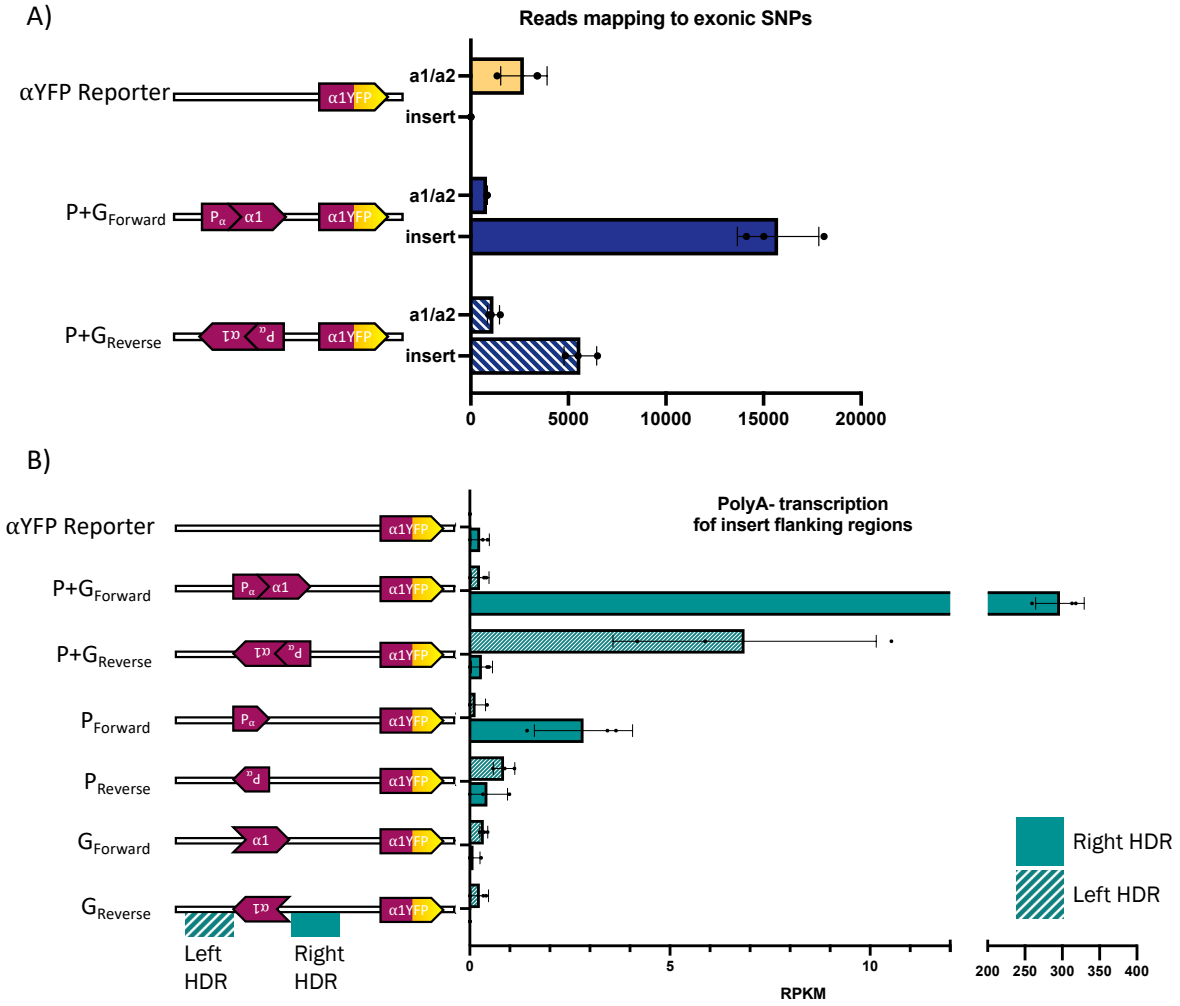

Figure 3.10: Quantification of transcription of inserted fragments.

A) *Quantification of variant specific expression from PolyA + RNA-seq; reads were mapped to a custom genome with one copy of Hba-a1, reads from the inserted fragments (insert) or endogenous (a1/a2) genes were distinguished by base identity at an exonic SNP.* ***B)*** *Quantification of transcription from insert flanking regions from PolyA- RNA-seq; reads mapping to either the left side (Left HDR, hashed) or the right side (Right HDR, solid) were quantified by Featurecounts and normalised with RPKM. Unidirectional transcription is observed, such that forward fragments transcribe into the right side, whilst reverse fragments transcribe to the left. The data represent n=3 independent differentiations with one clone per model.*

3.2.4 Insertion of transcribed fragments show insulator-like accumulation of cohesin

To determine if there were changes in the distribution of cohesin that would be consistent with the inserts acting as a block to the linear progression of the loop extrusion machinery, I performed cohesin ChIP-seq (Rad21). I chose to assay the inserts with the greatest insulator-like effect; PGF, PGR and PF. Cohesin distribution was largely similar across the locus in all the tested clones, with cohesin peaks detected at the enhancer elements, CTCF sites and active promoters (**Figure 3.11**); however, when looking at the changes surrounding the insertion site, there was an increase in cohesin accumulation in all inserts when compared to the parental reporter cell line (**Figure 3.11**). The greatest increase in cohesin is seen at the PGF insert, with a peak observed upstream of the insert corresponding to the direction of transcription (**Figure 3.11**). Similarly, cohesin accumulation is higher downstream of the PGR insert, also coinciding with the direction of transcription of the insert. Upon insertion of the PF insert, there is a slight increase in cohesin accumulation, less than that of the PGF insert, and again in the direction of transcription. To summarise; cohesin accumulates in the direction of transcription of the inserts and the level of cohesin accumulation correlates with the transcriptional output and insulator strength of each insert.

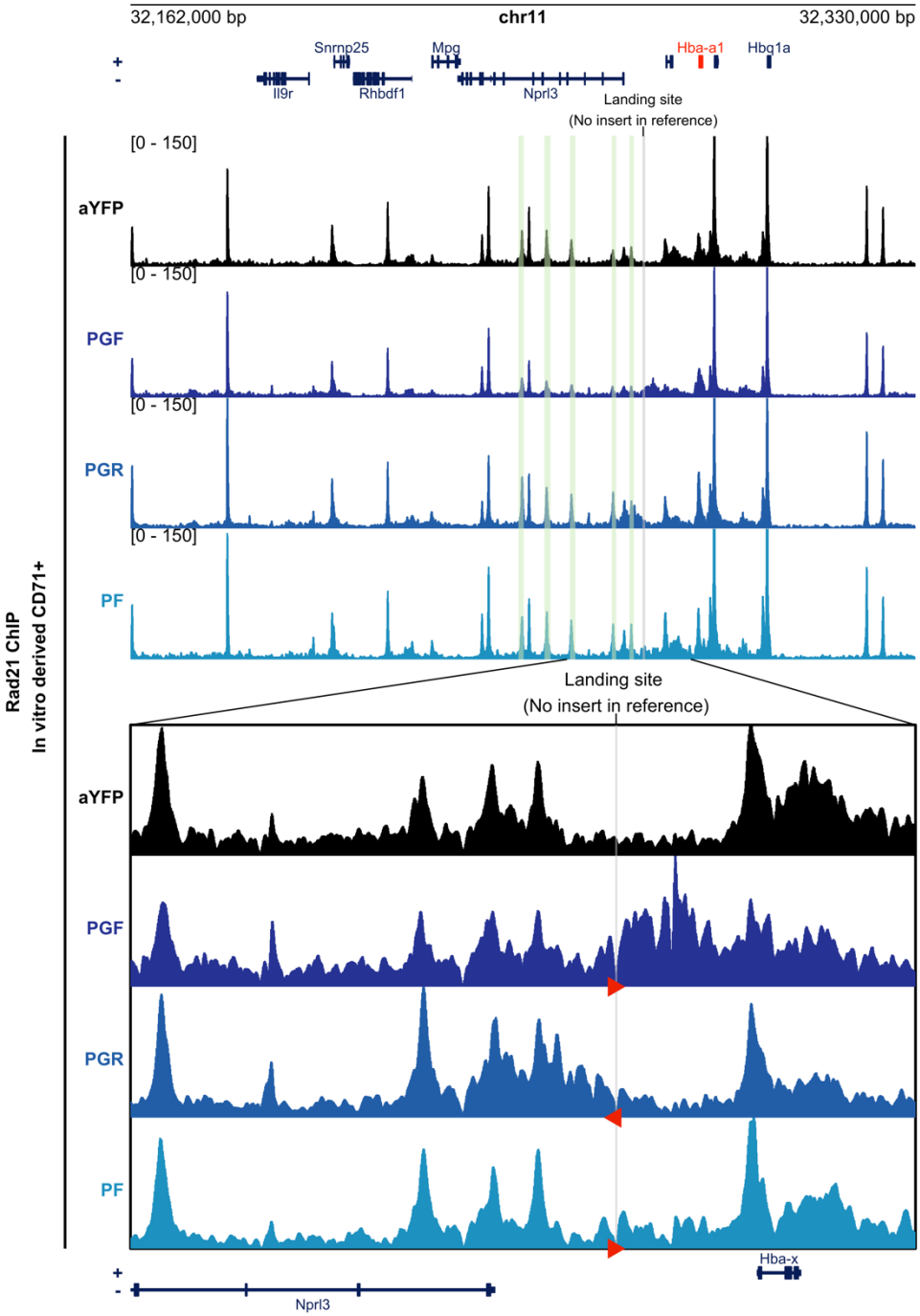

Figure 3.11: Cohesin accumulation across the α-globin locus in CD71+ erythroid cells derived from edited reporter cells.

RPKM normalised Rad21 ChIP-seq derived from CD71+ material from each of the indicated models aligned to a custom genome with only one copy of Hba-a1 in its native position, based on the mm10 reference (see page 42 Custom genome reference). Top panel shows the locus (mm10: 32162000-32330000) whilst the bottom panel zooms in on the insertion site at the locus (custom mm10: 32255300-32282000). The landing site is highlighted in grey and the direction of the insert indicated by a red arrow. Bigwigs are visualised as mean of data from n=2 independent differentiations of a representative clone per genotype. Key as in ***Figure 3.3****.*

3.2.5 Cohesin accumulates at active enhancer elements and transcription start sites genome-wide

Whilst cohesin accumulation at transcription start sites has been previously reported (Busslinger *et al*., 2017a; Hua *et al*., 2021), I wanted to check if this could be observed in the genome wide Rad21-ChIP from *in vitro* derived CD71+ erythroid cells and to determine if this accumulation was limited to the α-globin locus. Using Rad21 ChIP-seq I generated from non-reporter WT cells, I performed genome-wide analysis on the localisation of cohesin relative to regulatory elements. There was a significant accumulation of cohesin at TSS of genes (**Figure 3.12a**). Using total RNA-seq from WT *in vitro* derived CD71+ erythroid cells (Rosa Stopler, unpublished), I categorised TSSs according to whether they were inactive or active (RPKM > 1 in 2 replicates). Active TSSs (n=9,219) accumulated more cohesin than inactive TSSs (n=17,009) (**Figure 3.12b**). In addition, over 70% of the top 10,000 Rad21-accumulating TSS were active (**Figure 3.12c**). This together supports that active transcription is associated with cohesin accumulation. As displayed in the α-globin locus, cohesin accumulates at the composite elements of the super enhancer (**Figure 3.12d**); this may be a result of loading of cohesin at enhancer elements (Hanssen *et al*., 2017; Hua *et al*., 2021; Rinzema *et al*., 2022; Stolper *et al*., 2023). To confirm this across other loci, I collected the 1963 erythroid-specific enhancer elements previously classified by Hay *et al*., (2016), and plotted their cohesin levels. Cohesin is indeed enriched at these active enhancer elements. This data displays how cohesin does not exclusively accumulate at CTCF sites but is also found at active cis-regulatory elements.

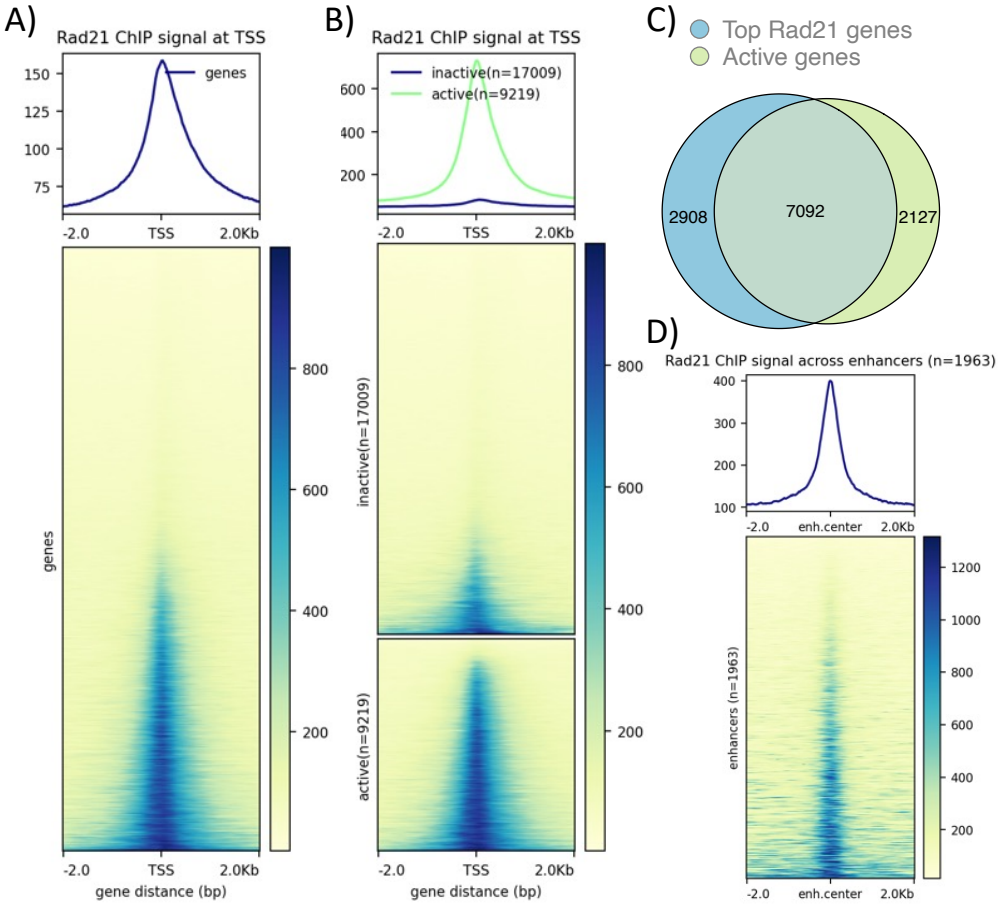

Figure 3.12: Cohesin accumulation at cis-regulatory elements genome-wide.

*Heatmaps and summary profiles displaying Rad21 ChIP-seq signal strengths at transcription start sites (TSS) in CD71+ erythroid cells derived from in vitro differentiation of WT mESC, for **A)** all annotated genes (n= 26228), **B)** inactive (n= 17009) and active genes (n=9219) and **D)** enhancer elements (n=1963). **C)** Euler plot displaying the overlap of the top 10,000 Rad21 accumulating genes and active genes.*

3.2.6 Preliminary Tiled-C Capture data indicate there may be an increase in interactions between the engineered landing site and enhancers

If the inserted elements were indeed acting as insulator elements, intercepting the enhancers, and preventing correct interaction with the endogenous target genes, I would expect a gain of interactions between the enhancers and insertion site, possibly resulting in reduced interactions with the native promoters. To determine if there was such a change in interactions, I performed Tiled-C across the α-globin locus in CD71+ erythroid cells derived from PGF, PGR and PF cell lines, where insulator function was found. Tiled-C is similar in principle to NG-CaptureC but utilises a larger pool of biotinylated oligonucleotides which cover a 3Mb region of the α-globin locus to enrich for contacts within the region (Oudelaar *et al*., 2020). I faced technical challenges when attempting the assay on the *in vitro*

derived CD71+ erythroid cells; electrophoresis of input DNA suggested there was non-specific degradation of input DNA which compromises the downstream steps (digestion and religation) and the quality of the 3C library (**Figure 2.11**). Also, because of a low yield of DNA after the first capture, I opted to perform single capture rather than the conventional double capture to avoid reducing the complexity of the final multiplexed library pool. With these limitations in mind, I consider the data I am presenting here to be preliminary and optimisation to improve the resolution of the interaction maps is required to confirm the observations. When analysing the Tiled-C data as a virtual capture from the R2 enhancer viewpoint, an increase in interactions can be observed between the enhancer and the landing site in the PGF and PF models relative to the parental reporter (**Figure 3.13**).

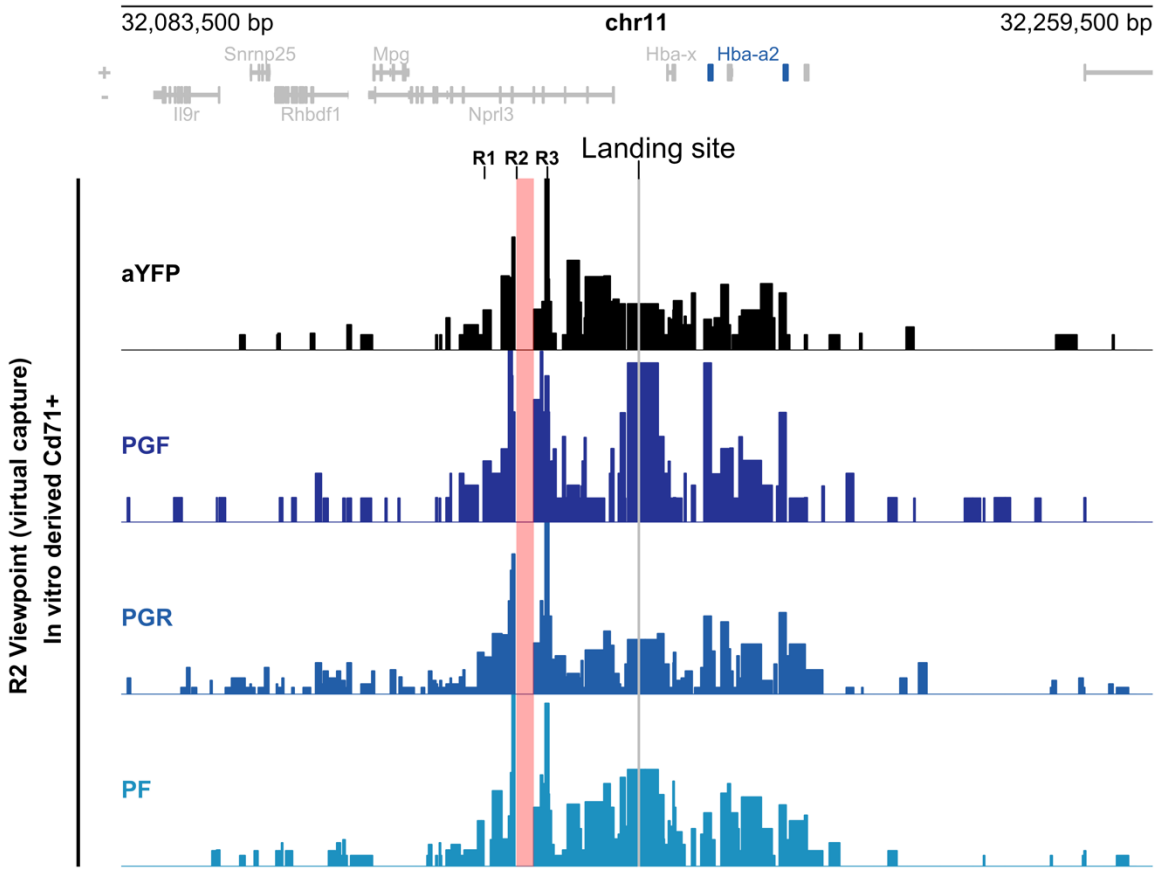

Figure 3.13: Preliminary chromatin conformation capture from CD71+ erythroid cells derived from edited reporter cells.

Virtual capture from Tiled-C data, using R2 as the viewpoint. The exclusion zone around the viewpoint is shaded in red. Profiles represent the mean number of normalised unique interactions per restriction fragment from n=2 independent differentiations. 'αYFP' = parental reporter with no insert; PGF = Promoter and gene body in the native orientation; PGR= Promoter and gene body in the reverse orientation; PF = Promoter in native orientation. Landing site is highlighted in grey.

3.3 Discussion

When placed between enhancer elements and endogenous genes, transcribing fragments of the α-globin gene display insulator behaviour, leading to a reduction in native gene expression. The strength of the insulation is well correlated with the transcriptional output and accumulation of cohesin at the newly inserted site. Preliminary data also suggests an increase in enhancer interactions with the inserted fragments. Together this data supports the hypothesis that transcribed sequences, such as the newly inserted α-globin genes, can act as insulators.

There are two potential models. The first involves mutually exclusive promoter competition, such that the inserts outcompete the endogenous genes for enhancer contact. The second is that actively transcribing units may act as partial barriers to directional tracking of cohesin-mediated loop extrusion. These two models are difficult to untangle and arguably both are supported by different aspects of the data presented.

Due to the increased proximity to the enhancer elements, the inserted α-globin gene (PGF & PGR) is far more activated than the native genes. This agrees with a recent study which confirmed that enhancer action declines with increased genomic distance to the target gene (Rinzema *et al*., 2022; Zuin *et al*., 2022). Notably, enhancer proximity does not always determine transcriptional output, as exemplified by the enhancer proximal embryonic *Hba-x* gene, which is silent in definitive erythropoiesis despite its proximity to the active superenhancer (King *et al*., 2021; Peschle *et al*., 1985). The huge discrepancy in transcription between ectopic and native genes could be interpreted as the insert 'outcompeting' the native genes for enhancer contact. Other instances of promoter competition *in cis* have been described where an active promoter located between an enhancer and another, more distal promoter causes reduced activity of the distal promoter (Cho *et al*., 2018; De Gobbi, 2006). The latter was observed in the α-globin locus; an SNV gave rise to a *de novo* promoter located between the enhancers and α-globin genes, leading to significant disruption of α-globin expression and development of α-thalassemia, similar to the insultation effect described in this chapter (Bozhilov *et al*., 2021; De Gobbi, 2006). This was associated with reduced contact between the enhancers and α-globin genes and a respective gain of interactions between the enhancers and SNV. Interestingly, α-globin expression was rescued when the SNV was placed just upstream of the enhancer elements, and yet still within the sub-TAD. Given that the *de novo* promoter had equal accessibility to the enhancers, in both upstream and downstream positions, the observation that only

the downstream position caused an insulation effect indicates that there is directionality to the underlying mechanism.

This directionality could be described by the second model in which actively transcribing units may act as obstacles to cohesin-mediated loop extrusion. When considering this model in relation to the data presented here, the key to disentangling the models is the observed redistribution of cohesin. As described previously, cohesin is the driver of domain formation (Rao *et al*., 2017). In the α-globin locus, cohesin is likely loaded at the enhancer elements and extrudes DNA; the cohesin translocating to the 5' end of the locus is halted by HS-38/39 CTCF sites (Barrington *et al*., 2019; Stolper *et al*., 2023; Vian *et al*., 2018), whilst 3' translocation encounters the inserted fragments. As displayed in cohesin ChIP-seq, the inserted transcribing fragments accumulate cohesin. This is not limited to the inserts at the modified α-globin locus but can also be observed genome-wide at active TSSs and enhancer elements, in line with previous literature (Busslinger *et al*., 2017; Hua *et al*., 2021). Admittedly, the genome wide cohesin profiles at TSS presented in this chapter would be reinforced by sub-categorising TSS, to those with and without CTCF binding and further determining transcriptional activity by incorporating active chromatin signatures, such as H3K4me3 and ATAC. Whilst an interesting observation that cohesin is present at *cis*-regulatory elements, it is important to consider that accumulation of cohesin on active enhancers and promoters may also be due to active loading of cohesin.

Transcriptional machinery is not limited to the RNA polymerase II initiation complex. Other apparatus such as capping, splicing and termination machinery is also loaded during processing and maturation of mRNA. Additionally, in the presence of the enhancers, multiprotein Mediator complexes are also loaded at the promoter. These bulky, macromolecular transcription complexes assembled at the promoter may form an obstacle to cohesin machinery and cause stalling or disruption of extrusion, such that a loop more frequently terminates at the site of transcription. This is potentially how the 'promoter with gene' (PGF, PFR) showed such strong insultation strength in the reporter assay, whilst 'promoter-only' inserts (PF, PR) were marginally weaker in strength, as transcripts from these inserts do not have the downstream motifs necessary for mRNA processing but can initiate transcription.

This model is corroborated by other observations, for example, whilst CTCF sites form the majority of TAD boundaries in mammalian cells, actively transcribing genes are also enriched at TAD boundaries (Bonev *et al*., 2017; Dixon *et al*., 2012). Many TAD boundaries in *D. melanogaster* are likewise associated with active genes (Ramírez *et al*., 2018). Evidence that the minichromosome maintenance (MCM) complex can affect cohesin processivity *in vitro* supports that cohesin is not solely halted by

CTCF (Dequeker *et al*., 2022). Cohesin processivity was likewise disrupted by R-loop structures which are associated with, but not exclusive to, transcription (H. Zhang *et al*., 2023). RNAPII was also reported to act as a moving barrier to loop extrusion in bacterial cells (Brandão *et al*., 2019) and induced transcription was associated with the formation of domain boundaries at activated genes in *S. cerevisiae* (Jeppsson *et al*., 2022). A similar conclusion was drawn from analysis of CTCF/Wapl double knock-out experiments in mammalian cells, which led to the emergence of cohesin-dependent contacts at active genes (Banigan *et al*., 2023). Finally, Micro-C analysis following degron-mediated depletion of RNAPII revealed a loss of enhancer-promoter loops, further supporting a connection between transcription and the 3D genome (S. Zhang *et al*., 2023).

Whilst these correlations indicate a relationship between insulator activity and transcription, this data falls short of verifying whether an actively transcribing gene can functionally insulate. However, there are some data that more closely address this question. During mouse neural differentiation, TAD boundaries were formed concomitant with activation of NPC-specific genes, even in the absence of CTCF; however, domain formation was not observed when the genes were activated upon CRISPRa (dCas9:VP64-p65-Hsf1) in mESCs (Konermann *et al*., 2015). In a different model, CRISPRa-mediated activation of gene expression did lead to chromatin remodelling at the induced gene (Chahar *et al*., 2022); however, this observation is caveated by the presence of dCas9:VP64-p65-Rta on the DNA which itself may also contribute to disruption of cohesin translocation (H. Zhang *et al*., 2023).

The two mechanisms discussed here to describe the insultation effect of the transcribing fragments may also not be mutually exclusive. Due to their proximity to the enhancers, the inserts may indeed 'outcompete' the endogenous genes for enhancer contact, Mediator binding and transcription factors. Simultaneously, transcription and all the proteins involved may itself act as a physical barrier for the extruding cohesin, and thus further preventing the interaction between enhancers and the endogenous promoters. Further molecular characterisation, such as PolII or Mediator ChIPs may be helpful in interpreting the observation.

The data presented in this chapter provides evidence that transcribing genes can act as insulators, and in the context of the α-globin locus, likely normally form the domain boundary, in the absence of the 3' CTCF sites.

3.4 Acknowledgments for Chapter 3

Dr Felice Tsang was instrumental in this project, having provided me with the fully validated parental αYFP cell line, training me on optimised protocols for FACs and differentiations and for consulting on design of targeting constructs. The CTCF boundary assay concept and reagents were initiated by Dr Mira Kassouf, Dr Hele Francis, Dr Rosa Stopler, Muhammad Hanifi and developed by Dr Felice Tsang.

Chapter 4: Inversion of a single copy of the α-globin gene relative to the super enhancer

4.1 Introduction

When considering enhancer-promoter interactions, one overlooked factor is the relative orientation of elements. This aspect is likely under-represented because individual enhancer elements are widely reported to act in an orientation-independent manner (Banerji *et al*., 1983; Mercola *et al*., 1983). This is certainly the case in reporter assays and in some genomic contexts; however, this appears not to hold for all super-enhancers. We have recently reported on the analysis of an inversion of the entire α-globin super-enhancer, in which the whole region was inverted with proximal and proximal CTCF sites removed (Preprint Kassouf *et al.*, 2022). Genes normally lying upstream of the super-enhancer in the direction of the re-oriented enhancer cluster increased in expression concomitantly with gained interactions between their promoters and α-globin super-enhancer elements whilst the α-globin gene expression decreased along with reduced interaction with their super-enhancer (**Figure 4.1**). In contrast to orientation-independent paradigm of individual enhancer elements, this data shows that this well characterised super-enhancer acts in an orientation-dependant fashion. This may be due to how each element within the super-enhancer contributes the function of the cluster differently; R1 and R2 are 'activating elements', which contribute the majority of enhancing potential, whilst R3, Rm and R4 are considered to be 'facilitators' and while having no intrinsic enhancer activity are required for full activity of the super enhancer (Blayney *et al.*, 2022, in publication). The facilitator elements are most proximal to the target globin genes, and it has been shown that their position within the cluster has an impact on overall transcriptional output. This, therefore, suggests that facilitators may have a role in directing and orientating the activating potential of the super-enhancer, perhaps by creating an optimum assembly and structure of the Mediator complex, PIC and RNAPII. Upon inversion, these facilitators now lie proximal to the 5' genes, thereby directing activity towards the genes that normally lie upstream of the α-globin and contributing to the specificity of enhancer/promoter interaction. Together this shows that, in the case of the α-globin locus, there is inherent sequence directionality to the super-enhancer function; raising the question of whether the orientation of genes relative to the super-enhancer may also affect gene expression.

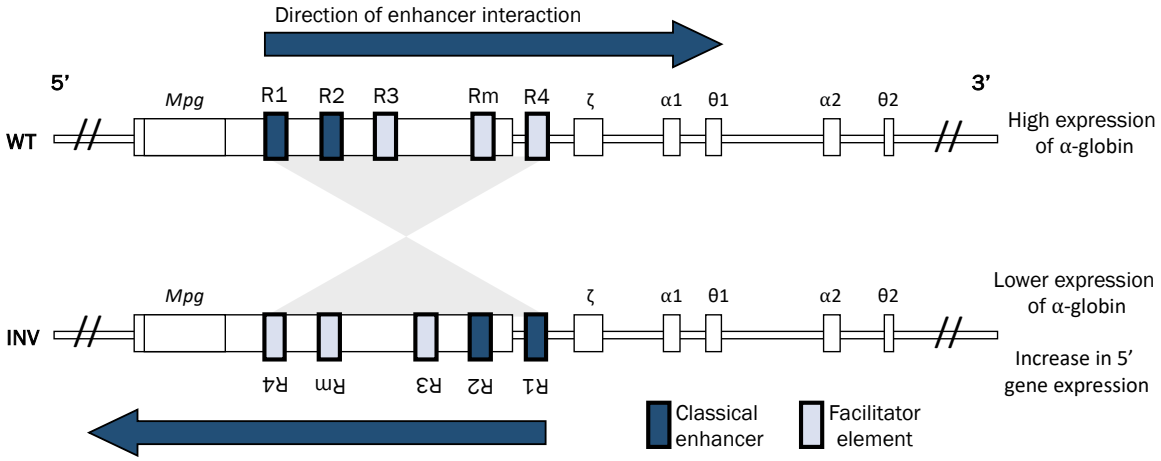

Figure 4.1: Inversion of the α-globin super-enhancer reveals interactions and target gene expression is orientation dependent.

Summary of findings from Kassouf et al., (Preprint 2022). The α-globin genes are controlled by a super-enhancer consisting of classical enhancer elements (R1 and R2, dark blue) and facilitator elements (R3, Rm and R4, light blue). In the WT α-globin locus, the enhancers show preferential interactions with their target genes at the 3' end of the locus. Upon inversion of the entire super-enhancer cluster, the direction of interaction flips, to instead interact and activate the genes at the 5' end of the locus, and concurrently α-globin expression decreases.

Evidence from evolutionary conservation of globin gene loci and other loci organisation suggests that the orientation of regulatory elements may be important. The α-globin and β-globin loci in mammals have the same linear organisation of elements, such that the target genes are downstream of their enhancers with their transcription directed away from the enhancers (Philipsen and Hardison, 2018). This is not unexpected as they both originated from the same duplicated locus during evolution. When considering this from a larger evolutionary perspective, i.e. in other animals including humans, platypuses, chickens and frogs; the organisation of globin-like loci is further conserved (Hardison, 2012). This level of conservation across multiple examples suggests that the linear organisation and relative orientation of elements have a role in optimising super-enhancer induced gene expression.

As discussed in Chapter 3, when a copy of the α-globin gene was inserted in either orientation ectopically, expression of the natively orientated gene was massively higher than that of the reverse (**Figure 3.10a**). This suggested that orientation does dictate the level of enhancer-induced expression, however due to presence of other copies of the gene, it was difficult to investigate the phenomenon further in that model. Therefore, in this chapter, I address the question of whether the native mouse α-globin genes respond to the super-enhancer in an orientation-specific manner, by inverting the orientation of the α-globin gene in its native position to assess changes in expression and cohesin distribution to gain mechanistic understanding.

4.2 Results

4.2.1 Generating an Hba-a1 inversion by non-homologous end joining (NHEJ)

Due to the homology of the 3' region of the α-globin locus, including large regions of homology around the genes themselves, it was important to simplify the locus such that a single copy of the α-globin gene could be individually manipulated. Using CRISPR-Cas9 mediated HDR, I deleted a 16.1kb region containing *Hbq-1a*, *Hba-a2* and *Hbq-1b* in hemizygous mESCs, to generate a locus with one copy of *Hba-a1* (α1-only) (**Figure 4.2a**). This 16.1kb deletion encompasses the CTCF sites within the *Hbq-1a* and *Hbq-1b,* but as we have previously reported, deletion of these sites alone doesn't lead to changes in α-globin expression or changes in interaction profile (Harrold *et al*., 2020). I therefore concluded that loss of CTCF sites in the deletion would have a negligible impact on the integrity of the locus. The deletion was verified using PCRs across the expected breakpoint, such that a product was only generated in the presence of the deletion (**Figure 4.2c**).

Following this, I targeted CRISPR-Cas9 with 2 sgRNAs targeting approximately 1kb upstream and downstream of the remaining *Hba-a1* gene. This generates double strand breaks which are then repaired by endogenous Non-Homologous End Joining (NHEJ) machinery (Materials and Methods) (**Figure 4.2a**); whilst the majority of products would be deletions, in a small proportion of cases the fragment released by these cuts is inverted during re-insertion (Blayney *et al*., 2020). Two successfully verified clones for *Hba-a1* inversion (α1-INV) were generated and verified using a combination of orientation specific PCRs and Sanger sequencing (0.7 % targeting efficiency, 2 positive clones / 288 single clone colonies tested) (**Figure 4.2b**). The breakpoints were also checked using Sasquatch (Schwessinger *et al*., 2017) to confirm no new TF binding sites were generated. As mentioned, the inverted sequence was designed to include flanking regions around *Hba-a1* following detailed characterisation of the underlying sequences; 1.4 kb upstream of the TSS and 900bp downstream of the stop codon. The aim of this design was to conserve native UTRs, distal promoter elements and the termination signals relative to the gene; therefore, changes in expression and interactions are solely due to the inversion and not to disruption of underlying elements. Additionally, due to the small size of *Hba-a1* (1kb) the relative change in distance between the enhancers and promoter is negligeable upon inversion, removing another potential confounding factor, as distance between enhancer-promoter pairs is known to affect expression (Rinzema *et al*., 2022; Zuin *et al*., 2022).

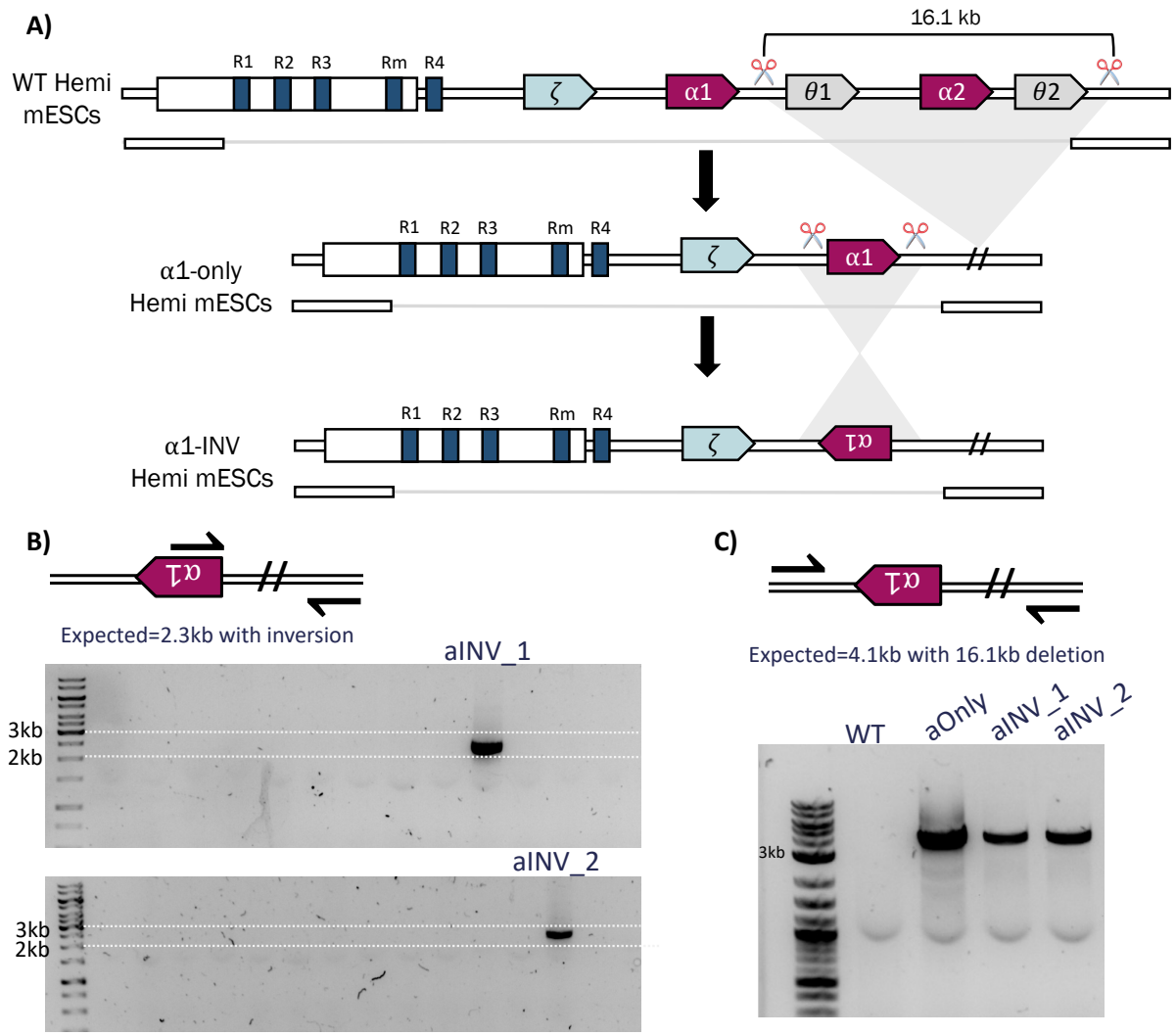

Figure 4.2: Schematic and data of gene editing for the inverted α-globin gene.

A) *Schematic of the parental line and the targeting strategy; the targeted mESC line is hemizygous for the α-globin locus with one deleted and one WT allele. CRISPR:Cas9 was used to delete a 16.1 kb region at the 3' end of the locus spanning Hbq-1a (ϑ1), Hba-a2 (α1), Hbq-1b (ϑ2, Δ[Hbq-1a, Hba-a2, Hbq-1b]) to create the α1-only model. The remaining copy of the Hba-a1 gene was then inverted using NHEJ following double stranded nicking.* ***B)*** *SYBR-Safe stained agarose gel showing the amplicon from orientation-specific PCR used for screening positive clones (aINV_1, aINV_2); as expected, an inversion and 16.1kb deletion resulted in a product of 2.3kb.* ***C)*** *A 4.1kb PCR product is expected in models with the large deletion (aOnly, α-globin gene in native orientation and aINV_1, aINV_2, the verified inverted clones) and was used for Sanger sequencing (not shown).*

In addition, to confirm the integrity of the locus following genome editing and to check if the locus was accessible during erythropoiesis, I performed ATAC-seq on *in vitro* derived CD71+ erythroid cells (**Figure 4.3**). To note, *Hba-a1* and *Hba-a2* are identical in sequence (apart from one SNP) and so when aligning reads to an unedited reference genome, reads from the single copy of *Hba-a1* do align to *Hba-a2* as an artifact. Despite this, the loss of reads from the region does confirm the deletion as well as confirming that the enhancer elements were similarly accessible in both all models (**Figure 4.3**).

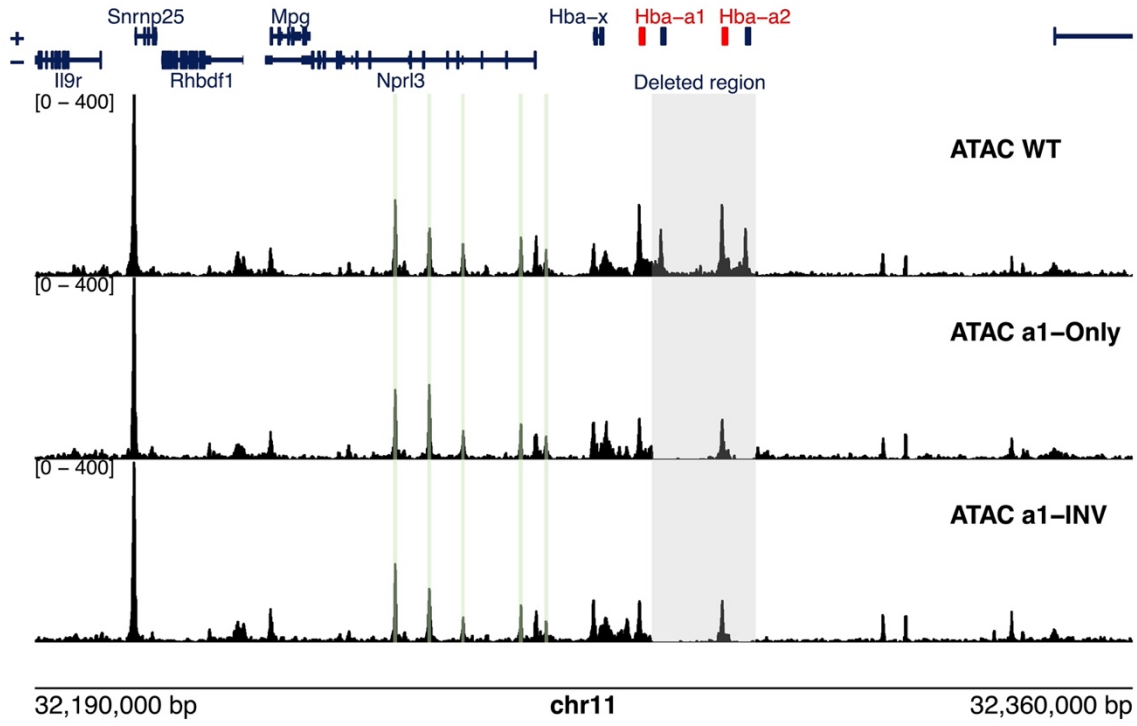

Figure 4.3: Chromatin accessibility of cis-regulatory elements in edited α-globin loci.

All material was collected from in vitro derived CD71+ erythroid cells. ATAC-seq tracks for hemizygous WT, α1-only (α-globin gene in native orientation) and α1-INV (average from 2 verified clones) models aligned to mm10 reference genome. Enhancer elements are highlighted in green and the expected 16.1kb deletion highlighted in grey. Reads mapping to the second copy of Hba-a2 in the deletion region are artifacts of alignment. Both copies of the Hba-a1/2 gene annotations are highlighted in red.

4.2.2 Inversion of a single copy of the α-globin gene, relative to the superenhancer, leads to reduction in expression

To determine if orientation may affect gene expression, RNA was isolated from CD71+ erythroid cells derived following *in vitro* differentiation of the edited mESCs and assayed by RT-qPCR (**Figure 4.4**). Material from WT hemizygous CD71+ cells was used a positive control and a benchmark of expression for the single copy *Hba-a* locus. Expression of α-globin in α1-only was reduced compared to that of WT levels; this is consistent with the loss of a copy of *Hba-a2* (**Figure 4.4a**). Surprisingly, upon inversion of the gene (α1-INV), α-globin decreases by approximately 30 % compared to the natively orientated α1-only (**Figure 4.4a**). This shows that the α-globin gene does indeed have a preferential orientation with respect to its enhancers for optimal expression. Whilst this decrease may appear relatively small,

considering α-globin transcripts are produced in the magnitude of 10^4 transcripts per cell, a 30 % reduction results in a considerable loss of thousands of transcripts.

I then checked whether the other genes in the locus were affected by the changes at *Hba-a1*. Firstly, the large deletion, Δ[*Hbq-1a, Hba-a2, Hbq-1b*], present in α1-only did not change the expression of all other genes with respect to WT, namely; ζ-globin, *Snrnp25, Rhbdf1, Nprl3* and *Mpg* (**Figure 4.4b-f**). In this case, the loss of one of the target genes does not lead to promiscuous enhancer activity at the other genes of the locus. Upon inversion of the remaining copy of α-globin, no change in ζ-globin and the other genes in the locus was observed either (**Figure 4.4b-f**).

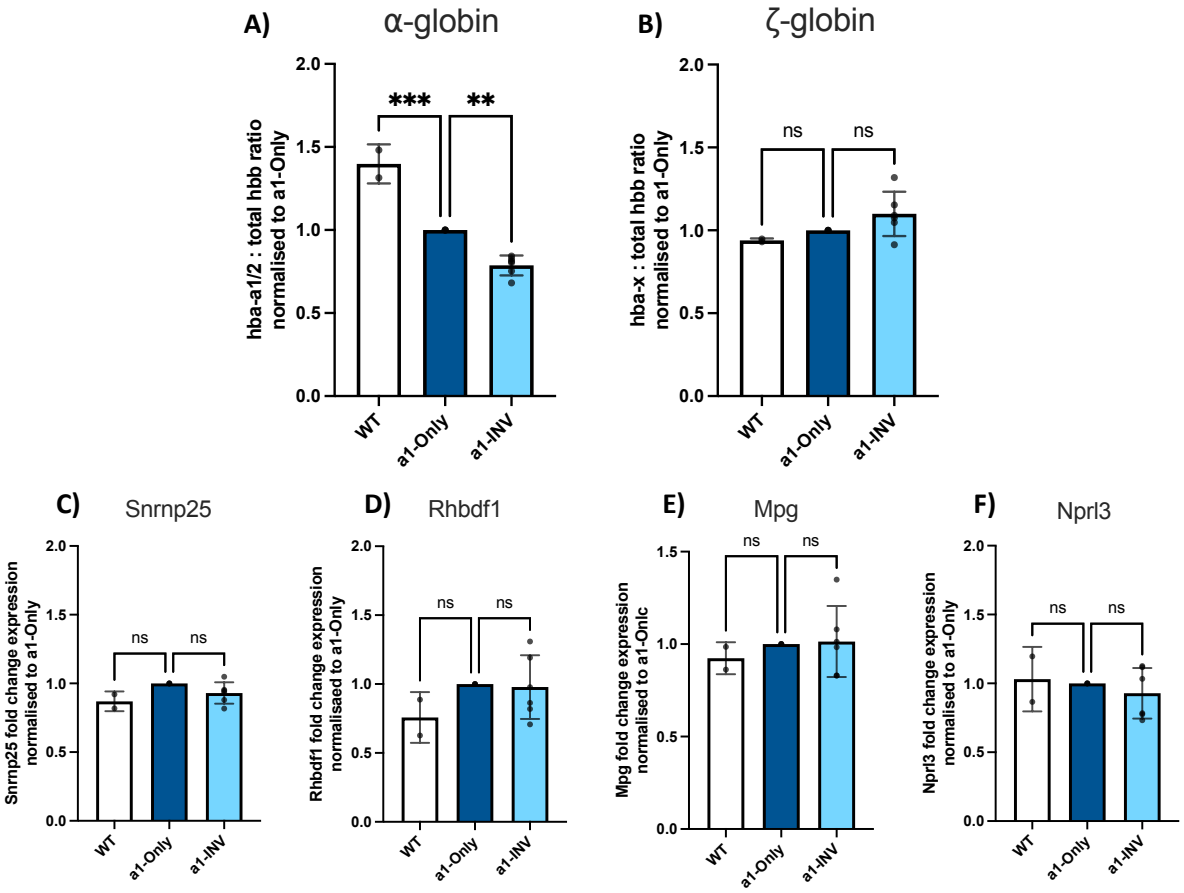

Figure 4.4: α-globin expression decreases upon inversion, whilst neighbouring genes are unaffected.

All material was collected from in vitro derived CD71+ erythroid cells. RT-qPCR quantification of mature ***A)*** *Hba-a1/2 and* ***B)*** *Hba-x transcripts against the mean output of the β-globin locus (Hbb-b1/2, Hbb-bh1 and Hbb-y) relative to RPS18 and normalised to α1-only. Rt-qPCR quantification of mature* ***C)*** *Snrnp25,* ***D)*** *Rhbdf1,* ***E)*** *Mpg and* ***F)*** *Nprl3 relative to RPS18 and normalised to α1-only. Results are from 3 replicate differentiations with the parental clone of α1-only and 2 clones of α1-INV. Error bars display standard deviation and significance calculated using Unpaired t-tests (ns = P > 0.05, * P ≤ 0.05, ** P ≤ 0.01, *** P ≤ 0.001).*

4.2.3 Gene inversion leads to changes in cohesin accumulation

Given the current understanding on how enhancer/promoter proximity results in transcriptional output, the reduction in expression due to orientation was an unexpected result. This observation suggests a mechanism in which the linear organisation of regulatory elements including the sequence orientation plays a role in expression. Cohesin is known to translocate across chromatin in a linear fashion; we have also observed that transcription interferes with its translocation (Chapter 3) as also observed in other reports (Busslinger *et al*., 2017b; S. Zhang *et al*., 2023). Therefore, I considered that cohesin-mediated loop extrusion may be a contributing factor in interfering with expression. I performed Rad21 ChIP-seq on CD71+ erythroid cells derived from both α1-only and α1-INV clones (**Figure 4.5**). Cohesin distribution across the locus is almost identical when looking at peaks over CTCF sites, enhancer elements and other promoters in the locus. However, there is a distinct change in cohesin at the inverted α-globin gene. In α1-only, there is a peak of Rad21 at the TSS of *Hba-a1,* which then decays gradually past the gene body into the downstream (3') region (**Figure 4.5** zoomed panel). In comparison, at the inverted gene, there is still build up at the TSS, however the decay does not occur to the downstream (3') region but appears to mirror the distribution in α1-only, instead accumulating at the upstream (5') region. This is admittedly complicated by the presence of *Hba-x* in the upstream region, which is also expressed and accumulates Rad21, however the loss of Rad21 from the downstream region is clear in α1-INV. This evokes a boundary effect imposed by the TTS as well as TSS, as previously reported (Valton *et al*., 2022), stalling the cohesin that is preferentially loading and translocating from the enhancer cluster towards the α-globin genes (Preprint Stolper *et al.*, 2023). In the process, this stalling may impact the enhancer-promoter interaction required for optimal expression. An assessment of the three-dimensional conformation of the locus in both the native and inverted gene states would shed light on any structural impact the inversion may have on the interaction domain.

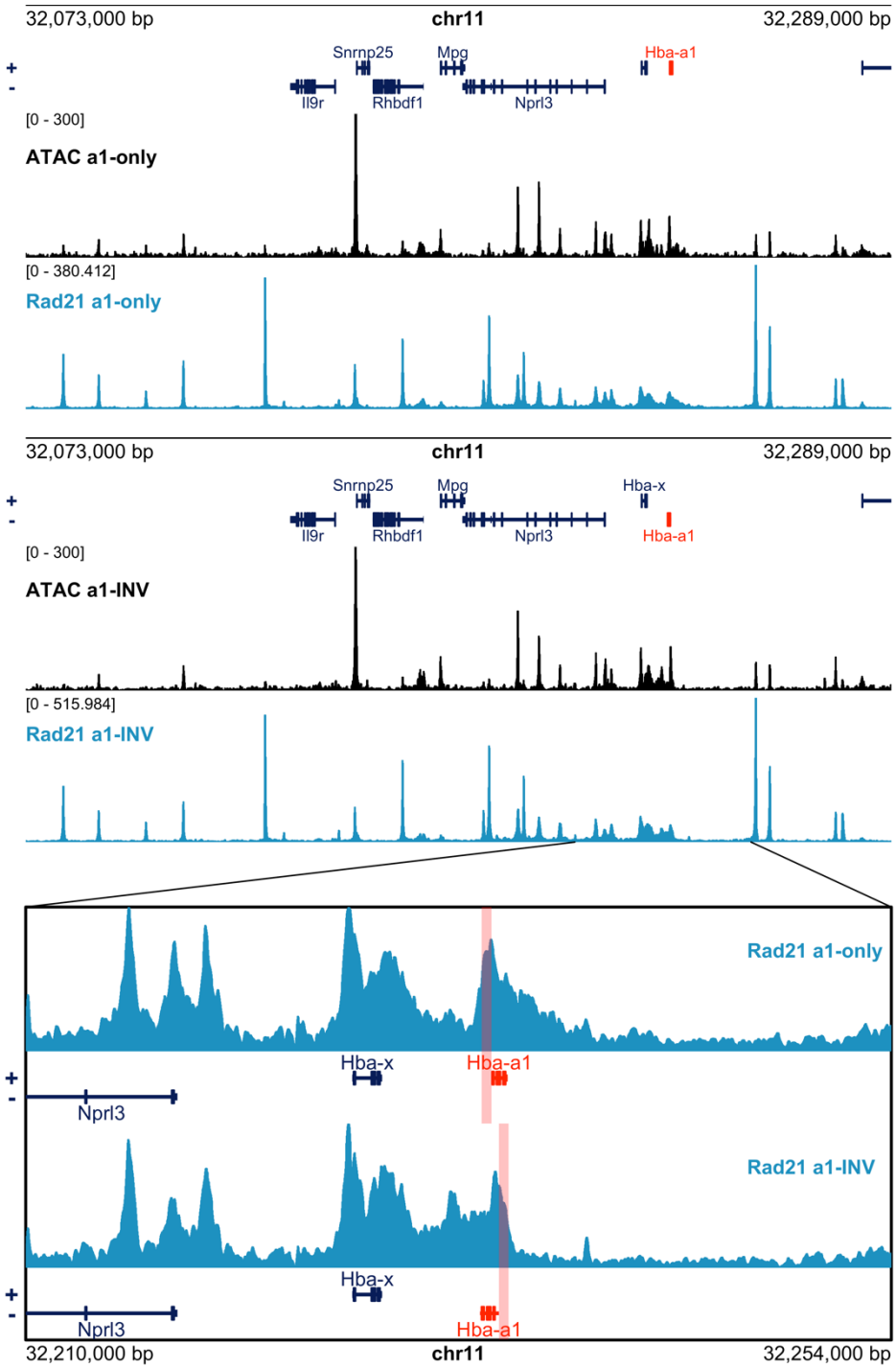

Figure 4.5: Inversion of Hba-a1 causes redistribution of cohesin at the gene.

All data is generated from in vitro derived CD71+ erythroid cells and plotted against custom genome sequences (either a1-only or a1-INV) created from the mm39 reference. All gene annotations are plotted with respect to their strand (top is +, bottom is – strand) and Hba-a1 is highlighted in red. ATAC-seq tracks are shown in black and Rad21 ChIP-seq shown in blue. Zoomed region shows Rad21 ChIP-seq as above, for coordinates chr11:32210000-32254000 (custom genome generated from the mm39 reference): Rad21 ChIP-seq for α1-only is shown in the top panel, α1-INV in the bottom panel. The TSS+500bp upstream is highlighted in red in both panels. Note the position changes in each model due to the orientation of Hba-a1.

4.3 Discussion

Inversion of the α-globin gene in its native position leads to a slight but significant reduction in gene expression, providing direct evidence that gene orientation with respect to its cognate super enhancer affects level of enhancer-induced expression. This reduction is also associated with a redistribution of cohesin across the gene, suggesting a mechanistic role of cohesin-mediated loop extrusion.

Cohesin accumulation around the natively orientated and inverted gene, in both cases, is consistent with their respective directions of transcription; with accumulation decreasing gradually from the TSS. As presented in Chapter 3, this same disparity in expression of the ectopic α-globin fragments can be observed, such that the natively orientated inserts are expressed more than the reverse (**Figure 3.10)**. Similarly, cohesin accumulation around the insert site (**Figure 3.11**) is consistent with the profiles seen at the endogenous position (**Figure 4.5)**, however is more apparent in this chapter due to more accurate alignment of ChIP-seq reads. With these observations, I re-assessed the genome-wide profiles of cohesin around active TSS (**Figure 3.12b**); importantly, the profiles are plotted such that the direction of transcription are aligned. Interestingly, the profiles are not perfectly symmetrical and there is a very slight bias in accumulation direction of transcription, consistent with the profile observed in **Figure 4.5.** Whilst ChIP-seq only provides a population snapshot, the profiles of Rad21 binding are suggestive of cohesin moving with or been pushed along by transcriptional machinery, consistent with previous reports in yeast (Brandão *et al*., 2019; Glynn *et al*., 2004; Lengronne *et al*., 2004) and in mammalian cells (Busslinger *et al*., 2017).

Therefore, transcription of the inverted gene may be affected by the linear tracking of cohesin (**Figure 4.6a**). Cohesin is present at enhancers, likely actively loaded by NIPBL, to then translocate bi-directionally (Chapter 3 Discussion). Due to the presence of the strong enhancer proximal CTCF sites at the α-globin locus, cohesin can be considered to track almost exclusively unidirectionally, toward the α-globin genes (Stolper *et al.*, 2023, preprint). In the native locus, transcription from the α-globin genes is congruent with translocation of cohesin from the enhancers. As discussed in Chapter 3, the transcription machineries in this scenario may act also as steric obstacle for the extruding cohesin, but this may help to increase the frequency of the enhancer-promoter interaction and enhance transcription of the most proximal gene. When the gene is inverted, transcription instead proceeds in the opposite direction to cohesin progression from the enhancers. Considering the large number of protein complexes associated with transcription, such as RNAPII, splicing apparatus, capping, and termination machinery, this could feasibly create an obstacle to cohesin tracking which reduces the

probability of cohesin completing a full trajectory, leading to fewer enhancer-promoter contacts and the reduction in transcription. On the other hand, the extruding cohesin may equally act as a roadblock for the transcription elongation machinery and reduce the overall transcription efficiency.

Alternatively, this could be caused by lock-key type interaction between the super-enhancer associated Mediator complex and the PIC at the promoter; such that the gene must be in the native orientation for correct positioning of the Mediator to facilitate RNAPII release (Mediator reviewed in Richter *et al.*, 2022) with the reverse orientation disfavouring RNAPII firing (**Figure 4.6b**). However, this seems unlikely, as flexibility in the DNA strand should allow for adjustments in topology to correct for the inversion. *In silico* polymer-dynamics modelling may be able to discern if DNA topology could facilitate this correction, however modelling with the required complexity to simulate this is currently unfeasible.

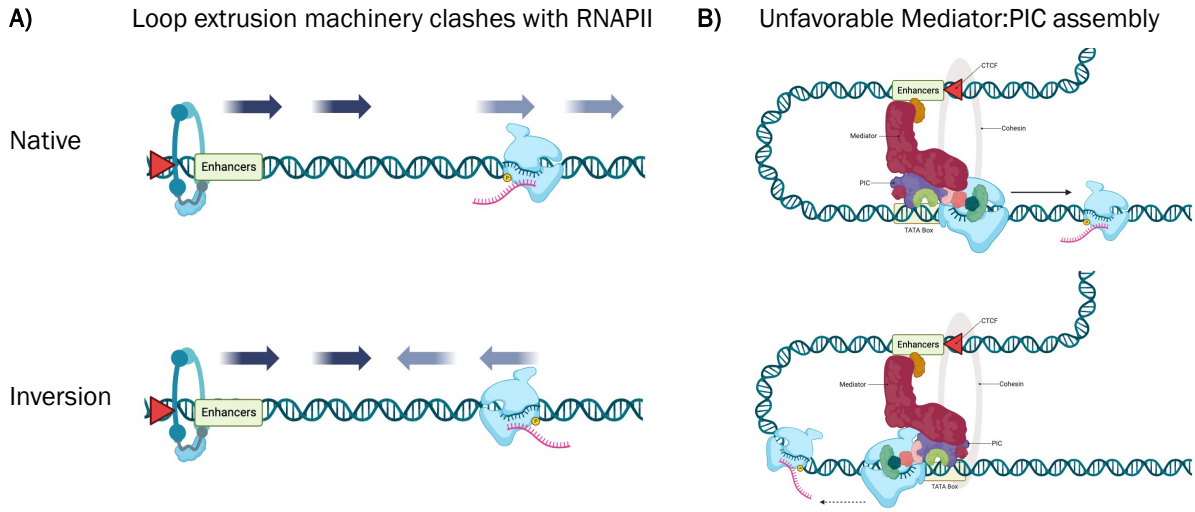

Figure 4.6: Proposed mechanisms of orientation-dependant enhancer induced transcription.

A) Linear tracking of loop extrusion: in the native orientation transcriptional machinery (light blue arrows) moves together with cohesin (dark blue arrows) (top); whilst upon inversion, transcriptional machinery (light blue arrows) converges with cohesion (dark blue arrows), creating obstacles to one another and decreasing the probability for enhancer-promoter interaction (bottom). ***B)*** *Unfavourable Mediator:PIC assemblies: in the native orientation the Mediator complex bridges to the Pre-Initiation Complex (PIC) such that Mediator binds the RNAPII-CTD leading to phosphorylation and release (top); in the reverse orientation, correct assembly may not be achieved as frequently as the PIC is not in the optimal position for Mediator interaction, however this could be overcome by additional DNA intra-looping for example.*

Due to the difficulties in performing 3C in the *in vitro* derived erythroid cells, as previously mentioned in Chapter 3, it remains to be determined if there are changes in contact frequency between the enhancers and inverted gene. However, it's unclear if this would help in determining the underlying

mechanism. Of course, disruption of cohesin-extrusion could change interaction frequencies; however, Mediator is also reported to act as an architectural bridge between enhancers and promoters (Ramasamy *et al*., 2023), and this could feasibly lead to changes in interaction as the Mediator:PIC complex is unfavourably assembled. Determining the distribution of PolII by ChIP may shed light on the relationship between cohesin and PolII termination and off-loading, but this remains to be done. Hence, it remains unclear how mechanistically the gene inversion leads to compromised transcription.

Of relevance, the β-globin genes, controlled by a cluster of enhancer elements (LCR) which also classify as a super-enhancer, were inverted by Cre-mediated recombination in large, randomly integrated transgenic inserts in mice. Depending on the erythroid developmental stage, adult β-globin expression was unaffected (definitive) or massively upregulated (primitive) with the inversion, whilst foetal γ-globin expression was reduced in primitive erythrocytes and abolished in foetal erythrocytes in the inversion (Tanimoto *et al*., 1999). Various study design factors confounded the study interpretation; the entire gene cluster was inverted, hence changing the genomic distance between the LCR and each of the genes which could also affect expression levels. So direct effect of the inversion on individual genes remains unclear. Similarly, inversion of *Myh* genes relative to their super-enhancer cluster showed a decrease in expression, however once again, the inversion resulted in changes of super enhancer proximity (Dos Santos *et al*., 2022). Recently, Kovina *et al.,* generated a series of constructs within the α-globin locus in a chicken erythroid cell line (HD3) in which they introduced an RFP reporter under the control of zebrafish (*Danio rerio*) globin promoters. When the inserted reporter's transcription was opposite to the direction of endogenous chicken globin gene transcription, there was a drastic absence of reporter expression. When reporter transcription coincided with endogenous transcription, the reporter generated high levels of expression. This presents promising evidence of preferential orientation of transcription, however there were also some limitations. Firstly, the reporter was always under control of a pair (β-globin and α-globin) of divergently orientated promoters, meaning all constructs could initiate in either direction. Secondly, the promoters and reporter were transgenes and non-native to the chicken model they were introduced into, therefore they maybe under different control with respect to their native environment. Lastly, the mechanism of how preferential orientation was facilitated remained undetermined. Therefore, it remained unclear as to how gene orientation may affect enhancer-driven expression.

Given the widely held view that enhancers function on their cognate promoters independently of orientation, this is a surprising result. Admittedly, the effect on expression is subtle and in the context

of the α-globin locus, in which the super-enhancer shows strong inherent directionality (Kassouf *et al*., 2022); however, it does suggest that there is a functional role in the organisation of elements, such that native orientation optimises enhancer-promoter communication. Whilst this is the case in the α-globin locus, it should be considered that other loci may not show similar preference; however, to the best of my knowledge, no other single gene inversions, without confounding positional effects, have been reported.

Whilst simple in set up, the findings presented here provide direct evidence that the orientation of a target gene with respect to a super-enhancer, does contribute to its level of activation, contesting a long-held assumption that promoter and enhancers interact independently of their orientation.

4.4 Acknowledgments for Chapter 4

This project was completed with Noortje van Dijk, a visiting MSc student from the Department of Experimental Haematology (UMCG), University of Groningen, who I had the privilege to supervise and work with. Phillip Hubiltz (Weatherall Genome Engineering Facility) consulted on targeting designs and created all the targeting constructs used in this chapter.

Chapter 5: Deletion of CTCF sites within the α-globin sub-TAD

5.1 Introduction

CTCF bound sites are known to contribute to the insulation of enhancer interactions, limiting them to within TADs and sub-TADs; this has been observed in the murine α-globin locus (**Figure 5.1**). A significant role for such genome compartmentalisation has been strengthened by specific examples of an apparent functional role in the regulation of gene expression (Dowen *et al*., 2014; Hanssen *et al*., 2017; Lee *et al*., 2017; Lupiáñez *et al*., 2015; Narendra *et al*., 2015; Symmons *et al*., 2016; Vos *et al*., 2021). Acute depletion of CTCF causes changes in the formation of TADs and sub-TADs but relatively few changes in gene expression, challenging the significance of the relationship between CTCF-mediated genome structure and function (Hsieh *et al*., 2020; Hyle *et al*., 2019; Nora *et al*., 2017; Xu *et al*., 2021). To address this in detail, we previously generated mouse models with informative deletions of CTCF sites within and located either side of the α-globin locus. These deletions subtly changed interactions within the domain and in some cases, (e.g., deletion of HS-38/39) they changed expression of genes lying upstream (5' of the locus). These changes in expression resulted from a loss of insulator function. Importantly, α-globin expression itself was not affected upon deletion of any of these CTCF sites individually or in informative combinations (Hanssen *et al*., 2017; Harrold *et al*., 2020) (**Table 5.1**).

Together, these studies showed that the upstream (5') boundary of the sub-TAD (HS-38/39) normally insulates the genes lying 5' of the locus from the effects of the α-globin enhancer. By contrast, the downstream boundary as judged by Capture-C experiments, did not correspond to the downstream CTCF elements. Rather, the downstream boundary corresponded more closely to the most 3' α-globin gene (Chapter 3). Although none of the CTCF site deletions on their own or in combination affected α-globin expression, the question remained as to whether CTCF sites within and surrounding the α-globin sub-TAD are needed at all to set up and correctly maintain α-globin expression. To answer this, I generated a model in which all CTCF sites within the α-globin sub-TAD are deleted.

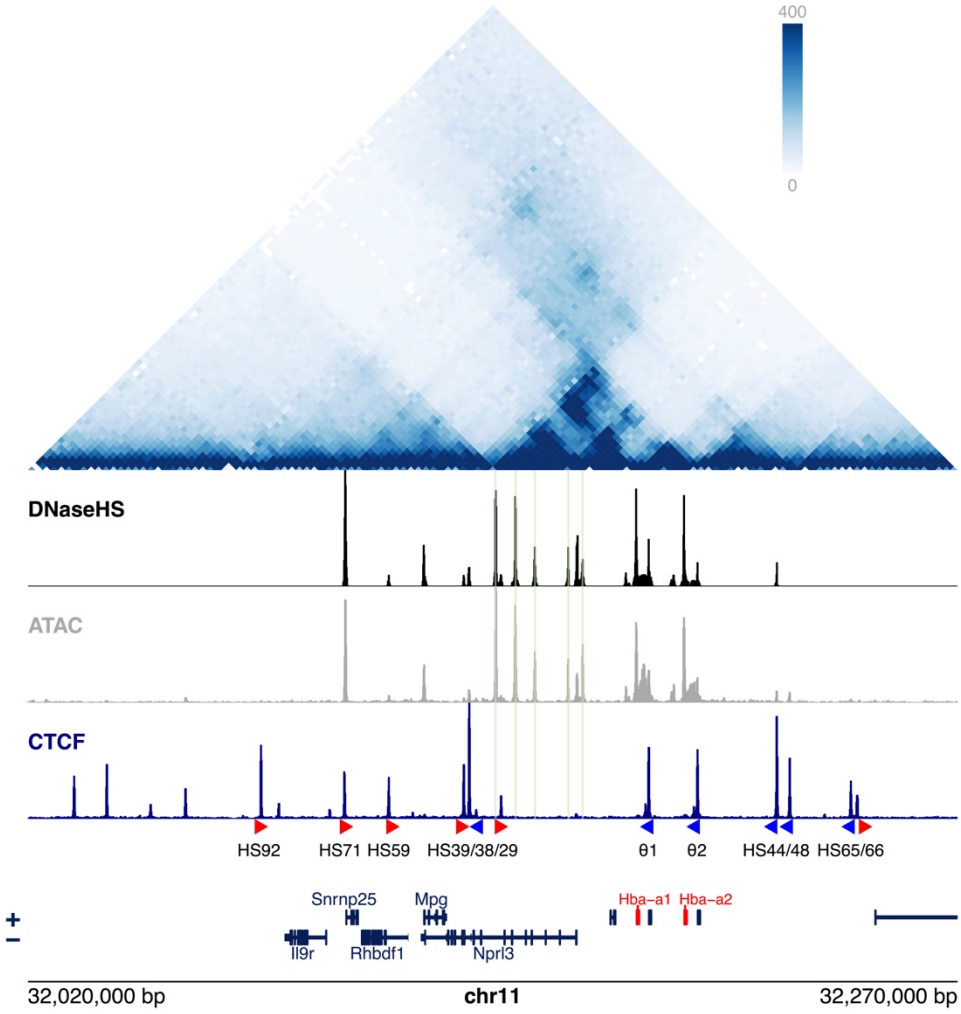

Figure 5.1: Structure of the α-globin locus TADs and underlying elements.

Tiled-C contact matrix displaying fine-structure of the α-globin locus at 2kb resolution replotted from (Oudelaar et al., 2020a) with DNase I hypersensitive sites sequencing (DNase-seq), ATAC-seq and CTCF ChIP-seq tracks plotted below, published previously (Hanssen et al., 2017). All tracks were derived from primary mature erythroid cells. Major CTCF sites and their orientation are highlighted with arrows and enhancer elements are highlighted in green (coordinates mm9: 32020000-32270000).

Table 5.1: Previously reported CTCF binding site deletions in the α-globin locus			
Deletions	**Side**	**Change in α-globin expression**	**Reported in**
ΔHS-39	5’	No Significant Change	(Hanssen *et al.*, 2017)
ΔHS-38	5’	No Significant Change	
ΔHS-38, HS-39	5’	No Significant Change	
ΔHS-29	5’	No Significant Change	
Δθ1	3’	No Significant Change	(Harrold *et al.*, 2020, preprint)
Δθ2	3’	No Significant Change	
Δθ1, θ2	3’	No Significant Change	
ΔHS+44, HS+48	3’	No Significant Change	
Δθ1,θ2,HS+44,HS+48	3’	No Significant Change	Unpublished

5.2 Results

5.2.1 Creating a CTCF-null (ΔCTCF) α-globin locus using RMGR

To determine the combinatorial role of CTCF binding sites (CBS) in the α-globin locus, I deleted all individual CTCF binding sites in the α-globin locus. As multiple deletions would require multiple rounds of genomic editing using CRISPR:Cas9 HDR, I was concerned at the possibility that repeat editing would decrease the pluripotency of the mESCs. I therefore opted for an alternative method of genomic engineering: Recombinase Mediated Genomic Replacement (RMGR). This allows the entire ~86kb region of the α-globin locus to be exchanged with any DNA fragment with complementary loxP/lox511 sites. In collaboration with Prof Jef Boeke (New York), we used large-scale DNA assembly to synthesize custom loci for exchange with the RMGR-ready locus contained within bacterial artificial chromosomes (BACs) (**Figure 5.2**).

We have previously used this technique to great success to generate the homozygous R2-only mouse model, in which a BAC was synthesized with R1, R3, Rm and R4 deleted, introduced into mESCs and used to successfully raise a mouse line (Blayney *et al*., 2022, in publication). During generation of the model, the low efficiency of the RMGR technique was improved by a series of optimizations; including use of lipofection in the place of electroporation and engineering the RMGR-ready cell line, such that the remaining 'non receptive' WT allele was deleted creating a hemizygous line with one allele of the α-globin locus amenable to RMGR (Hemizygous RMGR-ready). This increased the efficiency of targeting by greatly decreasing the possibility of recombination events at the exchanged locus. In addition, this allows ease of downstream analysis, removing the need to discern between signals from the WT allele and those of the introduced synthetic locus. The RMGR technique has great potential for investigating a range of biological questions by facilitating the creation of complex variants of the α-globin locus, such as exploring the effects of large inversions, large-scale rearrangements, and multiple combinatorial deletions of elements. A previous permutation of this method was used to 'humanise' the murine α-globin locus (Wallace *et al*., 2007). Similar large-scale synthetic approaches have been utilised recently to determine the contributions of regions to gene expression within the *HoxA* locus (Pinglay *et al*., 2022) and *Sox2* locus (Brosh *et al*., 2023).

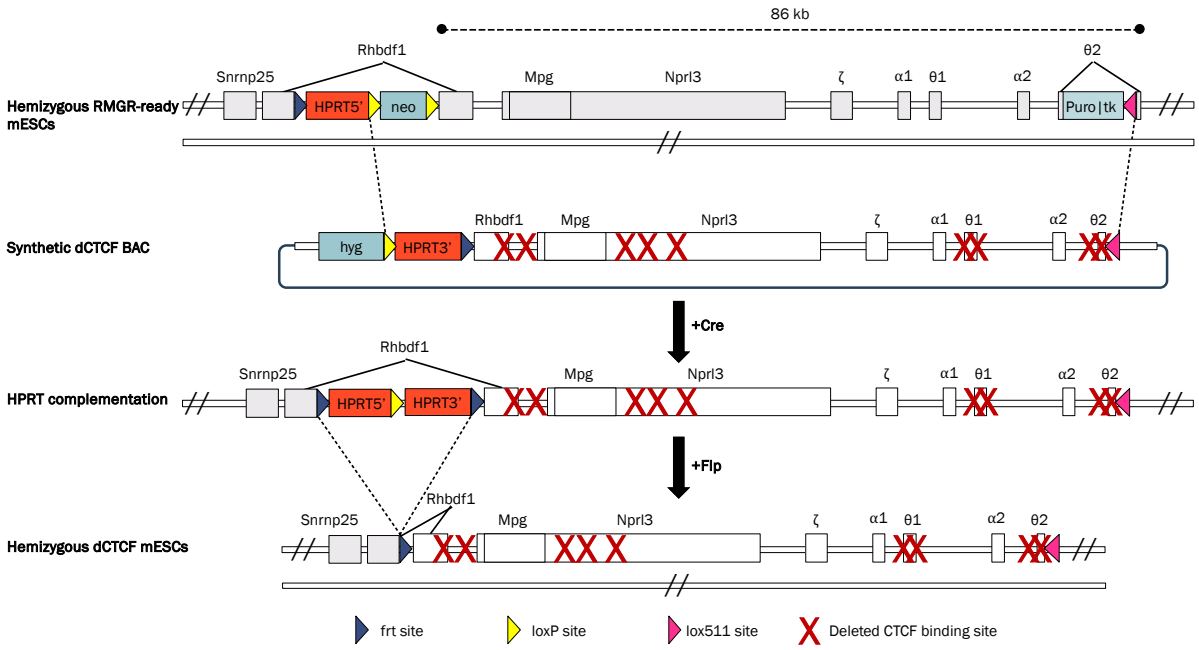

Figure 5.2: Schematic of the genome engineering approach followed to create the CTCF null model.

Parental hemizygous RMGR-ready cells, with Hprt-Δ3'|neomycinR cassette and PuromycinR|tk cassette integrated into the 5' and 3' regions of the α-globin locus (grey). A synthetic locus (white) with all CTCF sites deleted was created in a BAC, with complementary hygromycinR|Hprt-Δ5' cassette. The synthetic locus integrates into the RMGR-ready genome with the use of loxP/lox511:CRE-recombinase activity. The Hprt is removed by Frt:Flippase recombination.

The synthetic ΔCTCF α-globin locus was designed with deletions spanning each site's short CTCF binding consensus sequence as determined by their DNase hypersensitivity footprints previously characterized by our group (Hanssen *et al*., 2017). We have previously observed that the CTCF sites located in *Hbq1a* and *Hbq1b* (θ1 and θ2, respectively) could be separated into 2 distinct CTCF peaks in CTCF ChIP-seq approximately 1kb apart; therefore, these sites were treated independently of one another, as major and minor sites. The deletions were limited to those within the RMGR-ready exchange region and so external sites which remained untargeted in this model include HS-94 (5'), HS-71 (5'), HS+44 (3'), HS+48 (3'), HS+65 (3') and HS+66 (3') (**Figure 5.3**). This a limitation of the ΔCTCF model however, the external/flanking sites are predicted to have low contribution to the interactions within the locus.

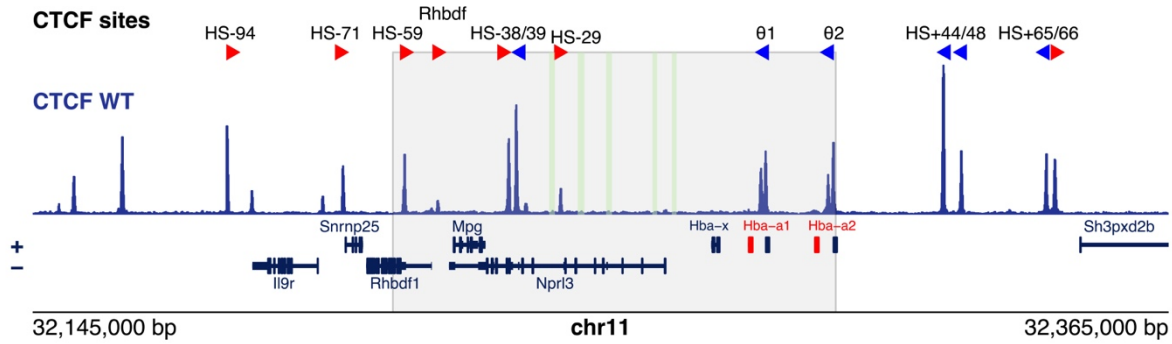

Figure 5.3: Location of CTCF binding sites within the α-globin locus.

Names of CTCF sites (HS: Hyper-Sensitive sites with the number referring to the bp position counting from the ζ-globin TSS with negative values upstream and positive values downstream of the TSS) are noted above their CTCF ChIP-seq peaks shown in dark blue. Orientation of CTCF sites shown as blue and red arrows and enhancer elements are highlighted in green. Grey box highlights the region replaced by RMGR (chr11:32214754-32300533 mm10).

The synthetic ΔCTCF α-globin locus was produced as a BAC and checked by long read sequencing by our collaborator Brendan Camellato in Jef Boeke's group (New York). To reduce instances of structural rearrangements in bacteria, the BACs are kept as single copy vectors, therefore large bacterial preparations are needed to obtain the mass of DNA required for transfection. I propagated and expanded the BAC containing the synthetic locus in large bacterial liquid cultures under antibiotic selection. I extracted BAC DNA by a CsCl:EtBr density gradient, which allowed separation of the BAC from bacterial genomic DNA (**Figure 5.4a**). Following DNA "clean up", restriction digests were performed to determine the integrity of the BAC and to ensure no large rearrangements had occurred during culture. The BAC was co-transfected with a plasmid expressing CRE recombinase into hemizygous RMGR-ready mESCs using lipofection (LTX). It was important that the transfection was done within 2 days of BAC DNA purification as the very large DNA fragment produced is sensitive to degradation, leading to lower transfection efficiency. Even so, only 2 colonies were recovered following HAT selection and HT recovery; and only 1 clone had PCR products consistent with complete exchange at both RMGR exchange 3' and 5' breakpoint sites (**Figure 5.4b**). The exchange of native locus to the ΔCTCF α-globin locus was further probed using PCRs covering the CTCF site deletions and Sanger sequencing of these products confirmed the presence of the expected CTCF site deletions (data not shown).

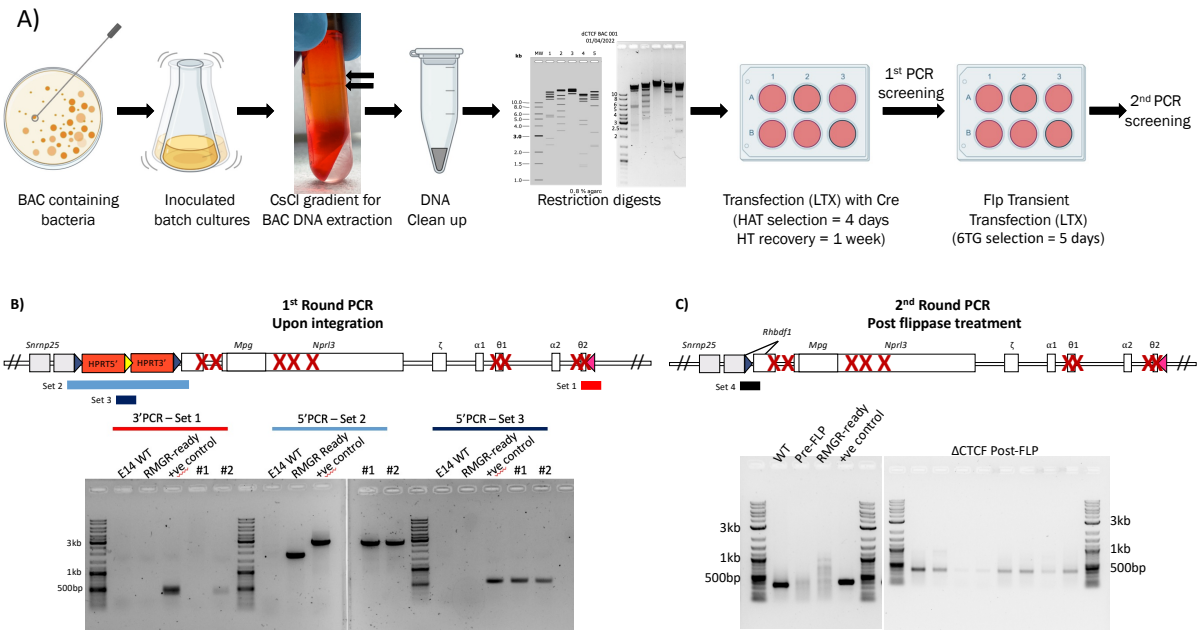

Figure 5.4: Preparation and transfection of the synthetic ΔCTCF α-globin BAC into RMGR-ready mESCs.

A) Schematic of a BAC DNA preparation from bacterial cultures; following quality checks using restriction digestion, the BAC was introduced into RMGR-ready mESCs using lipofection and transient CRE expression and selection. After the first round of PCR genotyping and expansion, the HPRT cassette was removed using transient Flippase (Flp) transfection. ***B)*** *First round of PCR screening; Set 1 (3422 & 3424) amplifies across the 3' breakpoint (528 bp expected with BAC integration); Set 2 (Tile56_Fw_1 & Tile57_Rv_1) amplifies across the completed HPRT cassette (expected 2.8 kb with BAC integration, 1.8 kb expected in the parental RMGR-ready control); Set 3 (HFR2_3 & HFR2_4) amplifies within the completed HPRT cassette (expected 598 bp with BAC integration). Clone #2 had all the expected products indicating the BAC had integrated properly.* ***C)*** *Second round of PCR screening performed after flippase transfection; Set 4 (Tile51_Fw_1 & Tile57_Rv_1) amplifies the regions flanking the HPRT cassette (expected 419 bp with HPRT removal).*

The next step in validation was to confirm the activity of the locus and the loss of CTCF loading at the deleted sites. The PCR validated clone was expanded and differentiated *in vitro* following our established protocol to generate erythroid cells (Francis *et al*., 2022). To confirm integration of the synthetic ΔCTCF locus and deletion of the CTCF sites, I performed CTCF ChIP-seq. This confirmed the expected loss of CTCF binding across the α-globin locus with the external CTCF sites unaffected by the exchange (**Figure 5.6**). With this confirmation, I completed the last stage of genome engineering by removing the *Hprt* selection cassette using *Flp*:frt recombination facilitated by the homotypic FRT sites flanking the *Hprt* gene and transient transfection of a *Flippase* expressing plasmid. This step removed the remaining machinery associated with RMGR, resulting in a near-seamless exchanged locus, with only 1 lox site (34bp) and 1 frt site (31bp) left in the final version of the engineered locus (**Figure 5.4c**).

I then performed ATAC-seq from the CD71+ erythroid cells derived from the ΔCTCF mESCs to determine the open chromatin landscape of the locus (**Figure 5.6**); comparing the ΔCTCF ATAC-seq to a hemizygous WT control, the enhancer elements and α-globin and ζ-globin genes have equivalent regions of chromatin accessibility, confirming the integrity of the ΔCTCF α-globin locus and that the major elements were active.

5.2.2 CTCF null α-globin locus presents with a reduction in α-globin expression

Having confirmed the integrity of the ΔCTCF α-globin locus, I performed RT-qPCR from RNA extracted from the CD71+ *in vitro* derived erythroid cells to determine how the CTCF deletions affected α-globin expression. In the CTCF-null locus, there was a ~60% reduction in α-globin expression (**Figure 5.5a**). This reduction was also observed for the embryonic ζ-globin expression (**Figure 5.5a**); this contrasts with previous results, wherein single or pairwise deletion of 5' or 3' CTCF sites resulted in no significant changes in α-globin expression. With the reduction in globin expression, there is a complementary increase in expression of external flanking genes, *Rhbdf1* and *Mpg* (**Figure 5.5b**) – this is in line with increase in inappropriate activation of the flanking genes seen with ΔHS-38/39 (Hanssen *et al.*, 2017).

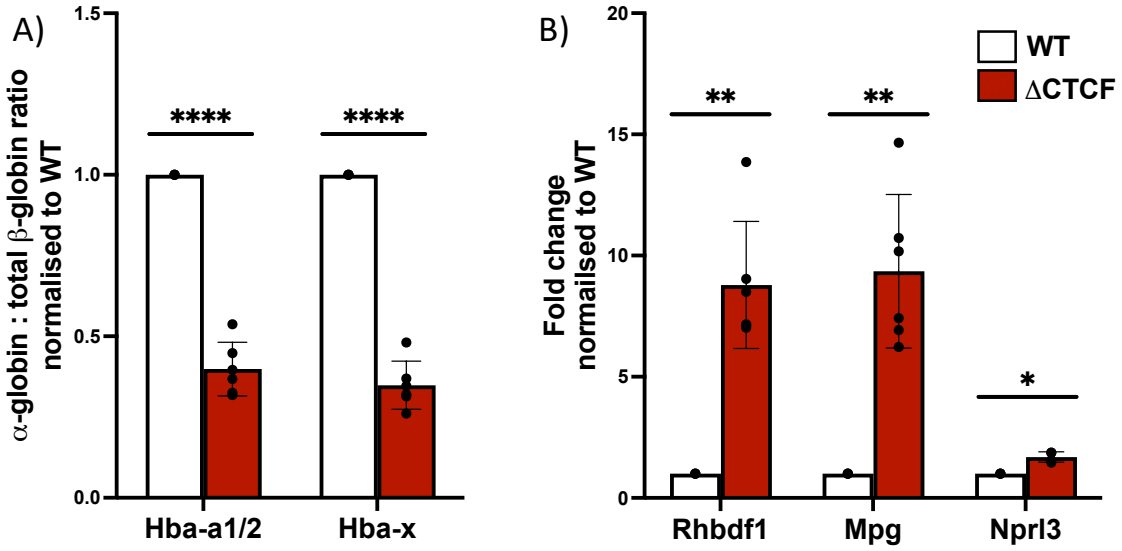

Figure 5.5: Gene expression changes in ΔCTCF in vitro derived erythrocytes.

All material was collected from in vitro derived CD71+ erythroid cells. ***A)*** *RT-qPCR quantification of mature Hba-a1/2 and Hba-x transcripts against the mean output of the β-globin locus (Hbb-b1/2, Hbb-bh1 and Hbb-y) relative to RPS18 and normalised to WT.* ***B)*** *RT-qPCR quantification of mature transcripts from Rhbdf1, Mpg and Nprl3 genes relative to RPS18 and normalised against WT. Results are from 3 replicate differentiations with 2 of ΔCTCF α-globin clones. Error bars display standard deviation and significance calculated using Unpaired t-tests (ns P > 0.05, * P ≤ 0.05, ** P ≤ 0.01, **** P ≤ 0.0001).*

5.2.3 Cohesin is redistributed to other CTCF sites whilst presence at enhancers and genes is unaffected at the ΔCTCF α-globin locus

Cohesin mediated loop extrusion is considered one of the main mechanisms by which interaction domains are formed. CTCF sites accumulate cohesin, supporting their role in chromatin structure. To determine how cohesin distribution is changed across the locus with CTCF site deletions, I performed Rad21 ChIP-seq on CD71+ erythroid cells derived from ΔCTCF α-globin mESCs (**Figure 5.6**). As expected, cohesin accumulation is lost at the deleted CTCF sites; notably there was a slight increase in Rad21 accumulation at HS-71 at the 5' and HS+44/48 sites at the 3' of the locus. Rad21 peaks at the enhancers and genes appear to be unaffected. The pattern of cohesin across the ΔCTCF α-globin locus is in line with cohesin loading at the enhancers and genes which then tracks outward to the flanks of the locus; without the internal CTCF sites, cohesin continues further in a bi-directional manner to be halted at the first sites encountered after the deleted CTCF sites. Rad21 accumulation is also observed at other gene promoters, including *Snrnp25, Mpg* and *Nprl3,* which have also increased in expression in the ΔCTCF α-globin locus. The consistent Rad21 accumulation across the active elements in the locus in the absence of CTCF sites is in line with the continued, albeit compromised, expression of the globin genes, suggesting the enhancers can still interact with their target genes.

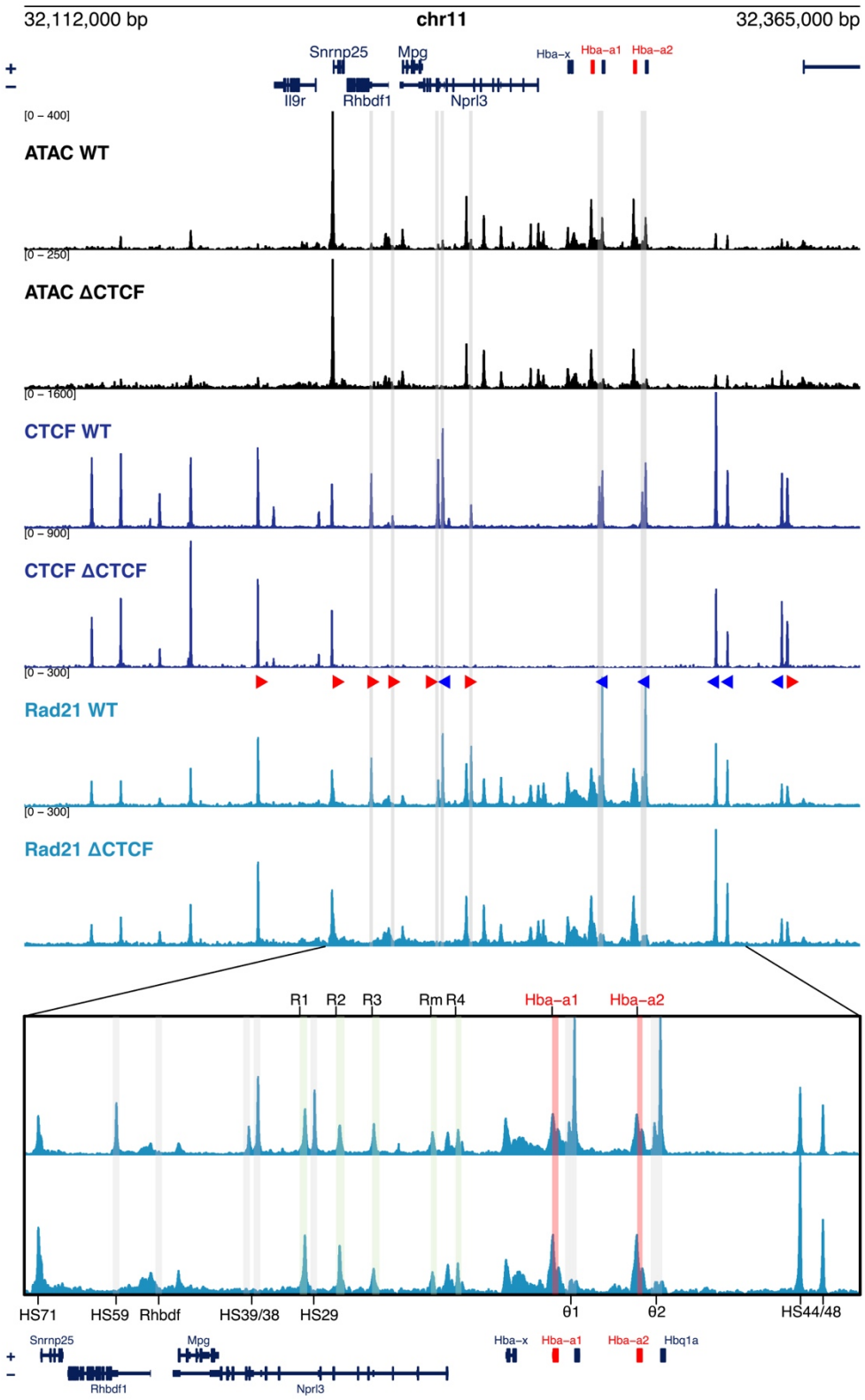

Figure 5.6: Molecular characterisation of the ΔCTCF α-globin locus.
All data is generated from in vitro derived CD71+ erythroid cells and plotted against the reference mm10 genome. UCSC tracks for data from WT and ΔCTCF α-globin clones (ΔCTCF), Top: ATAC-seq is shown in black. Middle: CTCF ChIP-seq peaks shown in dark blue. Bottom: Rad21 ChIP-seq shown in light blue. Zoomed region shows Rad21 ChIP-seq as above, for coordinates 32203000-32330500 (mm10). Expected CTCF site deletion regions in the ΔCTCFα-globin model are highlighted in grey, enhancer elements in green and α-globin genes in red.

5.3 Discussion

This is the first report of an α-globin model in which all CTCF sites were deleted simultaneously. As predicted, 5' non-erythroid genes were up-regulated in the absence of what we have previously characterised as the α-globin upstream boundary, the 5'CTCF sites (HS-38/39). However, surprisingly, the ΔCTCF α-globin model revealed a reduction in globin gene expression that was not observed in all previous combinatorial CTCF deletions (Hanssen *et al*., 2017; Harrold *et al*., 2020). Before we proceed to interpreting the unexpected phenotype, it is worth noting that more sequence validation is required to secure the engineered model's integrity. A possibility that there might be unexpected sequence deviations arising following genomic introduction of the synthetic locus, such as rearrangements or SNVs, that cannot be observed in the short-PCR validations or short-read NGS methods presented so far. In the previous generation of a RMGR derived model, linked-read sequencing was used to sequence across the entire locus (Blayney *et al*., 2022, in publication); unfortunately, this service was discontinued during the generation of the ΔCTCF α-globin model. Alternative methods are being actively considered, such as baited sequencing (Brosh *et al*., 2021) or Cas9-targeted Long-Read Sequencing (Walsh *et al*., 2021; Wongsurawat *et al*., 2020), as well as outsourcing to Pacbio to confirm the integrity of the locus. Despite this, as shown throughout this chapter, the most noticeable differences in NGS chromatin characterization assays (ATAC-seq and ChIP-seq) performed are at the expected CTCF sites and so it is likely the effects observed are due to the CTCF deletions. Following such targeted sequencing, a mouse model will be generated to allow analysis of primary erythroid tissues and allowing direct comparison to other *in vivo* mouse models of CTCF deletions (Hanssen *et al*., 2017; Harrold *et al*., 2020).

If the locus integrity is confirmed, we have considered various possible interpretations for the ΔCTCF α-globin data. As discussed in the previous chapter (Chapter 3), the predominant insulation at the 3' end of the locus appears to be at the transcribed α-globin genes, hence the 5' CTCF sites in this model will likely be more informative in interpreting data from the ΔCTCF α-globin model. For example, in the 5'CTCF site deletion studies, the ΔHS-38/39 model retained the HS-29 site, and the ΔHS-29 model retained HS-38/39 sites; in the absence of one, the other site may compensate. In other words, the CTCF sites may be functionally redundant. The HS-38/39/29 sites are all proximal to the one of the strongest enhancer elements (R1), and perhaps at least one enhancer-proximal CTCF site is sufficient for correct tethering and maximal activation of α-globin expression. If this was the case, I would predict a ΔHS-38/39/29 model would show decreased α-globin expression. Complementarily, a single

CTCF site (such as HS-38) could be inserted back into the ΔCTCF α-globin locus to determine if expression is restored.

The observation that cohesin appears to redistribute to the next available external CTCF sites (5'HS71 and 3'HS44) further supports what has been previously reported, that CTCF sites act redundantly (Despang *et al*., 2019). This is an aspect of this study that could be further explored; as the external CTCF sites found outside of the RMGR exchanged region were unchanged, there may be insultation and structural role provided by these sites that preserves α-globin expression, albeit insufficiently. Hence, to determine if enhancer-promoter interactions are truly independent of CTCF-anchors, the remaining CTCF sites would have to be deleted, or the α-globin sub-TAD transported away from its genomic locus and tested for its independent ability to self-regulate with or without the CTCF sites examined.

Whilst it was surprising that ΔCTCF would lead to changes in expression, CTCF deletion models in other loci have revealed similar changes in expression. Double and triple deletions of CTCF sites at either the 5' or 3' of the β-globin locus were shown to lead to a decrease in expression of β-globin genes; however, these results were confounded by the sequences deleted across clones being variable (Kang *et al*., 2021). Furthermore, a CTCF site deletion series in the *Sox9–Kcnj2* locus similarly observed a slight but significant reduction in *Sox9* expression only upon deletion of all bordering and intra-TAD CTCF sites, despite gradual fusion of neighbouring TADs at the loss of each CTCF site (Despang *et al*., 2019). Together, this data and the data presented in this chapter, suggest that CTCF sites are not essential for establishment of enhancer-promoter interactions, however function to provide robustness and precision to enhancer-induced gene expression.

How then may this be achieved? TADs are not stable structures in cells, they are instead dynamic and transient. Therefore, the loss of expression observed in the ΔCTCF α-globin model may be due to a reduction in the frequency of or duration of enhancer-promoter interactions. This is supported by live imaging studies which tracked chromatin dynamics upon loss of CTCF binding (Gabriele *et al*., 2022; Mach *et al*., 2022; Platania *et al*., 2023). Both Gabriele *et al*. and Mach *et al*. concluded that CTCF-anchored DNA loops are longer lived than non-anchored loops, whilst Platania *et al*. observed increase in mobility of chromatin upon CTCF site deletion. Altogether, these studies support that CTCF-anchored loops constrain DNA loops and lengthen the time in which associated enhancers and promoters can interact in 3D space. It would be interesting to determine how transcription dynamics are affected in the ΔCTCF α-globin model; this would require measurement of initiation frequency and

burst intensity. To do so, the ΔCTCF α-globin model would need to be engineered further to integrate the machinery required to perform live-imaging of nascent transcription. We have previously implemented such a technique to measure transcriptional dynamics during differentiation (Jeziorska *et al*., 2022). This would provide insight into how a presumably relaxed chromatin structure the ΔCTCF α-globin model affects enhancer-induced transcription. In addition, 3C interaction profiles would confirm changes in interaction frequencies in the ΔCTCF α-globin model; I would predict that the strength of enhancer-promoter interactions would decrease, and the domain itself will likely be enlarged to be anchored at the remaining flanking CTCF sites (HS-71 and HS+44). Whilst Tiled-C was performed (data not shown), similar technical issues as reported in Chapter 3 were experienced, and so these changes remain to be determined.

Overall, the deletion of all CTCF sites within the α-globin sub-TAD leads to overall lower, but not eradicated, enhancer activity at their cognate promoters. This could be due to a decreased potential contact frequency between the enhancers and their targets, due to an increased dynamics and mobility of elements upon removal of all primary tethers.

5.4 Acknowledgements for Chapter 5

Dr Helena Francis generated the RMGR-ready hemizygous mESCs. The synthetic ΔCTCF α-globin BAC sequence was designed by Dr Helena Francis and Dr Leslie Mitchell. Brendan Camellato (Jef Boeke's Lab) synthesised and performed initial sequence verification on the resultant BAC (Jef Boeke's Lab). Susannah Holliman characterised the Δ[θ1/θ2/HS-44/HS-48] (CTCF quad deletion) model from primary erythrocytes.

Chapter 6: General Discussion

6.1 Summary of Findings

In this thesis, I have explored how the order and orientation of regulatory elements can contribute to enhancer-induced activation of a cognate promoter within the context of the α-globin locus. I found that in addition to the well-known CTCF binding elements that may act as insulators, a newly inserted actively transcribing α-globin gene located between the natural enhancers and promoters may also acts as an insulator. Of interest, this newly inserted gene was expressed at higher levels than the more distal native α-globin genes. Together these results showed that not only was there a functional overlap between different classes of regulatory elements, but their linear order along the chromosome also affected their activities. Based on these observations, I investigated how relative orientation of an inserted α-globin gene with respect to its enhancer could influence its level of expression and found transcription was affected. Finally, I confirmed that CTCF sites in the α-globin locus are not essential for the establishment of enhancer-promoter communication, however, CTCF elements do provide robustness to expression likely by limiting enhancer activity to target genes and supporting contact.

6.2 Orientation and position of *cis*-regulatory elements contribute to multifactorial enhancer-promoter communication

These results highlight the multifactorial nature of enhancer-promoter communication, wherein the transcriptional output emerges as a composite outcome of many governing regulatory mechanisms. These factors include, but are not limited to: element accessibility (King *et al*., 2021), genomic distance (Rinzema *et al*., 2022; Zuin *et al*., 2022), biochemical 'compatibility' between elements (Bergman *et al*., 2022; Martinez-Ara *et al*., 2022); tethering factors (Pachano *et al*., 2021); a shared regulatory domain (Symmons *et al*., 2014); insultation by CTCF (Hanssen *et al*., 2017; Narendra *et al*., 2015); and trans-acting transcription factors. Whilst these can be used as guidelines, it is rational to predict that each genomic locus will be regulated by a unique contribution of these factors and in some cases, there may be redundant mechanisms.

With data from the gene inversion presented in Chapter 4 and super-enhancer inversion (Preprint Kassouf *et al*., 2022), the orientation of genomic regulatory elements with respect to each other can now also be included as a contributing factor to the regulation of mammalian loci. However, it is worth

noting that the orientation of a promoter may not always determine enhancer responsiveness – for example, the *de novo* promoter generated by an α-thalassemia-linked SNV does not show differential expression when inverted whilst its insulation strength is affected (Bozhilov *et al*., 2021). This suggests that the effect of the inversion may be promoter-dependant and perhaps is itself a combination of sequence motif grammar, conformation of recruited protein complexes and interaction of transcription with cohesin translocation. To fine-tune these findings, it may therefore be interesting to probe the promoter functional polarity further by inverting for example single motifs within the α-globin promoter. Whilst my data suggests an interplay between cohesin accumulation and transcription, mechanistic understanding of how these machineries interact may be more difficult to disentangle, largely due to our inability to capture the highly processive nature of proteins *in situ.* One such consideration should be that all data presented in this thesis is both static and at a population level. As discussed previously, enhancer-promoter interactions should be considered as dynamic and transient, as shown in super resolution microscopy wherein a 150 kb CTCF-CTCF loop is infrequently fully formed and with a lifetime of ~15 minutes (Gabriele *et al*., 2022; Mach *et al*., 2022). Therefore, all the governing rules mentioned above may be interpreted as contributing to increasing the likelihood and probability of correct enhancer-promoter contact.

The position of elements has also emerged as a parameter that influences enhancer-promoter interactions in my work (Chapter 3) as well as Blayney *et al* (BioRxiv, in press) and in other's work (Rinzema *et al*., 2022; Zuin *et al*., 2022). In Chapter 3, I have shown the ectopic insertion of the α-globin gene between the α-globin enhancer cluster and the endogenous genes lead to enhanced expression of the insert at the expense of expression of the endogenous gene. Similarly, Blayney *et al* identified 'facilitator' elements within the α-globin enhancer cluster that likewise function in a position-dependent manner. The facilitator elements (R3, Rm and R4) present a hierarchical function that is tied to their position with respect to the other *cis*-regulatory elements, with the facilitator in a promoter proximal position being optimal for gene expression. These findings pose a challenge when trying to reconcile them with a 'transcriptional hub' model of transcriptional regulation.

6.3 Challenging the current 'hub' model of enhancer-promoter interaction

Upon establishment of an enhancer-promoter interaction, a 'transcriptional hub' may be formed which may trap RNAPII and other coactivators required for gene expression in a high local concentration. This has been a useful model in combining different observations, such as multiple 'simultaneous' contacts of enhancers (Oudelaar *et al*., 2019) and sub-nuclear 'transcription factories'.

However, this model fails to address why orientation dependent expression is observed, as described in Chapters 3 and 4. For example, if transcription factors are free to diffuse between elements, then the orientation of elements would not be expected to affect transcription as TF binding should be equal in either orientation. Further, with the observation that activation potential of genes decreases 5' to 3' across the locus and that there is an associated change in the profile of cohesin accumulation across the genes, I therefore favour a model involving unidirectional tracking of cohesin (**Figure 6.1**). Even though cohesin does not seem essential for maintaining α-globin expression once established and is thought to act as an initial driving force of domain formation (Prepint Stolper *et al*., 2023) it continues to associate with the chromatin (as evident in ChIP-seq) and presumably continues to translocate throughout the locus. Given the growing evidence in the literature that transcription affects cohesin and domain formation (Busslinger *et al*., 2017; S. Zhang *et al*., 2023) it is feasible that transcription and cohesin translocation may physically interact. As previously described, due to the strong CTCF sites proximal to the enhancer, we can assume that cohesin translocation across the α-globin cluster largely travels uni-directionally, bringing the enhancers to each gene copy progressively. It is possible that at actively transcribed genes, due to the presence of transcription machineries, cohesin translocation is stalled and retained at each gene copy sequentially, reducing the probability of enhancer interaction at more distal genes (**Figure 6.1**). In the case of a gene in the reverse orientation, transcription still occurs and can still act equally as a barrier to extrusion but may itself be disrupted by oncoming cohesin complexes travelling in the opposing direction.

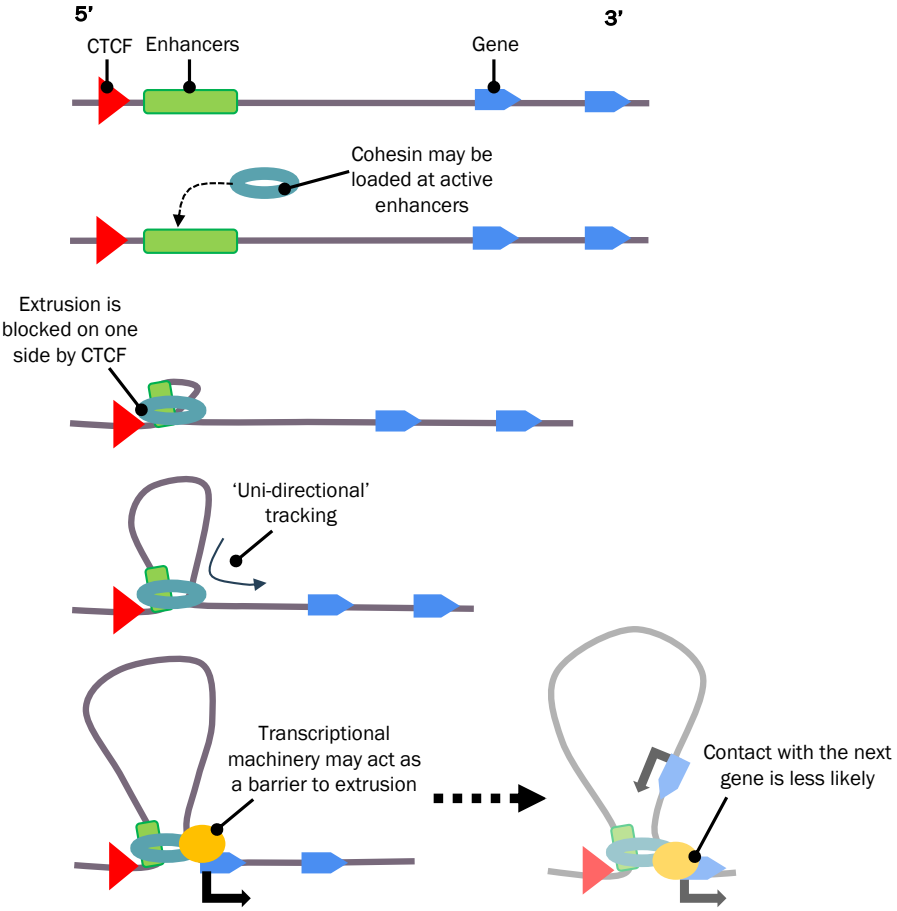

Figure 6.1: Model for how cohesin translocation interacts with transcription across the α-globin locus.

On the other hand, it is difficult to dismiss that there may be a topological restraint on the assembly of Mediator/co-activator complexes when the gene is inverted. In support of this is the observation that 'microcompartments' containing enhancer-promoter pairs in Micro-C mapping are often unaffected by cohesin depletion (Aljahani *et al*., 2022; Goel *et al*., 2023; Hsieh *et al*., 2022), thus may be more dependent on protein-protein interactions, such as between transcription factors or unidentified tethering agents. Once again, due to the processive nature of these complexes, it remains unclear what underlies orientation dependant expression.

6.4 Future work

The studies presented here are limited by the fact that I have only sampled from one timepoint in *in vitro* differentiation. Whilst earlier time points in differentiation could be used to emulate earlier stages in differentiation, it remains unclear whether these timepoints reliably model earlier erythropoiesis *in vivo.* We have previously characterised the mouse α-globin locus at different stages of erythropoiesis from primary erythrocytes (e.g. S0-S5), which uncovered the dynamics of enhancer-

promoter regulation (Oudelaar *et al*., 2020a). Over the course of erythropoiesis, the enhancer elements bind TFs and become progressively accessible whilst the genes remain silenced; in later stages, the formation of the erythroid-specific sub-TAD can be visualised and is concomitant with an increase in α-globin expression. Given our current hypothesis that cohesin-mediated extrusion may be essential in earlier stages of expression and that it may be loaded at enhancers (Preprint Stolper *et al*., 2023), it would be interesting to determine where and when cohesin accumulation occurs at these different stages. Such temporal resolution of domain formation of the ΔCTCF model may help to understand at what stage gene expression is aided by CTCF-cohesin contact. Following verification of the locus, we plan to generate a mouse model from the ΔCTCF mESCs I've generated here. This will no doubt open more avenues of study to determine mechanistically how gene expression is reduced in the absence of CTCF sites. Of course, mapping contact frequencies is a priority to confirm if enhancer-promoter interaction is reduced in the absence of the CTCF sites.

The unfortunate quality of the 3C data generated so far limits my conclusions, as I cannot yet fully determine the changes in interaction frequencies in the models I've generated. Whilst Tiled-C is a powerful technique, with the potential to see genomic architecture at high resolution from multiple viewpoints, it remains limited in resolution to ~1kb, dictated by restriction enzyme digestion of the genome. Of course, the highest resolution 3C methods at the time of writing this thesis is Micro-C and its variants (Hsieh *et al*., 2020, 2015; Hua *et al*., 2021; Krietenstein *et al*., 2020). These methods are yet to be performed successfully using material derived from the *in vitro* differentiation system due to the large number of cells needed for input. Current refinement of these methods and lowering of cell input may allow the examination of fine-scale changes in interaction profiles in our system and will shed light on the dynamics of these interactions in various scenarios.

With regards to the boundary assay (Chapter 3), the decision to use the native α-globin gene sequence as the insert was due to the aim of testing if the α-globin genes could act as insulators or boundary-like elements. Due to the sequence similarities between the elements, it required costly NGS based techniques to determine expression levels of the inserts. Potential development of this assay may benefit from the incorporation of a secondary fluorophore (such as RFP) at the terminus of the inserted gene sequence, similar to the *Hba-a1:mVenus* used as the primary readout of the assay, to allow more rapid quantification of expression from the insert. As discussed in Chapter 3, the insulation effect could be described by two mechanisms: an enhancer blocking activity dependant on cohesin or promoter competition. These two mechanisms may not be mutually exclusive, though it could be considered that promoter competition should be alleviated by using a promoter with lesser

'specificity' for the enhancer. Therefore, a potential direction for this project to further refine the observation and untangle the boundary versus competition argument would be to introduce promoters with lesser specificity for the α-globin enhancers; this could be used to tune the level of expression from the inserts to further correlate transcriptional output to insulation strength whilst simultaneously removing the element of enhancer-promoter competition.

Of course, determining which promoters to use would be an interesting topic in itself as there is evidence from MPSAs to suggest there is specificity between promoter and enhancer sequences (Arnold *et al*., 2017; Bergman *et al*., 2022; Haberle *et al*., 2019; Martinez-Ara *et al*., 2022). However, it remains unclear what drives compatibility between mammalian enhancers and promoters, as variably expressed gene promoters appear to be activated by enhancers agonistic of enhancer identity. Whist informative and capable of testing multiple combinations of enhancer-promoter pairs, these assays are limited in several ways, for example, the assays are often episomal; the sequences assayed are sometimes generated by random genomic sonication and therefore may not be biologically significant; and the 'enhancer-promoter' pairs are often assembled at short distances. These caveats may therefore obscure some of the nuances of compatibility. Whilst not presented in the thesis, I did begin setting up a promoter screen in the context of the α-globin locus, using genome integrated promoters in a fixed position in the locus, in the presence and absence of the super-enhancer. Further development of this assay may shed light on enhancer-promoter specificity in a developmentally relevant context.

6.5 Conclusion

Whilst already characterised in great depth, the α-globin locus continues to be a valuable model in which to dissect governing rules of gene regulation. In this thesis, I have genetically manipulated the locus to rearrange and reposition cis-regulatory elements revealing further how genomic syntax contributes to gene regulation.

Chapter 7: References

Albitar, M., Peschle, C., Liebhaber, S., 1989. Theta, zeta, and epsilon globin messenger RNAs are expressed in adults. Blood 74, 629–637. https://doi.org/10.1182/blood.V74.2.629.629

Aljahani, A., Hua, P., Karpinska, M.A., Quililan, K., Davies, J.O.J., Oudelaar, A.M., 2022. Analysis of sub-kilobase chromatin topology reveals nano-scale regulatory interactions with variable dependence on cohesin and CTCF. Nat Commun 13, 2139. https://doi.org/10.1038/s41467-022-29696-5

Arnold, C.D., Zabidi, M.A., Pagani, M., Rath, M., Schernhuber, K., Kazmar, T., Stark, A., 2017. Genome-wide assessment of sequence-intrinsic enhancer responsiveness at single-base-pair resolution. Nat Biotechnol 35, 136–144. https://doi.org/10.1038/nbt.3739

Badat, M., Davies, J.O.J., Fisher, C.A., Downes, D.J., Rose, A., Glenthøj, A.B., Van Beers, E.J., Harteveld, C.L., Higgs, D.R., 2021. A remarkable case of HbH disease illustrates the relative contributions of the α-globin enhancers to gene expression. Blood 137, 572–575. https://doi.org/10.1182/blood.2020006680

Banerji, J., Olson, L., Schaffner, W., 1983. A lymphocyte-specific cellular enhancer is located downstream of the joining region in immunoglobulin heavy chain genes. Cell 33, 729–740. https://doi.org/10.1016/0092-8674(83)90015-6

Banigan, E.J., Tang, W., Van Den Berg, A.A., Stocsits, R.R., Wutz, G., Brandão, H.B., Busslinger, G.A., Peters, J.-M., Mirny, L.A., 2023. Transcription shapes 3D chromatin organization by interacting with loop extrusion. Proc. Natl. Acad. Sci. U.S.A. 120, e2210480120. https://doi.org/10.1073/pnas.2210480120

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. Nat Commun 10, 2908. https://doi.org/10.1038/s41467-019-10725-9

Barshad, G., Lewis, J.J., Chivu, A.G., Abuhashem, A., Krietenstein, N., Rice, E.J., Ma, Y., Wang, Z., Rando, O.J., Hadjantonakis, A.-K., Danko, C.G., 2023. RNA polymerase II dynamics shape enhancer–promoter interactions. Nat Genet 55, 1370–1380. https://doi.org/10.1038/s41588-023-01442-7

Beagan, J.A., Phillips-Cremins, J.E., 2020. On the existence and functionality of topologically associating domains. Nat Genet 52, 8–16. https://doi.org/10.1038/s41588-019-0561-1

Beagrie, R.A., Scialdone, A., Schueler, M., Kraemer, D.C.A., Chotalia, M., Xie, S.Q., Barbieri, M., De Santiago, I., Lavitas, L.-M., Branco, M.R., Fraser, J., Dostie, J., Game, L., Dillon, N., Edwards, P.A.W., Nicodemi, M., Pombo, A., 2017. Complex multi-enhancer contacts captured by genome architecture mapping. Nature 543, 519–524. https://doi.org/10.1038/nature21411

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

Bergman, D.T., Jones, T.R., Liu, V., Ray, J., Jagoda, E., Siraj, L., Kang, H.Y., Nasser, J., Kane, M., Rios, A., Nguyen, T.H., Grossman, S.R., Fulco, C.P., Lander, E.S., Engreitz, J.M., 2022. Compatibility rules of human enhancer and promoter sequences. Nature 607, 176–184. https://doi.org/10.1038/s41586-022-04877-w

Bintu, B., Mateo, L.J., Su, J.-H., Sinnott-Armstrong, N.A., Parker, M., Kinrot, S., Yamaya, K., Boettiger, A.N., Zhuang, X., 2018. Super-resolution chromatin tracing reveals domains and cooperative interactions in single cells. Science 362, eaau1783. https://doi.org/10.1126/science.aau1783

Blayney, J., Foster, E., Jagielowicz, M., Kreuzer, M., Morotti, M., Reglinski, K., Xiao, J., Hublitz, P., 2020. Unexpectedly High Levels of Inverted Re-Insertions Using Paired sgRNAs for Genomic Deletions. MPs 3, 53. https://doi.org/10.3390/mps3030053

Blayney, J., Francis, H., Camellato, B., Mitchell, L., Stolper, R., Boeke, J., Higgs, D., Kassouf, M., 2022. Super-enhancers require a combination of classical enhancers and novel facilitator elements to drive high levels of gene expression (preprint). Genetics. https://doi.org/10.1101/2022.06.20.496856

Bonev, B., Mendelson Cohen, N., Szabo, Q., Fritsch, L., Papadopoulos, G.L., Lubling, Y., Xu, X., Lv, X., Hugnot, J.-P., Tanay, A., Cavalli, G., 2017. Multiscale 3D Genome Rewiring during Mouse Neural Development. Cell 171, 557-572.e24. https://doi.org/10.1016/j.cell.2017.09.043

Bozhilov, Y.K., Downes, D.J., Telenius, J., Marieke Oudelaar, A., Olivier, E.N., Mountford, J.C., Hughes, J.R., Gibbons, R.J., Higgs, D.R., 2021. A gain-of-function single nucleotide variant creates a new promoter which acts as an orientation-dependent enhancer-blocker. Nat Commun 12, 3806. https://doi.org/10.1038/s41467-021-23980-6

Brandão, H.B., Paul, P., Van Den Berg, A.A., Rudner, D.Z., Wang, X., Mirny, L.A., 2019. RNA polymerases as moving barriers to condensin loop extrusion. Proc. Natl. Acad. Sci. U.S.A. 116, 20489–20499. https://doi.org/10.1073/pnas.1907009116

Brosh, R., Coelho, C., Ribeiro-dos-Santos, A.M., Ellis, G., Hogan, M.S., Ashe, H.J., Somogyi, N., Ordoñez, R., Luther, R.D., Huang, E., Boeke, J.D., Maurano, M.T., 2023. Synthetic regulatory genomics uncovers enhancer context dependence at the Sox2 locus. Molecular Cell 83, 1140-1152.e7. https://doi.org/10.1016/j.molcel.2023.02.027

Brosh, R., Laurent, J.M., Ordoñez, R., Huang, E., Hogan, M.S., Hitchcock, A.M., Mitchell, L.A., Pinglay, S., Cadley, J.A., Luther, R.D., Truong, D.M., Boeke, J.D., Maurano, M.T., 2021. A versatile platform for locus-scale genome rewriting and verification. Proc. Natl. Acad. Sci. U.S.A. 118, e2023952118. https://doi.org/10.1073/pnas.2023952118

Brown, J.M., Roberts, N.A., Graham, B., Waithe, D., Lagerholm, C., Telenius, J.M., De Ornellas, S., Oudelaar, A.M., Scott, C., Szczerbal, I., Babbs, C., Kassouf, M.T., Hughes, J.R., Higgs, D.R., Buckle, V.J., 2018. A tissue-specific self-interacting chromatin domain forms independently of enhancer-promoter interactions. Nat Commun 9, 3849. https://doi.org/10.1038/s41467-018-06248-4

Buenrostro, J.D., Wu, B., Chang, H.Y., Greenleaf, W.J., 2015. ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide. CP Molecular Biology 109. https://doi.org/10.1002/0471142727.mb2129s109

Busslinger, G.A., Stocsits, R.R., Van Der Lelij, P., Axelsson, E., Tedeschi, A., Galjart, N., Peters, J.-M., 2017. Cohesin is positioned in mammalian genomes by transcription, CTCF and Wapl. Nature 544, 503–507. https://doi.org/10.1038/nature22063

Calderon, L., Weiss, F.D., Beagan, J.A., Oliveira, M.S., Georgieva, R., Wang, Y.-F., Carroll, T.S., Dharmalingam, G., Gong, W., Tossell, K., De Paola, V., Whilding, C., Ungless, M.A., Fisher, A.G., Phillips-Cremins, J.E., Merkenschlager, M., 2022. Cohesin-dependence of neuronal gene expression relates to chromatin loop length. eLife 11, e76539. https://doi.org/10.7554/eLife.76539

Chahar, S., Zouari, Y.B., Salari, H., Molitor, A.M., Kobi, D., Maroquenne, M., Erb, C., Mossler, A., Karasu, N., Jost, D., Sexton, T., 2022. Context-dependent transcriptional remodeling of TADs during differentiation (preprint). Genomics. https://doi.org/10.1101/2022.07.01.498405

Chen, Y., Lun, A.T.L., Smyth, G.K., 2016. From reads to genes to pathways: differential expression analysis of RNA-Seq experiments using Rsubread and the edgeR quasi-likelihood pipeline. F1000Res 5, 1438. https://doi.org/10.12688/f1000research.8987.2

Cho, S.W., Xu, J., Sun, R., Mumbach, M.R., Carter, A.C., Chen, Y.G., Yost, K.E., Kim, J., He, J., Nevins, S.A., Chin, S.-F., Caldas, C., Liu, S.J., Horlbeck, M.A., Lim, D.A., Weissman, J.S., Curtis, C., Chang, H.Y., 2018. Promoter of lncRNA Gene PVT1 Is a Tumor-Suppressor DNA Boundary Element. Cell 173, 1398-1412.e22. https://doi.org/10.1016/j.cell.2018.03.068

Ciosk, R., Shirayama, M., Shevchenko, Anna, Tanaka, T., Toth, A., Shevchenko, Andrej, Nasmyth, K., 2000. Cohesin's Binding to Chromosomes Depends on a Separate Complex Consisting of Scc2 and Scc4 Proteins. Molecular Cell 5, 243–254. https://doi.org/10.1016/S1097-2765(00)80420-7

Core, L.J., Waterfall, J.J., Lis, J.T., 2008. Nascent RNA Sequencing Reveals Widespread Pausing and Divergent Initiation at Human Promoters. Science 322, 1845–1848. https://doi.org/10.1126/science.1162228

Cova, G., Glaser, J., Schöpflin, R., Prada-Medina, C.A., Ali, S., Franke, M., Falcone, R., Federer, M., Ponzi, E., Ficarella, R., Novara, F., Wittler, L., Timmermann, B., Gentile, M., Zuffardi, O., Spielmann, M., Mundlos, S., 2023. Combinatorial effects on gene expression at the Lbx1/Fgf8

locus resolve split-hand/foot malformation type 3. Nat Commun 14, 1475. https://doi.org/10.1038/s41467-023-37057-z

Cremer, T., Cremer, C., 2001. Chromosome territories, nuclear architecture and gene regulation in mammalian cells. Nat Rev Genet 2, 292–301. https://doi.org/10.1038/35066075

Cuartero, S., Weiss, F.D., Dharmalingam, G., Guo, Y., Ing-Simmons, E., Masella, S., Robles-Rebollo, I., Xiao, X., Wang, Y.-F., Barozzi, I., Djeghloul, D., Amano, M.T., Niskanen, H., Petretto, E., Dowell, R.D., Tachibana, K., Kaikkonen, M.U., Nasmyth, K.A., Lenhard, B., Natoli, G., Fisher, A.G., Merkenschlager, M., 2018. Control of inducible gene expression links cohesin to hematopoietic progenitor self-renewal and differentiation. Nat Immunol 19, 932–941. https://doi.org/10.1038/s41590-018-0184-1

Davidson, I.F., Bauer, B., Goetz, D., Tang, W., Wutz, G., Peters, J.-M., 2019. DNA loop extrusion by human cohesin. Science 366, 1338–1345. https://doi.org/10.1126/science.aaz3418

Davidson, I.F., Peters, J.-M., 2021. Genome folding through loop extrusion by SMC complexes. Nat Rev Mol Cell Biol 22, 445–464. https://doi.org/10.1038/s41580-021-00349-7

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

De Bruijn, S.E., Fiorentino, A., Ottaviani, D., Fanucchi, S., Melo, U.S., Corral-Serrano, J.C., Mulders, T., Georgiou, M., Rivolta, C., Pontikos, N., Arno, G., Roberts, L., Greenberg, J., Albert, S., Gilissen, C., Aben, M., Rebello, G., Mead, S., Raymond, F.L., Corominas, J., Smith, C.E.L., Kremer, H., Downes, S., Black, G.C., Webster, A.R., Inglehearn, C.F., Van Den Born, L.I., Koenekoop, R.K., Michaelides, M., Ramesar, R.S., Hoyng, C.B., Mundlos, S., Mhlanga, M.M., Cremers, F.P.M., Cheetham, M.E., Roosing, S., Hardcastle, A.J., 2020. Structural Variants Create New Topological-Associated Domains and Ectopic Retinal Enhancer-Gene Contact in Dominant Retinitis Pigmentosa. The American Journal of Human Genetics 107, 802–814. https://doi.org/10.1016/j.ajhg.2020.09.002

De Gobbi, M., 2006. A Regulatory SNP Causes a Human Genetic Disease by Creating a New Transcriptional Promoter. Science 312, 1215–1217. https://doi.org/10.1126/science.1126431

Dekker, J., Rippe, K., Dekker, M., Kleckner, N., 2002. Capturing Chromosome Conformation. Science 295, 1306–1311. https://doi.org/10.1126/science.1067799

Deng, D., Yan, C., Pan, X., Mahfouz, M., Wang, J., Zhu, J.-K., Shi, Y., Yan, N., 2012. Structural Basis for Sequence-Specific Recognition of DNA by TAL Effectors. Science 335, 720–723. https://doi.org/10.1126/science.1215670

Dequeker, B.J.H., Scherr, M.J., Brandão, H.B., Gassler, J., Powell, S., Gaspar, I., Flyamer, I.M., Lalic, A., Tang, W., Stocsits, R., Davidson, I.F., Peters, J.-M., Duderstadt, K.E., Mirny, L.A., Tachibana, K., 2022. MCM complexes are barriers that restrict cohesin-mediated loop extrusion. Nature 606, 197–203. https://doi.org/10.1038/s41586-022-04730-0

Despang, A., Schöpflin, R., Franke, M., Ali, S., Jerković, I., Paliou, C., Chan, W.-L., Timmermann, B., Wittler, L., Vingron, M., Mundlos, S., Ibrahim, D.M., 2019. Functional dissection of the Sox9–Kcnj2 locus identifies nonessential and instructive roles of TAD architecture. Nat Genet 51, 1263–1271. https://doi.org/10.1038/s41588-019-0466-z

de Wit, E., Vos, E.S.M., Holwerda, S.J.B., Valdes-Quezada, C., Verstegen, M.J.A.M., Teunissen, H., Splinter, E., Wijchers, P.J., Krijger, P.H.L., de Laat, W., 2015. CTCF Binding Polarity Determines Chromatin Looping. Molecular Cell 60, 676–684. https://doi.org/10.1016/j.molcel.2015.09.023

Dixon, J.R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., Hu, M., Liu, J.S., Ren, B., 2012. Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature 485, 376–380. https://doi.org/10.1038/nature11082

Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., Gingeras, T.R., 2013. STAR: ultrafast universal RNA-seq aligner. Bioinformatics 29, 15–21. https://doi.org/10.1093/bioinformatics/bts635

Dong, H.Y., Wilkes, S., Yang, H., 2011. CD71 is Selectively and Ubiquitously Expressed at High Levels in Erythroid Precursors of All Maturation Stages: A Comparative Immunochemical

Study With Glycophorin A and Hemoglobin A. American Journal of Surgical Pathology 35, 723–732. https://doi.org/10.1097/PAS.0b013e31821247a8

Dos Santos, M., Backer, S., Auradé, F., Wong, M.M.-K., Wurmser, M., Pierre, R., Langa, F., Do Cruzeiro, M., Schmitt, A., Concordet, J.-P., Sotiropoulos, A., Jeffrey Dilworth, F., Noordermeer, D., Relaix, F., Sakakibara, I., Maire, P., 2022. A fast Myosin super enhancer dictates muscle fiber phenotype through competitive interactions with Myosin genes. Nat Commun 13, 1039. https://doi.org/10.1038/s41467-022-28666-1

Dowen, J.M., Fan, Z.P., Hnisz, D., Ren, G., Abraham, B.J., Zhang, L.N., Weintraub, A.S., Schuijers, J., Lee, T.I., Zhao, K., Young, R.A., 2014. Control of Cell Identity Genes Occurs in Insulated Neighborhoods in Mammalian Chromosomes. Cell 159, 374–387. https://doi.org/10.1016/j.cell.2014.09.030

Downes, D.J., Smith, A.L., Karpinska, M.A., Velychko, T., Rue-Albrecht, K., Sims, D., Milne, T.A., Davies, J.O.J., Oudelaar, A.M., Hughes, J.R., 2022. Capture-C: a modular and flexible approach for high-resolution chromosome conformation capture. Nat Protoc 17, 445–475. https://doi.org/10.1038/s41596-021-00651-w

Faure, A.J., Schmidt, D., Watt, S., Schwalie, P.C., Wilson, M.D., Xu, H., Ramsay, R.G., Odom, D.T., Flicek, P., 2012. Cohesin regulates tissue-specific expression by stabilizing highly occupied *cis* -regulatory modules. Genome Res. 22, 2163–2175. https://doi.org/10.1101/gr.136507.111

Finn, E.H., Pegoraro, G., Brandão, H.B., Valton, A.-L., Oomen, M.E., Dekker, J., Mirny, L., Misteli, T., 2019. Extensive Heterogeneity and Intrinsic Variation in Spatial Genome Organization. Cell 176, 1502-1515.e10. https://doi.org/10.1016/j.cell.2019.01.020

Flavahan, W.A., Drier, Y., Liau, B.B., Gillespie, S.M., Venteicher, A.S., Stemmer-Rachamimov, A.O., Suvà, M.L., Bernstein, B.E., 2016. Insulator dysfunction and oncogene activation in IDH mutant gliomas. Nature 529, 110–114. https://doi.org/10.1038/nature16490

Francis, H., 2019. Functional dissection of a single enhancer at the mouse α-globin locus (DPhil). University of Oxford.

Francis, H.S., Harold, C.L., Beagrie, R.A., King, A.J., Gosden, M.E., Blayney, J.W., Jeziorska, D.M., Babbs, C., Higgs, D.R., Kassouf, M.T., 2022. Scalable in vitro production of defined mouse erythroblasts. PLoS ONE 17, e0261950. https://doi.org/10.1371/journal.pone.0261950

Franke, M., Ibrahim, D.M., Andrey, G., Schwarzer, W., Heinrich, V., Schöpflin, R., Kraft, K., Kempfer, R., Jerković, I., Chan, W.-L., Spielmann, M., Timmermann, B., Wittler, L., Kurth, I., Cambiaso, P., Zuffardi, O., Houge, G., Lambie, L., Brancati, F., Pombo, A., Vingron, M., Spitz, F., Mundlos, S., 2016. Formation of new chromatin domains determines pathogenicity of genomic duplications. Nature 538, 265–269. https://doi.org/10.1038/nature19800

Frith, M.C., Pheasant, M., Mattick, J.S., 2005. Genomics: The amazing complexity of the human transcriptome. Eur J Hum Genet 13, 894–897. https://doi.org/10.1038/sj.ejhg.5201459

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

Furlong, E.E.M., Levine, M., 2018. Developmental enhancers and chromosome topology. Science 361, 1341–1345. https://doi.org/10.1126/science.aau0320

Gabriele, M., Brandão, H.B., Grosse-Holz, S., Jha, A., Dailey, G.M., Cattoglio, C., Hsieh, T.-H.S., Mirny, L., Zechner, C., Hansen, A.S., 2022. Dynamics of CTCF- and cohesin-mediated chromatin looping revealed by live-cell imaging. Science 376, 496–501. https://doi.org/10.1126/science.abn6583

Ganji, M., Shaltiel, I.A., Bisht, S., Kim, E., Kalichava, A., Haering, C.H., Dekker, C., 2018. Real-time imaging of DNA loop extrusion by condensin. Science 360, 102–105. https://doi.org/10.1126/science.aar7831

Georgiades, E., Harrold, C.L., Roberts, N., Kassouf, M., Riva, S.G., Sanders, E., Francis, H.S., Blayney, J., Oudelaar, A.M., Milne, T.A., Higgs, D.R., Hughes, J., 2023. Active regulatory elements recruit cohesin to establish cell-specific chromatin domains (preprint). Genomics. https://doi.org/10.1101/2023.10.13.562171

Glynn, E.F., Megee, P.C., Yu, H.-G., Mistrot, C., Unal, E., Koshland, D.E., DeRisi, J.L., Gerton, J.L., 2004. Genome-Wide Mapping of the Cohesin Complex in the Yeast Saccharomyces cerevisiae. PLoS Biol 2, e259. https://doi.org/10.1371/journal.pbio.0020259

Goel, V.Y., Huseyin, M.K., Hansen, A.S., 2023. Region Capture Micro-C reveals coalescence of enhancers and promoters into nested microcompartments. Nat Genet 55, 1048–1056. https://doi.org/10.1038/s41588-023-01391-1

Guo, Y., Al-Jibury, E., Garcia-Millan, R., Ntagiantas, K., King, J.W.D., Nash, A.J., Galjart, N., Lenhard, B., Rueckert, D., Fisher, A.G., Pruessner, G., Merkenschlager, M., 2022. Chromatin jets define the properties of cohesin-driven in vivo loop extrusion. Molecular Cell 82, 3769-3780.e5. https://doi.org/10.1016/j.molcel.2022.09.003

Guo, Y., Xu, Q., Canzio, D., Shou, J., Li, J., Gorkin, D.U., Jung, I., Wu, H., Zhai, Y., Tang, Y., Lu, Y., Wu, Y., Jia, Z., Li, W., Zhang, M.Q., Ren, B., Krainer, A.R., Maniatis, T., Wu, Q., 2015. CRISPR Inversion of CTCF Sites Alters Genome Topology and Enhancer/Promoter Function. Cell 162, 900–910. https://doi.org/10.1016/j.cell.2015.07.038

Haarhuis, J.H.I., Van Der Weide, R.H., Blomen, V.A., Yáñez-Cuna, J.O., Amendola, M., Van Ruiten, M.S., Krijger, P.H.L., Teunissen, H., Medema, R.H., Van Steensel, B., Brummelkamp, T.R., De Wit, E., Rowland, B.D., 2017. The Cohesin Release Factor WAPL Restricts Chromatin Loop Extension. Cell 169, 693-707.e14. https://doi.org/10.1016/j.cell.2017.04.013

Haberle, V., Arnold, C.D., Pagani, M., Rath, M., Schernhuber, K., Stark, A., 2019. Transcriptional cofactors display specificity for distinct types of core promoters. Nature 570, 122–126. https://doi.org/10.1038/s41586-019-1210-7

Haberle, V., Stark, A., 2018. Eukaryotic core promoters and the functional basis of transcription initiation. Nat Rev Mol Cell Biol 19, 621–637. https://doi.org/10.1038/s41580-018-0028-8

Haering, C.H., Löwe, J., Hochwagen, A., Nasmyth, K., 2002. Molecular Architecture of SMC Proteins and the Yeast Cohesin Complex. Molecular Cell 9, 773–788. https://doi.org/10.1016/S1097-2765(02)00515-4

Hanssen, L.L.P., Kassouf, M.T., Oudelaar, A.M., Biggs, D., Preece, C., Downes, D.J., Gosden, M., Sharpe, J.A., Sloane-Stanley, J.A., Hughes, J.R., Davies, B., Higgs, D.R., 2017. Tissue-specific CTCF–cohesin-mediated chromatin architecture delimits enhancer interactions and function in vivo. Nat Cell Biol 19, 952–961. https://doi.org/10.1038/ncb3573

Hardison, R.C., 2012. Evolution of Hemoglobin and Its Genes. Cold Spring Harbor Perspectives in Medicine 2, a011627–a011627. https://doi.org/10.1101/cshperspect.a011627

Harrold, C.L., Gosden, M.E., Hanssen, L.L.P., Stolper, R.J., Downes, D.J., Telenius, J.M., Biggs, D., Preece, C., Alghadban, S., Sharpe, J.A., Davies, B., Sloane-Stanley, J.A., Kassouf, M.T., Hughes, J.R., Higgs, D.R., 2020. A functional overlap between actively transcribed genes and chromatin boundary elements (preprint). Genomics. https://doi.org/10.1101/2020.07.01.182089

Hay, D., Hughes, J.R., Babbs, C., Davies, J.O.J., Graham, B.J., Hanssen, L.L.P., Kassouf, M.T., Oudelaar, A.M., Sharpe, J.A., Suciu, M.C., Telenius, J., Williams, R., Rode, C., Li, P.-S., Pennacchio, L.A., Sloane-Stanley, J.A., Ayyub, H., Butler, S., Sauka-Spengler, T., Gibbons, R.J., Smith, A.J.H., Wood, W.G., Higgs, D.R., 2016. Genetic dissection of the α-globin super-enhancer in vivo. Nat Genet 48, 895–903. https://doi.org/10.1038/ng.3605

Heinz, S., Texari, L., Hayes, M.G.B., Urbanowski, M., Chang, M.W., Givarkes, N., Rialdi, A., White, K.M., Albrecht, R.A., Pache, L., Marazzi, I., García-Sastre, A., Shaw, M.L., Benner, C., 2018. Transcription Elongation Can Affect Genome 3D Structure. Cell 174, 1522-1536.e22. https://doi.org/10.1016/j.cell.2018.07.047

Hildebrand, E.M., Dekker, J., 2020. Mechanisms and Functions of Chromosome Compartmentalization. Trends in Biochemical Sciences 45, 385–396. https://doi.org/10.1016/j.tibs.2020.01.002

Hnisz, D., Weintraub, A.S., Day, D.S., Valton, A.-L., Bak, R.O., Li, C.H., Goldmann, J., Lajoie, B.R., Fan, Z.P., Sigova, A.A., Reddy, J., Borges-Rivera, D., Lee, T.I., Jaenisch, R., Porteus, M.H., Dekker, J., Young, R.A., 2016. Activation of proto-oncogenes by disruption of chromosome neighborhoods. Science 351, 1454–1458. https://doi.org/10.1126/science.aad9024

Hsieh, T.-H.S., Cattoglio, C., Slobodyanyuk, E., Hansen, A.S., Darzacq, X., Tjian, R., 2022. Enhancer–promoter interactions and transcription are largely maintained upon acute loss of CTCF, cohesin, WAPL or YY1. Nat Genet 54, 1919–1932. https://doi.org/10.1038/s41588-022-01223-8

Hsieh, T.-H.S., Cattoglio, C., Slobodyanyuk, E., Hansen, A.S., Rando, O.J., Tjian, R., Darzacq, X., 2020. Resolving the 3D Landscape of Transcription-Linked Mammalian Chromatin Folding. Molecular Cell 78, 539-553.e8. https://doi.org/10.1016/j.molcel.2020.03.002

Hsieh, T.-H.S., Weiner, A., Lajoie, B., Dekker, J., Friedman, N., Rando, O.J., 2015. Mapping Nucleosome Resolution Chromosome Folding in Yeast by Micro-C. Cell 162, 108–119. https://doi.org/10.1016/j.cell.2015.05.048

Hua, P., Badat, M., Hanssen, L.L.P., Hentges, L.D., Crump, N., Downes, D.J., Jeziorska, D.M., Oudelaar, A.M., Schwessinger, R., Taylor, S., Milne, T.A., Hughes, J.R., Higgs, D.R., Davies, J.O.J., 2021. Defining genome architecture at base-pair resolution. Nature 595, 125–129. https://doi.org/10.1038/s41586-021-03639-4

Hughes, J.R., Cheng, J.-F., Ventress, N., Prabhakar, S., Clark, K., Anguita, E., De Gobbi, M., de Jong, P., Rubin, E., Higgs, D.R., 2005. Annotation of cis-regulatory elements by identification, subclassification, and functional assessment of multispecies conserved sequences. Proceedings of the National Academy of Sciences 102, 9830–9835. https://doi.org/10.1073/pnas.0503401102

Hughes, J.R., Roberts, N., McGowan, S., Hay, D., Giannoulatou, E., Lynch, M., De Gobbi, M., Taylor, S., Gibbons, R., Higgs, D.R., 2014. Analysis of hundreds of cis-regulatory landscapes at high resolution in a single, high-throughput experiment. Nat Genet 46, 205–212. https://doi.org/10.1038/ng.2871

Hyle, J., Zhang, Y., Wright, S., Xu, B., Shao, Y., Easton, J., Tian, L., Feng, R., Xu, P., Li, C., 2019. Acute depletion of CTCF directly affects MYC regulation through loss of enhancer–promoter looping. Nucleic Acids Research 47, 6699–6713. https://doi.org/10.1093/nar/gkz462

Jeppsson, K., Sakata, T., Nakato, R., Milanova, S., Shirahige, K., Björkegren, C., 2022. Cohesin-dependent chromosome loop extrusion is limited by transcription and stalled replication forks. Sci. Adv. 8, eabn7063. https://doi.org/10.1126/sciadv.abn7063

Jeziorska, D.M., Tunnacliffe, E.A.J., Brown, J.M., Ayyub, H., Sloane-Stanley, J., Sharpe, J.A., Lagerholm, B.C., Babbs, C., Smith, A.J.H., Buckle, V.J., Higgs, D.R., 2022. On-microscope staging of live cells reveals changes in the dynamics of transcriptional bursting during differentiation. Nat Commun 13, 6641. https://doi.org/10.1038/s41467-022-33977-4

Jiang, Y., Huang, J., Lun, K., Li, B., Zheng, H., Li, Y., Zhou, R., Duan, W., Wang, C., Feng, Y., Yao, H., Li, C., Ji, X., 2020. Genome-wide analyses of chromatin interactions after the loss of Pol I, Pol II, and Pol III. Genome Biol 21, 158. https://doi.org/10.1186/s13059-020-02067-3

Kagey, M.H., Newman, J.J., Bilodeau, S., Zhan, Y., Orlando, D.A., Van Berkum, N.L., Ebmeier, C.C., Goossens, J., Rahl, P.B., Levine, S.S., Taatjes, D.J., Dekker, J., Young, R.A., 2010. Mediator and cohesin connect gene expression and chromatin architecture. Nature 467, 430–435. https://doi.org/10.1038/nature09380

Kane, L., Williamson, I., Flyamer, I.M., Kumar, Y., Hill, R.E., Lettice, L.A., Bickmore, W.A., 2022. Cohesin is required for long-range enhancer action at the Shh locus. Nat Struct Mol Biol 29, 891–897. https://doi.org/10.1038/s41594-022-00821-8

Kang, J., Kim, Y.W., Park, S., Kang, Y., Kim, A., 2021. Multiple CTCF sites cooperate with each other to maintain a TAD for enhancer–promoter interaction in the β-globin locus. FASEB j. 35. https://doi.org/10.1096/fj.202100105RR

Kassouf, M.T., Francis, H.S., Gosden, M., Suciu, M.C., Downes, D.J., Harrold, C., Larke, M., Oudelaar, M., Cornell, L., Blayney, J., Telenius, J., Xella, B., Shen, Y., Sousos, N., Sharpe, J.A., Sloane-Stanley, J., Smith, A., Babbs, C., Hughes, J.R., Higgs, D.R., 2022. Multipartite super-enhancers function in an orientation-dependent manner (preprint). Molecular Biology. https://doi.org/10.1101/2022.07.14.499999

Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringle, T.H., Zahler, A.M., Haussler, A.D., 2002. The Human Genome Browser at UCSC. Genome Res. 12, 996–1006. https://doi.org/10.1101/gr.229102

Kikuta, H., Laplante, M., Navratilova, P., Komisarczuk, A.Z., Engström, P.G., Fredman, D., Akalin, A., Caccamo, M., Sealy, I., Howe, K., Ghislain, J., Pezeron, G., Mourrain, P., Ellingsen, S., Oates, A.C., Thisse, C., Thisse, B., Foucher, I., Adolf, B., Geling, A., Lenhard, B., Becker, T.S., 2007. Genomic regulatory blocks encompass multiple neighboring genes and maintain conserved synteny in vertebrates. Genome Res. 17, 545–555. https://doi.org/10.1101/gr.6086307

Kim, Y., Shi, Z., Zhang, H., Finkelstein, I.J., Yu, H., 2019. Human cohesin compacts DNA by loop extrusion. Science 366, 1345–1349. https://doi.org/10.1126/science.aaz4475

King, A.J., Songdej, D., Downes, D.J., Beagrie, R.A., Liu, S., Buckley, M., Hua, P., Suciu, M.C., Marieke Oudelaar, A., Hanssen, L.L.P., Jeziorska, D., Roberts, N., Carpenter, S.J., Francis, H., Telenius, J., Olijnik, A.-A., Sharpe, J.A., Sloane-Stanley, J., Eglinton, J., Kassouf, M.T., Orkin, S.H., Pennacchio, L.A., Davies, J.O.J., Hughes, J.R., Higgs, D.R., Babbs, C., 2021. Reactivation of a developmentally silenced embryonic globin gene. Nat Commun 12, 4439. https://doi.org/10.1038/s41467-021-24402-3

Konermann, S., Brigham, M.D., Trevino, A.E., Joung, J., Abudayyeh, O.O., Barcena, C., Hsu, P.D., Habib, N., Gootenberg, J.S., Nishimasu, H., Nureki, O., Zhang, F., 2015. Genome-scale transcriptional activation by an engineered CRISPR-Cas9 complex. Nature 517, 583–588. https://doi.org/10.1038/nature14136

Kovina, A.P., Petrova, N.V., Komkov, D.S., Dashinimaev, E.B., Razin, S.V., 2022. Regulatory systems of chicken alpha-globin gene domain suppress bidirectional transcription. Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms 1865, 194850. https://doi.org/10.1016/j.bbagrm.2022.194850

Kowalczyk, M.S., Hughes, J.R., Garrick, D., Lynch, M.D., Sharpe, J.A., Sloane-Stanley, J.A., McGowan, S.J., De Gobbi, M., Hosseini, M., Vernimmen, D., Brown, J.M., Gray, N.E., Collavin, L., Gibbons, R.J., Flint, J., Taylor, S., Buckle, V.J., Milne, T.A., Wood, W.G., Higgs, D.R., 2012. Intragenic Enhancers Act as Alternative Promoters. Molecular Cell 45, 447–458. https://doi.org/10.1016/j.molcel.2011.12.021

Kramer, N.E., Davis, E.S., Wenger, C.D., Deoudes, E.M., Parker, S.M., Love, M.I., Phanstiel, D.H., 2022. Plotgardener: cultivating precise multi-panel figures in R. Bioinformatics 38, 2042–2045. https://doi.org/10.1093/bioinformatics/btac057

Krietenstein, N., Abraham, S., Venev, S.V., Abdennur, N., Gibcus, J., Hsieh, T.-H.S., Parsi, K.M., Yang, L., Maehr, R., Mirny, L.A., Dekker, J., Rando, O.J., 2020. Ultrastructural Details of Mammalian Chromosome Architecture. Molecular Cell 78, 554-565.e7. https://doi.org/10.1016/j.molcel.2020.03.003

Kueng, S., Hegemann, B., Peters, B.H., Lipp, J.J., Schleiffer, A., Mechtler, K., Peters, J.-M., 2006. Wapl Controls the Dynamic Association of Cohesin with Chromatin. Cell 127, 955–967. https://doi.org/10.1016/j.cell.2006.09.040

Langmead, B., Salzberg, S.L., 2012. Fast gapped-read alignment with Bowtie 2. Nat Methods 9, 357–359. https://doi.org/10.1038/nmeth.1923

Larke, M.S.C., Schwessinger, R., Nojima, T., Telenius, J., Beagrie, R.A., Downes, D.J., Oudelaar, A.M., Truch, J., Graham, B., Bender, M.A., Proudfoot, N.J., Higgs, D.R., Hughes, J.R., 2021. Enhancers predominantly regulate gene expression during differentiation via transcription initiation. Molecular Cell 81, 983-997.e7. https://doi.org/10.1016/j.molcel.2021.01.002

Lawrence, M., Huber, W., Pagès, H., Aboyoun, P., Carlson, M., Gentleman, R., Morgan, M.T., Carey, V.J., 2013. Software for Computing and Annotating Genomic Ranges. PLoS Comput Biol 9, e1003118. https://doi.org/10.1371/journal.pcbi.1003118

Lee, H.K., Willi, M., Wang, C., Yang, C.M., Smith, H.E., Liu, C., Hennighausen, L., 2017. Functional assessment of CTCF sites at cytokine-sensing mammary enhancers using CRISPR/Cas9 gene editing in mice. Nucleic Acids Research 45, 4606–4618. https://doi.org/10.1093/nar/gkx185

Lengronne, A., Katou, Y., Mori, S., Yokobayashi, S., Kelly, G.P., Itoh, T., Watanabe, Y., Shirahige, K., Uhlmann, F., 2004. Cohesin relocation from sites of chromosomal loading to places of convergent transcription. Nature 430, 573–578. https://doi.org/10.1038/nature02742

Lettice, L.A., 2003. A long-range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly. Human Molecular Genetics 12, 1725–1735. https://doi.org/10.1093/hmg/ddg180

Li, H., 2011a. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics 27, 2987–2993. https://doi.org/10.1093/bioinformatics/btr509

Li, H., 2011b. Improving SNP discovery by base alignment quality. Bioinformatics 27, 1157–1158. https://doi.org/10.1093/bioinformatics/btr076

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., 1000 Genome Project Data Processing Subgroup, 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25, 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Liao, Y., Smyth, G.K., Shi, W., 2014. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics 30, 923–930. https://doi.org/10.1093/bioinformatics/btt656

Lieberman-Aiden, E., Van Berkum, N.L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B.R., Sabo, P.J., Dorschner, M.O., Sandstrom, R., Bernstein, B., Bender, M.A., Groudine, M., Gnirke, A., Stamatoyannopoulos, J., Mirny, L.A., Lander, E.S., Dekker, J., 2009. Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome. Science 326, 289–293. https://doi.org/10.1126/science.1181369

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

Lupiáñez, D.G., Kraft, K., Heinrich, V., Krawitz, P., Brancati, F., Klopocki, E., Horn, D., Kayserili, H., Opitz, J.M., Laxova, R., Santos-Simarro, F., Gilbert-Dussardier, B., Wittler, L., Borschiwer, M., Haas, S.A., Osterwalder, M., Franke, M., Timmermann, B., Hecht, J., Spielmann, M., Visel, A., Mundlos, S., 2015. Disruptions of Topological Chromatin Domains Cause Pathogenic Rewiring of Gene-Enhancer Interactions. Cell 161, 1012–1025. https://doi.org/10.1016/j.cell.2015.04.004

Mach, P., Kos, P.I., Zhan, Y., Cramard, J., Gaudin, S., Tünnermann, J., Marchi, E., Eglinger, J., Zuin, J., Kryzhanovska, M., Smallwood, S., Gelman, L., Roth, G., Nora, E.P., Tiana, G., Giorgetti, L., 2022. Cohesin and CTCF control the dynamics of chromosome folding. Nat Genet 54, 1907–1918. https://doi.org/10.1038/s41588-022-01232-7

Martinez-Ara, M., Comoglio, F., van Arensbergen, J., van Steensel, B., 2022. Systematic analysis of intrinsic enhancer-promoter compatibility in the mouse genome. Molecular Cell 82, 2519-2531.e6. https://doi.org/10.1016/j.molcel.2022.04.009

Mercola, M., Wang, X.-F., Olsen, J., Calame, K., 1983. Transcriptional Enhancer Elements in the Mouse Immunoglobulin Heavy Chain Locus. Science 221, 663–665. https://doi.org/10.1126/science.6306772

Mitchell, L.A., McCulloch, L.H., Pinglay, S., Berger, H., Bosco, N., Brosh, R., Bulajić, M., Huang, E., Hogan, M.S., Martin, J.A., Mazzoni, E.O., Davoli, T., Maurano, M.T., Boeke, J.D., 2021. *De novo* assembly and delivery to mouse cells of a 101 kb functional human gene. Genetics 218, iyab038. https://doi.org/10.1093/genetics/iyab038

Nagano, T., Lubling, Y., Stevens, T.J., Schoenfelder, S., Yaffe, E., Dean, W., Laue, E.D., Tanay, A., Fraser, P., 2013. Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. Nature 502, 59–64. https://doi.org/10.1038/nature12593

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

Nasmyth, K., Haering, C.H., 2009. Cohesin: Its Roles and Mechanisms. Annu. Rev. Genet. 43, 525–558. https://doi.org/10.1146/annurev-genet-102108-134233

Nichols, J., Evans, E.P., Smith, A.G., 1990. Establishment of germ-line-competent embryonic stem (ES) cells using differentiation inhibiting activity. Development 110, 1341–1348. https://doi.org/10.1242/dev.110.4.1341

Nora, E.P., Goloborodko, A., Valton, A.-L., Gibcus, J.H., Uebersohn, A., Abdennur, N., Dekker, J., Mirny, L.A., Bruneau, B.G., 2017. Targeted Degradation of CTCF Decouples Local Insulation of Chromosome Domains from Genomic Compartmentalization. Cell 169, 930-944.e22. https://doi.org/10.1016/j.cell.2017.05.004

Nora, E.P., Lajoie, B.R., Schulz, E.G., Giorgetti, L., Okamoto, I., Servant, N., Piolot, T., Van Berkum, N.L., Meisig, J., Sedat, J., Gribnau, J., Barillot, E., Blüthgen, N., Dekker, J., Heard, E., 2012. Spatial partitioning of the regulatory landscape of the X-inactivation centre. Nature 485, 381–385. https://doi.org/10.1038/nature11049

Nuebler, J., Fudenberg, G., Imakaev, M., Abdennur, N., Mirny, L.A., 2018. Chromatin organization by an interplay of loop extrusion and compartmental segregation. Proc. Natl. Acad. Sci. U.S.A. 115. https://doi.org/10.1073/pnas.1717730115

Oudelaar, A.M., Beagrie, R.A., Gosden, M., De Ornellas, S., Georgiades, E., Kerry, J., Hidalgo, D., Carrelha, J., Shivalingam, A., El-Sagheer, A.H., Telenius, J.M., Brown, T., Buckle, V.J., Socolovsky, M., Higgs, D.R., Hughes, J.R., 2020a. Dynamics of the 4D genome during in vivo lineage specification and differentiation. Nat Commun 11, 2722. https://doi.org/10.1038/s41467-020-16598-7

Oudelaar, A.M., Beagrie, R.A., Kassouf, M.T., Higgs, D.R., 2021. The mouse alpha-globin cluster: a paradigm for studying genome regulation and organization. Current Opinion in Genetics & Development 67, 18–24. https://doi.org/10.1016/j.gde.2020.10.003

Oudelaar, A.M., Harrold, C.L., Hanssen, L.L.P., Telenius, J.M., Higgs, D.R., Hughes, J.R., 2019. A revised model for promoter competition based on multi-way chromatin interactions at the α-globin locus. Nat Commun 10, 5412. https://doi.org/10.1038/s41467-019-13404-x

Pachano, T., Sánchez-Gaya, V., Ealo, T., Mariner-Faulí, M., Bleckwehl, T., Asenjo, H.G., Respuela, P., Cruz-Molina, S., Muñoz-San Martín, M., Haro, E., Van IJcken, W.F.J., Landeira, D., Rada-Iglesias, A., 2021. Orphan CpG islands amplify poised enhancer regulatory activity and determine target gene responsiveness. Nat Genet 53, 1036–1049. https://doi.org/10.1038/s41588-021-00888-x

Pagès, H., 2017. BSgenome. https://doi.org/10.18129/B9.BIOC.BSGENOME

Palis, J., 2014. Primitive and definitive erythropoiesis in mammals. Front. Physiol. 5. https://doi.org/10.3389/fphys.2014.00003

Papadopoulos, P., Gutiérrez, L., Van Der Linden, R., Kong-A-San, J., Maas, A., Drabek, D., Patrinos, G.P., Philipsen, S., Grosveld, F., 2012. A Dual Reporter Mouse Model of the Human β-Globin Locus: Applications and Limitations. PLoS ONE 7, e51272. https://doi.org/10.1371/journal.pone.0051272

Parelho, V., Hadjur, S., Spivakov, M., Leleu, M., Sauer, S., Gregson, H.C., Jarmuz, A., Canzonetta, C., Webster, Z., Nesterova, T., Cobb, B.S., Yokomori, K., Dillon, N., Aragon, L., Fisher, A.G., Merkenschlager, M., 2008. Cohesins Functionally Associate with CTCF on Mammalian Chromosome Arms. Cell 132, 422–433. https://doi.org/10.1016/j.cell.2008.01.011

Pennacchio, L.A., Ahituv, N., Moses, A.M., Prabhakar, S., Nobrega, M.A., Shoukry, M., Minovitsky, S., Dubchak, I., Holt, A., Lewis, K.D., Plajzer-Frick, I., Akiyama, J., De Val, S., Afzal, V., Black, B.L., Couronne, O., Eisen, M.B., Visel, A., Rubin, E.M., 2006. In vivo enhancer analysis of human conserved non-coding sequences. Nature 444, 499–502. https://doi.org/10.1038/nature05295

Peric-Hupkes, D., Meuleman, W., Pagie, L., Bruggeman, S.W.M., Solovei, I., Brugman, W., Gräf, S., Flicek, P., Kerkhoven, R.M., Van Lohuizen, M., Reinders, M., Wessels, L., Van Steensel, B., 2010. Molecular Maps of the Reorganization of Genome-Nuclear Lamina Interactions during Differentiation. Molecular Cell 38, 603–613. https://doi.org/10.1016/j.molcel.2010.03.016

Peschle, C., Mavilio, F., Carè, A., Migliaccio, G., Migliaccio, A.R., Salvo, G., Samoggia, P., Petti, S., Guerriero, R., Marinucci, M., Lazzaro, D., Russo, G., Mastroberardino, G., 1985. Haemoglobin switching in human embryos: asynchrony of ζ → α and ε → γ-globin switches in primitive and definitive erythropoietic lineage. Nature 313, 235–238. https://doi.org/10.1038/313235a0

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

Phillips-Cremins, J.E., Sauria, M.E.G., Sanyal, A., Gerasimova, T.I., Lajoie, B.R., Bell, J.S.K., Ong, C.-T., Hookway, T.A., Guo, C., Sun, Y., Bland, M.J., Wagstaff, W., Dalton, S., McDevitt, T.C., Sen, R., Dekker, J., Taylor, J., Corces, V.G., 2013. Architectural Protein Subclasses Shape 3D Organization of Genomes during Lineage Commitment. Cell 153, 1281–1295. https://doi.org/10.1016/j.cell.2013.04.053

Pinglay, S., Bulajić, M., Rahe, D.P., Huang, E., Brosh, R., Mamrak, N.E., King, B.R., German, S., Cadley, J.A., Rieber, L., Easo, N., Lionnet, T., Mahony, S., Maurano, M.T., Holt, L.J., Mazzoni, E.O., Boeke, J.D., 2022. Synthetic regulatory reconstitution reveals principles of mammalian *Hox* cluster regulation. Science 377, eabk2820. https://doi.org/10.1126/science.abk2820

Platania, A., Erb, C., Barbieri, M., Molcrette, B., Grandgirard, E., De Kort, M.A., Meaburn, K., Taylor, T., Shchuka, V.M., Kocanova, S., Monteiro Oliveira, G., Mitchell, J.A., Soutoglou, E., Lenstra, T.L., Molina, N., Papantonis, A., Bystricky, K., Sexton, T., 2023. Competition between transcription and loop extrusion modulates promoter and enhancer dynamics (preprint). Cell Biology. https://doi.org/10.1101/2023.04.25.538222

Preker, P., Nielsen, J., Kammler, S., Lykke-Andersen, S., Christensen, M.S., Mapendano, C.K., Schierup, M.H., Jensen, T.H., 2008. RNA Exosome Depletion Reveals Transcription Upstream of Active Human Promoters. Science 322, 1851–1854. https://doi.org/10.1126/science.1164096

Quinlan, A.R., Hall, I.M., 2010. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26, 841–842. https://doi.org/10.1093/bioinformatics/btq033

Quinodoz, S.A., Ollikainen, N., Tabak, B., Palla, A., Schmidt, J.M., Detmar, E., Lai, M.M., Shishkin, A.A., Bhat, P., Takei, Y., Trinh, V., Aznauryan, E., Russell, P., Cheng, C., Jovanovic, M., Chow, A., Cai, L., McDonel, P., Garber, M., Guttman, M., 2018. Higher-Order Inter-chromosomal Hubs Shape 3D Genome Organization in the Nucleus. Cell 174, 744-757.e24. https://doi.org/10.1016/j.cell.2018.05.024

Ramasamy, S., Aljahani, A., Karpinska, M.A., Cao, T.B.N., Velychko, T., Cruz, J.N., Lidschreiber, M., Oudelaar, A.M., 2023. The Mediator complex regulates enhancer-promoter interactions. Nat Struct Mol Biol 30, 991–1000. https://doi.org/10.1038/s41594-023-01027-2

Ramírez, F., Bhardwaj, V., Arrigoni, L., Lam, K.C., Grüning, B.A., Villaveces, J., Habermann, B., Akhtar, A., Manke, T., 2018. High-resolution TADs reveal DNA sequences underlying genome organization in flies. Nat Commun 9, 189. https://doi.org/10.1038/s41467-017-02525-w

Ramírez, F., Ryan, D.P., Grüning, B., Bhardwaj, V., Kilpert, F., Richter, A.S., Heyne, S., Dündar, F., Manke, T., 2016. deepTools2: a next generation web server for deep-sequencing data analysis. Nucleic Acids Res 44, W160–W165. https://doi.org/10.1093/nar/gkw257

Raney, B.J., Dreszer, T.R., Barber, G.P., Clawson, H., Fujita, P.A., Wang, T., Nguyen, N., Paten, B., Zweig, A.S., Karolchik, D., Kent, W.J., 2014. Track data hubs enable visualization of user-defined genome-wide annotations on the UCSC Genome Browser. Bioinformatics 30, 1003–1005. https://doi.org/10.1093/bioinformatics/btt637

Rao, S.S.P., Huang, S.-C., Glenn St Hilaire, B., Engreitz, J.M., Perez, E.M., Kieffer-Kwon, K.-R., Sanborn, A.L., Johnstone, S.E., Bascom, G.D., Bochkov, I.D., Huang, X., Shamim, M.S., Shin, J., Turner, D., Ye, Z., Omer, A.D., Robinson, J.T., Schlick, T., Bernstein, B.E., Casellas, R., Lander, E.S., Aiden, E.L., 2017. Cohesin Loss Eliminates All Loop Domains. Cell 171, 305-320.e24. https://doi.org/10.1016/j.cell.2017.09.026

Rao, S.S.P., Huntley, M.H., Durand, N.C., Stamenova, E.K., Bochkov, I.D., Robinson, J.T., Sanborn, A.L., Machol, I., Omer, A.D., Lander, E.S., Aiden, E.L., 2014. A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping. Cell 159, 1665–1680. https://doi.org/10.1016/j.cell.2014.11.021

Redolfi, J., Zhan, Y., Valdes-Quezada, C., Kryzhanovska, M., Guerreiro, I., Iesmantavicius, V., Pollex, T., Grand, R.S., Mulugeta, E., Kind, J., Tiana, G., Smallwood, S.A., De Laat, W., Giorgetti, L., 2019. DamC reveals principles of chromatin folding in vivo without crosslinking and ligation. Nat Struct Mol Biol 26, 471–480. https://doi.org/10.1038/s41594-019-0231-0

Rhodes, J., Mazza, D., Nasmyth, K., Uphoff, S., 2017. Scc2/Nipbl hops between chromosomal cohesin rings after loading. eLife 6, e30000. https://doi.org/10.7554/eLife.30000

Richter, W.F., Nayak, S., Iwasa, J., Taatjes, D.J., 2022. The Mediator complex as a master regulator of transcription by RNA polymerase II. Nat Rev Mol Cell Biol 23, 732–749. https://doi.org/10.1038/s41580-022-00498-3

Rinzema, N.J., Sofiadis, K., Tjalsma, S.J.D., Verstegen, M.J.A.M., Oz, Y., Valdes-Quezada, C., Felder, A.-K., Filipovska, T., Van Der Elst, S., De Andrade Dos Ramos, Z., Han, R., Krijger, P.H.L., De Laat, W., 2022. Building regulatory landscapes reveals that an enhancer can recruit cohesin to create contact domains, engage CTCF sites and activate distant genes. Nat Struct Mol Biol 29, 563–574. https://doi.org/10.1038/s41594-022-00787-7

Rubio, E.D., Reiss, D.J., Welcsh, P.L., Disteche, C.M., Filippova, G.N., Baliga, N.S., Aebersold, R., Ranish, J.A., Krumm, A., 2008. CTCF physically links cohesin to chromatin. Proc. Natl. Acad. Sci. U.S.A. 105, 8309–8314. https://doi.org/10.1073/pnas.0801273105

Sanborn, A.L., Rao, S.S.P., Huang, S.-C., Durand, N.C., Huntley, M.H., Jewett, A.I., Bochkov, I.D., Chinnappan, D., Cutkosky, A., Li, J., Geeting, K.P., Gnirke, A., Melnikov, A., McKenna, D., Stamenova, E.K., Lander, E.S., Aiden, E.L., 2015. Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes. Proc Natl Acad Sci USA 112, E6456–E6465. https://doi.org/10.1073/pnas.1518552112

Schmitt, A.D., Hu, M., Jung, I., Xu, Z., Qiu, Y., Tan, C.L., Li, Y., Lin, S., Lin, Y., Barr, C.L., Ren, B., 2016. A Compendium of Chromatin Contact Maps Reveals Spatially Active Regions in the Human Genome. Cell Reports 17, 2042–2059. https://doi.org/10.1016/j.celrep.2016.10.061

Schwarzer, W., Abdennur, N., Goloborodko, A., Pekowska, A., Fudenberg, G., Loe-Mie, Y., Fonseca, N.A., Huber, W., Haering, C.H., Mirny, L., Spitz, F., 2017. Two independent modes of chromatin organization revealed by cohesin removal. Nature 551, 51–56. https://doi.org/10.1038/nature24281

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

Seila, A.C., Calabrese, J.M., Levine, S.S., Yeo, G.W., Rahl, P.B., Flynn, R.A., Young, R.A., Sharp, P.A., 2008. Divergent Transcription from Active Promoters. Science 322, 1849–1851. https://doi.org/10.1126/science.1162253

Servant, N., Varoquaux, N., Lajoie, B.R., Viara, E., Chen, C.-J., Vert, J.-P., Heard, E., Dekker, J., Barillot, E., 2015. HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. Genome Biol 16, 259. https://doi.org/10.1186/s13059-015-0831-x

Sexton, T., Yaffe, E., Kenigsberg, E., Bantignies, F., Leblanc, B., Hoichman, M., Parrinello, H., Tanay, A., Cavalli, G., 2012. Three-Dimensional Folding and Functional Organization Principles of the Drosophila Genome. Cell 148, 458–472. https://doi.org/10.1016/j.cell.2012.01.010

Smith, A.G., 1991. Culture and differentiation of embryonic stem cells. Journal of Tissue Culture Methods 13, 89–94. https://doi.org/10.1007/BF01666137

Smith, A.J.H., Xian, J., Richardson, M., Johnstone, K.A., Rabbitts, P.H., 2002. Cre-loxP chromosome engineering of a targeted deletion in the mouse corresponding to the 3p21.3 region of homozygous loss in human tumours. Oncogene 21, 4521–4529. https://doi.org/10.1038/sj.onc.1205530

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

Stevens, T.J., Lando, D., Basu, S., Atkinson, L.P., Cao, Y., Lee, S.F., Leeb, M., Wohlfahrt, K.J., Boucher, W., O'Shaughnessy-Kirwan, A., Cramard, J., Faure, A.J., Ralser, M., Blanco, E., Morey, L., Sansó, M., Palayret, M.G.S., Lehner, B., Di Croce, L., Wutz, A., Hendrich, B., Klenerman, D., Laue, E.D., 2017. 3D structures of individual mammalian genomes studied by single-cell Hi-C. Nature 544, 59–64. https://doi.org/10.1038/nature21429

Stolper, R.J., Tsang, F.H., Georgiades, E., Hansen, L.L.P., Downes, D.J., Harrold, C.L., Hughes, J.R., Beagrie, R.A., Davies, B., Kassouf, M.T., Higgs, D.R., 2023. Loop extrusion by cohesin plays a key role in enhancer-activated gene expression during differentiation (preprint). Molecular Biology. https://doi.org/10.1101/2023.09.07.556660

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

Symmons, O., Uslu, V.V., Tsujimura, T., Ruf, S., Nassari, S., Schwarzer, W., Ettwiller, L., Spitz, F., 2014. Functional and topological characteristics of mammalian regulatory domains. Genome Res. 24, 390–400. https://doi.org/10.1101/gr.163519.113

Tallack, M.R., Whitington, T., Shan Yuen, W., Wainwright, E.N., Keys, J.R., Gardiner, B.B., Nourbakhsh, E., Cloonan, N., Grimmond, S.M., Bailey, T.L., Perkins, A.C., 2010. A global role for KLF1 in erythropoiesis revealed by ChIP-seq in primary erythroid cells. Genome Res. 20, 1052–1063. https://doi.org/10.1101/gr.106575.110

Tanimoto, K., Liu, Q., Bungert, J., Engel, J.D., 1999. Effects of altered gene order or orientation of the locus control region on human β-globin gene expression in mice. Nature 398, 344–348. https://doi.org/10.1038/18698

Tsang, F.H., Stolper, R.J., Hanifi, M., Cornell, L.J., Francis, H.S., Davies, B., Higgs, D.R., Kassouf, M.T., 2023. The characteristics of CTCF binding sequences contribute to enhancer blocking activity (preprint). Molecular Biology. https://doi.org/10.1101/2023.09.06.556325

Ugozzoli, L.A., Latorra, D., Pucket, R., Arar, K., Hamby, K., 2004. Real-time genotyping with oligonucleotide probes containing locked nucleic acids. Analytical Biochemistry 324, 143–152. https://doi.org/10.1016/j.ab.2003.09.003

Valton, A.-L., Venev, S.V., Mair, B., Khokhar, E.S., Tong, A.H.Y., Usaj, M., Chan, K., Pai, A.A., Moffat, J., Dekker, J., 2022. A cohesin traffic pattern genetically linked to gene regulation. Nat Struct Mol Biol 29, 1239–1251. https://doi.org/10.1038/s41594-022-00890-9

Vian, L., Pękowska, A., Rao, S.S.P., Kieffer-Kwon, K.-R., Jung, S., Baranello, L., Huang, S.-C., El Khattabi, L., Dose, M., Pruett, N., Sanborn, A.L., Canela, A., Maman, Y., Oksanen, A., Resch, W., Li, X., Lee, B., Kovalchuk, A.L., Tang, Z., Nelson, S., Di Pierro, M., Cheng, R.R., Machol, I., St Hilaire, B.G., Durand, N.C., Shamim, M.S., Stamenova, E.K., Onuchic, J.N., Ruan, Y., Nussenzweig, A., Levens, D., Aiden, E.L., Casellas, R., 2018. The Energetics and Physiological Impact of Cohesin Extrusion. Cell 173, 1165-1178.e20. https://doi.org/10.1016/j.cell.2018.03.072

Vietri Rudan, M., Barrington, C., Henderson, S., Ernst, C., Odom, D.T., Tanay, A., Hadjur, S., 2015. Comparative Hi-C Reveals that CTCF Underlies Evolution of Chromosomal Domain Architecture. Cell Reports 10, 1297–1309. https://doi.org/10.1016/j.celrep.2015.02.004

Vos, E.S.M., Valdes-Quezada, C., Huang, Y., Allahyar, A., Verstegen, M.J.A.M., Felder, A.-K., Van Der Vegt, F., Uijttewaal, E.C.H., Krijger, P.H.L., De Laat, W., 2021. Interplay between CTCF boundaries and a super enhancer controls cohesin extrusion trajectories and gene expression. Molecular Cell 81, 3082-3095.e6. https://doi.org/10.1016/j.molcel.2021.06.008

Wallace, H.A.C., Marques-Kranc, F., Richardson, M., Luna-Crespo, F., Sharpe, J.A., Hughes, J., Wood, W.G., Higgs, D.R., Smith, A.J.H., 2007. Manipulating the Mouse Genome to Engineer Precise Functional Syntenic Replacements with Human Sequence. Cell 128, 197–209. https://doi.org/10.1016/j.cell.2006.11.044

Walsh, T., Casadei, S., Munson, K.M., Eng, M., Mandell, J.B., Gulsuner, S., King, M.-C., 2021. CRISPR–Cas9/long-read sequencing approach to identify cryptic mutations in *BRCA1* and other tumour suppressor genes. J Med Genet 58, 850–852. https://doi.org/10.1136/jmedgenet-2020-107320

Wang, S., Su, J.-H., Beliveau, B.J., Bintu, B., Moffitt, J.R., Wu, C., Zhuang, X., 2016. Spatial organization of chromatin domains and compartments in single chromosomes. Science 353, 598–602. https://doi.org/10.1126/science.aaf8084

Wendt, K.S., Peters, J.-M., 2009. How cohesin and CTCF cooperate in regulating gene expression. Chromosome Res 17, 201–214. https://doi.org/10.1007/s10577-008-9017-7

Wendt, K.S., Yoshida, K., Itoh, T., Bando, M., Koch, B., Schirghuber, E., Tsutsumi, S., Nagae, G., Ishihara, K., Mishiro, T., Yahata, K., Imamoto, F., Aburatani, H., Nakao, M., Imamoto, N., Maeshima, K., Shirahige, K., Peters, J.-M., 2008. Cohesin mediates transcriptional insulation by CCCTC-binding factor. Nature 451, 796–801. https://doi.org/10.1038/nature06634

Wongsurawat, T., Jenjaroenpun, P., De Loose, A., Alkam, D., Ussery, D.W., Nookaew, I., Leung, Y.-K., Ho, S.-M., Day, J.D., Rodriguez, A., 2020. A novel Cas9-targeted long-read assay for simultaneous detection of IDH1/2 mutations and clinically relevant MGMT methylation in fresh biopsies of diffuse glioma. acta neuropathol commun 8, 87. https://doi.org/10.1186/s40478-020-00963-0

Wutz, G., Várnai, C., Nagasaka, K., Cisneros, D.A., Stocsits, R.R., Tang, W., Schoenfelder, S., Jessberger, G., Muhar, M., Hossain, M.J., Walther, N., Koch, B., Kueblbeck, M., Ellenberg, J., Zuber, J., Fraser, P., Peters, J., 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. https://doi.org/10.15252/embj.201798004

Xu, B., Wang, H., Wright, S., Hyle, J., Zhang, Y., Shao, Y., Niu, M., Fan, Y., Rosikiewicz, W., Djekidel, M.N., Peng, J., Lu, R., Li, C., 2021. Acute depletion of CTCF rewires genome-wide chromatin accessibility. Genome Biol 22, 244. https://doi.org/10.1186/s13059-021-02466-0

Zabidi, M.A., Arnold, C.D., 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. https://doi.org/10.1038/nature13994

Zerbino, D.R., Johnson, N., Juettemann, T., Wilder, S.P., Flicek, P., 2014. WiggleTools: parallel processing of large collections of genome-wide datasets for visualization and statistical analysis. Bioinformatics 30, 1008–1009. https://doi.org/10.1093/bioinformatics/btt737

Zhang, H., Shi, Z., Banigan, E.J., Kim, Y., Yu, H., Bai, X., Finkelstein, I.J., 2023. CTCF and R-loops are boundaries of cohesin-mediated DNA looping. Molecular Cell 83, 2856-2871.e8. https://doi.org/10.1016/j.molcel.2023.07.006

Zhang, S., Übelmesser, N., Barbieri, M., Papantonis, A., 2023a. Enhancer–promoter contact formation requires RNAPII and antagonizes loop extrusion. Nat Genet 55, 832–840. https://doi.org/10.1038/s41588-023-01364-4

Zhang, S., Übelmesser, N., Barbieri, M., Papantonis, A., 2023b. Enhancer–promoter contact formation requires RNAPII and antagonizes loop extrusion. Nat Genet 55, 832–840. https://doi.org/10.1038/s41588-023-01364-4

Zhang, S., Übelmesser, N., Josipovic, N., Forte, G., Slotman, J.A., Chiang, M., Gothe, H.J., Gusmao, E.G., Becker, C., Altmüller, J., Houtsmuller, A.B., Roukos, V., Wendt, K.S., Marenduzzo, D., Papantonis, A., 2021. RNA polymerase II is required for spatial chromatin reorganization following exit from mitosis. Sci. Adv. 7, eabg8205. https://doi.org/10.1126/sciadv.abg8205

Zijlstra, W.G., Buursma, A., Meeuwsen-van Der Roest, W.P., 1991. Absorption spectra of human fetal and adult oxyhemoglobin, de-oxyhemoglobin, carboxyhemoglobin, and methemoglobin. Clinical Chemistry 37, 1633–1638. https://doi.org/10.1093/clinchem/37.9.1633

Zuin, J., Roth, G., Zhan, Y., Cramard, J., Redolfi, J., Piskadlo, E., Mach, P., Kryzhanovska, M., Tihanyi, G., Kohler, H., Eder, M., Leemans, C., Van Steensel, B., Meister, P., Smallwood, S., Giorgetti, L., 2022. Nonlinear control of transcription through enhancer–promoter interactions. Nature 604, 571–577. https://doi.org/10.1038/s41586-022-04570-y