

Functional dissection of a single enhancer at the mouse α -globin locus

A thesis submitted for the degree of Doctor in Philosophy



Helena Francis

Exeter College

Wellcome Doctoral Training Programme in Chromosome and Developmental Biology

Weatherall Institute of Molecular Medicine

University of Oxford

Trinity Term 2019

Abstract

Enhancers are cis-regulatory DNA elements that bind transcription factors and drive expression of target genes in a spatio-temporal manner. Despite their collective importance in cell fate specification, up to now enhancers have often been studied in assays that ignore the chromosomal context required for their function *in vivo*. In my DPhil, I aim to probe a single enhancer sequence *in situ*, using the strongest enhancer at the mouse α -globin locus as a model. The project addresses four main issues. Firstly, I test the sufficiency of a single element (R2) to activate gene expression when removed from its super-enhancer context. Impaired activation of the α -globin genes in this model highlights the importance of sequence context in enhancer function. Secondly, I describe the characterisation of an *in vitro* erythroid differentiation system to enable the analysis of novel genetic mutants of the α -globin locus. Thirdly, I present *in vitro* data from a number of informative enhancer combinations of the α -globin super-enhancer. Results support findings from the *in vivo* single-enhancer model and highlight a potential role for seemingly inactive elements in super-enhancer function. Finally, I zoom in from the level of a complete gene locus to that of a single 238 bp transcription factor binding platform at the core of the R2 enhancer, to determine the rules governing enhancer function at the sequence level. Ultimately, I aim to determine which of the models of transcription factor binding and element co-operation are most fitting at this locus and probe the principles of transcriptional activation.

Acknowledgements

Firstly, I would like to thank my supervisors Mira Kassouf and Doug Higgs for giving me the opportunity to study under their guidance. Particular thanks go to Mira for the countless hours selflessly committed to mentoring me as an experimentalist and communicator, always with compassion and understanding. My thanks go to Doug for his endless support, infectious enthusiasm for science and optimism even through the most persistently failing experiments, and for sharing with me the potential of the α -globin locus.

To the members of the Higgs, Hughes and Buckle labs I am grateful for the technical assistance, scientific debate and moral support that has been plentiful throughout my time in the lab. Special thanks go to Chris Babbs for guiding me through the world of bacterial artificial chromosomes, Rob Beagrie for always being a fountain of knowledge for new technologies and Andrew King and Danuta Jeziorska for the creation of a fluorescent cell line that has been pivotal for my project. I also thank the masters of next-generation sequencing Damien Downes, Ron Schwessinger and Steph Hill for helping me with a range of methods and always lending me reagents when needed. Lastly, Liz Rose supports us all tremendously in our work so my thanks go to her for her patience and kindness.

This project would not have been possible without the fruitful collaboration we have established with Professor Jef Boeke and his lab members, especially Leslie Mitchell and Brendan Camellato. They achieved the technical feat of synthesising a challenging locus from scratch and were patient through our early technical difficulties. I hope they agree that the effort was worth it; I am positive that there are many more exciting results to be had from this endeavour. On this side of the pond, the work in this thesis could not have been completed without the core facilities at the Weatherall Institute of Molecular Medicine. For this I particularly thank Jackie Sharpe and the Transgenics Group, Kevin Clark and the Flow Cytometry Facility, and Philip Hublitz and the Genome Engineering Facility. Their hours of dedicated work, outstanding expertise and kindness have been invaluable.

For their guidance, both scientific and practical, I would like to thank a number of other group leaders: Andrew Smith, Jim Hughes, Philip Hublitz, Neil Brockdorff, Catherine Porcher and Tudor Fulga. Finally, for their relentless teamwork, support and friendship, I would like to thank Matt Gosden, Rosa Stolper and my fellow “terrible twin” Caz Harrold, who have somehow put up with me throughout this process and helped me in too many ways to count. I can’t imagine a better lab family to have shared this experience with.

For supporting me outside of the lab I thank my parents, my sisters and my friends. Collectively they have provided litres of tea, mopped up litres of tears and always reminded me that there is more to life than failed PCRs.

This work was generously funded by Wellcome and the Medical Research Council UK.

Table of Contents

Abstract	2
Acknowledgements	3
Table of Contents	4
List of figures	7
List of Abbreviations	10
Chapter 1 – Introduction.....	11
1.1 The composition of the eukaryotic genome	11
<i>Functions of the non-coding genome</i>	<i>12</i>
1.2 Enhancers as regulatory elements	13
<i>The role of enhancers in gene regulation.....</i>	<i>13</i>
<i>Mechanisms of enhancer function.....</i>	<i>13</i>
1.3 How to identify enhancers.....	15
<i>Predicting enhancers by molecular signature</i>	<i>15</i>
<i>Predicting enhancers by sequence conservation</i>	<i>16</i>
<i>Testing enhancer function in reporter assays and in situ</i>	<i>17</i>
1.4 Enhancers in clusters	18
<i>Genetic dissection of super-enhancers.....</i>	<i>20</i>
1.5 Enhancers in 3D space	23
<i>Point-to-point contact</i>	<i>23</i>
<i>Self-interacting active domains</i>	<i>24</i>
1.6 Enhancers as base pair sequences	27
<i>The components of an enhancer sequence</i>	<i>27</i>
<i>The organisation of an enhancer sequence</i>	<i>28</i>
1.7 The mouse α-globin locus as a model for enhancer function.....	31
<i>The globin loci as experimental models.....</i>	<i>31</i>
<i>The mouse α-globin locus.....</i>	<i>32</i>
<i>The mouse α-globin super-enhancer.....</i>	<i>34</i>
1.8 Outline for this thesis	35
Chapter 2 – Methods	37
2.1 mESC genetic engineering	37
<i>Synthetic RMGR.....</i>	<i>37</i>
<i>CRISPR/Cas9 editing.....</i>	<i>39</i>
<i>Screening.....</i>	<i>46</i>
2.2 Cell culture and isolation	50
<i>mES cell culture and differentiation</i>	<i>50</i>
<i>Isolation of in vitro-derived erythroid cells.....</i>	<i>51</i>
<i>Mouse model generation and breeding.....</i>	<i>53</i>
<i>Isolation of primary erythroid cells from embryos</i>	<i>54</i>

<i>Preparation of primary cells for single-cell FACS.....</i>	55
<i>Imaging cell cultures, EBs and cytopins.....</i>	56
2.3 Measurement of gene expression	57
<i>RNA extraction and RT-PCR.....</i>	57
<i>Single cell RT-PCR</i>	58
<i>YFP tag design and clonal readout.....</i>	62
<i>Pooled R2 variant sequencing.....</i>	63
2.4 NGS assays	67
<i>ATAC-seq and ChIP-seq.....</i>	67
<i>Motif analysis.....</i>	69
<i>Capture-C</i>	69
<i>BAC sequencing.....</i>	72
<i>10x linked-read analysis.....</i>	73
Chapter 3 – A super-enhancer reduced to a single element	74
3.1 Introduction.....	74
<i>Previous functional dissection of the mouse α-globin super-enhancer</i>	74
<i>Predictions for the R2-only model</i>	76
3.2 Results.....	78
<i>The local activity of the mouse α-globin super-enhancer</i>	78
<i>How to engineer multiple genetic elements in a single transfection step</i>	82
<i>The activity of the R2 element stripped of its super-enhancer context</i>	97
3.3 Discussion	102
<i>The impact of technology availability on scientific approach.....</i>	102
<i>Addition $\neq$ subtraction.....</i>	104
Acknowledgements.....	106
Chapter 4 – The embryoid body as a source of mouse erythroid cells <i>in vitro</i>	107
4.1 Introduction.....	107
<i>In vitro differentiation of stem cells as an experimental tool.....</i>	107
<i>Mouse erythropoiesis in a dish</i>	108
4.2 Results.....	111
<i>Isolation of erythroid cells from day 7 EBs.....</i>	111
<i>EB-derived erythroid cells are uniformly embryonic-like in nature.....</i>	115
<i>EB differentiation as a scalable tool for erythrocyte generation</i>	123
4.3 Discussion	131
<i>Requirements for an effective in vitro differentiation system</i>	131
Data availability	135
Acknowledgements.....	135
Chapter 5 – Rebuilding the α-globin super-enhancer from the bottom-up	136
5.1 Introduction.....	136
5.2 Results.....	137
<i>Generating a mESC line hemizygous for the α-globin locus.....</i>	137
<i>Validating enhancer mutants in vitro.....</i>	143
<i>A novel set of enhancer combination models</i>	147

5.3 Discussion	153
<i>A platform for the dissection of a regulatory domain.....</i>	<i>153</i>
<i>The importance of context in addressing enhancer function.....</i>	<i>154</i>
<i>Element co-operation through molecular environment</i>	<i>157</i>
<i>Element co-operation through co-linearity</i>	<i>158</i>
<i>The significance of enhancer sequence for function.....</i>	<i>159</i>
Data availability	161
Acknowledgements.....	161
Chapter 6 – Dissection of the R2 multispecies conserved sequence	162
6.1 Introduction.....	162
<i>The R2 multispecies conserved sequence</i>	<i>162</i>
<i>Existing methods for regulatory sequence dissection</i>	<i>166</i>
6.2 Results.....	169
<i>Development of a fluorescence-based screen to monitor α-globin expression</i>	<i>169</i>
<i>Screen #1: comparing elements of the α-globin super-enhancer</i>	<i>183</i>
<i>Screen #2: gross-scale dissection of the R2 enhancer with targeted in situ mutations</i>	<i>185</i>
<i>Screen #3: precise perturbation of predicted TF binding sites in the R2 enhancer.</i>	<i>191</i>
<i>Screen #4: variation of conserved Gata1-Nf-e2 spacing</i>	<i>203</i>
<i>The effect of mutational burden on R2 function</i>	<i>206</i>
6.3 Discussion	208
<i>A novel screening method to test mammalian regulatory sequences in situ</i>	<i>208</i>
<i>Context is as critical at an intra-enhancer resolution as for the inter-enhancer scale</i>	<i>209</i>
<i>The R2 MCS appears to be a flexible platform for TF recruitment.....</i>	<i>210</i>
<i>Evidence for TF density as a driver of enhancer function</i>	<i>212</i>
Acknowledgements.....	216
Chapter 7 – Final conclusions	217
7.1 Summary of findings.....	217
7.2 Contribution to the understanding of genome regulation	219
<i>The value in targeting endogenous loci at various resolutions</i>	<i>219</i>
<i>Variation in the roles of enhancer elements.....</i>	<i>221</i>
<i>The contributions of individual TFs to a single enhancer.....</i>	<i>224</i>
7.3 Next steps.....	226
Appendix - Genetic models used in this thesis.....	228
References	229

List of figures

Chapter 1 - Introduction

Figure 1.1: A range of possible functional relationships between active enhancer pairs. ...	20
Figure 1.2: Mechanisms to bring enhancers and promoters into physical proximity.....	26
Figure 1.3: The importance of motif organisation in enhancer sequence construction.	28
Figure 1.4: The structure of the interferon- β enhanceosome.	30
Figure 1.5: Molecular characterisation of the mouse α -globin locus.	33

Chapter 2 - Methods

Figure 2.1: Schematic diagrams of HDR donors.....	41
Figure 2.2: GFP/mRuby-positive selection for transfection of mESCs with CRISPR/Cas9 plasmids.	45
Figure 2.3: CD71-positive selection of EB-derived erythrocytes by FACS.....	53
Figure 2.4: MGG stained cytospin preparations of circulating blood collected from E9.5 and E10.5 embryos.	55
Figure 2.5: Division of a YFP-positive EB-derived population into eight gates to screen pooled R2 variant sequences.....	64
Figure 2.6: Estimation of YFP median fluorescence readings by NGS variant demultiplexing.	65
Figure 2.7: MEME-generated position weight matrices used for TF motif calling.	69

Chapter 3 - A super-enhancer reduced to a single element

Figure 3.1: α -globin gene transcription in single- and double-enhancer deletion models..	75
Figure 3.2: Prediction for α -globin expression levels across active enhancer deletion mouse models according to an additive model of super-enhancer function.....	77
Figure 3.3: Promoters within a 5.6 Mb region containing the α -globin locus do not show specific interactions with the super-enhancer elements in mouse erythrocytes.	81
Figure 3.4: Schematic for RMGR targeting strategy to generate the R2-only model in mESCs.	84
Figure 3.5: Workflow for the Synthetic RMGR procedure.	85
Figure 3.6: Enhancer deletion coordinates remove all sequence functionality.....	86
Figure 3.7: Schematic view of the eSWAP-In method (provided by Professor Jef Boeke).	88
Figure 3.8: Validation of synthetic BACs of WT and R2-only α -globin loci.....	90
Figure 3.9: Correct RMGR-exchanged clones were identified by PCR.....	93
Figure 3.10: The R2-only BAC was integrated into mESCs without structural rearrangements.	95
Figure 3.11: R2-only locus elements become accessible as expected in vivo.	99
Figure 3.12: Expression of the α - and ζ -globin genes is severely impaired in the R2-only model.	101

Chapter 4 - The embryoid body as a source of mouse erythroid cells *in vitro*

Figure 4.1: Mouse erythropoiesis occurs in three distinct waves, each with a different pattern of globin expression.	110
Figure 4.2: EBs can be directly exploited as a source of erythroid cells.	112
Figure 4.3: Column-based, high-purity isolation of erythroid cells from day 7 EBs.	114
Figure 4.4: An EB-derived erythroid population resembles embryonic primary red cells by ATAC-seq.	116
Figure 4.5: The globin loci are marked with active histone modifications in EB-derived erythrocytes.	117
Figure 4.6: Characteristics of erythroid cells used in single-cell RT-PCR by immunophenotype and tissue type.	120
Figure 4.7: EB-derived erythroid cells are homogeneously embryonic-like by globin expression at a single-cell level.	121
Figure 4.8: EB-derived single cells recapitulate an embryonic transcriptional programme.	122
Figure 4.9: 96-well EBs vary in size and number depending on mESC plating density..	124
Figure 4.10: Optimal plating density for 96-well EB differentiation is 200-400 cells per well.	126
Figure 4.11: EB differentiation can be scaled down to a 96-well format without compromising the erythroid output.	127
Figure 4.12: EBs differentiated in 96-well format produce primitive-like erythrocytes by ATAC-seq.	129
Chapter 5 - Rebuilding the α-globin super-enhancer from the bottom-up	
Figure 5.1: Design to delete the WT α -globin locus from RMGR-ready mESCs.	140
Figure 5.2: mESCs hemizygous for the α -globin locus undergo normal <i>in vitro</i> erythroid differentiation.	141
Figure 5.3: A hemizygous version of the mouse α -globin locus functions as normal.	142
Figure 5.4: Design for the production of hemizygous mESCs for $\Delta R1$, $\Delta R2$ and $\Delta R1R2$ models of the α -globin super-enhancer.	144
Figure 5.5: Chromatin accessibility results from hemizygous R1/R2 enhancer deletions <i>in vitro</i> mimic <i>in vivo</i> data.	145
Figure 5.6: Gene expression results from hemizygous R1/R2 enhancer deletions <i>in vitro</i> mimic <i>in vivo</i> data.	146
Figure 5.7: Design for the production of hemizygous mESCs for R1/R2/R4 insertion models of the α -globin super-enhancer.	148
Figure 5.8: mESCs hemizygous for the ΔSE model of the α -globin locus undergo normal <i>in vitro</i> erythroid differentiation.	149
Figure 5.9: Chromatin accessibility at the α -globin locus for hemizygous R1/R2/R4 enhancer insertion models differentiated <i>in vitro</i>	150
Figure 5.10: Chromatin accessibility at the β -globin locus for hemizygous R1/R2/R4 enhancer insertion models differentiated <i>in vitro</i>	151
Figure 5.11: Gene expression results from hemizygous R1/R2/R4 enhancer insertion models differentiated <i>in vitro</i>	152

Figure 5.12: The R1/R2 enhancers upregulate α -globin transcription to different extents depending on the integrity of the super-enhancer cluster.....	154
Figure 5.13: Differential binding of erythroid-specific TFs at elements of the mouse α -globin super-enhancer.	160
Chapter 6 - Dissection of the R2 multispecies conserved sequence	
Figure 6.1: Alignment of the R2 MCS in 18 vertebrate species.	164
Figure 6.2: <i>In silico</i> dissection of the R2 MCS.	165
Figure 6.3: Overview of a screen designed to dissect the R2 MCS.	170
Figure 6.4: Fluorescence of a YFP tag as a proxy for α -globin gene expression.....	172
Figure 6.5: Generation and <i>in vitro</i> differentiation of the R2-screening mESC line.....	174
Figure 6.6: The R2-screening model has normal chromatin at the globin loci.	176
Figure 6.7: Scheme for R2 variant replacement of the endogenous R2 enhancer by Cas9-directed HDR.	178
Figure 6.8: WT and Δ MCS R2 integrations by HDR result in functional and non-functional R2 elements, as expected.	179
Figure 6.9: Design of a clonal screen for α -YFP fluorescence levels.	182
Figure 6.10: α -YFP median results for R1 or R4 integrated in place of R2 (Screen #1)..	185
Figure 6.11: ChIP-seq for erythroid-specific TFs in EB-derived erythroid cells.....	187
Figure 6.12: Design for Screen #2: a gross-scale interrogation of R2 sequence composition.	188
Figure 6.13: Results from R2 Screen #2.	190
Figure 6.14: Design of Screen #3: an R2 screen to disrupt TF binding sites alone and in combinations.	192
Figure 6.15: Measurement of approximate median fluorescence values from R2 variant pools via gated YFP populations.....	194
Figure 6.16: Representations of individual variant sequences in Screen #3.....	196
Figure 6.17: The impact of single motif mutations on the function of the R2 MCS (R2 Screen #3).....	197
Figure 6.18: The impact of combined motif mutations on the function of the R2 MCS (R2 Screen #3).	199
Figure 6.19: The effect of Gata1-Nf-e2 motif spacing on R2 function (Screen #4).	205
Figure 6.20: The effect of R2 mutation depends on the disruption of TF motifs and not on the extent of sequence mutation in number of bp.	207
Figure 6.21: Erythroid-specific TF motifs are enriched at the R1 and R2 enhancers.	214
Chapter 7 - Final conclusions	
Figure 7.1: Proposed functional segregation of the elements of the mouse α -globin super-enhancer.....	222
Appendix - Genetic models used in this thesis	
Figure A.1: Graphical summary of the genetic models used in this thesis.	228

List of Abbreviations

3C	Chromosome conformation capture
APH	Acetylphenylhydrazine
ATAC	Assay for transposase-accessible chromatin
BAC	Bacterial artificial chromosome
CDF	Cumulative density function
ChIP	Chromatin immunoprecipitation
CI	Confidence interval
CRISPR	Clustered regularly interspaced short palindromic repeats
CRISPRa	CRISPR activation
CRISPRi	CRISPR inactivation
CTCF	CCCTC-binding factor
DHS	DNase I hypersensitive site
E9.5/E10.5	Embryonic day 9.5/10.5
E14	E14-TG2a.IV (mESC clone)
EB	Embryoid body
FACS	Fluorescence-activated cell sorting
FCS	Fetal calf serum
FIMO	Find Individual Motif Occurrences
Flp	Flippase
frt	Flippase recognition target
HAT	Hypoxanthine-aminopterin-thymidine
HDR	Homology-directed repair
Hprt	Hypoxanthine phosphoribosyltransferase
HT	Hypoxanthine-thymidine
MACS2	Model-based Analysis of ChIP-Seq
MCS	Multispecies conserved sequence
MEME	Multiple Em for Motif Elicitation
mESC	Mouse embryonic stem cell
MGG	May–Grünwald–Giemsa
MPRA	Massively parallel reporter assay
NGS	Next-generation sequencing
NLS	Nuclear localisation signal
PAM	Protospacer adjacent motif
PCA	Principal component analysis
PolII	RNA polymerase II
RMGR	Recombinase-Mediated Genomic Replacement
RPKM	Reads per kilobase per million mapped reads
rpm	Rotations per minute
RT-PCR	Real-time PCR
SE	Super-enhancer
TF	Transcription factor
tk	Thymidine kinase
UCSC	University of California Santa Cruz
UTR	Untranslated region
WT	Wild type
YFP	Yellow fluorescent protein

Chapter 1 – Introduction

1.1 The composition of the eukaryotic genome

An organism's genome can be thought of as its blueprint, containing the necessary instructions to build and maintain an entire organism from a single cell. The primary functional output of any genome comes from the expression of protein-coding genes. In prokaryotes, the number of genes generally dictates the resulting genome size in number of base pairs (Gregory, 2005). This is to be expected in a system in which gene expression is largely coordinated at the level of gene promoters. In eukaryotes, on the other hand, gene number is not predictive of genome size; in general, the number of genes plateaus for large genomes. The primary reason for this distinction is the abundance of non-coding DNA in complex organisms: non-coding sequences make up 11% of the genome in type strain *E. coli* (Meier-Kolthoff et al., 2014) but over 98% in human (Mouse ENCODE Consortium, 2012).

The importance of non-coding DNA is highlighted by the relationship between gene number and organism complexity. Given that the number of genes of an organism dictates the range of possible molecular functions, one might predict that, across evolution, gene number would scale with complexity. Yet, the human genome consists of only ~20,000 protein-coding genes; the same number as the nematode worm *C. elegans* (Hillier et al., 2005; International Human Genome Sequencing Consortium, 2004). This suggests that complexity is more likely driven by the spatio-temporal regulation of a modest number of genes than by absolute gene number. However, the abundance of non-coding DNA in itself

is not a direct indicator of organism complexity, either. The genome of the pufferfish *F. rubripes*, for example, contains the required coding content to produce a complex vertebrate with a genome that is an order of magnitude smaller than the human genome (Aparicio et al., 2002). A major challenge of post-genomics biology is to distinguish useful non-coding DNA from “junk”. As an ambitious target for this work, Human Genome Project-Write aims to produce a minimal complete genome sequence, in which any non-functional or redundant sequences are removed (Boeke et al., 2016). With this model it will be possible to test the true required ratio for non-coding vs coding sequence content in mammalian development.

Functions of the non-coding genome

Non-coding regions of the genome include sequences required at all levels of genetic regulation, from the production of transcripts and downstream processing steps to chromatin folding and replication. Fundamentally, the primary non-coding units involved in gene regulation can be broadly divided into three categories: gene promoters, enhancers and boundary elements. As for many concepts in biology, the distinctions between these categories are not absolute, and many units of the genome demonstrate characteristics associated with more than one, or indeed all of these categories (Andersson et al., 2015; Shen et al., 2012). Work in this thesis focuses on the role of well-defined enhancers as *cis*-regulatory elements of the mammalian genome, using the mouse as an experimental model.

1.2 Enhancers as regulatory elements

The role of enhancers in gene regulation

Since the coining of the term in 1981 (Banerji et al., 1981), the generalised model for the function of enhancers has become one of local activation, whereby sequence elements become accessible in relevant cell types and activate spatiotemporally-appropriate gene transcription *in cis*. The tissue-specific accessibility of enhancer elements near to upregulated genes is well-documented (Heinz et al., 2015; Long et al., 2016), but the pattern of their activity is not intuitive, due to the fact that enhancers can function independently of orientation and over large genomic distances (Shlyueva et al., 2014). Furthermore, enhancers can “skip” the most proximal promoters in a manner that is not fully understood, but may involve sequence compatibility or chromatin environment (Guo et al., 2012a; Haberle et al., 2019; Xi and Beer, 2018; Zabidi et al., 2015). Together, these features make the functions of enhancers extremely versatile and difficult to predict.

Mechanisms of enhancer function

Enhancers are thought to exert their activating potential on target genes through the direct binding of transcription factors (TFs) to the enhancer DNA sequences. Such binding events drive the recruitment of cofactors such as Mediator (Kagey et al., 2010), exemplified by the direct interaction of the Gata1 TF with the Mediator Med1 subunit in mouse erythroid cells (Stumpf et al., 2006). Mechanistically, the molecular consequences of TF/cofactor binding thought to be involved in gene activation can include nucleosome repositioning and/or polycomb eviction (Meers et al., 2019); the recruitment of PolII to enhancer and promoter

regions (Koch et al., 2011; Vernimmen et al., 2007); PolII pause release (Henriques et al., 2018); promoted transcriptional elongation (Cramer, 2019; Heinz et al., 2015; Vernimmen et al., 2011); eRNA transcription (Lam et al., 2014; Natoli and Andrau, 2012); domain formation via cohesin loading (Kagey et al., 2010; Zuin et al., 2014) and nuclear repositioning (Osborne et al., 2004; Ragoczy et al., 2006). The difficulty in probing just a single potential function at a time means that an understanding of the causal relationships between these mechanisms is still lacking.

1.3 How to identify enhancers

Fundamentally, there are four characteristics that can be used to classify regions of the genome as *bona fide* enhancers. First and second are descriptive features (chromatin environment and evolutionary conservation) that can be used to predict putative enhancer regions. Third and fourth are assessments of sequence function: predicted elements can be tested for the potential to enhance gene transcription using reporter assays and for function in their native chromosomal context using genetic models.

Predicting enhancers by molecular signature

By chromatin environment, efforts to catalogue the repertoire of *cis*-regulatory elements across a wide range of tissues in human and mouse have identified a total of 2.9 million and 1.5 million DNase I hypersensitive sites, respectively (DHSs; ENCODE Project Consortium, 2012; Thurman et al., 2012; Yue et al., 2014). This amounts to ~1% of the genome being accessible in any one cell type. Chromatin features such as histone modifications H3K27ac and H3K4me1 (Creyghton et al., 2010; Heintzman et al., 2007) and binding of co-activating proteins such as p300 and Mediator (Kagey et al., 2010; Visel et al., 2009) added to DNase-seq profiles hone in on likely active enhancers. By this analysis, there are a predicted ~300,000-400,000 enhancers in both mouse and human genomes (Shen et al., 2012; Yue et al., 2014; Zhu et al., 2013), though only a fraction will be active in any one cell type. This figure is similar to the total number of regions bound by tissue-specific TFs in mouse (Yue et al., 2014).

Alternative molecular signatures can also be used to predict the locations of active enhancer sequences. For example, the generation of transcripts from enhancers (eRNAs; De Santa et al., 2010; Kim et al., 2010) can be used in element prediction, although this captures a far smaller number of regions than protein binding analyses in the human genome (~45,000 total enhancers by the presence of bidirectional, capped eRNAs; Andersson et al., 2014). The binding of tissue-specific TFs and the resulting generation of DNase I footprints serve as further indicators of potential enhancers (Heinz et al., 2015; Stergachis et al., 2014). However, all of these biochemical attributes are neither a guarantee of function for an enhancer, nor a guide to the associated gene (Xi and Beer, 2018).

Predicting enhancers by sequence conservation

Sequence conservation of non-coding regions provides a predictive tool to identify putative enhancers (Pennacchio et al., 2006; Woolfe et al., 2004). Due to the importance of enhancer function in gene regulation, these sequences are evolutionarily constrained in terms of mutations when compared to the rest of the human genome (di Iulio et al., 2018). Of predicted enhancers in the mouse genome (as defined by the chromatin environment), 80% have identifiable homologues in the human genome (Yue et al., 2014). This represents a level of element conservation in line with that of gene promoter regions (80% by the same analysis; Yue et al., 2014). On the other hand, it should be noted that element conservation rapidly falls off with evolutionary distance; only 10% of mouse enhancers have homologues in zebrafish (Plessy et al., 2005). In line with this observation, it is thought that sequence conservation may be stronger for gene regulatory networks than for specific enhancer sequences (Stergachis et al., 2014; Yue et al., 2014). Therefore, while the presence of

sequence conservation can be associated with active elements, it is not a consistent indicator of enhancer function.

Testing enhancer function in reporter assays and in situ

For an element to be defined as an enhancer, it must be able to upregulate gene expression. Methods to test the sufficiency of putative enhancers in a high-throughput, unbiased manner typically rely on their juxtaposition with a reporter gene fused with a minimal promoter sequence. *In vitro*, sequences can be tested in a single cell type by transient transfection of reporter plasmids in highly scalable experiments termed massively parallel reporter assays (MPRAs; Levo and Segal, 2014). However, the function of sequences in reporter assays is not always instructive of activity within a native chromosomal context in which elements are bound by rules of DNA folding, chromatin environment and effects of element positioning (Splinter and de Laat, 2011). To test sequence function in a chromatinised context, therefore, sequences must be integrated randomly or directedly into the genome. Such lower scale experiments can be performed *in vitro* or can additionally be used to determine tissue-specificity *in vivo* by scoring whole embryos for visible expression of the *lacZ* gene product (Pennacchio et al., 2006) or by fluorescence *in situ* hybridisation (Kvon et al., 2014). Finally, while reporter assays are able to test the potential for a sequence to activate gene expression in a certain tissue, the essentiality of enhancers can be determined only through genetic manipulation of native regulatory domains.

1.4 Enhancers in clusters

Enhancers that act together to upregulate expression of a gene or genes can be found in clusters throughout the genome. The combinatorial use of differentially active enhancers at the same locus means that a single gene can be upregulated to different extents across a range of tissue types and/or in tightly regulated populations of cells (Maeda and Karch, 2011; Osterwalder et al., 2018). Alternatively, clusters of active elements can work together to increase the activity or robustness of gene expression (Long et al., 2016; Spitz and Furlong, 2012). The spatial relationships between enhancer elements can be highly conserved over even hundreds of kb, providing additional support for their role in activating a specific target gene *in cis* (Nobrega et al., 2003; Philipsen and Hardison, 2018). For regions containing multiple elements, a greater number of conserved sites and greater extent of sequence conservation through evolution are both associated with a higher level of target gene expression (Berthelot et al., 2018).

The finding that enhancers can act in clusters is not a recent observation; the well-characterised enhancer region of the β -globin locus containing 5 closely spaced enhancer elements, for example, was first described over 30 years ago (Grosveld et al., 1987). The classification of clusters or extended regions of enhancer activity has previously been attempted by multiple alternative criteria including the ability to negate effects of position variegation (locus control regions (LCRs), Grosveld et al., 1987); active sequence length (stretch enhancers, Parker et al., 2013); or the juxtaposition of multiple elements (composite enhancers, Bauer et al., 2013).

Recently, many enhancer clusters have instead been described as “super-enhancers”, based on the enrichment of active marks such as Med1 and H3K27ac (Whyte et al., 2013). Because the definition of a super-enhancer refers to the abundance of binding by ChIP-seq, rather than on the density of DHSs, super-enhancers can constitute anything from a single accessible element (15% of super-enhancers in mouse embryonic stem cells (mESCs); Pott and Lieb, 2015) to an active stretch of 50 kb. Super-enhancers are proposed to be important for the definition of cell identity, due to their proximity to genes used in lineage specification. They are also found to be enriched for sequence variants associated with human disease (Hnisz et al., 2013). However, many super-enhancers are active across a wide range of cell types, indicating that they may act more generally to promote gene expression at robust or high levels (Ryu et al., 2019).

In the original description of the super-enhancer concept, it was proposed that such structures could confer a level of activity that exceeds the sum of its constituent parts (individual elements separated pairwise by no more than 12.5 kb), due to the enhanced binding of active markers at these regions (Whyte et al., 2013). However, as the comparison was made against so-called “typical enhancers” with different sequence characteristics from elements of super-enhancers, it remains possible that super-enhancers can, instead, represent clusters of independently strong enhancers. Thus, controversy has arisen over whether super-enhancers truly act in a synergistic manner or, rather, are additive collections of strong enhancers. Pairwise between elements, a large range of functional relationships are possible, making dissection of the super-enhancer logic a challenging task (Figure 1.1).

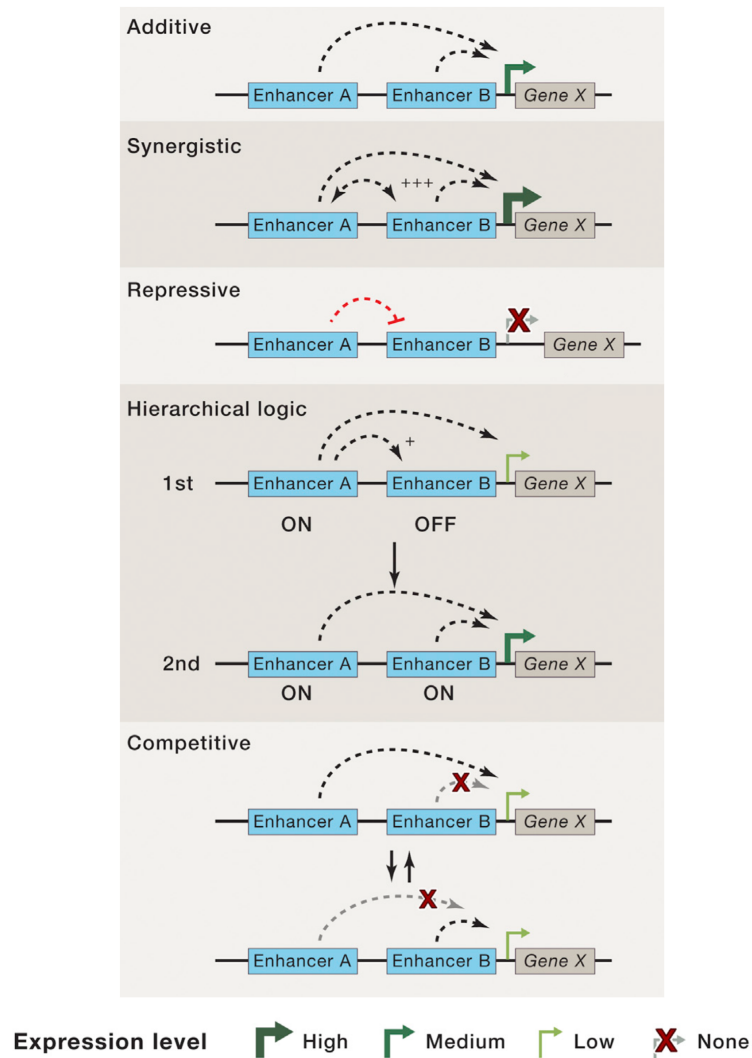


Figure 1.1: A range of possible functional relationships between active enhancer pairs.

Figure adapted from Long et al., 2016.

Genetic dissection of super-enhancers

A number of enhancer clusters labelled as super-enhancers have been genetically dissected to test for the level of synergy between constituent parts (Pott and Lieb, 2015). Expression data from single- and double-enhancer deletions at the mouse α -globin locus in the Higgs group best reflected an additive model of super-enhancer function (Hay et al., 2016). Other examples for additive-like enhancer relationships include the mouse Dppa5a gene locus

(Moorthy et al., 2017), β -globin LCR (Bender et al., 2012) and Bcl11a composite enhancer (Canver et al., 2015). On the other hand, results from the dissection of other endogenous loci have suggested a more hierarchical relationship between elements, in which deletion of a strong element appears to abrogate function of the others. Examples of this model include mouse loci for the miRNA290-295, Prdm14 and Sik1 genes (Hnisz et al., 2015) and the human Bcl11a composite enhancer (Canver et al., 2015). A similar pattern was seen at the super-enhancer for the mouse Slc25a37 gene but, in this case, the enhancer deletion with the strongest effect simultaneously deleted two prominent binding sites for the boundary protein CCCTC-binding factor (CTCF); thus, the experiment is confounded by the effect of domain structure (Huang et al., 2016). The effects of TF motif mutations at the mouse Wap gene locus also suggested a more hierarchical interaction between elements, though this was still insufficient to disprove a linear model for enhancer co-operation (Dukler et al., 2016; Shin et al., 2016). Genome-wide, it has been suggested that any contribution of enhancer hierarchy within a cluster can be predicted from interaction frequencies with the target gene promoter and is associated with the binding of CTCF to critical elements and thus with wider locus organisation (Huang et al., 2018).

Instead of implicating element synergy, results from other loci in *Drosophila* and mouse show evidence of enhancer redundancy, in which so-called “shadow” or “secondary” elements can appear non-functional until gene expression is tested in conditions of stress (Frankel et al., 2010; Montavon et al., 2011; Perry et al., 2010). Redundancy between elements may also create an environment for the development of new enhancer functions while retaining robust gene expression (Long et al., 2016).

The super-enhancer description provides a classification for the most strongly bound enhancer elements or clusters of elements in a single tissue type. By design, these regions are likely to be enriched for TF binding and, thus, perform biologically important functions. Therefore, I will refer to active clusters of enhancers with this terminology through this thesis.

1.5 Enhancers in 3D space

For enhancers or clusters of enhancers to exert their activity on distal gene promoters, there must be a physical proximity between these regions of the DNA molecule. In many cases, the manner in which this proximity is set up has consequences for enhancer-promoter specificity. Furthermore, potential functional coupling between the structure and activity of a regulatory landscape has implications for our understanding of what constitutes an enhancer element: traditional definitions dictate that enhancers upregulate gene expression directly but, within the context of a regulatory landscape, it is possible that some elements may exert their activity primarily in the coordination of large-scale DNA interactions. To understand the significance of 3D interactions in the potential roles of an enhancer, we must consider the interactions of an actively transcribing gene locus as a whole and review the mechanisms by which these interactions come about. A graphical summary of the discussed mechanisms is shown in Figure 1.2.

Point-to-point contact

Conceptually, the simplest model to rationalise how enhancers might exert their activity on long-distance promoters is through direct physical contact via DNA looping. Specific enhancer-promoter loops could be maintained by protein-protein interactions between TFs bound at each site. A number of *trans*-acting factors such as CTCF and Ldb1 have been proposed as potential mediators of such direct contacts through the formation of protein dimers (Deng et al., 2012; Rubio et al., 2008). However, both in terms of the mechanisms of setting up such interactions and through experimental observations of interaction frequencies across active gene loci, it is unlikely that the specific interaction of two points

in this way should occur without contribution from the surrounding sequence. Results from imaging and chromosome conformation capture (3C) methods suggest that spatial proximity of elements within a domain appears to be dynamic and diffuse, rather than fixed and punctate (imaging examples include Alexander et al., 2019; Benabdallah et al., 2019; Brown et al., 2018; Finn et al., 2019; Mateo et al., 2019; 3C examples include Davies et al., 2016; Dostie et al., 2006; van de Werken et al., 2012). This suggests that specific enhancer-promoter looping is not the primary driver of domain formation and maintenance. Examples in which point-to-point looping may be beneficially stabilised by direct protein-protein interactions include situations involving promoter skipping such as at the human Protocadherin locus (Guo et al., 2012a) and *Antennapedia* gene complex in *Drosophila* (Calhoun et al., 2002).

Self-interacting active domains

Genome-wide, interaction frequencies typically extend across large regions encompassing multiple active elements and are delimited by frequently interacting boundary regions, enriched for proteins such as CTCF (Ghirlando and Felsenfeld, 2016; Vietri-Rudan and Hadjur, 2015). This conformation is suggestive of the formation of self-interacting loops, which can be established prior to the onset of gene expression and in the absence of normal transcription activity (Benabdallah et al., 2019; Brown et al., 2018; Ghavi-Helm et al., 2014; Hay et al., 2016; Oudelaar et al., 2019). The proposed loop extrusion model provides a likely mechanism for the formation of self-interacting domains (Fudenberg et al., 2016). Evidence in support of the loop extrusion model is extensive from *in vivo* patterns of genome organisation (Fudenberg et al., 2016; Sanborn et al., 2015), perturbations of the involved machinery (Nuebler et al., 2018) and genetic disruption of flanking CTCF sites (examples

include Guo et al., 2015; Hanssen et al., 2017; Lupiáñez et al., 2015; Rodriguez-Carballo et al., 2017). The observed enrichment of extruding factors at active sequence elements and the tissue-specific nature of domain loops suggests that enhancers may be implicated in the formation of tissue-specific loops (Kagey et al., 2010; Schwarzer et al., 2017; Zuin et al., 2014).

A loop extrusion model for active domain formation allows for multiple elements within a linear sequence region to interact simultaneously. However, it does not account for the observation that interactions within a domain are not uniform: instead, contact frequencies are enriched between enhancer and promoter elements and are able to skip particular genes. In a transcription hub model, functional interactions between enhancers and promoters are stabilised by the favourable clustering of active factors bound at each sequence. This is supported by results from imaging and 3C methods capable of capturing multiway interactions, in which the interaction of multiple active elements together within a domain is seen to be enriched in active tissues (Allahyar et al., 2018; Oudelaar et al., 2018).

As TFs and associated cofactors tend to contain regions of high structural disorder, it has recently been proposed that clustering may involve extensive multivalent like-with-like interactions, reminiscent of phase separated condensates (Banani et al., 2017; Hnisz et al., 2017). This is corroborated by imaging studies in which proteins such as Oct4, Med1 and PolII can be seen to cluster at actively transcribing genes (Boija et al., 2018; Cho et al., 2018). It has long been recognised that regions of heterochromatin often cluster in 3D space by similar mechanisms (Pirrotta and Li, 2012; Strom et al., 2017). Indeed, interactions with active, inactive or Polycomb-repressed environments can be driven by the forced binding of factors such as an acidic peptide, a disordered peptide or TF (Nanog, YY1); a histone

methyltransferase (SUV39H1); or a component of the Polycomb Repressive Complex 2 (EZH2), respectively (Chong et al., 2018; Therizols et al., 2014; Weintraub et al., 2017; Wijchers et al., 2016). Several features such as high local TF concentrations and diffusion rates make phase condensation an appealing molecular model for transcription hub formation, but the difficulty in probing the condensate definition without large-scale disruption of the system leave it unclear as to what the driving mechanisms are in hub formation and function.

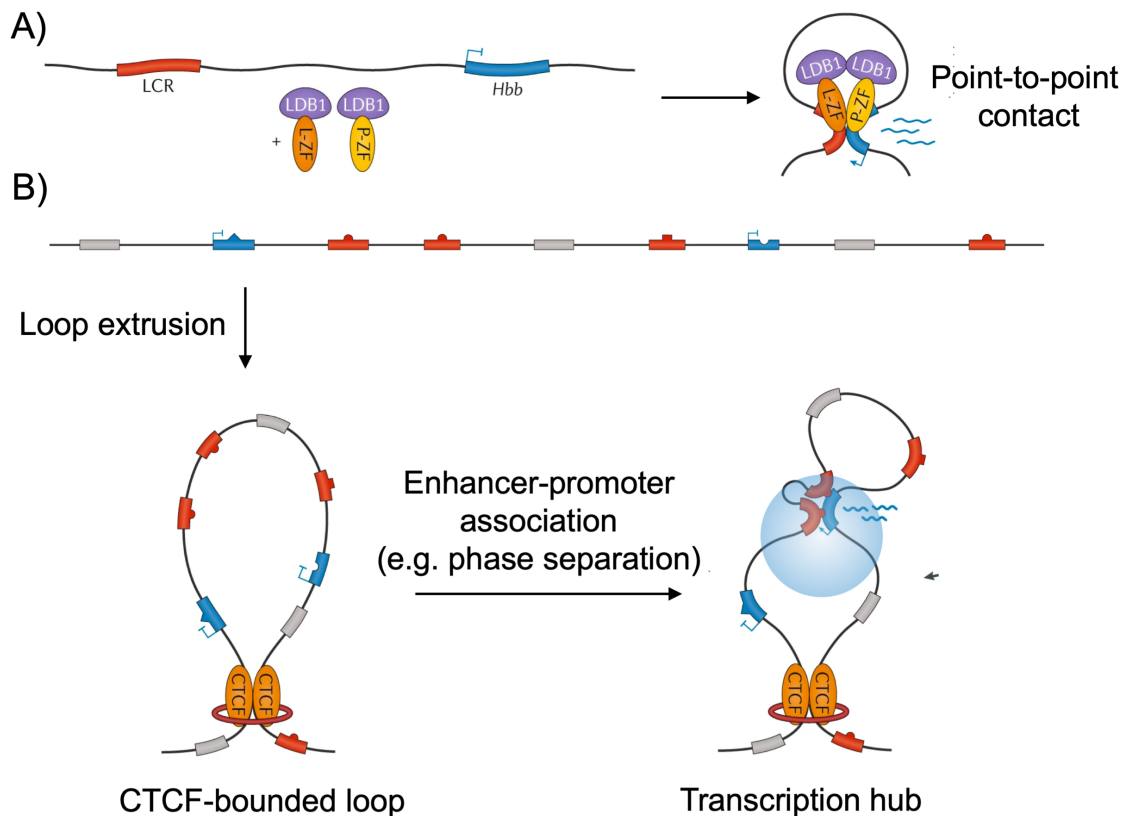


Figure 1.2: Mechanisms to bring enhancers and promoters into physical proximity.

A) Illustration of the point-to-point enhancer-promoter contacts induced in Deng et al., 2012. In this example, contacts were driven by homo-dimerisation of an Ldb1 moiety, fused to zinc-finger proteins labelled L-ZF and P-ZF to indicate binding sites at the LCR and *Hbb* promoter regions, respectively. **B)** Illustration of the current working model for the formation of domain loops and transcription hubs within them. Note that arrows highlight the effects of the loop extrusion and enhancer-promoter association steps and are not indicative of a fixed series of events in vivo, where these processes occur dynamically. Figure adapted from Schoenfelder and Fraser, 2019.

1.6 Enhancers as base pair sequences

The components of an enhancer sequence

Historically, it has been challenging to predict *cis*-regulatory elements from sequence due to the lack of easily-identifiable features: enhancers can activate target genes from extreme distances, with apparent disregard for intervening genes or other regulatory elements, and contain no universal sequence motifs (Pennacchio et al., 2013). One strategy to predict enhancers purely based on sequence would be to search for their component TF motifs in a given tissue, clusters of which are associated with active elements (Berman et al., 2002; Gotea et al., 2010). TFs typically interact directly with motifs of 6-12 bp in length (Tugrul et al., 2015) according to different temporal or spatial cues. Yet, the combinatorial nature of enhancer sequences means that they can be composed of multiple sequences that, individually, offer poor binding affinity and thus may be excluded from recognised motif searches (Jaeger et al., 2010; Ramos and Barolo, 2013). In addition, even high-affinity motifs can show large discrepancies between cases when TFs bind individually or as heterodimers, making these difficult to accurately curate and annotate (Jolma et al., 2015).

As well as encoding tissue-specificity, enhancer sequences can contain the necessary information to dictate the specificity of enhancer-promoter interactions. Sequence features of this type can be difficult to disentangle from studies of the whole genome, due to the large number of uncontrolled variables. However, within controlled sequence assays, it is possible to test pairs of enhancer-promoter sequences for sequence compatibility and enhancer function (Arnold et al., 2013; Zabidi et al., 2015).

The organisation of an enhancer sequence

Together, TF binding sites provide potential for a highly co-operative mechanism of gene activation: if multiple TFs are able to bind an enhancer simultaneously, protein-protein as well as protein-DNA interactions can be used to modulate gene activation. To what extent the underlying DNA sequence dictates or constrains these interactions remains unclear, though several models have been proposed. These range from the flexible billboard model, in which TF binding to consensus sequences occurs independently of surrounding sequence, to the strictly invariant enhanceosome, which is intolerant of even minor changes in sequence (Figure 1.3; Arnosti and Kulkarni, 2005; Junion et al., 2012; Long et al., 2016; Thanos and Maniatis, 1995).

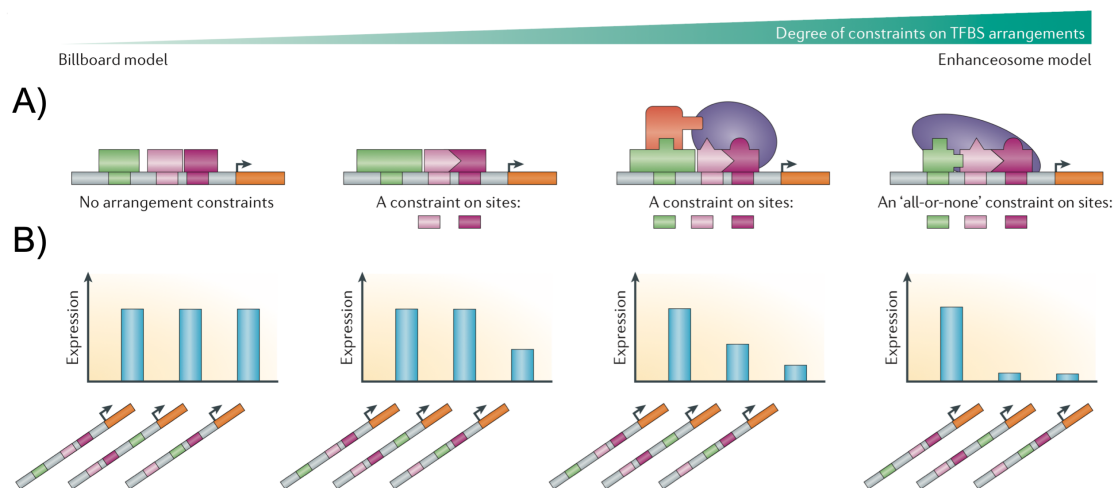


Figure 1.3: The importance of motif organisation in enhancer sequence construction.

A) Constraints of motif arrangement in enhancer sequence construction on a scale from a flexible billboard to an inflexible enhanceosome (above). The more specific protein-protein interactions are required for sequence binding, the more constraints there are on the primary DNA sequence. **B)** Corresponding effects of motif arrangement constraints on the functional output of enhancers with three different motif arrangements. Figure adapted from Levo and Segal, 2014.

The billboard model is supported by the observations that enhancer elements and their clusters can demonstrate high levels of redundancy and can act as flexible platforms for the evolution of new binding events (Long et al., 2016). These features make enhancers amenable to continuing sequence evolution and phenotypic robustness. Through evolution, enhancer sequences tend to show high conservation of *trans*-factor binding sites, with greater variance seen in intervening sequences, consistent with a “landing pad” structure for relatively flexible TF binding (Gotea et al., 2010). As an exception to the rule, one group of over 250 non-coding sequences, termed “ultraconserved”, show remarkable invariance in long (>200 bp) stretches of DNA between human and mouse genomes (Visel et al., 2008). As of yet, no unique function has been allocated to this group of sequences beyond the scope of typical enhancers (Dickel et al., 2018). Thus, it appears that the majority of mammalian enhancers are conserved primarily in their *cis*-regulatory logic, rather than by precise alignment of entire regulatory element sequences (Stergachis et al., 2014).

Primary evidence for the enhanceosome concept, on the other hand, comes from the human interferon- β enhancer. The complete length of this enhancer is highly conserved across a range of mammalian species (Panne et al., 2007) and function of the sequence has been shown experimentally to be extremely intolerant to even small mutations (Panne, 2008). A combined atomic structure of the recruited protein complex demonstrates why there is such inflexibility in this sequence: binding footprints for the constituent TFs are constrained to the extent that motifs for the core factors are overlapping (Figure 1.4).

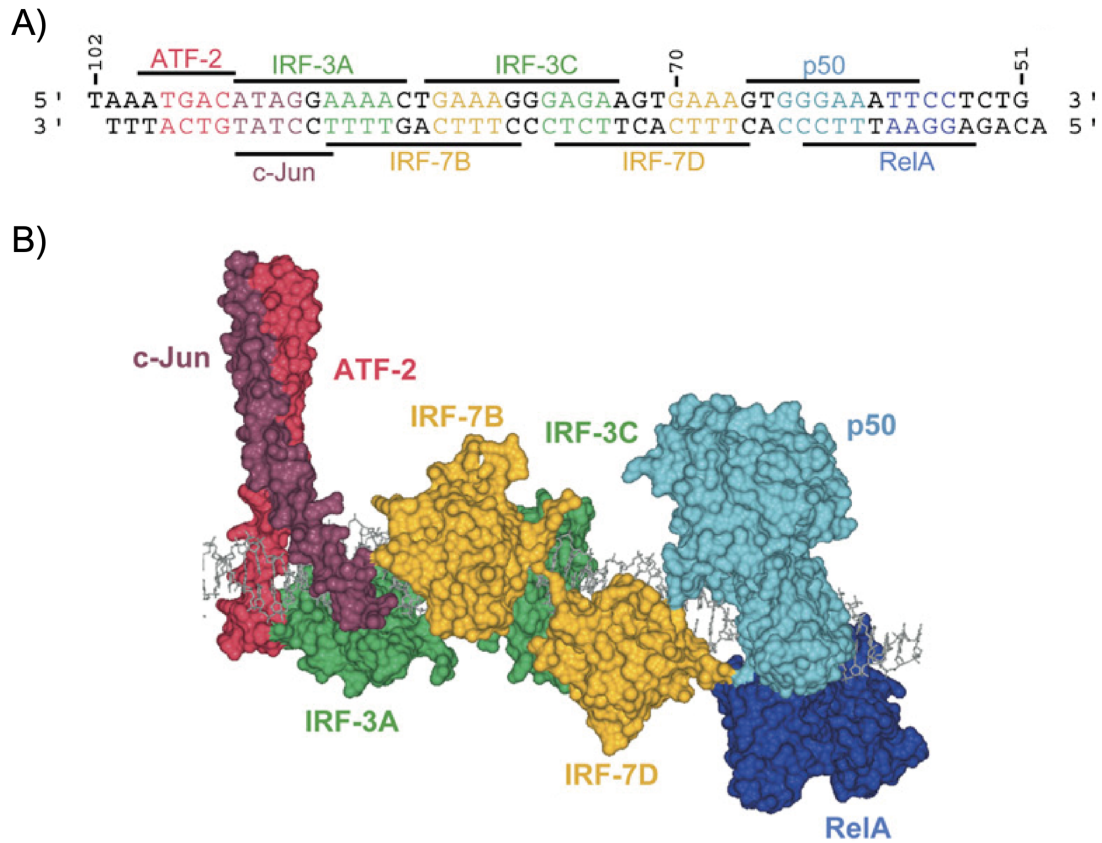


Figure 1.4: The structure of the interferon- β enhanceosome.

A) A 52 bp sequence containing the interferon- β enhancer. Motifs for bound TFs are highlighted by colour. Extensive overlap is seen between these motifs. **B)** Constructed atomic structure of the interferon- β enhanceosome bound by the TFs indicated in part A). Figure adapted from Panne et al., 2007.

There are many published examples of co-operative protein-protein interactions influencing binding to regulatory DNA sequences (for example, Grossman et al., 2017; Slattery et al., 2011; Yanez-Cuna et al., 2012). However, as yet, the interferon- β enhancer remains the most extreme example of multiple fixed protein-protein-DNA interactions in the literature. Conceptually, the presence of sequences that are so intolerant to mutations is functionally detrimental. One argument for the existence of the enhanceosome is in amplifying signal from TFs that are present in limiting concentration (Papatsenko and Levine, 2007). If this is the case, it may be that such regimented sequence organisation would only be required in a small number of instances throughout the genome.

1.7 The mouse α -globin locus as a model for enhancer function

The globin loci as experimental models

Historically, study of the mammalian globins has been instrumental across a wide range of research topics. Reasons for their use in experiments range from the practical in terms of the abundance of available material (both the relative ease in purifying a phenotypically pure population and the high concentration of globin mRNA and protein molecules in erythroid tissues) to the conceptual in the benefits of drawing direct causal links between genetic mutations and mis-regulated gene expression at these loci. The wealth of available data for the globin loci have also made them ideal test beds for a number of key technological developments, from endogenous gene disruption by homologous recombination (Smithies et al., 1985) to more recent advances in the pursuit of effective gene therapies (CRISPR Therapeutics, 2019; Davis et al., 2019).

The continued value in scrutinising a single locus in an age of large-scale genomics should not be underestimated. Many of the key discoveries in the study of gene regulation have been made possible by our in-depth understanding of sequence regions such as the globin loci; such findings can often be extrapolated genome-wide (Higgs et al., 2006; Higgs, 2013). In studying the principles of gene regulation, the α -globin locus offers an experimental platform with well-defined on-off states in populations that can be easily purified from primary tissue, a large dynamic range of transcriptional output and a regulatory domain that has been genetically dissected for its constituent parts. Furthermore, genetic modifications that perturb α -globin expression have no known consequences for erythropoiesis, and the regulatory elements of the locus all become accessible only in the erythroid lineage. Hence,

a complete range of gene expression can be tested in *bona fide* erythrocytes, without confounding issues due to indirect effects on cell fate.

The mouse α -globin locus

The structure of the α -globin locus is highly conserved throughout evolution. The functional embryonic and adult globin genes (ζ -globin/*Hba-x* and α -globin/*Hba-a1/2*, respectively) are contained within a 135 kb syntenic region that can be traced through 500 million years of evolution (Hughes et al., 2005). Using next-generation sequencing (NGS) and genetic assays, both *in vitro* and *in vivo*, the chromosomal region containing the mouse α -globin genes has been extensively characterised for its *cis*-regulatory elements (Figure 1.5). The well-defined α -globin super-enhancer consists of five individual elements, four of which (R1/R2/R3/Rm) are positioned within introns of the upstream *Nprl3* gene. The final enhancer (R4) is intergenic, located downstream of the *Nprl3* promoter. Elements were named based on their conservation across 22 vertebrate species (Hughes et al., 2005); the Rm sequence is found in the mouse genome but not in human.

The mouse α -globin genes and super-enhancer are contained within an erythroid-specific ~65 kb self-interacting domain, as shown by 3C methods (Figure 1.5; Hughes et al., 2014). The domain is demarcated by convergently oriented CTCF motifs, which are bound in both erythroid and non-erythroid cell types (Hanssen et al., 2017). Careful dissection of domain formation through erythropoiesis shows that chromatin organisation follows the opening of individual elements at the locus and is strengthened through erythropoiesis as gene expression becomes activated (Oudelaar et al., 2019). The overarching domain structure is able to form in genetic models lacking an intact α -globin super-enhancer or the *Hba-a1/2*

genes, though specific enhancer-promoter contacts are disrupted in these conditions (Brown et al., 2018; Hay et al., 2016).

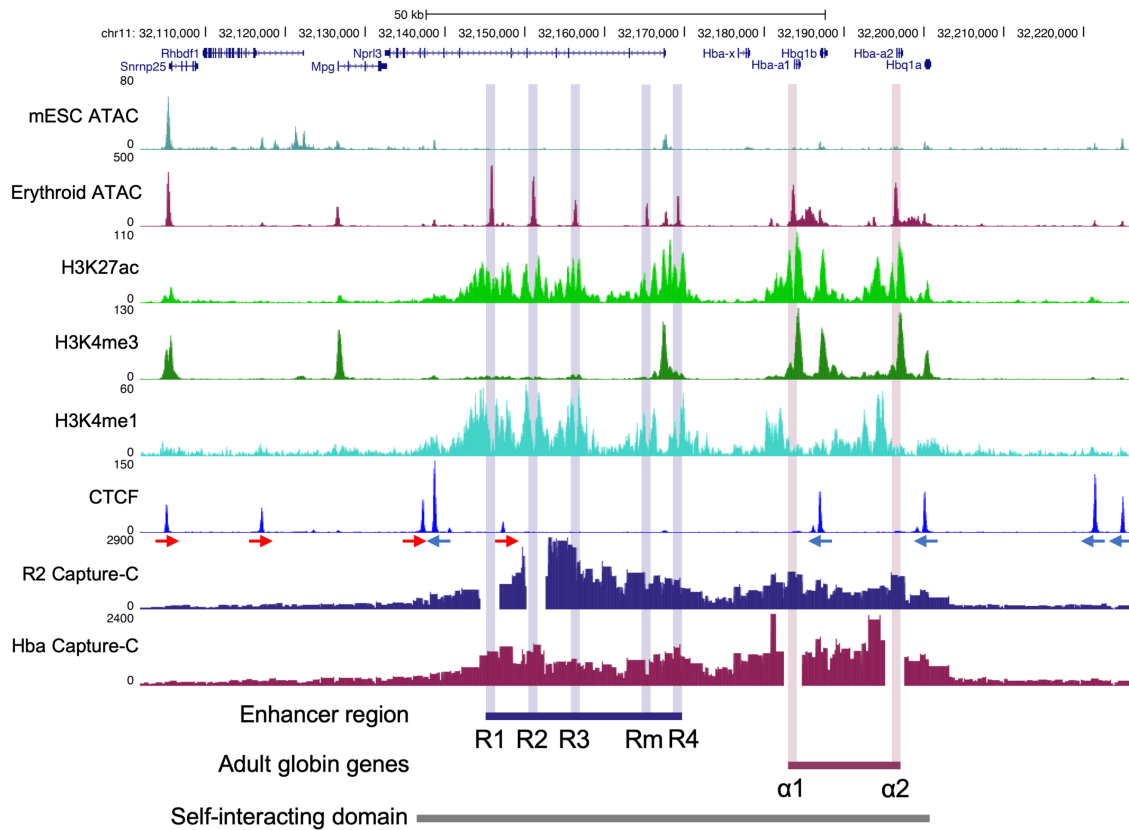


Figure 1.5: Molecular characterisation of the mouse α -globin locus.

Data from NGS assays are shown aligned to the mm9 mouse genome assembly. Top to bottom: Assay for transposase-accessible chromatin (ATAC-seq) tracks for mESCs and spleen-derived erythrocytes; ChIP-seq data from mouse spleen-derived erythrocytes for active histone marks H3K27ac (associated with active sequences), H3K4me3 (associated with active promoters) and H3K4me1 (associated with active enhancers); ChIP-seq data for mouse spleen-derived erythrocytes for CTCF (arrows below indicate the orientation of underlying CTCF motifs); Capture-C data for mouse spleen-derived erythrocytes from the viewpoint of the R2 enhancer and combined *Hba-a1/2* genes. The five enhancer elements (R1/R2/R3/Rm/R4) are highlighted in blue and the two adult α -globin genes (*Hba-a1/2*) in red. The erythroid-specific self-interacting domain is annotated in grey. All shown data have been previously published in Davies et al., 2016; Hanssen et al., 2017; Kowalczyk et al., 2012.

The mouse α -globin super-enhancer

The five elements of the α -globin super-enhancer become accessible in a coordinated manner through erythropoiesis but contribute to gene activation to different extents (Hay et al., 2016; Oudelaar et al., 2019). In mouse models carrying deletions of the enhancer elements, individual elements appear to confer different levels of activity to gene expression; R2 appears to be the most strongly activating element (50%; Hay et al., 2016). The sequence of the R2 element is also the most extensively conserved through evolution (Hughes et al., 2005; Philipson and Hardison, 2018). At the human α -globin locus, the prominence of R2 in the function of the cluster is even more extreme: its role as the major regulatory element of the locus is supported by the 97% reduction in human α -globin output in an interspecific hybrid mouse erythroleukaemia cell mutant (Bernet et al., 1995; Vernimmen et al., 2007), a humanised mouse mutant with comparable results (Wallace et al., 2007) and α -thalassemia phenotypes in patients carrying deletions of this region (Coelho et al., 2010; Higgs and Wood, 2008; Sollaino et al., 2010; Viprakasit et al., 2006).

1.8 Outline for this thesis

In this thesis I describe the functional dissection of the strongest enhancer of the mouse α -globin locus (R2) at a range of resolutions, from within its native chromosomal context.

Firstly, I ask whether the α -globin super-enhancer can be reduced to a single element by creating a homozygous mouse model devoid of the native R1/R3/Rm/R4 sequences. Thus far, *in situ* genetic dissections of enhancer clusters have primarily been performed by iterative element deletions from native loci, typically deleting one or two elements at a time. By instead posing the question of enhancer sufficiency from the perspective of sequence insertion into a null background, I ask whether positional context has a role to play in the activity of the R2 element (Chapter 3).

Secondly, I establish an embryoid body-based mESC differentiation system to recreate erythroid-specific phenotypes *in vitro*. I characterise the isolated erythroid population in terms of molecular phenotype and determine which *in vivo* cell state it best reflects (Chapter 4). As an experimental tool, this allows me to test genetic mutants of the α -globin super-enhancer in a manner that faithfully recapitulates *in vivo* phenotypes. I apply the mESC differentiation system to a novel set of minimal enhancer combination models to compare the activity of R2 in isolation to its activity when the seemingly inactive R3/Rm/R4 elements are present (Chapter 5).

Finally, I interrogate the sequence makeup of the R2 enhancer in a multiplexed *in situ* sequence screen for libraries of custom-designed variants. Although many previous studies have dissected regulatory sequences for their regulatory logic in high-throughput transgenic

assays, here I test the true ability for a range of sequences to act as distal enhancers from within a chromosomal context, using the *in vitro* erythroid differentiation tool described above. For this purpose, I outline the development of a screening approach based on flow cytometry, in which the recorded fluorescence intensity of an endogenous tag construct acts as a proxy for native α -globin gene expression. By mutating six predicted erythroid-specific TF motifs within the core R2 sequence, I test the importance of motif number, diversity and organisation in creating a strong enhancer and evaluate which of the existing models for regulatory sequence architecture is most fitting for the R2 element (Chapter 6).

My work highlights the value of a number of novel tools in genetically teasing apart the function of regulatory elements within their native chromosomal context. By successfully recreating *in vitro* a number of *in vivo* genetic mutants of the mouse α -globin locus with a newly characterised differentiation system, I lay out a platform to analyse regulatory elements and landscapes in an erythroid cell state, at a range of resolutions. Together, my thesis aims to disentangle the contributions of sequence and context in the function of a highly active *cis*-regulatory element (R2), within a well-characterised regulatory domain.

Chapter 2 – Methods

2.1 mESC genetic engineering

Synthetic RMGR

Bacterial artificial chromosomes (BACs) were synthesised from a wild type (WT) template construct containing the lox sites and hypoxanthine phosphoribosyltransferase (*Hprt*) split gene required for recombinase-mediated genomic replacement (RMGR). This template construct was generated by Dr Christian Babbs from a mouse BAC (RP23-469I8; BACPAC Resources Centre, Children's Hospital Oakland Research Institute; Osoegawa et al., 2000) by sequential shearing and lox integration steps using λ -red-mediated recombination (Zhang et al., 1998). Lox sites were integrated 85 kb apart, flanking the α -globin regulatory region (85,145 bp; Chr11:32,115,389-32,200,533). Additional synthetic fragments were used to create enhancer sequence deletions and to introduce a section including exon four of the *Rhbdf1* gene, missing in the template BAC but present in the exchange region of the RMGR-ready mESCs (635 bp; Chr11:32,114,754-32,115,388).

BACs containing synthetic versions of the mouse α -globin locus were received as transformed bacteria samples and grown up in LB under selection with kanamycin (20 μ g/ml) and ampicillin (50 μ g/ml). 1 L overnight cultures were cleared for cell debris using Plasmid Maxi Kit reagents P1-3 (Qiagen) as instructed for low-copy plasmids, with precipitated material removed by filtration. DNA was precipitated from the cleared supernatant using 1:1 isopropanol and washed with 70% ethanol, then purified by caesium

chloride centrifugation (Maniatis et al., 1982). BAC preparations were quantified by the Qubit dsDNA Broad-Range Assay (Thermo Fisher) and checked for sequence integrity by restriction digest and visualisation on a 1% agarose gel.

Purified BAC DNA was co-transfected with a pCAGGS-Cre-IRES*Spuro* plasmid (Smith et al., 2002) into RMGR-ready mESCs by lipofection using the Lipofectamine LTX reagent (Thermo Fisher). Transfections were performed in 6-well format by plating freshly trypsinised mES cells in 2 ml culture media supplemented with the transfection mix prepared to manufacturer's instructions: 5 µl LTX reagent, 2 µg plasmid DNA (1.5 µg BAC plus 0.5 µg pCAGGS-Cre-IRES*Spuro*), 2 µl PLUS reagent and 250 µl Opti-MEM (Thermo Fisher). Media was changed after 24 hours to remove the lipofection reagents and selection for complementation of the *Hprt* gene was started after an additional 24 hours with 1x HAT supplement (0.1 mM hypoxanthine, 0.4 mM aminopterin, 0.016 mM thymidine). Selection was continued for one-two weeks, at which point surviving colonies were picked into individual wells and HAT selection replaced with HT (0.1 mM hypoxanthine, 0.016 mM thymidine) recovery media.

Alternative transfection methods were tested with the following conditions. The BioRad Gene Pulser was used as previously published (Wallace et al., 2007). The BioRad Gene Pulser XCell and Thermo Fisher Neon Transfection System were used according to manufacturer guidelines; Neon electroporation conditions were Voltage 1400 V, Width 10 am, Pulses 3. The Lonza Nucleofector 2b was used with the Mouse Embryonic Stem Cell Nucleofector Kit, using programme A-023. FuGENE 6 lipofection was performed according to manufacturer guidelines, using 20 µl reagent per 10⁶ cells.

Colonies were expanded for stock freezing and genotyping before proceeding with downstream experiments. For *in vitro* experiments, *Hprt* gene removal was achieved through transfection of a transient flippase (Flp)-expressing plasmid (Schaff et al., 2001), clonal plating, colony picking, expansion and screening by selection with 10 μ M 6-thioguanine and PCR.

CRISPR/Cas9 editing

All clustered regularly interspaced short palindromic repeats (CRISPR)/Cas9 targeting strategies were designed, prepared and tested by Dr Philip Hublitz and colleagues in the Genome Engineering Core Facility at the Weatherall Institute of Molecular Medicine. Guide RNAs were cloned into a pSpCas9(BB)-2A-GFP (pX458) vector, a gift from Feng Zhang (Addgene plasmid #48138; <http://n2t.net/addgene:48138>; RRID:Addgene_48138) or a modified vector with the GFP tag exchanged for an mRuby cassette (pX458-Ruby; Kredel et al., 2009). Sequences of guide RNA target sequences are shown in Table 2.1.

Application	Target	Vector	Guide RNA sequence
Allele deletion	Left breakpoint	pX458	ACAAGATGCTCTCCGCTGT
	Right breakpoint	pX458-Ruby	AGGACAAATCCCGAAGGTC
R1 deletion	Left breakpoint	pX458	GTGACCATAAGTTGATTAG
	Right breakpoint	pX458-Ruby	GTACATACGCCGGGCGTGG
R2 deletion	Left breakpoint	pX458-Ruby	TACCTCCAAGGTTTTGCTC
	Right breakpoint	pX458	GCCGTGACACTTCATGCTCA
R2 variant targeting	Left breakpoint	pX458-Ruby	TACCTCCAAGGTTTTGCTC
Table 2.1: Guide RNA targeting sequences.			

Homology-directed repair (HDR) donor vectors for insertion into the R1 enhancer position were designed to carry at least 500 bp homology arms flanking the region for manipulation, the only additional modifications being inactivation of relevant Cas9 protospacer adjacent motif (PAM) sequences. Vectors were purchased as GeneART Gene Synthesis custom designs (Thermo Fisher; Figure 2.1).

To generate a library of R2 variant HDR vectors, a generic backbone containing the desired homology arms and puromycin / thymidine kinase selection cassettes flanking BsaI Type IIS restriction sites was ordered as a GeneART unit (Figure 2.1). Selection cassettes were positioned downstream of the R2 targeting region and Cas9 cut site to ensure that recombination of the entire, highly homologous, variant R2 sequence was selected for by the cassettes. This backbone also carries modifications to inactivate relevant PAM sequences, potential new footprint sequences involving an integrated Rox site as determined by the Sasquatch prediction tool (Schwessinger et al., 2017) and an unwanted BsaI site. R2 variant inserts were synthesised as Gene Fragments (Twist Bioscience) and cloned either individually or as pools into the generic HDR backbone by Golden Gate Assembly (NEB). In all cases, final vectors were transformed into Stellar Competent Cells in 20 µl reactions (Takara Bio Inc.). Transformed bacteria were plated and grown up in selective LB as required.

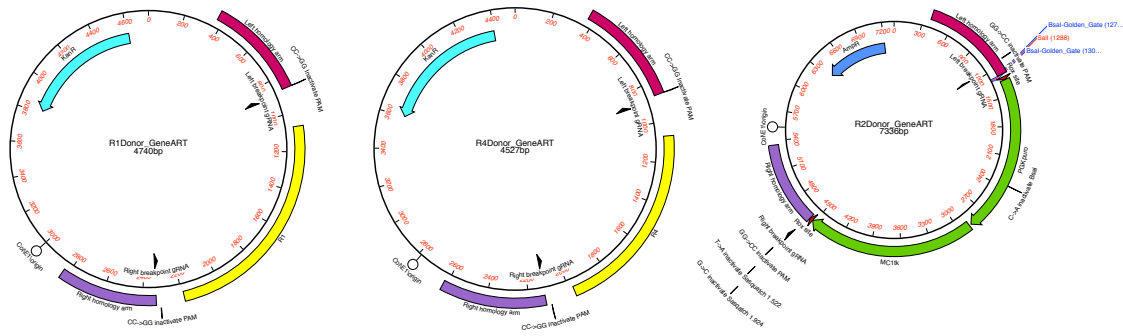


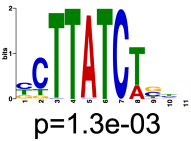
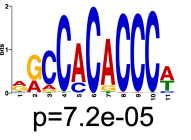
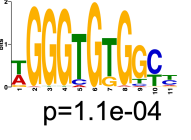
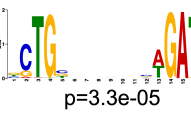
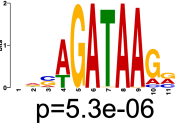
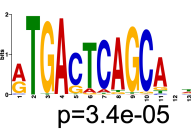
Figure 2.1: Schematic diagrams of HDR donors.

Variants for the R2 Screen #2 relied on the blocks defined in Table 2.2. Variants for R2 Screen #3 used combinations of the mutations specified in Tables 2.3 and Table 2.4. Spacing variants for R2 Screen #4 are shown in Table 2.5. Mutations were not predicted to create new TF motifs or footprints by Find Individual Motif Occurrences (FIMO) and Sasquatch analysis, using a shoulder-to-footprint ratio of 1.4 and visual inspection in cases above this threshold.

Block	Sequence	Length
A	ACCCATCTGGA <u>ACCTATCAGTG</u>	22 bp
B	ACCATAGTCAACAGCAGGT <u>GTACACACCCAGG</u> CCA <u>AGGGTGGAGCA</u>	46 bp
C	GACCACTGTGGGATCTATGGAGAT <u>GCTTGAACG</u> <u>AGCAGATAACT</u> AA	46 bp
D	GCCAAGC <u>ATGACTCAGAGTTTC</u> TAGAGGCCAC	32 bp
E	TAG <u>GGACTGCTGAGTAAT</u> ACTTGGGGG	26 bp
F	TACAGAGTCAGAAAGGAAAGGACAAATGGTAC <u>CACTGATTAGG</u> ACCTCTGACGCTGTTTTCCCATC	66 bp

Table 2.2: Blocks used in R2 variant sequence Screen #2.

Underlined sequences indicate conserved regions. Matches for conserved regions to erythroid motifs are highlighted in green (Gata1), yellow (Tal1 half E-box/Gata1 composite motif), pink (Nf-e2) and cyan (Klf1). One additional, unconserved Klf1 motif is highlighted in block B.

TF motif	Coordinates	Variant	Sequence
 p=1.3e-03	MCS:12-22	WT	TGGAACCTATCAGTGACCA
		Mutant 1	TGGAACCTAAACAGTGACCA
		Mutant 2	TGGAACCAATTGAGTGACCA
Block A Peak	MCS:34-36	WT	TCAACAGCAGG
		Mutant 1	TCAAGTGCAGG
 p=7.2e-05	MCS:42-52	WT	AGGTGTACACACCCAGGCC
		Mutant 1	AGGTGTACAGTCCGAGGCC
 p=1.1e-04	MCS:58-68	WT	GCCAAGGGTGGAGCAGACC
		Mutant 1	GCCAAGGCAGGAGCAGACC
		Mutant 2	GCCAACCGTGAGCAGACC
Block C Peak	MCS:73-79	WT	GACC <u>ACTGTGG</u> GATC
		Mutant 1	GACCTCTGACGGATC
 p=3.3e-05	MCS:93-110	WT	GATGCTTGAACGAGCAGATAA
		Mutant 1	GAAGGAAGAACGACCAGATAA
 p=5.3e-06	MCS:102-112	WT	AACGAGCAGATAACTAAGC
		Mutant 1	AACGAGCACTTAACTAAGC
		Mutant 2	AACGAGCAGAATTCTAAGC
 p=3.4e-05	MCS:122-136	WT	AAGCATGACTCAGAGTTTCTAGA
		Mutant 1	AAGCAACTCTCAGAGTTTCTAGA
		Mutant 2	AAGCAAGTCACTGAGTTTCTAGA

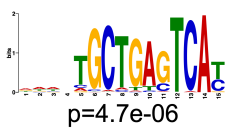
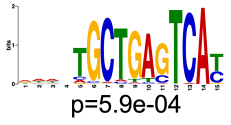
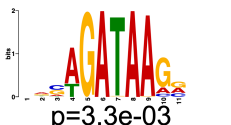
Block E Nf-e2  p=4.7e-06	MCS:149-163	WT	ACTAGGACTGCTGAGTAATACTT
		Mutant 1	ACTAGGACTG <u>GAGAGAA</u> TTACTT
		Mutant 2	ACTAGGACTG <u>GACAGAA</u> TTACTT
Block F Nf-e2  p=5.9e-04	MCS:169-183	WT	CTTGGGGGTACAGAGTCAGAAAG
		Mutant1	CTTGGGGGTACACAG <u>ACT</u> GAAAG
Block F Gata1  p=3.3e-03	MCS:205-215	WT	GTACCACTGATTAGGACCT
		Mutant 1	GTACCACTGT <u>ATT</u> GGACCT
		Mutant 2	GTACCACT <u>CTTA</u> AGGACCT

Table 2.3: Mutations used in R2 variant Screen #3.

By column: TF motif refers to the predicted motif or conserved peak sequence for mutation; relevant consensus sequences are shown below with the match p-value by FIMO analysis. For Klf1 motifs, “L” and “R” refer to the left (unconserved) and right (conserved) motifs within block B. Coordinates refers to the motif position within the R2 MCS. Variant indicates the sequence type. Sequence shows the motif match or conserved peak region (underlined) with mutated bases highlighted in yellow.

Variant	Gata1			Nf-e2		Klf1	
	A	C	F	D	E	B L	B R
ΔGata1	Mut 2	Mut 1	Mut 1				
ΔNf-e2				Mut 2	Mut 1		
ΔKlf1						Mut 1	Mut 2
Gata1-only				Mut 2	Mut 1	Mut 1	Mut 2
Nf-e2-only	Mut 2	Mut 1	Mut 1			Mut 1	Mut 2
Klf1-only	Mut 2	Mut 1	Mut 1	Mut 2	Mut 2		
ΔTF	Mut 2	Mut 1	Mut 1	Mut 2	Mut 2	Mut 1	Mut 2

Table 2.4: Combinations of mutations used in R2 Screen #3 variants.

Motifs are arranged by type and, within type, block. “Mut 1” or “Mut 2” refers to the mutation sequences in Table 2.3.

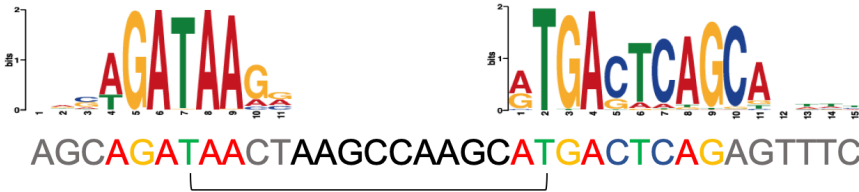
Variant	Sequence			
WT	 AGCAGATAACTAAGCCAAGCATGACTCAGAGTTTC			
5a	AGCAGAT	AA	C	A
5b	AGCAGAT	AA	G	A
10a	AGCAGAT	AACT	AA	GCA
10b	AGCAGAT	AACT	AGCA	A
15a	AGCAGAT	AACT	GCAACGACA	A
15b	AGCAGAT	AACTA	GCAACGA	CA
20a	AGCAGAT	AACTAAGCCAAGC		
		ATGCT	A	
20b	AGCAGAT	AACTA		
		GACCTGAGATCCA	A	
25a	AGCAGAT	AACTAAGCCAAGC	A	
		AAGCCAAGCA		
25b	AGCAGAT	AACT	GAATCGAC	
		CGAACAGAAC	CA	

Table 2.5: IIb spacing variant sequences used in R2 Screen #4.

Conserved Gata1 and Nf-e2 motifs are shown above the WT sequence and matching base pairs are coloured accordingly in all variant sequences. The varied IIb interval is highlighted with a bracket. Base pairs included in the motif region but without a significant match to the consensus sequence are shown in grey; other base pairs are shown in black and aligned to either side of the interval where applicable.

All plasmids were prepared by Plasmid Mini/Midi Kit column purification (Qiagen) from overnight cultures in LB media selected with ampicillin (50 µg/ml) or kanamycin (20 µg/ml) as required. Purified plasmid DNA was quantified by Nanodrop spectrophotometry (Thermo Fisher). Custom sequences were verified by Sanger sequencing prior to transfection into mES cells; guide RNA cutting efficiency was tested using the Surveyor nuclease assay (Qiu et al., 2004).

Plasmids were transfected into mES cells by lipofection as described for RMGR targeting, scaling up or down depending on the application and using a 3:1 HDR donor:guide plasmid(s) ratio by mass, when required. For the R2 variant screen, pools up to 120 individual HDR donors were co-transfected simultaneously by mixing verified plasmids at an equal ratio by mass prior to transfection. Subsequent transfection of a Dre recombinase plasmid to remove the puromycin / thymidine kinase selection cassettes was performed using the same lipofection protocol.

Successful transfection of CRISPR/Cas9 targeting plasmids was selected for by fluorescence activated cell sorting (FACS) of mESCs positive for GFP and/or mRuby fluorescence, as required (Figure 2.2). Integration of R2 variant sequences with associated resistance cassettes was positively selected for by treatment of cells with puromycin (Table 2.6). Following cassette excision by Dre recombinase, cells were negatively selected for cassette removal by treatment with ganciclovir (Table 2.6).

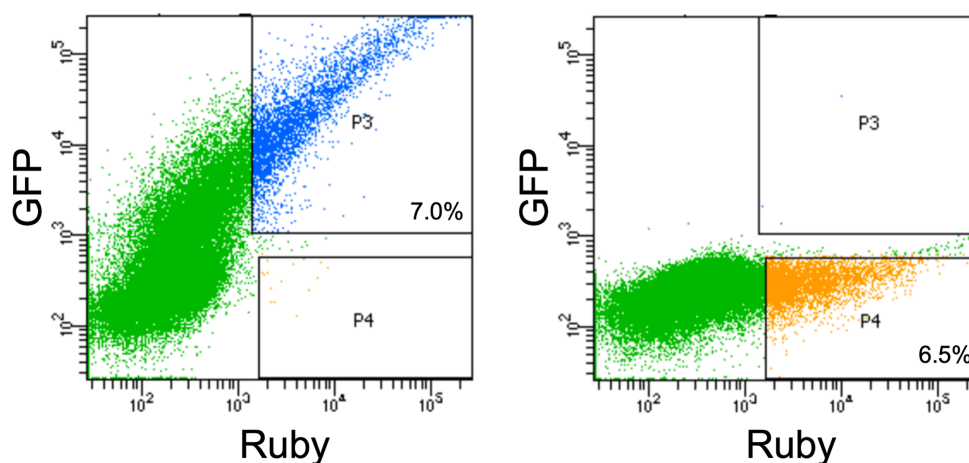


Figure 2.2: GFP/mRuby-positive selection for transfection of mESCs with CRISPR/Cas9 plasmids.

Left: representative flow cytometry plot for selection of double-positive GFP/mRuby mESCs. Right: flow cytometry plot for selection of single-positive mRuby mESCs.

Screening

Screening for desired genetic modifications was predominantly performed by PCR with or without Sanger sequencing, using the primers summarised in Table 2.7. Products were sequenced when annotating exact Cas9 cut sites or validating new screening strategies. Genomic DNA was obtained by incubating 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 (Thermo Fisher) at 37°C overnight. DNA was then either prepared by phenol-chloroform extraction, precipitated with isopropanol or added directly to a PCR reaction after Proteinase K inactivation (95°C, 5 minutes).

In the case of the R2-only model, the structural integrity of the locus was probed further with linked-read library preparation (10x Genomics) followed by Illumina sequencing, all performed at the Oxford Genomics Centre, Wellcome Centre for Human Genetics (Oxford, UK). See below section for further details: NGS sequencing.

For high-throughput clonal screening of integrated R2 variants, a PCR product over an incorporated restriction site was directly digested by ScaI (NEB), taking 7 µl into a 20 µl reaction. Non-digested and digested products were run out on an agarose gel to identify clones with a position-specific variant integration event. To confirm the integrity of expected variant sequences, the same PCR products were also cleaned up by treatment with Exonuclease I (NEB) and Shrimp Alkaline Phosphatase (NEB) and indexed for its well-of-origin in a second targeted PCR for NGS, as described (Table 2.8; Nussbaum et al., 2018). This approach uses custom indices to uniquely label PCR products from each clone of a 96-well plate by their row and column of origin.

For pooled screening of integrated R2 variants, see the below section: Measurement of gene expression.

Application	Selection	Concentration	Days before selection	Days of selection
RMGR <i>Hprt</i> ⁺	HAT	1x	2	7-14
RMGR <i>Hprt</i> ⁻	6-Thioguanine	10 μ M	5	5
R2 variants +	Puromycin	1 μ M	1	3
R2 variants <i>tk</i> ⁻	Ganciclovir	3 μ M	5	5

Table 2.6: Conditions for small molecule selection of mESC populations.

Application	Product	F/R	Sequence
RMGR	RMGR+ Over lox511	F	CTGTATTGAGCTCTGTGCGACATAAC TTCGTATAATGTATACT
		R	AGCAGCATCACACATCAGGCACAC ATGTAC
	R2-only+ Over R4	F	AGCTCTGAAATGGACGTGGT
		R	GAGCCGTCCTGTGTATGTGA
	RMGR- R1 internal	F	GGGCACAGCAAAAGAGGAAA
		R	AATGGTCCTTGTCTCCAG
	<i>Hprt</i> ⁺ Over loxP	F	ACTGGGGCTGAATCACATGA
		R	ACTCTTTCCCCTTGACCCAG
	Flp internal	F	ACGGAACAGCAATCAAGAGAGCCA
		R	TCGATCCTACCCCTTGCGCTAAA
Allele deletion	Over deletion	F	TGGATCTGTCCCATAGCAGAGGG
		R	TCCTGGGTACCATCCATCTTGGC
	Left junction	F	TGGATCTGTCCCATAGCAGAGGG
		R	TGGATCTGTCCCATAGCAGAGGG
	Right junction	F	ATCCCGGTCCTACCTCACTC
		R	TGGTGGAGTCTTCAGGGTCT
	Into α -YFP	F	GTGATTTGGTGCAAACTG

		R	CGTTGTGGCTGTTGTAGTTG
	Over α -deletion	F	GTGATTTGGTGCAAACTG
		R	AGGGTATCAGCCACCAATC
	Left RMGR cassette+	F	GGAGCTGGAGAAAGAGCAGA
		R	TCCACCTTCTCTTCCCACAC
	Left RMGR cassette-	F	GGAGCTGGAGAAAGAGCAGA
		R	GGGAGGGTTGAGGAGAGAAA
	Right RMGR cassette+	F	TTTATGGTTCGTGGGGGTTA
		R	ATACACCGAGGCATGACACA
	Right RMGR cassette-	F	ACTTCCCACACTTGGACCTG
		R	ATACACCGAGGCATGACACA
R1 deletion	Over deletion	F	TGTGTGTACAGTTTCCAGGAGGG
		R	CTTCCCTTAGGAAGGGATCTCCC
	Over deletion (large)	F	TTTCGGGAGCTGGGAGTAGA
		R	TTGAACCGTTAAACGCAGCAG
	R1 internal	F	GGGCACAGCAAAAGAGGAAA
		R	AATGGTCCTTGTCTCTCCAG
R2 deletion	Over deletion	F	AGGACTTCCTCTCCAGTACCC
		R	CCTTTAGTGTATGTCATTGAGCC
	R2 internal	F	CTCCTGGCCCATGTCTCTC
		R	TAACAGGCACAGGGTCACTT
R3 deletion	Over deletion	F	GAAGGCAGGAAAAGCACAGG
		R	TGAAGCATGTCTGTGGCTCT
Rm deletion	Over deletion	F	GCGTAGTCTAACTCCAGGCA
		R	AGCTGATTCCGTGAGAGACC
R4	Over deletion	F	AGCTCTGAAATGGACGTGGT
		R	GAGCCGTCCTGTGTATGTGA
R1 insertion	Into R1	F	CCGGGACTCTAGCCAACAG
		R	GACTCGAATGCTGTGCCAAG
R4 insertion	Into R4	F	CCCACAATAGCTGCCAAGGA
		R	CAACTTCCGACCCAGTGTGA
R2 variants	Over R2	F	CTCCTGGCCCATGTCTCTC

		R	GCCTGTATGAGCCAAGATGC
	Variant indexing	F	CTCCTGGCCCATGTCTCTC
		R	TAACAGGCACAGGGTCACTT

Table 2.7: Primers used for genome engineering screening.

ID	Barcode	Primer
iR_A	TATAGCCT	GAT <u>TATAGCCT</u> CTCCTGGCCCATGTCTCTC
iR_B	ATAGAGGC	GAT <u>ATAGAGGC</u> CTCCTGGCCCATGTCTCTC
iR_C	CCTATCCT	GAT <u>CCTATCCT</u> CTCCTGGCCCATGTCTCTC
iR_D	GGCTCTGA	GAT <u>GGCTCTGA</u> CTCCTGGCCCATGTCTCTC
iR_E	AGGCGAAG	GAT <u>AGGCGAAG</u> CTCCTGGCCCATGTCTCTC
iR_F	TAATCTTA	GAT <u>TAATCTTA</u> CTCCTGGCCCATGTCTCTC
iR_G	CAGGACGT	GAT <u>CAGGACGT</u> CTCCTGGCCCATGTCTCTC
iR_H	GTAAGTAC	GAT <u>GTAAGTAC</u> CTCCTGGCCCATGTCTCTC
iC_1	ATTACTCG	GAT <u>ATTACTCG</u> TAAACAGGCACAGGGTCACTT
iC_2	TCCGGAGA	GAT <u>TCCGGAGA</u> TAAACAGGCACAGGGTCACTT
iC_3	CGCTCATT	GAT <u>CGCTCATT</u> TAAACAGGCACAGGGTCACTT
iC_4	GAGATTCC	GAT <u>GAGATTCC</u> TAAACAGGCACAGGGTCACTT
iC_5	ATTCAGAA	GAT <u>ATTCAGAA</u> TAAACAGGCACAGGGTCACTT
iC_6	GAATTCGT	GAT <u>GAATTCGT</u> TAAACAGGCACAGGGTCACTT
iC_7	CTGAAGCT	GAT <u>CTGAAGCT</u> TAAACAGGCACAGGGTCACTT
iC_8	TAATGCGC	GAT <u>TAATGCGC</u> TAAACAGGCACAGGGTCACTT
iC_9	CGGCTATG	GAT <u>CGGCTATG</u> TAAACAGGCACAGGGTCACTT
iC_1	TCCGCGAA	GAT <u>TCCGCGAA</u> TAAACAGGCACAGGGTCACTT
iC_11	TCTCGCGC	GAT <u>TCTCGCGC</u> TAAACAGGCACAGGGTCACTT
iC_12	AGCGATAG	GAT <u>AGCGATAG</u> TAAACAGGCACAGGGTCACTT

Table 2.8: Primers used for indexed R2 variant NGS screening.

Underlined sequences represent unique barcodes used to label each row and column of a 96-well plate.

2.2 Cell culture and isolation

mES cell culture and differentiation

E14-TG2a.IV (E14) mESCs were maintained in gelatinised plates by standard methods (Smith, 1991; Nichols et al., 1990). 24-48 h prior to differentiation, cells were induced for differentiation by passaging into base media (Iscove's modified Dulbecco's medium (IMDM), 1.4×10^{-4} M monothioglycerol (Sigma-Aldrich) and 50 U/ml penicillin-streptomycin (Thermo Fisher)) supplemented with 15% heat-inactivated fetal calf serum (Δ FCS, Thermo Fisher) and 1000 U/ml LIF.

For embryoid body (EB) generation, cells were disaggregated by trypsinisation and quenched in base media (as above) supplemented with 10% Δ FCS. Differentiation media was prepared fresh on the day of differentiation by supplementing base media (as above) with 15% Δ FCS, 5% protein-free hybridoma medium (Thermo Fisher), 2 mM L-glutamine (Thermo Fisher), 50 μ g/ml L-ascorbic acid (Sigma Aldrich), 3×10^{-4} M monothioglycerol and 300 μ g/ml human transferrin (Sigma Aldrich). Cells were plated in either triple vent petri dishes (Thermo Fisher) or flat-bottom, uncoated 96-well plates (Thermo Fisher) at the following densities: for 10 cm dishes, 3×10^4 cells were added to 10 ml differentiation media; for 96-well plates, 200-400 cells were added to 200 μ l differentiation media. EBs were left to differentiate for up to seven days without disruption except for gentle manual shaking every few days to disrupt cell sticking.

EBs from 10 cm dishes were harvested by collection of the entire plate contents into tubes and a single wash step in PBS, spinning EBs at 1000 rotations per minute (rpm) for 5 minutes

to ensure complete recovery. After PBS removal, EBs were disaggregated in 0.25% trypsin (0.3 ml per 10 cm dish) for 3 minutes at 37°C with continuous manual shaking to prevent collection at the bottom of the tube. Trypsin was quenched with an equal volume of FCS and a single-cell suspension obtained through titration. Cells were collected by spinning at 1200 rpm for 5 minutes and resuspended as needed.

EBs from 96-well plates were prepared for analysis by performing a single PBS wash and disaggregation in a V-bottom 96-well plate. Whole EBs were transferred and washed using a multichannel pipette, resting tips on the kink of the V-bottom plate to avoid accidental EB suction during media removal steps. Disaggregation was performed using 50 µl 0.25% trypsin per well and quenched with an equal volume of FCS. Single-cell suspensions were obtained by pipetting up and down 20 times with the multichannel pipette. Wells were topped up to 300 µl and put directly for analysis by flow cytometry or first stained in 20 µl staining buffer (PBS with 10% FCS; 500 µl per 10⁷ cells), using the antibodies in Table 2.9, as required.

Isolation of in vitro-derived erythroid cells

Bulk populations of CD71-positive cells were isolated by magnetic column separation (LS Column, Miltenyi), according to the manufacturer's instructions. Briefly, cells were labelled for CD71 in staining buffer for 20 minutes at 4°C, rolling, then washed by adding staining buffer (1 ml per 10⁷ cells) and spinning. After supernatant removal, cells were incubated with MACS anti-FITC separation microbeads (Miltenyi; 10 µl per 10⁷ cells) in ice-cold separation buffer (PBS plus 0.5% bovine serum albumin (BSA) and 2 mM EDTA; 90 µl per

10^7 cells) for 15 minutes at 4°C, rolling, and washed by adding separation buffer (1 ml per 10^7 cells) and spinning.

Bead-labelled cells were resuspended in 500 µl cold separation buffer and added to a pre-equilibrated LS column. The negative fraction was washed through with two flushes of 3 ml cold separation buffer and the positive fraction collected by forcing cells from the column in 5 ml separation buffer. After spinning and supernatant removal, cells were resuspended in staining buffer as needed for downstream processing. Population purity and selection efficiency were determined by flow cytometry.

For applications requiring low cell numbers or additional immunophenotype selection steps, cell populations were isolated by FACS, using antibodies listed in Table 2.9 (Figure 2.3).

Marker	Fluor	Catalogue number	Working concentration
CD71	FITC	eBioscience 11-0711-85	2.5 µg/ml
CD71	APC	eBioscience 17-0711-80	0.1 µg/ml
Ter119	PE	BD Pharmingen 553673	1 µg/ml
c-kit	PE-Cy7	eBioscience 25-1171-82	2 µg/ml
Live/dead	Hoescht 33258	Invitrogen H3569	1 µg/ml
Table 2.9: Details of fluor-conjugated antibodies used for cell isolation and immunophenotyping.			

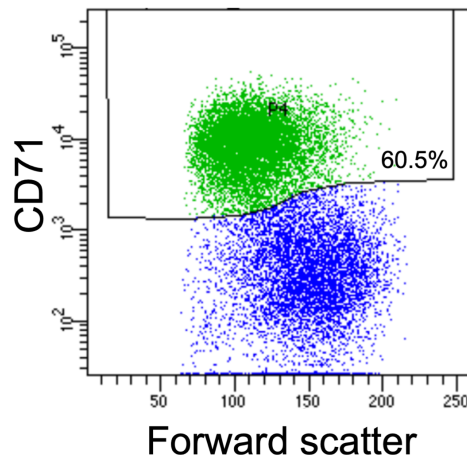


Figure 2.3: CD71-positive selection of EB-derived erythrocytes by FACS.

Flow cytometry plot for the isolation of CD71-positive erythrocytes (green). The erythroid phenotype was gated by both CD71 and size (forward scatter) to provide the best discrimination of the positive population.

Mouse model generation and breeding

Mouse model generation and handling was performed by the Mouse Transgenics Core Facility at the Weatherall Institute of Molecular Medicine. All work was conducted under Home Office licences after ethical review of the University of Oxford. In their hands, genetically engineered mES cells were screened for correct chromosome counts by karyotyping, microinjected into blastocysts from C57BL/6 mice and implanted into pseudo-pregnant females. Three resulting chimeric males were back-crossed with WT females. Between three litters from one chimera, 10 pups were identified with germline transmission, as assessed by agouti coat colour derived from the E14 mES cell background. Of these 10, five pups were confirmed to be heterozygous for the *Hprt*⁺ R2-only modification by PCR-based genotyping. Heterozygotes were crossed to a Flp-expressing line to promote recombinase-mediated excision of the *Hprt* cassette. *Hprt*⁻ R2-only heterozygotes were then back-crossed and inter-crossed to establish a Flp-negative heterozygous line and set up timed matings for embryo dissection.

Genotyping was performed on material from non-erythroid tissue: embryonic material (dissected embryos), ear notches (live pups) or brain (dead pups). DNA from ear notches and embryos was prepared using the DIRECTPCR-EAR PEQGOLD reagent (VWR) with 500 µg/ml Proteinase K treatment at 55°C overnight (Thermo Fisher). After Proteinase K inactivation for 5 minutes at 95°C, samples were spun at maximum speed for 3 minutes and supernatant added directly to standard PCR reactions using IMMOLASE DNA Polymerase (Bioline) supplemented with 1M betaine. Brain tissue was lysed overnight in a buffer of 50 mM Tris pH 8.0, 100 mM EDTA, 100 mM NaCl and 1% SDS with 50 µg/ml Proteinase K, then DNA was obtained by standard phenol-chloroform extraction and ethanol precipitation. Genotyping was performed using PCR primers detailed above for mESC clone screening.

Isolation of primary erythroid cells from embryos

Pregnant female mice were dissected at embryonic day 9.5 or 10.5 (E9.5/E10.5) after observation of a vaginal plug. Embryos including yolk sac material were individually collected into tubes containing PBS and manually disrupted to promote erythroid cell release from the circulation. The erythrocyte-containing supernatant was aspirated into fresh tubes and remaining embryo material used for genotyping as described above. Cells obtained in this way were found to be over 85% pure by flow cytometry for CD71/Ter119 and May–Grünwald–Giemsa staining (MGG; Figure 2.4).

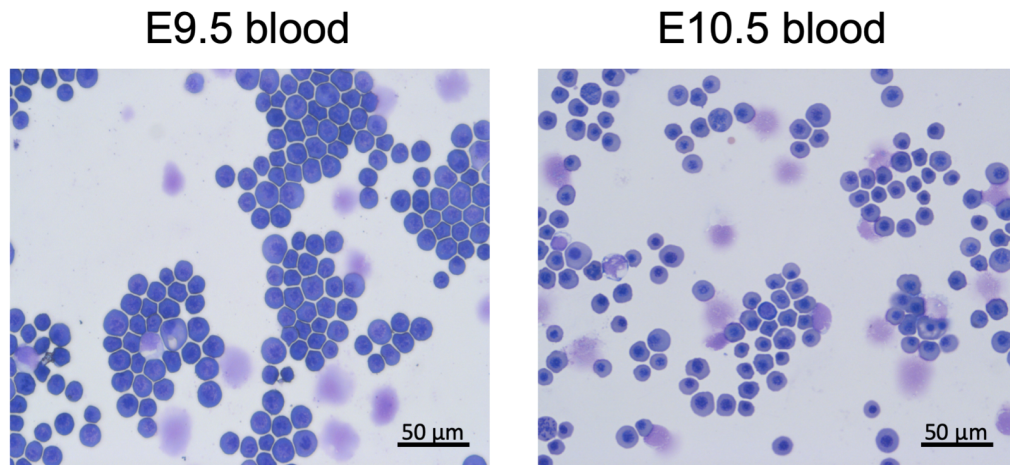


Figure 2.4: MGG stained cytopsin preparations of circulating blood collected from E9.5 and E10.5 embryos.

Cells at this timepoint are uniformly of the large basophilic erythroblast stage by morphology.

Preparation of primary cells for single-cell FACS

Embryos were dissected as above to obtain erythroid-enriched cells from E9.5/E10.5/E11.5. Fetal livers were dissected from embryos at E12.5 and disaggregated in ice cold staining buffer (0.2 % BSA and 5 mM glucose in PBS). Cells from spleens of acetylphenylhydrazine (APH)-treated adult mice (Spivak et al., 1973) were collected by manual tissue disruption in PBS supplemented with 10% FCS, then collected by centrifugation and resuspended in staining buffer to be processed in the same manner as the fetal liver tissue.

ChromePure Rabbit IgG (Jackson, 015-000-003) was added to a final concentration of 200 μg/ml to block Fc receptors. Cells were stained with Ter119:APC (BD; 1:200), CD71:PE-Cy7 (Biolegend; 1:1500), CD117:BB700 (BD; 1:100) and Lin:FITC (equal parts CD41, B220, CD3e, Mac-1 and Gr-1, all BD FITC conjugated) on ice for 45 min in the dark. Cells were washed twice with 1 ml ice-cold staining buffer then resuspended in FACS running

buffer (staining buffer plus 5 mM EDTA). Immediately before sorting, Hoechst (Invitrogen) was added at 1:10000 dilution).

Imaging cell cultures, EBs and cytopins

Colour images of cultured cells and EBs were captured with an EVOS cell imaging station (Thermo Fisher). Fluorescence microscopy images of day 7 EBs were captured with an IX51 instrument (OLYMPUS) using a filter for the yellow fluorescent protein (YFP) fluor and overlaid with brightfield images obtained in parallel using the Fiji software package (Schindelin et al., 2012). Cytopsin cell preparations were obtained by directing 10^5 cells at glass slides by centrifugation in 150 μ l PBS with 10% FCS for 4 minutes at 400 rpm. Hematek staining was performed according to manufacturer's instructions (Siemens).

2.3 Measurement of gene expression

RNA extraction and RT-PCR

For expression analyses, samples of 10^5 - 2×10^6 cells were lysed in TRI Reagent (Sigma) and immediately frozen at -80°C . RNA was extracted using a Direct-zol MicroPrep kit (Zymo Research) with a 30 minute DNase step on-column at room temperature. cDNA was then generated with Superscript III First-Strand Synthesis SuperMix (Life Technologies), including an RNase step to degrade remaining template molecules. Real-time PCR (RT-PCR) was performed with Fast SYBR Green Master Mix (Thermo Fisher) using the primers summarised in Table 2.10. Data were normalised to the 18S ribosomal gene or relevant β -globin genes using the ΔC_T method, taking the mean of multiple control C_T values when applicable. Results were plotted as individual points or averaged bar charts in R (R Core Team, 2017), using the ggplot2 package (Wickham, 2009) and colours from the ColorBrewer suite (www.ColorBrewer.org). RNA quality was routinely checked by Tapestation analysis (Agilent) and reverse transcriptase enzyme negative controls were included to confirm that DNase treatment was adequate.

Gene	Primer	Sequence	Primer target
<i>Hba-a1/2</i>	Forward	CTGGGGAAGACAAAAGCAAC	Exon 1
	Reverse	GCCGTGGCTTACATCAAAGT	Exon 2
<i>Hba-x</i>	Forward	ATGCGGTTAAGAGCATCGAC	Exon 2
	Reverse	GGGACAGGAGCTTGAAGTTG	Exon 2/3 junction
<i>Hbb-b1/2</i>	Forward	ACGATCATATTGCCAGGAG	Exon 3
	Reverse	ATGGCCTGAATCACTTGGAC	Exon 2
<i>Hbb-bh1</i>	Forward	TGGACAACCTCAAGGAGACC	Exon 2
	Reverse	ACCTCTGGGGTGAATTCCTT	Exon 3
<i>Hbb-y</i>	Forward	TGGCCTGTGGAGTAAGGTCAA	Exon 1
	Reverse	GAAGCAGAGGACAAGTTCCCA	Exon 2
<i>Rn18s</i>	Forward	GTAACCCGTTGAACCCCAT	Exon 1
	Reverse	CCATCCAATCGGTAGTAGCG	Exon 1
Table 2.10: Primers used for RT-PCR.			

Single cell RT-PCR

185 single erythroid cells were isolated by indexed FACS (16 cells each from mouse embryos at E9.5, E10.5 and E11.5; 42 cells each from spleens of APH-treated adult mice and 48 cells each from E12.5 embryos and EBs). Cells were sorted into wells of a 96-well plate containing 5.1 µl lysis buffer (5 µl 2x reaction buffer from SuperScript™ III One-Step RT-PCR with Platinum™ Taq kit (Thermo Fisher), 0.1 µl SUPERase In™ RNase Inhibitor (Thermo Fisher)). Plates were sealed, briefly centrifuged and frozen at -20 °C. Subsequent steps were performed by Dr Rob Beagrie. TaqMan assay mastermix was made by pooling 6.6 µl of 43 TaqMan assays (Table 2.11) and adding 376 µl TE buffer. Plates were defrosted and 4.9 µl amplification mix was added to each well (1.2 µl TE buffer (Invitrogen), 2.5 µl TaqMan assay mastermix, 1.2 µl RT/Taq enzyme mix from SuperScript™ III One-Step RT-PCR with Platinum™ Taq kit (Thermo Fisher)) except for one well containing a cell from

EBs, in which Platinum Taq was substituted for RT/Taq mix to serve as a no-RTase control. cDNA synthesis and sequence-specific preamplification was then carried out (reverse transcription at 50 °C for 15 min; RTase inactivation at 95 °C for 2 min; specific target amplification by 22 cycles of 95 °C for 15 s then 60 °C for 4 min). An RNA standard curve was also prepared at this point using the same mastermix and amplification program (see below). Plates were stored at -20 °C until analysis. Pre-amplified cDNA was diluted 5x with TE buffer and analysed using Universal PCR Master Mix (Applied Biosystems) and individual Taqman gene expression assays (Life Technologies, Table 2.11), on the Biomark System (Fluidigm) using two 192.24 Dynamic Arrays as per manufacturers protocol, except that elongation time was increased from 60 s to 75 s to accommodate longer custom gene-expression assays for *Hba-a1/2* and *Hbb-bt/s/1/2* (Table 2.12).

T7 polymerase was used to *in vitro* transcribe RNA from cDNA clones of *Cox6c*, *Sec61g*, *Oaz1*, *Hba-a1*, *Hba-x*, *Hbb-bt*, *Hbb-y* and *Hbb-bhl*. RNA was run on a 6% acrylamide TBE-urea gel (Life Technologies) and bands corresponding to full-length transcript were isolated from the gel. RNA was extracted by incubating crushed gel pieces with elution buffer (500 mM Ammonium acetate, 1 mM EDTA-KOH pH 8.0) for 3 hours at room temperature before ethanol precipitation, resuspension in TE buffer and quantification using RNA Quantifluor assay (Promega) and RNA Tapestation (Agilent). Each RNA was diluted to 16.8 pM concentration and the 8 transcripts were pooled, aliquoted and stored at -80 °C. When performing reverse transcription and pre-amplification on sorted single cells, one aliquot of RNA standard was also defrosted, and diluted 10x. Five more serial dilutions were created by diluting 6x in TE buffer, then 2 µl was taken from each of the six standard curve dilutions and pre-amplified using the same mastermix as used for single cells (see above).

Data analysis was performed using Python (v3.5.2; Matplotlib v2.0.2; Numpy v1.14.5; Pandas v0.23.3; scikit-learn v0.19.2; Scipy v1.1.0). C_T values beyond the limit of detection or marked as “Fail” by the instrument were set to the limit of detection value ($C_T = 40$). 21 cells were excluded from further analysis, 15 due to > 80% of assays being undetected and 6 due to outlier C_T values. Raw C_T values were transformed into expression space by calculating 2^{-C_T} . Three housekeeping genes (*Cox6c*, *Myl6* and *Sec61g*) were run on both 192.24 chips. The expression of these three genes was scaled to a range of 0 to 1, then the mean of the three scaled expression values was calculated for each cell and used to normalise expression values for the remaining 21 assays on each chip. This normalisation was performed separately for the two chips before merging the data together for analysis.

Gene Name	Assay ID
<i>Abcg4</i>	Mm00507247_m1
<i>Aqp1</i>	Mm01326466_m1
<i>Aqp8</i>	Mm00431846_m1
<i>Arid3a</i>	Mm00492248_m1
<i>Ccr2</i>	Mm00438270_m1
<i>Cited2</i>	Mm00516121_m1
<i>Cox6c</i>	Mm00835813_g1
<i>Cpeb4</i>	Mm01193440_m1
<i>Ddx21</i>	Mm00497941_m1
<i>Foxh1</i>	Mm00514851_m1
<i>H2afy</i>	Mm01337623_gH
<i>Hba-a1/a2</i>	Custom (IDT)
<i>Hba-x</i>	Mm00439255_m1
<i>Hbb-bh1</i>	Mm00433932_g1
<i>Hbb-bh2</i>	Mm01273444_g1

<i>Hbb-bt/bs/b1/b2</i>	Custom (IDT)
<i>Hbb-y</i>	Mm00433936_g1
<i>Hbq1a</i>	Mm00731011_s1
<i>Hbq1b</i>	Mm02747875_s1
<i>Hspe1</i>	Mm00434083_m1
<i>Il9r</i>	Mm00434313_m1
<i>Klf3</i>	Mm00492956_m1
<i>Lmo4</i>	Mm00495373_m1
<i>Mbd2</i>	Mm00521967_m1
<i>Mpg</i>	Mm00447872_m1
<i>Mpo</i>	Mm01298424_m1
<i>Myc</i>	Mm00487804_m1
<i>Myl6</i>	Mm02342525_g1
<i>Nprl3</i>	Mm01193449_m1
<i>Oaz1</i>	Mm01611061_g1
<i>Pbx1</i>	Mm04207617_m1
<i>Podxl</i>	Mm00449829_m1
<i>Pogz</i>	Mm00554475_m1
<i>Prtn3</i>	Mm00478323_m1
<i>Rhbdf1</i>	Mm00711711_m1
<i>Rhd</i>	Mm00456910_m1
<i>Sec61g</i>	Mm01613099_g1
<i>Selk</i>	Mm00785961_s1
<i>Smarca4</i>	Mm01151944_m1
<i>Snrnp25</i>	Mm00547218_m1
<i>Sox4</i>	Mm00486320_s1
<i>Sox6</i>	Mm01274768_m1
<i>Tmem14c</i>	Mm00481276_m1

<i>Vamp5</i>	Mm00444144_m1
<i>Zbtb7a</i>	Mm00657132_m1
Table 2.11: Taqman probes used for single-cell RT-PCR.	

Assay	Left primer	Right primer	Internal oligonucleotide
<i>Hba-a1/a2</i>	TGAAGCCCTGG AAAGGATGT	TGAAATCGGCAG GGTGGT	TGGATCCCGTCAACTT CAAGCTCCT
<i>Hbb-bt/bs/b1/b2</i>	CCGATGAAGTT GGTGGTGAG	ACATGCAGCTTG TCACAGTG	CCTGGGCAGGCTGCTG GTTG
Table 2.12: Custom Taqman designs used for single-cell RT-PCR.			

YFP tag design and clonal readout

A mouse ES cell line carrying an mVenus YFP tag (Nagai et al., 2002; Shaner et al., 2005) on the enhancer-proximal copy of the α -globin gene was designed and generated by Dr Danuta Jeziorska and Dr Andrew King. YFP was chosen due to its compatibility with the absorbance spectrum for the haemoglobin protein and a nuclear localisation signal was added to further reduce the chance of signal quenching from cytoplasmic haemoglobin (Papadopoulos et al., 2012; Zijlstra et al., 1991). The median fluorescence level of this tag for erythroid populations of interest was used as a proxy for gene expression by subjecting tagged mESC clones to *in vitro* differentiation into EBs as described above and flow cytometry with an Attune NxT instrument (Thermo Fisher). Subsequent analyses were performed with the FlowJo software package (BD).

Recordings from 96-well plates were made by combining EBs from three replicate plates into a single well for washing and disaggregation steps (described above). Two individual recordings were made for each experimental replicate (i.e. six replicate plates in total, split

into two recordings combined from three plates each). To remove unreliable results, data from individual wells were discarded if A) the difference between duplicate fluorescence recordings was greater than 300 (arbitrary units) or B) both recordings measured fewer than 1000 cells in total.

Pooled R2 variant sequencing

Pooled screening of integrated R2 variants was performed by subdividing the YFP+ erythroid population from EBs generated from pools of targeted mESCs into eight gates (Figure 2.5). An overview of the analysis workflow is shown in Figure 2.6. Resulting populations were lysed with 1 µl lysis buffer (above) per 1000 cells and digested overnight at 37°C. After Proteinase K inactivation, 2 µl of each sample was taken into a 20 µl PCR reaction to amplify a region of R2 that extends beyond the homology arms of the HDR donor vector pool, to specifically amplify only correctly integrated copies of the R2 variant library. 1 µl of each of the resulting products was cleaned up by Exonuclease I / Shrimp Alkaline Phosphatase treatment (as above) and directly used in a targeted indexing PCR, re-purposing the 96-well plate indexing barcodes to differentially index products from individual YFP gates (Table 2.8; Nussbaum et al., 2018). Products from each gate were pooled equally using quantitation on an agarose gel with the Fiji software package (Schindelin et al., 2012). NGS libraries were prepared and sequenced as previously described (Nussbaum et al., 2018). Indexed NGS files for individual YFP fluorescence gates were flashed and demultiplexed using an in-house pipeline as described, with settings `--ampliconSize 400 --minPCRcount 10 --minLen 270 --minIdentity 85 --flashOverlap 10 --minMatch 200` (Nussbaum et al., 2018).

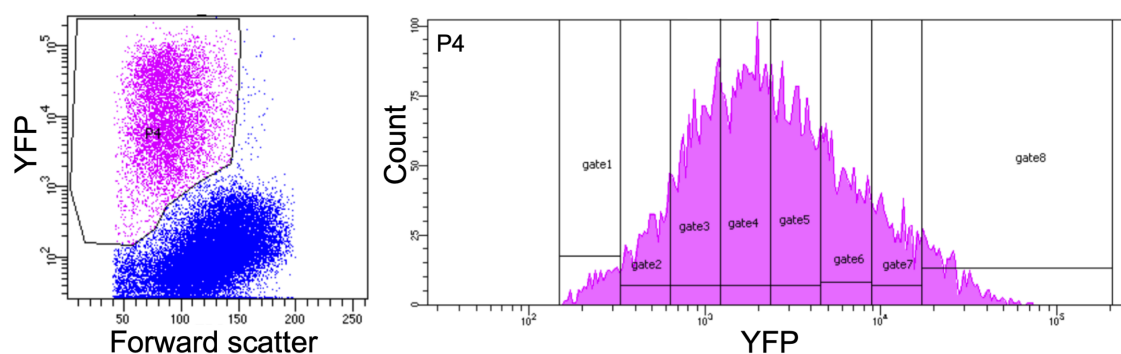


Figure 2.5: Division of a YFP-positive EB-derived population into eight gates to screen pooled R2 variant sequences.

Left: Flow cytometry plot showing the gating strategy to select the YFP-positive population from day 7 EBs. Selection was gated based on YFP fluorescence and cell size (forward scatter) for the best discrimination of the positive population. Right: segregation of the YFP-positive population into eight gates for sorting into individual tubes and subsequent R2 variant genotyping.

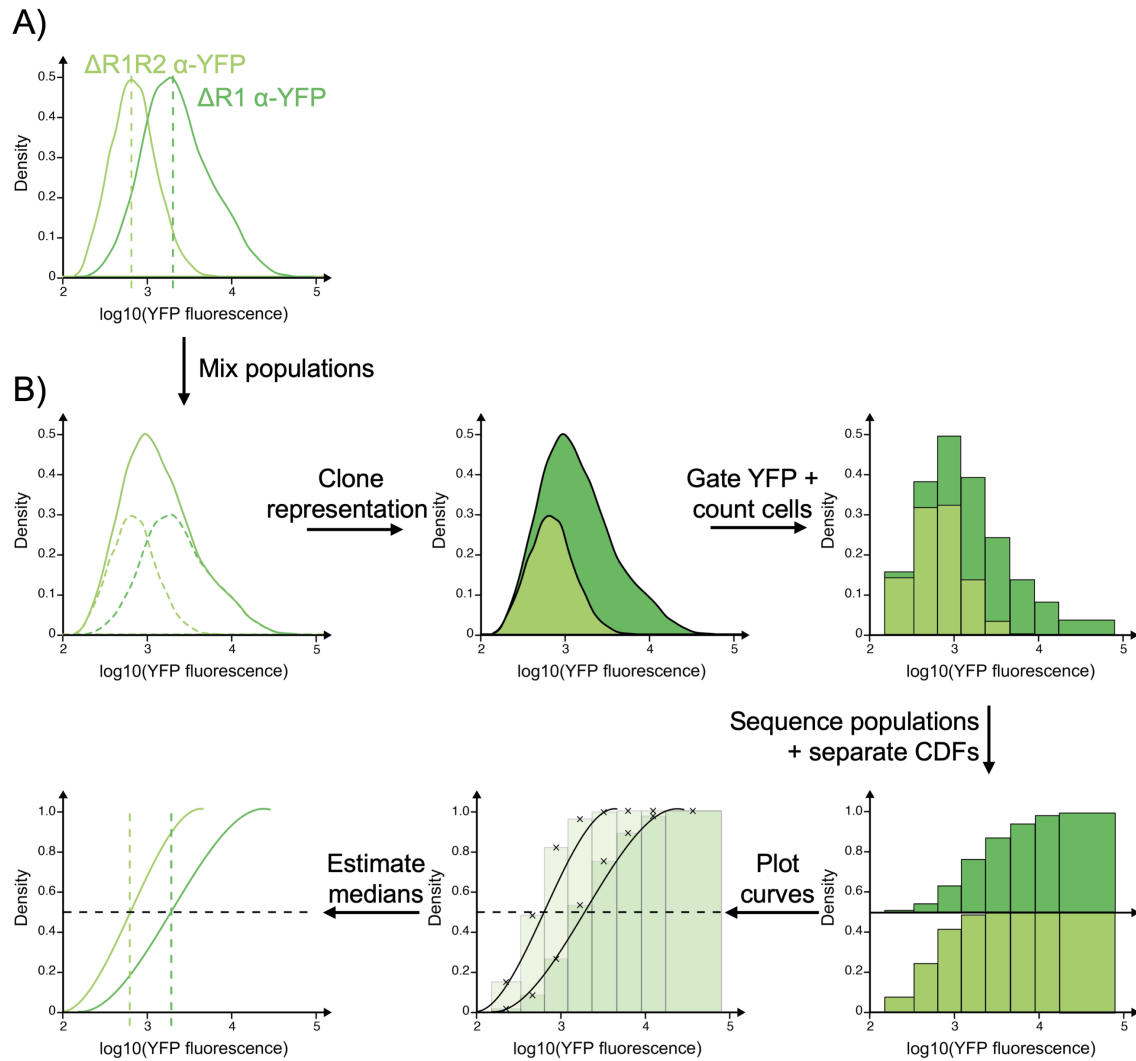


Figure 2.6: Estimation of YFP median fluorescence readings by NGS variant demultiplexing.

A) Erythroid-specific $\alpha\text{-YFP}$ fluorescence distributions for ΔR1 and ΔR1R2 models as shown in Figure 6.4. **B)** Demonstrative pooling and subsequent YFP median estimations of clones with different fluorescence levels. Pooling populations produces a combined fluorescence distribution (top left). This can be considered in terms of additive representation from the two individual clones (top middle). As a result, eight gated populations for YFP intensity have different representations of the two individual clones by cell counts, which can be inferred from NGS genotyping data from each population (top right). Cumulative density function approximations can then be plotted for each clone (bottom right) and sigmoidal curves plotted (bottom middle). Finally, intersection points between cumulative density curves and $y=0.5$ can be used to quantify approximate median YFP values (bottom left). Data shown are from simulated pooling, gating and cell counts and curve fitting of individual flow cytometry recordings from $\Delta\text{R1 } \alpha\text{-YFP}$ and $\Delta\text{R1R2 } \alpha\text{-YFP}$ clones.

Sequence attributes specific to each of the R2 variants included in a single experiment were then used to search for and count matches across the reconstructed R2 sequences, using custom scripts in R or bash. PCR counts per variant per gate were scaled to the recorded number of cells sorted in each gate to give an *in silico*-generated “cell count” for each data point. Results were filtered to remove data from any variants represented by fewer than 8 cells across all gates or variants for which no reads were detected in any one of gates 1-5. Cumulative cell counts were plotted as a % of the total cell count for each variant across the eight sorted gates and sigmoidal curves fitted in R. The point of intersection of each cumulative density function with $y=0.5$ was recorded as the estimated median. This was performed for each technical replicate and the combined cell counts across all two or four independent experiments to provide an indication of the most accurate expectation value and the range of noise inherent in the protocol. Resulting estimated medians were plotted for each variant in R (R Core Team, 2017), using the ggplot2 package (Wickham, 2009) and colours from the ColorBrewer suite (www.ColorBrewer.org).

2.4 NGS assays

ATAC-seq and ChIP-seq

Assay for transposase-accessible chromatin (ATAC)-seq was performed on $3\text{--}7 \times 10^5$ cells from target populations as previously described (Buenrostro et al., 2013, Hay et al., 2016). Chromatin immunoprecipitation (ChIP)-seq was performed using the ChIP Assay Kit (Millipore) as described previously (Hanssen et al., 2017) on aliquots of 5×10^6 CD71-positive cells derived from two separate differentiation experiments per antibody. Chromatin fragmentation was found to be optimal for 8 minutes rather than 15-20 minutes as described, accounted for by the lower cell number used. Antibodies used for ChIP-seq in this thesis are summarised in Table 2.13.

Enrichment by ChIP was assessed by RT-PCR prior to library indexing and NGS, using Fast SYBR Green Master Mix (Thermo Fisher) and the primers summarised in Table 2.14. Enrichment was calculated by comparing the amplification of immunoprecipitated genomic DNA with a dilution series of unprecipitated input material (5%, 1%, 0.2%, 0.04%) for a region of predicted enrichment and a non-specific control region. Primer pairs were validated by standard curve and melt curve analysis prior to use.

Data for ATAC-seq and ChIP-seq were analysed with an in-house pipeline as described (Hay et al., 2016; Telenius and Hughes, 2018). In each case, alignment files from two or three biological replicates were individually normalized to reads per kilobase per million mapped reads (RPKM) using the deepTools package (Ramírez et al., 2016) and averaged

using bedtools (Quinlan, 2014) with the awk bash command. Tracks were visualised with the University of California Santa Cruz (UCSC) genome browser (Kent et al., 2002).

To compare datasets genome-wide, triplicate data for each tissue of interest were peak-called with the Model-based Analysis of ChIP-Seq tool (MACS2, Zhang et al., 2008), using default parameters. Principal component analysis (PCA) was then performed using the DiffBind package in R (R Core Team, 2017).

Protein target	Antibody	Working concentration
H3K27ac	Abcam ab4729	0.5 µg/ml
H3K4me1	Abcam ab8895	0.3 µg/ml
H3K4me3	Abcam 8580	2.5 µg/ml
CTCF	Merck Millipore 07-729	5 µg/ml
Gata1	Abcam ab11852	3.25 µg/ml
Nf-e2	Santa Cruz sc-22827	1 µg/ml
Table 2.13: Antibodies used for ChIP.		

Product	Applications	F/R	Sequence
Chr15 region	H3K27ac -ve	F	GTAAGATGCCATTGTCCTGACC
	H3K4me1 -ve	R	TGTGTTCTCCTCCCTTCCTT
<i>Hba</i> promoter	H3K27ac +ve	F	CTAGCTTCCCCACCACCAAG
	H3K4me3 +ve	R	CTTCTTGCCGTGACCCTTGA
R2 enhancer	H3K4me1 +ve	F	GTACAGAGTCAGAAAGGAAAGGACAA
	Gata1 +ve	R	GGCAAATAAACAGGATGGGAAAAC
<i>Nprl3</i> gene body	CTCF -ve	F	CAGAGCTCCCAGACAACCAG
		R	AGCACACCCAGGGTTCTCTA

HS-38 CTCF	CTCF -ve	F	TGAGAAGGCTGGCCTTTGAG
		R	CCCAGGGAATGAATGCCAGT
Table 2.14: Primers for ChIP-RT-PCR.			

Motif analysis

Position weight matrices for selected erythroid TFs were generated using the Multiple Em for Motif Elicitation (MEME) tool (Bailey and Elkan, 1994) with peak calls from ChIP-seq data for primary mouse erythrocytes (Figure 5.13; Hay et al., 2016). Peak call datasets were produced using the MACS2 tool (Zhang et al., 2008) with default settings and filtered to take the highest 2000 peaks by motif enrichment for Gata1, Tal1 and Nf-e2 motifs or the top 975 peaks for Klf1, based on an estimation of several thousand peaks in previous ChIP-seq experiments for Gata1, Tal1 and Nf-e2 (Han et al., 2015; Steiner et al., 2010) and ~1000 peaks in a previous peak call of the Klf1 ChIP-seq data used here (Tallack et al., 2010). Appropriate motifs were selected based on similarity to published consensus sequences (Figure 2.7; Han et al., 2015; Igarashi et al., 1994; Tallack et al., 2010). Motif searches were performed using the FIMO tool (Grant et al., 2011) as described in the main text.

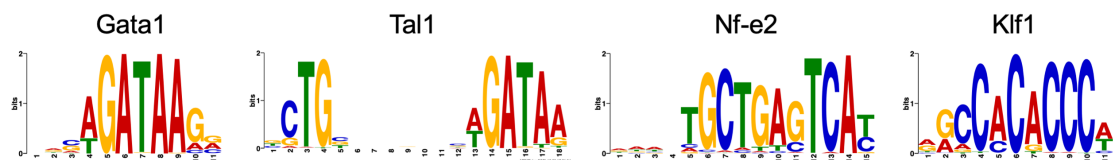


Figure 2.7: MEME-generated position weight matrices used for TF motif calling.

Capture-C

Next-generation Capture-C was performed as described (Davies et al., 2016) on samples of 2×10^7 Ter119+ cells isolated from WT C57BL/6 APH-treated adult mouse spleen (Spivak

et al., 1973). A summary of capture probes used to enrich for sequences of all gene promoters in a 5.6 Mb region surrounding the α -globin locus (except for the *Acyp2* gene for which no suitable capture probes could be designed) is shown in Table 2.15; these were quality filtered using the Capsequm tool (Hughes et al., 2014). Data were analysed using an in-house pipeline as previously described (Davies et al., 2016) and visualised as an overlay track in the UCSC genome browser (Kent et al., 2002).

Gene	Probe	Chr11 bp start	Chr11 bp end
<i>Pnpt1</i>	Left	29030334	29030454
	Right	29030768	29030888
<i>Smek2</i>	Left	29072384	29072504
	Right	29072733	29072853
<i>Ccdc104</i>	Left	29146517	29146637
	Right	29148002	29148122
<i>Ccdc88a</i>	Left	29271992	29272112
	Right	29274612	29274732
<i>Prorsd1</i>	Left	29414607	29414727
	Right	29415708	29415828
<i>Mtif2</i>	Left	29426038	29426158
	Right	29426926	29427046
<i>Rps27a</i>	Left	29447865	29447985
	Right	29448378	29448498
<i>Rtn4</i>	Left	29592839	29592959
	Right	29594046	29594166
<i>Eml6</i>	Single	29925992	29926112
<i>Spnb2</i>	Left	30098164	30098284
	Right	30098422	30098542
<i>Psme4</i>	Left	30671587	30671707
	Right	30671958	30672078
<i>Gpr75</i>	Left	30785163	30785283

	Right	30785458	30785578
<i>Erlec1</i>	Left	30853716	30853836
	Right	30855248	30855368
<i>Chac2</i>	Left	30885726	30885846
	Right	30886508	30886628
<i>Stc2</i>	Left	31269517	31269637
	Right	31270137	31270257
<i>Bod1</i>	Left	31570728	31570848
	Right	31572074	31572194
<i>Cpeb4</i>	Left	31769089	31769209
	Right	31771337	31771457
<i>Nsg2</i>	Left	31899865	31899985
	Right	31900315	31900435
<i>Snrnp25</i>	Left	32104921	32105041
	Right	32105415	32105535
<i>Rhbdf1</i>	Left	32122089	32122209
	Right	32122573	32122693
<i>Nprl3</i>	Left	32167508	32167628
	Right	32168252	32168372
<i>Mpg</i>	Left	32126341	32126461
	Right	32126554	32126674
<i>Il9r</i>	Left	32099819	32099939
	Right	32100179	32100299
<i>Hba-x</i>	Left	32176201	32176321
	Right	32176810	32176930
<i>Hba-a1</i>	Left	32182969	32183089
<i>Hba-a2</i>	Left	32195804	32195924
<i>Hba-a1/2</i>	Right	32196518	32196638
<i>Sh3pxd2b</i>	Single	32247535	32247655
<i>Ubt2</i>	Left	32354547	32354667
	Right	32355831	32355951
<i>Efcab9</i>	Left	32427057	32427177

	Right	32427737	32427857
<i>Stk10</i>	Left	32431986	32432106
	Right	32433397	32433517
<i>Fgf18</i>	Single	33048574	33048694
<i>Npm1</i>	Left	33063163	33063283
	Right	33063470	33063590
<i>Tlx3</i>	Left	33103236	33103356
	Right	33103547	33103667
<i>Ranbp17</i>	Left	33412705	33412825
	Right	33413585	33413705
<i>Gabrp</i>	Left	33478834	33478954
	Right	33479046	33479166
<i>Kcnip1</i>	Left	33742084	33742204
	Right	33743535	33743655
<i>Kcnmb1</i>	Left	33862918	33863038
	Right	33863447	33863567
<i>Lcp2</i>	Left	33946817	33946937
	Right	33947072	33947192
<i>Foxi1</i>	Left	34107595	34107715
	Right	34109082	34109202
<i>Dock2</i>	Left	34596916	34597036
	Right	34597560	34597680
<i>Ccdc99</i>	Single	34644236	34644356

Table 2.15: Genome coordinates of oligonucleotide probes for Capture-C (mm9).

Two probes of 120 bp each were designed to the DpnII fragment for the promoter region of each gene, labelled left and right. Where individual probes did not pass filtering requirements, a single probe was used to target the DpnII fragment, labelled single.

BAC sequencing

BACs were prepared for short-read sequencing by sonicating caesium chloride-purified DNA with a Covaris S220 focused ultrasonicator as described for the Capture-C protocol

(Davies et al., 2016). Sonicated DNA was purified using Ampure XP SPRI beads (Beckman Coulter) according to the manufacturer's protocol and assessed with the DNA1000 Screen Tape (Agilent Technologies). Resulting fragments were around 200 bp in size. Libraries were indexed using the NEBNext DNA Library Prep Kit for Illumina (NEB) and sequenced with a Miseq instrument (Illumina). Reads were aligned to BAC design sequences using the Bowtie 2 tool (Langmead and Salzberg, 2012). Alignment files were sorted and variants were called using the SAMtools suite (Li et al., 2009; Li, 2011).

10x linked-read analysis

High molecular weight genomic DNA was isolated for linked-read library preparation by a modified version of the regular protocol used for genotyping PCRs (described above). The procedure consisted of lysis of a large number of cells (10^7) overnight, precipitation with isopropanol, extraction using phenol:chloroform and washes with 70% ethanol. Extreme caution was taken in each step to avoid DNA shearing; a Pasteur pipette was used for transfer steps. DNA quality was tested by genomic DNA Tapestation analysis (Agilent). Linked-read libraries were prepared and sequenced according to the manufacturer's instructions (10x Genomics). Fastq files were analysed using the Long Ranger tool (10x Genomics), using mm9 as the reference genome with freebayes variant calling (Garrison and Marth, 2012). The resulting loupe file was viewed with Loupe 2.1.2 software (10x Genomics).

Chapter 3 – A super-enhancer reduced to a single element

3.1 Introduction

Previous functional dissection of the mouse α -globin super-enhancer

There are several proposed mechanisms for the formation and co-operative function of clusters of enhancer elements (Long et al., 2016; Pott and Lieb, 2015). But to what extent are co-operative relationships of proposed super-enhancers realised in normal gene function? Recent work in the Higgs lab genetically interrogated the mouse α -globin super-enhancer *in vivo* with a series of single-enhancer deletions and two double-enhancer deletions (Hay et al., 2016). Analysis of *ex vivo* cultured erythrocytes demonstrated that the elements of the super-enhancer became accessible during erythropoiesis irrespective of the presence or absence of the surrounding enhancers. The formation of the erythroid-specific self-interacting domain was also unaffected by all tested enhancer deletions. These results indicated that the chromatin structure of the locus was not dependent on the presence of a complete enhancer cluster.

Quantitation of transcriptional output from the α -globin genes further suggested that the enhancers may contribute independently to gene regulation (Figure 3.1). The R1 and R2 enhancers were identified as the most prominent activators of gene expression based upon their deletion phenotypes, whereas the R3 and Rm enhancer deletion models showed no notable phenotype. The R4 enhancer deletion model showed a small but significant reduction in α -globin expression. These data are in-line with predicted relative enhancer

activities by TF binding, activity in transient and stable enhancer assays and sequence conservation (Hay et al., 2016; Hughes et al., 2005).

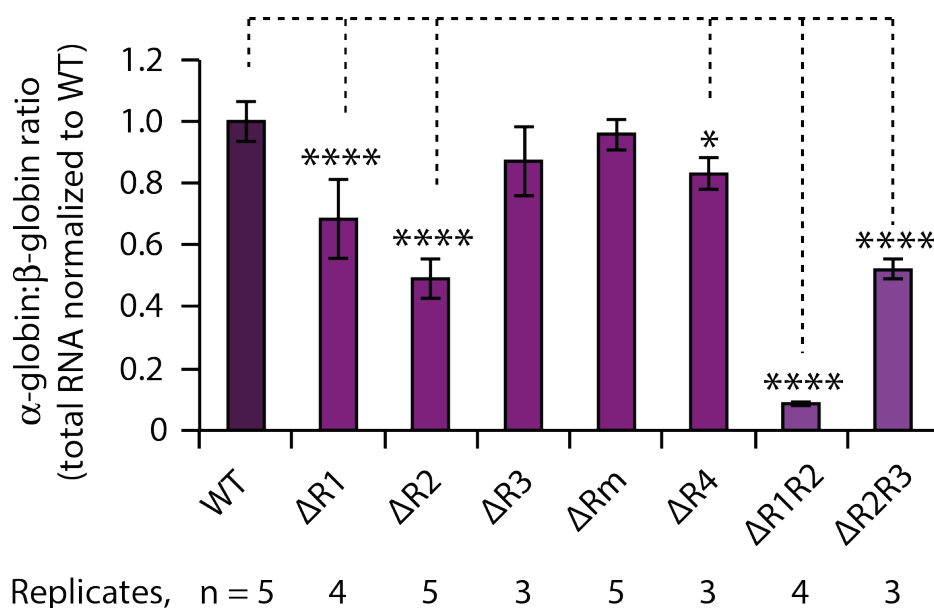


Figure 3.1: α -globin gene transcription in single- and double-enhancer deletion models.

Quantitation of mature adult *Hba-a1/2* α -globin transcripts normalised to the mature adult *Hbb-b1/2* β -globin gene for single- and double-enhancer deletion models. Data were obtained by Nanostring transcript quantitation of material from intermediate erythrocytes derived from E12.5 mouse fetal liver. Figure adapted from Hay et al., 2016.

Results from two combined deletion models, $\Delta R1R2$ and $\Delta R2R3$, showed a remarkable additive property in their phenotypes. Each double-enhancer mutant was attenuated for α -globin expression as if the effects of deleting the two removed enhancers had been summed. This was strong evidence that the elements of the α -globin super-enhancer are able to form and act independently from each other, each with different defined contributions to target gene expression. Rather than suggesting a synergistic model for super-enhancer function with inherent properties beyond the sum of its parts, these data fit best with an additive model for individual activities in a cluster of enhancers (Dukler et al., 2016; Pott and Lieb, 2015).

Predictions for the R2-only model

In this chapter, I set about creating a mouse model in which the activity of the R2 enhancer is tested individually, in a model devoid of the other elements of the super-enhancer. In a purely additive model of super-enhancer activity, the contributions of individual enhancer elements to gene activity would not necessarily be equivalent but would sum to 1. Based on analysis of the gene transcription data shown in Figure 3.2, this model holds for the α -globin locus. By individual element deletions, three enhancers (R1, R2 and R4) show different but individually significant contributions to α -globin gene activation (Hay et al., 2016), while the remaining two elements, R3 and Rm, show no detectable activity. In numbers, R1 deletion reduces expression by a relative proportion of 0.32, R2 by 0.51 and R4 by 0.17. Taking these values as the active element contributions, the sum of enhancer activities does, indeed, sum to 1 ($0.32 + 0.51 + 0.17 = 1$; Figure 3.2).

Expression measurements from double enhancer deletions of R1/R2 and R2/R3 also roughly conform to this model: given the apparent inactivity of the R3 and Rm elements, the $\Delta R1R2$ model should be driven by R4 alone (0.17 expression level predicted) and is measured at 0.09; $\Delta R2R3$ should be driven by R1 and R4 ($0.32 + 0.17 = 0.49$ expression level predicted) and is measured at 0.52. The disparity between the predicted and actual values for $\Delta R1R2$ mice may be indicative of the potential oversimplification of a purely additive model of gene activation when considering transcriptional activation over a wide range of expression levels. Such a shortcoming becoming apparent for the genetic mutant in which the two most active enhancers of the locus are deleted is perhaps unsurprising, given the molecular complexity of gene activation.

While mindful of these considerations, an additive model of super-enhancer function would predict an R2-only mouse to express α -globin to a level of 0.51 relative to wild type (WT). As this reflects only a modest change in gene expression, I concluded that the R2-only mouse model would likely express α -globin in line with the additive prediction and, therefore, be viable into adulthood as a homozygote.

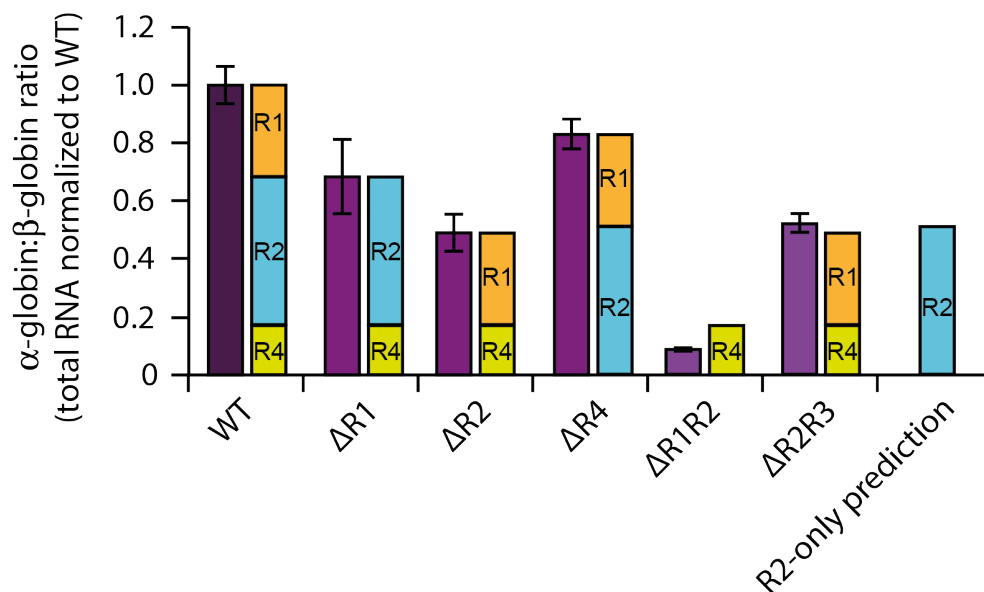


Figure 3.2: Prediction for α -globin expression levels across active enhancer deletion mouse models according to an additive model of super-enhancer function.

Purple bars represent data from Hay et al., 2016; accompanying stacked bars show predicted expression levels based on defined contributions from R1, R2 and R4. Remaining super-enhancer elements R3 and Rm were not found to actively contribute to α -globin expression and so are not shown.

In this chapter, I first outline the case for using the mouse α -globin super-enhancer as a model for local gene activation by testing the specificity of its activity within a 5 Mb window. Next, I describe the genome engineering procedure used to generate a mouse model with deletions of four spatially separated enhancer elements. Finally, I present *in vivo* data for the R2-only mouse and evaluate its fit against the predicted outcome.

3.2 Results

The local activity of the mouse α -globin super-enhancer

Previous experiments from the Higgs group have established that the syntenic region containing the α -globin genes contains all the necessary elements to coordinate appropriate gene expression (Higgs et al., 1990; Wallace et al., 2007). Before genetically interrogating the α -globin super-enhancer, it was important to establish, in addition, that the syntenic region acts only on the predicted globin genes, such that genetic manipulations would not create confounding phenotypes due to indirect effects. In other words, it is required that the activity of the regulatory domain is not only necessary and sufficient for expression of the α -globin genes, but also specific and unique.

Mouse models harbouring mutations of the α -globin regulatory domain have shown that non-globin genes in and immediately surrounding the α -globin self-interacting domain can be affected by perturbations of enhancer and boundary elements. RNA-seq data from Δ R1R2 erythroid cells showed no significant transcriptional changes except for α -globin (*Hba-a1/2* genes) and two additional genes at the locus: *Mpg* and *Il9r* (Martin Larke, unpublished data) and no significant global changes in element accessibility by assay for transposase-accessible chromatin using sequencing (ATAC-seq; Hay et al., 2016). Parallel work altering the 3D structure of the locus by CTCF site deletion gave similar results: the only significantly altered expression levels were those of the nearby *Mpg*, *Rhbdf1* and *Snrnp25* genes (Hanssen et al., 2017). The closest genes on the 3' side of the locus are not expressed in normal erythroid tissue and do not become activated in any of the published models or upon disruption of CTCF binding on that side (Yue et al., 2014; Caroline Harrold,

unpublished data). Encouragingly, no confounding phenotype issues were evident as a result of altered local gene expression of the 5' genes; even the severely deleterious $\Delta R1R2$ enhancer deletion mouse model allows for the isolation of *bona fide* erythroid cells from mouse embryos and fetal livers, despite α -globin expression being reduced to 10% of WT levels (Hay et al., 2016).

Beyond the genes immediately surrounding the α -globin locus, it is possible that long-range enhancer-promoter interactions with the α -globin super-enhancer may exist, and could be affected by novel mutations. In human erythroid cells, although not functionally critical for the production of erythroid cells, expression of the *NME4* gene located 300 kb from the α -globin cluster is affected by disruption of the α -globin genes or enhancers, with an accompanying change in direct interactions with the region (Lower et al., 2009). Interactions can also be detected with a gene located 625 kb from the cluster, *FAM173a* (Hughes et al., 2013).

As exemplified by the case for *NME4*, the accepted dogma for enhancer activity involves physical proximity between enhancers and their target promoters. The observed correlation between increased interaction frequency and upregulated gene expression is unequivocal, regardless of the underlying mechanisms concerning the specificity, establishment and maintenance of those interactions (Dekker and Misteli, 2015; Vermunt et al., 2019). Therefore, any genes likely to be susceptible to changes in gene expression upon mutation of enhancer elements would be expected to demonstrate interactions in WT tissue that could become perturbed in genetic mutants.

To confirm that additional promoters would be unaffected by novel manipulation of the mouse super-enhancer via altering long-range interactions, I examined existing Capture-C data from the R2 enhancer (Oudelaar et al., 2018) and performed a reciprocal Capture-C experiment for gene promoters within a 5 Mb region encompassing the α -globin locus in Ter119+ adult mouse erythrocytes (Figure 3.3A). Capture-C is a chromosome conformation capture (3C) method that can be used to study enhancer-promoter interactions at high resolution and, thus, be used to predict functional contacts in a tissue-specific manner (Davies et al., 2016; Davies et al., 2017). Interaction profiles from several promoters exhibited specific interactions with surrounding accessible elements. However, despite data from the R2 enhancer and α -globin genes showing apparent interactions with regions upstream of the gene locus, none of the promoters tested outside of the α -globin self-interacting domain, as defined previously by Capture-C and demarcated by CTCF sites (Hanssen et al., 2017; Oudelaar et al., 2018), showed specific interaction with the α -globin regulatory elements (Figure 3.3B). On the contrary, the closest genes on either side of the self-interacting domain demonstrated a relative avoidance of the region in a similar pattern to that observed for CTCF sites outside of the locus boundaries (Oudelaar et al., 2018). Therefore, I concluded that manipulations of the super-enhancer would be unlikely to affect surrounding genes other than those previously identified. As those genes (*Il9r*, *Snrnp25*, *Rhbdf1*, *Mpg*) are likely to be bystanders due to their proximity and have not previously been implicated in influencing any erythroid-specific phenotypes, I concluded that I could proceed with locus disruption and be confident in reporting direct effects on the α - and ζ -globin genes.

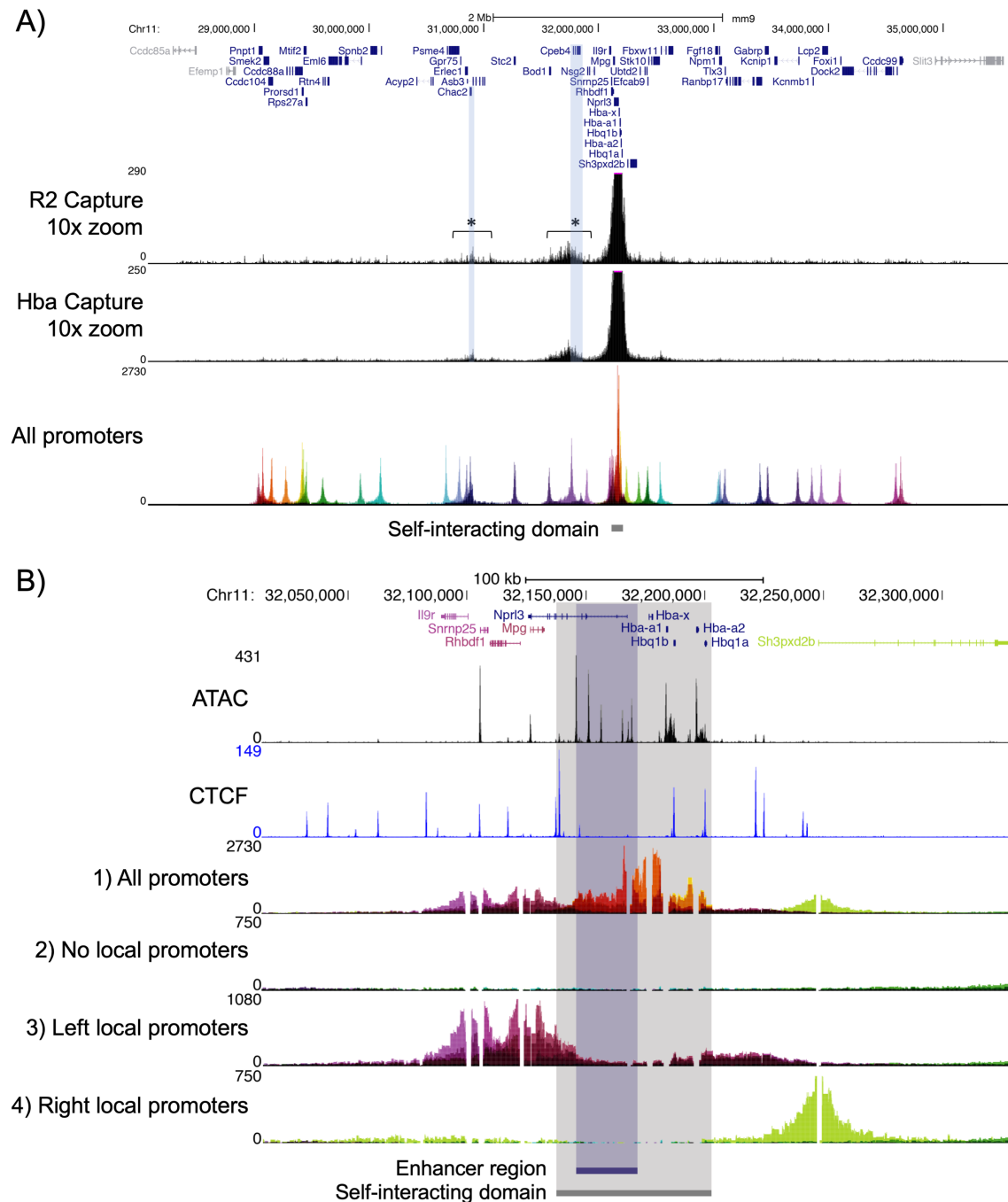


Figure 3.3: Promoters within a 5.6 Mb region containing the α -globin locus do not show specific interactions with the super-enhancer elements in mouse erythrocytes.

A) Top and middle: 10x zoomed Capture-C tracks from the R2 enhancer element and *Hba* gene promoters (*Hba-a1* and *Hba-a2* combined). Tracks, originally with data ranging from 0-2500 on the y-axis, are scaled to show only the weakest 10% of the signal such that low-enrichment, long-range interactions can be visualised. Areas of enriched interactions outside of the α -globin locus containing the *Cpeb4* and *Chac2* gene regions (shaded with blue boxes) are highlighted with brackets and asterisks (*). Bottom: Capture-C track overlay for all promoter capture points within a 5 Mb window (chr11: 29,030,751-34,647,143), using a rolling colour scheme to highlight results from the individual promoter elements. **B)** ATAC-seq (black), CTCF ChIP-seq (blue) and high-resolution Capture-C tracks from the 5 Mb promoter capture experiment in Ter119-positive mouse erythrocytes.

erythrocytes, viewed over the α -globin locus region. Capture-C overlays from top to bottom: 1) All capture points shown; 2) No local promoters shown: all promoters shown except for *Il9r*, *Snrnp25*, *Rhbdf1*, *Mpg*, *Nprl3*, *Hba-x*, *Hba-a1*, *Hba-a2* and *Sh3pxd2b*; 3) as 2) but with *Il9r*, *Snrnp25*, *Rhbdf1* and *Mpg* promoters included; 4) As 2) but with the *Sh3pxd2b* promoter included. Tracks are scaled to the maximal peak height shown except for 2), which is scaled the same as 4). The enhancer region and self-interacting domain are highlighted in blue and grey, respectively. All Capture-C tracks are merged from three independent biological replicates; ATAC-seq and ChIP-seq data are averaged over two biological replicates as previously published (Hanssen et al., 2017).

How to engineer multiple genetic elements in a single transfection step

Depending on the number of manipulations required, dissecting regulatory landscapes *in situ* generally relies on single or iterative rounds of genome engineering. Clustered regularly interspaced short palindromic repeats (CRISPR)/Cas9 technology makes it possible to delete or modify target regions of a single base pair to hundreds of kilobases in mESCs with high efficiency, but more complex designs involving homology-directed repair (HDR) are typically limited to less than 10 kb (Li et al., 2014; Wang et al., 2015). To create a mESC line with four of the five native mouse α -globin enhancer elements deleted, the simplest approach from a design perspective would be to perform four sequential element deletions. However, in the interests of time and maintaining cell pluripotency for mouse model generation, I pursued an alternative, recombination-based targeting strategy to replace the entire locus in a single exchange event.

The recombinase-mediated genomic replacement (RMGR) method was first described in a collaboration between the Higgs group and Dr Andrew Smith (University of Edinburgh) to create a “humanised” mouse model of the α -globin locus (Wallace et al., 2007). The strategy relies on the orientation-specific exchange of one copy of the native locus for an externally engineered version contained within a bacterial artificial chromosome (BAC). Purpose-

designed “RMGR-ready” mESCs were originally generated from WT E14 mESCs in two sequential homologous recombination events to specifically integrate heterotypic lox sites at either side of the locus (Wallace et al., 2007). A single copy of a loxP sequence, along with antibiotic resistance cassettes for positive selection, was integrated at the 5’ side of the locus within the fourth intron of the *Rhbdf1* gene; a similar cassette containing a lox511 sequence was integrated at the 3’ side within the second intron of the *Hbq1a* gene. To ensure that both heterozygous targeting events had taken place on the same allele, clones were tested for the efficiency of exchange of a small positive control donor carrying the appropriate lox sites but with no sequence in between. In the resulting selected mESC line, the second allele remains intact with no lox sites.

Exchange of the native locus with the engineered BAC in RMGR-ready mESCs is catalysed by orientation-specific Cre-directed recombination at the heterotypic lox sites flanking the locus and is enriched for by the complementation of a selectable hypoxanthine phosphoribosyltransferase (*Hprt*) gene cassette at the 5’ loxP site. The *Hprt* cassette is later excised by flippase (Flp)-mediated recombination between flanking flippase recognition target (ftr) sites. The result is a near-scarless exchange of the locus of interest into the native α -globin chromosomal context. A schematic depicting how this targeting approach can be used to produce a heterozygous version of the R2-only model in mESCs is shown in Figure 3.4.

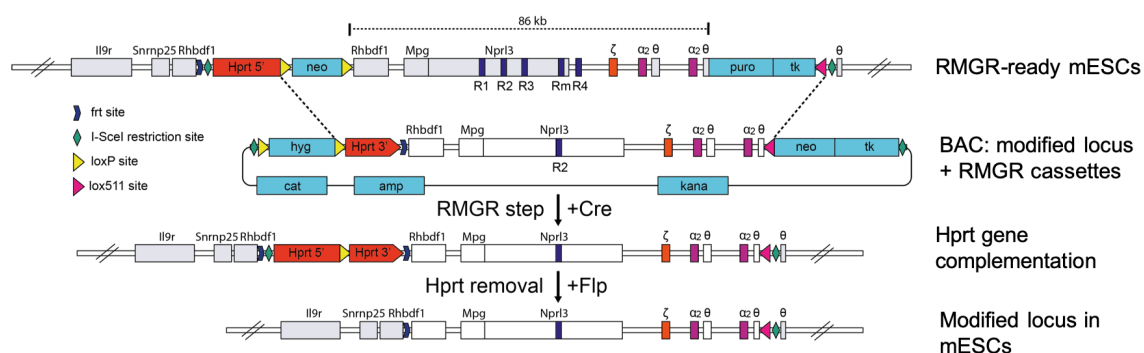


Figure 3.4: Schematic for RMGR targeting strategy to generate the R2-only model in mESCs.

LoxP and Lox511 sites for locus exchange are depicted by triangles in yellow and pink, respectively. ftr sites for *Hprt* gene excision are shown in navy blue. Restriction sites and resistance gene cassettes (shown in teal and light blue, respectively) were used in the initial RMGR setup or are used for BAC growth in selective LB: amp = ampicillin resistance gene; cat = chloramphenicol resistance gene; hyg = hygromycin resistance gene; kana = kanamycin resistance gene; neo = neomycin resistance gene; puro = puromycin resistance gene; tk = thymidine kinase.

Conventionally, sequence modifications of large DNA constructs such as BACs would be performed by iterative recombineering strategies in bacteria (Zhang et al., 1998). However, the number of selection steps needed to make four sequential enhancer deletions made this design technically challenging. Instead, I embarked on a collaboration with Dr Leslie Mitchell from Professor Jef Boeke's group (New York University, USA) to repurpose a large-scale DNA assembly technology, originally developed to generate yeast chromosomes from scratch, to create custom version of the mouse α -globin locus for use with the RMGR strategy (Richardson et al., 2017). The result would be a mouse model with an entirely synthetic version of the α -globin locus. I will refer to this combined strategy hereafter as Synthetic RMGR.

A workflow for the Synthetic RMGR strategy is shown in Figure 3.5. Briefly, the process begins with design of the locus of interest, which is then synthesised and assembled into a BAC. This is followed by DNA purification and exchange into RMGR-ready mESCs.

Correctly targeted mESC clones are selected for *Hprt* gene function and validated by PCR screening, then targeted for *Hprt* cassette removal. Resulting mESCs are used to generate erythroid cells *in vivo* through production of mouse models or *in vitro* by a method detailed in Chapter 4 and, finally, analysed for the resulting molecular phenotype.

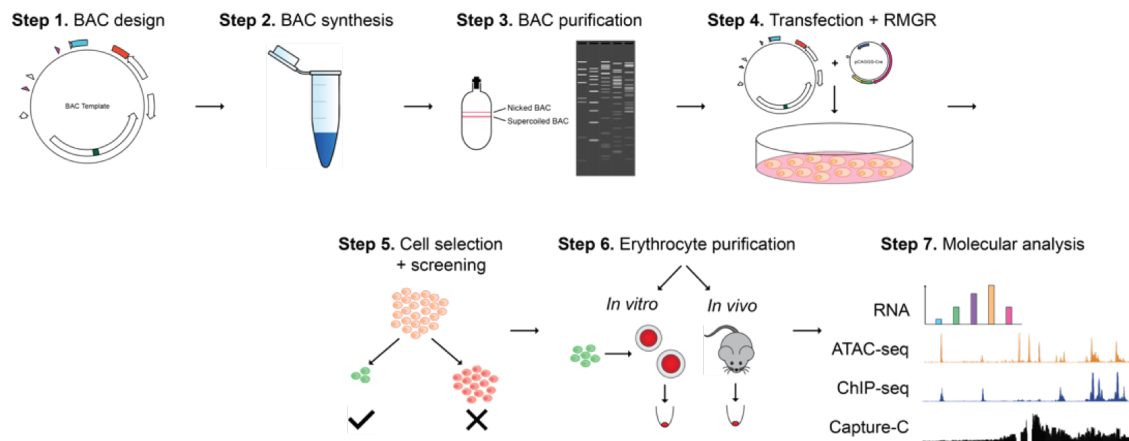


Figure 3.5: Workflow for the Synthetic RMGR procedure.

In generating R2-only mESCs, the workflow went as follows:

Step 1. BAC design

In the first instance, I designed BACs for WT and R2-only versions of the mouse α -globin locus, based on the sequence of a WT BAC previously generated and sequence-verified in the lab by Dr Christian Babbs. This template BAC contains the loxP and lox511 recombination sites and *Hprt* split gene necessary for RMGR, flanking a span of 86 kb that contains all of the regulatory elements and genes required for α -globin gene expression (Figure 3.4). To create the R2-only model, I designed individual enhancer element deletions to exactly reproduce those used previously (Hay et al., 2016). To confirm that these deletions removed the entire active regions for each element, I aligned their coordinates with DHS peaks, DNase-seq digital footprinting data and multispecies conserved sequence (MCS)

regions from primary mouse erythrocytes (Hughes et al., 2005; Maria Suci, unpublished data). DNase-seq peaks can be used to identify accessible regions of chromatin including enhancers (Thurman et al., 2012), while sequence protection data from digital footprinting of these data can be used to predict the sites of DNA-protein interactions, thus providing an indication of sequences that are likely to be functional (Hesselberth et al., 2009). Guided by these data, the prescribed deletions remove the entire functional units for each element (Figure 3.6A-B).

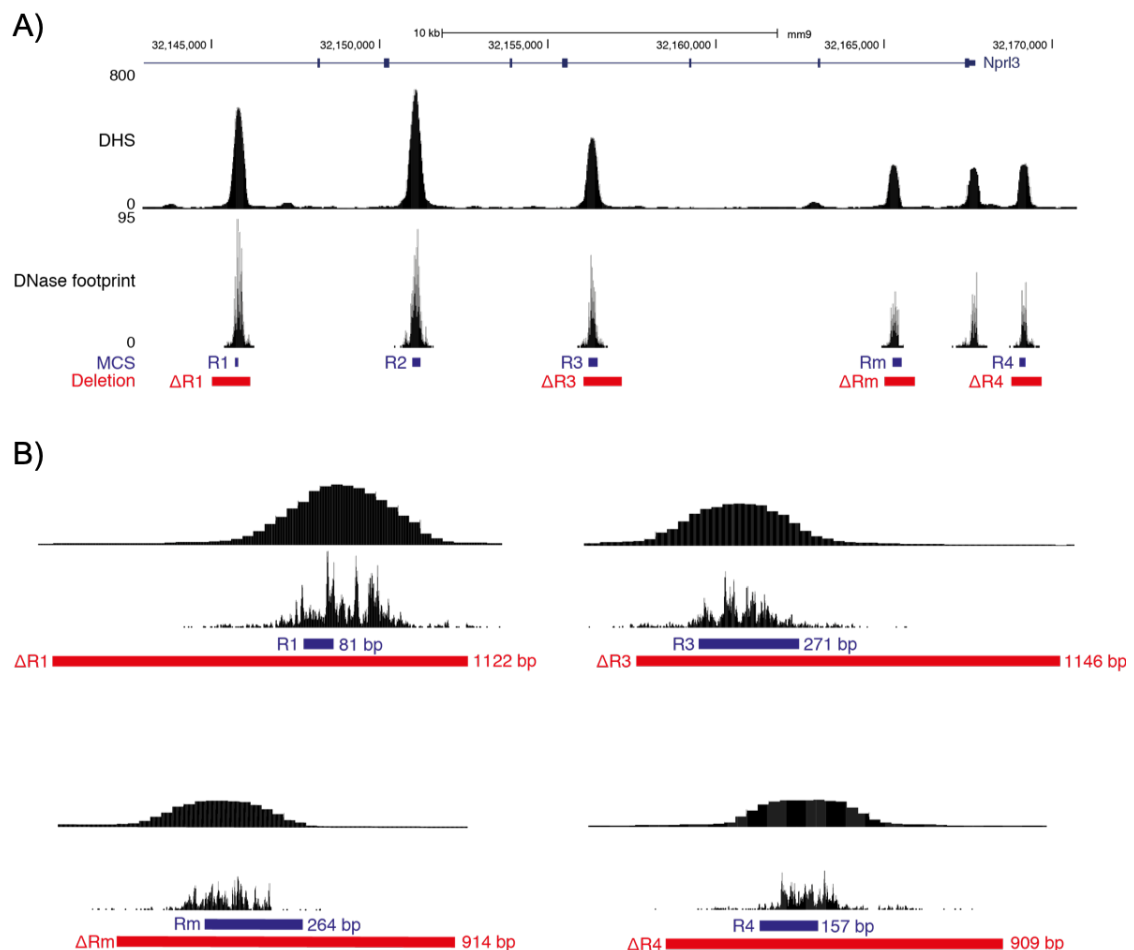


Figure 3.6: Enhancer deletion coordinates remove all sequence functionality.

A) Overview of the α -globin enhancer deletions used in Hay et al., 2016 aligned to DNase-seq and digital footprinting data from erythroid cells from embryonic day 10.5 (E10.5; Maria Suci, unpublished). MCS regions and deleted sequences are highlighted in blue and red, respectively. **B)** High resolution view of enhancer deletion regions.

Step 2. BAC synthesis

For both BACs (WT and R2-only), small (~3 kb) DNA segments spanning the 92 kb lox-locus-lox sequence region were constructed by Dr Leslie Mitchell, using a tiled PCR-based method. These were combined with synthetic inserts (100 bp – 1 kb) used to integrate the designed deletion junctions. Dr Mitchell assembled these segments into larger blocks by their “eSWAP-In” method, which involves iterative homologous recombination steps in *Saccharomyces cerevisiae* (Figure 3.7). Assemblon generation in this way was described for the construction of yeast chromosomes and, more recently, in the proof-of-principal constructions of the 101 kb human *HPRT1* gene locus a 119 kb region of the human genome containing the *HTRA1* and *ARMS2* genes (Richardson et al., 2017; Mitchell et al., 2019; Laurent et al., 2019). After retrieval of the bespoke sequences into a BAC backbone and transformation into bacteria for DNA purification, Dr Mitchell validated the locus sequences by PCR over the assembly junctions and long-read sequencing by PACBIO technology (Korlach et al., 2010). Sequencing results identified a small number of single base mutations in the assembly; these are described in the following section.

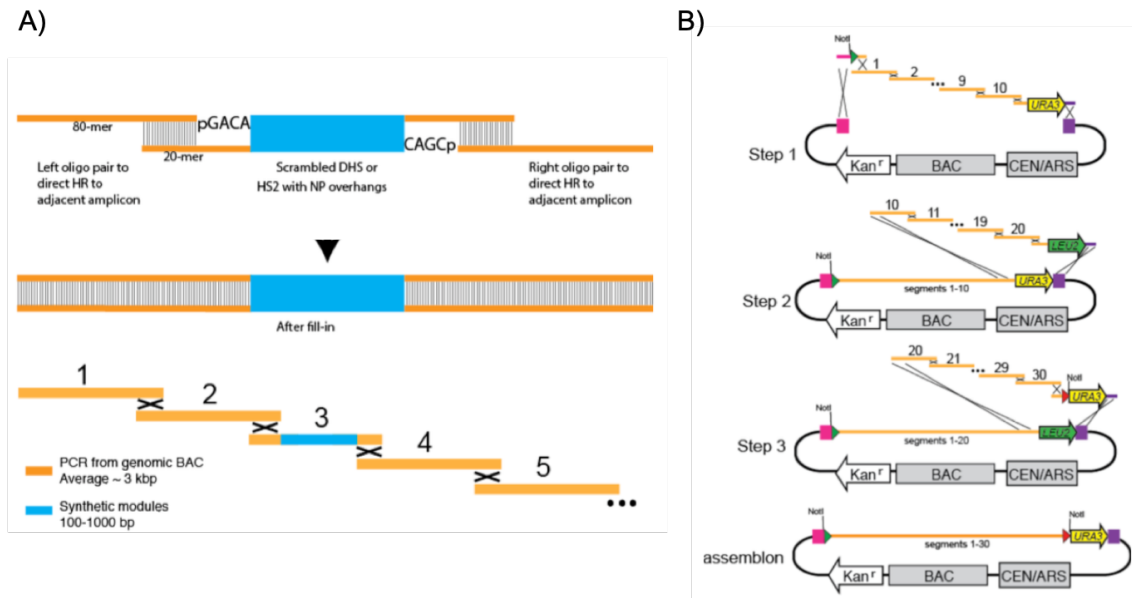


Figure 3.7: Schematic view of the eSWAP-In method (provided by Professor Jef Boeke).

A) Short segments of DNA are generated by PCR from a target backbone or ordered as synthetic constructs with matching overhangs to allow for homologous recombination. **B)** Homologous recombination is performed in yeast with blocks of up to 10 segments in a single step. Alternating selection for the *URA3* and *LEU2* genes is used to perform subsequent assembly steps. The resulting BAC contains the assembled locus, modules necessary for replication in both yeast and *E. Coli* (BAC / CEN/ARS, respectively) and selectable markers including the kanamycin resistance gene (*Kan^r*).

Step 3. BAC purification

I cultured the WT and R2-only BACs in *E. Coli* under conditions of antibiotic selection and isolated pure DNA by caesium chloride purification. I performed restriction digests to confirm that the BACs were intact and had not undergone any large-scale rearrangements (Figure 3.8A). To confirm that the sequence was correct at the base-pair level, I prepared libraries for Illumina short-read sequencing and aligned reads to the expected design sequences. The small number of variations I detected were in agreement with those detected by PACBIO sequencing before reagent shipping and largely comprised short repeat sections in intergenic regions which may be due to PCR infidelity or errors from the sequencing method, known to have low fidelity over repeat regions (Pfeiffer et al., 2018). Summaries

of *bona fide* variants for each BAC are shown in Tables 3.1 and 3.2. All variants are either between genes or found in introns not overlapping any recognised regulatory elements or splice sites with one exception: a single T->C substitution in the R3 enhancer element in the WT BAC. This base is not predicted to fall within a TF binding site by conservation analysis or DNase footprinting (Figure 3.8B) and the detected mutation is not predicted to damage the footprint of the surrounding 13 bp by analysis with the Sasquatch prediction tool (damage scores all below 0.2; Schwessinger et al., 2017). Crucially, I detected no errors in the sequences needed for RMGR.

Position	Type of variant	Reference	Variant	Description
5679	Substitution	C	T	<i>Rhbdf1</i> exon 2 - synonymous
27421	Substitution	C	T	<i>Nprl3</i> intron
31550	Substitution	C	T	<i>Nprl3</i> intron
35798	Substitution	C	T	<i>Nprl3</i> intron
39735	Substitution	C	T	<i>Nprl3</i> exon 5 – synonymous
45789	Substitution	T	C	R3 enhancer
68981	Substitution	T	C	Intergenic
74052	Substitution	G	A	Intergenic
Table 3.1: Variants called in the WT synthetic BAC vs provided designs.				

Position	Type of variant	Reference	Variant	Description
6145	Substitution	C	T	<i>Rhbdf1</i> intron
44425	Deletion	30 bp	-	<i>Nprl3</i> intron
46911	Substitution	G	A	<i>Nprl3</i> intron
58887	Substitution	G	C	Intergenic
Table 3.2: Variants called in the R2-only model synthetic BAC vs provided designs.				

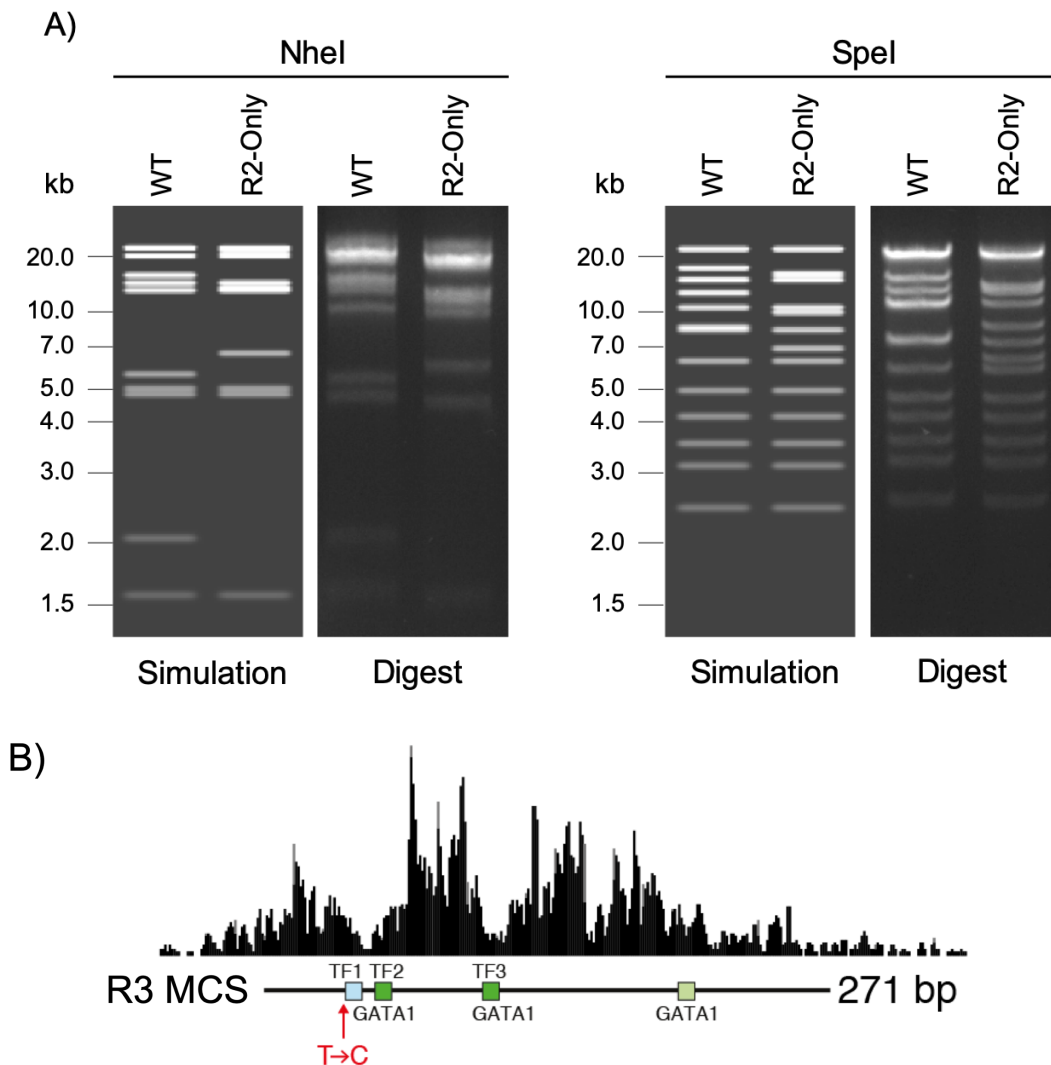


Figure 3.8: Validation of synthetic BACs of WT and R2-only α -globin loci.

A) Comparison of simulated and real restriction digests of the two BACs with restriction enzymes *NheI* and *SpeI*. Gel simulations were based on *in silico* digests of BAC design files. **B)** A synthesis error in the R3 enhancer sequence of the WT BAC is not predicted to be deleterious for function. Top: DNase-seq digital footprinting as in Figure 3.6; Bottom: schematic of the R3 enhancer MCS annotated with the sites of three conserved regions (TF1-3; Hughes et al., 2005) and an additional motif for the Gata1 TF. The T→C substitution is positioned immediately upstream of the conserved TF1 region and highlighted in red.

Step 4. Transfection + RMGR

The most limiting step in successful targeting of large DNA constructs into the genome is transfection and exchange into the cells of interest. This is primarily because the efficient delivery of constructs to mammalian nuclei by electroporation or lipid-based methods is

greatly attenuated with increasing plasmid size (Hornstein et al., 2016; Kreiss et al., 1999), in addition to the observation that transfection of large constructs can be associated with high cell toxicity (Lesueur et al., 2016). By established protocols for BAC transfection into mESCs, described below, extreme inefficiency (1-10 correct clones per 100 μ g DNA transfected into $\sim 50 \times 10^6$ cells) renders BAC exchange unfeasible for routine experimentation. The most prominent challenge is on the scale of DNA preparations required per transfection; this is made particularly technically taxing by the fact that BACs must be maintained at a single copy per bacterial cell to avoid structural rearrangements and must be transfected immediately upon purification to avoid degradation of the high molecular weight DNA molecules. Improvements in transfection efficiency would, therefore, be beneficial for the entire BAC delivery process.

In the original RMGR protocol, and many similar experimental designs, BACs have been delivered into mESCs by electroporation. All studies have reported low levels of targeting efficiency (Lee et al., 2014; Testa et al., 2003; Tomishima et al., 2006; Yang and Seed, 2003). To make this protocol viable for more routine delivery of BACs, I tested a range of popular methods for transfection of mammalian cells with DNA. Electroporation methods did not produce any surviving clones in my hands (Table 3.3). Instead, I optimised a method using lipid co-delivery; this has been shown to be possible for large DNA and potentially less damaging to BAC structure (McLenachan et al., 2007). With the resulting lipofection protocol, 1-3 correctly targeted clones can now be obtained with the transfection of 10^6 cells using just 2.5 μ g total DNA (3:1 BAC:Cre plasmid ratio by mass), representing a ~ 40 -fold improvement in efficiency. In this instance, correct targeting refers to successful Cre-mediated exchange at both 5' and 3' ends of the locus by *Hprt* gene selection and PCR, as described in Step 5.

Type	Method	# Cells transfected	[DNA] (3:1 BAC:Cre)	# Colonies recovered
Electroporation	BioRad	35x10 ⁶	100 ug	0
Electroporation	BioRad XCell	10x10 ⁶	30 ug	0
Electroporation	BioRad XCell	2x 9x10 ⁶	2x 30 ug	0
Electroporation	Neon	10 ⁶ each	6/12/30 ug	0
Electroporation	Neon	10 ⁶ each	2 ug	0
Electroporation	Nucleofector	5x10 ⁶	10 ug	0
		5x10 ⁶	15 ug	2 (0 correct)
		5x10 ⁶	20 ug	0
Electroporation	Nucleofector	5x10 ⁶ each	6/9/13/15 ug	0
Electroporation	Nucleofector	5x10 ⁶ each	13/16/16/20 µg	0
Electroporation	Nucleofector	5x10 ⁶ each	13/26 µg	0
Lipofection	FuGENE 6	10 ⁶	4/7/20 µg	0
Lipofection	FuGENE 6	5.7x10 ⁶	70 µg	13 (3 correct)
Lipofection	FuGENE 6	5x10 ⁶ each	2x 27 µg	0
Lipofection	FuGENE 6	8.4x10 ⁶	60 µg	0
Lipofection	LTX	6x10 ⁶	15 µg	35 (10 correct)

Table 3.3: Summary of BAC transfection attempts.

The # colonies recovered column refers to the number of colonies surviving HAT selection; numbers in brackets indicate how many resistant colonies were correctly targeted at both RMGR exchange junctions by PCR as described in Step 5. BioRad = BioRad Gene Pulser; BioRad XCell = BioRad Gene Pulser XCell; Neon = Thermo Fisher Neon Transfection System; Nucleofector = Lonza Nucleofector 2b; Fugene 6 = Promega FuGENE 6 Transfection Reagent; LTX = Thermo Fisher Lipofectamine LTX Reagent. For details of transfection protocols, see Methods (Chapter 2).

Step 5. Cell selection + screening

Initial selection for successful RMGR is based on complementation of the 5' *Hprt* gene and is performed by treatment of cells with hypoxanthine-aminopterin-thymidine (HAT). This process is extremely stringent; all clones surviving selection for integration of the synthetic

BACs were positive for the complete *Hprt* gene by PCR. Next, to ensure that locus exchange was complete for HAT-resistant colonies, I performed validation PCRs over the left loxP and right lox511 exchange sites and each of the enhancer elements (Figure 3.9A-B). Expected PCR products in targeted mESCs confirmed the desired exchange occurred in around 1/3 of HAT-resistant clones. Validated clones were tested for normal chromosome counts by karyotyping analysis; one clone containing the R2-only model locus was selected for microinjection into mouse blastocysts.

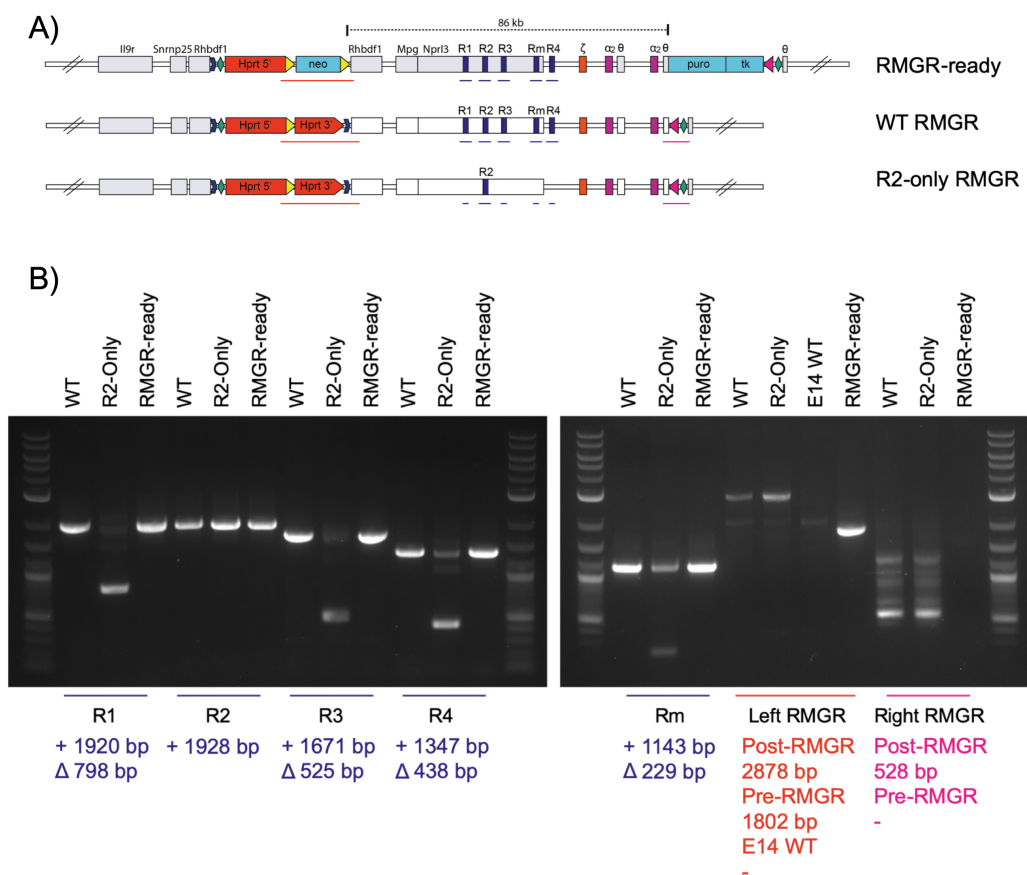


Figure 3.9: Correct RMGR-exchanged clones were identified by PCR.

A) Schematic of RMGR-ready, WT BAC-exchanged and R2-only BAC-exchanged mESC loci with PCR positions annotated with coloured bars. Sequence features are annotated as in Figure 3.4. **B)** PCR screening results for material from single mESC clones of heterozygous WT BAC RMGR, R2-only RMGR and RMGR-ready genotypes. Post-RMGR products are detected in WT and R2-only exchanged clones. Full length PCR products over individual enhancer elements are detected in the WT exchanged and RMGR-ready cells; faint bands are also amplified from the untargeted allele in the R2-only exchanged cells. Short PCR products indicating the presence of enhancer deletions are only detected in R2-only exchanged material.

One technical concern with targeting an entire locus in this manner is the possibility for large-scale rearrangements to go undetected with the short-length PCR validation and sequencing methods. The α -globin locus presents a particular challenge here due to the gene duplication which leaves a stretch of sequence open to recombination-induced rearrangements in the yeast used for assembly, the bacteria used for DNA preparation (although this is mostly mitigated by the use of single-copy BAC vectors) or in the process of locus exchange. To confirm that the locus was structurally intact in the final targeted mESCs selected for microinjection into mouse blastocysts, I prepared high molecular weight genomic DNA for linked-read sequencing (10X Genomics). In this method, contiguous DNA molecules are labelled with the same unique barcode sequence. Combined genome-wide sequencing reads can then be mapped both to the reference genome and tied back to the single molecule from which it derived, meaning that intact stretches of hundreds to thousands of kb can be reconstructed from short-read sequencing data (Kitzman, 2016). De-multiplexing of the linked-read data into two haplotypes clearly separated the WT and RMGR alleles from the targeted mESCs, with no additional structural changes detected (Figure 3.10).

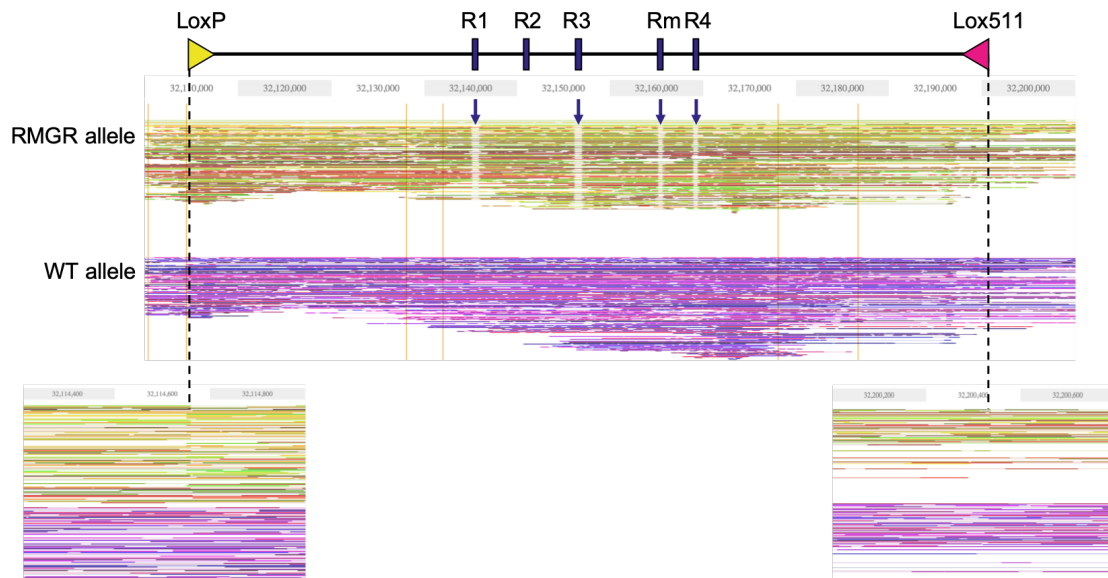


Figure 3.10: The R2-only BAC was integrated into mESCs without structural rearrangements.

Top: schematic of the RMGR region with enhancer elements and RMGR lox exchange sites annotated. Middle: phased linked-reads for the R2-only RMGR and WT alleles, visualised with proprietary Loupe software from 10X Genomics. Enhancer element deletions are clearly visible by sequencing dropout in the RMGR allele, as indicated by arrows. Bottom: junctions where lox sites remain in the R2-only allele are seen as mapped read ends in the RMGR allele, but not in the WT allele where there are no lox sites. Junctions are visible as vertical interruptions to mapped reads, aligned with dashed black lines from the schematic above.

Step 6. Erythrocyte purification

Following mESC targeting, the intention was to create an *in vivo* R2-only model to study the effect of such a modification on endogenous α -globin expression in primary erythroid cells. As such, validated R2-only model mESCs were microinjected into mouse blastocysts and resulting chimeras were back-crossed to WT mice, then bred with a line carrying a random integration of the Flp gene to remove the RMGR *Hprt* selection cassette. Heterozygous R2-only mice were viable with no notable phenotype. Heterozygotes with the *Hprt* cassette successfully removed, as determined by PCR, were crossed together and with WT mice to remove the eFlp gene for future crosses.

Of the three Het x Het crosses left to litter until now, there have been no homozygous R2-only mice surviving for longer than a week after birth (Table 3.4). It is, therefore, possible that the R2-only mouse suffers from a more severe reduction in α -globin expression than predicted. Based on these results, I instead focused on obtaining erythroid material from an earlier developmental timepoint where homozygous mice were more likely to be viable. As such, I obtained primary erythroid cells from embryonic day 9.5 or 10.5 (E9.5/E10.5). Although these cells are not directly comparable to the E12.5 fetal liver-derived erythrocytes used in the original super-enhancer dissection publication (Hay et al., 2016), data from the Higgs lab suggest that the α -globin genes are regulated similarly across embryonic and fetal/adult erythroid tissues; the embryo-specific ζ -globin gene is also affected by enhancer deletions but to a lesser extent than the adult genes (Duantida Songdej, unpublished data). I therefore reasoned that this remained an appropriate cell system to test the question of enhancer additivity *in vivo*. Although a high level of mortality was observed for homozygotes even at this early stage, I was able to obtain erythroid material from a total of two homozygotes, 16 heterozygotes and four WT littermates (Table 3.4).

Stage	Cross	Total	WT	Het	Hom	Notes
Live mice	Het x Het 1	4	2	0	2 dead	1 st Hom stillborn 2 nd Hom died at 1 week 2x WT + 2nd Hom Flp+
Live mice	Het x Het 2	3	0	3	0	2x Het Flp+
Live mice	Het x Het 3	3	0	3	0	
E10.5	Het x Het 4	8	2	5	1	
E10.5	Het x Het 5	8	2 + 1 dead	3	2 dead	
E9.5	Het x Het 6	11	0	8 + 1 dead	1 + 1 dead	
	Metric	Total	WT	Het	Hom	
	Total	37	7	23	7	
	Total live	30	6	22	2	
	% Live	81.1	85.7	95.7	28.6	
	Total embryos	27	5	17	5	
	Live embryos	22	4	16	2	
	% Live embryos	81.5	80	94.1	40	

Table 3.4: Summary of RMGR Het x Het offspring genotypes.

Counts are shown for genotypes of live offspring and dissected embryos. The Stage column indicates the timepoint at which mice were genotyped. Notes indicate the stages at which born homozygous mice were lost and where mice were carriers for the Flp gene (used for *Hprt* cassette excision).

The activity of the R2 element stripped of its super-enhancer context

Step 7. Molecular analysis

Using the E9.5/E10.5 erythroid material described above, I performed ATAC-seq to assess the impact of super-enhancer disruption on the open chromatin landscape of the α -globin

locus. By this method, the major locus elements became accessible as expected, including the R2 enhancer and α - and ζ -globin genes (Figure 3.11A). Notable, however, was a visible reduction in accessibility over the α - and ζ -globin gene regions when comparing between littermates from three independent harvests. Although ATAC-seq is generally regarded as a semi-quantitative technique, read counts normalised against the β -globin locus showed a reproducible decrease in reads from WT to heterozygous to homozygous embryos across the three harvests (Figure 3.11B). In contrast, read counts for the R2 enhancer did not show this same reduction: instead, the element had the same read count, if not higher, in embryos carrying one or two copies of the R2-only locus. Together these data indicate that reduction of the super-enhancer to a single element did not cause a complete loss of accessibility at the surrounding elements, but may have impacted the isolated chromatin opening and associated activity of the globin genes.

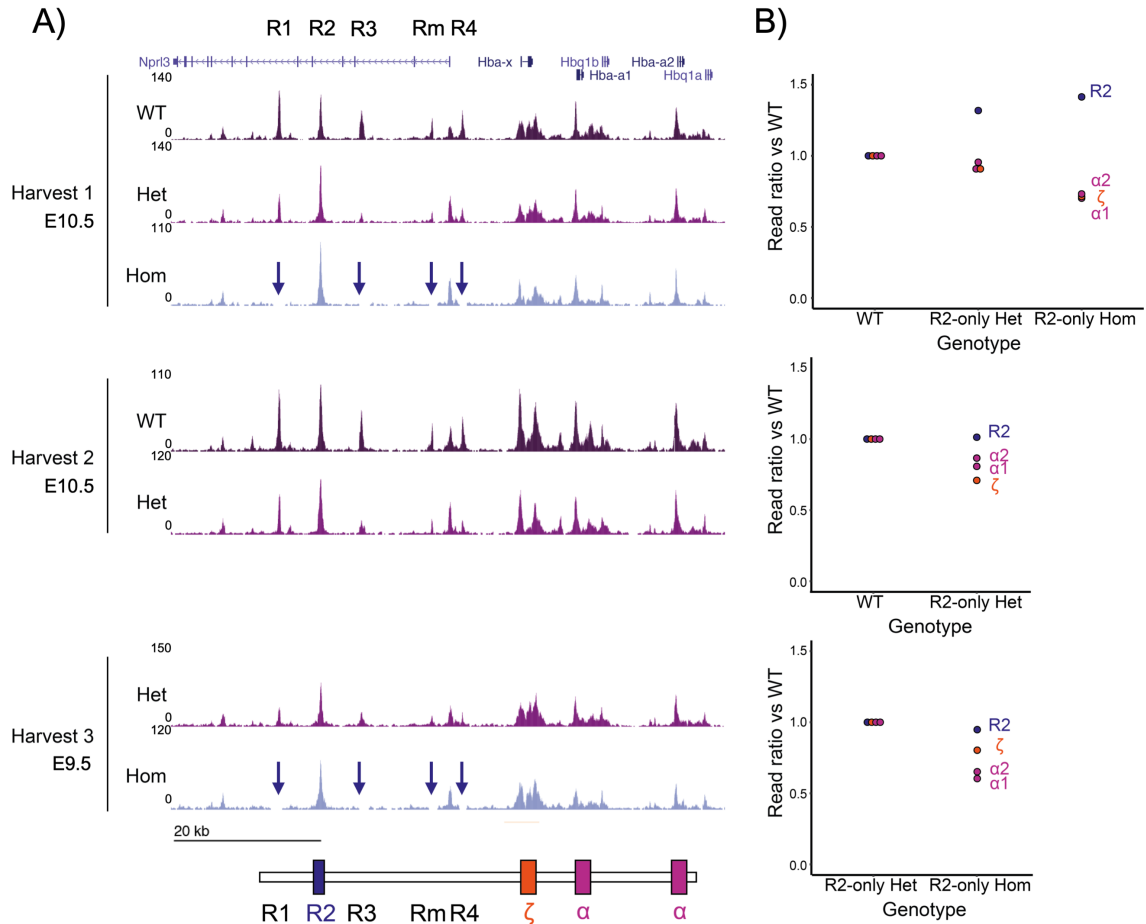


Figure 3.11: R2-only locus elements become accessible as expected in vivo.

A) ATAC-seq tracks from embryonic erythroid cells from individual WT, heterozygous (Het) and homozygous (Hom) mice for the R2-only model, as available, from three embryo harvest dates (Harvest 1/2/3). Embryos are compared directly to littermates to mitigate the effects of small differences in developmental staging. To negate the effect of inconsistent noise levels from the ATAC-seq assay on normalised peak height, reads per kilobase per million mapped reads (RPKM)-normalised tracks are visualised at scales that result in equivalent peak heights at a control locus containing histone genes (chr13:23,830,000-23,860,000). **B)** Plots detailing the number of ATAC-seq reads aligning to the α - and ζ -globin gene regions (chr11:32175940-32179679; 32183170-32188139; 32196010-32200979) normalised to counts from the β -globin genes (chr7: 110955100-110962869; 110976080-110976869; 110988240-110992769; 110996780-111001949) for each embryo (pink / orange dots), and reads aligning to the R2 enhancer (chr11:32150660-32151659), normalised to counts from the β -globin locus control region (chr7:111007173-111026984) for each embryo (blue dots).

To assess the effect of the R2-only locus modification on gene expression of the active α - and ζ -globin genes, I performed real-time PCR (RT-PCR) on cDNA from each sample. In addition to the samples harvested here, I included cDNA from homozygous $\Delta R1$ E10.5 embryonic blood for context. As discussed in the introduction to this chapter and shown in Figure 3.2, expression of the globin genes in homozygous $\Delta R1$ mice is predicted to be driven primarily by the R2 enhancer (Hay et al., 2016).

To our surprise, but in line with the findings of impaired viability, results from R2-only homozygotes showed a much more severe phenotype than empirically predicted (Figure 3.12A). Instead of the expected ~ 0.5 relative α -globin output, R2-only embryos retained an average of only 0.15 compared to WT. Heterozygote embryos gave an intermediate reading, with an output of around 0.55 of WT levels. Expression of ζ -globin was also impaired, with homozygotes producing only around 0.5 of WT levels, while results from the $\Delta R1$ model were comparable to WT embryos (Figure 3.12B).

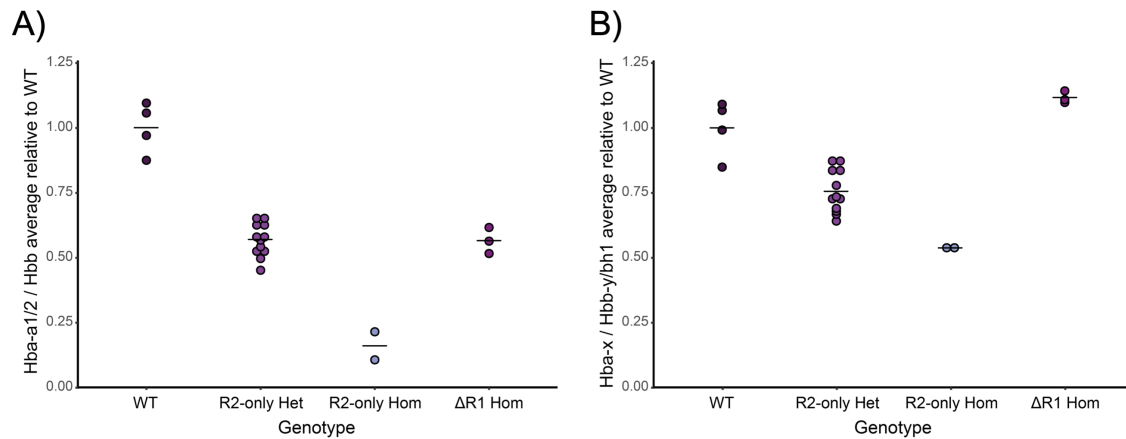


Figure 3.12: Expression of the α - and ζ -globin genes is severely impaired in the R2-only model.

A) Quantitation of mature adult *Hba-a1/2* α -globin transcripts compared to the transcriptional output from the β -globin locus (C_T values averaged across *Hbb-b1/2*, *Hbb-y* and *Hbb-bh1* genes) for WT, Het and Hom embryos for the R2-only model and homozygous $\Delta R1$ embryos. Normalising to the β -globin genes accounts for the variation in embryo timepoint and erythrocyte maturation across harvest dates. **B)** Quantitation of mature embryonic *Hba-x* transcripts compared to the embryonic output from the β -globin locus (C_T values averaged across *Hbb-y* and *Hbb-bh1* genes) for the same models as in part A). Data points each represent an individual embryo and are shown relative to mean WT values; mean results for each genotype are illustrated with a horizontal black bar.

These surprising findings called into question the existing additive model for super-enhancer function at this locus and may explain why no viable adult homozygous R2-only mice have been observed.

3.3 Discussion

The impact of technology availability on scientific approach

Our ability to dissect the function of regulatory elements in the genome has, by necessity, been dictated by the molecular toolkit available. Before targeted mutation of endogenous elements became possible, much of this work was performed using reporter assays in heterologous systems, followed by transgenes and large, randomly integrated units such as BACs. These experimental designs are not able to probe the function of an element within the native chromosomal context and are subject to inappropriate packaging within nucleosomes, interacting with surrounding elements as dictated by regulatory domains and conforming to larger-scale restrictions of chromatin structure, thus often resulting in large variation between replicates and unreliable results (Bender et al., 2000b; Long et al., 2016).

Since the dawn of endogenous gene targeting, investigating the importance of multiple elements in a regulatory domain *in situ* has depended on the availability of precise genome engineering strategies (examples include Bahr et al., 2018; Bender et al., 2012; Darbellay et al., 2019; Hay et al., 2016; Kraft et al., 2019; Shin et al., 2016). Due to the time-intensive nature of performing iterative rounds of genome engineering, these approaches have largely been used to dissect super-enhancer clusters and regulatory domains from a “top-down” perspective. Typical experiments would start with a WT chromosome and involve perturbation of just one or two elements in series, leaving the majority of the regulatory landscape intact. Thus, these experiments tend to probe elements for their necessity and redundancy for locus function, rather than their sufficiency. Such experimental approaches do not lend themselves to the reciprocal experiment where elements would be inserted

individually into a null locus (a “bottom-up” approach); as such, experiments of this nature have scarcely been performed.

Here, by applying a genome engineering method that is not restricted in the number of manipulations to be made in series, the rule book for how to approach biologically interesting questions can be reimaged. As methods for large-scale DNA synthesis and targeting become more extensively automated, this extends to ideas such as the ambitious ‘Human Genome Project-Write’ becoming accessible (Boeke et al., 2016; Hughes and Ellington, 2017). The BACs used here were assembled with a platform capable of performing 30,000 RT-PCRs per day at a cost of 7.4 cents each. Other workcell components include a colony picker, barcode reader, bulk loader, robotic centrifuge, and automated sealers and peelers, as well as robotic arms to coordinate plate transfers.

Combined with the ever-falling costs of DNA sequencing (Wetterstrand, 2019), technologies such as these will make it feasible to test a near-limitless number of locus variations. The bottleneck of transfection into mammalian cells has been somewhat ameliorated here with the introduction of a lipofection-based protocol; there is undoubtedly room for further improvement or automation to this step, in future. In Chapter 4, I describe a miniaturised *in vitro* differentiation protocol which goes on to streamline the downstream processing stages of targeted clones. In Chapter 6, I develop a fluorescence-based readout to test the effect of locus manipulations on α -globin gene output in a high-throughput manner. Consequently, the experimental design used here is a valuable proof-of-principle for a new mode of experimental genetics, in addition to providing an interesting result in itself.

Addition $\neq$ subtraction

By homozygous deletion *in vivo*, the R2 enhancer appears to contribute ~50% of the transcriptional activation of the α -globin gene (Hay et al., 2016). Yet, here, addition of the R2 enhancer to a neutral platform is capable of a maximum of ~15% gene activation. When dissecting the super-enhancer from the top-down, pairs of deletions appeared to reduce α -globin expression in a manner resembling an additive model. So, now when approaching the question from the bottom-up, why does enhancer addition not equal subtraction?

From a general perspective, it is possible that one or more synergistic models of enhancer co-operation may be fitting at this locus when it is reduced to its minimal components. There are several possible mechanisms for how the observed data might be realised at a molecular level; these will be discussed in more detail in Chapter 5.

By ATAC-seq, the locus elements appear to become accessible as normal, although with reduced peak heights over the globin gene regions. This result is in-line with findings from the α - and β -globin loci, where deletions of selected elements, including removal of the entire β -globin LCR, does not abrogate the accessibility of chromatin at remaining elements (Bender et al., 2000a; Bender et al., 2000b; Hay et al., 2016). Whether the reduced accessibility at the globin genes represents a lower accessibility probability at a single-cell basis or more persistent on/off fates at the level of cell populations requires further investigation.

One additional observation from the RT-PCR data from the R2-only mouse model is the limited effect of the enhancer elements on the embryonic ζ -globin gene. The differential

regulation of embryonic and adult globin genes in particular has been a topic of intense scrutiny for several decades, and remains a key focus of the Higgs group. The trend seen here is in line with previous enhancer deletions at the mouse α - and β -globin loci (Bender et al., 2001; Duantida Songdej, unpublished) and results from the humanised mouse model (Wallace et al., 2007). Potential mechanisms for differential adult/embryonic gene regulation include the extent and nature of regulatory protein binding at the relevant gene promoters, enhancer-promoter interaction compatibilities and contrasting chromatin environments (Blobel et al., 2015; Vinjamur et al., 2018). For the purpose of this project, the differential regulation of the α - and ζ -globin genes will not be probed further, but is important to note when considering effects on the erythroid cell phenotype.

To interrogate the question of element co-operation with an extended series of enhancer element combinations of the α -globin super-enhancer, I could no longer rely on the production of numerous *in vivo* models. Instead, I required a higher-throughput method for erythrocyte generation and readout. In Chapter 4, I describe the development of a scalable method to produce and isolate erythroid cells *in vitro* for this purpose. In Chapter 5, I apply this system to the question of super-enhancer co-operativity, by replicating data from existing *in vivo* models and creating new, as yet untested, enhancer combinations.

Acknowledgements

Maria Suciu performed the DNase-seq experiment, analysis and digital footprinting from E10.5 blood. Christian Babbs produced the template BAC from which the synthetic constructs were amplified and performed invaluable initial caesium chloride preparations and BAC transfection experiments. Leslie Mitchell designed and performed the BAC synthesis in its entirety. The lipofection method optimised here was based on a protocol from Associate Professor Ben Davies (Wellcome Centre for Human Genetics, Oxford). Damien Downes assisted with the sample preparation, experiment coordination and data analysis for the linked-read experiment (experiment performed by the Oxford Genomics Centre, Wellcome Centre for Human Genetics). Jacqueline Sharpe, Jackie Sloane-Stanley and Hannah Salisbury from the Transgenics Core Facility at the Weatherall Institute of Molecular Medicine performed karyotyping of mESCs and all subsequent mouse handling from blastocyst injections to maintenance and breeding. Christian Babbs performed dissections of E10.5 embryos.

Chapter 4 – The embryoid body as a source of mouse erythroid cells *in vitro*

4.1 Introduction

In vitro differentiation of stem cells as an experimental tool

When tackling questions of mammalian genetics, the use of mouse models has long been the gold standard (Capecchi, 2005; Glaser et al., 2005; Morse, 2007; Parker and Peterson, 2018). No other mammalian system offers the same relative ease of breeding and maintenance, the completeness of a phenotypic readout that runs throughout development and the consistency of using inbred strains which removes the confounding effects of genetic background. However, as methods at all stages of the experimental pipeline, from genetic engineering through to downstream assays, increase in availability and throughput, the universal use of mouse models becomes impractical and unethical. This is particularly true of the α -globin locus where, as yet, no significant effects of genetic manipulation have been observed outside of the well-characterised erythroid lineage.

When considering *in vitro* alternatives to animal models, embryonic or pluripotent stem cells are a valuable resource for many applications. *In vitro* differentiation of stem cells offers unparalleled access to a range of molecularly-defined cell types from a single clone of origin. This permits experimentation in desired tissues of interest while reducing our reliance on immortalised cell lines which inevitably have a corrupted genetic background. In the mouse system, mESCs offer the advantages of an “off-state” control cell type and the versatility to

create either *in vitro* tissues or mouse models, as required. The need for well-established, robust *in vitro* differentiation protocols becomes clear when considering the increasingly high-throughput nature of experimental designs in molecular biology. It is critical that derived cells faithfully recapitulate *in vivo* counterparts in a reproducible manner and, from a practical perspective, that they can be produced at the scale required for analysis.

Protocols for mESC differentiation *in vitro* often involve the production of embryoid bodies (EBs). The general principle behind EB generation from mESCs involves growing cells in conditions that promote three-dimensional growth in the absence of leukaemia inhibitory factor (LIF) or embryonic feeder cells (Keller, 2005). Resulting spheres, typically up to 1 mm in diameter, are heterogeneous populations of cells with representation from all three germ layers (Spangler et al., 2018). The output from EBs can be enriched for specific lineages of choice through altering the differentiation microenvironment by the addition of small molecules or alternative plating strategies. Chemical modifications include using different base media and growth factor additives, while culture conditions can be substituted for hanging droplet-style culture, suspension culture, attachment to extracellular matrices or stromal cells, or incubation in bioreactors (Bratt-Leal et al., 2009; Kurosawa, 2007).

Mouse erythropoiesis in a dish

Conditions for haematopoietic mouse EB differentiation were first described several decades ago (Doetschman et al., 1985; Keller et al., 1993; Schmitt et al., 1991). Colony replating detailed the emergence of a range of progenitors within EBs from day four onwards (Keller et al., 1993). This was subsequently shown to closely reflect the situation *in vivo* (Ditadi et al., 2016; Palis et al., 1999). Key to this work was the finding that, in addition to

the emergence of erythroid progenitor cells within EBs, authors reported spontaneous haemoglobinisation of EBs within a similar timeframe.

While in early studies haemoglobinisation was observed in ~1% of EBs (Doetschman et al., 1985), fine-tuning of culture conditions has now increased erythrocyte-containing EB representation to 100% (Stefanska et al., 2017). Interrogation of the detected globin gene products by isoelectric focusing (Doetschman et al., 1985), RNase protection (Wiles and Keller, 1991) and reverse transcription-PCR data (Keller et al., 1993) identified the presence of embryonic globins (*Hba-x*, *Hbb-y*, *Hbb-bhl* genes), suggesting that these erythrocytes most closely reflect an embryonic (primitive) cell state. *In vivo*, the embryonic erythroid lineage represents the first wave of erythrocyte differentiation originating from the developing yolk sac blood islands. This is followed by two fetal/adult (definitive) erythropoietic programmes, ultimately taking place in the fetal liver and bone marrow, respectively, in which the embryonic globin genes are silenced (Figure 4.1). However, detection of the embryonic globins within EBs does not necessarily mean that all EB-derived erythroid cells are embryonic-like in nature: adult globins (*Hba-a1/2* and *Hbb-b1/2* genes) can be detected across all waves of differentiation. Up to now, therefore, it has been unclear as to whether the EB's inherent erythroid population was homogenously embryonic-like or an embryonic-adult mixture, and to what extent the molecular phenotype of the chromatin environment matches that of primary erythroid cells.

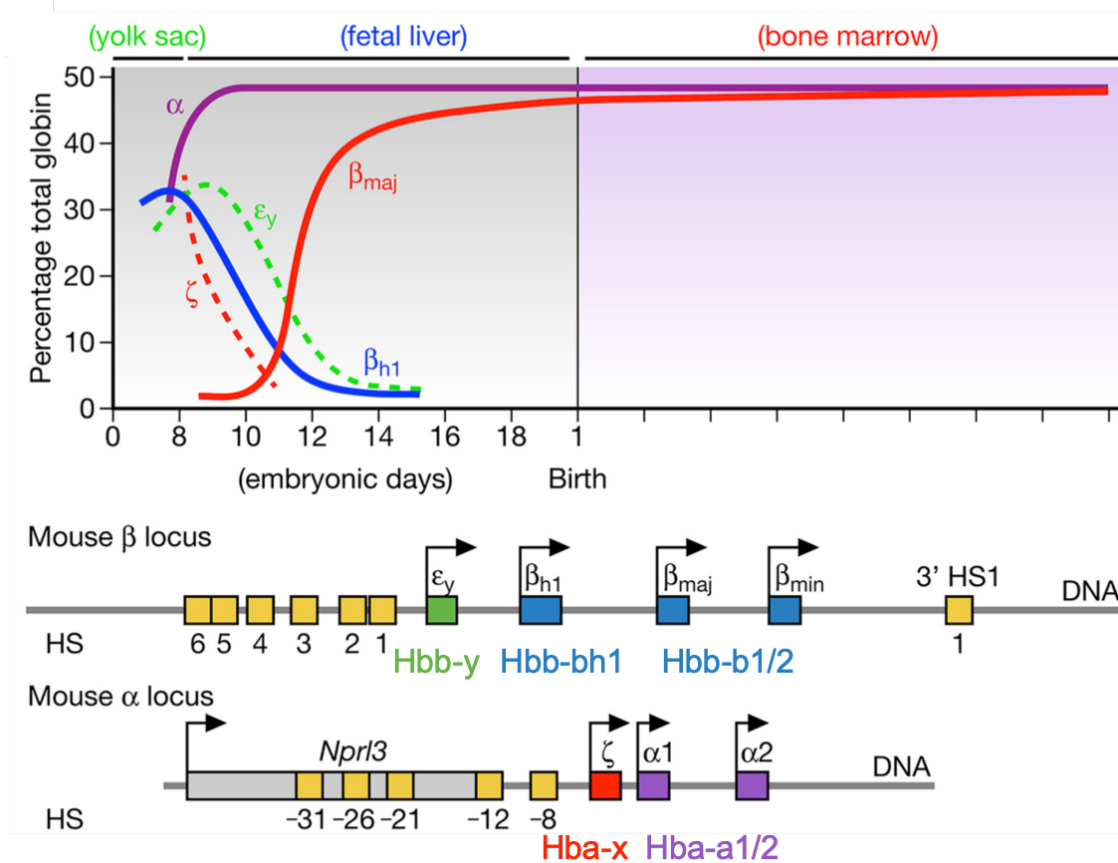


Figure 4.1: Mouse erythropoiesis occurs in three distinct waves, each with a different pattern of globin expression.

Top: relative globin gene representation through mouse development. Bottom: schematic diagrams of the mouse α - and β -globin loci. Globins are annotated with both protein and gene names for clarity. Figure adapted from Blobel et al., 2015.

In this chapter, I describe the exploitation of the EB's natural haematopoietic potential as a robust source of erythroid cells from mESCs *in vitro*. Firstly, I demonstrate that a pure erythroid population can be isolated from day 7 EBs by immunophenotype; secondly, I interrogate the molecular characteristics of the selected cells to assess their suitability as a biological model system; finally, I describe a protocol to miniaturise differentiation, opening the possibility for higher throughput studies on tens to hundreds of models in a single experiment.

4.2 Results

Isolation of erythroid cells from day 7 EBs

In designing a strategy for erythrocyte isolation from EBs, we sought a harvest timepoint that would provide the greatest number of mature red cells. We found that EB disaggregation is detrimental to cell viability beyond day 7 of differentiation, so focused our efforts on timepoints up to and including this limit. Previously published data show that globin expression increases through differentiation (Doetschman et al., 1985; Keller et al., 1993; Wiles and Keller, 1991). We confirmed this by RT-PCR at days 4-7 of differentiation (Figure 4.2A). Immunophenotyping by flow cytometry showed an accompanying increase in the red cell surface markers CD71 and Ter119 (Figure 4.2B; Fraser et al., 2007; Socolovsky et al., 2001); by day 7, EBs were ~40% positive for CD71, with roughly a third of this population also positive for Ter119.

To identify the most appropriate selection method for erythroid cells from whole EBs, we assayed expression of α -globin across CD71 and Ter119-defined populations with a mES cell line in which the enhancer-proximal copy of the α -globin gene is tagged with a yellow fluorescent protein (YFP) tag, mVenus (Nagai et al., 2002, Shaner et al., 2005). The YFP⁺ and CD71⁺ populations were found to overlap completely by flow cytometry (Figure 4.2C). Therefore, we reasoned that CD71 would be the marker for the most complete isolation of erythrocytes from whole EBs.

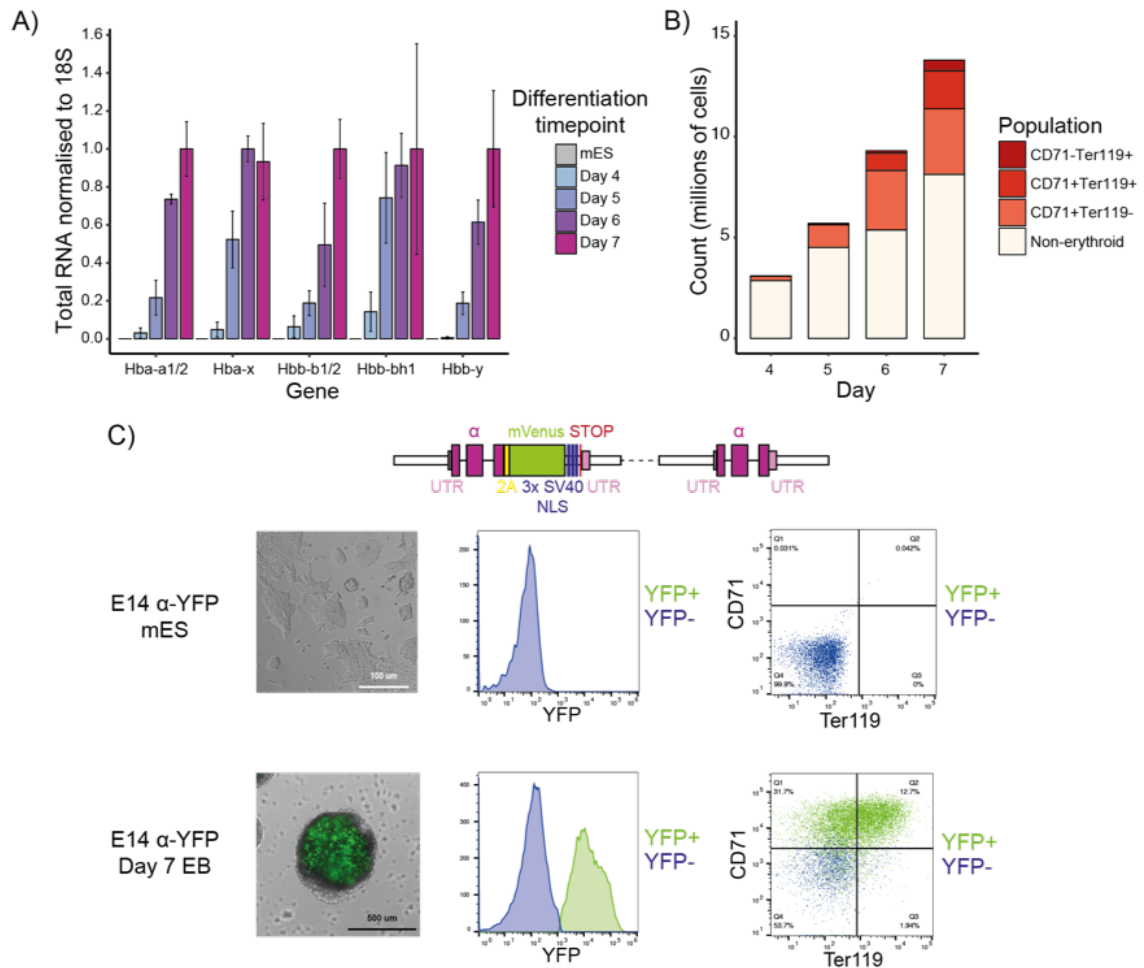


Figure 4.2: EBs can be directly exploited as a source of erythroid cells.

A) RT-PCR of mature mouse globin transcripts in mES cells and EBs between days 4-7 of differentiation, normalised to the 18S housekeeping gene. Levels are shown relative to the maximal detected expression for each gene. **B)** Cell counts for immunophenotypically-defined populations through differentiation. Data shown are from a single, typical differentiation in 10 cm dish format. **C)** Fluorescence levels of an α -globin-YFP tag in mES cells and day 7 EBs. Top: a schematic of the tag shows the insertion site after the final exon of the gene with a 2A self-cleaving peptide sequence in yellow, the mVenus coding region in green, nuclear localisation signal (NLS) repeats in blue, the STOP codon in red and the untranslated regions (UTRs) in pink. Bottom: brightfield images of mESCs and a single EB are overlaid with YFP fluorescence signal (left). Flow cytometry histograms for YFP fluorescence demonstrate the presence of an α -globin-positive population in day 7 EBs (coloured in green; middle). Flow cytometry plots for the erythroid markers CD71 and Ter119 show the overlap of YFP+ population from the histogram with the CD71+ cell populations in day 7 EBs (YFP+ cells labelled green as in the histogram; right).

Identification of the erythroid population by a cell surface marker makes cell isolation possible by FACS for small cell numbers. For larger-scale applications demanding up to 10^7 cells we optimised a magnetic column-based purification method that relies on the validated CD71 marker. We saw that selection in this manner, while resulting in modest cell losses (typical CD71+ recovery was 50-60%), could be scaled up to provide a large, pure (>96% CD71+) erythroid population for downstream analyses (Figure 4.3A). By cellular morphology with May–Grünwald–Giemsa (MGG) staining, we concluded that isolated erythroid cells are predominantly of the pro-erythroblast and basophilic erythroblast stages, in line with immunophenotyping results (Figure 4.3B).

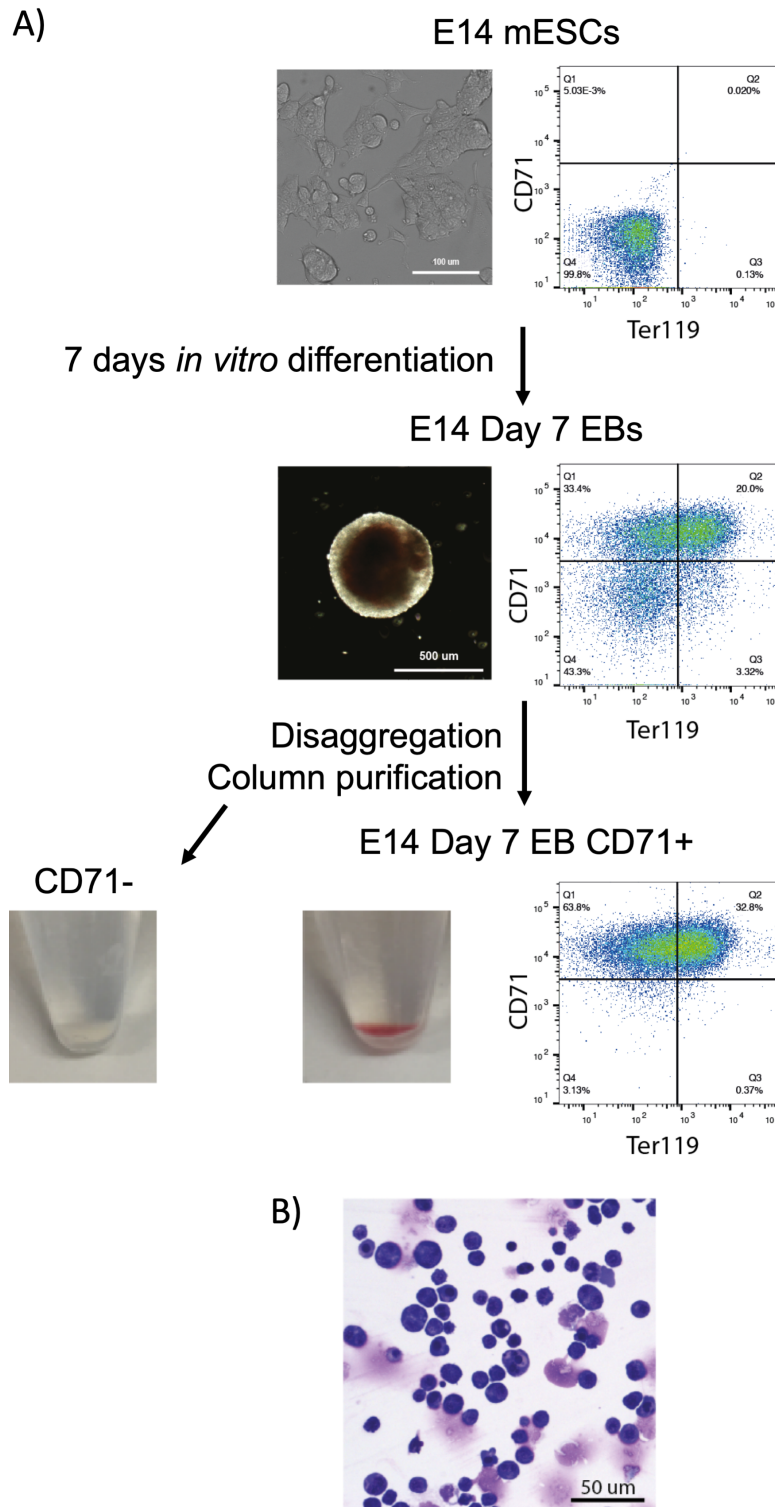


Figure 4.3: Column-based, high-purity isolation of erythroid cells from day 7 EBs.

A) Brightfield microscopy images, pellet images and flow cytometry plots for CD71 and Ter119 erythroid markers are shown for a typical EB differentiation and harvest. **B)** MGG-stained cell preparation of CD71+ cells from day 7 EBs, demonstrating that the population is pure for early differentiation stage erythrocytes.

EB-derived erythroid cells are uniformly embryonic-like in nature

Previous work has categorised the EB erythroid output as embryonic-like at a population level on the basis of detectable embryonic globin gene products. To further investigate the predicted embryonic-like nature of EB-derived erythroid cells, we performed a more detailed analysis of the associated molecular phenotype, including ATAC-seq (Buenrostro et al., 2013) and chromatin immunoprecipitation with sequencing (ChIP-seq; Johnson et al., 2007) for a number of informative histone modifications associated with active sequence elements (H3K27ac, H3K4me1 and H3K4me3; Calo and Wysocka, 2013).

Comparison of ATAC-seq data at the α - and β -globin loci with existing results from embryonic and adult mouse primary erythrocytes (E10.5 blood and Ter119+ cells selected from acetylphenylhydrazine (APH)-treated adult spleen, respectively) supported the findings of embryonic globin expression in these cells: we observed accessibility peaks over the embryonic genes specifically in the embryo and EB samples but not in the adult tissue (Figure 4.4A). We performed principal component analysis (PCA) on all ATAC-seq peaks from each dataset and found that EB similarity to embryonic tissue was also true genome-wide (Figure 4.4B).

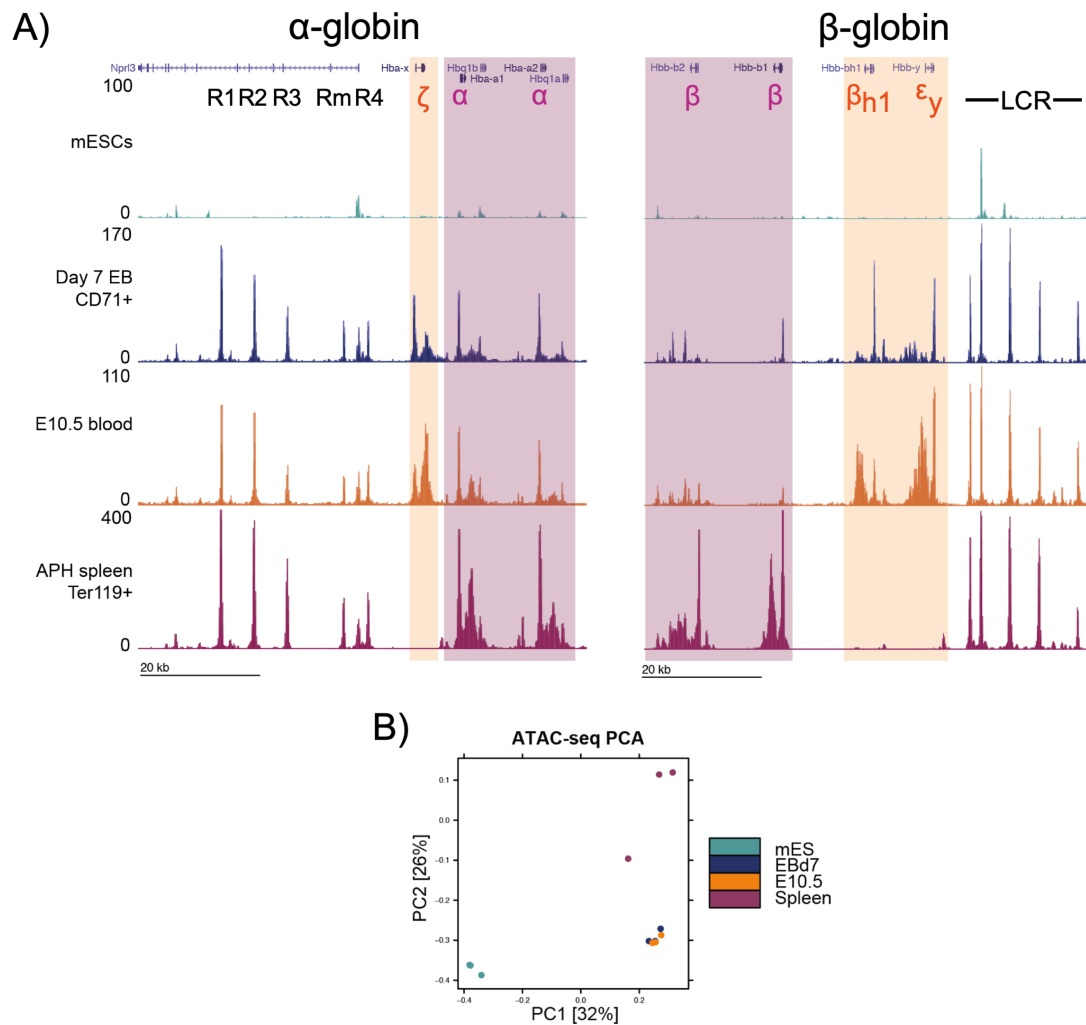


Figure 4.4: An EB-derived erythroid population resembles embryonic primary red cells by ATAC-seq.

A) ATAC-seq tracks for the α - and β -globin loci (chr11:32,131,800-32,204,799 and chr7:110,952,000-111,027,999, respectively) from mES cells and erythrocytes from day 7 EBs, E10.5 embryonic blood and Ter119+ erythrocytes from adult APH-treated spleen. Tracks are RPKM-normalised and averaged over three biological replicates. Embryo-specific globin genes are highlighted in orange; adult globin genes are highlighted in pink. Other visible peaks represent erythroid-specific enhancer elements. **B)** PCA plot of genome-wide ATAC-seq peaks from the tissues shown above. EB-derived cells cluster with E10.5 primary erythroid cells, away from mESCs and spleen-derived erythrocytes.

ChIP-seq for active histone modifications showed a pattern consistent with the ATAC-seq data and previous element classification of enhancer and promoter elements (Figure 4.5; Hughes et al., 2005). Data closely resembled the pattern seen for active histone marks in primary embryonic erythroid cells (Andrew King, unpublished data). Together, these

chromatin feature annotations further supported the argument for EBs generating embryonic-like erythroid cells at a population level.

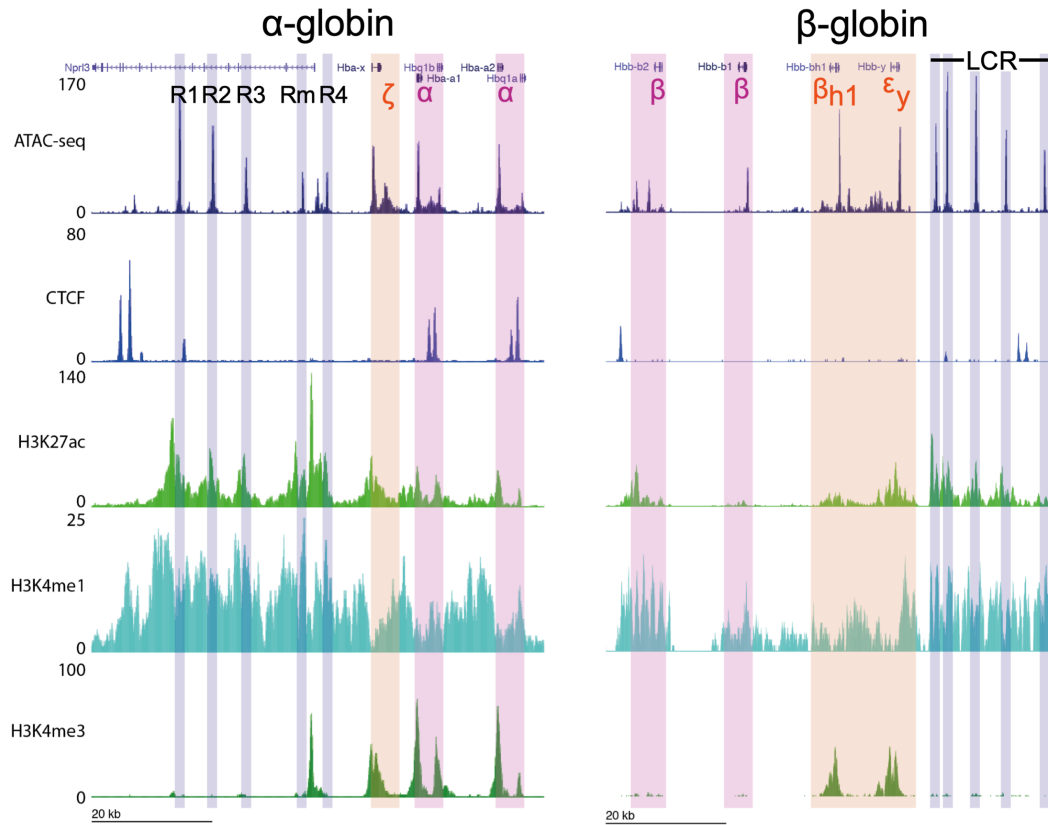


Figure 4.5: The globin loci are marked with active histone modifications in EB-derived erythrocytes.

ATAC-seq and ChIP-seq tracks for the indicated histone marks covering the same regions as in Figure 4.4. ChIP-seq data are shown as average tracks across two biological replicates, normalised by RPKM. Annotated enhancer elements are highlighted in blue, embryo-specific genes in orange and adult genes in pink.

Finally, to test whether the identified embryonic-like characteristics were representative of a pure cell population, rather than an artefact of mixed population-based assays, we performed single-cell RT-PCR for 40 genes with the potential to distinguish embryonic from fetal/adult lineages, including the globin genes. Table 4.1 highlights the justification for the inclusion of each probe in the set. In total, we obtained data from 149 cells from different immunophenotypic populations of embryonic and fetal/adult lineages (Figure 4.6). Briefly,

cells from adult APH-treated spleen and E12.5 fetal liver were sorted based upon CD71/Ter119 immunophenotyping into S1, S2 and S3 populations in accordance with the well-characterised separation of erythroid differentiation stages (Fraser et al., 2007; Pop et al., 2010). Cells from day 7 EBs were sorted into S1 and S3 populations based on the same scale. The increase in Ter119 signal from gate S1 to S3 is associated with increased erythroid maturation in both embryonic and adult tissues. Embryonic erythrocytes from E9.5/E10.5/E11.5 were sorted as single populations based on the presence of the CD71 and Ter119 markers; use of cells from three separate timepoints reflects the synchronous differentiation of primitive erythroid cells *in vivo*.

Gene	ErythronDB hit?	Other
<i>Hba-a1/2</i>		Globin
<i>Hba-x</i>	Y	Globin
<i>Hbb-bh1</i>	Y	Globin
<i>Hbb-bh2</i>		Globin
<i>Hbb-b1/2</i>	Y	Globin
<i>Hbb-y</i>	Y	Globin
<i>Hbq1a</i>		Globin
<i>Hbq1b</i>		Globin
<i>Cpeb4</i>	Y	α -locus
<i>Il9r</i>		α -locus
<i>Mpg</i>		α -locus
<i>Nprl3</i>		α -locus
<i>Rhbdfl</i>		α -locus
<i>Snrnp25</i>	Y	α -locus
<i>Abcg4</i>	Y	
<i>Aquaporin1</i>	Y	
<i>Aquaporin8</i>	Y	
<i>Arid3a</i>	Y	
<i>Ccr2</i>	Y	
<i>Cited2</i>	Y	
<i>Cox6c</i>		Housekeeping
<i>Foxh1</i>	Y	

<i>H2afy</i>	Y	
<i>Hspe1</i>	Y	
<i>Klf3</i>	Y	
<i>Lmo4</i>	Y	
<i>Mbd2</i>	Y	Sankaran et al., 2010
<i>Mpo</i>	Y	
<i>Myc</i>		
<i>Myl6</i>		Housekeeping
<i>Pbx1</i>	Y	
<i>Podxl</i>	Y	
<i>Pogz</i>	Y	
<i>Prtn3</i>	Y	
<i>Rhd</i>	Y	
<i>Sec61g</i>		Housekeeping
<i>Smarca4</i>	Y	Sankaran et al., 2010
<i>Sox4</i>	Y	
<i>Sox6</i>	Y	Sankaran et al., 2010
<i>Tmem14c</i>	Y	
<i>Vamp5</i>	Y	
<i>Zbtb7a</i>	Y	Masuda et al., 2016
Table 4.1: List of gene probes used in the single-cell RT-PCR experiment. Probes were chosen to detect differential expression between embryonic and adult erythroid cells of different maturation stages. ErythronDB is a database of erythroid gene expression (Kingsley et al., 2012).		

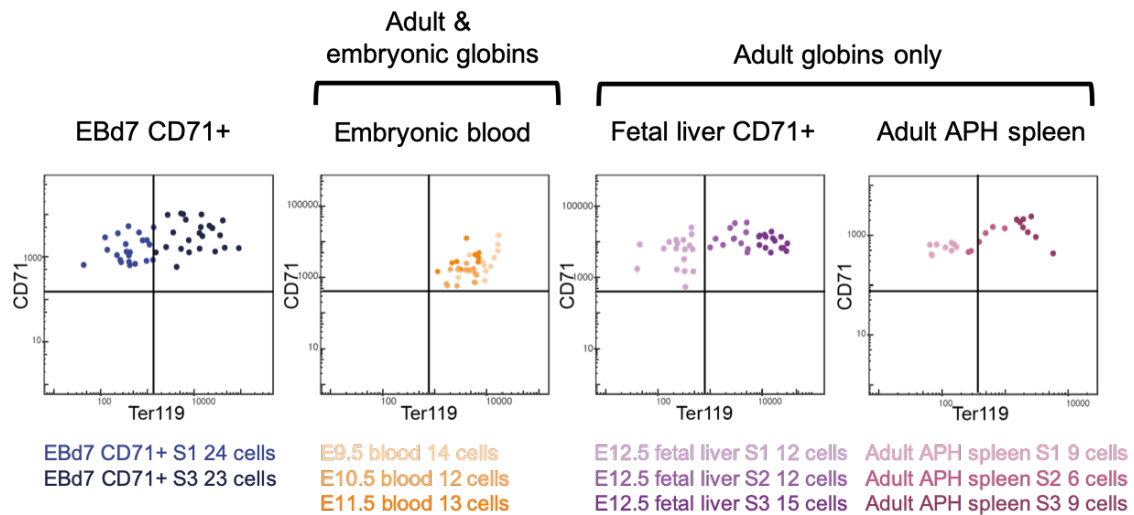


Figure 4.6: Characteristics of erythroid cells used in single-cell RT-PCR by immunophenotype and tissue type.

Flow cytometry plots are shown for CD71 and Ter119 erythroid markers for EBs and three primary tissue types; each primary tissue is labelled above with the expected globin expression. All sorted cells were positive for CD71; where applicable the population was further segregated into gates S1-S3 based on the level of Ter119 detection (Pop et al., 2010). Higher Ter119 labelling is associated with a more differentiated stage of erythropoiesis.

We detected expression of both embryonic globins (*Hba-x* and *Hbb-y*) and adult globins (predominantly *Hba-a1/2*) in all selected EB cells, indicating that the population is homogeneously embryonic-like in nature (Figure 4.7). Transcript levels for the embryonic globins increased with developmental time between E9.5-E11.5 stages *in vivo*; from EB-derived cells we observed transcript levels lower than any *in vivo* sample. Thus, while globin expression from EB-derived cells most closely resembled those of E9.5 primary erythrocytes, it is possible that they may best reflect an even earlier, untested stage of erythroid maturation (E7.5-E8.5) or that the full extent of globin expression is not realised in this model. We also saw a difference in expression associated with Ter119 level in the EB-derived populations, comparable with those seen for the fetal liver and spleen tissues or between E9.5-11.5 maturation stages. This suggests that the CD71+ cell selection from EBs

could be further refined to isolate a less-mature or more-mature state of erythroid differentiation based on Ter119 levels, if needed.

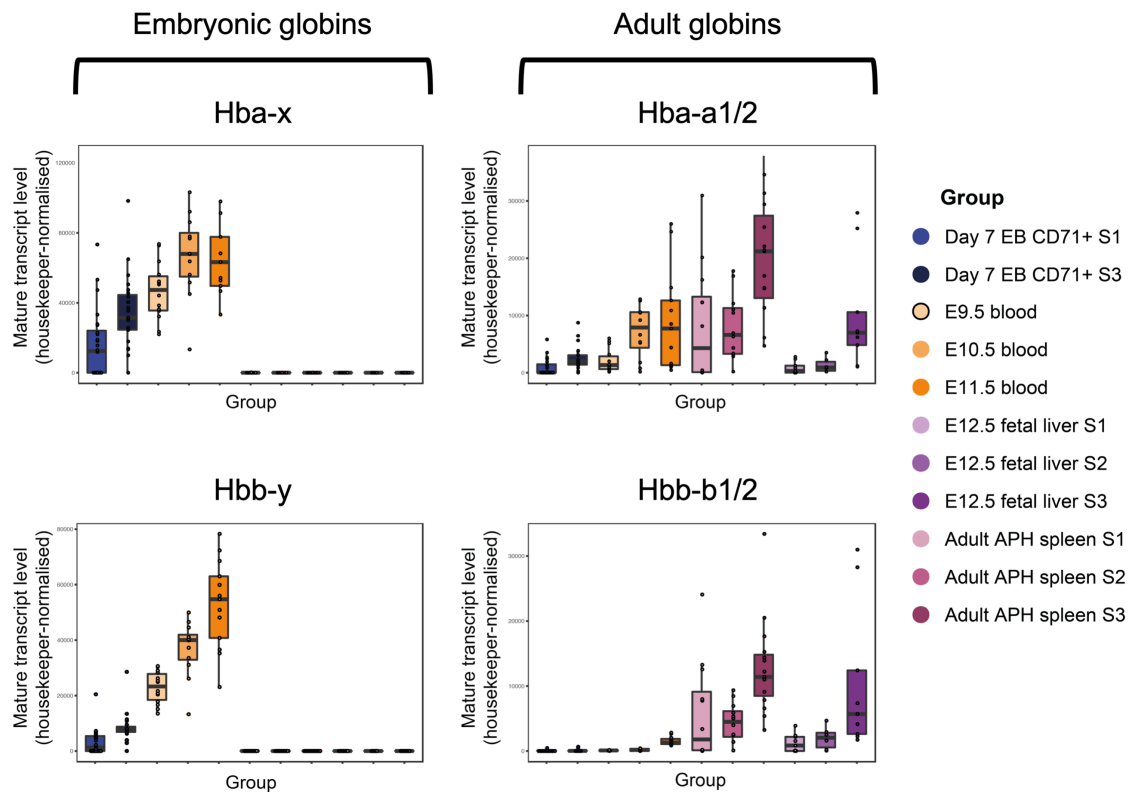


Figure 4.7: EB-derived erythroid cells are homogeneously embryonic-like by globin expression at a single-cell level.

Expression of embryonic and adult globins, normalised to housekeeping genes *Cox6c*, *Myf6* and *Sec61g* is shown for the populations described in Figure 4.6.

While expression of the globin genes provides the most striking difference between embryonic and fetal/adult erythroid cells, other genes we included for single-cell RT-PCR were also able to broadly distinguish the major populations. PCA of all housekeeping genes-normalised expression measurements clearly separated the fetal liver and adult APH-spleen cells from embryonic E10.5 and EB-derived erythrocytes (Figure 4.8A). Furthermore, PCA of the non-globin genes alone was able to segregate the cells in a similar manner, albeit to a lesser extent (Figure 4.8B). Multiple probes showed higher expression for either embryonic (EB and embryonic blood) or fetal/adult (fetal liver and spleen) cell types, indicating that

EB-derived cells resemble embryonic red cells at the level of the wider transcriptional programme: not just at the globin loci (Figure 4.8C-D).

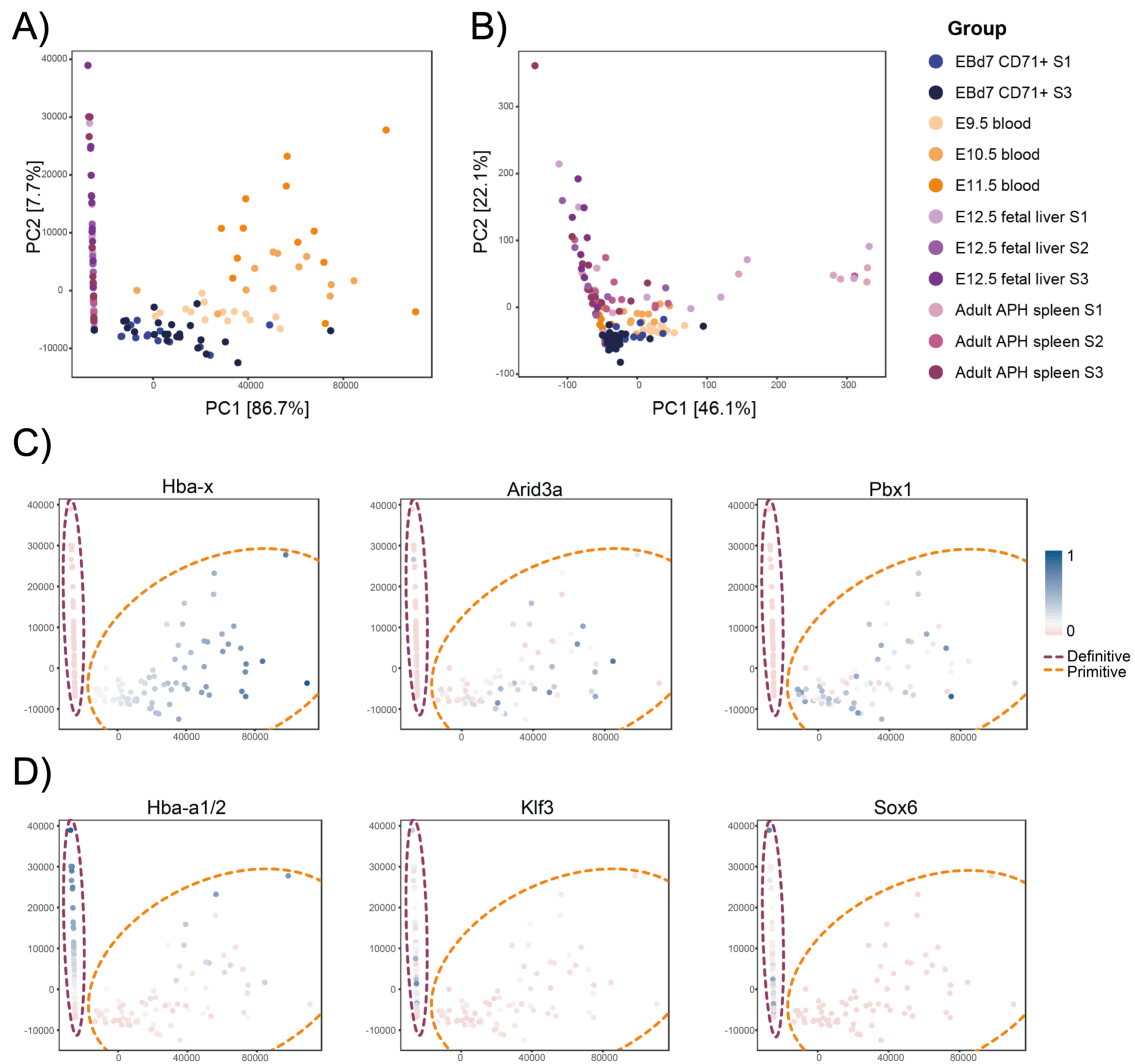


Figure 4.8: EB-derived single cells recapitulate an embryonic transcriptional programme.

A) PCA of housekeeping genes-normalised expression levels for the complete probe set.
B) PCA of housekeeping genes-normalised expression levels for non-globin probes.
C, D) Expression of six example genes where enrichment is seen in either embryonic or adult cells, respectively. Expression levels for the indicated genes are overlaid on the PCA plot from part A) with colour density (red to blue) representing a proportion of the maximal detected expression level.

EB differentiation as a scalable tool for erythrocyte generation

Experiments described thus far in this chapter were performed with a conventional, bulk EB plating strategy in 10 cm dishes and scaled up by increasing the number of replicate plates, where necessary. While such large-scale culture is routine and valuable for a small number of clones at a time where large cell numbers are required, differentiation of up to tens or hundreds of clones in parallel for smaller-scale experiments would be impractical in this format. This is particularly the case for screening-style assays employing large numbers of genetic manipulations, or for testing multiple small molecules in a high-throughput manner.

To make differentiation scalable for a broader range of applications, we optimised conditions for EB generation in a 96-well plate. By plating a range of cell numbers (100-1200) into 200 µl differentiation media in individual wells of uncoated plasticware, we found that EBs could be formed as seen for bulk cultures (Figure 4.9A). The initial plating density of mESCs was critical for the characteristics of EB formation: higher cell densities at the time of plating resulted in the formation of fewer, larger EBs (Figure 4.9B-C).

The differences in EB aggregation properties had consequences for erythropoiesis: the percentage of CD71⁺ erythroid cells decreased with increasing EB size (Figure 4.9D). These data are in line with previous findings about the nature of erythropoiesis within EBs and the observation that particularly dense cultures of EBs favour cardiac rather than haematopoietic differentiation (Lee et al., 2011). Successful differentiation of other lineages of all germ layers within EBs has also been found to be directly influenced by EB size (Messana et al., 2008; Niebruegge et al., 2009; Valamehr et al., 2008).

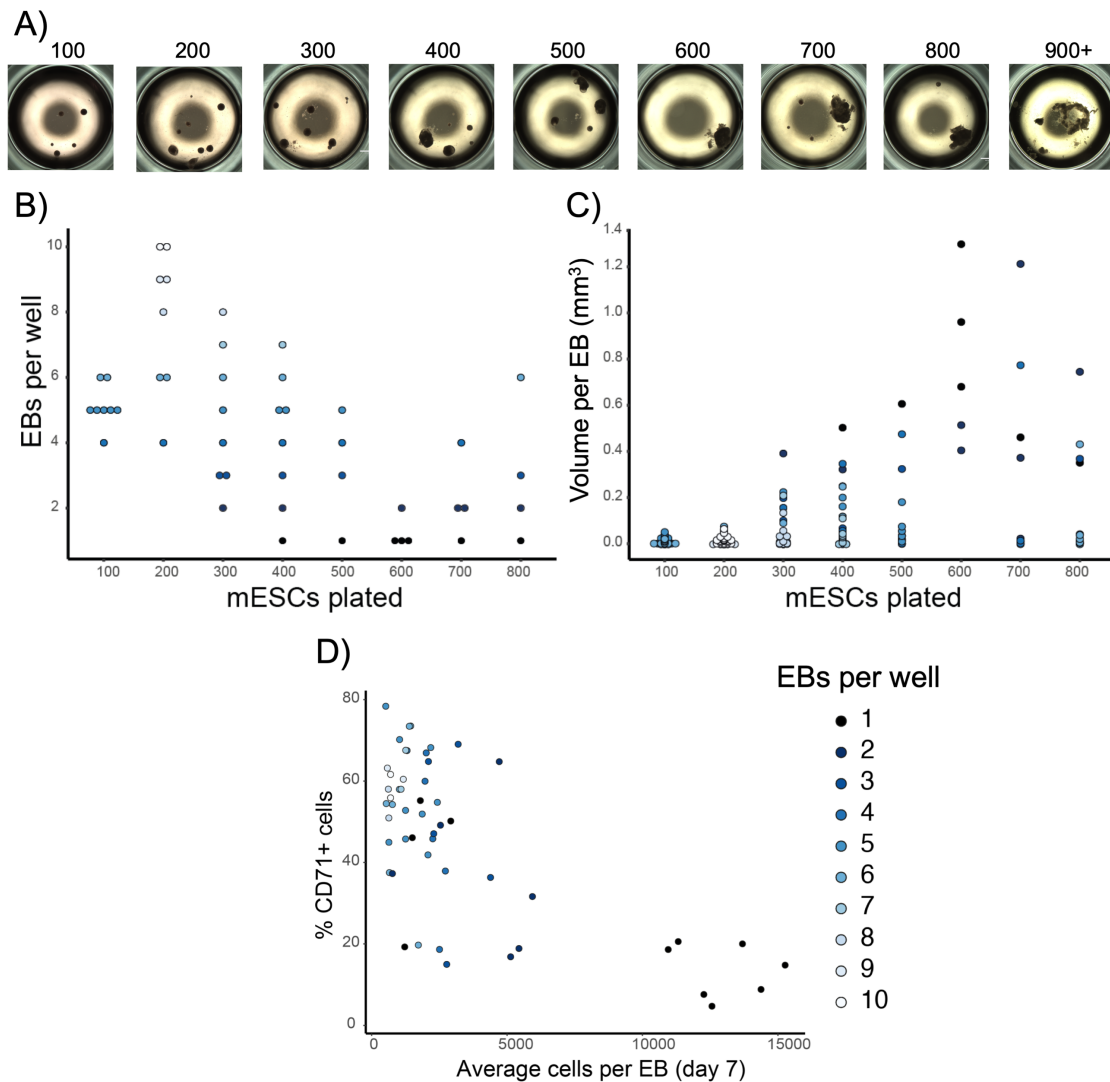


Figure 4.9: 96-well EBs vary in size and number depending on mESC plating density.

A) Example brightfield images of day 7 EBs from wells plated with the indicated number of cells. **B)** Total number of EBs compared to the number of mESCs plated, per well. Each dot represents data from an individual well. The number of individual EBs formed per well decreases with increasing mESC plating density. Above 900 cells per well, EB structure became abnormal to the extent where irregular masses were entirely non-spherical and could not be accurately counted. **C)** Individual approximate EB volumes plotted against mESC plating density. Each dot represents data from an individual EB. Volume estimates were calculated from EB diameter measurements by microscopy. The volume per EB increases with mESC plating density. **D)** Percentage CD71+ cells from day 7 EB 96-well wells plotted against the average number of cells per EB for each well. Dots are coloured in all plots by the number of EBs in each recorded well, as indicated in part B).

In designing a miniaturised differentiation system to probe questions of gene regulation, the most important feature to optimise was the generation of a population of CD71+ *bona fide* erythroid cells that could be reliably experimentally tested in each well. A popular solution to standardise differentiation and reduce well-to-well variation in high-throughput differentiation systems is to optimise conditions for uniform EB growth, typically through the use of hanging droplet culture or controlled single-EB conditions (Choi et al., 2010; Dang et al., 2004; Flamier et al., 2017; Ng et al., 2005). However, key to the study of gene regulation is the fact that extreme variation in mature transcript levels between individual cells is seen even in theoretically homogeneous populations such as a single culture of mESCs (Faure et al., 2017; Mantsoki et al., 2016), and is also apparent in our single-cell RT-PCR results for the globin genes among immunophenotypically-defined populations derived from bulk culture EBs (Figure 4.7). We have established here that single, large EBs show attenuated erythrocyte production and are, therefore, not appropriate for this experimental design (Figure 4.9D). We, therefore, set out to determine the optimal conditions to generate the greatest number of CD71+ erythroid cells per well, in order to minimise the effect of potential variation in data from small populations. Based on results from Figure 4.9, this was likely to reflect conditions resulting in a large number of small EBs per well.

By plating density, we found to be the optimal conditions for maximal erythroid cell output by CD71 staining to be 200-400 mESCs per well (Figure 4.10). This was the compromise between increasing total cell numbers and a falling percentage of erythroid cell representation with increasing mESC plating density. EB-derived erythrocytes from the resulting miniaturised protocol are comparable to those differentiated in 10 cm dish format, as judged by flow cytometry and cell morphology (Figure 4.11A-B).

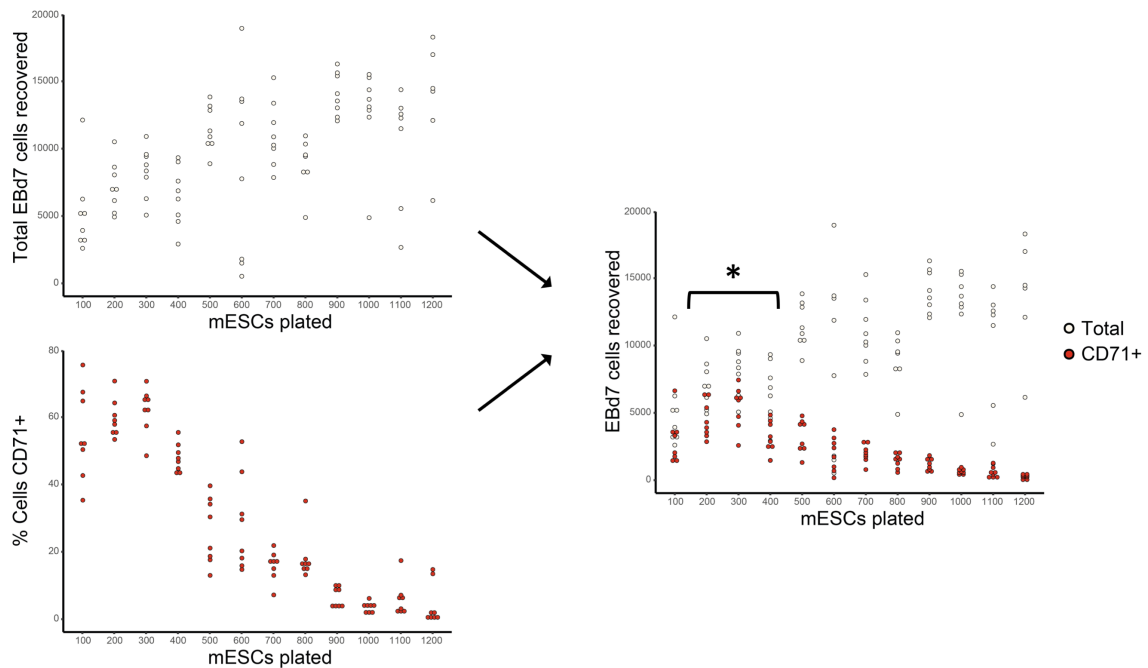


Figure 4.10: Optimal plating density for 96-well EB differentiation is 200-400 cells per well.

Left: plots detailing the total number of EB-derived cells recovered per well, each represented as a dot, for a range of plating densities and the associated percentages of CD71+ cells. Total and CD71+ cell counts were both obtained by flow cytometry. Right: resulting number of total cells and CD71+ cells recovered per well. The optimal plating density for maximal CD71+ cell output is highlighted with a bracket and asterisk (*).

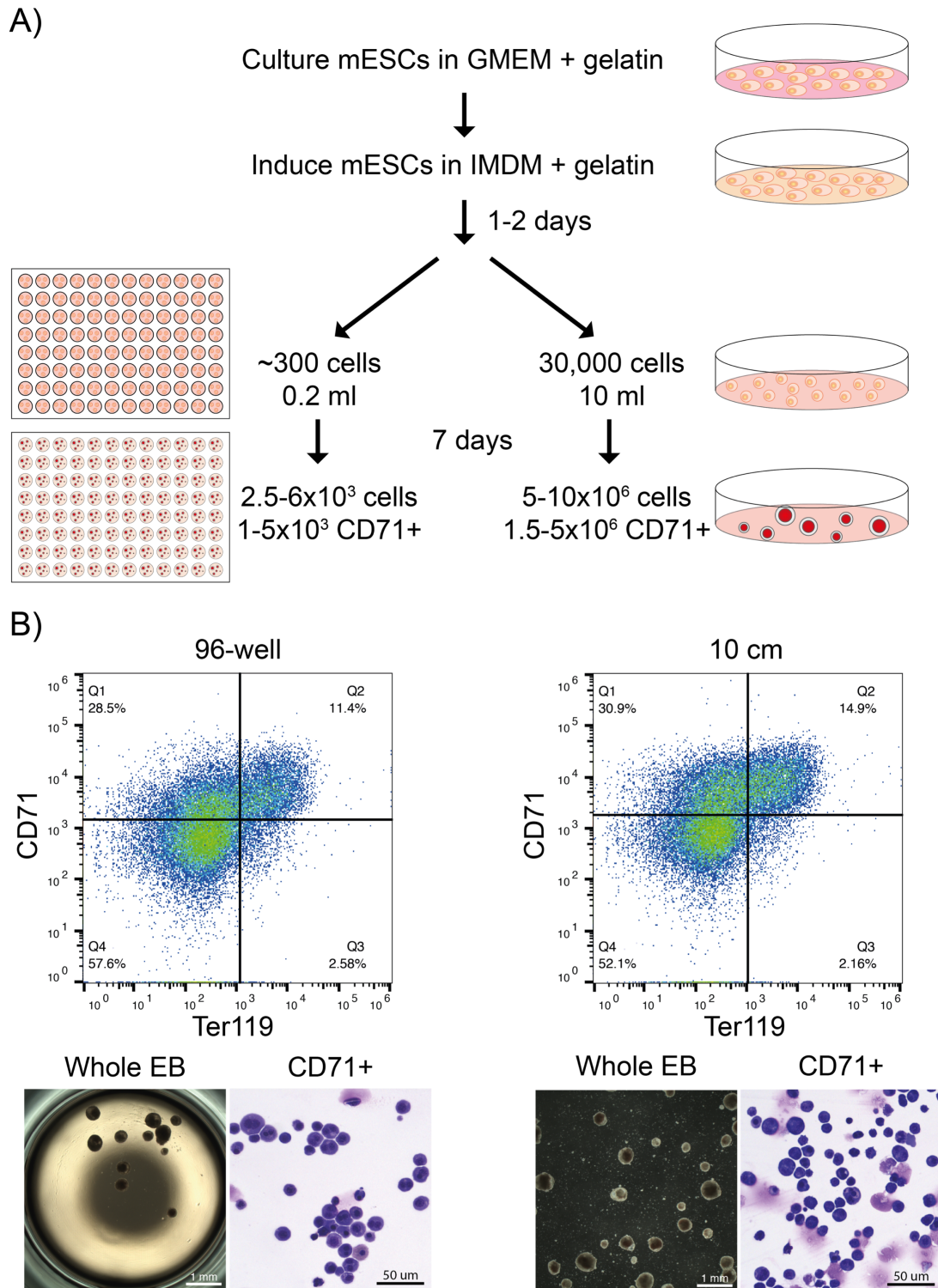


Figure 4.11: EB differentiation can be scaled down to a 96-well format without compromising the erythroid output.

A) Comparison of protocols for haematopoietic EB differentiation in 10 cm dish and 96-well plate formats. **B)** Comparison of erythroid outputs from EB differentiation at two scales by flow cytometry for erythroid markers, gross-scale microscopy of whole EBs and cellular MGG staining. Whole EB images show a single 96-well and a sample of a 10 cm dish, respectively.

To further validate the equivalence of 96-well EB-derived erythroid cells to those from routine 10cm dish culture, we applied column CD71+ purification to a pool of 96-well EBs and performed ATAC-seq. As for conventionally cultured EB-derived cells, the 96-well EB-derived erythroid cells showed specific chromatin accessibility over all of the globin genes and clustered together with primitive cells by genome-wide analysis of ATAC-seq data (Figure 4.12A-B). Together, these results indicate that it is possible to scale down the haematopoietic EB differentiation procedure into to a miniaturised 96-well format, without compromising the efficiency and faithfulness of differentiation.

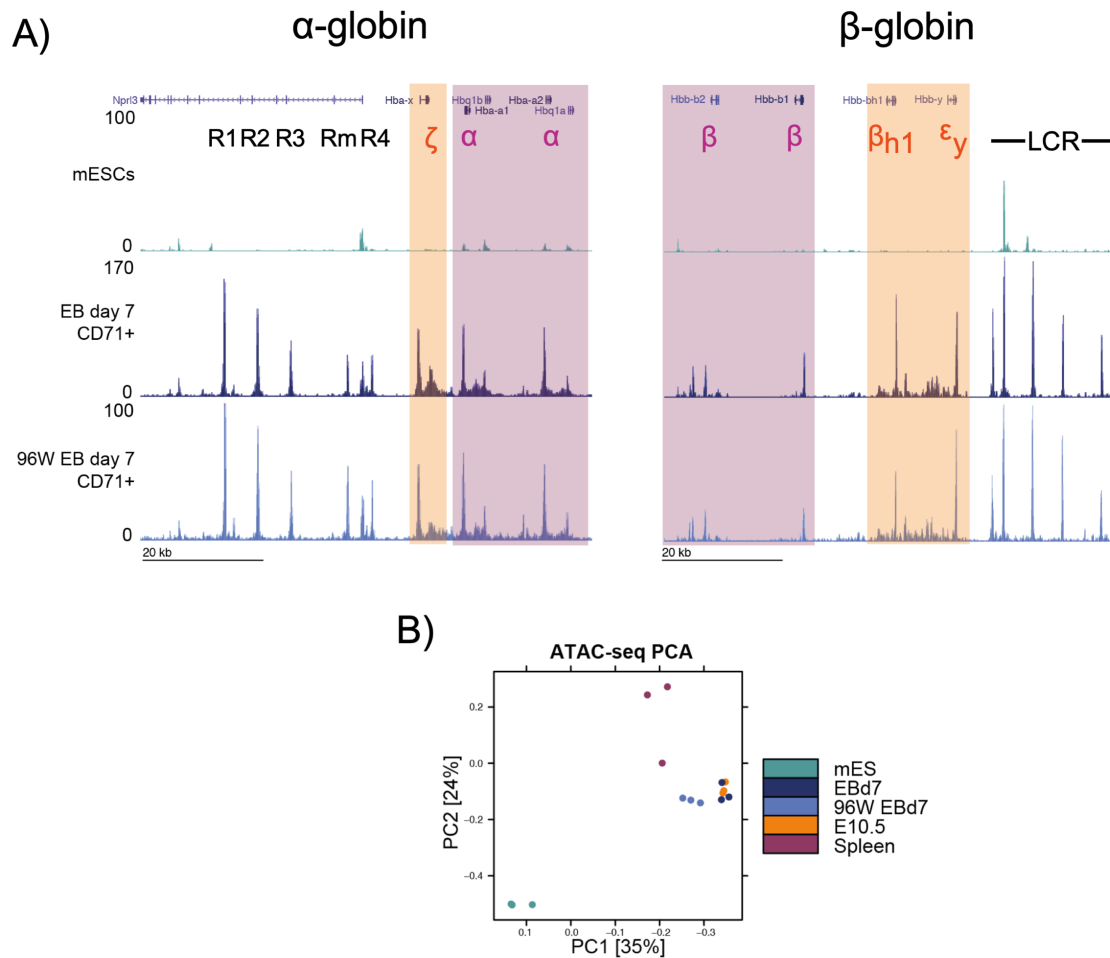


Figure 4.12: EBs differentiated in 96-well format produce primitive-like erythrocytes by ATAC-seq.

A) ATAC-seq tracks for the α - and β -globin loci with globin genes highlighted as in Figure 4.3. Day 7 EBs differentiated in 96-well plates (96W EB day 7) produce an erythroid output comparable to material generated in bulk culture (EB day 7). **B)** PCA of all ATAC-seq peaks from mES cells, EB-derived erythroid cells and primary cells. Cells from 96-well miniaturised EB culture cluster with bulk day 7 EB-derived cells and E10.5 embryonic erythrocytes.

Together, these results validate tools to exploit the spontaneous differentiation of erythroid cells within EBs as a readily-accessible and scalable erythroid cell population ($10^2 - 10^7$ cells/clone/plating). By selecting for erythroid cells expressing the transferrin receptor (CD71), we can routinely isolate and characterise a large, homogeneous population of uniformly embryonic-like mouse erythroid cells. Finally, we present a protocol for

miniaturisation of the differentiation protocol, which offers the potential to perform high-throughput studies on tens to hundreds of experimental models in a single experiment.

4.3 Discussion

Requirements for an effective in vitro differentiation system

When deciding upon an appropriate system for *in vitro* experimentation, there are several basic requirements that must be addressed.

The ideal *in vitro* system should be:

1) Amenable to clonal genetic engineering

For cells to be genetically engineered, potentially multiple times over, it must be possible to maintain them in an actively dividing state through targeting, single-cell cloning and expansion for screening. While possible for most cell lines, the same cannot be said for many populations derived and expanded directly from primary tissue (Diepstraten and Hart, 2019).

Mouse ESCs are an ideal cell type for genome engineering, and have been used in this way through a range of method developments: from homologous recombination and viral integration strategies to zinc-finger nuclease approaches and CRISPR/Cas9 technologies. Clones can be rapidly expanded or separated to single cell-derived colonies and routinely maintained in the hundreds by a single researcher.

2) Homogeneous in cell identity

When studying the effects of genetic manipulation, it is key that any observed effect can be traced back to the altered gene or genes. In an *in vitro* context, this means either culturing a homogenous cell type such that any changes in differentiation state can be monitored

directly, or having a robustly selectable population from a cultured mixture, which is not susceptible to changes in cell state under the genetic manipulation conditions of choice.

By their nature, EBs are inherently a mixed populations of cells comprising outputs from all three germ layers of early embryo development. To overcome this issue, we have developed a stringent selection method that is able to isolate just α -globin-expressing erythroid cells. Although variation does remain within this population, for example in the availability of the Ter119 cell surface marker, we have shown that all cells are embryonic-like in nature.

3) Scalable

Current techniques in molecular biology allow for complex methods to be performed on impressively small numbers of cells, meaning that tens to hundreds of genetically- or chemically-modified samples can reasonably be processed in a single experiment. However, there remain some protocols for which a large number of cells is required, due to the population-based nature of the assay (e.g. Capture-C, ChIP-seq, protein quantitation). The ideal *in vitro* system should, therefore, be applicable to both small numbers of cells for a large number of clones, or large numbers of cells for a few select models at a time in a time- and cost-efficient manner. While generally not an issue for tissue-specific cell lines, systems involving stem cell differentiation often require the use of multiple costly cytokines or complex plating strategies that prohibit bulk experiments.

The system described here can generate embryonic-like mouse erythroid cells for most existing experimental assays. By plating in a bulk 10 cm dish format, we demonstrate an inexpensive, single-plating protocol to derive tens of millions of erythroid cells in a single experiment. These numbers make samples amenable to a wide range of molecular biology

techniques including assays for DNA accessibility (ATAC-seq; $\sim 10^5$ cells), DNA binding events (ChIP-seq; 10^6 - 10^7 cells) and chromatin confirmation (Capture-C; $<10^7$ cells). At the other end of the spectrum, the system can be used to multiplex the differentiation of hundreds of clones for low-scale analyses in parallel. These experiments would be unachievable if constrained to the use of mouse models. High-throughput readouts such as flow cytometry are ideally suited to this system; as such, this differentiation format could conceivably be applied to screens for genetic or chemical manipulations.

4) Genetically and phenotypically stable

Culturing cells outside of an organism for any period of time removes the requirement for them to function in a life-sustaining manner. The newfound selection pressure to divide rapidly can induce everything from simple gene mutations up to changes in karyotype, which together make phenotype results unreliable and difficult to reproduce between experiments. This effect is well-documented in cells that have undergone immortalisation steps, such as the human HEL, K562 and HUDEP-2 haematopoietic cell lines and mouse erythroleukemia cells (Diepstraten and Hart, 2019; Mamaeva and Tsvileneva, 1985; Miller et al., 2002; Preisler et al., 1976).

Mouse ESCs are unique in that their continuing “normal” function can be confirmed through their injection into mouse blastocysts and contribution to a viable organism. As cultured cell lines, they are relatively resistant to clonal drift and can undergo multiple re-cloning steps without affecting their karyotype or pluripotency. This is demonstrated in Chapter 3 where we generated a mouse model from mESCs that had been through four successive rounds of genomic engineering. Multiple clones of the same genotype should be used, nonetheless,

especially when performing genome engineering that could have unexpected “off-target” effects.

5) Faithful to *in vivo* phenotypes

All *in vitro* experiments must be designed with the caveats of the system in mind. As such, genetic modifications must be targeted in a cell type where a direct effect on the phenotype is anticipated. However, even with these considerations made, for many cell lines the exact state of the *in vitro* population is unclear or native signalling processes are impaired to the extent that even well-established genotype-phenotype links can be prevented from being faithfully reproduced (Fernández-Nestosa et al., 2008; Kotkow and Orkin, 1996; Skarpidi et al., 1998).

The recapitulation of erythroid-specific phenotypes using the mESC-EB differentiation system will be explored in Chapter 5 where I reproduce data from existing variations of the α -globin super-enhancer *in vitro*.

Together, I describe an *in vitro* mESC differentiation method that can be used to study manipulations of the α -globin locus by faithfully mimicking embryonic blood development. In Chapter 5, I proceed to use the system to study a number of novel variations of the α -globin enhancer cluster.

Data availability

Spleen ATAC-seq: Hanssen et al., 2017

E10.5 ATAC-seq: King et al., in preparation

All other data in this chapter are presented in a manuscript ready for submission to eBlood (Francis et al., in preparation).

Acknowledgements

Experiments reported in this chapter were performed by myself with members of the Higgs and Hughes groups. Mira Kassouf introduced the EB system to the group and developed the column-based selection approach for erythrocyte isolation. Danuta Jeziorska and Andrew King designed and generated the α -globin-YFP mESC line. Mira Kassouf and Matthew Gosden performed ATAC-seq from E14 WT EB day 7 CD71+ cells. Caroline Harrold performed ATAC-seq on E14 mESCs and collaborated with me for the ChIP-seq of histone modifications in E14 WT EB day 7 CD71+ cells. Christian Babbs collected embryonic tissue for the single-cell RT-PCR experiment; Robert Beagrie performed the Biomark experiment and analysed the raw data for plotting.

Chapter 5 – Rebuilding the α -globin super-enhancer from the bottom-up

5.1 Introduction

So far in this thesis, I have reported an unexpected result from an *in vivo* single-enhancer model of the mouse α -globin super-enhancer and described the development of an *in vitro* (mESC-EB) system that can be used to study this genetic model and others. In this chapter, I validate the EB-derived erythroid population as a system to faithfully recreate established genotype-phenotype links *in vitro*. I proceed to analyse additional, previously untested combinations of the α -globin super-enhancer elements to ask what is meant by our definition of “active” and “inactive” elements. For clarity, a schematic summary of the modified iterations of the α -globin locus used in this chapter (and throughout this thesis) is provided in the included Appendix.

5.2 Results

Generating a mESC line hemizygous for the α -globin locus

The Synthetic RMGR system described in Chapter 3 can be used to produce heterozygous mESC lines carrying a range of conceptually novel manipulations of the α -globin regulatory landscape. However, the value of this tool in studying gene regulation is limited *in vivo* by the effect of mutations on the viability of mice when bred to homozygosity (as seen with the R2-only model in Chapter 3) and *in vitro* by the presence of a remaining WT allele. To make post-RMGR mESCs amenable to analysis *in vitro*, I needed either to develop a strategy to make lines homozygous for the designed locus of interest, or remove the WT allele.

Interference of the untargeted allele is a common issue in mouse genome engineering that has yet to be easily resolved (Glaser et al., 2005). In practice, the primary options were A) to introduce lox sites into the remaining WT allele of RMGR-ready mESCs (described in Chapter 3) to allow for parallel RMGR of both alleles; B) to introduce regularly spaced single nucleotide variants into the synthetic locus design to allow for de-multiplexing of synthetic and WT alleles from downstream NGS and gene expression data or C) to delete the WT locus. The low efficiency of the RMGR targeting step made option A) technically unfeasible. Option B) was also not desirable due to the density of variants required to completely segregate data from the two alleles (~1 variant per 75 bp for compatibility with conventional short-read sequencing assays) and the potential for effects of the aberrant sequence to mask true phenotypes.

Option C), on the other hand, could be realised with alternative strategies. In one, I would target haploid mESCs to re-create the RMGR exchange locus in a cell line with only one copy of each chromosome (Wutz, 2014). However, these cultures have been shown to acquire a diploid karyotype over time and require regular fluorescence activated cell sorting (FACS)-based purification to maintain the appropriate chromosome number. Such karyotype instability would be technically demanding when processing tens of clones for a single experiment and could add a confounding effect to results from gene expression analyses (Wutz, 2014). In a more targeted strategy, I would specifically remove the WT α -globin locus from otherwise normal RMGR-ready mESCs. A similar strategy has been employed previously to study allele-specific transcription at the human protocadherin gene cluster (Canzio et al., 2019). Although resulting cells would have an imbalance in dosage of the genes within the RMGR segment (*Rhbdf1*, *Mpg*, *Nprl3*, *Hba-x*, *Hba-a1*, *Hbqlb*, *Hba-a2* and *Hbqla*), previous data suggest that this would not impact cell viability or differentiation. Indeed, such large-scale deletions occur naturally in the human population and result in no discernible traits other than α -thalassemia (Higgs, 2001; Horsley et al., 2001). Expression of the *Rhbdf1*, *Mpg* and *Nprl3* genes and *Hbqlb* and *Hbqla* pseudogenes has been perturbed in previous studies with no identified consequences for cell fate (Hanssen et al., 2017; Kowalczyk et al., 2012; Philipsen and Hardison, 2018; Mira Kassouf, unpublished data). For the globin genes, severely reduced expression *in vivo* results in an α -thalassemia phenotype (reduced mean blood cell volume and haemoglobin measurements and increased reticulocyte counts) but still allows for isolation of erythroid cells, despite the imbalance in α - and β -like globin chains (Hay et al., 2016). I therefore reasoned that the simplest way to remove any confounding influence of the WT locus on phenotypic readout from the RMGR allele would be to delete it in its entirety from existing RMGR-ready mESCs, creating a hemizygous mESC line for the α -globin locus.

I designed a strategy to delete the WT locus in RMGR-ready mESCs using CRISPR/Cas9. I chose single guide RNA target intergenic sequences flanking the α -globin locus, such that deletion of the intervening 117 kb would remove the entire self-interacting domain and flanking CTCF sites (Figure 5.1A). This deletion also removes the entire (86 kb) sequence exchanged using RMGR. When assayed by PCR over the junction created by the deletion, this appeared to be a remarkably efficient process, with 12/96 clones (13%) selected for uptake of targeting plasmids carrying a deletion of at least one copy of the α -globin locus. Deletion clones were viable and showed no growth abnormalities when compared to untargeted cells. I selected clones for the presence of an intact RMGR locus, confirmed deletion of the WT locus and checked the integrity of the predicted Cas9 cut sites on the RMGR chromosome by PCR (Figure 5.1B).

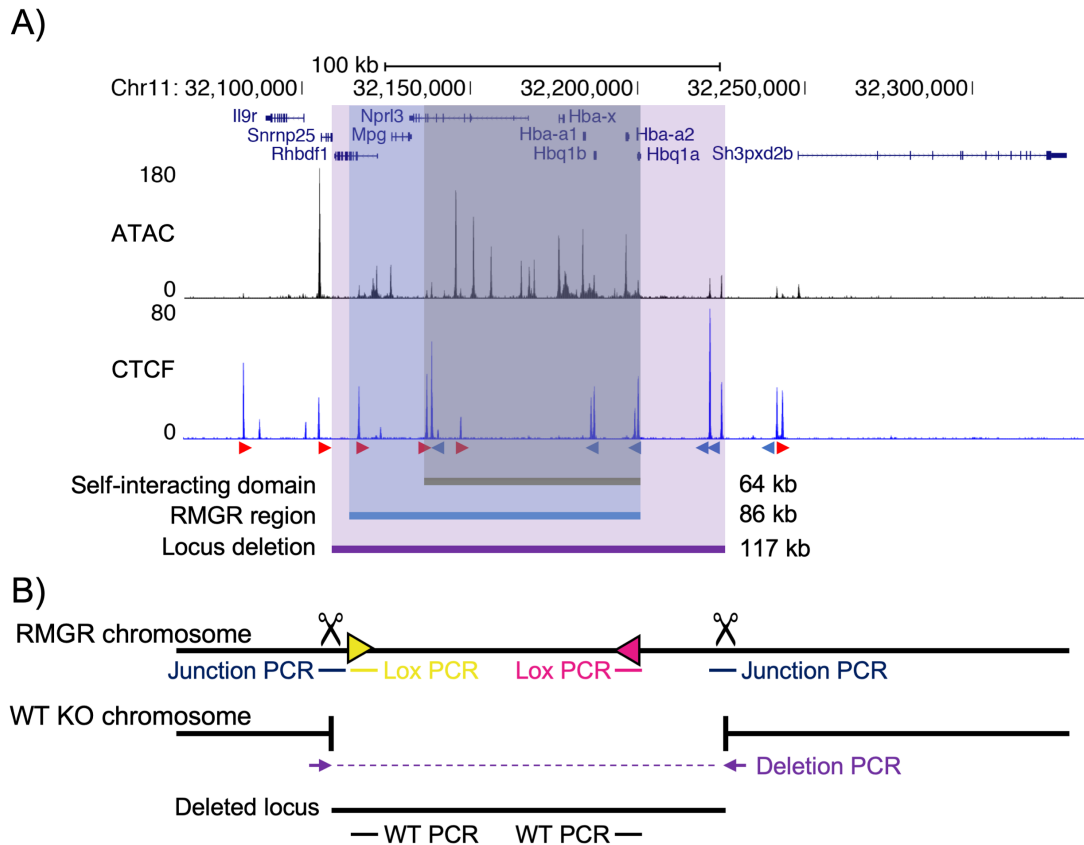


Figure 5.1: Design to delete the WT α -globin locus from RMGR-ready mESCs.

A) Tracks of ATAC-seq and CTCF ChIP-seq data from day 7 EB-derived erythrocytes. Gene annotations from the University of California Santa Cruz (UCSC) genome browser are shown at the top. Regions containing the self-interacting domain, RMGR exchange points and designed locus deletion are highlighted in grey, blue and purple, respectively, and annotated with sizes in kb. Coordinates for locus deletion: chr11:32,109,272-32,226,122; 116,851 bp. **B)** To-scale schematic of the genotyping PCRs used in mESC clone validation. Junction indels and precise deletion coordinates were identified by Sanger sequencing of PCR products. The amplification of these products indicated that the RMGR chromosome remained in its native orientation, without an inversion event (often seen in CRISPR/Cas9 targeting experiments). Lox PCRs were confirmed as positive for RMGR locus retention and WT PCRs as negative to ensure that the deleted locus segment had not re-integrated elsewhere in the genome.

To test the effect of α -globin locus hemizyosity on erythropoiesis, I took cells successfully targeted by Synthetic RMGR using the WT BAC and deleted for the WT locus through bulk *in vitro* differentiation. These cells are referred to hereafter as “WT hemizygous”. WT hemizygous EBs formed normally in terms of their aggregation properties and

immunophenotype (Figure 5.2A-B). Subsequently isolated CD71+ erythroid cells appeared to be normal by MGG staining (Figure 5.2C).

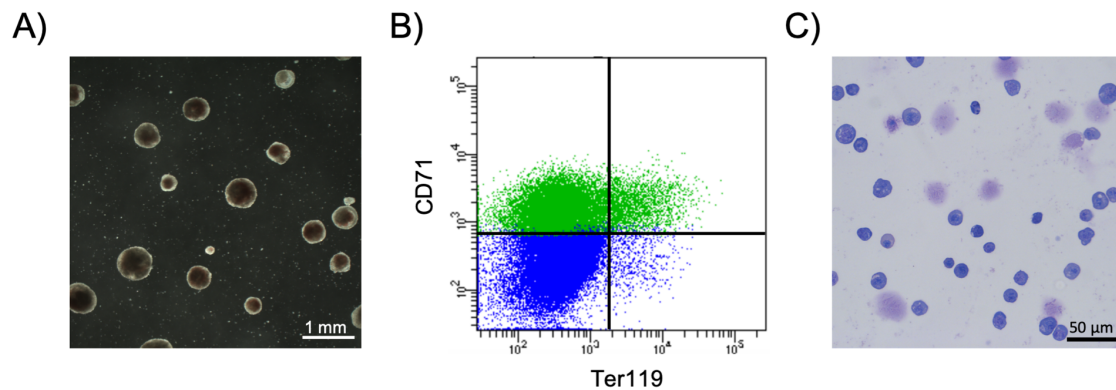


Figure 5.2: mESCs hemizygous for the α -globin locus undergo normal *in vitro* erythroid differentiation.

A) Brightfield image of day 7 EBs from WT hemizygous mESCs. **B)** Immunostaining for erythroid CD71 and Ter119 cell surface markers from disaggregated EBs shown in part A). CD71+ erythroid cells are shown in green; non-erythroid cells are shown in blue. **C)** MGG-stained cell preparation of the green CD71+ erythroid population sorted in part B). By all analyses, EBs from hemizygous mESCs developed as normal (Figure 4.3).

To test the effect of hemizygosity on the integrity of the α -globin locus itself, I performed genome-wide ATAC-seq and RT-PCR for the globin genes on EB-derived erythrocytes. ATAC-seq peaks were reduced in height specifically in the deleted region of the α -globin locus, but showed no visible change at the β -globin locus (Figure 5.3A-B). This observation was corroborated by a reduction in relative read counts between the two loci (Figure 5.3C). By RT-PCR, expression levels of the *Hba-x* and *Hba-a1/2* genes were also reduced as expected to ~50% (58% for both genes; Figure 5.3D). Together, these data demonstrate that mESCs hemizygous for the α -globin locus are viable and capable of normal erythroid differentiation *in vitro*. The single-copy α -globin locus becomes accessible and transcribed as normal. The resulting cell line is a stable platform with which it is possible to directly assess the effects of novel mutations of the α -globin super-enhancer *in vitro*.

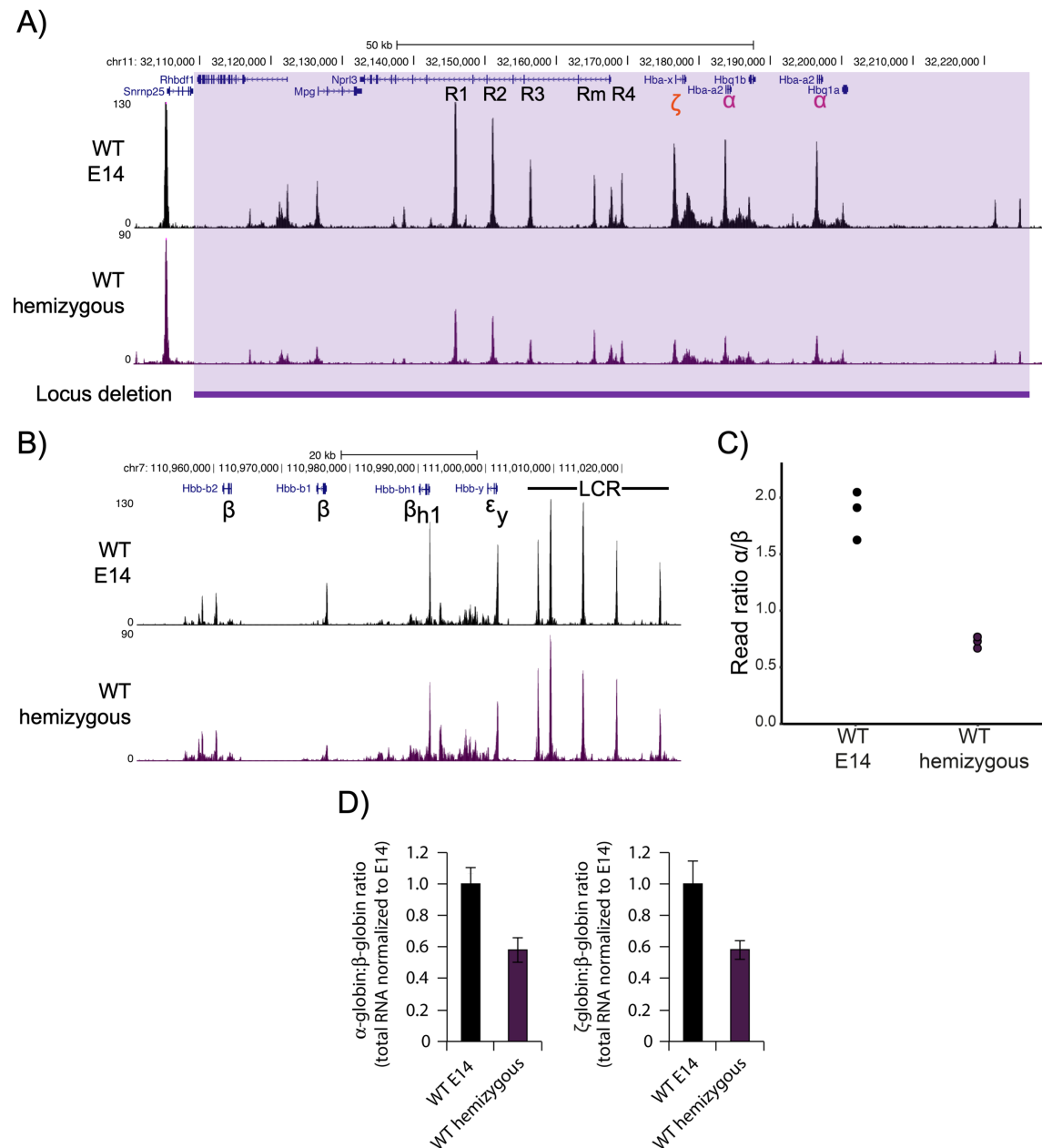


Figure 5.3: A hemizygous version of the mouse α -globin locus functions as normal. **A)** ATAC-seq tracks for the α -globin locus from WT E14 and WT hemizygous EB-derived erythrocytes. The hemizygous region shows peak heights reduced in height and is highlighted in purple. Tracks are normalised to RPKM and averaged over three independent differentiations (E14) or three independently targeted clones, differentiated in parallel (WT hemizygous). To negate the effect of inconsistent noise levels on normalised peak height, tracks are visualised at scales that result in equivalent peak heights at a control locus containing histone genes (Chr13:23,830,000-23,860,000). **B)** As part A) but for the β -globin locus. **C)** Quantitation of ATAC-seq reads within the hemizygous region divided by the reads from a region of the same length encompassing the β -globin locus (chr7:110,941,000-111,057,850; 116,851 bp). The relative proportion of reads is reduced to below half in the hemizygous population (38% by comparison of average counts). **D)** RT-PCR of mature α - and ζ -globin transcript levels compared to

output from the β -globin locus from WT E14 and WT hemizygous EB-derived erythrocytes. α -globin refers to the combined output from *Hba-a1* and *Hba-a2* genes; ζ -globin refers to the output from the *Hba-x* gene. β -globin output is calculated by taking the mean C_T values for all active genes at the locus: *Hbb-b1/2*, *Hbb-y* and *Hbb-bh1*. Data are normalised to E14. Results are the averages of three independent differentiations (E14) or three independently targeted clones, differentiated in parallel (WT hemizygous); error bars represent standard deviation of the mean.

Validating enhancer mutants in vitro

Before embarking on analysing new super-enhancer mutants in the hemizygous *in vitro* system, it was important to establish that phenotypes from existing *in vivo* models could be faithfully recreated in this way. Starting with the WT hemizygous cells described above, I created three genetic mutants previously studied *in vivo* by creating single- and double-enhancer deletions of the two strongest enhancers of the cluster: R1 and R2 (Figure 5.4A). The resulting hemizygous models (WT, Δ R1, Δ R2 and Δ R1R2) would test the fidelity of the *in vitro* system to *in vivo* phenotypes across a wide range of activity levels (9-100% native α -globin expression; Hay et al., 2016; Figure 3.1). Instead of targeting these models by Synthetic RMGR with a bespoke BAC for each design, I used CRISPR/Cas9 targeting as a more efficient way to perform these simple genome edits. I designed guide RNA recognition sites to remove the multispecies conserved sequence and footprint region for each enhancer (Figure 5.4B-C). I screened clones by PCR and Sanger sequencing over the predicted deletion junctions. Efficiency of the deletion procedure was high among cells selected for uptake of targeting plasmids: 13/60 clones for the R1 KO step (22%) and 23/96 clones for R2 KO (24%). I confirmed that selected clones were pure and did not carry random integrations of the deleted enhancers by performing PCRs internal to the deletion regions (data not shown).

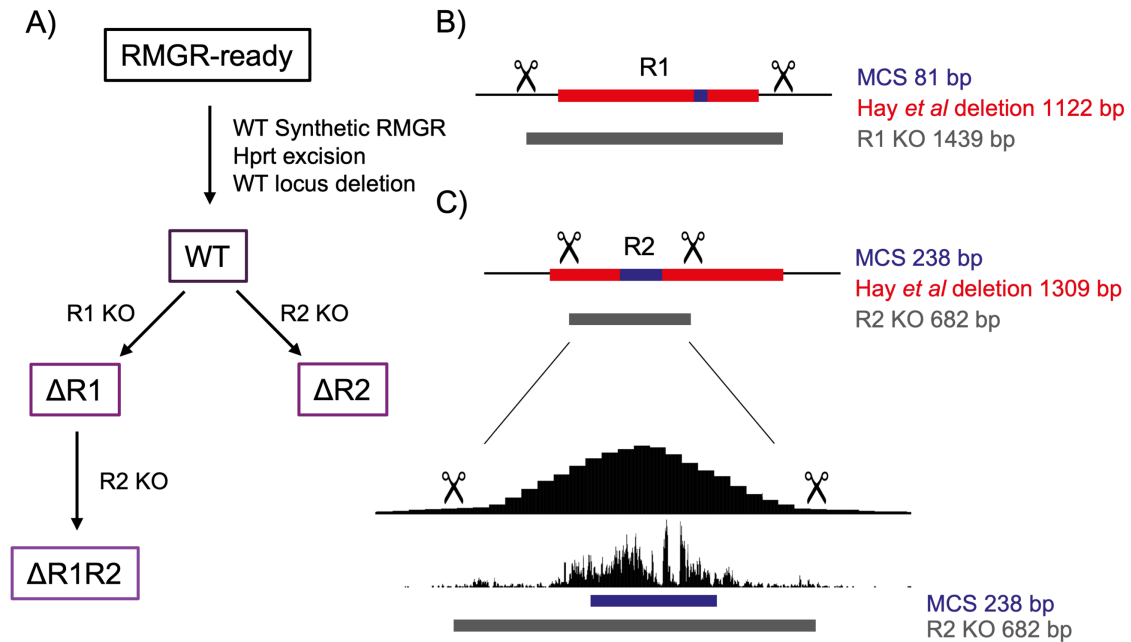


Figure 5.4: Design for the production of hemizygous mESCs for $\Delta R1$, $\Delta R2$ and $\Delta R1R2$ models of the α -globin super-enhancer.

A) Workflow for the generation of specified super-enhancer mutants. “WT” here represents WT hemizygous cells described previously, such that enhancer deletions (R1 KO / R2 KO) are targeted to the single remaining copy of the α -globin locus. B) Schematic for the R1 KO targeting step relative to the deletion used previously and detailed in Figure 3.3. C) Schematic for the R2 KO targeting step relative to the deletion used in Hay et al., 2016 and DNase-seq accessibility and digital footprinting data as shown for the other enhancer elements in Figure 3.3. Although smaller than the deletion published previously, the R2 KO deletion removes the entire footprint region and MCS. In parts B) and C) scissors indicate the positions of Cas9 target cut sites.

After selecting three independently targeted clones for each of the hemizygous $\Delta R1$, $\Delta R2$ and $\Delta R1R2$ mESC models, I isolated CD71⁺ erythroid cells from day 7 EBs and subjected them to ATAC-seq analysis (Figure 5.5A-B). At the α -globin locus, deletion models showed loss of ATAC-seq peaks over the deleted enhancers and reduced peak heights over the globin genes, as seen previously in primary erythroid cells (Hay et al., 2016). I observed no changes in element accessibility at the β -globin locus, suggesting that the effect on accessibility of the α - and ζ -globin gene bodies was a specific result of regulatory landscape disruption, rather than an artefact of impaired *in vitro* erythropoiesis.

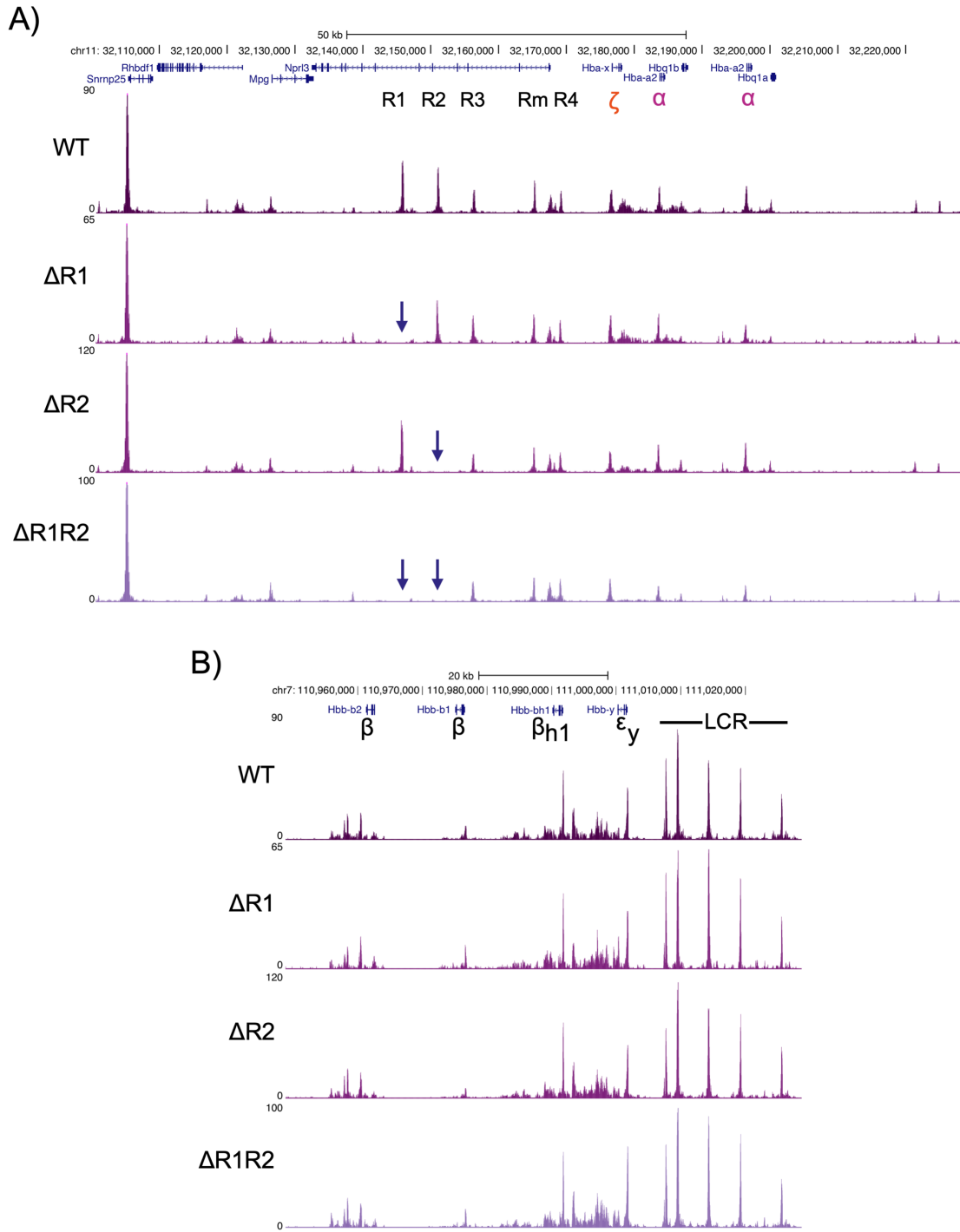


Figure 5.5: Chromatin accessibility results from hemizygous R1/R2 enhancer deletions *in vitro* mimic *in vivo* data.

A, B) ATAC-seq tracks of the α -globin locus (A) and β -globin locus (B) for WT, $\Delta R1$, $\Delta R2$ and $\Delta R1R2$ hemizygous CD71⁺ EB-derived erythrocytes. Super-enhancer/LCR elements and functional globin genes are annotated for each locus and, where applicable, deleted enhancers are highlighted with blue arrows. All active regions become accessible as expected, independently of other elements. Only peak heights over the α - and ζ -globin genes are affected by super-enhancer mutation. Tracks are normalised by RPKM and

averaged over three independently targeted clones for each mutation, differentiated in parallel. Tracks are scaled as described in Figure 5.3.

Expression of the α -globin genes in $\Delta R1$, $\Delta R2$ and $\Delta R1R2$ models has been assessed previously for primary cells of both fetal and embryonic lineages (Hay et al., 2016; Duantida Songdej, unpublished data). To validate the *in vitro* erythroid differentiation system and hemizygous targeting platform as faithfully recapitulating *in vivo* phenotypes, I performed RT-PCR for mature globin transcript levels in bulk EB-derived erythrocytes for the WT, $\Delta R1$, $\Delta R2$ and $\Delta R1R2$ hemizygous models. Comparison with data from primary cells showed that the trend was uniform across all cell types. Variation in the quantified expression levels could be attributed to the fact that EB-derived cells were assessed using a different method (RT-PCR compared to Nanostring technology) and reflect an earlier stage in erythroid maturation than the other tissue types, as discussed in Chapter 4.

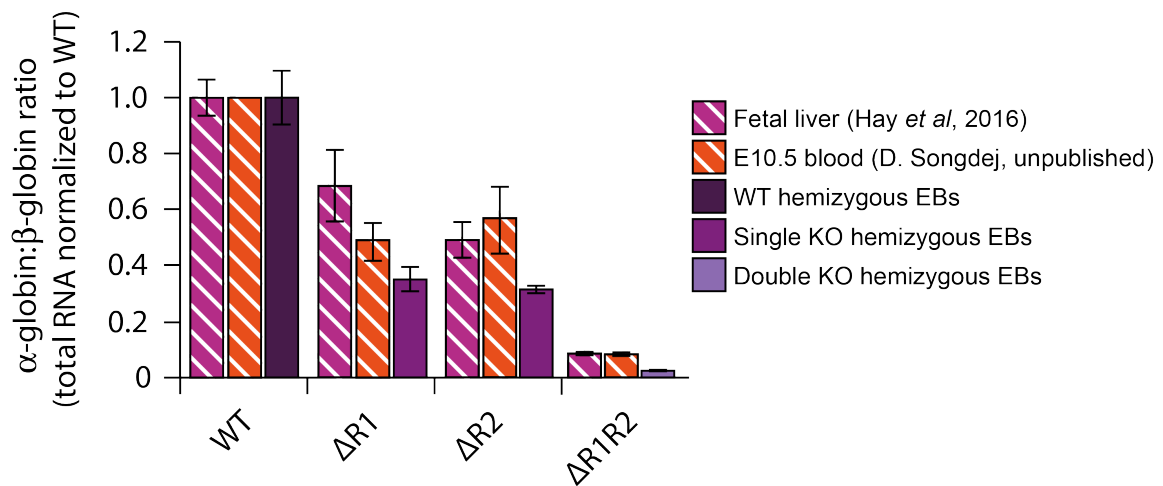


Figure 5.6: Gene expression results from hemizygous R1/R2 enhancer deletions *in vitro* mimic *in vivo* data.

Mature transcript levels are shown for α -globin normalised to the active β -like globins for each tissue: Nanostring data from fetal liver-derived cells are normalised to *Hbb-b1/2*; Nanostring data from E10.5 blood cells are normalised to the total output of *Hbb-b1/2*, *Hbb-y* and *Hbb-bh1*; RT-PCR data from hemizygous EBs are normalised to mean C_T values for *Hbb-b1/2*, *Hbb-y* and *Hbb-bh1*. *In vivo* data are compared to WT mice. *In vitro* hemizygous EB mutants are compared to WT hemizygous EBs. Bars represent mean

expression values from at least three biological replicates or individually targeted clones normalised to WT. Error bars represent standard deviation of the mean. By one-way ANOVA ($p < 1e-6$) with Tukey post-hoc test, all models showed different transcript levels in the *in vitro* system compared to WT ($p < 1e-5$) and the $\Delta R1R2$ model was different from both $\Delta R1$ and $\Delta R2$ models ($p < 1e-3$).

Together, I concluded that the consistency of chromatin and expression data between primary and *in vitro*-derived cells in all the examined models rendered the hemizygous *in vitro* system a reliable tool to further dissect the α -globin super-enhancer.

A novel set of enhancer combination models

The unexpected results from the R2-only model in Chapter 3 highlighted the value in rebuilding a super-enhancer from the bottom-up, to complement results from the top-down approach. The hemizygous *in vitro* system lends itself perfectly to the analysis of new single- or double-enhancer models, as genome engineering steps can easily be targeted to the single remaining copy of the α -globin locus and phenotypes can be assessed without the confounding issues of organism lethality. Starting with Synthetic RMGR of the R2-only BAC into hemizygous RMGR-ready mESCs, I recreated the R2-only locus in this cell line and, in addition, created a deletion of the complete super-enhancer (ΔSE) through subsequent deletion of the R2 enhancer by CRISPR/Cas9. In parallel, I created a double-enhancer R1R2-only model by insertion of R1 into its native position in the R2-only model by CRISPR/Cas9 with HDR and a similar model in which R1 is replaced with the R4 enhancer sequence (the R4R2-only model). By the same method, I inserted R1 into its native position in the ΔSE model to create an R1-only model. A schematic for the designed targeting steps is shown in Figure 5.7A-C. I screened mESC clones for RMGR and KO steps

by HAT selection and PCR as described above and in Chapter 3. I screened clones for R1 and R4 insertion steps by PCR across integration junctions.

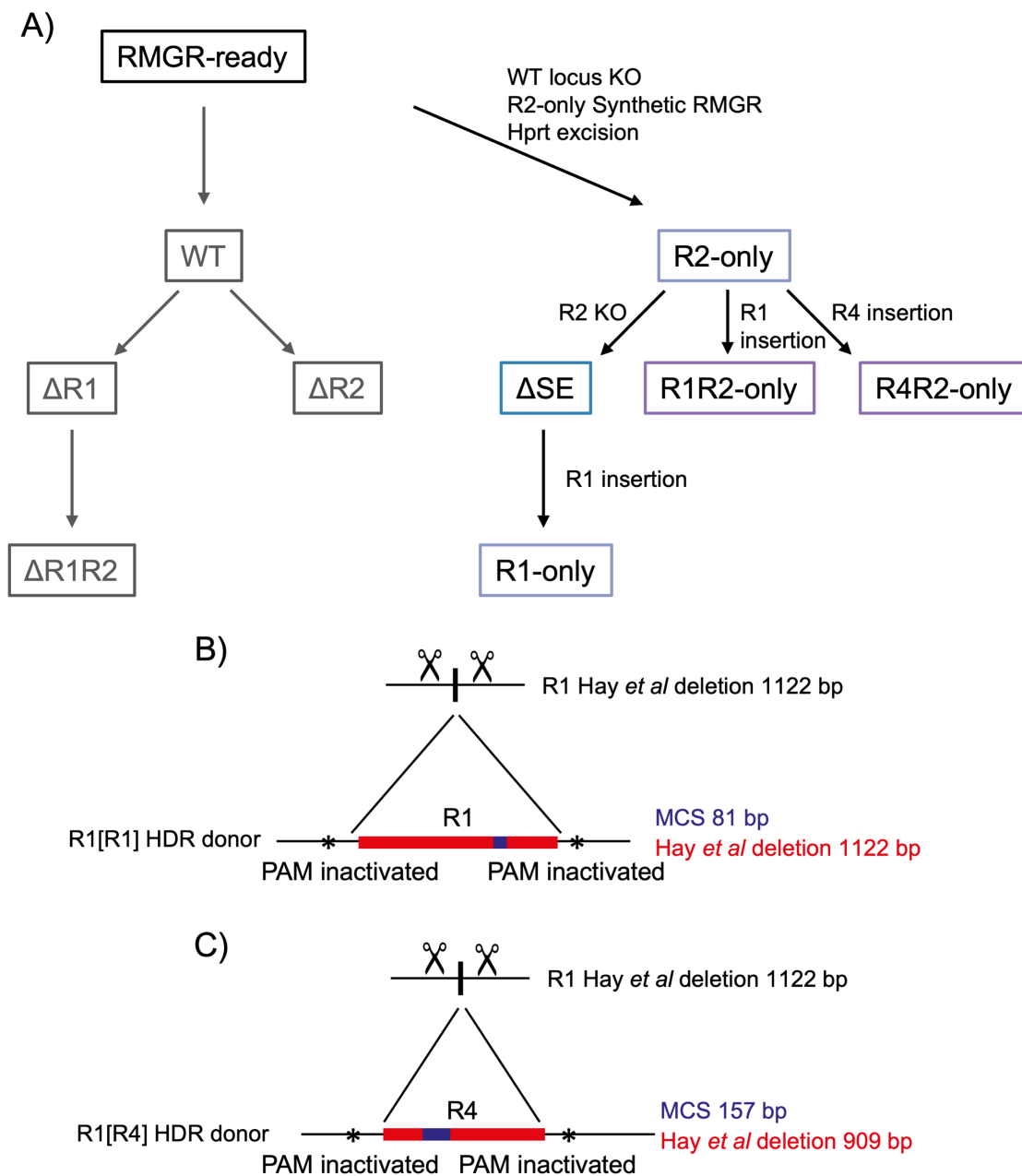


Figure 5.7: Design for the production of hemizygous mESCs for R1/R2/R4 insertion models of the α -globin super-enhancer.

A) Workflow for the generation of specified super-enhancer mutants. WT locus deletion was performed prior to Synthetic RMGR targeting of the R2-only model, such that all models from this point onwards are hemizygous: additional targeting steps were performed on the single remaining copy of the α -globin locus. **B)** Schematic for the R1 insertion targeting step relative to the R1 deletion present in R2-only mESCs and detailed in Figure 3.3. R1 insertion was targeted through CRISPR/Cas9 combined with HDR. Cas9 target cut sites are indicated with scissors and guide RNA inactivating mutations are

highlighted on homology arms of the donor vector (protospacer adjacent motif (PAM) inactivated). C) As part B) for the R4 insertion targeting step. The complete R4 footprint region contained within the region previously deleted (Figure 3.3) was inserted into the position of the deleted R1 enhancer.

I subjected three independently targeted clones for each of the hemizygous Δ SE, R1-only, R2-only, R1R2-only and R4R2-only mESC models to *in vitro* differentiation and CD71+ erythroid cell isolation. All models underwent normal erythroid differentiation despite the presumed imbalance in α : β chain ratios, including the most extreme Δ SE model (Figure 5.8A-C). I first assessed clones by ATAC-seq for the globin loci (Figure 5.9; Figure 5.10). At the α -globin locus, all models showed loss of peaks over the deleted enhancers and reduced peak heights over the globin genes, as seen previously for the R2-only model in primary erythroid cells (Figure 3.8). For the Δ SE model, this amounted to a complete loss of signal across the gene bodies and extended 3' regions beyond the globin genes (Figure 5.9). I observed no differences in element accessibility at the β -globin locus (Figure 5.10).

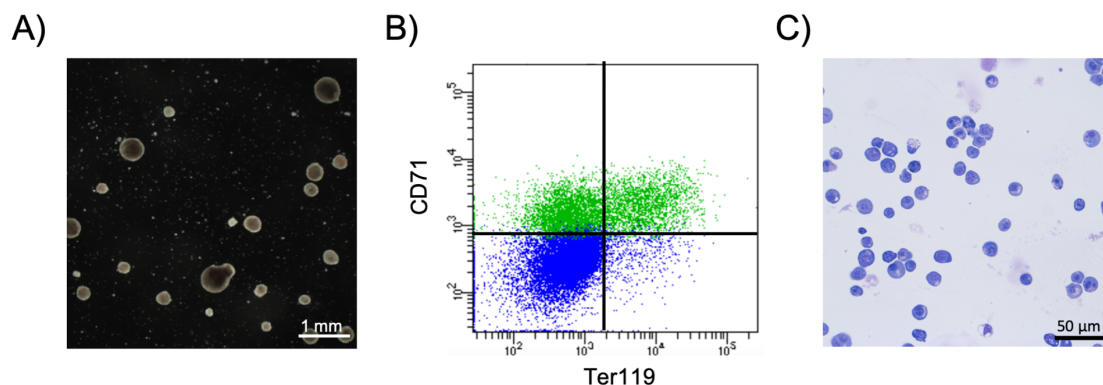


Figure 5.8: mESCs hemizygous for the Δ SE model of the α -globin locus undergo normal *in vitro* erythroid differentiation.

A) Brightfield image of day 7 EBs from Δ SE hemizygous mESCs. B) Immunostaining for erythroid CD71 and Ter119 cell surface markers from disaggregated EBs shown in part A). CD71+ erythroid cells are shown in green; non-erythroid cells are shown in blue. C) MGG-stained cell preparation of the green CD71+ erythroid population sorted in part B).

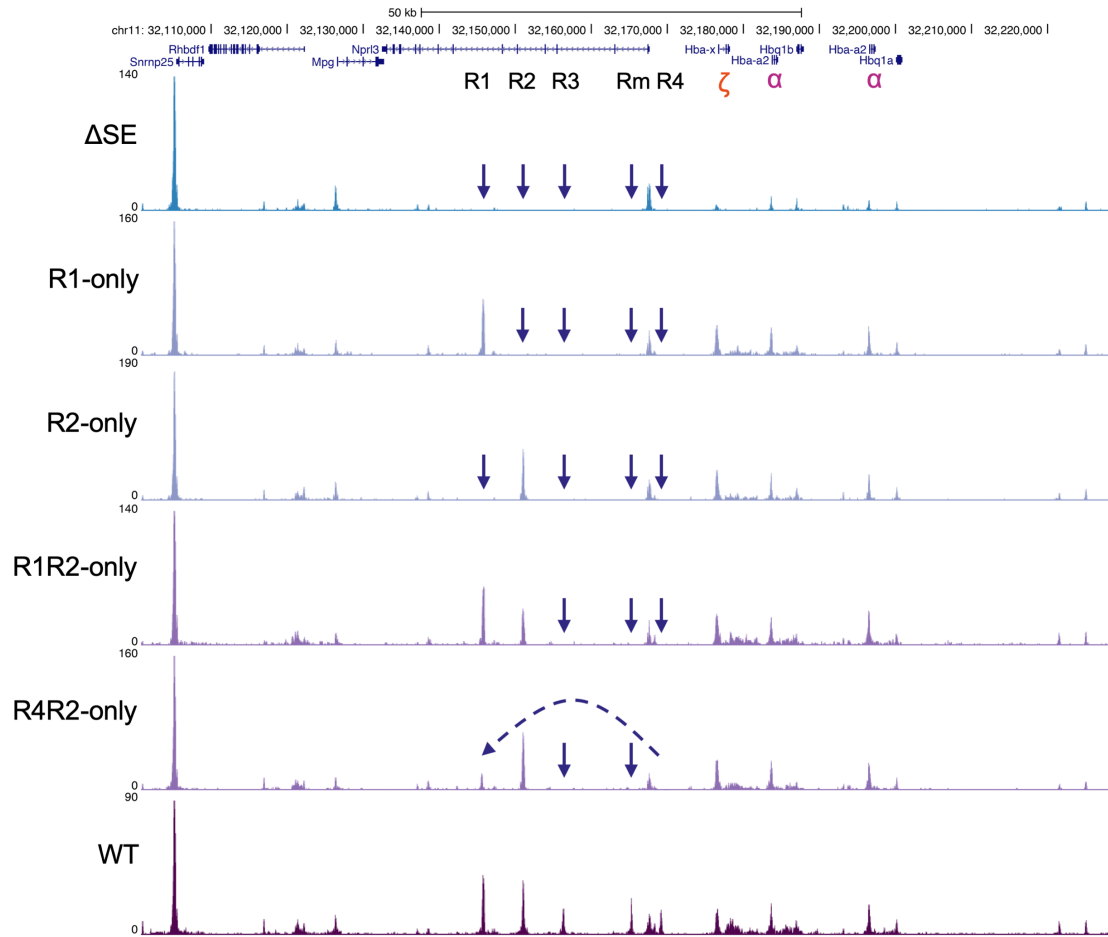


Figure 5.9: Chromatin accessibility at the α -globin locus for hemizygous R1/R2/R4 enhancer insertion models differentiated *in vitro*.

ATAC-seq tracks of the α -globin locus for hemizygous Δ SE, R1-only, R2-only, R1R2-only and R4R2-only models of the α -globin locus from CD71+ EB-derived erythrocytes, compared to the track from WT hemizygous erythrocytes (Figure 5.3). Where applicable, deleted enhancers are highlighted with blue arrows. The R4 enhancer integration into the R1 position is indicated with a blue dashed arrow. As in Figure 5.5, all active regions become accessible as expected, independently of other elements. Only peak height over the α - and ζ -globin genes is affected by super-enhancer mutation. Tracks are normalised by RPKM and averaged over three independently targeted clones for each mutation, differentiated in parallel. Tracks are scaled as described in Figure 5.3.

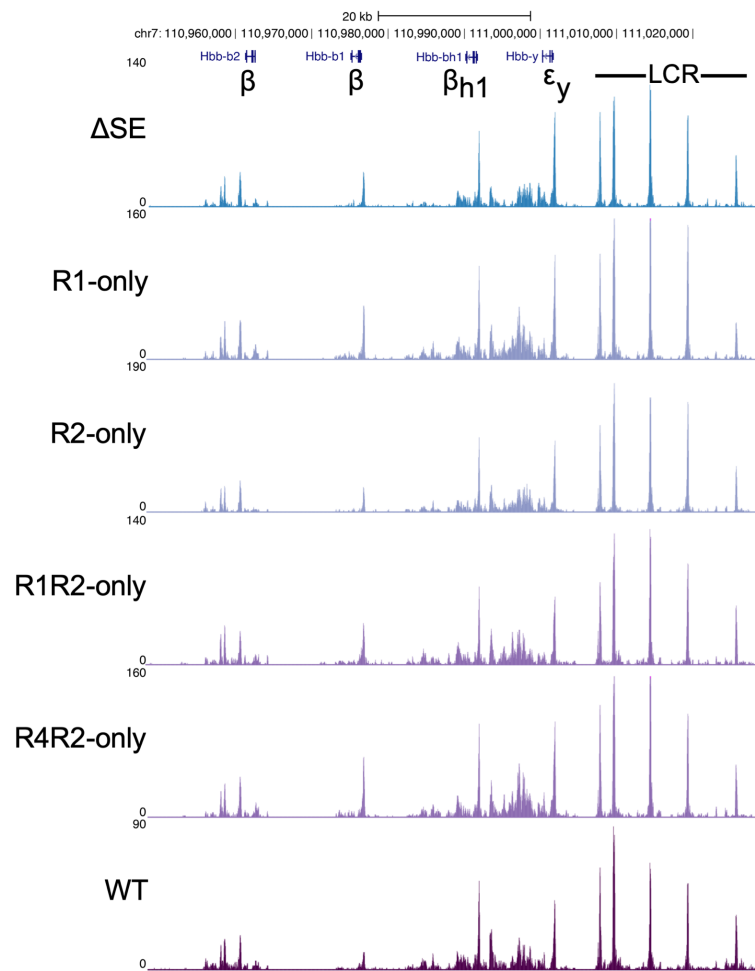


Figure 5.10: Chromatin accessibility at the β -globin locus for hemizygous R1/R2/R4 enhancer insertion models differentiated *in vitro*.

As Figure 5.9 for the β -globin locus. No changes are seen in β -globin accessibility in mutants for the α -globin super-enhancer.

To test the effect of complete super-enhancer deletion and subsequent enhancer insertions on expression of the α -globin genes, I performed RT-PCR on CD71+ erythroid cells from day 7 EBs of the above hemizygous mESC clones, in parallel with WT hemizygous clones described previously (Figure 5.11). Deletion of all elements of the α -globin super-enhancer abolished transcription of the globin genes to baseline levels (0.1% of WT). In isolation as single enhancer elements, R1 and R2 enhancers show similar levels of gene activation (mean values 12% and 10% of WT expression, respectively). Together, their activity appears to be enhanced beyond purely additive contributions (mean 43% of WT expression). Addition of

the weaker R4 enhancer to the R2-only model, in the position of R1, gives no significant increase in activity compared to the R2-only model (mean 16% of WT expression).

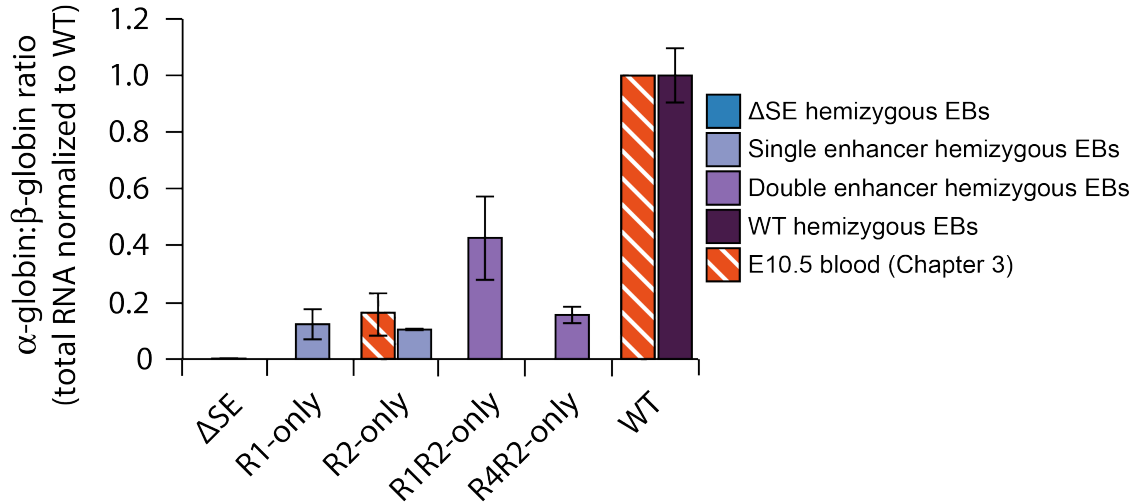


Figure 5.11: Gene expression results from hemizygous R1/R2/R4 enhancer insertion models differentiated *in vitro*.

Mature α -globin transcript levels are shown for RT-PCR data from hemizygous EBs and E10.5 blood, where available, normalised to mean C_T values for *Hbb-b1/2*, *Hbb-y* and *Hbb-bh1*. Bars represent mean expression values from at least three biological replicates or clones normalised to WT. Error bars represent standard deviation of the mean. Results from the WT and R2-only E10.5 blood samples are shown an alternative representation of data in Figure 3.9. Results show relative increases in α -globin expression upon the addition of sequential enhancer elements to the Δ SE baseline. By one-way ANOVA ($p < 1e-8$) with Tukey post-hoc test, all models showed different transcript levels compared to the WT integrated sequence ($p < 1e-5$). The R1R2-only model was significantly different from both the R1-only and R2-only models ($p < 0.002$), whereas the R4R2-only model was not.

Together, these results represent the first attempt to begin re-constructing the function of a super-enhancer from the bottom-up within its native chromosomal context. Results further question the concept of super-enhancer additivity and lay out a system to further probe the function of the regulatory landscape of the mouse α -globin locus.

5.3 Discussion

A platform for the dissection of a regulatory domain

In this chapter, I have used a mESC cell line that permits genome engineering at a range of resolutions; at one end of the spectrum, I used Synthetic RMGR to produce custom locus designs at the scale of up to 100 kb to tackle questions of large-scale element positioning and context. At the other end, I used CRISPR/Cas9 to efficiently tackle questions regarding the importance of individual elements or their sequences. RT-PCR from *in vitro*-differentiated erythroid populations (described in Chapter 4) was sufficiently sensitive to ascertain the resulting effects of genetic manipulation on gene expression. I have shown that results generated using this platform are reliable and faithful to *in vivo* phenotypes for a range of mutations ($\Delta R1$, $\Delta R2$, $\Delta R1R2$, R2-only) and that it is feasible to test multiple models in parallel in a single experiment. Thus, I have developed a system to fully analyse the regulatory landscape of the well-characterised mouse α -globin locus in its native chromosomal context and in an appropriate cell type, without confounding issues of animal viability, experimental scale or position-effect variegation.

This platform naturally lends itself to expansion into higher-throughput analyses using the miniaturised differentiation protocol outlined in Figure 4.8. In Chapter 6, I describe one such application in the development of a fluorescence-based screening approach to study gene regulation. Combined with automation steps to alleviate the burden of manual processing steps, this platform has the potential to query genetic mutants at scales and resolutions that provide a more complete picture of relationship between structure and function.

The importance of context in addressing enhancer function

Results from this chapter build upon the finding in Chapter 3 that the nominal activity of a single enhancer element differs based on how it is tested; the R2-only mouse model, expected to recover 50% of the α -globin transcriptional activation, provided only 15% of WT activity when present in isolation. With the novel super-enhancer mutants generated here, it is informative to examine the discrepancy in findings from two parallel strategies; building a functional enhancer unit from a null background (Δ SE) or from a non-functional background containing R3, Rm and R4 enhancers (R3/Rm/R4; Figure 5.12).

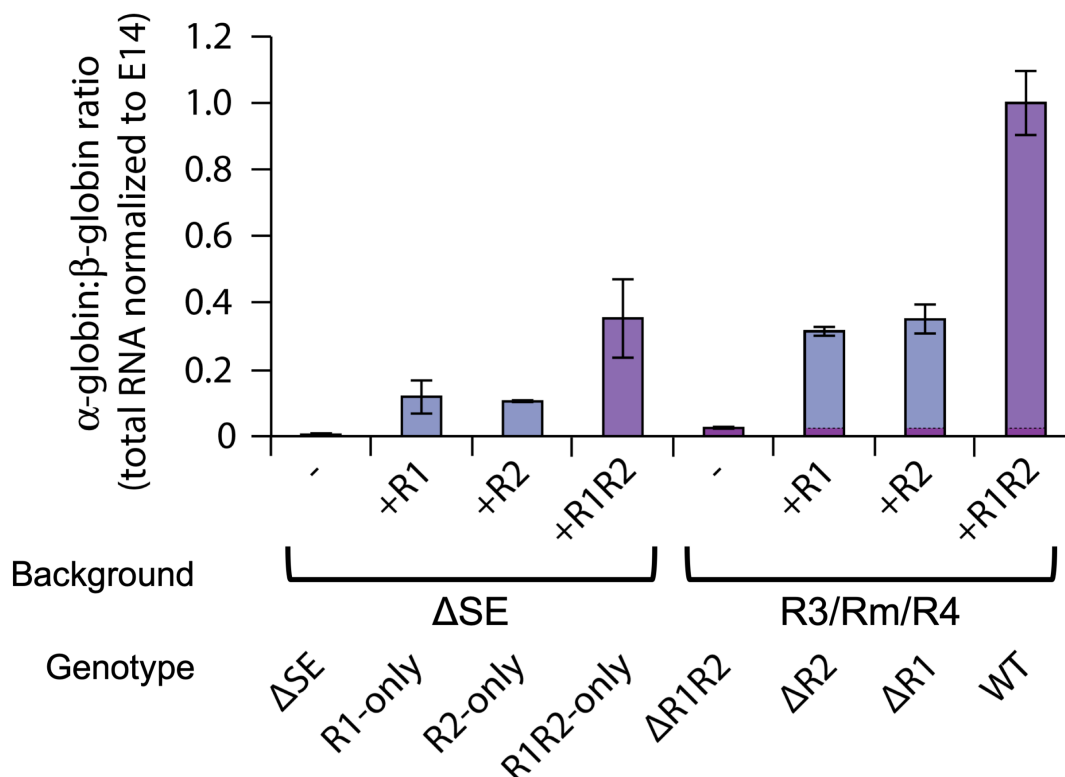


Figure 5.12: The R1/R2 enhancers upregulate α -globin transcription to different extents depending on the integrity of the super-enhancer cluster.

RT-PCR data for mature α -globin transcripts normalised to β -globin are combined from Figure 5.6 and Figure 5.11. The α -locus backgrounds for enhancer insertions are designated with brackets and additional enhancer insertions are labelled for each bar. Genotype names used for these enhancer combinations throughout this chapter are shown below for clarity. In both Δ SE and R3/Rm/R4 backgrounds, insertion of the R1 or R2

enhancer resulted in upregulated α -globin gene expression and insertion of both elements had a greater effect than either element in isolation. However, the magnitudes of relative upregulation were larger for insertions into the R3/Rm/R4 background. Bars are coloured by the number of inserted enhancers above background. For the R3/Rm/R4 background, a small dark coloured bar is overlaid with complete bars to highlight the maximal contribution of the R3/Rm/R4 enhancers in an additive model of super-enhancer enhancer function.

I first established an *in vitro* null background mESC model; I demonstrated that complete removal of the endogenous α -globin super-enhancer elements reduces transcription to baseline levels in *in vitro*-isolated erythroid cells (0.1% WT; Figure 5.11). This was in-line with, if not more extreme than, results from *in vivo* deletion of the LCR at the β -globin locus (Bender et al., 2000a; 1-8% WT expression was maintained in primary cells by gel-based reverse transcription-PCR). The effect of compound deletion of R1 and R2 enhancers (Δ R1R2), the main contributors to the alpha globin expression (~30 and 50% respectively *in vivo*; Hay et al., 2016), was recapitulated *in vitro* and established the non-functional background R3/Rm/R4 model. Insertion of the R1 or R2 enhancers into the null (Δ SE) or non-functional (R3/Rm/R4) backgrounds resulted in very different outputs. On a null background, R1 or R2 produced relatively small contributions to α -globin gene expression (~10%) as seen in the R2-only mouse model. Yet, the contribution of R1 or R2 on a non-functional background which contained previously characterized but non-functional putative regulatory elements (R3/Rm/R4) was enhanced to ~40%. The presence of both R1 and R2 on a null background did not recapitulate the addition of R1 and R2 to the R3/Rm/R4 background (i.e. the WT super-enhancer configuration). It appeared for all added enhancer combinations (R1/R2/R1R2) that the gene activation potential of the strong R1/R2 elements was amplified by the presence of the R3/Rm/R4 background.

As we continue to investigate the emergent properties of a super-enhancer, it is important to note that the increase in α -globin expression upon integration of a single enhancer element into a null background, while smaller than perhaps anticipated, nevertheless represents an upregulation of two orders of magnitude compared to baseline levels (~10% WT expression in R1-only and R2-only models compared to 0.1% Δ SE baseline expression). Thus, just as the R1 and R2 enhancers were able to upregulate gene expression in previous transgenic enhancer assays and embryo gene reporter systems (Bouhassira et al., 1997; Gourdon et al., 1995; Hay et al., 2016), they are also able to do so within their native chromosomal contexts, albeit to a relatively low level.

The tools I have developed allowed the creation of multiple genetic models that provide a direct comparison of top-down and bottom-up approaches in the genetic dissection of a super-enhancer. It becomes clear that the presence of what we considered as non-functional 3' enhancers results in a greater ability for R1 and R2 to upregulate α -globin transcription, in absolute terms (Figure 5.12). How does the presence of two apparently non-functional elements (R3 and Rm) and one weakly functional element (R4) account for more than a two-fold change in α -globin output from the two strongest elements of the cluster? What are these elements providing that is missing in the R1R2-only model? To begin to answer these questions, we must consider the proposed mechanisms for how the elements of the α -globin super-enhancer might function together. Below, I outline how existing data for this locus align with possible mechanisms for element co-operation in gene regulation.

Element co-operation through molecular environment

In recent years, data from 3C-based methods and high-resolution cell imaging technology have highlighted the importance of spatial context in gene regulation (de Wit and de Laat, 2012; Dekker et al., 2017; Oudelaar et al., 2017; Vermunt et al., 2019). Through the formation of self-interacting domains, it is possible for multiple *cis*-regulatory elements to contribute simultaneously to the local molecular environment of a transcribing gene.

In the case of the R3/Rm/R4 enhancers, mechanisms for the propagation of transcriptional enhancer activity of R1/R2 through local molecular changes could be manifested in a number of ways. By Tri-C, a one-to-many 3C technique, the native R2 enhancer is known to demonstrate multi-way interactions with the other enhancer elements and globin gene promoters specifically in mouse erythroid cells (Oudelaar et al., 2018). Therefore, binding of TFs to R3/Rm/R4 may support element interactions within a transcription hub by adding to the valency of possible element-element interactions (Hnisz et al., 2017). Within the constrained nuclear space these elements could also increase the availability of TFs or cofactors that may be otherwise limiting, or induce changes in local chromatin modifications (Liu and Tjian, 2018; Whyte et al., 2013). Alternatively, binding events at these elements may confer enhancer-promoter specificity to the R1/R2 enhancers, for example through the compatibility of the associated protein-protein interactions as exemplified by the differential activity of a range of cofactors at promoter sequences (Haberle et al., 2019).

Element co-operation through co-linearity

The importance of positional context in the co-ordination of gene expression is well-documented in the literature, such as at the Hox cluster (Lemons and McGinnis, 2006) and globin loci (Philipsen and Hardison, 2018). Transgenic experiments indicate that both gene sequence and positioning of genes within clusters have roles to play in these spatially-defined patterns of expression (Peterson and Stamatoyannopoulos, 1993). At the mouse α -globin locus, orientation-specific activity observed upon inversion of the enhancer cluster suggests that regulatory elements can also be functionally dependent on the co-linearity of elements within a domain (Mira Kassouf, unpublished data).

One way in which the requirement for a specified syntax of regulatory elements could be manifested functionally is through a tracking mechanism, for example through the process of loop extrusion by the cohesin moiety or through transcription by Pol II. In the case of the mouse α -globin locus, although there is no existing evidence of contiguous transcription between the enhancers and promoters, both cohesin and Pol II have been shown to accumulate at the α -globin super-enhancer in an erythroid-specific manner (Hanssen et al., 2017). An additional model for co-linear element function involves a binding cascade of activating protein factors, anchored at enhancers and directed to target promoters (Furlong and Levine, 2018; Matharu and Ahituv, 2015). In this model, it is easy to surmise how the placement of three additional TF-bound elements between the active R1 and R2 enhancers and the α -globin gene promoters might help to spatially direct the accumulation of active factors towards the target genes.

In determining which models best describe the mode of super-enhancer function at the α -globin locus, we are limited by technology. Existing methods fail to capture the sensitivity and real-time dynamics needed to provide a causative link between structure and function at a locus-specific level. However, further interrogation of the molecular characteristics of the super-enhancer may provide insight into the nature of element co-operation. In parallel, functional genetics experiments can be used to probe whether the proposed spatial relationships between elements are critical for correct gene regulation and what sequence features are necessary to create those elements.

The significance of enhancer sequence for function

In this chapter, I demonstrated that, in the R1R2-only and R4R2-only models, the individual strength of the sequence in the R1 position dictated the level of transcriptional upregulation compared to the R2-only model. This indicates that the addition of the R1 enhancer to the R2-only model did not purely act as an additional accessible unit and implies that the binding of prescribed sequence-specific TFs is required for optimal super-enhancer function. In Chapter 6, I explore this idea further with an in-depth sequence dissection of the conserved R2 enhancer sequence.

Investigating potential *trans*-acting candidates for the differential functions of R4 and R1, there are several TFs known to bind differentially, and likely many more besides. Comparing each of the mouse α -globin super-enhancer elements for binding of key erythroid TFs points at R1 and R2 being the most bound elements and the only elements bound by the Nf-e2 factor (Figure 5.13; Hay et al., 2016). Indeed, by binding of these TFs, it appears unsurprising that the R3/Rm/R4 enhancers confer minimal gene activation potential to the

α -globin genes and perhaps surprising that they contribute any additional activity to the R1R2-only model. How the relative binding affinities and combinations of TFs at the individual enhancer elements contribute to their function has yet to be resolved.

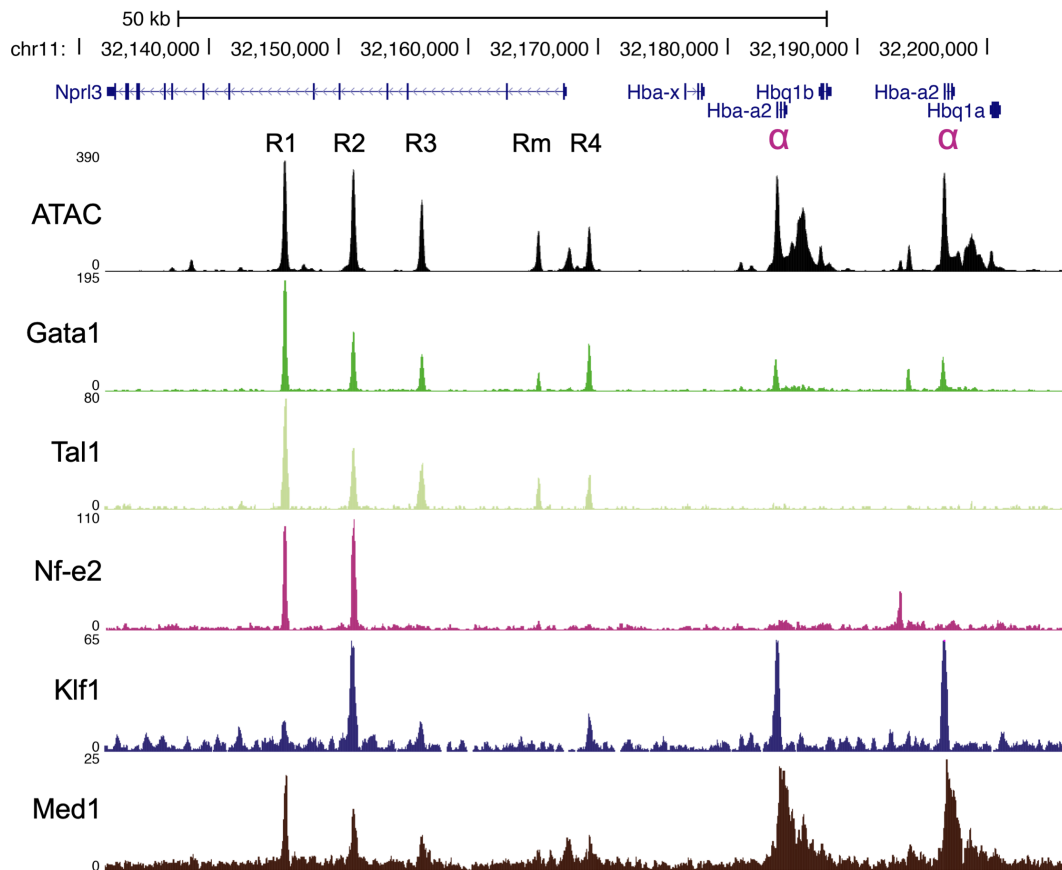


Figure 5.13: Differential binding of erythroid-specific TFs at elements of the mouse α -globin super-enhancer.

ATAC-seq and ChIP-seq tracks from primary mouse erythrocytes. Tracks are normalised by RPKM and averaged between replicates, where appropriate (Gata1 n=2, Tal1 n=3, Nf-e2 n= 2, Klf1 n=1, Med1 n=1).

Irrespective of the underlying mechanisms, from the results in this chapter I conclude that the apparent activity of an individual enhancer element is dependent on its context within a regulatory landscape, even when its immediate surrounding sequences and relative positioning to target genes is unaffected. The most prominent questions to approach in the light of these data are the re-examination of the function of the three 3' enhancer elements (R3, Rm and R4), thought to have minimal or no effect on transcriptional activation, and the

discrimination of sequence and context in assigning enhancer activity. The question of enhancer element function will continue to be studied in subsequent projects in the Higgs group. As for the dissection of enhancers as binding platforms for TFs, in Chapter 6 I begin tackling the question of sequence constraint on enhancer function by probing the structure of the 238 bp R2 MCS at high resolution.

Data availability

Spleen ATAC-seq: Hanssen et al., 2017

Primary erythroid Gata1 ChIP-seq: Martin Larke, unpublished

Primary erythroid Nf-e2 ChIP-seq: Hughes et al., 2014

Primary erythroid Tal1 ChIP-seq: Pimkin et al., 2014

Primary erythroid Klf1 ChIP-seq: Tallack et al., 2010

Acknowledgements

I performed and analysed all experiments described in this chapter. Philip Hublitz (Genome Engineering Facility, Weatherall Institute of Molecular Medicine, University of Oxford) assisted with design and production of CRISPR/Cas9 targeting vectors. Kevin Clark and Sally Clark (Flow Cytometry Facility, Weatherall Institute of Molecular Medicine, University of Oxford) performed FACS sorts.

Chapter 6 – Dissection of the R2 multispecies conserved sequence

6.1 Introduction

The R2 multispecies conserved sequence

In previous chapters of this thesis, I have tested elements of the mouse α -globin super-enhancer as complete functional units. Through biochemical assays and the study of genetic models that disrupt the locus at the level of complete enhancer units, the R2 enhancer (as a ~1 kb accessible region of DNA) is established as a major contributor to α -globin gene regulation. In this chapter, I test the importance of sequence within the context of the R2 enhancer position by focusing in on the functional core of the R2 sequence and ask what roles TF number, type and grammar play in creating an active enhancer.

The sequence of the R2 element of the α -globin locus has been highly conserved through evolution (Hughes et al., 2005; Philipsen and Hardison, 2018). Alignment of sequences from 18 vertebrate species defines a 238 bp MCS and clearly distinguishes short conserved sections from intervening, divergent DNA within this region (Figure 6.1), consistent with the structure of a “landing pad” for TF binding at conserved sequences (Gotea et al., 2010). Indeed, several of the conserved regions of the R2 MCS can be identified as known motifs for erythroid TFs (Gata1, Tal1, Klf1 and Nf-e2; Figure 6.2A-B), each of which is required for normal erythropoiesis *in vivo* (Fujiwara et al., 1996; Perkins et al., 1995; Shivdasani and Orkin, 1995; Shivdasani et al., 1995). Mutations of these factors in human patients are

associated with a variety of haematological disorders (Ferreira et al., 2005; Gasiorek and Blank, 2015; Perkins et al., 2016). Binding of the Gata1, Tal1, Klf1 and Nf-e2 TFs to R2 in mature primary mouse erythrocytes has been shown previously by ChIP-seq (Figure 5.13). Further support for binding at these motifs comes from DNase I footprinting data, where marked footprints can be seen for the central Gata1 and Nf-e2 sequences (conserved motifs 3-5), in agreement with footprints previously determined by gel electrophoresis (Gourdon et al., 1995). By gel electrophoresis, additional footprints are observed at the positions of conserved Gata1 motifs 1 and 6.

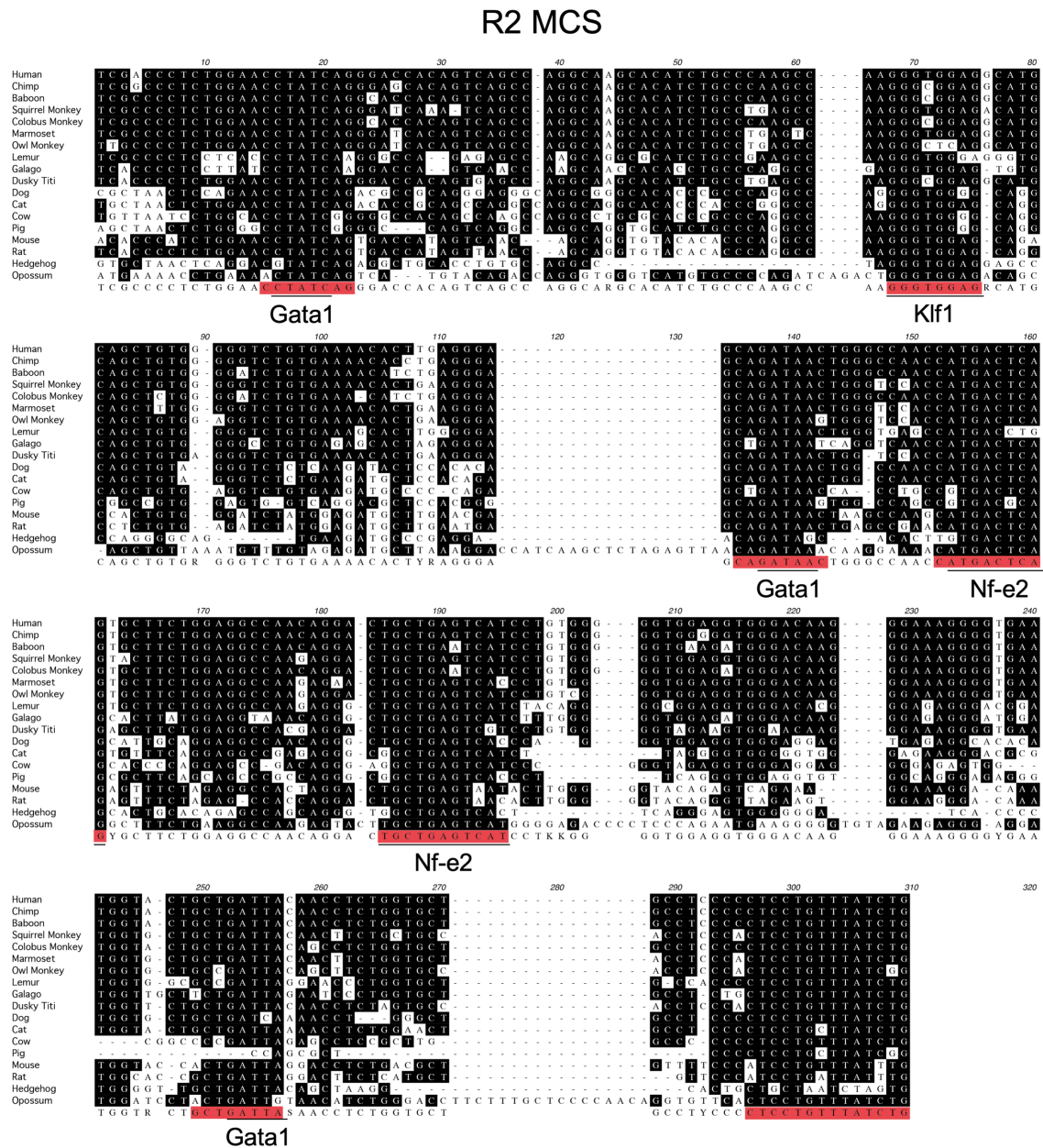


Figure 6.1: Alignment of the R2 MCS in 18 vertebrate species.

Sequences for each species are shown aligned to a consensus sequence (bottom). Bases matching the consensus are highlighted in black; mismatching bases are highlighted in white. Strongly conserved regions are highlighted in red in the consensus sequence row. Sequences in these red highlighted regions can be identified as erythroid TF motifs (underlined and labelled), indicating that they are likely to represent sequences necessary for enhancer function. Figure adapted from Hughes et al., 2005.

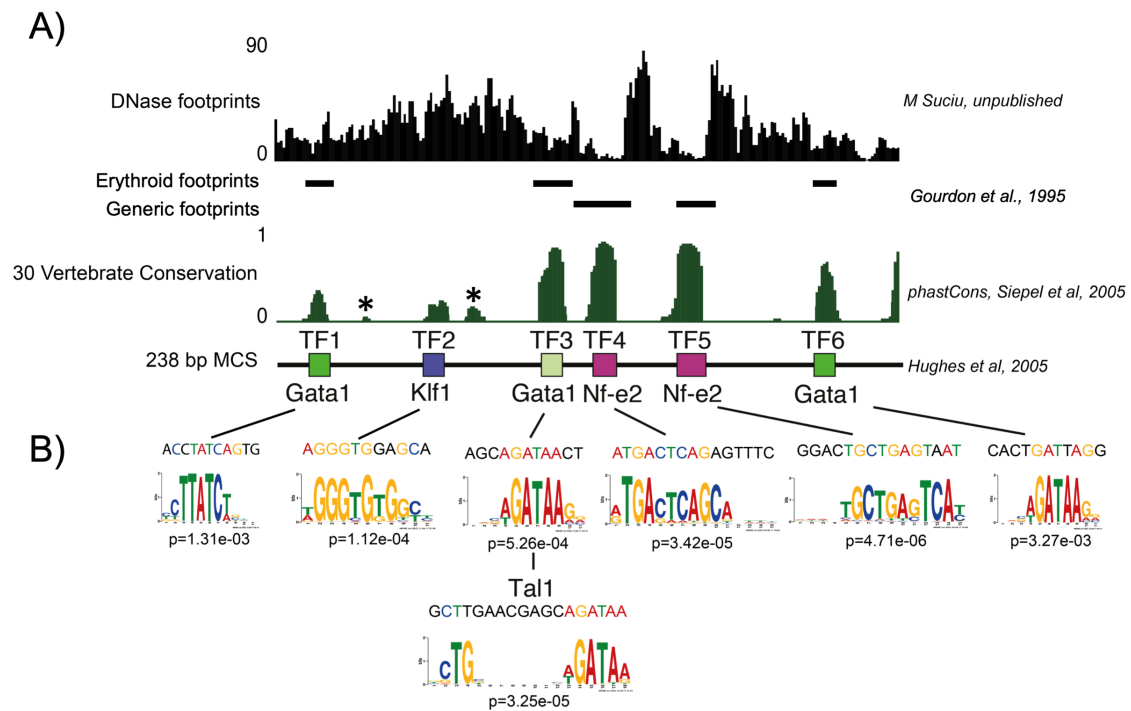


Figure 6.2: *In silico* dissection of the R2 MCS.

A) Alignment of existing data to predict functional regions of the R2 MCS. Top: DNase footprinting data from mouse erythrocytes from E10.5 embryos. Middle 1: Footprints previously determined by gel electrophoresis using erythroid (mouse erythroleukaemia) and non-erythroid (HeLa) cell nuclear extracts. Middle 2: PhastCons conservation scores across the MCS (Siepel et al., 2005). Areas of enriched conservation across the 30 included vertebrate species overlap with those identified in a specific study of the α -globin locus (Figure 6.1; Hughes et al., 2005) with the exception of two additional peaks, highlighted with asterisks (*). These regions do not correspond to motifs for Gata1, Tal1, Nf-e2 or Klf1. Bottom: schematic illustration of the conserved regions highlighted in Figure 6.1. Each predicted motif is labelled by number (TF1-6) and represented by a box, coloured according to the matched TF motif. **B)** Detailed alignments of conserved regions to predicted erythroid-specific consensus motifs. Position weight matrices were generated by analysis of ChIP-seq data from mature primary mouse erythrocytes using the Multiple Em for Motif Elicitation (MEME) tool (Figure 5.13; Bailey and Elkan, 1994). Motif matches were assessed using the Find Individual Motif Occurrences (FIMO) tool (Grant et al., 2011). P-values represent the match probability of tested sequences to the provided position weight matrices.

Addressing the nature, prevalence and relative positioning of TF sequences within the R2 region would shed light on the rules underlying the creation of a strong enhancer. Comparing between elements of the α -globin super-enhancer, is the activity of R2 driven by its position or its sequence? Within the sequence, is it the number of recruited TFs? The precise

coordination of protein-protein interactions? Or perhaps the presence of a single key motif? To deconstruct these questions experimentally, a strategy for *in situ* genetic screening is required.

Existing methods for regulatory sequence dissection

The genetic dissection of regulatory sequences is not a new concept. Experimental designs have evolved with the availability of genome engineering and sequencing technologies from methods testing small numbers of sequences for a particular active element, to large-scale screens testing thousands to millions of sequences in parallel. Two main categories exist for large-scale screening designs: firstly, those that either test entire classes of elements or genomes for a particular function; secondly, those that dissect a particular region of interest at high resolution.

To test entire genomes or element subsets for a particular function in an unbiased manner, sequences are typically tested for enhancer or promoter activity by transient transfection or integration into a stable landing pad, adjacent to a region to be transcribed for functional readout (Levo and Segal, 2014). Functional output is typically measured by quantifying enrichment of sequences that are able to upregulate self-transcription or transcription of an associated barcode. Such assays have been performed *in vitro* and *in vivo*, across a range of organisms (Arnold et al., 2013; Arnold et al., 2016; Dickel et al., 2014; Gisselbrecht et al., 2013; Grossman et al., 2017; Melnikov et al., 2012; Patwardhan et al., 2012; Smith et al., 2013). While these screens identify sequences with the potential to upregulate transcription, they do not take into account the ability of sequences to function within the context of a structured chromatin environment.

To dissect selected sequences at high resolution, endogenous elements can be functionally tested for necessity by screening involving programmable nucleases, such as CRISPR/Cas9. Experiments of this type are generally designed to test for guide RNA enrichment upon pooled mutation of a large number of elements genome-wide (Montalbano et al., 2017) or dissection of a specific sequence by saturation mutagenesis (Canver et al., 2015; Lopes et al., 2016; Vierstra et al., 2015). Outcomes are selected by phenotypes such as induced cell lethality. The additional development of tools such as CRISPR inactivation (CRISPRi) and CRISPR activation (CRISPRa) make guide RNA libraries versatile tools to study a range of endogenous elements in an unbiased manner (Montalbano et al., 2017). However, the inability to control desired mutations of interest is a serious limitation with this system. Mutations are limited to regions with appropriate recognition sequences and resulting indels are typically small, but can frequently be unusably large in certain cell types, including mESCs (Owens et al., 2019).

To dissect the functional sequence of the R2 enhancer *in situ*, two unconventional features were critical for the screen design. Firstly, I needed to assess expression of the distal α -globin genes upon mutation of the R2 sequence, i.e. to physically separate the sites of sequence targeting and functional readout without the use of a viability phenotype. Similar designs in the literature have used fluorescent reporter constructs to assess distal sequence function (Kinney et al., 2010; Kosuri et al., 2013; Sharon et al., 2012). Secondly, I needed to replace the endogenous copy of R2 with designed variant sequences of interest, as seamlessly as possible. CRISPR/Cas9-directed HDR has previously been used for this purpose (Findlay et al., 2014).

Combining a method for endogenous integration of designed R2 variant sequences with a fluorescence readout from the native target α -globin gene would allow me to fully dissect the sequence constraints of the R2 MCS. In this chapter, I describe the development of a screening method with the required attributes, and its application to unpicking the contributions of the conserved TF motifs to R2 enhancer function.

6.2 Results

Development of a fluorescence-based screen to monitor α -globin expression

The R2 MCS is 238 bp in length and contains six conserved regions likely to represent key TF binding motifs. When considering how to genetically dissect such a sequence, it is clear that testing a handful of sequences in a conventional, clone-based experiment would be insufficient to unpick the sequence rules governing enhancer function. To simply mutate each binding site individually in triplicate would already require the production and analysis of 18 individual clones, making targeting, screening, erythroid cell isolation from EBs and RT-PCR technically demanding and time-consuming. Expanding experiments to a saturation mutagenesis level or to probe questions of sequence combinations, permutations and grammar would rapidly push this number into the hundreds and make experiments unfeasible with current methods. I therefore embarked on developing an *in vitro*, fluorescence-based screen in which the effect of R2 mutation on α -globin gene expression could be quantified in a high-throughput manner. The screen design was based around the use of an existing fluorescently tagged mESC line, mentioned in Chapter 4. This line carries a YFP tag on one copy of the *Hba-a1* gene and shows erythroid-specific fluorescence (Figure 4.2C).

A schematic of the basic screen design is shown in Figure 6.3. This highlights the major points for screen optimisation when designing the experiment. Briefly, I first had to determine whether measurement of YFP fluorescence levels from the existing YFP-tagged mESC line could serve as a reliable proxy for α -globin transcription levels. Secondly, I needed to design and test a system for R2 sequence replacement. Finally, I needed to

determine how to record YFP measurements from individual variant sequences, be that as a clonal approach or pooled variant strategy.

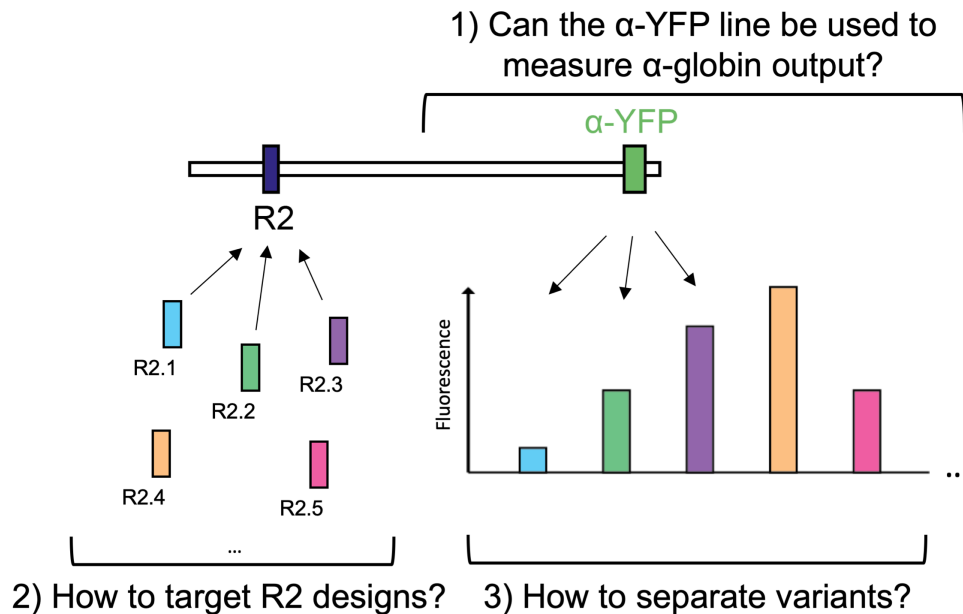


Figure 6.3: Overview of a screen designed to dissect the R2 MCS.

Schematic depicts the targeting of designed variants into the endogenous R2 position, followed by measurement of the effect on α -globin gene expression through the fluorescence level of a YFP tag on the α -globin gene. The major questions to answer in developing this as a viable screening system are highlighted and discussed in the main text.

1. Can the α -YFP mESC line be used to measure α -globin output?

I started with the YFP-tagged α -globin mESC line mentioned in Chapter 4 (Figure 4.2C), hereafter referred to as the α -YFP mESC line. Similar cell models have been developed previously to study activation of genes at the β -globin locus (McColl et al., 2014; Papadopoulos et al., 2012; Peterson et al., 2014). To test whether recordings of fluorescence levels from these cells, differentiated into day 7 EBs, could be used as a proxy for the measurement of α -globin gene transcription, I generated triplicate clones of two additional mESC models that would provide a range of α -globin expression levels: Δ R1 α -YFP and Δ R1R2 α -YFP. These models were targeted as described previously (Figure 5.4). The

resulting lines harbour the YFP tag on one copy of the locus and are expected to produce ~65% and ~95% reductions in α -globin transcript levels compared to WT, respectively (Figure 5.6).

By immunophenotype, day 7 EBs from the described panel of α -YFP mESC lines underwent normal erythroid differentiation (Figure 6.4A). By flow cytometry of the CD71+ erythroid population from each sample, YFP fluorescence distributions followed a pattern reminiscent of the expected α -globin expression levels, with the intensity progressively reduced with the loss of R1 and R2 enhancers (Figure 6.4B). Fluorescence histograms show considerable overlap between clones, but they can be clearly distinguished by their median measurements.

To test how well differences in these median fluorescence measurements mimic changes in transcript levels, I performed RT-PCR for transcripts of α -globin (*Hba-a1/2* genes) and the YFP tag construct (mVenus) in isolated CD71+ erythroid populations from all YFP-tagged samples. As depicted in Figure 4.2C, design of the tag is such that a single transcript should be produced from the *Hba-a1* gene through to the end of the mVenus coding region; the resulting polypeptide is self-cleaved by a 2A peptide sequence between the functional units (Kim et al., 2011). Thus, the expression of the YFP tag should be directly related to expression of the α -globin gene. As expected, relative transcript levels were indistinguishable for α -globin and YFP coding regions and followed the expected pattern for R1/R2 enhancer deletion clones (Figure 6.4C). In matching YFP fluorescence output with gene transcript levels, fluorescence medians showed a good correlation with RT-PCR data across the full range of expression ($R^2=0.89$ by linear regression; Figure 6.4D).

Therefore, I concluded that the median fluorescence level from an endogenous α -YFP tag constituted a reliable proxy for α -globin gene transcription.

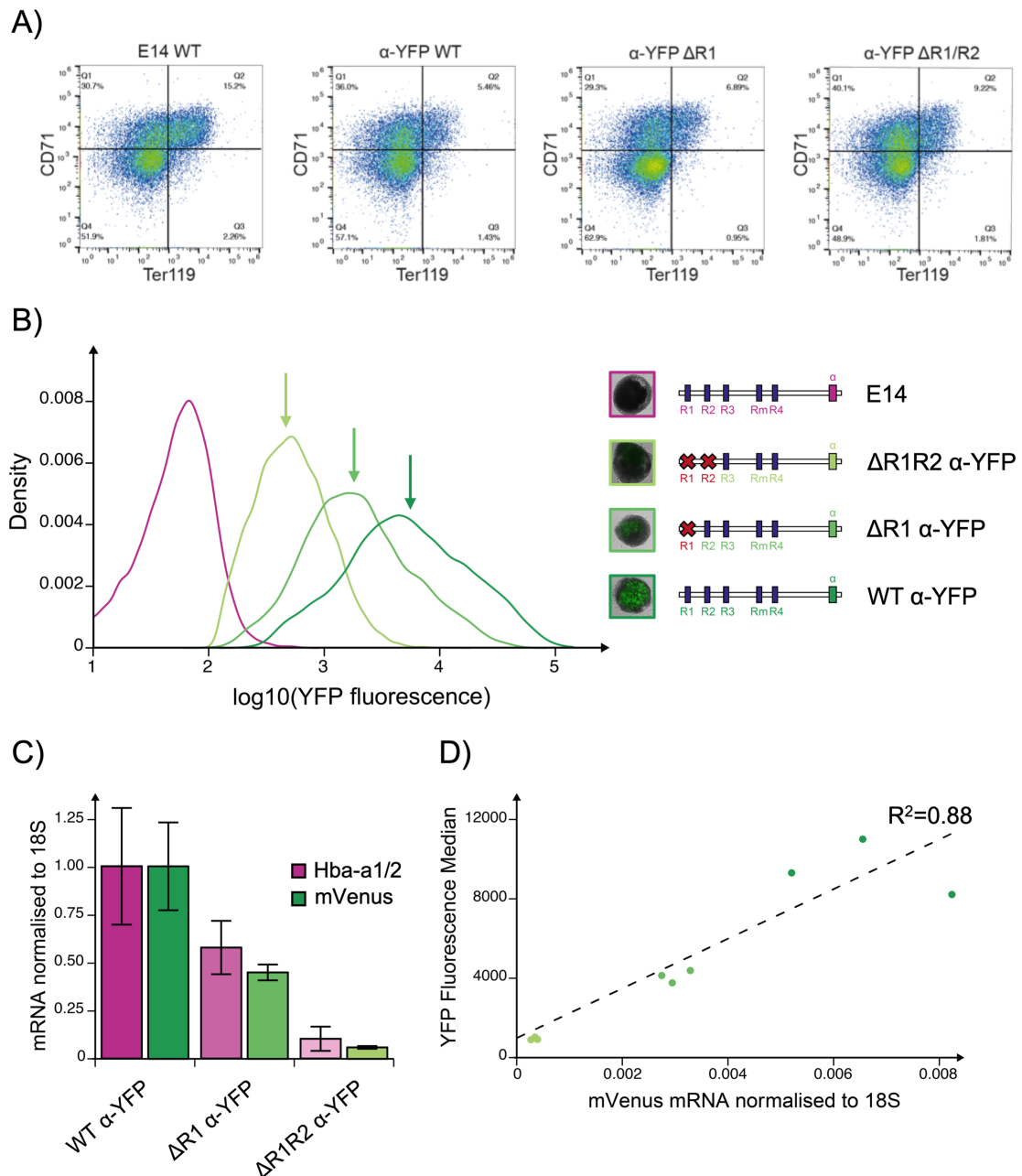


Figure 6.4: Fluorescence of a YFP tag as a proxy for α -globin gene expression.

A) Flow cytometry plots for CD71 and Ter119 erythroid markers for day 7 EBs generated from a panel of WT and YFP-tagged mESC lines: untagged E14 WT, tagged WT α -YFP, tagged $\Delta R1$ α -YFP and tagged $\Delta R1R2$ α -YFP. Shown are representative plots from a single clone of each genotype. All models show normal *in vitro* erythropoiesis by immunophenotype. **B)** Left: flow cytometry histograms for erythroid-specific YFP fluorescence from the models shown in part A). The YFP tag is present on a single copy of the endogenous *Hba-a1* gene; the sequence composition is shown in Figure 4.2C. Distributions represent combined YFP fluorescence histograms for three independent

clones for each genotype, differentiated in parallel as bulk-plated EBs. Median fluorescence levels for each population are highlighted with appropriately coloured arrows. Right: YFP/brightfield overlay images of single day 7 EBs of each genotype are shown with colour-coded schematics of each YFP-tagged locus. Increasing YFP signal is observed from $\Delta R1R2$ α -YFP to WT α -YFP clones. **C)** RT-PCR quantitation of transcripts for the α -globin genes (*Hba-a1/2*) and the YFP tag construct (mVenus) from YFP-tagged clones compared to 18S and normalised to WT by the $\Delta\Delta C_T$ method. Relative transcript levels were consistent between the tagged *Hba* gene and the mVenus tag construct and in line with previous results of R1/R2 enhancer deletions (Figure 5.6). Bars represent mean data from three independently targeted clones for enhancer deletions or three clones each derived from a single cell of the original WT α -YFP line, all differentiated in parallel; error bars represent standard deviation of the mean. **D)** Scatter plot of median YFP fluorescence readings against mVenus mRNA levels from samples analysed in part B). Each coloured dot represents the result from a single clone of the corresponding genotype. There is a strong positive correlation between datasets, as indicated by the linear regression line of best fit.

To focus the R2 screen directly on the sequence of the R2 enhancer and avoid potential compensating interference from the other major active element of the cluster, I decided to use the $\Delta R1$ α -YFP mESCs described above for the R2 sequence screen design. In addition, as for the super-enhancer dissection in Chapter 5, I wanted the capability to direct R2 mutations to the YFP-tagged copy of the α -globin locus, without interference from the remaining WT chromosome. Therefore, using the strategy outlined in Figure 5.1, I deleted the WT α -globin locus from the $\Delta R1$ α -YFP mESC line. In this case, I screened for retention of the YFP-tagged locus and deletion of the untagged locus by PCR across the insertion point of the YFP tag. A summary of these required genome edits is shown in Figure 6.5A. Resulting hemizygous $\Delta R1$ α -YFP mESCs, which I will now refer to as R2-screening mESCs, can be differentiated into EBs and show the same erythroid specificity for YFP fluorescence signal as seen for the original α -YFP line (Figure 6.5B-D; Figure 4.2). By flow cytometry, the YFP fluorescence distribution is indistinguishable from that of the $\Delta R1$ α -YFP line from which it is derived (Figure 6.5E).

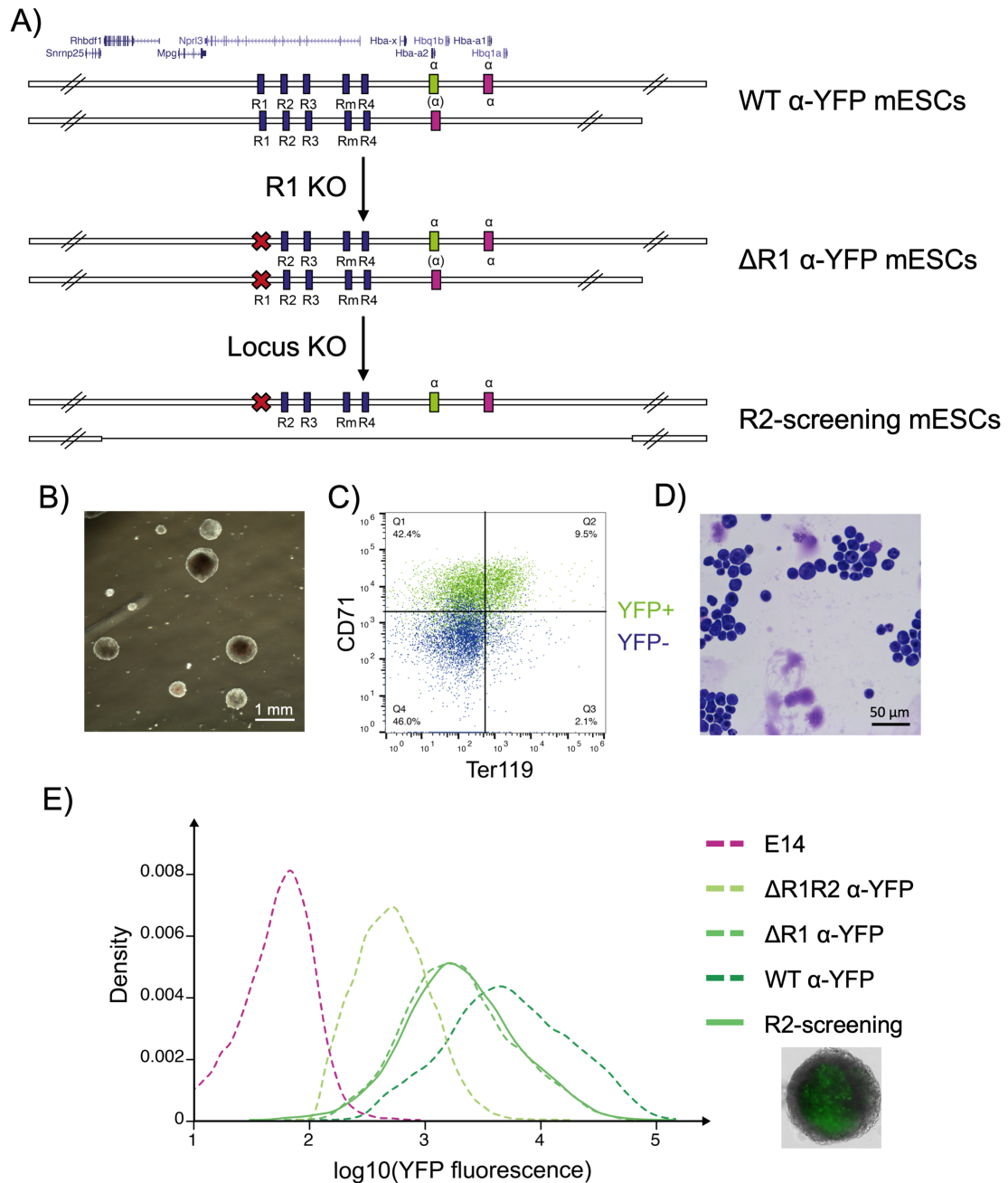


Figure 6.5: Generation and *in vitro* differentiation of the R2-screening mESC line.

A) Schematic for genome engineering steps applied to the α -YFP mESC line to generate R2-screening mESCs. Both copies of chromosome 11 are shown for clarity. The (α) symbol on the untagged chromosome represents the result of a deletion event between the duplicated *Hba* genes via Cas9-mediated cutting of the homologous sequence during the YFP targeting step. The *Hba-a2* gene on the YFP-tagged locus is intact with no indels. **B)** Brightfield microscopy image of whole day 7 EBs of the R2-screening line. **C)** Flow cytometry of YFP+ and YFP- populations from R2-screening day 7 EBs for CD71 and Ter119 erythroid markers. The YFP+ population is erythroid-specific, overlapping with the CD71 erythroid marker. **D)** MGG-stained cell preparation of CD71+ cells from day 7 EBs of the R2-screening model showing early stage erythroid cells. **E)** Flow cytometry

histogram plot for erythroid-specific YFP fluorescence from day 7 EBs, reproduced from Figure 6.4B with dashed lines and overlaid with a solid green histogram from the R2-screening model. The YFP distribution for the R2-screening model represents combined histograms from three independent bulk-plated EB differentiations of the same clone. As expected, the R2-screening histogram overlaps with that of the $\Delta R1$ α -YFP line from which it was derived. A YFP/brightfield overlay image of a single EB expressing YFP is shown below its label in the graph legend.

Finally, I validated EB-derived erythrocytes from the R2-screening model for their chromatin environment at the globin loci by ATAC-seq (Figure 6.6A-B). Tracks showed no abnormal peaks, including when data were mapped directly to the mVenus coding region. Relative peak heights were comparable to those seen for the hemizygous $\Delta R1$ model (Figure 5.5A). As shown in Chapter 5 (Figure 5.3C), the read density over the α -globin locus compared to the β -globin locus was reduced by roughly half in the R2-screening line due to hemizyosity (57% reduction; Figure 6.6C).

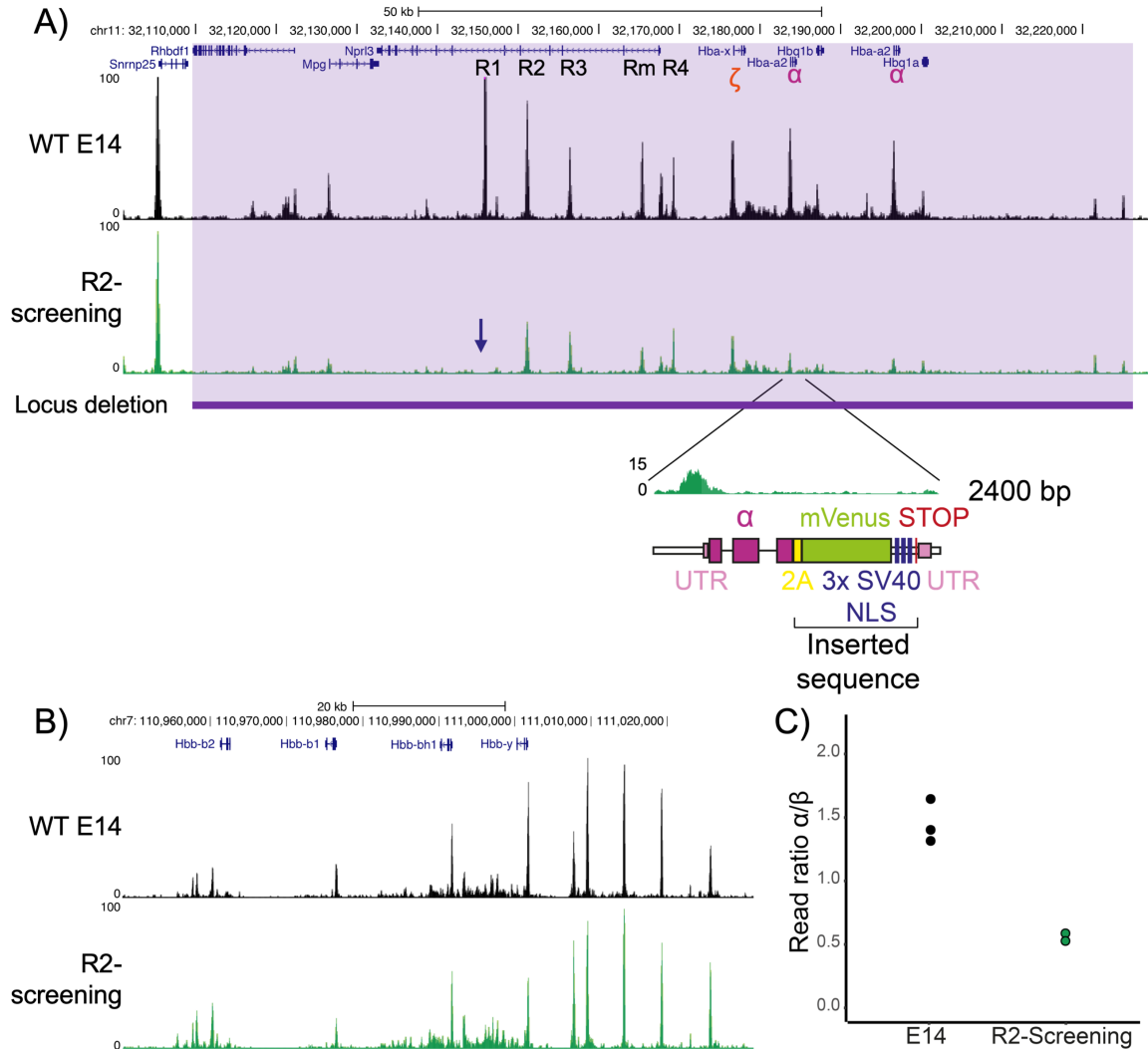


Figure 6.6: The R2-screening model has normal chromatin at the globin loci.

A) Top: ATAC-seq tracks of the α -globin locus from EB-derived erythrocytes generated in 96-well plates from WT E14 and R2-screening mESCs, aligned to the mm9 reference genome. Tracks are normalised by RPKM and averaged for 2 (R2-screening) or 3 replicates (WT E14) from individual differentiations. The R1 enhancer deletion in the R2-screening line is highlighted with an arrow. Bottom right: ATAC-seq data aligned to the YFP-tagged *Hba-a1* gene region. The peak corresponds to the *Hba-a1* promoter peak seen in the complete track; the schematic below represents a to-scale annotation of the downstream tag sequence. This demonstrates that no additional peaks are created over the mVenus sequence. **B)** As a control for the R2-screening mESC system, ATAC-seq tracks are shown as part A) for the β -globin locus. **C)** Quantitation of ATAC-seq reads within the hemizygous mm9-mappable region of the α -globin locus with R1 deleted divided by the reads from a region of the same length encompassing the β -globin locus (chr7:110,941,000-111,056,731; 115,731 bp). The relative proportion of reads is reduced to below half in the hemizygous population (43% by comparison of average counts).

2. How to replace endogenous R2 with designed variants?

In choosing a method to target the endogenous R2 sequence with rationally designed sequence variants, there were two main options: CRISPR/Cas9-directed HDR or integration of a landing pad for recombinase-mediated sequence exchange. The landing pad method, although likely to be high in efficiency, would require additional cell targeting to integrate lox sites (or similar) and may invalidate the system's future use with the Synthetic RMGR system due to the fact that the most popular, incompatible loxP, lox511 and frt recognition sites are already used in the RMGR system (Branda and Dymecki, 2004; Glaser et al., 2005). Instead, I took results from previous targeting of mESCs in the Higgs group as an indication that modifying the R2 sequence by HDR would be sufficiently efficient for sequence screening purposes (10-50% of clones homozygously targeted for HDR, depending on selection conditions; Rosa Stolper and Mira Kassouf, unpublished data).

To boost the selectable efficiency of R2 replacement, I designed a CRISPR/Cas9-directed HDR strategy that involved integration of an excisable positive/negative antibiotic resistance cassette for puromycin and thymidine kinase. A schematic overview of the targeting process is shown in Figure 6.7. To test the HDR methodology, I generated and transfected WT and null iterations of the exchangeable sequence, containing the intact R2 MCS or a complete deletion of this 238 bp (Δ MCS). I screened resulting clones for the presence of an integrated HDR donor by PCR of the single remaining copy of R2 over the targeted homology arms and digestion of an inserted restriction site (ScaI). By ScaI digestion, 16/96 clones (16.7%) of clones were positive for an HDR event. Clones of R2-screening mESCs that have been targeted for a particular variant sequence will hereafter be referred to as R2[Variant] for clarity.

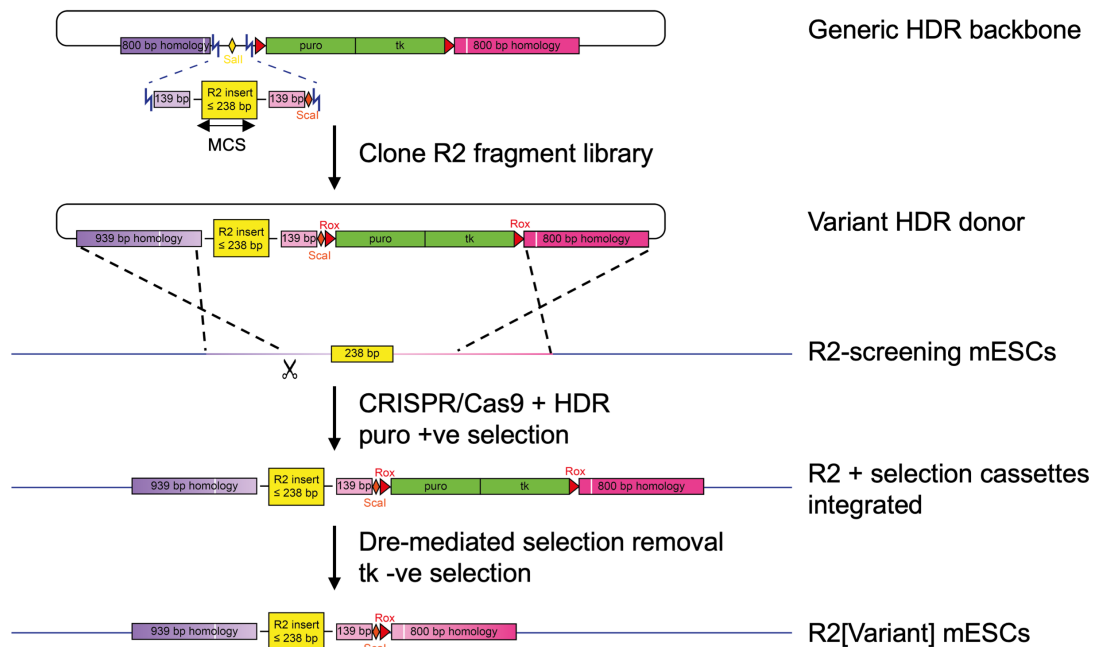


Figure 6.7: Scheme for R2 variant replacement of the endogenous R2 enhancer by Cas9-directed HDR.

See Methods for a complete description of the targeting process. Briefly, this involves cloning of variant sequences of the native R2 MCS along with flanking sequences into a generic custom HDR backbone, transfection into R2-screening mESCs and positive selection for resulting integrations, transfection of Dre recombinase and negative selection for resulting excision of the resistance cassettes. The result is an integrated complete sequence of custom design with minimal sequence scarring. Clones generated in this way are labelled R2[Variant] to indicate that they are derived from the R2-screening line and have been targeted for R2 replacement. Key sequences are labelled directly except for the BsaI restriction sites used for cloning (blue staggered lines), sites of guide RNA target sequence inactivation (white vertical lines within HDR donor homology arms) and the designed Cas9 cut site (black scissors image).

Finally, I needed to validate that the MCS could be seamlessly replaced by HDR to give the full expected dynamic range of YFP fluorescence levels. I therefore tested WT and Δ MCS HDR-exchanged clones for α -globin locus integrity and gene activation. I confirmed by ATAC-seq that integration of the WT R2 sequence (with the accompanying ScaI site, Rox recombination site and guide RNA inactivating mutations) re-created the accessibility pattern of the parent line (Figure 6.8A). Integration of the Δ MCS sequence, on the other hand, abolished the R2 accessibility peak and led to reduced peak height over the globin

genes, as seen previously for $\Delta R1R2$ models (Figure 5.5; Hay et al., 2016). Neither integration affected DNA accessibility at the β -globin locus (Figure 6.8A). By flow cytometry, erythroid-specific YFP fluorescence distributions showed good overlap between the $\Delta R1$ and $\Delta R1R2$ α -YFP lines compared to the R2[WT] and R2[Δ MCS] lines (R2-screening cells targeted with WT and Δ MCS variant sequences), respectively (Figure 6.8B). These results demonstrate that the replacement of the endogenous R2 sequence with WT and Δ MCS sequences is able to capture the expected dynamic range of YFP fluorescence levels as seen for $\Delta R1$ and $\Delta R1R2$ models.

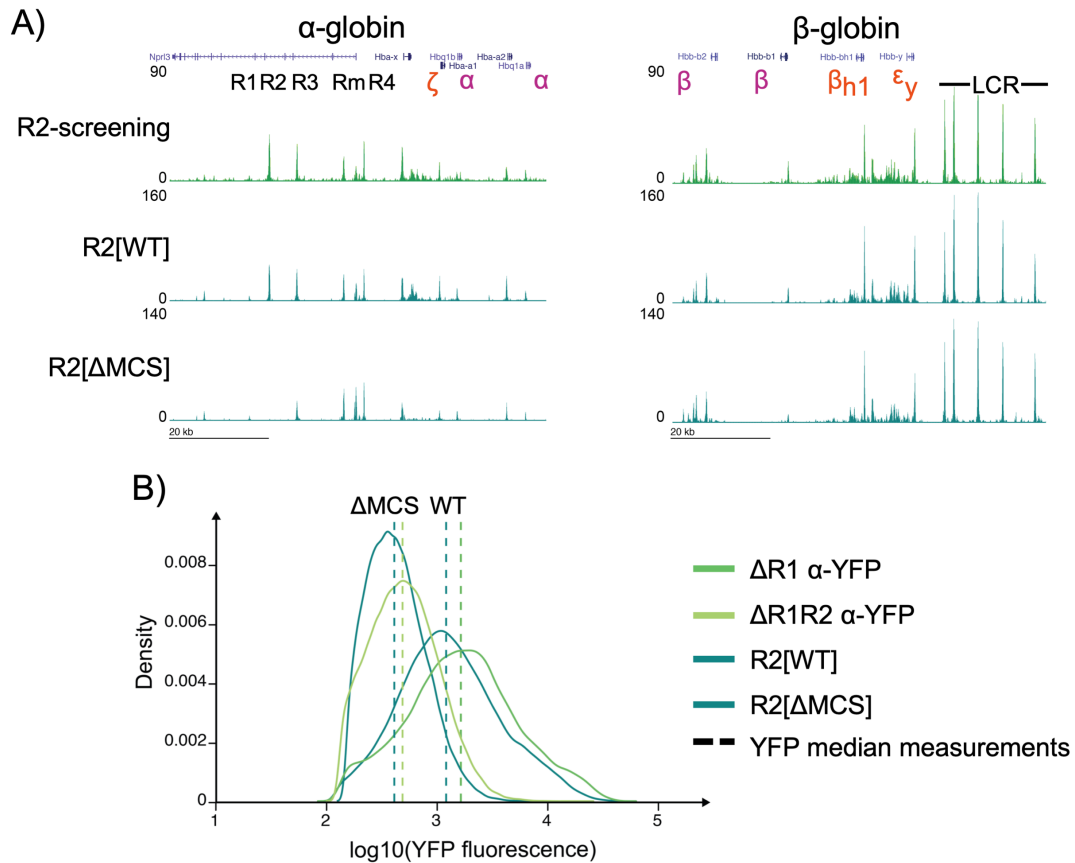


Figure 6.8: WT and Δ MCS R2 integrations by HDR result in functional and non-functional R2 elements, as expected.

A) ATAC-seq tracks of the globin loci derived from bulk-plated EB-derived erythrocytes for HDR-replaced WT and Δ MCS variants of the R2 enhancer sequence in R2-screening mESCs (R2[WT] and R2[Δ MCS], respectively), compared to the parental line (R2-screening). Tracks are normalised by RPKM and averaged for three individually targeted

clones, differentiated in parallel. Regions shown: chr11:32,131,800-32,204,799 and chr7:110,952,000-111,027,999. Tracks are scaled as described in Figure 5.3. **B)** YFP fluorescence histograms for CD71+ erythroid populations from 96-well EBs for $\Delta R1$ α -YFP and $\Delta R1R2$ α -YFP models and corresponding R2[WT] and R2[Δ MCS] clones. Medians for the HDR-targeted populations are highlighted by dashed lines for clarity. Distributions represent combined data from at least 12 independent 96-well recordings from at least four independent clones for the R2[Variant] models.

I concluded that a CRISPR/Cas9-directed HDR was a viable method to target the endogenous R2 sequence of the R2-screening mESC line with designed variants of my choice and that the resulting R2[Variant] cells are a reliable reporter model.

3. How to reliably isolate and characterise R2 variants?

A conventional strategy for a genetic screen of the R2 sequence would involve the segregation of individually targeted clones after HDR and resistance cassette excision, followed by clonal differentiation and YFP fluorescence measurement. The miniaturised *in vitro* differentiation system developed in Chapter 4 allows for this clonal strategy at a high-throughput scale (Figure 4.11).

One concern with the use of the miniaturised EB differentiation protocol for this purpose was potential well-to-well variation. Variability in EB size and haemoglobinisation is common in both large- and small-scale cultures (Wilson et al., 2014). The resulting impact on globin gene expression from single cells (as seen in Figure 4.7) is generally averaged out when plating and analysing populations at a large scale (such as in a 10 cm dish format), minimising any potential plate-to-plate experimental noise. When harvesting just a few EBs at a time, however (such as in a 96-well format), the potential for well-to-well differentiation artefacts to cause poor sampling of the CD71+ erythroid population could be severe. Although I optimised plating density conditions to maximise the CD71+ cell output per well,

I did not know whether this number of erythrocytes ($\sim 1.5 \times 10^3$ CD71+ cells from 1-10 EBs per well) would be sufficient to provide a robust readout of gene expression.

To that end, I studied the variation in fluorescence levels from 96-well EB wells derived from a number of α -YFP mESC lines used above (Figure 6.4) to calibrate the YFP fluorescence-mRNA correlation. By combining a range of 1-6 wells of a 96-well plate per recording, I found that combining EBs from three wells (representing a of total $\sim 12,000$ CD71+ cells) was sufficient to lower the variation between readings: combining further wells beyond this number did not improve the data quality (Figure 6.9A). At this sensitivity, the presence of functional and non-functional R2 sequences could clearly be separated by resulting α -YFP fluorescence levels, meaning that it would be possible to distinguish R2 variants at least into these two categories, and possibly into finer tuned brackets within this dynamic range.

Based on these results, I put together a workflow for a high-throughput, clone-based screening assay, whereby individually targeted mESC colonies could, in parallel, be differentiated *in vitro* in 96-well plates and measured for their YFP fluorescence output, genotyped for the integrated R2 variant and stored for further analysis (Figure 6.9B).

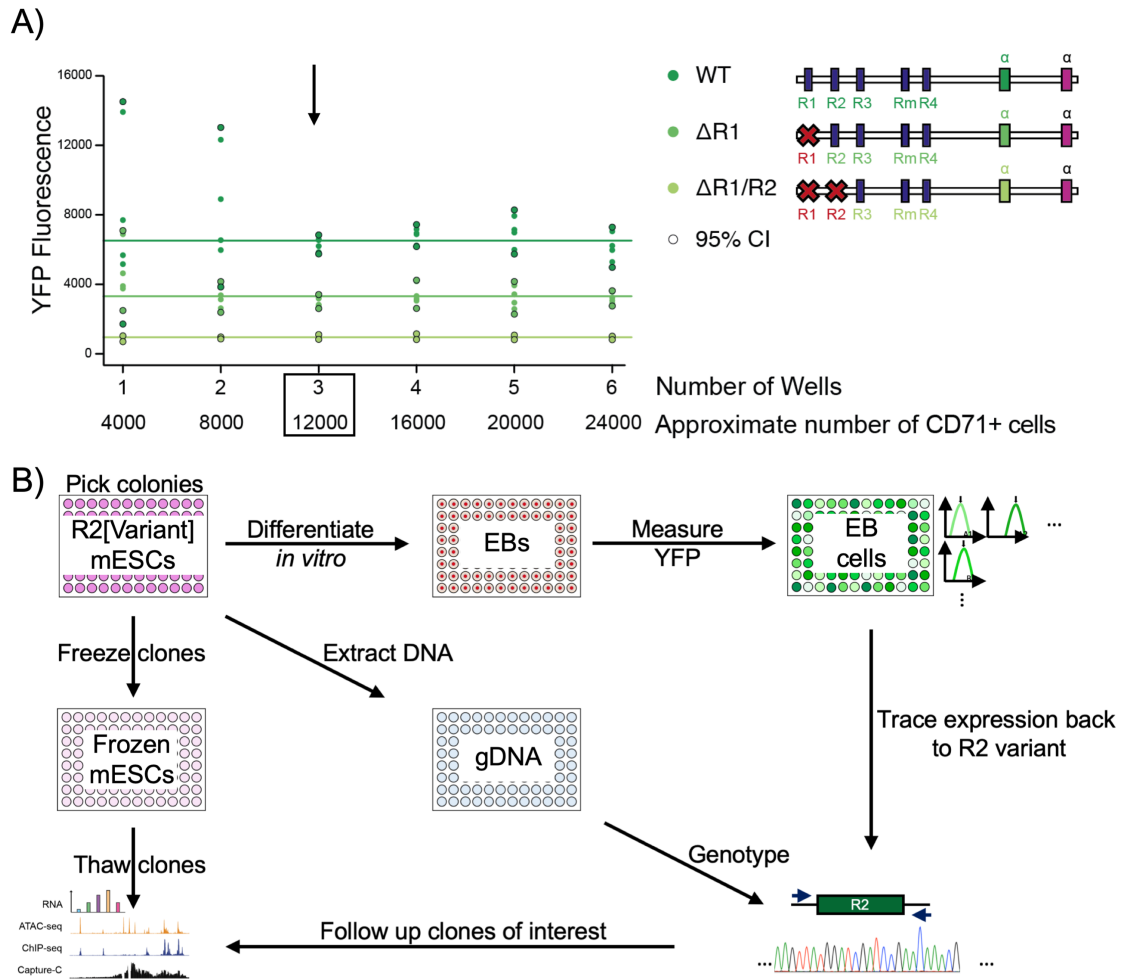


Figure 6.9: Design of a clonal screen for α -YFP fluorescence levels.

A) Right: schematic diagrams of α -YFP enhancer deletion loci used in this experiment. Left: median YFP fluorescence levels for erythroid populations from three differentiated mES cell lines with or without deletion of the R1/R2 enhancer(s) are plotted against the number of wells combined for flow cytometry analysis. Average YFP fluorescence levels for each differentiated mESC line are shown with horizontal lines. Dots with black borders indicate upper and lower bounds of 95% confidence intervals (CIs) from the shown data sample of 4 recordings each. The variation in fluorescence readings and size of CIs is decreased by combining three wells (indicated by an arrow) but no further improvement is seen by combining additional wells. **B)** Schematic overview of the clonal screening approach. The EB cells label refers to disaggregated EBs; whole EBs are recorded by flow cytometry and the YFP fluorescence measurement made for the CD71+ erythroid population only.

In summary, I answered the questions posed at the start of this design section and demonstrated that:

1. The R2-screening mESC line provides a single-copy, YFP-tagged locus that fluoresces in an erythroid-specific manner, dependent on the activity of the α -globin super-enhancer.
2. CRISPR/Cas9-directed HDR with excisable small molecule selection cassettes results in near-seamless, relatively efficient integration of sequences that can dictate the complete dynamic range R2 element function.
3. A clonal approach can be used to measure YFP fluorescence by combining multiple wells from a miniaturised differentiation system. Using this system, it is possible to reliably record results from tens to hundreds of individual clones.

With the technical issues solved, I went on to test the sequence constraints of the endogenous R2 enhancer in a high-throughput manner, in its native context.

Screen #1: comparing elements of the α -globin super-enhancer

In Chapters 3 and 5 I showed that the context of the R2 enhancer within a complete super-enhancer is important for its function. When in isolation, the R2 enhancer was barely able to activate α -globin expression to 15% of WT levels, but within the native super-enhancer it contributes ~50% to WT transcriptional output. Before delving into the grammar of the R2 sequence, I wanted to further gauge the effect of context on the function of R2: is the

element in the R2 position the strongest element of the cluster because it has the strongest position? Or because it has a strong sequence?

To answer these questions, I designed HDR donors to replace R2 with R1 or R4 enhancers of the α -globin super-enhancer. In each case, the entire MCS of the enhancer element was used. If sequence is predictive of relative enhancer activity within the cluster when comparing single enhancers, results from Chapter 5 suggest that the R1 sequence should activate α -globin transcription to roughly the same level as R2 (based on similar results from R1-only/R2-only and Δ R1/ Δ R2 models; Figure 5.6, Figure 5.11). The R4 sequence, on the other hand, is predicted to have little to no activity, based on the minor effect of its deletion *in vivo* (Figure 3.1) and its undetectable activity in the R4R2-only model (Figure 5.11).

By flow cytometry, median α -YFP measurements for the R1 and R4 enhancers aligned with the WT and Δ MCS R2 clones, respectively (Figure 6.10A-B). As in the case for the R1R2-only and R4R2-only models (the R1R2-only model conferred a boost in transcriptional activity when compared to the R2-only model, whereas the R4R2-only model did not), this observation suggests that, within a single enhancer position, the primary sequences of individual elements determine the strength of their effect on α -globin gene expression.

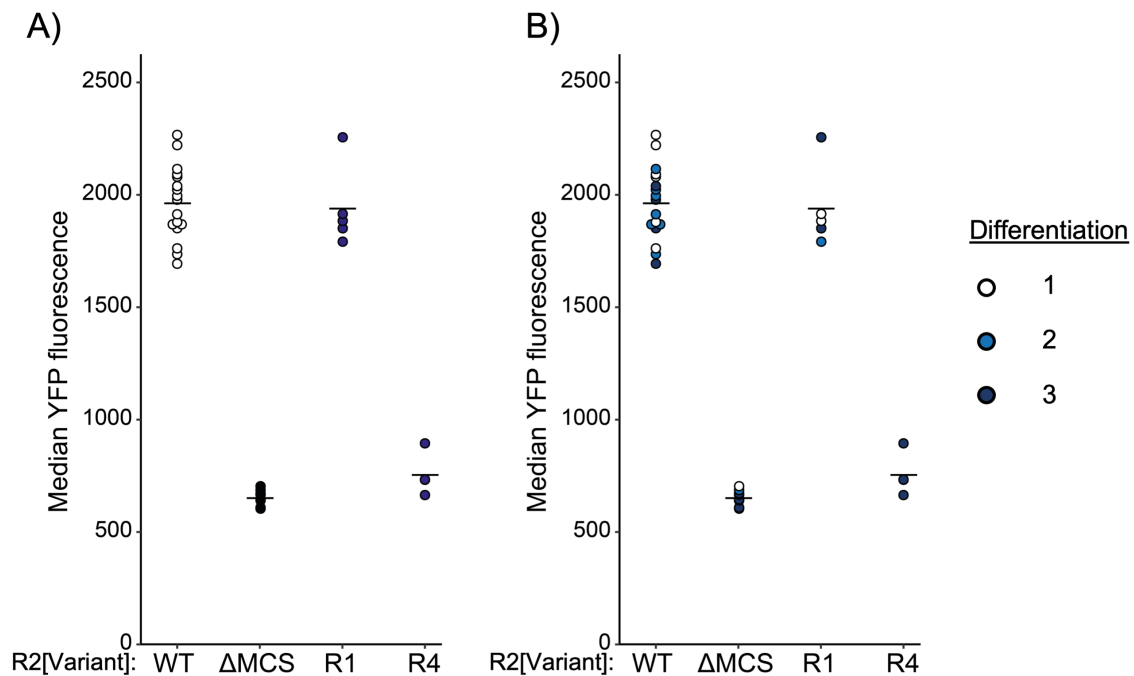


Figure 6.10: α -YFP median results for R1 or R4 integrated in place of R2 (Screen #1).

A) Median YFP fluorescence recordings for EB-derived erythrocytes for R2[WT], R2[Δ MCS], R2[R1] and R2[R4] clones in which the R2 element of R2-screening mESCs has been replaced with the indicated sequences. Individual dots each represent the average median fluorescence from two 96-well recordings, filtered for replicate equivalence and the number of reporter cells (see Methods for details). Plots represent data from at least 2 individually targeted clones per variant. Horizontal lines indicate mean results for each variant. By one-way ANOVA ($p < 1e-15$) with Tukey post-hoc test, measurements for R2[R1] were different from R2[Δ MCS] ($p < 1e-12$) but not from R2[WT]. Conversely, measurements for R2[R4] were different from R2[WT] ($p < 1e-12$) but not R2[Δ MCS]. **B)** Results from part A), coloured by differentiation date. The overlap of colours demonstrates that results were consistent across independent differentiations and clones.

I concluded that the nature of the sequence in the R2 position is critical for its function as an enhancer element.

Screen #2: gross-scale dissection of the R2 enhancer with targeted in situ mutations

In silico sequence analysis of the R2 MCS identified six predicted erythroid-specific TF binding sites (Figure 6.2). However, before proceeding with experimental dissection of these

predicted binding sites, I wanted to confirm that the annotated TFs, known to bind to the R2 sequence in adult mouse erythrocytes (Figure 5.13), are also bound in the embryonic-like EB-derived erythroid population. I successfully confirmed the binding of Gata1 and Nf-e2 at R2 by ChIP-seq (Figure 6.11A). I attempted ChIP-seq for Klf1 and Tal1 TFs in parallel, unsuccessfully. I obtained no DNA enrichment genome-wide with the Klf1 antibody; data from the Tal1 antibody showed a poor signal-to-noise ratio but showed potential enriched peaks at the R1 and R2 enhancers (Figure 6.11B). Additional optimisation of ChIP conditions would be needed to consolidate this observation. Given the similarity of binding at the α -globin super-enhancer between adult and embryonic erythrocytes for the successful ChIPs, I proceeded with the assumption that Tal1 and Klf1 were also likely to be bound to R2 in EB-derived erythrocytes (as in Ter119+ primary erythrocytes; Figure 5.13).

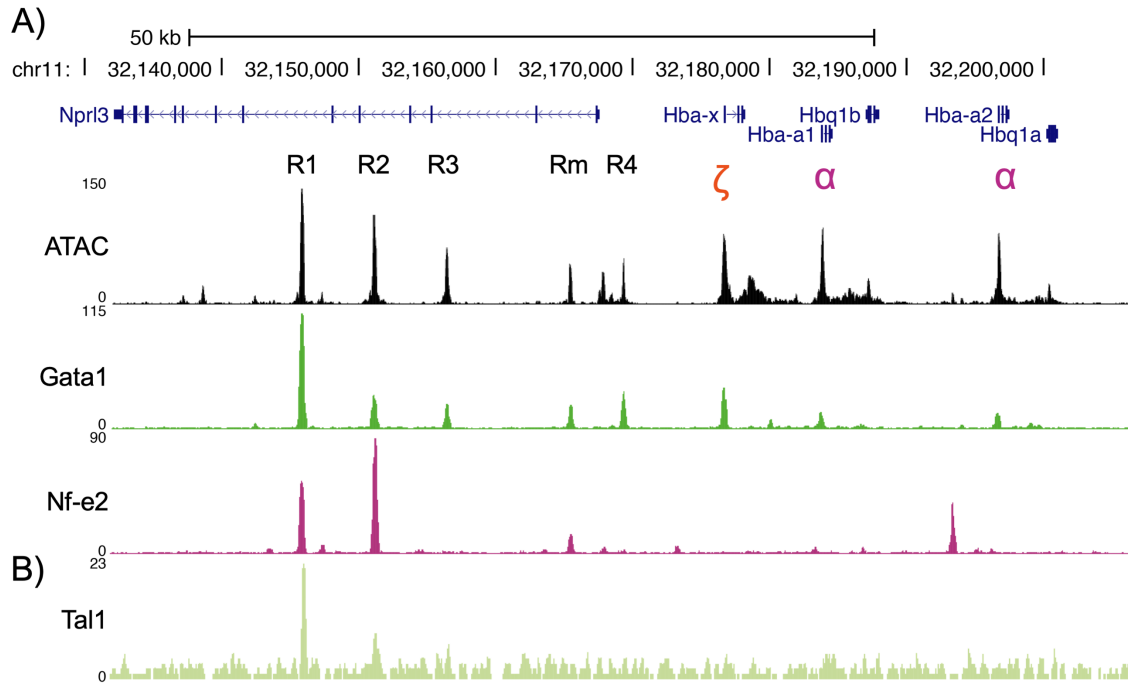


Figure 6.11: ChIP-seq for erythroid-specific TFs in EB-derived erythroid cells.

A) ChIP-seq tracks from WT E14 EB-derived erythrocytes are shown for TFs Gata1 and Nf-e2. ATAC-seq data from WT EB-derived erythrocytes from Figure 4.4 are shown for reference. Tracks for Gata1 and Nf-e2 are normalised by RPKM and averaged over 2 replicates from independent differentiation experiments. With the exception of a new Gata1 peak at the ζ -globin gene promoter (*Hba-x*) and more prominent peak for Nf-e2 at the Rm enhancer, tracks for Gata1 and Nf-e2 are comparable to those generated from erythrocytes derived from adult mouse APH-treated spleen (Figure 5.13). Differential Gata1 signal at the *Hba-x* promoter likely reflects the fact that expression of this gene is restricted to embryonic erythroid cells (such as those derived from EBs). The Nf-e2 peak at Rm may be identifiable here due to a greater signal-to-noise ratio than for the adult ChIP-seq data. **B)** ChIP-seq track for a single replicate of EB-derived erythrocytes for the Tal1 TF. Data are presented as raw mapped reads; these are not of high enough quality to assess differential binding but show potential TF binding at the R1 and R2 enhancers.

For an initial interrogation of the R2 sequence, I chose to divide the R2 MCS into six consecutive blocks, each containing a single predicted TF binding site, and test their individual contributions to R2 function (Figure 6.12A). I created a library of 12 HDR donors, each containing an R2 variant in which a single block was deleted from the WT sequence or inserted into a Δ MCS background (Figure 6.12B). I confirmed that the new junctions formed over deleted regions did not create novel predicted footprints by the published

Sasquatch tool (Schwessinger et al., 2017) or motifs for Gata1/Nf-e2/Klf1/Tal1 by FIMO analysis (Grant et al., 2011). I targeted R2-screening mESCs with the cloned library of HDR donors in a single transfection and screened resulting colonies by an NGS-based genotyping approach (see Methods for details). I generated EBs from successfully targeted clones in 96-well format over four independent differentiations and recorded the median YFP fluorescence level for each clone.

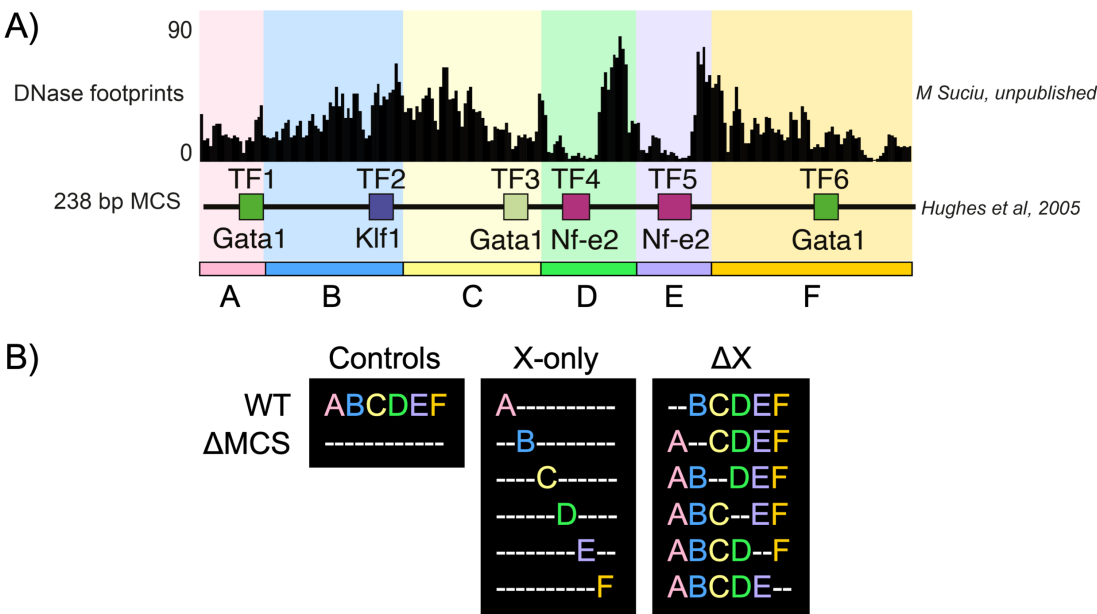


Figure 6.12: Design for Screen #2: a gross-scale interrogation of R2 sequence composition.

A) *In silico* R2 MCS analysis from Figure 6.2 annotated with designed “blocks” that divide the sequence into six regions, each containing a predicted erythroid-specific TF binding site. **B)** Summary of the R2 variant designs to be included in the pilot screen. Letters refer to the blocks annotated in part A).

Immediately apparent from the results was a split between the variants constructed with a single block remaining (X-only) and those with a single block removed (ΔX; Figure 6.13A-B). X-only clones all gave YFP fluorescence levels that were indistinguishable from the non-functional ΔMCS sequence, indicating that any single block in isolation does not harbour an enhancer function. The ΔX clones, on the other hand, all showed higher

fluorescence levels than the Δ MCS baseline but lower levels than the WT sequence, indicating that they were all functionally compromised by the removal of a single block. The extent of functional impairment varied depending on which block was deleted; the deletion with the greatest effect was Δ B, which represents predicted removal of the only conserved Klf1 binding site. The Gata1-targeted Δ A and Δ C models had the next greatest functional impairment, followed by the Nf-e2-targeted Δ D and Δ E and, finally, the flanking Gata1-targeted Δ F. Considered as a whole, these results indicate that all of the designed blocks are individually necessary for maximal R2 function (albeit to different extents) but none is sufficient.

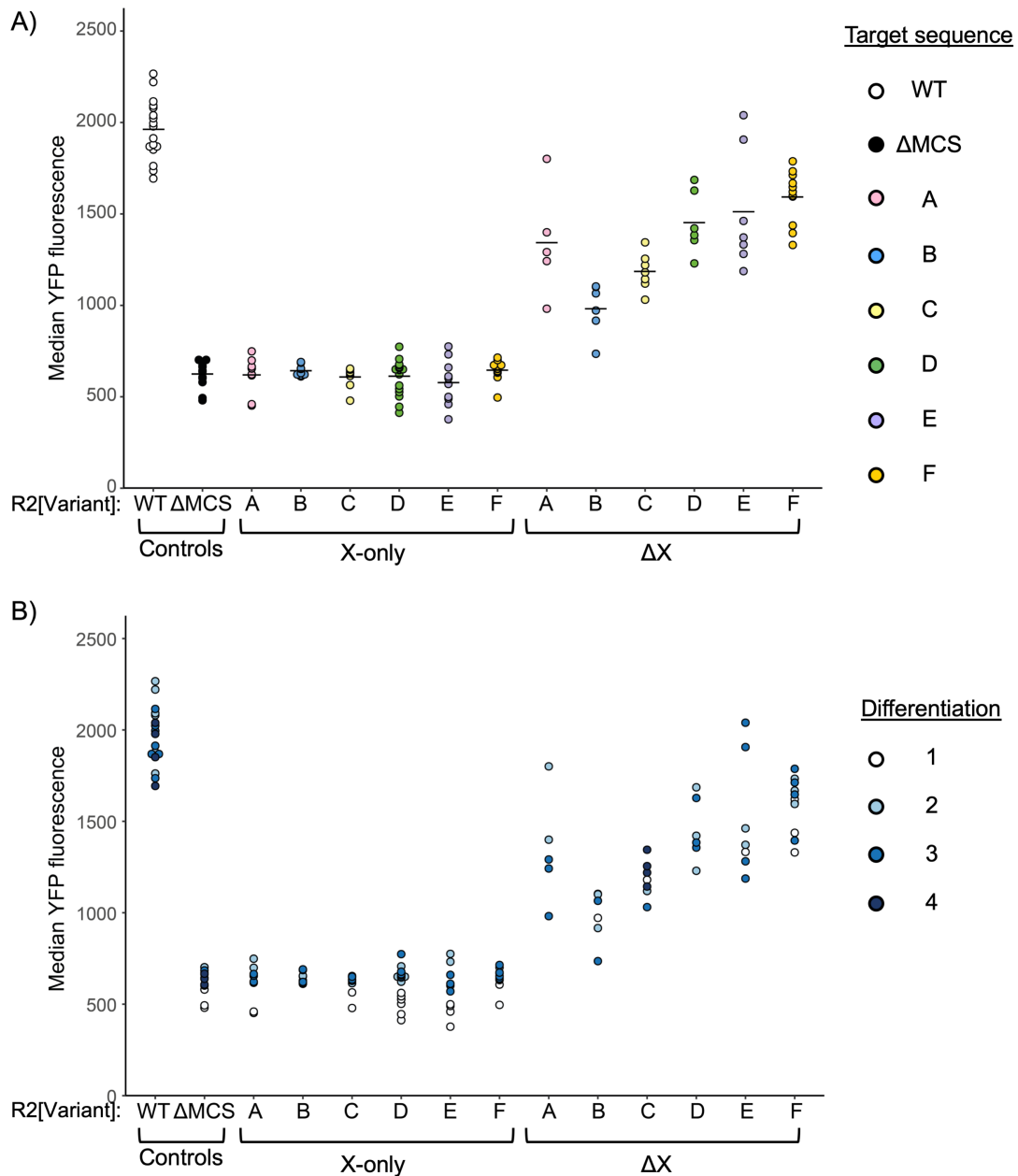


Figure 6.13: Results from R2 Screen #2.

A) Median YFP fluorescence recordings for all R2[X-only] and R2[ΔX] clones, compared to R2[WT] and R2[ΔMCS] clones from Screen #1. Individual dots represent the average median fluorescence from two 96-well recordings, filtered for replicate equivalence and the number of reporter cells (see Methods for details). Data are combined from at least 2 individually targeted clones per variant, each differentiated in up to 4 individual experiments. Horizontal lines indicate mean results for each variant. By one-way ANOVA ($p < 1e-15$) with Tukey post-hoc test, all variants showed significantly different fluorescence levels compared to the WT integrated sequence ($p < 1e-7$) and levels from the ΔX variants were also all statistically different from those of ΔMCS sequence ($p < 1e-4$).

B) Results from part A), coloured by differentiation date. As for Figure 6.10, the overlap of colours demonstrates that results were consistent across independent differentiations and clones.

Screen #3: precise perturbation of predicted TF binding sites in the R2 enhancer

One interpretation from the single-block deletion results of Screen #2 would be that diversity of TF binding is key for R2 function: removal of block B, containing the only conserved Klf1 motif, resulted in the greatest reduction in R2-directed activation of α -globin expression. However, the gross-scale modifications used in this initial study are a blunt tool that may have masked the intricacies of the system, such as the presence of additional bound motifs or sequence grammar. To test whether results from the gross-scale modifications used in the pilot screen could be recapitulated by direct disruption of the contained conserved TF motifs, I designed two mutant variants for each motif (Figure 6.14A). Mutations substituted 1-5 key bases within the predicted motifs while preserving GC content. Using the same tools as in Screen #2, I confirmed that the designed mutations abolished motif matches (Grant et al., 2011) and did not create novel predicted footprints (Schwessinger et al., 2017). In parallel, I designed variants to mutate three additional motifs identified by FIMO (Figure 6.14B; Grant et al., 2011), two additional peaks for PhastCons conservation score (Figure 6.14C; Siepel et al., 2005) and combinations of motif mutations to either remove all occurrences of one or more single motif types (Figure 6.14D-F).

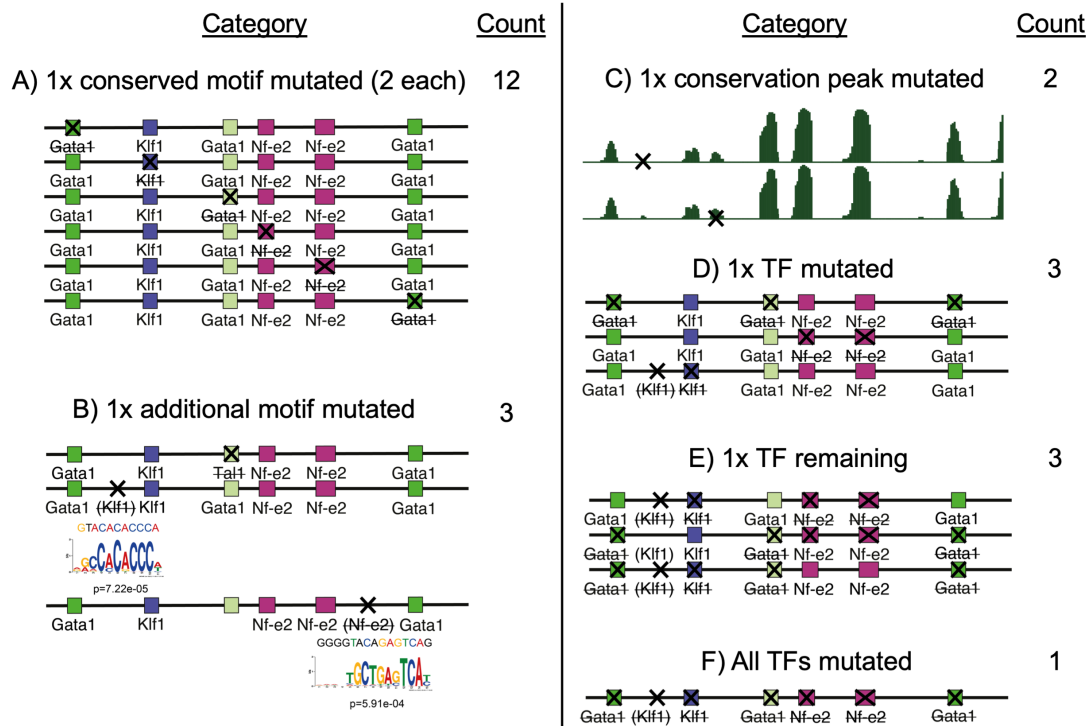


Figure 6.14: Design of Screen #3: an R2 screen to disrupt TF binding sites alone and in combinations.

Schematic illustrations of all 24 variants designed for Screen #3, categorised into groups; the total number of variants for each group are shown under “Count”. Mutations of 1-5 bp are indicated for all variants with black crosses. **A)** Variants in which a single predicted motif was mutated. Two variants were designed for each motif with alternative sequence mutations. **B)** Variants mutating one additional, non-conserved TF binding site; matches to the proposed TF motifs are shown below (Klf1, Nf-e2) or in Figure 6.2 (Tal1). **C)** Variants that mutate a single conservation peak; PhastCons data are shown as in Figure 6.2. **D)** Variants with combinatorial mutations for all motifs of a single type. In the Δ Klf1 variant, both the conserved and non-conserved sequences were mutated due to the near-perfect motif match for the non-conserved site. **E)** Variants with combination mutations for all motifs of two factors, leaving one intact. **F)** A variant in which all predicted motifs are mutated.

When tackling an experiment with 24 variant sequences, the modest efficiency of the CRISPR/Cas9-directed HDR protocol (~16%) became a limiting factor. To make the YFP-based screening approach feasible for large-scale screens, I optimised a novel screening approach, in which I could demultiplex individual variant sequences from a pool of R2[Variant] mESCs.

An example analysis output for a pooled fluorescence screen is shown in Figure 6.15A-B. Full details for the analysis workflow are provided in the Methods section of this thesis. Briefly, this relies on the fact that the median YFP fluorescence level, instead of being thought of as the halfway point in a histogram representation, can be expressed as the point at which a cumulative density function (CDF) for the fluorescence distribution reaches 0.5 (Figure 6.15A). By dividing and sorting the YFP-positive population into gates based on fluorescence intensity, the contributions of individual variants to different fluorescence levels can be determined by NGS genotyping over the R2 position. A sigmoidal curve, fitted to data points obtained in this way, can be used as an approximation for the CDF for each contributing variant sequence. A median YFP fluorescence measurement can, correspondingly, be estimated by the intersection of this curve with the line where density=0.5 (Figure 6.15B). In reality, this process is inherently noisy but is able to recapitulate raw data when replicate experiments are performed. Thus, I have developed an analysis pipeline to approximate median YFP fluorescence values for individual variants from a pool of multiple R2[Variant] clones.

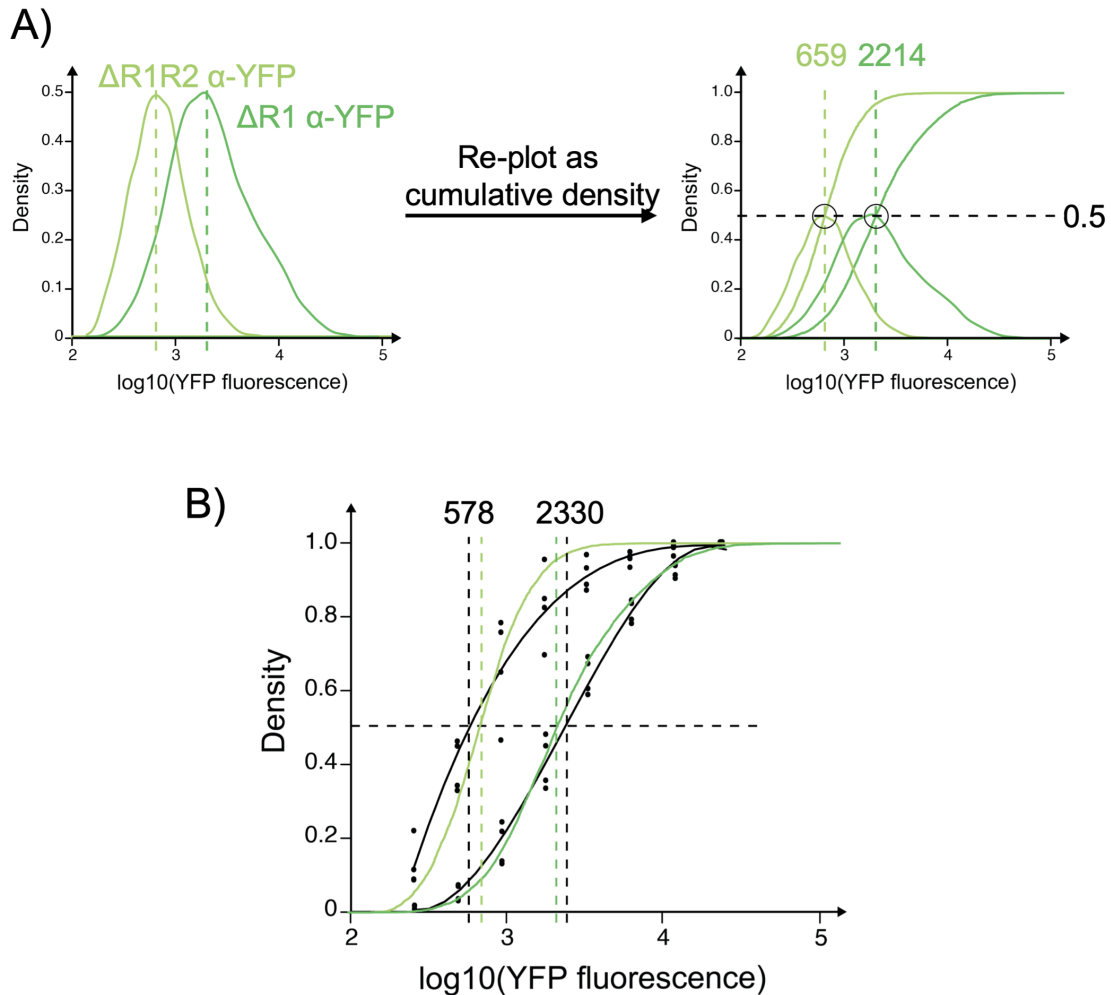


Figure 6.15: Measurement of approximate median fluorescence values from R2 variant pools via gated YFP populations.

A) Erythroid-specific α -YFP fluorescence distributions for $\Delta R1$ and $\Delta R1R2$ models as shown in Figure 6.4 (left) can be transformed into CDFs (right). Green vertical dashed lines highlight the median YFP levels for each model. The black horizontal dashed line on the CDF plot represents the point at which the CDF value is equal to 0.5. Intersections between CDF curves and the black dashed line are highlighted with circles and occur at median YFP values, as seen by co-intersection of the green dashed lines. Green numbers above quantify median YFP measurements for each model. **B)** Demonstration of the analysis method for clones carrying functional (R2[WT]) and non-functional (R2[All TFs mutated]; Figure 6.14) variants of the R2 sequence (see Figure 6.17 for complete results from Screen #3). Plotting approximated CDF curves from four replicate experiments provided good estimates of median YFP values when compared to clonal CDFs from $\Delta R1$ α -YFP and $\Delta R1R2$ α -YFP models.

One concern when studying pools of sequences in a single experiment is that library complexity should be maintained throughout the targeting and analysis procedure. After targeting R2-screening mESCs with the library of HDR donors for Screen #3, I confirmed that all sequences were represented to an acceptable level by comparing estimated cell counts for each R2[Variant] cell type after NGS genotyping of YFP gated populations. As a conservative estimate, the total representation of HDR-exchanged variants should be ~10% (based on the ~16% targeting efficiency seen in screens #1 and #2). Therefore, assuming equal integration of pooled HDR donors, the included variants should each represent $10\% \div 24 = \sim 0.4\%$ of the final estimated cell counts. To filter out variants with unacceptably low estimated cell counts, I set the minimum cut-off as 8 cells (0.05% of the 16,000 cells purified by gate and genotyped). According to these metrics, all variants were suitably represented in the total cell population across all four replicate experiments (Figure 6.16). The spread of representations was most even when counts from all experiments were combined. I performed one additional data filtering step to ensure that curve fitting was not influenced by missing data points; 7/96 total samples were removed in this process (see Methods for details). For the remaining 89 data points, I proceeded to calculate approximate YFP values as described above.

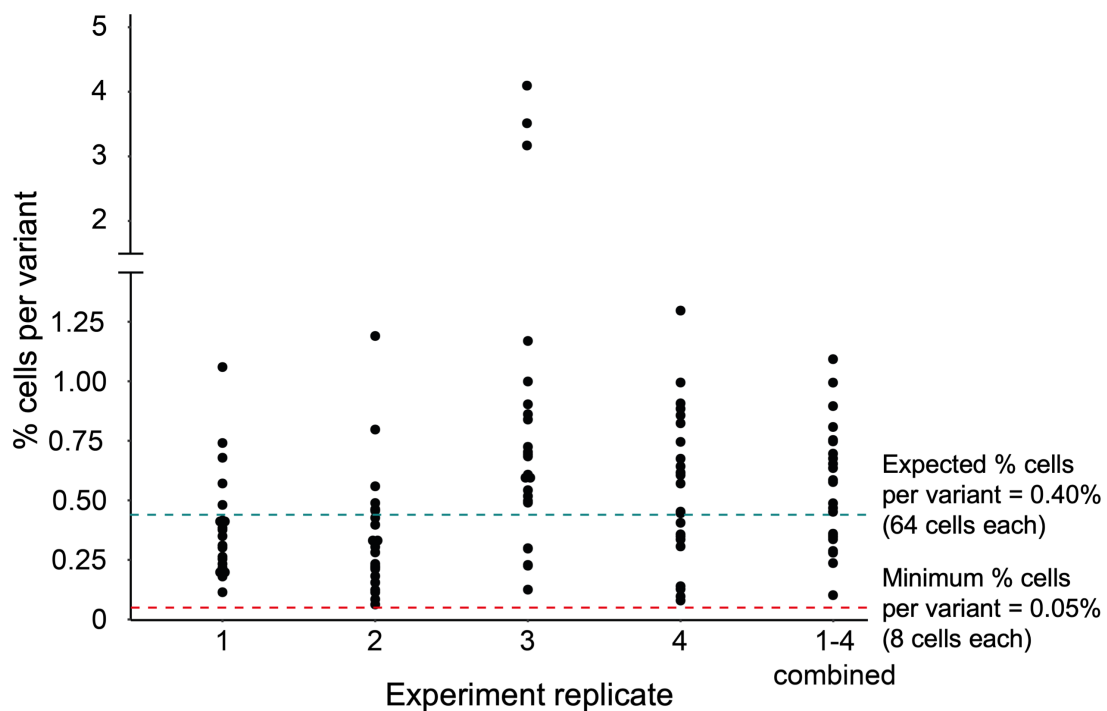


Figure 6.16: Representations of individual variant sequences in Screen #3.

Variant sequence representations are depicted as percentages of total estimated cell counts across the YFP gated population. Each dot represents the quantitation for an individual variant sequence. The blue dashed line indicates the expected representation of each sequence if all variants are present in an equal ratio. The red dashed line indicates the minimum lower limit for cell representation of each individual variant. Data are shown for four individually acquired datasets and a combined estimated cell count incorporating all replicates.

Results for the variants in Screen #3 are grouped by the sequence block(s) they affect (as defined in Figure 6.12A). For single motif disruptions, results are compared to clonal data from Screen #2 for functional (WT) and non-functional (Δ MCS) sequences and relevant Δ X/X-only models (Figure 6.17A-F). Schematic illustrations of the labelled mutations are shown in the same panel for clarity (Figure 6.17G). For combinatorial variants in which, for example, all sites for a particular TF are mutated in series, results are shown adjacent to a summary of the single mutations of that type (Figure 4.18A). As for the single motif disruptions, an illustration of labelled variants is included in the same panel for clarity (Figure 4.18B). Below, I summarise results from each motif by type.

horizontal bars represent median α -YFP estimations obtained by combining estimated cell counts from all replicates. N=4 for all variants except B ~~Klf1~~ 1 (N=3), C ~~Tal1~~ (N=3), and F ~~Gata1~~ 2 (N=3), where single replicates did not pass filtering criteria. Variant sequences are labelled and numbered for the included mutations, corresponding to the diagrams in part G); for each block, the major conserved motif was mutated with two alternative sequences. In part B), combinatorial mutant variants Δ Klf1 and Klf1-only are included for comparison. By one-way ANOVA ($p < 1e-8$) with Tukey post-hoc test, median α -YFP recordings that were reduced compared to results from R2[WT] clones are highlighted with an asterisk (*, $p < 0.06$). **G)** Illustrative key to the variants labelled in parts A-F).

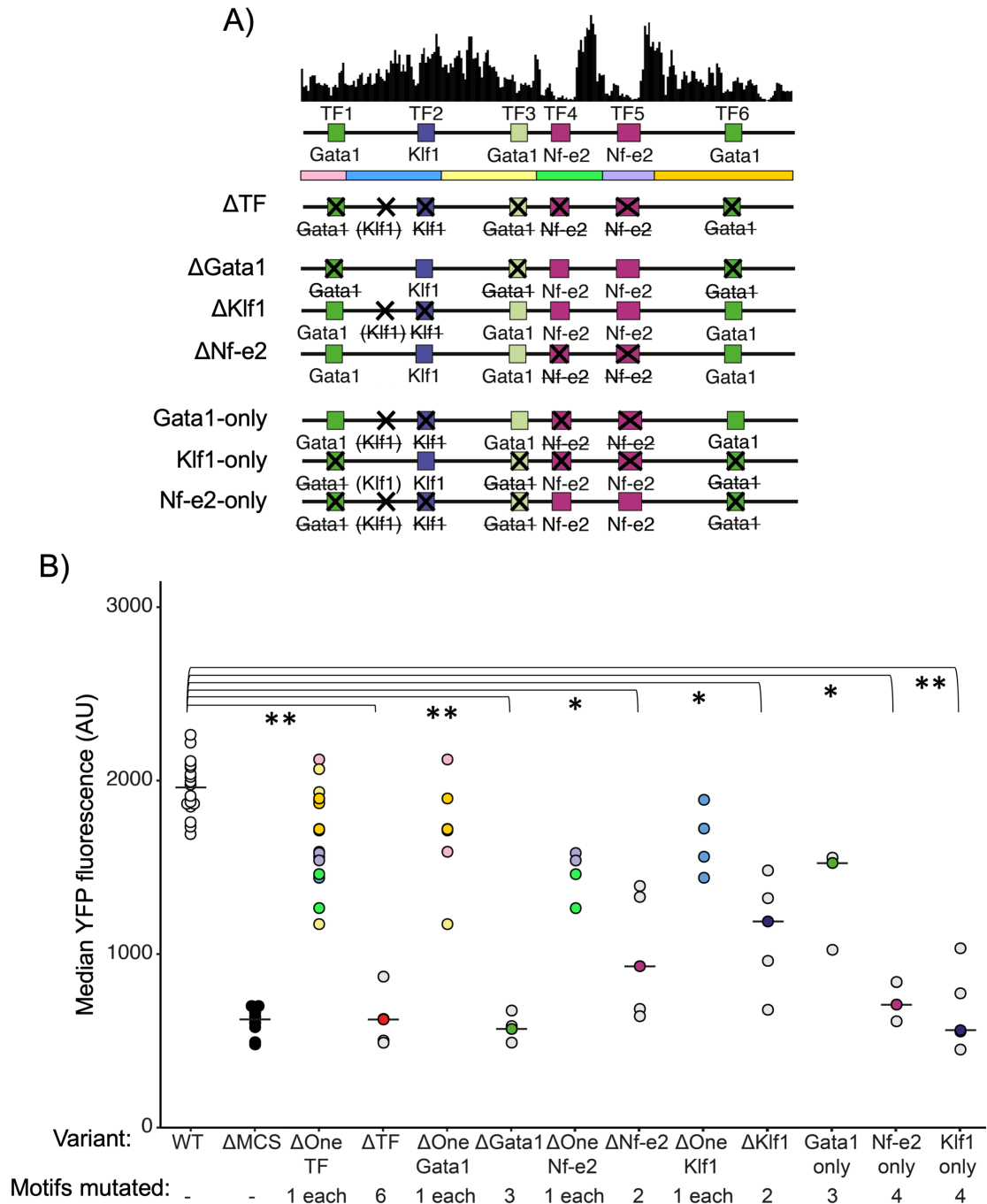


Figure 6.18: The impact of combined motif mutations on the function of the R2 MCS (R2 Screen #3).

A) Illustrative key for combinatorial variants. **B)** Estimated median α -YFP fluorescence levels for 7 R2 motif combinatorial variant sequences; grey dots represent results from individual experiments and coloured dots with horizontal bars represent median α -YFP estimations obtained by combining estimated cell counts from all replicates. Results for R2[WT] and R2[ΔMCS] clones from Screen #1 and coloured dots for single motif mutations from Figure 6.17, categorised by TF motif type (ΔOne TF), are shown for context. The number of site mutations used to generate each combined variant is shown below (motifs mutated). N=4 for all variants except ΔGata1 (N=3), Gata1-only (N=2) and

Nf-e2-only (N=3), where individual replicates did not pass filtering criteria. By one-way ANOVA ($p < 1e-8$) with Tukey post-hoc test, median α -YFP recordings that were statistically different to results from R2[WT] clones are highlighted with asterisks (*, $p < 1e-02$; **, $p < 1e-04$).

Gata1

Point mutations of the individual Gata1 motif in blocks A, C and F did not consistently replicate the effects seen from Gata1 block deletions. For blocks A and C, results from two alternative mutations of the Gata1 motif could not be reconciled (Figure 6.17A, C). Only one of two mutations in each case had a similar effect as deletion of the entire block and only one of these represented a statistically significant reduction in α -YFP levels (block C ~~Gata1~~ 1). For block F, the less extreme effect of block deletion was partially recapitulated by both Gata1 motif mutations (Figure 6.17F). Therefore, whereas point mutations of individual Gata1 motifs did not consistently replicate the effects of entire block deletions, the functional impairment seen in 4/6 Gata1 motif mutant variants (though only statistically significant in one instance) indicates that some level of functional activity can be attributed to these sequences.

Mutation of all three conserved Gata1 sites in series, on the other hand, resulted in a complete loss of function of the R2 sequence (Figure 6.18A). This was the strongest effect of any of the included Δ TF models for a specific factor. Although the loss of Gata1 binding has not been biochemically confirmed by ChIP-seq, this suggests that Gata1 is required for R2 function. Conversely, leaving Gata1 motifs as the only remaining conserved sites gave the greatest residual function of any individual TF model (Gata1-only; Figure 6.18A). Two data points for the Gata1-only variant were removed in the data filtering process, meaning that additional replicate experiments should be performed to consolidate this finding.

Tal1

There is one predicted Tal1 motif in the R2 MCS, associated with the Gata1 motif in Block C. Despite ChIP-seq data suggesting that Tal1 is likely to be bound at this site in erythrocytes (Figure 5.13; Figure 6.11B), specific disruption of the Tal1 binding motif did not have any discernible effect on R2 activity in this assay (Figure 6.17C). Analysis of Tal1 binding by ChIP-seq would establish whether its recruitment is compromised by the included mutation. Instead, it seems likely that Tal1 is recruited to R2 indirectly, as has been shown previously to be a possible mode of binding (Kassouf et al., 2008; Kassouf et al., 2010).

Nf-e2

There are two conserved Nf-e2 sites in the R2 MCS, assigned to blocks D and E. Both conserved sites have strong matches to the established binding motif (Igarashi et al., 1994) and remarkably pronounced footprints by DNase-seq (Figure 6.2). In this experiment, two alternative mutations of each conserved Nf-e2 motif had the same impact on R2 activity as deletion of the entire blocks (ΔD and ΔE ; Figure 6.17D-E). This strongly suggests that the binding of Nf-e2 to motifs in blocks D and E is necessary for function of the combined 58 bp region of the R2 enhancer MCS. One additional, non-conserved potential match for the Nf-e2 motif by FIMO analysis is found within Block F ($p=5.9e-04$; Grant et al., 2011), though the exact motif is not predicted to be bound based on results from a gel shift assay (Andrews et al., 1993). Mutation of this sequence had no detectable effect on R2 function, suggesting that it is either not bound or not functional (Figure 6.17F).

Mutation of the two conserved Nf-e2 sites in series resulted in greater R2 functional disruption than that of either site in isolation (Figure 6.18A). This is in line with published data from direct interrogation of these two sites in an *in vivo* mouse model, where

replacement of a region containing both Nf-e2 sites with a lox sequence resulted in partial loss of function of the R2 sequence (Loyd et al., 2003). The functional significance of Nf-e2 recruitment to the R2 sequence (to be confirmed by ChIP-seq) is not seen when the motifs are present in isolation; a model with all conserved binding sites mutated except for Nf-e2 showed no gain in function compared to the null Δ TF model (Figure 6.18A).

Klf1

There are two motif matches for Klf1 in the R2 MCS region, both positioned within block B. One site is conserved and the other, while only present in mouse and rat genomes (Figure 6.1), shows a near-perfect match to the Klf1 consensus sequence (Figure 6.14; Tallack et al., 2010). Disruption of the non-conserved site did not have any significant impact on R2 function (Figure 6.17B). Mutation of the conserved site, however, resulted in a modest decrease in R2 function in data from two alternative mutation constructs (Figure 6.17B). Pending confirmation by ChIP-seq, these data suggest that the conserved site is more critical to R2 function.

Mutation of both predicted Klf1 sites in series resulted in a more pronounced decrease in R2 function, recapitulating deletion of the entire B block (Δ B; Figure 6.17B). Therefore, it appears that compromising the binding of Klf1 through targeted motif mutations accounts for the severe functional impairment seen in the Δ B model. Although the potential for the R2 MCS to recruit Klf1 seems to be critical for its function, the presence of Klf1 binding sites alone is not sufficient to create a functional enhancer sequence. A variant harbouring mutations for all conserved sites except for Klf1 (Klf1-only) was not functional; YFP fluorescence levels from this model were indistinguishable from recordings for Δ TF or B-only sequences (Figure 6.18A).

Together, it appears that, while mutations of the predicted TF motifs are generally able to partially reproduce the effects of block deletions, there may be additional binding events or sequence grammar restrictions that contribute to the function of complete sequence blocks. To test the entire R2 MCS in an unbiased manner, I performed a screen in which I tiled a 6 bp deletion across the entire sequence in steps of 2 bp. Analysis for this experiment is in progress; results may highlight additional regions of enriched function that are not marked by the sequence conservation and motif characteristics used here to make motif predictions. To probe the question of sequence grammar, I chose to manipulate a single motif spacing interval in a controlled manner, described below.

Screen #4: variation of conserved Gata1-Nf-e2 spacing

To test the tolerance of the R2 enhancer to changes in sequence grammar, rather than motif integrity, I designed a series of enhancer variants carrying a range of spacing distances between the central Gata1 and Nf-e2 motifs (associated with blocks C and D, respectively; interval IIb, Figure 6.19A). The Gata1-Nf-e2 spacing was previously found to be exactly 15 bp in a range of organisms spanning 100 million years of evolution (Figure 6.19B). The HDR donor vectors for Screen #4 represent 10 sequence variants: 2 variants each representing 5 bp, 10 bp, 15 bp, 20 bp or 25 bp spacing between the centres of the Gata1 and Nf-e2 motifs (Figure 6.19A). The complete Gata1 and Nf-e2 motifs and GC-content are maintained in all designed variants. No additional predicted footprints or motif matches were created in any variants, as described for Screen #2. I analysed the functions of the pooled variants as described for Screen #3 (Figure 6.15). I removed one designed variant from the analysis due to low sequence representation (15b; Figure 6.19C).

Unexpectedly, variation of the native 15 bp IIb spacing by up to 10 bp in either direction had no effect on R2 sequence function (Figure 6.19D). Estimated median α -YFP fluorescence levels from all variant sequences showed no changes across two individual experiments. These results suggest that, despite the conserved IIb lengths seen in nature, this region of the R2 enhancer is remarkably tolerant of changes in structure and not constrained by rules such as helical phasing (Huang et al., 2012).

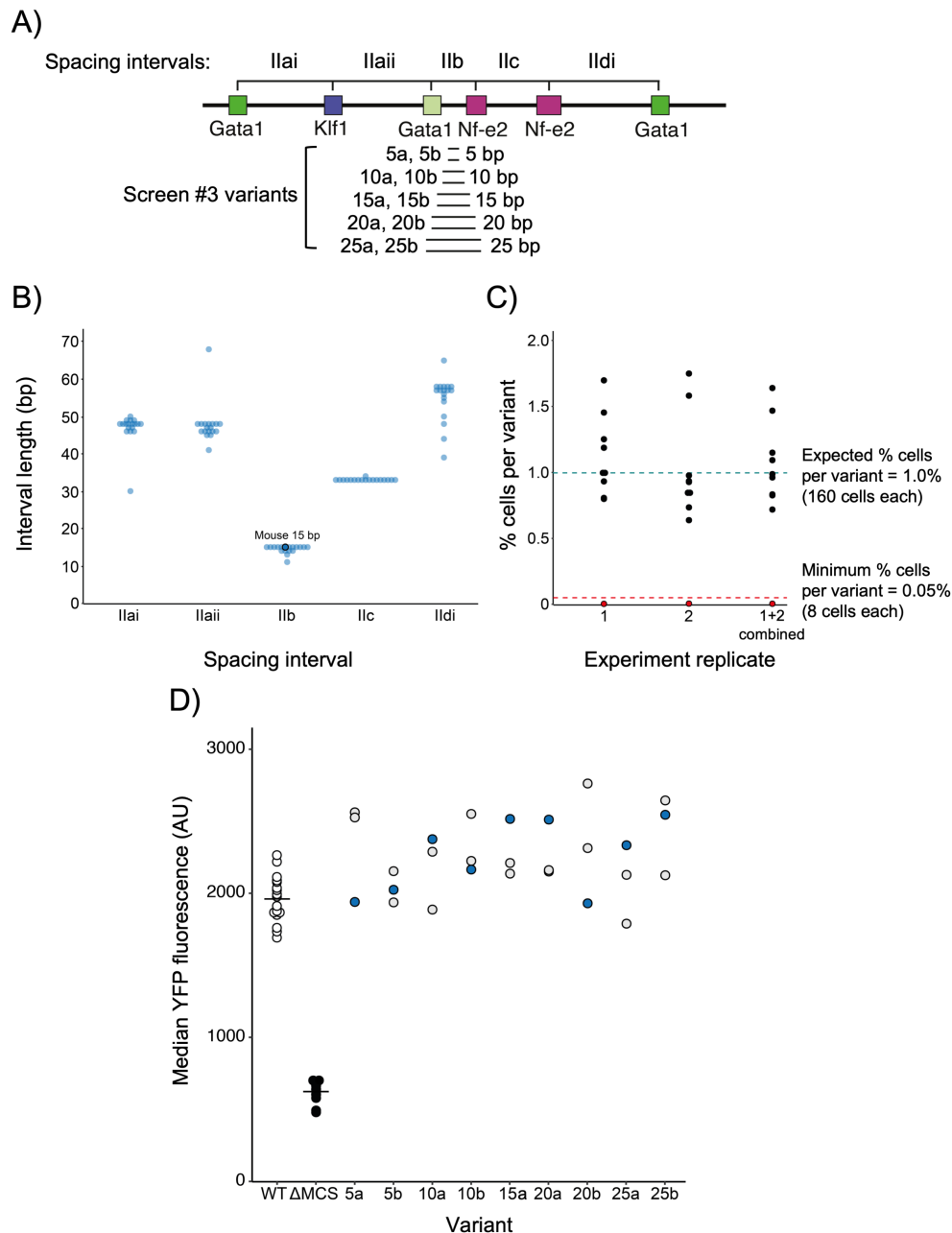


Figure 6.19: The effect of Gata1-Nf-e2 motif spacing on R2 function (Screen #4).

A) Schematic illustration of the R2 MCS and annotated motif-motif distances. The IIb spacing is modified in Screen #4 as indicated by the included variant lengths. **B)** Plot of motif-motif distances in up to 18 species where sequences were conserved (Hughes et al., 2005). IIb and IIc distances appear to be constrained through evolution. **C)** Representation of Screen #4 variants across two replicate experiments (1, 2) and the combined dataset (1+2). Blue and red dashed lines indicate the expected and minimum cell numbers required for analysis. One variant sequence (15b) was not represented to a sufficient level for analysis (data point below the red dotted line) and was removed from subsequent analyses. **D)** Estimated median α -YFP fluorescence levels for 9 R2 variant sequences specified in part A) and labelled on the x-axis, shown with R2[WT] and R2[Δ MCS] results from Screen #1 for context. No changes in α -YFP level were observed.

The effect of mutational burden on R2 function

One caveat in the design of this R2 sequence screen is the incremental accumulation of base pair mutations compared to the native R2 MCS. It is therefore possible that the observed impacts of sequences such as Δ Gata1, Klf1-only and Nf-e2-only are due to more extensive general disruption of the WT sequence, rather than a consequence of disrupting the desired TF binding events. While not necessarily critical for flexible models of TF binding (Arnosti and Kulkarni, 2005), effects of this mutational burden could influence TF binding by influencing DNA shape or novel binding specificities of combined TFs (Levo and Segal, 2014). To control for this factor, I assessed existing variants for their observed effect on R2 activity with respect to: 1) the number of predicted TF motifs and 2) the number of base pairs mutated. As expected, the mutation of a greater number of TF motifs resulted in a greater effect on enhancer activity (Figure 6.20A). The one clear outlier to this trend was for the Gata1-only variant (4 TF motifs mutated), as discussed above. Next, to disentangle the effect of general mutational burden from increasing TF motif disruption, I focused on variants where 0 or 1 predicted TF motif(s) were affected. I plotted the functional effect of each variant against the number of mutated base pairs and observed no trend between the number of mutated base pairs and effect on enhancer function (Figure 6.20B). Together, these results suggest that mutational burden is unlikely to be consequential for R2 function in this study. Thus, results from the sequence screens can be evaluated in terms of the specific binding events they target.

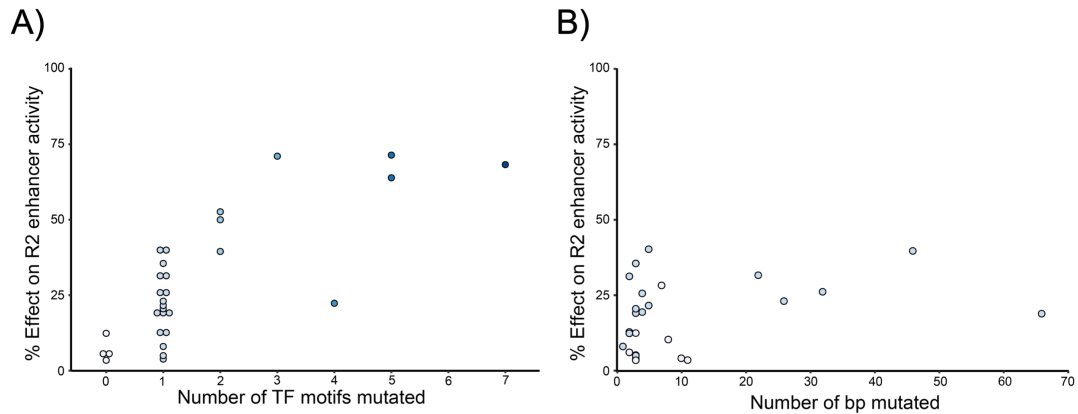


Figure 6.20: The effect of R2 mutation depends on the disruption of TF motifs and not on the extent of sequence mutation in number of bp.

A) The effect of sequence mutation on R2 enhancer function plotted against the number of predicted TF motifs mutated for each of the variants in Screens #2 and #3. Effects are represented for the combined average approximate median α -YFP for each variant as a percentage change compared to R2[WT] and R2[Δ MCS]. Motifs counted were the 6 conserved motifs highlighted in Figure 6.2 and the additional Klf1 consensus sequence match in block B. The Tal1 half E-box was not included. Variants with 0 motifs mutated were the spacing variants from Screen #4. **B)** The effect of sequence mutation on R2 enhancer function plotted against the number of base pairs mutated for variants disrupting 0 or 1 of the predicted TF motifs included in part A). Base pairs were counted as mutated if they were substituted or deleted compared to WT. No trend was seen between effect size and mutational burden.

6.3 Discussion

A novel screening method to test mammalian regulatory sequences in situ

In this chapter, I have described the development of a fluorescent reporter-based screening method that can probe the function of any regulatory element associated with an identified target gene, from its endogenous position. Considerations when applying this method at the scale of a few to hundreds of variants to other regions of the genome are the amenability of the cell type and target sequence to editing by CRISPR/Cas9-directed HDR (Zhang et al., 2017) and the length restriction of available high-output NGS technologies for designed variant screening (currently 600 bp for Illumina short-read sequencing technology; Illumina, 2019). A further restriction to consider when determining the throughput of the experimental design is the number of reads needed to reliably analyse gated YFP populations. In the design for a higher throughput experiment reported in this chapter (Screen #3), the number of variants (24) was modest compared to the large-scale experiments performed in other studies (Levo and Segal, 2014; Slattery et al., 2014), but an unbiased screen for tiled deletions was successful in testing 117 sequence variants in parallel (data analysis in progress). With minor technical modifications of the existing workflow used here, throughput could reasonably be pushed to testing around 1000 variants in a single experiment. Therefore, while not suitable for exhaustive sequence dissections, this protocol represents a scalable system to test sequence function *in situ*, in a near-scarless and controlled manner.

Context is as critical at an intra-enhancer resolution as for the inter-enhancer scale

Results from Screen #2 corroborate findings from earlier chapters: the activity of a sequence unit, be it a complete regulatory element or single TF motif, is dependent on its context; a non-redundant function of an element revealed by its deletion (subtraction) was a poor predictor of sequence sufficiency when tested in isolation (addition). In Chapters 3 and 5, the predicted activity of the R2 enhancer was far less when studied in isolation compared to its native context within a cluster; when R2 was inserted into a null background (in the R2-only model), it contributed 15% of WT activity, whereas in the Δ R2 background in which the other elements are present, it contributed 50% of WT output. In this chapter, replacement of the R2 MCS with a single TF block removed R2 enhancer function, whereas removal of a single block caused functional impairments of up to 50% (Δ B).

The observation that no single TF motif preserved in the MCS in isolation constituted a functional enhancer is not surprising. By random chance, the Gata1 motif (GATAA) is predicted to occur once every $4^5=1024$ bp, or 62 times in the α -globin self-interacting domain (in reality it occurs 38 times in this region, as shown in Figure 5.20A-B). Clearly, not every incidence of the motif is stably bound and able to produce a functional output. In fact, data from the human ENCODE project suggest that 98% of TF binding events are restricted to within DHSs, despite an abundance of motifs throughout the genome (ENCODE Project Consortium, 2012). The same principle applies to all specific protein-DNA interactions, which are typically centred on motifs of 6-12 bp in length (Tugrul et al., 2015). It is well established that, instead, the combination of multiple sequence-directed binding events is key in producing a functional regulatory sequence (Levo and Segal, 2014;

Long et al., 2016). But, if no single TF binding site is sufficient to create an active enhancer, what are the constraints on sequence construction that make or break a functional element?

The R2 MCS appears to be a flexible platform for TF recruitment

Models for regulatory sequence composition typically lie on a spectrum between a highly constrained enhanceosome (Thanos and Maniatis, 1995) and a flexible billboard (Arnosti and Kulkarni, 2005; Levo and Segal, 2014). In a completely flexible model, binding of individual TFs occurs independently, meaning that the relative orientation, order and spacing of contributing motifs is not important for sequence function. On the other hand, in an enhanceosome model, the complete enhancer sequence is optimally constructed for protein-protein interactions such that it is intolerant to even minor modifications (Panne, 2008). It should be noted, however, that evolution of specific motif-motif grammar can occur not only to produce maximally active sequences, but also to fine tune sub-optimal enhancer activities when critical for gene dosage, such as during cell fate decisions (Farley et al., 2015). For higher order, multicellular organisms a modular system likely fits somewhere between the two models; co-operative protein-protein interactions can be either constrained by sequence grammar as exemplified by the emergence of dimerised TF motifs (Jolma et al., 2015), or flexible to stabilise binding and generate the required structural or molecular changes to promote the recruitment of downstream co-factors (Spitz and Furlong, 2012).

Several findings point to R2 being a flexible sequence (with few constraints on the sequence grammar). Existing conservation alignment data for the R2 MCS across a range of species show that regions between conserved TF motifs have diverged significantly through

evolution, but the predicted recruitment of trans-acting factors to identified motifs remains relatively constant (Figure 6.1; Philipsen and Hardison, 2018).

In this chapter, I tested the tolerance of the R2 MCS to disrupted sequence integrity experimentally. Multiple results corroborate the idea of R2 as a flexible binding platform. Firstly, in Screen #3, two alternative mutations of each Nf-e2 site of the R2 MCS were able to produce the same impact on R2 function as deletion of the entire corresponding blocks (Figure 6.17D-E). Thus, deletions of 32 bp (ΔD) and 26 bp (ΔE) in length had no more impact on R2 activity than individual substitutions of 3-5 bp. Similarly, specific mutation of the two Klf1 motifs found within block B in series (6 bp substituted in total) resulted in a phenotype comparable to that of the ΔB block deletion model (46 bp deletion; Figure 6.16B).

A direct test for the role of sequence grammar on function was in Screen #4, where no effect was observed when modifying the conserved distance between two motifs by up to 10 bp (Figure 6.19D). This indicates that constraints of spacing-dependent co-operative binding or helical phasing reported for other binding events (He et al., 2000; Panne, 2008; Sharon et al., 2012; Takahashi et al., 1986) are not important at this site. This finding is in line with genome-wide ChIP-seq data from human K562 cells where no consistent spacing was found between bound Gata1 and Nf-e2 motifs (Guo et al., 2012b).

It is interesting to note that, even in examples where motif-motif associations appear to be strongly selected for across a genome or through evolution, critical functional significance to the observed grammar is not always realised. For example, modification of an apparently coordinated motif structure for Dorsal/Twist TFs in *Drosophila* resulted in a modest effect on absolute transcription levels but did not alter tissue-specificity of activity (Arnosti et al.,

1996). Several studies have highlighted the pitfalls of looking genome-wide for grammar associations such as Bicoid/Hunchback motifs in *Drosophila* or Oct4/Sox2 in human, finding that it is possible for motif-motif associations to be enriched in the genome due to repeated integrations of a transposable element carrying this configuration of sequences or an evolutionary bias for deletion events to bring motifs into close proximity over time (Avsec et al., 2019; Lusk and Eisen, 2010). On the other hand, several motif-motif associations have been experimentally validated in large-scale transgenic experiments (Grossman et al., 2017; Sharon et al., 2012). In more targeted experiments in *Drosophila*, certain rearrangements of TF motifs in an enhancer for the Pax2 gene result in an altered pattern of activity from cone cells to photoreceptors by a closely positioned GFP reporter gene (Swanson et al., 2010) and inversion of a GATA site in the Cecropin regulatory region leads to severely attenuated expression by imaging of a downstream lacZ fusion gene (Senger et al., 2004). Therefore, it is possible for grammar rules to impact regulatory sequence function, but restricted interactions may not be as prevalent as suggested by conserved motif-motif spacing.

Here, results point towards the core of the R2 enhancer acting as a flexible landing pad for TF binding that collectively drives gene activation. But what are the requirements for factor recruitment in creating an active sequence?

Evidence for TF density as a driver of enhancer function

Examination of the wider α -globin locus for the distributions of the motifs discussed in this chapter shows an informative pattern; although motifs for each of the factors are identified throughout the region, there is specific enrichment of motifs within the R1 and R2 footprint

regions (Figure 6.21A-E). One might predict from this annotation that R1 and R2 would be the strongest functional elements of the α -globin super-enhancer cluster, in a similar manner to the observed differential binding of TFs between the elements by ChIP-seq (Figure 5.13). The trend seen in Figure 6.20A, in which the number of mutated TF motifs correlated with the impact of variants on R2 function, further suggests that the number of motifs within a regulatory element can contribute to its activity. The observation that R1 and R2 sequences both contain multiple motifs for the same TF is not unusual; homotypic clustering of motifs is a common feature of vertebrate enhancers and can be used as an indicator for element activity (Gotea et al., 2010). The recruitment of the same factor to multiple motifs within the same sequence region may contribute benefits of improved diffusion between sites (due to intramolecular vs intermolecular kinetics), co-operative binding and functional redundancy (Burz et al., 1998; Sharon et al., 2012; Weingarten-Gabbay and Segal, 2014).

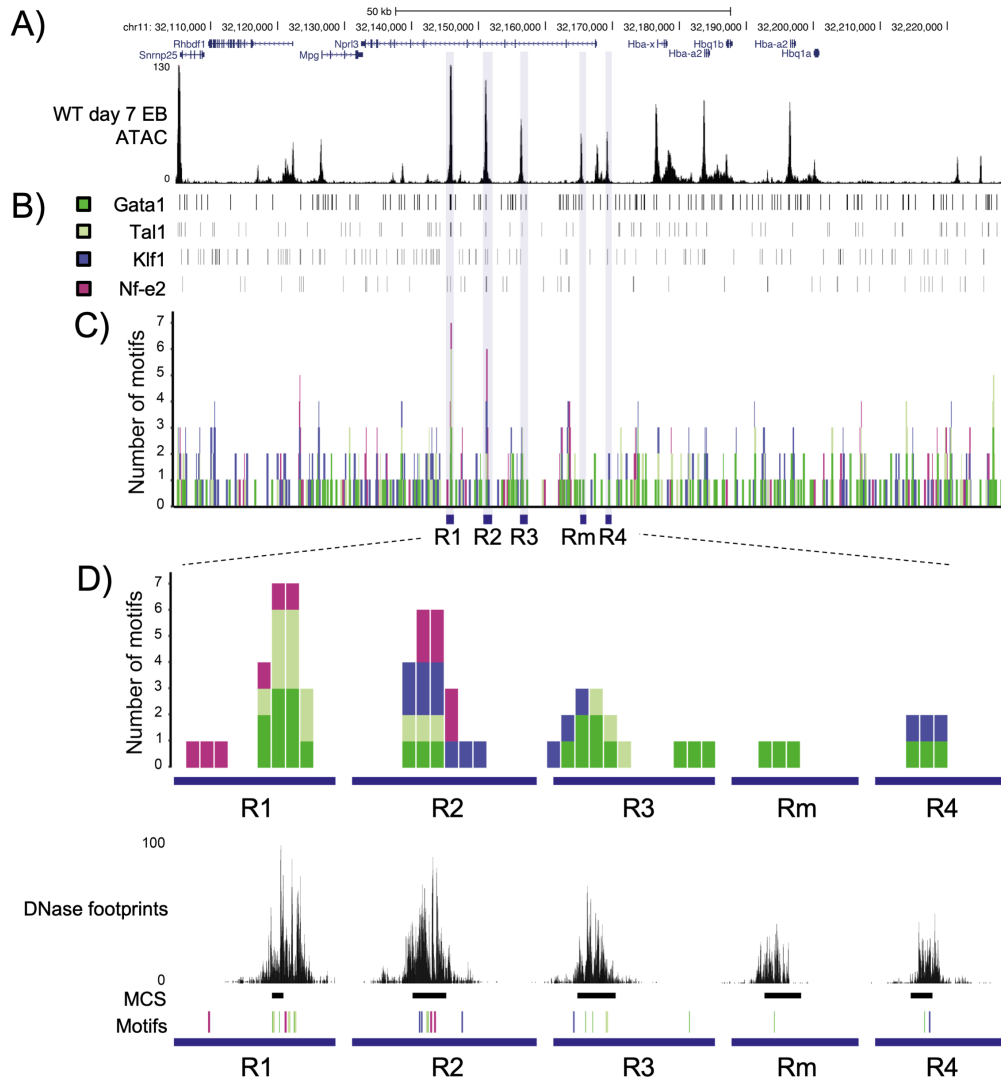


Figure 6.21: Erythroid-specific TF motifs are enriched at the R1 and R2 enhancers.

A) ATAC-seq data from bulk EB-derived erythrocytes, as shown in Figure 4.4. **B)** Positions of erythroid-specific TF motifs across the mouse α -globin locus, designated by vertical bars and aligned with the track in part A). Bars are coloured by the match score to provided position weight matrices by FIMO analysis (Grant et al., 2011); darker colours represent a better motif match. P-value cut-offs for each motif were selected to ensure the presence of at least one motif in each region at the locus known to be bound by ChIP-seq or, in the case of Gata1 (GATAA), for an exact motif match: $p=2e-03$ (Gata1), $p=2e-04$ (Tal1), $p=1.2e-04$ (Klf1) and $p=1e-04$ (Nf-e2). **C)** Counts for the motifs shown in part B) across 300 bp regions, staggered at 100 bp intervals. Bars are stacked and coloured for the corresponding motifs labelled in part B). **D)** Zoomed views of motif counts for elements of the α -globin super-enhancer. Blue horizontal bars under each cluster of motifs represent the extents of deletions used previously to remove the entire functional enhancer sequences, as labelled (R1/R2/R3/Rm/R4; Figure 3.6). **E)** DNase digital footprinting data for the regions shown in part D) from primary embryonic mouse erythrocytes (as shown in Figure 3.6), aligned with MCS sequences (Hughes et al., 2005) and motif matches and coloured as in part B).

The functional significance of individual TF binding is tackled in Screen #3; Gata1 motifs were found to be the most critical of those studied for R2 function. The Δ Gata1 model showed functional impairment comparable to the Δ TF model (where all predicted motifs were disrupted), whereas Δ Klf1 and Δ Nf-e2 variants retained partial R2 activity. Reciprocally, the functional output of the Gata1-only model was greater than for the Klf1-only and Nf-e2-only models. The variegated effect of TF motif disruption on R2 function indicates that the nature and activity of recruited TFs are key for enhancer function.

From the data acquired in the R2 enhancer screens, it appears that the number, functional contribution and diversity of TF motifs are important for R2 function. These findings, though limited to the dissection of a 238 bp sequence, may be informative when further probing the rules governing the function of an active enhancer sequence, such as spacing between each motif. As mentioned earlier in this chapter, binding site disruptions have not been functionally tested by ChIP-seq. To further understand the rules for creating a functional enhancer sequence beyond a specific element at a specific locus, it would be necessary to test a greater number of guided predictions. While this could be attempted empirically with rational designs based on existing results, a more effective way to detect patterns in an unbiased manner would be to employ machine learning methods. For example, models for enhancer function can be trained based on the positioning of motif occurrences or ChIP-seq peaks (Berman et al., 2002; Zinzen et al., 2009). Alternatively, proxies for function such as DNA accessibility or enhancer-promoter contacts can be trained using a range of characteristics to assign predictions based on sequence alone (Kelley et al., 2016; Schwessinger et al. 2019; Whalen et al., 2016; Zhou and Troyanskaya, 2015). Data from the variants screened here and future test sequences would provide valuable validation for any such model.

In addition to dissecting the endogenous regulatory code, there are conceptually interesting questions that could be further addressed with the synthetic enhancer approach described in this chapter. For example, would an R2 sequence of 6 Gata1 motifs be functional? Is it possible to enhance R2 activity further by adding more motifs? At what point does the activity of a single element become saturated? The combination of an unbiased sequence prediction tool trained with a wealth of genome-wide data and an *in vitro* experimental test platform presents an opportune way to answer these questions and to fully understand the sequence constraints of a short regulatory sequence in the R2 position.

Acknowledgements

I performed and analysed all experiments described in this chapter. Philip Hublitz (Genome Engineering Facility, Weatherall Institute of Molecular Medicine, University of Oxford) assisted with design and production of CRISPR/Cas9 targeting vectors. Kevin Clark (Flow Cytometry Facility, Weatherall Institute of Molecular Medicine, University of Oxford) performed FACS sorts.

Chapter 7 – Final conclusions

7.1 Summary of findings

In this thesis, I have attempted to disentangle the importance of context and sequence in the function of a single enhancer element within the α -globin super-enhancer. My key findings for each aspect are summarised below.

On enhancer context:

- The activity of the R2 enhancer element is attenuated when isolated from its super-enhancer context *in vivo* (Figure 3.12)
- Optimal function of the R1 and R2 enhancers is halved in the absence of the individually redundant R3/Rm/R4 elements (Figure 5.12)
- The native R2 position can be used as a test bed for transcription enhancing activity (Figure 6.10)

On enhancer sequence:

- The density of transcription factor motifs is an indicator of sequence activity across the mouse α -globin locus (Figure 6.21)
- Both the type and number of TF motifs appear to be important for R2 function (Figure 6.18; Figure 6.20)
- The precise grammar of sequences within R2 can be flexible (Figure 6.19)

In addition to these findings, I present a number of technical developments that will allow for further dissection of the α -globin regulatory landscape:

- Targeted integration of large-scale synthetic DNA sequences into engineered mESCs to test iterative designs of a complete regulatory domain (~100 kb) in a native chromosomal context (Synthetic RMGR; Figure 3.4-3.5)
- Scalable and robust isolation of homogeneously primitive erythrocytes from an *in vitro* differentiation protocol for WT or engineered mESCs (Figure 4.11)
- A high-throughput fluorescence reporter-based sequence screening method to test the activity of hundreds of rationally designed regulatory sequence variants *in situ*, in a single experiment (Figure 6.4)

7.2 Contribution to the understanding of genome regulation

The value in targeting endogenous loci at various resolutions

Non-coding regions of the mammalian genome have been previously described as “regulatory jungles” (Ruf et al., 2011). This description refers to the observation that genes at alternative positions within a small sequence window can show vastly different patterns of gene expression, depending on the relative contributions of various local *cis*-regulatory elements. The relative contributions of *cis*-regulatory elements depends on sequence compatibility but also on effects of physical availability, defined by distance, context and chromatin state within a regulatory domain (Splinter and de Laat, 2011). Over the past decade, much emphasis has been put on understanding the physical characteristics of active regions of the genome that affect the ability of elements to functionally interact. Ultimately, however, zooming into the fractal globule of genome structure, the interactions that set up formation of active structures are driven by the underlying DNA sequence. Therefore, to understand how active sequences are able to exert their function, with the awareness of such a complex physical system, it is critical to dissect the DNA code in its native context.

Prior to the work performed in this thesis, the testing of custom regulatory landscapes (as coined in Spitz et al., 2003 at the *Hoxd* locus) had been limited to studies using ectopic integration sites to test custom DNA constructs. Results show high levels of variation dependent on the site of integration and are subject to issues of improper copy number and phenotypic interference from endogenous genes when a stable integration site is used (discussed in Bender et al., 2000a; de Krom et al., 2002; Glaser et al., 2005). In parallel,

studies targeting endogenous loci have been limited by available genome engineering methods to the creation of mutants from a top-down approach. Through impressive perseverance, many groups have successfully created a wide range of models for individual gene loci (examples include Bahr et al., 2018; Hnisz et al., 2015; Rodriguez-Carballo et al., 2017), but the creation of new models has been dependent on the stepwise modification of existing mutants, making them susceptible to a growing burden of potential off-target effects and consequences of long-term cell culture. In my work, I removed these potential confounding factors by harnessing technical advances in large-scale DNA synthesis (in collaboration with Professor Jef Boeke) and genome engineering tools to target ~90 kb custom designs of the mouse α -globin regulatory domain to the native chromosomal context in purposefully designed mESCs.

At a higher sequence resolution, the targeted genetic dissection of individual elements has, thus far, been limited for similar reasons. Existing high-throughput analyses of regulatory sequences have generally been performed by juxtaposing test sequences with transcribed units for functional readout, in the context of transiently transfected plasmids or a random site for stable integration (reviewed in the introduction to Chapter 6). Therefore, reporter assays are able to ascertain the intrinsic activity of a sequence; whether this is exerted or not depends on the native chromatin context. Using the *in situ* R2 screening method described in this thesis, sequences can be tested not only for their ability to bind factors for gene activation but also for their ability to promote local chromatin opening and long-range gene activation, as is required for the vast majority of enhancer-like elements in the genome. The observation that the R1 and R4 α -globin enhancers perform as expected when integrated into the R2 position suggests that this system acts as a valid test for enhancer function, outperforming traditional plasmid-based assays for enhancer potential (Splinter and de Laat,

2011). The flexibility of the Cas9-directed HDR targeting method used in this procedure means that further screens could be designed to test sequences in alternative positions, allowing users to investigate the influence of molecular context on sequence function.

Variation in the roles of enhancer elements

Equivalence of *cis*-regulatory elements in terms of biochemical features does not signify equivalence of sequences at a functional level. For example, elements classified as strong or weak enhancers by chromatin features or conservation analyses activate a minimal promoter to a range of extents (Kwasnieski et al., 2014; Pennacchio et al., 2006). This either suggests that enhancers with similar molecular features can have a range of activities or that the primary function of some enhancers *in situ* may not be in direct activation of gene expression.

A key finding from Chapter 5 was that the strongest elements of the α -globin super-enhancer (R1 and R2) were not able to fully activate gene expression without the presence of the three downstream, seemingly non-functional elements (R3, Rm and R4). Analysis of the motif content and binding of a number of erythroid TFs to the α -globin locus *in vivo* suggests a functional segregation of these elements into two groups. This predicted segregation is corroborated by activity in experiments testing sequence contribution to α -globin gene expression and the effects of element deletions (Figure 7.1).

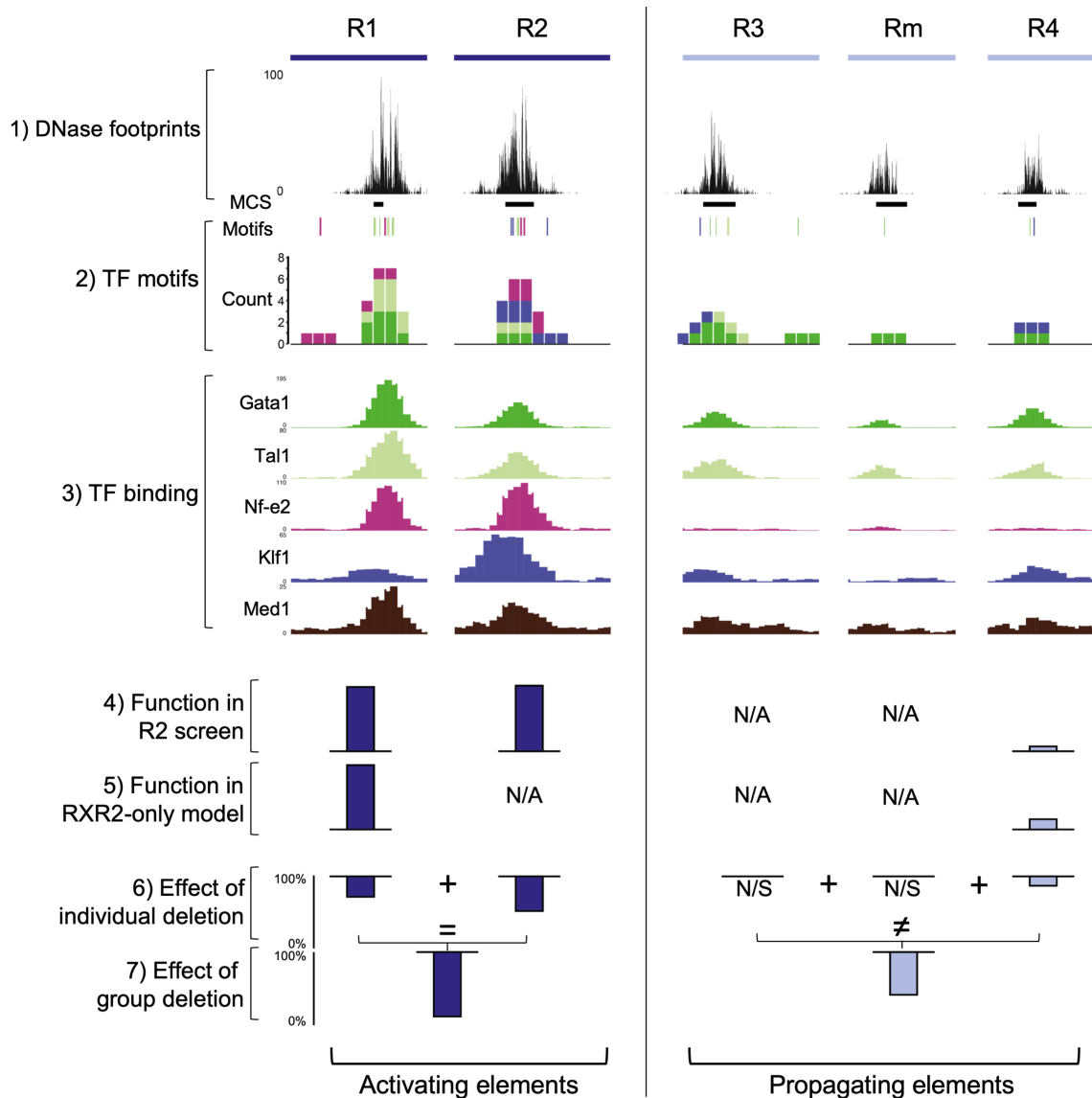


Figure 7.1: Proposed functional segregation of the elements of the mouse α -globin super-enhancer.

A summary of data in support of the classification of α -globin super-enhancer components into activating (R1/R2) and non-activating elements of unknown function (R3/Rm/R4). Data refer to the element labelled at the top of each column. Top to bottom: 1) DNase-seq digital footprinting and MCS annotations as shown in Figure 3.6. 2) Erythroid TF motifs aligned to footprinted regions, shown as motif maps and windowed counts over 300 bp regions (Figure 6.21). 3) Erythroid TF binding by ChIP-seq in primary mouse erythrocytes (Figure 5.13). 4) Relative activity of R1, R2 and R4 MCS regions in an α -YFP enhancer activity assay in the native R2 position. Bars are scaled to results from the R2 sequence (adapted from Figure 6.10A). 5) Relative activity of R1 and R4 sequences in the native R1 position when added to a hemizygous R2-only model. Bars are scaled to represent the added activity compared to the R2-only model (adapted from Figure 5.11). 6-7) The effects of single- and grouped-element deletions on α -globin expression. Bars represent the reduction in maximal expression in vivo except for the grouped R3/Rm/R4 deletion which was performed in the hemizygous in vitro system only (adapted from Figures 3.1

and 5.11). The effect of element deletions appears to be additive for the activating elements; the effect of deletion of individual propagating elements, on the other hand, appears to be minimal but their removal as a group has a greater effect. N/A, data not available; N/S, effect non-significant.

The appearance of weakly active, accessible elements near strong enhancers is a common feature of the genome. For example, creation of a single new binding site for the MYB TF at the human Tal1 gene locus is able to open additional local sequence regions to create a super-enhancer from normally neutral DNA (Mansour et al., 2014). Similarly, experiments from the Higgs and Hughes groups have shown that integration of the R2 element into a neutral part of the genome is sufficient to open several new peaks by ATAC-seq (Caroline Harrold, unpublished data). The secondary opening of primed sequences when in close proximity to active elements is suggested to be involved in the evolution of new enhancer functions and clusters (Long et al., 2016). The development of induced accessible regions into conserved elements near genes that are strongly expressed or critical for cell fate decisions may reflect an acquisition of sequence robustness.

In the case of the α -globin super-enhancer, the weaker R3, Rm and R4 elements show sequence conservation among a range of species and become accessible independently of the presence of R1 and R2, indicating that their presence has likely been selected as beneficial for effective gene expression. Thus, rather than being non-functional, the R3, Rm and R4 elements, while not conforming to the classical definition of enhancers, may represent an alternative type of *cis*-regulatory element. Together, I propose a model in which the R1 and R2 enhancers act as the primary transcriptional activators, while the R3, Rm and R4 elements act to propagate the signal to the α -globin genes. Mechanistically, this could be manifested through contributions to the local molecular environment of the self-interacting domain or through a mechanism involving the co-linearity of elements. A range of

chromatin assays could be used to test the former; impaired α -globin gene expression seen in an existing mouse model carrying an inversion of the enhancer cluster suggests at least partial contribution of the latter (Mira Kassouf, unpublished data).

The contributions of individual TFs to a single enhancer

Scanning the elements of the α -globin super-enhancer for erythroid TF motifs identified a trend for increased element activity with motif abundance. Functional screening of the R2 enhancer sequence identified that both the number and type of conserved motifs were important for maximal element function. In terms of the range of TFs recruited to the sequence, Gata1 motifs were identified as being most critical for enhancer activity. Removal of conserved sites for Klf1 and Nf-e2, on the other hand, affected the level of sequence activity but did not abolish R2 function. This suggests that individual TFs can have different roles in enhancer function. Beyond the selected TFs studied here, it is entirely possible (and indeed, likely) that more TFs are able to bind the R2 sequence in erythrocytes. A search for potential binders from the JASPAR mammalian motif database (Mathelier et al., 2016), for example, identifies over 50 matches to recognised consensus sequences by FIMO analysis ($p < 1e-4$; data not shown). Modelling of TF binding to the mouse R1 enhancer has previously predicted the possible binding of 11 factors simultaneously without steric hinderance (Maria Suci thesis, unpublished). Results from a tiled deletion screen will pinpoint any necessary motifs not yet identified; to exhaustively catalogue the wealth of binding events at the R2 sequence an unbiased proteomics approach would be needed (such as locus proximity labelling, Myers et al., 2018).

Of the factors tested here, Gata1 offers two modes of activity that may of importance. Firstly, it is thought that Gata1 is able to act as a “pioneer factor” to open inaccessible chromatin, possibly aided by the preceding binding of Gata2 (Anguita et al., 2004; May et al., 2013; Zaret and Carroll, 2011). In this way, the recruitment of Gata1 may be necessary for additional TFs to bind at the R2 sequence. Secondly, the association of Tal1 binding with Gata1 and observation that the predicted Tal1 motif within R2 is not necessary for binding suggests that Gata1 may be able to recruit additional factors to R2 via protein-protein interactions (Kassouf et al., 2010). At this point, these suggestions are purely speculative but, in future, assays such as ATAC-seq or ChIP-seq could be applied on selected R2 variant sequences to test the mechanistic consequences of different TF recruitment. In addition, the creation of synthetic enhancers from the bottom-up could test the need for additional, non-conserved binding events in the construction of the R2 sequence and identify the maximum allowed spacing between TF motifs to still be incorporated into the active enhancer region.

7.3 Next steps

The next steps in outlining the function of the super-enhancer as a whole will focus on interpreting the function of the R3, Rm and R4 elements. This can be approached from the angles of mechanistic study of existing models (such as imaging or 3C techniques) or the testing of novel locus constructs. What happens, for example, if the non-functional elements are replaced with highly active enhancer sequences? Or *vice versa*? Is the order of elements within their native positions important for super-enhancer function? Does the super-enhancer work if the elements are juxtaposed into a single composite enhancer? Or is a single point-to-point enhancer-promoter configuration inefficient for the physics of DNA looping? And so on.

From the perspective of the R2 enhancer specifically, the positional context within the cluster remains to be fully tested by integrating it specifically into the locations of each of the other native elements and vice versa. From a sequence perspective, the most valuable experiments will involve the *in silico* modelling of R2 sequence function and the mechanistic interpretation of new and existing sequences using various chromatin assays. The use of computational approaches will aid the generation of testable predictions for functional sequence construction beyond easily identifiable trends, while improvements on existing experimental methods (e.g. high-resolution TF binding analyses, He et al., 2015; Rhee and Pugh, 2012; single-molecule tracking, Liu and Tjian, 2018; single-cell chromatin assays, Wang et al., 2019) will allow for greater functional interpretation of the trends found by sequence screening.

Beyond the scope of the R1/R2/R3/Rm/R4 elements, many interesting questions remain for the regulation of the α -globin locus. Is it relevant, for example, that the position of the enhancers relative to the *Nprl3* gene is conserved through evolution (Philipsen and Hardison, 2018)? Would the super-enhancer-promoter hub form in the absence of flanking CTCF sites? How do mutations of the cluster affect the dynamics of domain formation and gene expression through erythropoiesis?

The tools developed and data reported in this thesis lay the groundwork for a wide range of future projects. It is my hope that the tools and principles explored here will be informative not only to the molecular understanding of the mouse α -globin locus, but to the fundamental question of how genes are regulated.

Appendix - Genetic models used in this thesis

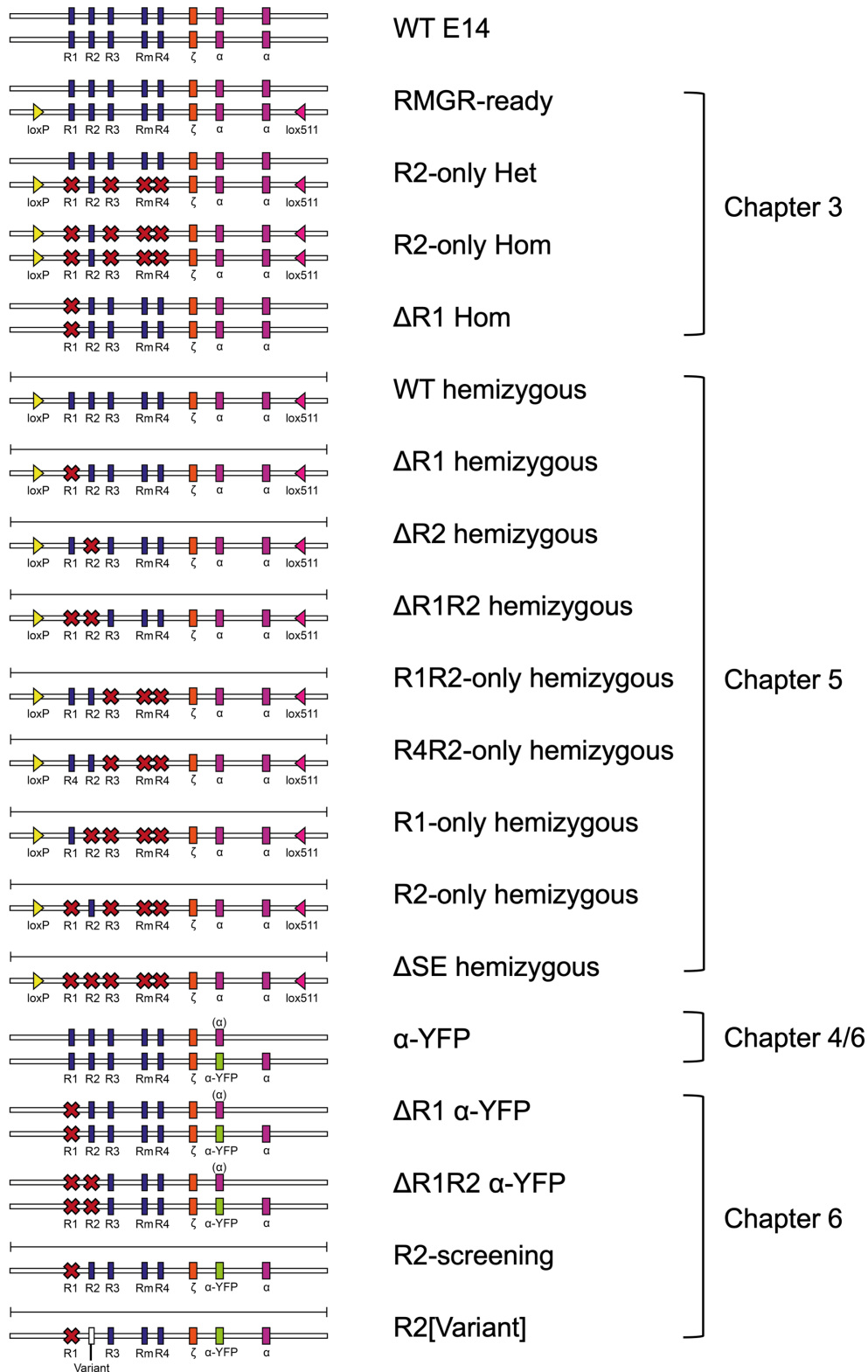


Figure A.1: Graphical summary of the genetic models used in this thesis.

Models are grouped by the chapter(s) in which they feature and annotated with key features. Deleted loci in hemizygous models are represented by horizontal bars.

References

- Alexander, J.M. et al. (2019). Live-cell imaging reveals enhancer-dependent Sox2 transcription in the absence of enhancer proximity. *eLife* 8, e41769.
- Allahyar, A., et al. (2018). Enhancer hubs and loop collisions identified from single-allele topologies. *Nature Genetics* 50, 1151–1160.
- ENCODE Project Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature* 489, 57–74.
- Andersson, R., et al. (2014). An atlas of active enhancers across human cell types and tissues. *Nature* 507, 455–461.
- Andersson, R., Sandelin, A., Danko, C.G. (2015). A unified architecture of transcriptional regulatory elements. *Trends in Genetics* 31, 426–433.
- Andrews, N.C., Erdjument-Bromage, H., Davidson, M.B., Tempst, P., Orkin, S.H. (1993). Erythroid transcription factor NF-E2 is a haematopoietic-specific basic-leucine zipper protein. *Nature* 362, 722–728.
- Anguita, E., et al. (2004). Globin gene activation during haemopoiesis is driven by protein complexes nucleated by GATA-1 and GATA-2. *Embo Journal* 23, 2841–2852.
- Arnold, C.D., et al. (2013). Genome-wide quantitative enhancer activity maps identified by STARR-seq. *Science* 339, 1074–1077.
- Arnold, C.D., et al. (2016). Genome-wide assessment of sequence-intrinsic enhancer responsiveness at single-base-pair resolution. *Nature Biotechnology* 35, 136–144.
- Aparicio, S., et al. (2002). Whole-genome shotgun assembly and analysis of the genome of *Fugu rubripes*. *Science (New York, N.Y.)* 297, 1301–1310.
- Arnosti, D.N., Barolo, S., Levine, M., Small, S. (1996). The eve stripe 2 enhancer employs multiple modes of transcriptional synergy. *Development* 122, 205.
- Arnosti, D.N., Kulkarni, M.M. (2005). Transcriptional enhancers: Intelligent enhanceosomes or flexible billboards? *J. Cell. Biochem.* 94, 890–898.
- Avsec, Ž., et al. (2019). Deep learning at base-resolution reveals motif syntax of the cis-regulatory code. *bioRxiv* 737981.
- Bahr, C., et al. (2018). A Myc enhancer cluster regulates normal and leukaemic haematopoietic stem cell hierarchies. *Nature* 553, 515–520.
- Bailey, T.L., Elkan, C. (1994). Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc Int Conf Intell Syst Mol Biol* 2, 28–36.

- Banani, S.F., Lee, H.O., Hyman, A.A., Rosen, M.K. (2017). Biomolecular condensates: organizers of cellular biochemistry. *Nature Reviews Molecular Cell Biology* 18, 285–298.
- Banerji, J., Rusconi, S., Schaffner, W. (1981). Expression of a beta-globin gene is enhanced by remote SV40 DNA sequences. *Cell* 27, 299–308.
- Barrington, C., et al. (2019). Enhancer accessibility and CTCF occupancy underlie asymmetric TAD architecture and cell type specific genome topology. *Nature Communications* 10, 2908.
- Bauer, D.E., et al. (2013). An erythroid enhancer of BCL11A subject to genetic variation determines fetal hemoglobin level. *Science (New York, N.Y.)* 342, 253–257.
- Benabdallah, N.S., et al. (2019). Decreased Enhancer-Promoter Proximity Accompanying Enhancer Activation. *Molecular Cell* 1–20.
- Bender, M.A., Bulger, M., Close, J., Groudine, M. (2000a). Beta-globin gene switching and DNase I sensitivity of the endogenous beta-globin locus in mice do not require the locus control region. *Molecular Cell* 5, 387–393.
- Bender, M.A., et al. (2000b). Independent formation of DnaseI hypersensitive sites in the murine beta-globin locus control region. *Blood* 95, 3600–3604.
- Bender, M.A., et al. (2001). Targeted deletion of 5′HS1 and 5′HS4 of the beta-globin locus control region reveals additive activity of the DNaseI hypersensitive sites. *Blood* 98, 2022–2027.
- Bender, M.A., et al. (2012). The hypersensitive sites of the murine β -globin locus control region act independently to affect nuclear localization and transcriptional elongation. *Blood* 119, 3820–3827.
- Berman, B.P., et al. (2002). Exploiting transcription factor binding site clustering to identify cis-regulatory modules involved in pattern formation in the Drosophila genome. *Proc. Natl. Acad. Sci. U.S.A.* 99, 757–762.
- Bernet, A., et al. (1995). Targeted inactivation of the major positive regulatory element (HS-40) of the human alpha-globin gene locus. *Blood* 86, 1202–1211.
- Berthelot, C., Villar, D., Horvath, J.E., Odom, D.T., Flicek, P. (2018). Complexity and conservation of regulatory landscapes underlie evolutionary resilience of mammalian gene expression. *Nat Ecol Evol* 2, 152–163.
- Blobel, G.A., et al. (2015). An international effort to cure a global health problem: A report on the 19th Hemoglobin Switching Conference. *Experimental Hematology* 43, 821–837.
- Boeke, J.D., et al. (2016). GENOME ENGINEERING. The Genome Project-Write. *Science* 353, 126–127.

- Boija, A., et al. (2018). Transcription Factors Activate Genes through the Phase-Separation Capacity of Their Activation Domains. *Cell* 175, 1842–1855.e16.
- Bouhassira, E.E., et al. (1997). Properties of the mouse alpha-globin HS-26: relationship to HS-40, the major enhancer of human alpha-globin gene expression. *Am. J. Hematol.* 54, 30–39.
- Branda, C.S., Dymecki, S.M. (2004). Talking about a Revolution: The Impact of Site-Specific Recombinases on Genetic Analyses in Mice. *Developmental Cell* 6, 7–28.
- Bratt-Leal, A.M., Carpenedo, R.L., McDevitt, T.C. (2009). Engineering the embryoid body microenvironment to direct embryonic stem cell differentiation. *Biotechnol. Prog.* 25, 43–51.
- Brown, J.M., et al. (2018). A tissue-specific self-interacting chromatin domain forms independently of enhancer-promoter interactions. *Nature Communications* 9, 3849–15.
- Buenrostro, J.D., Giresi, P.G., Zaba, L.C., Chang, H.Y., Greenleaf, W.J. (2013). Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nature Methods* 10, 1213–1218.
- Burz, D.S., Rivera-Pomar, R., Jackle, H., Hanes, S.D. (1998). Cooperative DNA-binding by Bicoid provides a mechanism for threshold-dependent gene activation in the *Drosophila* embryo. *Embo Journal* 17, 5998–6009.
- Calhoun, V.C., Stathopoulos, A., Levine, M. (2002). Promoter-proximal tethering elements regulate enhancer-promoter specificity in the *Drosophila Antennapedia* complex. *Proc Natl Acad Sci USA* 99, 9243.
- Calo, E., Wysocka, J. (2013). Modification of enhancer chromatin: what, how, and why? *Molecular Cell* 49, 825–837.
- Canver, M.C., et al. (2015). BCL11A enhancer dissection by Cas9-mediated in situ saturating mutagenesis. *Nature* 527, 192–197.
- Canzio, D., et al. (2019). Antisense lncRNA Transcription Mediates DNA Demethylation to Drive Stochastic Protocadherin alpha Promoter Choice. *Cell* 177, 639–653.e15.
- Capecchi, M.R. (2005). Gene targeting in mice: functional analysis of the mammalian genome for the twenty-first century. *Nature Reviews Genetics* 6, 507–512.
- Cho, W.-K., et al. (2018). Mediator and RNA polymerase II clusters associate in transcription-dependent condensates. *Science (New York, N.Y.)* 361, 412–415.
- Choi, Y.Y., et al. (2010). Controlled-size embryoid body formation in concave microwell arrays. *Biomaterials* 31, 4296–4303.
- Chong, S., et al. (2018). Imaging dynamic and selective low-complexity domain interactions that control gene transcription. *Science* 361, eaar2555.

- Coelho, A., Picanço, I., Seuanez, F., Seixas, M.T., Faustino, P. (2010). Novel large deletions in the human α -globin gene cluster: Clarifying the HS-40 long-range regulatory role in the native chromosome environment. *Blood Cells, Molecules, and Diseases* 45, 147–153.
- Cramer, P. (2019). Organization and regulation of gene transcription. *Nature* 573, 45–54.
- Creyghton, M.P., et al. (2010). Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proceedings of the National Academy of Sciences* 107, 21931–21936.
- CRISPR Therapeutics (2019). CRISPR Therapeutics and Vertex Announce Progress in Clinical Development Programs for the Investigational CRISPR/Cas9 Gene-Editing Therapy CTX001. URL <http://ir.crisprtx.com/news-releases/news-release-details/crispr-therapeutics-and-vertex-announce-progress-clinical> (accessed 15.9.19).
- Dang, S.M., Gerecht-Nir, S., Chen, J., Itskovitz-Eldor, J., Zandstra, P.W. (2004). Controlled, scalable embryonic stem cell differentiation culture. *Stem Cells* 22, 275–282.
- Darbellay, F., et al. (2019). The constrained architecture of mammalian Hox gene clusters. *Proc. Natl. Acad. Sci. U.S.A.* 116, 13424–13433.
- Davies, J.O.J., Oudelaar, A.M., Higgs, D.R., Hughes, J.R. (2017). How best to identify chromosomal interactions: a comparison of approaches. *Nature Methods* 14, 125–134.
- Davies, J.O.J., et al. (2016). Multiplexed analysis of chromosome conformation at vastly improved sensitivity. *Nature Methods* 13, 74–80.
- Davis, R., Gurumurthy, A., Hossain, M.A., Gunn, E.M., Bungert, J. (2019). Engineering Globin Gene Expression. *Mol Ther Methods Clin Dev* 12, 102–110.
- de Krom, M., van de Corput, M., Lindern, von, M., Grosveld, F., Strouboulis, J. (2002). Stochastic Patterns in Globin Gene Expression Are Established prior to Transcriptional Activation and Are Clonally Inherited. *Molecular Cell* 9, 1319–1326.
- De Santa, F., et al. (2010). A Large Fraction of Extragenic RNA Pol II Transcription Sites Overlap Enhancers. *PLoS Biol.* 8, e1000384.
- de Wit, E., de Laat, W. (2012). A decade of 3C technologies: insights into nuclear organization. *Genes & Development* 26, 11–24.
- Dekker, J., et al. (2017). The 4D nucleome project. *Nature* 549, 219–226.
- Dekker, J., Misteli, T. (2015). Long-Range Chromatin Interactions. *Cold Spring Harbor Perspectives in Biology* 7, a019356.
- Deng, W., et al. (2012). Controlling long-range genomic interactions at a native locus by targeted tethering of a looping factor. *Cell* 149, 1233–1244.

- di Iulio, et al. (2018). The human noncoding genome defined by genetic diversity. *Nature Genetics* *50*, 333–337.
- Dickel, D.E., et al. (2018). Ultraconserved Enhancers Are Required for Normal Development. *Cell* *172*, 491–499.e15.
- Dickel, D.E., et al. (2014). Function-based identification of mammalian enhancers using site-specific integration. *Nature Methods* *11*, 566–571.
- Diepstraten, S.T., Hart, A.H. (2019). Modelling human haemoglobin switching. *Blood Rev* *33*, 11–23.
- Ditadi, A., Sturgeon, C.M., Keller, G. (2016). A view of human haematopoietic development from the Petri dish. *Nature Reviews Molecular Cell Biology* *18*, 56–67.
- Doetschman, T.C., Eistetter, H., Katz, M., Schmidt, W., Kemler, R. (1985). The in vitro development of blastocyst-derived embryonic stem cell lines: formation of visceral yolk sac, blood islands and myocardium. *J Embryol Exp Morphol* *87*, 27–45.
- Dostie, J., et al. (2006). Chromosome Conformation Capture Carbon Copy (5C): a massively parallel solution for mapping interactions between genomic elements. *Genome Res.* *16*, 1299–1309.
- Dukler, N., Gulko, B., Huang, Y.-F., Siepel, A. (2016). Is a super-enhancer greater than the sum of its parts? *Nature Genetics* *49*, 2–3.
- Farley, E.K., et al. (2015). Suboptimization of developmental enhancers. *Science* *350*, 325–328.
- Faure, A.J., Schmiedel, J.M., Lehner, B. (2017). Systematic Analysis of the Determinants of Gene Expression Noise in Embryonic Stem Cells. *Cell Syst* *5*, 471–484.e4.
- Fernández-Nestosa, M.J., Hernández, P., Schwartzman, J.B., Krimer, D.B. (2008). PU.1 is dispensable to block erythroid differentiation in Friend erythroleukemia cells. *Leuk. Res.* *32*, 121–130.
- Ferreira, R., Ohneda, K., Yamamoto, M., Philipsen, S. (2005). GATA1 Function, a Paradigm for Transcription Factors in Hematopoiesis. *Mol. Cell. Biol.* *25*, 1215.
- Findlay, G.M., Boyle, E.A., Hause, R.J., Klein, J.C., Shendure, J. (2014). Saturation editing of genomic regions by multiplex homology-directed repair. *Nature* *513*, 120–123.
- Finn, E.H., et al. (2019). Extensive Heterogeneity and Intrinsic Variation in Spatial Genome Organization. *Cell* *176*, 1502–1515.e10.
- Flamier, A., Singh, S., Rasmussen, T.P. (2017). A standardized human embryoid body platform for the detection and analysis of teratogens. *PLOS ONE* *12*, e0171101.

- Frankel, N., et al. (2010). Phenotypic robustness conferred by apparently redundant transcriptional enhancers. *Nature* *466*, 490–493.
- Fraser, S.T., Isern, J., Baron, M.H. (2007). Maturation and enucleation of primitive erythroblasts during mouse embryogenesis is accompanied by changes in cell-surface antigen expression. *Blood* *109*, 343–352.
- Fudenberg, G., et al. (2016). Formation of Chromosomal Domains by Loop Extrusion. *Cell Reports* *15*, 2038–2049.
- Fujiwara, Y., Browne, C.P., Cunniff, K., Goff, S.C., Orkin, S.H. (1996). Arrested development of embryonic red cell precursors in mouse embryos lacking transcription factor GATA-1. *Proc Natl Acad Sci USA* *93*, 12355.
- Furlong, E.E.M., Levine, M. (2018). Developmental enhancers and chromosome topology. *Science* *361*, 1341–1345.
- Garrison, E., Marth, G. (2012). Haplotype-based variant detection from short-read sequencing. *arXiv*.
- Gasiorek, J.J., Blank, V. (2015). Regulation and function of the NFE2 transcription factor in hematopoietic and non-hematopoietic cells. *Cellular and Molecular Life Sciences* *72*, 2323–2335.
- Ghavi-Helm, et al. (2014). Enhancer loops appear stable during development and are associated with paused polymerase. *Nature* *512*, 96–100.
- Ghirlando, R., Felsenfeld, G. (2016). CTCF: making the right connections. *Genes & Development* *30*, 881–891.
- Gisselbrecht, S.S., et al. (2013). Highly parallel assays of tissue-specific enhancers in whole *Drosophila* embryos. *Nature Methods* *10*, 774–80.
- Glaser, S., Anastassiadis, K., Stewart, A.F. (2005). Current issues in mouse genome engineering. *Nature Genetics* *37*, 1187–1193.
- Gotea, V., et al. (2010). Homotypic clusters of transcription factor binding sites are a key component of human promoters and enhancers. *Genome Res.* *20*, 565–577.
- Gourdon, G., Sharpe, J.A., Higgs, D.R., Wood, W.G. (1995). The mouse alpha-globin locus regulatory element. *Blood* *86*, 766–775.
- Grant, C.E., Bailey, T.L., Noble, W.S. (2011). FIMO: scanning for occurrences of a given motif. *bioinformatics* *27*, 1017–1018.
- Gregory, T.R. (2005). Synergy between sequence and size in large-scale genomics. *Nature Reviews Genetics* *6*, 699–708.

- Grossman, S.R., et al. (2017). Systematic dissection of genomic features determining transcription factor binding and enhancer function. *Proc. Natl. Acad. Sci. U.S.A.* *114*, E1291–E1300.
- Grosveld, F., van Assendelft, G.B., Greaves, D.R., Kollias, G. (1987). Position-independent, high-level expression of the human beta-globin gene in transgenic mice. *Cell* *51*, 975–985.
- Guo, Ya, et al. (2012a). CTCF/cohesin-mediated DNA looping is required for protocadherin α promoter choice. *Proceedings of the National Academy of Sciences* *109*, 21081–21086.
- Guo, Ya, et al. (2015). CRISPR Inversion of CTCF Sites Alters Genome Topology and Enhancer/Promoter Function. *Cell* *162*, 900–910.
- Guo, Yuchun, Mahony, S., Gifford, D.K. (2012b). High resolution genome wide binding event finding and motif discovery reveals transcription factor spatial binding constraints. *PLoS Comput. Biol.* *8*, e1002638.
- Haberle, V., et al. (2019). Transcriptional cofactors display specificity for distinct types of core promoters. *Nature* *570*, 122–126.
- Han, G.C., et al. (2015). Genome-wide organization of GATA1 and TAL1 determined at high resolution. *Molecular and Cellular Biology* *36*, 157–52.
- Hanssen, L.L., et al. (2017). Tissue-specific CTCF–cohesin-mediated chromatin architecture delimits enhancer interactions and function in vivo. *Nature Cell Biology* *13*, 74.
- Hay, D., et al. (2016). Genetic dissection of the α -globin super-enhancer in vivo. *Nature Genetics* *48*, 895–903.
- He, Q., Johnston, J., Zeitlinger, J. (2015). ChIP-nexus enables improved detection of in vivo transcription factor binding footprints. *Nature Biotechnology* *33*, 395–401.
- He, X., Hohn, T., Futterer, J. (2000). Transcriptional activation of the rice tungro bacilliform virus gene is critically dependent on an activator element located immediately upstream of the TATA box. *The Journal of Biological Chemistry* *275*, 11799–11808.
- Heintzman, N.D., et al. (2007). Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nature Genetics* *39*, 311–318.
- Heinz, S., Romanoski, C.E., Benner, C., Glass, C.K. (2015). The selection and function of cell type-specific enhancers. *Nature Reviews Molecular Cell Biology* *16*, 144–154.
- Henriques, T., et al. (2018). Widespread transcriptional pausing and elongation control at enhancers. *Genes & Development* *32*, 26–41.

- Hesselberth, J.R., et al. (2009). Global mapping of protein-DNA interactions in vivo by digital genomic footprinting. *Nature Methods* 6, 283–289.
- Higgs, D.R. (2001). The Molecular Basis of α -Thalassemia. In: Disorders of Hemoglobin. Cambridge University Press.
- Higgs, D.R. (2013). The molecular basis of α -thalassemia. *Cold Spring Harbor Perspectives in Medicine* 3, a011718–a011718.
- Higgs, D.R., et al. (2006). How transcriptional and epigenetic programmes are played out on an individual mammalian gene cluster during lineage commitment and differentiation. *Biochemical Society Symposium* 73, 11–22.
- Higgs, D.R., Wood, W.G. (2008). Long-range regulation of alpha globin gene expression during erythropoiesis. *Current Opinion in Hematology* 15, 176–183.
- Higgs, D.R., et al. (1990). A major positive regulatory region located far upstream of the human alpha-globin gene locus. *Genes & Development* 4, 1588–1601.
- Hillier, L.W., et al. (2005). Genomics in *C. elegans*: so many genes, such a little worm. *Genome Res.* 15, 1651–1660.
- Hnisz, D., et al. (2013). Super-enhancers in the control of cell identity and disease. *Cell* 155, 934–947.
- Hnisz, D., et al. (2015). Convergence of Developmental and Oncogenic Signaling Pathways at Transcriptional Super-Enhancers. *Molecular Cell* 58, 362–370.
- Hnisz, D., Shrinivas, K., Young, R.A., Chakraborty, A.K., Sharp, P.A. (2017). A Phase Separation Model for Transcriptional Control. *Cell* 169, 13–23.
- Hornstein, B.D., Roman, D., Arévalo-Soliz, L.M., Engevik, M.A., Zechiedrich, L. (2016). Effects of Circular DNA Length on Transfection Efficiency by Electroporation into HeLa Cells. *PLOS ONE* 11, e0167537.
- Horsley, S.W., et al. (2001). Monosomy for the most telomeric, gene-rich region of the short arm of human chromosome 16 causes minimal phenotypic effects. *European Journal of Human Genetics* 9, 217–225.
- Huang, J., et al. (2018). Dissecting super-enhancer hierarchy based on chromatin interactions. *Nature Communications* 9, 943.
- Huang, J., et al. (2016). Dynamic Control of Enhancer Repertoires Drives Lineage and Stage-Specific Transcription during Hematopoiesis. *Developmental Cell* 36, 9–23.
- Huang, Q., et al. (2012). Distance and Helical Phase Dependence of Synergistic Transcription Activation in cis-Regulatory Module. *PLOS ONE* 7, e31198.

- Hughes, J.R., et al. (2005). Annotation of cis-regulatory elements by identification, subclassification, and functional assessment of multispecies conserved sequences. *Proc. Natl. Acad. Sci. U.S.A.* *102*, 9830–9835.
- Hughes, J.R., et al. (2013). High-resolution analysis of cis-acting regulatory networks at the α -globin locus. *Philosophical Transactions of the Royal Society B: Biological Sciences* *368*, 20120361–20120361.
- Hughes, J.R., et al. (2014). Analysis of hundreds of cis-regulatory landscapes at high resolution in a single, high-throughput experiment. *Nature Genetics* *46*, 205–212.
- Hughes, R.A., Ellington, A.D. (2017). Synthetic DNA Synthesis and Assembly: Putting the Synthetic in Synthetic Biology. *Cold Spring Harbor Perspectives in Biology* *9*, a023812.
- Igarashi, K., et al. (1994). Regulation of transcription by dimerization of erythroid factor NF-E2 p45 with small Maf proteins. *Nature* *367*, 568–572.
- Illumina (2019). Illumina sequencing platforms: September 2019. URL <https://www.illumina.com/systems/sequencing-platforms.html> (accessed 15.9.19).
- International Human Genome Sequencing Consortium (2004). Finishing the euchromatic sequence of the human genome. *Nature* *431*, 931–945.
- Jaeger, S.A., et al. (2010). Conservation and regulatory associations of a wide affinity range of mouse transcription factor binding sites. *Genomics* *95*, 185–195.
- Johnson, D.S., Mortazavi, A., Myers, R.M., Wold, B. (2007). Genome-wide mapping of in vivo protein-DNA interactions. *Science (New York, N.Y.)* *316*, 1497–1502.
- Jolma, A., et al. (2015). DNA-dependent formation of transcription factor pairs alters their binding specificity. *Nature* *527*, 384–388.
- Junion, G., et al. (2012). A transcription factor collective defines cardiac cell fate and reflects lineage history. *Cell* *148*, 473–486.
- Kagey, M.H., et al. (2010). Mediator and cohesin connect gene expression and chromatin architecture. *Nature* *467*, 430–435.
- Kassouf, M.T., Chagraoui, H., Vyas, P., Porcher, C. (2008). Differential use of SCL/TAL-1 DNA-binding domain in developmental hematopoiesis. *Blood* *112*, 1056–1067.
- Kassouf, M.T., et al. (2010). Genome-wide identification of TAL1's functional targets: insights into its mechanisms of action in primary erythroid cells. *Genome Res.* *20*, 1064–1083.
- Keller, G. (2005). Embryonic stem cell differentiation: emergence of a new era in biology and medicine. *Genes & Development* *19*, 1129–1155.

- Keller, G., Kennedy, M., Papayannopoulou, T., Wiles, M.V. (1993). Hematopoietic commitment during embryonic stem cell differentiation in culture. *Molecular and Cellular Biology* *13*, 473–486.
- Kelley, D.R., Snoek, J., Rinn, J.L. (2016). Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks. *Genome Res.* *26*, 990–999.
- Kent, W.J., et al. (2002). The human genome browser at UCSC. *Genome Res.* *12*, 996–1006.
- Kim, J.H., et al. (2011). High Cleavage Efficiency of a 2A Peptide Derived from Porcine Teschovirus-1 in Human Cell Lines, Zebrafish and Mice. *PLOS ONE* *6*, e18556.
- Kim, T.-K., et al. (2010). Widespread transcription at neuronal activity-regulated enhancers. *Nature* *465*, 182–187.
- Kingsley P.D., et al. (2012). Ontogeny of erythroid gene expression. *eBlood* *121*, 5–13.
- Kinney, J.B., Murugan, A., Callan, C.G.J., Cox, E.C. (2010). Using deep sequencing to characterize the biophysical mechanism of a transcriptional regulatory sequence. *Proceedings of the National Academy of Sciences* *107*, 9158–9163.
- Kitzman, J.O. (2016). Haplotypes drop by drop. *Nature Biotechnology* *34*, 296–298.
- Koch, F., et al. (2011). Transcription initiation platforms and GTF recruitment at tissue-specific enhancers and promoters. *Nature Publishing Group* *18*, 956–963.
- Korlach, J et al. (2010). Real-time DNA sequencing from single polymerase molecules. *Meth. Enzymol.* *472*, 431–455.
- Kosuri, S., et al. (2013). Composability of regulatory sequences controlling transcription and translation in *Escherichia coli* *Proc Natl Acad Sci USA* *110*, 14024.
- Kotkow, K.J., Orkin, S.H. (1996). Complexity of the erythroid transcription factor NF-E2 as revealed by gene targeting of the mouse p18 NF-E2 locus. *Proceedings of the National Academy of Sciences* *93*, 3514–3518.
- Kowalczyk, M.S., et al. (2012). Intragenic Enhancers Act as Alternative Promoters. *Molecular Cell* *45*, 447–458.
- Kraft, K., et al. (2019). Serial genomic inversions induce tissue-specific architectural stripes, gene misexpression and congenital malformations. *Nature Cell Biology* *21*, 305–310.
- Kredel, S., et al. (2009). mRuby, a Bright Monomeric Red Fluorescent Protein for Labeling of Subcellular Structures. *PLOS ONE* *4*, e4391.

- Kreiss, P., et al. (1999). Plasmid DNA size does not affect the physicochemical properties of lipoplexes but modulates gene transfer efficiency. *Nucleic Acids Research* 27, 3792–3798.
- Kurosawa, H. (2007). Methods for inducing embryoid body formation: in vitro differentiation system of embryonic stem cells. *J. Biosci. Bioeng.* 103, 389–398.
- Kvon, E.Z., et al. (2014). Genome-scale functional characterization of *Drosophila* developmental enhancers in vivo. *Nature* 512, 91–95.
- Kwasnieski, J.C., Fiore, C., Chaudhari, H.G., Cohen, B.A. (2014). High-throughput functional testing of ENCODE segmentation predictions. *Genome Res.* 24, 1595–1602.
- Lam, M.T.Y., Li, W., Rosenfeld, M.G., Glass, C.K. (2014). Enhancer RNAs and regulated transcriptional programs. *Trends Biochem. Sci.* 39, 170–182.
- Langmead, B., Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods* 9, 357–359.
- Laurent, J.M., et al. (2019). Big DNA as a tool to dissect an age-related macular degeneration-associated haplotype. *Precis Clin Med* 2, 1–7.
- Lee, E.-C., et al. (2014). Complete humanization of the mouse immunoglobulin loci enables efficient therapeutic antibody discovery. *Nature Biotechnology* 32, 356–363.
- Lee, M.Y., et al. (2011). High density cultures of embryoid bodies enhanced cardiac differentiation of murine embryonic stem cells. *Biochem. Biophys. Res. Commun.* 416, 51–57.
- Lemons, D., McGinnis, W. (2006). Genomic Evolution of Hox Gene Clusters. *Science* 313, 1918.
- Lesueur, L.L., Mir, L.M., André, F.M. (2016). Overcoming the Specific Toxicity of Large Plasmids Electrotransfer in Primary Cells In Vitro. *Mol Ther Nucleic Acids* 5, e291.
- Levo, M., Segal, E. (2014). In pursuit of design principles of regulatory sequences. *Nature Reviews Genetics* 15, 453–468.
- Li, H. (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *bioinformatics* 27, 2987–2993.
- Li, H., et al. (2009). The Sequence Alignment/Map format and SAMtools. *bioinformatics* 25, 2078–2079.
- Li, K., Wang, G., Andersen, T., Zhou, P., Pu, W.T. (2014). Optimization of genome engineering approaches with the CRISPR/Cas9 system. *PLOS ONE* 9, e105779.

- Liu, Z., Tjian, R. (2018). Visualizing transcription factor dynamics in living cells. *J. Cell Biol.* *217*, 1181–1191.
- Long, H.K., Prescott, S.L., Wysocka, J. (2016). Ever-Changing Landscapes: Transcriptional Enhancers in Development and Evolution. *Cell* *167*, 1170–1187.
- Lopes, R., Korkmaz, G., Agami, R. (2016). Applying CRISPR-Cas9 tools to identify and characterize transcriptional enhancers. *Nature Reviews Molecular Cell Biology* *17*, 597–604.
- Lower, K.M., et al. (2009). Adventitious changes in long-range gene expression caused by polymorphic structural variation and promoter competition. *Proc. Natl. Acad. Sci. U.S.A.* *106*, 21771–21776.
- Loyd, M.R., Okamoto, Y., Randall, M.S., Ney, P.A. (2003). Role of AP1/NFE2 binding sites in endogenous alpha-globin gene transcription. *Blood* *102*, 4223–4228.
- Lupiáñez, D.G., et al. (2015). Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. *Cell* *161*, 1012–1025.
- Lusk, R.W., Eisen, M.B. (2010). Evolutionary Mirages: Selection on Binding Site Composition Creates the Illusion of Conserved Grammars in *Drosophila* Enhancers. *PLoS genetics* *6*, e1000829.
- Maeda, R.K., Karch, F. (2011). Gene expression in time and space: additive vs hierarchical organization of cis-regulatory regions. *Chromosomes and expression mechanisms* *21*, 187–193.
- Mamaeva, S.E., Tsvileneva, N.N. (1985). A study of chromosome content of Friend virus-induced mouse erythroleukemia cells (clone M2) via karyotype reconstruction. *Cancer Genet. Cytogenet.* *16*, 199–205.
- Maniatis, T., Fritsch, E.F., Sambrook, J. (1982). *Molecular Cloning. A Laboratory Manual*. Cold Spring Harbor Laboratory, New York.
- Mansour, M.R., et al. (2014). Oncogene regulation. An oncogenic super-enhancer formed through somatic mutation of a noncoding intergenic element. *Science (New York, N.Y.)* *346*, 1373–1377.
- Mantsoki, A., Devailly, G., Joshi, A. (2016). Gene expression variability in mammalian embryonic stem cells using single cell RNA-seq data. *Comput Biol Chem* *63*, 52–61.
- Masuda, T., et al. (2016). Transcription factors LRF and BCL11A independently repress expression of fetal hemoglobin. *Science* *351*, 285–289.
- Mateo, L.J., et al. (2019). Visualizing DNA folding and RNA in embryos at single-cell resolution. *Nature* *568*, 49–54.

- Matharu, N., Ahituv, N. (2015). Minor Loops in Major Folds: Enhancer–Promoter Looping, Chromatin Restructuring, and Their Association with Transcriptional Regulation and Disease. *PLoS genetics* *11*, e1005640.
- Mathelier, A., et al. (2016). JASPAR 2016: a major expansion and update of the open-access database of transcription factor binding profiles. *Nucleic Acids Research* *44*, D110–5.
- May, G., et al. (2013). Dynamic analysis of gene expression and genome-wide transcription factor binding during lineage specification of multipotent progenitors. *Stem Cell* *13*, 754–768.
- McColl, B., et al. (2014). An in vivo model for analysis of developmental erythropoiesis and globin gene regulation. *FASEB J.* *28*, 2306–2317.
- McLenachan, S., Sarsero, J.P., Ioannou, P.A. (2007). Flow-cytometric analysis of mouse embryonic stem cell lipofection using small and large DNA constructs. *Genomics* *89*, 708–720.
- Meers, M.P., Janssens, D.H., Henikoff, S. (2019). Pioneer Factor-Nucleosome Binding Events during Differentiation Are Motif Encoded. *Molecular Cell* *75*, 562–575.e5.
- Meier-Kolthoff, J.P., et al. (2014). Complete genome sequence of DSM 30083(T), the type strain (U5/41(T)) of *Escherichia coli*, and a proposal for delineating subspecies in microbial taxonomy. *Stand Genomic Sci* *9*, 2–2.
- Melnikov, A., et al. (2012). Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. *Nature Biotechnology* *30*, 271–277.
- Messana, J.M., Hwang, N.S., Coburn, J., Elisseeff, J.H., Zhang, Z. (2008). Size of the embryoid body influences chondrogenesis of mouse embryonic stem cells. *J Tissue Eng Regen Med* *2*, 499–506.
- Miller, D.A., Tantravahi, R., Miller, O.J. (2002). Karyotype of Friend Virus-induced Mouse Erythroleukemia Cells. *Cancer Genet Cytogenet.* *16*, 199-205.
- Mitchell, L.A., Mcculloch, L.H., Boeke, J.D. (2019). De novo assembly, delivery and expression of a 101 kb human gene in mouse cells. *BioRxiv* 423426.
- Montalbano, A., Canver, M.C., Sanjana, N.E. (2017). High-Throughput Approaches to Pinpoint Function within the Noncoding Genome. *Molecular Cell* *68*, 44–59.
- Montavon, T., et al. (2011). A regulatory archipelago controls Hox genes transcription in digits. *Cell* *147*, 1132–1145.
- Moorthy, S.D., et al. (2017). Enhancers and super-enhancers have an equivalent regulatory role in embryonic stem cells through regulation of single or multiple genes. *Genome Res.* *27*, 246–258.

- Morse, H.C.I. (2007). Building a Better Mouse: One Hundred Years of Genetics and Biology. In: Building a Better Mouse: One Hundred Years of Genetics and Biology. pp. 1–11.
- Mouse ENCODE Consortium (2012). An encyclopedia of mouse DNA elements (Mouse ENCODE). *Genome Biol.* *13*, 418–5.
- Myers, S.A., et al. (2018). Discovery of proteins associated with a predefined genomic locus via dCas9-APEX-mediated proximity labeling. *Nature Methods* *15*, 437–439.
- Nagai, T., et al. (2002). A variant of yellow fluorescent protein with fast and efficient maturation for cell-biological applications. *Nature Biotechnology* *20*, 87–90.
- Natoli, G., Andrau, J.-C. (2012). Noncoding transcription at enhancers: general principles and functional models. *Annual Review of Genetics* *46*, 1–19.
- Ng, E.S., Davis, R.P., Azzola, L., Stanley, E.G., Elefanty, A.G. (2005). Forced aggregation of defined numbers of human embryonic stem cells into embryoid bodies fosters robust, reproducible hematopoietic differentiation. *Blood* *106*, 1601–1603.
- 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.
- Niebruegge, S., et al. (2009). Generation of human embryonic stem cell-derived mesoderm and cardiac cells using size-specified aggregates in an oxygen-controlled bioreactor. *Biotechnol. Bioeng.* *102*, 493–507.
- Nobrega, M.A., Ovcharenko, I., Afzal, V., Rubin, E.M. (2003). Scanning human gene deserts for long-range enhancers. *Science (New York, N.Y.)* *302*, 413.
- 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*, E6697–E6706.
- Nussbaum, L., et al. (2018). High-Throughput Genotyping of CRISPR/Cas Edited Cells in 96-Well Plates. *Methods Protoc* *1*, 29.
- Osborne, C.S., et al. (2004). Active genes dynamically colocalize to shared sites of ongoing transcription. *Nature Genetics* *36*, 1065–1071.
- Osoegawa, K., et al. (2000). Bacterial artificial chromosome libraries for mouse sequencing and functional analysis. *Genome Res.* *10*, 116–128.
- Osterwalder, M., et al. (2018). Enhancer redundancy provides phenotypic robustness in mammalian development. *Nature* *554*, 239–243.
- Oudelaar, A.M., et al. (2019). Dissection of the 4D chromatin structure of the α -globin locus through *in vivo* erythroid differentiation with extreme spatial and temporal resolution. *bioRxiv* 763763.

- Oudelaar, A.M., et al. (2018). Single-allele chromatin interactions identify regulatory hubs in dynamic compartmentalized domains. *Nature Genetics* *50*, 1744–1751.
- Oudelaar, A.M., et al. (2017). Between form and function: the complexity of genome folding. *Human Molecular Genetics* *26*, R208–R215.
- Owens, D.D.G., et al. (2019). Microhomologies are prevalent at Cas9-induced larger deletions. *Nucleic Acids Research* *47*, 7402–7417.
- Palis, J., Robertson, S., Kennedy, M., Wall, C., Keller, G. (1999). Development of erythroid and myeloid progenitors in the yolk sac and embryo proper of the mouse. *Development* *126*, 5073–5084.
- Panne, D. (2008). The enhanceosome. *Curr. Opin. Struct. Biol.* *18*, 236–242.
- Panne, D., Maniatis, T., Harrison, S.C. (2007). An atomic model of the interferon-beta enhanceosome. *Cell* *129*, 1111–1123.
- Papadopoulos, P., et al. (2012). A dual reporter mouse model of the human β -globin locus: applications and limitations. *PLOS ONE* *7*, e51272.
- Papatsenko, D., Levine, M. (2007). A rationale for the enhanceosome and other evolutionarily constrained enhancers. *Current Biology* *17*, R955–R957.
- Parker, M.P., Peterson, K.R. (2018). Mouse Models of Erythropoiesis and Associated Diseases. *Methods Mol. Biol.* *1698*, 37–65.
- Parker, S.C.J., et al. (2013). Chromatin stretch enhancer states drive cell-specific gene regulation and harbor human disease risk variants. *Proc Natl Acad Sci USA* *110*, 17921.
- Patwardhan, R.P., et al. (2012). Massively parallel functional dissection of mammalian enhancers in vivo. *Nature Biotechnology* *30*, 265–270.
- Pennacchio, L.A., et al. (2006). In vivo enhancer analysis of human conserved non-coding sequences. *Nature* *444*, 499–502.
- Pennacchio, L.A., Bickmore, W., Dean, A., Nobrega, M.A., Bejerano, G. (2013). Enhancers: five essential questions. *Nature Reviews Genetics* *14*, 288–295.
- Perkins, A., et al. (2016). Krüppeling erythropoiesis: an unexpected broad spectrum of human red blood cell disorders due to KLF1 variants. *Blood* *127*, 1856–1862.
- Perkins, A.C., Sharpe, A.H., Orkin, S.H. (1995). Lethal beta-thalassaemia in mice lacking the erythroid CACCC-transcription factor EKLF. *Nature* *375*, 318–322.
- Perry, M.W., Boettiger, A.N., Bothma, J.P., Levine, M. (2010). Shadow enhancers foster robustness of *Drosophila* gastrulation. *Current Biology* *20*, 1562–1567.

- Peterson, K.R., et al. (2014). A cell-based high-throughput screen for novel chemical inducers of fetal hemoglobin for treatment of hemoglobinopathies. *PLOS ONE* 9, e107006.
- Peterson, K.R., Stamatoyannopoulos, G. (1993). Role of gene order in developmental control of human gamma- and beta-globin gene expression. *Molecular and Cellular Biology* 13, 4836–4843.
- Pfeiffer, F., et al. (2018). Systematic evaluation of error rates and causes in short samples in next-generation sequencing. *Sci Rep* 8, 10950–14.
- Philipsen, S., Hardison, R.C. (2018). Evolution of hemoglobin loci and their regulatory elements. *Blood Cells, Molecules, and Diseases* 70, 2–12.
- Pimkin, M., et al. (2014). Divergent functions of hematopoietic transcription factors in lineage priming and differentiation during erythro-megakaryopoiesis. *Genome Res.* 24, 1932–1944.
- Pirrotta, V., Li, H.-B. (2012). A view of nuclear Polycomb bodies. *Chromosomes and expression mechanisms* 22, 101–109.
- Plessy, C., Dickmeis, T., Chalmel, F., Strahle, U. (2005). Enhancer sequence conservation between vertebrates is favoured in developmental regulator genes. *Trends in Genetics* 21, 207–210.
- Pop, R., et al. (2010). A key commitment step in erythropoiesis is synchronized with the cell cycle clock through mutual inhibition between PU.1 and S-phase progression. *PLoS Biol.* 8, e1000484.
- Pott, S., Lieb, J.D. (2015). What are super-enhancers? *Nature Genetics* 47, 8–12.
- Preisler, H.D., Shiraishi, Y., Mori, M., Sandberg, A.A. (1976). Clones of Friend leukemia cells: differences in karyotypes and responsiveness to inducers of differentiation. *Cell Differ.* 5, 207–216.
- Qiu, P., et al. (2004). Mutation detection using Surveyor nuclease. *BioTechniques* 36, 702–707.
- Quinlan, A.R. (2014). BEDTools: The Swiss-Army Tool for Genome Feature Analysis. *Curr Protoc Bioinformatics* 47, 11.12.1–34.
- Ragoczy, T., Bender, M.A., Telling, A., Byron, R., Groudine, M. (2006). The locus control region is required for association of the murine beta-globin locus with engaged transcription factories during erythroid maturation. *Genes & Development* 20, 1447–1457.
- Ramírez, F., et al. (2016). deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Research* 44, W160–5.

- Ramos, A.I., Barolo, S. (2013). Low-affinity transcription factor binding sites shape morphogen responses and enhancer evolution. *Philosophical Transactions of the Royal Society B: Biological Sciences* 368, 20130018.
- Rhee, H.S., Pugh, B.F. (2012). ChIP-exo method for identifying genomic location of DNA-binding proteins with near-single-nucleotide accuracy. *Curr Protoc Mol Biol* Chapter 21, Unit 21.24.
- Richardson, S.M., et al. (2017). Design of a synthetic yeast genome. *Science* 355, 1040–1044.
- Rodriguez-Carballo, E., et al. (2017). The HoxD cluster is a dynamic and resilient TAD boundary controlling the segregation of antagonistic regulatory landscapes. *Genes & Development* 31, 2264–2281.
- Rubio, E.D., et al. (2008). CTCF physically links cohesin to chromatin. *Proc. Natl. Acad. Sci. U.S.A.* 105, 8309–8314.
- Ruf, S., et al. (2011). Large-scale analysis of the regulatory architecture of the mouse genome with a transposon-associated sensor. *Nature Genetics* 43, 379–386.
- Ryu, J., Kim, H., Yang, D., Lee, A.J., Jung, I. (2019). A new class of constitutively active super-enhancers is associated with fast recovery of 3D chromatin loops. *BMC Bioinformatics* 20, 127.
- Sanborn, A.L., et al. (2015). Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes. *Proc. Natl. Acad. Sci. U.S.A.* 112, E6456–65.
- Sankaran, V.G., Xu, J., Orkin, S.H. (2010). Advances in the understanding of haemoglobin switching. *British Journal of Haematology* 149, 181–194.
- Schaft, J., Ashery-Padan, R., van der Hoeven, F., Gruss, P., Stewart, A.F. (2001). Efficient FLP recombination in mouse ES cells and oocytes. *Genesis* 31, 6–10.
- Schindelin, J., et al. (2012). Fiji: an open-source platform for biological-image analysis. *Nature Methods* 9, 676–682.
- Schmitt, R.M., Bruyns, E., Snodgrass, H.R. (1991). Hematopoietic development of embryonic stem cells in vitro: cytokine and receptor gene expression. *Genes & Development* 5, 728–740.
- Schoenfelder, S., Fraser, P. (2019). Long-range enhancer-promoter contacts in gene expression control. *Nature Reviews Genetics* 20, 437–455.
- Schwarzer, W., et al. (2017). Two independent modes of chromatin organization revealed by cohesin removal. *Nature* 551, 51–56.
- Schwessinger, R., et al. (2017). Sasquatch: predicting the impact of regulatory SNPs on transcription factor binding from cell- and tissue-specific DNase footprints. *Genome Res.* 27, 1730–1742.

- Schwessinger, R. et al. (2019). DeepC: Predicting chromatin interactions using megabase scaled deep neural networks and transfer learning. *BioRxiv* 724005.
- Senger, K., et al. (2004). Immunity regulatory DNAs share common organizational features in *Drosophila*. *Molecular Cell* *13*, 19–32.
- Shaner, N.C., Steinbach, P.A., Tsien, R.Y. (2005). A guide to choosing fluorescent proteins. *Nature Methods* *2*, 905–909.
- Sharon, E., et al. (2012). Inferring gene regulatory logic from high-throughput measurements of thousands of systematically designed promoters. *Nature Biotechnology* *30*, 521–530.
- Shen, Y., et al. (2012). A map of the cis-regulatory sequences in the mouse genome. *Nature* *488*, 116–120.
- Shin, H.Y., et al. (2016). Hierarchy within the mammary STAT5-driven Wap super-enhancer. *Nature Genetics* *48*, 904–911.
- Shivdasani, R.A., Mayer, E.L., Orkin, S.H. (1995). Absence of blood formation in mice lacking the T-cell leukaemia oncoprotein tal-1/SCL. *Nature* *373*, 432–434.
- Shivdasani, R.A., Orkin, S.H. (1995). Erythropoiesis and globin gene expression in mice lacking the transcription factor NF-E2. *Proc Natl Acad Sci USA* *92*, 8690.
- Shlyueva, D., Stampfel, G., Stark, A. (2014). Transcriptional enhancers: from properties to genome-wide predictions. *Nature Reviews Genetics* *15*, 272–286.
- Siepel, A., et al. (2005). Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* *15*, 1034–1050.
- Skarpidi, E., Vassilopoulos, G., Stamatoyannopoulos, G., Li, Q. (1998). Comparison of expression of human globin genes transferred into mouse erythroleukemia cells and in transgenic mice. *Blood* *92*, 3416–3421.
- Slattery, M., et al. (2011). Cofactor binding evokes latent differences in DNA binding specificity between Hox proteins. *Cell* *147*, 1270–1282.
- Slattery, M., et al. (2014). Absence of a simple code: how transcription factors read the genome. *Trends Biochem. Sci.* *39*, 381–399.
- Smith, A.G. (1991). Culture and differentiation of embryonic stem cells. *Journal of tissue culture methods* *13*, 89–94.
- 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.

- Smith, R.P., et al. (2013). Massively parallel decoding of mammalian regulatory sequences supports a flexible organizational model. *Nature Genetics* 45, 1021–1028.
- Smithies, O., Gregg, R.G., Boggs, S.S., Koralewski, M.A., Kucherlapati, R.S. (1985). Insertion of DNA sequences into the human chromosomal beta-globin locus by homologous recombination. *Nature* 317, 230–234.
- Socolovsky, M., et al. (2001). Ineffective erythropoiesis in Stat5a(-/-)5b(-/-) mice due to decreased survival of early erythroblasts. *Blood* 98, 3261–3273.
- Sollaino, M.C., et al. (2010). Homozygous deletion of the major alpha-globin regulatory element (MCS-R2) responsible for a severe case of hemoglobin H disease. *Blood* 116, 2193–2194.
- Spangler, A., Su, E.Y., Craft, A.M., Cahan, P. (2018). A single cell transcriptional portrait of embryoid body differentiation and comparison to progenitors of the developing embryo. *Stem Cell Res* 31, 201–215.
- Spitz, F., Furlong, E.E.M. (2012). Transcription factors: from enhancer binding to developmental control. *Nature Reviews Genetics* 13, 613–626.
- Spitz, F., Gonzalez, F., Duboule, D. (2003). A global control region defines a chromosomal regulatory landscape containing the HoxD cluster. *Cell* 113, 405–417.
- Spivak, J.L., Toretti, D., Dickerman, H.W. (1973). Effect of phenylhydrazine-induced hemolytic anemia on nuclear RNA polymerase activity of the mouse spleen. *Blood* 42, 257–266.
- Splinter, E., de Laat, W. (2011). The complex transcription regulatory landscape of our genome: control in three dimensions. *Embo Journal* 30, 4345–4355.
- Stefanska, M., et al. (2017). Primitive erythrocytes are generated from hemogenic endothelial cells. *Sci Rep* 7, 6401.
- Steiner, L.A., et al. (2010). Dynamic CO-Localization of GATA1, NFE2, and EKLF and Changes in Gene Expression During Hematopoiesis. *Blood* 116, 741.
- Stergachis, A.B., et al. (2014). Conservation of trans-acting circuitry during mammalian regulatory evolution. *Nature* 515, 365–370.
- Strom, A.R., et al. (2017). Phase separation drives heterochromatin domain formation. *Nature* 547, 241–245.
- Stumpf, M., et al. (2006). The mediator complex functions as a coactivator for GATA-1 in erythropoiesis via subunit Med1/TRAP220. *Proceedings of the National Academy of Sciences* 103, 18504–18509.
- Swanson, C.I., Evans, N.C., Barolo, S. (2010). Structural rules and complex regulatory circuitry constrain expression of a Notch- and EGFR-regulated eye enhancer. *Developmental Cell* 18, 359–370.

- Takahashi, K., et al. (1986). Requirement of stereospecific alignments for initiation from the simian virus 40 early promoter. *Nature* *319*, 121–126.
- Tallack, M.R., et al. (2010). A global role for KLF1 in erythropoiesis revealed by ChIP-seq in primary erythroid cells. *Genome Res.* *20*, 1052–1063.
- R Core Team (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL <http://www.R-project.org/> (accessed 15.9.19)
- Telenius, J., Hughes, J.R. (2018). NGseqBasic - a single-command UNIX tool for ATAC-seq, DNaseI-seq, Cut-and-Run, and ChIP-seq data mapping, high-resolution visualisation, and quality control. *bioRxiv* 393413.
- Testa, G., et al. (2003). Engineering the mouse genome with bacterial artificial chromosomes to create multipurpose alleles. *Nature Biotechnology* *21*, 443–447.
- Thanos, D., Maniatis, T. (1995). Virus induction of human IFN β gene expression requires the assembly of an enhanceosome. *Cell* *83*, 1091–1100.
- Therizols, P., et al. (2014). Chromatin decondensation is sufficient to alter nuclear organization in embryonic stem cells. *Science* *346*, 1238.
- Thurman, R.E., et al. (2012). The accessible chromatin landscape of the human genome. *Nature* *489*, 75–82.
- Tomishima, M.J., Hadjantonakis, A.-K., Gong, S., Studer, L. (2006). Production of Green Fluorescent Protein Transgenic Embryonic Stem Cells Using the GENSAT Bacterial Artificial Chromosome Library. *Stem Cells* *25*, 39–45.
- Tugrul, M., Paixao, T., Barton, N.H., Tkacik, G. (2015). Dynamics of Transcription Factor Binding Site Evolution. *PLoS genetics* *11*, e1005639.
- Valamehr, B., et al. (2008). Hydrophobic surfaces for enhanced differentiation of embryonic stem cell-derived embryoid bodies. *Proc. Natl. Acad. Sci. U.S.A.* *105*, 14459–14464.
- van de Werken, H.J.G., et al. (2012). Robust 4C-seq data analysis to screen for regulatory DNA interactions. *Nature Methods* *9*, 969–972.
- van Steensel, B., Furlong, E.E.M. (2019). The role of transcription in shaping the spatial organization of the genome. *Nature Reviews Molecular Cell Biology* *20*, 290.
- Vermunt, M.W., Zhang, D., Blobel, G.A. (2019). The interdependence of gene-regulatory elements and the 3D genome. *J. Cell Biol.* *218*, 12–26.
- Vernimmen, D., De Gobbi, M., Sloane-Stanley, J.A., Wood, W.G., Higgs, D.R. (2007). Long-range chromosomal interactions regulate the timing of the transition between poised and active gene expression. *Embo Journal* *26*, 2041–2051.

- Vernimmen, D., et al. (2011). Polycomb eviction as a new distant enhancer function. *Genes & Development* 25, 1583–1588.
- Vian, L., et al. (2018). The Energetics and Physiological Impact of Cohesin Extrusion. *Cell* 175, 292–294.
- Vierstra, J., et al. (2015). Functional footprinting of regulatory DNA. *Nature Methods* 12, 927–930.
- Vietri-Rudan, M., Hadjur, S. (2015). Genetic Tailors: CTCF and Cohesin Shape the Genome During Evolution. *Trends in Genetics* 31, 651–660.
- Vinjamur, D.S., Bauer, D.E., Orkin, S.H. (2018). Recent progress in understanding and manipulating haemoglobin switching for the haemoglobinopathies. *British Journal of Haematology* 180, 630–643.
- Viprakasit, V., et al. (2006). A novel deletion causing alpha thalassemia clarifies the importance of the major human alpha globin regulatory element. *Blood* 107, 3811–3812.
- Visel, A., et al. (2009). ChIP-seq accurately predicts tissue-specific activity of enhancers. *Nature* 457, 854–858.
- Visel, A., et al. (2008). Ultraconservation identifies a small subset of extremely constrained developmental enhancers. *Nature Genetics* 40, 158–160.
- Wallace, H.A.C., et al. (2007). Manipulating the Mouse Genome to Engineer Precise Functional Syntenic Replacements with Human Sequence. *Cell* 128, 197–209.
- Wang, B., et al. (2015). Highly efficient CRISPR/HDR-mediated knock-in for mouse embryonic stem cells and zygotes. *BioTechniques* 59, 201–2–204–206–8.
- Wang, Q., et al. (2019). CoBATCH for High-Throughput Single-Cell Epigenomic Profiling. *Molecular Cell* 76, 206–216.
- Weingarten-Gabbay, S., Segal, E. (2014). The grammar of transcriptional regulation. *Hum. Genet.* 133, 701–711.
- Weintraub, A.S., et al. (2017). YY1 Is a Structural Regulator of Enhancer-Promoter Loops. *Cell* 171, 1573–1588.e28.
- Wetterstrand, K.A. (2019). DNA Sequencing Costs: Data. URL <https://www.genome.gov/about-genomics/fact-sheets/DNA-Sequencing-Costs-Data> (accessed 15.9.19).
- Whalen, S., Truty, R.M., Pollard, K.S. (2016). Enhancer-promoter interactions are encoded by complex genomic signatures on looping chromatin. *Nature Genetics* 48, 488–496.

- Whyte, W.A., et al. (2013). Master transcription factors and mediator establish super-enhancers at key cell identity genes. *Cell* *153*, 307–319.
- Wickham, H. (2009). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag, New York.
- Wijchers, P.J., et al. (2016). Cause and Consequence of Tethering a SubTAD to Different Nuclear Compartments. *Molecular Cell* *61*, 461–473.
- Wiles, M.V., Keller, G. (1991). Multiple hematopoietic lineages develop from embryonic stem (ES) cells in culture. *Development* *111*, 259–267.
- Wilson, J.L., et al. (2014). Single-cell analysis of embryoid body heterogeneity using microfluidic trapping array. *Biomed Microdevices* *16*, 79–90.
- Woolfe, A., et al. (2004). Highly Conserved Non-Coding Sequences Are Associated with Vertebrate Development. *PLoS Biol.* *3*, e7.
- Wutz, A. (2014). Haploid Mouse Embryonic Stem Cells: Rapid Genetic Screening and Germline Transmission. *Annu. Rev. Cell Dev. Biol.* *30*, 705–722.
- Xi, W., Beer, M.A. (2018). Local epigenomic state cannot discriminate interacting and non-interacting enhancer-promoter pairs with high accuracy. *PLoS Comput. Biol.* *14*, e1006625.
- Yanez-Cuna, J.O., Dinh, H.Q., Kvon, E.Z., Shlyueva, D., Stark, A. (2012). Uncovering cis-regulatory sequence requirements for context-specific transcription factor binding. *Genome Res.* *22*, 2018–2030.
- Yang, Y., Seed, B. (2003). Site-specific gene targeting in mouse embryonic stem cells with intact bacterial artificial chromosomes. *Nature Biotechnology* *21*, 447–451.
- Yue, F., et al. (2014). A comparative encyclopedia of DNA elements in the mouse genome. *Nature* *515*, 355–364.
- Zabidi, M.A., et al. (2015). Enhancer-core-promoter specificity separates developmental and housekeeping gene regulation. *Nature* *518*, 556–559.
- Zaret, K.S., Carroll, J.S. (2011). Pioneer transcription factors: establishing competence for gene expression. *Genes & Development* *25*, 2227–2241.
- Zhang, J.-P., et al. (2017). Efficient precise knockin with a double cut HDR donor after CRISPR/Cas9-mediated double-stranded DNA cleavage. *Genome Biol.* *18*, 35.
- Zhang, Y., Buchholz, F., Muirers, J.P., Stewart, A.F. (1998). A new logic for DNA engineering using recombination in *Escherichia coli*. *Nature Genetics* *20*, 123–128.
- Zhang, Y., et al. (2008). Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* *9*, R137.

- Zhou, J., Troyanskaya, O.G. (2015). Predicting effects of noncoding variants with deep learning-based sequence model. *Nature Methods* *12*, 931–934.
- Zhu, J., et al. (2013). Genome-wide chromatin state transitions associated with developmental and environmental cues. *Cell* *152*, 642–654.
- Zijlstra, W.G., Buursma, A., Meeuwssen-van der Roest, W.P. (1991). Absorption spectra of human fetal and adult oxyhemoglobin, de-oxyhemoglobin, carboxyhemoglobin, and methemoglobin. *Clin Chem* *37*, 1633–1638.
- Zinzen, R.P., Girardot, C., Gagneur, J., Braun, M., Furlong, E.E.M. (2009). Combinatorial binding predicts spatio-temporal cis-regulatory activity. *Nature* *462*, 65–70.
- Zuin, J., et al. (2014). A cohesin-independent role for NIPBL at promoters provides insights in CdLS. *PLoS genetics* *10*, e1004153.