

The role of distal regulatory elements in *α -globin* transcription

Deborah Hay
Balliol College

MRC Molecular Haematology Unit
Weatherall Institute of Molecular Medicine
University of Oxford

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

Hilary Term 2014

Abstract: The role of regulatory elements in alpha globin transcription

Understanding the mechanisms of transcriptional control by *cis*-acting regulatory elements is one of the major goals in current molecular biology. Estimates of enhancer numbers in mammalian genomes extend to hundreds of thousands, outnumbering genes by an order of magnitude or more. Yet current high-throughput methods for enhancer identification give only indirect insights into their function, and are unable to demonstrate heterogeneity, redundancy, or co-operativity between regulatory elements.

The *α-globin* locus has been extensively characterized with regard to transcriptional output, chromatin marks and chromosome conformation throughout erythroid maturation, and offers an excellent model for the investigation of enhancer function. Such analysis at this locus has previously been undertaken in heterologous transgenic models, including a 'humanized' mouse bearing the entire human *α-globin* cluster.

Through the investigation of a rare naturally occurring mutation in a patient with thalassaemia, we demonstrate that heterologous transgenic models do not accurately recapitulate normal transcriptional control at this locus. This allows us to revise the prevailing view of human *α-globin* transcription control, in which a reliance on a single regulatory element (HS-40, or R2) was assumed.

An analysis of the transcriptional contribution of the five regulatory elements at this locus (R1-4 and R[m]) has therefore been undertaken, using knockout mouse models. Assessment of the haematological impact of each regulatory element knockout demonstrates that none is indispensable for *α-globin* transcription; however, analysis of double knockouts proves functional non-equivalence of the elements, with R1 and R2 being critically required for viability.

Quantification of nascent transcripts from each model shows a significantly different contribution to *α-globin* transcription from each of the five putative enhancers, with two providing no transcriptional contribution despite chromatin marks and transcription factor binding profiles suggestive of enhancer activity. R1 and R2 are defined as the minor and major enhancers for *α-globin*; their deletion in concert also shows an unexpected impact on nascent transcription more widely across chromosome 11, challenging current views on enhancer-promoter specificity.

This locus-specific study highlights heterogeneity in regulatory element biology that would be undetectable in current genome-wide studies. The models generated provide a platform for the further investigation of general principles of enhancer function, and will also permit us to examine the regulation of *α-globin* transcription in non-erythroid tissues.

Acknowledgements

I am very grateful to Doug Higgs for giving me the opportunity to join his lab, and for his extraordinary kindness and encouragement. It is a privilege to work under his supervision.

All members of the Higgs and Gibbons labs have helped me with their expertise, support and good humour – particularly Jim Hughes, whose advice and guidance have been invaluable.

Andrew Smith has been consistently generous with his time and great expertise, and I feel very fortunate to have worked under his expert tuition.

Jackie Sharpe, Jackie Sloane Stanley and Sue Butler have been closely involved throughout the project; as well as their vital role in generating and managing the transgenic mice, they have provided much general guidance and support.

It has also been a pleasure to collaborate with Paula Faustino and Andreia Coelho in the investigation of a case of thalassaemia that opened the way to the rest of the project.

Finally, I thank my husband, Chris, to whom the thesis is dedicated, and to whom I owe more than I could possibly repay.

Abbreviations

µg	Microgram
µl	Microlitre
3C	Chromosome conformation capture
4-TU	4- thiouridine
APRT	Adenine phosphoribosyl transferase
ATRX	Alpha thalassaemia/mental retardation syndrome, X-linked
BAC	Bacterial artificial chromosome
BCB	Brilliant Cresyl Blue
BSA	Bovine serum albumin
CD	Cluster of differentiation
ChIA-PET	Chromatin Interaction Analysis by Paired-End Tag Sequencing
ChIP	Chromatin immunoprecipitation
Cre	Cyclic recombinase
CRISPR	Clustered Regularly Interspaced Short Palindromic Repeats
CTCF	CCCTC-binding factor
dCTP	Deoxycytidine triphosphate
DEPC	Diethylpyrocarbonate
DIG	Digoxigenin
DMEM	Dulbecco's modified Eagle's medium
DMSO	Dimethyl sulfoxide
DNA	Deoxyribonucleic acid
DNase	deoxyribonuclease
dNTP	Deoxyribonucleotide triphosphate
dpc	Days post coitum
DTT	Dithiothreitol
EDTA	Ethylenediaminetetraacetic
ENCODE	Encyclopedia of DNA Elements
Epo	Erythropoietin
ESC	Embryonic stem cell
FACS	Fluorescence-activated cell sorting
FCS	Fetal calf serum
FIAU	1-(2'-deoxy-2'-fluoro-1-β-D-arabinofuranosyl)-5-iodouracil
FISH	Fluorescence in situ hybridization
G418	Geneticin
Gb	Gigabase
GMEM	Glasgow Minimum Essential Medium
GTP	Guanosine-5'-triphosphate
GWAS	Genome wide association study
h	Hours
Hb	Haemoglobin
HbH	Haemoglobin H
Het.	Heterozygote/heterozygous
HMBA	hexamethylene bis-acetamide

Hom.	Homozygote/homozygous
HPDP	Hexyl-3'-(2'-pyridyldithio)propionamide
IMDM	Iscoe's Modified Dulbecco's Medium
Kb	Kilobase
KO	Knockout
LB	Luria broth
LCR	Locus control region
Mb	Megabase
MCH	Mean cell haemoglobin
MCS	Multi-species conserved sequence
MCV	Mean cell volume
MEL	Mouse erythroleukaemia (cell)
MHU	Molecular haematology unit
Min	Minute
ml	Millilitre
MLPA	Multiplex ligation-dependent probe amplification
mM	Millimol
mRNA	Messenger RNA
NCBI	National Center for Biotechnology Information
PBS	Phosphate buffered saline
PCR	Polymerase chain reaction
PE	Phycoerythrin
PolII	RNA polymerase II
qPCR	Quantitative Polymerase chain reaction
RNA	Ribonucleic acid
RPM	Revolutions per minute
RPMI	Roswell Park Memorial Institute medium
SCF	Stem cell factor
SD	Standard deviation
SDS	Sodium dodecyl-sulfate
SND	Smallest natural deletion
SNP	Single nucleotide polymorphism
SOC	Super Optimal broth with catabolite repression
SSC	Saline sodium citrate
TAD	Transcription activation domain
TAE	Tris-acetate-EDTA
TCR	T cell receptor
TE	Tris-EDTA
TPO	Thrombopoietin
U	Units
UCSC	University of California Santa Cruz
UV	Ultraviolet
V	Volts
WIMM	Weatherall Institute of Molecular Medicine
WT	Wild type
YAC	Yeast artificial chromosome

Table of Contents

1. Introduction	2
1.1 <i>The regulation of eukaryotic gene expression</i>	2
1.2 <i>Defining and Identifying Enhancers</i>	5
1.2.1 The original approach to definition and identification of enhancers	5
1.2.2 Improving on the enhancer assay?	10
1.3 <i>From identification to function</i>	18
1.3.1 How to establish if an enhancer is active	21
1.3.2 Enhancer/promoter interactions and their specificity	22
1.3.3. Function in isolation vs orchestrated activity	27
1.3.4 Are all these elements equal?	31
1.3.5 What do enhancers do to influence gene expression?	35
1.4 <i>The clinical significance of regulatory elements</i>	37
1.5 <i>Unanswered questions</i>	39
1.6 <i>Regulatory elements at the α-globin locus</i>	41
1.7 <i>Outline of the Thesis</i>	43
2. Materials and Methods	46
2.1 <i>DNA procedures</i>	46
2.1.1 DNA extraction	46
2.1.2 Plasmid DNA extraction	47
2.2 <i>RNA procedures</i>	47
2.2.1 RNA extraction	47
2.2.2 DNase treatment	48
2.2.3 cDNA synthesis	48
2.2.4 Metabolic labelling and isolation of nascent RNA	49
2.2.5 RNA quality control	50
2.3 <i>Cloning techniques and bacterial culture</i>	50
2.3.1 Assembly of targeting constructs	50
2.3.2. Cloning into shuttle vectors	51
2.3.3 Transformation of competent cells	51
2.3.4 Cultures and archives of E Coli	51
2.3.5 Gel extraction of DNA fragments	52
2.3.6 DNA ligation	52
2.3.7 Validation of targeting constructs	52
2.4 <i>Handling mammalian cells</i>	52
2.4.1 Primary erythroid culture (Fibach culture)	52
2.4.2 Mouse embryonic stem cell targeting	53
2.4.3 Karyotyping of targeted ES cells	54
2.4.4 In vitro Cre recombination in targeted ES cells	54
2.4.5 Fetal liver cultures	55

2.4.6 Megakaryocyte expansion from fetal livers.....	55
2.5 <i>Animal handling and obtaining samples for analysis</i>	56
2.5.1 Blastocyst injection and chimera generation	56
2.5.2 Harvesting erythroid tissue from mouse models	57
2.6. <i>Chromatin techniques</i>	57
2.6.1 Chromatin immunoprecipitation	57
2.6.2 Chromosome conformation capture	58
2.7. <i>Expression analysis</i>	58
2.7.1 Blood counts, cytopsins and blood films	58
2.7.2 Real-time PCR	59
2.7.3 RNA Fluorescence in situ hybridisation	60
2.7.4 Nanostring RNA quantitation.....	61
2.8. <i>Other techniques</i>	61
2.8.1 Bisulphite sequencing and nested PCR amplification	61
2.8.2 FACS analysis.....	61
2.8.3 Southern blotting.....	62
2.8.4 Colony blotting	63
3. A novel mutation at the α-globin locus	65
3.1 <i>Introduction</i>	65
3.1.1 Naturally occurring deletions at the α -globin locus.....	65
3.1.2 HS-40 as the major regulatory element at the α -globin locus	70
3.1.3 Conclusions from previous studies, and state of the literature.....	74
3.2. <i>An isolated deletion of HS-40 in homozygosity</i>	77
3.3. <i>Results</i>	81
3.3.1 Phenotypes of MC and RC: expression in primary erythroid cells	81
3.3.2 Clarification of the absence of HS-40.....	82
3.3.3 The 39bp insert: absence of canonical transcription factor binding sites .	83
3.3.4 Assessment of the chromatin environment of the mutant α -globin locus	85
3.3.5 Confirming the normality of the remaining regulatory elements	86
3.3.6 Enhancer promoter looping in the mutant chromosome.....	89
3.3.7 Investigating the discrepancy between transcription in the human and the humanized mouse	92
3.4. <i>Discussion</i>	94
3.4.1 Human, mouse and humanized mouse: general comments on interspecific analysis of enhancer function	94
3.4.2 Implications for studies at the α -globin locus.....	97
Chapter 4: Generating mouse models	102
4.1 <i>Introduction</i>	102
4.1.1 Overview: dissecting enhancer function at the mouse α -globin locus ...	103
4.1.2 Phenotypes of previously generated mouse models.....	104
4.2 <i>Completing a series of knockout models</i>	107
4.2.1 R1 in mouse erythroid cells	109
4.2.2 Creation of a targeting vector for deleting of mouse R1, HS-31.....	112
4.2.3 Creation of a targeted ES cell line for R1 deletion	114
4.2.4 Generation of mice homozygous for Δ R1.....	117
4.3 <i>Generation of the ΔR1/ΔR2 mouse</i>	121

4.3.1 A strategy for the generation of a double knockout	122
4.3 Generation of the $\Delta R2/\Delta R3$ mouse	121
4.4.1 A critical mass of enhancers, or a critical role for R1 and R2?	128
4.4.2 Generation of the $\Delta R2/\Delta R3$ mouse	128
5. Characterizing mouse models of enhancer knockouts	133
5.1 Introduction	133
5.2 Results	134
5.2.1 Viability of single knockout models	134
5.2.2 Viability of double knockout models	136
5.2.3 Haematological phenotypes	136
5.2.4 Assessing α -globin mRNA levels by qPCR	144
5.2.5 Assessing α -globin mRNA levels by RNase protection assay	145
5.3 Discussion – limitations in the initial phenotype assessment	147
6. Nascent transcription from knockout mouse models	150
6.1 Introduction	150
6.2 Results	153
6.2.1. Choosing a system to investigate: fetal liver cultures	153
6.2.2. Collecting newly transcribed RNA	155
6.2.4 Validating samples of nascent RNA obtained	160
6.2.3 Quantifying nascent RNA	161
6.2.5 Choosing control genes for nanostring analysis	167
6.2.6 Nanostring: technical and biological replicates, and normalization	168
6.2.7 Impact on other genes within the α -globin cluster	174
6.3 Other roles for conserved elements at the α -globin locus?	176
6.4 Summary	182
Chapter 7: Final Conclusions	186
7.1 Summary of Findings	186
7.2 Wider implications	188
Appendices and References	190
Appendix 1: Primers	191
Appendix 2: Nanostring probes	193
References	195

Table of Figures

Figure 1.1 The proportion of the genome occupied by coding and non-coding sequence.....	3
Figure 1.2 Schematic representation of Banerji's enhancer assay.....	6
Figure 1.3 Ultrastructure of the SV40 enhancer showing the location of transcription factor binding sites.	7
Figure 1.4 Schematic representation of failings of the enhancer assay and intragenically, interspersed among several distinct promoters.	10
Figure 1.5 Sequence conservation at the α -globin locus	12
Figure 1.6 Schematic representation of ChIP peaks for enhancers and promoters.	
Figure 1.7 Conformation change at the β -globin locus	23
Figure 1.8 The principle of chromosome confirmation capture technology	24
Figure 3.1 α^0 thalassaemia resulting from upstream deletions in the α -globin cluster.	
Figure 3.2 Putative regulatory elements across the α -globin locus	69
Figure 3.3 Humanized mouse syntenic region.....	71
Figure 3.4 Expression of human α -globin in the presence and absence of HS-40, in MEL hybrids	73
Figure 3.5 A schematic representation of the organisation of the α -globin locus in human and mouse	76
Figure 3.6 Deletion of the murine equivalent of HS-40 (HS-26) has an unexpectedly mild haematological phenotype	76
Figure 3.7 The (aa) ^{ALT} mutation	78
Figure 3.8 ($\alpha\alpha$) ^{ALT} mutation in homozygosity – clinical data	80
Figure 3.9 α -globin expression in a heterozygote (RC) and homozygote (MC) for the HS-40 deletion (aa) ^{ALT}	82
Figure 3.10 HS-40 is absent in patient MC.....	83
Figure 3.11 Histone modifications across the locus in the ($\alpha\alpha$) ^{ALT} chromosome Error! Bookmark not defined.	
Figure 3.12 qPCR validation of ChIP material, including primers and probes specific for the 39bp insert of mutation ($\alpha\alpha$) ^{ALT}	88
Figure 3.13 Maintained erythroid specific chromosome looping in the absence of HS-40.	91
Figure 3.14 Summary of bisuphite sequencing results for regulatory elements MCS-R1-3.....	93
Figure 3.15 The role of R2 in α -globin expression in mouse, human and humanized mouse	100
Figure 4.1 Summary of enhancer knockout models.	105
Figure 4.2 Features of the R1 element.	110
Figure 4.3 R1 knockout construct in pNTFlox.	113
Figure 4.4 Restriction enzyme digests confirming correct targeting construct Δ R1-pNTFlox.....	114
Figure 4.5 Schematic representation of targeting plan for the R1 knockout, and Southern blot strategy to detect correctly targeted clones.	115
Figure 4.6 Screening with R probe for Δ R1.....	115
Figure 4.7 Southern blots analysis of clone 1D11 used for blastocyst injection.....	116

Figure 4.8 Karyotyping targeted clones for blastocyst injection.	117
Figure 4.9 Screening strategy and Southern blots to determine $\Delta R1$ floxed x cre mice.	119
Figure 4.10 PCR to confirm appropriate Cre recombination of targeted alleles	120
Figure 4.11 Confirming homozygosity among $\Delta R1$ mice	121
Figure 4.12 Strategy for re-targeting the $\Delta R2$ line with <i>pNTFlox</i> $\Delta R1$	123
Figure 4.13 Homologous recombination of the $\Delta R1$ pNTFlox construct to an ES cell line previously targeted with a construct for the deletion of R1.....	124
Figure 4.14 Southern blot using Bmt1 digest for $\Delta R2$, $\Delta R1$	126
Figure 4.15 Agouti offspring of $\Delta R2$, $\Delta R1$ chimeras.....	127
Figure 4.16 Screening strategy and Southern blot to confirm <i>in cis</i> targeting of $\Delta R3$ pNTFlox to $\Delta R2$ line.	129
Figure 4.17 Southern blot to clarify discrete deletions of R2 and R3 following <i>in vitro</i> Cre recombination.	130
Figure 5.1 $\Delta R1/\Delta R2$ heterozygote (left) and homozygote (right) littermates, taken at e14.5..	137
Figure 5.2 Peripheral blood films from mice homozygous for regulatory element knockouts.	139
Figure 5.3 Peripheral blood haemoglobin and red cell indices from mice homozygous for regulatory element knockouts.	139
Figure 5.5 Peripheral blood reticulocyte counts from mice homozygous for regulatory element knockouts.	140
Figure 5.6 Sample reticulocyte preparations from peripheral blood of mice homozygous for $\Delta R1$ and $\Delta R(m)$	141
Figure 5.7 Haematological analysis of $\Delta R1/\Delta R2$ heterozygotes.....	142
Figure 5.8 qPCR is insufficiently sensitive to detect fractional changes in α -globin expression.....	145
Figure 5.9 RNase protection assay is insufficiently sensitive to detect fractional changes in α -globin expression..	146
Figure 6.1 Cytospins demonstrate synchronous maturation of fetal liver erythroid cultures over 48 hours in maturation medium.....	154
Figure 6.2 Comparison between morphology and immunophenotype in WT and $\Delta R2$ homozygote fetal liver cultures, at three points during erythroid maturation.....	156
Figure 6.3 Immunophenotypic comparison of WT and mutant fetal liver cultures using Ter119, CD71 and CD44.	158
Figure 6.4 Workflow summary: obtaining nascent RNA from fetal liver cultures ...	159
Figure 6.5 ABI RNA Bioanalyser traces for nascent RNA.....	162
Figure 6.6 Nascent RNA sequencing across the α -globin locus. .. Error! Bookmark not defined.	
Figure 6.7 Nascent RNA sequencing across the <i>Slc25a37</i> locus. . Error! Bookmark not defined.	
Figure 6.8 Assessing the suitability of control genes chosen for nanostring expression analysis.	169
Figure 6.9 Nanostring quantitation of nascent RNA from fetal liver cultures for each mouse model.	171
Figure 6.10 RNA FISH validation of nascent transcript quantitation by nanostring .	171
Figure 6.11 Nascent RNA quantitation late in erythroid maturation.	173

.....

Figure 6.12 RNA FISH analysis of the changing likelihood of active globin loci during erythroid maturation.....	173
Figure 6.13 Deletion of α -globin regulatory elements impacts distant gene expression.....	175
Figure 6.14 <i>α-globin</i> in vascular endothelium.....	179
Figure 6.15 <i>α-globin</i> in megakaryocytes?.....	181

Chapter 1

Introduction

1. Introduction

Many of the most significant open questions in biology reflect our incomplete understanding of how gene expression is regulated. Whether in the specification of cell fate, the processes of development, the responses to environmental stimuli, or the subversion of normal gene expression in disease, we are currently unable to explain how the genome deploys its resources. Moreover, cataloguing genome sequences alone brings us no closer to an understanding of how each terminally differentiated cell can express only those genes needed for its normal function, with effective silencing of the rest. Not only is this a fundamental problem in basic biology, it is an obstacle in the approach to the modulation of gene expression for therapeutic purposes. Efforts to understand the regulation of transcription are therefore prime amongst the goals of current biology.

1.1 The regulation of eukaryotic gene expression

The human genome is estimated to contain approximately 20,000 genes (Collins et al. 2004, Pertea & Salzberg 2010) - a number very similar to that seen in *C. elegans*

(The *C. elegans* Sequencing Consortium 1998, Chen et al. 2004). However, while the simpler organism fits its coding sequences into a genome of only 100Mb, the human genome extends to approximately 3.2Gb of sequence.

The paradox of gene number versus organismal complexity was pointed out as early as 1966 by Ohno and colleagues (Ohno and Atkin 1966). Ohno quotes Ohta and Kimura's only slight over-estimate that 6% of the human genome would be responsible for protein coding sequences (Ohta and Kimura, 1971) – the current estimate being nearer 1.2% (ENCODE Consortium, 2013) - and suggested that much of the rest might comprise a historical record of “the remains of extinct genes”, or the evidence of “nature's experiments which [have] failed” (Ohno, 1972). That the non-coding sequence may have function beyond providing space and raw material for the acquisition of potentially useful mutations through natural selection was not considered in Ohno's original treatise, and the possibility that the non-coding sequence itself might better reflect organismal complexity (figure 1.1) was not raised.

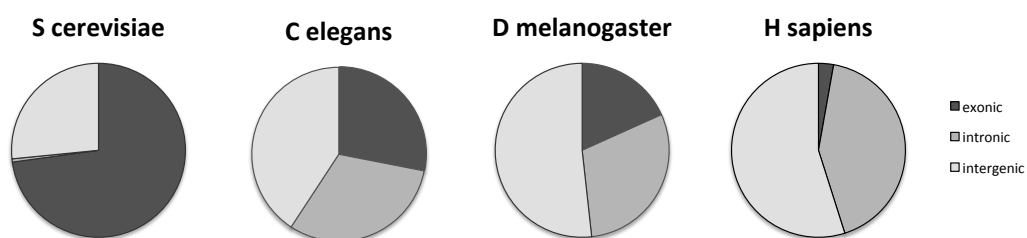


Figure 1.1 The proportion of the genome occupied by coding and non-coding sequence. (Data from Alexander et al. 2010)

.....

It has since become clear that non-coding sequences contain a panoply of genomic elements with likely regulatory function – with almost bewildering numbers being defined in higher organisms (ENCODE Consortium, 2013). The implication is that the complexity of higher organisms may be achieved in part by the more pleiotropic use of their coding sequences – that is, through varying the expression of their coding sequences in different spatio-temporal patterns. In order to accomplish this, precise regulatory control must be exerted over gene expression through mechanisms that are currently understood at only the most basic level.

The regulation of the eukaryotic genome has been the subject of intense investigation for at least forty years. Attempts to find the “eukaryotic operon” – a means of genetic regulation comparable to Jacob and Monod’s Lac operon in bacteria – led to increasing awareness of both structural and sequence dependent constraints on gene expression. The eukaryotic promoter was identified in 1977 (Maxam et al. 1977) and sparked a series of discoveries of *cis* regulatory elements of increasing complexity.

At its simplest, the classification of regulatory elements includes promoters, enhancers (Banerji et al. 1981), locus control regions (Li et al. 2002), insulators (Kellum and Schedl, 1991) and silencers (Laimins et al. 1986). In each case, however, our understanding of the function of these elements remains imperfect, and increasingly diverse roles in transcription control are being uncovered for even the best characterized among them. In many cases, the very definitions that separate

.....

one *cis* element from another are being eroded (e.g. Kowalczyk, Hughes, Garrick, et al. 2012b, Arnold et al. 2013).

Many of the difficulties with the analysis of *cis* regulatory elements relate to the experimental strategies used to identify, locate, define and characterize them. In this introductory chapter, I shall highlight the current limits of our understanding of regulatory element function, with particular focus on enhancers. I will discuss the inherent limitations of both genome-wide and heterologous experimental systems with respect to defining and characterizing regulatory elements, and highlight some of the difficulties and contradictions these studies raise. In introducing the framework and aims of the thesis, I will make the case for locus-specific studies in our attempts to expand our appreciation of transcriptional regulation.

1.2 Defining and Identifying Enhancers

1.2.1 The original approach to definition and identification of enhancers

The formal definition of enhancers is directly referable to the experimental methods used in their initial identification. An enhancer is a genomic sequence capable of up-regulating expression from a linked reporter gene, irrespective of its position relative to the gene's promoter, in terms of either distance or orientation. Banerji and colleagues (Banerji et al. 1981) were the first to describe these elements, isolating a sequence from the SV40 virus which was capable of reproducibly up-regulating expression of a linked rabbit *β -globin* gene in HeLa cells (figure 1.2).

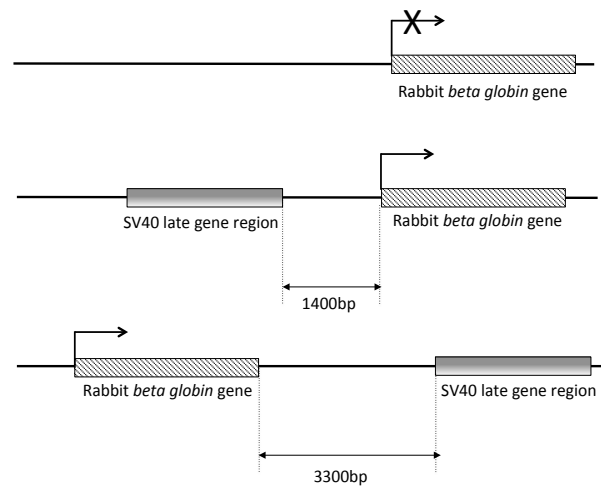


Figure 1.2 Schematic representation of Banerji's enhancer assay. A construct bearing the β -globin gene is transcribed only in the presence of the SV40 late gene region – independent of its orientation and distance from the gene's promoter.

Their significance was immediately apparent, providing insight into a layer of potential transcriptional regulation more flexible than the interaction between general transcription factors and a gene's promoter. The 'enhancer' activity of the SV40 sequence was subsequently found to be equally effective in driving transcription from the human β -globin gene (Moschonas et al. 1981) and from a number of histone genes (Banerji et al. 1981), suggesting that its ability to increase transcriptional output was intrinsic to the sequence itself rather than consequent upon interaction with a specific promoter. Further characterization of this sequence demonstrated binding sites for both tissue-specific and general transcription factors, e.g. Sp1, AP1, AP2, Oct1, and NF- κ B (figure 1.3).

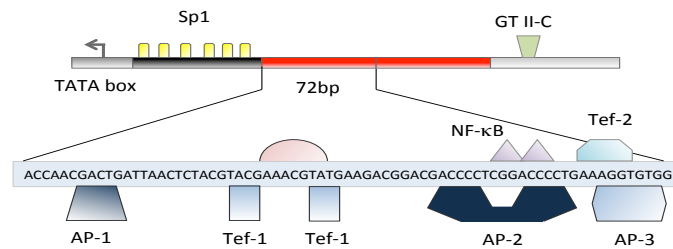


Figure 1.3 Ultrastructure of the SV40 enhancer showing the location of transcription factor binding sites.

Enhancer activity was also demonstrated for a similar sequence in the polyoma virus, and the same team shortly afterward described a *cis* acting element with corresponding activity in the mammalian genome (an immunoglobulin heavy chain enhancer, capable of stimulating increased transcription from a linked rabbit β -globin gene or SV40 T-antigen, even at several thousand base pairs distance [Banerji et al. 1983]). Using this example, Banerji and colleagues were able to implicate enhancer elements in the pathological up-regulation of gene expression in certain disease states (e.g. over-expression of *myc* in the context of translocations which place it in the vicinity of the immunoglobulin heavy chain enhancer, as in Burkitt lymphoma [Adams et al. 1985]).

However, the observations of Banerji and colleagues rely on some fundamentally artificial experimental conditions to try and define a functional regulatory element. Their assay clearly defines a sequence with a biochemical function, and it remains the gold standard for the identification and validation of enhancers today (e.g. Attanasio et al. 2013, Heinz et al. 2013); however, extrapolation from the output of the rather spare enhancer assay to a similar function in the complexity of the

.....

genome is not addressed, and remains tacitly assumed in many studies that employ this assay.

Yet the differences between the assessment of enhancer function in the classical reporter assay and in the native genomic setting are numerous.

First, Banerji's original assay assessed transcriptional output in a single cell line – a method still used in the luciferase assay. Leaving aside the wider objections to the use of immortalized lines in the investigation of biological processes, enhancers that rely on tissue-specific transcription factors for their activation may be manifest only in cell lines expressing those factors. A negative outcome from the enhancer assay need not mean that a sequence would not have enhancer activity in a different cell line.

When the enhancer assay is used to generate stable lines, or assessed in the whole embryo setting, there are further complications. The random site of integration of the reporter construct raises the possibility of the reporter gene being subject to the influence of nearby, native enhancers, giving false positive results or inappropriate patterns of activation, or being silenced by local chromatin effects, thereby giving false negatives.

Next, the assay's output may vary depending on whether one or more enhancers are included in the construct. The comprehensive analysis of the *fgf8* locus (Marinić et al. 2013) provides a good example of this. Individual enhancers from this locus gave

.....

a quite different pattern of gene expression when interrogated individually contrasted with testing *en bloc*. Moreover, the location of the reporter gene in the midst of the bank of *fgf8* enhancers was noted to influence its transcriptional output – suggesting that the critical requirement for enhancer independence of distance or orientation relative to its target gene might be simplistic. Work from the Engel group (Tanimoto et al. 1999), which demonstrates a transcriptional difference to inversion of the enhancers at the *β-globin* locus, also supports this latter point.

The reporter construct itself may cause problems: the typical enhancer assay still uses minimal and heterologous promoters (as in, for example, Attanasio et al. 2013 and Heinz et al. 2013) though Marinic and colleagues demonstrate a differing pattern of reporter gene expression under the native *fgf8* promoter and the heterologous *β-globin* promoter. There is support for this finding in studies of the *β-globin* regulatory region (of which more below): Guy and colleagues (Guy et al. 1996) show that this region is capable of driving tissue specific transcription of the linked *β-globin* genes, but not of linked *lacZ* genes in the same system.

Finally, with a more detailed understanding of the nature of enhancers comes the appreciation that many of their key characteristics are absent from the original enhancer assay. The chromatin environment, for example, has been proposed to be a consistent feature of enhancers genome wide (Heintzman et al. 2007) and is discussed in more detail below; yet this is a dimension necessarily absent from the classical enhancer assay.

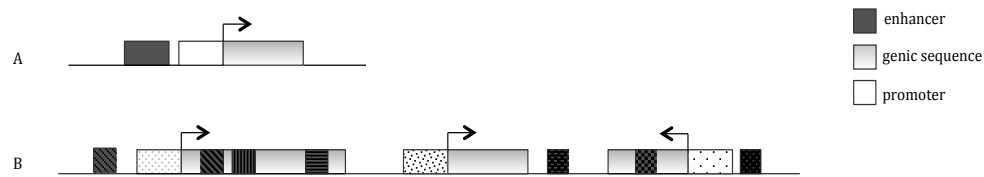


Figure 1.4 Schematic representation of failings of the enhancer assay

- A) A single enhancer, with a heterologous promoter and reporter gene, typical of the enhancer assay;
 B) A theoretical locus with banks of distinct enhancers located intergenically and intragenically, interspersed among several distinct promoters.

Figure 1.4 gives a schematic representation of the shortcomings of the enhancer assay. Banerji and colleagues have given an operational definition of a biochemical phenomenon, but left us with only imperfect methods for extending their observation in a biologically authentic way. Can we refine the functional definition of an enhancer *in vivo*? How common are they? Where are they to be found in the genome? What are their normal roles? After more than thirty years of progress that has revolutionized our approach to the study of enhancers, and despite an unprecedented collegiate effort across molecular biology laboratories worldwide, these questions are still the subject of active debate.

1.2.2 Improving on the enhancer assay?

Several methods are available to try and locate enhancers in their native setting, most taking advantage of advances in sequencing to extend their potential genome-wide. Even the most efficient and high throughput of systems for generating reporter constructs must defer to genome wide strategies for the identification of

.....

regulatory elements, though each method has its own limitations. The most commonly used of these methods are considered in more detail below.

Sequence Conservation

Since the genome is subject to persistent random mutation and replication error, sequences demonstrating conservation despite evolutionary diversification are likely to be functionally significant. As well as the expected conservation of coding sequences, non-coding regions including elements required for regulated gene expression will also be subject to evolutionary constraint. Although less than 2% of the genome represents coding sequence, comparison of human, rat, mouse and dog suggests that nearer 5% is under positive evolutionary selection (Broad Institute Sequencing Platform and Whole Genome Assembly Team et al. 2012).

Aligning this technique with additional genomic data (e.g. chromatin modifications, transcription factor binding and RNA sequence – of which more below), comparison across 29 species including human, rat and sloth, identified as many as a million putative regulatory elements in mammalian genomes.

More focused studies at individual loci have been able to describe the fine scale of evolutionary sequence constraint and, through the identification of regions of conserved synteny, to provide a comprehensive map of all the likely regulatory elements for a given gene (e.g. Figure 1.5, adapted from Hughes et al. 2005).

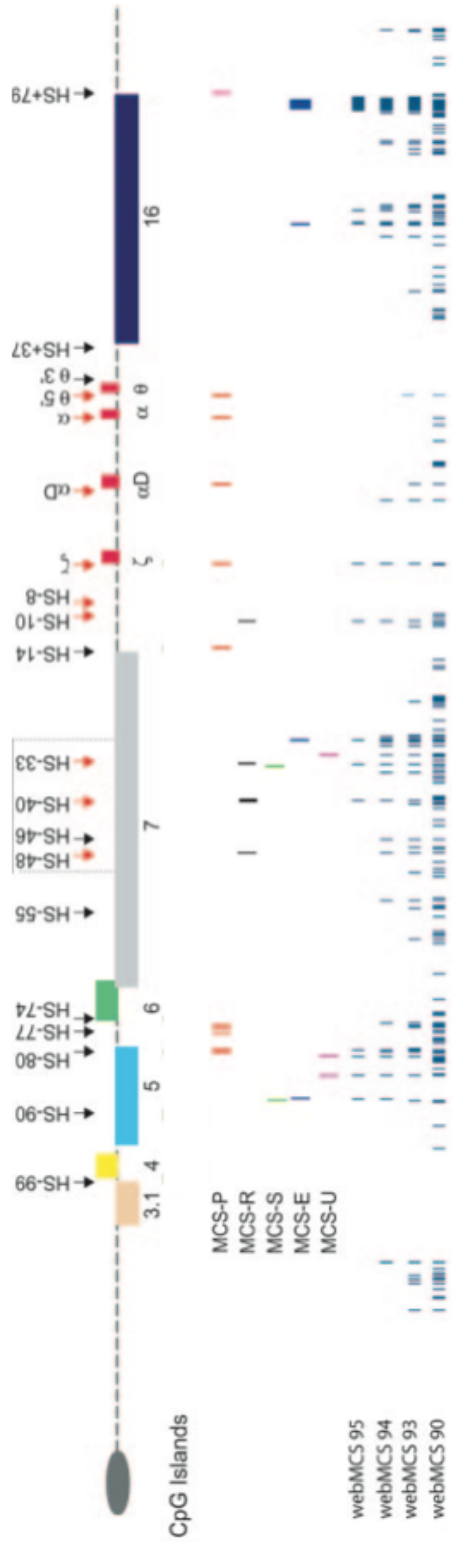


Figure 1.5 (from Hughes et al 2005) Using sequence conservation to define regulatory elements at the alpha globin cluster

Sequence conservation, assessed in 22 species across 500M years of evolution, demonstrates putative regulatory elements at the alpha globin locus. Coding sequences in a region of conserved syntenic regions are highlighted by coloured boxes; conserved non-coding regions, including promoters (MCS-P), regulatory elements (MCS-R), alternative exons (MCS-E) and splice sites (MCS-S) are identified and validated with webMCS data.

.....

The use of phylogenetic foot-printing to define regulatory elements clearly offers advantages for the identification of regulatory elements, but it not without hazards. Most obviously, patterns of gene expression are not absolutely conserved between different species (Wittkopp et al. 2004); this may result in potentially functional elements being overlooked. Further, although the conservation of an amino acid sequence requires the close conservation of coding sequence, there is evidence to suggest that the maintenance of enhancer function from species to species requires less stringent conservation at the sequence level (Ludwig et al. 2000; Costas et al. 2003, Ludwig et al. 2005). It has been argued that enhancers' ability to tolerate a degree of sequence variation without loss of function might contribute to their significance in evolutionary terms (Wray 2007, Frankel et al. 2010).

In addition, non-coding sequence conservation can tell us only that a genomic element has been evolutionarily preserved: it cannot define the reason for its preservation. Therefore, although it implies function, evolutionary conservation alone cannot expand upon what that function might be. Critical regulatory elements other than enhancers (e.g. CTCF sites) will also be subject to constraint, and will also be identified by phylogenetic foot-printing.

DNase I sensitivity

DNase I sensitivity is observed in chromatin which is relatively depleted in nucleosomes, and which is therefore more likely to be bound by *trans*-acting factors. Consequently, the DNase I assay has been widely used to locate possible regulatory

.....

elements (e.g. Varshavsky et al. 1978, Parslow & Granner 1983). As with phylogenetic foot-printing, it is hampered by its relative non-specificity, since all *cis* acting elements capable of binding *trans*-acting factors will demonstrate DNase I sensitivity, including enhancers, promoters, CTCF sites and insulator elements. Sequencing DNase I tracks to increasing depths permits the identification of DNase I 'footprints', which represent the location of binding of individual *trans* acting factors within nucleosome-depleted regions. This in turn allows the identification of individual *trans*-acting factors bound within any given nucleosome-depleted region through their canonical binding sequence. When allied with other data sets (such as the location of transcription start sites), DNase I sequencing is a powerful tool for the identification of potential regulatory elements genome wide.

As with phylogenetic foot-printing, data obtained using DNase I sequencing provides some of the more spectacular estimates of how many regulatory elements there might be in mammalian cell types. The ENCODE consortium locate approximately 1.5 million DNase I sites that are distant from transcriptional start sites (and which are therefore assumed not to be promoters) in 79 cell types (Thurman et al. 2013). What proportion of this extraordinary superfluity of apparent regulatory elements might represent true enhancers is open to debate: certainly, not all are functional in the gold standard enhancer assay of Banerji and colleagues.

Chromatin Immunoprecipitation

Both phylogenetic techniques and DNase I sensitivity rely on proxy markers of transcription factor binding. The direct interrogation of the binding patterns of *trans*

.....

acting factors is permitted by chromatin immunoprecipitation (ChIP) (Solomon et al. 1988). This allows the identification of both genic and non-genic regions bound by transcription factors, and can also identify regions bearing specific histone modifications. In both cases, the corollary is that the presence of such features implies functional regulatory significance. This is the basis of the striking (but contested) claim from the ENCODE consortium that >80% of the mammalian genome is 'functional' (Consortium, 2013).

ChIP for transcription factor binding at locations distant from the transcriptional start site (e.g. Hallikas et al. 2006) addresses the identification of regulatory elements most directly. The major disadvantage here is the extent to which transcription factors are tissue-specific, thereby identifying only those enhancers active in the tissue in question. Relatively few ChIP targets permit a more comprehensive approach to enhancer identification: these include the general transcription factor, p300 or the activated form of PolII (e.g. Visel et al. 2009).

Even where general transcription factor binding is tested, the technical limitations of ChIP must be borne in mind when interpreting the data it yields. As with the DNase I assay, ChIP is currently able to provide information on the binding of *trans* acting factors only in a population of cells, and cannot take account of cell-to-cell variation. Populations of cells that are in any way uncontrolled (for example, with respect to the cell cycle) may therefore manifest significant variation, which is obscured by the technique itself. This is one possible explanation for some of the more surprising conclusions that have been drawn from ChIP-seq datasets: examining the data for

.....

the ENCODE consortium for specific known enhancer elements suggests that a remarkable number of transcription factors are able to bind to a very limited sequence (see also chapter 4).

A less tissue-specific, and more generic approach to enhancer identification was proposed by Heintzman and colleagues, who were among the first to suggest a histone modification shorthand for the detection of enhancers genome-wide (Heintzman et al. 2007). First identifying putative enhancers by the co-localization of DNase-I sensitivity and binding of the general transcription factor p300, Heintzman and colleagues noted that these sequences were characterized by a strong enrichment of H3K4 monomethylation. This mark, by contrast, was typically depleted at promoters, where H3K4 trimethylation was more prominent. Despite subsequent modifications (e.g. data from Pekowska and colleagues, 2011, suggesting a role for H3K4 trimethylation specifically at active enhancers - possibly of significance when we consider the transcription of regulatory elements themselves) this observation has become central to the identification of enhancers genome-wide: the ratio of H3K4 monomethylation to trimethylation is now widely used as a means of identifying regulatory elements and classifying them as enhancers or promoters (Figure 1.6). It is of note, however, that the other proposed regulatory elements (e.g. insulators, silencers, CTCF sites) are not incorporated into Heintzman's dichotomy.

With further work, the identification of enhancers based on chromatin marks has become increasingly nuanced. Rada-Inglesias and colleagues used H3K27 marks to

.....

separate ‘active’ from ‘poised’ enhancers (Rada-Iglesias et al. 2012). Using ES cells, this group identified a group of enhancers characterized by a preponderance of H3K27 acetylation, and contrasted them with enhancers bearing principally H3K27 trimethylation. The former were proposed as ‘active’ enhancers driving gene expression in ES cells, while the latter were suggested to be ‘poised’ enhancers not yet driving active gene expression. During the process of differentiation of ES cells into neuroepithelial tissues, several enhancers that had previously been marked as ‘poised’ gained the H3K27acetylation mark, as the gene with which they were associated became expressed. Similar work has also highlighted the existence of ‘permissive’ enhancer states – i.e. those in which an apparently active enhancer is associated with a polycomb-repressed promoter (e.g. Taberlay et al. 2011). In both cases, these data are clearly dependent on the ability accurately to link an enhancer with the gene it regulates (see also section 1.3.2).

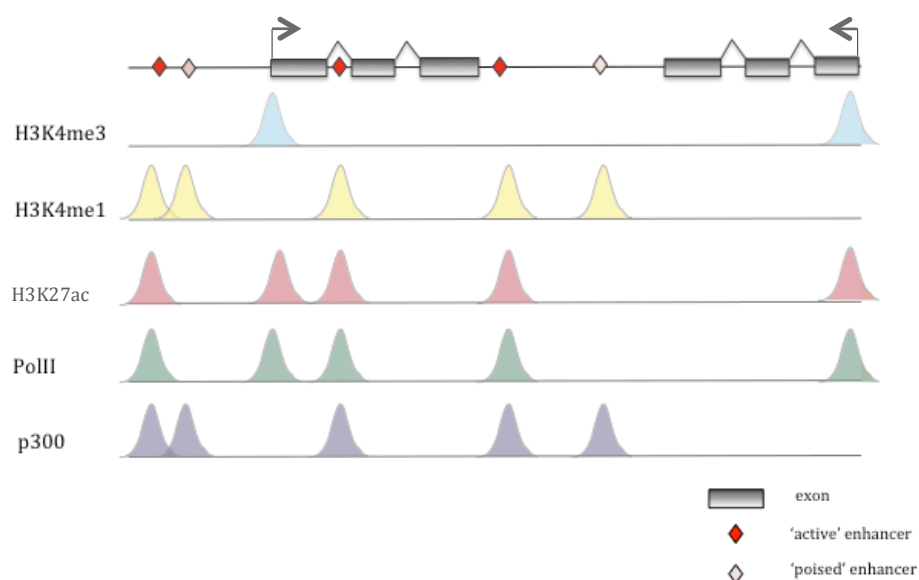


Figure 1.6 Schematic representation of ChIP peaks for enhancers and promoters. While promoters are predominantly marked by H3K4me3, enhancers typically bear H3K4me1. Poised and active enhancers may be distinguished by H3K27 acetylation.

.....

With the accrual of more detailed ChIP-seq datasets, other histone marks have been proposed as indicative of specific functional roles (e.g. H4K16 acetylation, associated with transcriptional elongation) with varying degrees of experimental support. Whether the combinatorial pattern of histone modification will add further subtlety to the ChIP based identification and classification of regulatory elements remains to be determined (reviewed in Rando 2012).

Despite the increasing sophistication of genome-wide methods for the identification of enhancer elements, none of these methods assesses their functionality, and the gold standard for the validation of enhancer elements remains the low-throughput generation of assays linking putative enhancer elements to reporter genes. As well as the faults intrinsic to the enhancer assay itself, this validation relies upon the assumption that the elements identified by these two very different methods are equivalent. They may not be: studies that use enhancer assays to validate the identification of enhancers by genome-wide methods invariably find a proportion that do not up-regulate the expression of a linked gene (Table 1.1) Neither methodology is ideal; it is not clear which provides the more biologically meaningful data; and that they are aligned often without question compounds the problem.

1.3 From identification to function

One of the principal criticisms of the genome-wide studies outlined above is their inability to link attempts to identify enhancers with assessment of their function, and

Table 1.1 Concordance between genome-wide methods for enhancer identification and the results of reporter assay

Study	Method	Number of Enhancers identified	Positivity validated by reporter assay
Heintzmann <i>et al.</i> , 2009	Histone modification ChIP	36,589	78% positive (7/9)
Visel <i>et al.</i> , 2009	p300 ChIP	5,209	87% (75/86)
Pennacchio, 2006	Sequence conservation	*	46% (246/528)
Yip <i>et al.</i> , 2012	Histone modification ChIP	13,539	30% (6/20) in mouse; 70% (19/27) in fish
Narlikar <i>et al.</i> , 2010	Transcription factor binding	41,930	62% (16/26)
Crawford <i>et al.</i> , 2006	DNase I sensitivity	14,190	1.2% (1/77)
Heinz <i>et al.</i> , 2013	Lineage specific, combinatorial transcription factor binding	932	90% (18/20)
Attanasio <i>et al.</i> , 2013	p300 ChIP seq	4399	59% (121/205)

.....

thus the inability to provide a meaningful validation of a class of regulatory elements. This has been addressed by the study of knockout models of individual enhancer elements, as well as by the characterization of naturally-occurring extragenic mutations in human disease, both of which provide direct functional evidence to support a critical role for enhancer elements in normal gene expression.

Among the best-described examples of naturally occurring mutations which identify the location of functional regulatory elements are the rare cases of thalassaemia that helped to define the *β -globin* locus control region and the *α -globin* major regulatory element (Kioussis et al. 2004, Curtin et al. 1985, Nicholls et al. 1987, Hatton et al. 1990). In these cases, reduced expression the *globin* genes is observed in the context of an intact coding sequence, with the implication that distant deletions segregating with the phenotype are responsible for the reduced transcription through impaired enhancer activity. Such naturally occurring deletions provide the opportunity to assess the significance of these elements in their native chromosomal setting, without experimental manipulation. Quantifying the impact of deletions on normal gene transcription avoids many of the technical shortcomings of the enhancer assay, while the sub-cloning of enhancer regions identified by such naturally occurring deletions and their interrogation in typical enhancer assays (e.g. Q. Li & Stamatoyannopoulos 1994) provides some validation for the *in vivo* significance of elements tested using this method. Extending this method to other loci of interest through engineered knockouts in animal models permits flexibility, but prior to the development of high throughout gene targeting methodologies (e.g.

.....

CRISPRs) this was a labour-intensive and painstaking method for the study of enhancers.

Such locus specific studies raise important points. First, any study of regulatory elements is incomplete without a functional dimension; but essential for this is the ability to determine which enhancers act on which gene promoters. At the *α -globin* locus, for example, the regulatory elements for the globin gene cluster lie within a bank of several widely expressed genes; confidently linking the right enhancer with the right promoter is critical to the assessment of each regulatory element's activity. Second, do these elements work alone or in concert? The layout of both the *α -* and *β -globin* loci suggests the latter, but neither the genome-wide methods for enhancer identification, the reporter assay's functional test nor simple knockouts of individual regulatory elements can confirm this. And finally, are all enhancers equal? While genome-wide studies resolutely ignore the possibility that the multiplicity of regulatory elements they identify might harbour significant functional variation, the results of locus specific studies suggest that there may be more complexity in these elements than previously thought.

1.3.1 How to establish if an enhancer is active

A pre-requisite for understanding whether an enhancer is active *in vivo* is an appreciation of the gene(s) upon which it exerts its function. Although there are data suggesting that active enhancers are themselves transcribed (De Santa et al. 2010) the best marker for enhancer activity available at present is the increased transcription of their associated genes. This leads into critical questions in enhancer

.....

biology: how does an enhancer select its target promoter(s) while preventing promiscuous activation of other, possibly more proximally located promoters, and how is it mechanistically able to influence the function of a promoter which may be hundred of kilobases distant? The logic determining enhancer-promoter specificity is still obscure, but without a method for confidently assigning enhancers to their promoters, functional studies *in vivo* will be extremely difficult to perform. An allied question is therefore how to assign enhancers to their promoters experimentally, ideally using high throughput genome-wide technologies.

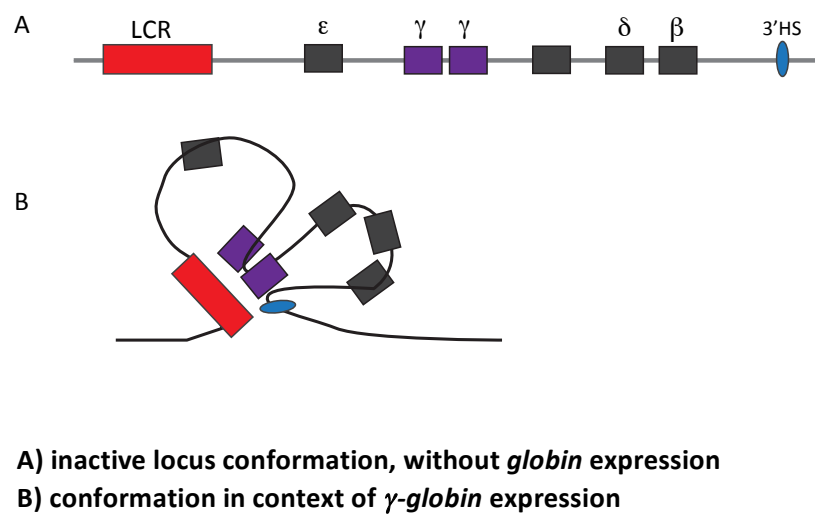
1.3.2 Enhancer/promoter interactions and their specificity

Where enhancer–promoter interactions have been convincingly demonstrated, studies of specific loci have demonstrated that enhancers may be located up to megabases away from the genes that they control, their apparent promoter specificity demonstrated by their ability ‘ignore’ several intervening promoters (the ultimate example being the action of the enhancer for *SHH*, located 1.5Mb away from its promoter in an intron of the *LMBR1* gene, with the promoter of *RNF32* between the enhancer and its ‘correct’ target). How this specificity is achieved is not clear, though the possibility of dimerization of transcription factors bound at both the enhancer and promoter has been suggested. How an enhancer even locates its promoter across large sequences remains to be determined, with the three dimensional organization of the genome being understood to only a limited extent.

Tolhuis and colleagues (Tolhuis et al. 2002) used chromosome conformation capture (3C) technology (Dekker 2002) to demonstrate that the *β-globin* locus adopts a

conformation which changes between committed erythroid progenitors that have yet to express globin and terminally differentiated erythroid cells which are actively expressing globin (Figure 1.7). The developmental switch in the expression of the β -globin genes is accompanied by a switch the interaction between the regulatory elements: the fetal γ -globin gene is looped out of the active chromatin hub, and the adult β -globin genes are included. This process has been validated by more direct means of visualizing chromosome interactions, including RNA FISH/TRAP (Carter et al. 2002).

Figure 1.7 **Conformation change at the β -globin locus**



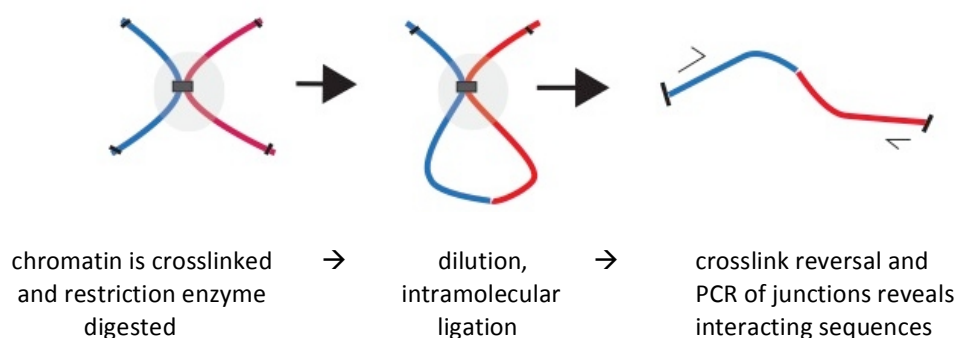
3C technology has subsequently been applied to many other gene loci to demonstrate the interactions between putative regulatory elements and their associated promoters in the context of active gene expression. Particularly well-described examples include the *c-kit* locus (Jing et al. 2008) during the process of erythropoiesis. *Kit* is expressed in early erythropoiesis, but is down-regulated during

.....

later stages of erythroid maturation. The looping of distal GATA2-bound enhancer to the *Kit* promoter during early erythroid development gives one chromosomal conformation, which is lost as the GATA switch takes place and *Kit* expression ceases.

The technique itself, as shown schematically in figure 1.8, has several disadvantages. It can detect interaction only between a single promoter and chosen fragments of surrounding sequence – introducing bias and hampering the discovery of previously unsuspected regulatory elements, as well as confining its utility to a single gene locus at any time. Its sensitivity is limited by the size of fragment generated by restriction enzyme digest. It is also unable to detect interacting elements that are in very close proximity to the promoter, due to a high frequency of background interactions – and, since regulatory elements appear not infrequently to be located in a gene's first intron, this is a potentially significant limitation.

Figure 1.8 The principle of chromosome conformation capture technology



Linking chromosome conformation capture with high throughput sequencing has given rise to several variations (table 1.2) that aim to describe the interactions

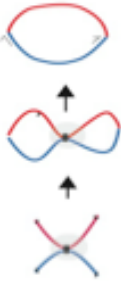
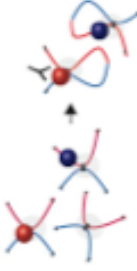
Basic 3C strategy 							
DNA is cross-linked, restriction-enzyme digested and subject to intramolecular ligation under dilute conditions. On reversal of crosslinking, quantitation of ligation between fragments is determined. The interaction likelihood between two chosen loci is estimated using qPCR to quantify the number of ligated fragments.							
4C		5C		HIC		ChIA-PET	
Uses inverse to locate loci interacting with a single known anchor point in a unbiased manner	Uses universal primers ligated to junction sites creates a sequencing library	Uses biotin tagging / dCTP to fill in junction sites, which are then sequenced	Uses a ChIP stage to identify junctions occurring in proximity to a specific protein or histone modification.	Genome-wide (all vs all)	Genome-wide (all vs all)	Genome-wide (all vs all)	Limited resolution by sequencing
Detects genome wide interactions of given locus. Unbiased. Limited to a single anchor.	Detects genome wide interactions among several selected loci. Limited resolution by sequencing or array	Genome-wide (all vs all) Limited resolution by sequencing	Genome-wide (all vs all) Limited resolution by sequencing				

Table 1.2. Chromosome conformation capture and its variants

.....

between regulatory elements and their contacts genome-wide. The use of “all-to-all” methods to interrogate chromosome conformation suggests that, rather than a series of isolated enhancer-promoter contacts, there are broader regions of increased intrachromosomal interaction. Such topologically associated domains (TADS) have been purported to define the boundaries of local transcriptional control, incorporating *cis* acting elements for coordinately regulated genes. Studies assessing TADs in classic model loci have demonstrated that they can correspond to the interactions between a promoter and its distal enhancers, and can alter at different developmental stages according to the nature of the enhancer-promoter interactions active at that time (Andrey et al. 2013).

The tentative conclusions of the various ‘C’ experiments support a model of the genome in which non-random chromosomal contacts subdivide of the nucleus into functional domains. Without single cell studies, it is difficult to be certain that the chromosomal interactions identified are more than simply the random ‘sampling’ of nearby chromatin by an enhancer before it locates its ‘true’ target.

Just as we do not know how an enhancer locates its target promoter, so we still do not understand how it avoids the inappropriate activation of other regulatory elements. Early experiments in *Drosophila* (Ohtsuki et al. 1998; Butler 2001) suggest that core promoters which feature TATA boxes and those which feature downstream promoter elements for the binding of TFIID have a differential likelihood of being activated by specific enhancers. Known specific enhancer-promoter interactions in

.....

Drosophila were also the source of data to suggest that proximal promoter tethering elements might be able to restrain enhancers from inappropriate promoter activation. (Calhoun et al. 2002). The (perhaps simplistic) view of insulator elements as boundaries to the influence of regulatory elements has also been proposed as a means for the constraint of enhancer-promoter interactions.

Blobel and colleagues (Deng et al. 2012) take this further by experimentally manipulating the machinery required for enhancer-promoter interaction. They report the use of zinc finger nucleases to tether the transcription factor Idb1 to the *β -globin* promoter in erythroblasts engineered not to express GATA1. Normally GATA1 null red cells express no *β -globin*; however, it was possible to rescue the locus back to full levels of globin expression by tethering Idb1 to the promoter. While GATA1 null cells without the promoter-tethered Idb1 failed to show looping between the promoter and the *β -globin* regulatory region, the promoter-bound Idb1 enabled the reinstatement of promoter-enhancer interactions. The implication is that, at this locus at least, the binding of tissue specific transcription factors to promoters as well as enhancers is needed for normal regulation of gene expression, and for the detection of normal enhancer function, and might even be required to determine promoter-enhancer specificity through dimerization.

1.3.3.Function in isolation vs orchestrated activity

The concept that enhancers may function as part of a group in controlling a given gene's expression is not amenable to experimental assessment using genome wide strategies, yet there are growing data to support this additional layer of complexity

.....

in enhancer function. Returning to the *β-globin* locus, the concept of the locus control region was proposed as a solution to the paradox that individual regulatory elements at this locus were not able to recapitulate normal *β-globin* expression in reporter assays. The group as a whole, by contrast, was able to induce *β-globin* expression in a tissue-specific, copy number-dependent fashion, irrespective of the site of genomic integration of the reporter construct. These criteria, a rather more stringent variation on the theme of the enhancer assay, were proposed to describe a distinct class of regulatory element, with the ability to open chromatin as a precursor to transcription being proposed as a mechanistic explanation (G. Li et al. 1998)

The *β-globin* locus features the best studied of locus control regions, with several groups attempting to dissect the contribution of its individual components, whether independent, additive or synergistic. Although important, these studies have presented a variety of challenges to interpretation. A significant difficulty has been the contradictory results obtained from studies performed in chimeric and native genomic settings, as well as lack of concordance between classical enhancer assays and gene targeting experiments. For example, the HS3 component of the *β-globin* LCR does not function reliably in reporter assays, but has been shown to drive expression of globin genes when incorporated with the entire *β-globin* locus. Here, HS3 was found to have a disproportionate impact on *γ-globin* transcription, (Fraser et al. 1993). Deletion of HS3 from a *β-globin* cluster YAC grossly disrupts the expression of both adult and embryonic *β-globin* genes (e.g. Bungert et al. 1995). However, upon deletion from the native murine locus by gene targeting, loss of the

.....

equivalent element results in a much more modest reduction of globin expression (Hug et al. 1996, Bender et al. 2012).

Similar inconsistencies are observed for studies of the LCR as a whole. Despite studies in transgenic systems support a role for chromatin opening in LCR function, targeted deletions of the entire LCR from the mouse *β -globin* cluster fail to impact on its DNase sensitivity – raising the possibility that, as with enhancers, there may be gap between the *in vitro* and *in vivo* assessment of LCR function.

Other counterintuitive findings that remain without a plausible explanation have included the observation that that a large deletion of a hypersensitive site (2.4kb) has an effect less pronounced than the loss of a smaller core section of the same site (Fang et al. 2005). Add to this the influence of integration sites that has plagued the quantitative analysis of the transcriptional impact of each element in reporter assays, and the relatively blunt tools available for quantitation of gene expression in the 1980-1990's when the bulk of this work was done, and it can be seen that reaching a coherent view of the function of the LCR and its constituent parts on the basis of these studies is not straightforward.

The *β -globin* cluster is not the only LCR that has been subject to investigation. LCR activity has been described at a number of other loci, including the *TCR α/β* locus, the immunoglobulin heavy chain locus and the *CD2* locus, again with roles in chromatin opening and demethylation. In each case, chromatin-related position

.....

effects in the expression of a reporter gene could be abrogated only by incorporation of the full complement of constituent enhancers.

Although bearing similarities to enhancers in many respects, the concept of the LCR further complicates our appreciation of transcriptional regulation. The individual components of the LCR would be identified as enhancers by the genome-wide techniques of section 1.2.2; but their full function cannot be inferred from these studies. Further, their impact appears to extend beyond transcription factor binding, with influence on the wider chromosomal context. That the results of *in vivo* and *in vitro* assays do not always concur underscores the importance of testing these elements in a biologically authentic setting.

The concept of the locus control region was revisited and expanded in 2013 by Young and colleagues (Whyte et al. 2013), to generate the concept of the ‘super-enhancer’. These were defined as a collection of enhancers, with individual elements located within 11kb of each other, characterized by very high levels of mediator binding. It was suggested (perhaps without sufficient empirical support) that these grouped elements constituted a functional unit, typically employed in the control of lineage specific or lineage-determining genes. Leaving aside this study’s questionable use of transcription factor binding sites alone as a means of identifying enhancers (a strategy which will inevitably yield a population of elements including promoters) a key flaw was the inability of the group to link putative enhancers with their cognate promoters confidently. As highlighted in section 1.3.2, this is a critical failing in common with many other studies purporting to assign functional

.....

significance to enhancers, and is only redoubled here by the assumption that adjacent components within a 'super-enhancer' will act on the same promoter.

However, the concept of the super-enhancer (and the LCR, since these may be different faces of the same biological phenomenon) does raise some important points. If the function of an enhancer depends on its genomic context, the validity of the reporter assay becomes even more questionable, and the conclusions we draw from it become increasingly laden with caveats. It also begins to acknowledge the question, again frequently overlooked, of why so many regulatory elements (hundreds of thousands, even at conservative estimates) are needed to control the expression of only 20,000 genes in the mammalian genome.

1.3.4 Are all these elements equal?

Whether acting alone or as part of a larger unit, it is clear that not all non-coding sequences are essential for an organism's normal function. Nobrega and colleagues (Nobrega et al. 2004) engineered deletions of 1511kb and 835kb in gene deserts in mouse, including over 1200 noncoding sequences with conservation between mouse and human, but were unable to demonstrate any significant impact of the loss in terms of morphology, growth, lifespan, or reproductive capacity of the animals they generated. Moreover, the expression of nearby genes was influenced to only a trivial degree – though the very long-range regulation of gene expression means that assessing only the nearby genes may miss the main activity of any specific element. The paradoxical implication is that these elements, despite their sequence

.....

conservation between species, are dispensable for normal gene expression, or that there is substantial redundancy between elements.

The concept that every enhancer might not be critically required for an organism's normal function has also been suggested from studies at specific mammalian loci. At the murine *IgH* locus, four *cis* regulatory elements with lymphoid specific DNase I hypersensitivity have been shown to interact with the *IgH* constant chain genes and play a role in immunoglobulin class switch recombination. The elements (named from 5' to 3', HS3a, HS1-2, HS3b and HS4) comprise a *cis* regulatory complex with LCR activity. While a combined deletion of the 3'-most elements, HS-4 and HS3b, is capable of profound disruption of *IgH* expression and class switch recombination (Pinaud et al. 2001), isolated deletion of either HS4 or HS3b has minimal impact on *IgH* expression or class switch recombination (Vincent-Fabert et al. 2009; Bebin et al. 2010). The isolated loss of HS3a and HS1-2 are also without significant impact (Rouaud et al. 2012). It is of note that all of these experiments were performed using precise gene targeting of the native murine *IgH* locus; they cast doubt upon earlier work in the relatively artificial setting of the reporter assay in which synergy was reported between the *IgH* locus enhancers (Ong et al. 1998, Chauveau et al. 1998). Similarly at the *TCR γ* cluster, where two conserved enhancer elements have been identified, deletion of both is needed to abrogate transcription. Deletion of either element in isolation has a negligible effect of *TCR γ* transcription (Xiong et al. 2013).

.....

While interesting, deletions at such complex genetic loci are perhaps not the optimal way to analyse the roles of individual enhancers. The number of potential constant region transcriptional targets complicates the assessment of transcriptional output from the *IgH* locus, and class switch recombination is necessarily a blunt phenotypic readout. More readily interpreted are similar studies at the *β -globin* locus, the transcriptional readout of which lends itself more easily to the assessment of enhancer activity. The most comprehensive analysis of the *β -globin* locus has been undertaken in mouse, by Bender and colleagues (Bender et al. 2012). This study again used gene targeting *in vivo* to delete each of the regulatory elements at the *β -globin* locus, using qPCR for adult β -globin chains in primary erythroid tissue as a measure of enhancer impact. No single regulatory element was deemed to be without function, though some were reported as being responsible for only very small contribution to the total of transcribed globin chains (with reductions of as little as 3% relative to normal globin mRNA levels.) Whether it is possible to make dogmatic statements on the biological significance of such minimal reductions in *β -globin* transcription measured by qPCR is questionable (see also chapter 5). Further, qualitative rather than merely quantitative differences between the roles of individual elements in the *β -globin* LCR were suggested, with different elements having a disproportionate role in, for example, the initiation of transcription and transcriptional elongation.

Further examples of functional variety and even redundancy exist. *Pax3* expression in the neural crest is controlled by a neural crest element directly upstream of the coding sequence, and a further regulatory element in the gene's fourth intron.

.....

These are capable of directing appropriate spatio-temporal transcription when incorporated into reporter assays, but targeted deletion of either element in isolation results in viable mice with normal *Pax3* expression and unaltered neural crest formation. Degenhardt and colleagues use this example to highlight apparent redundancy in regulatory elements, and counsel against confusing transgenic assays for sufficiency with proper tests of necessity (Degenhardt et al. 2010).

Degenhardt and colleagues also discuss the possibility that additional genetic or environmental perturbations may have revealed critical functions in these otherwise apparently redundant regulatory elements. This has been investigated most rigorously in *Drosophila*, giving rise to the concept of shadow enhancers.

Several genes in *Drosophila*, most notably those involved in developmental patterning, appear to be regulated by a bank of structurally similar transcriptional enhancers, each with the ability to direct a similar pattern of gene expression in reporter assays (e.g. Zeitlinger et al. 2007).

In careful experiments undertaken using the *shavenbaby* (*ovo*) locus as a model, Frankel and colleagues provided an excellent example of enhancer redundancy, demonstrating that the loss of individual enhancers at the locus was without effect in embryos grown in optimal conditions (Frankel et al. 2010). However, loss of a single enhancer did result in a consistent abnormal phenotype when embryos were raised in suboptimal conditions – whether environmental (low temperatures) or genetic (fewer copies of the gene *Wingless*). Their conclusion was that shadow

.....

enhancers were necessary for ‘phenotypic robustness’, or the canalization of phenotype despite variation in genetic or environmental conditions. Similar work showed that, in conditions of extreme temperature or in the context of loss of one of the maternal copies of *Dorsal*, two enhancers were required for the normal expression of the zinc finger transcription factor encoded by *snail*. In embryos not subject to the same environmental or genetic challenges, a single copy was sufficient. (Dunipace et al. 2011)

Thus, while apparently redundant, a selection pressure to maintain shadow enhancers at the population level would result in their sequence conservation. The same principles outlined to support the conservation of redundant coding sequences (e.g. Nowak et al. 1997) may also be applied to regulatory elements. This further complicates attempts to assign function to enhancers identified using genome-wide technologies: it is hard to see how the study of shadow enhancers could have been undertaken without low throughput painstaking locus-specific studies.

1.3.5 What do enhancers do to influence gene expression?

Having determined that an element marked as an enhancer may be functional, the mechanisms by which it influences transcription must be determined. Locus specific studies have suggested a variety of roles for enhancers, summarized below in table 1.3.

Table 1.3 Roles of regulatory elements in transcriptional control

Function	Test Locus	Reference
Opening chromatin	<i>HBB</i>	Fromm et al. 2011
Polycomb eviction	<i>HBA</i>	Vernimmen et al. 2011
Recruitment of general transcription factors and PolII	<i>HBA</i> <i>HBB</i>	Vernimmen et al. 2007 Drissen et al. 2004
Adoption of active chromosomal conformation	<i>HBA</i> <i>HBB</i>	Vernimmen et al. 2009 Tolhuis et al. 2002 Vakoc et al. 2005
Positioning of the locus in nuclear space	<i>HBB</i>	Francastel et al. 2001 Ragoczy et al 2006
Promoting transcriptional elongation	<i>HBB</i>	Bender et al. 2012 Sawado et al. 2003
Transcriptional memory?	<i>CD4</i>	Chong et al. 2010

However, as well as regulating transcription of their associated genes, enhancers are themselves subject to transcription. The phenomenon was first described at the HS2 element of the *β -globin* locus control region (Routledge and Proudfoot, 2002), both in transgenic constructs and *in vivo*. Although no functional role for either the transcript or the process of transcription could be proven, there was an association between the presence of enhancer transcripts and globin gene expression.

This work has been expanded genome wide in studies such as that by Kim and colleagues (Kim et al. 2010): here, transcription was identified from enhancers with a strong correlation to expression of nearby genes (note the assumption regarding the targets for these transcribed enhancers).

Similar work by de Santa and colleagues (De Santa et al. 2010) defined PolII binding sites outwith promoters, the majority of which bore the typical chromatin marks of

.....

enhancers. This was supported by Wang et al. (Wang et al. 2011), who also found that transcription of enhancers to be a good guide to their activity across the genome (with the usual caveats concerning enhancer-promoter linkage).

It is not yet clear whether transcription of enhancers is critical to their function, or whether it is simply an epiphenomenon – the inevitable consequence of PIC assembly at the enhancer. Various groups have proposed role for both the transcript and the process of transcription itself; experiments to differentiate between these will not be straightforward to design and control, but may provide important insights into the way enhancers work.

1.5 The clinical significance of regulatory elements

If, as the discussion above suggests, regulatory elements are indeed critical for normal tissue specific and developmental stage specific profiles of gene expression, we might anticipate that defects within them would be responsible for genetic disease. Table 1.4 shows some of the known syndromes resulting from disruption of *cis* regulatory elements (Kleinjan & Coutinho 2009). These diseases have several features in common. Most strikingly, they are all exceedingly rare – an observation which may in itself be taken as support for arguments concerning the redundancy of regulatory elements. They are predominantly diseases characterized by developmental defects or disruption of body patterning; there are some data to suggest that the regulatory elements implicated in developmental gene expression are particularly well conserved and, by extension, defects within these elements may be more likely to result in significant genetic disease (Bejerano 2004, Sironi 2005).

Most involve sizeable deletions, with point mutations being relatively rare. In table 1.4, the disorders of globin synthesis stand out as by far the ‘simplest’ of phenotypes resulting from enhancer loss, emphasizing their position as an attractive potential model for the study of enhancer function.

Table 1. Regulatory element dysfunction and disease (adapted from Kleinjan 2009)

Enhancer defect	Disease	Gene	Reference
Loss of element	Van Buchem disease	<i>SOST</i>	Loots et al. 2005
	Mesomelic dysplasia	<i>SHOX</i>	Sabherwal et al. 2006
	Leri-Weill dyschondrosteosis	<i>SHOX</i>	Fukami et al. 2005
	α thalassaemia	<i>HBA</i>	Hatton et al. 1990
	β thalassaemia	<i>HBB</i>	Kioussis et al. 2004
Translocation	Aniridia	<i>PAX6</i>	Fantes et al. 1995
	Saethre-Chotzen syndrome	<i>TWIST</i>	Cai et al. 2003
	Campomelic dysplasia	<i>SOX9</i>	Leipoldt et al. 2006
	Holoprosencephaly (HPE2)	<i>RUNX2</i>	Fernandez et al. 2005
	Holoprosencephaly (HPE3)	<i>HOXD</i>	Spitz et al. 2002
	Townes-Brocks syndrome	<i>SIX3</i>	Wallis et al. 1999
	Iridogoniodysgenesis	<i>FOXC1</i>	Davies et al. 1999
	Peter’s anomaly	<i>TGFb2</i>	David et al. 2003
	Perlizauer-Merzbacher disease	<i>PLP1</i>	Muncke et al. 2004
Point mutations	Pre-axial polydactyly	<i>SHH</i>	Lettice et al. 2002
	Pierre Robin sequence	<i>SOX9</i>	Benko et al. 2009
	Holoprosencephaly (HPE3)	<i>SHH</i>	Spitz et al. 2002
	Non-syndromic cleft lip/palate	<i>IRF6</i>	Rahimov et al. 2008
Promoter down-regulation	α thalassaemia	<i>HBA</i>	Tufarelli et al. 2003
	Lynch syndrome	<i>MSH2</i>	Ligtenberg et al. 2008
Decoy promoter formation	α thalassaemia	<i>HBA</i>	De Gobbi et al 2006
Duplication	Pre-axial polydactyly	<i>SHH</i>	Klopocki et al. 2008

.....

The rarity of enhancer-based Mendelian disorders perhaps fits with the complex and often redundant roles these elements play in controlling gene expression. The combination of some internal sequence flexibility within individual enhancers and the ability, at least in some cases, for enhancers to overlap in their function may permit some regulatory variants to become quite prevalent in populations. The common disease variant hypothesis, which was generated to explain the high population frequency of non-Mendelian disorders such as cardiovascular disease, states that such sequence variants may have a minor effect in isolation, but a much more significant impact when inherited in combination with other similar variants.

In accordance with this hypothesis, GWAS studies in common non-Mendelian disorders have frequently highlighted single nucleotide polymorphisms in non-coding regions. Visel and colleagues (Visel et al. 2009) suggest that 40% of such GWAS hits are in non-coding sequence; where these have been experimentally dissected, several have shown evidence of enhancer activity (e.g. *TCF7L2* in type II diabetes, and *EIF3H* in colorectal cancer (Savic et al. 2011; Pittman et al. 2010). Their role in the pathogenesis of some of the commonest diseases adds further urgency to efforts to understand enhancer function.

1.4 Unanswered questions

The discussion above highlights some of the key problems with the study of regulatory elements. We can define enhancer function, but only in experimental systems far removed from the normal function of each element; we can map the location of a series of genomic elements with consistent features that we term

.....

regulatory elements, but we are unable empirically to assign a function to the group as a whole; and we can align these two kinds of study only imperfectly and for a fraction of the total number of genetic loci. Thus even the most rudimentary framework for a study of regulatory elements - namely the ability to define what they are and where they can be found – is far from secure. Further, our ability to link a regulatory element's features with its activity is hampered by the limitations of techniques to assign each element to its cognate promoter. We do not know if the elements identified by genome scanning methods are all functional enhancers; we do not whether they are functionally equivalent or diverse; and for the most part we do not know if they act in isolation or cooperatively. Meanwhile their importance is highlighted by studies that implicate them in the very process of evolution, and GWAS studies implicate them in the causation of diseases that are responsible for a vast burden of morbidity and mortality worldwide.

Despite the potential of genomics methods for the rapid acquisition of large datasets, they will not permit us to solve these problems. The dismissal of computational methods as “low input, high throughput, no output” (from Brenner 2009) is perhaps consciously extreme, but it is clear that the experimental dissection of individual well-described loci will still be required to define the functional aspects of regulatory elements.

1.6 Regulatory elements at the α -globin locus

As several examples in the preceding discussion testify, the *globin* loci have been the model for the analysis of many principles of enhancer function. There is also a clear medical imperative for the study of these loci: the globin disorders comprise the commonest monogenic disorders worldwide, with a burden of morbidity and mortality that is borne principally by the poorest countries in the world. Understanding the mechanisms by which the *globin* genes are turned on and off is likely to be of major clinical significance for the management of thalassaemia and sickle cell anaemia.

In addition, from a basic scientific perspective, the *globin* loci lend themselves readily to the study of transcription. Each has genes expressed in red blood cells, which are easily obtained from patients and experimental model organisms, and α - and β -globin are both expressed in a predictable fashion during the process of erythropoiesis – a process which can be closely observed in primary cells *ex vivo*. Finally, experimental studies have been authenticated through the investigation of the impact of naturally occurring deletions at both loci that produce a relatively simple clinical phenotype. For these reasons, studies of the *globin* loci provide a perspective on transcriptional regulation that would otherwise be difficult to achieve.

As described in section 1.2.2, the α -globin locus is found within region of synteny of ~135kb has been defined by sequence comparison across 22 species (Hughes et al.

.....

2005), covering points of divergence spanning 500 million years of evolution. Non-coding sequences within this region showing >75% sequence conservation in the mammalian species studied were defined as MCSs (Multispecies Conserved Sequences) and, following exclusion of promoters, splice sites and previously undefined exons, four regulatory elements (MCS-R1-4) were defined. These are located within or adjacent to the housekeeping gene *NPRL3*, and correspond to previously noted erythroid-specific DNase1 hypersensitive sites. Their biological significance is supported by the presence of conserved transcription factor binding sites, with both sequence and spacing between binding sites also appearing subject to evolutionary constraint.

The functional significance of the upstream regulatory elements is further emphasized by a series of naturally occurring mutations resulting in α -thalassaemia. Most cases of α -thalassaemia arise from deletions in the *α -globin* genes themselves (denoted α^+ if one of the two *α -globin* genes on a given chromosome 16 is intact, and α^0 if both genes are lost); but a number of rare cases of α -thalassaemia have been described in which both *α -globin* genes *in cis* are intact and where deletions affecting upstream regulatory elements are instead responsible (see chapter 3).

These elements, and remainder of the syntenic region have been very well characterized in terms of the chromatin landscape, with several histone modifications being mapped across the locus and beyond. CTCF binding sites have been identified, as have binding site for the major erythroid-specific and general transcription factors. The three dimensional configuration of the locus in erythroid

.....

and non-erythroid cells, both globin expressing and non-globin expressing has been studied using several methods in both globin expressing and non-expressing developmental stages.

The α -globin locus therefore offers an excellent opportunity to attempt to dissect some of the questions highlighted in this introductory chapter. How do the four putative elements at the α -globin locus control transcription from this gene cluster? By what criteria do they fulfill the definition of enhancers? Do they act alone, in synergy, or redundantly? Are their effects equivalent, or is there underlying diversity in the function of these apparently similar elements?

1.7 Outline of the Thesis

This thesis attempts to build on previously existing studies at the α -globin locus to dissect the contributions of individual regulatory elements to α -globin transcription.

To address the questions above, I shall first present the characterization of a rare naturally occurring upstream deletion in the regulatory region of the α -globin locus that has enabled us to revise an established view of enhancer function at this locus. This permits a critical re-evaluation of the models currently available for the analysis of transcriptional regulation at the α -globin locus, and highlights the current state of our understanding of how enhancers control α -globin expression (chapter 3).

I will describe, and try to justify, our approach to a comprehensive dissection of the *cis* regulatory machinery at the locus, through the generation of a series of single

.....

and double knockout mouse models for each of the α -globin enhancers (chapter 4), before describing the limitations to our assessment of their phenotypes using currently available techniques (chapter 5). Finally, I will report the re-assessment of each model using a newly developed method of nascent RNA quantitation by nanostring, and assess the wider influence of these elements on non-globin genes, and in non-erythroid tissues.

Chapter 2

Materials and Methods

2. Materials and Methods

2.1 DNA procedures

2.1.1 DNA extraction

A standard phenol-chloroform extraction method was used to isolate DNA. For DNA extraction from cell pellets, cells were lysed for 4-6 hours in Tris, 10mM, pH 8.0, NaCl 10mM, EDTA 10mM, and 0.5% SDS, supplemented with proteinase K prior to use. For genomic DNA from mouse tail-clips or ear notch samples, material was first incubated for 6-8 hours in 'tail lysis buffer' (50mM Tris pH 8.0, 100mM EDTA, 100mM NaCl and 1% SDS, supplemented prior to use with 50µg proteinase K per ml of lysis buffer). DNA was precipitated with 4M NaCl and 100% ethanol, before washing with 70% ethanol and redissolving in TE buffer.

2.1.2 Plasmid DNA extraction

Plasmid DNA was isolated from overnight bacterial cultures using the Qiagen Quicklyse Miniprep kit (Qiagen 27405), which employs the alkaline lysis technique for separating plasmid and chromosomal DNA. Bacterial cells from 2ml of overnight culture were pelleted for each miniprep.

Larger volume plasmid preparations were made using the Qiagen maxiprep kit (Qiagen 12162). 200ml of Luria broth (LB) is inoculated with up to 2ml of starter culture in log phase growth. Pelleted cells are then subject to alkaline lysis, and plasmid DNA bound to an anion exchange tip prior to medium salt washes and high salt elution. The final plasmid yield is precipitated with isopropanol and redissolved in TE.

2.2 RNA procedures

2.2.1 RNA extraction

This was performed using Tri-reagent (Sigma T9424) according to the manufacturer's protocol. Briefly, 5×10^6 cells were lysed in tri-reagent by pipetting with a P1000 tip. Following a five minute incubation at room temperature to permit complete dissociation nucleoprotein complexes, 0.2ml of chloroform was added per ml of tri-reagent and mixed vigorously. Following centrifugation at 3200RPM for 30 minutes at 4°C, the aqueous phase was added to 0.5ml of isopropanol per ml of tri-reagent. RNA was precipitated following a 10 minute incubation at room temperature by centrifugation at 3200RPM for 10 minutes at 4°C. RNA pellets were then washed

.....

with 70% ethanol in DEPC treated water, air dried for ten minutes and redissolved in DEPC treated water.

2.2.2 DNase treatment

Turbo DNA-free (Ambion AM1907) was used to minimize DNA contamination of RNA samples, according to the manufacturer's protocol. 2-3U of recombinant DNase I were added to a 50µl RNA sample, supplemented with 10X DNaseI buffer. Samples were incubated at 37°C for 30 minutes before 0.1 volume of inactivation reagent was added, and the sample centrifuged at 10000xg for 1.5 minutes. The supernatant, containing treated RNA was quantified using the Nanodrop spectrophotometer (Thermo scientific).

2.2.3 cDNA synthesis

Superscript III (Invitrogen 18080-093) was used for first strand cDNA synthesis. 1µg of total RNA was used per reaction. To this, 150ng of random primers (Invitrogen 48190-011) and 1µl of 10mM dNTP mix were added and heated to 65°C for five minutes before placing immediately on ice for a further five minutes. 4µl of first strand buffer, 1µl of 0.1M DTT, 1µl of RNaseOUT (Invitrogen 10777-019) and 1µl of Superscript II reverse transcriptase were added and mixed by pipetting. Samples were incubated at 25°C for 5 minutes; 50°C for 60 minutes; then inactivated at 70°C for 15 minutes. Each 20µl reaction was diluted with sigma water to 200µl for real-time PCR reactions.

2.2.4 Metabolic labelling and isolation of nascent RNA

The protocol for pulldown of newly synthesized RNA from fetal liver cultures was adapted from protocols published by the Regev (Rabani et al. 2011) and Dolken groups (Friedel & Dölken 2009). 4-thiouridine, a nucleoside analogue, is added to liquid cultures of primary erythroid cells at a final concentration of 500 μ M, and the culture is incubated for 30 minutes at 37°C, 5% CO₂. Following incubation, the cells are pelleted by centrifugation at 1000RPM for 5 minutes, and washed in PBS. The cell pellet is lysed in tri-reagent, and a standard RNA extraction is performed.

To isolated labelled RNA, 100 μ g of total RNA is incubated with 200 μ l HPDP biotin (Pierce, EZ-Link Biotin-HPDP, 21341, made up to 1mg/ml in dimethylformamide), and 100 μ l of Biotinylation buffer (100mM Tris pH 7.4, 10mM EDTA) in a 1ml reaction, for 1.5 hours at room temperature with rotation. The pyridylthiol biotin reacts with reduced thiols to form reversible disulphide bonds between the labelled transcript and the biotin tag.

Unbound biotin label is removed with a chloroform extraction, and the biotinylated RNA is then incubated with streptavidin magnetic beads (mMacs from Miltenyi 130-074-101) for 15 minutes at room temperature with rotation. The streptavidin/biotin-bound transcripts are then washed through mMACS columns, with 3x 65°C washes and 3 room temperature washes with wash buffer (100mM Tris pH7.5, 10mM EDTA, 1M NaCl and 0.1% Tween20). Labelled RNA is cleaved from the biotin/streptavidin bound magnetic beads by 2x elution phases using freshly made

.....

100mM dithiothreitol. RNA is eluted directly into the lysis buffer of the Qiagen RNeasy RNA clean-up kit. The manufacturer's instructions are followed to yield labelled RNA only in 20µl of DEPC water.

2.2.5 RNA quality control

Representative RNA samples used for downstream analysis were subject to quality control using the Agilent Bioanalyser. For total RNA samples, the manufacturer's protocol was followed for the RNA nano-chip. For labelled RNA, which is obtained in much smaller quantities, the Pico-chip system was used. For further details of the use of the bioanalyser with nascent samples, see also chapter 5.

2.3 Cloning techniques and bacterial culture

2.3.1 Assembly of targeting constructs

Genomic coordinates for the regions to be deleted are given throughout from Build 37 (mm9) of the NCBI and the Mouse genome Sequencing Consortium, and were accessed using the UCSC genome browser. Sequences were imported into MacVector software programme (MacVector Inc. 2010). The incorporated software packages for PCR and sequencing primer design (Primer3) were used throughout. PCR amplification of homology arms was performed using Roche Expand High Fidelity Taq (Roche 04 743 725 001). A typical 50µl reaction would comprise: 5µl 10X buffer, 2.5µl dNTPs at 10mM, 1µl of each of forward and reverse primers at 20pM/µl, 0.75µl of polymerase, 100ng of genomic DNA template, and sigma water to 50µl. Cycling conditions were used according to the manufacturer's protocol.

2.3.2. Cloning into shuttle vectors

PCR products were subject to purification using the Qiagen PCR purification column kit (Qiagen 28104) before ligation into the shuttle vector pGemT-easy (Promega A1360). Ligations were typically performed at 4°C overnight.

2.3.3 Transformation of competent cells

Heat shock transformation of competent cells (DH5 α subcloning efficiency, Invitrogen 04 743 725 001) was performed by incubation of 5 μ l (~10ng) of DNA with 50 μ l of competent cells on ice for 30 minutes followed by a 45 second heat shock at 42°C. Cells were then incubated on ice for a further two minutes before being incubated in 900 μ l SOC medium (super-optimal broth with catabolite repression, Invitrogen 15544-034) for 1.5 hours at 37°C with shaking. Cells were plated out at varying concentrations on agar supplemented with ampicillin and grown overnight at 37°C.

2.3.4 Cultures and archives of *E Coli*

Single colonies were picked under aseptic conditions and grown in 4ml of Luria Broth (stocks prepared and autoclaved by Sue Harper, Molecular Haematology Unit). All inoculated cultures were grown at 37°C overnight with shaking at 220RPM.

Appropriate cultures were stored as glycerol stocks (0.6ml of culture mixed with 0.4ml of 100% glycerol in sterile cryovials, stored at -80°C

2.3.5 Gel extraction of DNA fragments

pGemT-easy cloned fragments which showed no PCR amplification errors by sequencing were subject to digestion with appropriate restriction enzymes to excise the amplified sequences. Digested plasmids were run on a 0.8% agarose gel, and the required fragments were isolated and purified using the Qiagen Gel Purification kit (Qiagen 28704) according to manufacturer's protocol.

2.3.6 DNA ligation

Gel purified fragments were ligated into targeting vectors by blunted ended ligation. Linearized plasmid had its 5' phosphate removed using Roche rapid alkaline phosphatase (04898133001) prior to incubation with insert DNA in a molar ratio of 1:3, with T4 ligase at 16°C overnight.

2.3.7 Validation of targeting constructs

Completed constructs were validated by a series of diagnostic restriction enzyme digests, followed by complete sequencing using primers designed at approximately 300bp intervals around the complete plasmid. Sequencing primers were designed using the 'Primers' features in MacVector.

2.4 Handling mammalian cells

2.4.1 Primary erythroid culture (Fibach culture)

The two-phase liquid culture from 50ml of venous blood was performed by J Sloane-Stanley, according to published protocols (e.g. Fibach et al. 1989). Briefly, mononuclear cells are isolated from peripheral blood by centrifugation on a

.....

histopaque gradient. Cells are then cultured in DMEM supplemented with L-Glutamine 2mM, 10% fetal calf serum, 1mg/ml cyclosporin A and 10% conditioned medium from cultures of 5637 bladder carcinoma cell lines (ATCC) for 6 days at 37°C, with 5% CO₂. For phase II, non-adherent cells are harvested by aspiration and washed in DMEM, before resuspending in an equal volume of phase II medium (DMEM plus 30% FCS, 1U/ml recombinant human erythropoietin, beta mercaptoethanol 1x10⁻⁵M, 1µl per ml culture, dexamethasone 1x10⁻⁶M final concentration, holo-transferrin 0.3µg/ml final concentration, recombinant human stem cell factor 10ng/ml, and 1% final concentration of BSA). Samples for analysis were taken at the intermediate Fibach stage (day 8-11 of phase II).

2.4.2 Mouse embryonic stem cell targeting

129/Ola mouse embryonic stem cells were cultured without feeder layers on gelatinized plates, in accordance with standard protocols. 100µg of linearized targeting construct at 1mg/ml was electroporated into 1x10⁸ ES cells, as a single cell suspension in PBS, using the Biorad Gene Pulser (0.8kV, capacitance 3mF, time constant 0.1s). Electroporated cells were plated out at differing concentrations on gelatinised 10cm cell culture plates. Positive selection with G418 was added after 48 hours, with fresh selection medium added every 48 hours. Negative selection with FIAU commenced at 96 hours. Doubly resistant colonies were manually isolated using a P20 Gilson pipette under direct visualization and plated into U-bottomed 96-well plates. Individual clones were then trypsinized, disaggregated and transferred to gelatinized 96 well plates for expansion in selective medium. Once nearing confluence, colonies were again trypsinized and split between two replica plates:

.....

one expanded for DNA analysis by Southern blot, and one frozen for selection of correct clones for karyotyping and blastocyst injection. For cryostorage, cells were suspended in freezing medium (complete medium +10% DMSO and 20% FCS) and immediately transferred to -70°C. After 24-48 hours at -70°C, cells were transferred in dry ice to a liquid nitrogen cryopreservation bank.

2.4.3 Karyotyping of targeted ES cells

Karyotyping was performed according to standard protocol. Colcemid is added to cells at 0.1µg/ml of medium, and incubated at 37°C for 2-2.5 hours. Cells are then washed twice with PBS and trypsinized, and the washed pellet resuspended in 5ml of 0.075M KCl at 37°C, and incubated at room temperature for 10 minutes. The cells are centrifuged at 600RPM for 5 minutes and resuspended in 1ml of ice-cold fixative solution (3:1 methanol: acetic acid). After a further 600RPM 5minute centrifugation step, cells are again resuspended in approximately 500ml of cold fixative solution. The cell suspension was held in a Pasteur pipette, 10-20cm above a glass slide: individual droplets discharged from this height formed metaphase spreads, which were stained with Giemsa using the Hematek 1000 automated staining instrument. 40 metaphase spreads were examined for each clone to be injected.

2.4.4 In vitro Cre recombination in targeted ES cells

Corrected targeted clones were expanded to 1×10^8 cell/ml. Electroporation with 100µg Cre plasmid was performed as above. Colonies were isolated and replica plated prior to commencement of G418 selection. Clones susceptible to G418 selection were deemed to have undergone appropriate Cre recombination and were

.....

therefore expanded from replica plates. PCR screening to establish the nature of the recombination event was followed by formal Southern blot characterisation.

2.4.5 Fetal liver cultures

This two phase culture system allows for the expansion of primary erythroid cells from murine fetal livers. An initial expansion phase permits cell numbers in excess of 1×10^8 to be obtained from single fetal livers, while the subsequent maturation phase permits the semi-synchronous maturation of erythroid progenitors to mature red cells.

The livers of mice at embryonic day 12.5, 13.5 or 14.5, according to the requirement of the protocol, were disaggregated using a P1000 tip in 2ml of Stempro medium (Invitrogen 10640-019) plus nutrient supplement, L-glutamine 2mM, erythropoietin 1U/ml, SCF 50ng/ml and dexamethasone ($1 \mu\text{M}$ final concentration). Cells were kept at $1 \times 10^6/\text{ml}$ of supplemented StemPro medium during the expansion phase, which may be continued for up to six days without spontaneous maturation. Maturation was permitted once sufficient numbers of cells had been obtained during the expansion phase, by pelleting cells at 1000RPM for 5 minutes, washing in PBS and transferring to Stempro medium supplemented with 5U/ml erythropoietin and transferrin $0.5 \mu\text{g}/\text{ml}$ only.

2.4.6 Megakaryocyte expansion from fetal livers

Livers from E14.5 fetal mice are isolated and disaggregated using a P1000 tip. Enrichment for CD117+ progenitor cells is performed using anti-CD117 magnetic

.....

beads (Miltenyi Biotec) and LS positive selection Macs columns (Miltenyi). Enriched cells are passed through a 70µm cell filter and suspended at 1×10^6 /ml of IMDM with recombinant murine SCF and TPO. After 5-7 days of expansion, cells are pelleted by centrifugation at 1000RPM for 5 minutes, and washed three times with unsupplemented IMDM, before transferring to IMDM supplemented with TPO only. Cells are allowed to mature in IMDM-TPO for a further 5 days. Megakaryocytes are isolated by lineage depletion (BD Biosciences) and CD41 positive selection on either Macs Large Cell columns (Miltenyi Biotec) or FACS. The purity of all cells obtained was ascertained by FACS for CD41 positivity, Ter119 and c-kit negativity.

2.5 Animal handling and obtaining samples for analysis

All mouse handling was performed by Dr Jackie Sharpe, Ms Jackie Sloane Stanley, Ms Christina Rode and Mrs Sue Butler from the Molecular Haematology Unit Transgenics Group, according to Home Office regulation and under an approved licence. Where necessary, mice were sacrificed by cervical dislocation and embryos were sacrificed by decapitation.

2.5.1 *Blastocyst injection and chimera generation*

This was performed by C. Rode and J Sloane-Stanley. ES cells from correctly targeted clones were injected into day 3.5 blastocysts from C57BL/6 mice and implanted into pseudo-pregnant females. Chimeric male offspring, once obtained, were mated with C57BL/6 females to generate F1 litters with germline transmission, with a view to crossing with GATA/Cre expressing mice ($\Delta R1$ only) to yield mice heterozygous for the knockout allele.

2.5.2 Harvesting erythroid tissue from mouse models

For chromatin immunoprecipitation and chromosome conformation capture, erythroid material was obtained from both fetal liver cultures and adult mouse spleen. For the latter, mice were treated with phenylhydrazine (APH) by intraperitoneal injection on three occasions over days 1 and 2 to induce an acute haemolytic anaemia and expand extramedullary erythropoiesis. On day 6 following treatment, the mouse is sacrificed and the enlarged spleen removed.

Erythroid cells are isolated from APH treated spleens by incubation with Ter-119-PE (BD biosciences) and anti-PE microbeads (Miltenyi biotec), prior to positive selection using an AutoMacs cell separator. Whether used for ChIP or 3C, selected cells were returned to a 37°C incubator in RPMI+10% FCS after selection for at least 30 minutes before downstream experiments. Positive and negative fractions were examined for appropriate enrichment by FACS analysis.

2.6. Chromatin techniques

2.6.1 Chromatin immunoprecipitation

This was performed using antibodies obtained from Santa Cruz (RNA PolII), Upstate (H3K4me3 and pan-H4Ac), and from Dr Catherine Porcher, Molecular Haematology Unit, WIMM (SCL). Chromatin was cross-linked with 1% formaldehyde for 10 minutes and neutralised with 1M glycine prior to washing in cold PBS, and cell lysis in SDS lysis buffer (Millipore ChIP kit, 17-295). Cross-linked chromatin was sonicated using the Diagenode Bioruptor with a pulse duration of 30 seconds/minute for 12 minutes. A pre-clearing phase with protein A agarose beads preceded overnight incubation

.....

with antibody, at empirically determined concentration. After an overnight incubation, antibody-bound chromatin fragments are isolated by incubating with protein A agarose beads for one hour. Washes of increasing stringency (low salt, high salt, lithium chloride immune complex buffers) remove non-specific binding, and chromatin is then eluted with 1% SDS, 0.1M NaHCO₃. Cross-links are reversed at 45°C with simultaneous proteinase K treatment to yield DNA, which is then phenol-chloroform extracted and ethanol precipitated. Immunoprecipitated DNA is analysed by qPCR, hybridized to a custom tiled array, or used for high throughput sequencing (ChIP-seq).

2.6.2 Chromosome conformation capture

Chromatin is cross-linked with 2% formaldehyde for 10 minutes, neutralised with 1M glycine, and washed and lysed as for ChIP. The SDS lysis buffer is neutralised with 20% triton X prior to addition of 400U high concentration HindIII restriction enzyme and overnight digestion. Cross-linked and digested chromatin is then ligated using T4 ligase and incubated at 16° for four hours. 600U (30µl) of proteinase K is added and crosslinks are reversed overnight at 65°C. The interacting fragments are extracted using a standard phenol chloroform DNA extraction, and ligated fragments are analysed by qPCR, using primers and probes previously designed by Dr D Vernimmen (Vernimmen et al. 2009, Vernimmen et al. 2007).

2.7 Expression analysis

2.7.1 Blood counts, cytopins and blood films

.....

Blood was obtained from adult (>10week old) mice by tail venesection where the mouse was to be preserved, or by aspiration of great vessels following sacrifice by cervical dislocation. Blood was collected into heparinized capillary tubes and full blood counts were assayed using the Horiba Medical Scil Vet abc Plus+ instrument.

Blood smears were obtained from a single drop of heparinised blood spread across a glass microscope slide. Approximately 50,000 cells in a volume of approximately 200 μ l were used to prepare cytopins, and were centrifuged for 4 minutes at 300RPM in the Shandon Cytospin 4 instrument. Both films and cytopins were stained with a Wright Giemsa stain using the Hema-tek 100 automated staining instrument.

2.7.2 Real-time PCR

Both Taqman and Sybr green methods were used for real time PCR analysis. For the former, reactions comprised 12.5 μ l Taqman Universal Mastermix, 1.25 μ l 20X primer mix, and template DNA, with Sigma water to 25 μ l final volume. Primers for α and β globin (both human and murine) and for control genes ribosomal 18S and β actin, were commercially available from ABI. Custom primers (e.g. $\alpha\alpha$ (ALT), see chapter 3) were designed using the Primer express software from Applied Biosystems.

Reactions were set up in duplicate and were performed on the Applied Biosystems Sequence Detection System 7000 thermocycler; data were analysed using the intrinsic software provided by Applied Biosystems.

.....

Real time PCR assessment of unspliced RNA was performed using Sybr Green amplicons previously designed and available as laboratory reagents.

2.7.3 RNA Fluorescence in situ hybridisation

This was performed by Dr Jill Brown and Dr Veronica Buckle, who have kindly provided the following experimental details. Poly-L-lysine coated slides were prepared in RNase-free conditions. Aliquots of 1×10^6 cells were spotted onto slides and allowed to settle for 90 seconds, before fixing in 4% formaldehyde, 5% glacial acetic acid in saline for 20 minutes. Slides were washed three times in PBS for 5 seconds. For hybridisation, slides were rinsed in tri-saline (0.15M NaCl, 0.1M tri-HCl pH7.5), then placed in pepsin (Sigma P7000) with HCl (0.01M pH2) at 37°C for 4 minutes. After incubation, slides were washed in RNase free water, fixed in 3.7% formaldehyde for 5 minutes, and washed again in PBS for 10 minutes. Slides were then dehydrated using solution of ascending concentrations of ethanol, and air dried. 50ng of DIG labelled oligo probes for alpha and beta globin in 12µl of hybridisation buffer were then spotted onto each coverslip, mounted onto the slides and hybridised overnight at 37°C.

After overnight incubation, slides are washed in 2xSSC at 37°C for 15 minutes, and 100ml of blocking solution (1.35% blocking powder in tris-saline) added to the slide in a humid chamber for 30 minutes at room temperature. Appropriate fluorophore conjugated antibodies are prepared in blocking solution, and, once slides were washed in Tris-saline/Tween (0.15M NaCl, 0.1M Tris-HCl pH7.5, 0.05% Tween 20), the antibodies were applied and incubated at room temperature for 30 minutes.

.....

Slides were mounted in 40µl of vectashield/DAPI and visualised on the Olympus BX51 microscope. Images were obtained using the Radiance 2000 Confocal device and Laser Sharp software (Biorad).

2.7.4 Nanostring RNA quantitation

5µl of DNaseI treated RNA was added to 10µl of capture codeset, 10µl of hybridisation buffer, and 5µl of reporter codeset and gently mixed by pipetting. The hybridisation mix was incubated in a thermal cycler at 65°C for 12-30 hours until transfer to the NCounter prep station. Here, excess and unbound capture and reporter probes are washed off using an automated two step magnetic bead based purification method, and transcripts to be counted are applied to the streptavidin coated NCounter cartridge. The cartridge is then transferred to the NCounter data acquisition module, in which up to 1155 fields of view are counter per flow cell, to yield hundreds of thousands of counts of individual target transcripts.

2.8 Other techniques

2.8.1 Bisulphite sequencing and nested PCR amplification

Bisulphite conversion was performed according to the EZ DNA Methylation Gold kit standard protocol (Cambridge Biosciences) with subsequent nested PCR performed with HotStart Taq (Qiagen). All nested PCR primers were available as pre-existing lab reagents, and are shown in appendix 1.

2.8.2 FACS analysis

.....

This was principally performed by Jackie Sloane Stanley. Up to 1×10^6 cells were resuspended in FACS buffer (PBS+10% FCS) and incubated for 20 minutes at 4°C with the appropriate fluorophore conjugated antibody. FACS analysis was performed using the Beckman Coulter CyAn.

2.8.3 Southern blotting

Southern blotting was typically performed with 10µg of genomic DNA, digested overnight in a 50µl reaction with up to 1/10 volume of appropriate enzyme and 1µl of 50mM spermidine. Digested DNA was run on a 0.8% TAE gel overnight at 30-40V and visualised after ethidium staining. Gels were then denatured in 1M NaOH for 1-2 hours, and neutralised in a solution of 3M NaCl, 1M Tris-HCl pH7.5 for 2-4 hours. Gels were blotted overnight in 6xSSC using a zeta-probe nylon membrane (Biorad). Membranes were aired dried for 20 minutes before crosslinking in a UV Stratalinker (Stratagene) and baking at 80°C for three hours to fix DNA to the membrane.

Southern blot probes were labelled according to the Amersham Multiprime kit manufacturer's protocol. Briefly, 50ng of probe was denatured at 95°C with random primers, and end-labelled using 1µl of [α -P³²]dCTP (Perkin-Elmer) for 10 minutes at 37°C. Unincorporated label was removed using an Illustra microSpin G-50 column (GE healthcare), and the labelled probe was denatured at 95°C for five minutes before addition to 10ml of PerfectHyb hybridisation buffer, preheated to 65°C. Membranes for hybridisation were briefly soaked in 2xSSC before a 5 minute pre-hybridisation in 5ml PerfectHyb buffer without probe, at 65°C. Hybridisation buffer

.....

with label is added to this mix, and hybridisation permitted for a period from 4h – overnight.

Hybridised membranes are washed for 10 minutes in room temperature 2xSSC, then 0.1xSSX, 0.1%SDS heated to 58-65°C (probe dependent, empirically determined) for a further 10 minutes, with shaking. Washed membranes are exposed with Kodak autoradiographic film for 2-7 days, and processed using the Kodak Xograph processor.

2.8.4 Colony blotting

The colony blotting technique used was developed by Prof T Rabbitts (Buluwela et al. 1989). Hybond N membranes are labelled and firmly applied to the surface of an agar plate containing the colonies to be blotted. The orientation of the membrane with respect to the plate is indelibly marked by using an inked needle to make an asymmetric pattern of marks through the membrane and agar. Bacterial colonies will peel off the plate with the membrane. The membrane is then dampened with 2xSSC, 5% SDS and microwaved at full power for approximately five minutes. The lysed bacterial colonies will be irreversibly cross-linked to the membrane, and the original plates if placed at 37°C for a few hours will re-establish the identical pattern of colonies. The filters are then hybridized as for Southern blot, with the exception that a 2-3 hour hybridization is sufficient, and short (1-2 hour) exposure will detect correctly targeted bacterial clones. Marking the plates' outlines and orientation marks on the autoradiograph will enable alignment of the original plates with the colonies on the autoradiograph.

Chapter 3

A novel mutation at the *α -globin* locus

3. A novel mutation at the α -globin locus

3.1 Introduction

3.1.1 Naturally occurring deletions at the α -globin locus

Most cases of α -thalassaemia arise from deletions in the α -globin genes themselves (denoted α^+ if one of the two α -globin genes on a given chromosome 16 is intact, and α^0 if both genes are affected). In a number of cases, however, patients with a clinical picture of α^0 thalassaemia have been described in whom both α -globin genes *in cis* are intact; in these cases, alternative explanations for the failure of appropriate transcription of α -globin have been sought, and have provided insights into novel mechanisms of genetic disease. For some of these cases, disruption of the upstream regulatory elements of the α -globin cluster is identified as responsible for the clinical phenotype.

.....

These cases are valuable: they provide a validation for the indirect attribution of regulatory activity to the deleted regions (for example, by sequence conservation or appropriate histone modifications), and do not expose the locus to any potentially confounding experimental manipulation. However, they are also rare. Figure 3.1 defines the number and extent of the regulatory regions mutations identified at the *α-globin* locus over a thirty year period.

The earliest examples, described in the 1980's and 1990's, delineated extensive deletions of the sequence upstream of the *α-globin* genes - see, for example patient RA in figure 3.1, reported by the Higgs lab in 1987 (Nicholls et al. 1987, Hatton et al. 1990). Patients with these deletions have a phenotype identical to that observed with a deletion of both duplicated *α-globin* genes *in cis*: there is a mild anaemia with hypochromic and microcytic red cell indices. The presence of occasional haemoglobin H bodies confirms the imbalance between α and β globin chain synthesis seen with a failure of *α-globin* transcription *in cis* (Phylipsen et al. 2010). This is supported in rare cases by the coinheritance of α^+ thalassaemia (i.e. deletion of the α_2 gene) on one copy of chromosome 16, and an upstream deletion of the regulatory region on the other: in these cases, a phenotype consistent with the loss of three of the four *α-globin* genes, giving HbH disease, is observed.

Evidence that these mutations result in a reduction of *α-globin* transcription *in cis* to basal levels comes from a variety of sources. Key among these has been the use of cell fusion to transfer the affected human chromosome 16 into mouse erythroleukaemia (MEL) cells to generate MEL cell hybrids (Deisseroth & Hendrick

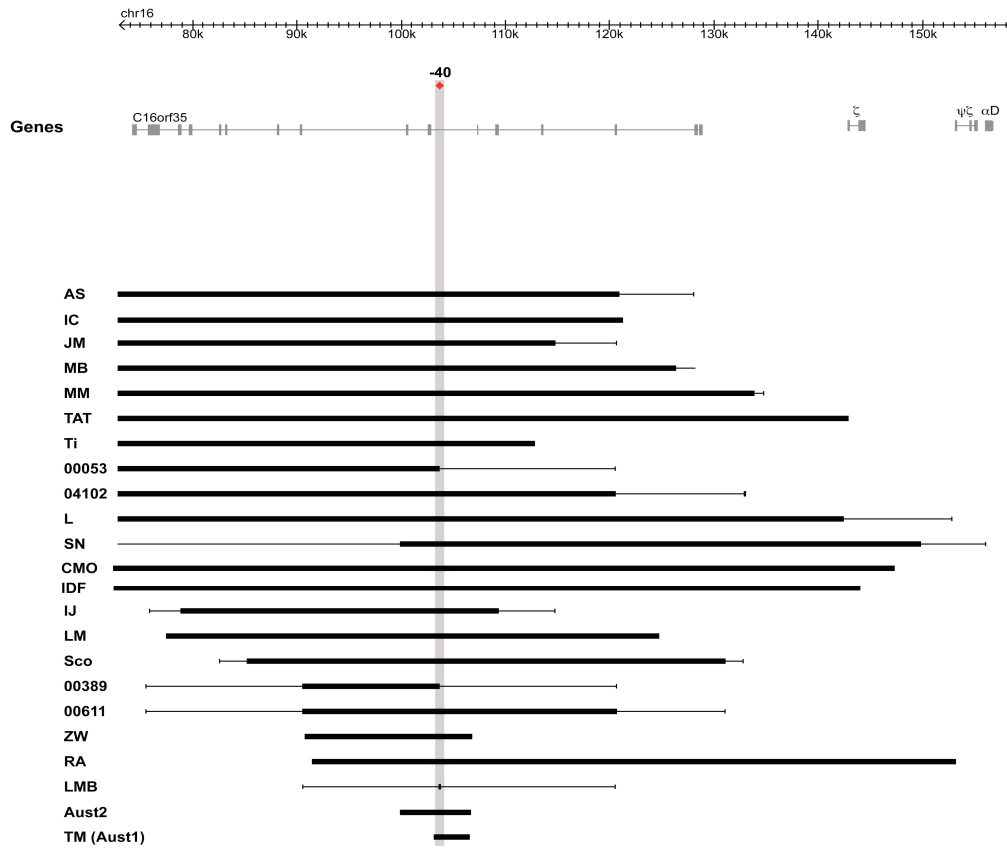


Figure 3.1 α^0 thalassaemia resulting from upstream deletions in the α -globin cluster. Upper track shows chromosome 16 with positions of *NPRL3* and the α -globin genes. The extent of individual mutations is shown by black bars; where uncertainty exists as to the exact breakpoints (due to use of MLPA probes for delineation of the mutation) the greatest possible extent is shown by thin grey bars. All mutations delete HS-40 (MCS-R2), highlighted by the vertical grey bar. Figure prepared by Dr Nicki Gray.

Table 3.1 Haematological characteristics of selected patients with upstream deletions

Patient	Hb	MCV	MCH	$\alpha:\beta$ ratio	HbH bodies	Reference
RA	13.2	64	22	0.71	Y	Hatton et al. 1990
Ro.A	12.0	71	22	0.66	Y	Hatton et al. 1990
ZW	11.3	65	19	0.71	Y	Viprakasit et al. 2006
TM	14	73	24	NS	N	Phylipsen et al. 2010

.....

1978). This method uses adenine phosphoribosyl transferase deficient MEL cells; since APRT is encoded on human chromosome 16, the cultivation of MEL hybrids generated by cell fusion in media in which adenine is the only source of purines permits the isolation of hybrid cells retaining human chromosome 16. Over 80% of cells cultured in this way contain the human chromosome 16, and subsequent work has demonstrated the maintained synthesis of human α -globin in this system on induction of terminal erythroid maturation with dimethyl sulphoxide or hemin and hexamethylenebisacetamide (HMBA).

Each of the upstream deletions described in figure 3.1 has been analysed in this way, through transfer of the affected chromosome to form a mutation-specific MEL hybrids. Mouse and human α -globin mRNA expression can then be quantified by RNase protection assay and real time quantitative PCR to assess the impact of individual upstream deletions. In each case, a complete failure of up-regulation of human α -globin expression was observed from the affected chromosome.

The conclusion is that some or all of the upstream regulatory region encompassed by these deletions within and around *NPRL3* is essential for α -globin transcription. Observational studies using many of the techniques outlined in chapter one have identified several putative regulatory elements within the boundaries of the upstream deletions (figure 3.2). Contrasting figures 3.1 and 3.2 demonstrates how the slowly expanding library of upstream deletions resulting in a similar clinical phenotype has permitted a progressive narrowing and refinement of this region thought to be critical for globin expression. All include the DNaseI hypersensitive

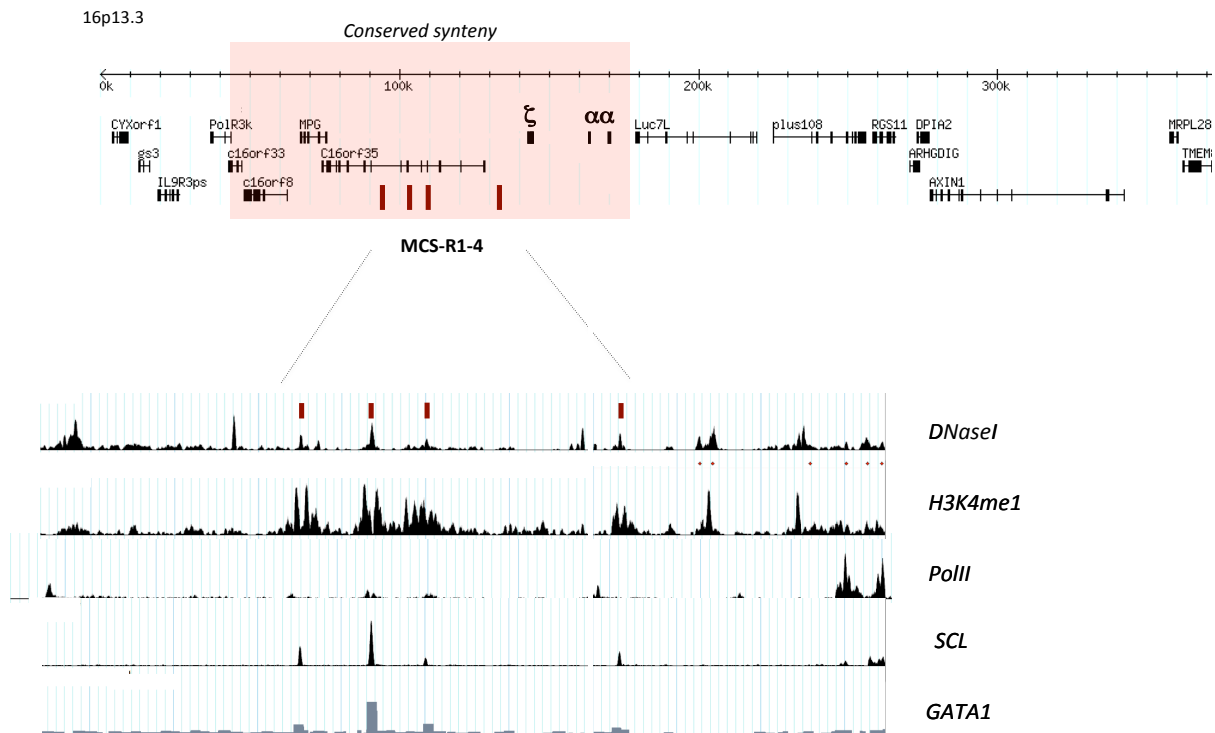


Figure 3.2 Putative regulatory elements across the α -globin locus

The upper panel shows the α -globin cluster and its wider chromosomal context, with region of conserved syntenicity highlighted in pink. The positions of the four conserved non-coding regulatory elements defined by Hughes *et al.* (2005) are shown by red bars.

The lower panel shows representative ChIP-seq tracks for DNase I, enhancer-associated chromatin mark H3K4me1, PolII and erythroid transcription factors SCL and GATA1 for the regulatory region, with the position of regulatory elements again marked with red bars.

.....

site 40kb upstream of the ζ -globin transcriptional start site, HS-40, the second multispecies conserved regulatory element described in (Hughes et al. 2005).

3.1.2 HS-40 as the major regulatory element at the α -globin locus

A series of experimental findings has supported the contention that HS-40 is the major, if not the sole, regulatory element at the α -globin locus. The α -globin genes are not appropriately expressed in transgenic systems in the absence of the upstream regulatory region, providing evidence that some, if not all of the regulatory region is required for normal α -globin transcription. A systematic dissection of the regulatory region, testing small restriction fragments in MEL cell lines, enabled the identification of a region between -37kb and -47kb upstream of the ζ -globin promoter which was able to drive α -globin transcription to a level similar to that seen from the intact α -globin cluster (Higgs et al. 1990). Analysis of this region at a finer scale (Jarman et al. 1991) narrowed the regulatory activity to a region of 350bp over HS-40, complete with GATA-1 and AP-1 transcription factor binding sites.

These studies, in transgenic systems, suggest that the element at HS-40 is sufficient for α -globin expression. Each naturally occurring deletion appears to include HS-40, adding weight to the proposition that it is also critically necessary for α -globin transcription. This question is addressed more precisely by studies in the 'humanized' mouse (Wallace et al. 2007). This model, shown schematically in figure 3.3, was generated using recombinase-mediated cassette exchange to replace the

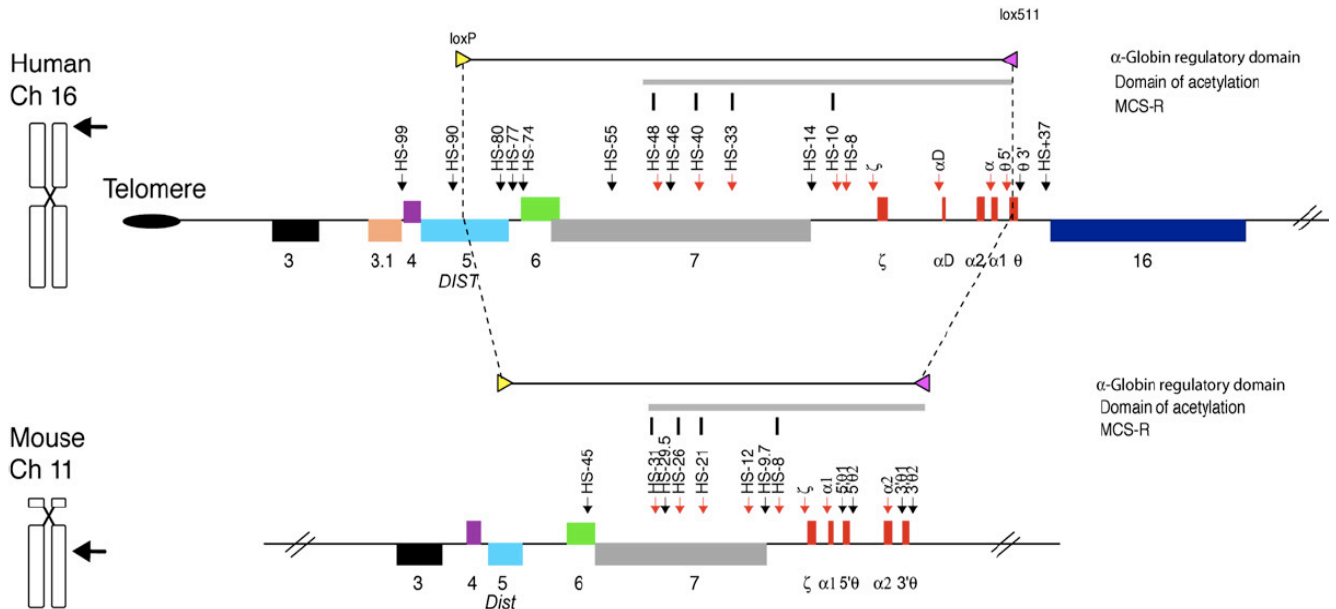


Figure 3.3 Humanized mouse syntenic region. From Wallace et al, 2007

The region of conserved syntenicity between murine and human α -globin loci is shown by a black bar, flanked by *lox* sites, indicating the extent of recombinase-mediated genomic exchange employed to generate the 'humanized' mouse. Erythroid specific hypersensitive sites are shown with red arrows; constitutive DNaseI hypersensitive sites are indicated by black arrows. Genes are denoted thus: 3, *IL9RP3*; 3.1, *POLR3K*; 4, *c16orf33*; 5, *c16orf8*; 6, *MPG*; 7, *NPRL3*; 16, *LUC7L*.

entire murine α -globin cluster with the syntenic human region. Not only does this model express human α -globin coordinately with murine α -globin, but deletion of HS-40 from the humanized locus abrogates the expression of human α -globin. In widely accepted transgenic models, therefore, there is excellent support for the regulatory element at HS-40 being both necessary and sufficient for control of α -globin expression.

With regard to the role of HS-40 in its native setting, two Australian patients described by Phylipsen *et al.* (Phylipsen *et al.* 2010) are of particular relevance. One patient, TM, carries a 3.3.kb deletion encompassing most of the fifth intron of *NPRL3*, removing HS-40 entirely, but also adding a 39 bp insertion of apparently orphan sequence between the breakpoints flanking the HS-40 deletion. The 3' breakpoint is located within an Alu repeat; the 5' breakpoint is not located within any recognized repeat region. The second patient featured a 6.7kb deletion, including HS-40 and extending into the coding sequence of the *NPRL3* gene.

Both of these patients have hypochromia and microcytosis similar to that observed in patients with larger upstream deletions. When the affected HS-40-deleted chromosome from patient TM was incorporated into MEL hybrids no human α -globin mRNA is detected (figure 3.4), consistent with the hypothesis that HS-40 is critically required for α -globin expression. Distinct from previously described and more extensive upstream deletions, HbH bodies were not observed on supravital staining with brilliant cresyl blue in either of the patients with isolated HS-40

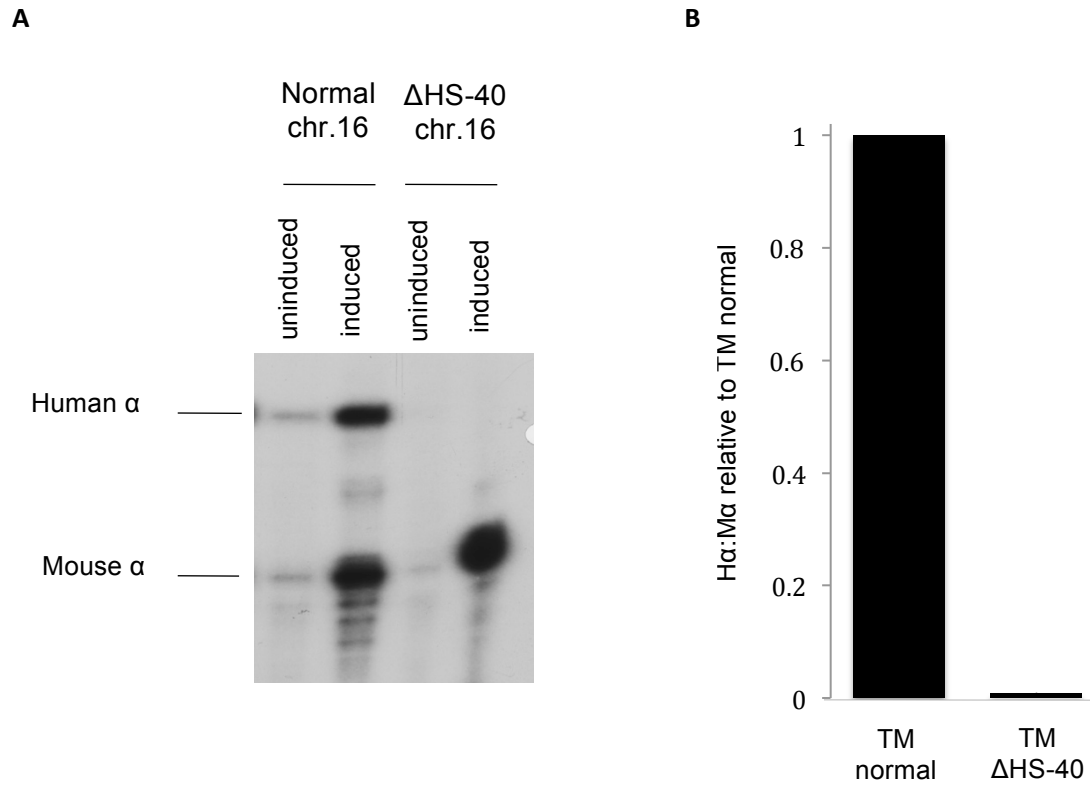


Figure 3.4 Expression of human α -globin in the presence and absence of HS-40, in MEL hybrids

A. RNase protection assay (Dr J Sharpe) contrasting the human and mouse α -globin mRNA from the normal and mutant chromosome 16 from patient TM, heterozygous for the subsequently named $(\alpha\alpha)^{ALT}$ deletion. For each hybrid, assays have been performed on both induced and uninduced MEL cells demonstrating the failure of α -globin transcription in from the mutant chromosome.

B. Real time PCR analysis from the normal and mutant chromosome 16 in MEL hybrids.

.....

deletions described by Phylipsen *et al.* Although striking in the context of the other upstream deletions, it is difficult to draw conclusions from a technique which may be technically troublesome and time consuming, and which is inevitably limited by inter-observer variation.

The deletions shown in figure 3.1 are not described in homozygosity. First, this is unsurprising for such rare alleles; but further, individuals homozygous for these alleles would be expected to have a failure of α -globin transcription from both copies of chromosome 16, and thus to die *in utero* from Hb Barts hydrops fetalis. In all but the most isolated deletion of HS-40 (patient TM) disruption caused to the coding sequence of the gene *NPRL3* in homozygosity would also be expected to be incompatible with survival (Kowalczyk, Hughes, Babbs, et al. 2012a): mice lacking the orthologous gene *nprl3* die *in utero* from a combination of cardiac defects, gastroschisis, and generalized midline dysraphism.

3.1.3 Conclusions from previous studies, and state of the literature

Data from naturally occurring mutations and their validation in heterologous transgenic systems led to the logical conclusion that HS-40 is critically required for α -globin expression, and that none of the other apparent regulatory elements are able to compensate for its absence. This tenet has become entrenched in the literature of transcriptional regulation – with the α -globin locus frequently cited in contrast to LCR-controlled transcription at the β -globin locus.

.....

However, there remains the caveat that all the supporting experimental studies have been performed in transgenic systems: with the human chromosome incorporated into the MEL hybrid cell line, and the human *α -globin* cluster incorporated into the mouse chromosome 11. Although the data supporting this conclusion appear firm, there are a few points that stand out as requiring explanation:

1. As reported by Anguita and colleagues in 2002, mice homozygous for deletion of the element equivalent to HS-40, namely HS-26 (figure 3.5), survive to adulthood, breed normally, and have only non-significant trends toward microcytosis and hypochromia (figure 3.6) (Anguita 2002). Although there are differences between the murine and human *α -globin* loci that might be proposed to explain this discrepancy (e.g. the presence of a murine specific regulatory element at HS-12) the apparent difference in regulatory element function is surprising in the context of the close conservation of regulatory machinery between the mouse and human *α -globin* loci.
2. Not all mutations in the locus appear to be completely recapitulated in our transgenic experimental models. For example, the Vanuatu mutation (a substitution creating a new GATA site upstream of the *α -globin* promoters, which is capable of co-opting enhancer activity from HS-40 and thus reducing *α -globin* transcription [De Gobbi et al. 2006]) does not reduce *α -globin* expression in the MEL hybrid system.

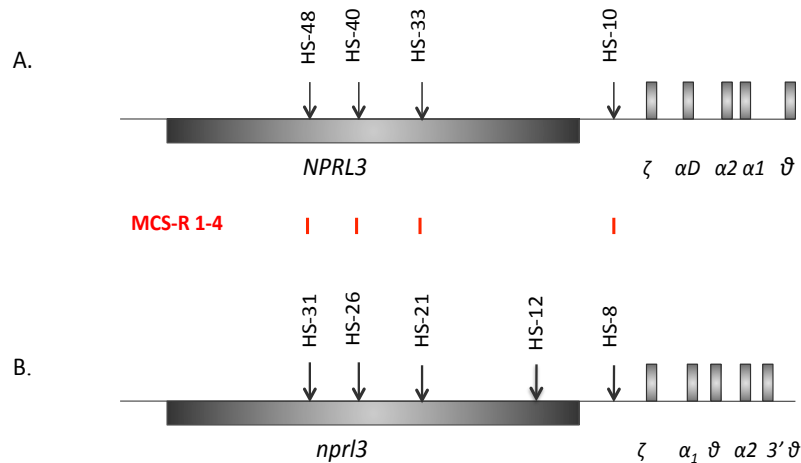


Figure 3.5 A schematic representation of the organisation of the α -globin locus in human (A) and mouse (B). The locations of erythroid specific DNase sensitive elements in the introns and immediate 5' region of the *NPRL3* gene are shown by arrows; those corresponding to multispecies conserved regulatory sequences 1-4 are highlighted by red bars. The α -globin genes are shown by grey blocks.

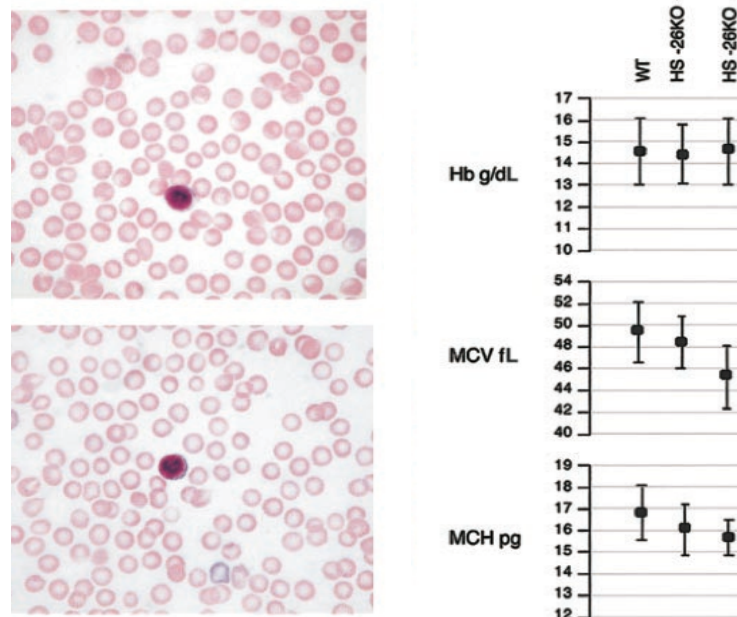


Figure 3.6 Deletion of the murine equivalent of HS-40 (HS-26) has an unexpectedly mild haematological phenotype (figure from Anguita et al., 2002)

Top left: Peripheral blood smear from a wild type mouse

Bottom left: Peripheral blood smear from a Δ HS-26 mouse,

Right: Haemoglobin and red cell indices for the Δ HS-26 mouse, demonstrating only non-significant trends toward microcytosis and hypochromia.

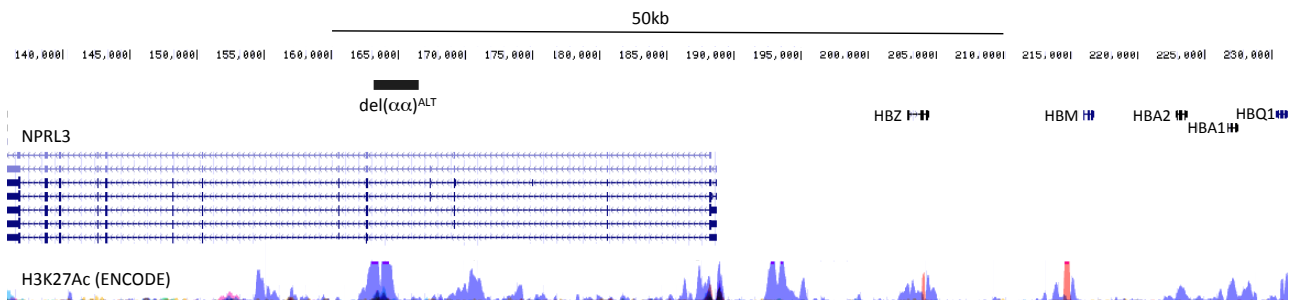
3. Mutations of ATRX, associated with α -thalassaemia in human patients, do not reduce α -globin expression in the humanized mouse model (Higgs and Gibbons groups unpublished data).

To these outstanding issues must be added the hitherto poorly explained lack of HbH bodies in the blood of patients with isolated HS-40 deletions; and, more fundamentally, if HS-40 genuinely is the only element required for full α -globin expression, what are the roles of the remaining three highly conserved regulatory elements at the human locus?

3.2. An isolated deletion of HS-40 in homozygosity

MC is a Portuguese patient who is homozygous for an isolated deletion of HS-40 termed $(\alpha\alpha)^{\text{ALT}}$ by Paula Faustino and colleagues (Coelho et al. 2010). The $(\alpha\alpha)^{\text{ALT}}$ deletion removes nearly the complete 5th intron of *NPRL3*, including the whole of HS-40, but leaves the coding sequence of *NPRL3* intact. Coelho and colleagues defined the breakpoints of the mutation (chr16:163191 to chr16:166555, using build hg19), and also detected the presence of a 39 base pair orphan sequence - also present in the earlier allele reported in patient TM from the series of Western Australian cases of α thalassaemia (Phylipsen et al. 2010) (Figure 3.7). The identical genetic description of these two independently reported patients strongly suggests a common origin, though this has not been further investigated. However, of critical interest in the case of MC is the presence of the mutation in homozygosity.

A.



B.

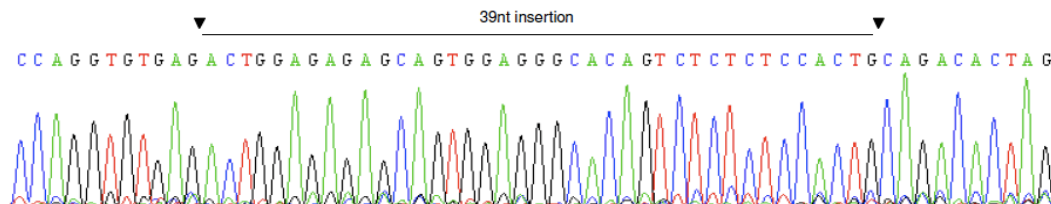


Figure 3.7 The ($\alpha\alpha$)^{ALT} mutation

A. The location and extent of the mutation on the UCSC genome browser. The H3K27 acetylation ENCODE track demonstrates that the mutation deletes the regulatory element HS-40, but leaves the coding sequence of *NPRL3* intact.

B. The 39 base pair insert of apparently orphan DNA between breakpoints which delete 3.3kb sequence in the 6th intron of *NPRL3* (from Coelho *et al*, 2010). An identical insert between identical breakpoints is identified in the previously identified patient TM (figure 3.4)

.....

If previous conclusions regarding the exclusive ability of HS-40 to drive α -globin transcription were true, the expected failure of α -globin expression from both copies of chromosome 16 in MC would have resulted in death *in utero*; MC is, however, fit and well, and indeed has never needed medical support for thalassaemia.

Patient MC's kindred is shown in figure 3.8, along with his haematological parameters, peripheral blood smear and haemoglobin H preparation. MC has a clinical phenotype consistent with HbH disease, leading to the unorthodox conclusion that the locus must be capable of synthesizing significant amounts of α -globin even in the absence of HS-40. MC's daughter has a milder haematological phenotype, consistent with that previously found in patient TM who also bears the identical mutation in heterozygosity (Phylipsen et al. 2010).

Several possible hypotheses may explain the unexpected findings. If HS-40 genuinely is the only element capable of driving α -globin expression, we would expect to find some other abnormalities present in patient MC:

- HS-40 might be transposed to a different location within the genome.
- A new active element may be formed by the breakpoints of the deletion, or by the orphan sequence between the breakpoints, to compensate for the loss of HS-40.
- An entirely new regulatory feature maybe formed elsewhere, capable of bearing the role of HS-40.
- There could be coincidental alterations in the other conserved elements that allow them to assume an active regulatory role.

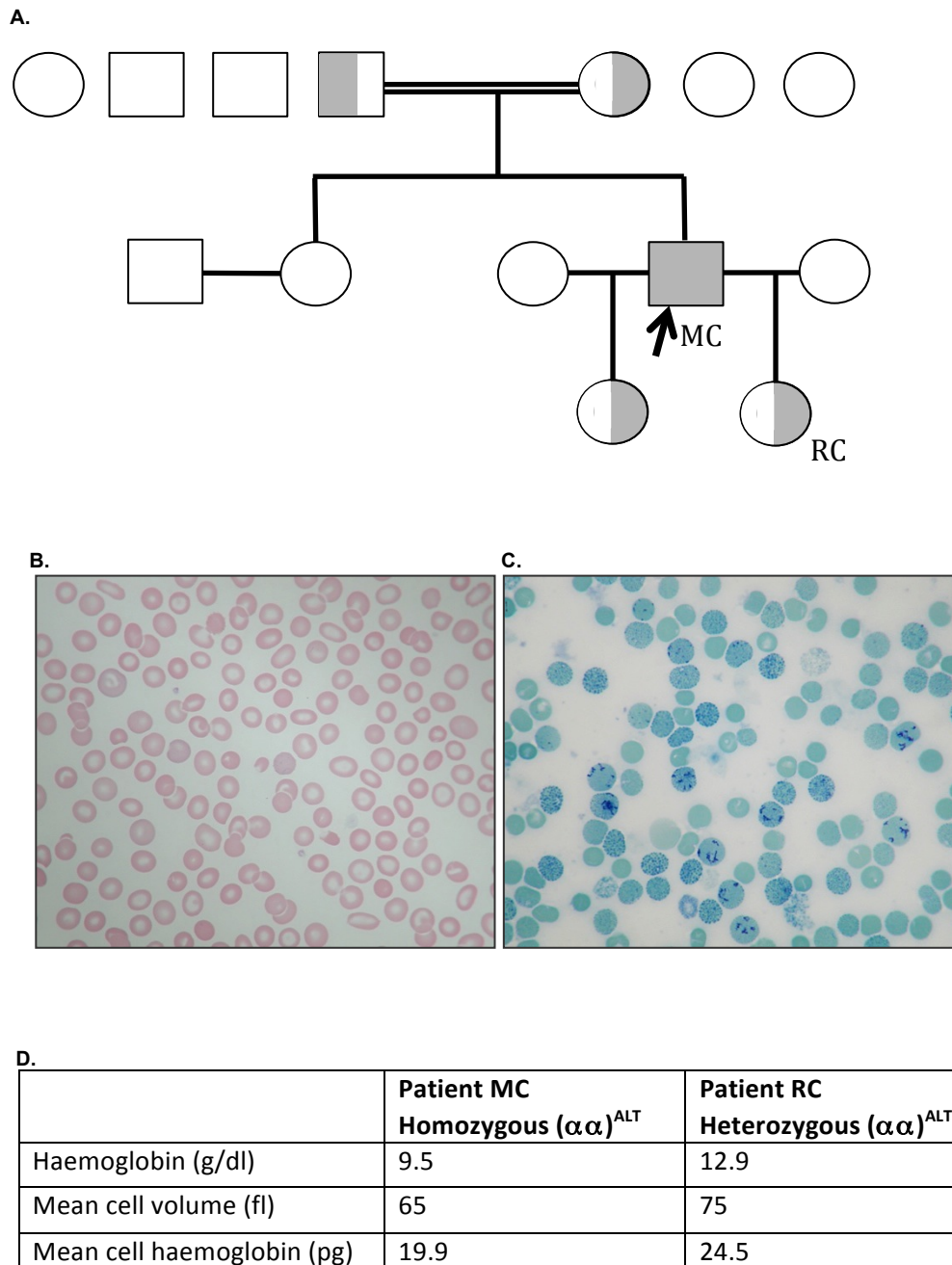


Figure 3.8 $(\alpha\alpha)^{ALT}$ mutation in homozygosity – clinical data

A. Kindred showing patient MC (arrowed) and his heterozygote daughters. Both daughters are alive and well, and there is no haematological morbidity in the wider family.

B. Peripheral blood smear, Giemsa stained; and **C.** peripheral blood HbH preparation, brilliant cresyl blue stained, from patient MC. Peripheral blood erythrocytes show anisocytosis and poikilocytosis with some irregularly contracted cells. HbH preparation shows 51% HbH inclusions.

D. Haemoglobin and red cell parameters from homozygote (MC) and heterozygote (RC) for the $(\alpha\alpha)^{ALT}$ mutation. MC's red blood cell indices are consistent with HbH disease. RC's red cell indices are consistent with those of previously described patient TM, also heterozygous for the identical mutation.

.....

The alternative explanation is that our transgenic models do not accurately represent transcriptional control of the *α-globin* locus, and that we need to revise long-held views on the regulation of this gene cluster.

To address these hypotheses, a more detailed characterization of MC's mutation is warranted to exclude the presence of additional confounding mutations that might obscure the interpretation of the isolated HS-40 deletion. This characterization forms the basis of the rest of this chapter.

3.3 Results

3.3.1 Phenotypes of MC and RC: expression analysis from primary erythroid cells

Peripheral blood samples (50-75ml) from a normal control, patient MC and his daughter RC were subject to expansion using the two-phase culture system described by Fibach and colleagues (Fibach et al. 1989). The morphology of cells grown to intermediate normoblast stage (day 8 of phase II) is comparable between each sample.

Total RNA was extracted, and α and β globin transcripts were quantified by semi-quantitative real time PCR (Figure 3.9). MC has approximately 30% *α-globin* mRNA compared with a non-thalassaemic control, and his heterozygote daughter RC 55%. This confirms a meaningful degree of *α-globin* transcription from the Δ HS-40 allele – a finding incompatible with our previous understanding of *α-globin* transcriptional control.

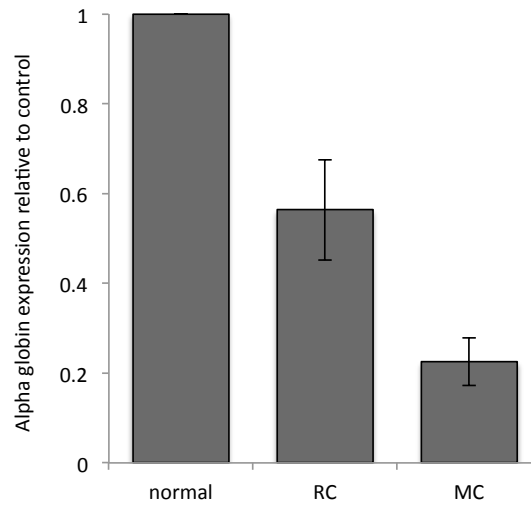


Figure 3.9 α -globin expression in a heterozygote (RC) and homozygote (MC) for the HS-40 deletion ($\alpha\alpha$)^{ALT}. α -globin is expressed relative to a non-thalassaemic control.

3.3.2 Clarification of the absence of HS-40

While the MLPA analysis undertaken by Coelho and colleagues suggests that HS-40 is absent from its correct chromosomal context in patient MC, it is important to confirm that it is not simply present in an ectopic location – especially since some studies suggest that HS-40 is able to function from outwith its normal position (and indeed outside the syntenic region altogether) (Vernimmen et al. 2009).

Southern blot, using a probe to HS-40 (amplified by PCR from genomic DNA, primers 5'-AACAAGAAAACCAGCAGGCTCC-3' and 5'-TGTAAGTCCATCCAGGTGTGAGTTC-3' giving a 528bp product chr16:163380-163907 on build hg19) confirms the absence of HS-40 from patient MC. Figure 3.10 shows patient MC, with his heterozygous

daughter RC and a non-thalassaemic individual as controls, along with the location of the HS-40 probe and the Southern blot design. This demonstrates that the intact regulatory element HS-40 is not present in the genome of patient MC, albeit with the caveat that it is not possible, using this 528bp probe, to establish whether an abbreviated or rearranged version of HS-40 is present elsewhere.

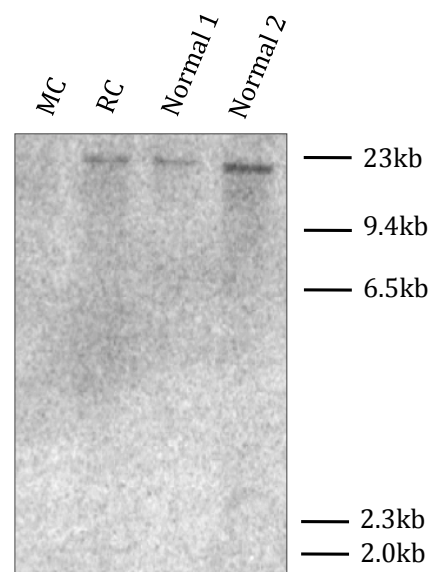


Figure 3.10 HS-40 is absent in patient MC. A probe for the core of MCS-R2, HS-40 gives an appropriate 19kb band on BglII digest of genomic DNA from normal controls and patient RC (heterozygous for the $(\alpha\alpha)^{ALT}$ mutation). The probe does not hybridize to genomic DNA from patient MC (homozygous for the $[\alpha\alpha]^{ALT}$ mutation).

3.3.3 The 39 bp insert: absence of canonical transcription factor binding sites

The orphan sequence located between the breakpoints of the deletion may itself have regulatory potential, thereby explaining the maintained α -globin expression in the absence of HS-40. To investigate this possibility, we used PROMO (<http://algeeb.lsi.upc.es>), and Cluster-Buster (<http://zlab.bu.edu/cluster-buster>) to

.....

attempt to locate canonical binding motifs for general transcription factors, and for erythroid factors NFE-2, GATA1, KLF-1 and SCL (using the canonical binding motifs 5'-TGA(G/C)TCA 3' for NFE-2, 5'-WGATAR-3' for GATA1, 5'-CCN-CNC-CCN-3' for KLF-1, and 5'-CANNTG-3' for SCL). None was discovered. Sp1 sites are found within the 39bp region, though their near ubiquity across the genome means that their presence alone in the 39bp insert cannot be taken as evidence of formation of a functional unit.

To assess whether the 39bp insert might originate in a genomic location of potential regulatory significance, we attempted to define its source. We were unable to locate any genomic origin for the element with UCSC basic local alignment search tools in human genome build hg19. Since intra-species variation might mean the fragment is not present in the reference genome, we also interrogated sequence alignment mapping files for all normal controls available in the laboratory. Again, no evidence to suggest the source of the 39bp fragment was found.

MC is Portuguese, but none of the individual genomes we searched for the 39bp fragment were of Portuguese origin. We considered the possibility that ethnically matched genomes might be a better source for the investigation of the source of the 39bp insert. We therefore attempted Southern blot using three genomic DNA samples from individuals of Portuguese origin, and MC as a positive control, with an end-labelled 39bp oligonucleotide as a probe. Unfortunately neither this very short probe, nor longer versions produced by end-to-end ligation of identical oligos, gave a clear signal in the positive control lane. The attempt was therefore abandoned after

efforts to optimize labelling, hybridization and washing conditions. To date the genomic source of the 39bp insert between the breakpoints in $(\alpha\alpha)^{ALT}$ remains obscure.

3.3.4 Assessment of the chromatin environment of the mutant α -globin locus

Our efforts to demonstrate that the sequence between the breakpoints of $(\alpha\alpha)^{ALT}$ is not functionally significant have so far extended only to the detection of known canonical transcription factor motifs. The possibility that unrecognized or unknown transcription factors may bind here cannot be excluded using *in silico* methods alone. Further, the possibility that entirely separate, new regulatory elements are present at the $(\alpha\alpha)^{ALT}$ cluster with the ability to assume the role of the absent regulatory element HS-40 must also be excluded.

Therefore, as a proxy for locating such new functional elements, we undertook chromatin immunoprecipitation for erythroid specific transcription factors SCL and GATA1, and for 'active' chromatin marks H4 acetylation, H3K4me1 and H3K27Acetylation. Since it is recognized as being present on transcriptionally active enhancers as well as promoters (Pekowska et al. 2011) we also assessed H3K4me3. Immunoprecipitated material was hybridized onto custom tiled arrays for analysis (Figure 3.11)

Being absent from the reference genome, the orphan sequence between the breakpoints of the $(\alpha\alpha)^{ALT}$ deletion is not assessed by the methodologies described above. To address this, qPCR primers and probes were designed using the 'Primer

.....

Design Express' software v3.0.1 from Applied Biosystems to capture this sequence specifically (primers FWD 5' AGA ACC TGC CCT GAG CAA AG 3'; REV 5' CGA TGG AGG GTG CAG TCT CTC T 3' – this sequence is specific to the 39bp orphan sequence in its correct setting in this particular mutation; and probe: 5' CCC TGC TAG TGT CTG C 3') and enrichment was contrasted with other positive and negative controls across the *α-globin* locus. (Figure 3.12).

It remains possible that functionally active elements might be present but unmarked by the chromatin modifications described above. We have not been able to exclude this possibility formally. A comparison between the DNase I hypersensitive sites in patient MC and in a non-thalassaemic control may have helped; unfortunately we were limited by the number of cells it is possible to amplify using the Fibach two stage culture method from peripheral blood samples – which was too small to permit DNase I analysis. In order formally to exclude this possibility, we are currently attempting to generate iPS cells from EBV transformed cells of patient MC and his daughter with a view to developing a better supply of erythroid material. However, proposing a second, independent genetic lesion in order to align MC's phenotype with the data from heterologous transgenic models seems an unnecessarily complex explanation for our observations.

3.3.5 Confirming the normality of the remaining regulatory elements

While chromatin immunoprecipitation will help to detect wholly new regions of active chromatin, it will not enable us to detect functional changes within elements

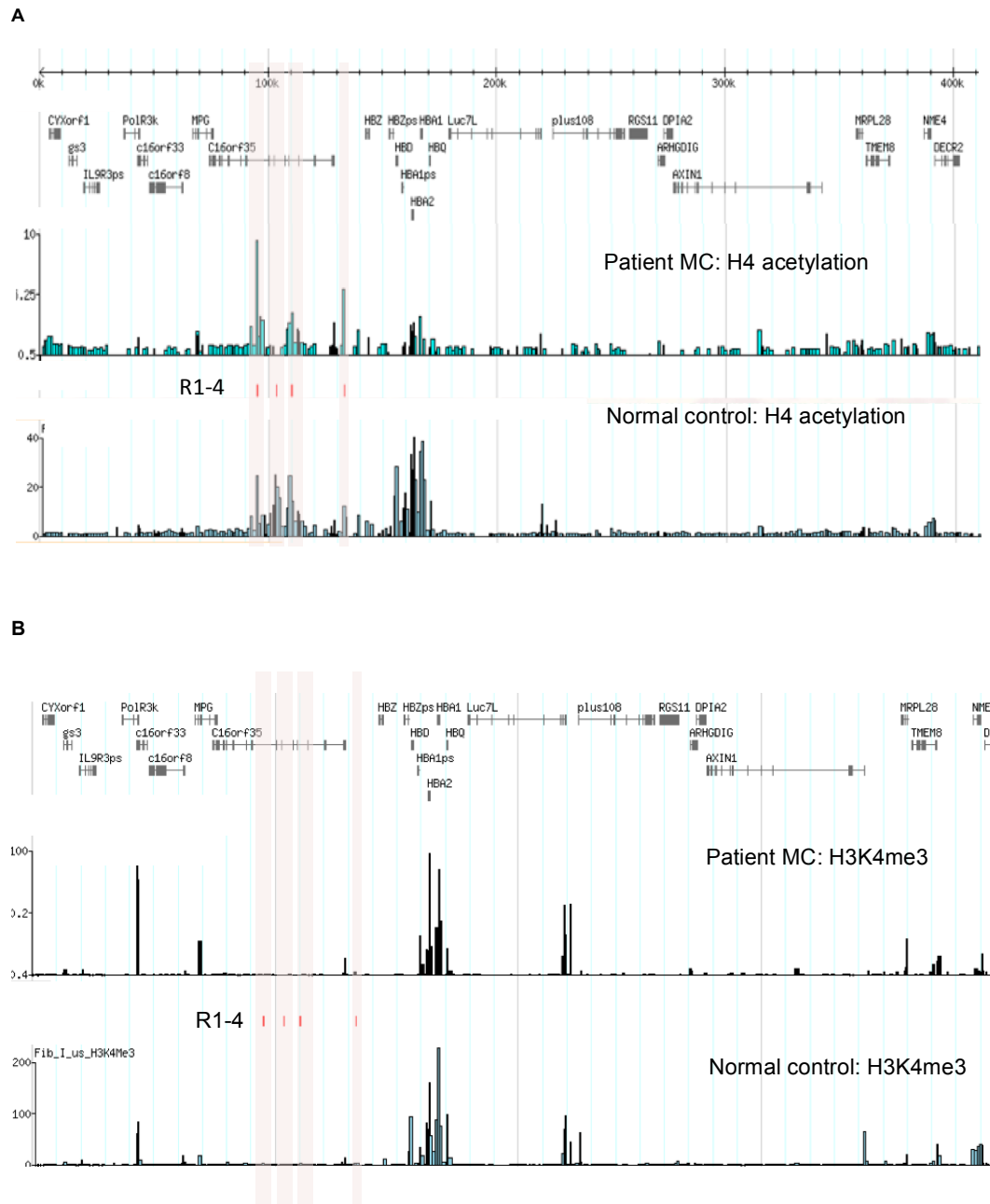


Figure 3.11 Histone modifications across the locus in the $\text{del}(\alpha\alpha)^{\text{ALT}}$ chromosome

Assessment of **(A)** panH4 acetylation and **(B)** H3K4me3 across the locus by chromatin immunoprecipitation – comparison between intermediate Fibach culture chromatin from patient MC (upper track in each panel) and a normal control. Note that the peak over HS-40 is absent in the profile for patient MC, but that no new peaks are found. These data provide no evidence to suggest novel regulatory regions which might adopt the role of HS-40 in its absence.

Red bars between the tracks for each ChIP indicate the location of the multispecies conserved elements MCS-R1-4, including R2 (HS-40)

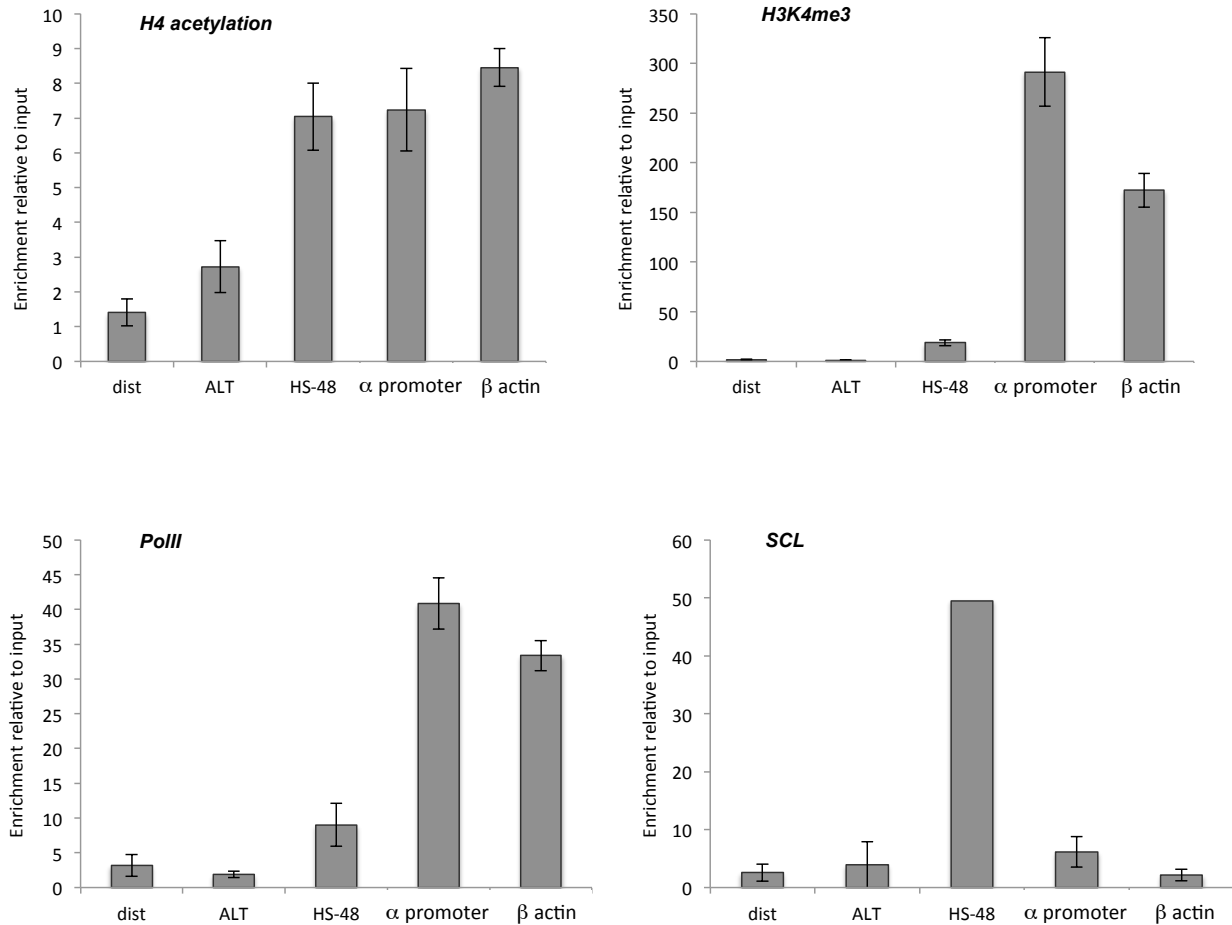


Figure 3.12 qPCR validation of ChIP material, including primers and probes specific for the 39bp insert of mutation ($\alpha\alpha$)^{ALT}. The Y axis represents enrichment relative to ChIP input, normalised to 18S. Amplicons are labelled thus: *dist*, intronic amplicon within the Rhomboid gene *RHBDF1*, used as a negative control; *ALT*, amplicon specific for 39bp insert; α promoter, α -globin promoter; *HS-48*, the human MCS-R1; β actin, a control amplicon over the β actin promoter.

that have already formed. For example, single nucleotide polymorphisms within the other DNase I hypersensitive sites may be responsible for functional changes in their ability to influence gene expression (e.g. (Milton et al. 2013)). To address this, 300-500bp sequences over the remaining intact regulatory regions HS-48, HS-33 and HS-10 were amplified by high fidelity PCR and the products cloned into the pGemT-easy sequencing plasmid in DH5 α cells. Three clones from each PCR were sequenced to ensure that PCR errors had not been incorporated, and the resulting sequence was compared with the reference genome.

Three SNPs were identified within the regulatory regions:

<i>Regulatory element</i>	<i>PCR coordinates (hg 18)</i>	<i>Result in patient MC</i>
HS-10	Chr16:132945 -133532	as hg18
HS-33	Chr16:109631 – 110394	SNP rs 11866877 (G/A)
HS-48	Chr16:94840 - 95907	SNP rs798612 (A/G) SNP rs72526676 (-/TC)

None of these SNPs has been identified on a comprehensive genome wide screen of variants that influence red cell parameters (van der Harst 2013). This suggests, though cannot prove, that their impact on globin transcription is negligible.

3.3.6 Enhancer promoter looping in the mutant chromosome

The physical interaction between regulatory elements and the promoters of their cognate genes is observed in the context of active transcription. Thus, in non-erythroid cells there is no physical interaction between the α -globin promoters and regulatory elements as determined by chromosome conformation capture (3C),

.....

while in actively expressing erythroid cells this measure of the physical apposition of enhancer and promoters suggests their proximity, and has been interpreted as indicating interaction between enhancer and promoter elements (Vernimmen et al. 2007). In transgenic models of human α -globin expression, HS-40 has been found to be critically required for looping of the regulatory region to the α -globin promoters; reintroduction of the element in an ectopic position prompted the formation of a different profile of interaction, this time between HS-40 in its new position and the globin promoters (Vernimmen et al. 2009). The interpretation of these data suggested that HS-40 is both necessary and sufficient for loop formation, and implied that such interaction was a necessary overture to transcription initiation.

We were therefore interested to see the impact of a naturally occurring deletion of HS-40 in human on chromosome conformation. Figure 3.13 shows the results of 3C assays for patient MC (homozygous for the HS-40 deletion) and RC (his heterozygous daughter) in both erythroid cells and non-erythroid controls (EBV immortalized lymphoid cells). qPCR analysis demonstrates a significantly increased interaction between the residual regulatory elements and the α -globin promoter in erythroid cells, even in the absence of HS-40 – in direct contradiction of the work seen in the transgenic system of the humanized mouse (Vernimmen et al. 2009).

In summary, while the $(\alpha\alpha)^{\text{ALT}}$ mutation when characterized in the transgenic setting of the MEL hybrid has neither enhancer-promoter looping, nor significant α -globin transcription, the identical mutation assessed in the native human *trans* acting milieu manifests quite different features. Transcription from the α -globin genes is

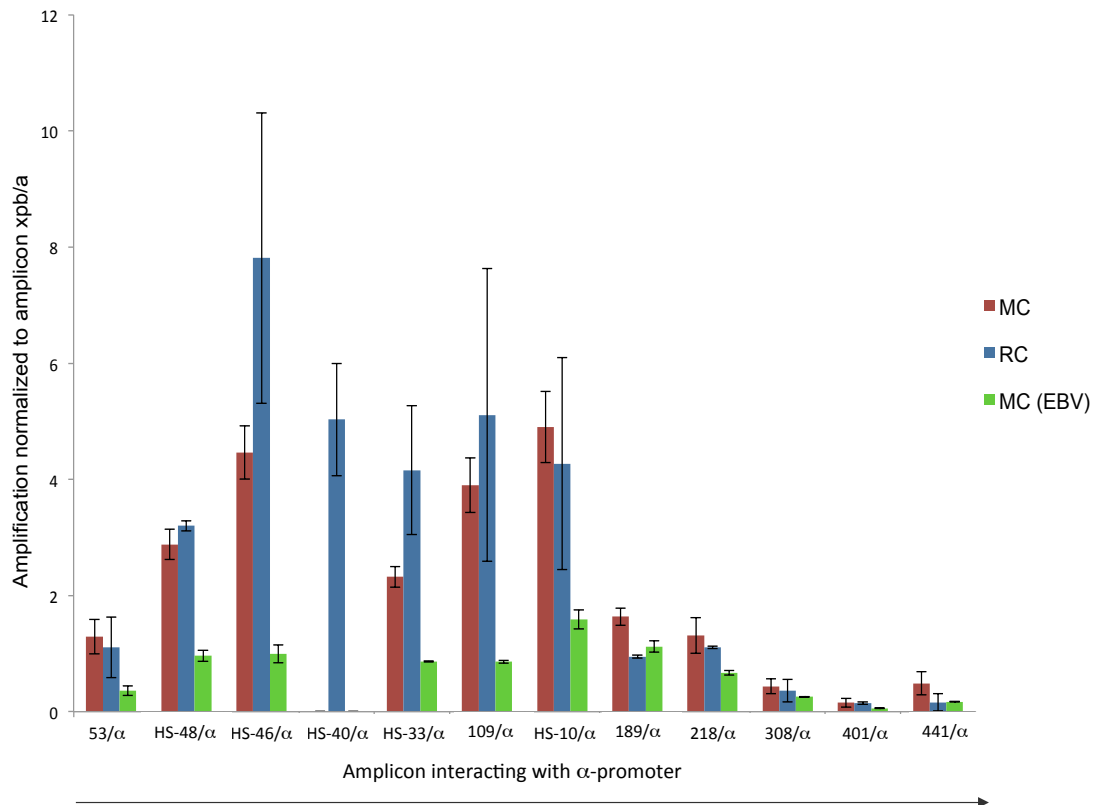


Figure 3.13 Maintained erythroid specific chromosome looping in the absence of HS-40.

The degree of interaction with the α -globin promoters, relative to control, is shown for positions along chromosome 16, from amplicon 53 (the promoter of the upstream gene *Dist*), through the regulatory elements HS-48, HS-40, HS-33 and HS-10, to downstream amplicons 189, 218, 380, 401 and 441 (progressing centromerically from gene *LUC7L*, through *LOC100134368*, to *RAB11FIP3* none of which is expected to interact with the α -promoter).

Patient MC, homozygous for the HS-40 deletion, shows intact chromosome looping between the residual regulatory element and the α -globin promoter (red bars). RC, his daughter, heterozygous for the deletion is shown as a positive control (blue bars). The erythroid specific nature of the interaction is suggested by the failure of loop formation in chromatin from EBV lines derived from patient MC (green bars). Note that no HS-40 amplicon interaction is documented for MC.

Error bars reflect standard deviation of the mean in three PCR quantifications of the same 3C experiment.

reduced, but not lost; and, consistent with this, loop formation persists. This clear distinction between the effects of this deletion in the native human setting, and in the heterologous transgenic systems of either somatic cell interspecific hybrid or humanized mouse would have been difficult to identify without the unusual finding of a patient homozygous for this mutation.

3.3.7 Investigating the discrepancy between transcription in the human and the humanized mouse

The different transcriptional output of identical chromosomes in the native human and heterologous murine *trans*-acting environments requires explanation, not least because there are many models which incorporate human sequence into the genomic context of a model organism for the investigation of mechanisms of gene expression (as reviewed in Devoy et al. 2011).

There are differences in methylation between the human α -globin cluster in its native context and its relatively hypomethylated status in the humanized mouse (Supachai Ekwattanakit and others, unpublished observations); how these generalized methylation differences arise and the impact they have on gene expression are currently the subject of ongoing work by other members of the laboratory. However, it is possible that changes in methylation status at the regulatory elements might influence their activity and contribute to the different impact of the HS-40 deletion in these two settings.

To investigate this, populations of erythroid cells were obtained from normal human (intermediate erythroblasts from day 11 Fibach cultures) and humanized mouse (ter119+ cells from the spleens of phenylhydrazine-treated mice). DNA extracted from each cell population was subject to bisulphite modification using the EZ DNA methylation Gold Kit from Zymo research, with subsequent PCR amplification using bisulphite conversion sensitive primers for each of the conserved elements. The resulting PCR products were cloned into pGemT-easy cloning vector, and the inserts sequenced using T7 and SP6 primers. For the PCR product of each regulatory element, at least 20 clones were sequenced from both human and humanized material.

Results are given in figure 3.14. There is no consistent difference between methylation status at the regulatory elements in the humanised mouse compared with the native human regulatory elements, arguing against this being a significant factor in the different behaviour of the two.

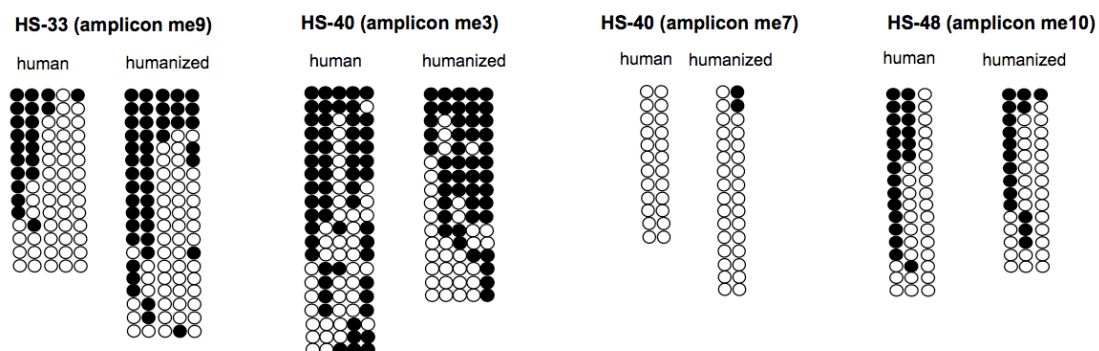


Figure 3.14 Summary of bisulphite sequencing results for regulatory elements MCS-R1-3 in primary human erythroid tissue and Ter119+ selected erythroid cells from the spleen of phenylhydrazine treated homozygous humanized mice. Each circle represents a CpG from the amplified genomic sequence. Filled circles represent methylated CpGs (maintained after bisulfite conversion); unfilled circles represent unmethylated CpGs.

.....

While these two cell populations used are the standards employed by the laboratory (and more widely) for examining human and mouse erythropoiesis, it should be recognised that the two populations may not be exactly comparable. The cells derived from mouse spleen arise in the context of stress erythropoiesis, which may have features distinct from those seen in the steady state samples taken for the generation of Fibach cultures (e.g. Coleman et al. 1969, R J Cole 1974). Similarly, primary erythroid cells that have been maintained in culture for ten days, as in the case of Fibach cultures, may undergo biological changes making them distinct from primary erythroid cells *in vivo*. These factors must be borne in mind when interpreting the results of the methylation assays.

3.4 Discussion

3.4.1 Human, mouse and humanized mouse: general comments on interspecific analysis of enhancer function

The prevailing view in evolutionary biology is that interspecies differences in phenotype occurring secondary to changes in gene expression are effected not by alterations in the coding sequence of genes but rather by changes in the mechanics of their regulation. (King & Wilson 1975; McLean et al. 2012). The extent to which these differences rely on the selection of alterations in the structure and function of *cis* elements, or variations in the activity of *trans*-acting factors is unclear.

In our assessment of the function of the human regulatory elements in MEL hybrid cell lines, identical *cis* acting sequences (indeed, identical chromosomes) give

.....

different transcriptional outputs in the context of distinct *trans* acting environments. The interpretation that no element aside from HS-40 was capable of driving α -globin expression was based on an assumption that human *cis* acting sequences would respond identically to murine and human *trans* acting factors.

There are excellent studies examining the distinct *cis* and *trans* acting environments of related but distinct species of *Drosophila* (e.g. Landry 2005); studies which address the potential impact of *trans* acting factors alone in interspecies gene expression do exist, but all are based on the classical enhancer assay and are beset by its intrinsic limitations. Ariza-Cosano and colleagues (Ariza-Cosano et al. 2012) tested a series of human enhancer elements (identified by Pennacchio and colleagues) in mouse and zebra fish models, and demonstrated distinct expression profiles between the two species in 36% of their tested elements. While these ‘dramatically different’ expression profiles are perhaps due in part to variations in the integration positions of tested elements, the ‘species-specific expression domains’ observed in up to 83% of tested elements were consistent across a number of stable transgenic lines, which argues against a purely positional effect. We would echo Ariza-Cosano’s note of caution in the investigation of human enhancers in different model organisms.

In terms of assessing the impact of interspecies differences in *trans* acting factors on transcription, the analysis of thalassaemia-causing upstream mutations in the MEL cell line offers a model superior to that described in Ariza-Costano’s paper. Not only are each of the upstream deletions investigated in an identical chromosomal context, this is also their correct chromosomal context. The entire human

.....

chromosome 16 is intact in these cell lines, diminishing the possibility that local disturbances of the *cis* acting environment might be responsible for any deviations from the normal regulatory function. The humanized mouse also pre-empts the possible criticisms of the malignant nature of the MEL cell line: the entire syntenic region around the α -globin locus is included in non-immortalized, primary cells, and, while the wider chromosomal context might not be identical to that of the normal human chromosome 16, the apparent defects in regulatory function are equivalent between this model and the somatic cell hybrids.

Together these data suggest that sequence-independent factors are able to influence the regulatory machinery. Those chromatin modifications which have been assessed to date in these systems (De Gobbi et al. 2007, Wallace et al. 2007), are equivalent suggesting that hitherto unconsidered chromatin factors, or *trans* acting factors, are responsible for the differential regulatory activity, and thus the different expression profiles. Additional erythroid specific transcription factors, not yet described, have been suggested by studies combining DNase I sensitivity and motif analysis (Xiong et al. 2013). That relatively suppressed α -globin transcription is seen from both MEL hybrids and the humanized mouse allele even in the absence of regulatory element deletions suggests that these aspects contribute to normal levels of α -globin expression.

The contention that the additive transgenic models do not faithfully recapitulate normal globin regulation is also supported by the unexpected behaviour of other known mutations which result in thalassaemia in humans, yet appear innocuous in

.....

the transgenic models. These include the Vanuatu mutation (De Gobbi et al. 2006) and ATRX mutations, discussed above (unpublished observations, Higgs and Gibbons groups).

It should be emphasized that these data are valid only for the α -globin locus, and we are unable to extrapolate to comment upon representations of regulatory machinery in heterologous models more generally; but they do advise caution in the interpretation of enhancer function in interspecific settings.

3.4.2 Implications for studies at the α -globin locus

The question of why four regulatory elements appear to have been conserved across millions of years of evolutionary diversification when only one has the ability to up-regulate α -globin expression can be seen to be over-simplistic. Instead of a single major regulatory element at the α -globin locus, it appears that at least one other element, and maybe all four elements, are functional. Mechanisms for determining the critical elements needed for globin expression must be reconsidered: the failure of expression of human α -globin from the MEL hybrids can be seen to be an unreliable indicator of the role of the residual elements present in patients with upstream deletions causing thalassaemia. However, all of the remaining upstream deletions shown in figure 3.1 remove at least multispecies conserved elements R1 and R2, and appear, at a gross level, phenotypically similar. All feature occasional HbH bodies on supravital staining with brilliant cresyl blue, just as would be expected in patients with α^0 thalassaemia trait due to deletion of both α -globin genes *in cis*. Using this relatively blunt phenotypic readout, it could be concluded that both R1

.....

and R2 are needed for full α -globin expression, but that in their absence the remaining elements are incapable of up-regulating α -globin expression to a clinically significant degree. But a much more quantitative method assessing the role of each regulatory element must be established; at present our most quantitative comment on the human α -globin locus is that HS-40, MCS R2, contributes 'most' of the enhancer activity of the locus.

Other potential *cis* elements also need to be taken into account. For example, a CTCF binding site is situated between R1 and 2 in the human α -globin locus, and is left intact in $(\alpha\alpha)^{\text{ALT}}$ deletion: it could be argued that the combination of deletions of upstream enhancers and the CTCF site might be responsible for the apparent failure of α -globin expression *in cis*. Alternatively, the larger deletions encompassing R1 and 2 may disrupt the wider structure of the chromosome without implicating the specific enhancer function of R1. Perhaps deletion of both elements directly influences the formation of other regulator sites at the locus.

From a broader perspective, several of the questions raised in chapter one focus on our current inability, using genome-wide sequencing methods, to dissect the functional heterogeneity of regulatory elements and to explore the possible redundancy between regulatory elements at a given locus. The hypothesis that there is differential activity among the globin enhancers provides a tractable system in which to investigate this question.

3.5 Summary and Conclusions

The work described in this chapter is summarized in figure 3.15. We have used the characterization of a rare naturally occurring mutation at the *α-globin* locus to address inconsistencies in our previous interpretation of its transcriptional regulation. We have demonstrated the HS-40 is not the only element capable of driving transcription of *α-globin* in its native setting, and have clarified the apparent inconsistency between murine and human *α-globin* regulation, and as well as the previously documented failures of the experimental systems employed to investigate *α-globin* transcription.

There is now a need for a systematic dissection of enhancer function at this locus. The humanized mouse model, though elegant, cannot be readily used for this purpose, since transcriptional control in this model reflects neither normal human nor murine transcriptional control. Figure 3.15 demonstrates that the native mouse locus bears similarity to the human locus in terms of transcriptional control. Examination of this locus provides the opportunity to evaluate the contribution of each individual regulatory element, as a model of general transcriptional control; it also provides the opportunity formally to examine the equivalence or otherwise of the human and murine *α-globin* loci, which have previously been considered to be differently regulated, and thus to clarify the best experimental model of globin transcription control.

The functional dissection of the regulatory machinery of mouse *α-globin* locus comprises the remainder of this thesis.

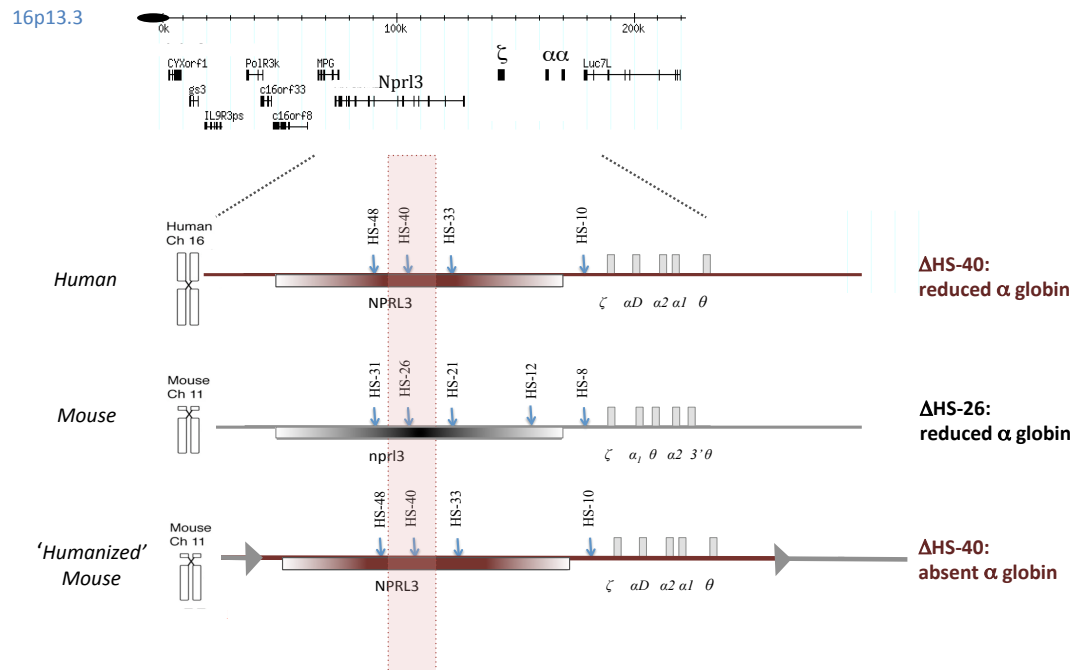


Figure 3.15 The role of R2 in α -globin expression in mouse, human and humanized mouse
Deletion of HS-40 and its equivalent, HS-26, in mouse reduces but does not abrogate α -globin transcription *in vivo*. By contrast, in heterologous systems such as the humanized mouse and MEL hybrid cell lines, the loss of HS-40 results in a complete failure of α -globin transcription. The heterologous transgenic models therefore represent neither human nor mouse transcriptional control, while native and human loci appear to respond similarly to the loss of R2.

Chapter 4

Generating Mouse Models

Chapter 4: Generating mouse models

4.1 Introduction

The characterization of the Portuguese HS-40 deletion, $(\alpha\alpha)^{\text{ALT}}$, described in chapter 3 led to the following conclusions:

- The regulation control of globin transcription is more complex than the current literature suggests: specifically, there is now evidence to support a role for elements other than R2 (HS-40, and its mouse equivalent HS-26) in both human and mouse *α -globin* transcriptional control.
- There is an opportunity to dissect the contribution of series of putative regulatory elements to transcriptional control, using the *α -globin* locus as a test bed.
- Heterologous transgenic models are unsuitable experimental tools for this task, since we have demonstrated that they do not reflect the normal mechanics of transcriptional control of either the host or the target organism, at least at this locus.

.....

Prior to the characterization of the $(\alpha\alpha)^{\text{ALT}}$ it could have been argued that the regulation of α -globin transcription in the mouse was atypical; the study of the rare mutation presented in chapter 3 highlights the similar phenotypes observed in both human and mouse with an isolated loss of R2, supporting the contention that the two loci are equivalently regulated, and that the study of one may give insights into the regulation of the other. The native, non-transgenic mouse locus therefore provides an experimental model for the investigation of how a complex bank of enhancers might contribute to transcriptional regulation. In this chapter, I will describe the experimental platform developed to answer the following questions:

1. What are the contributions of the individual regulatory elements to α -globin transcription in the mouse?
2. Do the regulatory elements at the mouse α -globin locus work in a redundant, additive or synergistic manner?
3. Are the apparently similar mouse regulatory elements functionally equivalent?
4. Can we use engineered mutations in the mouse to confirm or disprove the hypothesis that mouse and human α -globin regulation are equivalent?

4.1.1 Overview: dissecting enhancer function at the mouse α -globin locus

The murine α -globin locus is shown schematically in figure 4.1. For each human element (HS-48, HS-40, HS-33 and HS-10) there is an equivalent mouse element (HS-31, HS-26, HS-21 and HS-8). For clarity and ease of description throughout the rest

.....

of the following chapters, these will be referred to as elements R1, R2, R3 and R4, as shown in figure 4.1. The additional element present in mouse but not human, located at HS-12, will be termed R(m), with 'm' denoting 'mouse'.

The extent of deletions which had already been generated in the Higgs laboratory by gene targeting in embryonic stem cells followed by knockout mouse generation are shown as grey boxes. 'SND' denotes 'smallest natural deletion' – a term now obsolete as more and more circumscribed deletions affecting the upstream regulatory elements have been identified. The deletion SND removes regulatory elements R1, R2 and R3; it leaves intact the upstream gene *mpg* but disrupts the coding sequence of *nprl3*. While the disruption of this gene has little impact in heterozygosity, it is known that deletion of its promoter in mouse is associated with significant developmental abnormalities, such that mice homozygous for the SND mutations are not viable for analysis. The remaining deletions have been engineered in isolation.

4.1.2 Phenotypes of previously generated mouse models

Work undertaken by other laboratory members including E.Anguita, J.Hughes, M.Kowalczyk, and P-S Lee has defined the phenotypes of the previously generated knockout mice. The loss of R1-3 in the SND deletion is incompatible with survival in homozygosity, with mice dying at approximately E13.5. Of the remaining single knockouts, all are viable in homozygosity. None of the individual deletions has a significant impact on haemoglobin level or red cell indices (mean cell volume or mean cell haemoglobin) in adult mice, though homozygosity for $\Delta R2$ ($\Delta HS-26$) did

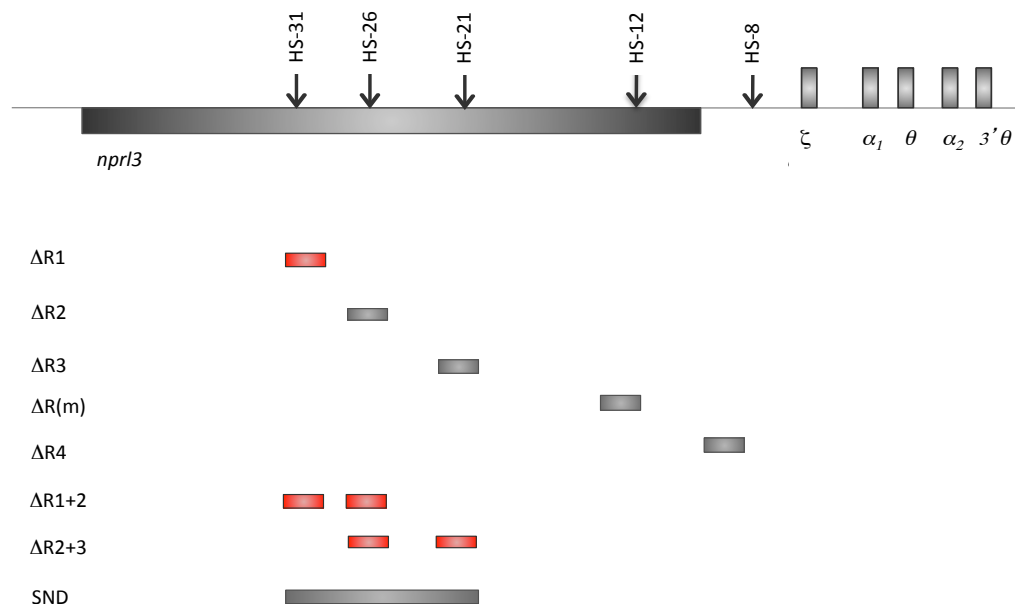


Figure 4.1 Summary of enhancer knockout models.

The α -globin cluster is shown schematically at the top of the figure, with *nprl3* upstream of the α -globin genes. The position of the globin regulatory elements is given by black arrows, and their position relative to the promoter of ζ -globin is given in kb. Grey boxes show the position and extent of deletions already engineered by previous laboratory members; red boxes show the position and approximate extent of deletions created as part of this project. Only one, the previously generated SND mutation, disrupts the coding sequence of *nprl3*.

.....

result in a non-significant apparent reduction in both MCV and MCH. It is of note that, since these mice were generated between the years of 2000 and 2009, that different instruments were used for blood count analysis for different animals; in chapter 5, therefore, all mouse models are reanalysed contemporaneously to minimize the impact of technical inconsistencies in blood count assessment.

Previously, there have been technical difficulties in the use of real time PCR for the assessment of *α -globin* mRNA levels in these models: both Anguita (2002) and Kowalczyk (2009, DPhil thesis, University of Oxford) have found qPCR to be an unhelpful method for the assessment of the impact of enhancer deletion on globin expression. This may be a consequence of the inherent limitations of qPCR in terms of sensitivity: while helpful in detecting fold changes, it is not an optimal technique for detecting fractional changes in expression (a 50% reduction in expression, for example, corresponds to only 0.5 cycles of amplification – a frequently quoted interval for an adequate technical replicate [e.g. <http://genomique.irc.ca/resources>, <http://www.genomeweb.com/node/918926>]). Furthermore, the very high levels of *α -globin* expression in red cells (at least 85% of the red cell's cytoplasmic mRNA by mass [D. A. Jackson et al. 2000]) in conjunction with the unusually long half life of the mRNA (Lodish & Small 1976, Rodgers et al. 2002) makes selection of control genes awkward. Small changes in the expression of control genes may have disproportionate effects on the quantification of the much more abundant globin transcripts.

.....

Consequently, the quantification of α -globin transcripts has been performed by RNase protection assay – a complex technique, which relies in its final step on the quantitation of autoradiographic bands to determine an $\alpha:\beta$ globin mRNA ratio. Although this has a long heritage, it is laborious and, critically, of limited sensitivity. Estimates of standard deviation for replicates are between 10 and 20%.

Previous laboratory members have therefore generated an excellent resource in terms of enhancer knockout mouse models, but the analysis of these models had been limited by the technologies available, and deletions key to an understanding of how α -globin transcription is controlled are missing from the series. My aims in the following chapters are to outline the generation of three additional models to complete and expand the series described above (chapter 4) and to develop more accurate ways of assessing the impact of enhancer deletion on globin transcription (chapters 5 and 6).

4.2 Completing a series of knockout models

Patients with deletions affecting incorporating both R1 and R2 are clinically indistinguishable from those in whom both α -globin genes are absent *in cis* – having microcytic and hypochromic red cell indices and occasional HbH bodies on supravital staining. Reliance on the expression of α -globin mRNA from the affected chromosome in the context of MEL interspecific hybrid cell lines had been put forward as evidence that no α -globin was transcribed *in cis*; our improved understanding of the limitations of this system now makes it difficult to be so bold about the precise level of transcription from alleles so affected, but the phenotype

.....

of affected patients is consistent with their having at best very low levels of α -globin transcription *in cis*. The same can be said for the SND mouse, in which R1, 2 and 3 are deleted. By contrast, a patient with a deletion of R2 alone, or a mouse with an equivalent engineered deletion, have reduced but not absent α -globin transcription *in cis*.

Therefore we may hypothesize that R1 has a role in normal *globin* transcription. As the sole element not yet targeted in the mouse locus, creating a knockout of this enhancer by gene targeting will complete the series of murine α -globin enhancer deletions.

A better appreciation of the congruence between the human and murine loci requires an assessment of the phenotype of a mouse harbouring deletions of R1 and R2, the elements which appear critically required for α -globin transcription in human. This may be accomplished by retargeting the previously generated Δ R2 (Δ HS-26) ES cell line with the construct required to produce Δ R1, (Δ HS-31), thereby generating a double knockout line *in cis*. By extrapolation from the effects of naturally occurring human deletions, we would predict that such a deletion would largely abrogate α -globin expression *in cis*, and therefore be incompatible with survival in homozygosity.

Finally, useful information about the mechanism by which enhancers act together to up-regulate globin expression may be obtained by generating an alternative double knockout line. Could a deletion of any two enhancers cripple *globin* expression *in*

.....

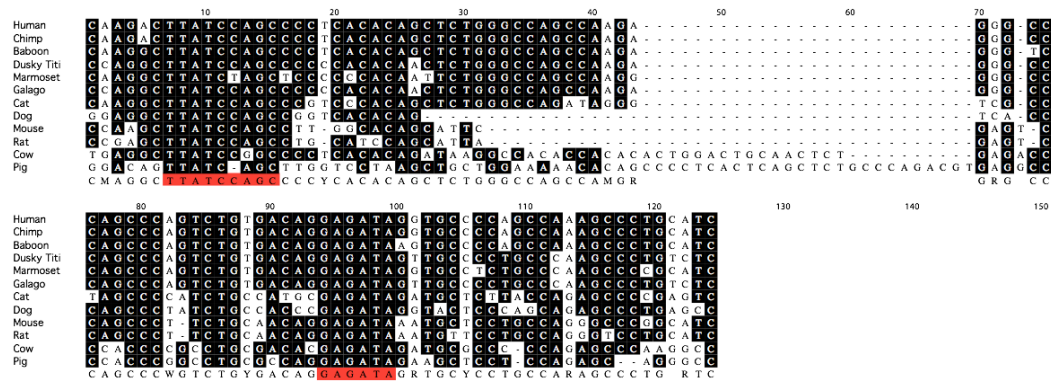
cis, or is there a particular requirement for R1 and R2? If so, in what might that particular requirement subsist? These questions may be addressed by the generation of a mouse model with a deletion of R2 and R3 *in cis*. Each of these models is shown schematically in figure 4.1, with the extent of planned deletions models represented by red boxes.

Although the generation of these enhancer knockout models is of interest in the appreciation of regulation at the α -globin locus, there is also a wider significance to the assessment of this series of models. As discussed in chapter 1, the extraordinary abundance of enhancers in the genome raises some important questions. How does a bank of enhancers work together to regulate the expression of its cognate promoter? How might we begin to classify enhancers based on their apparent levels of activity in promoting gene expression? Since relatively few loci have been dissected at this level, the complete series of knockout models provides a test bed for the further investigation of general rules of enhancer function.

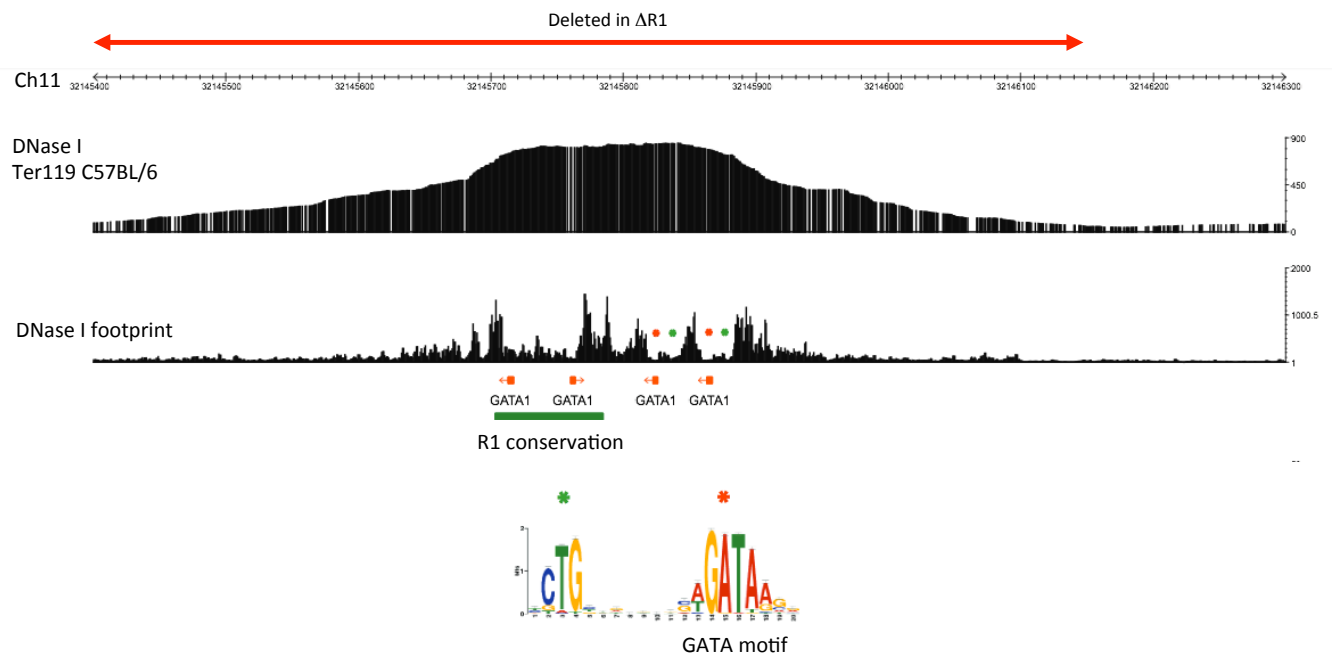
4.2.1 R1 in mouse erythroid cells

The mouse R1, HS-31, is a complex regulatory element. Its definition by conservation alone highlights two canonical GATA1 sites and associated half-E box elements consistent with SCL binding, but closer inspection of the region using DNase I footprinting highlights additional, likely functional regions (Figure 4.2). Chromatin immunoprecipitation data from mature erythroid (Ter119 positive) cells provide evidence for the binding of several erythroid specific and general transcription factors to the R1 element. ChIP for GATA1, SCL, NF-E2 and (to a lesser

A.

Multi-species sequence conservation at R1 (from Hughes *et al*, 2005)

B.

**Figure 4.2 Features of the R1 element.**

A. From Hughes *et al*, 2005: the sequence conservation across R1, demonstrating two highly conserved GATA sites.

B. Location of the R1 DNase I sensitive site on mouse chromosome 11 (top track), with the region deleted in this knockout shown by the red bar. Beneath this, the R1 DNase I footprint, showing complexity beyond the region of conservation (highlighted by the green bar), including additional GATA sites and E box sites (red and green asterisks).

.....

extent) KLF suggest binding of each of these erythroid specific factors to R1. The limitations of ChIP as a technique must be borne in mind, however: since chromosome conformation capture data suggest the physical interaction between the individual regulatory elements, there is the possibility of detecting transcription factors from other interacting regulatory elements cross-linked to R1 as a consequence of looping between enhancer groups and their associated promoters. Further, there are other, unanswered questions about the significance of ChIP data that describe transcription factor binding to individual elements. Analysis of ENCODE data from K562 cells suggests that R1 alone bears at least 14 separate transcription factors tested by ChIP, with R2 featuring an even more remarkable 37 different transcription factors in K562 cells alone – a surprising burden of transcription factor binding for an active sequence of no more than 800bp in length, and an observation which may suggest additional complexity in the function of transcription factor activity at enhancers than had previously been suspected. For instance, do transcription factor binding profiles change during the process of generating a transcript? Do different elements bind to different subsets of transcribing cells? Does the profile of binding alter during the cell cycle, or at different stages of erythroid maturation? Further, not all genome-wide ChIP tracks are reported along with contemporaneously generated input tracks, leading to the inevitable criticism that some ChIP datasets are barely more informative than DNase I hypersensitivity tracks. Although none of these questions has been definitively answered to date, there is sufficient basis to suggest that R1 is an active and relevant regulatory sequence in transcription in erythroid cells and for *α -globin* in particular.

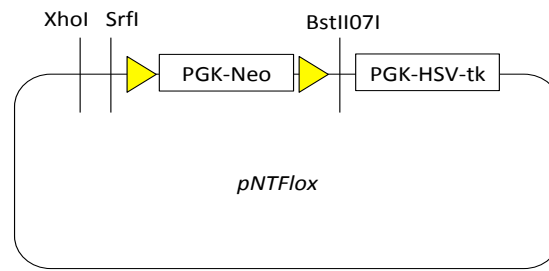
4.2.2 Creation and validation of a targeting vector for deletion of mouse R1, HS-31

The region targeted for deletion was selected using four major criteria: first, the region must contain all of the conserved sequence identified by Hughes et al. by sequence conservation (Hughes et al. 2005); secondly, it should contain within its boundaries the chromatin immunoprecipitation peaks for erythroid specific transcription factors known to bind to murine R1 (SCL, GATA1, NFE2), as well as the peak of DNase I sensitivity found at this locus; thirdly, it should not disrupt the coding sequence of *nprl3*; and finally, there should be avoidance of any impingement on other active elements within the locus (e.g. CTCF sites).

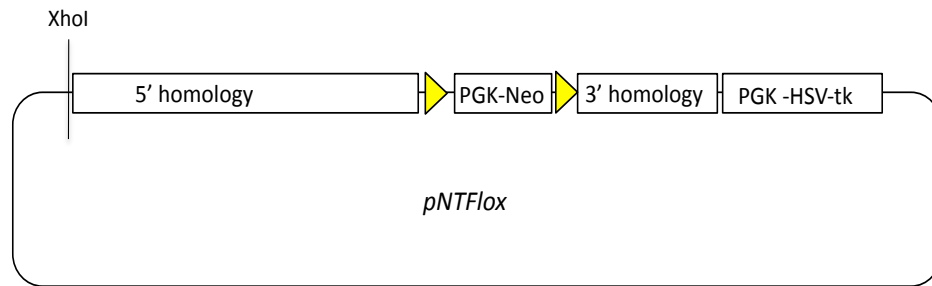
In addition, the deletion is designed to be comparable in size to the previously generated enhancer knockouts at the locus, in light of the apparent reported discrepancies between deletions of varying extent at the *β -globin* locus (J. D. Jackson et al. 1996). To satisfy all of the above targets a region of approximately 1kb over R1 was identified: its location and extent are shown as a red bar in figure 4.2

For effective homologous recombination, regions of 1577bp and 6092bp, to the 3' and 5' respectively, of the region to be deleted were identified as suitable for PCR amplification. These regions were amplified from a mouse 129/Ola specific *α -globin* BAC using high fidelity Taq polymerase, and sequenced in the shuttle vector pGemT-easy, before excision with EcoR1, blunting with T4 polymerase, and ligation into the unique Srf1 and Bst1107I sites in the pNTFlox targeting vector. Further details are given in chapter 2 (Methods). The final targeting vector was verified by restriction

A.



B.



C.

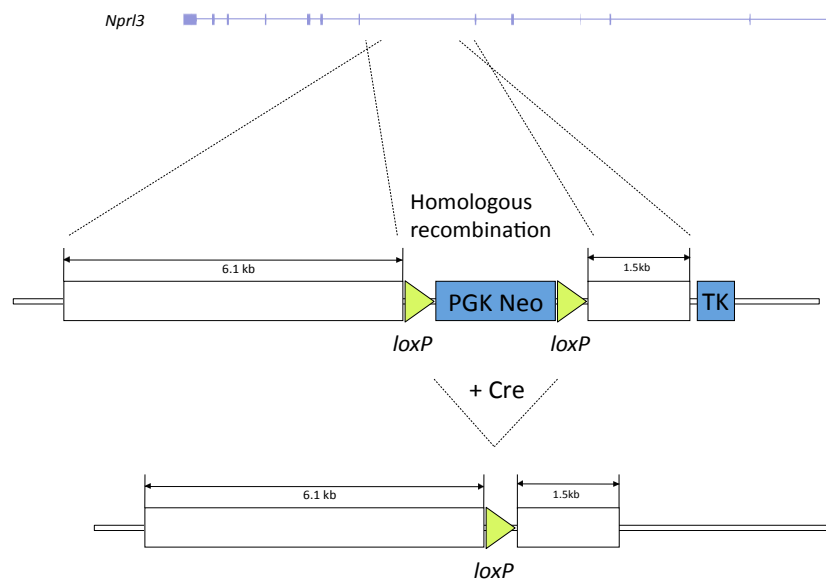


Figure 4.3 R1 knockout construct in pNTFlox.

A. pNTFlox targeting plasmid, containing SrfI and BstII07I restriction enzyme sites for the insertion of homology arms for the deletion of R1. These flank a floxed neomycin resistance gene driven by the PGK promoter (PGK-Neo).

B. Homology arms flank the PGK-Neo cassette; a PGK promoter-driven herpes simplex thymidine kinase cassette (PGK-HSV-TK) is located outside the homology arms to permit negative selection. An unique XhoI site for construct linearization prior to electroporation is shown.

C. Targeting strategy to *Nprl3* and subsequent Cre recombination.

enzyme digestions (figure 4.4) and by sequencing of the entire vector. For details of sequencing primers, see appendix 1.

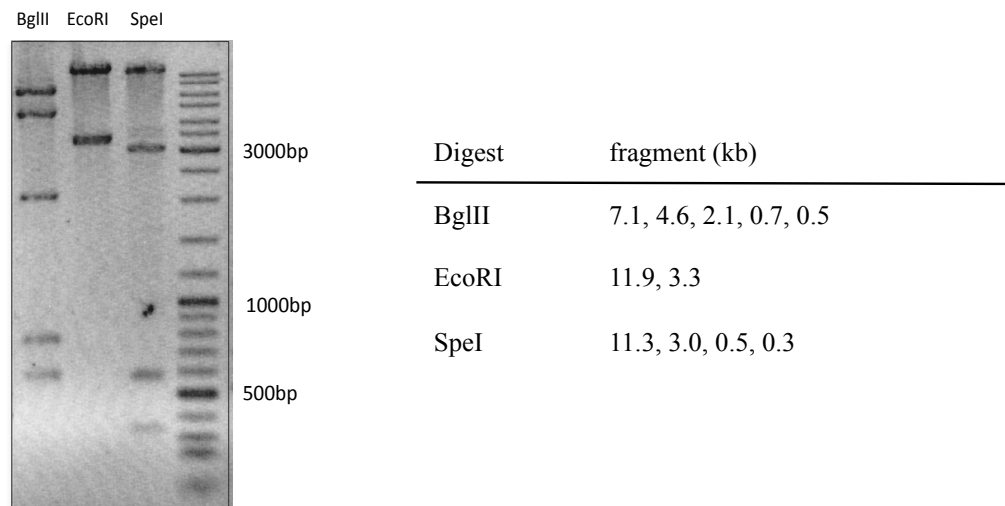


Figure 4.4 | Restriction enzyme digests confirming correct targeting construct $\Delta R1$ -pNTFlox

4.2.3 Creation of a targeted ES cell line for R1 deletion

Sequenced and purified vector was linearized at its unique XhoI site, and purified prior to electroporation into 129/Ola ES cells, which were then cultured in GMEM as described in chapter 2. Positive selection with G418 and negative selection with FIAU were used to isolate cells incorporating the construct. The former selects clones that have incorporated the neomycin resistance cassette, while the latter aims to exclude clones which have incorporated sequence outwith the regions of homology (and which are therefore less likely to be correctly targeted). The screening strategy and results for clones surviving double selection are shown in figure 4.5 and 4.6. Correctly targeted clones were expanded for more detailed Southern blot and sequencing analysis (figure 4.7)

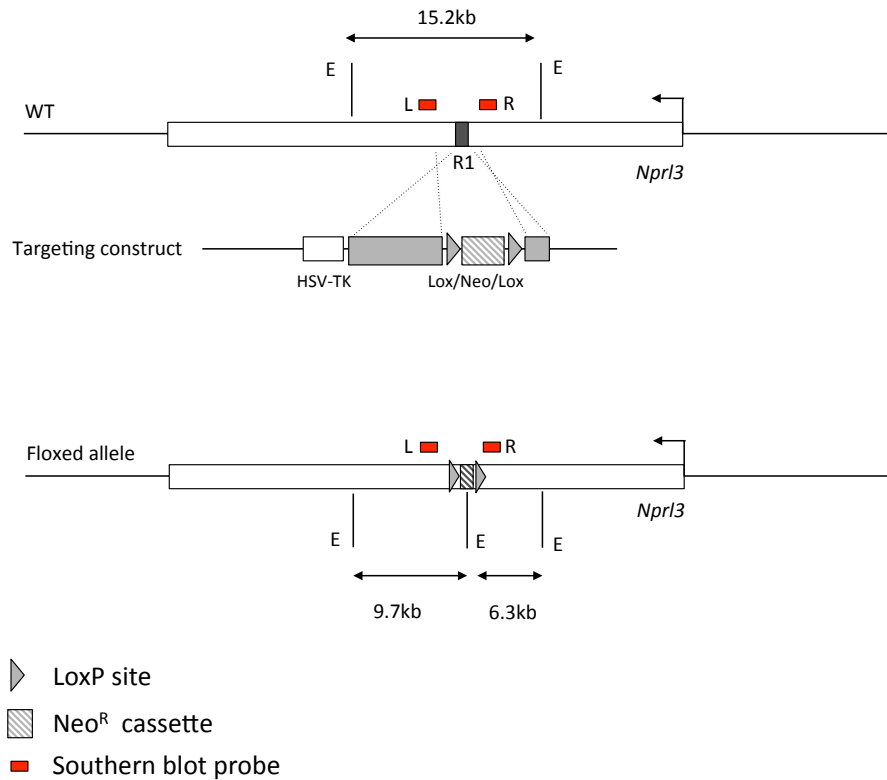


Figure 4.5 Schematic representation of targeting plan for the R1 knockout, and Southern blot strategy to detect correctly targeted clones.

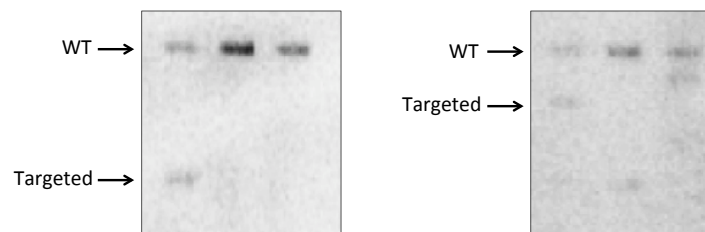
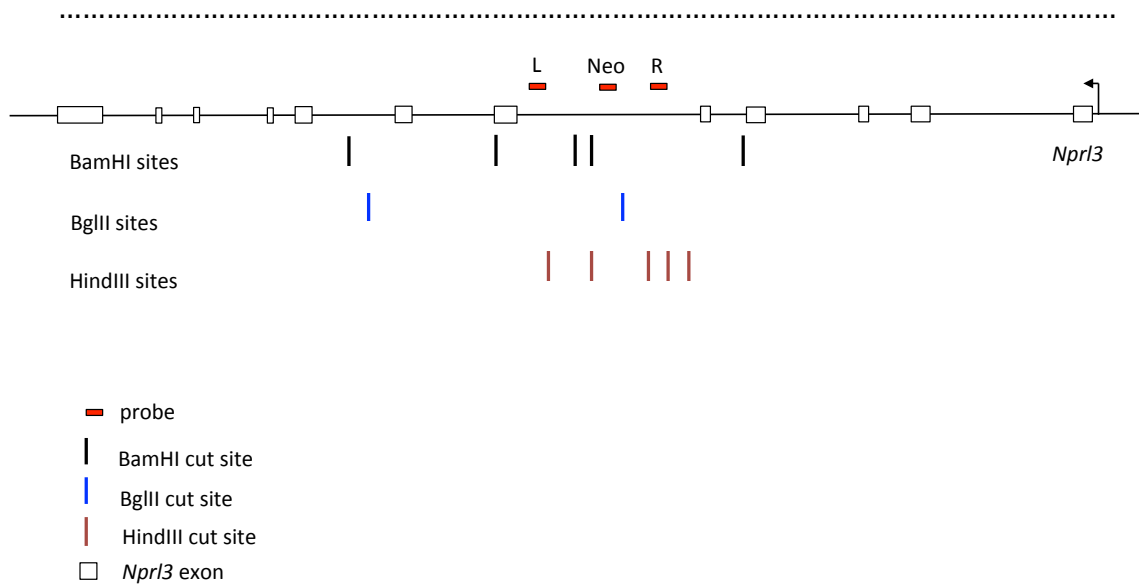


Figure 4.6 Screening with R probe: clone 1/D11 (*lane 1 left blot*) gives correct size bands for WT (15.2kb) and targeted (6.3kb) alleles; re-hybridising (*right blot*) with L probe gives expected 15.2kb and 9.7kb bands. Lanes 2 and 3 are digests from incorrectly targeted clones.



Expected band sizes:

	BamHI	BglII	HindIII	SacI
Probe R	11.1, 17	7.7, 6.8	0.8	6.1, 5.2
Probe L	17, 7.4	7.7, 6.8	1.4	1.4, 4.4
Neo^R probe	7.4	7.7	3.0	6.1

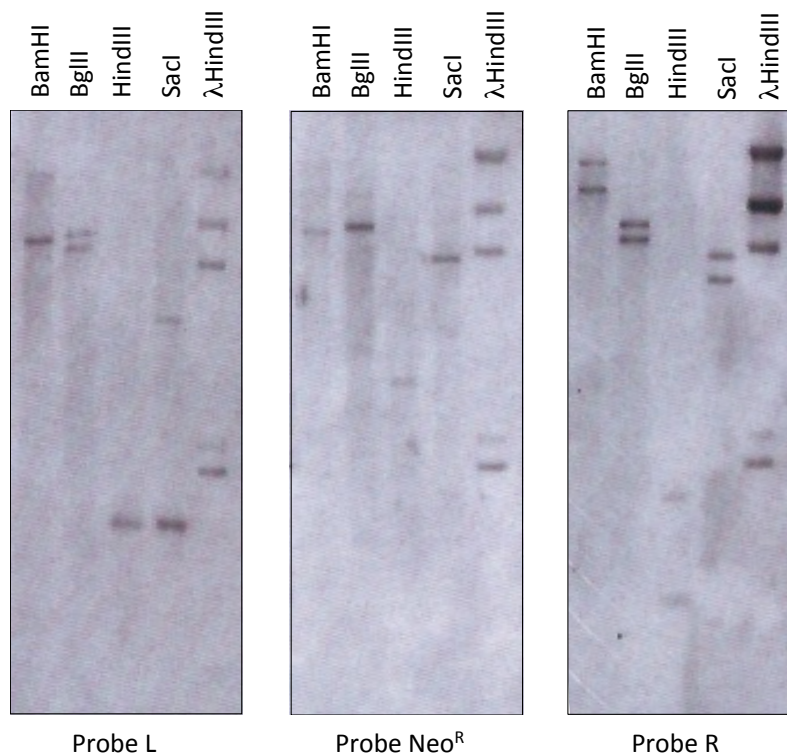


Figure 4.7 Southern blots analysis of clone 1D11 used for blastocyst injection

Left blot: Probe L, digests BamHI, BglII, HindIII, SacI

Central blot: Probe Neo^R, digests BamHI, BglII, HindIII, SacI

Right blot: Probe R, digests BamHI, BglII, HindIII, SacI

Figure 4.8 shows typical metaphase spreads from two clones karyotyped prior to blastocyst injection. At least twenty metaphase spreads were counted for each clone used, and a full complement of 40 chromosomes was confirmed in 95% or more of the spreads. Four chimeric mice were generated by blastocyst injection (injections performed by Ms Christina Rode, MHU transgenics facility).

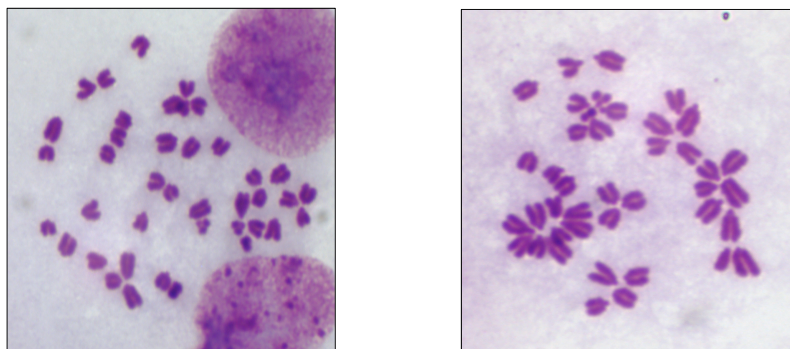


Figure 4.8 Karyotyping targeted clones for blastocyst injection. Clones 1D11 and 4F2 for $\Delta R1$ were both found to have a normal complement of chromosomes on >95% of 40 metaphase spreads examined.

4.2.4 Generation of mice homozygous for $\Delta R1$

Matings with C57Bl/6 mice resulted in germ-line transmission in three agouti offspring, confirmed by PCR for the neomycin-containing cassette in the correct context using tail-clip and ear-notch derived DNA. Southern blot for the neomycin-containing cassette confirms its presence in the expected context in agouti mice (WT/ $\Delta R1^{\text{flox}}$).

.....

The impact of the neomycin resistance cassette on gene expression at targeted alleles has been previously described (Anguita et al. 2002, Fiering et al. 1995). For example, in the $\Delta R2$ ($\Delta HS-26$) mouse, while homozygotes for the deleted enhancer had only mild effects on red cell size and haemoglobinization, mice with alleles still harbouring the neomycin resistance cassette in place of the R2 element were much more severely affected. It has been postulated, though never proven, that the strong PGK promoter in the neomycin resistance cassette competes with the α -globin promoter to co-opt enhancer activity from the remaining α -globin gene regulatory elements. The continued presence of the neomycin resistance cassette therefore severely hampers the accurate estimation of the effect of enhancer knockout, and its removal prior to analysis is critical.

Mice bearing the germ-line targeted allele were therefore crossed with mice expressing *Cre* recombinase under the control of the GATA-1 promoter to generate mice heterozygous for the deleted allele. The GATA-1-cre mouse was initially generated with the aim of permitting myeloid-specific recombination; however, its expression very early during embryonic development results in near universal recombination (Mao et al. 1999). The offspring of GATA1-Cre x WT/ $\Delta R1^{flox}$ had genomic DNA from ear notches or tail clips examined by both PCR and Southern blot to confirm appropriate recombination, using a strategy shown in figure 4.9 and 4.10. Mice with appropriately recombined alleles following exposure to *Cre* recombinase were interbred to yield mice homozygous for the deletion of R1, with genotyping performed by PCR and Southern blots (figure 4.11).

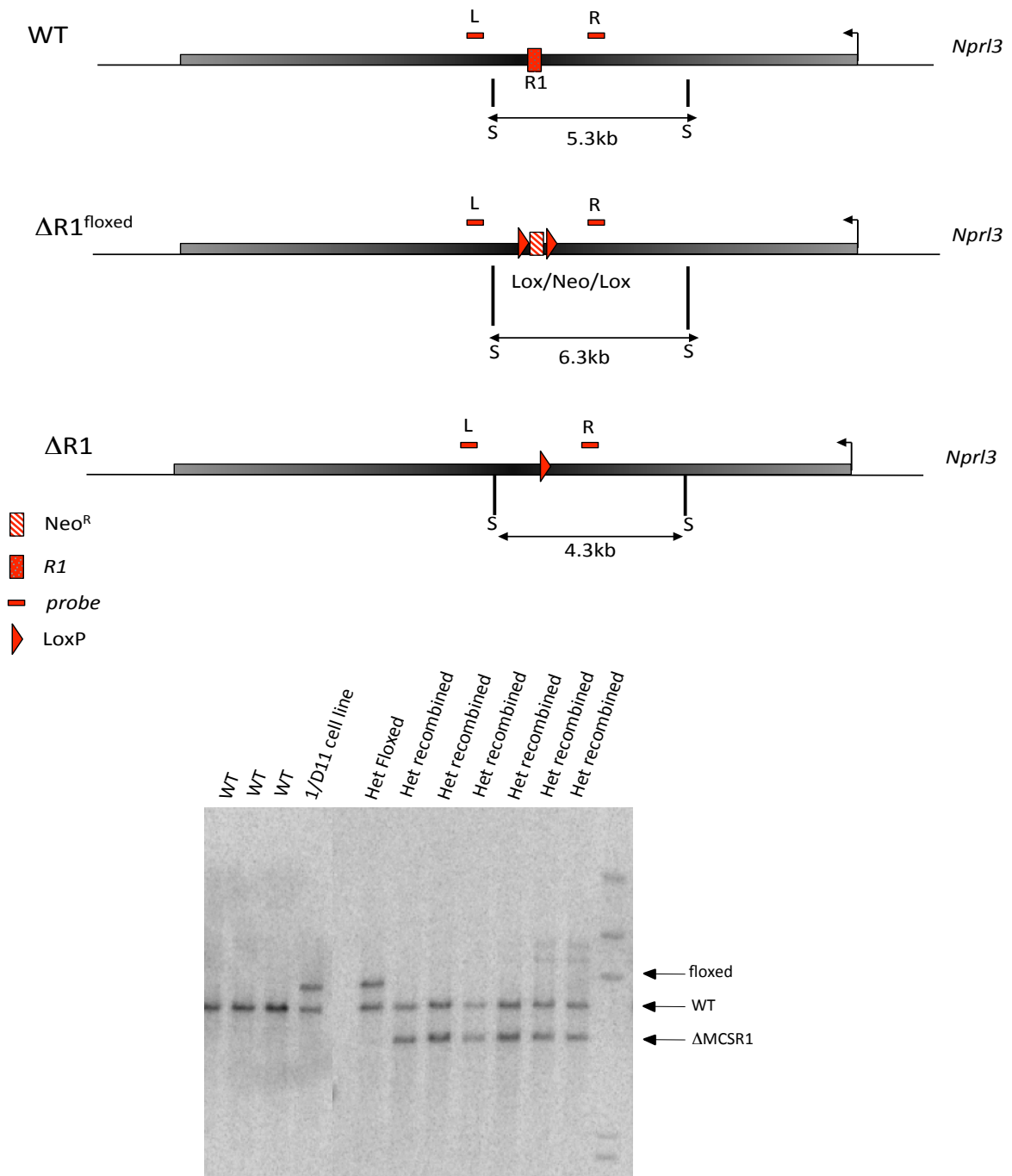


Figure 4.9 Screening strategy and Southern blots to determine $\Delta R1$ floxed $\times$ cre mice.

Upper panel, Southern strategy: figure shows structure of *nprl3*, with the position of Southern blot probes *L* and *R* shown in red. The digest for this blot is with *SacI*, with the position of cutting shown by *S*. Probe *R* is used.

Lower panel: Southern blot of using WT mouse DNA, E14 cell line DNA from purified targeted clone 1D11 (giving the expected band for the floxed allele), and DNA from tail clips of the litter of a $\Delta R1^{fl/WT} \times$ GATA-Cre cross.

Note the additional higher molecular weight bands in the last two sample lanes: *SacI*, although not classically methylation sensitive, is sensitive to methylation of the first cytosine in the extended cut site sequence GAGCTC. This extended cut site sequence is manifest in this digest, and would explain the weak additional bands in this blot.

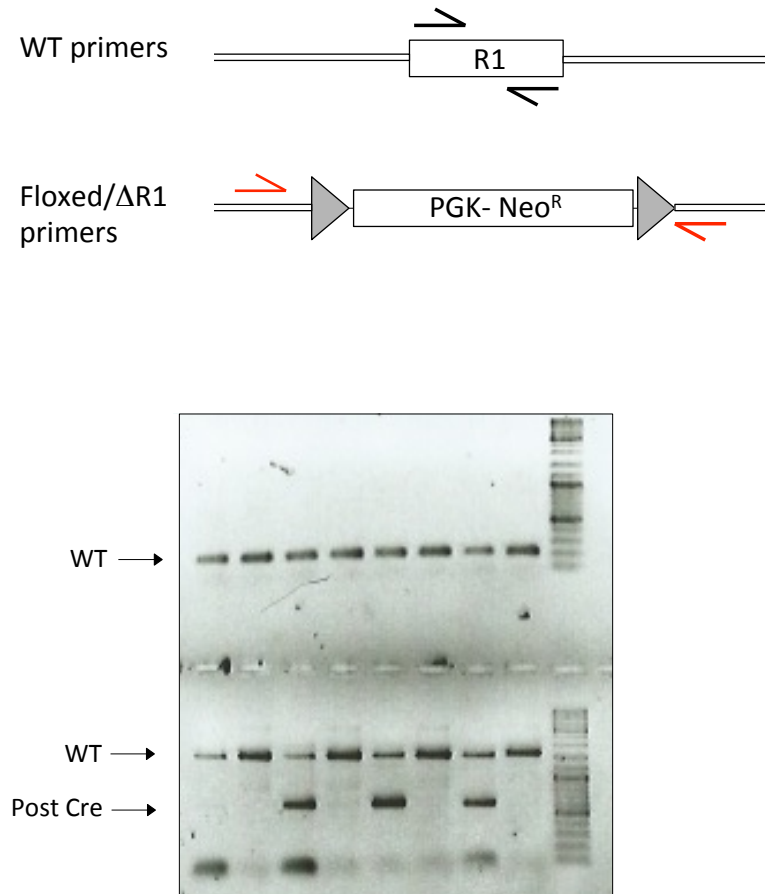


Figure 4.10 PCR to confirm appropriate Cre recombination of targeted alleles

Upper half of gel: WT amplicon, internal to deleted region (black primers in schema above)

Lower half of gel: Post Cre amplicon, with primers outwith deleted region (red primers in schema above)

PCR products

WT amplicon product size	210bp
Post Cre amplicon in WT allele	1583bp
Post Cre amplicon in Floxed allele	2467bp
Post Cre amplicon in Recombined (KO) allele	607bp

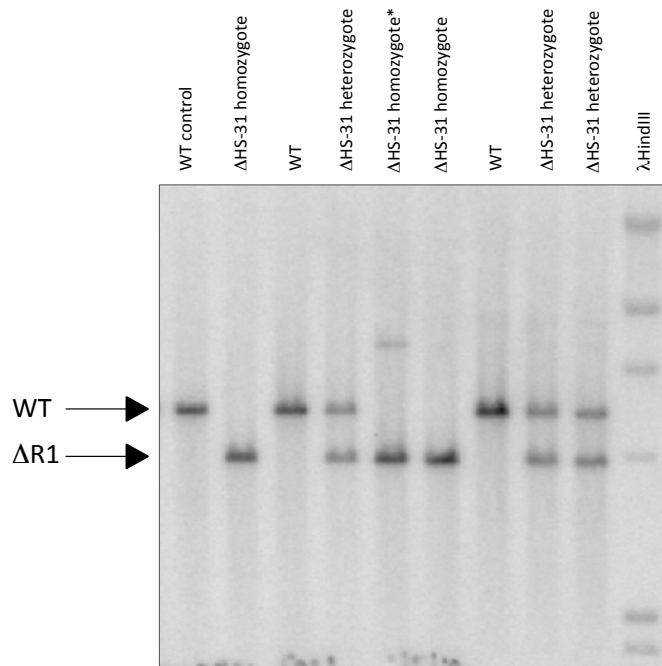


Figure 4.11 Confirming homozygosity among $\Delta R1$ mice. Southern blot, using probes, *SacI* digest and strategy as for figure 4.9. Lanes 2 and 6 show digests from homozygotes for the $\Delta R1$ allele

4.3 Generation of the $\Delta R1/\Delta R2$ mouse

Although it is difficult absolutely to quantify the transcriptional output of alleles which bear upstream deletions of the regulatory elements in human, there is no phenotypic difference between patients in whom R1, 2 and 3 are deleted, and a patient in whom only R1 and R2 are lost (Phylipsen et al. 2010, Viprakasit 2006). The conclusion that R1 and R2 might comprise the elements critically required for α -globin expression can be addressed in the mouse by retargeting *in cis* the vector described above, for deletion of R1, to the ES cell line already heterozygous for a deletion of R2. As well as enabling us to address the hypothesis that the murine and human loci are equivalently regulated, this also gives us an excellent platform from

.....

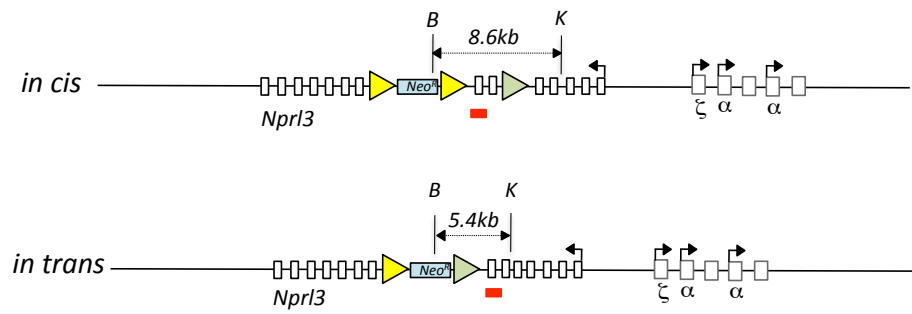
which to ask questions regarding the interactions of regulatory elements in transcriptional control.

4.3.1 A strategy for the generation of a double knockout

Figure 4.12 demonstrates the strategy for retargeting the $\Delta R2$ cell line. This line is heterozygous for the replacement of R2 with a neomycin resistance marker – the same marker required for selection of clones targeted by *pNTFLox- $\Delta R1$* . Therefore, it was necessary to clarify the orientation of the LoxP sites in the targeted allele, and to excise the Neo^R cassette *in vitro* prior to retargeting. This had been performed by Dr Eduardo Anguita shortly after generation of the vector, and the recombined cell line heterozygous for $\Delta R2$ had been in cryostorage for several years. In order to ensure the health of the recombined ES cell line, and to clarify that *in vitro* Cre did not give any phenotypic difference from mice subjected to GATA1-Cre recombination *in vivo*, the recombined cell line was used in blastocyst injections to generate a new series of $\Delta R2$ homozygote mice, the analysis of which forms part of the next chapter.

Retargeting of the $\Delta R2$ cell line was performed as described for the $\Delta R1$ mouse, in section 4.2. To ensure targeting of the R1 deletion *in cis* to the deletion of R2, the Southern blot strategy shown in figure 4.12 was employed.

The resulting double-targeted allele is represented in figure 4.13 along with the possible consequences of Cre recombination. Of these, only two will result in G418 sensitivity, and will survive antibiotic selection; of these, only one leaves intact the coding sequence of *nprl3*. While the maximal deletion shown in figure 4.13 may



Southern blot with 3' probe

WT	11.9kb
ΔHS-26 only	15kb
ΔHS-31 ^{flox} only	5.4kb
ΔHS-26 + ΔHS-31 ^{flox} <i>in cis</i>	8.6kb

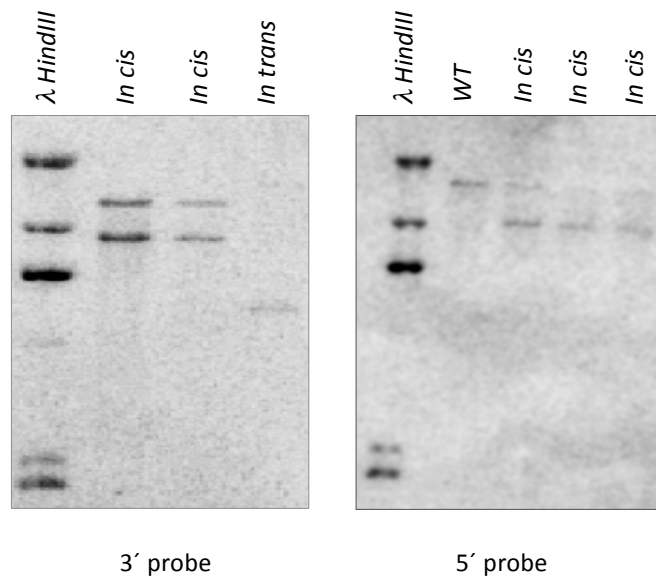


Figure 4.12 Strategy for re-targeting the ΔR2 line with *pNTFlox* ΔR1

The upper panel shows that the strategy for screening using probe R. Expected band sizes with probes R and L are shown in the central table, with sample blots confirming correctly targeted samples.

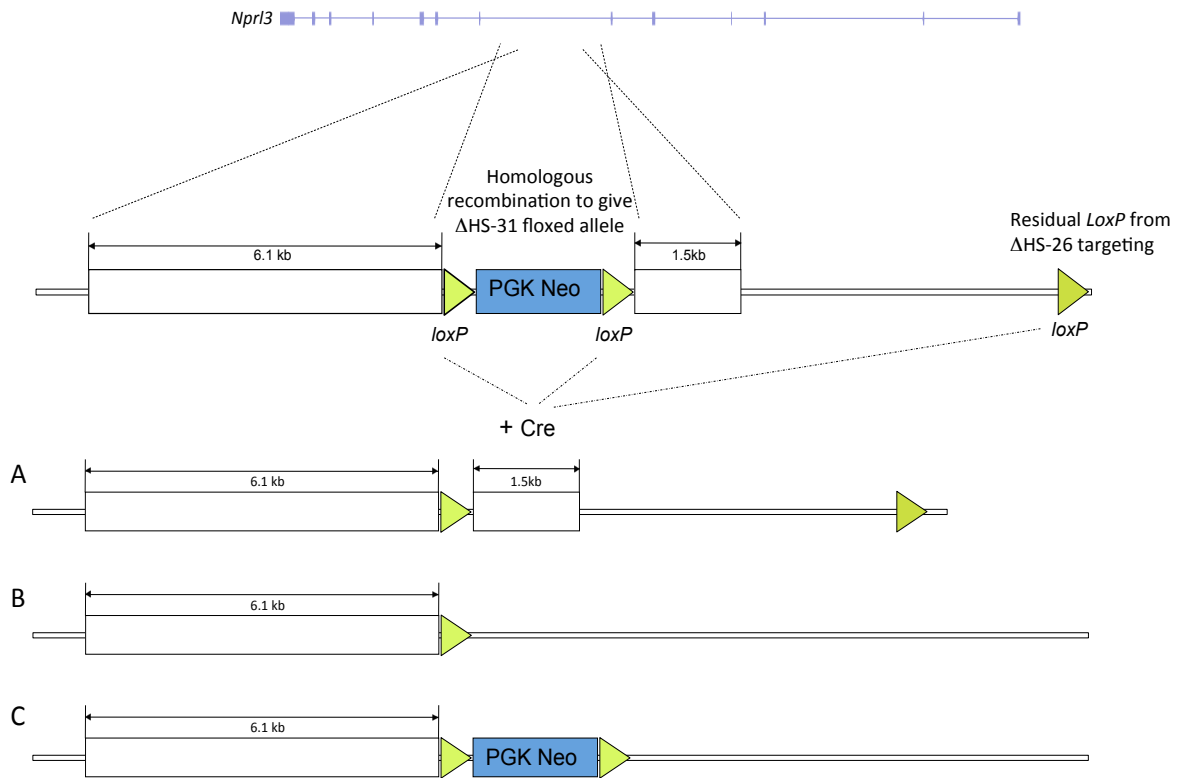


Figure 4.13 Homologous recombination of the Δ R1 pNTFlox construct to an ES cell line previously targeted with a construct for the deletion of R2. *In vitro* recombination of the HS-26 targeted line prior to retargeting with R1 pNTFlox will excise the resistance marker employed in R2 targeting, and leave a single *LoxP* *in situ*. A second round of *in vitro* recombination after correct targeting with the Δ R1 construct yields three potential recombination products.

Of these products, C will retain its neomycin resistance, and will therefore be selected against. Both A and B will be neomycin sensitive; while B will disrupt the coding sequence of *nprl3*, A does not, giving the desired discrete deletions of both R1 and R2.

.....

represent more accurately the naturally occurring deletions found in patients with alpha thalassaemia, the disruption of *nprl3* will be predicted to obscure the phenotype of homozygotes. In addition, since our question focuses on the role of the enhancers, it is important to minimize the disruption of other potentially significant regulatory elements. A CTCF binding site is present between R1 and R2 in both human and murine α -globin loci, and is of unknown significance in the control of *globin* expression. The discrete deletion in figure 4.13 allows the maintenance of this element and well as the exons of *nprl3*.

The prolonged exposure of targeted alleles to cre recombinase under the GATA1 promoter is likely to yield the terminal (maximal) recombination product. Furthermore, the generation and screening of mice with potentially incorrectly recombined alleles would be unnecessarily time-consuming, as well as costly in terms of mouse numbers. Therefore, we elected to undertake a further round of *in vitro* cre recombination to attempt to obtain the discrete recombination product shown in figure 4.13 prior to blastocyst injection.

An initial attempt with pCAGGS-Cre electroporated into the double-targeted ES cell clones yielded only the terminal (maximal) recombination product as screened by PCR. A Cre-recombinase driven by a weaker promoter (pMC-Cre) gave a mix of both discrete and maximal deletions (figure 4.14). Of these, two clones were chosen, karyotyped, and used in blastocyst injection as described in section 4.2 above. Figure 4.15 shows a Southern blot from tail DNA of agouti offspring of double KO chimeras.

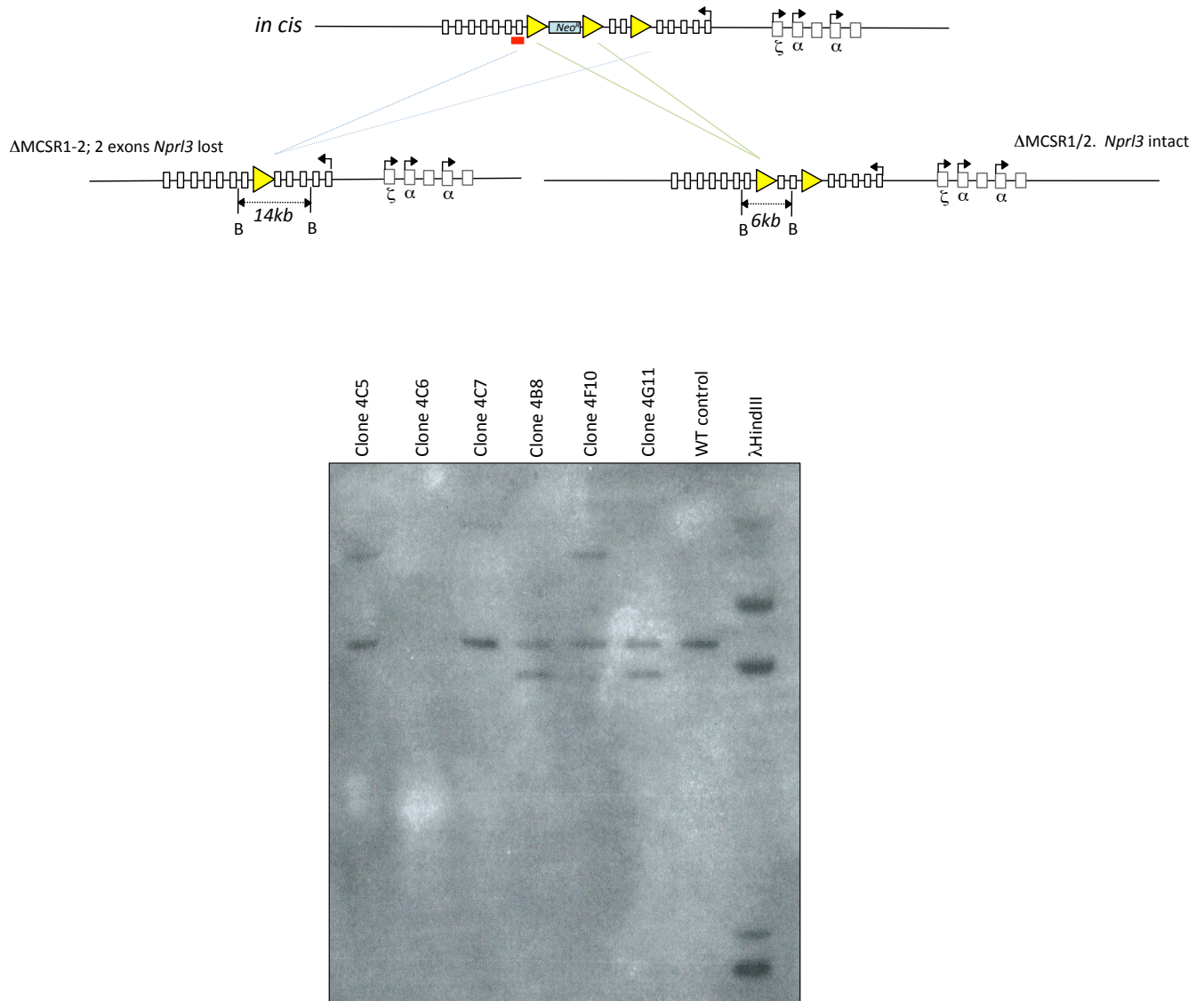


Figure 4.14 Southern blot using Bmt1 digest (cut sites marked 'B') and *L* probe (shown as red box) differentiates between WT, maximum deletion, and discrete deletion alleles.

Band sizes:
 WT: 7.1kb
 Maximal deletion: 14kb
 Discrete deletion: 6.1kb

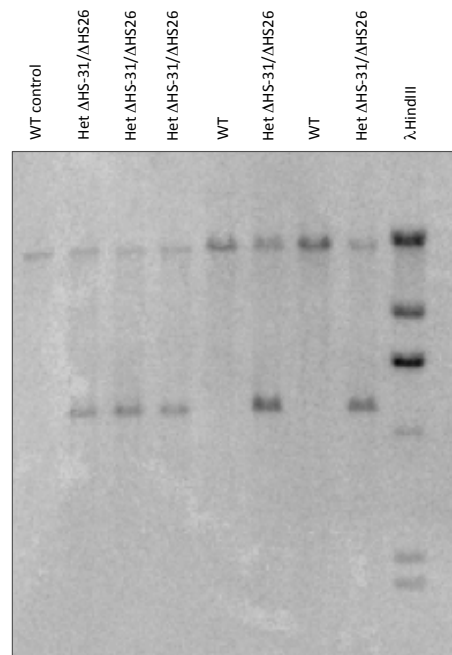
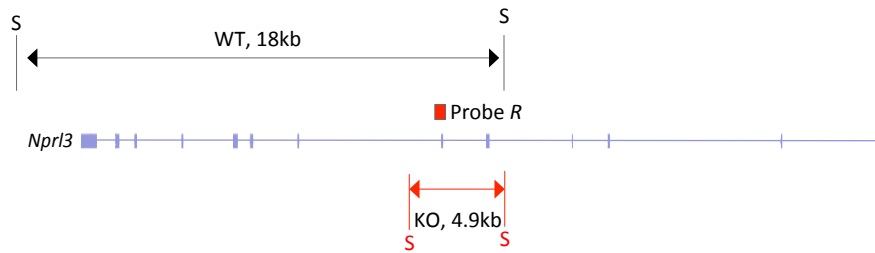


Figure 4.15 Agouti offspring of Δ R2, Δ R1 chimeras

Southern strategy is shown: Digest is with SpeI. A new SpeI site is present in the multiple cloning region left behind as a footprint of targeting following Cre recombination of the targeted Δ R1/ Δ R2 allele. This gives the shorter band with probe *R* compared with WT alleles. Expected band sizes: WT 18kb, deletion 4.9kb

4.4. Generation of the $\Delta R2/\Delta R3$ mouse

4.4.1 A critical mass of enhancers, or a critical role for R1 and R2?

That all the patient mutations described to date involve both R1 and R2 has been used to suggest that there is a specific, critical role for these two elements in driving α -globin expression. While this may be correct, other hypotheses are equally viable: it may equally be that deletion of, for example, elements R3 and R4 in conjunction, or R2 and R3 together, might also cripple the up-regulation of α -globin expression. (That patients with these mutations have never been described comprises scant evidence; and the maintenance of these elements through evolutionary history is likely to imply functional significance). Is there a specific significance to elements R1 and R2, or is there simply a requirement for a critical mass of enhancers at the α -globin locus?

4.4.2 Generation of the $\Delta R2/\Delta R3$ mouse

In order to address this, I have used reagents already available in the lab to generate a further double knockout by retargeting *in cis* the Cre-recombined $\Delta R2$ line with a plasmid for the deletion of R3 (HS-21), using a strategy identical to that described in section 4.3. A schematic summary, outline of the screening strategy and representative results are shown in figure 4.16.

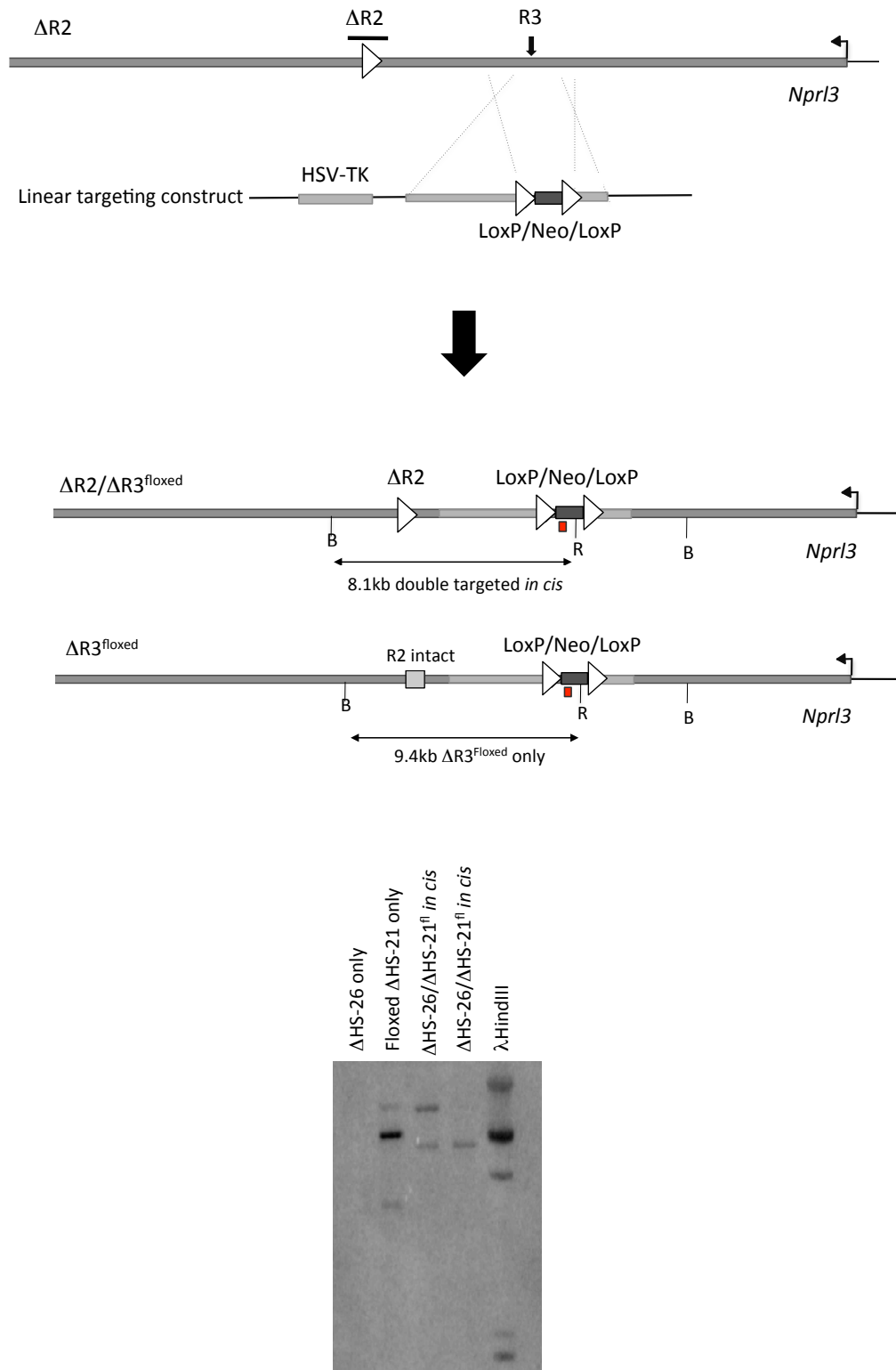


Figure 4.16 Screening strategy and Southern blot to confirm *in cis* targeting of $\Delta R3$ pNTFlox to $\Delta R2$ line. Lane 1 ($\Delta R2$ only) lacks a band, since the probe used is specific to the Neo^R cassette. Only cells from the clone tested in lane 4 were used for further analysis, due to the additional unexpected bands seen in lane 3.

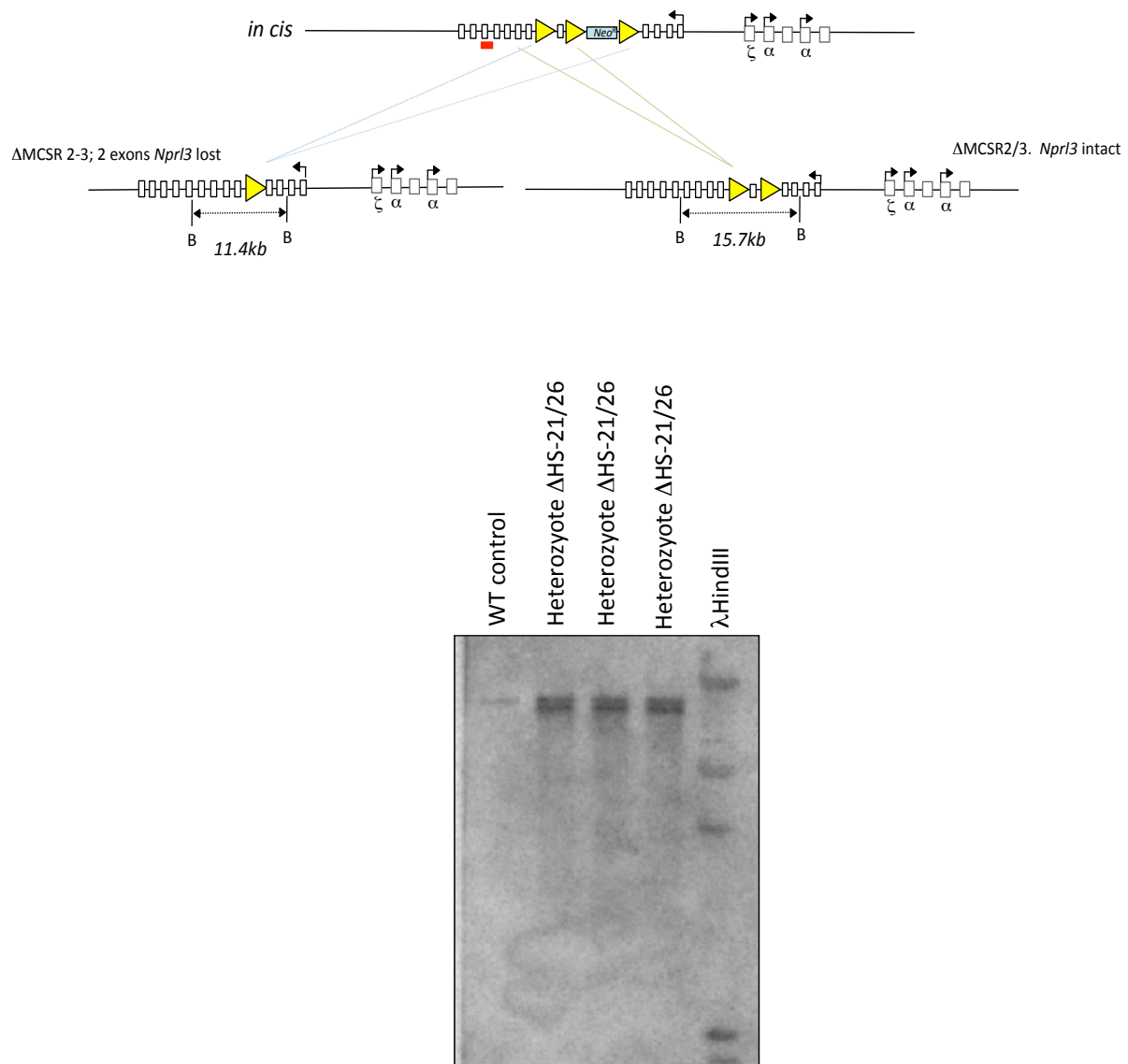


Figure 4.17 Southern blot to clarify discrete deletions of R2 and R3 following *in vitro* Cre recombination. Digest BamHI, probe L. Expected band sizes: WT 17kb; maximal deletion 11.4kb; discrete deletions 15.7kb

4.5 Chapter Summary

Chapter 3 highlights that heterologous models for the study of *α -globin* do not accurately recapitulate normal transcriptional control, and underlines our limited understanding of the contribution of the conserved regulatory elements in *α -globin* expression. In this chapter, I have outlined our rationale for addressing this question, by completing a series of knockouts at the mouse *α -globin* locus. This series should:

1. Define the elements critically required for *α -globin* expression in the mouse;
2. Permit functional comparison between the murine and human *α -globin* loci;
3. Permit the investigation of interaction between the *α -globin* enhancers.

The analysis of the phenotypes of the mice generated forms the basis of the next chapter.

Chapter 5

Characterizing Mouse Models

5. Characterizing mouse models of enhancer knockouts

5.1 Introduction

Each of the five regulatory elements deleted from the mouse *α-globin* locus would be highlighted as an enhancer using the criteria employed for the typical genome wide assays of enhancer detection. It is not clear, however, whether each element acts in an equivalent manner, or whether there is significant redundancy between them. Can this collection of *cis* acting elements (and other *cis* elements like them) be functionally sub-classified? In order to address this question, it is necessary to have a clear understanding of the impact of each element's deletion on *α-globin* transcription.

For each of the mouse models described in chapter 4, the initial characterization was performed through investigation of its viability and haematological phenotype, before focusing on the effect on expression of *α-globin*. This chapter reports the

.....

initial characterization of each model, and highlights the shortcomings of the methods used.

5.2 Results

5.2.1 Viability of single knockout models

In the light of previously reported deleterious effects of the neomycin resistance cassette on local transcription, knockout models were examined only following Cre recombination to excise the resistance cassette.

The viability of single knockouts of R2, R3, R(m) and R4 have been previously described (Anguita et al (Anguita 2002); Kowalczyk, Hughes, Lee, unpublished observations). That mice homozygous for these deletions in isolation survive to adulthood and breed normally implies that there is no essential role for these elements in α -globin transcription, or in erythropoiesis more generally. Each one, in isolation, is dispensable. Table 5.1 gives data regarding the litter sizes and composition of heterozygote x heterozygote breedings for each model, undertaken during this study. In some pedigrees heterozygote crosses were infrequently performed (because obtaining homozygous knockouts was the desired outcome), potentially giving misleading results (e.g. Δ R3 and Δ R4). Where fewer than three litters were available a similar assessment of heterozygote x homozygote crosses was examined to provide additional information (table 5.2).

The newly developed knockout model of R1 (HS-31) was also found to be viable in homozygosity. Litter sizes from homozygote matings were equivalent to those from

Litter size from Het x Het crosses

Model	$\Delta R1$	$\Delta R2$	$\Delta R3$	$\Delta R(m)$	$\Delta R4$	$\Delta R1+\Delta R2$	$\Delta R2+\Delta R3$
Number of pups	7.0	7.5	8.0	11	7.5	5.3	6.5

Percentage Homozygotes, Heterozygotes and WT mice in litters from Het x Het crosses

	$\Delta R1$	$\Delta R2$	$\Delta R3$	$\Delta R(m)$	$\Delta R4$	$\Delta R1+\Delta R2$	$\Delta R2+\Delta R3$
WT	28	18	25	18	14	23	31
Het	43	53	25	55	33	77	31
Hom	28	29	50	27	53	0	38
χ^2	0.143	0.857	6.000	0.273	6.467	10.700	2.077
P	0.9311	0.6514	0.0498	0.8725	0.0394	0.0047	0.3540

Table 5.1 Variation in litter size and homozygote numbers from heterozygote x heterozygote crosses for each of the mouse regulatory element knockouts

Percentage Homozygotes and Heterozygotes in litters from Het x Hom crosses

	$\Delta R1$	$\Delta R3$	$\Delta R(m)$	$\Delta R4$
Het	61	62	50	46
Hom	39	38	50	54
χ^2	0.692	0.500	0.000	0.077
P	0.4054	0.4795	1.000	0.7815

Table 5.2 Litter composition for single knockout homozygote x heterozygote crosses

.....

wild type matings, and the proportions of WT, heterozygotes and homozygotes were as expected, suggesting no defect *in utero*.

5.2.2 Viability of double knockout models

Heterozygotes for the $\Delta R1/\Delta R2$ knockout line were viable as expected, and were able to breed. The only kindred in which a reliable difference was observed between actual and expected litter composition of heterozygote crosses was for the double knockout $\Delta R1/\Delta R2$. This was highly statistically significant ($P=0.0047$). Litter sizes were also typically smaller than those of wild type or single knockout heterozygote crosses. The sacrifice of pregnant females following timed matings suggested that homozygotes for the double deletion were dying *in utero* at ~13.5 days (table 5.3, figure 5.1). This is consistent with embryonic lethality at a time corresponding to the switch from primitive to definitive erythropoiesis. However, that fetuses observed at this time were also small and pale implies defective erythropoiesis prior to that point, and would be consistent with a role for one or both elements in embryonic erythropoiesis also.

By contrast, homozygotes for the $\Delta R2/\Delta R3$ double deletion were born in the appropriate Mendelian ratio, and were viable to adulthood.

5.2.3 Haematological phenotypes

The haematology of mice varies with age and, if inbred, with strain (Russell et al 1951). We have chosen to examine the haematological phenotype of all mice at age

Percentage Homozygotes, Heterozygotes and WT mice in litters from $\Delta R1/\Delta R2$ Het x Het crosses

	$\Delta R1/\Delta R2$ – term litters	$\Delta R1/\Delta R2$ – in litters at 12.5dpc
WT	23	30
Het	77	48
Hom	0	22
χ^2	10.700	0.333
<i>P</i>	0.0047	0.8465

Table 5.3 $\Delta R1/\Delta R2$ heterozygote x heterozygote pregnancies: homozygotes found in expected numbers in pregnancies sacrificed at 12.5dpc**Figure 5.1** $\Delta R1/\Delta R2$ heterozygote (left) and homozygote (right) littermates, taken at e14.5. The much smaller and paler homozygote is already in the process of being resorbed.

.....

>8 weeks, when there is evidence to suggest there is minimal inter-individual variation (Biology of the Laboratory Mouse, 2nd edition, Jackson Laboratory Members). Since all mice are generated on the same complex background, we would not anticipate any impact from strain variation. We have also analysed a balanced mix of males and females, though there is little data supporting sex-variation in haematological parameters in mice (Wirth-Dzięciołowska et al. 2009). Where possible for the analysis of individual mouse lines, we have used normal littermates as a control. However, to compare the mouse models with each other, we have pooled wild type samples from different pedigrees.

The blood counts of homozygote knockouts of R2, R3 and R(m) had been previously established (Anguita et al. 2002; Kowalczyk, Hughes, Lee, unpublished observations), but were repeated using the *Horiba Medical Scil Vet abc Plus+* instrument to ensure that technical variables were not responsible for any differences observed.

To these data are added new results from the $\Delta R1$ homozygotes, $\Delta R2/\Delta R3$ homozygote knockout and the $\Delta R1/\Delta R2$ heterozygotes. We have been unable to obtain blood from live homozygotes with the double deletion of R1 and R2, prior to embryonic day 13.5; for this model, heterozygote mice only were analysed.

It is notable that for none of the single knockouts does inspection of the peripheral blood films for each model demonstrate any detectable deviation from wild type (figure 5.2). Likewise, no single knockout reduces haemoglobin concentrations in the

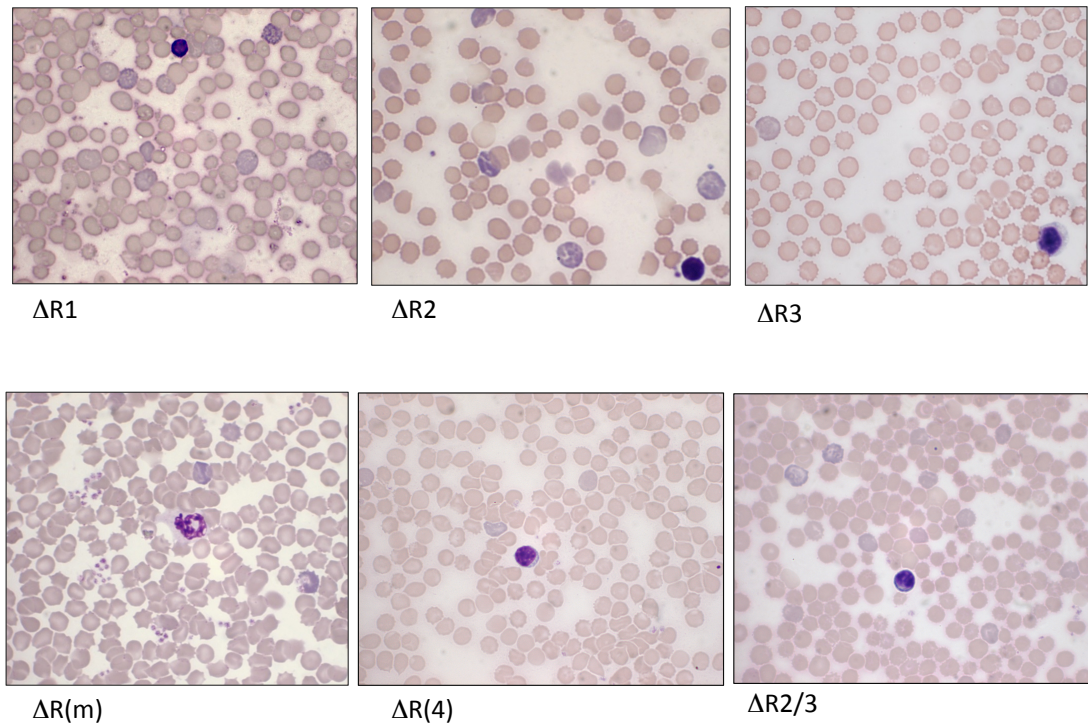


Figure 5.2 Peripheral blood films from mice homozygous for regulatory element knockouts.

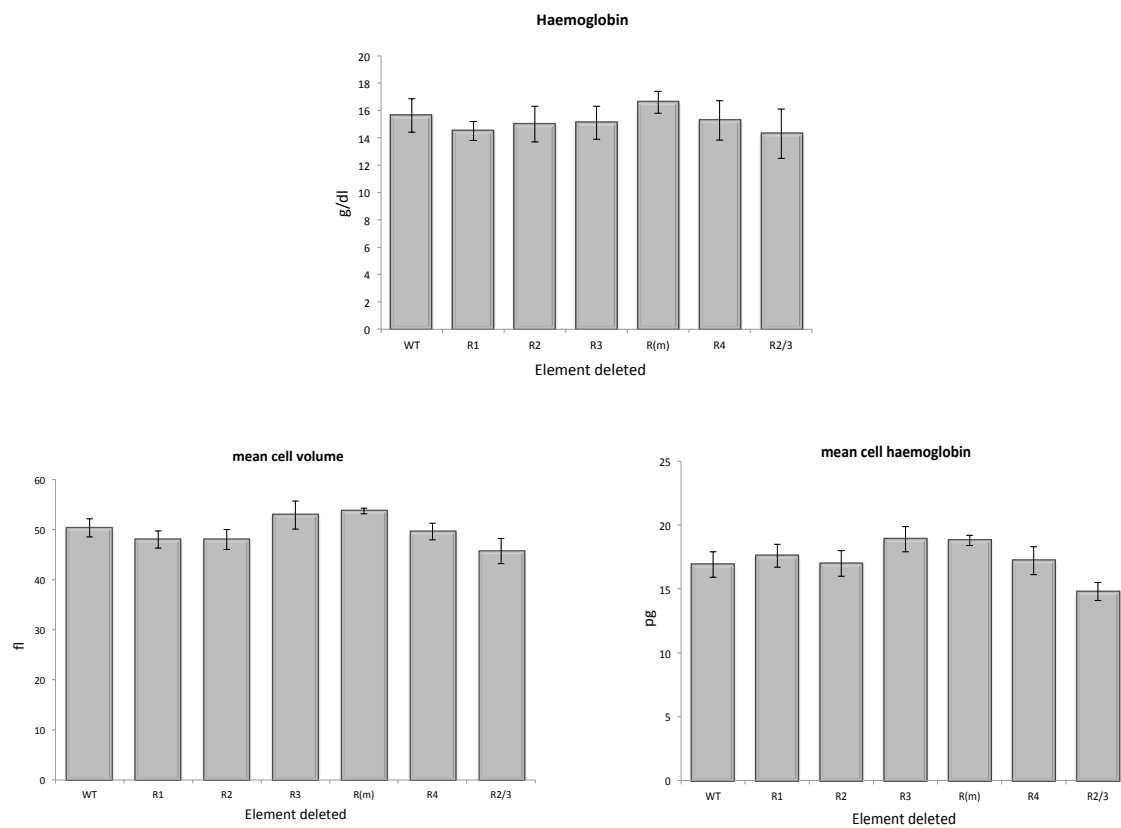


Figure 5.3 Peripheral blood haemoglobin and red cell indices from mice homozygous for regulatory element knockouts.

blood of adult mice (figure 5.3), nor do they result in a statistically significant reduction in either mean cell volume or mean cell haemoglobin - though the $\Delta R2$ ($\Delta HS-26$) homozygote does have a non-significant trend toward a lower MCV and MCH as reported previously (Anguita et al 2002).

Reticulocyte counts in peripheral blood films stained with brilliant cresyl blue reflect the rate of red cell formation, and thus give an indication of the degree of erythroid stress for each model. Counts are made by assessors blinded to each mouse's genotype, and are then recounted by a second assessor independently. Figures 5.5 and 5.6 demonstrate the degree of variation in reticulocyte numbers in homozygote knockout models, with a statistically significant increase in reticulocyte numbers in $\Delta R1$ homozygotes relative to WT. This has previously been demonstrated for $\Delta R2$ also (Anguita et al 2002), and suggests a haematologically significant impact of these deletions.

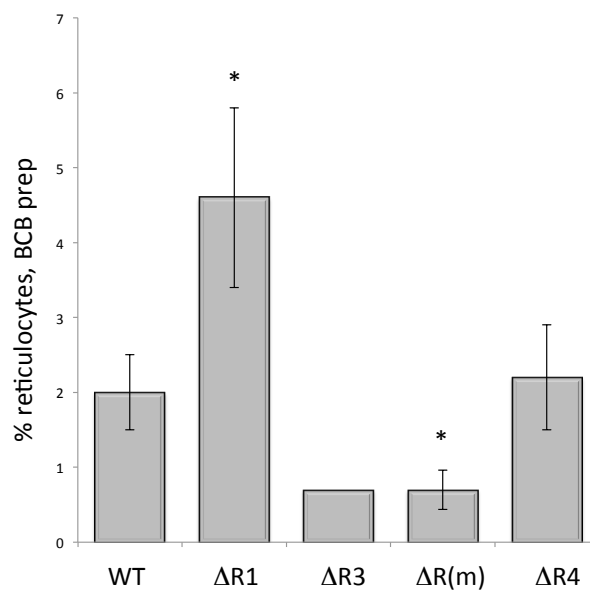


Figure 5.5 Peripheral blood reticulocyte counts from mice homozygous for regulatory element knockouts, as determined by blinded assessment of blood films stained with brilliant cresyl blue.

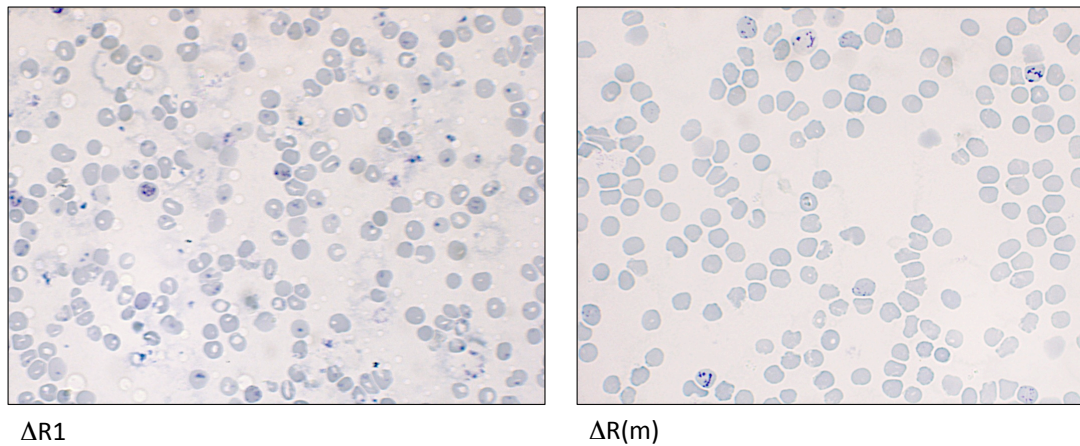


Figure 5.6 Sample reticulocyte preparations (brilliant cresyl blue stain) from peripheral blood of mice homozygous for $\Delta R1$ and $\Delta R(m)$

The haematological double knockout $\Delta R1/\Delta R2$ has been assessed in heterozygosity, since homozygotes do not survive to birth. The results are shown in figure 5.7. As well as increased reticulocyte numbers, $\Delta R1/\Delta R2$ heterozygotes also have a statistically significant reduction in mean cell volume and mean cell haemoglobin, consistent with thalassaemia.

It is worth emphasizing that the double knockout of $\Delta R2/\Delta R3$ is no more severely affected than the $\Delta R2$ knockout alone; in contrast to mice which are merely heterozygous for the $\Delta R1/\Delta R2$ deletion, homozygotes for $\Delta R2/\Delta R3$ have no significant reduction in either haemoglobin or red cell indices.

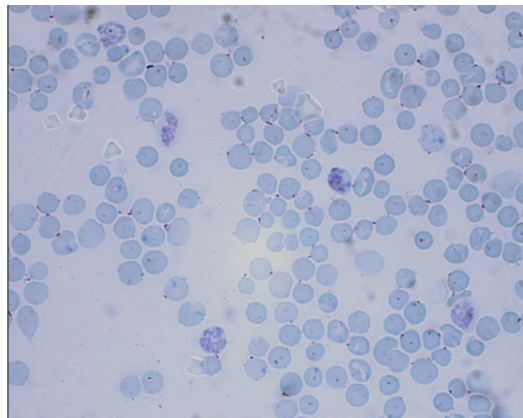
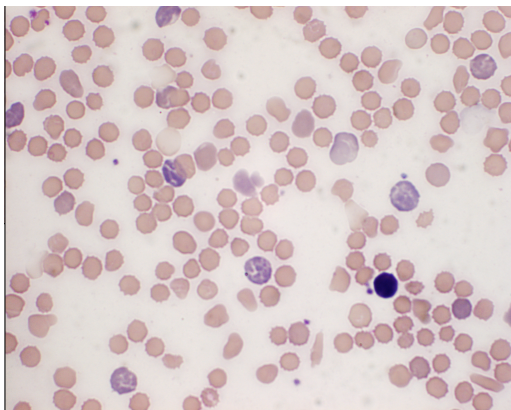
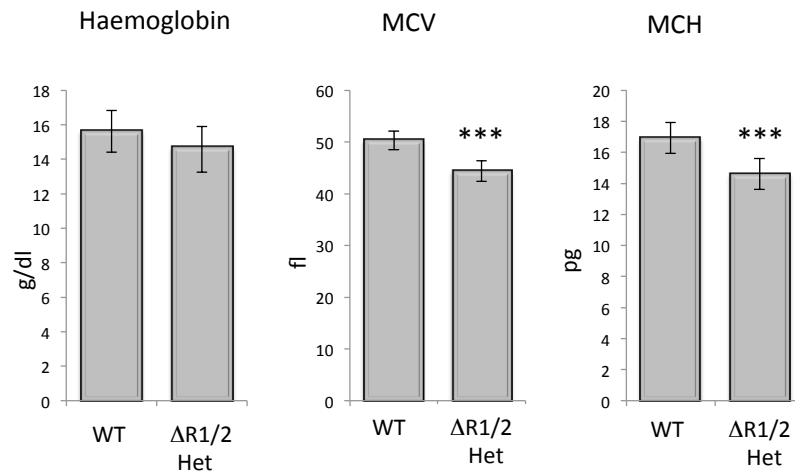


Figure 5.7 Haematological analysis of $\Delta R1/\Delta R2$ heterozygotes.

Top: peripheral blood haemoglobin and red cell indices. Bottom left: peripheral blood film showing anisocytosis and polychromasia. Bottom right: reticulocyte preparation (brilliant cresyl blue) showing increased numbers of reticulocytes. All features are consistent with a thalassaemic phenotype.

.....

Even this most basic of phenotypic assessments goes some way to answering key questions about the *α-globin* regulatory elements. First, we can conclude that none of the enhancers is required absolutely for adequate erythropoiesis. That no significant impact from any individual knockout is seen confirms that none of these enhancers has a critical role that cannot be co-opted by one of its neighbouring elements. Further, we can conclude that, rather than a critical mass of enhancers being needed for viability and normal haematological features, there is evidence to support a distinct and non-redundant role for R1 and R2.

The full blood count is a physiologically relevant outcome measure. In addition, there is little risk of perturbation of the system in the process of collecting or interpreting these data (contrast some of the other methods we have employed below). However, it is at best a blunt tool for the assessment of impact of enhancer deletion: taking the example of patients with *α*-thalassaemia trait, a reduction of *α-globin* expression to 50% of normal values (e.g. by deletion of both *α-globin* genes *in cis* in the common South East Asia deletion) will have at most a trivial impact on the haematological phenotype. This is entirely consistent with the well-studied phenomenon of canalization, in which a phenotype is preserved in the face of a genetic change, through compensatory mechanisms at either a genetic or systems level.

Therefore, using blood counts to assess these models will establish only the most marked of disruption of globin transcription, and more sensitive methods of readout

.....

are required before we can conclude that deletion of any regulatory element is without effect.

5.2.4 Assessing α -globin mRNA levels by qPCR

Quantification of α -globin mRNA will provide a more direct assessment of the impact of enhancer deletion on the α -globin locus. However, the very high levels at which globin mRNA is detected in erythroid cells makes the identification of suitable normalizers for qPCR challenging. Using genes expressed at substantially lower levels as normalizers, small changes in normalizer transcription may result in hefty (but misleading) 'corrections' to the C_t readout of test transcripts. Commonly used normalizers such as GAPDH and β -actin may amplify several qPCR cycles behind the globin transcripts, and are therefore suboptimal normalizers.

Here we have used ribosomal 18S as a normalizer in the light of its high level transcription. It also has the merit of being essentially stable during the process of erythropoiesis (Merryweather-Clarke et al. 2011). Notwithstanding, figure 5.8 shows the substantial difficulties with using qPCR to try to assess the impact of enhancer deletions on globin transcription.

Appropriate technical replicates in a single qPCR plate should be <0.5 cycles apart – corresponding to a difference in expression of ~50%. Thus, while qPCR is excellent for determining fold-changes in gene expression, its ability to detect fractional changes in a reliable fashion is questionable. Even with the most accurate experimental technique, it is unlikely that qPCR can reliably detect fractional changes

of much less than 50% - yet such fractional changes are all that is expected with any of the knockout models we have been able to analyse as adult mice.

The substantial inter-assay variation shown in figure 5.8 demonstrates that, although there is no statistically significant difference between wild type and single enhancer knockouts or even $\Delta R1/\Delta R2$ heterozygotes, the technique is likely to be insufficiently sensitive for our aims in this experiment.

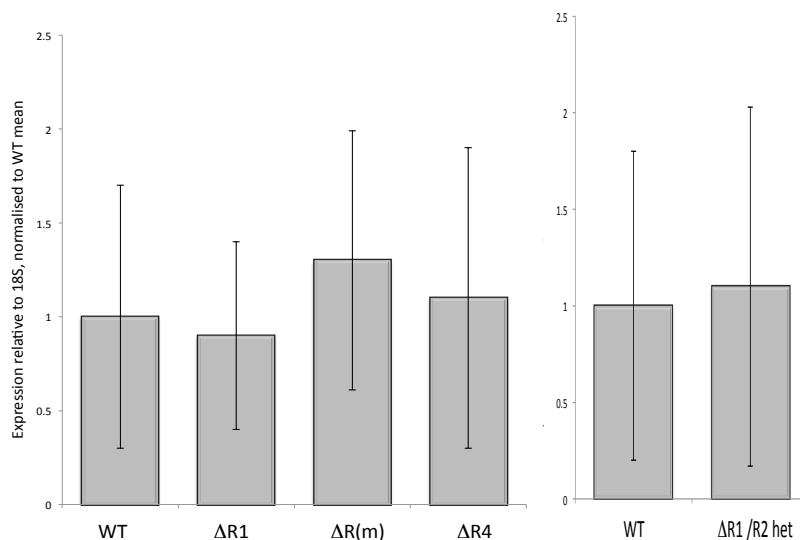


Figure 5.8 qPCR is insufficiently sensitive to detect fractional changes in α -globin expression

The left hand histogram shows attempts to quantify α -globin expression from the peripheral blood of WT mice and homozygotes for single enhancer knockouts; the right hand panel shows WT vs heterozygotes for the $\Delta R1/\Delta R2$ double knockout.

5.2.5 Assessing α -globin mRNA levels by RNase protection assay

This technique involves labelling single stranded RNA molecules of interest, with subsequent detection by autoradiography. Although very helpful in confirming the presence or absence of transcripts of interest (see, for example, the discussion of α -

globin expression in MEL hybrids in chapter 3), its ability to quantify small-scale differences is limited. The quantification of bands is performed autoradiographically; a large component of subjective judgement is needed to define the extent of the bands to be quantified. This, combined with the relatively laborious nature of the technique, and its requirement for exposure to GTP with a very high specific activity, has made this procedure largely outmoded.

RNase protection assay has been performed on the previously generated mouse models, with results summarized in figure 5.9.

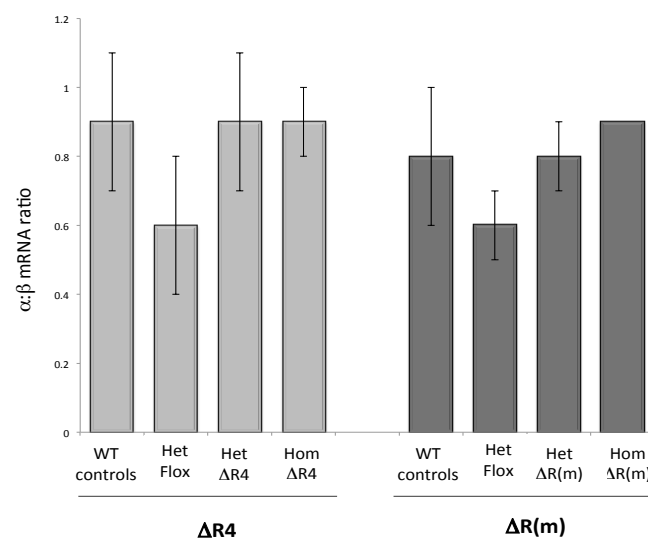


Figure 5.9 RNase protection assay is insufficiently sensitive to detect fractional changes in α -globin expression. There is no statistically significant difference between WT mice and those with either floxed or knockout alleles for R4 or R(m), or for R2 and R3 (data not shown).

These support the observation that there is limited impact from deletion of the regulatory elements R2, R3, R4 and R(m), but a standard deviation of 10-20% can be

.....

seen for $\alpha:\beta$ mRNA ratios. Although this method purports to be more sensitive than qPCR in this context, and does not require the amplification phase of qPCR (which may itself distort the accurate quantification of target genes), it still cannot provide the sensitivity needed to permit the detection of minor changes in transcription as a consequence of enhancer deletion.

5.3 Discussion – limitations in the initial phenotype assessment

The analysis of primary haematological data has some advantages: unlike most methods of measurement, the process of obtaining the data does not impact upon the system being measured. Its major weakness is that it is not able to take account of the potential compensatory mechanisms which will tend to minimize the impact of each knockout on gene expression.

We can draw the following conclusions, however:

1. There is no single element without which the *α -globin* locus is unable to be transcribed. No single element is absolutely required for globin transcription. Whether the loss of individual elements can, on their own, contribute to a reduction in *α -globin* expression cannot be discerned from these data.
2. There is clear evidence of non-equivalence between the elements. The double knockout of R1 and R2 has a notably more severe phenotype than the double knockout of R2 and R3. Rather than a critical mass of enhancers being required for globin synthesis, there is specific significance to either or both of R1 and R2.

-
3. Haematological phenotype is too blunt a tool to be suitable for addressing the exact impact of each enhancer deletion on transcription from the α -*globin* locus, and our current means of assessing transcription output are also of limited usefulness.

Therefore, in order to establish in detail the impact of each deletion on transcription, a more sensitive method needs to be established. This is the subject of chapter 6.

Chapter 6

Nascent transcription from mouse models

6. Nascent transcription from knockout mouse models

6.1 Introduction

There is a paradox in the phenotypes of the knockout mice described in the preceding chapter. Why should these regulatory elements, each bearing the marks of active enhancers, be subject to evolutionary conservation (with the implication of function that entails) yet be entirely dispensable from a phenotypic perspective?

This is not a situation confined to the α -globin locus. As described in chapter 1, apparently dispensable enhancers have been described at the β -globin locus, the *IgH* locus and, most comprehensively analysed, at several loci in the *Drosophila* genome.

From first principles, there are several possible explanations:

- Akin to the ‘shadow enhancers’ defined in *Drosophila*, conserved elements might be needed for preservation of full function in the context of genetic or

environmental perturbation. However, what such perturbation might be at the *globin* loci is not clear. Haemorrhage and haemolysis (induced, for example by malaria) are evolutionarily important erythroid stressors – but the physiological response to these is found at the level of red cell progenitor proliferation, through the influence of erythropoietin, rather than up-regulation of expression at individual *globin* loci. The interrogation of haematological parameters following the pharmacological induction of haemolysis has been performed in several of our mouse models, with no additional impact demonstrated relative to wild type (data not shown).

- A threshold number of transcripts may be required for a normal haematological phenotype – with not all enhancers contributing sufficiently to the total transcriptional output to impact upon this threshold. Above a certain level of α -*globin* transcription, the number of transcripts may no longer be the rate-limiting factor in globin chain synthesis. Simply, our phenotypic readout may be too well buffered to detect the role of minor elements in transcriptional regulation.
- qPCR and RNase protection assay with their 10-20% minimum standard deviation may be too blunt a tool to detect minor contributions to α -*globin* transcription.
- There may be genuine redundancy between elements. Clearly this cannot apply absolutely to all elements: Deletion of R1 and R2 together is beyond compensation by other elements. If there is redundancy, it can only be in the context of a functional hierarchy across the elements.

-
- The elements might function differently at different stages of erythroid development or maturation. There is no evidence to guide us on the latter; with regard to the former, we can only comment that no individual element is absolutely required for ζ -globin expression, since all homozygotes survive to adulthood. We may argue on the basis of the data presented in chapter 5 that R1 and R2 play a role in embryonic globin synthesis - but they cannot be critically required since fetuses homozygous for the double knockout survive beyond the point of switch from primitive to definitive erythropoiesis.
 - The elements could function differently in different tissues – explaining their conservation in tissues where they appear to be without an active role. At the time of generating and interrogating these models, the red cell was the only known tissue in which α -globin was expressed. This has subsequently been shown to be an oversimplification.

Not all of these possibilities are experimentally tractable within the confines of this thesis. For example, a true appreciation of redundancy at the locus would involve the generation of an exhaustive set of combinatorial knockouts that is neither practicable nor financially feasible. We might, however, improve the sensitivity of our assessment of enhancer function by improving the precision of our question: instead of examining a phenotypic outcome of enhancer deletion at the organismal level, or quantifying accumulated total RNA, we could instead assay direct transcriptional output by quantifying nascent transcripts. By concentrating on nascent transcription, we may avoid any potential threshold effect in accumulated mRNA and thus in haematological phenotype.

.....

This chapter describes the development of a technique to quantify nascent transcripts from the α -globin locus, and presents results for each knockout model.

6.2 Results

6.2.1. Choosing a system to investigate: fetal liver cultures

Since the $\Delta R1/\Delta R2$ homozygote does not survive to birth, we have chosen to use erythroid culture from fetal liver to analyse the impact of each enhancer knockout on globin gene expression. This system, described in chapter 2, also has the advantage of permitting metabolic labelling of nascent transcripts.

Fetal liver cultures have the potential to yield synchronously maturing erythroid cells. Although the cellular population is mixed at the time of collection, the more mature cells are progressively lost while earlier progenitors proliferate in SCF-containing expansion medium. After expansion, cells are induced to undergo terminal differentiation by transfer into a medium containing increased concentration of erythropoietin and transferrin, and no SCF (see chapter 2 for details). Figure 6.1 shows that there is an ordered progression through erythroid maturation after changing (at $t=0h$) from expansion medium to differentiation medium. We have noted, however, that prolonged periods of expansion lead to increasing spontaneous differentiation, and loss of the synchronicity of maturation.

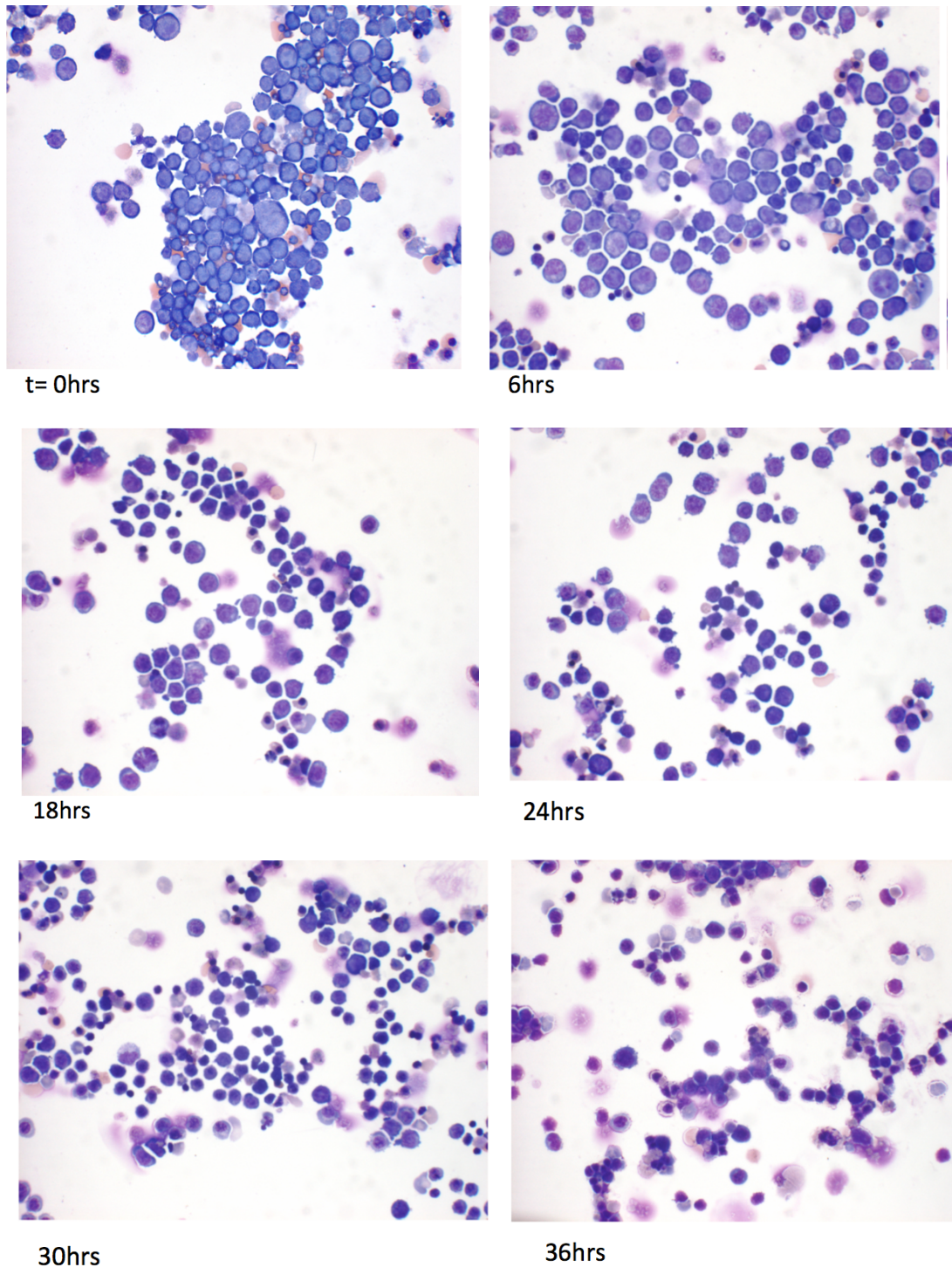


Figure 6.1 Cytospins demonstrate synchronous maturation of fetal liver erythroid cultures over 48 hours in maturation medium, from a uniform population of pro-erythroblasts (t=0 hours after adding maturation medium) through basophilic and orthochromic normoblasts (18-24 hours) to late erythroblasts with pyknotic nuclei (36 hours).

.....

Samples are, comparable between wild type and enhancer knockout models, as shown morphologically in figure 6.2. Immunophenotypic analysis of each sample shows a heterogeneous composition with regard to Ter119 and CD71 expression; more homogeneous cultures may be obtained by the depletion of mature erythroid cells from the fetal liver cultures prior to expansion and cultures expanded from Ter119 negative cells have a substantially purer composition. However such cultures are substantially smaller in volume, meaning that that this strategy was unsuitable for yielding the large cell numbers required for nascent RNA collection.

Figure 6.3 shows immunophenotypic profile for fetal liver cells from WT and two knockout models at three time points during erythroid differentiation. There are some inconsistencies from culture to culture in the expression of CD44; this implies some variation in the maturation of the cultures that is not evident from morphological inspection alone. CD71 and Ter119 expression are relatively stable.

6.2.2. Collecting newly transcribed RNA

Several possible proxy measures for nascent RNA have been described. Isolating transcripts which have not undergone polyadenylation will select a proportion of newly transcribed RNA – in addition to a variety of other species of transcript (Yang et al. 2011). Unspliced transcripts may also be quantified, though it is recognized that the process of transcription and splicing take place simultaneously in many cases (Montes et al. 2012). Khodor and colleagues (Khodor et al. 2011) suggest an alternative method of obtaining newly synthesized RNA by isolating only nuclear

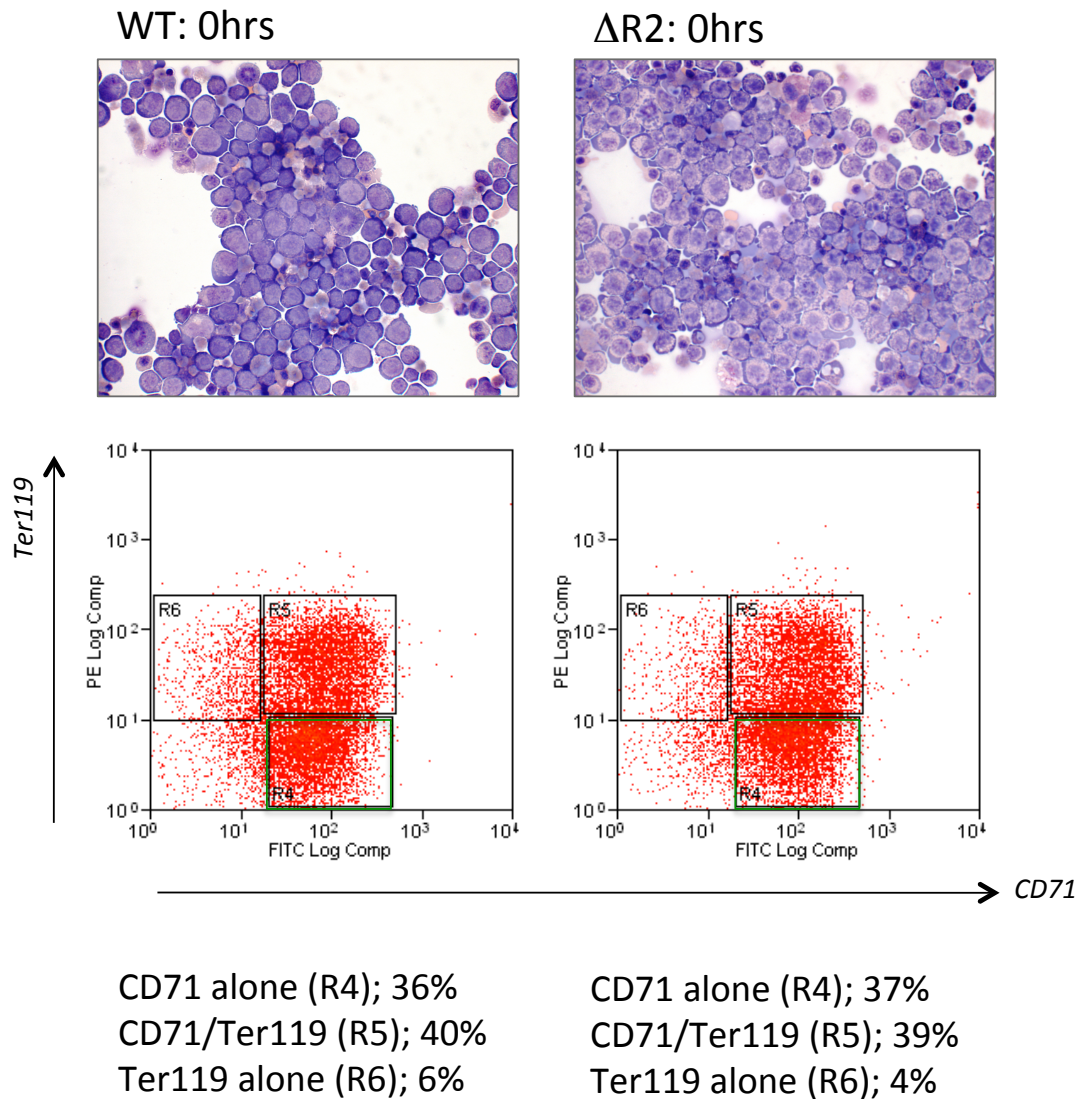


Figure 6.2 Comparison between morphology and immunophenotype in WT and $\Delta R2$ homozygote fetal liver cultures, at three points during erythroid maturation (where transfer to differentiation medium takes place at 0 hours): this page t=0h; next page t=24 and t=48h.

FACS plots show Ter119 and CD71 expression. Ter119 expression is gained during erythroid maturation. CD71 (transferrin receptor) is present on early and intermediate normoblasts, but is lost on mature red blood cells. No single sample has a wholly uniform population of maturing erythroid cells.

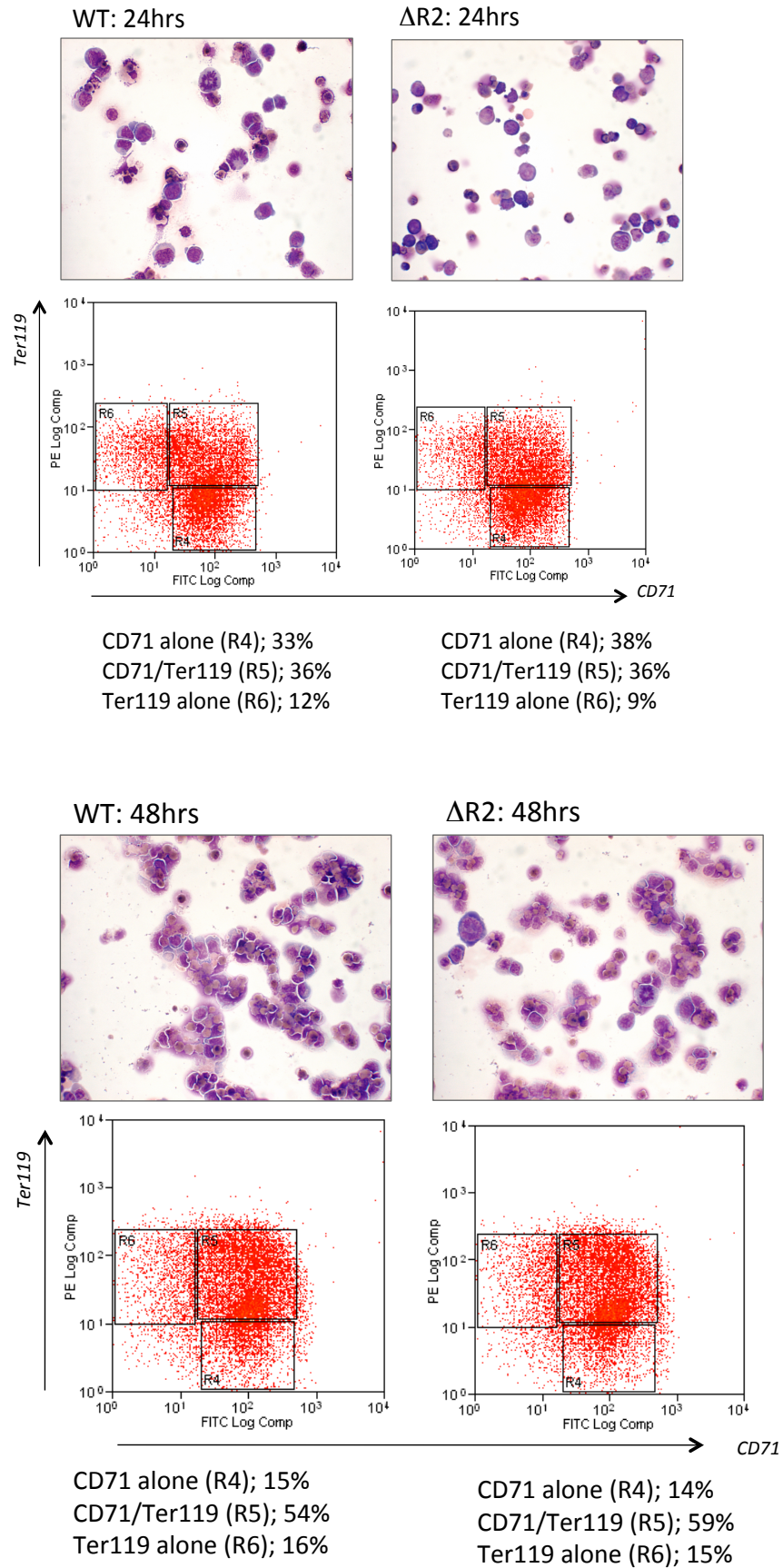


Figure 6.2 (continued)

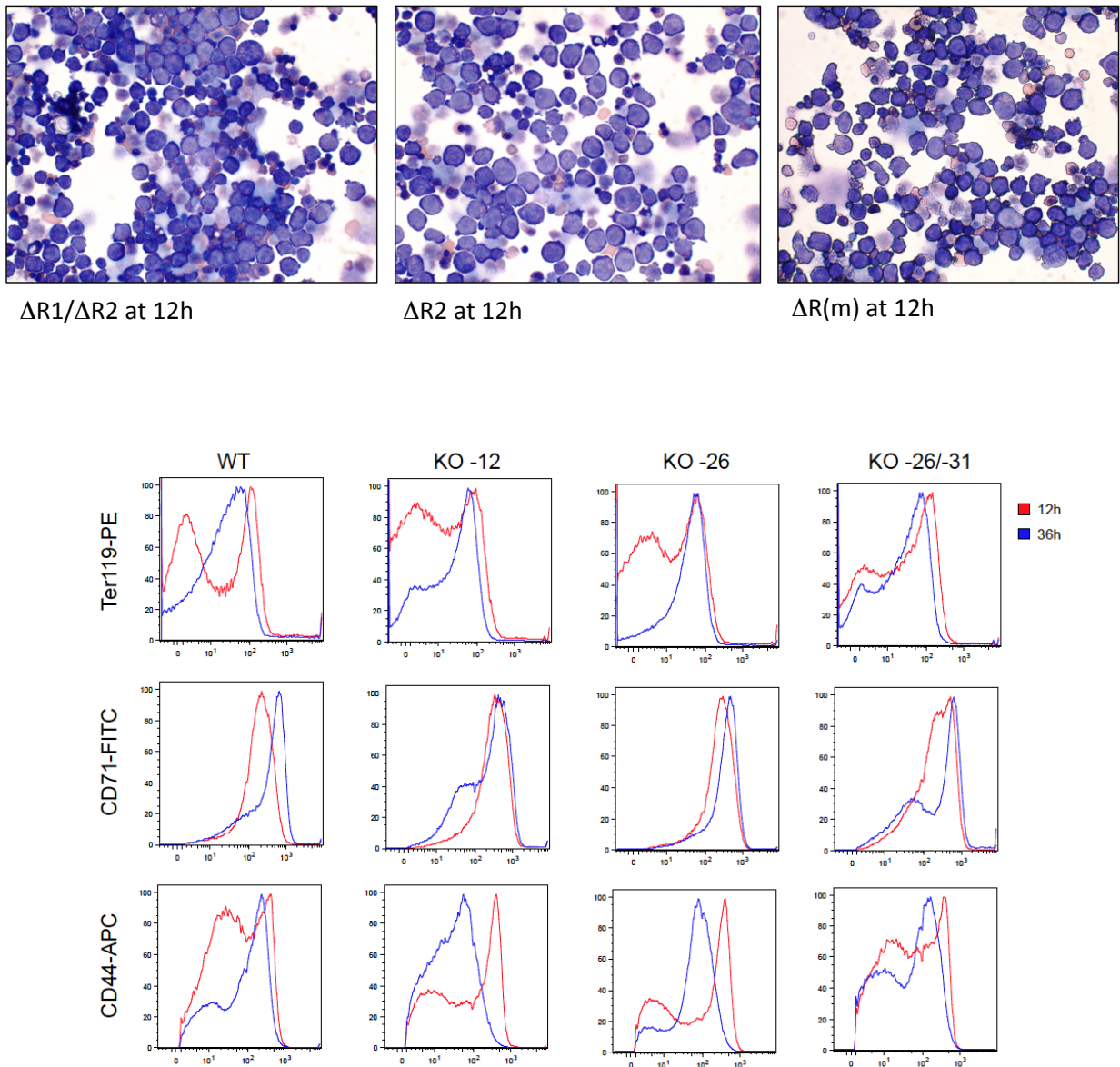


Figure 6.3 Immunophenotypic comparison of WT and mutant fetal liver cultures using Ter119, CD71 and CD44 at 12 and 36 hours following transfer from expansion to differentiation medium.

transcripts, rather than including total cellular RNA (the cytoplasmic proportion of which is mature and awaiting translation). Global run-on methods have also been reported (Core et al. 2008), but are technically challenging.

For technical simplicity and the closest approximation to newly synthesized mRNA we have chosen to use metabolic labelling of primary erythroid cells in liquid culture followed by selective purification of only labelled transcripts from total extracted RNA. The methodology is described in detail in chapter 2, and the workflow is shown in figure 6.4.

Add 4-thiouridine at final concentration of 500mM

Incubate for 30 minutes at 37C

Pellet cells and lyse in Tri-reagent

Standard RNA extraction

HPDP-biotinylate labelled RNA

Elute with DTT, and clean up RNA

Figure 6.4 Workflow summary for obtaining nascent RNA from fetal liver cultures

.....

Briefly, cells are labelled for a predetermined period of time (we chose 30 minute to 1 hour intervals) with a 500mmol final concentration of 4-thiouridine, which is incorporated into the newly transcribed RNA in place of uridine. Total RNA is then extracted, and incubated with hexylpyridyl dithiopropionamide (HPDP)-biotin, in order to biotinylate the labelled RNA specifically. This is accomplished by a disulphide exchange between the labelled RNA's thiol group and the HPDP reagent's 2-pyridylthiol group. The result is a release of pyridine 2-thione, and formation of a biotinylated transcript with a disulphide bridge between the biotin and the transcript itself. Following pull-down on a streptavidin column, the nascent transcripts may be eluted by DTT-mediated reduction of the disulphide bridge.

This technique should capture both polyadenylated and polyA-negative transcripts produced in the interval of labelling; it should also capture both spliced and unspliced new transcripts. Further, the nascent RNA transcripts obtained are free of biotin (contrast technologies such as the Click-It nascent RNA isolation kit), making it suitable for downstream applications which require a biotinylation dependent stage (e.g. Nanostring).

6.2.4 Validating samples of nascent RNA obtained

All samples obtained using the nascent pulldown technique were analysed using the Agilent RNA PicoChip. Representative traces are shown in figure 6.5. Clearly, these are quite dissimilar to the traces obtained with good quality total RNA; however, since the RNA Bioanalyser process relies on an appropriate ratio between levels of

.....

ribosomal RNA detected to calculate an RNA integrity number, it is unlikely that newly synthesized RNA species will be present in the same proportion and distribution.

There are no published data to demonstrate the expected bioanalyser trace for labelled nascent RNA. There are, however, published traces for nascent RNA collected using other means (Khodor et al. 2011) This is shown alongside the representative trace in figure 6.5 for comparison.

To ensure that the sample we obtained using the method described above genuinely represents nascent transcription, the material obtained from a one-hour period of labelling was used for library preparation and RNA-seq analysis. Sample tracks from this experiment are shown in figures 6.6 and 6.7, which cover the *α -globin* cluster and *band 3* genes. Sequence alignment was performed by Dr Jim Hughes of the MRC Molecular Haematology Unit, WIMM. In both cases, stranded nascent transcription is evident (including intronic species). The short RNAs associated with enhancer transcription (Kowalczyk, Hughes, Garrick, et al. 2012b) are visible, but relatively poorly represented, suggesting their short half life compared to other nascent transcripts.

6.2.3 Quantifying nascent RNA

The shortcoming of qPCR analysis for quantifying *α -globin* message include the requirement for post collection conversion to cDNA, with an amplification phase that may distort the relative abundance of test and control mRNA; and the intrinsic

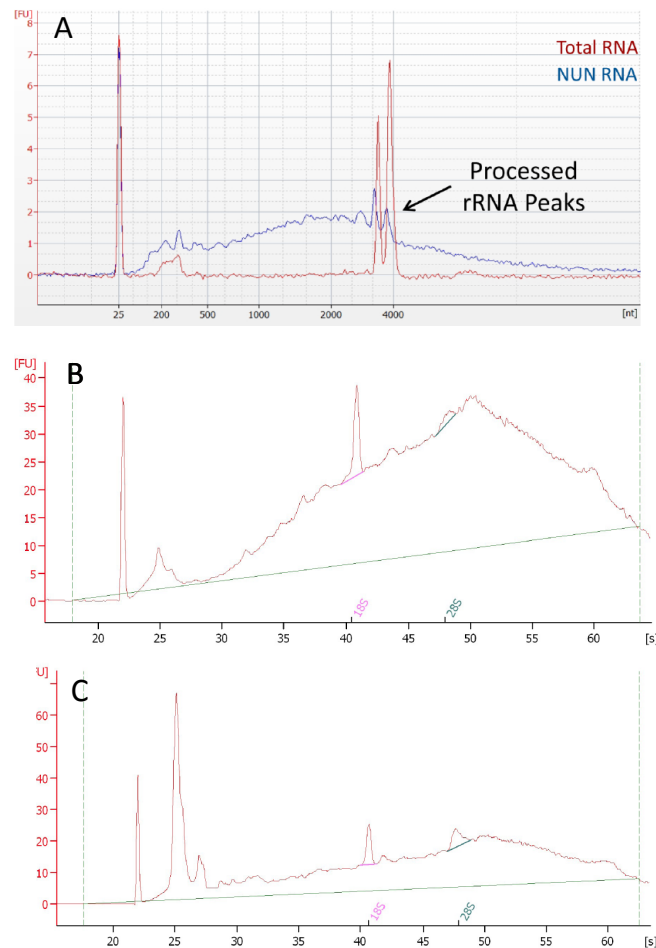


Figure 6.5 ABI RNA Bioanalyser traces for nascent RNA

A) shows a nascent RNA trace published by Khodor and colleagues (2012), where RNA was isolated from nuclear transcripts only. Here the red trace reflects total RNA, and the blue trace shows nascent RNA obtained by lysing nuclei in a solution of NaCl, urea and NP-40 (NUN).

B) and **C)** show sample traces from WT and $\Delta R1/\Delta R2$ fetal liver nascent RNA pulldowns respectively

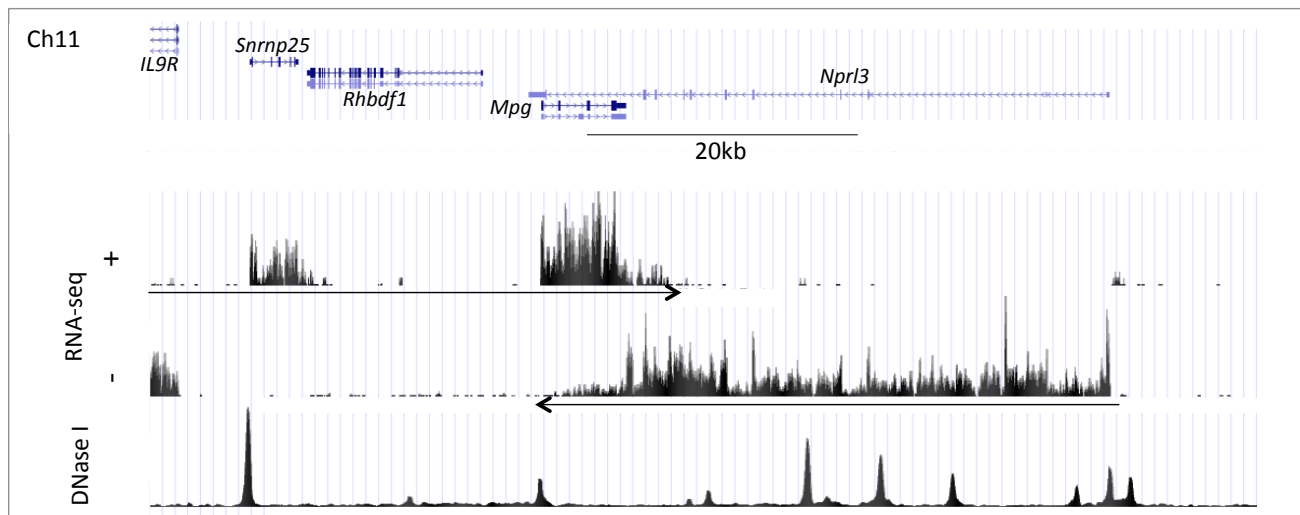


Figure 6.6 Nascent RNA sequencing across the α -globin locus. Above the tracks, refseq gene positions are shown in blue.

Top track: forward strand sequencing, with direction of transcription highlighted by the right-facing black arrow.

Middle track: reverse strand sequence, with left-facing arrow highlighting the direction of transcription. Note that the gene *Rhbdf1* is not transcribed in erythroid cells.

Bottom track: DNase I sensitivity in mouse erythroid cells, included to highlight the just-visible enhancer associated transcripts in the forward sequencing track.

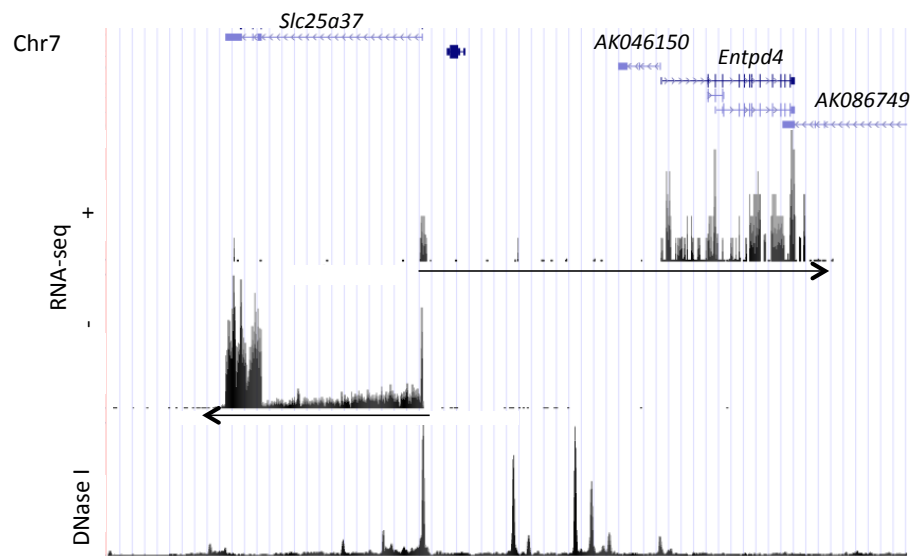


Figure 6.7 Nascent RNA sequencing across the *Slc25a37* locus. Above the tracks, refseq gene positions are shown in blue.

Top track: forward strand sequencing, with direction of transcription highlighted by the right-facing black arrow.

Middle track: reverse strand sequencing, with direction of transcription highlighted by the left-facing black arrow.

Bottom track: DNase I sensitivity across the same locus

.....

limitation to the sensitivity of the technique in terms of fractional rather than fold changes. This prompted us to consider using nanostring technology for RNA quantitation.

The principle behind nanostring technology is outlined in chapter 2; however, briefly, we hybridized nascent RNA to a codeset including biotinylated capture and labelled reporter probes by incubation at 65°C for 12 hours. The labelled sample is applied to the streptavidin-coated nanostring cartridge and excess probes are washed off. Finally, transcripts bound to the cartridge are subject to an electric field to ensure their six-colour barcodes are extended in a linear fashion and orientated in the same direction for readout. The NCounter data analyser then images the cartridge to which the probes are immobilized, and counts the number of each unique colour barcode. The final readout is reported as an integer, the simple number of RNA molecules of each type present on the cartridge.

In addition to avoiding the failings of qPCR highlighted above, nanostring has the advantage that its readout pertains directly to transcript numbers (rather than proxy marks such as C_t , subject in turn to a variety of mathematical manipulations). In addition, nanostring permits the multiplexed quantification of a large number of target genes (several hundred if required), and up to twelve samples, simultaneously. This would not be possible on a single qPCR plate, preventing the direct comparison of so large a range of samples and targets.

.....

Since splicing is expected to take place contemporaneously with transcription, it is necessary to design codesets that will detect both spliced and unspliced transcripts (i.e. probes confined to exons) in order to obtain an accurate estimation of the rate of transcription. The location of these probes is given in the appendix. We have also attempted to minimize the potential bias that might arise for detecting short versus long transcripts, by designing all probes within the first two exons of each gene.

Exonic probes will also detect genomic DNA. However, three steps should limit the possible contamination of the quantified material by DNA: first, total RNA extracted from labelled cultures is DNase treated; second, the nascent pulldown would not be expected to yield any DNA; and finally, the annealing process takes place at 65°C, with no denaturing (95°C) step incorporated.

Given the opportunities for multiplexed quantitation that the nanostring platform offers, we chose to expand our analysis beyond $\alpha:\beta$ *globin* ratios alone. Appendix 2 shows all the genes the transcription of which we have chosen to assess by nanostring. These include the full complement of *globin* genes, including adult and embryonic genes, and those of uncertain function (e.g. *θ -globin*); the genes located within a 2Mb window to either side of the *globin* genes, including those which are not typically expressed in erythroid cells; and a series of control genes.

6.2.5 Choosing control genes for nanostring analysis

The group of control genes was chosen to meet two criteria: first, genes should be expressed in erythroid cells, ideally at levels not too dissimilar from the *globin* genes themselves; and their expression should be stable at different stages of erythroid maturation.

With regard to the former, there are few gene products expressed in erythroid cells to such extremely high levels as α -globin. β -globin is one potential control; other highly expressed erythroid genes include *band3* (used as a control by Zon and colleagues for globin quantitation in zebra fish experiments, e.g. Ganis et al. 2012), *GypA* and *carbonic anhydrase*. Non-erythroid specific controls include *RP18S* and β -actin, both widely used comparators in qPCR studies.

There are few data to guide us regarding the stability of expression of genes during erythroid maturation in mouse. Some individual genes have well described kinetics of expression relative to α -globin (e.g. β -globin and *glycophorin A*); for others, the nature of their expression kinetics during erythroid maturation is obscure, and is potentially of relevance for the normalization of α -globin expression data.

The question of expression profiles in erythroid progenitors and maturing red cells has been addressed by Merryweather-Clark and colleagues in human primary erythroid cells (Merryweather-Clarke et al. 2011). Since no similar dataset currently exists for mouse erythropoiesis, we were obliged to extrapolate from the human to

.....

the mouse in choosing genes that we expected to be stably expressed during erythroid maturation.

Using wild type samples, the expression of control genes was evaluated, and is described in figure 6.8 relative to *β -globin*. Those samples with R2 values less than 0.9 were discarded, and the remainder used to control for the amount of RNA loaded for nanostring analysis.

6.2.6 Nanostring: technical and biological replicates, and normalization

Up to 100ng of nascent RNA for each sample was used for nanostring analysis. The raw data are subject to minimal manipulation following acquisition. First, a negative threshold value is obtained, using the negative controls supplied in each nanostring codeset. These give an estimate of the expected false positive threshold, and all readings within 2SD of this value are discarded. Gene expression with transcript counts consistently at or below this level are not considered further.

The geometric mean of all control transcripts for a given sample is calculated. This is calculated for all 12 samples in a given nanostring cartridge, and the mean of all twelve values is calculated. The difference between the lane specific mean and the cross-lane average is used to calculate a lane-specific normalization factor, which normalizes for the amount of RNA loaded in each sample. This permits comparison of values across the different samples tested.

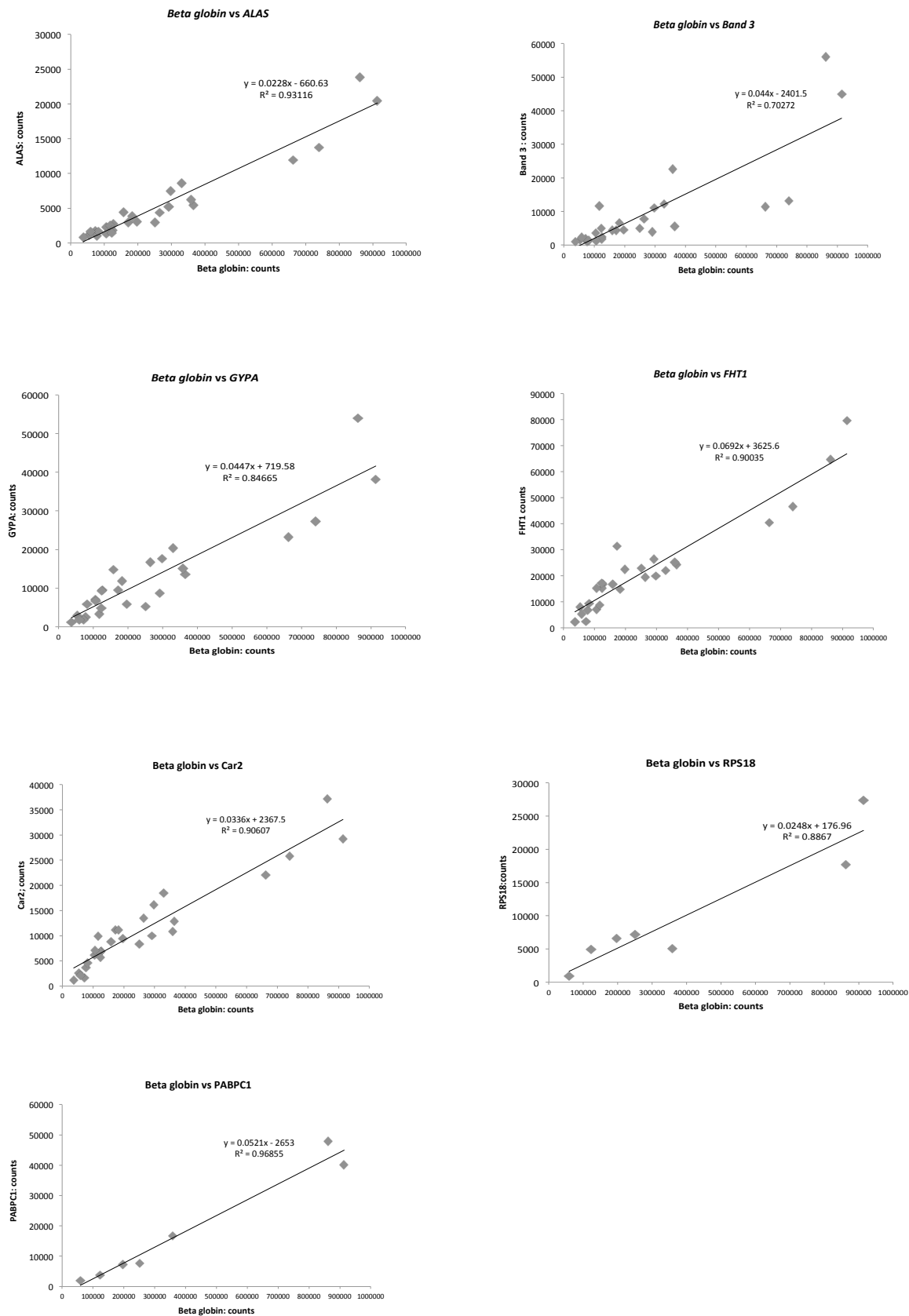


Figure 6.8 Assessing the suitability of control genes chosen for nanostring expression analysis.

.....

Data for all mouse models generated as part of this project, and for previously generated models, are shown in figure 6.9.

Since the methodology has not previously been used for the interrogation of our mouse models, validation with RNA FISH was undertaken with selected models (FISH analysis performed by Drs Veronica Buckle and Dr Jill Brown, MRC Molecular Haematology Unit). The results support those from nanostring analysis of nascent RNA, and are shown in figure 6.10.

Both sets of data suggest that the double deletion of R1 and R2 reduces *α-globin* expression to <10% of normal values. Both R2 and R1 deletions in isolation have a reproducible and significant impact on globin expression following 12h of culture in differentiation medium, with reductions of 50% and 25% respectively. R4 has a minor impact that would be best assessed following additional cultures and further nanostring assays. The deletion of the remaining elements appear not to impact significantly on *globin* expression; although there is no statistically significant difference between WT and $\Delta R(m)$ or $\Delta R3$ models, the mean *α-globin* transcript level in the latter two is surprisingly higher than in WT. We might attempt to explain these findings by reference to the observation that the regulatory elements can act as erythroid-specific alternative promoters for the generation of short (though fully spliced and polyadenylated) transcripts of the *nprl3* gene (Kowalczyk, Hughes,

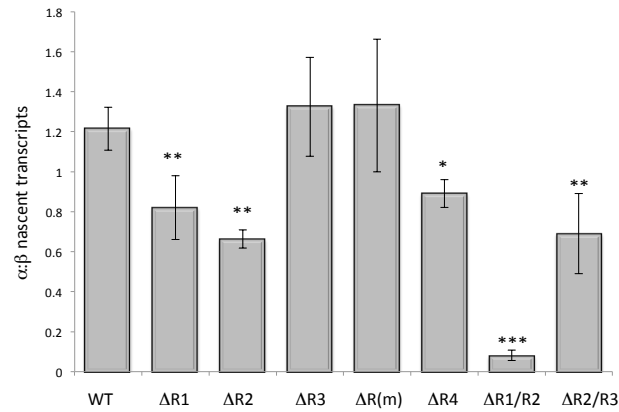


Figure 6.9 Nanostring quantitation of nascent RNA from fetal liver cultures for each mouse model, expressed as a ratio of $\alpha:\beta$ globin. Samples are drawn at 12h following transfer of cultures to differentiation medium.

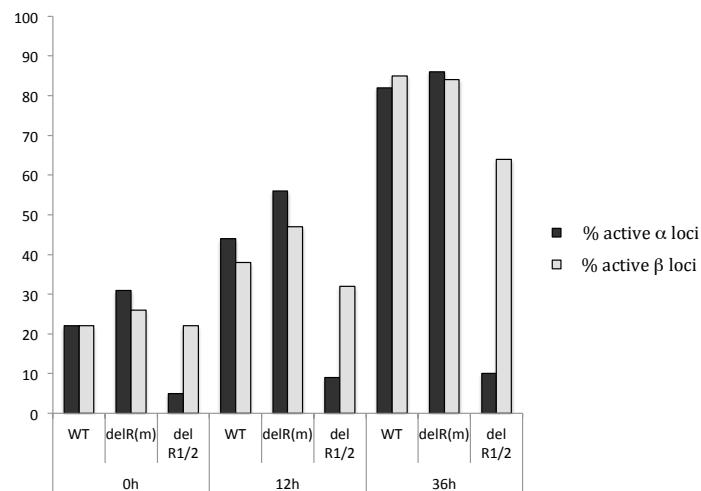


Figure 6.10 RNA FISH validation of nascent transcript quantitation by nanostring

RNA FISH was performed by Drs V Buckle and J Brown (MRC Molecular Haematology Unit) on fetal liver cultures from WT, $\Delta R1/\Delta R2$ and $\Delta R(m)$ homozygote embryos, at 0h, 12 and 36h following the switch to differentiation medium. Supporting nanostring data, a reduction of expressing α -globin loci to approximately 10% of WT was seen in double R1/R2 knockouts, while R(m) deletion had no significant effect. Beta globin loci are used as a control.

.....

Garrick, et al. 2012b). It is conceivable that the 'strong' α -globin enhancers, R1 and R2, may be able to drive expression from the remaining elements acting as alternative promoters. With the deletion of these alternative promoters, there would be greater enhancer activity to focus upon the α -globin promoters proper.

The first nanostring analysis was undertaken on samples labelled at 12h following the switch to differentiation medium, in cells with the morphological features of intermediate normoblasts. A second round of assays was performed on more mature cells, at 30 hours into the differentiation process. The results are shown in figure 6.11. Surprisingly, a much less clear pattern of transcriptional change is observed, with only the double knockout $\Delta R1/\Delta R2$ having a significant impact on α -globin expression (note that data are not given for the double deletion $\Delta R2/\Delta R3$, for which hybridization to the nanostring cartridge failed on this occasion).

It is interesting to speculate on why such a change has occurred in the impact of regulatory element deletion during erythroid maturation. Figure 6.12 shows that the percentage of active α - and β -globin loci climbs during the process of erythroid maturation. A 'hit and run' model of enhancer activation, in which regulatory elements influence the likelihood of transcriptional activation, but are not required to maintain ongoing transcription once a locus is activated, would fit with these data (Sen & Grosschedl 2010, Chong et al. 2010). Empirically, this could be tested by quantifying nascent transcription in a model with a conditional deletion of R1 and R2 following complete erythroid maturation.

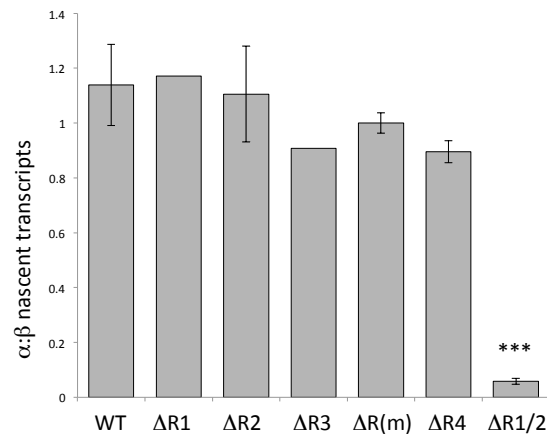


Figure 6.11 Nascent RNA quantitation late in erythroid maturation. Expressed as $\alpha:\beta$ *globin* transcript ratios. Samples drawn at 36 hours following the switch to differentiation medium.

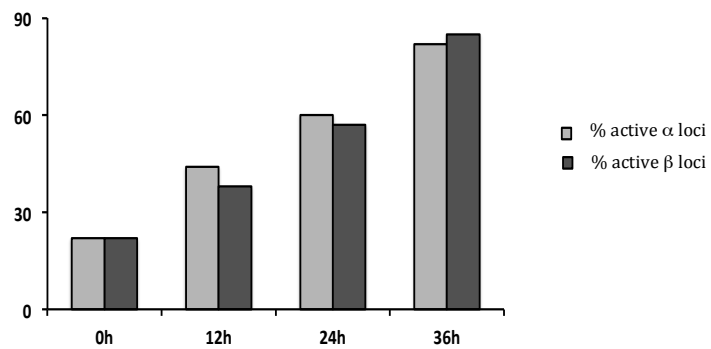


Figure 6.12 RNA FISH analysis of the changing likelihood of active *globin loci* during erythroid maturation. From WT mice. FISH performed by Drs V Buckle and J Brown.

6.2.7 Impact on other genes within the α -globin cluster and beyond

It has previously been observed that there is an erythroid specific up-regulation of the expression of *Nprl3*, attributed to the presence of the highly active α -globin regulatory elements within and adjacent to this gene in erythroid tissues. This is confirmed in our analysis, which demonstrates a reduction in *nprl3* transcript numbers in the context of $\Delta R1$, $\Delta R2$ and (most markedly) double $\Delta R1/\Delta R2$ knockouts (figure 6.13).

The other genes within the globin cluster are variably expressed during development. The embryonic ξ -globin gene is expressed to only a trivial degree in the fetal liver cultures at day 12.5, and therefore it is not possible to assess the impact of enhancer deletions of ξ -globin expression. θ -globin expression is noted, however, albeit at much lower levels than α -globin itself; nevertheless a clear and statistically significant reduction in θ -globin expression is noted in the absence of the R1, R2 or both regulatory elements.

The capacity of Nanostring to permit multiplexed analysis of gene expression has been exploited to assess the impact of enhancer deletion on the transcription of surrounding genes.

Expression of several of the surrounding genes (e.g. *Nsg2*, *Rhbdf1*) was detected at too low a level to permit further assessment by Nanostring. However, where

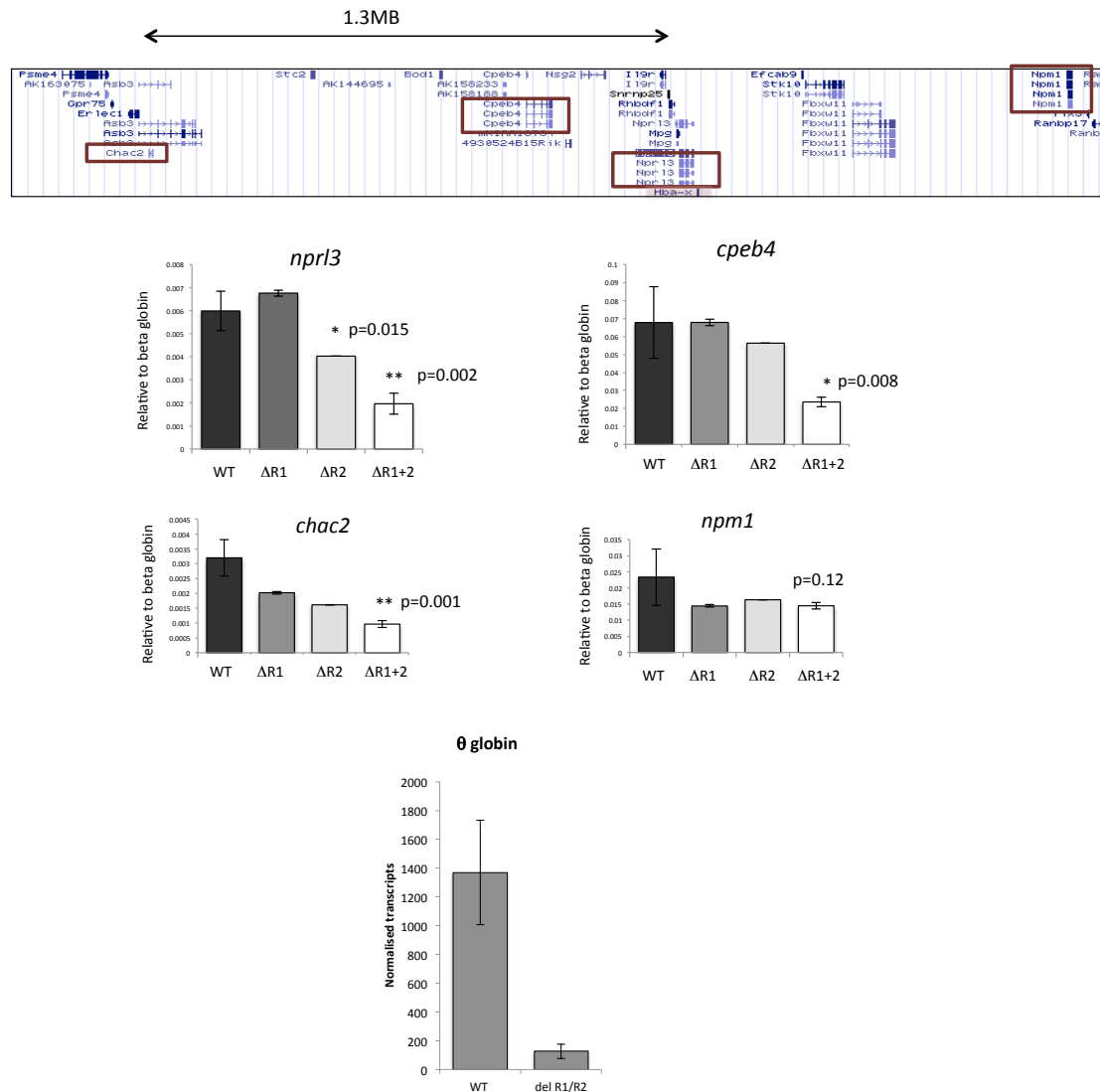


Figure 6.13 Deletion of α -globin regulatory elements impacts distant gene expression

Upper panel: genome browser track highlighting four sample genes distant to the α -globin locus. Genes up to 1.3MB away (*Chac2*) are influenced by the loss of the major regulatory elements R1 and R2. Not all genes are equivalently affected (e.g. *Npm1*, which has no significant change in transcription in response to R1 and R2's loss. Theta globin expression is also impacted by the double knockout of R1 and R2.

.....

transcripts are detected at sufficiently high levels, several genes appear to have their transcription profiles influenced by the loss of the most active regulatory elements. *Cpeb4*, and *Chac2* are two particularly clear examples of diminished expression in the absence of the major regulatory elements at the α -globin locus. Figure 6.13 demonstrates that the apparent impact of globin element depletion extends over a region of >1.3MB.

Clarification of the impact of regulatory element deletion on wider transcription is currently under further investigation by independent methods (nascent RNA-sequencing).

6.3 Other roles for conserved elements at the α -globin locus?

The data presented in this chapter support the contention that there are major and minor regulatory elements for α -globin transcription, R2 and R1 respectively. This is consistent with the data from naturally occurring mutations in patients with thalassaemia.

We have been unable to demonstrate, using our nascent transcription quantitation, a role for R3, or R(m); only a limited role is supported for R4. As previously discussed, these elements are still highly conserved, and contain GATA and SCL transcription factors, implying functional significance. The following hypotheses need to be considered:

-
- These elements may be implicated in *globin* transcription in the context of environmental or genetics stressors of which we are currently unaware.
 - They may impact on α -*globin* transcription at other developmental stages – specifically on ζ -*globin* expression. The models are currently being used to generate primitive erythroid cultures from day 8 embryos (Dr C Babbs, MRC Molecular Haematology Unit) to address this question directly. However, the normal survival of all models apart from $\Delta R1/\Delta R2$ suggests that there will not be a major contribution of the residual elements to α -*globin* transcription in early erythropoiesis.
 - They may be different contributions of the *globin* regulatory elements in different tissues. At the time the project was started, there was no evidence for α -*globin* expression in tissues other than erythroid cells; reports suggesting its expression in other tissues (e.g. Nishi et al. 2008, Biagioli et al. 2009, Marcos Almaraz et al. 2012) had been attributed to blood contamination in the case of primary cells, and to abnormal expression in the case of cell lines. Recent studies have suggested a rather more complex pattern of expression of α -*globin* in tissues including endothelial cells – where a functional role for globin has also been suggested (Straub et al. 2013). It is theoretically possible that different regulatory elements may be responsible for α -*globin* transcription in different tissues.

.....

The report from Straub et al, describing a role for α -globin in controlling the flux of nitric oxide across the myo-endothelial junction, and thus being critical to the control of vascular reactivity, is of particular interest. Could the residual regulatory elements be implicated in the control of *α -globin* transcription in this previously unrecognized non-erythroid setting?

There are no publically available datasets demonstrating DNase sensitivity in murine vascular endothelium for us to interrogate. In human umbilical vein endothelial cells, DNase sensitivity is demonstrable only for element R2 (figure 6.14). Since we have demonstrated that murine and human *α -globin* regulation appear to follow very similar patterns, we have chosen to compare vascular reactivity in response to nitric oxide signalling in the vascular endothelium of both wild type mice and those homozygous for a knockout of murine R2. Preliminary results, produced in collaboration with Dr Gillian Douglas and Mr Surawee Chuaiphichai of the Wellcome Trust Centre for Human Genetics, are shown in figure 6.14. ACh-induced vasodilatation (panel C), which is mediated through nitric oxide, is essentially obliterated in the Δ R2 mouse. SLIGRL is a PAR-2 agonist which vasodilates in a manner partially dependent on nitric oxide, and exposure to this agent gives a blunted response in the Δ R2 mouse relative to WT. Meanwhile nitric oxide independent mechanisms of vasodilation (e.g. sodium nitroprusside mediated, panel E) are unaffected. While superficially these suggest a marked vascular phenotype for the Δ R2 mouse, with significant dysregulation of the nitric oxide control of vascular tone, there is the possibility that simple vascular endothelial disruption in the Δ R2

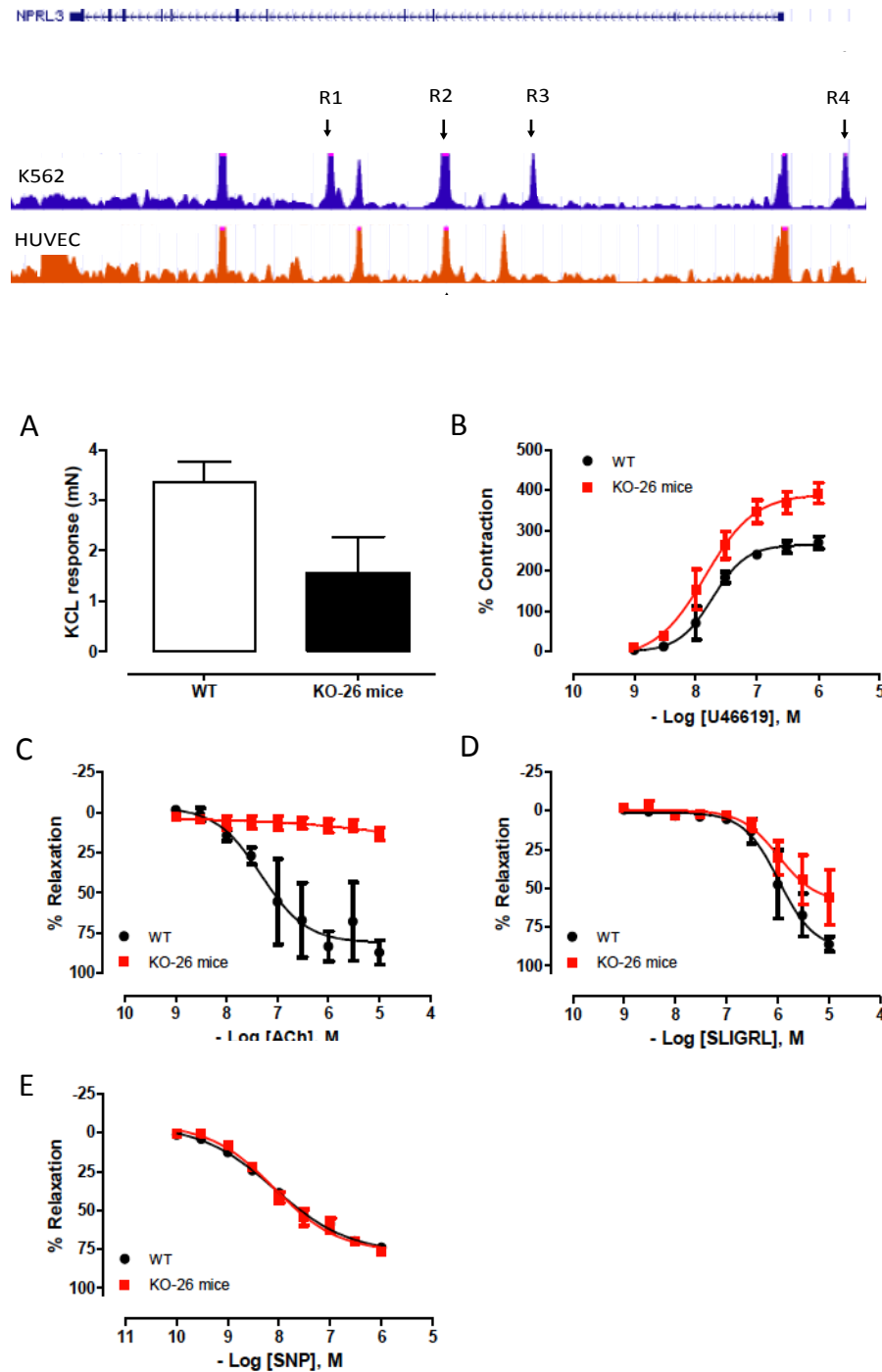


Figure 6.14 | α -globin in vascular endothelium.

Upper panel: DNase I tracks from K562 cells (blue) and HUVEC (orange) over the α -globin regulatory region. The position of the human regulatory elements for α -globin are indicated by arrows. Only R2 is marked as DNase I sensitive in HUVEC.

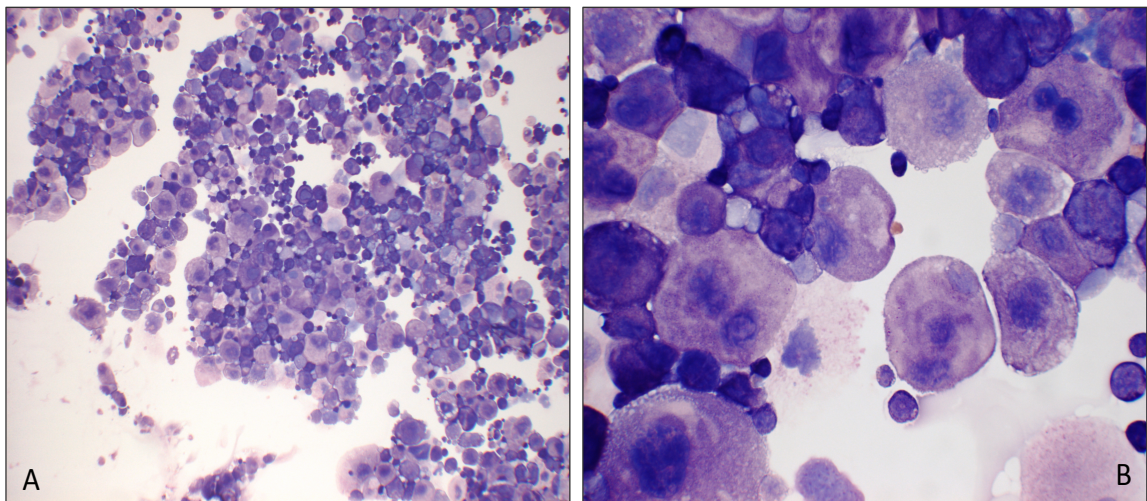
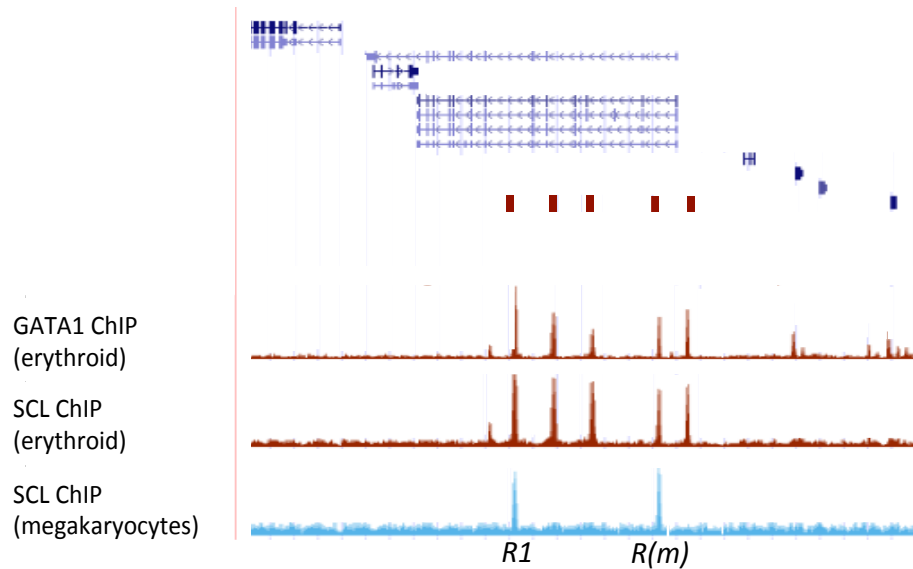
A-E: Vascular reactivity studies (performed by Dr G Douglas and Mr S Chuaiphichai, Wellcome Trust Centre for Human Genetics, Oxford). Vasoactive agents: U46619 - thromboxane A₂ (TP) receptor agonist; ACh - acetylcholinesterase; SLIGRL - PAR-2 agonist; SNP - sodium nitroprusside (see text).

.....

mouse vessels may be responsible for the results seen. At the time of writing, work is ongoing to clarify the significance or otherwise of these findings.

Vascular endothelium is not the only tissue in which *α -globin* appears to be expressed. Publically available RNA sequencing datasets from the Weiss group, Pennsylvania State University (paper in press) demonstrate that *α -globin*, but not *β -globin* is also expressed in megakaryocytes. There is no known role for α -globin in this context. Although DNase I tracks for megakaryocytes are not publically available, there are ChIP-seq datasets describing the binding of key transcription factors in this cell type. These show binding of SCL to R1 and R(m) in megakaryocytes – an apparently specific finding since SCL binding sites are also located within the remaining regulatory elements where no ChIP seq peak is demonstrable (figure 6.15).

We have attempted to use the fetal liver culture system to generate pure populations of megakaryocytes to clarify the role of these elements in particular in *α -globin* expression in this cell type. Fetal liver cultures contain very small numbers of megakaryocytes as isolated by FACS, using CD61+/Ter119-/c-kit- gate to collect megakaryocytes while excluding erythroid cells and blasts. Using this strategy alone, too few cells were collected to enable a representative readout by nanostring. In addition, even the most stringent of sorting strategies will inevitably yield some minor contamination by red cells – problematic in view of the extremely high *α -globin* transcription in red cells. Using the method employed by the Weiss lab (a variation on the fetal liver culture system described for erythroid cells, in which



Megakaryocyte enrichment: Phase II culture, post lin-depletion. A. x20, B. x50

Figure 6.15 α -globin in megakaryocytes?

Upper panel shows GATA1 and SCL ChIP tracks from murine erythroid cells, and SCL ChIP from murine megakaryocytes (publically available dataset, Weiss laboratory) over the α -globin locus.

Lower panel shows enrichment of fetal liver cultures for megakaryocytes following expansion with SCF and TPO, and differentiation in high dose TPO. Residual erythroid cells are still visible morphologically.

.....

thrombopoietin replaces erythropoietin) an expanded population of megakaryocytes is obtained – but in our hands this process still yields a significant degree of erythroid contamination, visible morphologically in figure 6.15. At the time of writing, work is ongoing to try to obtain purer megakaryocyte populations for experimental assessment of the impact of regulatory element deletion on *α-globin* expression in this cell type.

6.4 Summary

In this chapter, I have described and validated the use of nascent transcript quantitation by nanostring to investigate the impact of regulatory element deletions on *α-globin* transcription. Distinct from the canalized haematological phenotypes observed on peripheral blood counts, and with improved sensitivity compared with qPCR or RNase protection assay measurement of total mRNA, this method has allowed us to confidently define a role for R1 and R2 in *α-globin* transcription, with a possible minor role for R4.

Results from our analysis of the impact of regulatory element deletion at different stages of erythroid maturation would be consistent with a ‘hit and run’ model of transcriptional activation, since the impact of regulatory element deletion is less marked in more mature erythroid cells than in earlier committed erythroblasts. This has previously been observed in the report of the phenotype of $\Delta R2$ mice (Anguita et al. 2002): *α:β-globin* mRNA ratios were near normal in peripheral blood, but skewed toward *β-globin* in bone marrow, where a higher proportion of immature erythroid

.....

cells are found. Further work in more cleanly defined populations of maturing erythroid cells will be needed to clarify this observation, and work is ongoing in the laboratory to scale down the nascent RNA collection technique to permit analysis of cell populations sorted by Ter119, CD71 and CD44.

The impact of regulatory element deletion on the wider transcriptome is also under further investigation. The question of whether deletion of the major regulatory elements for *α -globin* impacts more widely on the general transcriptional landscape is experimentally tractable using this technique only if there is certainty with regard to the stability of expression of normalizing genes. The follow-up of this intriguing possibility is therefore being undertaken by RNA-seq in WT and knockout models.

Additional critical controls are also needed to clarify the mechanism behind this apparent wider transcriptional dysregulation. Fetal liver erythropoietic cells homozygous for deletion of R1 and R2 are taken from mice that would have been expected to die *in utero* within the next 24-48 hours, and may therefore manifest widespread transcriptional changes that are causally distinct from any regulatory element deletion. Therefore, to control for this possibility, equivalent experiments are currently being undertaken using a mouse model homozygous for a deletion of the *α -globin* genes themselves. This mouse would be haematologically comparable to the Δ R1/ Δ R2 knockout, and would be expected to die *in utero* at a similar gestational stage, but would have no alteration to the regulatory machinery implicated in our study.

.....

Finally, the hypothesis that the apparently non-functional or redundant *α -globin* regulatory elements might play a role in primitive rather than definitive erythropoiesis, or that they may act in non-erythroid tissues has not been disproven, but remains under investigation.

Chapter 7

Final Conclusions

Chapter 7: Final Conclusions

7.1 Summary of Findings

The work described in this thesis builds on an extensive background of study at the human and mouse *α-globin* locus. The study of a new upstream deletion of regulatory elements at the human locus provided the impetus to review our previous assumptions regarding *α-globin* transcriptional regulation, and led to the completion of a series of knockout models that provide the experimental tools to address several questions in enhancer biology.

Our experimental conclusions include the following:

- HS-40 (R2) is not the sole regulatory element at the human *α-globin* locus.
- Contrary to previous reports, the mouse and human *α-globin* loci appear equivalently regulated.

-
- Heterospecific models, including the MEL hybrid and humanized mouse, do not accurately reflect either mouse or human transcriptional regulation at this locus.
 - Assessment of the impact of enhancer knockouts by blood count, or by qPCR/RNase protection assay quantitation of total accumulated *globin* mRNA are unable to demonstrate a difference between WT and mutant.
 - R1 and R2 contribute most significantly to *α-globin* nascent transcription, with a possible minor contribution from R4.
 - Deletion of R3 and R(m) has little or no impact on *α-globin* expression in erythroid cells
 - The impact of enhancer deletion appears more pronounced in early erythroid maturation.
 - Loss of R1 and R2 impacts on the transcription of genes outside the *α-globin* cluster.

Findings that remain under investigation include the potential regulation of *α-globin* in primitive erythropoiesis, as well as in vascular endothelium and in megakaryocytes. The knockout models generated as part of this project will be important in addressing the role of regulatory elements in these settings, giving us the potential to produce an evaluation of enhancer function that is more detailed than that undertaken at any other mammalian gene locus. More broadly, the regulation of globin expression in non-erythroid tissues may provide clues to its physiological functions in these settings – which are to date defined only in vascular endothelial cells.

7.2 Wider implications

The implications of the work described in this thesis extend beyond the *α -globin* locus. We have demonstrated, using a sensitive method for transcriptional quantitation, that apparently similar regulatory elements have widely differing abilities to enhance transcription. All of the four conserved elements at the *α -globin* locus, and the murine specific element R(m), would be labelled as enhancers using the histone marks identified by Heintzman and colleagues; R1-4 all produce enhancer-originating transcripts; all bind the general transcription factor p300; and all harbour active PolII. Using genome-wide techniques for enhancer identification, therefore, these elements are indistinguishable. Yet at least two of the five appear to be wholly dispensable for *α -globin* transcription. The question follows: how many of the hundreds of thousands of putative regulatory elements genome-wide are actually required for gene transcription? Can we find a means for distinguishing between functional and apparently redundant regulatory elements, for example using a histone code shorthand? While none of the histone marks we have interrogated to date permit us to make this distinction, we are currently investigating the possibility that H4K16 acetylation might distinguish between these elements in accordance with their functional capacity (data publically available from the laboratory of H.Lodish, MIT demonstrate H4K16ac at HS-26 only in non-*globin* expressing erythroid cells, with its subsequent acquisition at all the elements). Clarification of this at the *α -globin* locus will be needed before we can consider extending the observation to other well-studied loci (e.g. *β -globin*) or testing genome-wide.

.....

The data we have accrued may also be helpful in determining the mechanics of transcriptional activation by enhancer elements. ChIP-seq peaks for the erythroid specific transcription factor KLF are consistently higher at R2 than at the other regulatory elements, which may suggest a critical role for this element in the initiation of transcriptional activation by enhancers. Targeted inactivation of this site may enable us to start dissecting the series of steps required for activation of transcription by enhancers. As mentioned in chapter 6, our models also provide an excellent system for the assessment of continuous *versus* transient enhancer requirements for ongoing transcription from this locus.

The question of why a locus requires a series of enhancers for the activation of gene expression is likely to be answered by arguments of evolutionary fitness. Akin to the phenomenon of ‘shadow enhancers’ in *Drosophila*, it is possible that a bank of enhancers may contribute to transcriptional precision, or phenotypic robustness; even regulatory elements giving only very small transcriptional contributions may be subject to selection at the population level. Detailed kinetic studies of globin expression during the process of erythropoiesis in the knockouts we have generated may provide experimental evidence to support this hypothesis.

We would argue that it is only through the careful analysis of test loci that the major outstanding issues in transcriptional regulation can truly be addressed. We hope that this study and the further questions that it generates will contribute to this endeavour.

.....

Appendices and References

Appendix 1: Primers

Role	Name	Sequence
R1 homology amplification	MCSR1 5HA fwd	AAGGAGCAGAACAAAAGGGACC
R1 homology amplification	MCSR1 5HA rev	TGTACCTAGCACATGGTTGTACC
R1 homology amplification	MCSR1 3HA fwd	AGACAGGGTTTCTCTGTGTAGCCC
R1 homology amplification	MCSR1 3HA rev	CGTAATAGTCTGTGCTCTGGGAGAG
R1 sequencing	HS-31 seq section 2f	TTCCGCCACAAGAATGAGGG
R1 sequencing	HS-31 seq section 2r	TATCTGGTCACACTTAGAGCCACG
R1 sequencing	HS-31 seq section 3f	TGTGCTGAGCTGGAAGATCCAC
R1 sequencing	HS-31 seq section 3r	TGGTGGTTCATGCCTTTAATCC
R1 sequencing	HS-31 seq section 4f	CCTGTACCAAGTAAGCATGGAACAC
R1 sequencing	HS-31 seq section 4r	CACCCTTAATCCCAGCTCTTAGG
R1 sequencing	HS-31 seq section 5f	GTGACAGGGTTCTCTGTATAGCCG
R1 sequencing	HS-31 seq section 5r	GCTCTGTTCTCAGCACACAAAGAAG
R1 sequencing	HS-31 seq section 6f	ACTGTAGGAGAACTGGGCAATGAG
R1 sequencing	HS-31 seq section 6r	CTGACGGGCTGAATGGATTATATG
R1 sequencing	HS-31 seq section 7f	TGTTGGCTAGGGGAAAACGC
R1 sequencing	HS-31 seq section 7r	GGGCAGTATTAGGTAGCAAACCTGG
R1 sequencing	HS-31 seq section 8f	CATCTCCATCACTCAAGTCCAGC
R1 sequencing	HS-31 seq section 8r	TCTCGCCACATCCCCTTTAAC
R1 sequencing	HS-31 seq section 9f	CAGGGCAGACTAGGAGTTAAAGGG
R1 sequencing	HS-31 seq section 9r	GGTTGGTTGTGAGCTTTGCATTC
R1 sequencing	HS-31 seq section 10f	CAGGGTTTAATGTCAAGATGTGGG
R1 sequencing	HS-31 seq section 10r	CATTTGTCACGTCTGCACGAC
R1 sequencing	HS-31 seq section 11f	TCGTGCAGGACGTGACAAATG
R1 sequencing	HS-31 seq section 11r	GCCATGATGGATACTTTCTCGG
R1 sequencing	HS-31 seq section 12f	GCAGGACGTGACAAATGGAAGTAG
R1 sequencing	HS-31 seq section 12r	CGCCAAGCTCTTCAGCAATATCAC
R1 sequencing	HS-31 seq section 13f	CCTGCTTGCCGAATATCATGG
R1 sequencing	HS-31 seq section 13r	ATTAACAGAGCCTGGGGTTGGG
R1 sequencing	HS-31 seq section 14f	AGACATCTTCAGGCTTGTCTTGC
R1 sequencing	HS-31 seq section 14r	CAGTTTGTACCTCTCTGCTGTATC
R1 sequencing	HS-31 seq section 15f	TGTGTGATGAGGCTGGAGTGAATC
R1 sequencing	mcs 5'ha seq1	AGCAGAACAAAAGGGACC
R1 sequencing	mcs 5'ha seq2	AAACGAAAACAAGGGGGC
R1 sequencing	mcs 5'ha seq3	TTACCACTCAAGTCTCTGC
R1 sequencing	mcs 5'ha seq4	AGGTTCTTCACAGCAGACG
R1 sequencing	mcs 5'ha seq5	TGGCAGTGAGTGTCTTTACC
R1 sequencing	mcs 5'ha seq6	CTGTCTCTGCCACCTCATAC
R1 sequencing	mcs 5'ha seq 7	GGGGGAAGTGTCTTTCTAAAC
R1 sequencing	mcs 5'ha seq 8	GAGAACTGGGCAATGAGTAATC
R1 sequencing	mcs 5'ha seq 9	GAAGGCAATCTGCTAAAGC
R1 sequencing	mcs 5'ha seq10	ATGGCTTCAGGCTCCGTTC
R1 sequencing	mcs 5'ha seq 11	ACTATGGTGGGCATTTGCG
R1 sequencing	mcs 5'ha seq12	TAATCCATTACGCCGTC
R1 sequencing	mcs 5'ha seq 13	GGGCAGAATAGGGCTTATC
R1 sequencing	mcs 5'ha seq 14	TCTACCTGGGCAGTATCATC
R1 sequencing	mcs 5'ha seq 15	CACCTTGTGACTTTCTCGTAAC
R1 sequencing	mcs 5'ha seq 16	TCTTCCAGAACTCATCCCG

R1 sequencing	mcs 5'ha seq 17	CCAGTGC GGGAAGATAAC
R1 sequencing	mcs 5'ha seq 18	TAAAGGGGATGTGGCGAGAG
R1 sequencing	mcs 5'ha seq 19	AAGGGTTACTGTGCTGCG
R1 sequencing	mcs 5'ha seq 20	ACAGGGTGGCTATTGAAAC
R1 sequencing	mcs 3'ha seq 1	CCTGGAACCTACTCTGTAGACC
R1 sequencing	mcs 3'ha seq 2	GGAACATTCAGAGACATCTTCAGG
R1 sequencing	mcs 3'ha seq 3	TCATTCTCTGCCATTC
R1 sequencing	mcs 3'ha seq 4	GGGTAACCTCCTAAGGGATTC
R1 sequencing	mcs 3'ha seq 5	CCACAACCTCTGTCTTTTAGG
R1 sequencing	mcs 3'ha seq 6	GATGAGGCTGGAGTGAATC
R1 sequencing	mcs 3'ha seq 7	AGAAGTGTGGCTCTCAACC
R1 sequencing	mcs 3'ha seq 8	CCACTAACAAACAGGGTATCTG
R1 sequencing	mcs 3'ha seq 9	AGTCAAGCCCTGTCTACATAAG
R1 sequencing	mcs 3'ha seq 10	CACACACCTGTATCAAGTCACC
R1 sequencing	mcs 3'ha seq 11	TTCCCAATACTTGCTGACC
ΔR2 genotyping	HS-26 post cre fwd	TGCAGGACAGTAGCAATTCGG
ΔR2 genotyping	HS-26 post cre rev	GTATGAGCCAAGATGCAAACCC
R1 genotyping	chk pos -31 clone fwd1	TTGCTCCTGCCGAGAAAGTATC
R1 genotyping	chk pos -31 clone rev1	ATTGAGGTGGACATCACTGTTGC
R1 genotyping	chk pos -31 clone fwd 2	GCCGCTTTTCTGGATTCATCG
R1 genotyping	chk pos -31 clone rev 2	CAGTGCCAAGTTCTTTACCACCTTC
R2 genotyping	26 only 3' F	TCTTGGTCGCTCTCCTTACAGC
R2 genotyping	26 only 3' R	GTTATCTTTCCCGCACTGGATG
R2 genotyping	26 only 5' F	ACATCTTTGACCCCAGTCCCTC
R2 genotyping	26 only 5' R	AACTGCATGTTGACCATAGCCAG
ΔR1 sequencing	31 3' breakpoint F	CCTTGGGACCTCAAACAAAACCTC
ΔR1 sequencing	31 3' breakpoint R	AGGCGGATTCTGTAGTTTCGAGG
ΔR2/ ΔR3 genotype	21 26 SMALL CRE F2	AGGGTGGCATTCTATCAGTGAC
ΔR2/ ΔR3 genotype	21 26 SMALL CRE R2	CGACAGACAGGGAGTTTTTTAGTGG
ΔR2/ ΔR3 screening	21_26 screen fwd	GCCTTCTATCGCCTTCTTGACG
ΔR2/ ΔR3 screening	21_26 screen rev	GAGCCACAGTAGGTTTTCAGGGTAG
ΔR1/ΔR2 screening	maxdel 31 26 seq F	GAAAGGGTTACTGTGCTGCG
ΔR1/ΔR2 screening	maxdel 31 26 seq R	GGATACATGCTTTTGGGAGGC
ΔR2/ ΔR3 screening	maxdel 21 26 seq F	TGCTGGCTGACTTCCTCATCTTC
ΔR2/ ΔR3 screening	maxdel 21 26 seq R	TGCTGGCTGACTTCCTCATCTTC

Appendix 2: Nanostring probes

PN(CP;RP) shows the position of codeset probes and reporter probes respectively

Tm: melting temperature of codeset probe and reporter probe.

NSID: accession number

Gene	PN(CP;RP)	Tm_CP	Tm_RP	NSID
Actn4	371715;271715	85	84	NM_021895.2:269
Acyp2	371716;271716	85	77	NM_029344.3:220
Alas2	322864;222864	80	82	NM_009653.3:615
Asb3	371750;271750	80	81	NM_023906.3:835
Bod1	371723;271723	85	85	NM_001024919.1:369
Calm2	371724;271724	86	85	NM_007589.5:28
Car2	371725;271725	85	78	NM_009801.4:437
Chac2	371753;271753	81	83	NM_026527.2:294
Cpeb4	371752;271752	82	80	NM_026252.3:560
Efcab9	371754;271754	82	80	NM_027031.3:355
Erlec1	371751;271751	81	82	NM_025745.3:940
Fbxw11	371727;271727	84	80	NM_134015.2:276
Fgf18	371745;271745	83	82	NM_008005.1:505
Fth1	320313;220313	85	85	NM_010239.1:446
Gabrp	371728;271728	84	82	NM_146017.3:484
Gpr75	371755;271755	82	82	NM_175490.4:600
Gypa	371729;271729	81	79	NM_010369.3:494
Hba-a1	371714;271714	86	84	NM_008218.2:26
Hba-x	371747;271747	85	85	NM_010405.4:246
Hbb-b1	313467;213467	86	85	NM_008220.4:196
Hbb-bh1	371730;271730	84	85	NM_008219.3:31
Hbb-y	371722;271722	85	84	NM_008221.4:34
Hbq1b	371721;271721	84	79	NM_001033981.3:441
Il9r	371731;271731	85	81	NM_001134458.1:320
Kcnp1	371743;271743	86	85	NM_027398.3:459
Laptm4a	371733;271733	84	81	NM_008640.2:533
Mpg	371719;271719	86	84	NM_010822.3:276
Npm1	371744;271744	82	85	NM_008722.2:6
Nprl3	371736;271736	79	81	NM_181569.2:274
Nsg2	371737;271737	86	84	NM_008741.3:187
Pabpc1	386474;286474	80	82	NM_008774.3:2212
Paip2	346812;246812	80	79	NM_026420.2:130
Psme4	371738;271738	80	84	NM_134013.3:1286
Ranbp17	371739;271739	84	82	NM_023146.2:397
Rhbdf1	371740;271740	85	84	NM_010117.2:680
Rps18	342174;242174	92	77	NM_011296.2:420
Sh3pxd2b	371717;271717	84	85	NM_177364.3:536
Shmt2	371742;271742	84	82	NM_001252316.1:54
Slc4a1	371718;271718	85	85	NM_011403.2:565
Snrrnp25	371720;271720	85	85	NM_030093.3:133

.....

Stc2	329493;229493	82	80	NM_011491.3:360
Stk10	371746;271746	81	82	NM_009288.2:115
Tlx3	371749;271749	86	84	NM_019916.2:277
Ubtd2	371741;271741	83	86	NM_173784.3:186

References

- Adams, J.M. et al., 1985. The c-myc oncogene driven by immunoglobulin enhancers induces lymphoid malignancy in transgenic mice. *Nature*, 318(6046), pp.533–538.
- Alexander, R.P. et al., 2010. Annotating non-coding regions of the genome. *Nature Reviews Genetics*, 11(8), pp.559–571.
- Andrey, G. et al., 2013. A Switch Between Topological Domains Underlies HoxD Genes Collinearity in Mouse Limbs. *Science*, 340(6137), pp.1234167–1234167.
- Anguita, E., 2002. Deletion of the mouse alpha -globin regulatory element (HS -26) has an unexpectedly mild phenotype. *Blood*, 100(10), pp.3450–3456.
- Ariza-Cosano, A. et al., 2012. Differences in enhancer activity in mouse and zebrafish reporter assays are often associated with changes in gene expression. *BMC Genomics*, 13(1), pp.1–1.
- Arnold, C.D. et al., 2013. Genome-Wide Quantitative Enhancer Activity Maps Identified by STARR-seq. *Science*, 339(6123), pp.1074–1077.
- Attanasio, C. et al., 2013. Fine Tuning of Craniofacial Morphology by Distant-Acting Enhancers. *Science*, 342(6157), pp.1241006–1241006.
- Banerji, J.J., Olson, L.L. & Schaffner, W.W., 1983. A lymphocyte-specific cellular enhancer is located downstream of the joining region in immunoglobulin heavy chain genes. *Cell*, 33(3), pp.729–740.
- Banerji, J.J., Rusconi, S.S. & Schaffner, W.W., 1981. Expression of a beta-globin gene is enhanced by remote SV40 DNA sequences. *Cell*, 27(2 Pt 1), pp.299–308.
- Bebin, A.G. et al., 2010. In Vivo Redundant Function of the 3' IgH Regulatory Element HS3b in the Mouse. *The Journal of Immunology*, 184(7), pp.3710–3717.
- Bejerano, G., 2004. Ultraconserved Elements in the Human Genome. *Science*, 304(5675), pp.1321–1325.
- Bender, M.A. et al., 2012. The hypersensitive sites of the murine -globin locus control region act independently to affect nuclear localization and transcriptional elongation. *Blood*, 119(16), pp.3820–3827.

- Benko, S. et al., 2009. Highly conserved non-coding elements on either side of SOX9 associated with Pierre Robin sequence. *Nature Genetics*, 41(3), pp.359–364.
- Biagioli, M. et al., 2009. Unexpected expression of alpha- and beta-globin in mesencephalic dopaminergic neurons and glial cells. *Proceedings of the National Academy of Sciences of the United States of America*, 106(36), pp.15454–15459.
- Brenner, S., 2009. Sequences and consequences. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 365(1537), pp.207–212.
- Broad Institute Sequencing Platform and Whole Genome Assembly Team. 2012. A high-resolution map of human evolutionary constraint using 29 mammals. *Nature*, 478(7370), pp.476–481.
- Buluwela, L. et al., 1989. A rapid procedure for colony screening using nylon filters. *Nucleic Acids Research*, 17(1), pp.452–452.
- Bungert, J. et al., 1995. Synergistic regulation of human beta-globin gene switching by locus control region elements HS3 and HS4. *Genes & development*, 9(24), pp.3083–3096.
- Butler, J.E.F., 2001. Enhancer-promoter specificity mediated by DPE or TATA core promoter motifs. *Genes & development*, 15(19), pp.2515–2519.
- Cai, J. et al., 2003. Increased risk for developmental delay in Saethre-Chotzen syndrome is associated with TWIST deletions: an improved strategy for TWIST mutation screening. *Human Genetics*, 114(1), pp.68–76.
- Calhoun, V.C., Stathopoulos, A. & Levine, M., 2002. Promoter-proximal tethering elements regulate enhancer-promoter specificity in the *Drosophila* Antennapedia complex. *Proceedings of the National Academy of Sciences*, 99(14), pp.9243–9247.
- Carter, D. et al., 2002. Long-range chromatin regulatory interactions in vivo. *Nature Genetics*, 32(4), pp.623–626.
- Chauveau, C., Pinaud, E. & Cogne, M., 1998. Synergies between regulatory elements of the immunoglobulin heavy chain locus and its palindromic 3' locus control region. *European journal of immunology*, 28(10), pp.3048–3056.
- Chen, N., 2004. WormBase: a comprehensive data resource for *Caenorhabditis* biology and genomics. *Nucleic Acids Research*, 33(Database issue), pp.D383–D389.
- Chong et al., 2010. Epigenetic propagation of CD4 expression is established by the Cd4 proximal enhancer in helper T cells. *Genes & development*, 24(7), pp.659–669.

- Coelho, A. et al., 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(2), pp.147–153.
- Coleman, D., Russell, E. & Levin, E., 1969. Enzymatic studies of the hemopoietic defect in flexed mice. *Genetics*, 61(3), pp.631–642.
- Collins, F.S. et al., 2004. Finishing the euchromatic sequence of the human genome. *Nature*, 431(7011), pp.931–945.
- Consortium, T.E.P. et al., 2013. An integrated encyclopedia of DNA elements in the human genome. *Nature*, 488(7414), pp.57–74.
- Core, L.J., Waterfall, J.J. & Lis, J.T., 2008. Nascent RNA Sequencing Reveals Widespread Pausing and Divergent Initiation at Human Promoters. *Science*, 322(5909), pp.1845–1848.
- Costas, J., Casares, F. & Vieira, J., 2003. Turnover of binding sites for transcription factors involved in early *Drosophila* development. *Gene*, 310, pp.215–220.
- Curtin, P. et al., 1985. A distant gene deletion affects beta-globin gene function in an atypical gamma delta beta-thalassemia. *Journal of Clinical Investigation*, 76(4), pp.1554–1558.
- David, D. et al., 2003. Molecular characterization of a familial translocation implicates disruption of HDAC9 and possible position effect on TGF β 2 in the pathogenesis of Peters' anomaly. *Genomics*, 81(5), pp.489–503.
- Davies, A.F. et al., 1999. An interstitial deletion of 6p24-p25 proximal to the FKHL7 locus and including AP-2alpha that affects anterior eye chamber development. *Journal of Medical Genetics*, 36(9), pp.708–710.
- De Gobbi, M., 2006. A Regulatory SNP Causes a Human Genetic Disease by Creating a New Transcriptional Promoter. *Science*, 312(5777), pp.1215–1217.
- De Gobbi, M. et al., 2007. Tissue-specific histone modification and transcription factor binding in globin gene expression. *Blood*, 110(13), pp.4503–4510.
- De Santa, F. et al., 2010. A Large Fraction of Extragenic RNA Pol II Transcription Sites Overlap Enhancers J. S. Mattick, ed. *PLoS Biology*, 8(5), p.e1000384.
- Degenhardt, K.R. et al., 2010. Distinct enhancers at the Pax3 locus can function redundantly to regulate neural tube and neural crest expressions. *Developmental Biology*, 339(2), pp.519–527.
- Deisseroth, A. & Hendrick, D., 1978. Human alpha-globin gene expression following chromosomal dependent gene transfer into mouse erythroleukemia cells. *Cell*, 15(1), pp.55–63.

-
- Dekker, J., 2002. Capturing Chromosome Conformation. *Science*, 295(5558), pp.1306–1311.
- Deng, W. et al., 2012. Controlling Long-Range Genomic Interactions at a Native Locus by Targeted Tethering of a Looping Factor. *Cell*, 149(6), pp.1233–1244.
- Devoy, A. et al., 2011. Genomically humanized mice: technologies and promises. *Nature Reviews Genetics*, 13(1), pp.14–20.
- Drissen, R., 2004. The active spatial organization of the γ -globin locus requires the transcription factor EKLF. *Genes & development*, 18(20), pp.2485–2490.
- Dunipace, L., Ozdemir, A. & Stathopoulos, A., 2011. Complex interactions between cis-regulatory modules in native conformation are critical for *Drosophila* snail expression. *Development*, 138(20), pp.4566–4566.
- Fang, X. et al., 2005. Synergistic and additive properties of the beta-globin locus control region (LCR) revealed by 5' HS3 deletion mutations: implication for LCR chromatin architecture. *Molecular and cellular biology*, 25(16), pp.7033–7041.
- Fantes, J. et al., 1995. Aniridia-associated cytogenetic rearrangements suggest that a position effect may cause the mutant phenotype. *Human Molecular Genetics*, 4(3), pp.415–422.
- Fernandez, B.A. et al., 2005. Holoprosencephaly and cleidocranial dysplasia in a patient due to two position-effect mutations: case report and review of the literature. *Clinical Genetics*, 68(4), pp.349–359.
- Fibach, E. et al., 1989. Proliferation and maturation of human erythroid progenitors in liquid culture. *Blood*, 73(1), pp.100–103.
- Fiering, S. et al., 1995. Targeted deletion of 5'HS2 of the murine beta-globin LCR reveals that it is not essential for proper regulation of the beta-globin locus. *Genes & development*, 9(18), pp.2203–2213.
- Francastel, C., Magis, W. & Groudine, M., 2001. Nuclear relocation of a transactivator subunit precedes target gene activation. *Proceedings of the National Academy of Sciences*, 98(21), pp.12120–12125.
- Frankel, N. et al., 2010. Phenotypic robustness conferred by apparently redundant transcriptional enhancers. *Nature*, 466(7305), pp.490–493.
- Fraser, P. et al., 1993. Each hypersensitive site of the human beta-globin locus control region confers a different developmental pattern of expression on the globin genes. *Genes & development*, 7(1), pp.106–113.

-
- Friedel, C.C. & Dölken, L., 2009. Metabolic tagging and purification of nascent RNA: implications for transcriptomics. *Molecular BioSystems*, 5(11), p.1271.
- Fromm, G. et al., 2011. An embryonic stage-specific enhancer within the murine α -globin locus mediates domain-wide histone hyperacetylation. *Blood*, 117(19), pp.5207–5214.
- Fukami M et al., 2005. Microdeletion in the SHOX 3 Region Associated With Skeletal Phenotypes of Langer Mesomelic Dysplasia in a 45,X/46,X,r(X) Infant and Leri–Weill Dyschondrosteosis in her 46,XX Mother: Implication for the SHOX Enhancer. *American Journal of Medical Genetics* 137A:72–76
- Ganis, J.J. et al., 2012. Zebrafish globin switching occurs in two developmental stages and is controlled by the LCR. *Developmental Biology*, 366(2), pp.185–194.
- Guy, L.G. et al., 1996. The beta-globin locus control region enhances transcription of but does not confer position-independent expression onto the lacZ gene in transgenic mice. *The EMBO journal*, 15(14), pp.3713–3721.
- Hallikas, O. et al., 2006. Genome-wide Prediction of Mammalian Enhancers Based on Analysis of Transcription-Factor Binding Affinity. *Cell*, 124(1), pp.47–59.
- Hatton, C.S. et al., 1990. Alpha-thalassemia caused by a large (62 kb) deletion upstream of the human alpha globin gene cluster. *Blood*, 76(1), pp.221–227.
- Heintzman, N.D. et al., 2007. Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nature Genetics*, 39(3), pp.311–318.
- Heinz, S. et al., 2013. Effect of natural genetic variation on enhancer selection and function. *Nature*, pp.1–20.
- Higgs, D.R. et al., 1990. A major positive regulatory region located far upstream of the human alpha-globin gene locus. *Genes & development*, 4(9), pp.1588–1601.
- Hug, B.A. et al., 1996. Analysis of mice containing a targeted deletion of beta-globin locus control region 5' hypersensitive site 3. *Molecular and cellular biology*, 16(6), pp.2906–2912.
- Hughes, J.R. et al., 2005. Annotation of cis-regulatory elements by identification, subclassification, and functional assessment of multispecies conserved sequences. *Proceedings of the National Academy of Sciences*, 102(28), pp.9830–9835.
- Jackson, D.A., Pombo, A. & Iborra, F., 2000. The balance sheet for transcription: an analysis of nuclear RNA metabolism in mammalian cells. *FASEB journal : official publication of the Federation of American Societies for Experimental Biology*, 14(2), pp.242–254.

-
- Jackson, J.D. et al., 1996. Role of DNA sequences outside the cores of DNase hypersensitive sites (HSs) in functions of the beta-globin locus control region. Domain opening and synergism between HS2 and HS3. *The Journal of biological chemistry*, 271(20), pp.11871–11878.
- Jarman, A.P. et al., 1991. Characterization of the major regulatory element upstream of the human alpha-globin gene cluster. *Molecular and cellular biology*, 11(9), pp.4679–4689.
- Jing, H. et al., 2008. Exchange of GATA Factors Mediates Transitions in Looped Chromatin Organization at a Developmentally Regulated Gene Locus. *Molecular cell*, 29(2), pp.232–242.
- Kellum, R. & Schedl, P., 1991. A position-effect assay for boundaries of higher order chromosomal domains. *Cell*, 64(5), pp.941–950.
- Khodor, Y.L. et al., 2011. Nascent-seq indicates widespread cotranscriptional pre-mRNA splicing in *Drosophila*. *Genes & development*, 25(23), pp.2502–2512.
- Kim, T.-K. et al., 2010. Widespread transcription at neuronal activity-regulated enhancers. *Nature*, 465(7295), pp.182–187.
- King, M.-C. & Wilson, A.C., 1975. Evolution at two levels in humans and chimpanzees. *Science*, 188(4184), pp.107–116.
- Kioussis, D. et al., 2004. beta globin gene inactivation by DNA translocation in gamma beta thalassaemia. *Nature*, 306, pp.662–666.
- Kleinjan, D.J. & Coutinho, P., 2009. Cis-ruption mechanisms: disruption of cis-regulatory control as a cause of human genetic disease. *Briefings in functional genomics & proteomics*, 8(4), pp.317–332.
- Klopocki, E. et al., 2008. A microduplication of the long range SHH limb regulator (ZRS) is associated with triphalangeal thumb-polysyndactyly syndrome. *Journal of Medical Genetics*, 45(6), pp.370–375.
- Kowalczyk, M.S., Hughes, J.R., Babbs, C., et al., 2012a. Nprl3 is required for normal development of the cardiovascular system. *Mammalian Genome*, 23(7-8), pp.404–415.
- Kowalczyk, M.S., Hughes, J.R., Garrick, D., et al., 2012b. Intragenic Enhancers Act as Alternative Promoters. *Molecular cell*, 45(4), pp.447–458.
- Laimins, L., Holmgren-König, M. & Khoury, G., 1986. Transcriptional “silencer” element in rat repetitive sequences associated with the rat insulin 1 gene locus. *Proceedings of the National Academy of Sciences*, 83(10), pp.3151–3155.

- Landry, C.R., 2005. Compensatory cis-trans Evolution and the Dysregulation of Gene Expression in Interspecific Hybrids of *Drosophila*. *Genetics*, 171(4), pp.1813–1822.
- Leipoldt, M. et al., 2006. Two novel translocation breakpoints upstream of SOX9 define borders of the proximal and distal breakpoint cluster region in campomelic dysplasia. *Clinical Genetics*, 71(1), pp.67–75.
- Lettice, L.A. et al., 2002. Disruption of a long-range cis-acting regulator for Shh causes preaxial polydactyly. *Proceedings of the National Academy of Sciences*, 99(11), pp.7548–7553.
- Li, G. et al., 1998. Individual LCR hypersensitive sites cooperate to generate an open chromatin domain spanning the human beta-globin locus. *Genes to cells*, 3(7), pp.415–429.
- Li, Q., 2002. Locus control regions. *Blood*, 100(9), pp.3077–3086.
- Li, Q. & Stamatoyannopoulos, G., 1994. Hypersensitive site 5 of the human beta locus control region functions as a chromatin insulator. *Blood*, 84(5), pp.1399–1401.
- Ligtenberg, M.J.L. et al., 2008. Heritable somatic methylation and inactivation of MSH2 in families with Lynch syndrome due to deletion of the 3' exons of TACSTD1. *Nature Genetics*, 41(1), pp.112–117.
- Lodish, H.F. & Small, B., 1976. Different lifetimes of reticulocyte messenger RNA. *Cell*, 7(1), pp.59–65.
- Loots et al. 2005 Genomic deletion of a long-range bone enhancer misregulates sclerostin in Van Buchem disease. *Genome Res.* 15: 928-935
- Ludwig, M. et al., 2000. Evidence for stabilizing selection in a eukaryotic enhancer element. *Nature*, 403, pp.564–567.
- Ludwig, M.Z. et al., 2005. Functional Evolution of a cis-Regulatory Module. *PLoS Biology*, 3(4), p.e93.
- Mao, X., Fujiwara, Y. & Orkin, S.H., 1999. Improved reporter strain for monitoring Cre recombinase-mediated DNA excisions in mice. *Proceedings of the National Academy of Sciences*, 96(9), pp.5037–5042.
- Marcos Almaraz, M.T. et al., 2012. α -Haemoglobin regulates sympathoadrenal cell metabolism to maintain a catecholaminergic phenotype. *Biochemical Journal*, 441(3), pp.843–850.
- Marinić, M. et al., 2013. An Integrated Holo-Enhancer Unit Defines Tissue and Gene Specificity of the Fgf8 Regulatory Landscape. *Developmental Cell*, 24(5), pp.530–542.

- Maxam, A. et al., 1977. Promotor region for yeast 5S ribosomal RNA. *Nature*, 267 pp.643-645.
- McLean, C.Y. et al., 2012. Human-specific loss of regulatory DNA and the evolution of human-specific traits. *Nature*, 471(7337), pp.216–219.
- Merryweather-Clarke, A.T. et al., 2011. Global gene expression analysis of human erythroid progenitors. *Blood*, 117(13), pp.e96–e108.
- Milton, J.N. et al., 2013. Genetic determinants of haemolysis in sickle cell anaemia. *British journal of haematology*, 161(2), pp.270–278.
- Montes, M. et al., 2012. Functional coupling of transcription and splicing. *Gene*, 501(2), pp.104–117.
- Moschonas, N. et al., 1981. Structure and expression of a cloned β^0 thalassaemic globin gene. *Nucleic Acids Research*, 9(17), pp.4391–4402.
- Muncke, N., 2004. Position effect on PLP1 may cause a subset of Pelizaeus-Merzbacher disease symptoms. *Journal of Medical Genetics*, 41(12), pp.e121–e121.
- Nicholls, R.D., Fischel-Ghodsian, N. & Higgs, D.R., 1987. Recombination at the human alpha-globin gene cluster: sequence features and topological constraints. *Cell*, 49(3), pp.369–378.
- Nishi, H. et al., 2008. Hemoglobin Is Expressed by Mesangial Cells and Reduces Oxidant Stress. *Journal of the American Society of Nephrology*, 19(8), pp.1500–1508.
- nobrega, M.A., Llewellyn, M. & Rainey, P.B., 2004. megabase deletions of gene deserts result in viable mice. *Nature*, 431(7011), pp.984–988.
- Nowak, M.A. et al., 1997. Evolution of genetic redundancy. *Nature*, 388(6638), pp.167–171.
- Ohta and Kimura, 1971. Functional Organization of Genetic Material as a Product of Molecular Evolution K. M, ed., pp.1–2.
- Ohtsuki, S., Levine, M. & Cai, H.N., 1998. Different core promoters possess distinct regulatory activities in the *Drosophila* embryo. *Genes & development*, 12(4), pp.547–556.
- Ong, J. et al., 1998. 3' IgH enhancer elements shift synergistic interactions during B cell development. *The Journal of Immunology*, 160(10), pp.4896–4903.

-
- Parslow, T.G. & Granner, D.K., 1983. Structure of a nuclease-sensitive region inside the immunoglobulin kappa gene: evidence for a role in gene regulation. *Nucleic Acids Research*, 11(14), pp.4775–4792.
- Pekowska, A. et al., 2011. H3K4 tri-methylation provides an epigenetic signature of active enhancers. *The EMBO journal*, 30(20), pp.4198–4210.
- Pertea, M. & Salzberg, S.L., 2010. Between a chicken and a grape: estimating the number of human genes. *Genome Biology*, 11(5), p.206.
- Phylipsen, M. et al., 2010. Thalassemia in Western Australia: 11 novel deletions characterized by Multiplex Ligation-dependent Probe Amplification. *Blood Cells, Molecules, and Diseases*, 44(3), pp.146–151.
- Pinaud, E. et al., 2001. Localization of the 3' IgH locus elements that effect long-distance regulation of class switch recombination. *Immunity*, 15(2), pp.187–199.
- Pittman, A.M. et al., 2010. Allelic Variation at the 8q23.3 Colorectal Cancer Risk Locus Functions as a Cis-Acting Regulator of EIF3H J. M. Ford, ed. *PLoS Genetics*, 6(9), p.e1001126.
- R J Cole, J.G.R.G.T., 1974. Disturbed haem and globin synthesis in reticulocytes of prenatal flexed-tail anaemic mice. *Genetics Research*, 23 pp125-135
- Rabani, M. et al., 2011. Metabolic labeling of RNA uncovers principles of RNA production and degradation dynamics in mammalian cells. *Nature Biotechnology*, 29(5), pp.436–442.
- Rada-Iglesias, A. et al., 2012. A unique chromatin signature uncovers early developmental enhancers in humans. *Nature*, 470(7333), pp.279–283.
- Ragoczy, T., 2006. The locus control region is required for association of the murine beta-globin locus with engaged transcription factories during erythroid maturation. *Genes & development*, 20(11), pp.1447–1457.
- Rahimov, F. et al., 2008. Disruption of an AP-2 α binding site in an IRF6 enhancer is associated with cleft lip. *Nature Genetics*, 40(11), pp.1341–1347.
- Rando, O.J., 2012. Combinatorial complexity in chromatin structure and function: revisiting the histone code. *Current Opinion in Genetics & Development*, 22(2), pp.148–155.
- Rodgers, N.D., Wang, Z. & Kiledjian, M., 2002. Regulated alpha-globin mRNA decay is a cytoplasmic event proceeding through 3'-to-5' exosome-dependent decapping. *Rna*, 8(12), pp.1526–1537.

-
- Rouaud, P. et al., 2012. Enhancers Located in Heavy Chain Regulatory Region (hs3a, hs1,2, hs3b, and hs4) Are Dispensable for Diversity of VDJ Recombination. *Journal of Biological Chemistry*, 287(11), pp.8356–8360.
- Routledge, SJ & Proudfoot, NJ., 2002. Definition of transcriptional promoters at the human beta globin locus control region. *Journal of Molecular Biology*, 323(4), pp.601–611
- Sabherwal et al. 2007. Long-range conserved non-coding SHOX sequences regulate expression in developing chicken limb and are associated with short stature phenotypes in human patients. *Human Molecular Genetics* 16 (2) 210–222
- Savic, D. et al., 2011. Alterations in TCF7L2 expression define its role as a key regulator of glucose metabolism. *Genome Research*, 21(9), pp.1417–1425.
- Sawado, T., 2003. The beta -globin locus control region (LCR) functions primarily by enhancing the transition from transcription initiation to elongation. *Genes & development*, 17(8), pp.1009–1018.
- Sen, R. & Grosschedl, R., 2010. Memories of lost enhancers. *Genes & development*, 24(10), pp.973–979.
- Sironi, M., 2005. Analysis of intronic conserved elements indicates that functional complexity might represent a major source of negative selection on non-coding sequences. *Human Molecular Genetics*, 14(17), pp.2533–2546.
- Solomon, M.J., Larsen, P.L. & Varshavsky, A., 1988. Mapping protein-DNA interactions in vivo with formaldehyde: evidence that histone H4 is retained on a highly transcribed gene. *Cell*, 53(6), pp.937–947.
- Spitz, F. et al., 2002. A t(2;8) Balanced Translocation with Breakpoints Near the Human HOXD Complex Causes Mesomelic Dysplasia and Vertebral Defects. *Genomics*, 79(4), pp.493–498.
- Straub, A.C. et al., 2013. Endothelial cell expression of haemoglobin α regulates nitric oxide signalling. *Nature*, 491(7424), pp.473–477.
- Taberlay, P.C. et al., 2011. Polycomb-Repressed Genes Have Permissive Enhancers that Initiate Reprogramming. *Cell*, 147(6), pp.1283–1294.
- Tanimoto, K. et al., 1999. Effects of altered gene order or orientation of the locus control region on human β -globin gene expression in mice. *Nature*, 398(6725), pp.344–348.
- The C. elegans Sequencing Consortium, 1998. Genome Sequence of the Nematode C. elegans: A Platform for Investigating Biology. *Science*, 282(5396), pp.2012–2018.

-
- Thurman, R.E. et al., 2013. The accessible chromatin landscape of the human genome. *Nature*, 488(7414), pp.75–82.
- Tolhuis, B. et al., 2002. Looping and interaction between hypersensitive sites in the active beta-globin locus. *Molecular cell*, 10(6), pp.1453–1465.
- Tufarelli, C. et al., 2003. Transcription of antisense RNA leading to gene silencing and methylation as a novel cause of human genetic disease. *Nature Genetics*, 34(2), pp.157–165.
- Vakoc, C.R. et al., 2005. Proximity among Distant Regulatory Elements at the β -Globin Locus Requires GATA-1 and FOG-1. *Molecular cell*, 17(3), pp.453–462.
- van der Harst, P., 2013. Seventy-five genetic loci influencing the human red blood cell. *Nature*, 492(7429), pp.369–375.
- Varshavsky, A.J., Sundin, O.H. & Bohn, M.J., 1978. SV40 viral minichromosome: preferential exposure of the origin of replication as probed by restriction endonucleases. *Nucleic Acids Research*, 5(10), pp.3469–3477.
- Vernimmen, D. et al., 2009. Chromosome looping at the human α -globin locus is mediated via the major upstream regulatory element (HS- 40). *Blood*, 114(19), pp.4253–4260.
- Vernimmen, D. et al., 2011. Polycomb eviction as a new distant enhancer function. *Genes & development*, 25(15), pp.1583–1588.
- Vernimmen, D. et al., 2007. Long-range chromosomal interactions regulate the timing of the transition between poised and active gene expression. *The EMBO journal*, 26(8), pp.2041–2051.
- Vincent-Fabert, C. et al., 2009. Ig Synthesis and Class Switching Do Not Require the Presence of the hs4 Enhancer in the 3' IgH Regulatory Region. *The Journal of Immunology*, 182(11), pp.6926–6932.
- Viprakasit, V., 2006. A novel deletion causing thalassemia clarifies the importance of the major human alpha globin regulatory element. *Blood*, 107(9), pp.3811–3812.
- Visel, A. et al., 2009. ChIP-seq accurately predicts tissue-specific activity of enhancers. *Nature*, 457(7231), pp.854–858.
- Wallace, H.A.C. et al., 2007. Manipulating the Mouse Genome to Engineer Precise Functional Syntenic Replacements with Human Sequence. *Cell*, 128(1), pp.197–209.
- Wallis, D.E. et al., 1999. Mutations in the homeodomain of the human SIX3 gene cause holoprosencephaly. *Nature Genetics*, 22(2), pp.196–198.

-
- Wang, D. et al., 2011. Reprogramming transcription by distinct classes of enhancers functionally defined by eRNA. *Nature*, 474(7351), pp.390–394.
- Whyte, W.A. et al., 2013. Master Transcription Factors and Mediator Establish Super-Enhancers at Key Cell Identity Genes. *Cell*, 153(2), pp.307–319.
- Wirth-Dzięciołowska, E., Karaszewska, K.P. & Smolińska, M.G., 2009. Selected peripheral blood cell parameters in twelve inbred strains of laboratory mice. *Animal Science Papers and Reports*, 27(1), pp.69–77.
- Wittkopp, P.J., Haerum, B.K. & Clark, A.G., 2004. Evolutionary changes in cis and trans gene regulation. *Nature*, 430(6995), pp.85–88.
- Wray, G.A., 2007. The evolutionary significance of cis-regulatory mutations. *Nature Reviews Genetics*, 8(3), pp.206–216.
- Xiong, Q. et al., 2013. Comprehensive characterization of erythroid-specific enhancers in the genomic regions of human Kruppel-like factors. *BMC Genomics*, 14(1), pp.1–1.
- Yang, L. et al., 2011. Genomewide characterization of non-polyadenylated RNAs. *Genome Biology*, 12(2), p.R16.
- Zeitlinger, J. et al., 2007. Whole-genome ChIP-chip analysis of Dorsal, Twist, and Snail suggests integration of diverse patterning processes in the *Drosophila* embryo. *Genes & development*, 21(4), pp.385–390.