Analysis of a gain-of-function mutation that produces a new transcriptional unit in the α-globin locus

Yavor Krasimirov Bozhilov

Lincoln College

MRC Molecular Haematology Unit
Weatherall Institute of Molecular Medicine



Thesis submitted for the degree of Doctor of Philosophy
University of Oxford

Michaelmas 2017

To my family,

For the unconditional support.

Abstract

The expression of the human α-globin locus is regulated by a set of well characterized enhancer elements. Mutations in the α-globin protein coding genes or their regulatory elements can cause a form of anaemia called α-thalassemia which results from the downregulation of α-globin in red blood cells. There are rare cases where this downregulation is caused by genetic defects outside the know elements.

This thesis deals with a genetic mutation found in an intergenic region in the α-globin locus outside of any know regulatory site. The individuals with the mutation show the typical features of α-thalassemia but have also acquired the expression of a new transcript in their red blood cells. This transcript originates near a single base pair substitution which has been associated with the phenotype and is located upstream of the α-globin genes. The mutation is a transition from the wild type Thymidine to a Cytosine, which generates a new binding motif for the key erythroid transcription factor GATA1. Previous studies of an affected individual's primary erythroid cells demonstrate enrichment of GATA1 in the region of the mutation, along with the recruitment of TAL1 and RNA Polymerase II and an increase in acetylation of histone H3.

The thesis tests the hypothesis that the new transcriptional unit may utilise the established transcription regulatory landscape of the α-globin locus for its own expression. This would result in competition with the α-globin promoters for enhancer activity, thereby causing a downregulation of α-globin transcription.

The first objective of this thesis is to recapitulate the previously observed phenotype in a cell line system. This model is then used to address the question of how the normally un-transcribed sequence has acquired the ability to be expressed and lastly, this thesis explores the molecular mechanisms that cause the downregulation of α-globin.

To this end an iPSC-based erythroid differentiation system was characterized. This system appears to recapitulate key aspects of erythropoiesis and generates a highly pure culture containing a large number of erythroid-like cells. Next, patient iPSC lines containing the mutation were generated and differentiated along the erythroid pathway. These cells faithfully express the novel transcript and show a reduction in α-globin mRNA levels. Expression of the new transcript is accompanied by the emergence of a new region of open chromatin at the site of the mutation along with the recruitment of GATA1, KLF1 and RNA Polymerase II and the enrichment of the promoter mark H3K4me3.

Chromatin conformation analysis was used to assess the proximity between the elements in the α-globin locus in the presence of the mutation. This demonstrates increased interaction between the enhancers and the site of the new transcriptional unit and reduced interaction of the α-globin promoters with their enhancers, suggesting competition for enhancer activity.

Finally, to demonstrate that the variant is necessary for the observed phenotype a patient line was genetically engineered to correct the mutation. This rescues the phenotype by reducing the expression of the new transcript, abolishing the accessible region and upregulating α-globin transcription. In addition, the mutation was introduced in a wild

type iPSC line to recreate the phenotype. This results in the emergence of the patient-specific region of accessible chromatin, expression of the novel transcript and downregulation of α-globin. These observations prove that the single base pair change is sufficient to cause the formation of the new transcriptional unit and that it is the causal mutation for this form of α-thalassemia.

Acknowledgements

I would like to express my gratitude to my supervisors Prof Douglas Higgs and Prof Richard Gibbons for all the support and advice throughout the project over the last four years. They have created a stimulating and nurturing environment in the laboratory which has helped me maximise my potential as a scientist. I would also like to thank Prof Catherine Porcher and Prof Rob Klose on my thesis committee for the discussions and guidance.

I feel lucky to have benefited from the knowledge and expertise of the people with whom I have at one time or another tackled the obstacles of this project. I am indebted to Dr Lee Carpenter for helping me improve my iPSC culture technique and for guiding me through many experiments. I am grateful to Dr Emmanuel Olivier and Dr Jo Mountford for sharing their differentiation protocol as this has made the entire project possible. I would like to thank Dr Chris Fisher and Dr David Clynes for their help and advice. Throughout my DPhil I have relied on Dr Jelena Telenius, Ms Marieke Oudelaar, Ms Julia Truch, Mr Martin Larke, Mr Ron Schwessinger and Dr Maria Suciu for help with my bioinformatics analysis and I would like to express my appreciation for their time and patience. I highly value the contributions of Prof Jim Hughes for his insights into the project and Dr Christian Babbs for his expert advice with all matters concerning molecular biology. I would also like to thank Dr Jackie Sloane-Stanley for invaluable assistance with tissue culture work.

This thesis has been improved through many discussions with members of the Higgs and Gibbons laboratories both past and present and other members of the institute: Prof Andrew Smith, Dr Lars Hanssen, Dr Philip Hublitz, Dr Mira Kassouf, Dr James Davies, Dr Bryony Graham, Dr Diu Nguyen, to name but a few. I feel lucky to have had Dr Anthony Cheong as a friend throughout the course of my PhD. I would also like to acknowledge Dr Andrew King's involvement in making the late nights at work more tolerable.

Finally, none of this would have been possible without the support and love I have received from my family. My heartfelt gratitude goes to my parents who have supported me through many hardships and to my partner Veronika who has been by my side for the last seven years and always finds a way to make me smile.

Благодаря ви!

List of abbreviations

3C	Chromosome conformation capture
A	Adenine
APC	Allophycocyanin
ATAC	Assay for transposition of native chromatin
BCL11A	B-cell lymphoma/leukaemia 11A protein
BFU-E	Burst forming unit-erythroid
BRE	TFIIB recognition element
bHLH	basic helix-loop-helix
bp	base pair
C	Cytosine
Cas	CRISPR-associated
CBP	CREB binding protein
CD	Cluster of differentiation
cDNA	Complementary DNA
CFU-E	Colony forming unit-erythroid
ChIP	Chromatin Immunoprecipitation
CNV	Copy number variation
CN-LOH	Copy-neutral loss of heterozygosity
CpG	Cytosine guanine dinucleotide
CRISPR	Clustered, regularly interspaced, short palindromic repeat
crRNA	CRISPR RNA
CRM	*cis*-regulatory module
CTCF	CCCTC-binding factor
c-MYC	MYC proto-oncogene, bHLH transcription factor
DCE	the downstream core element
DNA	Deoxyribonucleic acid
DNaseI	Deoxyribonuclease I
dNTP	2'-deoxynucleotides 5'-triphosphate
DPE	Downstream promoter element
DSB	Double strand break
EB	Embryoid body
EBV cells	Epstein–Barr virus transformed lymphoblastoid cells
EKLF	Erythroid Kruppel-like factor
EMP	Erythro-myeloid progenitor
EPO	Erythropoietin
ES/ESC	Embryonic stem/ embryonic stem cell
ETO2	Eight-Twenty-One 2
ETS	E26 transformation-specific
F	Fusion protein
FACS	Fluorescence activated cell sorting
FCS	Fetal calf serum
FITC	Fluoresceinisothiocyanate
FOG-1	Friend of Gata 1
FOX	Forkhead box

G	Guanine
GATA1	GATA-binding factor1
G/C	Guanine-cytosine
GFP	Green fluorescent protein
GTF	General transcription factor
GWAS	Genome-wide association study
H	Histone
H3ac	Histone H3 acetylation
H3K4me1	Histone H3 lysine 4 monomethylation
H3K4me3	Histone H3 lysine 4 trimethylation
H3K27ac	Histone H3 lysine 27 acetylation
H4ac	Histone H4 acetylation
HAT	Histone acetyl transferases
Hb	Haemoglobin
HBA	α-globin
HbA	Adult haemoglobin
HBB	β-globin
HBE	ε-globin
HbF	Foetal haemoglobin
HBG	γ-globin
HbH	Haemoglobin H
HBM	μ-globin
HBQ	θ-globin
HBZ	ζ-globin
HDAC	Histone deacetylase
HDR	Homology directed repair
HN	Hemagglutinin-neuraminidase
HS	Hypersensitive
HSC	Haematopoietic stem cell
ICM	Inner cell mass
IF	Immunofluorescence
IgG	Immunoglobulin G
Inr	Initiator promoter element
iPSC	Induced pluripotent stem cell
K	Lysine
Kb	Kilo bases
KLF1	Kruppel-like factor 1
KLF4	Kruppel-like factor 4
LCR	Locus Control Regions
LDB1	LOB domain-containing protein 1
LIN28A	lin-28 homolog A
LINE	long interspersed nuclear element
LMO2	LIM domain only 2
LTR	Long terminal repeat
M	Matrix protein
MACS	Magnetic activated cell sorting
MEF	Mouse embryonic fibroblast

MCS	Multispecies conserved sequences
mg	Milligram
min	Minute
ml	Millilitre
mM	Millimolar
MNC	Mononuclear cell
MOI	Multiplicity-of-infection
mRNA	Messenger ribonucleic acid
MTE	the motif ten element
MYB	MYB proto-oncogene, transcription factor
NANOG	Nanog homeobox
NF-E2	Nuclear factor-erythroid 2
ng	Nanogram
NGS	Next generation sequencing
NHEJ	Non-homologous end joining
NP	nucleocapsid protein
NPRL3	NPR3 like, GATOR1 complex subunit
NuRD	Nucleosome Remodeling Deacetylase
OCT4	Octamer-binding transcription factor 4
p300	Histone acetyltransferase p300
PAM	Protospacer-adjacent motif
PB	Peripheral blood
PBMC	Peripheral blood mononuclear cells
PBS	Phosphate-buffered saline
PCR	Polymerase chain reaction
PE	Phycoerythrin
PI	Protease inhibitor
PRC	Polycomb repressive complex
PSC	Pluripotent stem cells
PTM	Post-translational modification
P/S	Penicillin/Streptomycin
qRT-PCR/qPCR	Quantitative real-time polymerase chain reaction
RBC	Red blood cell
RNA	Ribonucleic acid
RNA Pol II	RNA polymerase II
RPL13A	Ribosomal Protein L13a
RPS18	Ribosomal protein S18
RT	Reverse transcription
SCF	Stem cell factor
SCL	Stem cell leukaemia
SCNT	Somatic cell nuclear transfer
SD	Standard deviation
SEA	Southeast Asian
Seq	Sequencing
SeV	Sendai virus
sgRNA	Short guide RNA
SNP	Single nucleotide polymorphism

SNV	Single nucleotide variant
SOX2	SRY (sex determining region Y)-box 2
SSEA4	Stage-specific embryonic antigen-4
SWI/SNF	SWItch/Sucrose Non-Fermentable
T	Thymidine
TAF9	TATA-Box Binding Protein Associated Factor 9
TAD	Topologically associating domains
TAL1	T-cell acute lymphocytic leukemia protein 1
TALEN	Transcription Activator-Like Effector Nucleases
TF	Transcription factor
TFBS	Transcription factor binding site
TR	Tandem repeat
tracrRNA	Trans-activating CRISPR RNA
TRA-1-60	Pluripotency marker antibody recognizing epitopes on podocalyxin
TSS	Transcription start site
T-ALL	T-cell acute lymphoblastic leukaemia
U	Uracil
VAN	Vanuatu
VNTR	Variable number of tandem repeat
WT	Wild type
XCPE1	x core promoter element 1
ZFN	Zinc finger nuclease
^{o}C	Degrees Celsius
μg	Microgram
μl	Microlitre

Table of content

Abstract……iii
Acknowledgements……vi
List of abbreviations……viii
Chapter 1: Introduction……1
1.1 Transcription regulation……2
1.2 α-globin gene cluster and regulation……16
1.3 α-thalassemia……22
1.4 Vanuatu mutation……24
1.5 Pluripotent stem cells, reprogramming, and the use of iPSCs as a system to model human genetic diseases……30
1.6 Summary, Hypothesis and Aim……39
Chapter 2: Materials and Methods……41
2.1 Cell culture techniques……42
2.1.1 Induced Pluripotent Stem Cell Maintenance……42
2.1.2 Reprogramming…… 43
2.1.3 Erythroid differentiation of iPSCs……45
2.1.4 Genome editing of iPSCs……47
2.2 Cell biology techniques……50
2.2.1 Flow cytometry……50
2.2.2 Immunofluorescence (IF) ……50
2.2.3 Wright stain of cytospins for morphology analysis……51
2.3 Molecular Biology techniques……52
2.3.1 RNA isolation and cDNA generation……52
2.3.2 Quantitative real-time polymerase chain reaction (qRT-PCR)……52
2.3.3 DNA isolation……54
2.3.4 Genotyping……55
2.3.5 Assessing edited clones for large indels……56
2.3.6 Copy number variation analysis……58
2.3.7 Detection of SeV vectors……58
2.3.8 Chromatin Immuno-Precipitation followed by Next Generation Sequencing……59
*2.3.9 Assay for transposase-accessible chromatin sequencing (ATAC-Seq)*61
2.3.10 Next Generation Capture C (published in Davies et al., 2016)……62
2.3.11Preparation of short guide RNA (sgRNA) expression construct.……63
2.4 Computational analysis……65
2.4.1 Alignment, normalization and visualization……65
Chapter 3: Reprogramming and iPSC quality control……66
3.1 Introduction……67
3.2 Reprogramming expanded CD34+ cells and EBV cells using Sendai vectors……71
3.3 CNV analysis……75
3.4 Pluripotency markers and vector detection……81
3.4.1 Immunofluorescence (IF) for pluripotency markers……82
3.4.2 RT-PCR for Vector DNA……83

3.5 Concussion.....85
Chapter 4: Characterizing the Erythroid differentiation system.....87
4.1 Introduction.....88
4.2 Differentiation, Expansion, FACS and Morphologic analysis.....93
4.3 Gene expression.....99
4.4 Analysis of open chromatin regions.....101
4.5 Chromatin mark analysis of the α-globin cluster in normal iPSCs.....103
4.6 Analysis of Chromatin conformation at the α-globin locus by Capture C.....105
4.7 Analysis of common types of α-thalassemia iPSC lines.....107
4.8 Conclusion.....109
Chapter 5: Recapitulating the Vanuatu phenotype.....113
5.1 Introduction.....114
5.2 Expression analysis of patient derived iPSC lines differentiated along the erythroid pathway.....117
5.3 Chromatin mark analysis of the a-globin locus.....121
5.4 Analysis of open chromatin regions by ATAC-seq.....128
5.5 Chromatin conformation analysis (Capture C) 130
5.6 Conclusion.....134
Chapter 6: Editing human iPSC lines.....136
6.1 Introduction to CRISPR.....137
6.2 Editing strategy.....142
6.3 Editing iPSC lines.....155
6.4 Conclusion.....162
Chapter 7: Rescuing and recreating the Vanuatu phenotype.....164
7.1 Introduction.....165
7.2 Expression analysis of gene edited iPSC lines differentiated down the erythroid pathway..... 166
7.3 Analysis of open chromatin regions (ATAC-seq) of edited iPSC lines differentiated down the erythroid pathway.....169
7.4 Conclusion..... 171
Chapter 8: Discussion.....173
8.1 Modelling an erythroid disease in an iPSC-based differentiation system.....174
8.2 Transcription factor binding at the site of the Vanuatu mutation.....177
8.3 The nature of the GATAν transcriptional unit.....181
8.4 Model of interaction between GATAn and the a-globin enhancers.....188
8.5 Summary.....193
Appendix.....194
References.....208

Chapter 1: Introduction

1.1 Transcription regulation

Multicellular organisms use gene regulation to determine cell differentiation and morphogenesis. This leads to the establishment of different types of specialised cells, with different gene expression profiles, that have arisen from the same genomic content. The expression of eukaryotic genes is controlled at many levels, including transcription as a major influence, to ensure the appropriate genes are switched on and off at the correct stages in development and differentiation. In eukaryotic cells, DNA is tightly packed around histones forming nucleosomes and is condensed into chromatin. The compaction limits the amount of DNA available for the binding of many regulatory factors (Li et al., 2007) making the chromatin structure intrinsically repressive, this in turn stabilizes gene expression states and prohibits unregulated cell fate transitions. The regulation of transcription is achieved by the interaction between chromatin, transcription factors, transcription co-regulators and their 3-dimensional organization in the nucleus.

Chromatin compaction

The smallest structural unit of DNA packaging is the nucleosome. It is comprised of 146 bases of DNA wound around an octamer composed of two pairs of core histones H2A, H2B, H3 and H4. The nucleosomes can be further compacted into higher-order chromatin structures with the help of additional factors such as histone chaperones and nucleosome remodelling factors (Kornberg and Lorch, 1999). Chromatin organisation can directly dictate gene expression by virtue of nucleosome density. If it is high, the DNA strands

become difficult to access by the transcription machinery (Knezetic and Luse, 1986; Lorch et al., 1987). In this way chromatin can act as a barrier that needs to be actively overcome for transcriptional activation to occur, and cells possess mechanisms to achieve this by locally remodelling the accessibility of their chromatin to switch genes on or off (Bell et al., 2011). Those mechanisms involve chromatin remodelling and histone modifying proteins, however, some transcription factors (TFs) referred to as pioneer transcription factors can access DNA in regions of compacted chromatin to create a permissive environment in which other factors may bind (Zaret and Carroll, 2011; Zaret and Mango, 2016).

Chromatin remodellers are multiprotein complexes that use energy from ATP hydrolysis to redistribute nucleosomes along the DNA template to form either inaccessible nucleosome-dense regions or accessible nucleosome-depleted regions (Saha et al., 2006). In addition, the human genome holds more than 150 proteins that are involved in histone modification (Khare et al., 2012) and at least 130 unique histone post-translational modifications (PTMs) (Tan et al., 2011). Most of the known PTMs are found on the N-terminal tails of core histones, however, modifications have also been identified in the globular domains of histone proteins. There are several mechanisms by which histone PTMs are thought to influence chromatin structure. PTMs can modulate the net charge of histone molecules to alter the DNA to nucleosome or inter-nucleosomal interactions and modify chromatin accessibility. Specific histone modifications can be recognized and bound by regulatory proteins that contain specialized domains (e.g. PHD-finger, bromo- and chromo- domains, etc.). Conversely, histone PTMs can serve to inhibit the interaction between specific factors and chromatin (Bannister and Kouzarides, 2011). The pattern of

histone tail modifications can correlate with specific features of the underlying DNA. A good example is the tri-methylation of the lysine residue at position 4 of histone variant H3 (H3K4me3) which is found at promoters (Bernstein et al., 2005; Guenther et al., 2007; Kim et al., 2005; Mikkelsen et al., 2007). In the context of chromatin accessibility, the acetylation status of lysine residues on histone H3 and H4 tails is believed to act as an important regulatory system. Acetylated lysine residues lose their positive charge and reduce the interactions between histones and the negatively charged DNA backbone and in this way, may adapt chromatin structure to a more permissive environment (Bannister and Kouzarides, 2011).

Cis-regulatory elements

Covalent histone modifications are associated with different degrees of DNA accessibility, and increased accessibility at regulatory sequences is associated with gene transcription (Higgs et al., 2007; Lelli et al., 2012). These sequences are found within *cis*-regulatory elements (see *fig.* 1.1.1) and govern the rules which ensure the correct pattern of gene expression required for development (Buecker and Wysocka, 2012; Bulger and Groudine, 2011; Levine, 2010).

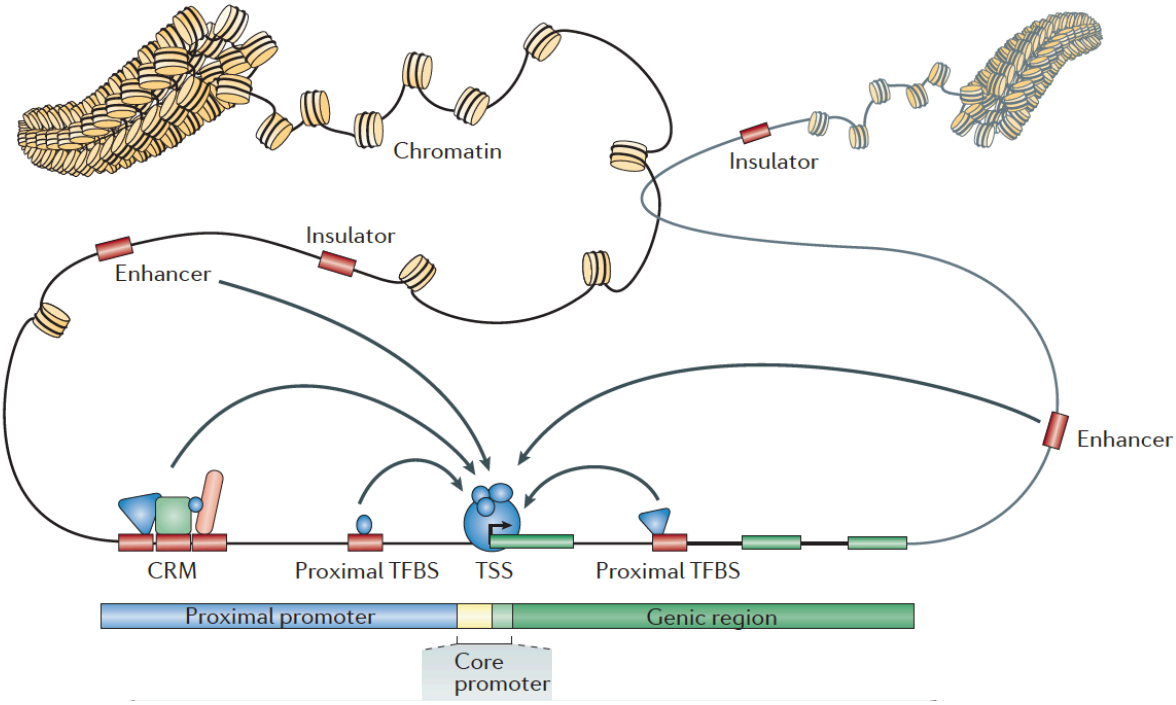


Figure 1.1.1 A summary of *cis*-regulatory elements.
Chromatin can be highly compacted and inaccessible, or it can be accessible to proteins. The boundaries between the different states of compaction can be marked by insulators. The regulatory elements around the transcription start site (TSS) can be divided into the core promoter which contains the TSS, and the upstream proximal promoter region. Initiation of transcription is facilitated by the binding of sequence-specific regulatory proteins that are referred to as transcription factors (TFs). The TFs recognize and bind to specific binding sites (TFBSs) which are located near the TSS or can be a long distance from it (e.g. in enhancers) and can also occur in clusters giving rise to *cis*-regulatory modules (CRMs). Bound TFs recruit RNA polymerase II and co-activators to form complexes that stabilise the transcription-initiation machinery and enable gene expression. Figure was adapted from Lenhard et al., 2012.

The transcription of a gene starts at the transcription start site (TSS) which lies within a region called the promoter. In addition to the TSS the promoter also contains sequence elements required for the initiation of transcription by the DNA-dependent RNA polymerase. These sequence motifs are bound by transcription factors in a sequence-specific manner. The initiation of transcription requires the action of a specialized group of general TFs (GTFs) that bind to several interchangeable sequence elements near the TSS in a region called the core promoter (Lenhard et al., 2012; Thomas and Chiang, 2006).

When the GTFs bind the core promoter, they recruit RNA polymerase and initiate transcription (Lee and Young, 2000). The genes that are transcribed by RNA polymerase II (RNA Pol II) are very diverse in terms of their expression levels and their context specificity. This calls for a highly specialized and dynamic type of transcription control. A major component of this diversity is the specification of different classes of RNA Pol II promoters with different architectures that determine their function and form of regulation (Sandelin et al., 2007; Valen and Sandelin, 2011). Metazoan promoters, in particular, are complex and their regulatory elements can be spread over large distances. Additionally, the number of regulatory elements in multi cellular eukaryotes is higher than in single cell eukaryotes, reflecting the more diverse regulatory tasks imposed by multicellularity, such as the development and maintenance of specialised tissues and cell to cell communication (Lenhard et al., 2012).

Metazoan gene regulation relies predominantly on the combination of promoters and multiple distal regulatory elements including enhancers and promoters (Levine and Tjian, 2003). The most prominent and highly studied distal regulatory elements are the enhancers. Enhancers are DNA sequences that contain multiple transcription factor binding motifs and are usually found within non-coding regions of the genome. They were originally defined as sequences that can activate transcription independent of their orientation, location or distance with respect to their corresponding promoters. A definition that was based on *in vivo* reporter assays and *in vitro* reconstituted biochemical systems (Banerji et al., 1981; Dynan and Tjian 1985; Moreau et al., 1981; Sergeant et al., 1984; Treisman and Maniatis, 1985). Recent genetic and genomic approaches have clarified the properties of enhancers. We know they can be found at various distances

from their target genes, anywhere between 1kb to over 1000 kb, with multiple enhancer elements often regulating the expression of a single gene (de Laat and Duboule, 2013; Maston et al., 2006). Enhancers have a characteristic chromatin signature: namely, high chromatin accessibility, enrichment of H3K4me1 and H3K27Ac, and low enrichment of H3K4me3 (Shlyueva et al., 2014). The current model for enhancer function dictates that enhancers act as clusters of motif sequences capable of binding combinations of transcription factors that interact with other transcription factors and co-activators (e.g. components of the Mediator complex) to help recruit RNA Pol II to the promoter and influence various aspects of transcription (Malik and Roeder, 2010; Maston et al., 2006). Various models have been proposed to explain how enhancers facilitate this function. One model called the "tracking model" (Blackwood and Kadonaga, 1998; Hatzis and Talianidis, 2002; Tuan et al., 1992) suggests that components of the transcription machinery are first assembled at enhancers, then disengage from them and track along the DNA, possibly with RNA Pol II serving as a molecular motor, to reach the relevant promoters. At that point, the complete transcriptional apparatus is assembled and transcription of that gene is initiated. However, the function of enhancer regions appears to be mediated through physical interaction with promoters. Therefore, a second model put forward to explain how remote regulatory elements communicate with their target genes proposes that enhancers and promoters interact directly via DNA binding multiprotein complexes, with the intervening chromatin looping out, to facilitate the delivery of enhancer bound TFs and co-activators, and promote the recruitment of the transcription machinery (Bulger and Groudine, 2011; Ong and Corces, 2011; Schoenfelder et al., 2010). When it was first proposed, this "looping model" was based on observations of bacterial and phage repressor proteins, which function by homo-multimerization when bound to different

regulatory sites. Here electron microscopy was used to visually demonstrate that the DNA between the binding sites loops out (Ptashne, 1986). This model assumes a certain level of flexibility of the chromatin fibre which facilitates the random collision between chromatin fibres with a frequency determined by physical constraints and the stability of the formed loop depending on affinities between proteins bound to colliding DNA segments. A third model referred to as “facilitated tracking”, proposes that regulatory protein complexes assemble at *cis*-regulatory elements and stay attached to them as they move towards each other and to their target promoters by scanning the intervening chromatin, progressively forming a loop (Blackwood and Kadonaga, 1998; Wang et al., 2005). A recently proposed model of loop formation is based on a hypothetical extrusion complex, which consists of two DNA-binding subunits, each of which can extrude DNA in the opposite direction as they slide along chromatin. The loading of the complex to a single locus would facilitate the formation of a small chromatin loop and as chromatin is extruded through the tethered subunits, such that they move in opposite directions with respect to the genome, the loop would continue to grow until the complex encounters appropriately oriented boundary motifs (*fig*. 1.1.2) (Alipour and Marko, 2012; Nasmyth, 2001; Sanborn et al., 2015) a similar concept that relies on only one DNA extruding subunit has also been proposed (Fudenberg et al., 2016). It has been suggested that the “extrusion” model might involve the cohesin complex which forms a ring that can slide along chromatin and its interaction partner, the insulator-binding protein CTCF which has important general functions in transcription regulation and chromatin organization (Ong and Corces, 2014; Phillips-Cremins et al., 2013). The extrusion model can explain a wider range of observations compared to other models. In polymer simulations, the extrusion model can reproduce many of the finer-scale features of maps generated by the

chromatin conformation capture technique Hi-C. It explains the enrichment of architectural proteins at the boundaries of the sub-megabase scale self-interacting regions that partition interphase chromosomes of higher eukaryotes which are known as topologically associating domains (TADs). The model also explains the common orientation of CTCF biding motifs at the boundaries of chromatin loops that have overwhelmingly been found to form between convergent CTCF sites. In addition, loop extrusion explains the strong tendency of chromatin loops not to overlap or give rise to intra-chromosomal knots and inter-chromosomal entanglements as opposed to random diffusion which is likely to result in substantial overlap and entanglement (Fudenberg et al., 2016; Sanborn et al., 2015).

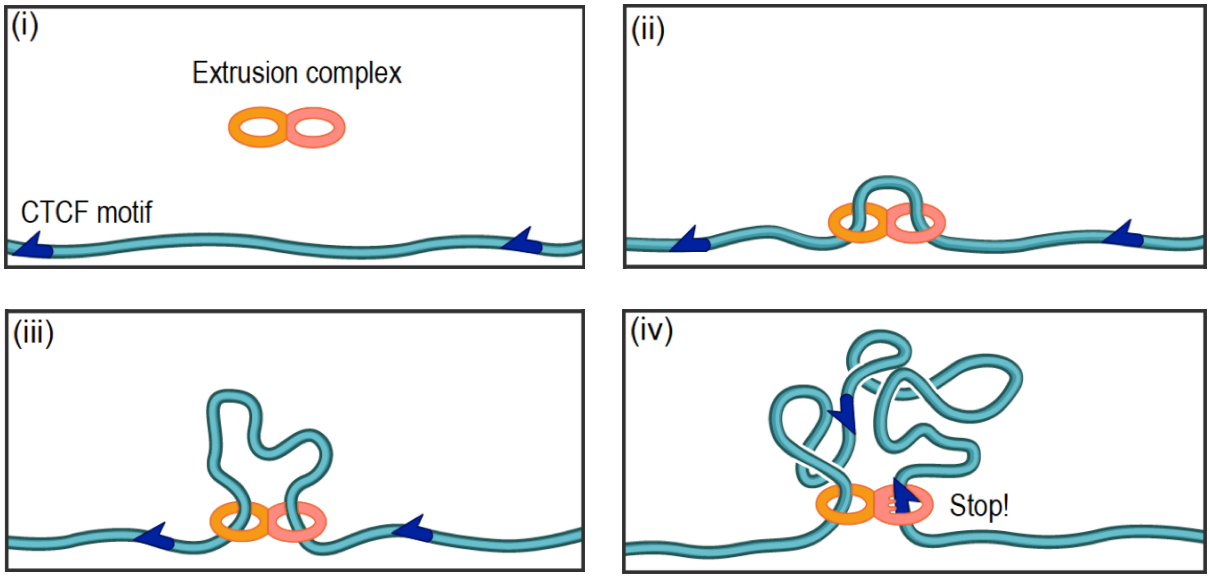


Figure 1.1.2 Loop formation based on the loop extrusion model.
The extrusion complex loads onto the chromatin fiber at a random locus, the act of lading results in the formation a short-range loop (i and ii). The subunits move in opposite directions along the fiber which results in the expansion of the loop and the formation of a domain by the extruded fiber (iii). When a subunit of the extrusion complex detects an appropriately positioned motif it stops sliding (iv). Figure was adapted from Sanborn et al., 2015.

Enhancers very often function in a tissue-specific manner, driven by specific transcription factors, and play key roles in regulating differentiation and development. The ability of a group of enhancers, and adjacent elements, to regulate high-level, tissue-specific gene expression has led to defining the Locus Control Regions (LCRs) (Li, Q., et al., 2002) and more recently the 'super-enhancers' (Hnisz et al., 2013; Lovén et al., 2013; Whyte et al., 2013). The first LCR to be described was the β-globin locus control region which demonstrated a potent and tissue-specific transcription-enhancing activity over an entire gene cluster and the ability to create and maintain chromatin accessibility when integrated into ectopic genomic locations (Grosveld et al., 1987). The β-globin LCR has also served as a model to provide the first compelling evidence in support of the chromatin-looping model for enhancer function (Tolhuis et al., 2002). Later, the regulatory sequences of the α-globin locus were extensively characterised. The cis-regulatory elements that control α-globin expression have been studied using a variety of approaches, including transgenic experiments (Higgs et al., 1990; Sharpe et al., 1992), expression analysis in interspecific hybrids (Sharpe et al., 1993; Vyas et al., 1992), comparative genomics (Flint et al., 2001; Hughes et al.,2005), chromatin analysis (Anguita et al., 2001; Anguita et al., 2005) and precise genome engineering (Hanssen et al., 2017; Hay et al., 2016; Wallace et al., 2007). This has built a comprehensive picture of the regulation of the α-globin locus (reviewed in section 1.2).

Transcription factors

Transcription factors are proteins with a modular structure that contain a DNA binding domain responsible for biding specific, short sequence motifs (Latchman, 1997).

Regulatory regions commonly contain multiple binding motifs whose order, orientation and spacing can play a key role in recruiting transcription factor complexes and determining modes of regulation (Spitz and Furlong, 2012). The transcription factors that are relevant for this study are the key erythroid factors KLF1 and GATA1 along with some of their most prominent binding partners, specifically TAL1 (also known as SCL1). A more comprehensive list of the transcription factors involved in erythropoiesis can be found at Higgs et al., 2008.

GATA1

GATA1 or GATA-binding factor 1 is a member of the GATA family of transcription factors with key functions in the hematopoietic lineage. In particular, GATA1 is a master transcription factor required for erythropoiesis which has been demonstrated by knock-out experiments (Fujiwara, et al., 1996; Pevny et al. 1991). It regulates genes necessary for normal erythropoiesis in a stage-specific manner and can act both as an activator and a repressor through interactions with co-factors as well as through PTMs to specific domains (Ferreira et al., 2005; Love et al., 2014; Welch et al., 2004). GATA1 uses its C-terminal C2C2 zinc finger domain to bind DNA at consensus sequence (A/T)GATA(A/G) (Ferreira et al., 2005). GATA1 predominantly binds at distal regulatory elements and at a minor proportion of proximal promoters (Fujiwara et al., 2009; Yu et al., 2009). This holds true for the human α-globin locus where GATA1 is found at the upstream enhancers but is relatively poorly enriched at the promoters (De Gobbi et al., 2007).

GATA1 regulates transcription through its many binding partners. The canonical co-factor of GATA1 is the transcriptional co-regulator Friend of GATA1 or FOG1. FOG1 binds GATA1 (Rodriguez, 2005), is an essential co-regulator of erythropoiesis (Tsang et al., 1998) and

its interaction with GATA1 is necessary for erythroid differentiation (Crispino et al., 1999). FOG1 can mediate both activation and repression of GATA1 target genes (Bresnick et al., 2010), one of the ways FOG-1 does this involves interaction with the NuRD chromatin remodelling complex (Gao et al., 2010; Hong et al., 2005; Miccio et al., 2010). GATA1 may also utilize FOG1 to promote chromatin looping (Jing et al., 2008; Kim et al., 2007; Vakoc et al., 2005). GATA1 can regulate genes independently of FOG1. For example, by interacting with the multimeric LDB1-complex. The complex consists of TAL1/E-protein heterodimer bound to GATA1 via LMO2 and LDB1 (Wadman et al., 1997). Transcriptional regulation by the LDB1-complex is thought to be achieved through facilitating interactions between distal regulatory elements and promoters (Love et al., 2014).

Through interaction with its various partners, GATA1 may be able to convert chromatin to a transcriptionally permissive environment. By recruiting histone acetyltransferase CBP, GATA1 may promote chromatin accessibility mediated via acetylation of histone H3 and H4 (Letting et al., 2003). In addition, GATA1 has been shown to cooperate with NF-E2 to contribute to the recruitment of BRG1, the ATPase component of SWI/SNF chromatin remodelling complex, and to the formation of accessible chromatin regions at the β-globin LCR (Woon Kim et al. 2011). The ability of GATA1 to recruit the chromatin remodeller BRG1 has been demonstrated previously with respect to BRG1 function in maintaining higher order chromatin structure (Hu et al., 2011; Kim et al., 2009a). In particular, BRG1 is required for long range interactions in the mouse α-globin locus (Kim et al., 2009b) where GATA1 is known to bind to both the enhancers and the globin promoters early on during erythroid differentiation (Anguita et al., 2001).

KLF1

KLF1 or Erythroid Kruppel-like factor (also known as EKLF) is a member of the Kruppel-like factor family (Bieker, 2001). KLF1 is a master erythroid transcription factor (Miller and Bieker, 1993; Perkins et al., 1995) which regulates the expression of erythroid-lineage specific genes, including the globin genes, genes controlling heme synthesis, and genes regulating the cell cycle, among many others (Hodge et al., 2006; Tallack et al., 2012). KLF1 uses its three C-terminal C2H2 zinc finger domains to bind DNA at a consensus sequence CC(A/C)C(A/G)CCCN, referred to as the extended CACC motif (Miller and Bieker, 1993). Like GATA1, the majority of KLF1 binding sites are found in distal regulatory elements while a minor fraction is found in the vicinity of TSSs (Tallack et al., 2010). In both the human and mouse α-globin locus KLF1 is found at a subset of the enhancer elements and occupies the globin promoters (Higgs et al., 2008).

KLF1 can interact with units of the basal transcription machinery, namely p62/TFIIH and TAF9, and this interaction has been shown to be important for the optimal activation of the β-globin promoter (Mas et al., 2011; Senguptaet al., 2009). KLF1 is an important factor regulating the developmental γ- to β-globin gene switch: mutations in KLF1 cause elevated levels of γ-globin expression in adult erythrocytes (Borg et al., 2010; Perkins et al., 1996). KLF1 interacts with CBP/p300 (Zhang and Bieker 1998) and the SWI/SNF complex (Armstrong et al., 1998; Kadam et al., 2000) and plays an important role in chromatin accessibility as a lack of KLF1 leads to loss of accessible sites in the β-globin locus and histone hypo-acetylation (Bottardi et al., 2006; Wijgerde et al., 1996). KLF1 is also required for the regulation of chromatin looping and the formation of a functional active chromatin hub. One possible mechanism for this is by interacting with GATA1/TAL1 in the LDB1-complex (Drissen et al., 2004; Kang et al., 2015; Tallack et al., 2012). KLF1 and GATA1 can

interact *in vitro* and exhibit functional synergy at the promoters of several erythroid genes (Gregory et al., 1996; Merika and Orkin, 1995). It is not established whether their physical interaction is required for this synergism. However, genome-wide studies show that GATA motifs can be found at a non-random distance from KLF1 binding sites (Tallack et al., 2010) and that a significant number of genes are co-occupied by GATA1, KLF1 and TAL1 and have suggested these TFs form a transcriptional core network (Li et al., 2013; Wontakal et al., 2012).

TAL1

TAL1 or T-cell acute lymphocytic leukaemia protein 1 is a member of the helix-loop-helix transcription factor family (Begley et al., 1989). TAL1 has an essential role in all haematopoietic lineages (Porcher et al., 1996) and more specifically is necessary for terminal erythroid maturation (Mikkola et al., 2003). TAL1 is known to function as a heterodimer with other basic helix-loop-helix factors called E-proteins and binds DNA at a consensus sequence CANNTG referred to as an E-box motif (Massari and Murre, 2000). In addition, TAL1/E-protein bind the non-DNA binding co-factors LMO2 and LDB1 to form the core LDB1 complex which is able to recruit additional factors. GATA1 is one of the best-described partners of the complex and TAL1 ChIP-seq analyses describe GATA motifs as the most frequent TAL1-bound sequences (Kassouf et al., 2010; Org et al., 2015; Palii et al., 2011; Wu et al., 2014). In addition to composite TAL1 and GATA1 binding sites (E-box-GATA1 motifs) being frequently identified (Li et al. 2013) half E-box-GATA motifs (CTG(N_{8-10})WGATAR) have been found in sequences bound by TAL1 (El Omari et al., 2013; Kassouf et al., 2010; Soler et al., 2010). In both the human and mouse α-globin locus, TAL1 is found at the enhancer elements but not at the globin promoters (Higgs et al., 2008).

The LDB1 complex can act both as an activator and a repressor through recruitment of additional factors. Co-factors such as p300/CBP and the repressor ETO2 interact directly with E-proteins and can switch between activating and repressive complexes through competitive binding (El Omari et al., 2013; Zhang et al., 2004). In erythropoiesis, TAL1 complexes can be repressive through association with GATA2 in early progenitors, while interaction with GATA1 can serve to activate genes necessary for terminal maturation (Fujiwara et al., 2009; Pimkin et al., 2014; Tripic et al., 2009). Lastly, the TAL1 complex is involved in chromatin architecture through chromatin looping. Chromatin looping is thought to be promoted through the dimerization of LDB1 to bring together enhancers and promoters. This event promotes recruitment of co-factors and RNA Pol II at promoters and leads to gene activation (Deng et al., 2012; Krivega et al., 2014).

The pivotal role of these transcription factors in erythroid specification and gene regulation has been highlighted in a recent trans-differentiation experiment. Forced expression of GATA1, TAL1, LMO2, and c-MYC was able to reprogram fibroblasts to cells with erythroid characteristics and gene expression patterns similar to those of primitive erythroblasts. Subsequently, the overexpression of KLF1 was necessary to induce the expression of adult haemoglobin in the transdifferentiated erythroid progenitors (Capellera-Garcia et al., 2016). This finding further defines GATA1 TAL1 and KLF1 as part of a conserved transcriptional program that governs erythroid cell fate specialization and maturation in mammals. In addition, their ability to direct fate conversion which involves the capacity to overcome epigenetic barriers and to establish cell type-specific transcriptional programs, suggests that some of them may act as or interact with pioneer factors (Iwafuchi-Doi and Zaret, 2014).

1.2 α-globin gene cluster and regulation

The α-globin locus is activated during erythropoiesis to ensure the production of α-globin chains which form part of the haemoglobin molecule necessary for the primary function of red blood cells (erythrocytes). Haemoglobin (Hb) is a highly specialised protein responsible for the delivery of oxygen from the lungs to tissues. In humans, haemoglobin is a tetramer consisting of two α-like and two β-like globin chains. Different isoforms of haemoglobin made up of different α-like and β-like globin chains are synthesised during development in accordance with the changing needs for oxygen transport associated with growth. Hb Gower-I ($\zeta_2\varepsilon_2$), Hb Gower-II ($\alpha_2\varepsilon_2$) and Hb Portland ($\zeta_2\gamma_2$) are produced during the embryonic stage of development and are then replaced by HbF ($\alpha_2\gamma_2$) in the fetal stage, which is replaced by HbA ($\alpha_2\beta_2$) and HbA2 ($\alpha_2\delta_2$) in the adult (*fig.* 1.2.1) (Higgs and Wood, 2011).

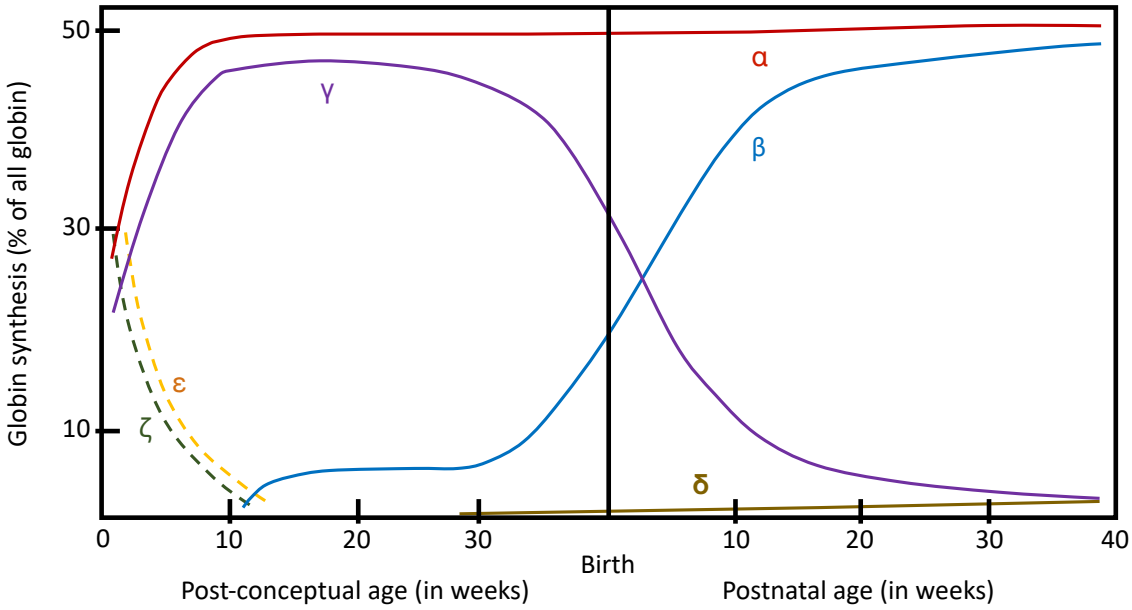


Figure 1.2.1 Expression of the human globin genes during development. The α-like ζ and β-like ε are the first globin genes to be expressed in the early embryo and are then switched to α and β-like γ in the fetal stage. The switch from γ to β takes place in the first months after birth (Higgs et al., 2012).

The human α-globin locus sits on the short arm of chromosome 16 (16p13.3) located ~150kb away from the telomere. The globin genes starting from the telomere are ζ-globin gene *HBZ*, minor globin gene μ *HBM*, then the two α-globin genes *HBA2* and *HBA1* and lastly the minor globin gene θ *HBQ*; there are also two pseudogenes in the locus ψζ *HBZP1* (between ζ and μ) and ψα1 *HBAP1* (between μ and α2). The *cis*-regulatory elements of the gene cluster have been identified using comparative sequence analysis and DNaseI hypersensitive site mapping (Higgs and Wood, 2008; Hughes et al., 2005). In the human, there are four distal enhancers that are required for the faithful tissue- and developmental stage-specific expression of α-globin. The enhancers are situated 10 (Multi-species Conserved Sequence (MCS)-R4), 33 (MCS-R3), 40 (MCS-R2) and 48 (MCS-R1) kb upstream of the α-globin gene promoters (*fig.* 1.2.2). Transgenic experiments have

confirmed that the expression of α–globin diminishes unless the genes are linked to the upstream elements, but the four enhancers differ with respect to their ability to upregulate the expression of α-globin (Hay et al., 2016; Higgs et al., 1990; Sharpe et al., 1992; Sharpe et al., 1993b).

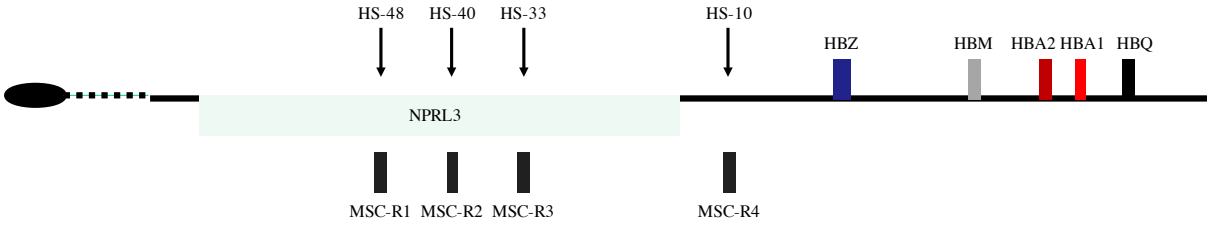


Figure 1.2.2 Schematic representation of the human α-globin locus.
16p telomere is located on the left. Upstream of the α-globin genes are the four enhancers: MCS-R1 to R3 lie in the NPRL3 gene and R4 is downstream of the NPRL3 promoter. The α-globin genes are expressed during human development in the same order in which they positioned along the locus, first ζ-globin gene HBZ (embryonic) and then the two α-globin genes HBA2 and HBA1 (foetal and adult); The minor globin genes μ-globin HBM and θ-globin HBQ do not take part in haemoglobin formation (Higgs et al., 2012).

The human α-globin cluster is located in an early replicating (Dhar et al., 1988; Epner et al., 1981; Goldman et al., 1984), gene dense region of the genome and is surrounded by housekeeping genes (Higgs et al., 2005). The locus is highly GC rich, with multiple CpG islands that are associated with the gene promoters and gene bodies (Fischel-Ghodsian et al., 1987). These CpG islands remain un-methylated in all cell types that have been tested (Bird et al., 1987). The α-globin genes are located in a region of constitutively open chromatin (Craddock et al., 1995) and need to be actively silenced in non-erythroid tissues. The transcription repressive polycomb complex PRC2 (Simon and Kingston, 2013)

and the establishment of H3K27me3 in non-erythroid cells fulfil this function (Garrick et al., 2008). When the α-globin locus is activated in erythroid cells the repression is alleviated by erasing H3K27me3 and eviction of PRC2 (Vernimmen et al., 2011). This is accompanied by an increase in the levels of the chromatin mark H3K4me3 at the α-globin promoters (De Gobbi et al., 2011).

The formation of chromatin accessibility and the binding of key tissue-specific transcription factors occurs independently at the α-globin promoters and the upstream elements (Craddock, et al., 1995; Vernimmen, et al., 2007). However, whereas in the absence of the α-globin promoters, RNA Pol II and components of the transcription machinery are still recruited to the upstream enhancers when the upstream elements are deleted, RNA Pol II recruitment to the globin promoters is severely reduced (Vernimmen et al., 2007). This supports a major role for the α-globin enhancers in recruitment of GTFs and RNA Pol II to the promoters. In pluripotent embryonic stem cells where key hematopoietic transcription factors are not expressed, none of the MCS elements are accessible or bound by any of the factors. By contrast in multipotent hematopoietic progenitors that express early hematopoietic factors MCS-R2 is bound by GATA2, NF-E2, and the TAL1 complex. By the time the cells are fully committed to the erythroid pathway GATA1 has replaced GATA2 at MCS-R2 and the chromatin accessibility over the α-globin promoters has increased (Anguita et al., 2004). Histone modification associated with activation appear in parallel with the binding of transcription factors. These include acetylation of histone H3 and H4 (H3ac and H4ac) with peaks of acetylation associated with the MCS elements and the globin promoters. In addition, mono-methylation of histone H3 at lysine residue 4 (H3K4me1) is enriched at enhancers and tri-methylation of

histone H3 at lysine residue 4 (H3K4me3) is predominantly found at promoters (Anguita et al., 2001; De Gobbi et al., 2007) (see *fig.* 1.2.3 and *tab.* 1.2.1). Even though the chromatin at the α-globin regulatory elements is accessible, expression begins only late in erythropoiesis (Brown et al., 2006) upon terminal maturation when RNA Pol II is recruited. Possible candidates for the onset of activation are the Sp-XKLF factors which are recruited in parallel with RNA Pol II at the locus: Sp1 and Sp3 are recruited to the promoters and KLF1 is recruited to both the promoters and to MCS-R2 (Vernimmen et al., 2007).

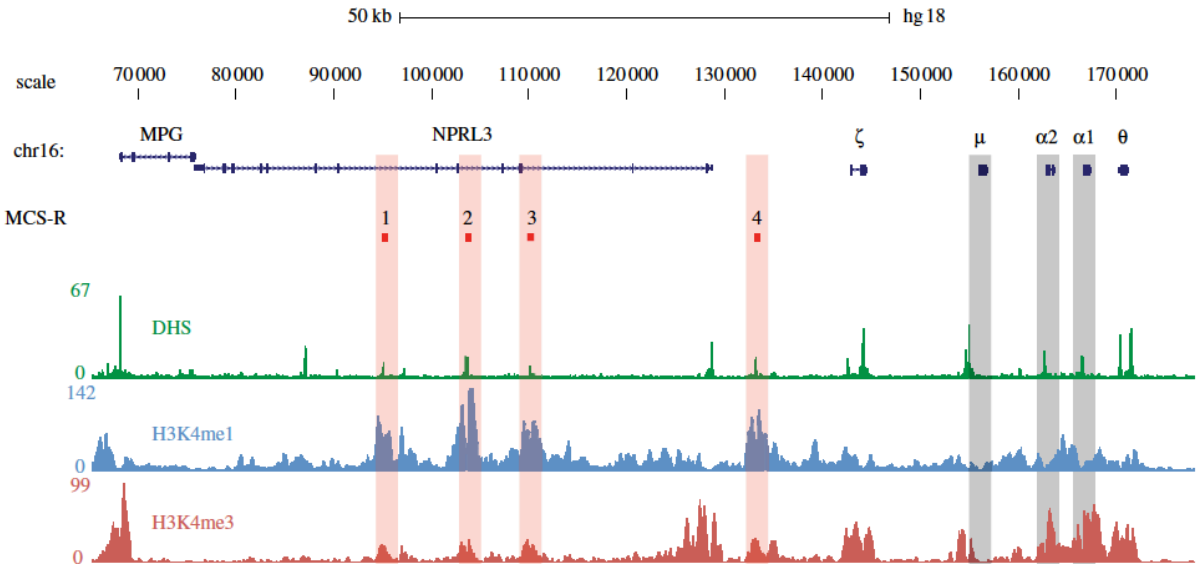


Figure 1.2.3 Chromatin at the cis-acting elements in the human α-globin locus
Tracks for DNaseI-seq are in green and ChIP-seq tracks for H3K4me1 and H3K4me3 are in light blue and in red, respectively. The tracks cover a 100kb region encompassing the α-globin locus on human chromosome 16 (position 70 to 170 kb on genome build hg18). RefSeq genes are displayed in dark blue. The α -globin enhancers (MCS-R1–4) are displayed on the MCS-R track and their alignment to the DNaseI-seq and ChIP-seq tracks is highlighted in red. Figure was adapted from (Hughes et al., 2013).

Activation of the human α-globin genes during erythroid differentiation consists of several key events. Namely, the generation of accessible chromatin over the regulatory regions in the locus and the recruitment of key transcription factors GATA1, KLF1, TAL1 together with E-protein E2A as part of the LDB1 complex and NF-E2 to the upstream enhancers of the α-globin cluster (De Gobbi et al., 2007). The chromatin is marked with modifications associated with transcriptional activation (table 1.2.1). In addition, GTFs and RNA Pol II are recruited to the α-globin promoters associated with physical interaction between enhancers and promoters (Hughes et al., 2013; Hughes et al., 2014; Vernimmen, et al., 2009; Vernimmen, 2014). These events ensure production of the large amounts of globin mRNA required for the proper function of erythroid cells in a stage and tissue-specific manner.

Multispecies conserved sequence Hypersensitive site	MCS-R1 HS -48	MCS-R2 HS -40	MCS-R3 HS -33	MCS-R4 HS -10	α-globin promoter
H4ac	++	++	++	++	++
H3ac	+	+	+	+	++
H3K4me1	++	++	++	++	+
H3K4me3	-	-	-	-	+++
RNA Pol II	+	+	+	+	+++
GATA1	+	+	+	+	-
TAL1	+	+	+	+	-
KLF1	+	+	+	-	+
NF-E2	+	+	-	-	-
Sp1	-	-	-	-	+
Sp3	-	-	-	-	+

Table 1.2.1 Chromatin composition at the human alpha globin locus. Figure was adapted from Higgs et al., 2008.

1.3 α-thalassemia

Haemoglobin is a tetrameric protein comprised of two α-like and two β-like globin chains ($\alpha_2\beta_2$) found in red blood cells. Reduced production of α-globin leads to the generation of a pool of excess β-like globin protein that can organize into tetramers (β_4, known as HbH). The integrity of the two copies of the α-globin genes ensures correct expression and the production of the appropriate amount of α-globin. Naturally occurring mutants of the endogenous human α-globin locus, which cause downregulation of α-globin expression that results in an imbalanced amount of α-globin chains give rise to the common human genetic disease α-thalassemia. The severity of the condition depends on the number of globin genes affected. α^+-thalassemia occurs when a mutation only partially downregulates expression from a chromosome, for example via deletion of a single α-globin gene (-α/). Carriers for α^+-thalassemia (e.g.-α/αα) are referred to phenotypically as "silent carriers". α^0-thalassaemia occurs when the expression from one chromosome is abolished (--/) phenotypically this leads to the α-thalassaemia trait (--/αα) in heterozygotes. Co-inheritance of both α^0 and α^+, i.e. losing expression from three genes (-α/--) leads to HbH disease in which HbH tetramers lead to reduced red blood cell survival. The most severe condition results from loss of all four copies of α-globin (--/--) which causes the Hb Bart's hydrops fetalis syndrome, a condition resulting from the formation of γ-globin tetramers (Haemoglobin Bart's) in the absence of α-chains in the red blood cells of the fetus (Higgs, 2009).

The mutations which cause α-thalassemia can result from deletions or less frequently from single point mutations or nucleotide insertions. α-thalassemia most commonly results from more than 30 large deletions ranging from 3 to over 100 kb, which remove one or both adult α-globin genes (Higgs, 2009). One rarer, small deletion of 970 bp that causes α-thalassemia has helped outline the sequences required for α-globin activation in and around the promoter (Poodt et al., 2006). There is also a number of rare deletions that encompass one or more of the MCS elements but leave the globin genes intact (Viprakasit et al., 2003; Viprakasit et al 2006). MCS-R2 is removed in all of these deletional mutations and there have been α-thalassemia deletions described that remove only MCS-R2 highlighting the importance of this element for the correct activation of globin gene expression (Phylipsen et al., 2010). Occasionally point mutations or nucleotide insertions in critical regions that control gene expression may cause, so-called, non-deletional α-thalassaemia. These mutations are commonly found in splicing and polyadenylation sites, termination codon and /or structural mutations (Higgs, 2009). In addition, a GWAS study which looked at the associations between genomic loci and red blood cell phenotypes has identified variants in MCS-R2 which have an effect on mean cell haemoglobin content (van der Harst et al., 2012). This highlights the fact that regulatory point mutations found in the general population may have an effect on α-globin production.

1.4 Vanuatu Mutation

A study of individuals with α-thalassemia from Vanuatu in Melanesia (De Gobbi et al., 2006) suggested an entirely new pathogenic mechanism affecting expression of the α-globin genes. This type of α-thalassemia was associated with a single-nucleotide mutation in an intergenic region lying between the α-globin genes and their upstream regulatory elements. The mutation creates a new transcriptional unit that appears to downregulate expression of the downstream α-globin genes. We refer to this as the "Vanuatu mutation".

The mutation was identified in a group of patients who had none of the known α-thalassemia mutations. Initially, the trait was linked in *cis* to the α-globin locus via a shared region containing a variable number of tandem repeats (VNTR) site around 8.5 kb away from the α-globin genes. The causal area was narrowed down to 3.7 kb using a tiled array to detect transcripts expressed from the sub-telomeric region of the short arm of chromosome 16. What was striking in the RNA expression profiles of patient primary erythroblasts was the identification of a novel 3.7 kb long region of transcription between 13 to 9.3 kb upstream of the *HBA2* promoter (coordinates 149,682 to 153,390 for genome hg18) (*fig.* 1.4.1). Quantifying the expression of the new transcript using qRT-PCR showed the expression was orders of magnitude higher in the mutant compared to the wildtype chromosome. In addition, expression of the closest gene *HBM* had decreased about 80-fold by qRT-PCR. *HBM* is an erythroid tissue specific gene that is transcribed at

approximately 0.1% of what has been measured for α–globin, but it does not produce any protein (Goh et al., 2005).

The DNA underlying the new transcribed region is poorly conserved and mostly non-coding with the exception of the pseudo gene ψζ *HBZP1* at the tail of the peak. Within that region there were only 7 single nucleotide variants (SNVs) that had not been identified previously in non-thalassemic individuals. Genetic linkage studies on non-thalassemic Melanesian individuals and individuals with this rare form of α-thalassemia trait pinpointed a single T to C transition as the potential causative SNV. Only the C allele of the variant (mutant C or wild type T located at base pair 209,709 of chromosome 16 in human genome assembly hg19; 149,709 for hg18) segregated with the thalassemia trait.

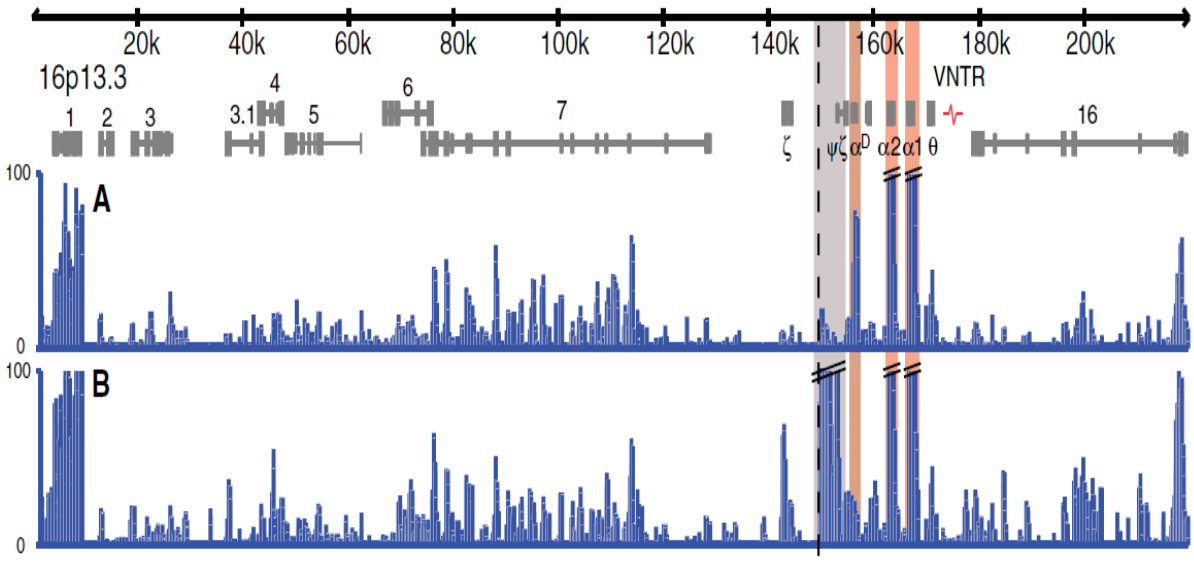


Figure 1.4.1 Measuring transcript abundance in primary erythroid cells using a tiled Affymetrix array covering the 16p sub-telomeric region. At the top, the globin genes are labelled and the VNTR in linkage is shown in red. The causal mutation is indicated with a dashed vertical line. The pattern of gene expression in primary erythroid cells obtained from a non-thalassemic individual homozygous for the wild-type T allele (A) and a patient homozygous for the mutant C allele (B) are displayed in blue. The 3.7kb area of the new transcriptional peak is highlighted in grey; the downstream genes *HBM* (α^D), *HBA2* (α2) and *HBA1* (α1) are highlighted in orange. This figure is adapted from De Gobbi et al., 2006.

In situ RNA hybridization and qRT-PCR was performed on erythroid cells from a patient homozygous for the C allele and a phenotype of HbH disease. This confirmed that the associated α-thalassemia was caused by defects in expression at the stage of transcription. The ratio of α-like to β-like globin transcripts (α/β mRNA ratio) can be used as an approximation for the severity of α-thalassemia and while this was found to be between 0.85 to 1.22 in wild type erythroid cells the patient had an α/β mRNA ratio of 0.46 indicting a severe reduction of α-globin transcription in accordance with the HbH disease phenotype. To dissect the molecular mechanism underlying the phenotype, the orientation of transcription of the new site was determined by strand specific RT-PCR and this showed that transcription proceeded towards *HBM*, in the same orientation as the two α-globin genes.

The mutation itself, a T to C transition, creates a GATA1 motif on the opposite DNA strand, turning the wild type sequence TAATAA (T allele) to TGATAA (C allele). Using electro-mobility gel shift assays for GATA1 binding coupled with an anti-GATA1 antibody it was shown that this GATA1 binds only to DNA that contains the mutation. To determine whether GATA1 binds to the site *in vivo* ChIP-qPCR was performed in patient and wild type primary erythroid cells using qPCR probes across the α-globin cluster. This data shows that GATA-1 occupancy is enriched at the C allele in patient cells (*fig.* 1.4.2). Additional ChIP experiments demonstrated that the C allele is also bound by TAL1 and by association the pentameric erythroid LDB1 complex. Furthermore, in the presence of the C allele, acetylation of histones H3 and H4 increases and RNA Pol II is recruited to the site of the mutation. These results confirmed that the binding of GATA1 at the newly created

motif is associated with the establishment of a new site of transcriptionally active chromatin in the α-globin locus (*fig.* 1.4.2).

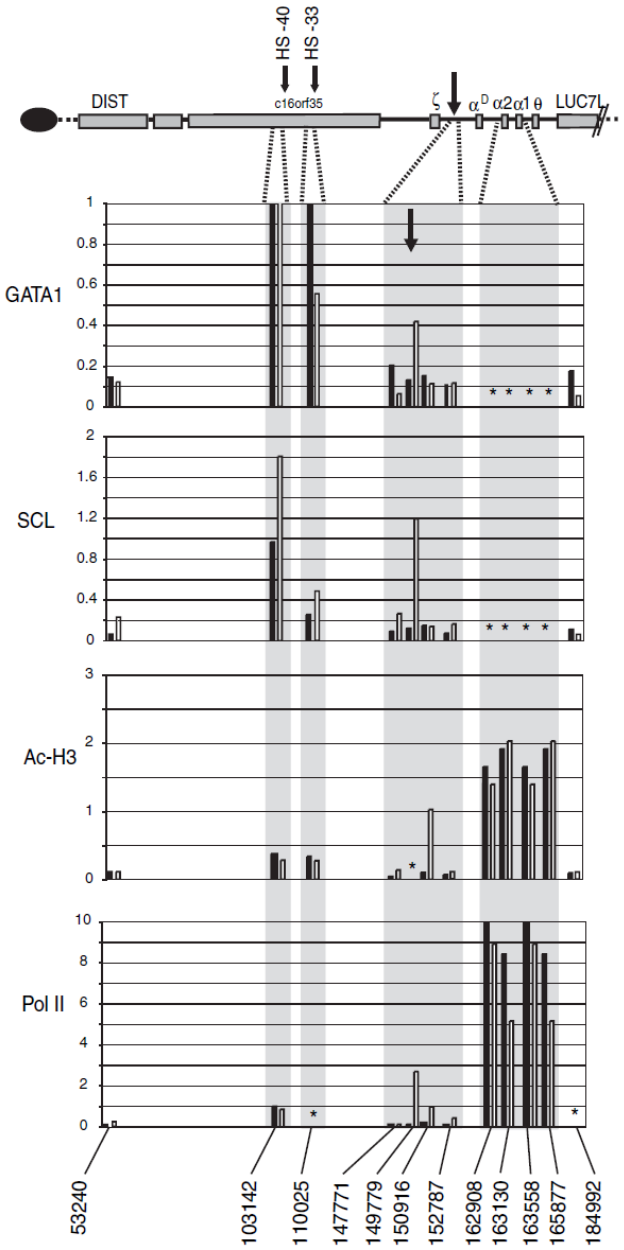


Figure 1.4.2 Chromatin immunoprecipitation of key erythroid transcription factors, PolII and active chromatin marks at the α-globin locus.
Above, the 16p telomeric region is schematically represented: globin genes are labelled along with two of the MSC elements: R2 (HS -40) and R3 (HS -33). Coordinates on hg18 for the qPCR probes are shown on the x axis at the bottom, the new GATA1 binding site is found at coordinate 149,709 indicated with a black arrow on the top. TFs and chromatin modifications are shown on the left. Enrichment in a non-thalassemic individual homozygous for the wild type T allele (black columns) and in a patient homozygous for the mutant C allele (white columns) is shown on the y axis. (Asterisks indicate where no data is available.) Figure was adapted from De Gobbi et al., 2006.

As discussed in section 1.3 of this chapter, the α-globin genes are expressed late in erythropoiesis after a well-defined order of events that involves transcription factors binding to enhancers and recruitment of the transcription machinery associated with the interaction of enhancers and α-globin promoters. Following from this, and the observation that insertion of constitutively active promoters in the transcribing α-globin locus can disrupt the expression of the globin genes (Esperet et al., 2000; Leder et al., 1997), it was argued that this new transcriptional unit might cause globin downregulation in a similar way. It is possible that this might result from preferential interaction between the new transcribed region and the α-globin enhancers by out-competing the interaction with the α-globin promoters. However, technological limitations and more importantly the unavailability of patient material precluded any further exploration of the phenotype which could have definitively proven the causal variant for the new type of α-thalassemia and elucidated the molecular mechanism behind it. Modelling the patient-specific genotype in an erythroid cell line system was the next logical step and is the topic of this thesis. I have chosen to do this using an induced pluripotent stem cell (iPSC)-based erythroid differentiation system.

1.5 Pluripotent stem cells, reprogramming, and the use of iPSCs as a system to model human genetic diseases

For over 10 years the field of iPSCs has been growing rapidly and has produced robust model systems that have given valuable contributions to a variety of areas, including transcription control and cell identity, drug screening and disease modeling. iPSCs combine the valuable traits of a cell line: robust and limitless expansion while maintaining cell identity; the ability to perform genome editing, and clonal expansion. They also have some of the desired traits of primary cells: a more stable karyotype that can be free of exogenous genetic material and the ability to produce a range of cell types that exhibit tissue-specific expression similar to their *in vivo* counterparts. For studying cellular material which is difficult to access due to physiological constraints (e.g. neurons and cardiomyocytes) or designing models for complex diseases human iPSCs have proven to be a valuable resource (Avior et al., 2016; Shi et al., 2017).

Pluripotency and iPSC origin

The term pluripotency refers to the capacity to differentiate into all cell types from tissues present in the embryo. Pluripotent cells can be isolated after the mammalian zygote develops into a blastocyst from a cluster of cells called the inner cell mass (ICM). During gastrulation, the cells from the ICM form the three germ layers from which all the tissues found in the adult organism are established (Alberts, 2008). ICM-derived cells isolated from mouse blastocysts and cultured *in vitro* can give rise to embryonic stem (ES) cells (Evans and Kaufman, 1981; Martin, 1981). The characteristic features of ES cells are

pluripotency and self-renewal - the ability to be expanded indefinitely in an undifferentiated state. ES cell lines have also been generated from human embryos produced by *in vitro* fertilization (Thomson et al., 1998).

Long before the establishment of ES cells, experiments were carried out to revert the differentiation process in adult cells back to a more pluripotent state via a process called reprogramming. Using a technique called somatic cell nuclear transfer (SCNT) pioneered in leopard frogs (Briggs and King, 1952) it was shown that transferring somatic cell nuclei from intestinal cells into an enucleated oocyte could result in the generation of an embryo in *Xenopus laevis* (Gurdon, 1962). Later this was confirmed in a mammalian system by cloning sheep resulting in the generation of an adult animal by SCNT (Wilmut et al., 1997). To remove any doubt that the resulting animals were not produced from cells that had not fully differentiated, mice were cloned from the nuclei of adult lymphocytes. These cells undergo genetic rearrangements of the immunoglobulin genes upon differentiation that can be used to distinguish them from other contaminating cells (Hochedlinger and Jaenisch 2002). Together these experiments unequivocally prove that nuclei from somatic cells can be reprogrammed to pluripotency in the appropriate environment and that the factors that mediate this reprogramming are contained within oocyte. The second important observation for reprogramming studies came from the use of pluripotent cells (embryonal carcinoma, embryonic stem and embryonic germ cells) to reprogram somatic cells by cell fusion (Cowan et al., 2005; Miller and Ruddle, 1976; Tada et al., 1997; Tada et al., 2001). The fusion produces tetraploid hybrid cells that display features characteristic of pluripotency and shows that pluripotent cells, like oocytes, harbour a reprogramming activity capable of reverting differentiation. Subsequent studies indicated that factors

which are present in the nucleus drive the reprogramming process during cell fusion and nuclear transfer (Do and Scholer, 2004; Egli et al., 2007), suggesting that TFs may be involved. Following from the logic that overexpression of key transcription factors could reconfigure the cell transcriptional network through a process called trans-differentiation (Davis et al., 1987; Kulessa et al., 1995; Xie et al., 2004), the first iPSCs were generated by directly reprogramming fully differentiated cells. This was done by overexpressing a defined set of factors: OCT4, SOX2, KLF4 and c-MYC in both human and mouse cells (Takahashi and Yamanaka 2006; Takahashi et al., 2007). The expression of the four transcription factors reactivates endogenous pluripotency genes and initiates the auto-regulatory circuit composed of OCT4, NANOG, and SOX2 that maintains the pluripotent state (Boyer et al., 2005; Takahashi and Yamanaka 2006). The resulting iPSCs exhibited the characteristic features of an ESC, namely pluripotency and self-renewal.

Benefits of iPSCs as a model system

Modelling human diseases *in vitro* can be achieved by several means. First, we will consider the available alternatives to iPSCs with particular emphasis on erythroid model systems.

Disease modelling can in theory be done using primary cells from individuals with the relevant genetic condition. However, for many cell types and or diseases this is not an option due to either limited numbers of suitable donors, limited cell proliferation and/or viability, and particularities of the disease state. For example, primary human mononuclear cells or CD34+ cells isolated from peripheral blood and bone marrow have

been utilized in *in vitro* culture systems supporting erythropoiesis that yield a limited number of mature erythroid cells (Fibach et al. 1993; Dolznig et al. 2005; Giarratana et al. 2005). The limited proliferative capacity of these cells prohibits any complex gene editing strategies that aim to establish a resource of clonal, genotype specific cells and isogenic controls in order to dissect the molecular mechanisms underpinning a phenotype.

Immortalized cell systems have been conventionally used to study a variety of disease mechanisms in part due to their ease of handling and extended proliferative capacity. Cancer studies often rely on tumour-derived cell lines isolated from patients. Additionally, patient cells originating from various types of relevant tissues can be immortalized and used for the modelling of genetic disorders. Recently there have been several studies that have generated immortalized lines from human erythroid cells (Kurita et al., 2013; Huang et al., 2014; Trakarnsanga et al., 2017) which have been an improvement over the commonly used leukemic cell line K562 (Lozzio and Lozzio, 1975). However, one of the issues with using immortalized cell lines is their tendency to have unstable karyotypes. This problem is best exemplified in one of the most commonly used cell lines (HeLa) where different populations of the cell line showed a variable gain and loss of genomic material which drastically affects gene expression and response to stimuli (Frattini et al., 2015). Chromosome instability giving rise to clonal variation becomes an even bigger issue when establishing lines with a particular genotype via gene editing which requires sub-cloning (a difficult problem to tackle which we have observed in our laboratory for some of the previously published immortalized erythroid cell lines).

There is considerable reliance on using animals to study disease. As mammalian genomes are highly evolutionarily conserved, mice have been most commonly used for disease modelling. This is in part because of the relative ease with which gene targeted homologous recombination can be used to precisely manipulate the mouse genome (Skarnes et al., 2011). However, since mice and humans diverged from their common ancestors around 100 million years ago, they have acquired significant genetic, developmental and physiological differences (Strachan, 1997; Mouse Genome Sequencing Consortium et al., 2002; Hamlin and Altschuld, 2011). The dissimilarities can obstruct the recapitulation of the phenotype of a human disease in the mouse. This is particularly true when studying mutations in non-conserved intergenic regions of the genome. There is the option of exchanging the syntenic regions of interest between human and mice to generate a so called humanized mouse model. In order to analyse aspects of gene regulation of human α-globin expression in primary cells, an experimental model was developed in which the human α-globin cluster replaces the mouse α-globin cluster (Wallace et al., 2007). This model has been instrumental in understanding various aspects of regulating the expression of the α-globin locus (De Gobbi et al., 2007; Vernimmen et al., 2009; Vernimmen et al., 2011; Lynch et al., 2012). In this model, the expression of the human α-globin genes is faithfully recapitulated during development, and the patterns of TF binding and the chromatin looping appear to be identical to those in primary human erythroblasts. However, the level of α-globin expression is at about half of what is expected. Additionally, the deletion of the MCS-R2 regulatory element virtually abolishes α-globin expression in the humanized mouse system, whereas in human erythroid cells lack of MCS-R2 is associated with strong downregulation but not complete absence of α-globin transcription (Coelho et al., 2010). It seems likely that this

discrepancy comes from species-specific differences between the mouse transcription factors and/or cofactors and the human *cis*-acting elements. This indicates that a fully equivalent expression of the human α-globin locus in the epigenetic environment of the mouse nucleus may not be achievable.

Human pluripotent stem cells (PSCs) can surpass abovementioned models for studying human diseases, because they combine three major advantages: they can self-renew functionally indefinitely, they have the potential to give rise to any cell type of the body and are essentially normal human primary cell lines. The capacity of human ESCs for multi-lineage hematopoietic differentiation was demonstrated using methods based on existing knowledge from mouse ESC and developmental studies (Kaufman et al., 2001), later more defined methods were developed that produced erythrocytes predominantly displaying characteristics of cells formed early in development (Vodyanik et al., 2005; Olivier et al., 2006; Zambidis et al., 2005; Chang et al., 2006; Qiu *et al., 2008;* Lu, S., et al., 2008). The potential of using ES cells for modelling human diseases via gene editing was utilised shortly after they were first isolated (Zwaka, and Thomson, 2003; Urbach et al., 2004). Human ES cells with wild type genotypes were also utilized as models, for example for the study of chromosomal disorders (Urbach and Benvenisty, 2009). Alternatively, ESCs were isolated from human embryos carrying specific mutations were identified by pre-implantation genetic diagnosis (Eiges et al., 2007; Mateizel et al., 2006). However, only a small group of disorders can be identified this way and for the remaining, gene editing would be necessary to facilitate the generation of disease models in human ESCs. Additionally, these models would be limited to studies where the causative loci for the disease are identified.

Some of the limitations and the ethical issues associated with human ESCs have been overcome by reprogramming somatic, patient-derived cells to pluripotency to generate an easily accessible source of PSCs. iPSCs were quickly utilised to generate human disease models which were used in drug screening studies. One key advantage over ESCs is that the model can be readily available by generating patient specific stem cells, this is especially important in studies of complex genetic diseases, which can often involve various unidentified loci.

Even though the molecular mechanisms that govern reprogramming by defined factors are not completely understood iPSC generation has proven to be extremely robust and versatile. Lines can be derived using a variety of methods (Shi et al., 2017), from multiple somatic cell types (Loh et al., 2009; Takahashi et al., 2007) and a wide age range of donors (Dimos et al., 2008; Takahashi et al., 2007). Even so, the iPSC system is not without its own issues. For example, iPSCs have been shown to accumulate karyotypic abnormalities and copy number variations while maintained in culture (Lund et al., 2012). Residual epigenetic memory from the original cell type that had undergone reprogramming may sometimes persist in the iPSCs and this can potentially affect their differentiation capacity (Bar-Nur et al., 2011; Kim et al., 2011; Ohi et al., 2011). Furthermore, iPSC clones from the same individual have been known to show variations in differentiation efficiency (Inoue et al., 2014). Another issue with iPSCs is that after differentiation the cells commonly exhibit immature characteristics of the respective cell type (Studer et al., 2015). These limitations need to be carefully assessed when using iPSCs as a disease model.

Several studies have reported the successful differentiation of human iPSCs along the erythroid pathway (Chang et al., 2011; Dias et al., 2011; Kobari et al., 2012; Lapillonne et al., 2010; Smith et al., 2013) with varying differentiation efficiency, yield and characteristics of the cells produced. However, almost all of these use feeders and/or serum at some stage of the process which could introduce unwanted variability in the differentiation system. Recently a robust, highly efficient and fully feeder and serum free method for erythroid differentiation of iPSCs was described promising to be of value for basic studies of erythropoiesis (Olivier et al.,2016).

Advancements in gene editing methodologies, and particularly the development of the CRISPR–Cas9 programmable endonuclease system, have accelerated the creation of genetically defined disease models using human iPSCs (Hockemeyer and Jaenisch, 2016). The CRISPR–Cas9 technology is not without its challenges, including efficiency and off-target mutagenesis, but the ability to produce isogenic iPSC lines between which the sole variable is a specific disease-associated mutation has the potential to create a well-defined genetically controlled system. Although, relatively high levels of off-target mutations were initially identified when using CRISPR–Cas9, these were described in immortalized cell lines (Fu et al., 2013). Subsequently, using whole genome sequencing, multiple laboratories have shown that off-targets are rare in primary human cells and this includes human iPSCs (Chang et al., 2015; Li et al., 2016; Park et al., 2015; Smith et al., 2014; Veres et al., 2014).

Currently, iPSCs have been established as an attractive system for disease modelling by being easily accessible, having robust differentiation protocols for many disease-relevant cell types and having access to efficient genome editing tools.

1.6 Summary, Hypothesis and Aim

This thesis studies a single nucleotide genetic variant (T to C transition) found in an intergenic region in the α-globin locus which we have called the Vanuatu mutation. The abnormal C variant was identified in a group of Melanesians who had α-thalassemia but didn't carry any previously known mutation associated with the disease. In these patients, the downregulation of the intact α-globin genes is accompanied by the establishment of a new transcriptional unit which initiates transcription near the position of the variant. It has been previously shown that the C allele generates a new GATA1 binding site, promotes the recruitment of TAL1 and RNA Pol II and causes a local increase in acetylation of histone H3.

I hypothesise that the new unit may use the established regulatory machinery of the α-globin genes and may compete with the α-globin promoters for enhancer activity.

The aim of this thesis is to characterize the Vanuatuan mutation which establishes a new transcriptional unit in the α-globin locus and to understand the molecular mechanisms which downregulate α-globin expression. To do this, there were three main objectives. First, to recapitulate the phenotype of the mutation in a human iPSC based erythroid system which regulates α-globin expression in a manner similar to primary cells and is suitable for molecular experiments requiring large cell numbers. Second, to understand the mechanism regulating transcription of the newly established unit. Third, to prove,

beyond doubt, through the use of gene editing, that the mutation was the causal variant for the transcriptional activation of the new unit and for the α-thalassemia phenotype.

Chapter 3 of the thesis describes the generation of patient specific iPSC from three individuals with common forms of α-thalassemia and two individuals carrying the C allele of the Vanuatu mutation (one homozygous, one heterozygous). Chapter 4 describes the characterization of a human iPSC based erythroid differentiation system by looking at cell morphology, surface markers, chromatin accessibility, transcription factor binding and expression at the α-globin locus, and its suitability to detect α-thalassemia. Chapter 5 illustrates the recapitulation of the phenotype observed in patients with the Vanuatu mutation and further characterises the nature of its transcription by assessing chromatin conformation. Chapter 6 deals with the optimization and establishment of a highly efficient protocol to perform homology directed repair (HDR)-mediated single nucleotide editing in human iPSCs. In chapter 7 the Vanuatu mutation is reverted to wildtype in patient iPSCs to rescue the thalassemia phenotype and wild type iPSC are edited to successfully recreate the Vanuatu mutation and its associated phenotype. In conclusion, I propose that the generation of a site of GATA1 binding at a specific region in the active α-globin locus is sufficient to cause transcriptional activation of a new transcriptional unit. The new GATA1 site appears to create a promoter which competes with the α-globin promoters by interacting with the regulatory elements in the α-globin locus thus reducing α-globin transcription and causing α-thalassemia. Finally, I discuss the possible reasons why this variant produces a promoter-like transcriptional unit and how the findings in this thesis fit within a proposed model of enhancer-promoter interactions.

Chapter 2. Materials and Methods

2.1 Cell culture techniques

2.1.1 Induced Pluripotent Stem Cell Maintenance

Human iPSCs were cultured at 37ºC 5% CO_2 in mTeSR1 medium (StemCell Technologies) on tissue culture treated plastic ware (Corning) coated with Human ESC Qualified Matrigel (Corning) for 2 hours at room temperature. Daily medium changes were performed until cultures were ready for passaging: 75-90% confluency. Cultures were passaged 1 in 8 every 3-4 days using a non-enzymatic method employing a 0.02% EDTA containing solution (Versene from Lonza). Briefly, iPSC colonies were washed with DPBS w/o Ca^{2+}/Mg^{2+}(Invitrogen) and incubated with Versene for 3min at 37ºC 5% CO_2. The Versene solution was then replaced with mTeSR1 medium and the cells were scraped off the dish using a 1 ml pipette tip to produce a suspension of 100-300 cell clumps. The clumps were seeded at the appropriate density onto Matrigel coated dishes in mTeSR1.

Frozen stocks were prepared by producing a suspension of 200-500 cell clumps using Versene and freezing $2x10^6$ cells in 0.5ml of 50% mTeSR1 30% Knock-Out Serum Replacement (Gibco) 10% Knock-Out DMEM (Life Technologies) 10% DMSO (SIGMA) using a Mr. Frosty freezing container (Nalgene) and subsequently stored in liquid Nitrogen vapor. Frozen stocks were resurrected in mTeSR1 medium containing 10µM ROCK inhibitor (RnD) for 24h and cultured on Matrigel coated plates.

Wild type human induced pluripotent stem cells were obtained from Oxford StemBANCC: lines SBAD-02-10, SBAD-03-01; and wild type iPSC line AH017-03 was kindly provided by the

Cader laboratory, WIMM, Oxford. Ethical approval for the study was granted by North West Research Ethics Committee of NHS National Research Ethics Services. The experimental procedures in this study were carried out in accordance with the approved guidelines. Informed consent was obtained for the generation of patient derived iPSC lines.

2.1.2 Reprogramming

Starting material for reprogramming to pluripotency was either EBV immortalized B-cells (a) or CD34+ isolated cells (b).

a) EBV cells were obtained by infecting whole blood with EBV virus and expanding the resulting populations for 3 months in RPMI (Gibco) with 15% FCS (Gibco). Immortalized lines were subsequently cultured in the same medium.

b) MNCs from human peripheral blood were isolated using Ficoll (SIGMA) and stained with ant-CD34 MACS beads (Miltenyi Biotec) at up 10^7cells/100µl staining buffer (PBS 2mM EDTA 0.5% BSA) for 20min on ice. Cells were then washed and separated using LS MACS columns (Miltenyi Biotec) according to manufacturer's protocol. The separated CD34+ cells were then expanded for 7 days in serum-free StemSpan SFEM II (Stemcell technologies) medium containing 100ng/ml Stem Sell Factor (Peprotech), 10ng/ml Interleukin-3 (Peprotech), 0.5IU/ml human recombinant Erythropoietin (Janssen's EPREX), 100U/ml Penicillin/Streptomycin (Gibco) and 2mM Glutamine (Gibco).

For reprogramming the CytoTune-iPS 2.0 Sendai Reprogramming Kit (Life Technologies) was used according to manufacturer's protocol. Briefly, 150,000 cells were infected using the CytoTuneR 2.0 Sendai reprogramming vectors at the appropriate MOI. Cells were incubated at 37°C 5%CO_2 overnight, after 24 hours the medium was replaced to remove the CytoTuneR 2.0 Sendai reprogramming vectors. 72 hours after infection the cells were plated on Mitomycin C inactivated MEF culture dishes in complete StemPRO-34 medium supplemented with GlutaMAX (both from Life Technologies). 7 days after infection the medium was changed to iPSC medium: 20% KnockOut Serum Replacement in KnockOut DMEM/F-12 supplemented with 100 µM MEM Non-Essential Amino Acids Solution, 1x GlutaMAX, 55µM β-mercaptoethanol, 11ng/ml bFGF (all from Life Technologies). From day 7 to 14 spent medium was replaced with a half medium change every day. After day 14 daily full medium changes were made and culture vessels were monitored for the establishment of large iPSC colonies. From day 13 onward Conditioned iPSC medium was used for the medium changes. To prepare conditioned iPSC medium mitotically inactivated MEFs were cultured in iPSC medium for 24h then the medium was removed and supplemented with 11ng/ml bFGF to make complete Conditioned iPSC medium. When iPSC colonies were ready for transfer (day 17-19) they were cut in 200-500 cell patches using a hypodermic needle and seeded onto a Matrigel coated plate in mTeSR1 medium supplemented with 10µM ROCK inhibitor for 24h. Subsequently the cells were maintained for around 2.5 months until they reached passage 15-20 and were stored in cryogenic store as outlined in 2.1.1.

To prepare mitotically inactivated MEFs passage 3 CF1 MEFs (PMEF-CFL from Millipore) were expanded for 3 passages in Advanced DMEM (Gibco) with 10% FCS, 1x GlutaMAX, 55µM β-mercaptoethanol. The cells were then treated with 30µM Mitomycin C for 3 hours at 37°C 5% CO_2 and then dissociated to a single cell suspension using TrypLE Select (Life Technologies). Inactivated MEFs were seeded at 3x10^{4}cells per cm^{2} on Gelatine (SIGMA) coated plates for final use.

2.1.3 Erythroid differentiation of iPSCs (provided by the Mountford Lab; see Olivier et al., 2016)

Prior to the initiation of differentiation protocol, hiPSCs were cultured on Vitronectin (Invitrogen) coated plates (Corning; TC treated) in STEMPRO hESC media (Invitrogen) supplemented with 20ng/ml bFGF (RnD). Cultures were maintained in this way for 5-6 passages at high confluency and split 1 in 4 when over 95% confluent (every 2-3 days) using Versine as previously outlined in 2.1.1.

Day 0. hiPSCs which have been grown to a monolayer on a Vitronectin coated 6 well plate in STEMPRO hESC medium were used as the starting material for differentiation. The monolayer was cut using STEMPRO EZPassage tool (Invitrogen) to small square patches. Embryoid bodies were generated from the patches at 1x10^{5} cells/ml in Stemline II medium (SIGMA) supplemented with 10ng/ml BMP4 (RnD), 10ng/ml VEGF (Peprotech), 10ng/ml Wnt3A

(Peprotech), 5ng/ml Activin A (Peprotech), 2μM Inhibitor VIII (Calbiochem) in Costar Ultra-low Attachment 6 well plates (Corning).

Day 2 (48hr EBs). 48 hours old EBs should be translucent and small and were supplemented with a new set of cytokines: 20ng/ml BMP4, 30ng/ml VEGF, 10ng/ml Wnt3A, 5ng/ml Activin A, 2μM Inhibitor VIII, 10ng/ml αFGF (Peprotech), 20ng/ml SCF (Invitrogen), 0.4ng/ml β-Estradiol (SIGMA)

Day 3 (72hr EBs). 72 hours old EBs were harvested and disaggregated using TrypLE Select then seeded at $1.6x10^4$ cells/cm^2 of a TC treated dish in 0.32ml/cm^2 Stemline II medium supplemented with 20ng/ml BMP4, 30ng/ml VEGF, 10ng/ml αFGF, 30ng/ml SCF, 0.4ng/ml β-Estradiol, 10ng/ml IGF2 (Peprotech), 10ng/ml TPO (Peprotech), 5μg/ml Heparin (SIGMA), 50μM IBMX (SIGMA). Media was supplemented with same cytokines again on day 5.

Day 7. By this stage the cultures had formed small cluster that are budding out non-adherent cells. The suspension cells and adherent clumps were harvested by mechanical force and split 1 in 2 in fresh Stemline II medium supplemented with 20ng/ml BMP4, 30ng/ml VEGF, 10ng/ml αFGF, 30ng/ml SCF, 0.4ng/ml β-Estradiol, 10ng/ml IGF2, 10ng/ml TPO, 5μg/ml Heparin, 50μM IBMX. Media was supplemented with same cytokines again on day 9.

Days 10 to 17. Suspension cells were harvested and counted, then centrifuged and reseeded at a density of up to $1x10^5$ cells/ml in Stemline II supplemented with 1μM Hydrocortisone (SIGMA), 50ng/ml SCF, 16.7ng/ml Flt3L (Peprotech), 6.7ng/ml BMP4, 6.7ng/ml IL3 (Peprotech), 6.7ng/ml IL11 (Peprotech), 3U/ml EPO (Janssen), 50μM IBMX. Samples were taken for FACS (CD34, CD71, CD235a), cytospins, and RNA. Cultures supplemented with

cytokines every 2 days and were split as necessary to keep the density under $1x10^6$ cells/ml, complete media change was performed on day 14.

Days 17-21. Cells were counted, harvested and re-suspended at $3\text{-}4x10^5$ cells/ml in Iscove's Basal Media (Biochrom) media supplemented with 1% BSA (SIGMA), 1x GlutaMAX, 10µg/ml Insulin (SIGMA), 200µg/ml holo-transferrin (SIGMA), 100µM β-mercaptoethanol, 0.5% Lipid Mixture 1 (SIGMA), 0.004% Ethanolamine (SIGMA), 1µM Hydrocortisone, 20ng/ml SCF, 20ng/ml IGF1 (Peprotech), 6.7ng/ml IL3, 6.7ng/ml IL11, 3U/ml EPO. On day 19 cultures were supplemented with the same cytokines and split 1 in 2. Samples were taken for FACS (CD71, CD235a), cytospins, RNA, etc.

2.1.4 Genome editing of iPSCs

Prior to editing, hiPSCs were cultured on Synthemax II-SC (Corning) coated plates (Corning; TC treated) in TeSR-E8 media (StemCell Technologies) supplemented with 20ng/ml bFGF (RnD). Cultures were maintained in this way for 1-2 passages and split 1 in 8 when over 70% confluent (every 3-4 days) using the previously outlined method (Versine).

A single cell suspension was created from one well of a 6 well plate of iPSCs grown in TeSR-E8 on Synthemax II-SC (70-80% confluence) using Accutase (Millipore). Single cells were harvested, re-suspended in mTeSR1 supplemented with 10µM ROCK inhibitor and counted (typically over 90% viable).

A lipofection based method was used to introduce DNA to the iPS cells: 32µl Fugene 6 (Promega) was combined with 270µl mTeSR1 media with 10µM ROCK inhibitor and incubated for 5min at room temperature. Then 4µg of pSpCas9(BB)-2A-Puro (PX459) V2.0 plasmid DNA plus 4µg single stranded DNA oligo donor was added to the Fugene6/medium solution and incubated for 15min at room temperature. Finally, $1x10^6$ iPSCs in single cell suspension in 1.7ml mTeSR1 media with 10µM ROCK inhibitor was combined with the Fugene6/medium/DNA mix and seeded onto a Matrigel coated well of a TC treated 6 well plate. The cells were incubated at 37°C 5% CO_2 for 24h. After 24h the medium was changed to mTeSR1 only and left for 6-8h to clear out the ROCK inhibitor. 30-32 hours post lipofection puromycin selection was initiated with 700-1000ng/ml puromycin (depending on cell line) for 24h. After 24h puromycin containing medium was replaced by mTeSR1 to allow cells to recover. The selected cells were expanded and split 1 in 3 when they reach 50-60% confluency in TeSR-E8 medium on Synthemax II-SC coated dishes using Versene.

Puromycin resistant cells grown in TeSR-E8 medium on Synthemax II-SC were brought to a single cell suspension (Accutase as above) and seeded at 800-1,000 cells per well of a 6 well plate in TeSR-E8 on Synthemax II-SC. After 7-9 days iPSC colonies were ready to be picked into a 96-well plate coated with Matrigel in mTeSR1 medium with 10µM ROCK inhibitor. The 96 well plate was split using Versene once cultures had reached 70-90% confluency to create 2 replica plates. One of the replica plates was used for genotyping (outlined in section 2.3) the other one was used to expand the correct clones in larger dishes and freeze them in liquid nitrogen as outlined in section 2.1.1.

For optimizing DNA delivery Amaxa 4D and Neon were also tested. Amaxa 4D nucelofection (Lonza) was performed using the P3 solutio kit following manufacturer's protocol and using programs CA137 and CB150. 4μg DNA were used for $1x10^6$ cells. Neon electroporation was performed following manufacturer's protocol and using 1400V for a 20ms pulse. 4μg DNA were used for $1x10^6$ cells.Following DNA delivery analysis of GFP positive cells was performed on a Cyan ADP (Beckman Coulter) analyzer using Summit v4.3 and subsequently using Flowing Software for gating on viable single cell events and representing the data as a dot plot

2.2 Cell biology techniques

2.2.1 Flow cytometry

$2x10^5$ cells were washed with PBS (Gibco), resuspended in 200µl of 2% bovine serum albumin (SIGMA) in PBS and labelled for 20 minutes on ice with the following anti-human antibodies: either single stained for allophycocyanin (APC) conjugated anti-CD34 (Miltenyl Biotech; 130-090-654) or double stained for fluorescein isothiocyanate (FITC) conjugated anti-CD71 (BD Biosciences; 555536) and phycoerythrin (PE) conjugated anti-CD235a (BD Biosciences; 340947). After staining cells were harvested and resuspended in PBS containing 0.02% Hoechst 33258 pentahydrate nucleic acid stain (Invitrogen) to stain dead cells. Analysis was performed on a Cyan ADP (Beckman Coulter) analyzer using Summit v4.3 and subsequently using Flowing Software for gating on viable single cell events and representing the data as a dot plot.

2.2.2 Immunofluorescence (IF)

iPSC lines at passage 15-20 were assessed for pluripotency marker expression using the PSC 4-Marker Immunocytochemistry Kit (Invitrogen) as per the manufacturer's protocol. Briefly clones of iPSCs were grown as individual colonies in separate wells of a Matrigel coated 96 well TC treated plate in mTeSR1 medium. The cells were washed with DPBS and fixed using

4% formaldehyde in DPBS, permeabilized using 1% Saponin in DPBS and stained with an anti-OCT4, anti-SOX2, anti-SSEA4 or anti-TRA-1-60 primary antibody in 3% BSA in DPBS for 3 hours at room temperature. The primary antibody stained cells were the stained with an appropriate fluorophore conjugated secondary antibody for 1h at room teperature. The cells were then imaged on a Nikon inverted fluorescence microscope.

2.2.3 Wright stain of cytospins for morphology analysis

Cytospins were prepared by centrifuging $8x10^4$ cells at 400rpm for 4 minutes onto a glass slide. The slides were air-dried for 5min at room temperature and subsequently stained with modified Wright stain using a Hemateck 03564647 slide strainer (BAYER). The slides were then imaged on a Nikon inverted microscope.

2.3 Molecular Biology techniques

2.3.1 RNA isolation and cDNA generation

RNA was isolated from $4x10^6$ cells using the RNeasy kit (Qiagen) according to the manufacturer's protocol. RNA concentration was determined using the Qubit RNA Broad Range assay kit (ThermoFisher Scientific) and RNA purity was assessed using Nanodrop ND-1000 spectrophotometer (ThermoFisher Scientific). RNA was stored at -70^oC for long term storage. Genomic DNA contamination was removed by DNase I treatment using DNase I Amplification Grade kit (ThermoFisher Scientific) on 1000ng of RNA. The equivalent of 800ng of DNA free RNA was reverse-transcribed using the Superscript III First Strand Synthesis kit (Life Technologies) and the cDNA solution was treated with RNase H according to manufacturer's protocol. cDNA was stored at -20^oC for long term storage.

2.3.2 Quantitative real-time polymerase chain reaction (qRT-PCR)

Commercially validated TaqMan assays (Applied Biosystems) were used in qRT-PCR experiments, these are outlined in Table 2.3.2.1. In addition, custom TaqMan assays were used for ChIP qPCR analysis and for detecting the novel transcript produced from the Vanuatu transcription unit (sequences for these are found in table 2.3.2.2). Ribosomal Protein S18 (RPS18) was used as the house keeping gene for normalization as it proved to be constant

throughout the course of differentiation in the iPSC system (see Appendix *fig.* A2.3.2.1 where RPS18 is compared to RPL13a, a previously published reference gene in human primary erythroid cells from Merryweather-Clarke et al., 2011). In all the assays, primers and probes span exon junctions. qRT-PCR was carried out in 10µL reaction volumes using the TaqMan Universal PCR Master Mix (ThermoFisher Scientific) as per manufacturers protocol. The reactions were carried out in triplicate in 96-well plates on the StepOnePlus Real-Time PCR System (ThermoFisher Scientific). Cycling conditions were as follows: 50°C for 2min and 95°C for 10 min followed by 40 cycles of 95°C for 15 seconds and 60°C for 1 min. Quantification of data was made using the delta Ct method where the relative abundance of the gene of interest (Ct of target) compared to a housekeeping control (Ct of reference) can be calculated as $2^{-(\text{Ct of target-Ct of reference})}$.

Table 2.3.2.1 List of TaqMan assays used for qRT-PCR

Target gene ID	**Target gene name**	**TaqMan assay ID**
HBA1/2	α-globin	Hs00361191_g1
HBB	β-globin	Hs00747223_g1
HBD	δ-globin	Hs00426283_m1
HBE1	ε-globin	Hs00362216_m1
HBG1/2	γ-globin	Hs00361131_g1
HBZ	ζ-globin	Hs00923579_m1
RPS18	Ribosomal protein S18	Hs01375212_g1

Table 2.3.2.2 List of custom TaqMan assay used for qRT-PCR

Target	Function	hg16 assay origin	Sequence 5' to 3'
HBA2 Promoter	Forward primer	162908-F	GGGCCGGCACTCTTCTG
	Probe	162908-P	CCCACAGACTCAGAGAGAACCCACCATG
	Reverse primer	162908-R	GGCCTTGACGTTGGTCTTGT
HS -40	Forward primer	103142-F	CAGGCTCCAGGCCCATATC
	Probe	103142-P	TGCCCAAGAGCTCCTTCTGCAACC
	Reverse primer	103142-R	CCTCCTGCACTGTCCTTTGAC
Vanuatu TaqMan	Forward primer	149736-F	GGAGGCGGGACACAAGACTA
	Probe	14973-P	CCTTGCACCAACAGCTTTTCACCTTTCTT
	Reverse primer mutant	149MUT-R	CAAGCCGTTGGAAATACCAAGAG
	Reverse primer WT	149736-R	CAAGCCATTGGAAATACCAAGAG
Negative 14A	Forward primer	82257-F	AACTGCCTTTGCACCCTCAT
	Probe	82257-P	TGTAGGCAGTCCCCAGTCCCAAAGG
	Reverse primer	82257-R	GCTTACCTTTTCTGGGCCTACA
Negative 14B	Forward primer	99295-F	CAGCCAATACAATCAGAACAACTGT
	Probe	99295-P	TTCATGGTATCCTCAAGGCTTGCTGCC
	Reverse primer	99295-R	GTTTGCGACACTGACTGTGTGA

2.3.3 DNA isolation

DNA was isolated using two methods: PureLink Genomic DNA kit (Invitrogen) (a) and Propanol precipitation method (b)

a) Up to $4x10^6$ cells were harvested and used for the PureLink Genomic DNA kit as per manufacturer's protocol

b) iPSC grown to a monolayer were lysed in the culture vessel using a 0.1% SDS solution with 200µg/ml Proteinase K and 40µg/ml RNase A and incubated at 37^oC overnight. Following incubation LiCl was added to 0.8M concentration and one

volume of isopropanol was added to the lysate. The culture vessel was centrifuged at 4°C for 30 minutes at 3,500 rpm, then the liquid was removed and the wells washed with 70% ethanol. The DNA was air-dried, then re-suspended in 0.1x TE buffer and stored at -20°C.

2.3.4 Genotyping

Genotyping was performed by two methods enzyme digestion (a) and Sanger sequencing (b) using a 294bp amplicon (GATAnew) that contains the Vanuatu mutation.

a) 100ng of genomic DNA was used to perform PCR of 294bp encompassing the Vanuatu region (190bp upstream of the mutation site and 103bp downstream). AmpliTaq Gold 360 PCR master mix (Life technologies) was used for the reaction as per manufacturer's protocol for 20µl reactions with cycling conditions as follows: 95°C for 10 minutes then 35 cycles of 95°C for 30 seconds, 61°C for 30 seconds and 72°C for 45 seconds, then a final extension 72°C for 7 minutes. 5µl of the PCR reaction was used for digestion with Hpy188I enzyme in CutSmart buffer (NEB) for 60 minutes at 37°C as per manufacturer's protocol. The digestions were run on 1.5% agarose (SIGMA) gel stained with SYBR Safe (Invitrogen).
b) 100ng of genomic DNA was used to perform PCR of 294bp encompassing the Vanuatu region (190bp upstream of the mutation site and 103bp downstream). Platinum PCR SuperMix (Invitrogen) was used for the reaction as per

manufacturer's protocol for 50µl reactions with cycling conditions as follows: 95°C for 2 minutes then 35 cycles of 95°C for 20 seconds, 61°C for 20 seconds and 72°C for 30 seconds, then a final extension 72°C for 2 minutes. The PCR product was purified using the QIAquick PCR purification kit (Qiagen) as per manufacturer's protocol. Sequencing of PCR product was performed using BigDye Terminator v3.1 chemistry in ABI-3730 DNA analyzer (Applied Biosystems). Sequences were visualized and analyzed using SnapGene software.

Primer pairs for the 294bp amplicon (GATAnew):

Amplicon		**Sequence 5' to 3'**
GATAnew	Forward primer	CATTGTTGAGATGATTTGCTGGAGACACACAGATG
	Reverse primer	GAAATACCAAGAGCTTCCTTGCACCAACAGCT

2.3.5 Assessing edited clones for large indels

Assessing correctly edited homozygous iPSC clones for large mutations which would not be picked up by the short (294bp) genotyping PCR was performed in two ways: a long range (1.8kb) PCR (a) and Southern blot (b)

a) 100ng of genomic DNA was used to perform PCR of 1782bp encompassing the Vanuatu region, adjacent upstream LINE2 element and downstream VNTR (591bp upstream of the mutation site and 1190bp downstream). TaKaRa LA PCR kit

Ver.2.1 (TAKARA BIO) was used for the reaction as per manufacturer's protocol for high GC template. The PCR products were run on 1.5% agarose gel stained with SYBR Safe.

b) 5µg of genomic DNA was digested by PvuII-HF and XbaI in CutSmart buffer (NEB) at 37°C overnight. The digested samples were run on 1% agarose gel on low voltage overnight. After electrophoresis, the gel was denatured for 1hr in1M NaOH and neutralised in 1M Tris, 3M NaCl, pH 7.4 for 1.5hr. DNA was transferred onto Zeta-probe membranes (Biorad) by upward capillary transfer. The membrane was baked for 2hrs in an 80°C oven. The probe was generated as a single stranded DNA oligo and labelled using the Megaprime Labelling Kit (Amersham, GE Healthcare). Hybridisation was carried out in PerfectHyb Plus buffer (SIGMA) at 65°C and the bands were visualized on autoradiography film (GE Healthcare).

Primer pairs for the 1782bp long-range PCR amplicon:

Amplicon		**Sequence 5' to 3'**
LR_GATA_PCR	Forward primer	GCTCCTCTTGAGCTTACATTCTAGTGGACTG
	Reverse primer	CAGCACCAAGTCTTTTCTCCCTGTAGCT

Sequence of the Southern blot probe 5' to 3':

CTTTCTCTCGGAGCCCTTGGAGAGGGGTATGCAAATATCCGTACTCTAAATATCCTCCATATACTGTG
TATTTCCTAAAATCAACAAGGACATTAGGCTGCACAGCCAGAGAACAACCATCAAAATCAGGTTAAT
ATTGATCCAAATCCA

2.3.6 Copy number variation analysis

Copy number variation (CNV) analysis was performed on high quality genomic DNA using the Infinium HD assay on Human CytoSNP 12 Beadchip v2.1 (Illumina) at the Wellcome Trust Centre for Human Genetics. Data was analyzed using the KaryoStudio software package (Illumina) to perform automatic normalisation and identify genomic aberrations utilising the default settings of the built-in cnvPartition algorithm. These parameters are designed to detect CNVs greater than 75 kb and CN-LOH regions larger than 1 mb.

2.3.7 Detection of SeV vectors

The presence of SeV reprogramming vectors from the CytoTune-iPS 2.0 Sendai Reprogramming Kit (Life Technologies) was detected following manufacturer's protocol for RT-PCR based detection. Briefly, a reverse transcription reaction was carried out using 1 µg of RNA and the SuperScript VILO cDNA Synthesis Kit (Life Technologies) following manufacturer's protocol. After RT 10 µL of cDNA were used for a PCR reaction using AccuPrime SuperMix I (Life Technologies) following manufacturer's protocol with cycling conditions as follows: 95°C for 2 minutes then 35 cycles of 95°C for 30 seconds, 55°C for 30 seconds and 72°C for 30 seconds, then a final extension 72°C for 5 minutes. RT-PCR primer sequences and the expected product sizes are found in table 2.3.7.1.

Table 2.3.7.1 List of primer pairs used for SeV vector detection by RT-PCR

Vector	Product size	Primer	Sequence 5' to 3'
SeV	181bp	Forward primer	GGATCACTAGGTGATATCGAGC
		Reverse primer	ACCAGACAAGAGTTTAAGAGATATGTATC
KOS	528bp	Forward primer	ATGCACCGCTACGACGTGAGCGC
		Reverse primer	ACCTTGACAATCCTGATGTGG
Klf4	410bp	Forward primer	TTCCTGCATGCCAGAGGAGCCC
		Reverse primer	AATGTATCGAAGGTGCTCAA
c-Myc	532bp	Forward primer	TAACTGACTAGCAGGCTTGTCG
		Reverse primer	TCCACATACAGTCCTGGATGATGATG

2.3.8 Chromatin Immuno-Precipitation followed by Next Generation Sequencing

$5x10^7$ cells were fixed in 20ml RPMI media (Gibco) with 10% FCS and 1% formaldehyde (SIGMA) for 10min at RT. The fixation reaction was terminated by the addition of glycine to 140mM final concentration. Fixed cells were spun down at 4°C, the supernatant was removed and the cells were washed with cold PBS with Protease inhibitor cocktail (PI) (ROCHE) and snap frozen in dry ice with ethanol.

Fixed cells were lysed at $5x10^7$ cells/ml cell lysis buffer (5mM PIPES (Gibco), 85mM KCl, 0.5% NP-40 (SIGMA)) + proteinase inhibitor (PI) on ice for 20min. The lysed fixed nuclei were spun down at 4°C, the supernatant was removed and the nuclei were re-suspended in nuclear lysis buffer (50mM Tris-HCl, 10mM EDTA (both from Life Technologies) 1% SDS (SIGMA)) +PI at a concentration of $1x10^8$ cells/ml, incubated on ice for 5min. The nuclear lysis was sonicated using the Covaris S220 Focused-ultrasonicator in 130µl microTUBEs. The sonicated lysate was

spun at max speed at 10°C to remove insoluble material. The clear fixed chromatin was diluted 1 in 10 with RIPA buffer w/o SDS +PI (10mM Tris-HCl, 1mM EDTA, 0.5mM EGTA (SIGMA), 1% Triton X-100 (SIGMA), 0.1% Na Deoxycholate (SIGMA), 140mM NaCl) and 1/10 of the diluted chromatin was set aside as input. The diluted chromatin was precleared twice with 10µl of a 1:1 mix of Protein A and Protein G Dynabeads (Invitrogen) by incubating at 4°C for 30mins and discarding the beads. The pre-cleared chromatin was incubated with the appropriate amount of antibody coupled to a 100µl 1:1 mix of Protein A and Protein G Dynabeads at 4°C overnight. Chromatin bound beads were washed twice with RIPA (10mM Tris-HCl, 1mM EDTA, 0.5mM EGTA, 1% Triton X-100, 0.1% SDS, 0.1% Na Deoxycholate, 140mM NaCl) +PI, then twice high salt RIPA (same as RIPA but with 500mM NaCl) +PI, once with LiCl RIPA (same as RIPA but with 250mM LiCl instead of NaCl) +PI and twice with TE buffer. Chromatin was eluted off beads using 20mM Tris-HCl, 5mM EDTA, 50mM NaCl, 1% SDS, 130µg/ml RNase A, 65µg/ml Proteinase K. Eluted chromatin and Input samples were incubated at 68°C overnight to de-crosslink DNA from protein. DNA was isolated by Phenol-Cholorform (SIGMA) extraction as per manufacturer's protocol and precipitate by adding 1 in 500 GlycoBlue co-precipitant (Life technologies), NaAc to 0.3M concentration and 2 volumes of ethanol, followed by centrifugation at maximum speed for 30 minutes at 4°C.

The ChIP library preparation was performed using the NEBNext Ultra DNA Library Preparation Kit for Illumina (NEB) with adaptations for low input DNA following manufacturer's protocol and subsequent size selection. Adapter stocks were diluted 1:10 and primer stocks were diluted 1:10. For size selection after library generation fragments over 500bp were bound to 0.6x AMPure XP Bead solution (Beckman Coulter) and discarded, the remaining fragments

were bound to 1x AMPure XP Bead solution and the DNA was isolated as per manufacturer's protocol. The concentration of resulting libraries was measured using the KAPA library quantification kit (KAPA Biosystems). The libraries were sequenced on the NextSeq platform using High-output 75cycle kits (Illumina), paired end sequencing.

Antibody used for RNA Polymerase II ChIP was the N-20 antibody SC-899 (Santa Cruz Biotechnology) that recognizes all RNAP II phosphoisoforms. Polyclonal anti-KLF1 antibody was kindly provided by the Perkins Laboratory, Translational Research Institute, Brisbane, Australia. Antibody used for H3K4me3 ChIP was the Polyclonal anti-trimethyl-Histone H3 (Lys4) antibody 07-473 (Merck Millipore). GATA1 ChIP was performed using anti-GATA1 antibody ab11852 (Abcam).

2.3.9 Assay for transposase-accessible chromatin sequencing (ATAC-Seq)

ATAC-seq was performed as previously published Buenrostro et al., 2013. $7x10^4$ cells were lysed and nuclei were isolated prior to tagmentation with Tn5 transposase (Illumina) for 30min at 37°C. Tagmented DNA was purified using the MinElute kit (Qiagen) and amplified using the NEBNext 2x Mastermix (NEB) and custom barcoded primers. Library profiles were visualized using a D1000 tape on the Tapestation analyzer (Agilent). The concentration of resulting libraries was measured using the KAPA library quantification kit. The libraries were sequenced on the NextSeq platform using High-output 75cycle kits, paired end sequencing.

2.3.10 Next Generation Capture C (published in Davies et al., 2016)

Next generation capture C was performed as previously published in Davies et al., 2016. Briefly, $1x10^7$ cells were fixed in 2% formaldehyde and lysed in 10 mM Tris 10 mM NaCl 0.2% NP40 +PI. The lysate was digested using NlaIII enzyme in CutSmart buffer (NEB) overnight, the digested chromatin was then ligated using T4 DNA HC ligase (Life technologies) overnight. Ligated chromatin was incubated with 100µg/ml RNase A, 50µg/ml Proteinase K at 65° overnight to de-crosslink DNA from protein. DNA was isolated by Phenol-Cholorform extraction and precipitate by adding NaAc to 0.3M concentration and 2 volumes of ethanol, followed by centrifugation at 3500rpm for 45 minutes at 4°C. 5 µg isolated ligated DNA was sonicated using the Covaris S220 Focused-ultrasonicator in 130µl microTUBEs in the following conditions: Duty cycle: 10%; ntensity: 5; Cycles per burst: 200; Time: 360 seconds. Libraries for capture were generated from the sonicated material using the NEBNext DNA library Prep Reagent Set (NEB) following the manufacturer's protocol up to the addition of adapters then the libraries were indexed using the Herculase II Fusion Polymerase kit (Agilent) and the NEBNext Multiplex Oligos for Illumina Primers (NEB) following the manufacturer's protocol. Capture was performed on pooled indexed libraries using biotinylated DNA probes (Table 2.3). 1.5µg of library DNA was hybridized with 13nmols of probe using the NimbleGen SeqCap EZ Reagent kit (Roche) following manufacturers protocol. Library fragments hybridized to biotinylated oligos were pulled down using M-270 Streptavidin Dynabeads (Life Technologies) following the NimbleGen SeqCap EZ protocol. DNA fragments bound to streptavidin beads

were amplified using the KAPA HiFi HotStart ReadyMix and POST-LM_PCR Oligos 1 and 2 following the NimbleGen SeqCap Library SR protocol. The library DNA was isolated using 1.8x AMPure XP Bead solution following manufacturer's protocol and the capture was repeated a second time. The concentration of resulting libraries was measured using the KAPA library quantification kit. The libraries were sequenced on the NextSeq platform using Mid-output 300cycle kits, paired end sequencing.

2.3.11Preparation of short guide RNA (sgRNA) expression construct.

Plasmid px459 v 2.0 (Addgene id 62988) was a gift from Feng Zhang (Ran et al., 2013). Cloning of sgRNA into plasmid px459 v2.0 was performed using BbsI (NEB) restriction enzyme to cut the vector. Top and bottom DNA oligos containing sgRNA sequence (10µl each of 100µM) were incubated at 99°C for 5minutes and cooled down slowly at room temperature to anneal. Double stranded sgRNA coding DNA was cloned into the pX459 backbone vector by ligation. The reaction was set up in 20µl final volume using T4 ligase (NEB) and 1x buffer with 20ng backbone vector (px459) and 25ng DNA fragment containing sgRNA sequence, incubated at room temperature for 1 hour and transformed in to *Escherichia coli*. Bacteria were plated on to agar plates containing ampicillin. Resistant colonies were picked and grown in liquid culture (Luria-Bertani broth with ampicillin). Plasmid DNA was isolated using QIAGEN EndoFree Plasmid Maxi Kit (Qiagen) according to manufacturer's protocol.

Table 2.3.11.1 List of DNA oligo pairs for sgRNA expression construct cloning into px459 v2.0

sgRNA	Orientation	Sequence 5' to 3'
1t	Forward oligo	caccGCTGTTATTTTCTAATAATG
	Reverse oligo	aaacCATTATTAGAAAATAACAGC
1c	Forward oligo	caccGCTGTTATTTTCTGATAATG
	Reverse oligo	aaacCATTATCAGAAAATAACAGC
3t	Forward oligo	caccGTATTAGAAAATAACAGCAC
	Reverse oligo	aaacGTGCTGTTATTTTCTAATAC
3c	Forward oligo	caccGTATCAGAAAATAACAGCAC
	Reverse oligo	aaacGTGCTGTTATTTTCTGATAC

2.4 Computational analysis

2.4.1 Alignment, normalization and visualization

Paired-end sequencing reads for ATAC-Seq, ChIP-seq and genomic DNA control samples were analysed using an in-house DNase1 and ChIP pipeline (curtesy of J. Telenius, CBRG, WIMM, University of Oxford). Paired-end sequencing reads for NG Capture C were analysed using an in-house Capture C pipeline (curtesy of J. Davies and J. Telenius, WIMM, University of Oxford). Reads were aligned to the human genome build: hg18 for Capture C and hg19 for ATAC and ChIP-seq; using bowtie 2 (version 2.3.2 http://bowtie-bio.sourceforge.net/bowtie2/index.shtml). Bowtie 2 was run with the –m flag set to 2 allowing reads to map twice to the genome in order to account for the high degree of sequence similarity of the duplicated globin genes. To exclude over-amplified PCR products the reads that map to the exact same position in the genome were collapsed into a single representative read. Ploidy regions were removed as they are strongly over-represented in high-throughput sequencing experiments because there is a large difference between the number of copies physically present in the real genome and that number in the genome build. This has the potential to cause difficulties when normalizing the data. Genome-wide tracks of the analysed samples were produced using the in-house pipeline and displayed in the UCSC Genome Browser. Data sets were normalised for the same number of reads (100 million).

Chapter 3: Reprogramming and iPSC quality control

3.1 Introduction

Reprograming differentiated cells to pluripotency requires the expression of OCT4, SOX2, KLF4 and c-MYC to restart the pluripotency network. Fully reprogrammed iPSCs can differentiate to produce cells derived from all three germ layers and in mice can generate fertile individuals using tetraploid complementation (Okita et al., 2007; Zhao et al., 2009). During somatic cell reprogramming the cells undergo global transcriptional and epigenetic changes that produce iPSCs which functionally and molecularly resemble ESCs (Koche et al., 2011; Maherali et al., 2007; Mikkelsen et al., 2008).

Initially, cells that had undergone reprogramming were identified using drug selection that was coupled to the expression of the gene Fbxo15, whose expression is ESCs specific but is not essential for their self-renewal (Takahashi and Yamanaka 2006). The first generation of iPSCs resembled ESCs, however, there were some critical differences. Their transcriptional and epigenetic profiles were not completely reset from the somatic to the ESC state and they could not contribute to the generation of viable chimeric mice, indicating incomplete reprogramming to pluripotency. Subsequent studies, show that endogenous reactivation of OCT4 or NANOG is a better indicator of successful reprogramming, giving rise to cells that resemble ESCs more closely (Maherali et al., 2007; Okita et al., 2007; Wernig et al., 2007). Additionally, allowing enough time for reprogramming to occur before starting selection produces more faithfully reprogrammed cells. This led to experiments demonstrating that when identifying

reprogrammed colonies morphological criteria can be a good guideline to obtain high-quality iPSCs (Blelloch et al., 2007; Maherali et al., 2007; Meissner et al., 2007).

iPSCs which have been faithfully reprogrammed, from a molecular perspective, have epigenetic signatures (histone marks, DNA methylation, reactivation of the silent X chromosome in female cell lines, etc.) and transcriptional profiles that are highly similar to those of ESCs (Maherali et al., 2007; Mikkelsen et al., 2008; Okita et al., 2007; Wernig et al., 2007). Functionally, completely reprogrammed iPSCs can contribute to the germline of chimeric mice and can also produce embryos that are entirely iPSC derived (Hanna et al., 2008; Kim et al., 2008; Wernig et al., 2007). However, it is necessary to point out that human and mouse iPSCs are not equivalent. Mouse ESCs and iPSCs are thought to represent a 'naive' state of pluripotency, resembling cells of the early epiblasts of the mouse pre-implantation embryo (Nichols and Smith, 2009) and have more homogeneous characteristics. In contrast, human ESCs and iPSCs are thought of as being at the 'primed' state, resembling cells of the late epiblast of the post-implantation embryo, and are heterogeneous with respect to both cell population and differentiation capabilities (Hough et al., 2014; Narsinh et al., 2011). Moreover, although iPSCs derived from mice seem to be nearly indistinguishable from mouse ESCs in their pluripotency characteristics (Boland et al., 2009; Kang et al., 2009; Zhao et al. 2009), human iPSCs are different in both expression and epigenetic profiles when compared to human ESCs (Yamanaka, 2012). The genetic background diversity of human iPSCs coupled with the observation that there are also clone-to-clone differences between human ESCs (Osafune et al., 2008), has made the characterization of the pluripotent state of human iPSCs even more difficult. Furthermore, reprogramming is thought to involve rare and

stochastic events which can lead to the generation of partially reprogrammed iPSCs that are prone to defective differentiation (Koyanagi-Aoi et al., 2013; Miura et al., 2009). These variables make the categorization and establishment of standardized human iPSCs difficult and highlights the necessity for the evaluation of pluripotency prior to using human iPSC.

The reprogramming of somatic cells to iPSCs has proven to be a versatile technique. iPSCs have been produced from an assortment of cell types that cover all three germ layers: blood (Hanna et al., 2008); pancreas (Stadtfeld et al., 2008); stomach and liver (Aoi et al., 2008); intestine and adrenals (Wernig et al., 2008); fibroblasts (Lowry et al., 2008; Park et al., 2008; Takahashi et al., 2007) and keratinocytes (Aasen et al., 2008; Maherali et al., 2008); brain (Eminli et al., 2008; Kim et al., 2008; Shi et al., 2008). Furthermore, somatic cells can be converted into iPSCs using different combinations of factors, for example: OCT4, SOX2, NANOG and LIN28, as done by Yu et al., 2007.

Originally, reprogramming methods involved transcription factor delivery via DNA integrating viruses - retroviruses (Takahashi and Yamanaka 2006; Takahashi et al., 2007) because they provided stable and high levels of transgene expression. This was, however, undesirable because of the risks associated with genomic integration of viral sequences (Cavazzana-Calvo et al., 2010; Howe et al., 2008; Ott, M., et al., 2006), namely aberrant activation or disruption of nearby genes, and the incomplete repression or reactivation of the reprogramming vectors. Additionally, it was found that the expression of reprogramming factors is required only transiently, and it subsequently needs to be silenced in the reprogrammed iPSCs to avoid interference with

differentiation (Okita et al., 2007; Okita et al., 2008; Wernig et al., 2007). To avoid these issues efficient, integration-free methods for cell reprogramming were developed to facilitate the transient expression of factors. These methods involve plasmids (Jia et al., 2010; Okita et al., 2008; Okita et al., 2011; Yu et al., 2009), transposons (Kaji et al., 2009; Woltjen et al., 2009; Yusa et al., 2009), adenoviruses (Stadtfeld et al., 2008b), an RNA virus-based delivery system (Sendai) (Fusaki et al., 2009; Nishimura et al., 2010), recombinant proteins (Kim et al., 2009c) or synthetic mRNA (Warren et al., 2010); even a combination of small-molecules can reprogram mouse fibroblasts (Hou et al., 2013). Currently, the use of episomal vectors and Sendai viruses is commonly utilized for the generation of iPSCs that are free of exogenous DNA.

For this project, I chose to use Sendai viral vectors to reprogram human somatic cells to iPSCs using the commercially available CytoTune-iPS 2.0 kit (Invitrogen).

3.2 Reprogramming expanded CD34+ cells and EBV cells using Sendai vectors

The sendai virus (SeV) is an enveloped virus with a single-chain RNA genome that codes for the structural proteins (NP- nucleocapsid protein and M- matrix protein); subunits of RNA polymerase (P and L); hemagglutinin-neuraminidase (HN), which recognizes sialic acid; and fusion protein (F), which mediates the fusion between the viral envelope and the membrane of the cells upon infection (Shioda et al., 1986). There are two main reasons for using SeV for reprogramming. First, they can infect a wide range of cell types from various animal species as the virus invades the cell by interacting with sialic acid that is found on the surface of many cells. Second, the viral vectors exist as RNA, do not go through a DNA intermediate and remain in the cytoplasm, this ensures that they will not integrate into the genome of the host (Fusaki et al., 2009; Nakanishi and Otsu, 2012; Seki et al., 2010). In addition, the vectors combine high transduction efficiency at low multiplicity-of-infection (MOI) with fast and high-level expression of the transgenes. The SeV vectors used in the CytoTune-iPS 2.0 kit consist of viral proteins NP, P, M, F, HN, and L, and the SeV genome RNA that is modified through the deletion of the F gene and the introduction of temperature sensitive mutations in genes coding for other viral proteins (Inoue et al., 2003; Li et al., 2000). These modifications render the vector incapable of producing infectious particles, thus preventing transmission of the virus between cells and in addition, reducing the propagation of the vectors inside daughter cells during division. This ensures that the vectors contained in the cytoplasm of infected cells are diluted out with subsequent divisions and are eventually cleared out in the established iPSC clones. The kit contains three vectors, one expressing OCT4, SOX2, and KLF4

(labelled KOS); one expressing c-MYC (labelled c-Myc); and one expressing additional KLF4 and RNA polymerase (labelled Klf4) to enhance reprogramming efficiency.

The high efficiency of reprogramming using SeV vectors allows the use of a relatively small amount of starting material (150,000 cells) for infection. In this study, the starting material for generating patient-specific iPSCs was either transformed lymphoblastoid (EBV) cells or expanded CD34+ cells. The process of reprogramming (see *fig.* 3.2.1) involves the infection (transduction) of a defined number of somatic cells with reprogramming-factor-expressing Sendai viral vectors at the appropriate MOI. Shortly after the introduction of the vectors, the cells, which are still in suspension, are cultured in dishes that are coated with mitotically inactivated mouse embryonic fibroblasts. As the cells begin to undergo the reprogramming process they become adherent and attach to the feeder layer. The cells undergoing reprogramming are subsequently cultured in a medium that supports the self-renewal and pluripotency of stem cells and within two weeks of infection the first morphologically distinctive iPSC colonies emerge (see *fig.* 3.2.2). On the third week after infection the new iPSC colonies are transferred to feeder-free conditions that promote the expansion of undifferentiated iPS cells. At this stage, morphological criteria are used to select the most suitable colonies for establishing iPSC lines. The colonies must consist of a monolayer with a well-defined edge, they must be compact and must be made of cells that have a high nuclear to cytoplasmic ratio (see *fig.* 3.2.2 Day 18).

Newly established lines were expanded for 15-20 passages to remove the inefficiently replicating viral vectors from the cytoplasm and were cryogenically stored in liquid nitrogen. The clones which could be successfully revived after cryo-store without major signs of differentiation were tested for the expression of pluripotency markers and the

presence of SeV reprogramming vectors and large copy number variations (CNVs) in their genomic content. The most suitable clones were selected and used for subsequent experiments.

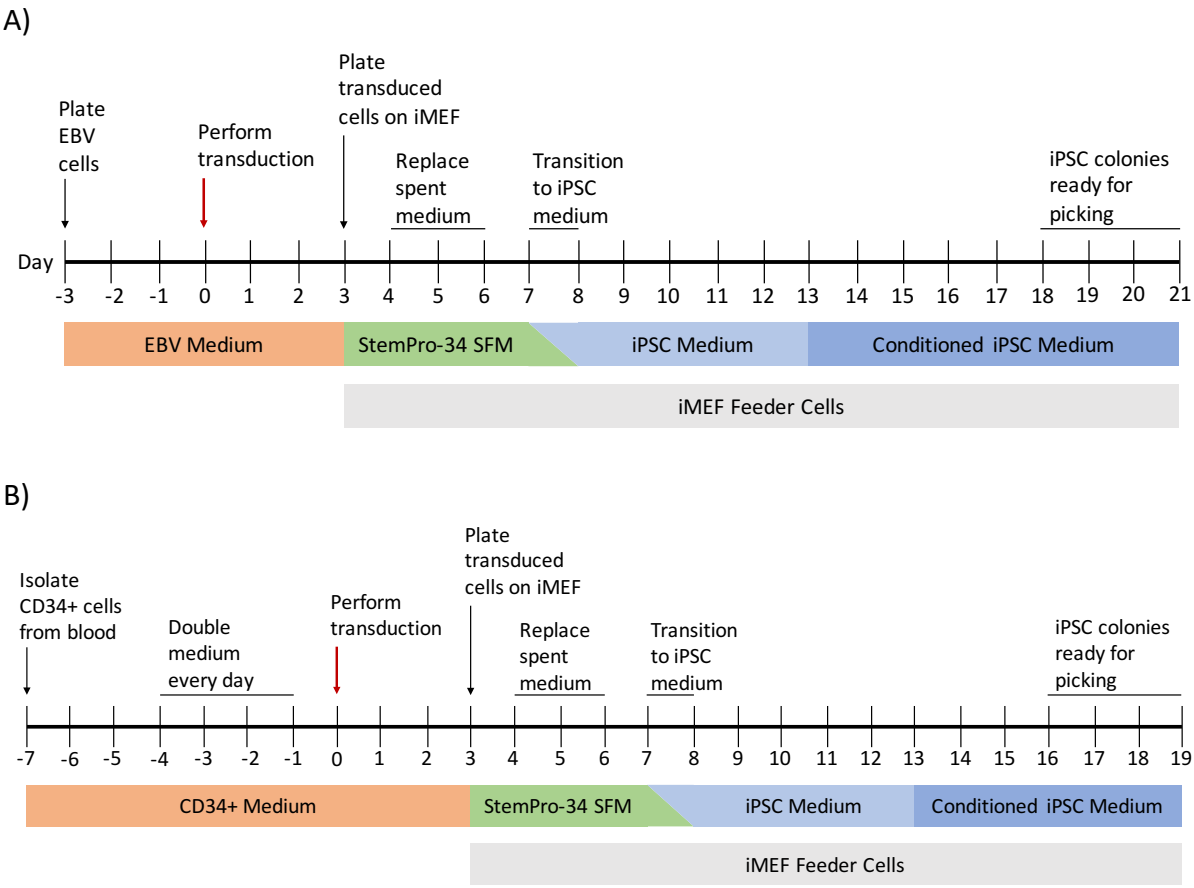
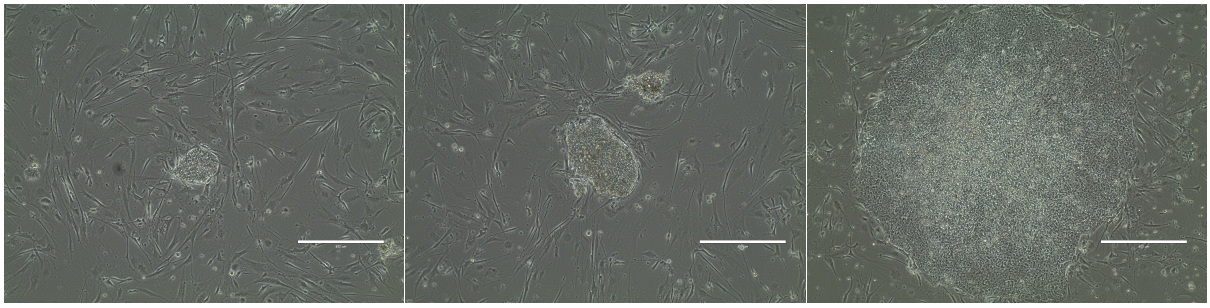


Figure 3.2.1 Reprogramming process for EBV cells and expanded CD34+ cells.
A) Reprogramming of patient EBV cells. EBV cells are cultured in EBV medium for 3 days and then infected with the reprogramming SeV vectors. After 3 more days in EBV medium, the cells are planted onto mitotically inactivated mouse embryonic fibroblasts (iMEF) in StemPro-34 serum free medium (SFM). On day 7 after infection half of the medium is replaced with iPSC medium and fully replaced the next day. From day 13 onward, cells are cultured in conditioned iPSC medium. By day 18 to 21 the iPSC colonies are ready to transfer to Matrigel coated plates and mTeSR1 medium.
B) Reprogramming of CD34 positive cells isolated from patients' peripheral blood. The cells are cultured in CD34+ expansion medium for 7 days and then infected with the reprogramming SeV vectors and expanded further in CD34+ expansion medium. After 3 more days, the cells are planted onto iMEF in StemPro-34 SFM. On day 7 after infection half of the medium is replaced with iPSC medium and fully replaced the following day. From day 13 iPSC colonies are cultured in conditioned iPSC medium. By day 16 to 19 the iPSC colonies are ready to transfer to Matrigel coated plates and cultured in mTeSR1.



Day 13 Day 15 Day 18

Figure 3.2.2 Representative images of an iPSC colony during the last stages of reprogramming until it is ready to be transferred to feeder free conditions (Matrigel coated plates and mTeSR medium). The size of the scale bar is 400μm; the day after infection is marked below the image; the original material used for the reprogramming of the colony depicted in the images was expanded patient CD34+ cells.

The iPSC lines generated for this project included:

- Four iPSC clones from an EBV line that was established from a patient with the Vanuatu phenotype that is homozygous for the C allele. These were labeled: LA01, LA02, LA12, LA13;
- Four iPSC clones from expanded CD34+ cells isolated from a patient with the Vanuatu phenotype that is heterozygous for the C allele. These were labeled: AD01, AD06, AD07, AD13 (clone AD07 displayed high levels of differentiation in late-stage passages and was discarded);
- Four iPSC clones each from expanded CD34+ cells isolated from three patients with α-thalassemia due to a common deletion that removes both copies of the α-globin genes referred to as the Southeast Asian (SEA) deletion (Higgs, 2009). The lines were marked with initials SS, KR, DR and the clones were labeled: SS03, SS06, SS07, SS09; KR01, KR03, KR08, KR12; DR01, DR06, DR07, DR16;

3.3 CNV analysis

Pluripotent stem cells can acquire genomic abnormalities when grown in culture. These abnormalities can range in size from full chromosome aneuploidy to sub-chromosomal aberrations, including point mutations (reviewed in Weissbein et al., 2014). Karyotypic abnormalities are commonly thought to arise as a consequence of adaptation to culturing conditions driven by positive selection (Baker et al., 2007; Draper et al., 2004). With this in mind, a positive correlation has been observed between passage number and karyotype abnormalities, however, abnormalities can also be found in cultures with a low low-passage number (Mayshar et al., 2010; Taapken et al., 2011). It is important to note that late-passage PSCs commonly harbour only a few types of aneuploidies and this implies that most chromosomal aberrations are under negative selection in culture. It has been suggested that the driving force behind the competitive advantage that specific aneuploidies have in culture conditions is abnormal gene expression (Baker et al., 2007; Ben-David and Benvenisty, 2012; Blum et al., 2009; Mayshar et al., 2010). During reprogramming chromosomal aberrations that are on a mega-base scale and smaller kilobase scale copy-number alterations can arise *de novo* or be amplified from a small population of parental somatic cells. This is exemplified by DNA array analysis which shows that low-passage human iPSC lines tend to harbour a larger amount of CNVs when compared to the parental cells from which they were established and also when compared to their later-passage iPSCs counterparts (Hussein et al., 2011; Laurent et al., 2011). This observation draws a parallel to what has been speculated for larger karyotype abnormalities. It is possible that the smaller CNVs tend to be disadvantageous

and are commonly selected against in culture. Furthermore, it should be noted that there are many naturally occurring CNVs and they are generally considered benign. Relatively large duplications and deletions are common in the general population with around 70% of healthy individuals found to harbour CNVs of at least 100kb (Itsara et al., 2009), indicating that only a small number of specific aberrations would be of potential concern. The commonly observed chromosomal aberrations in human PSCs are gains of chromosome 1, 12, 17, 20 and X and amplification of chromosome segments which contain genes that may provide growth advantage in culture, namely 20q11.21 and 12p13.31 (reviewed in Weissbein et al., 2014). Another form of structural variation that is found in the genome and can alter gene expression is copy-neutral loss of heterozygosity (CN-LOH). CN-LOH is characterized by a region of homozygosity that covers a part of a chromosome. If there are alleles in the affected region that contain recessive mutations or are part of a locus that is subject to imprinting, this may cause net changes in expression (Kearney et al., 2011). LOH would be normally considered inherited if found in iPSCs, however, the acquisition of LOH has been observed during the progression of many cancers (Makishima and Maciejewski, 2011). This suggests that in characterising iPSCs it is necessary to screen for CN-LOH events and that LOH regions may require further investigation.

While standard karyotyping is sufficient to establish the overall integrity of the genome, higher-resolution CNV analysis of newly generated iPSC colonies is also required to identify chromosomal abnormalities that may affect gene expression and potentially interfere with the maintenance of iPS cell identity, and differentiation potential. Between 5–10% of the genome of a normal individual contains CNVs averaging a few

kilobases (kb) in length (Zarrei et al., 2015) and very high-resolution arrays would generate large data sets that predominantly contain these small naturally occurring CNVs (Bernardini et al., 2010; Redon et al., 2006). For this reason, the HumanCytoSNP-12 array from Illumina was chosen to detect the presence of CNVs and LOH in the newly generated patient iPSCs after 2.5 months of culture (19-20 passages). This array identifies CNVs that are larger than 75 kb and LOH over 1 MB in size. This resolution is sufficient to perform general cytogenetic analysis of human iPSC using the KryoStudio package from Illumina's GenomeStudio software. The summary of CNV regions identified in the iPSC clones studied here can be found in table 3.3.1.1. The homozygous mutant Vanuatu patient iPSC clones LA01, -02, -12, -13 all share three CN-LOH regions: chromosome 1q31.1-3; chromosome 11p11.12-2; chromosome Xp11.1-22. These are quite likely to be inherited and can be explained by the fact that the patient comes from an isolated region with a small founding population. Two of the clones (LA02 and LA12) contained a 1.45 MB duplication in chromosome 16p13.3 which encompasses the α-globin gene locus. These clones were discarded because the duplication will confound the interpretation of gene expression experiments. Clones LA01 and LA13 were used for downstream experiments. From the heterozygous mutant Vanuatu patient, iPSC clones AD01, -06, -13 only clone AD01 did not have any large CNVs, the other two clones share a duplication in chromosome 9q34.13 which could have been present in the parental population and have been amplified during the expansion of the isolated CD34+ cells or it could have arisen during the very early stages of reprogramming. Clone AD13 has an additional duplication in chromosome 20q11.21 that has most likely arisen in culture. Only clone AD01 was used for experiments. The α-thalassemia lines all share a duplication in chromosome

10q11.21 and two of the lines (SS and DR) share a duplication in chromosome 5q35.3 which is not surprising since the donors belong to the same family. The 10q11.21 duplication has a frequency of 0.05% in the general population and the 5q35.3 occurs in 0.07% of healthy individuals according to the Database of Genomic Variants (MacDonald et al., 2014). This strongly suggests that these CNVs are of parental origin and are not likely to interfere with experimental results. Only lines SS06 and DR06 were discarded as they contained novel CNVs in chromosome 6p12.1 and 3p14.2 respectively.

Line	Chr	Cytobands	Start	Stop	Length (bp)	Type	Genes affected
LA01	1	q31.1 q31.2 q31.3	189,752,408	198,117,642	8,365,234	LOH	FAM5C; RGS18; RGS21; RGS1; RGS13; RGS2; UCHL5; TROVE2; TROVE2; GLRX2; GLRX2; CDC73; MIR1278; B3GALT2; KCNT2; CFH; CFHR3; CFHR1; CFHR4; CFHR2; CFHR5; F13B; ASPM; ZBTB41; CRB1; DENND1B; DENND1B; C1orf53; LHX9; LHX9;
	11	p11.2 p11.12	45,173,290	51,530,241	6,356,951	LOH	PRDM11; SYT13; CHST1; DKFZp779M0652; SLC35C1; SLC35C1; CRY2; CRY2; MAPK8IP1; C11orf94; PEX16; GYLTL1B; PHF21A; CREB3L1; DGKZ; DGKZ; DGKZ; DGKZ; MDK; MDK; MDK; CHRM4; AMBRA1; HARBI1; KIAA0652; ARHGAP1; ZNF408; F2; CKAP5; SNORD67; LRP4; C11orf49; ARFGAP2; PACSIN3; DDB2; ACP2; ACP2; NR1H3; NR1H3; MADD; MADD; MADD; MYBPC3; SPI1; SLC39A13; PSMC3; RAPSN; CUGBP1; PTPMT1; KBTBD4; NDUFS3; FAM180B; C1QTNF4; MTCH2; AGBL2; FNBP4; NUP160; PTPRJ; OR4B1; OR4X2; OR4X1; OR4S1; OR4C3; OR4C45; OR4A47; FOLH1; LOC440040; OR4C13; OR4C12; LOC441601; LOC646813;
	X	p11.22 p11.21 p11.1	54,774,832	58,339,545	3,564,713	LOH	ITIH5L; MAGED2; MAGED2; MAGED2; SNORA11; TRO; PFKFB1; APEX2; ALAS2; PAGE2B; PAGE2; FAM104B; FAM104B; PAGE5; PAGE5; PAGE3; MAGEH1; USP51; FOXR2; RRAGB; KLF8; KLF8; UBQLN2; DKFZp686L07201; LOC442454; SPIN3; SPIN3; SPIN2B; SPIN2A; FAAH2; ZXDB; ZXDA;
LA02	16	p13.3	105,320	1,560,111	1,454,791	Duplication	SNRNP25; RHBDF1; MPG; MPG; C16orf35; HBZ; HBM; HBA2; HBA1; HBQ1; LUC7L; LUC7L; ITFG3; RGS11; ARHGDIG; PDIA2; AXIN1; MRPL28; TMEM8A; LOC100134368; NME4; DECR2; RAB11FIP3; RAB11FIP3; C16orf10; SOLH; C16orf11; FLJ36208; PIGQ; RAB40C; WFIKKN1; C16orf13; FAM195A; WDR90; RHOT2; RHBDL1; STUB1; JMJD8; WDR24; FBXL16; METRN; FAM173A; CCDC78; HAGHL; HAGHL; NARFL; MSLN; MSLNL; MIR662; RPUSD1; CHTF18; GNG13; PRR25; LMF1; SOX8; LOC146336; SSTR5; C1QTNF8; CACNA1H;
	1	q31.1 q31.2 q31.3	189,752,408	198,117,642	8,365,234	LOH	FAM5C; RGS18; RGS21; RGS1; RGS13; RGS2; UCHL5; TROVE2; TROVE2; GLRX2; GLRX2; CDC73; MIR1278; B3GALT2; KCNT2; CFH; CFHR3; CFHR1; CFHR4; CFHR2; CFHR5; F13B; ASPM; ZBTB41; CRB1; DENND1B; DENND1B; C1orf53; LHX9; LHX9;
	11	p11.2 p11.12	45,173,290	51,530,241	6,356,951	LOH	PRDM11; SYT13; CHST1; DKFZp779M0652; SLC35C1; SLC35C1; CRY2; CRY2; MAPK8IP1; C11orf94; PEX16; GYLTL1B; PHF21A; CREB3L1; DGKZ; DGKZ; DGKZ; DGKZ; MDK; MDK; MDK; CHRM4; AMBRA1; HARBI1; KIAA0652; ARHGAP1; ZNF408; F2; CKAP5; SNORD67; LRP4; C11orf49; ARFGAP2; PACSIN3; DDB2; ACP2; ACP2; NR1H3; NR1H3; MADD; MADD; MADD; MYBPC3; SPI1; SLC39A13; PSMC3; RAPSN; CUGBP1; PTPMT1; KBTBD4; NDUFS3; FAM180B; C1QTNF4; MTCH2; AGBL2; FNBP4; NUP160; PTPRJ; OR4B1; OR4X2;
	X	p11.22 p11.21 p11.1	54,774,832	58,339,545	3,564,713	LOH	ITIH5L; MAGED2; MAGED2; MAGED2; SNORA11; TRO; PFKFB1; APEX2; ALAS2; PAGE2B; PAGE2; FAM104B; FAM104B; PAGE5; PAGE5; PAGE3; MAGEH1; USP51; FOXR2; RRAGB; KLF8; KLF8; UBQLN2; DKFZp686L07201; LOC442454; SPIN3; SPIN3; SPIN2B; SPIN2A; FAAH2; ZXDB; ZXDA;
LA12	16	p13.3	204,500	1,560,111	1,355,611	Duplication	HBZ; HBM; HBA2; HBA1; HBQ1; LUC7L; LUC7L; ITFG3; RGS11; ARHGDIG; PDIA2; AXIN1; MRPL28; TMEM8A; LOC100134368; NME4; DECR2; RAB11FIP3; RAB11FIP3; C16orf10; SOLH; C16orf11; FLJ36208; PIGQ; RAB40C; WFIKKN1; C16orf13; FAM195A; WDR90; RHOT2; RHBDL1; STUB1; JMJD8; WDR24; FBXL16; METRN; FAM173A; CCDC78; HAGHL; HAGHL; NARFL; MSLN; MSLNL; MIR662; RPUSD1; CHTF18; GNG13; PRR25; LMF1; SOX8; LOC146336; SSTR5; C1QTNF8; CACNA1H; TPSG1; TPSB2; TPSAB1; TPSD1; UBE2I;
	1	q31.1 q31.2 q31.3	189,752,408	198,117,642	8,365,234	LOH	FAM5C; RGS18; RGS21; RGS1; RGS13; RGS2; UCHL5; TROVE2; TROVE2; GLRX2; GLRX2; CDC73; MIR1278; B3GALT2; KCNT2; CFH; CFHR3; CFHR1; CFHR4; CFHR2; CFHR5; F13B; ASPM; ZBTB41; CRB1; DENND1B; DENND1B; C1orf53; LHX9; LHX9;
	11	p11.2 p11.12	45,173,290	49,357,043	4,183,753	LOH	PRDM11; SYT13; CHST1; DKFZp779M0652; SLC35C1; SLC35C1; CRY2; CRY2; MAPK8IP1; C11orf94; PEX16; GYLTL1B; PHF21A; CREB3L1; DGKZ; DGKZ; DGKZ; DGKZ; MDK; MDK; MDK; CHRM4; AMBRA1; HARBI1; KIAA0652; ARHGAP1; ZNF408; F2; CKAP5; SNORD67; LRP4; C11orf49; ARFGAP2; PACSIN3; DDB2; ACP2; ACP2; NR1H3; NR1H3; MADD; MADD; MADD; MYBPC3; SPI1; SLC39A13; PSMC3; RAPSN; CUGBP1;
	X	p11.22 p11.21 p11.1	54,774,832	58,339,545	3,564,713	LOH	ITIH5L; MAGED2; MAGED2; MAGED2; SNORA11; TRO; PFKFB1; APEX2; ALAS2; PAGE2B; PAGE2; FAM104B; FAM104B; PAGE5; PAGE5; PAGE3; MAGEH1; USP51; FOXR2; RRAGB; KLF8; KLF8; UBQLN2; DKFZp686L07201; LOC442454; SPIN3; SPIN3; SPIN2B; SPIN2A; FAAH2; ZXDB; ZXDA;
LA13	1	q31.1 q31.2 q31.3	189,752,408	198,117,642	8,365,234	LOH	FAM5C; RGS18; RGS21; RGS1; RGS13; RGS2; UCHL5; TROVE2; TROVE2; GLRX2; GLRX2; CDC73; MIR1278; B3GALT2; KCNT2; CFH; CFHR3; CFHR1; CFHR4; CFHR2; CFHR5; F13B; ASPM; ZBTB41; CRB1; DENND1B; DENND1B; C1orf53; LHX9; LHX9;
	11	p11.2 p11.12	45,173,290	51,530,241	6,356,951	LOH	PRDM11; SYT13; CHST1; DKFZp779M0652; SLC35C1; SLC35C1; CRY2; CRY2; MAPK8IP1; C11orf94; PEX16; GYLTL1B; PHF21A; CREB3L1; DGKZ; DGKZ; DGKZ; DGKZ; MDK; MDK; MDK; CHRM4; AMBRA1; HARBI1; KIAA0652; ARHGAP1; ZNF408; F2; CKAP5; SNORD67; LRP4; C11orf49; ARFGAP2; PACSIN3; DDB2; ACP2; ACP2; NR1H3; NR1H3; MADD; MADD; MADD; MYBPC3; SPI1; SLC39A13; PSMC3; RAPSN; CUGBP1; PTPMT1; KBTBD4; NDUFS3; FAM180B; C1QTNF4; MTCH2; AGBL2; FNBP4; NUP160; PTPRJ; OR4B1; OR4X2;
	X	p11.22 p11.21 p11.1	54,774,832	58,339,545	3,564,713	LOH	ITIH5L; MAGED2; MAGED2; MAGED2; SNORA11; TRO; PFKFB1; APEX2; ALAS2; PAGE2B; PAGE2; FAM104B; FAM104B; PAGE5; PAGE5; PAGE3; MAGEH1; USP51; FOXR2; RRAGB; KLF8; KLF8; UBQLN2; DKFZp686L07201; LOC442454; SPIN3; SPIN3; SPIN2B; SPIN2A; FAAH2; ZXDB; ZXDA;
AD01	no large CNVs						
AD06	9	q34.13	134,176,997	134,326,950	149,953	Duplication	PPAPDC3; BAT2L;
AD13	9	q34.13	134,176,997	134,326,950	149,953	Duplication	PPAPDC3; BAT2L;
	20	q11.21	29,843,182	30,523,480	680,298	Duplication	DEFB115; DEFB116; DEFB118; DEFB119; DEFB119; DEFB121; DEFB122; DEFB123; DEFB124; REM1; NCRNA00028; HM13; PSIMCT-1; ID1; COX4I2; BCL2L1; TPX2; MYLK2; FOXS1; DUSP15; TTLL9;
SS03	5	q35.3	178,722,356	179,014,740	292,384	Duplication	ADAMTS2; ADAMTS2; RUFY1; RUFY1;
	10	q11.21	42,810,783	42,929,705	118,922	Duplication	LOC441666;
SS06	5	q35.3	178,728,830	178,997,692	268,862	Duplication	ADAMTS2; ADAMTS2; RUFY1; RUFY1;
	6	p12.1	56,733,052	56,920,262	187,210	Duplication	BEND6; KIAA1586;
	10	q11.21	42,810,783	42,929,705	118,922	Duplication	LOC441666;
SS07	5	q35.3	178,722,356	178,997,692	275,336	Duplication	ADAMTS2; ADAMTS2; RUFY1; RUFY1;
	10	q11.21	42,810,783	42,929,705	118,922	Duplication	LOC441666;
SS09	5	q35.3	178,722,356	178,997,692	275,336	Duplication	ADAMTS2; ADAMTS2; RUFY1; RUFY1;
	10	q11.21	42,810,783	42,929,705	118,922	Duplication	LOC441666;
KR01	10	q11.21	42,810,783	42,929,705	118,922	Duplication	LOC441666;
KR03	10	q11.21	42,810,783	42,929,705	118,922	Duplication	LOC441666;
KR08	10	q11.21	42,810,783	42,929,705	118,922	Duplication	LOC441666;
KR12	10	q11.21	42,810,783	42,929,705	118,922	Duplication	LOC441666;
DR01	5	q35.3	178,722,356	179,014,740	292,384	Duplication	ADAMTS2; ADAMTS2; RUFY1; RUFY1;
	10	q11.21	42,810,783	42,929,705	118,922	Duplication	LOC441666;
DR06	3	p14.2	60,853,686	61,119,923	266,237	1 copy deletion	FHIT;
	5	q35.3	178,722,356	178,931,310	208,954	Duplication	ADAMTS2; ADAMTS2;
	10	q11.21	42,810,783	42,929,705	118,922	Duplication	LOC441666;
DR07	5	q35.3	178,722,356	178,997,692	275,336	Duplication	ADAMTS2; ADAMTS2; RUFY1; RUFY1;
	10	q11.21	42,810,783	42,947,627	136,844	Duplication	LOC441666;
DR16	5	q35.3	178,728,830	179,014,740	285,910	Duplication	ADAMTS2; ADAMTS2; RUFY1; RUFY1;
	10	q11.21	42,810,783	43,047,544	236,761	Duplication	LOC441666; LOC84856; ZNF37B;

Table 3.3.1.1 Large CNVs in human iPSC lines generated by reprogramming patient somatic cells. The table presents the copy number variation found in iPSC lines generated for this thesis. The first column contains the name of the line; the second is the chromosome in which the CNV is found; the third column is the cytoband/s which the CNV encompasses; the fourth and fifth column are the start and stop positions of the CNV region in human genome hg18; the sixth column indicates the type of mutation: LOH – loss of heterozygosity; Duplication - gain of one copy; 1 copy deletion - loss of one copy; the last column lists the genes affected by the abnormality.

3.4 Pluripotency markers and vector detection

It is important to carefully characterize iPSCs, because the efficiency of the reprogramming process is low and the quality of the generated cells depends on many factors. A variety of methods for the characterization of iPSCs have been developed. Commonly, the first indication of iPSC formation is the characteristic morphological features of iPSCs: cells which are small, have a large nucleus/cytoplasm ratio and form compact colonies that are defined by clear borders. There are many cellular and molecular methods that can be used in combination with morphology to further characterise these cells. One involves the detection of pluripotency marker proteins that are expressed in PSCs (Boulting et al., 2011; Brouwer et al., 2016). Because the markers are not necessarily expressed only in PSCs, a combination of multiple markers needs to be used. For the purpose of this thesis IF was used to determine the expression of OCT4 and SOX2, which together with NANOG form the core pluripotency network (Boyer et al., 2005). Tis was combined with the detection of PSC surface markers SSEA4 and TRA-1-60 was assessed (Nagano et al., 2008).

3.4.1 Immunofluorescence (IF) for pluripotency markers

Patient iPSC lines grown for 15-20 passages were fixed with formaldehyde and permeabilized as colonies onto tissue culture plates. The fixed cells were first stained with primary antibodies against OCT4, SOX2, SSEA4 and TRA-1-60 and then with appropriate fluorophore conjugated secondary antibodies (see Chapter 2). The fluorescent anybody labelled iPSC colonies were imaged using an inverted fluorescence microscope (*fig.* 3.4.1.1). All of the reprogrammed patient cell lines used for the work in this thesis express pluripotency markers characteristic of iPSCs: OCT4, SOX2, SSEA4 and TRA-1-60.

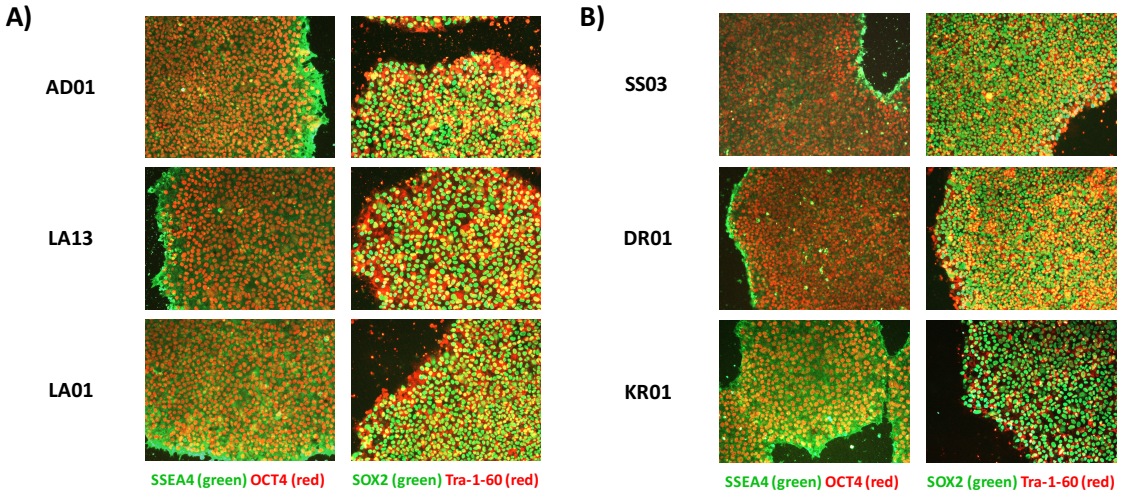


Figure 3.4.1.1 Immuno-fluorescence staining for pluripotency markers.
A) Pluripotency marker IF for iPSC lines AD01, LA13 and LA01. The first column shows overlaid staining for SSEA4 in green and OCT4 in red, the second column shows overlaid staining for SOX2 in green and Tra-1-60 in red; B) Pluripotency marker IF for lines SS03, DR01 and KR01. The first column shows overlaid staining for SSEA4 in green and OCT4 in red, the second column shows overlaid staining for SOX2 in green and Tra-1-60 in red.

3.4.2 RT-PCR for Vector DNA

The expression of reprogramming factors must be silenced in established iPSC lines in order to avoid interference with differentiation. When employing the non-integrative Sendai viral system this means that the vectors have to be cleared from the cytoplasm of the iPSCs before the lines can be used. This is accomplished by expanding the line for 15 to 20 passages as the replication inefficient SeV vectors cannot keep up with successive cellular division cycles. In iPSC lines (passage 15-20) that were selected for downstream experiments the presence of the SeV vectors was assessed by RT-PCR using primer pairs to detect each individual vector and one pair for the SeV backbone (*fig.* 3.4.2.1) (see Chapter 2 for details).

The presence of residual SeV vector was detected only in line DR01, heterozygous of the common SEA α-thalassemia mutation. This line was substituted with clone DR07 which shares the same donor with DR01. DR07 was assessed for pluripotency marker expression and presence of viral vectors and was found to be suitable for downstream experiments (Appendix *fig*. A3.4.2.1).

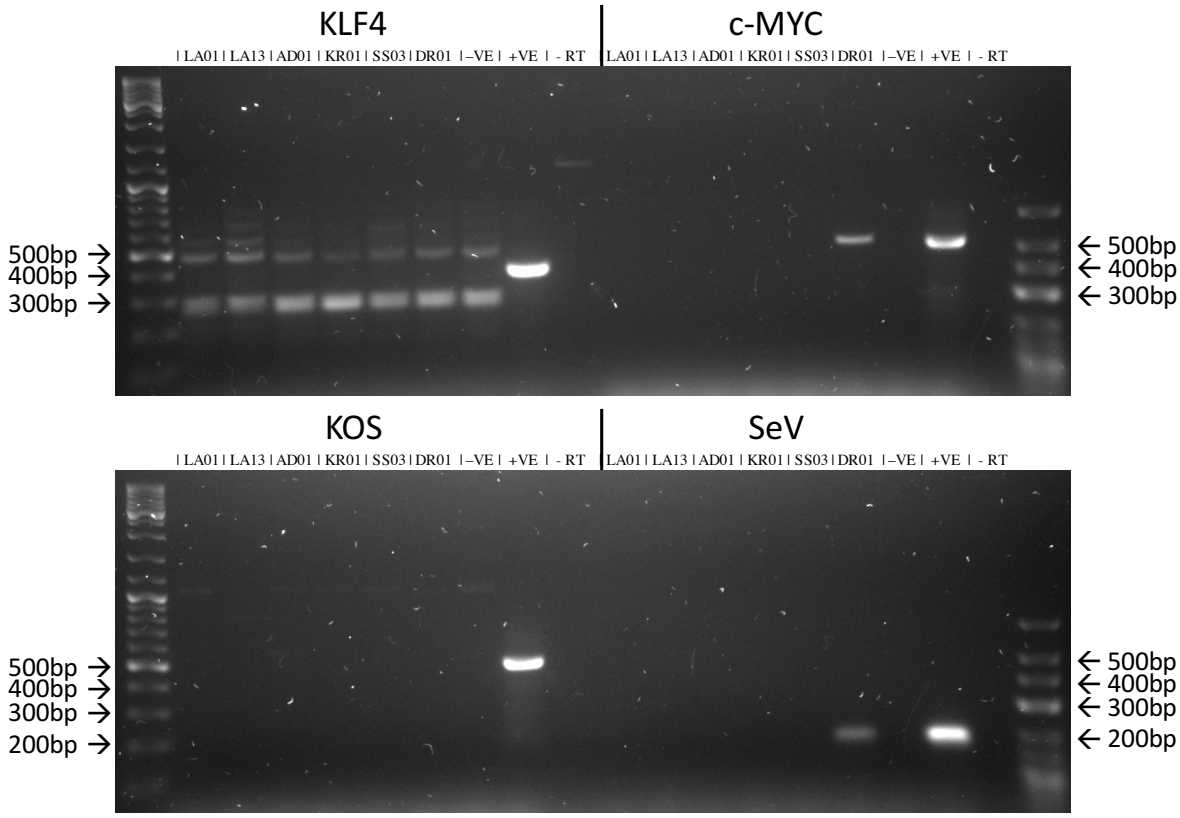


Figure 3.4.2.1 RT-PCR to detect the presence of vector in newly generated iPSC clones.
PCR primer pairs to specifically detect the KLF4, c-MYC, KOS (KLF4, OCT4, SOX2) expressing vectors and a primer pair to detect the SeV vector backbone were used on cDNA from line LA01, LA13, AD01, KR01, SS03 and DR02. For negative control (-ve) cDNA was used from an iPSC line which is known to be clear of SeV vectors. The positive control (+ve) was cDNA from cells undergoing reprogramming (day 3 after infection). In addition, a no RT (-RT) control was used to detect any DNA or reagent contamination. Expected band sizes: KLF4 – 410bp, c-MYC – 532bp, KOS – 528bp, SeV – 181bp. Ladders: left - 2 log; right - 100bp.

3.5 Conclusion

The reprogramming of differentiated cells to pluripotency generates cells that are phenotypically and functionally very similar to ES cells in that they can self-renew indefinitely and are pluripotent. Patient derived cells were reprogrammed using SeV viral vectors that express the reprogramming factors OCT4, SOX2, KLF4 and c-MYC. Prior to further analysis, the newly generated iPSCs were assessed for the expression of pluripotency markers, the silencing of reprogramming vectors and the lack of large detrimental genomic aberrations. Using IF the expression of pluripotency transcription factors OCT4 and SOX2 and cell surface markers SSEA4 and TRA-1-60 was demonstrated in the iPSC lines generated for the work done in this thesis. SNP array analysis confirmed the absence of potentially detrimental CNVs larger than 75kb which may have arisen during culture, with the caveat that smaller CNVs which may not have been detected can also cause changes in gene expression. Additionally, RT-PCR was used to demonstrate the absence of residual reprogramming vectors in the newly established lines. Before downstream experiments, the lines were confirmed to be mycoplasma negative. The summary of the characterization of patient-derived iPSCs is presented in table 3.5.1.

Line	CNV and LOH	Pluripotency marker expression	SeV vectors	Mycoplasma
LA01	None of concern	✓	None detected	Clean
LA13	None of concern	✓	None detected	Clean
AD01	None detected	✓	None detected	Clean
KR01	None of concern	✓	None detected	Clean
SS03	None of concern	✓	None detected	Clean
DR07	None of concern	✓	None detected	Clean

Table 3.5.1 Patient iPSC lines selected for differentiation along the erythroid pathway

Chapter 4: Characterizing the Erythroid differentiation system

4.1 Introduction

Erythropoiesis

Adult erythropoiesis is a highly regulated and complex multi-step process through which mature red blood cells (erythrocytes) are produced through the differentiation of hematopoietic stem cells (HSCs). There are two main categories of erythropoiesis: the first one is primitive erythropoiesis which is characterized by megaloblastic, nucleated erythrocytes that are enucleated after they enter circulation; the second one is definitive erythropoiesis where cells are smaller and remove their nucleus before entering circulation (Palis 2008).

Primitive erythropoiesis is established in the extra-embryonic yolk sac where the primitive erythroid progenitors are thought to be derived from a hemangioblast precursor and arise at 18 to 20 days of human gestation in structures called blood islands (Bloom and Bartelmez, 1940; Choi et al., 1998; Huber et al., 2004; Kennedy et al, 2007; Luckett, 1978; Zambidis et al., 2005). Early on, primitive erythroblasts express the embryonic ζ-globin gene (*HBZ*) from the α-globin locus and the embryonic ε-globin (*HBE*) gene and a small amount of the foetal γ-globin from the β-globin locus. A switch in expression from ζ-globin to the foetal/adult α-globin genes *HBA1* and *HBA2* and from ε-globin to the foetal γ-globin genes *HBG1* and *HBG2* takes place between 5 and 7 weeks of human gestation in the primitive erythroid compartment with the ζ to α switch preceding the ε to γ (Peschle et al., 1985). Definitive erythropoiesis on the other hand occurs in two waves. The first wave of definitive erythropoiesis occurs in the yolk sac

and starts with the establishment of the so-called burst-forming unit erythroid (BFU-E) cells. At the same time a multipotential hematopoietic progenitor cell emerges from the yolk sac: the definitive erythro-myeloid progenitor (EMP), which is believed to be the precursor of the BFU-E (Palis et al., 2014). EMPs are thought to originate from either a hemogenic endothelial or hemangioblast precursor (Hirschi, 2012). The EMP derived BFU-Es in the human embryo emerge at 4–5 weeks of gestation. They expand rapidly in the yolk sac and then enter circulation to ultimately seed the foetal liver, as it begins to form during development, thus establishing the earliest form of definitive erythropoiesis (Migliaccio et al., 1986). The mouse yolk sac-derived definitive erythroid cells predominantly express the foetal/adult β-globin, but also low levels of embryonic globin (McGrath et al., 2011). This is similar to early human definitive erythroid cells which produce the embryonic ε- globin, the foetal γ-globin plus low amounts of the adult β-globin and high amounts of α-globin at 6 weeks of gestation but only α-globin and γ-globin plus low β-globin from 8 weeks onward (Peschle et al., 1985). The first wave of definitive erythropoiesis is transient and is replaced by a permanent wave that is derived from hematopoietic stem cells (HSCs) which seed the liver at a similar time as the transient definitive wave and establish lifelong definitive erythropoiesis (McGrath and Palis, 2008; Palis, 2014). Before birth the site of definitive erythropoiesis begins to change from the foetal liver to the bone marrow where the HSCs will reside permanently throughout life (Charbord et al., 1996). In humans, with the onset of definitive erythropoiesis from the foetal liver, the production of the foetal haemoglobin, composed of γ-globin and α-globin, is markedly increased. Subsequently a fetal to adult haemoglobin switch occurs at the perinatal period, transitioning from γ-globin to the adult β-globin gene HBB (Ley et al, 1989; McGrath and Palis, 2008; Peschle et al, 1985;

Stamatoyannopoulos, 2005). An illustration summarising globin gene expression during primitive and definitive erythropoiesis can be found in figure 4.1.1.

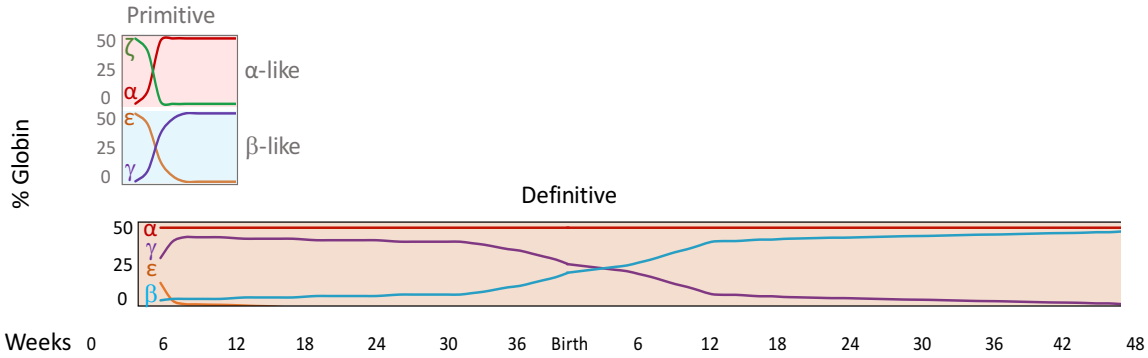


Figure 4.1.1 Globin gene expression during primitive and definitive erythropoiesis from conception to 48 weeks after birth. Vertical axis represents percent globin chains produced; Horizontal axis represents weeks from conception to birth and up to 48 weeks after birth.

During definitive erythropoiesis, the most immature erythroid lineage restricted progenitor is the BFU-E cell which is followed by the colony forming unit erythroid (CFU-E) cell. This classification stems from the ability of the progenitors to produce colonies composed of maturing erythroid cells in a semisolid medium (Palis 2014). Early on in differentiation these erythroid precursors express the transferrin receptor CD71 (Okumura et al., 1992). It takes 14 days for a human BFU-E to generate over a thousand haemoglobinised erythroblasts. Erythroblast differentiation can be divided in four stages and is accompanied by successive loss of the expression of CD71 and upregulation of sialoglycoprotein glycophorin A (CD235a) to reach maximal expression at the later stages of differentiation (Loken et al., 1987). In increasing order of maturation, the four stages of erythroblast differentiation are proerythroblast, basophilic erythroblast, polychromatophilic erythroblast and orthochromatophilic erythroblast (*fig.* 4.1.2).

During maturation, the erythroblast undergoes morphological and size changes with the proerythroblast cell being the largest. Following the proerythroblast stage, the cells undergo a reduction in size with their nucleus condensing and increasing its heterochromatin content. Additionally, the cytoplasm of the cells becomes increasingly acidophilic reflecting the increase of haemoglobin and decrease of RNA content. The last division that the orthochromatophilic erythroblast undergoes is asymmetric and results in enucleation and the creation of a pyrenocyte (a membrane bound excluded nucleus) and a reticulocyte, which contains the cytoplasm, organelles and haemoglobin. The reticulocyte continues to reduce in size by removing its organelles and undergoes cell membrane changes to produce a flexible, biconcave disc (Bell, Sallah and Diggs, 2005; Griffiths et al. 2012).

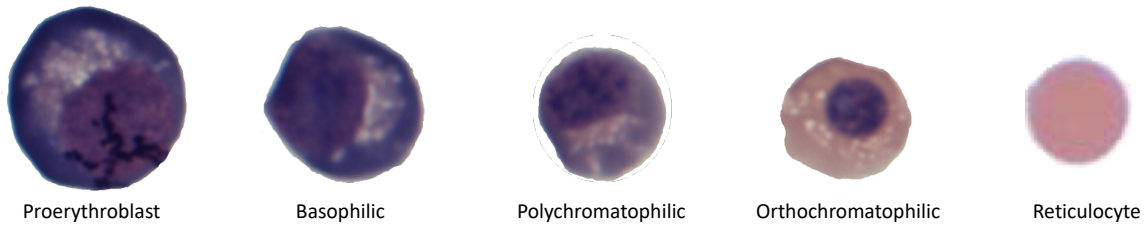


Figure 4.1.2 Morphology of stained human erythroblasts during maturation from Pro- to Orthochromatophilic –erythroblast and reticulocyte.

Erythroid differentiation of iPSCs

To accomplish the aim of the thesis, iPSCs had to be differentiated along the erythroid pathway to produce cells which activate the α-globin locus and regulate its expression in a manner similar to that seen in primary erythrocytes. The differentiation was

performed using an earlier version of the protocol published by Olivier et al., 2016. This earlier version doesn't include the use of two small molecules (StemRegenin - SR1 and Pluripotin - SC1) which the authors have shown, when differentiating human ESCs in this system, significantly reduce the production of embryonic globin chains. The protocol has proven to be robust and highly efficient and in addition is fully feeder and serum free which removes the batch to batch variability that can affect the differentiation process, creating a fully defined system.

4.2 Differentiation, Expansion, FACS and Morphologic analysis

The differentiation process begins with mechanically cutting a nearly confluent monolayer of iPSC into small squares of similar sizes (see Chapter 2 Materials and Methods). These squares generate small embryoid bodies (EBs) which are cultured in the presence of cytokines that promote mesoderm induction and prime hematopoietic commitment (Lengerke et al., 2008; Pearson et al., 2008; Pick et al., 2007). After 72 hours, the EBs are disaggregated and plated out at low density in medium containing factors that promote hematopoietic differentiation (Olivier et al., 2012; Perlingeiro et al., 2003; Zhang et al., 2006). The plated cells form aggregated adherent structures which begin releasing non-adherent cells in suspension at day 7 of differentiation (see fig. 4.2.1a), Olivier et al., 2016 have argued that these budding structures resemble haemogenic endothelium. Indeed, a parallel can be drawn between the formation of these iPSC-derived budding structures and the process by which blood forms from hemogenic endothelium. This is said to involve an endothelial-to-hematopoietic transition where the transitioning cell becomes round and detaches from the endothelial layer (Bertrand et al., 2010; Boisset et al., 2010; Eilken et al., 2009; Kissa and Herbomel, 2010; Lancrin et al., 2009). On day 10 of differentiation the non-adherent cells are collected and cultured in medium that promotes erythroid differentiation and proliferation. Subsequently the cells are collected at day 17 and placed in medium that promotes erythroid maturation (see Chapter 2 for details) and cultured until day 25.

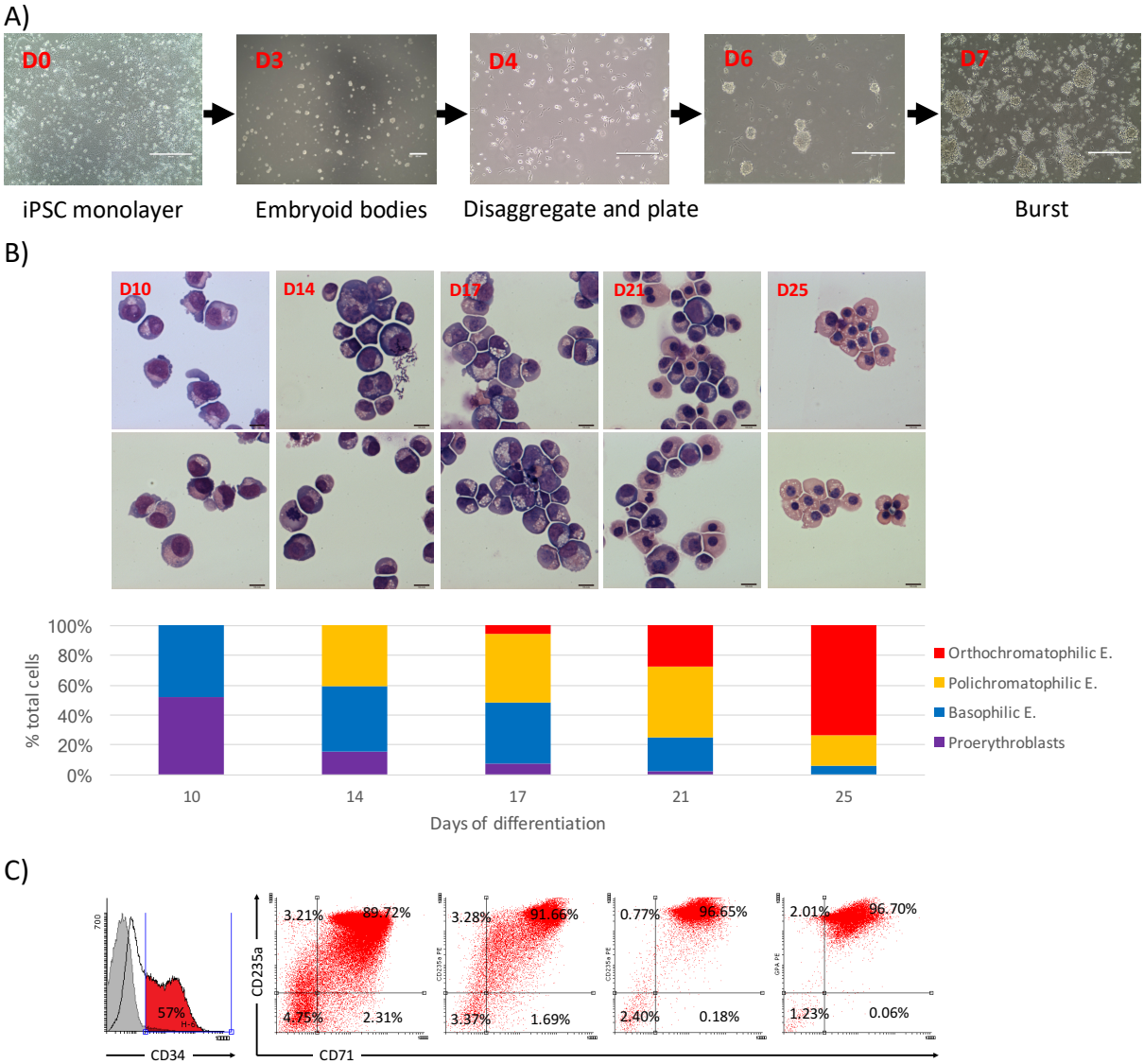


Figure 4.2.1 Morphology and FACS analysis of iPSC derived erythroid cells during differentiation
A) Phase contrast images left to right: iPSCs at day 0 of differentiation, fully grown EBs at day 3 of differentiation, adherent cells at day 4, one day after EB disaggregation, day 6 budding structures before they burst, bursting budding structures producing floating cells at day 7 of differentiation, white scale bar is 400µm. B) Top panel: modified Wright stain of floating non-adherent cells during differentiation days 10, 14, 17, 21 and 25, duplicates taken at random locations on same slide are placed horizontally, black scale bar is 10µm; Bottom panel: percentages of Pro-, Basophilic-, Polychromatophilic- and Orthochromatophilic- erythroblasts during differentiation days 10, 14, 17, 21 and 25. C) Representative FACS for CD34+ cells on day 10 and CD71+ and CD235a+ on days 14, 17, 21 and 25 of differentiation. Morphology and FACS analysis for other WT lines can be found in Appendix *fig*. A4.2.1.

The cells which are produced by the iPSC differentiation system between day 10 and day 25 closely resemble proerythroblasts, basophilic erythroblasts, polychromatophilic erythroblasts and orthochromatophilic erythroblasts, represented in different proportions as the culture progresses. Judging by morphology from day 10 to day 25 of differentiation (*fig*. 4.2.1b) the cells appear to be progressing through erythroblast maturation in a manner similar to that seen in primary erythroid cells. The culture of non-adherent cells starts off as a mixture of large proerythroblast-like cells (52%) and smaller basophilic erythroblast-like cells (48%). As the cells proceed through differentiation, they get smaller, their cytoplasm gradually turns more acidophilic and the nucleus condenses to form polychromatophilic erythroblasts-like cells which at day 14 and 17 represent 41-46% of the culture. Orthochromatophilic erythroid-like cells first appear around day 17 and represent 27% of the culture on day 21 and 73% on day 25. Very few cells appear to be of non-erythroid nature but those which appear to be non-erythroid are very large with lobulated nuclei resembling megakaryocytes. Of note, even at day 25 no enucleated erythroid cells could be readily identified in the cultures.

The hematopoietic and erythroid nature of the iPSC derived erythroid-like cells was further confirmed by immuno-phenotypical characterization using Fluorescence Activated Cell Sorting (FACS) analysis. At day 10 a proportion of the non-adherent cells express CD34, a marker which is associated with HSCs, which indicates the immature nature of the cells. This is of interest as CD34 in the erythroid lineage is not expressed after the CFU-E stage which is marked by the expression of CD71 (Higgs and Wood, 2011). Erythroid differentiation was confirmed by using antibodies against CD71 and CD235a. As previously discussed CD71 is expressed in early erythroid progenitors and is

progressively lost during late stage differentiation to become absent in reticulocytes and mature RBCs. In contrast, CD235a is a marker of more mature erythroid cells and first appears at the proerythroblast stage with its expression increasing throughout the rest of erythroid differentiation (Higgs and Wood, 2011). Both CD71 and CD235a are established on erythroid cells at the time when the presence of α-globin chains becomes detectable in differentiating BFU-Es. In the final stages of *ex vivo* differentiation of BFU-Es extracted from umbilical cord blood, when the culture contains mainly poly- and ortho-chromatophilic erythroblasts, the expression of CD71 becomes significantly reduced and the expression of CD235a has already reached its maximum and plateaued (Okumura et al., 1992). In the iPSC differentiation system both markers are present on the majority of non-adherent cells throughout differentiation (*fig*. 4.2.1c). If we compare the iPSC differentiation system and BFU-E differentiation, the equivalent stage when most of the culture is comprised of poly- and ortho-chromatophilic erythroblasts is between days 21 to 25 of iPSC differentiation. However, there is no significant change in CD71 expression, although the cells have visibly matured based on morphology.

This iPSC based erythroid protocol resulted in high levels of cell expansion with one iPSC being able to generate around 10,000 erythroid-like cells by day 21 of expansion and up to 15,000 cells by day 25 (*fig* 4.2.2a). This large expansion potential was absolutely essential for the completion of the project as many of the downstream experiments require tens of millions of erythroid cells. During the first 3 days of culture (EB formation) cell growth increased exponentially. From day 4 to 6 the cells were growing as attached clusters on the cell dish and from day 7 up to day 10 the culture was composed of both suspension and adherent cells. The cells in suspension were collected

while the leftover adherent cells were discarded. Only the non-adherent cells were counted on day 10 and for that reason a slight dent in the exponential curve can be observed at that time. The cells retain their exponential expansion from day 10 to day 16-17 at which point the growth of the culture decreases with further maturation of the erythroid cells. Between day 17 and day 21 the cell number only doubles and from day 21 to day 25 the culture number increases by less than 50%, at this stage the rate of cell death increases significantly (*fig* 4.2.2b). The cell culture kinetics from day 10 onwards is comparable to that of HSC based primary erythroid culture protocols with an early exponential phase and a decrease in expansion during late stage maturation (Giarratana et al., 2005; van den Akker et al., 2010).

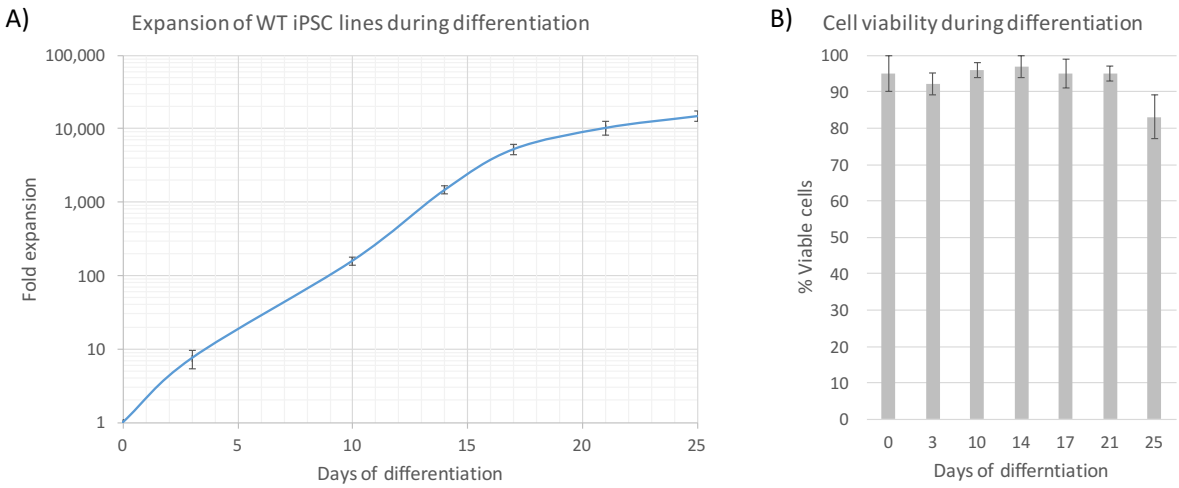


Fig 4.2.2 Expansion potential and cell viability during erythroid differentiation of three WT iPSCs lines.
A) Fold expansion during erythroid differentiation of wild type iPSCs up to day 25, error bars represent one standard deviation. Values are adjusted to represent the expansion of one starting iPSC. B) Viability during erythroid differentiation of wild type iPSCs up to day 25 error bars represent standard error.

The viability of the cells was high throughout the differentiation on average over 95% up to day 21, from day 21 onwards the viability dropped to 83% on day 25. To ensure that the apoptotic cells don't affect downstream analysis and to ensure that globins were actively expressed in the majority of the assayed cells, an intermediate stage in the culture protocol was chosen past which the cells would not be harvested for analysis: day 21.

4.3 Gene expression

Globin gene expression is a characteristic marker of erythroid differentiation and during erythropoiesis globin mRNA levels rise gradually from the proerythroblast stage. The expression of ζ-globin (*HBZ*), α-globin (*HBA1,2*), ε-globin (*HBE*), γ-globin (*HBG1,2*) and β-globin (*HBB*) was assessed using Quantitative PCR (qPCR with exon-spanning TaqMan probes) in three individual, wild-type iPSC lines differentiated along the erythroid pathway (*fig.* 4.3.1). During the early stages of differentiation, expression of the embryonic globins ζ and ε, is higher than expression of the fetal γ-globin and α-globin, while β-globin expression is repressed. As the culture progresses the level of globin expression goes up but there is also a marked switch between ζ and α-globin and a subtler shift in γ-globin over ε-globin, analogous to the ζ to α switch preceding the ε to γ switch in early primitive erythroid cells. Adult β-globin expression gradually increases during differentiation and at day 17, it is at a similar level as the expression of the reference control, the highly expressed ribosomal subunit gene RPS18. On day 21 β-globin expression is around 3-fold higher than RPS18 but it is still 10- to over 100-fold lower than that of the other globins. The high expression of embryonic and foetal globins further validates the erythroid nature of the cells in culture and provides some insight into their developmental background. However, the RNA expression analysis needs to be validated by measuring protein abundance to quantify individual globin chain levels.

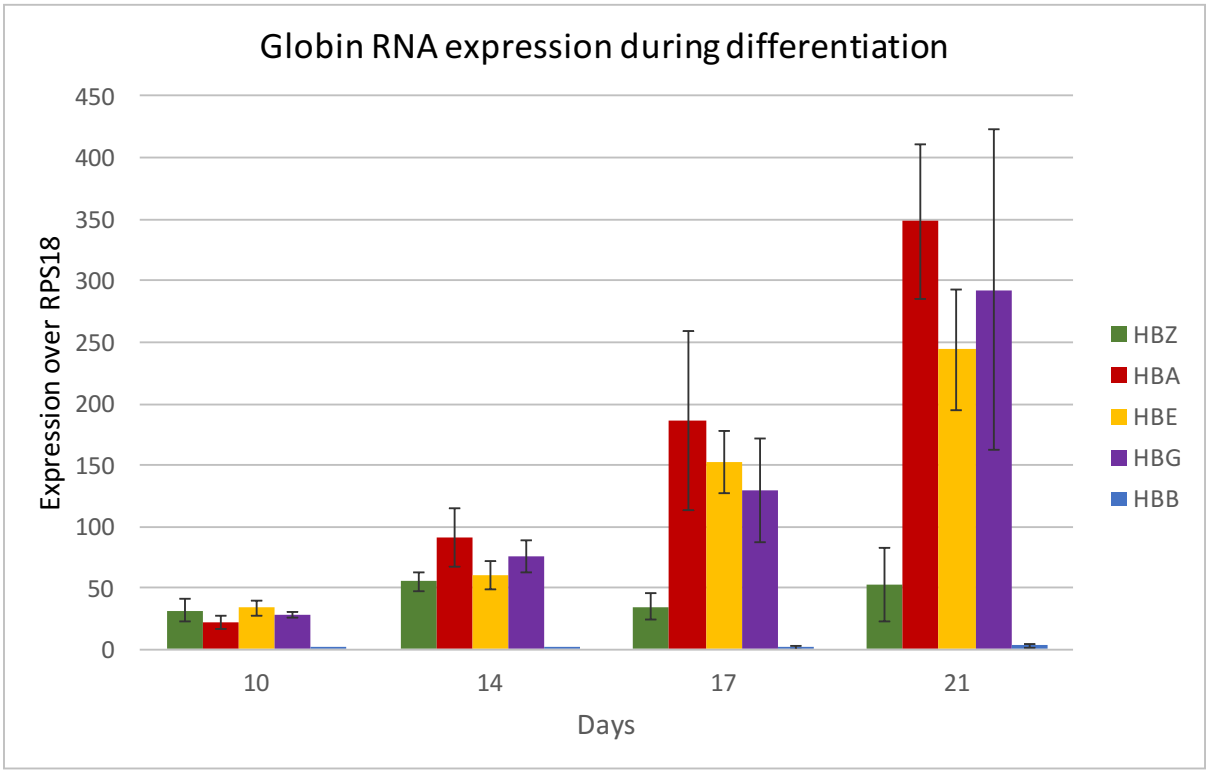


Figure 4.3.1 Globin RNA expression during differentiation
Expression of ζ-globin (*HBZ*), α-globin (*HBA1,2*), ε-globin (*HBE*), γ-globin (*HBG1,2*) and β-globin (*HBB*) RNA during days 10, 14, 17 and 21 of differentiation Mean expression levels relative to *RPS18* (ribosomal protein S18) from 3 independent wild-type iPSC lines quantified using RT-qPCR; error bars represent one standard deviation.

4.4 Analysis of open chromatin regions

One of the characteristic features of regulatory elements is that they are located in open-chromatin. Chromatin accessibility can be assayed by DNaseI as open chromatin is sensitive to endonuclease activity (Higgs et al., 1990) and this can be used to analyse accessibility genome-wide at high resolution (Neph et al., 2012; Song and Crawford, 2010). In addition to DNase-seq a new assay (ATAC-Seq) has been described for determining regions of increased chromatin accessibility. This assay combines ease of use, robustness and allows the use of low cell numbers (Buenrostro et al., 2013). Much like DNase-seq it relies on enzymatic activity but ATAC uses a modified Tn5transposase, which preferentially integrates at open chromatin regions.

We used ATAC-seq to assess the open chromatin regions in cells derived from the iPSC-based erythroid differentiation system. The assay was performed as described in Chapter 2, with input material of 60,000-75,000 wild type iPSC-derived erythroid-like cells harvested at day 21 of differentiation (fig. 4.4.1). For comparison, an ATAC-seq track is shown which has been previously generated in our laboratory (not published; "WT Primary erythroid cells" is not own work; figure generated in collaboration with Martin Larke, Higgs Laboratory, Oxford; primary erythroid cells grown as in Mettananda et al., 2017). The results at the α-globin locus show that the pattern from the ATAC-seq analysis of cells from day 21 of the iPSC-based erythroid differentiated system is comparable to the pattern obtained from ATAC-Seq performed on primary human erythroid cells. Most importantly, all of the regulatory regions (R1-R4) are accessible in

the iPSC system which suggests that the enhancers are within open chromatin and potentially available to regulate expression of the locus. In addition, there is a peak of accessible chromatin over the *HBZ* gene in iPSC-derived erythroid like cells which is not observed in primary adult erythroid cells. This is in agreement with the expression of the *HBZ* gene in the iPSC system observed by qPCR.

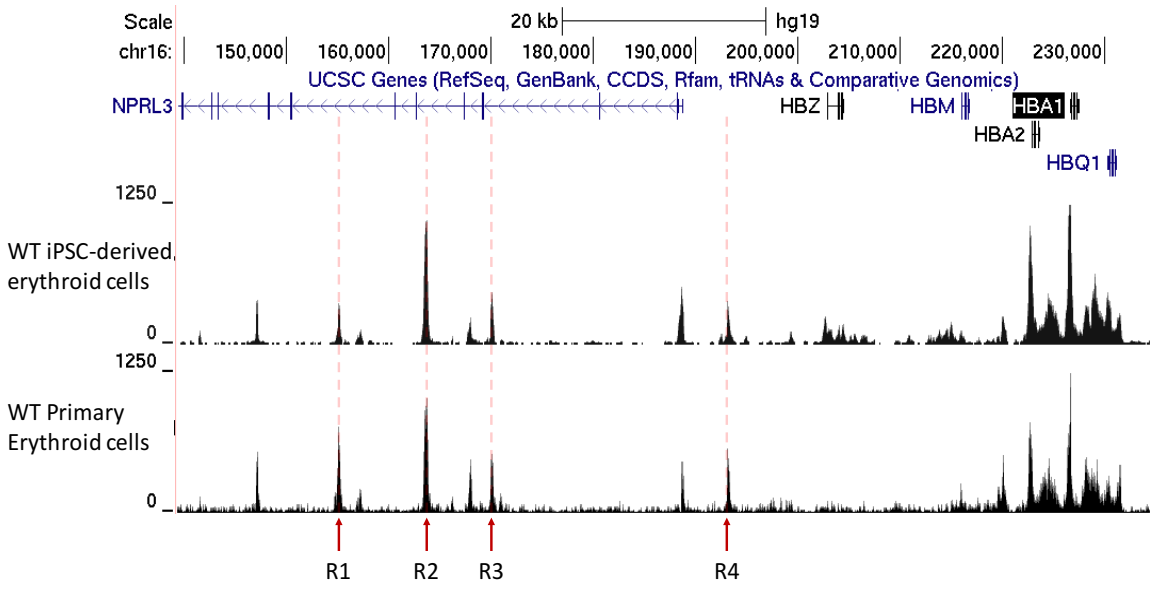


Figure 4.4.1 Chromatin accessibility in the α-globin locus.
UCSC tracks showing chromatin accessibility as assayed by ATAC-seq.
Top panel labelled "WT iPSC-derived erythroid cells": chromatin accessibility at the α-globin locus in day 21 iPSC derived erythroid-like cells using ATAC-seq; Bottom panel labelled "WT Primary Erythroid cells": chromatin accessibility at the α-globin locus from normal adult primary erythroid cells using ATAC-seq; Below the figure, the α-globin enhancers (MCS-R1 to R4) are indicted with red arrows: R1 to R4; Above the tracks the positions of the genes is indicated in blue.

4.5 Chromatin mark analysis of the α-globin cluster in normal iPSCs

Chromatin Immuno-Precipitation followed by next generation sequencing (ChIP-seq) was performed on iPSC-derived erythroid-like cells with a wild type genotype, harvested at day 21 of differentiation (see Chapter 2 for details). I have analysed the enrichment of the promoter mark H3K4me3, the level of RNA Pol II and the binding of the erythroid-specific transcription factors GATA1 and KLF1 (fig. 4.4.1). The data at the α-globin locus shows high levels of PolII over the two α-globin genes HBA1 and 2, moderate levels over *HBZ* and low levels over *HBM* as expected. The results show a pattern similar to that described by De Gobbi et al., 2007 for primary erythroid cells (see Chapter 1 *table* 1.2.1). Namely, the recruitment of key erythroid transcription factors GATA1 and KLF1 to the upstream α-globin enhancers with only KLF1 binding strongly to the globin promoters. In addition, the pattern of H3K4me3 fits that previously described for primary erythroid cells with H3K4me3 only enriched at the globin genes and not at their regulatory elements. These observations are further evidence in favour of the ability of the iPSC-derived erythroid-like cells to activate their α-globin locus in a manner similar to that of primary erythroid cells.

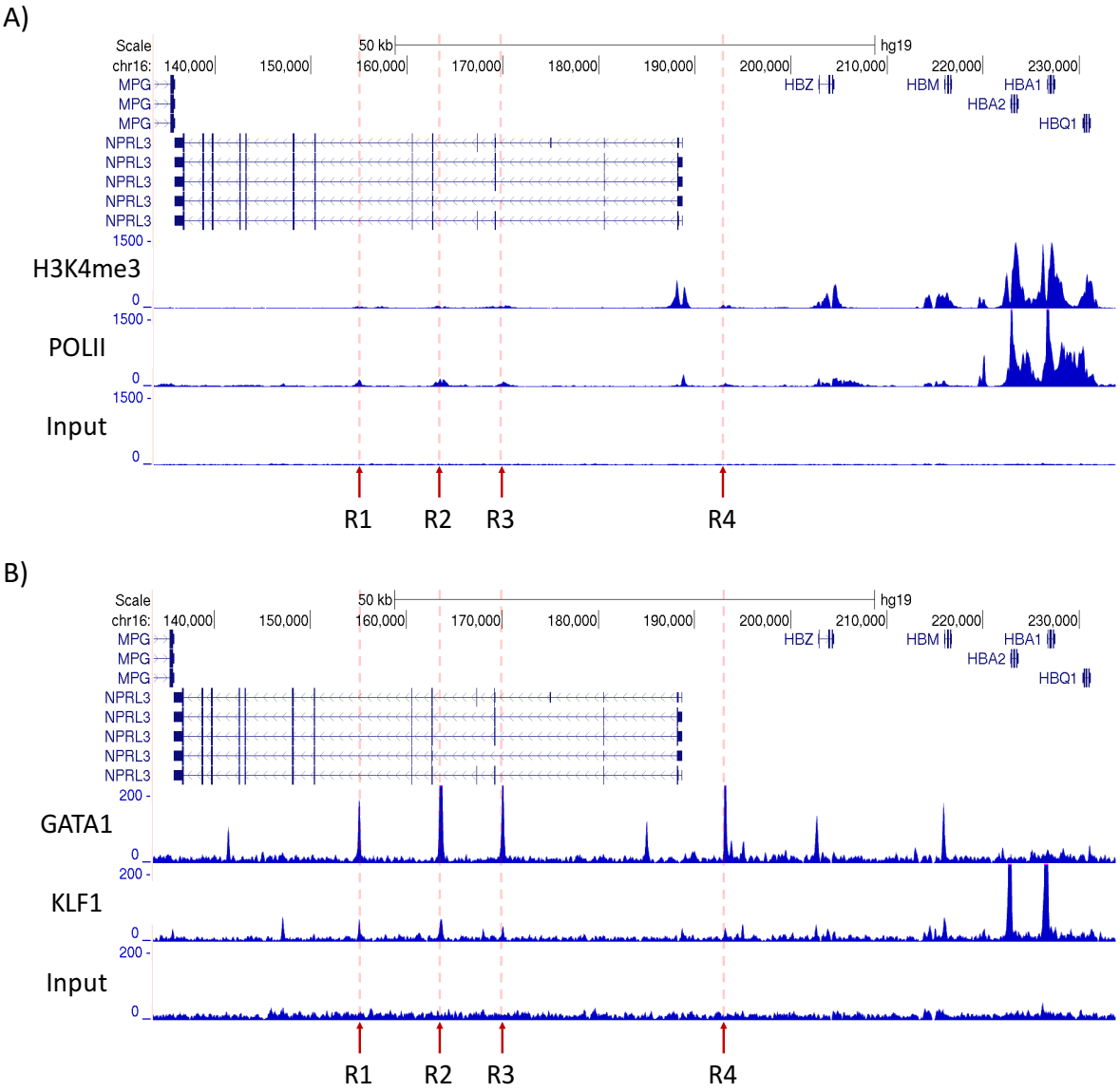


Figure 4.5.1 Chromatin composition of the α-globin locus in iPSC-derived erythroid-like cells. A) ChIP-seq tracks showing H3K4me3 histone mark and RNA polymerase II occupancy of the α-globin locus in day 21 iPSC derived erythroid-like cells, bottom track is genomic DNA input. B) ChIP-seq tracks showing GATA1 and KLF1 occupancy at the α-globin locus in day 21 iPSC derived erythroid-like cells, bottom track is genomic DNA input. All tracks are normalized to the same numbers of reads. Below the figure, the α-globin enhancers are indicated with red arrows: R1 to R4. Genes are above tracks in blue.

4.6 Analysis of Chromatin conformation at the α-globin locus by Capture C

I analysed chromatin interactions using the chromatin confirmation capture (3C) technique Capture-C (Davies et al., 2016) using $1x10^7$ iPSC derived erythroid-like cells with a wild type genotype, harvested at day 21 of differentiation (*fig*. 4.6.1) for details see Chapter 2. It has been demonstrated that activation of the α-globin locus is associated with physical interactions between enhancers and promoters and the recruitment of GTFs and RNA Pol II to the α-globin promoters (Vernimmen, et al., 2009; Hughes et al., 2013; Hughes et al., 2014; Vernimmen, 2014). The Capture-C data demonstrate that, much like their primary cell counterparts, the iPSC-derived erythroid-like cells show an interaction between the active α-globin genes and their enhancers. I assessed the interactions both from the viewpoint of the enhancers and the promoters and the results complement each other (*fig*. 4.6.1) showing increased proximity between regulatory regions R1 and R2 and the α-globin genes *HBA1* and *HBA2*.

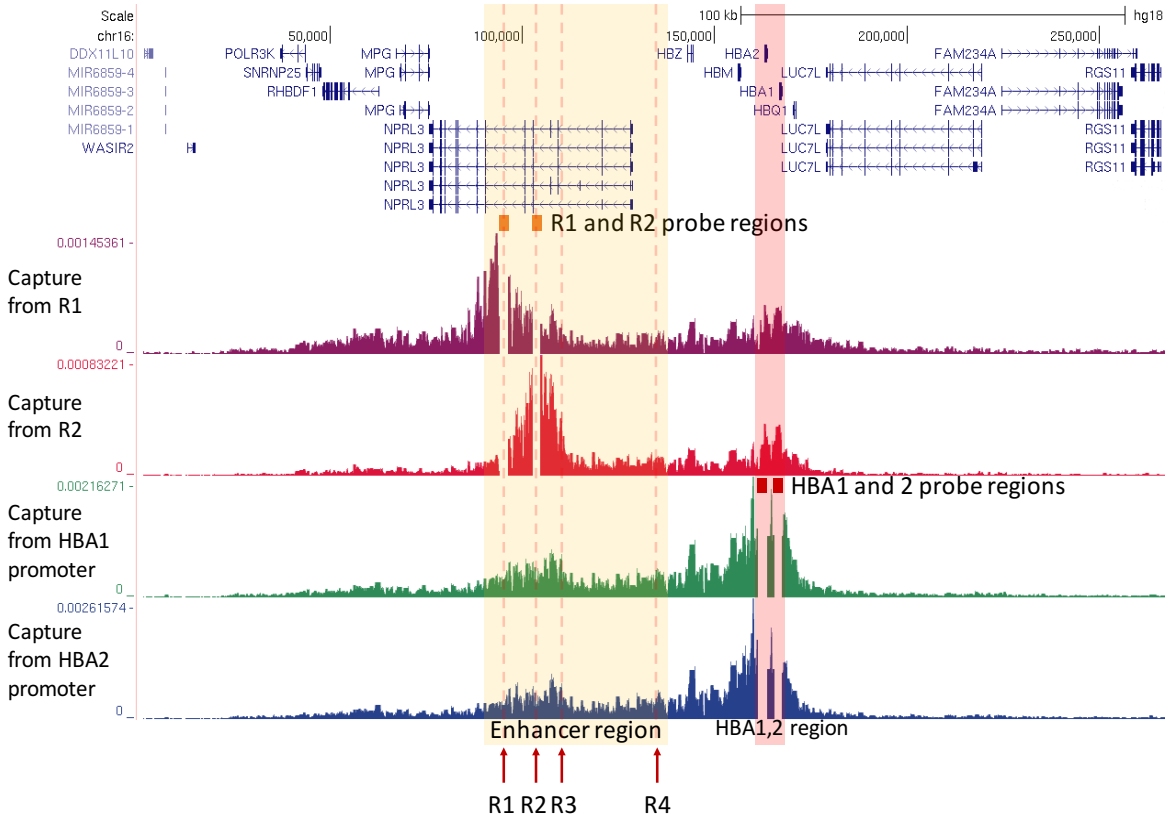


Figure 4.6.1 α-globin locus Capture-C profiles using day 21 wild type cells from the iPSC-based erythroid differentiation system.
Positive values indicate a higher level of interaction with the probed region;
First track represents chromatin interactions between the probe region encompassing the α-globin enhancer R1 and the rest of the α-globin locus. The second track represents chromatin interactions between the probe region encompassing the α-globin enhancer R2 and the rest of the α-globin locus; R1 and R2 probe regions are indicated in orange above the top track, the regions include a 1 kb exclusion zone from which the signal has been removed as it would be too strong and impede visualisation; Third track represents chromatin interactions between probe region encompassing the promoter of *HBA1* and the rest of the α-globin locus. The fourth track represents chromatin interactions between probe region encompassing the promoter of *HBA2* and the rest of the α-globin locus; HBA1 and HBA2 probe regions are indicated in red above the third track, the regions include a 1 kb exclusion zone from which the signal has been removed; Genes are indicated above the track in blue. Below the figure, the α-globin enhancers are indicated with red arrows: R1 to R4;

4.7 Analysis of common types of α-thalassemia iPSC lines

Since one aspect of the phenotype that this thesis aims to dissect is α-thalassemia I sought to prove that the system re-creates downregulation of α-globin RNA caused by a common α-thalassemia mutation which results in the loss of two of the four copies of α-globin and demonstrate that I can detect the expected change in globin RNA. For this purpose, 3 iPSC lines carrying a common α-thalassemia mutation (heterozygous for the (SEA) mutation) were differentiated along the erythroid pathway. The cell cultures were assessed for erythroid characteristics at day 21 (*fig*. 4.7.1a) and all three cultures were composed mainly of erythroid-like cells judging by cell morphology and expression of the late erythroid-specific marker CD235a (over 87% CD235a positive cells).

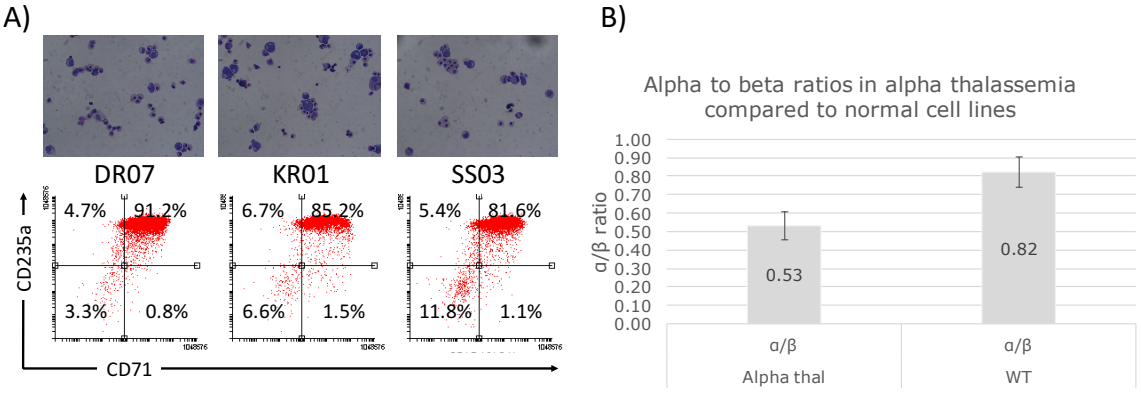


Figure 4.7.1 Alpha thalassemia iPSC lines differentiated along the erythroid pathway for 21 days
A) Top panel, modified Wright staining of day 21 iPSC-derived erythroid-like cells, left to right: line DR07, KR01, SS03; Bottom panel, FACS analysis for CD71 and CD235a of day 21 iPSC-derived erythroid-like cells, left to right: line DR07, KR01, SS03; B) average α/β ratios represented the sum of the expression over RPS18 of the α–like globin genes HBZ and HBA1/2 divided by the sum of the expression over RPS18 of β–like globins HBE, HBG1,2, and HBB. Values were obtained by qPCR and were taken from the three α-thalassemia lines (denoted Alpha thal) and from three wild-type lines (denoted WT). Error bars represent one standard deviation.

Using TaqMan based RT-qPCR the ratio of expression of α–like to β–like globin genes was determined to serve as a measure of the severity of α–thalassemia (*fig*. 4.7.1a). There is a marked decrease in the α/β ratio in the α–thalassemia lines when compared to the wild type lines. This indicates that the TaqMan-based qPCR approach is sensitive enough to detect a drop of α–globin expression of 50% or greater and that the erythroid differentiation system can faithfully reproduce the α–thalassemia phenotype.

4.8 Conclusion

The iPSC-based erythroid differentiation protocol assessed in this chapter produces large numbers of cells which morphologically and immuno-phenotypically resemble primary erythrocytes. These iPSC-derived erythroid-like cells produce high levels of globin mRNA and develop heterogeneous cultures which morphologically appear to progress through different stages of erythroblast maturation similar to that observed in primary erythroid cultures. There are, however, several particularities of the culture which differ from the expansion of primary erythroid cells that are noteworthy, and these are outlined in the following paragraphs.

It is of interest to consider the possible ontogeny of the erythroid-like cells produced by the iPSC-based differentiation system and understand which cells, from which wave of erythropoiesis they resemble the most. If we look at early erythroid progenitors, human BFU-Es require around 14 days to generate mature colonies which typically contain up to over a thousand erythroid cells (Palis, 2014). By comparison, *ex vivo* erythroid differentiation of HSCs isolated from peripheral blood typically result in over ten thousand-fold expansion in the same period of time (see Appendix *fig*. 4.8.1). Committed primitive erythroid progenitors are not capable of self-renewal in culture (England et al., 2011) and have a lower proliferative potential than BFU-Es, typically generating several hundred mature erythroid cells (Wong et al., 1986; Palis et al., 1999). It seems likely that iPSC-derived erythroid-like cell culture at day 7 gives rise to the earliest committed progenitors as this is the day the budding bodies start releasing

suspension cells and according to Olivier et al., 2016 that is the point at which almost all of the cells in the culture express CD34 and CD43. These markers have been used to define hematopoietic progenitors originating from differentiating human PSC (Vodyanik et al., 2006; Choi et al., 2009), in addition, there is evidence demonstrating that these CD43 expressing PSC-derived progenitors are biased towards the erythroid lineage (Kessel et al., 2017). However, from day 7 to days 21-25 differentiating cells in the system used for this thesis only expand 250-350-fold. If we take all evidence into account, judging by proliferative capacity, morphology and globin expression profiles the iPSC derived erythroid cells appear to be similar to the cells originating from the primitive or the transient definitive wave of erythropoiesis and are thus unlikely to be derived from cells which resemble long-term repopulating hematopoietic stem cells. Thus, the iPSC-derived erythroid cells appear to resemble erythrocytes derived from progenitors of embryonic or at most from EMPs of early fetal origin. An alternative explanation of the particular morphologic features, cell surface marker and globin expression patterns is that the iPSC-derived erythroid cells are experiencing a condition similar to stress erythropoiesis. Stress erythropoiesis is a major component of the systematic response that aims to increase oxygen delivery during tissue hypoxia and it results in the rapidly generation of large numbers of new erythrocytes. This condition is thought to involve changes in the cell cycle and terminal maturation of erythroid progenitors and is associated with elevated levels of fetal haemoglobin and dysregulation of erythroid cell surface markers, in particular the simultaneous expression of both immature and late stage markers (Alter et al.,1976; Alter, 1979; Luck et al., 2004; Mathias et al., 2000; Papayannopoulou et al., 1980; Stamatoyannopoulos et al., 1985). This could explain the fetal–like nature of the iPSC-generated erythroid cells

and the asynchrony between morphologic maturation and cell surface marker expression. The condition could be explained by the reliance on cytokines that have been associated with regulating stressed erythropoiesis, specifically BMP4, but also SCF and EPO which drive the expansion and differentiation of stressed erythroid progenitors (Paulson et al., 2011; Xiang et al., 2015). This is further supported by the requirement for BMP4 at the later stages of the differentiation culture, after mesoderm induction (personal communication with the authors of Olivier et al., 2016, unpublished). However, it should be noted that the cells are grown at 20% O_2 and it has been shown that hypoxic conditions (1% O_2) are necessary to elicit the maximal response of stress erythroid progenitors to BMP4 and SCF signalling in the mouse (Perry et al., 2007), suggesting that what is observed in the iPSC differentiation system could be a partial response. If this is true than reduced oxygen conditions may improve the expansion potential of the culture. Another important point to address is that during human erythropoiesis, the only globins that are restricted to specific categories are ζ-globin (restricted to primitive erythropoiesis) and β-globin (restricted to definitive erythropoiesis). However, in the iPSC system both of these globins are expressed at the same time, which either suggests that the timing of globin gene regulation is not faithfully recapitulated in the iPSC system or indicates a mixture of primitive- and definitive- like erythroid cells. Further work examining expression at the single cell level will improve our understanding of the nature of the iPSC-derived erythroid cells.

One of the main questions that we have to address is whether the model of α-globin regulation in cells derived from the iPSC-based erythroid differentiation system is the same as that observed in primary erythroid cells. I have assessed this by looking at

several features of the chromatin in the α-globin locus in cells obtained from the system. The ATAC-seq data shows that the enhancers of the α-globin locus are accessible, and the ChiP-seq data shows that they are bound by erythroid-specific transcription factors, indicating that these regulatory regions are poised or active. Tying these observations together with the chromatin conformation analysis, we see that the α-globin enhancers and promoters are in close proximity and interacting. This suggests that expression of the globin genes is under the regulation of their enhancers as observed for primary erythroid cells, and that the interaction with the regulatory elements upregulates the expression of α-globin to the high levels that have been observed by qPCR.

In conclusion, the analysed iPSC-based erythroid system is capable of producing large numbers of cells which, based on morphology and gene expression, are very similar to primary erythroid cells. In addition, the high output of α-globin and the analogous to *in vivo* regulation of expression, coupled with the ability of iPSCs to generate a stable source of isogeneic mutant and control cells lines, make the system ideal for dissecting the molecular mechanisms underlying the Vanuatu phenotype.

Chapter 5: Recapitulating the Vanuatu phenotype

5.1 Introduction

The main focus of this thesis is to understand the molecular mechanisms that underpin the transcriptional activation at the site of the Vanuatu mutation and why this affects the regulation of α-globin in erythroid cells. Previous work in the lab has already narrowed down the probable cause of the phenotype to a single-nucleotide mutation in an intergenic region between the α-globin genes and their upstream regulatory elements. The mutant allele creates a new binding site for the key erythroid factor GATA1 and also facilitates the recruitment of TAL1 and RNA Pol II in addition to promoting the enrichment of acetylated histone H3 at the site. This results in the transcriptional activation of the region around the Vanuatu mutation and somehow interferes with the expression of the downstream α-globin genes (*fig.* 5.1.1). The presence of the mutation on both alleles in a patient causes a drop in the ratio of α-like to β-like globin RNA transcripts from 0.85-1.22, found in wild type erythroid cells to 0.46 in the patient's red blood cells indicting a severe reduction of α-globin transcription.

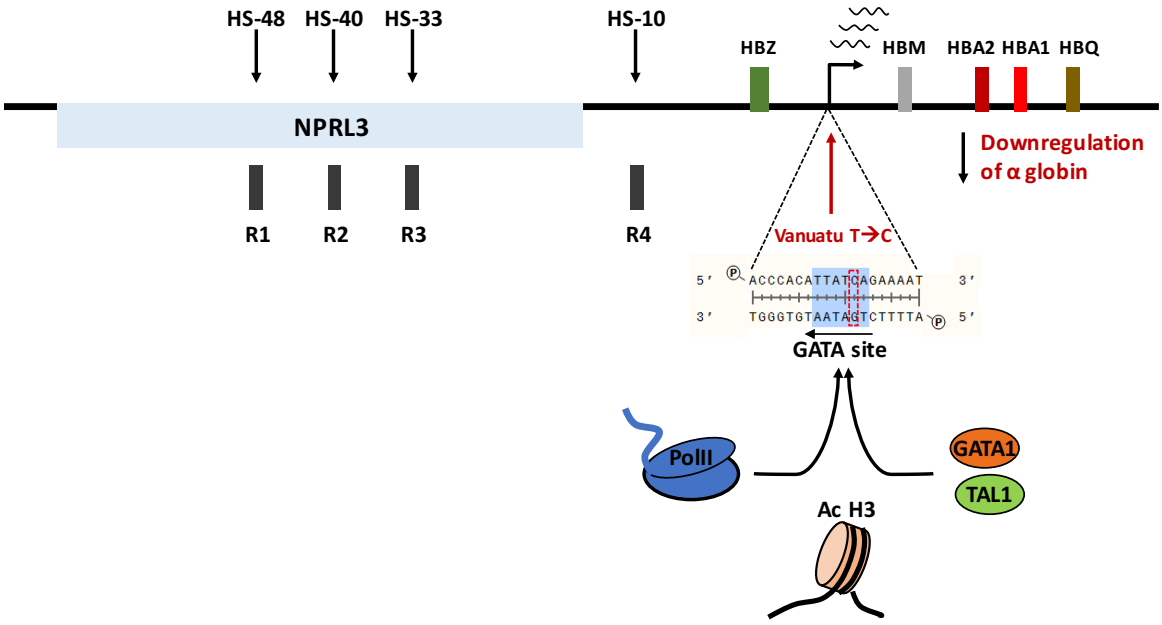


Figure 5.1.1 Schematic representation of the α-globin locus and the characteristic features of the Vanuatu mutation.
On the top, the α-globin locus is depicted covering genes: *NPRL3, HBZ, HBM, HBA2, HBA1* and *HBQ*, along with regulatory elements R1 to R4. The site of the Vanuatu mutation is indicated with a red arrow and expanded to show the nucleotide sequence 5' to 3', the mutant bp is boxed with a red dotted line. Below is an illustration of the chromatin features that have been established in primary erythroid cells derived from an affected patient De Gobbi et al., 2006, namely recruitment of GATA1, TAL1 and RNA POLII and the enrichment of acetylated histone H3.

The Vanuatu mutation is located at coordinate 209,709 for human genome hg19, or 149,709 for human genome hg18. It is a T to C transition that creates a potential binding motif for the key erythroid transcription factor GATA1, on one DNA strand, turning the wild type sequence TAATAA (T allele) to TGATAA (C allele). The new GATA motif is found in a 138bp long region of unique sequence, referred to in this thesis as the Vanuatu site. This unique sequence is surrounded by repetitive DNA elements (*fig.* 5.1.2). Upstream of the GATA site towards the telomere region of the short arm of chromosome 16 there is an inactive LINE element (a LINE2/L2 element) and downstream towards the centromere there is a G/C rich tandem repeat (TR). It is worth mentioning that the GATA1 binding motif is part of a half E-box-GATA motif (CTG-(N_8)-WGATAR), which is a

sequence found to be co-occupied by TAL1 (Kassouf et al., 2010; Soler et al., 2010; El Omari et al., 2013).

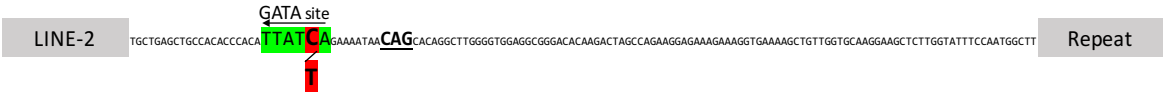


Figure 5.1.2 Vanuatu region sequence.
The 138bp of unique sequence around the Vanuatu mutation are depicted between the upstream LINE element and downstream repeat. The GATA site is highlighted in green and the mutant bp is in red. The upstream half E-box element is enlarged and underlined.

To study the effect of the Vanuatu mutation on the regulation of the α-globin locus I chose to model the disease in an iPSC-based erythroid differentiation system. Patient derived and wild type iPSC lines were differentiated down the erythroid pathway and harvested at day 21 of differentiation. These cells were analyzed for globin gene expression, chromatin accessibility and epigenetic marks. Additionally, chromatin conformation in the α-globin locus was assessed to determine whether the transcribed Vanuatu site establishes new interactions in the active α-globin locus.

5.2 Expression analysis of patient derived iPSC lines differentiated along the erythroid pathway

The mRNA levels of the α-globin (*HBA1,2*), ζ-globin (*HBZ*), ε-globin (*HBE*), and γ-globin (*HBG1,2*) genes was assessed using Quantitative PCR (qPCR) and exon spanning TaqMan probes. The expression of β-globin (*HBB*) was excluded from subsequent studies as it was negligible compared to the expression of the other globin genes and did not have a noticeable effect on the α/β-like ratios which were used to determine the severity of α-thalassemia. Additionally, the expression of the newly generated transcriptional unit at the site of the Vanuatu mutation which is labeled GATAν (or GATA- lowercase Greek letter Nu) was determined using a custom TaqMan probe designed to the unique region surrounding the newly generated GATA1 site (*fig.* 5.2.1). Expression analysis was performed in two iPSC clones derived from a patient homozygous for the Vanuatu mutation (LA01 and LA13), one iPSC line derived from a patient heterozygous for the Vanuatu mutation (AD01) and three individual wild-type iPSC lines (see Chapter 2) differentiated down the erythroid pathway; material was harvested at day 21 (*fig.* 5.2.2).



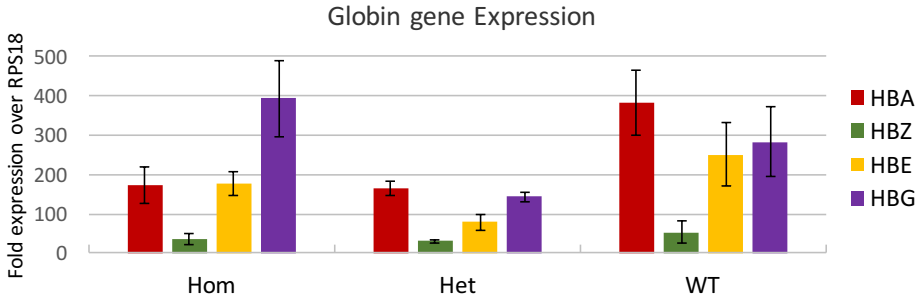
Figure 5.2.1 Vanuatu region sequence with custom TaqMan amplicon.
The 138bp of unique sequence around the Vanuatu mutation are positioned between the upstream LINE element and downstream repeat. The GATA site is highlighted in green and the mutant bp is in red. The TaqMan probe is indicated in purple under its corresponding sequence.

The wild type and homozygous Vanuatu mutant lines expressed similar levels of β-like globin genes with the homozygous Vanuatu clones expressing slightly more fetal (HBG) over embryonic (HBE) globin as compared with the wild type line (*fig.* 5.2.2a). However, the expression of the α-globin genes HBA1 and 2 was significantly reduced in the homozygous Vanuatu mutant when compared to the wild type lines as illustrated by the reduced α/β-like ratio from 0.83 ± 0.06 in the wild type to 0.37 ± 0.11 in the homozygous mutant (*fig.* 5.2.2b). The overall level of globin RNA in the heterozygous Vanuatu mutant was about half of that seen in the wild type lines but the α/β-like ratios were similar 0.87 ± 0.02 and 0.83 ± 0.06 respectively. It is possible that that the heterozygous line has a different differentiation potential when compared to the other lines and may be lagging behind by a few days, suggested by the overall reduced globin expression. The expression of the GATAν transcript in the homozygous Vanuatu clones was between 30- to 100-fold higher than the levels measured in wild type cells (*fig*. 5.2.2c). The expression of the GATAν transcript in the heterozygous Vanuatu clones was around 3-fold lower compared to the homozygous mutants. The reason for the large deviation was the fact that when GATAν expression was measured by qPCR the cDNA from wild type cells produced high Ct values ranging from 36 to 38. The calculated value for GATAν expression in wild type cells as a factor of *RPS18* mRNA levels was still included as relevant since the Ct values were below the no reverse transcriptase control.

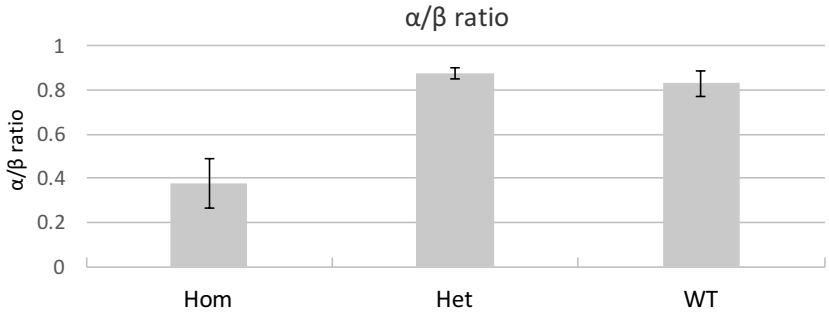
The low levels of the GATAν RNA measured compared to the reference gene *RPS18* can be due to a low level of transcription, but they could also be explained by several other factors. Firstly, the transcription start site of the new transcript has not been determined and the position of the TaqMan probe which is immediately adjacent to the

GATA site might not be optimal to determine the actual levels of expression. Second, qPCR might not be the correct technique to determine the actual rate of transcription as it is targeted at determining the level of stable transcripts and the stability of the GATAν transcript has not been determined yet.

A)



B)



C)

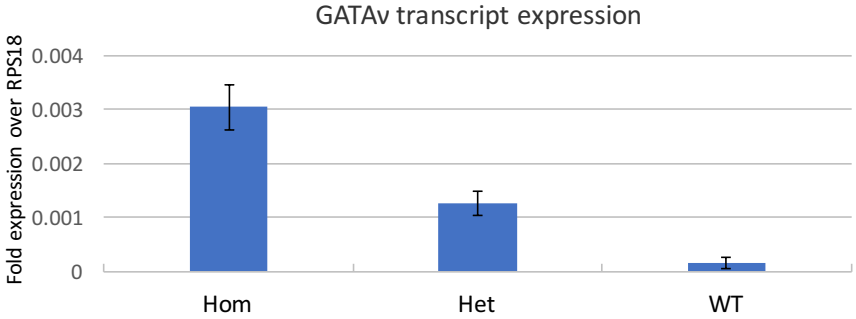


Figure 5.2.2 Expression analysis of patient and wild type iPSC derived erythroid-like cells
A) Expression of α-globin (*HBA1,2*), ζ-globin (*HBZ*), ε-globin (*HBE*) and γ-globin (*HBG1,2*) at day 21 of differentiation. Mean expression levels relative to ribosomal protein S18 (*RPS18*) from two patient lines homozygous for the Vanuatu mutation (Hom) in duplicate (n=4), one patient line

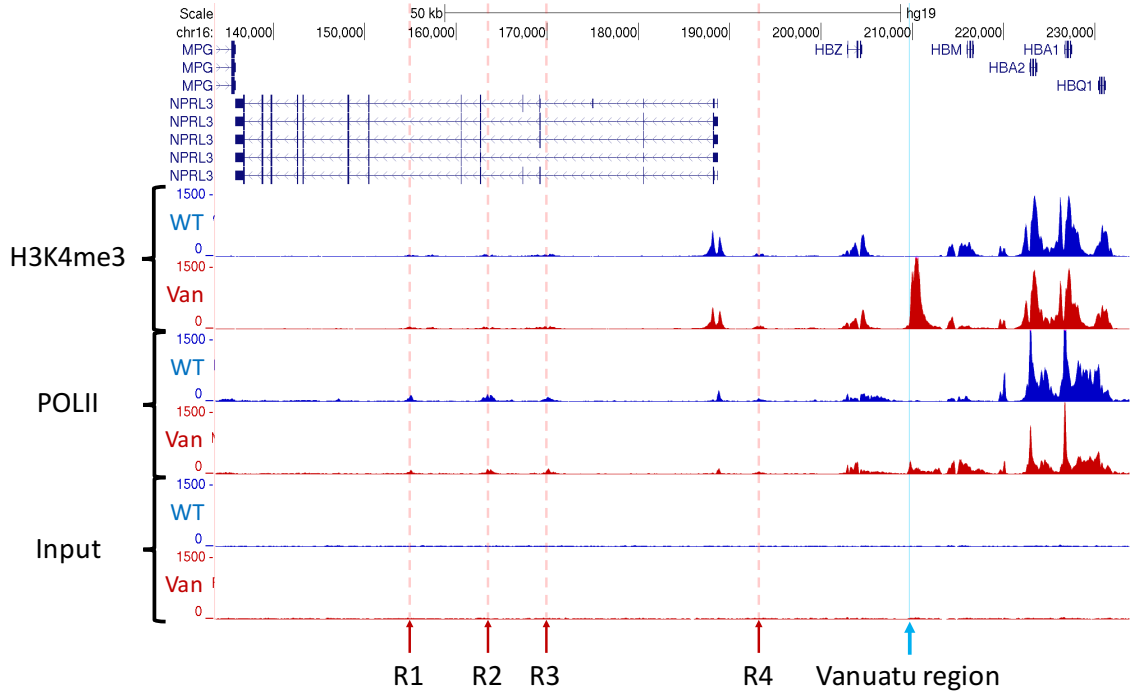
heterozygous for the Vanuatu mutation (Het) in triplicate (n=3) and three independent wild-type iPSC lines (n=3) quantified using RT-qPCR. Error bars represent one standard deviation.
B) Average α/β-like ratios represent the sum of the expression over *RPS18* of the α–like globin genes *HBZ* and *HBA1,2* divided by the sum of the expression over *RPS18* of β–like globins *HBE* and *HBG1,2* at day 21 of differentiation. Values were obtained by RT-qPCR from two patient lines homozygous for the Vanuatu mutation (Hom) in duplicate (n=4), one patient line heterozygous for the Vanuatu mutation (Het) in triplicate (n=3) and three independent wild-type iPSC lines (n=3); Error bars represent one standard deviation.
C) Expression of the new transcript generated by the Vanuatu mutation labeled GATAν at day 21 of differentiation. Mean expression levels relative to *RPS18* from two patient lines homozygous for the Vanuatu mutation (Hom) in duplicate (n=4), one patient line heterozygous for the Vanuatu mutation (Het) in triplicate (n=3) and three independent wild-type iPSC lines (n=3) quantified using RT-qPCR; Error bars represent one standard deviation.

5.3 Chromatin mark analysis of the α-globin locus

ChIP analysis was performed using iPSC derived erythroid-like cells with a wild type genotype and homozygous mutants for the Vanuatu variant (mutant C allele), harvested at day 21 of differentiation (for details see Chapter 2). ChIP-seq for the promoter mark H3K4me3 shows enrichment at the region of the Vanuatu mutation only in the presence of the mutant C allele (*fig*. 5.3.1a) and ChIP-qPCR shows that this enrichment is similar to that observed at the promoter of the *HBA2* gene (*fig*. 5.3.1b).

ChIP-seq for RNA Pol II shows recruitment at the region of the Vanuatu mutation only in the presence of the mutant C allele (*fig*. 5.3.1a). ChIP-qPCR was used to quantify this enrichment and shows that it is 10-fold higher than that observed in wild type cells at the same region and one sixth of the RNA Pol II enrichment observed at the promoter of the *HBA2* gene in wild type cells. ChIP-qPCR for RNA Pol II shows a statistically significant decrease of recruitment at the promoter of the *HBA2* gene in homozygous mutant (C allele) cells compared to the occupancy in wild type cells (*fig*. 5.3.1b). From work done by De Gobbi et al., 2006 we know that the transcription of the GATAν unit is unidirectional and propagates in the direction of *HBM*. The pattern of RNA Pol II occupancy in the patient derived cells corroborates this observation as RNA Pol II occupancy appears to propagate downstream of the putative TSS in the proposed direction of transcription.

A)



B)

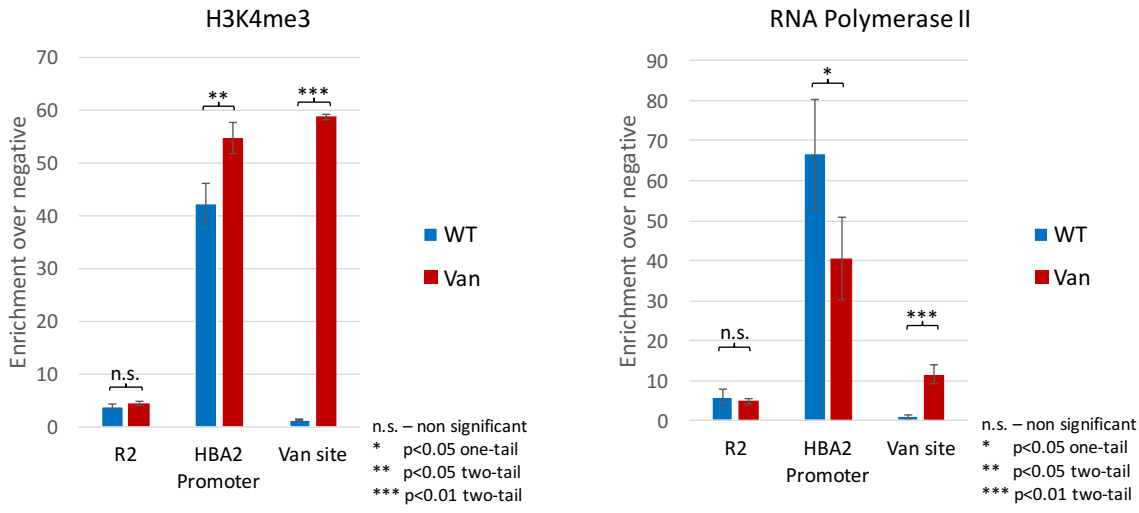


Figure 5.3.1 H3K4me3 and RNA Pol II occupancy in the α-globin locus in iPSC derived erythroid-like cells.

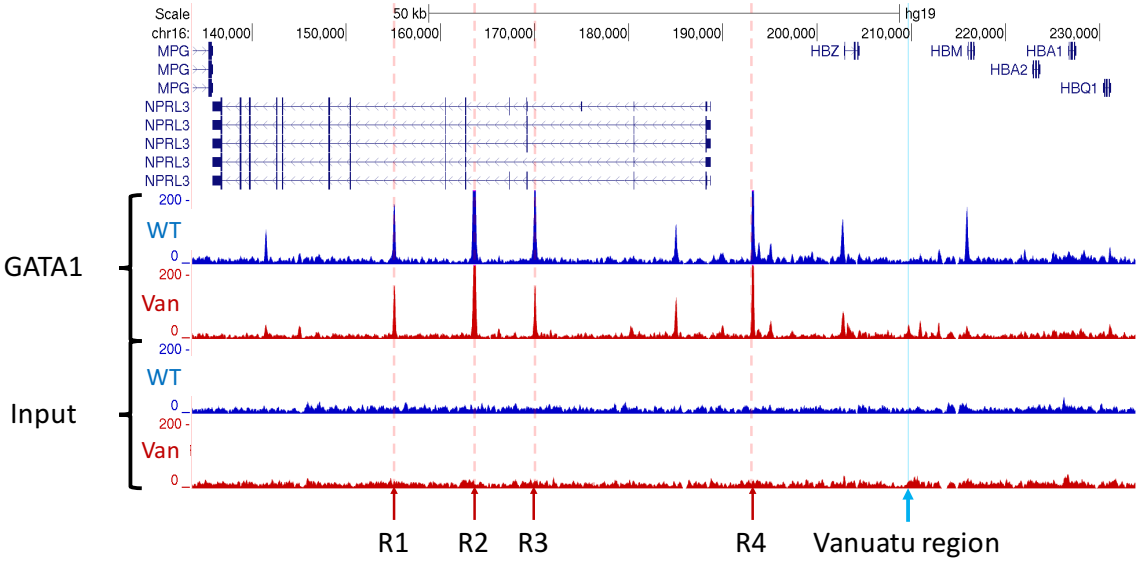
A) ChIP-seq tracks showing H3K4me3 histone mark and RNA polymerase II occupancy of the α-globin locus in day 21 iPSC derived erythroid-like cells from wild type cells (marked WT in blue) and homozygous Vanuatu mutant cells (marked Van in red), bottom track is genomic DNA input; All tracks are normalized to the same number of reads. Below the figure the α-globin enhancers are indicted with red arrows: R1 to R4 and the 138bp of unique sequence of the Vanuatu region is highlighted in light blue; Genes are above tracks in dark blue.

B) ChIP-qPCR results from H3K4me3 histone mark and RNA polymerase II enrichment at the α-globin locus in day 21 iPSC derived erythroid-like cells from wild type cells in triplicate (marked WT in blue) and homozygous Vanuatu mutant cells in triplicate (marked Van in red). The enrichment is measured over enhancer R2, the promoter of the *HBA2* gene and the region around the Vanuatu mutation. Values are represented as enrichment over the average value of

two negative sites found in an intergenic location on chromosome 16. Error bars represent one standard deviation. Statistical significance between samples is indicated by an asterisk above the bars on the histogram.

ChIP-seq for the key erythroid transcription factor GATA1 shows increased binding at the region of the Vanuatu mutation only in the presence of the mutant C allele (*fig*. 5.3.2a) and ChIP-qPCR for GATA1 demonstrates around 2-fold statistically significant enrichment of GATA1 over the mutant C allele as compared to the wild type T allele (*fig*. 5.3.2b). Similar to the 3-fold enrichment observed by De Gobbi et al., 2006. There appears to be a general redistribution of GATA1 occupancy in the α-globin locus in the vicinity of the Vanuatu mutation. In addition to the new peak of GATA1 binding over the mutant C variant there is a new peak of enrichment 1.5kb downstream of the mutation towards *HBM* and a severe reduction of the GATA1 peak at the promoter of the *HBM* gene. This observation is consistent with the downregulation of the *HBM* gene in patient primary erythroid cells and provides a clue to the possible mechanism by which the GATAν transcriptional unit could interfere with the activity of the α-globin promoters (discussed further in Chapter 8 section 8.4).

A)



B)

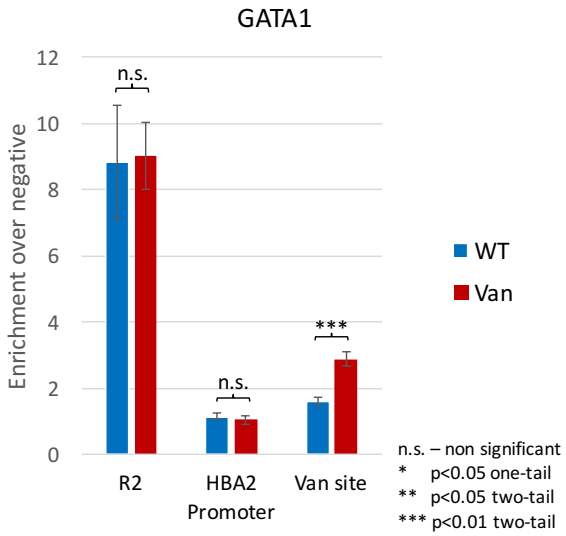


Figure 5.3.2 GATA1 occupancy in the α-globin locus in iPSC derived erythroid-like cells.
A) ChIP-seq tracks of GATA1 occupancy of the α-globin locus in day 21 iPSC derived erythroid-like cells from wild type cells (marked WT in blue) and homozygous Vanuatu mutant cells (marked Van in red), bottom track is genomic DNA input. All tracks are normalized to the same number of reads. Below the figure, the α-globin enhancers are indicated with red arrows: R1 to R4 and the 138bp of unique sequence of the Vanuatu region is highlighted in light blue. Genes are above tracks in dark blue.
B) ChIP-qPCR results from GATA1 enrichment at the α-globin locus in day 21 iPSC derived erythroid-like cells from wild type cells in triplicate (marked WT in blue) and homozygous Vanuatu mutant cells in triplicate (marked Van in red). The enrichment is measured over enhancer R2, the promoter of the *HBA2* gene and the region around the Vanuatu mutation; Values are represented as enrichment over the average value of two negative sites found in an intergenic location on chromosome 16. Error bars represent one standard deviation. Statistical significance between samples is indicated by an asterisk above the bars on the histogram.

The observation by De Gobbi et al., 2006 that TAL1 binds together with GATA1 to the mutant C allele in primary erythroid cells prompts the question of which other transcription factors might bind the site. To systematically narrow down the options we used a computational approach that was recently published by our lab (Schwessinger et al., 2017). The tool predicts the effect of point mutations on transcription factor binding in relevant tissue types and, as expected, was able to predict that reverting the mutant C allele to the wild-type T should have a large negative impact on GATA1 binding at the Vanuatu site in erythroid cells. By simulating mutations in the surrounding sequence, we identified several regions which could be potential transcription factor binding sites in erythroid cells and are present with the same sequence in both the wild type and mutant allele. After querying the sequences underneath the regions against the transcription factor motif database JASPAR two potential binding sites for the key erythroid transcription factor KLF1 were identified flanking the new GATA site in the Vanuatu region (*fig.* 5.3.3). The other major peaks contained binding motifs for the FOX and ETS family of transcription factors (see Appendix *fig*. A5.3.1). To test whether KLF1 binds to the mutant C allele ChIP was performed on day 21 differentiated erythroid-like cells derived from wild-type and homozygous Vanuatu mutant iPSC lines. ChIP-seq shows that the TF does indeed bind at the region of the Vanuatu mutation only in the presence of the mutant C allele (*fig*. 5.3.4a) and ChIP-qPCR demonstrates an over 8-fold statistically significant enrichment of KLF1 over the mutant C allele as compared to the wild type T allele (*fig*. 5.3.4b).

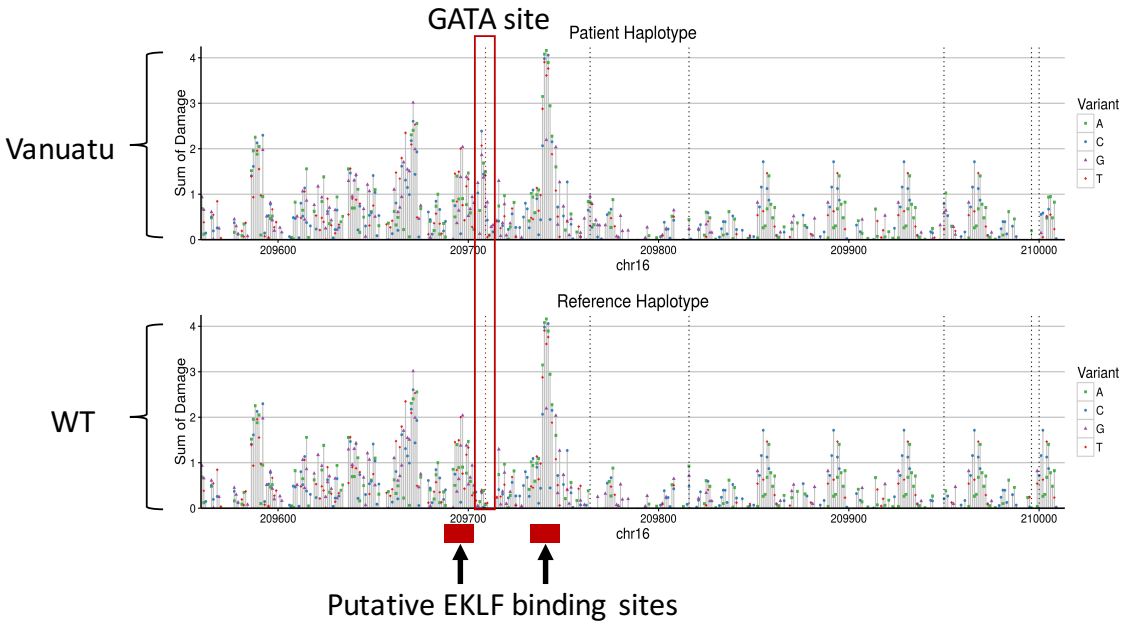
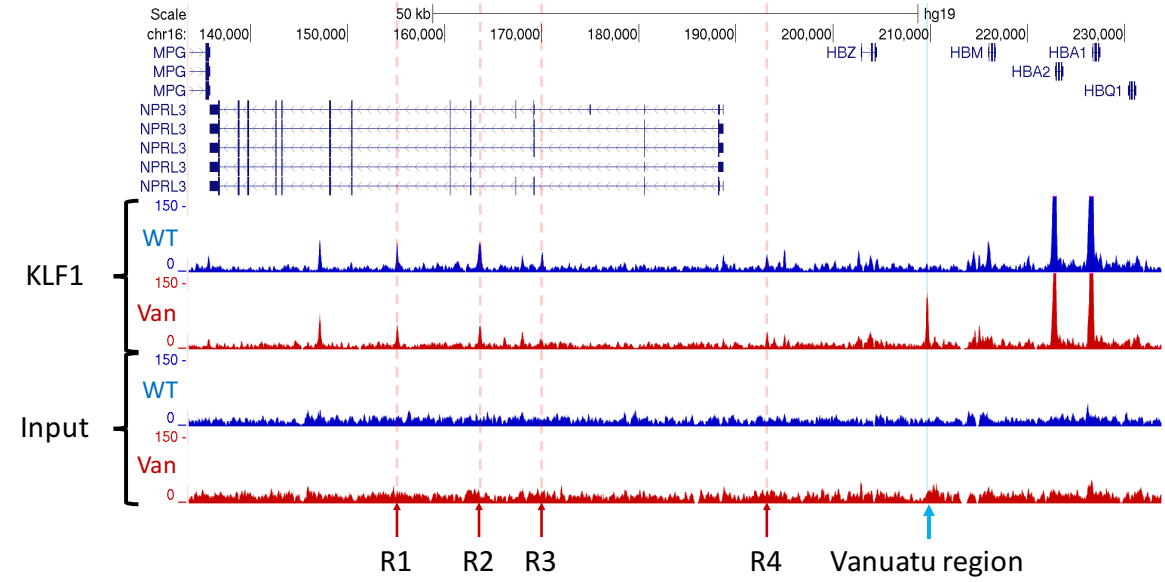


Figure 5.3.3 Sasquatch damage score predictions from *in silico* mutation, based on primary erythroid DNase-seq data comparing patient haplotype (Van on top) vs. reference haplotype (WT on bottom). Horizontal axis represents 466bp of sequence around the Vanuatu variant in chromosome 16 on human genome build hg19. Dotted lines indicate positions of variants which differ between the patient and the reference haplotype, the Vanuatu variant is boxed in red and indicated as GATA site at the top. Vertical axis represents damage score - the potential negative effect a mutation has on the binding affinity of a motif found to be occupied by a transcription factor in DNase-seq footprint data from primary erythroid cells. Each possible variation with respect to the haplotype is predicted. Clusters of high damage scores (peaks) indicate predicted, actively bound transcription factor binding sites in the context of open-chromatin in an erythroid tissue-specific manner.

A)



B)

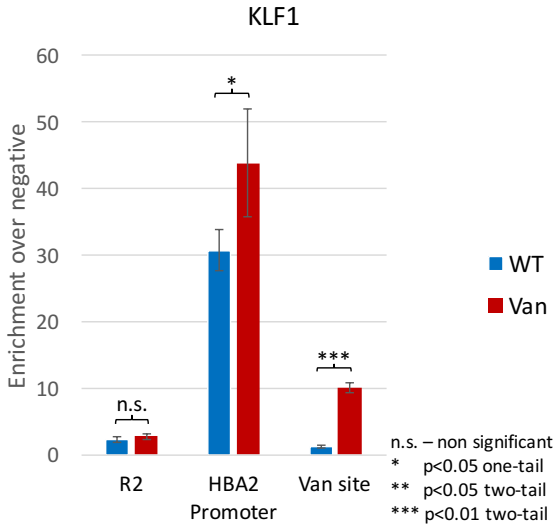


Figure 5.3.4 KLF1 occupancy in the α-globin locus in iPSC derived erythroid-like cells.
A) ChIP-seq tracks of KLF1 occupancy of the α-globin locus in day 21 iPSC derived erythroid-like cells from wild type cells (marked WT in blue) and homozygous Vanuatu mutant cells (marked Van in red), bottom track is genomic DNA input. All tracks are normalized to the same number of reads; Below the figure the α-globin enhancers are indicted with red arrows: R1 to R4 and the 138bp of unique sequence of the Vanuatu region is highlighted in light blue. Genes are above tracks in dark blue.
B) ChIP-qPCR results from KLF1 enrichment at the α-globin locus in day 21 iPSC derived erythroid-like cells from wild type cells in triplicate (marked WT in blue) and homozygous Vanuatu mutant cells in triplicate (marked Van in red); The enrichment is measured over enhancer R2, the promoter of the *HBA2* gene and the region around the Vanuatu mutation. Values are represented as enrichment over the average value of two negative sites found in an intergenic location on chromosome 16. Error bars represent one standard deviation. Statistical significance between samples is indicated by an asterisk above the bars on the histogram.

5.4 Analysis of open chromatin regions by ATAC-seq

ATAC-seq was used to assess the open chromatin regions in wild type and patient cells with the Vanuatu mutation. The assay was performed as described in Chapter 2, with input material of 60,000-75,000 of either undifferentiated iPSCs or iPSC derived erythroid-like cells harvested at day 21 of differentiation (fig. 5.4.1). Chromatin accessibility at the α-globin locus in undifferentiated iPSCs shows no activation of the upstream regulatory elements of the α-globin locus and little or no accessibility at the genes themselves in both wild type and patient cells homozygous for the mutant C allele. It is important also to note that just as in wild type cells there is no accessibility over the Vanuatu region in undifferentiated iPSC homozygous for the C allele. At day 21 of erythroid differentiation all lines show activation of the α-globin enhancers and genes, however, the patient lines show an additional peak over the Vanuatu region associated with activation of the GATAν transcriptional unit. The peak of chromatin accessibility over the Vanuatu site in the heterozygous mutant line is noticeably smaller that the peak observed in the homozygous mutant in agreement with the mutation dose-dependent nature of the phenotype. These data demonstrate that upon differentiation along the erythroid lineage, as the α-globin locus is activated, so is the GATAν transcriptional unit containing the mutant C allele suggesting a tissue specific nature for the new transcribed site.

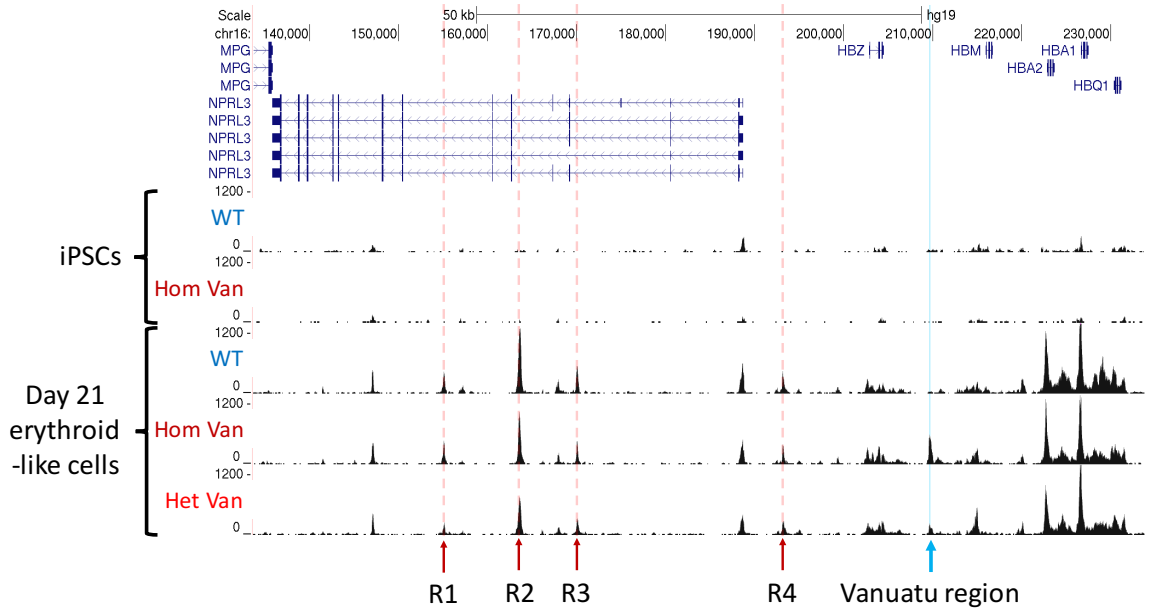


Figure 5.4.1 Chromatin accessibility in the α-globin locus.
UCSC tracks showing chromatin accessibility as assayed by ATAC-seq in the α-globin locus.
First two tracks show chromatin accessibility of undifferentiated iPSCs; last three tracks show chromatin accessibility in day 21 iPSC-derived erythroid-like cells; wild type is labeled as WT in blue, homozygous mutant for the Vanuatu variant is labeled as Hom Van in dark red, heterozygous mutant is labeled Het Van in light red; Below the figure the α-globin enhancers are indicated with red arrows: R1 to R4; Above the tracks the positions of the genes are indicated in blue;

5.5 Chromatin conformation analysis (Capture C)

Chromatin interactions in the α-globin locus of iPSC-derived erythroid–like cells at day 21 of differentiation were measured by Capture-C. The assay was performed on wild type and patient cells with the Vanuatu mutation, as described in Chapter 2, with input material of $1x10^7$ cells

When capturing with a probe that hybridizes to the Vanuatu region, the Capture C data show that in wild type cells the site of the Vanuatu mutation doesn't interact very strongly with the upstream regulatory elements in the α-globin locus, however in the patient cells, homozygous for the mutant C allele, there is a strong interaction between the region surrounding the Vanuatu variant and the α-globin enhancers (*fig.* 5.5.1). This is best visualized by subtracting the profile of the wild type cells from the profile of the patient cells (*fig.* 5.5.1 third track). In the subtraction tracks, a positive value indicates an increase in interaction in the patient line homozygous for the mutant C allele compared to the interaction found in a wild type line, and a negative value indicates a decrease in interaction in the patient line. Captures from the α-globin enhancers R1 and R2 confirm the increase in interaction between the enhancers and the Vanuatu region in cells from a patient homozygous for the C allele (*fig.* 5.5.2). In addition, the subtraction profiles of the R1 and R2 captures show that the increased interaction with the Vanuatu site is accompanied by a decrease in interaction between the α-globin genes and enhancers R1 and R2 in patient cells. This decrease in interaction between the α-globin genes and their enhancers has been confirmed in captures from the promoters of *HBA1* and *HBA2* (*fig.* 5.5.3).

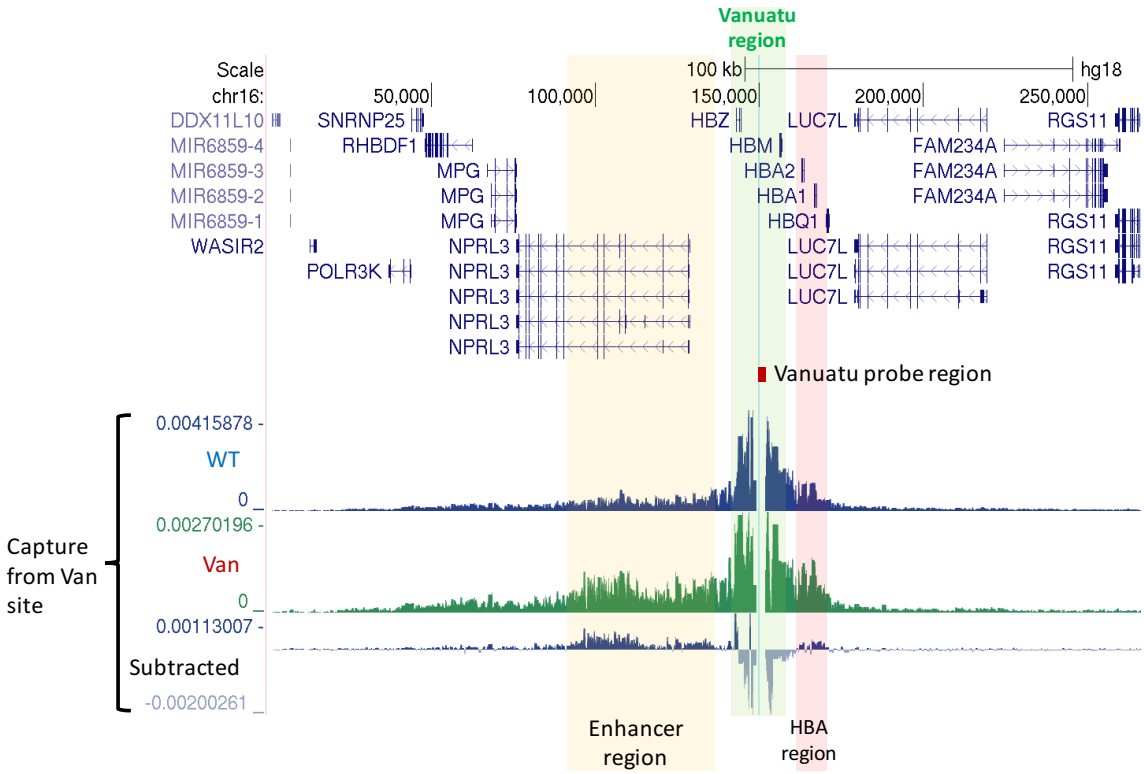


Figure 5.5.1 Capture-C from the Vanuatu region
Capture-C profiles of day 21 cells from the iPSC-based erythroid differentiation system. Positive values indicate higher level of interaction with the site of the Vanuatu mutation. First track shows chromatin interactions in wild type cells (WT). Second track shows chromatin interactions in mutant cells homozygous for the C allele (Van). Third track shows the profile of the patient line subtracted from the profile of the wild type line. The Vanuatu probe region is indicated in red above the top track, the regions include a 1 kb exclusion zone from which the signal has been removed. Genes are indicated above the track in blue; The region of the α-globin enhancers, the region surrounding the Vanuatu mutation and the region of the α-globin genes are highlighted in yellow, green and red respectively.

In the subtraction Capture-C tracks from captures from the Vanuatu site, in addition to the increased interaction with the α-globin enhancers the mutant C allele containing the Vanuatu site also appears to interact more strongly with the α-globin genes (*fig.* 5.5.1 third track). This suggests that the active GATAν transcriptional unit could be recruited into an open chromatin hub comprised of both the enhancers and α-globin genes from

which the Vanuatu site would normally be excluded when not accessible, in the absence of the mutation.

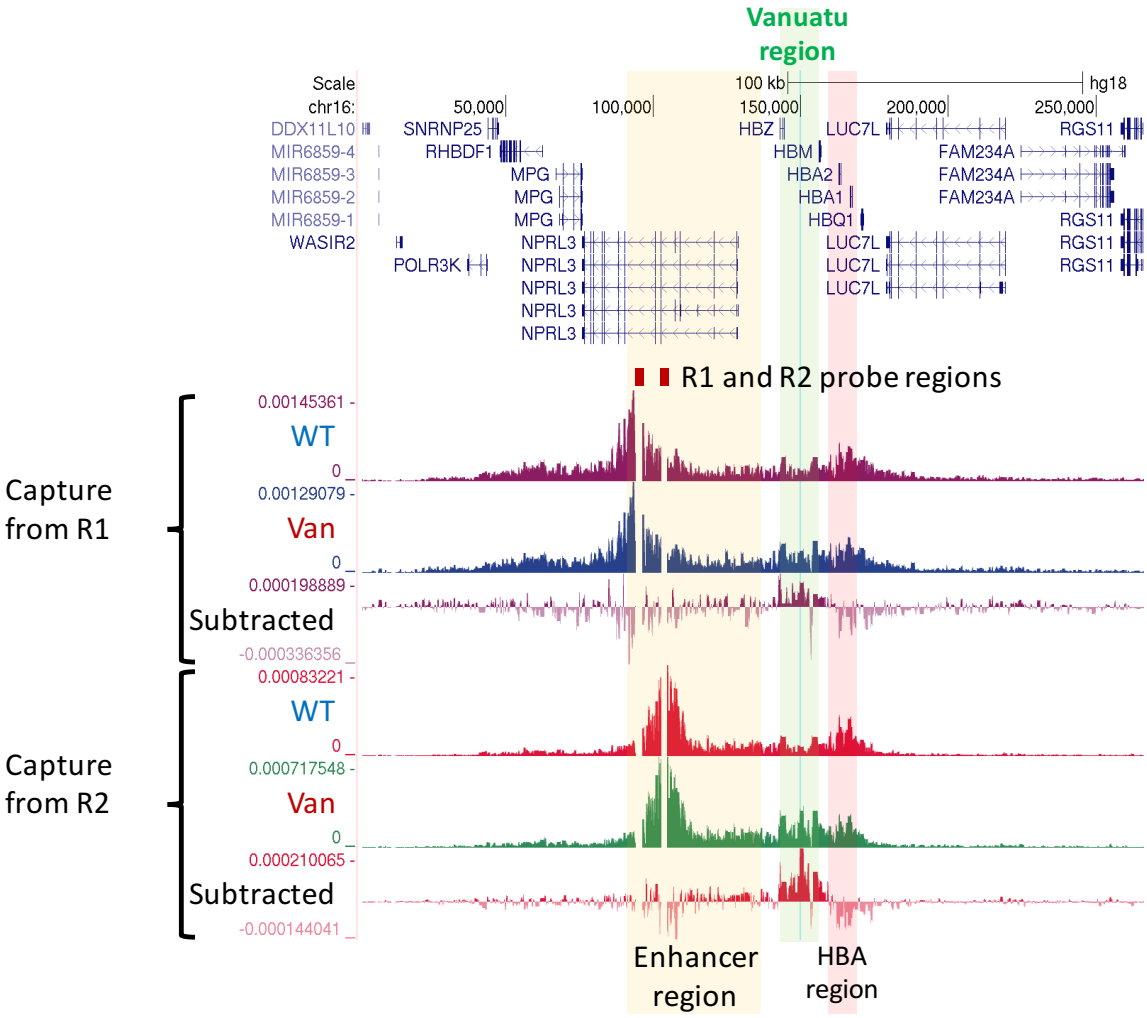


Figure 5.5.2 Capture-C from enhancers R1 and R2
Capture-C profiles of day 21 cells from the iPSC-based erythroid differentiation system. Positive values indicate higher level of interaction with the α-globin enhancers; The tracks are divided into two groups of threes: top group shows capture from R1, bottom group from R2. First track in a group shows chromatin interactions in wild type cells (WT); Second track in a group shows chromatin interactions in mutant cells homozygous for the C allele (Van). Third track in a group shows the profile of the patient line subtracted from the profile of the wild type line; The R1 and R2 probe regions (1kb exclusion zone) are indicated in red above the top track Genes are indicated above the track in blue. The region of the α-globin enhancers, the region surrounding the Vanuatu mutation and the region of the α-globin genes are highlighter in yellow, green and red respectively.

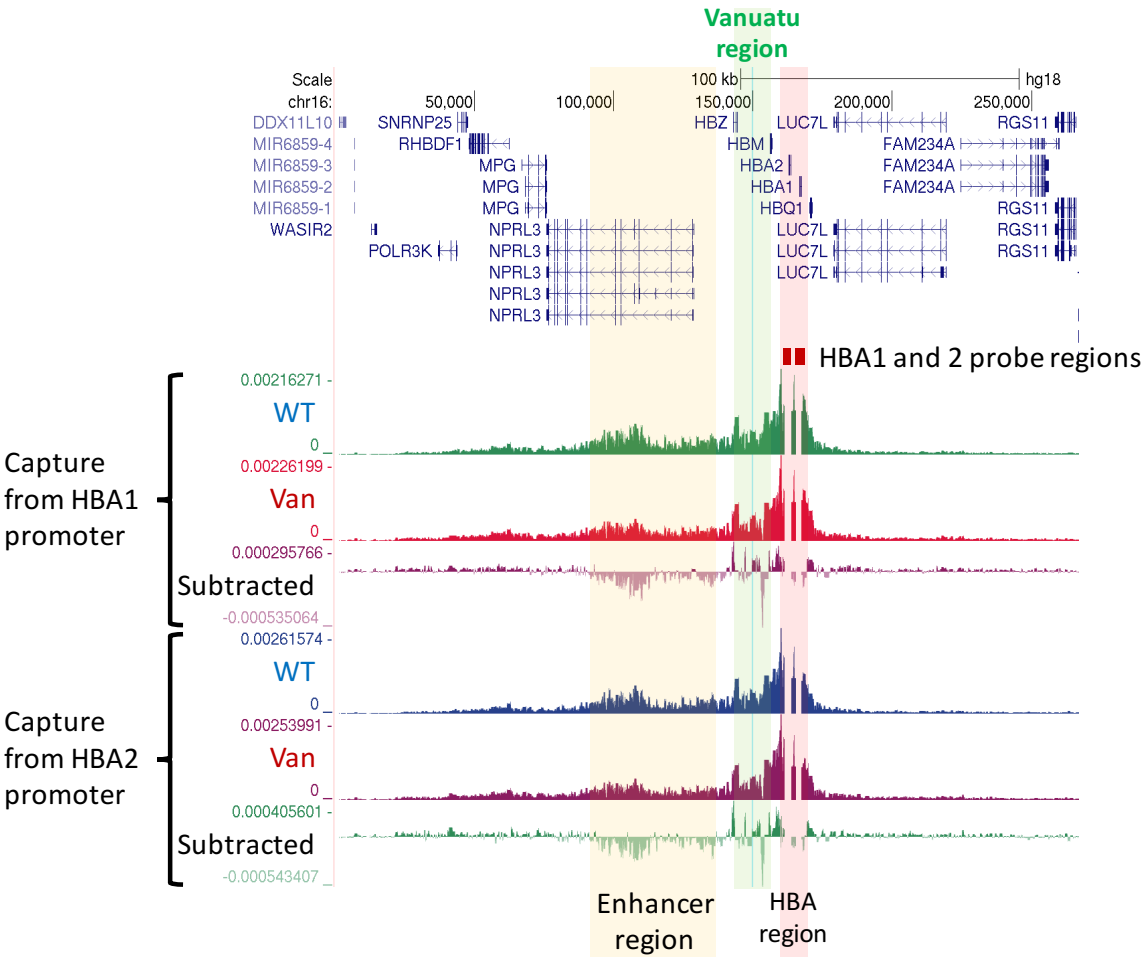


Figure 5.5.3 Capture C from *HBA1* and *HBA2* promoters

Capture C profiles of day 21 cells from the iPSC-based erythroid differentiation system. Positive values indicate higher level of interaction with the promoters of *HBA1* and *HBA2*; The tracks are divided into two groups of threes: top group shows capture from *HBA1*, bottom group from *HBA2*; First track in a group shows chromatin interactions in wild type cells (WT). Second track in a group shows chromatin interactions in mutant cells homozygous for the C allele (Van). Third track in a group shows the profile of the patient line subtracted from the profile of the wild type line. The HBA1 and HBA2 probe regions are indicated in red above the top track, the regions include a 1 kb exclusion zone from which the signal has been removed. Genes are indicated above the track in blue; The region of the α-globin enhancers, the region surrounding the Vanuatu mutation and the region of the α-globin genes are highlighter in yellow, green and red respectively.

5.6 Conclusion

Using an erythroid differentiation system to study the phenotype of a rare mutation described by De Gobbi et al., 2006 I have shown that iPSCs differentiated along the erythroid pathway produce cells that can recapitulate all the features of the disease that were described in primary erythroid cells.

During erythroid differentiation of iPSCs derived from patients homozygous for the mutant C allele of the Vanuatu variant, chromatin at the site of the mutation becomes accessible and transcriptionally active. This transcriptional activation is accompanied by the down regulation of α-globin. Similar to that shown for primary cells, GATA1 and RNA Pol II are recruited to the Vanuatu site in the iPSC derived erythroid–like cells that carry the mutant C allele. Additionally, ChIP experiments have demonstrated that the key erythroid transcription factor KLF1 also binds the site. This is of particular interest because, as discussed previously, a correlation between KLF1 and GATA1 ChIP-seq data has revealed extensive co-occupancy of KLF1 and GATA1 in the genome of erythroid cells. Coincidently, the half E-box-GATA1 motif at the heart of the Vanuatu site has also been identified in nearly a third of KLF1 ChIP-seq peak sequences which has led to the suggestion that an interaction between KLF1 and the TAL1-GATA1 containing complex may be a common feature at erythroid regulatory elements (Tallack et al., 2010). Together with the observation made by De Gobbi et al., 2006 that TAL1 is recruited to the C allele at the Vanuatu region in patient primary erythroid cells, this suggests that the particular motif grammar found at the Vanuatu site might create a strong erythroid-

specific transcriptional platform. The potential of this unit to be transcriptionally upregulated by other erythroid specific enhancers in the context of the erythroid tissue could be worth exploring.

Understanding the mechanism behind the transcriptional downregulation that is caused by the Vanuatu mutation is one of the main topics of this thesis. To achieve this, a better understanding of the mode of regulation exerted on the GATAν unit was necessary. The iPSC-based system was used to demonstrate that the novel transcriptional unit is enriched in the promoter mark H3K4me3 suggesting that it is a promoter. Additionally, the GATAν promoter-like element was found to interact with the enhancers of the α-globin locus and this in turn causes a reduction in the interaction between the enhancers and α-globin promoters in patient cells. This reduced interaction could explain the downregulation of α-globin observed in patient-derived erythroid cells and is consistent with the reduced levels of RNA Pol II at the promoter of HBA2 observed in patient cells differentiated in the iPSC system.

In conclusion, using an iPSC-based erythroid differentiation system I have modeled the phenotype caused by a rare mutation that generates a new transcriptional unit in the α-globin locus. ChIP data has shown that this new unit is bound by key erythroid transcription factors and has the chromatin signature of a promoter. Chromatin interaction analysis has provided evidence that this new transcriptional unit is under the regulation of the α-globin enhancers and competes with the α-globin promoters for enhancer activity. This competition could be the underlying cause of α-thalassemia observed in the individuals that carry the rare mutation.

Chapter 6: Editing human iPSC lines

6.1 Introduction to CRISPR

Sequence-specific programmable engineered nucleases have revolutionized biomedical research with wide applications in all aspects of genomics, including as tools for editing genomes to investigate gene regulation. These nucleases work by creating a double strand break (DSB) at chosen locations, subsequently the DSB is repaired by the cell and depending on the repair pathway can result in the creation of mutations at the targeted sites that can be designed to disrupt the function of a sequence or generate sequence-specific changes to the genetic material. The engineered nucleases that are currently used for gene editing can be classified into three main families: Zinc finger nucleases (ZFNs) and Transcription Activator-Like Effector Nucleases (TALENs) which function through protein–DNA interactions, and the RNA-guided nucleases clustered, regularly interspaced, short palindromic repeat (CRISPR)/CRISPR-associated (Cas). The main advantages programmable nucleases have compared to traditional genome editing include the increased precision of targeting, reduced off-target effects, and high efficiency of mutagenesis (Kim and Kim, 2014). The recently developed CRISPR-Cas system which employs the endonuclease Cas9 guided by small RNAs through base pairing with target DNA (Garneau et al., 2010) is easy to design, highly specific and efficient (Ran et al., 2013). For these reasons, I have chosen this approach to generate single base pair changes in iPSCs and thereby repair the Vanuatu mutation in patient-derived cell lines. Similarly, I have used this to create the mutation in wild type cell lines. This chapter discusses the genome editing strategy and generation of single base pair substitutions in iPSCs to either repair or generate the Vanuatu mutation.

CRISPR serves as an adaptive immune system found in many bacteria and the majority of characterized archaea. It works by acquiring short sequences of DNA from invading bacteriophages and plasmids, these are then transcribed into CRISPR RNAs (crRNAs) used to target the endonuclease activity of the system to the invading nucleic acid (Bolotin et al., 2005; Brouns et al., 2008; Garneau et al., 2010; Ishino et al., 1987; Koonin and Krupovic, 2015; Marraffini and Sontheimer, 2008; Mojica et al., 2005; Pourcel et al., 2005; Rath et al., 2015). Because the sequence specific nuclease activity of Cas9 is dependent largely on complementarity between the guide RNA and the targeted DNA sequence, it can be directed to new sites, simply by changing the sequence of its guiding RNA. For this reason, its programmable nuclease activity was engineered for genome editing. In nature, Cas9 finds its target site using an RNA duplex composed of the crRNA and the tracrRNA (trans-activating CRISPR RNA). The crRNA recognizes its target sequence through complementarity using around 20bp of Watson-Crick base-pairing, while the tracrRNA hybridizes with the crRNA and is incorporated with it into the nuclease (Barrangou et al., 2007; Deltcheva et al., 2011; Garneau et al., 2010; Gasiunas et al., 2012; Hale et al., 2009; Jinek et al., 2012). To repurpose the system for genome editing a vector was designed to ectopically expressed the Cas9 protein and a chimeric single guide RNA referred to as an sgRNA that was made by the fusion of the crRNA and tracrRNA. This was used for the generation of site specific DSBs (Jinek et al., 2012).

The only limiting feature in terms of targeting the nuclease activity is that the recognition site for the Cas9–sgRNA complex must be adjacent to a short sequence called a protospacer-adjacent motif (PAM) in order to induce cleavage within the base-pairing region. Genome engineering is most widely performed using Cas9 that has been adapted

from *Streptococcus pyogenes* (SP). It requires a relatively simple PAM sequence: NGG (or a weaker NAG) (Hsu et al., 2013) which is estimated to occur once every 8bp, on average, in the genome and cuts approximately 3nt away from the PAM (Jinek et al., 2012; Gasiunas et al., 2012;). This creates a platform that can be programmed for site-specific gene targeting by designing an approximately 20bp sequence of the sgRNA that recognizes the genomic target which must be adjacent to an NGG. When Cas9, guided by a specific sgRNA, recognises its target it creates a DSB which can stimulate repair through either the non-homologous end joining (NHEJ) pathway or the homology-directed repair (HDR) pathway. This promotes a significant increase in the rate of sequence modification at the targeted genomic location (Choulika et al., 1995; Plessis et al., 1992; Porteus, 2016; Rouet et al., 1994a, 1994b; Rudin et al., 1989). The error-prone NHEJ pathway causes insertional and deletional mutations (indels) at the region of the DSB (Lieber, 2010) while HDR can produce precise replacement of sequences through the use of a repair DNA template that contains the desired sequence changes (Choulika et al., 1995; Rudin et al., 1989).

The CRISPR technology has made the generation of disease models for genetic disorders much easier (reviewed in Doudna and Charpentier, 2014; Dow, 2015; Hsu et al., 2014) and, more importantly for the purpose of this thesis, many studies have published the use of Cas9-mediated genome editing in iPSCs (Li et al., 2015a, 2015b; Park et al., 2015; Smith et al., 2015; Song et al., 2015; Xie et al., 2014; Xu et al., 2015). Additionally, multiple reports have shown how the correction of a disease-associated mutation in iPSCs can improve the disease phenotype of differentiated cells (reviewed in Deleidi and Yu, 2016; Hotta and Yamanaka, 2015; Seah et al., 2015). Furthermore, the successful introduction

of mutations into wild-type iPSCs has also been described (Murai et al., 2016; Wen et al., 2014). Moreover, using CRISPR– Cas9 technology, researches have recently demonstrated the generation of genetically controlled isogenic iPSC lines between which the only variable is a specific disease-associated genetic risk variant. One such study described the generation of iPSC lines whose only difference was a disease-associated risk variant in a regulatory element and this enabled the functional dissection of the *cis*-acting effect of the variant (Soldner et al., 2016). These studies provide a framework around which to create a successful gene editing strategy for iPSC that can be used in differentiation experiments.

The CRISPR/Cas9 system has demonstrated great potential in the fields of basic research and therapeutics, however, there are still several aspects of the technology which require improvement, namely the specificity, the efficiency, and the spatio-temporal control of the nuclease activity. Despite the high efficiency of target-site DSB generation of Cas9, the initial reports for the efficiency of targeted genome modification in human iPSCs showed only 1–2% HDR-mediated editing (Mali et al., 2013; Yang et al., 2013). At this level the recovery of correctly targeted clones without introducing any form of selection becomes labour-intensive and inconsistent. Other studies have shown that the low transfection efficiency of the relatively large Cas9 expression vector is a major limiting factor on the yield of targeted clones in pluripotent stem cells (Li et al., 2014) and selection for Cas9-transfected cells, for example based on the co-expression of a fluorescent protein, increase the isolation of modified clones (Ding et al., 2013). However, cell sorting is not a preferred method of selection for stem cells as it causes unwanted stress to the cells and poses a contamination risks, thus other methods of

selection would be preferable. One more major concern with the conventionally utilised wild type Cas9 is the potential for off-target mutagenesis, as a sequence similar to the 20-bp targeting sgRNA sequence and the 3bp PAM may be found at a different location in the genome and could be recognized and cut by the nuclease (Fu et al., 2013; Hsu et al., 2013; Lin et al., 2014; O'Geen et al., 2015). However, the initial reports that had raised concerns about Cas9 off-target effects were primarily based on work done in immortalized cell lines; recent data from multiple whole genome sequencing studies done in normal primary cells, including iPSCs, show very little off-target mutagenesis and less cause for concern (Chang et al., 2015; Li et al., 2016; Park et al., 2015; Smith et al., 2014; Suzuki et al., 2014; Veres et al., 2014; Yang et al., 2014).

Because the work for this thesis required a smaller-scale editing experiment in human iPSC lines, I opted to use transient transfection to introduce a selectable plasmid expressing wild type Cas9 and a targeting sgRNA, plus a single stranded oligo donor (ssODN) to serve as a repair template to generate a single base pair substitution.

6.2 Editing strategy

The basic strategy to generate the single base pair substitution at the Vanuatu site in iPSCs as outlined in section 6.1 was to use a plasmid that expresses both the wild type Cas9 protein and sgRNA, and a drug resistance gene to select for cells which have taken up the plasmid and thus have site-specific nuclease activity generating a DSB at the Vanuatu site. A DNA template was used to promote homology directed repair of the break at the target site. This template consists of a short single-stranded oligonucleotide which contains the edited base. After the plasmid and oligo were introduced to the cells they were exposed to drug selection and then the selected cells were separated and expanded as single cells to form clonal lines. The clonal lines were screened for the presence of the desired mutation as well as any additional mutations that might have arisen at the α-globin locus during the editing process.

Several stages of this process needed to be optimized to successfully generate scarless single nucleotide editing in human iPSCs that could be used for differentiation experiments. The factors which determine the success and efficiency of editing are: the selection of plasmid and type of donor template; the design of the sgRNA, and the donor template; the method of delivering the exogenous DNA, and the screening process.

Choice of plasmid and donor template

The Cas9 plasmid selected was the px459 v2.0 which featured a U6-promoter-expressed sgRNA and a hybrid CMV enhancer/chicken β-actin promoter driving expression of Cas9 that is connected by the self-cleaving peptide P2A sequence to a protein that confers resistance to puromycin: puromycin-N-acetyltransferase (*fig*. 6.2.1) (Ran et al., 2013). Puromycin was chosen as a selectable marker as previous evidence indicated that even prolonged exposure of human iPSCs to puromycin did not affect the pluripotency of the cells (Ye et al., 2014). The choice of using a single-stranded oligonucleotide donor was also based on a previous successful study which managed to generate scarless substitutions of short DNA sequences in human iPSCs, albeit at relatively low HDR efficiency rates (Yang et al., 2013).

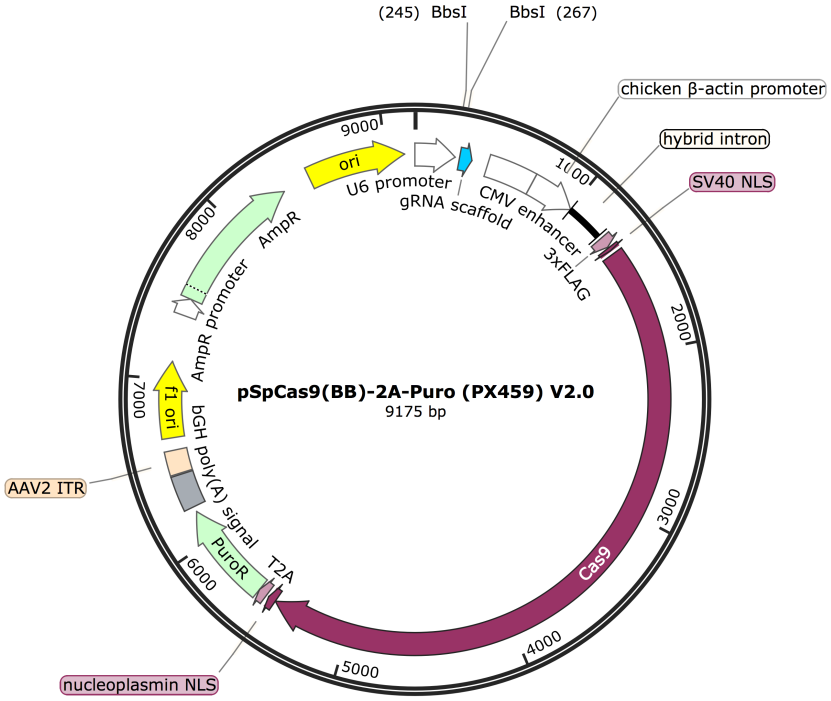


Figure 6.2.1 Map of px459 v2.0 plasmid
Map was generated using SnapGene software from the px459v2.0 sequence from Addgene-plasmid reference number 62988

Optimization of DNA delivery

To ensure the optimal DNA uptake by the iPSCs several nucleic acid delivery methods were tested (*fig.* 6.2.2). The results from the electroporation-based methods showed that both Amaxa 4D nucleofection and Neon transfection (see details of conditions in Chapter 2) were sensitive to the size of the transfected DNA. Electroporation seemed to be more efficient for the small 3.5kb pmaxGFP plasmid (Lonza) than the large 9.3kb Cas9 and GFP expressing px458 plasmid (Addgene ref.48138). However, this effect wasn't observed in the lipofection based method using Fugene 6 where both the large and small plasmids were taken up with similar efficiencies. Thus, Fugene 6 was chosen to introduce the Cas9 expressing, puromycin-selectable plasmid and ssODN into the human iPSC lines.

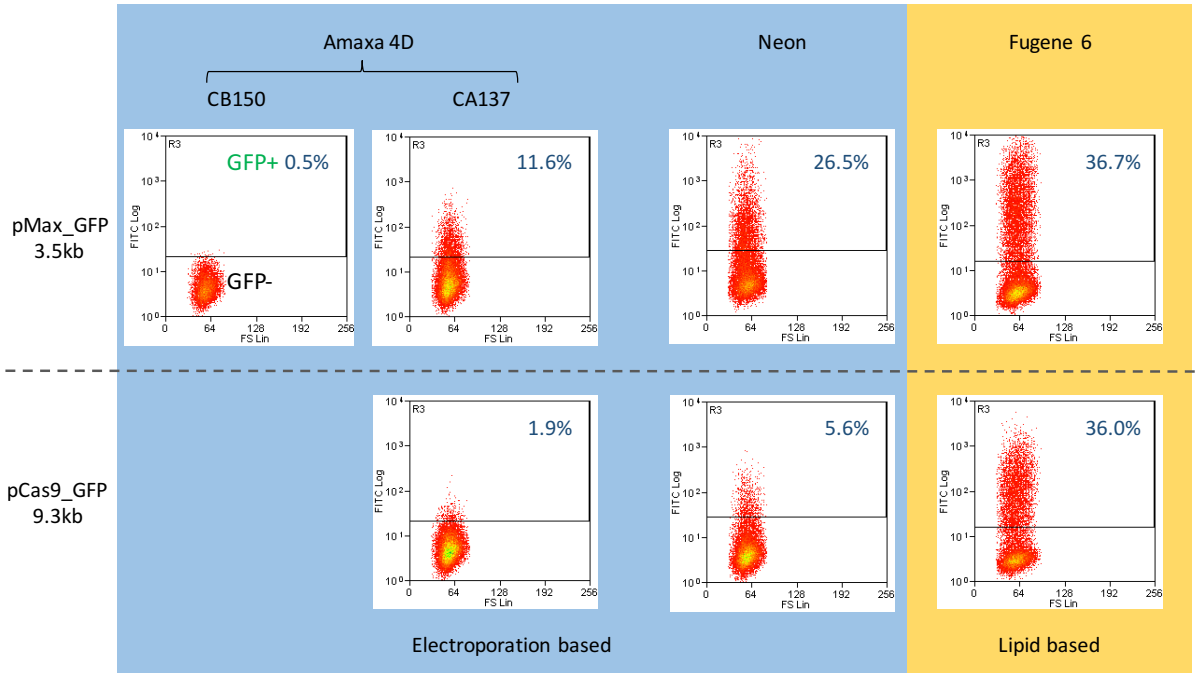


Figure 6.2.2 Optimization of DNA delivery in a wild type human iPSC line using GFP expressing plasmids.
Top panel shows experiments carried out with pMax_GFP plasmid (pmaxGFP from lonza), bottom panel shows experiments carried out with pCas9_GFP plasmid (px458). Electroporation-based transfections are shown on a blue background and lipid based transfections (lipofection) on orange. Data are presented in the form of density plots from FACS analysis measuring the proportion of GFP positive cells in a population 24h post transfection. The values for GFP positive cells are given in percentages in the top right-hand corner of the plot.

Screening strategy

A quick and efficient screening method was needed to distinguish between the mutant C allele or the wild type T allele at the Vanuatu site. Fortuitously, the Vanuatu mutation turns the wild type TTAGA sequence to TCAGA creating a recognition site for the restriction enzyme Hpy188I which cuts DNA at TCNGA sequences. This observation was used to design a restriction digestion-based screening strategy (*fig.* 6.2.3) that employed the generation of a PCR product using primers that flank the Hpy188I site: 191bp upstream and 103bp downstream. The PCR product was subsequently digested by

Hpy188I and the results were visualised using gel electrophoresis. This approach was easy to implement in large scale screening strategies of up to 96 individual clones at a time.

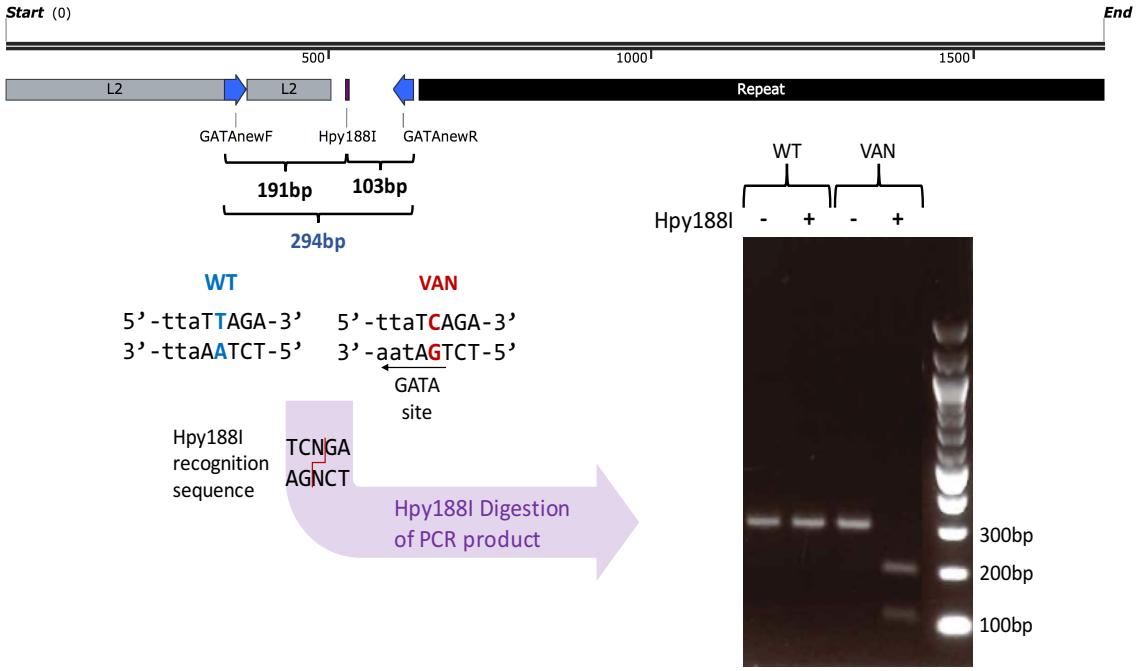


Figure 6.2.3. PCR product restriction digestion screening method.
The region around the Vanuatu site including the upstream LINE2 (L2) in grey and downstream repeat in black are presented on the top. The PCR primers flanking the Vanuatu site are shown as blue arrows labelled GATAnewF/R. Below are the sizes between the 5' ends of the primers and the Hpy188I cut site and the whole size of the PCR product. Underneath the PCR product sizes are the wild type (WT in blue) and mutant (VAN in red) sequences at the Hpy188I cut site. The cut site recognition sequence is in capitals. The gel image on the bottom right shows the result of the digestion reaction of PCR products generated from either a wild type (WT) or mutant line homozygous for the C allele (VAN).

sgRNA and ssODN design

The components of the gene editing system that have the biggest impact on the editing efficiency are the guide RNA and the repair template. Sequence-specific targeting of Cas9 to a genomic locus can be accomplished by designing an sgRNA that can pair with the target sequence which must be adjacent to a PAM. When designing sgRNAs, studies have

shown that there are some rules that should be followed to achieve optimal results. Firstly, all sgRNAs need to be followed by a PAM site, for Sp Cas9 the optimal sequence would be NGG (Hsu et al., 2013; Jinek et al., 2012). If the sgRNA is expressed from a U6 promoter the first nucleotide of the guide must be a G to ensure high level transcription, although A has been suggested to work in some contexts (Ma et al., 2014). It has been demonstrated that the Cas9 nuclease cuts with high efficiency even when there is a mismatch between the 5' G of the sgRNA and the target sequence. This feature of the nuclease is convenient in the cases where there is no G 18–25 bp upstream of NGG at the targeted region (Hwang et al., 2013; Shalem et al., 2014). Additionally, if a U6 promoter is used for expression, the sequence of the sgRNA cannot contain four or more successive U's as this causes a premature termination of the transcript (Nielsen et al., 2013). Shorter stretches of U's close to the 3' end of the sgRNA have also been shown to be unfavourable for Cas9 activity (Wang et al., 2014). In addition, long repeats of the same nucleotide should also be avoided as they can greatly decrease on-target nuclease activity (Gilbert et al., 2014).

With these rules in mind and by using the WGE genome editing tool which can filter through all possible 20nt gRNA adjacent to a PAM sequence using user defined parameters (Hodgkins et al., 2015) several guides were identified. Two of these (*fig.* 6.2.4) were selected for testing. The first guide, sgRNA1, had the lowest off-target effect prediction and cuts closest to the Vanuatu mutation: 4bp away. This feature is particularly desirable as previous research, done in iPSCs in particular, has shown that the efficiency of HDR-mediated base editing is inversely proportional to the distance from the cut site (Paquet et al., 2016). However, sgRNA1 contained multiple Us in addition to a stretch of

four Us in its sequence and therefore needed to be assessed for editing efficiency alongside another sgRNA. The other guide - sgRNA3 had an only slightly higher off-target effect prediction and didn't contain many Us in its sequence. SgRNA3 cuts 13bp away from the Vanuatu mutation but it didn't start with a G so the first base of the guide was replaced with a G. The predicted off-target regions can be found in Appendix *fig.* A6.2.1. Separate sgRNAs were designed to target each allele, one corresponding to the wild type T allele (labelled sgRNAt1 and sgRNAt3) and one corresponding to the mutant C allele (labelled sgRNAc1 and sgRNAc3). The sgRNA were cloned into the px459 v2.0 plasmid as described in Chapter 2 Materials and Methods.

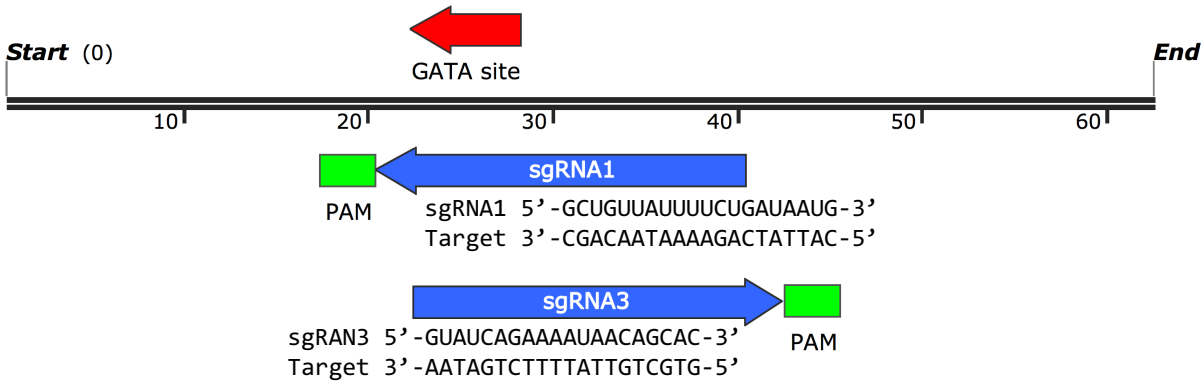


Figure 6.2.4. Short guide RNA for genome editing in iPSCs
The sgRNAs are shown in blue, the PAM motifs in green and the GATA1 site at the Vanuatu region is in red. Arrows indicate the orientation of each feature 5' to 3'. The RNA sequence and corresponding DNA target sequence of each guide are shown below the blue arrow. Showing sgRNAs for C allele.

The design of the ssODN was based on a study demonstrating that before complete dissociation of Cas9 from its DNA target, the 3' end of the cleaved strand that is not complementary to the sgRNA (i.e. the non-target strand) is released - creating a free 3'

single stranded DNA overhang (Richardson et al., 2016). By designing their single-stranded DNA donors to have a specific length of complementarity to the strand that is released first, the authors managed to increase the efficiency of HDR-based genome editing up to 60% in human cells. More specifically, they found that an asymmetric ssODN designed to anneal with 36bp to the released 3' strand that is found on the PAM-distal side of the cut, and 91bp of homology on the PAM-proximal side of the cut (*fig.* 6.2.5), supported high HDR frequencies. An additional option to the asymmetric design of the oligo was the formation of phosphorothioate bonds between the three nucleotides at the 5' and 3' end of the oligo. This type of modification has been used to protect the ssDNA from exonuclease digestion inside cells (Putney et al., 1981), which prolongs the duration over which the template is available to the cell for repair.

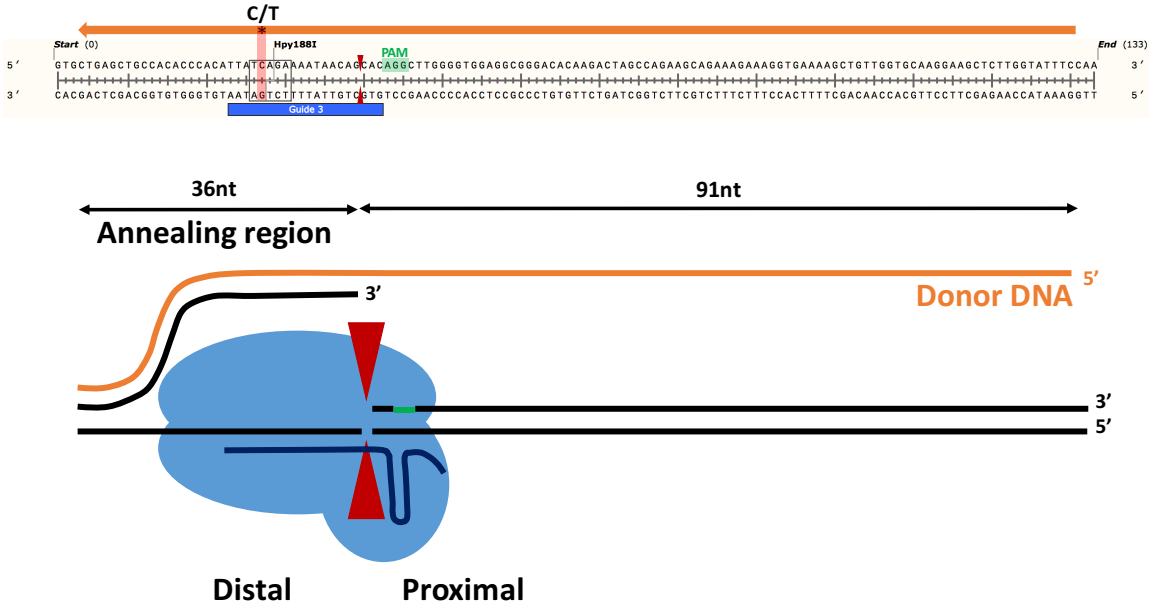


Figure 6.2.5. ssODN design for sgRNA3.
Top panel shows the genomic sequence for a part of the Vanuatu region. The mutant base pair is highlighted in red. On the bottom, a cartoon representation of the Cas9 DNA complex with the released PAM-distal 3' end which has base paired with the short 36bp stretch of the asymmetrical

ssODN. The sgRNA3 is shown in dark blue, Cas9 is in light blue, the PAM motifs in green and the cut site is indicated with red arrows, ssODN is in orange. Mutant C allele sequence is shown. Figure is based on Richardson et al., 2016.

Testing puromycin susceptibility in iPSCs

Before using puromycin for selection in iPSCs, the minimal concentration of the drug required to kill all of the cells present in a 70-80% confluent culture dish in 24h needed to be determined. This concentration was found to be around 1μg/ml based on confluency of an iPSCs monolayer after 24h of puromycin treatment (*tbl.* 6.2.1); the same concentration that was reported by Ye et al., 2014.

Puromycin (μg/ml)	**0.3**	**0.5**	**1.0**	**2.0**
Cells remaining after 24h	40-60%	10-20%	<1%	0%

Table 6.2.1 Puromycin susceptibility of iPSCs

Testing sgRNAs, oligo designs and different puromycin concentrations

In order to select the final conditions for the editing strategy, the efficiency of editing in cells after puromycin selection needed to be assessed. Lipofection was used to deliver different combinations of editing reagents into patient-derived iPSCs homozygous for the mutant C allele. The DNA delivered was the px459 v2.0 plasmid expressing either sgRNA1c or sgRNA3c that recognize the C allele and an asymmetric ssODN that had the T allele and was either phosphorothioated or not modified. 30h after lipofection the cells were subjected to puromycin selection for 24h using two different concentrations: 0.5 or 1μg/ml. Following selection, the cells were expanded and a proportion was harvested for DNA, while the majority were stored in liquid Nitrogen. The efficiency of editing was

analysed using the quick PCR product restriction digestion method that employs Hpy188I (*fig.* 6.2.6), outlined previously. The data shows that the use of a phosphorothioated oligo template when converting the C allele to a T generated more alleles protected from Hpy188I digestion than when using an unmodified oligo. The two guide RNAs were tested and sgRNA3 performed significantly better than sgRNA1, and the use of the higher dose of 1μg/ml puromycin for selection resulted in the generation of more protected alleles. To confirm that the editing in the bulk cell population was genuine and to quantify what proportion of the alleles were edited in a scarless manner Next Generation Sequencing (NGS) was performed. The NGS libraries were produced from the 294bp PCR amplification product (GATAnew) from the Vanuatu site (for details see Chapter 2) from patient cells edited using sgRNA3 and a phosphorothioated donor oligo and selected using 1μg/ml puromycin (*fig.* 6.2.7 LA01 C to T). This demonstrates that the protection of the Hpy188I site is not due to indels and that the editing efficiency for scarless allele conversion in the homozygous mutant patient iPS cell line (LA01) was over 27%. The same was performed for the wild type iPSCs (SBAD-02-01) but converting the T allele to a C, yielding HDR efficiency over 44% (*fig.* 6.2.7 SBAD-02-01 T to C). These high rates of HDR-mediated editing indicate that, assuming a normal random distribution of the alleles in the cell population, fewer than 48 clones (half a 96 well plate) would need to be isolated in order to establish a triplicate of homozygous edited iPSC clonal lines.

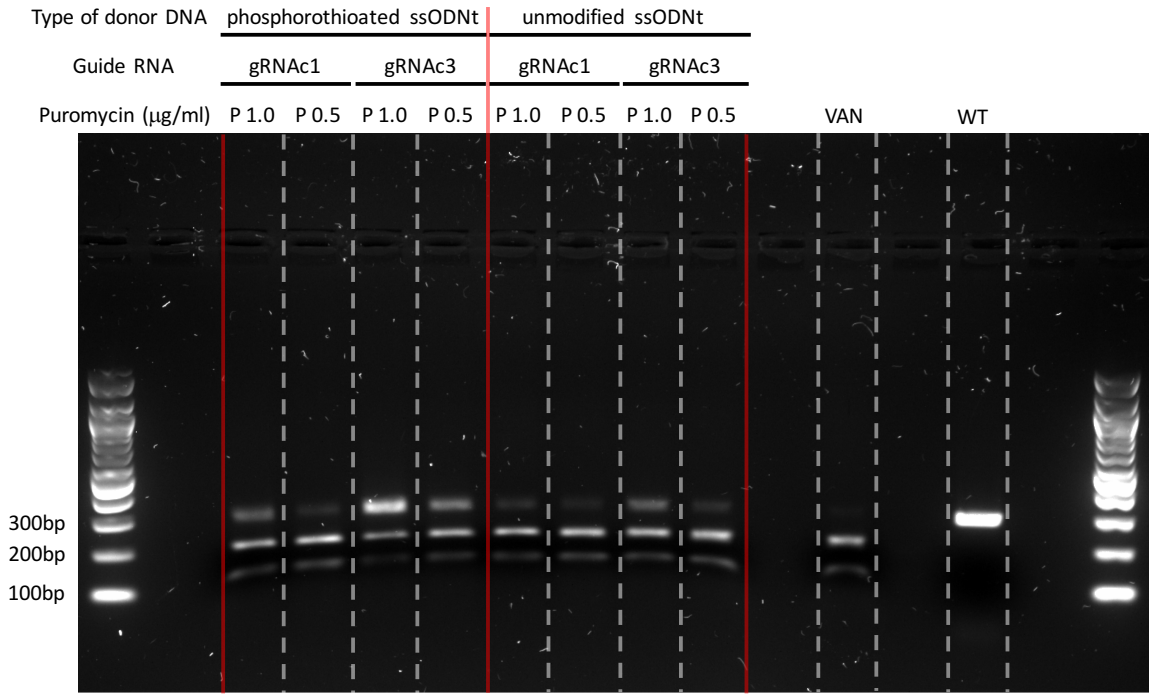


Figure 6.2.6 Optimizing guide, template and puromycin concentration
PCR product restriction digestion screening performed on bulk population of edited cells after selection. Top panel shows the different combinations of reagents: sgRNAc1 (G1) or sgRNAc3 (G3) using either phosphorothioated or unmodified ssODN and puromycin selection at two different concentrations (0.5 and 1.0 μg/ml). The particular editing experiment depicted was the repair of the mutant C allele to a T in a homozygous patient iPSC line (VAN).

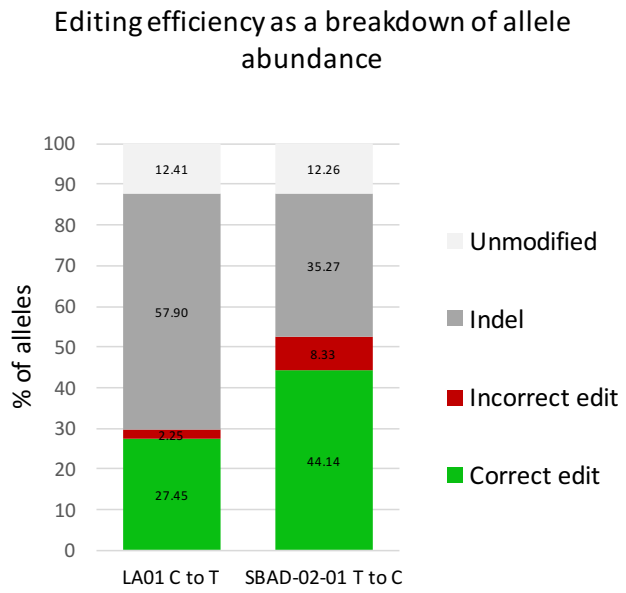


Figure 6.2.7 Editing efficiency as a breakdown of allele abundance (reads in NGS library) Unmodified (light grey) is the percentage of reads with the original sequence. Indel (dark grey) is the percentage of reads with a mutation other than the targeted single base pair substitution sequence. Incorrect edit (red) is the percentage of reads that have an additional mutation on top of the targeted substitution; Correct edit (green) is the percentage of alleles that only have the targeted substitution.

Complete editing strategy

The complete editing strategy (*fig.* 6.2.8) starts with the lipofection based delivery of the px459 v2.0 plasmid expressing sgRNA3 that is specific for either the C or T allele (sgRNAc3 or sgRNAt3) together with the appropriate phosphorothioated ssODN containing the appropriate T or C allele substitution. After 30h the cells were subjected to selection with 1μg/ml puromycin for 24h. The selected cells were harvested and divided into 3, one lot was stored in liquid nitrogen; one lot used for DNA isolation, and one lot was plated at a low density as single cells. The single cells were expanded to form colonies which are picked into wells of a 96 well plate. The plate was split in two replica plates one of which was harvested for DNA and screening and the other expanded or frozen in cryostorage.

The correctly edited lines were expanded and assessed for large mutations in the α-globin locus that might have been caused by the editing process.

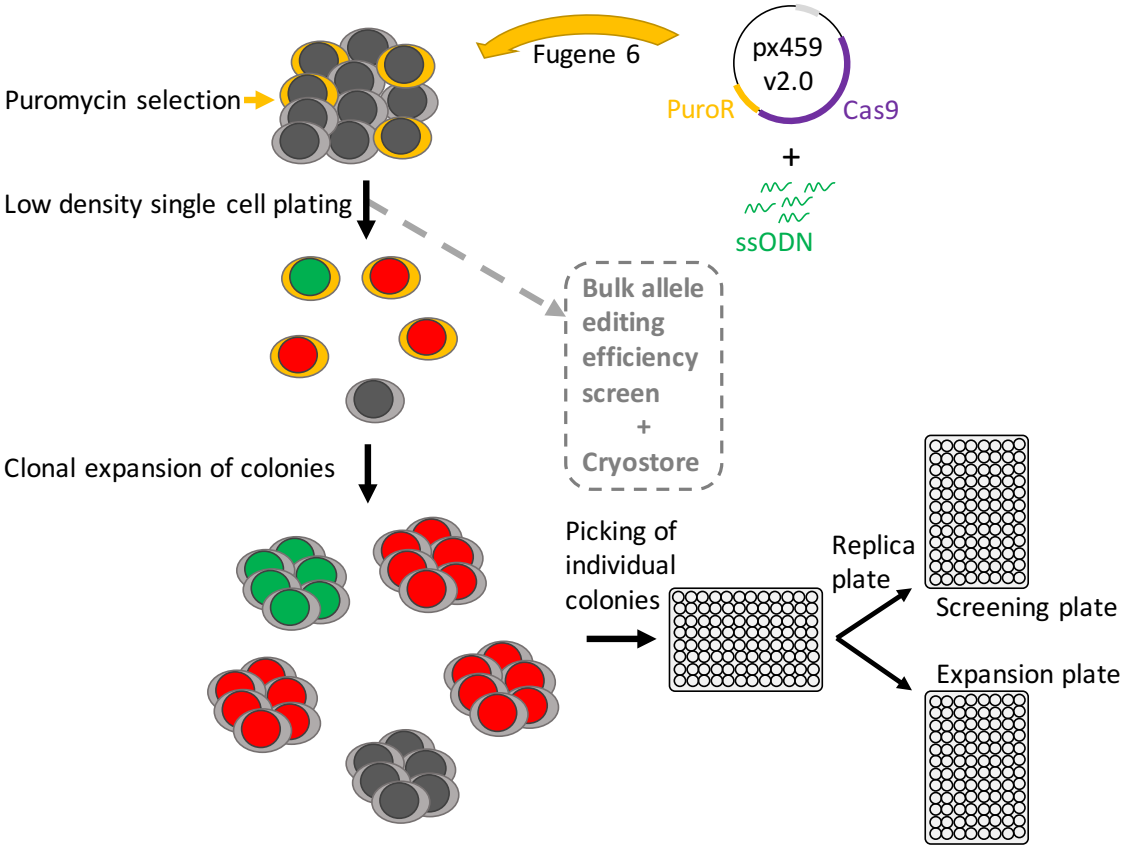


Figure 6.2.8 Schematic of complete editing strategy

6.3 Editing iPSC lines

To establish clean, genome edited iPSC clones it is necessary to have a quick, efficient and high-throughput method of screening large numbers of colonies. The Vanuatu PCR restriction digestion screen assay was scaled up to a 96-clone format (see Chapter 2) (*fig*. 6.3.1 showing the T to C allele conversion). After analysing the clones derived from editing of both the wild type and patient lines the rate of obtaining homozygous edited clones was found to be between 22-38% with around 5/6 of these being scarless edits, as assessed by Sanger sequencing, and only 16-26% of the clones were heterozygous edits (see Appendix *fig*. A6.3.1).

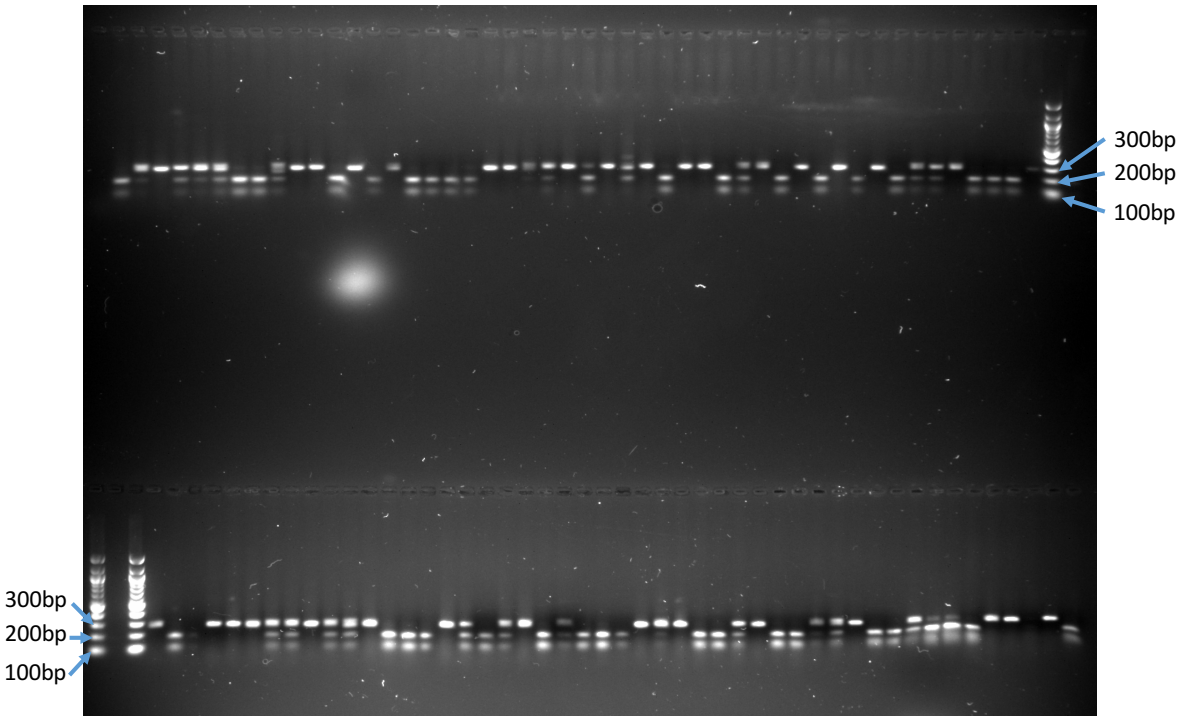


Figure 6.3.1 Recreating the Vanuatu mutation in a wild type iPSC line (T to C allele conversion) Vanuatu PCR restriction digestion of 96 individual clones grown from single cells.

In order to validate the scarless homozygous editing of the isolated iPSC clones three strategies for quality control were employed (*fig.6.3.2*) (see Chapter 2 for details). The first was a small scale but high-resolution screen based on Sanger sequencing of the 294bp amplicon generated from the Vanuatu site PCR using the GATAnewF/R primer set. Second, a long-range PCR strategy was employed to check for any medium sized mutations that would extend outside of the 294bp amplicon. Lastly, Southern blot was performed to check for large genomic aberrations in the region of the Vanuatu site in the isolated clones using a 150bp probe located around 1.2 kb away from the cut site of sgRNA3.

Sanger sequencing was used to determine whether a small indel was responsible for the loss of the restriction site rather than the intended base substitution. Such an indel at one of the alleles would result in a frameshift and a composite sequence of intensities in the Sanger sequencing chromatogram after the site of the mutation. Using long-range PCR and Southern blot the presence of a larger mutation that extends outside the 294bp Vanuatu site amplicon would result in the appearance of an additional band in the lane of the respective clone.

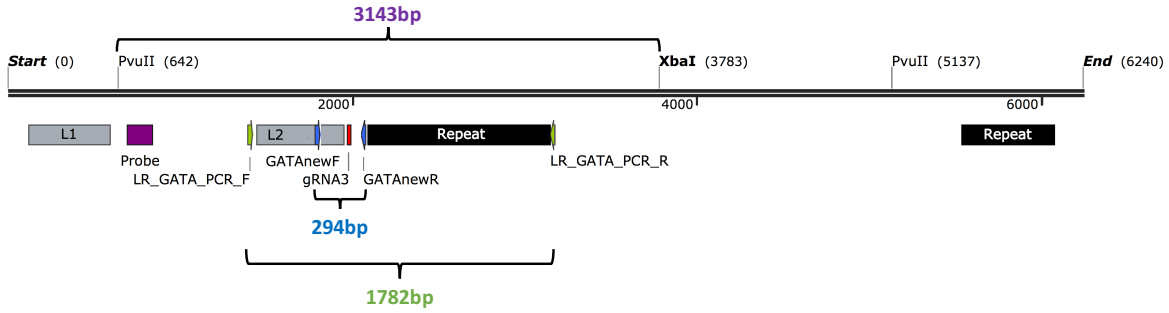


Figure 6.3.2. Screening for small, medium and large sized genomic abnormalities in the α-globin locus of edited cells.
The region (6240bp) around the Vanuatu site including the two upstream LINE elements: LINE2 (L2) and LINE1(L1) in grey; and two downstream simple repeat elements in black; plus, gRNA3 in red. The PCR primers flanking the Vanuatu site are shown as blue arrows labelled GATAnewF/R, below is the size of their PCR product (294bp). The PCR primers flanking the Vanuatu site, L2 and simple repeat elements are shown as green arrows labelled LR_GATA_PCR_F/R, below is the size of their PCR product (1782bp). The Southern blot probe is shown in purple, the restriction fragment size that would be bound by the probe after digestion with PvuII and XbaI is indicated on the top in purple (3143bp).

Sanger sequencing was performed on all of the isolated clones from edited wild type and patient cell lines, which were expanded without major signs of differentiation. Four clones from each allele conversion experiment were selected for differentiation along the erythroid pathway. The chromatograms of the reverse strand for the T to C conversion in wild type and the C to T conversion in patient iPSC lines are show in *fig.*6.3.3 and *fig.*6.3.4 respectively. All the chromatograms show a clear change to the corresponding edited allele with no detectable signs of contamination from a non-edited base.

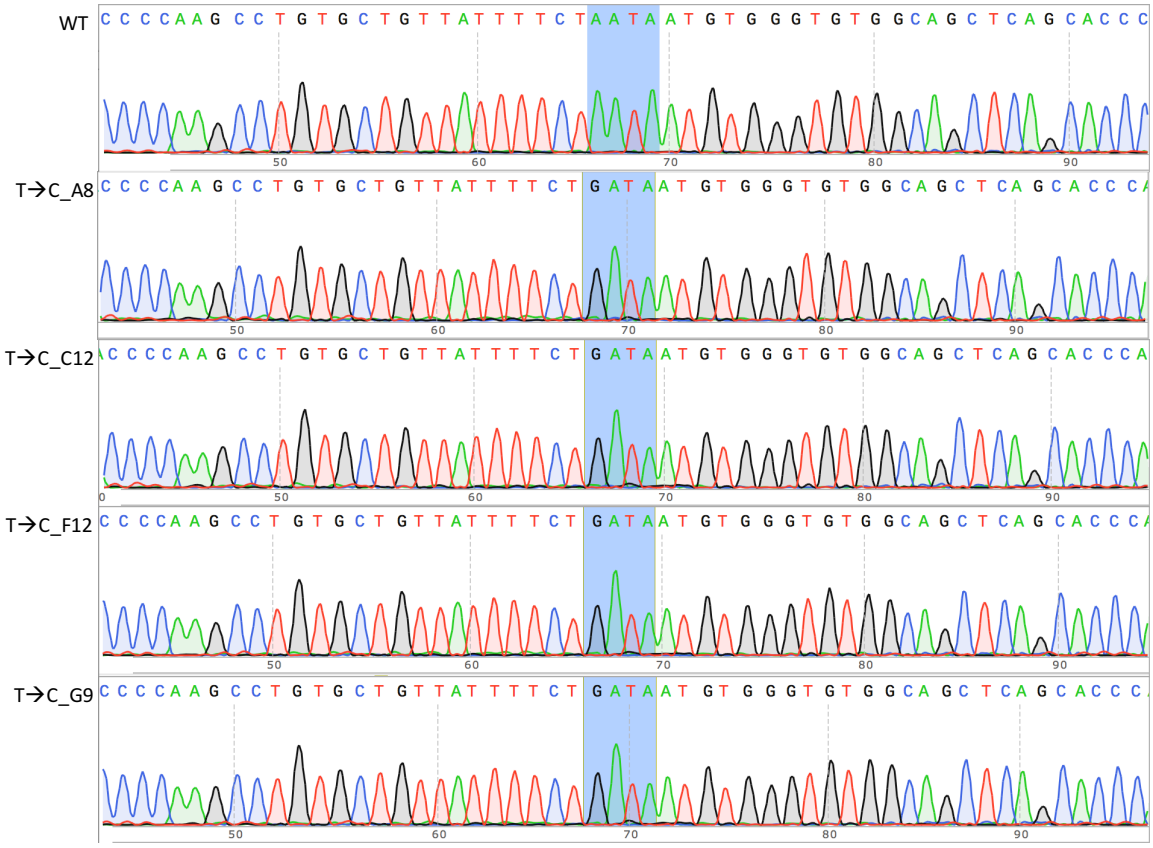


Figure 6.3.3. Chromatogram data from Sanger sequencing on T to C conversion lines.
Core GATA site shown 5’ to 3’ and is highlighted in blue; Names of lines are on the left side; Top chromatogram is from the same cell line (wild type-WT) before editing.

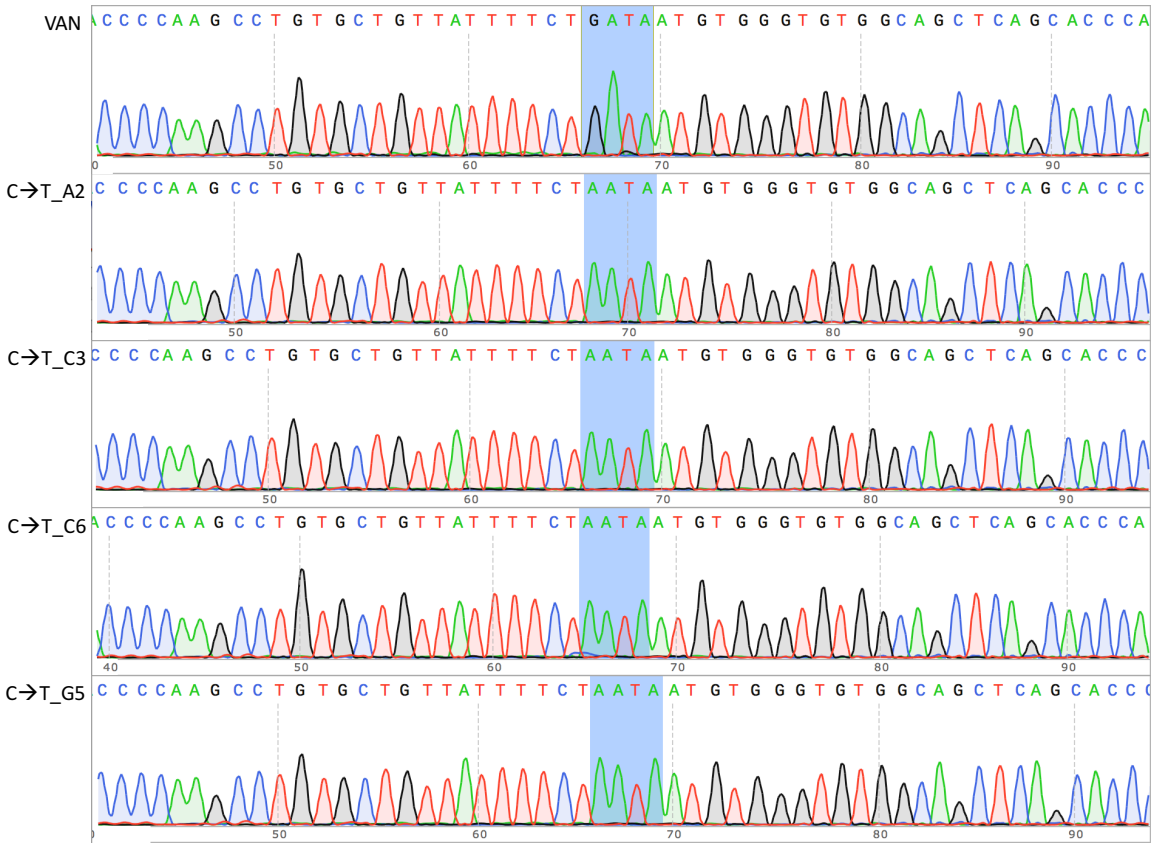


Figure 6.3.4. Chromatogram data from Sanger sequencing on C to T conversion lines.
Core GATA site shown 5' to 3' and is highlighted in blue; Names of lines are on the left side; Top chromatogram is from the same cell line (homozygous mutant patient line - VAN) before editing.

To confirm that there are no mutations that extend outside the 294bp Vanuatu amplicon present on one of the alleles of the iPSC lines, first a larger long-range PCR was performed on the 8 clones derived from the T to C and C to T conversions (*fig.*6.3.5). The long-range PCR screen showed no large indels up to 600bp upstream and 1.1 kb downstream of the cut site of sgRNA3. In addition to the PCR screens, Southern blotting (*fig.*6.3.6) has shown no detectable large-scale genomic aberrations at the region encompassing the Vanuatu site. Together these data suggest that the genome edited iPSC lines contain two identical alleles, homozygous for the desired base conversion and otherwise identical to the

parental isogenic control in the region of the Vanuatu site except for the one base pair edit.

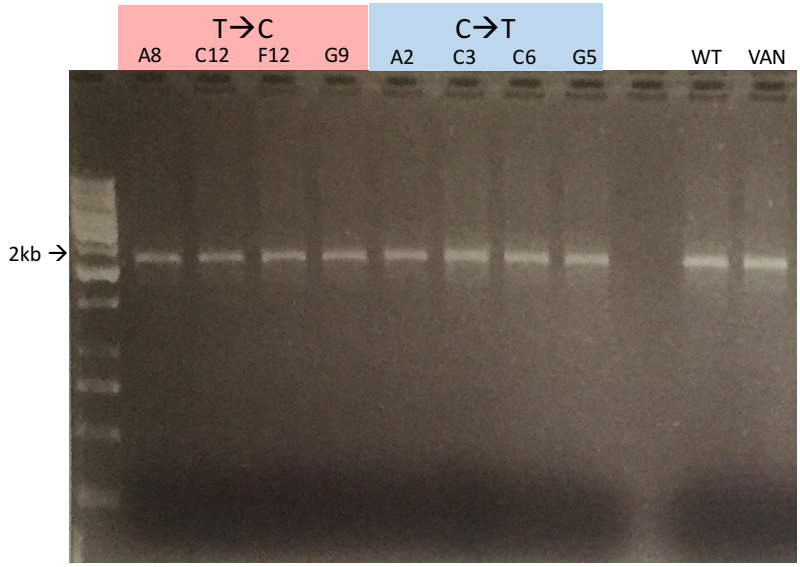


Figure 6.3.5 Long range PCR in the α-globin locus.
T to C conversion lines are highlighted in red; C to T conversion lines are highlighted in blue. Wild type parental line is labelled WT and homozygous mutant patient line is labelled VAN. WT product size is 1782bp

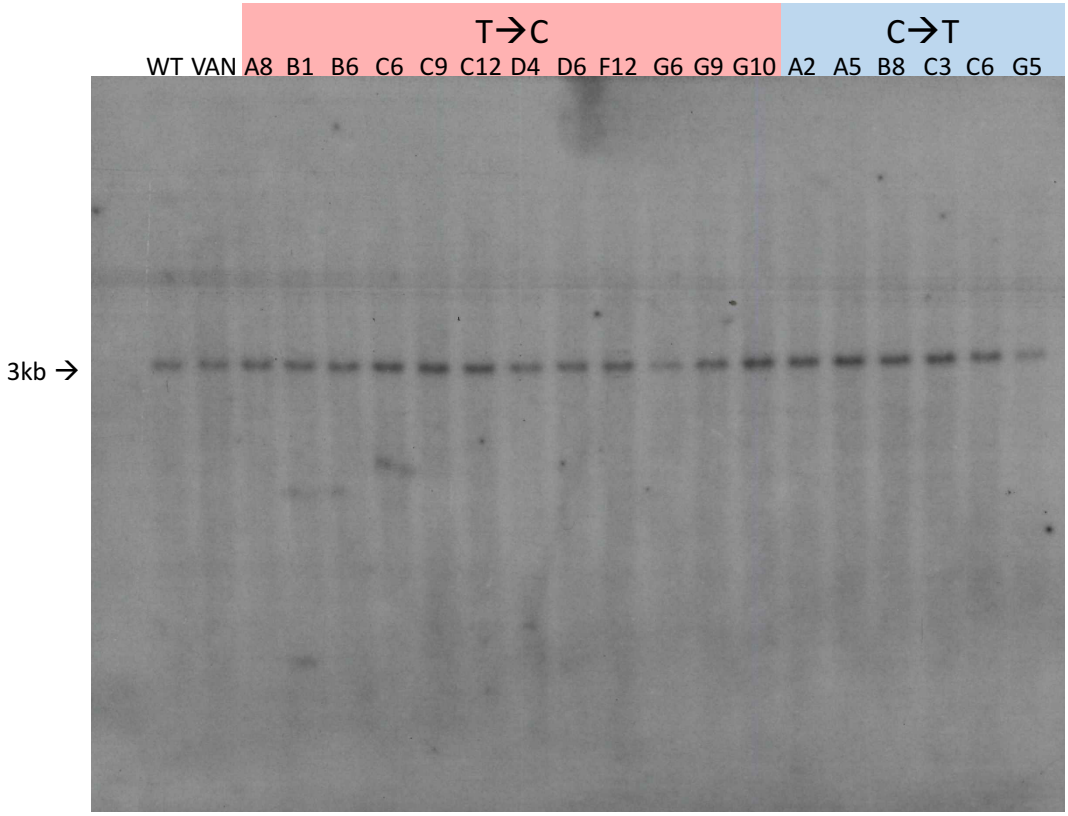


Figure 6.3.6 Southern Blot for PvuII, XbaI digestion fragment containing the Vanuatu site.
T to C conversion lines are highlighted in red; C to T conversion lines are highlighted in blue. Wild type parental line is labelled WT and homozygous mutant patient line is labelled VAN. WT product size is 3143bp.

6.4 Conclusion

The genome editing strategy employed in this thesis yielded high HDR-mediated single base pair conversion rates in iPSCs, which resulted in the generation of mutant and isogenic control lines that can be employed to answer precise biological questions. The protocol has been optimized to robustly and faithfully produce homozygous, scarlessly-edited iPSC clones at rates of 18-31% for the target site examined in this thesis, which was higher than other highly efficient, but transgenic, systems for introducing single nucleotide variants in iPSCs (González et al., 2014; Wang et al., 2017). The high efficiency of homozygous HDR-mediated editing makes the isolation of iPSC lines with a desired genotype easy and convenient. Using the outlined editing strategy, the mutant C allele of the Vanuatu variant in homozygous patient iPSCs was converted to the wild type T allele, repairing the Vanuatu mutation. Four clones generated from the C to T conversion were confirmed to contain the desired sequence at the Vanuatu site without the presence of additional local genomic aberrations. In addition, the wild type T allele of the Vanuatu variant in a wild type iPSC line was converted to the mutant C allele, replicating the Vanuatu mutation in wild type cells. Four clones from the T to C conversion were also screened extensively and shown to contain only the desired sequence at the Vanuatu site without additional local genomic rearrangements.

Several important observations came out during the process of optimization and screening. First, the clones isolated after selection, that had been screened for the presence of large and small undesirable mutations using several complementary

techniques, contained mostly small indels and appeared to be devoid of large genomic aberrations. This is in line with what has been shown by others for the CRISPR system (Canver et al., 2014; Paquet et al., 2016; Wang et al., 2014; Yang et al., 2013), where the majority of mutations at cut sites have been found to be smaller than 10bp and appear to be the result of microhomology-based repair (McVey and Lee, 2008; Symington and Gautier, 2011). Second, similarly to that observed in previous studies (Nielsen et al., 2013; Wang et al., 2014), the sgRNA that contained a long stretch of Us and was otherwise U rich did not perform as well as the sgRNA that fitted all the design criteria for optimal expression from the U6 promoter but had a cut site that was further away from the mutation site than the optimal distance for obtaining homozygous edited alleles, previously reported in iPSCs (Paquet et al., 2016). This suggests that a higher expression level of the sgRNA is preferable to the proximity of the mutation to the cut site, even when HDR-mediated gene editing is concerned. Another interesting observation was the difference in the balance between HDR and NHEJ in different iPSC lines. This was evident from the different percentages of HDR-derived alleles between the wild type and patient iPSC lines (52 and 30% respectively) with both cell lines having a similar rate of mutagenesis at the cut site (around 88%). This effect has been observed previously for different iPSCs lines but depended heavily on the target site (Mitzelfelt et al., 2017). The last important observation was that the use of a single stranded DNA donor that is protected from degradation inside the cell significantly increased the yield of HDR edited alleles in human iPSCs, possibly by prolonging the availability of the template to the repair machinery. This effect has also been observed in immortalized cell lines and rodents (Renaud et al., 2016).

Chapter 7: Rescuing and recreating the Vanuatu phenotype

7.1 Introduction

Previous work on the Vanuatu mutation by De Gobbi et al., 2006, performed in patient primary erythroid cells, suggests that the molecular characteristics of the phenotype correlate with the SNV that is located on the short arm of chromosome 16 at coordinate 209,709 for human genome build hg19. There is strong evidence that just changing the wild type sequence TAATAA (T allele) to the GATA motif TGATAA (C allele) is sufficient to open the chromatin around the site and cause transcriptional activation of the region.

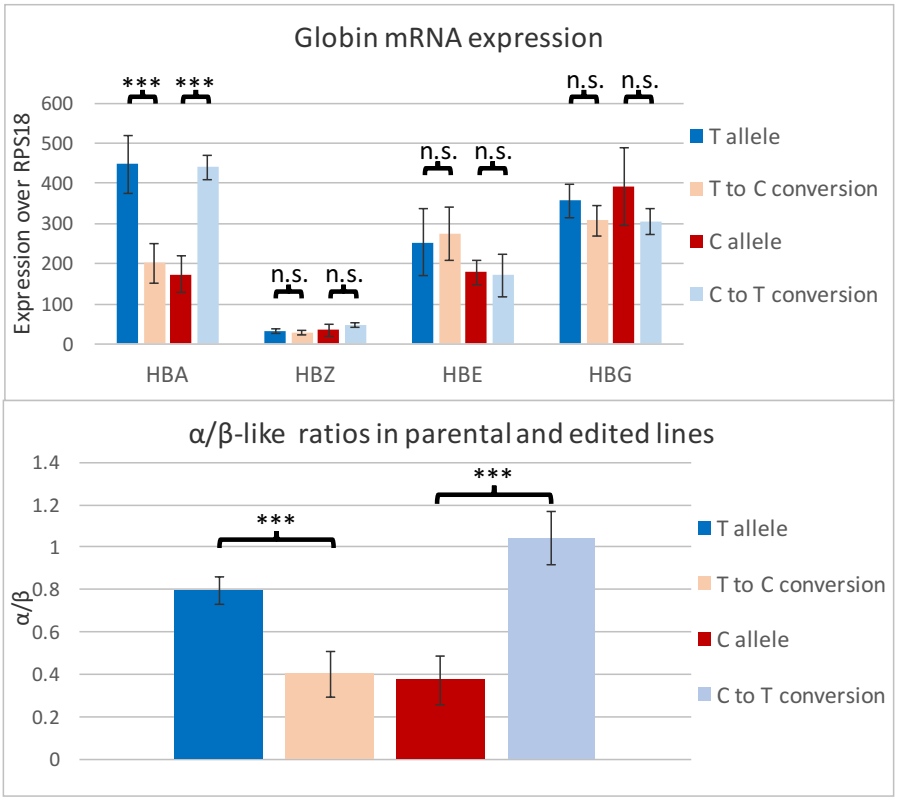
Using patient derived iPSCs and a human iPSC-based erythroid differentiation system I have recapitulated the molecular characteristics of the phenotype caused by the Vanuatu mutation. Following from this, in order to definitively prove that the causal variant of the Vanuatu phenotype is the SNV at position 209,709 (hg19) on chromosome 16, wild type and patient cells have been converted to carry the opposite allele by gene editing using the CRISPR/Cas9 system. Cell lines established from wild type iPSCs that have been edited to become homozygous for the mutant C allele and lines established from patient iPSCs that have been edited to become homozygous for the wild type T allele were differentiated down the erythroid pathway and analysed. To understand the function of the SNV in the two major characteristic features of the phenotype, namely the downregulation α-globin and the establishment of a novel transcriptional unit, expression of the globin genes as well as the GATAν transcript was measured and chromatin accessibility was analysed as an additional marker of active chromatin.

7.2 Expression analysis of gene edited iPSC lines differentiated down the erythroid pathway

The wild type, homozygous Vanuatu mutant and edited iPSC lines all expressed similar levels of β-like globin genes at day 21 of erythroid differentiation (*fig.* 7.2.1) (see Appendix *figs*. A7.2.1 and A 7.2.2 for morphology and cell surface marker expression at day 21). In addition, the expression of *HBZ* is also not significantly altered between the lines that contain the T allele and the lines that contain the C allele. However, the expression of the α-globin genes *HBA1* and *2* is significantly reduced in the lines which are homozygous for the mutant C allele when compared to the lines homozygous for the wild type T allele. This is illustrated by the reduction in the α/β-like ratio. The ratio of α-like to β-like globins drops from 0.79 ± 0.06 in the wild type isogenic control down to 0.40 ± 0.11 in the homozygous T to C allele conversion edited line. Complementary to this observation, the α/β ratio rises from 0.37 ± 0.11 in the isogenic control patient cells (homozygous for the C allele) to 1.04 ± 0.13 in the homozygous C to T allele conversion rescue line (*fig.* 7.2.1). This shows that the only determinant for the downregulation of α-globin in the Vanuatu mutation-specific α-thalassemia is the presence of the single nucleotide transition from T to C.

Expression of the GATAν transcript levels in iPSC-derived erythroid-like cells increases around 20-fold when the wild type T allele at the Vanuatu site is converted to the mutant C in the wild type genetic background. Similarly, the conversion of the C allele in the homozygous patient cells to a T results in over 10-fold downregulation of the GATAν transcript in differentiated cells (*fig.* 7.2.2). The caveat when measuring GATAν

expression by qPCR in the wild type setting (cells homozygous for the T allele) was the high Ct values obtained (over 35). However, the calculated value for GATAν expression cells homozygous for the T allele as a factor of *RPS18* mRNA levels was still regarded as relevant as the Ct values were below the no reverse transcriptase control. This observation shows that the presence of the C allele is necessary and sufficient within a wild-type background for the expression of the GATAν transcription unit.



n.s. – non significant
* p<0.05 one-tail
** p<0.05 two-tail
*** p<0.01 two-tail

Figure 7.2.1 Globin mRNA expression in edited iPSC lines and isogenic controls
Top histogram - expression of α-globin (*HBA1,2*), ζ-globin (*HBZ*), ε-globin (*HBE*) and γ-globin (*HBG1,2*) at day 21 of differentiation. Mean expression levels relative to ribosomal protein S18 (*RPS18*) from wild type iPSCs (T allele) n=3 (isogenic control), edited T to C conversion iPSCs homozygous for the Vanuatu mutation (T to C conversion) n=4, patient iPSCs homozygous for the Vanuatu mutation (C allele) n=4 (isogenic control) and edited C to T conversion iPSCs

homozygous for the T allele (C to T conversion) n=3, quantified using RT-qPCR. Bottom histogram - average α/β-like ratios represented the sum of the expression over *RPS18* of the α–like globin genes *HBZ* and *HBA1,2* divided by the sum of the expression over *RPS18* of β–like globins *HBE* and *HBG1,2* at day 21 of differentiation. Values were obtained by RT-qPCR from wild type iPSCs (T allele) n=3, edited T to C conversion iPSCs homozygous for the Vanuatu mutation (T to C conversion) n=4, patient iPSCs homozygous for the Vanuatu mutation (C allele) n=4 and edited C to T conversion iPSCs homozygous for the T allele (C to T conversion) n=4. Error bars represent one standard deviation; Statistical significance between samples is indicated by an asterisk above the bars on the histogram;

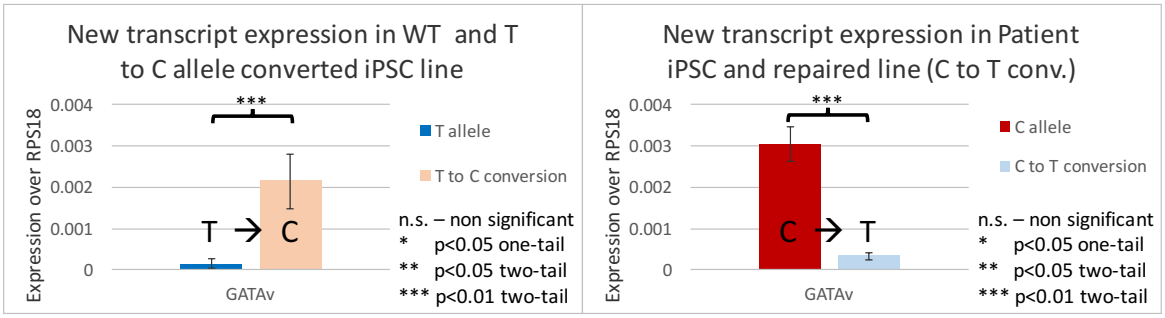


Figure 7.2.2 GATAν transcript expression in edited iPSC lines and isogenic controls
Expression of the new transcript generated by the Vanuatu mutation labeled GATAν at day 21 of differentiation. Mean expression levels relative to *RPS18* from wilt type iPSCs (T allele) n=3, edited T to C conversion iPSCs homozygous for the Vanuatu mutation (T to C conversion) n=4, patient iPSCs homozygous for the Vanuatu mutation (C allele) n=4 and edited C to T conversion iPSCs homozygous for the T allele (C to T conversion) n=4, quantified using RT-qPCR. Error bars represent one standard deviation. Statistical significance between samples is indicated by an asterisk above the bars on the histogram.

7.3 Analysis of open chromatin regions (ATAC-seq) of edited iPSC lines differentiated down the erythroid pathway

ATAC-seq was used to assess the open chromatin in the α-globin locus in edited and isogenic control iPSC lines. The assay was performed as described in Chapter 2, with input material of 60,000-75,000 of either undifferentiated iPSCs or iPSC derived erythroid-like cells harvested at day 21 of differentiation (*fig*. 7.3.1). Chromatin accessibility at the α-globin locus in undifferentiated iPSCs shows no activation of the GATAν transcription unit or the upstream regulatory elements of the α-globin locus and little to no accessibility at the genes themselves in wild type and in patient cells homozygous for the mutant C allele. At day 21 of erythroid differentiation the isogenic controls and the edited lines show the activation of the α-globin enhancers and genes, and the homozygous patient line shows an additional peak over the Vanuatu region which is missing in the wild type line. When the C allele in the patient line is converted to the wild type T the peak over the Vanuatu site is completely abolished indicating the absolute necessity of the C variant for the activation of the GATAν transcriptional unit. Conversely if the wild type T allele in wild type cells is converted to the mutant C allele the chromatin at the region of the Vanuatu site becomes accessible showing that the presence of the C variant of the allele is sufficient to direct the activation of the GATAν transcriptional unit in the erythroid lineage.

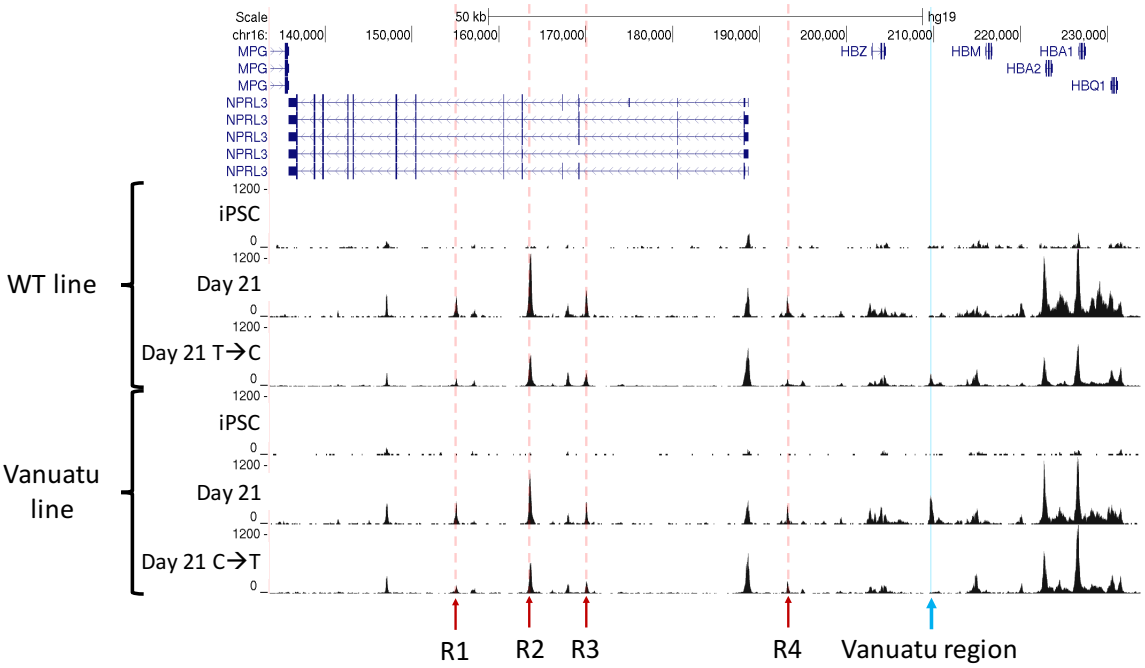


Figure 7.3.1 Chromatin accessibility in edited iPSC lines and isogenic controls
UCSC tracks showing chromatin accessibility as assayed by ATAC-seq in the α-globin locus. First two tracks show chromatin accessibility of a wild type iPSC line: undifferentiated (first) and differentiation day 21 iPSC derived erythroid-like cells (second); Third track is from day 21 iPSC derived erythroid-like cells from the T to C allele conversion gene editing experiment; Fourth and fifth tracks show chromatin accessibility of a patient iPSC line homozygous for the C allele: undifferentiated (fourth) and differentiation day 21 iPSC derived erythroid-like cells (fifth); sixth track is from day 21 iPSC derived erythroid-like cells from the C to T allele conversion gene editing experiment. Below the figure the α-globin enhancers are indicted with red arrows: R1 to R4; Vanuatu site is indicated with a light blue arrow; Above the tracks the positions of the genes are shown in blue.

7.4 Conclusion

Using genome editing I have shown that repairing the single mutant C allele of the Vanuatu SNV at position 209,709 (hg19) on chromosome 16 rescues the associated α-thalassemia phenotype and abolishes the transcriptionally active GATAν unit in iPSC-derived erythroid-like cells. This demonstrates that the presence of the C base is necessary for the establishment of the new unit. Along with previous ChIP data, this observation strongly suggests that the point mutation causes the recruitment of GATA1 to the newly created WGATAR motif. To definitively prove this, I would need to perform GATA1 ChIP to examine the binding of the TF to the Vanuatu site after the repair of the mutation, to determine if this results in a loss of GATA1 enrichment over the repaired allele. Further, I have shown that recreating the Vanuatu mutation in a wild type iPSC line, upon erythroid differentiation, recapitulates the phenotype observed in primary erythroid cells from homozygous mutant patients; namely the down-regulation of α-globin and the establishment of a new, transcriptionally active region over the Vanuatu site. This demonstrates that the T to C transition mutation at position 209,709 (hg19) on chromosome 16 is sufficient to generate a region of chromatin accessibility and transcriptionally activate the surrounding DNA. The combination of these results proves, without doubt, that this single T to C base change found in a group of α-thalassemia patients is the cause of the emergence of a new transcriptional unit in an intergenic region of the α-globin locus and the cause of the repression of the α-globin genes in patient erythroid cells.

These observations provide an important perspective for identifying functional SNVs that may influence gene expression. Clearly this will be of importance in further analysis of GWAS studies. There has been a focus on understanding the functional significance of variants located in regions that have been determined to display chromatin characteristics associated with gene regulation from available chromatin accessibility, ChIP-seq and other data, i.e. regions of the genome likely to contain *cis*-regulatory elements (Knight, 2014; Mathelier et al., 2015; Mulin et al., 2014; Schaub et al., 2012). However, the very rare gain-of-function regulatory SNV analysed in this thesis is not characterized by any obvious chromatin features at the wild type allele and is located in a non-conserved intergenic region of the α-globin locus that can be deleted without affecting the expression of α-globin (Winichagoon et al., 1982). This demonstrates that such variants should not be dismissed as of lower priority and that genome-wide profiling of the chromatin landscape of patient cells is an indispensable complementary tool to categorize the functional significance of non-coding variation.

Chapter 8: Discussion

The aim of this thesis was to characterize the molecular mechanisms by which a SNV creates a new transcriptional unit and affects expression of a distal group of genes using an iPSC based erythroid differentiation system. Understanding the processes that drive the characteristic features of the phenotype could potentially uncover basic mechanisms of promoter-enhancer interactions and potentially outline the different rules for the formation of promoters or enhancers.

8.1 Modelling an erythroid disease in an iPSC-based differentiation system

Experimental modelling of human disorders enables the definition of the underlying cellular and molecular mechanisms. When the molecular basis of a disease can be narrowed down to a defined genetic locus, modelling the phenotype may be facilitated by studying wildtype and disease associated genotypes in a suitable tissue-specific context. The availability of human iPSCs, which can be generated from individuals with disease-relevant genotypes, and have the potential to differentiate into virtually any cell type, can now help to overcome, for certain disorders, the limitations of animal and cell-line models as well as the scarcity of the appropriate patient derived material. In this study, the use of an iPSC-based erythroid differentiation system was employed to model a phenotype observed in a group of α-thalassemia patients that harbour a rare regulatory single nucleotide variant.

The iPSC-derived erythroid-like cells used in this thesis resemble primary erythrocytes both morphologically and immuno-phenotypically. They produce high levels of globin mRNA and express the embryonic ε-globin gene *HBE* and, to a lesser extent, ζ-globin gene *HBZ*, the fetal γ-globin genes *HBG1* and *HBG2*, the fetal/adult α-globin genes *HBA1* and *HBA2* and produce very low amounts of the adult β-globin gene *HBB*. The cells most closely resemble erythroid cells derived from either the primitive wave of erythropoiesis or from an early window during transient definitive erythropoiesis. This classification stems from the observation that they express embryonic globin mRNA but also transcribe fetal globin genes. The iPSC-derived erythroid-like cells are most likely undergoing a

condition analogous to stress erythropoiesis, evident by the asynchrony between morphological maturation and the expression of earlier cell surface markers at the final stages of differentiation. This could be the explanation for the fetal-like nature of the cells. In the iPSC-derived erythroid-like cells the *cis*-regulatory elements in the α-globin locus appear to regulate the globin genes in a manner similar to that observed in primary erythroid cells. This is supported by the observation that the chromatin of the four known α-globin enhancers and the globin genes is open and the DNA is appropriately bound by key erythroid TFs GATA1 and KLF1. Furthermore, the α-globin genes and their enhancers form the appropriate *cis*-regulatory interactions.

Technological limitations had previously precluded the further study of a gain-of-function regulatory SNV described by De Gobbi et al., 2006 to generate a new GATA1 binding site that gives rise to an aberrant transcriptional unit in an intergenic region of the α-globin locus and causes α-thalassemia. Using the iPSC-based erythroid differentiation system and patient cells reprogrammed to pluripotency we have overcome these previous limitations. In this study patient-derived iPSCs homozygous for the mutant C allele of the Vanuatu variant were differentiated along the erythroid pathway to recapitulate the features of the phenotype of the mutation that have been described in primary erythroid cells. This involves the recruitment of GATA1 and RNA Pol II and transcriptional activation at the Vanuatu mutation site, in addition to the down-regulation of α-globin.

The characteristics of the differentiation system and its ability to replicate various aspects of a disease associated phenotype suggest that the system not only has the potential to

be used to study the regulation of α-globin expression but also to model many erythroid-specific diseases.

8.2 Transcription factor binding at the site of the Vanuatu mutation

De Gobbi et al., 2006 reported the generation of a novel GATA1 binding site at the C allele of patient -derived Vanuatu mutant erythroid cells. In these cells, this was accompanied by the recruitment of another erythroid transcription factor TAL1 to the same region. This co-localization of GATA1 and TAL1 is not surprising as the two factors are known to work together to form an activating complex (Tripic et al., 2009). This raises the question: What other transcription factors are bound to the mutant C allele in patient-derived erythroid cells?

To address this question a tool that predicts the effect of point mutations on transcription factor binding in a tissue-specific manner (recently published by our lab Schwessinger et al., 2017) was used to perform *in silico* mutagenesis and predict possible TF binding sites in a 500bp region around the Vanuatu mutation in erythroid cells. This has identified two potential binding motifs for the key erythroid transcription factor KLF1 that flank the GATA motif. Subsequently the recruitment of KLF1 to the C allele of the Vanuatu variant was confirmed by ChIP. This observation is particularly interesting as we know that the introduction of a KLF1 binding site at the promoter of the fetal γ-globin is sufficient to upregulate expression of the gene in an adult erythroid environment where it is normally repressed (Weatherall et al., 1975; Wienert et al., 2017; Wood et al., 1982). In addition, KLF1 is essential for the establishment of chromatin accessibility and activating transcription in the β-globin locus (Armstrong et al., 1998; Gillemans et al., 1998; Kadam et al., 2000; Lee et al., 1999). Furthermore, KLF1 interacts with the histone

acetyltransferases CBP and p300 and this interaction has been shown to enhance its ability to activate transcription both *in vitro* and *in vivo* (Zhang and Bieker, 1998). These observations suggest that there might be a synergistic effect between KLF1 and GATA1 in the activation of the GATAν transcriptional unit and it is worth considering what that interaction might entail.

The genomic context of this particular TF motif may be a key feature explaining why this single SNP leads to transcriptional activation. The newly generated GATA1 binding site which is part of the TAL1 binding half E-box-GATA motif CTG-(N_8)-WGATAR is flanked by two KLF1 binding motifs. This indicates that the expression may be the result of the specific binding affinity, orientation, spacing and order of transcription factor binding sequences. It also presents the opportunity to ask whether GATA1 is capable of generating this actively transcribed region only due to the interaction with nearby transcription factors and to address questions regarding co-operative binding of transcription factors. One might speculate that the generation of the GATA1 binding site acts as a threshold event which increases the DNA-binding affinity of the whole erythroid transcription factor complex above the minimal required for transcriptional activation and that GATA1 binds only as part of that complex. An alternative explanation is that GATA1 is able to bind the DNA within a compacted chromatin environment and act as a pioneer factor. In this case GATA1 may play a role in opening the chromatin to allow the establishment of a transcription factor complex composed of KLF1 and perhaps the TAL1 containing multimeric LDB1 complex. Indeed, there are several lines of evidence that support the role of GATA1 as a pioneer factor. First, it has been demonstrated that the GATA TF family member GATA4, as well as the FOX TF family member FOXA1, are capable of binding their target sites when they are found in compacted chromatin *in vitro* and can increase

chromatin accessibility locally around the binding site (Cirillo et al. 2002). Furthermore, epigenetic reprogramming of cells from one fate to another appears to depend to some extent on the action of pioneer factors (Zaret and Mango, 2016) and many pioneer factors have been found in transcription factor cocktails that have been used for the differentiation of specific cell types. This is true for the pioneer factors GATA4 and FOXA3, which are among a group of transcription factors that facilitate the trans-differentiation of fibroblasts to hepatocytes (Huang et al., 2011). In line with its proposed pioneer nature, the ectopic expression of GATA1 in monocytes induces cell fate change to the erythroid and megakaryocyte programs (Visvader et al. 1992; Kulessa et al. 1995). Furthermore, ectopic GATA1 induces erythroid, eosinophilic or basophile-like characteristics when expressed in mouse granulocyte–monocyte progenitors (Heyworthet al., 2002). In addition, a combination of GATA1, TAL1, LMO2, and c-MYC is able to reprogram fibroblasts to erythroid progenitors (Capellera-Garcia et al., 2016).

Other DNA binding TFs could also be binding in the vicinity of the Vanuatu site and contribute to the activity of the transcriptional complex that forms over the Vanuatu mutation. Indeed, the *in silico* prediction of erythroid cells specific TF binging sites identified two additional strong candidates within 150bp of the site of the Vanuatu SNV: one covering a motif for the FOX- and one for the ETS- family of transcription factors. Along these lines, studies of other FOX TF family members have demonstrated that they can also open compacted chromatin (Cuesta et al., 2007; Hatta and Cirillo, 2007), suggesting that, if bound, the FOX site could be another major contributing factor to the activation of the GATAν transcriptional unit. Further ChIP experiments are required to determine the presence of other DNA binding TFs in the vicinity of the mutant C allele of the Vanuatu variant in erythroid cells.

The questions raised in this section can be addressed by generating combinations of mutants which reduce the binding affinity of the flanking TF binding sites and of the elements of the half E-box-GATA motif. This would allow the dissection of the GATAν transcriptional unit to determine whether, in addition to GATA1, there are other TF binding events that are essential for transcriptional activation and which aspect of the process they regulate.

8.3 The nature of the GATAν transcriptional unit

Here I have demonstrated that a single base pair substitution resulting in the generation of a GATA1 binding site in an intergenic region of the active α-globin locus is sufficient to cause the emergence of a new transcriptional unit - GATAν. GATAν is a promoter-like element that is marked by high levels of the histone modification H3K4me3 which has been associated with promoter regions (Barski et al., 2007; Heintzman et al., 2007; Mikkelsen et al., 2007) and previous data show that it exhibits unidirectional transcription (De Gobbi et al., 2006).

Disease causing regulatory variants may result in the creation of a gain-of-function regulatory element as is the case for the Vanuatu SNV or result in the destruction of a regulatory element like the loss-of-function regulatory SNV which alters the binding of transcription factor MSX1 and abrogates the function of a developmental enhancer implicated in a congenital condition known as Pierre Robin sequence (Benko et al., 2009). However, examples where single point mutations located in regulatory elements cause a disease are relatively rare. In general, SNPs associated with *cis*-regulatory elements show minor effects on expression and disease phenotypes and are able to explain only a small proportion of phenotypic variance (Antoniani et al., 2017; Fugger et al., 2012; Frazer et al., 2009). This is exemplified by the genetic variation within the enhancer of *BCL11A*, a repressor of HbF expression, where individual SNPs have a modest impact on the binding of TF, the expression of *BCL11A* and the downstream effect on HbF levels (Bauer et al., 2013). It is also worth pointing out the observation that it is common for single nucleotide

substitutions to only cause a modest loss of enhancer activity even they are found within critical motifs (Melnikov et al., 2012; Patwardhan et al., 2012). This highlights the importance of the GATAν promoter unit where a 1bp change dictates the difference between open chromatin and transcriptional activity or inaccessible chromatin and no expression.

Mutations in regulatory elements are responsible for several genetic and non-genetic haematological diseases (Antoniani et al., 2017). Among these is a gain-of-function mutation in an intergenic region 7.5 kb upstream of the promoter of *TAL1*, which generates a binding site for the TF MYB (Mansour et al., 2014). The mutations associated with the creation of the TF binding site are a series of heterozygous indels 2–18bp in size. The indels overlap at the same hotspot and are found in a subset of T-cell acute lymphoblastic leukaemia (T-ALL) patients. The mutations result in the recruitment of MYB and additional TFs and the gain of the activating histone modification H3K27ac, but unlike the Vanuatu SNV this gives rise to a strong enhancer-like unit which upregulates the expression of the *TAL1* gene. This example raises the question: Why has the generation of the new open chromatin region in the α-globin locus produced a promoter rather than an enhancer?

The RNA Pol II transcription system is a key component in the expression of protein-coding and many non-coding genes in eukaryotes. The initiation of RNA Pol II transcription is facilitated by a stretch of DNA known as the core promoter (for reviews, see Kadonaga, 2012; Lenhard, et al. 2012; Vo Ngoc et al., 2017). In the standard view of RNA Pol II promoter function the core promoter is made of a number of interchangeable sequence

elements that are positioned around the TSS and are bound by components of the transcription machinery. In the past, the core promoter was often regarded as a generic DNA element with a TATA box that functions universally at all genes. Bioinformatics analysis, however, has revealed that the majority of the genes in mammalian genomes don't fit within this simple model where a TATA box directs transcription at a TSS which initiates from a single nucleotide position. Now we know that a TATA box is only present in a small proportion of metazoan core promoters (10-20%), that about 70% of the promoters in the human genome are associated with CpG islands and that many genes contain multiple promoters that can possess multiple TSSs (Bajic et al., 2006; Carninci et al., 2006; FitzGerald et al., 2004; Sandelin et al., 2007; Tokusumi et al., 2007). Indeed, most of the highly transcribing RNA Pol II core promoters in the human genome, contain several closely positioned TSSs spread over 50–100bp (Frith et al., 2008). It has also become apparent that no core promoter element can be found universally at all promoters. Elements that have been previously identified, such as the initiator (Inr), the motif ten element (MTE), the downstream promoter element (DPE), the downstream core element (DCE), the TFIIB recognition element (BRE), and the x core promoter element 1 (XCPE1) are not found within a big proportion of the promoters in the mammalian genome (Butler and Kadonaga, 2002; Lee et al., 2005; Lim et al., 2004; Muller et al., 2007; Ohler et al., 2002; Smale and Kadonaga, 2003; Suzuki et al., 2001; Tokusumi et al., 2007). Moreover, many core promoters lack any of the known motifs, suggesting there are other, undiscovered, core promoter elements. What is certain is that there is a large diversity, in terms of composition and function, in the core promoter and that precise sequences at particular locations are probably required for core promoter function. These observations

need to be taken in consideration when exploring the potential reasons behind the specification of the GATAν transcriptional unit as a promoter.

The first possibility to consider is that the emergence of the GATAν unit could be a simple case of reactivating a pseudogene, but that seems unlikely as the region around the Vanuatu site is poorly conserved in vertebrates (see Appendix *fig*. A8.1.1). However, there could still be some fortuitous combination of known promoter sequence elements at the site. Analysis of the DNA surrounding the Vanuatu SNV revealed that there are also no CpG islands associated with the GATAν unit and a search for promoter sequence elements identified a limited set of motifs in the immediate vicinity (1000bp around the SNV) of the GATA site (see Appendix *fig*. A8.1.2). Unsurprisingly, there is no consensus TATA box, but there are a few appropriately spaced mammalian Inr and DPE elements that are also present in multiple copies in the tandem repeat (TR). The caveat of this analysis is the fact that we do not know where the GATAν unit initiates transcription, thus before any conclusive evaluation of putative core promoter elements at the Vanuatu site, the transcription start site/sites of the GATAν transcript need to be mapped.

Next, we discuss how the particular TF complex formed at the two novel units might determine their nature. In the case of the novel *TAL1* enhancer in T-ALL cells, the newly introduced MYB motif leads to MYB binding along with the recruitment of H3K27 acetyltransferase CBP and core components of a leukemogenic TF complex: RUNX1, GATA3, and TAL1. In certain human cancers, large regulatory elements defined by extensive acetylation of H3K27, referred to as super-enhancers, drive the expression of critical oncogenes. In T-ALL cells most endogenous super-enhancers are bound by MYB

and CBP and knockdown of MYB causes a reduction in many super-enhancer associated genes in T-ALL cells, this is evidence that MYB has a general role in the activity of super-enhancer in T-ALL (Kwiatkowski et al., 2014; Mansour et al., 2014; Sanda et al., 2012). In the instance of the GATAν promoter in erythroid cells, we know that, genome wide, the majority of KLF1, GATA1 and TAL1 binding sites are located at large distances from transcription initiation sites, which suggests long-distance regulation (Fujiwara et al., 2009; Palii et al., 2011; Tallack et al., 2010; Yu et al., 2009b). This is inconsistent with a promoter-specific TF complex formation over the Vanuatu region and more work defining the transcription factor network that regulates the activity of the GATAν promoter is required before any conclusions can be made.

Genomic position may play a role in determining the nature of the new chromatin accessible site over the Vanuatu SNV. The new GATA1 motif generated by the mutation is around 6 and 7kb from the nearest globin gene promoters (*HBM* upstream and *HBZ* downstream respectively) and is proximal to a LINE2 element (within 25bp upstream) and GC-rich TR (within 110bp downstream) known as the inter-zeta hyper variable region with a 36nt unit size (Goodbourn, S. et al., 1983). LINE elements: LINE1 (L1) and LINE2 (L2) are non-long terminal repeat retrotransposons (Goodier and Kazazian, 2008). The LINE1 retrotransposon clade has expanded greatly in mammals, reaching 17% of the human genome with a limited number of them still actively undergoing retrotransposition (Brouha et al., 2003). In contrast, LINE2 elements have lost their activity in placental mammals and the human genome harbours only highly defective copies of LINE2 retrotransposons (Furano et al., 2004; Lander et al., 2001; Lovsin et al., 2001).

Interestingly, retroviral elements can act both as enhancers to influence gene expression and also as promoters for many tissue-specific non-coding RNAs or as alternative promoters for the transcription of protein-coding mRNA (Goke and Hui Ng, 2016; Lenhard, et al. 2012). The retrotransposons that are utilized as promoters are commonly not highly conserved and the promoters they generate tend to be TATA-less and without other strong spatially constrained motifs. Studies have identified many human retrotransposon-driven TSSs (including LINE element-derived TSSs) most of which show a low level of transcription and are transcribed from a degenerate Inr element with transcripts often initiating from within the retrotransposon (Faulkner et al., 2009; Kapusta et al., 2013; Kelley and Rinn 2012). RNAs that are expressed from these retrotransposon derived promoters are often not polyadenylated (Faulkner and Carninci, 2009) and commonly confined to the nucleus (Plessy, C. et al., 2010). On the other hand, retrotransposons can also act as enhancers and a large number of retrotransposon-derived enhancers have been employed by embryonic stem cells and tissues that are linked to embryogenesis such as the placenta or endometrium. Moreover, specific classes of retrotransposons enriched at enhancers contain tissue-specific TF binding sites (Chuong et al., 2013; Kunarso et al., 2010; Lynch et al., 2011; Lynch et al., 2015).

Short G/C rich TR elements are enriched around the TSSs of human promoters (Sawaya et al., 2013) and TR with unit sizes up to 45bp present in promoters in the human genome have been linked to alteration in gene expression and DNA methylation, suggesting a link between TRs and local genome function (Quilez et al., 2016). TR associated with promoters have been found to serve as arrays for TF binding and increase in repeat size has been often associated with upregulation of expression (Gemayel et al., 2010). This implies a functional role for TRs at gene promoters.

In comparison to GATAν, in the novel TAL1 enhancer the new MYB binding site is more than 300bp away from an L1 element, around 7.5kb from the TAL1 promoter and there are no TRs at least 2kb in either direction. It is possible that the presence of the tandem repeat, the nature of the LINE element and the shorter distance between it and the erythroid transcription factor binding domain has influenced the nature of the GATAν transcription unit.

The summary of the observations outlined in sections 8.2 and 8.3 indicates that the establishment of the GATAν promoter-like unit is likely to be a combination of appropriately positioned repetitive elements and specific erythroid transcription factor motif grammar. Genetic experiments to define the minimal GATAν transcriptional unit capable of erythroid specific transcriptional activation will help to understand the specific features required for its expression.

8.4 Model of interaction between GATAν and the α-globin enhancers

Here I have demonstrated that the emergence of a new actively transcribing promoter-like unit (GATAν) in the α-globin locus causes a perturbation of the normal chromatin interactions in the gene cluster. Analysis of chromatin conformation has revealed that GATAν displays very strong interactions with the enhancers of the α-globin genes. This interaction causes a competition with the promoters of the α-globin genes *HBA1* and *HBA2* and results in reduced proximity between the endogenous α-globin genes and enhancers. The reduction in interaction is consistent with the downregulation of α-globin observed in patient primary erythroid cells and erythroid-like cells differentiated from patient-derived iPSCs.

This observation puts forward the question: Why would the enhancers of the α-globin genes interact preferentially with the GATAν promoter rather than their endogenous target genes?

To answer this question, we need to understand the mechanism by which the α-globin enhancers establish proximity-based interactions with their target genes. There are several observations that help uncover some principles of the interaction.

In Chapter 5 the occupancy of the key erythroid TF GATA1 was shown for the α-globin locus. In the wild iPSC-derived erythroid like cells there were GATA1 peaks associated with the *HBM* and *HBZ* genes. Both peaks were found at a similar distance away from the SNV (about 6-7kb) but only the enhancer distal peak in relation to the SNV - the *HBM* peak, was abolished in the patient cells and this reduction is not due to mutations in the

underlying sequence (see Appendix *fig.* A8.4.1). This is in line with what is known for patent primary erythroid cells where *HBM* is severely downregulated (De Gobbi et al., 2006). Additionally, on the enhancer distal side of the SNV (that is from the SNV towards the α-globin genes) we observed the most prominent features of the phenotype, namely the reduction of interaction between the enhancers the promoters of *HBA1* and *HBA2* and their downregulation. On the other hand, on the enhancer proximal side in relation to the SNV (that is from the SNV towards the enhancers) the *HBZ* associated GATA1 peak is preserved and *HBZ* expression is not affected by the presence of the C allele in the iPSC derived erythroid-like cells (summarised in *fig*. 8.4.1a). The presence of the C allele appears to divide the α-globin locus at least as far as transcription and GATA1 binding are concerned. This indicates the possibility that the effect of the mutation is established through a mechanism that acts in a polar manner where chromatin features between the enhancers and the mutation are preserved and features distally from the enhancers in relation to the mutation are negatively affected. The one mechanism that we have been able to link the mutation to is the formation of enhancer-promoter interactions. The polarity of the effect of the Vanuatu SNV suggests that there might be some sort of scanning mechanism involved in the activity of the α-globin enhancers and the establishment of interactions within the α-globin locus. This would best fit a model that explains the establishment of enhancer promoter interactions through a mechanism that involves participation of the intervening DNA sequences and cannot be explained by a simple looping mechanism that would predict random collisions between proteins bound to regulatory DNA elements.

The outlined observations would fit with the extrusion model of chromatin loop formation where a hypothetical extrusion complex would extrude DNA as it slides along chromatin

until it reaches appropriate motifs that would stop its progression and define the boundaries of the generated loop. The extrusion complex has been proposed to involve the cohesin complex which forms a ring that can slide along chromatin and the insulator-binding protein CTCF which has an important role in transcription regulation and the organization of chromatin (Alipour and Marko, 2012; Nasmyth, 2001; Sanborn et al., 2015; Fudenberg et al., 2016). In line with this proposition previous work in our laboratory has shown that, for the mouse α-globin locus at least, CTCF-cohesin boundaries delimit the region of chromatin within which enhancers have access to promoters, thereby defining the interactions within the locus (Hanssen et al., 2017).

If we speculate that the same is true for the human α-globin locus then it is likely that what is observed as a consequence of the Vanuatu SNV is a result of the nature of loop formation as predicted by the extrusion model. If that is the case and we assume a random chance of loading along the α-globin locus then depending on where the extrusion complex loads there could be one of two possible outcomes which would be dictated by the order in which the boundary elements are anchored to the complex (Enhancer first or HBA first). If the complex loads closer to the enhancer proximal boundary (Enhancer first) then it is more likely that it will bring together the Vanuatu region and the α-globin enhancers before it brings HBA2 and HBA1 into the loop; if it loads closer to the α-globin gene proximal boundary (HBA first) it is more likely that it will bring in closer proximity the α-globin enhancers and promoters as the Vanuatu region would be positioned more centrally inside the loop (*fig*. 8.4.1b). This would suggest that the enhancers would have a roughly equal chance (depending on distances in the chromatin fibre) to first come in contact with either the α-globin promoters or the GATAν promoter and establish interactions. Indeed, we do see a roughly equal number of interactions between the

enhancers and either the globin genes or GATAν (see Chapter 5). It is possible that this model based on random loading of the extrusion complex could account for the downregulation of HBA1 and HBA2 but it cannot account for the polarity effect of the Vanuatu SNV on GATA1 binding at the locus. The abolishment of the GATA1 peak at the *HBM* promoter fits better with a model where the extrusion complex is preferentially loaded closer to the enhancer region. This working model is still very simplistic as there are factors which it cannot account for, including the position and effect of additional regulatory elements found between the enhancers and the α-globin genes, including CTCF binding sites, or the relative distances between the elements in their native chromatin environment which would differ from the linear representation. Currently there is not enough evidence to conclude the validity of the proposed model.

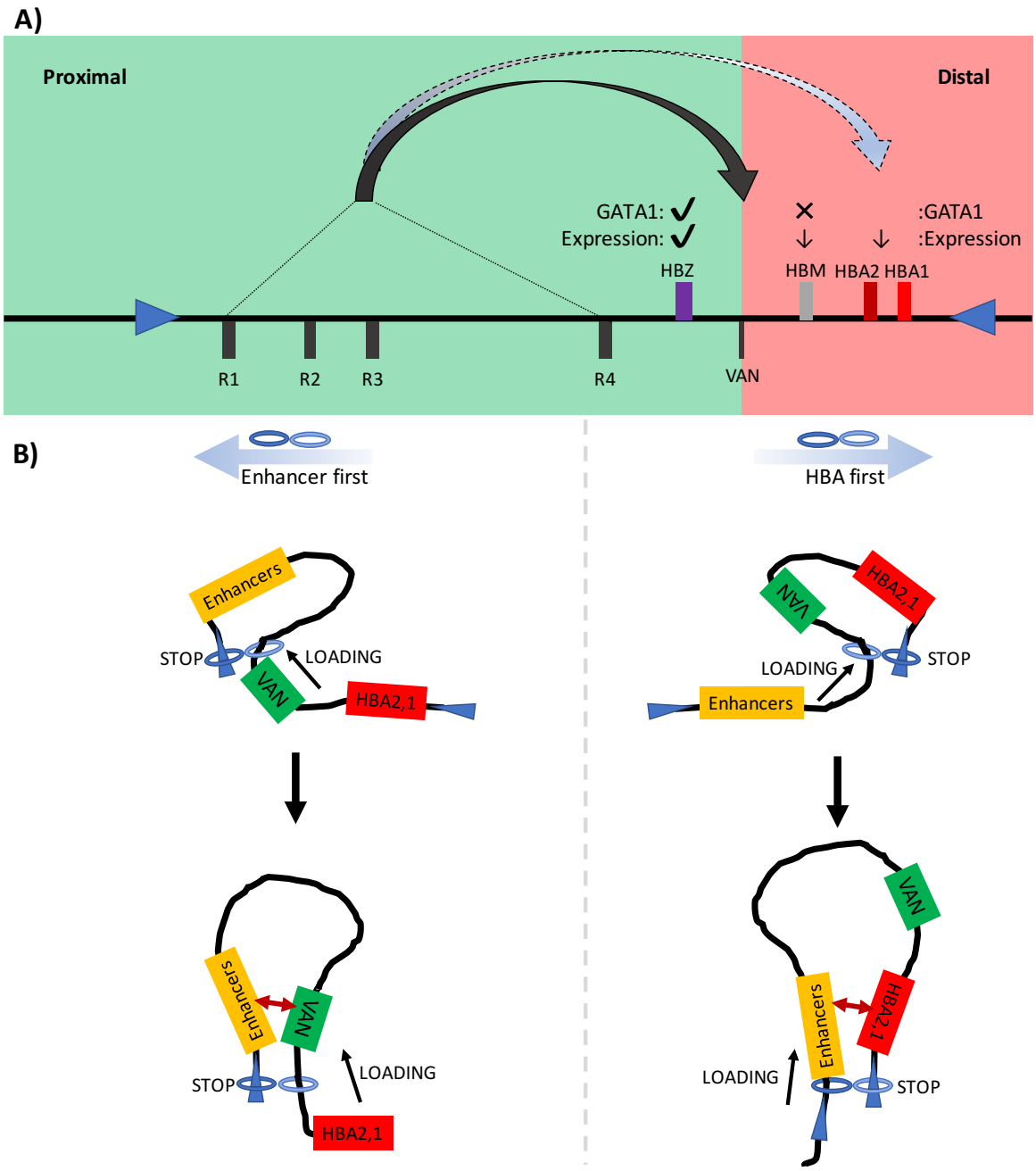


Figure 8.4.1 Working model of chromatin looping and the α-globin locus.
A) Polarity effect of Vanuatu mutation. Schematic of the α-globin locus with selected elements represented: as black rectangles multispecies conserved sequences R1 to R4 and the site of the Vanuatu SNV (VAN T→C), ζ-globin gene *HBZ* (green), μ-globin *HBM* (grey) and the two α-globin genes *HBA2* and *HBA1* (red). Enhancer proximal side is on a green background, enhancer distal side is on a red background. Arrows indicate interaction with the enhancers
B) Possible outcomes depending on the site of extrusion complex loading. Left: Enhancer first – loading of the extrusion complex closer to the enhancer proximal boundary would suggest that the chromatin loop would be anchored to the complex at the enhancer boundary and then the complex would continue to load the remainder of the locus into the loop. Right: HBA first - loading of the extrusion complex closer to the α-globin gene proximal boundary would suggest that the chromatin loop would be anchored to the complex at the α-globin gene boundary and then the complex would continue to load the remainder of the locus into the loop.

8.5 Summary

In this thesis, I have described the optimization of a highly efficient gene editing protocol for introducing point mutations in human iPSCs and the full methodology is available in Chapters 2 and 6. I have demonstrated that an iPSC-based erythroid differentiation system can be used to model a rare genetic disorder caused by a gain-of-function point mutation in the human α-globin locus. There I have demonstrated that the generation of a GATA1 binding site at a specific position in the α-globin locus is sufficient to cause the emergence of a new promoter - GATAν. The new promoter competes with the promoters of the α-globin genes *HBA1* and *HBA2* for enhancer activity and this is the likely cause of the reduced α-globin expression observed in patients. I propose the establishment of the GATAν promoter to be a combination of appropriately positioned repetitive elements and specific erythroid transcription factor motif grammar. Finally, I outline a simplistic working model of the establishment of enhancer promoter interactions in the α-globin locus based on this study and already existing models.

Appendix

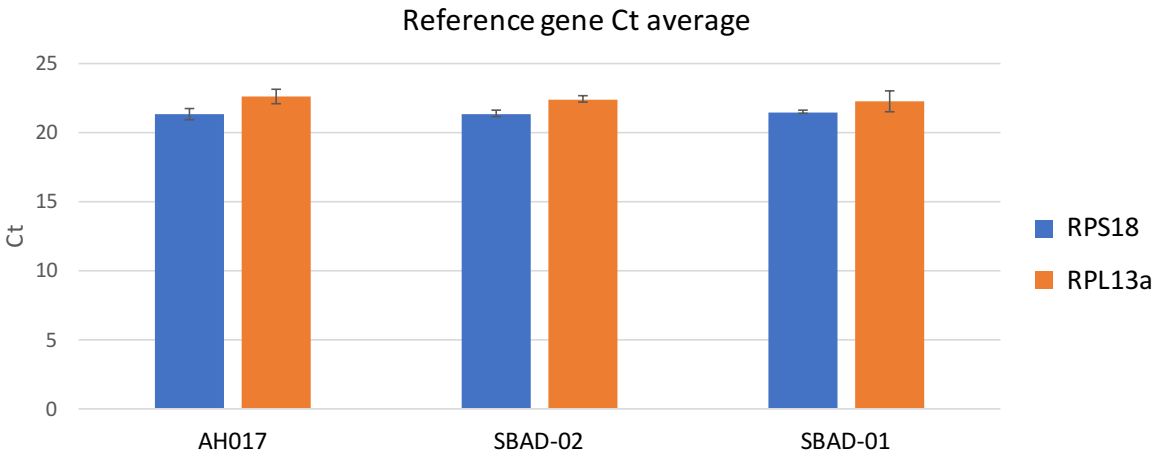


Figure A2.3.2.1 Comparison of CT values reference controls RPS18 and RPL13a for 3 different iPSC cell lines, values represent average from days 10, 14, 17 and 21 of definition. Equal amount of cDNA loaded for each sample. Error bar represents one standard error.
Values for RPS18 show less variance than those for RPL13a in the iPSC differentiation system.

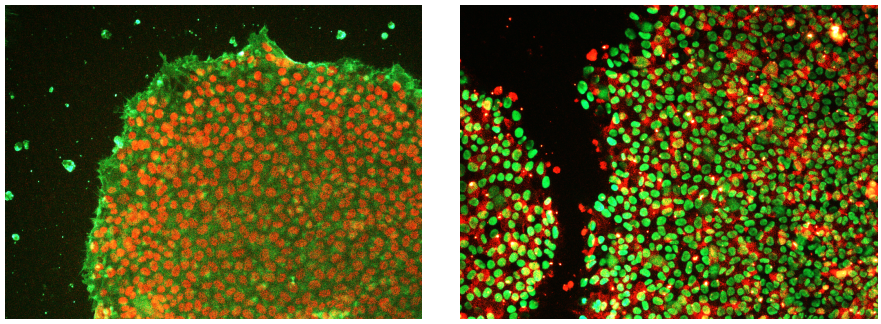
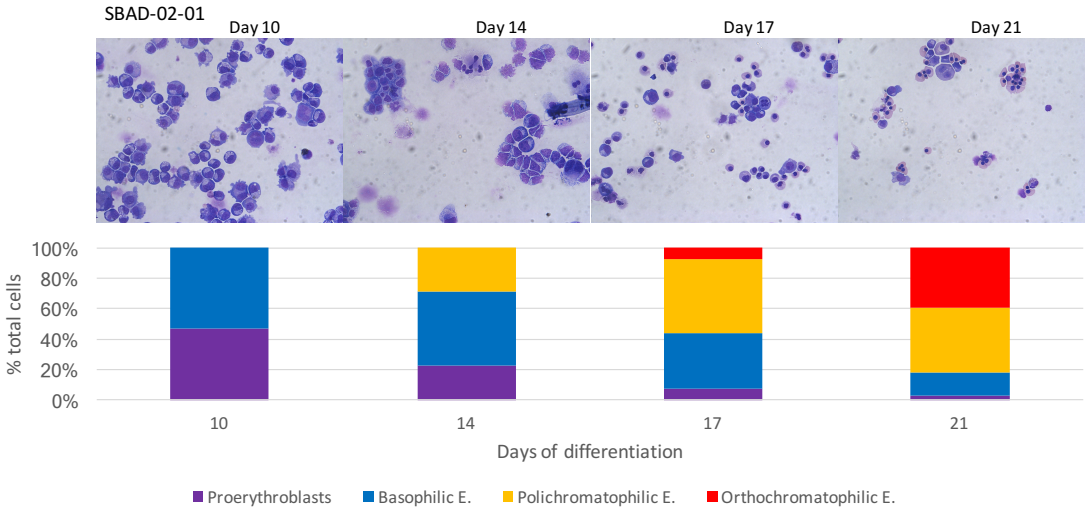
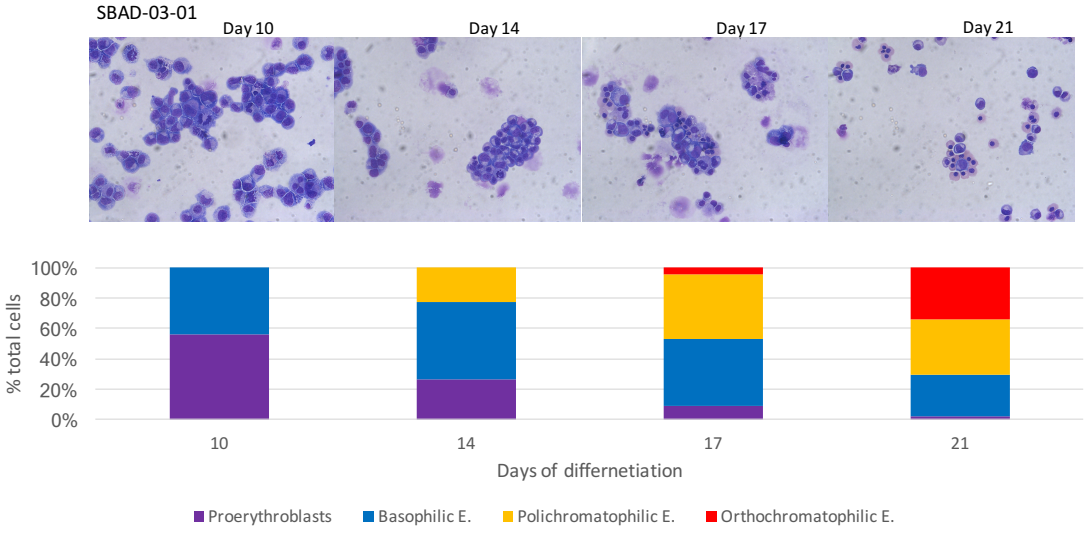


Figure A3.4.2.1 Pluripotency marker IF for line DR07.
The first column shows overlaid staining for SSEA4 in green and OCT4 in red, the second column shows overlaid staining for SOX2 in green and Tra-1-60 in red.

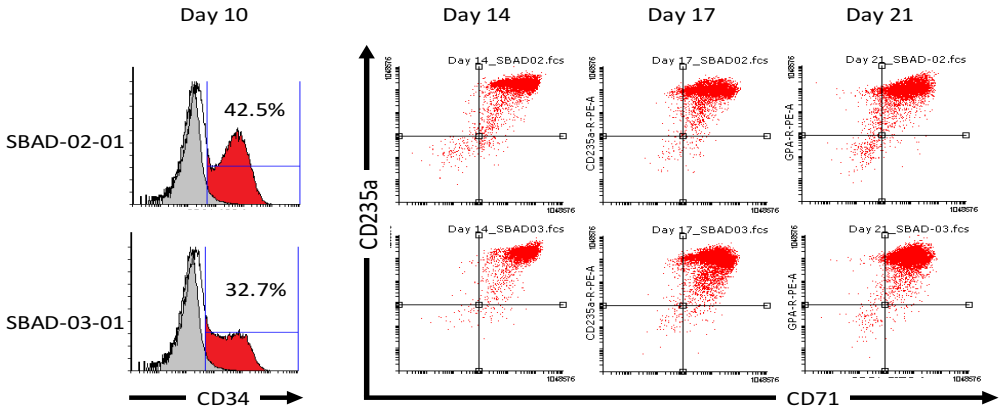
A)



B



C)



	Left Bottom: %	Right Bottom: %	Left Top: %	Right Top: %
Day 14_SBAD-02	2.6	1.1	0.7	95.6
Day 14_SBAD-03	1.8	1.0	0.6	96.6
Day 17_SBAD-02	1.7	1.5	3.5	93.4
Day 17_SBAD-03	1.4	1.1	1.6	96.0
Day 21_SBAD-02	6.3	0.7	5.4	87.6
Day 21_SBAD-03	2.3	0.5	1.9	95.3

Figure A4.2.1. Morphology and FACS analysis of iPSC derived erythroid cells (lines SBAD-02-01 and SBAD-03-01) during differentiation (days 10, 14, 17 and 21)
A) SBAD-02-01 Top panel: modified Wright stain of floating non-adherent cells during differentiation days 10, 14, 17 and 21; Bottom panel: percentages of Pro-, Basophilic-, Polychromatophilic- and Orthochromatophilic- erythroblasts during differentiation days 10, 14, 17 and 21. B) SBAD-02-01 Top panel: modified Wright stain of floating non-adherent cells during differentiation days 10, 14, 17 and 21; Bottom panel: percentages of Pro-, Basophilic-, Polychromatophilic- and Orthochromatophilic- erythroblasts during differentiation days 10, 14, 17 and 21. C) Representative FACS for CD34+ cells on day 10 and CD71+ and CD235a+ on days 14, 17 and 21 of differentiation for lines SBAD-02-01 and SBAD-03-01.

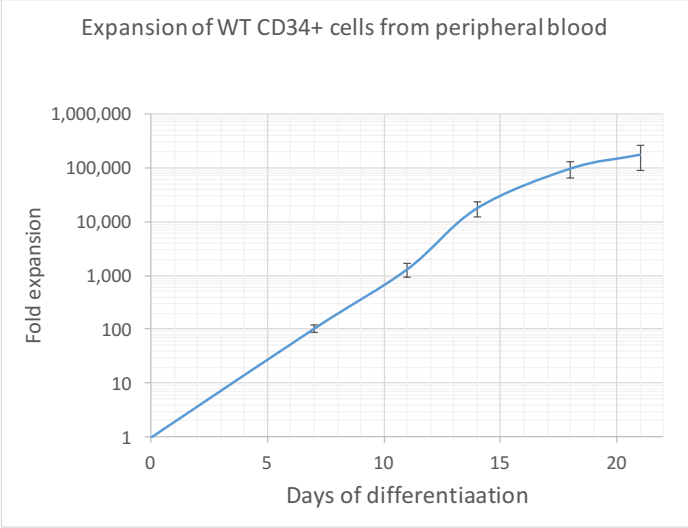


Figure A4.8.1. Fold expansion during erythroid differentiation of CD34+ cells isolated from adult peripheral blood up to day 21, error bars represent one standard deviation. Values are adjusted to represent the expansion of one starting CD34+ cell. Cells were expanded as in Mettananda et al., 2017.

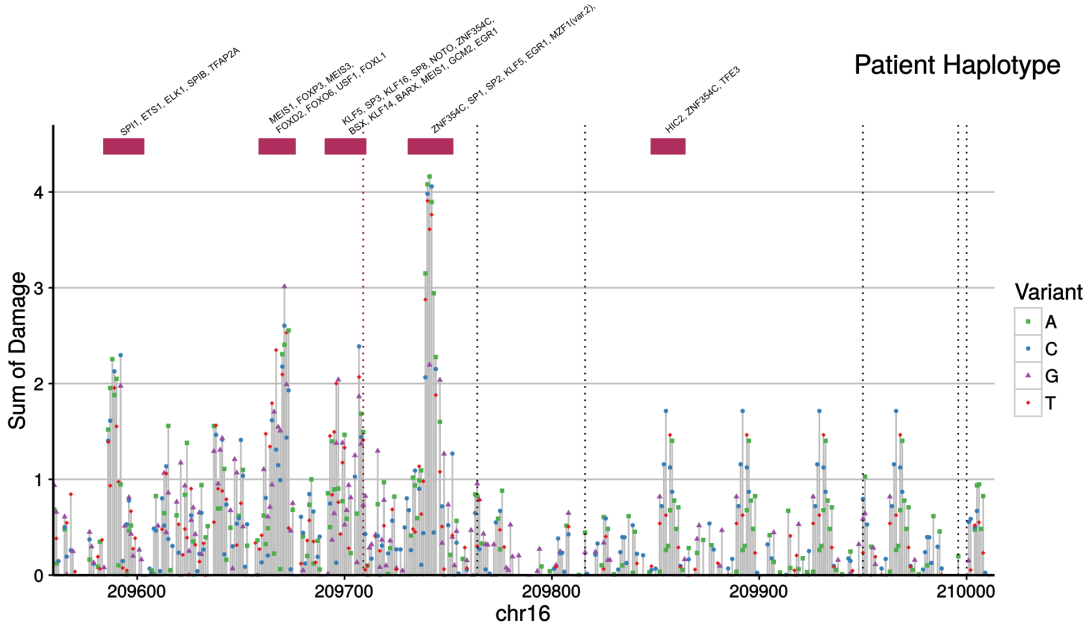


Figure A5.3.1 Sasquatch damage score predictions from *in silico* mutation, based on primary erythroid DNase-seq, for patient haplotype.
Horizontal axis represents 466bp of sequence around the Vanuatu variant in chromosome 16 on human genome build hg19. Dotted lines indicate positions of variants which differ between the patient and the reference haplotype, the Vanuatu variant is indicated as a red dotted line. Vertical axis represents damage score - the potential negative effect a mutation has on the binding affinity of a motif found to be occupied by a transcription factor in DNase-seq footprint data from primary erythroid cells. Each possible variation with respect to the haplotype is predicted. Clusters of high damage scores (peaks) indicate predicted, actively bound transcription factor binding sites in the context of open-chromatin in an erythroid tissue-specific manner. Above the significant peaks are the JASPAR motif predictions above in purple rectangles indicating the queried sequence.

A)

sgRNA	1t
Sequence	GCTGTTATTTTCTAATAATG
Orientation	Pam right
Off-target summary	{0: 0, 1: 0, 2: 0, 3: 19, 4: 269}
Total off-targets	288

WGE ID	Location	Sequence	Mismatches	Strand	Type
1132479404	Original CRISPR	GCTGTTATTTTCTAATAATG TGG (reversed)			Intronic
1132479404	16:159701-159723	GCTGTTATTTTCTAATAATG TGG	0	-	Intronic
1099624303	12:85049281-85049303	GTGGTTATTTGCTAATAATG AGG	3	-	Intronic
1066708844	10:38210992-38211014	GCTATAATTTTCAAATAATG TGG	3	-	Intergenic
951642829	3:24855171-24855193	GCTGTTAATTTTAAATAATG TGG	3	+	Intergenic
1065307172	10:24380259-24380281	GCTGTCACTTTCTTATAATG TGG	3	+	Intronic
1047276451	8:123409122-123409144	GCTGATATATTCTTATAATG TGG	3	+	Intronic
1058870486	9:109197537-109197559	GTTGCTATTTTCTACTAATG GGG	3	+	Intronic
910782933	1:90960883-90960905	GCAGTTATTTTCTTTTAATG AGG	3	-	Intronic
996420984	5:123262325-123262347	GCAGTTATTTTCTTTTAATG TGG	3	-	Intergenic
997681881	5:135762301-135762323	GCTGTTATTTGCTGAAAATG GGG	3	+	Intergenic
972152424	4:36110277-36110299	GCTGTTATCTTCTGATTATG TGG	3	-	Intronic
945502075	2:210588803-210588825	GCTGGTATTCTCTAATAAGG AGG	3	-	Intronic
987365512	5:17145090-17145112	GCTGTGATTTTGTAATAAAG AGG	3	-	Intronic
949630552	3:5920904-5920926	AATGTTATTTTCTAATAATC TGG	3	-	Intergenic
1134477123	16:14584077-14584099	GCTTCTATTTTCTAATAATA TGG	3	-	Intronic
1016369327	6:143355941-143355963	ACTGTGATTTTCTAATAATT GGG	3	+	Intergenic
1075262906	10:120978380-120978402	GCTGTTTATTTCTAATAATT TGG	3	+	Intergenic
1126063930	15:44810738-44810760	GCTCTTATTTTCTAAAAATT AGG	3	+	Intergenic
1165163182	19:33830659-33830681	GCTGATATTTTCTAATGATA AGG	3	+	Intergenic
940163498	2:150741019-150741041	GCTGTTATTTTCAAATGATC AGG	3	-	Intergenic

B)

sgRNA	1c
Sequence	GCTGTTATTTTCTGATAATGNGG
Orientation	Pam right
Off-target summary	{0: 0, 1: 1, 2: 2, 3: 26, 4: 238}
Total off-targets	267

WGE ID	Location	Sequence	Mismatches	Strand	Type
sgRNA1c	16:159701-159723	GCTGTTATTTTCTGATAATG NGG			Not set
1132479404	16:159701-159723	GCTGTTATTTTCTAATAATG TGG	1	-	Intronic
972152424	4:36110277-36110299	GCTGTTATCTTCTGATTATG TGG	2	-	Intronic
997681881	5:135762301-135762323	GCTGTTATTTGCTGAAAATG GGG	2	+	Intergenic
901434517	1:9238608-9238630	GTTGTCATTTACTGATAATG GGG	3	+	Intronic
910782933	1:90960883-90960905	GCAGTTATTTTCTTTTAATG AGG	3	-	Intronic
921135567	1:212256386-212256408	GCTGTCATTTGCTGAAAATG AGG	3	-	Intergenic
930384959	2:50682645-50682667	GGTTTTATCTTCTGATAATG AGG	3	-	Intronic
946406521	2:219494880-219494902	CCTGTTATTTTATGATGATG GGG	3	-	Intronic
964430229	3:156597960-156597982	GATCTAATTTTCTGATAATG AGG	3	+	Intergenic
977031169	4:91885380-91885402	TGTGTTATGTTCTGATAATG TGG	3	-	Intergenic
996420984	5:123262325-123262347	GCAGTTATTTTCTTTTAATG TGG	3	-	Intergenic
1000930007	5:167240135-167240157	GCTTGTATTTTCTGATACTG GGG	3	+	Intergenic
1019773533	7:2898531-2898553	GCAGCTATTTTGTGATAATG AGG	3	+	Intergenic
1047276451	8:123409122-123409144	GCTGATATATTCTTATAATG TGG	3	+	Intronic
1061654396	9:132077846-132077868	GTTGTTAATTTCTGATTATG GGG	3	-	Intronic
1064789603	10:18941541-18941563	CCTGTTTTTTTCTGATAAGG AGG	3	-	Intergenic
1065307172	10:24380259-24380281	GCTGTCACTTTCTTATAATG TGG	3	+	Intronic
1080754316	11:35180923-35180945	GCTTTTATTTTCTGAGAAGG TGG	3	+	Intronic
1089997787	11:122925446-122925468	GCTTTTATATTCTGAAAATG AGG	3	-	Intronic
1121375437	14:93405886-93405908	GCTGGTATCTTCTGATATTG GGG	3	-	Intronic
1125958635	15:43809965-43809987	ACTGTTATTTTCAGATAATT TGG	3	-	Intronic
1129125141	15:73433138-73433160	GCTGTTTTTTGTCTGATAATA AGG	3	+	Intergenic
1133145391	16:3781686-3781708	GATTTTATTTTCTGATATTG GGG	3	-	Intronic
1139330737	16:66187873-66187895	GTTGTTATTATCAGATAATG TGG	3	-	Intergenic
1153204859	18:2687861-2687883	TCTGTTTATTTCTGATAATG GGG	3	+	Intronic
1156902329	18:42314684-42314706	GCTGTAATTTTCTGGTAGTG AGG	3	+	Intergenic
1158926442	18:62268177-62268199	GCTTTTATTTACTGATCATG AGG	3	-	Intronic
1183929555	22:41228163-41228185	GCTGTTTTTTTCTGAACATG GGG	3	-	Intronic
1187557767	X:20368424-20368446	GCTGTACTTTCCTGATAATG AGG	3	-	Intergenic

C)

sgRNA	3t
Sequence	GTATTAGAAAATAACAGCAC
Orientation	Pam right
Off-target summary	{0: 0, 1: 1, 2: 2, 3: 24, 4: 292}
Total off-targets	319

WGE ID	Location	Sequence	Mismatches	Strand	Type
sgRNA3t	16:159706-159728	GTATTAGAAAATAACAGCAC NGG			Not set
1132479405	16:159706-159728	TTATTAGAAAATAACAGCAC AGG	1	+	Intronic
1125104517	15:35954776-35954798	GTATTAGAAAATAATACCAC GGG	2	-	Intergenic
1178246003	21:30953335-30953357	GGATTAGAAAATAACAGAAC AGG	2	-	Intergenic
902187361	1:14735358-14735380	GTCTTTGAAAACAACAGCAC GGG	3	-	Intronic
911878469	1:103200704-103200726	TTATTAGAAAGAAACAGCAC AGG	3	-	Intergenic
935466883	2:103409573-103409595	CTATTAATAATAACAGCAC TGG	3	+	Intergenic
938197565	2:129342874-129342896	GTATTGGAAAATAATAGCTC AGG	3	+	Intergenic
941612170	2:167675613-167675635	GTTTTGGAAAATAACATCAC AGG	3	-	Intergenic
943762049	2:191620592-191620614	ATGTAAGAAAATAACAGCAC TGG	3	-	Intergenic
955462833	3:59203691-59203713	GTATGAGAAAATAACAGAAA TGG	3	-	Intergenic
975459999	4:74640136-74640158	GTATTAGCAAATCCCAGCAC AGG	3	-	Intergenic
985275568	4:188234301-188234323	GTATTAGAAAATGAAAACAC AGG	3	+	Intergenic
986599204	5:9454537-9454559	GGATTGGAAAAAAACAGCAC TGG	3	-	Intronic
1020781674	7:12524154-12524176	TTATTAGAAAATAAAAGAAC AGG	3	-	Intergenic
1030284307	7:107809853-107809875	CTATTAAAAAATAACAGTAC AGG	3	+	Intergenic
1035744407	8:1951316-1951338	GCTTTAGAAAGTAACAGCAC CGG	3	+	Intronic
1068030083	10:51695745-51695767	GTAGTAGAAAACAACAGCAT AGG	3	+	Exonic
1117175456	14:53141581-53141603	GTGTTAGAAATTAACAACAC AGG	3	+	Intronic
1118552301	14:66967782-66967804	GTAGTAGAAAATAATAGCAG GGG	3	+	Intronic
1120290391	14:82562751-82562773	TTATTAGAAAAAAACAACAC AGG	3	+	Intergenic
1122396506	14:101436419-101436441	GTATGAGAAGCTAACAGCAC AGG	3	-	Intergenic
1145906556	17:28519546-28519568	GTATCAGCAAATACCAGCAC AGG	3	-	Intronic
1160102525	18:75936438-75936460	CTATTAGAAAATCTCAGCAC AGG	3	+	Intergenic
1188939949	X:36225385-36225407	GCACTAGAAAATGACAGCAC AGG	3	+	Intergenic
1190617853	X:52255650-52255672	CTTTTAGAAAATAAAAGCAC AGG	3	-	Intergenic
1197172531	X:123450418-123450440	GTAGTAGAAAAAAACAACAC AGG	3	+	Intronic
1198619963	X:138496439-138496461	GTATTGGAAAAAAACAGTAC CGG	3	-	Intergenic

D)

sgRNA	3c
Sequence	GTATCAGAAAATAACAGCAC
Orientation	Pam right
Off-target summary	{0: 0, 1: 0, 2: 2, 3: 22, 4: 239}
Total off-targets	263

WGE ID	Location	Sequence	Mismatches	Strand	Type
sgRNA3c	16:159706-159728	GTATCAGAAAATAACAGCAC NGG			Not set
1132479405	16:159706-159728	TTATTAGAAAATAACAGCAC AGG	2	+	Intronic
1145906556	17:28519546-28519568	GTATCAGCAAATACCAGCAC AGG	2	-	Intronic
909773920	1:79460557-79460579	GTATCTGAAAACAACAGCAA TGG	3	+	Intergenic
916042377	1:160972107-160972129	GTATCAGAAAATAAAAGGAA TGG	3	+	Intergenic
943091468	2:183380534-183380556	GAATCAGAAAATAGCAGAAC TGG	3	-	Intergenic
943762049	2:191620592-191620614	ATGTAAGAAAATAACAGCAC TGG	3	-	Intergenic
951864768	3:27295306-27295328	GTATCAGAATATAACAACAG GGG	3	-	Intronic
952398000	3:32938096-32938118	CTGTCAGAAAATAATAGCAC TGG	3	-	Intergenic
955462833	3:59203691-59203713	GTATGAGAAAATAACAGAAA TGG	3	-	Intergenic
982534655	4:156595150-156595172	GTAACAGAAAATAACTGAAC AGG	3	+	Intergenic
987247874	5:16067502-16067524	TTATCAGAAGATAACAGCTC GGG	3	-	Exonic
1007084121	6:39131066-39131088	CTATCAGCAAATAATAGCAC TGG	3	-	Intergenic
1015249027	6:131107423-131107445	GTATCATAGAATAACAGAAC TGG	3	+	Intergenic
1045643267	8:104275319-104275341	GTATCACAAAATAATATCAC AGG	3	+	Intergenic
1047108671	8:121764191-121764213	GGATCAGTAAATAACAGCCC TGG	3	-	Intergenic
1050702923	9:8361482-8361504	GTATCAGATTATAACAGCAT AGG	3	+	Intronic
1119341753	14:73885321-73885343	GTTTCAGAAAATGACAGGAC TGG	3	+	Exonic
1122396506	14:101436419-101436441	GTATGAGAAGCTAACAGCAC AGG	3	-	Intergenic
1125104517	15:35954776-35954798	GTATTAGAAAATAATACCAC GGG	3	-	Intergenic
1133170513	16:3979994-3980016	GTTTCAGAAAATCACACCAC AGG	3	+	Intronic
1135714864	16:24754490-24754512	GAATCAGAAAATAATAGCCC TGG	3	-	Intronic
1175055454	20:56193545-56193567	GTAGCAGAAAAGAAAAGCAC TGG	3	-	Intergenic
1178246003	21:30953335-30953357	GGATTAGAAAATAACAGAAC AGG	3	-	Intergenic
1186509846	X:10122574-10122596	GGGTCAGGAAATAACAGCAC CGG	3	-	Intronic

Figure A6.2.1. Off-target sites predicted for sgRNA1t (A), sgRNA1c(B), sgRNA3t(C), sgRNA3c(D) using the WGE genome editing tool (Hodgkins et al., 2015). A summary of the number of sites with 0,1,2,3 or 4 mismatches is presented at the top of each section, sites that have 3 mismatches or less are listed as sequences and their position is indicated on the left. *Note: original position of short guide RNAs is labelled as intronic due to miss notation by Ensembl fusing *HBZ* and *HBM.*

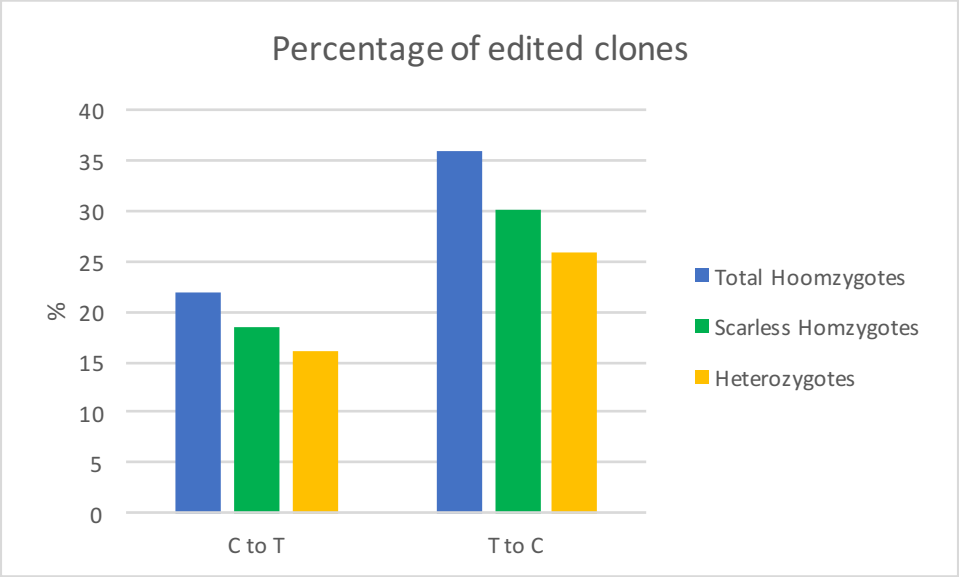
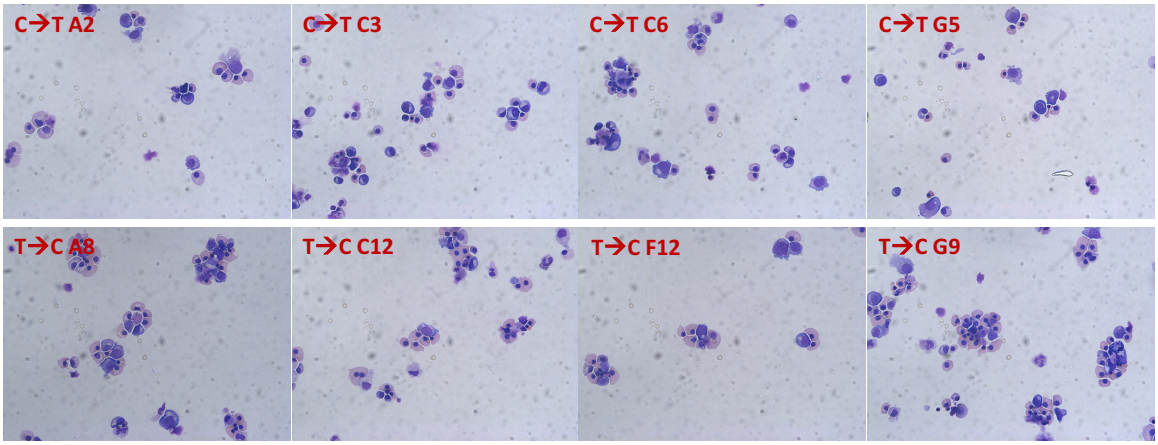
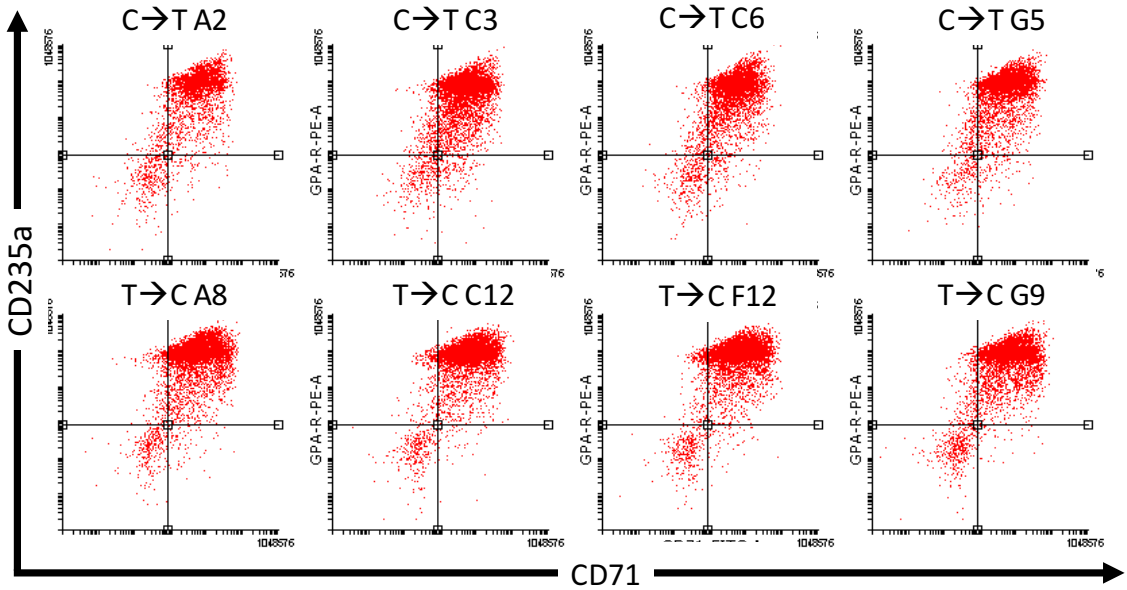


Figure A6.3.1. Percentage of homozygous and heterozygous edited clones from the C to T allele genome editing (iPSC line LA01) and from the T to C allele genome editing (iPSC line SBAD-02-01). All homozygotes are in blue, homozygotes without small indels are in green, all heterozygotes are in grey.

A)



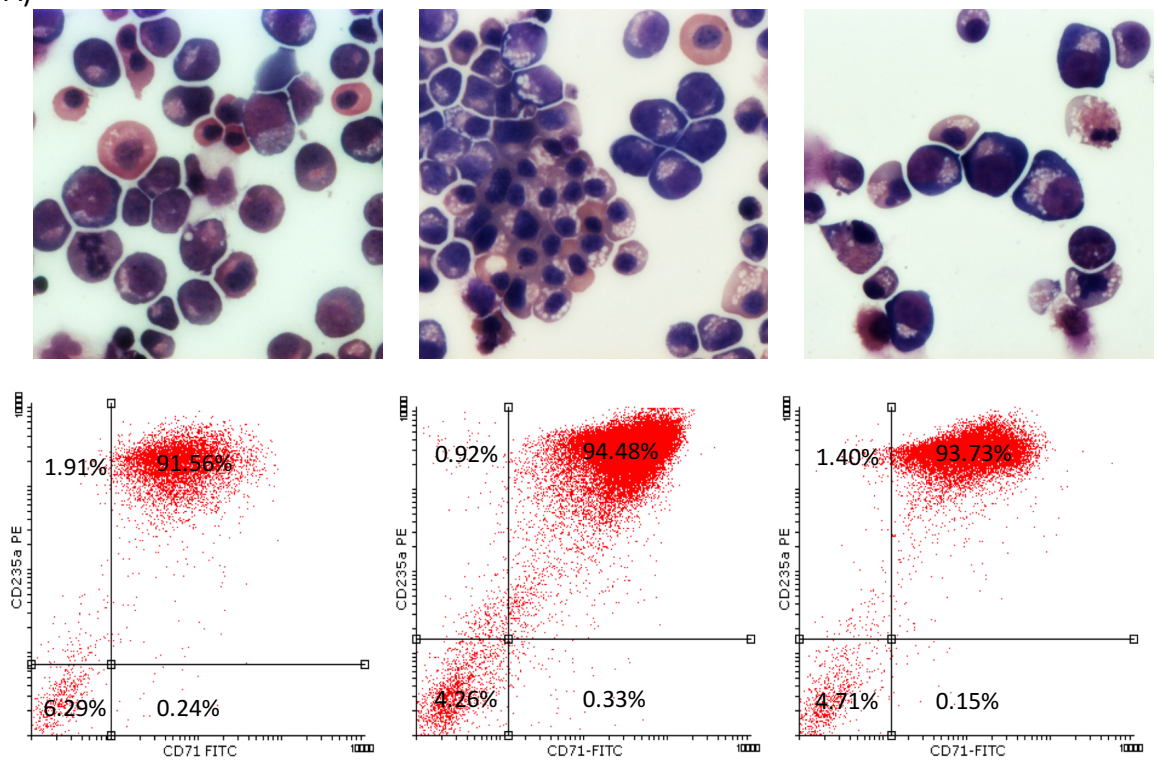
B)



	Left Bottom: %	Right Bottom: %	Left Top: %	Right Top: %
Day 21_CtoT A2	17.1	3.0	4.8	75.1
Day 21_CtoT C3	7.9	3.8	5.0	83.4
Day 21_CtoT C6	12.7	3.2	4.9	79.2
Day 21_CtoT G5	8.0	2.7	4.7	84.6
Day 21_TtoC A8	8.4	1.7	4.7	85.3
Day 21_TtoC C12	9.1	1.1	3.6	86.3
Day 21_TtoC F12	9.2	1.2	3.9	85.7
Day 21_TtoC G9	15.3	1.2	5.1	78.4

Figure A7.2.1 Morphology and FACS for day 21 for edited clones (C to T and T to C transition). A) Modified Wright stain of iPSC derived erythroid-like cells on differentiation day 21. B) Representative FACS analysis for CD71+ and CD235a+ on day 21 of differentiation.

A)



B)

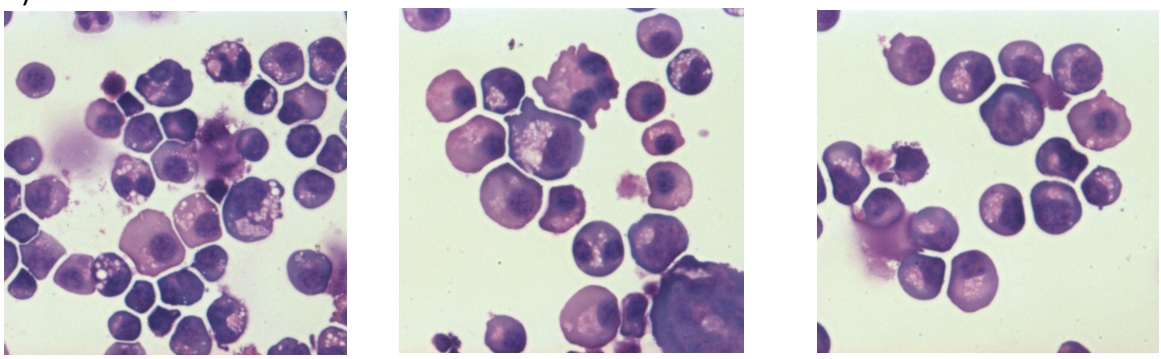
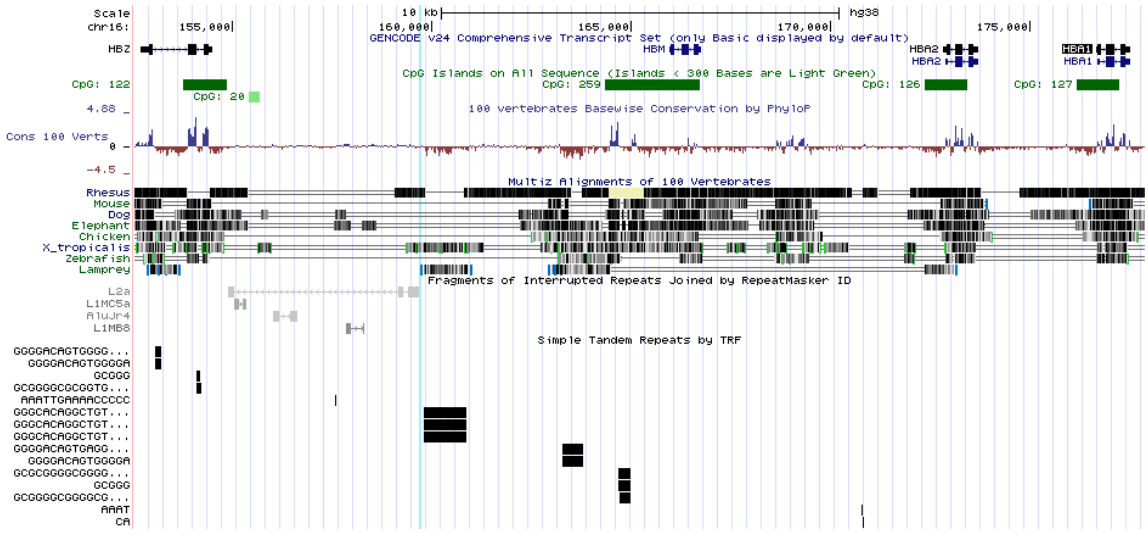


Figure A7.2.2 Morphology and FACS for day 21 for triplicates of erythroid differentiation of iPSC lines LA01 and SBAD-02-01 (isogenic controls).
A) Top panel: modified Wright stain of LA01 iPSC-derived erythroid-like cells on differentiation day 21. Bottom panel: representative FACS analysis for CD71+ and CD235a+ on day 21 of differentiation.
B) Modified Wright stain of SBAD-02-01 iPSC-derived erythroid-like cells on day 21 of differentiation.

A)



B)

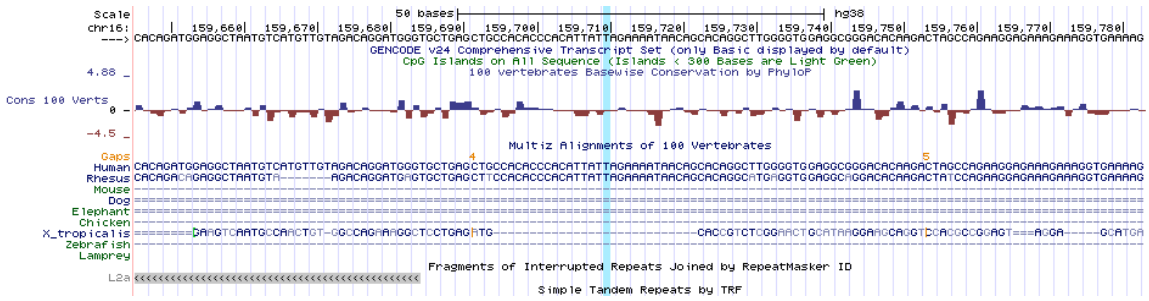


Figure A8.1.1. Repeat elements, CpG Islands and conservation at the α-globin locus.
UCSC genome browser was used to visualise different elements, the Vanuatu site is highlighted in blue. From top to bottom panel: Scale displaying chromosome position, RefSeq genes, CpG island positions in green, Conservation score among 100 vertebrates, regions of conservation is several vertebrate species, Interrupted repeats, Simple tandem repeats.
A) Region between *HBZ* and *HBA1*.
B) Zoomed in view of the Vanuatu SNV (highlighted in blue).

CTGTGGATAAAGAAAAGCAGGGCAAGTAGGGGCTTAGACAGAGGAGGGGAGGATTCAGATTT
TAAATGGGTTGGCCACTGTAGGTCTATTAACGTGGTGACATTTGAGGGAGTGGCAATACTAG
GGAAGGGGCTTCAGGGGAGTGGCCAGGAGCTAGGGATAGAGGGAGGGAGGACAGGAGGCCTT
GTCTGTCTTTTCCTCCATATGTAAGTTTCAGGAGTGAGTGGGGGGTGTCGAGGGTGCTGTGC
TCTCCGGCCTGAGCCTCAGGAAGGAAGGGCAGTAGTCAGGGATGCCAGGGAAGGACAGTGGA
GTAGGCTTTGTGGGGAACTTCACGGTTCCATTGTTGAGATGATTTGCTGGAGACACACAGAT
GAGGACATCAAATACATCCCTGGATCAGGCCCTGGGGCCTGAGTCCGGAAGAGAGGTCTGTA
TGGACACACCCATCAATGGGAGCACCAGGACACAGATGGAGGCTAATGTCATGTTGTAGACA
GGATGGGTGCTGAGCTGCCACACCCACATTATCAGAAAATAACAGCACAGGCTTGGGGTGGA
GGCGGGACACAAGACTAGCCAGAAGGAGAAAGAAAGGTGAAAAGCTGTTGGTGCAAGGAAGC
TCTTGGTATTTCCAATGGCTTgggcacaggctgtgagggtgcctgggacggcttgtggggca
caggctgcaagaggtgccaggacggcttgtggggcacaggttgtgagaggtgccctggacg
gcttgtggggcacaggctgtgagaggtgcccaggacggcttgtggggcacaggctgtgaggg
tgcccgggacggcttgtggggcacaggttgtgagaggtgcccgggacggcttgtggggcaca
ggtttcagaggtgcccgggacggcttgtggggcacaggttgtgagaggtgcccgggacggct
tgtgggacacaggttgtgagaggtgcctgggacggcttgtggggcacaggctgtgagggtgc
ctgggacg

Figure A8.1.2 Core promoter element scan of 1000bp region encompassing the Vanuatu mutation.
Elements Navigation Tool (ElemeNT) from the Juven-Gershon Lab (Sloutskin et al., 2015) was used with recommended settings. LINE2 element sequence is highlighted in grey, tandem repeat sequence is in lower case, the new GATA site is highlighted in green and the Vanuatu SNV is in red. Combinations of mammalian Inr element (blue) and DPE (purple) are underlined.

CCTTGGGGAGGTTCTAGCCCCTGAGCACCGGAGCTGCGGCCCGGGTGGAGCGGAGCAGTCCCGGGCCGGCCCG
CGGCGTCTCCTGGGGTCCTTGAGTCGGACGGGCGTTTGTGCGTCTCCCGGCTTCCCATATCGCACAAAGATTG
TCACTTCACTAAGCGTATTGGAAGCGTGTCGGGGCTCAGGGAACTTTTCCACAAAGCCTGACGTCCGAATCCC
GGGACTCTGGCAGCTACGGGGGTCCCTGAGGCCGGTCCCTCCCCGACTCCTAAGAGAGTAGGGGGTTTCCTGC
CCGGTGTTCTCTCTCCGGTTCCTCCCATGTGCTCCCTCCTGGCAGAGCAGTAACTTTACCCGAGGGGAGTAAA
CAGATGCCCCTAAAGTCTGCAGTAAAGGTGCCCACGCGCAACGGCGTGGGTCAATGCCAGAAACCCTGGGATC
CCGGAGGTCGAGGCCTCCACACAGACGGGAACCCGCGCTGGTTACGTTCCCCGGCGCAGGCCGAGGGTCCCCG
CGTTCCCGCCGCGCTCGGGCCGATAAGGACGGGCGGGGTGCCCGGAGGCTCTATAAGGAGGCCAGGGCGGCGG
GCGCGGCCCCCAGAGCACGTCAGGCGGCGCCATGCTCAGCGCCCAGGAGCGCGCCCAAATCGCGCAGGTCTGG
GACCTGATTGCGGGCCACGAGGCGCAATTCGGGGCGGAGCTGCTGCTCAGGTCGGTAGAGGCGGGGTCTCCGG
GAGCTCAGGGAGGTGGAGATGAGGGTTTTGGGCGCGTGGGCCGCCAACGCCATCCAAGGTCCTTCGGGTGCGG
ATCCCCGGGGCTCTGGGCGGTGTGGGCGCTAGTGAAGCCCCACGCAGCCGCCCTCCTCCCCGGTCACTGACCT
GGTCCTGCAGGCTCTTCACGGTGTACCCCAGCACCAAGGTCTACTTCCCGCACCTGAGCGCCTGCCAGGA

Figure A8.4.1. LA01 iPSC line genome sequence under HBM associated GATA1 peak. Only one SNV is present (in red) which does not disrupt any know GATA1 motif.

Word count: 38,000 words excluding bibliography, appendices, diagrams and tables

References

Aasen, T., et al., (2008). Efficient and rapid generation of induced pluripotent stem cells from human keratinocytes. *Nat. Biotechnol*. **26**, 1276-1284

Alberts, B., (2008). Molecular biology of the cell. 5th edition, Garland Science

Alipour, E, and Marko, J., (2012) Self-organization of domain structures by DNA-loop extruding enzymes. *Nucl Acids Res* **40**, 11202–11212

Alter, B., et al., (1976). Fetal erythropoiesis following bone marrow transplantation. *Blood* **48**, 843–853

Alter, B., (1979). Fetal erythropoiesis in stress hematopoiesis. *Exp Hematol.* **7**, 200–209

Anguita, E., et al., (2001). Identification of a conserved erythroidspecific domain of histone acetylation across the alpha-globin gene cluster. *PNAS* **98**, 12114-12119

Anguita, E., et al., (2004). Globin gene activation during haemopoiesis is driven by protein complexes nucleated by GATA-1 and GATA-2. *EMBO Journal* **23**, 2841-2852

Antoniani, C., et al., (2017). Concise Review: Epigenetic Regulation of Hematopoiesis: Biological Insights and Therapeutic Applications. *Stem Cells Transl Med*. **6**, 2106–2114

Aoi, T., et al., (2008). Generation of pluripotent stem cells from adult mouse liver and stomach cells. *Science* **321**, 699-702

Armstrong, J., et al., (1998). A SWI/SNF-related chromatin remodeling complex, E-RC1, is required for tissue-specific transcriptional regulation by EKLF in vitro. *Cell* **95**, 93–104

Avior, Y., et al., (2016). Pluripotent stem cells in disease modelling and drug discovery. *Nat. Rev. Mol. Cell Biol.* **17**, 170-182

Bajic, V., et al., (2006). Mice and men: their promoter properties. *PLoS Genet* **2**, e54

Baker, D., et al., (2007). Adaptation to culture of human embryonic stem cells and oncogenesis in vivo. *Nat. Biotechnol*. **25**, 207–215

Bannister, A. and Kouzarides, T. (2011). Regulation of chromatin by histone modifications. *Cell Research* **21**, 381-395

Banerji, J., et al., (1981). Expression of a beta-globin gene is enhanced by remote SV40 DNA sequences. *Cell* **27**, 299–308

Bar-Nur, O., et al., (2011). Epigenetic memory and preferential lineage-specific differentiation in induced pluripotent stem cells derived from human pancreatic islet beta cells. *Cell Stem Cell* **9**, 17–23

Barrangou, R., et al., (2007). CRISPR provides acquired resistance against viruses in prokaryotes. *Science* **315**, 1709–1712

Barski, A., et al., (2007). High-resolution profiling of histone methylations in the human genome. *Cell* **129**, 823–837

Bauer, D., et al., (2013). An erythroid enhancer of *BCL11A* subject to genetic variation determines fetal hemoglobin level. *Science* **342**, 253–257

Begley, C., et al., (1989). A new gene, SCL, is expressed during early haematopoiesis and encodes a differentiation-related DNA-binding motif. *PNAS* **86**, 10128-10132

Bell, A., Sallah, S., and Diggs, L., (2005). *The Morphology of Human Blood Cells.* 7th Edition, Abbott Park, Illinois: Abbott Laboratories

Bell, O., et al. (2011). Determinants and dynamics of genome accessibility. *Nature reviews. Genetics* **12**, 554-564

Ben-David, U., and Benvenisty, N., (2012). High prevalence of evolutionarily conserved and species-specific genomic aberrations in mouse pluripotent stem cells. *Stem Cells.* **30**, 612–622

Benko, S., et al. 2009. Highly conserved noncoding elements on either side of SOX9 associated with Pierre Robin sequence. *Nat Genet* **41**, 359–364

Bertrand, J., et al., (2010). Haematopoietic stem cells derive directly from aortic endothelium during development. *Nature* **464**, 108–111

Bernardini, L., et al., (2010). High-resolution SNP arrays in mental retardation diagnostics: how much do we gain? *Eur. J. Hum. Genet.* **18**, 178–185

Bernstein, B., et al., (2005). Genomic maps and comparative analysis of histone modifications in human and mouse. *Cell* **120**, 169–181

Bieker, J., (2001). Kruppel-like factors: Three fingers in many pies. *J. Biol. Chem*. **276**, 34355–34358

Bird, A., et al. (1987). Non-methylated CpG-rich islands at the human α-globin locus: implications for evolution of the α-globin pseudogene. *EMBO Journal* **6**, 999-1004

Blackwood E., and Kadonaga, J., (1998). Going the distance: a current view of enhancer action. *Science* **281**, 61-63

Blelloch, R., et al., (2007). Generation of induced pluripotent stem cells in the absence of drug selection. *Cell Stem Cell* **1**, 245-247

Bloom, W., and Bartelmez, G., (1940). Hematopoiesis in young human embryos. *Am. J. Anat.* **67**, 21–53

Blum, B., et al., (2009). The anti-apoptotic gene survivin contributes to teratoma formation by human embryonic stem cells. *Nat. Biotechnol*. **27**, 281–287

Boisset, J., et al., (2010). In vivo imaging of haematopoietic cells emerging from the mouse aortic endothelium. *Nature* **464**, 116–120

Boland, M., et al., (2009). Adult mice generated from induced pluripotent stem cells. *Nature* **461**, 91–94

Bolotin, A., et al., (2005). Clustered regularly interspaced short palindrome repeats (CRISPRs) have spacers of extrachromosomal origin. *Microbiology* **151**, 2551–2561

Borg, J., et al., (2010). Haploinsufficiency for the erythroid transcription factor KLF1 causes hereditary persistence of fetal hemoglobin. *Nature Publishing Group* **42**, 813-817

Brouha, B., et al., (2003). Hot L1s account for the bulk of retrotransposition in the human population. *PNAS* **100**, 5280 – 5285

Bottardi, S., et al., (2006). Lineagespecific activators affect beta-globin locus chromatin in multipotent hematopoietic progenitors. *EMBO Journal* **25**, 3586 –3595

Boulting, G., et al., (2011). A functionally characterized test set of human induced pluripotent stem cells. *Nature Biotechnology* **29**, 279-286

Boyer, L., et al., (2005). Core transcriptional regulatory circuitry in human embryonic stem cells. *Cell* **122**, 947 –956

Bresnick, E., et al., (2010). GATA switches as developmental drivers. *J. Biol. Chem.* **285**, 31087–31093

Briggs, R. and King, T., (1952). Transplantation of Living Nuclei From Blastula Cells into Enucleated Frogs' Eggs. *PNAS* **38**, 455-463

Brouns, S., et al., (2008). Small CRISPR RNAs guide antiviral defense in prokaryotes. *Science* **321**, 960–964

Brouwer, M., et al., (2016). Choices for Induction of Pluripotency: Recent Developments in Human Induced Pluripotent Stem Cell Reprogramming Strategies. *Stem Cell Rev and Rep* **12**, 54-72

Brown, J., et al., (2006). Co-regulated human globin genes are frequently in spatial proximity when active. *Journal of Cell Biology* **172**, 177–187

Buecker, C. and Wysocka, J. (2012). Enhancers as information integration hubs in development: lessons from genomics. *Trends in Genetics* **28**, 276–284

Buenrostro, J. et al., (2013). Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat. Methods* **10**, 1213-1218

Bulger, M. and Groudine, M. (2011). Functional and mechanistic diversity of distal transcription enhancers. *Cell* **144**, 327–339

Butler, J., and Kadonaga, J., (2002) The RNA polymerase II core promoter: a key component in the regulation of gene expression. *Genes Dev* **16**, 2583–2592

Canver, M., et al., (2014). Characterization of Genomic Deletion Efficiency Mediated by Clustered Regularly Interspaced Palindromic Repeats (CRISPR)/Cas9 Nuclease System in Mammalian Cells. *The Journal of Biological Chemistry* **289**, 21312-21324

Capellera-Garcia, S., et al., (2016). Defining the Minimal Factors Required for Erythropoiesis through Direct Lineage Conversion. *Cell Reports* **15**, 2550–2562

Carninci, P., et al., (2006). Genome-wide analysis of mammalian promoter architecture and evolution. *Nat Genet* **38**, 626–635

Cavazzana-Calvo, M., et al., (2010). Transfusion independence and HMGA2 activation after gene therapy of human beta-thalassaemia. *Nature* **467**, 318–322

Chang, C., et al., (2011). Production of Embryonic and Fetal-Like Red Blood Cells from Human Induced Pluripotent Stem Cells. *PLoS ONE* **6,** e25761

Chang, C., et al., (2015). Modeling human severe combined immunodeficiency and correction by CRISPR– Cas9‑enhanced gene targeting. *Cell Rep.* **12**, 1668–1677

Chang, K., et al., (2006). Definitive-like erythroid cells derived from human embryonic stem cells coexpress high levels of embryonic and fetal globins with little or no adult globin. *Blood* **108**, 1515–1523

Charbord, P., et al., (1996). Early ontogeny of the human marrow from long bones: an immunohistochemical study of hematopoiesis and its microenvironment. *Blood* **87**, 4109–4119

Choi, K., et al., (1998). A common precursor for hematopoietic and endothelial cells. *Development* **125**, 725–732

Choi, K. et al., (2009). Generation of mature human myelomonocytic cells through expansion and differentiation of pluripotent stem cell-derived lin-CD34+CD43+CD45+ progenitors. *J Clin Invest.* **119**, 2818-2829

Choulika, A., et al., (1995). Induction of homologous recombination in mammalian chromosomes by using the I-SceI system of Saccharomyces cerevisiae. *Mol. Cell. Biol.* **15**, 1968–1973

Chuong, E., et al., (2013). Endogenous retroviruses function as species-specific enhancer elements in the placenta. *Nat Genet* **45**, 325–329

Cirillo, L., et al., (2002). Opening of compacted chromatin by early developmental transcription factors HNF3 (FOXA) and GATA-4. *Mol. Cell.* **9**, 279–289

Coelho, A., et al., (2010). Novel large deletions in the human alpha-globin gene cluster: clarifying the HS-40 long-range regulatory role in the native chromosome environment. *Blood Cells Mol. Dis.* **45**, 147–153

Cowan, C., et al., (2005). Nuclear reprogramming of somatic cells after fusion with human embryonic stem cells. *Science* **309**, 1369-1373

Craddock, C., et al., (1995). Contrasting effects of alpha and beta globin regulatory elements on chromatin structure may be related to their different chromosomal environments. *EMBO Journal* **14**, 1718-1726

Crispino, J., et al., (1999). Use of Altered Specificity Mutants to Probe a Specific Protein - Protein Interaction in Differentiation: the GATA-1 : FOG Complex. *Molecular Cell* **3**, 219-228

Cuesta, I., et al., (2007). The forkhead factor FoxE1 binds to the thyroperoxidase promoter during thyroid cell differentiation and modifies compacted chromatin structure. *Mol Cell Biol* **27**, 7302–7314

Davies, J., et al., (2016). Multiplexed analysis of chromosome conformation at vastly improved sensitivity. *Nature Methods* **13**, 74-80

Davis, R., et al., (1987). Expression of a single transfected cDNA converts fibroblasts to myoblasts. *Cell* **51**, 987–1000

De Gobbi, M., et al., (2006). A Regulatory SNP Causes a Human Genetic Disease by Creating a New Transcriptional Promoter. *Science* **312**, 1215-1217

De Gobbi, M., et al., (2007). Tissue-specific histone modification and transcription factor binding in globin gene expression. *Blood* **110,** 4503-4510

De Gobbi, M., et al., (2011). Generation of bivalent chromatin domains during cell fate decisions. *Epigenetics & chromatin* **4**, 9

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

Deleidi, M., and Yu, C., (2016). Genome editing in pluripotent stem cells: research and therapeutic applications. Biochem. *Biophys. Res. Commun*. **473**, 665–674

Deltcheva, E., et al., (2011). CRISPR RNA maturation by trans-encoded small RNA and host factor RNase III. *Nature* **471**, 602–607

Deng, W., et al., (2012). Controlling long-range genomic interactions at a native locus by targeted tethering of a looping factor. *Cell* **149**, 1233-1244

Dhar, V., et al., (1988). The co-ordinate replication of the human β-globin gene domain reflects its transcriptional activity and nuclease hypersensitivity. *Mol. Cell Biol.* **8**, 4958-4965

Dias, J., et al., (2011). Generation of red blood cells from human induced pluripotent stem cells. *Stem Cells Dev*. **20**, 1639–1647

Dimos, J., et al. (2008). Induced pluripotent stem cells generated from patients with ALS can be differentiated into motor neurons. *Science* **321**, 1218–1221

Ding, Q., et al., (2013). Enhanced efficiency of human pluripotent stem cell genome editing through replacing TALENs with CRISPRs. *Cell Stem Cell* **12**, 393–394

Do, J., and Scholer, H., (2004). Nuclei of embryonic stem cells reprogram somatic cells. *Stem Cells* **22**, 941-949

Dolznig, H. et al., (2005). Expansion and differentiation of immature mouse and human hematopoietic progenitors. *Methods in molecular medicine* **105**, 323—344

Doudna, J., and Charpentier, E., (2014). The new frontier of genome engineering with CRISPR-Cas9. *Science* **346**, 1258096

Dow, L., (2015). Modeling disease in vivo with CRISPR/Cas9. Trends Mol. Med. 21:609–21

Draper, J., et al., (2004). Recurrent gain of chromosomes 17q and 12 in cultured human embryonic stem cells. *Nat. Biotechnol.* **22**, 53–54

Drissen, R. et al., (2004). The active spatial organization of the β-globin locus requires the transcription factor EKLF. *Genes and Development* **18**, 2485-2490

Dynan, W., and Tjian, R., (1985). Control of eukaryotic messenger RNA synthesis by sequence-specific DNAbinding proteins. *Nature* **316**, 774–778

Egli, D., et al., (2007). Developmental reprogramming after chromosome transfer into mitotic mouse zygotes. *Nature* **447**, 679-685

Eiges, R., et al., (2007). Developmental study of fragile X syndrome using human embryonic stem cells derived from preimplantation genetically diagnosed embryos. *Cell Stem Cell* **1**, 568–577

Eilken, H., et al., (2009). Continuous single-cell imaging of blood generation from haemogenic endothelium. *Nature* **457**, 896–900

El Omari, K., et al., (2013). Structural basis for LMO2-driven recruitment of the SCL:E47 bHLH heterodimer to hematopoietic-specific transcriptional targets. *Cell Reports* **4**, 135-147

Eminli, S., et al., (2008). Reprogramming of neural progenitor cells into induced pluripotent stem cells in the absence of exogenous Sox2 expression. *Stem Cells* **26**, 2467-2474

England, S., et al., (2011). Immature erythroblasts with extensive ex vivo self-renewal capacity emerge from the early mammalian fetus. *Blood* **117**, 2708–2717

Epner, E., et al., (1981). Replication of alpha and beta globin DNA sequences occurs during early S phase in murine erythroleukemia cells*. PNAS* **78**, 3058-3062

Esperet, C., et al., (2000). Non-erythroid Genes Inserted on Either Side of Human HS-40 Impair the Activation of Its Natural α-Globin Gene Targets without Being Themselves Preferentially Activated**.** *J. Biol. Chem.* **275**, 25831–25839

Evans, M., and Kaufman, M., (1981). Establishment in culture of pluripotential cells from mouse embryos. *Nature* **292**, 154-156

Faulkner, G., and Carninci, P., (2009). Altruistic functions for selfish DNA. *Cell Cycle* **8**, 2895–2900

Faulkner, G., et al., (2009). The regulated retrotransposon transcriptome of mammalian cells. *Nature Genet.* **41**, 563–571

Ferreira, R., et al., (2005). GATA1 Function, a Paradigm for Transcription Factors in Hematopoiesis. *Molecular and Cellular Biology* **25**, 1215–1227

Fibach, E. et al., (1993). Enhanced fetal hemoglobin production by phenylacetate and 4-phenylbutyrate in erythroid precursors derived from normal donors and patients with sickle cell anemia and beta-thalassemia. *Blood* **82**, 2203-2209

Frith, M., et al., (2008). A code for transcription initiation in mammalian genomes. *Genome Res* **18**, 1–12

Fischel-Ghodsian, N., et al. (1987). Unusual features of CpG-rich (HTF) islands in the human α-globin complex: association with non-functional pseudogenes and presence within the 3' portion of the ζ genes. *Nucleic Acids Research* **15**, 9215-25

FitzGerald, P., et al., (2004). Clustering of DNA sequences in human promoters. *Genome Res* **14**, 1562–1574

Flint, J., et al., (2001). Comparative genome analysis delimits a chromosomal domain and identifies key regulatory elements in the alpha globin cluster. *Hum. Mol. Genet.* **10**, 371-382

Frattini, A., et al. (2015). High variability of genomic instability and gene expression profiling in different HeLa clones. *Nature Scientific Reports* **5**, 15377

Frazer, K., et al., (2009). Human genetic variation and its contribution to complex traits. *Nature Reviews Genetics* **10**, 241–251

Fu, Y., et al., (2013). High-frequency off-target mutagenesis induced by CRISPR–Cas9 nucleases in human cells. *Nat. Biotechnol*. **31**, 822–826

Fudenberg, G., et al., (2016). Formation of chromosomal domains by loop extrusion. *Cell Rep.* **15**, 2038–2049

Fugger, L., et al., (2012). Genomewide Association Studies and Common Disease — Realizing Clinical Utility. *N Engl J Med* **367**, 2370-2371

Fujiwara, Y. et al., (1996). Arrested development of embryonic red cell precursors in mouse embryos lacking transcription factor GATA-1. *PNAS* **93**, 12355-12358

Fujiwara, T. et al., (2009). Discovering Hematopoietic Mechanisms through Genome-wide Analysis of GATA Factor Chromatin Occupancy. *Molecular Cell* **36**, 667- 681

Furano, A., et al., (2004). L1 (LINE-1) retrotransposon diversity differs dramatically between mammals and fish. *Trends Genet*. **20**, 9-14

Fusaki, N., et al., (2009). Efficient induction of transgene-free human pluripotent stem cells using a vector based on Sendai virus, an RNA virus that does not integrate into the host genome. *Proc. Jpn Acad. Ser. B Phys. Biol. Sci.* **85**, 348–362

Gao, Z., et al., (2010). FOG-1-mediated recruitment of NuRD is required for cell lineage re-inforcement during hematopoiesis. *EMBO Journal* **29**, 457–468

Garneau, J., et al., (2010). The CRISPR/Cas bacterial immune system cleaves bacteriophage and plasmid DNA. *Nature* **468**, 67-71

Garrick, D., et al., (2008). The role of the polycomb complex in silencing alpha-globin gene expression in nonerythroid cells. *Blood* **112**, 3889-3899

Gasiunas G, et al., (2012). Cas9-crRNA ribonucleoprotein complex mediates specific DNA cleavage for adaptive immunity in bacteria. *PNAS* **109**, E2579–2586

Gemayel R, et al., (2010). Variable tandem repeats accelerate evolution of coding and regulatory sequences. *Annu Rev Genet* **44**, 445–477

Giarratana, M., et al., (2005). Ex vivo generation of fully mature human red blood cells from hematopoietic stem cells. *Nature biotechnology* **23**, 69-74

Gilbert, L., et al., (2014). Genome-scale CRISPR-mediated control of gene repression and activation. *Cell* **159**, 647–661

Gillemans, N., et al., (1998). Altered DNA-binding specificity mutants of EKLF and Sp1 show that EKLF is an activator of the beta-globin locus control region in vivo. *Genes Dev*. **12**, 2863-2873

Goh, S., et al., (2005). A newly discovered human α-globin gene. *Blood* **106**, 1466-1472

Goke, J., and Hui Ng, H., (2016). CTRL+INSERT: retrotransposons and their contribution to regulation and innovation of the transcriptome. *EMBO reports* **17**, 1131-1144

Goldman, M., et al. (1984). Replication timing of genes and middle repetitive sequences. *Science* **224**, 686-692

González, F., et al., (2014). An iCRISPR platform for rapid, multiplexable and inducible genome editing in human pluripotent stem cells. *Cell Stem Cell* **15**, 215–226

Goodbourn, S. et al., (1983). Molecular basis of length polymorphism in the human C-globin gene complex. *PNAS* **80**, 5022-5026

Goodier, J., and Kazazian, H., (2008). Retrotransposons revisited: the restraint and rehabilitation of parasites. *Cell* **135**, 23-35

Gregory, R., et al., (1996). Functional interaction of GATA1 with erythroid Krüppel-like factor and SP1 at defined erythroid promoters. *Blood* **87**, 1793–1801

Grosveld, F., et al., (1987). Position-independent, high-level expression of the human beta-globin gene in transgenic mice. *Cell* **51**, 975-985

Guenther, M., et al., (2007). A chromatin landmark and transcription initiation at most promoters in human cells. *Cell* **130**, 77–88

Gurdon, J., (1962) The developmental capacity of nuclei taken from intestinal epithelium cells of feeding tadpoles. *Journal of embryology and experimental morphology* **10**, 622-640

Hale, C., et al., (2009). RNA-guided RNA cleavage by a CRISPR RNA-Cas protein complex. *Cell* **139**, 945–956

Hamlin, R., and Altschuld, R., (2011). Extrapolation from mouse to man. *Circ. Cardiovasc. Imaging* **4**, 2–4

Hanna, J., et al., (2008). Direct reprogramming of terminally differentiated mature B lymphocytes to pluripotency. *Cell* **133**, 250-264

Hanssen, L., et al., (2017). Tissue-specific CTCF-cohesin-mediated chromatin architecture delimits enhancer interactions and function in vivo. *Nat Cell Biol*. **19**, 952-961

Hatta, M., and Cirillo, L., et al., (2007). Chromatin opening and stable perturbation of core histone:DNA contacts by FoxO1. *J Biol Chem* **282**, 35583–35593

Hatzis, P., and Talianidis, I., (2002). Dynamics of enhancer-promoter communication during differentiation-induced gene activation. *Mol. Cell* **10**, 1467–1477

Hay, D., et al., (2016). Genetic dissection of the α-globin super-enhancer in vivo. *Nature Genetics* **48**, 895- 903

Heintzman, N., et al., (2007). Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat Genet* **39**, 311–318

Heyworth, C., et al., (2002). Transcription factormediated lineage switching reveals plasticity in primary committed progenitor cells. *EMBO J* **21**, 3770–3781

Higgs, D., et al., (1990). A major positive regulatory region located far upstream of the human alpha-globin gene locus. *Genes Dev.* **4**, 1588-1601

Higgs, D., et al., (2005). Understanding alpha-globin gene regulation: Aiming to improve the management of thalassemia. *Annals of the New York Academy of Sciences* **1054**, 92-102

Higgs, D., et al., (2007). Using Genomics to Study How Chromatin Influences Gene Expression*. Annual Rev. Genom. Human Genet*. **8**, 299–325

Higgs, D., et al., (2008). Long-range regulation of alpha-globin gene expression. *Adv. Genet.* **61**, 143–173

Higgs, D., et al., (2012). Thalassaemia. *Lancet* **379**, 373-383

Higgs, D., and Wood, W., (2008). Long-range regulation of alpha globin gene expression during erythropoiesis. *Current opinion in hematology* **15**, 176-183

Higgs, D., and Wood, W., (2011). Erythropoiesis. In Postgraduate Haematology (Edited by A. Hoffbrand, D. Catovsky, E. Tuddenham, and A. Green), 12-25. NJ: Wiley-Blackwell

Higgs, D., (2009). Section three 13. Molecular basis of α-thalassemia and 14 The pathophysiology and clinical features of α-thalassemia. *Disorders of Hemoglobin:*

Genetics, Pathophysiology, and Clinical Management **Second Edition** (Edited by M. H. Steinberg, B. G. Forget, D. R. Higgs and D. J. Weatherall), 241–295. Cambridge University Press, Cambridge.

Hirschi, K.K., 2012. Hemogenic endothelium during development and beyond. *Blood* **119**, 4823—4827

Hnisz, D. et al., (2013). Super-enhancers in the control of cell identity and disease. *Cell* **155**, 934–947

Hochedlinger, K., and Jaenisch, R., (2002). Monoclonal mice generated by nuclear transfer from mature B and T donor cells. *Nature* **415**, 1035-1038

Hockemeyer, D. and Jaenisch, R. (2016). Induced pluripotent stem cells meet genome editing. *Cell Stem Cell* **18**, 573–586

Hodge, D. et al., (2006). A global role for EKLF in definitive and primitive erythropoiesis. *Blood* **107**, 3359-3371

Hodgkins, A., et al., (2015). WGE: A CRISPR database for genome engineering. *Bioinformatics* **31**, 3078-3080

Hong, W., et al., (2005). FOG-1 recruits the NuRD repressor complex to mediate transcriptional repression by GATA-1. *EMBO Journal* **24**, 67–78

Hotta, A. and Yamanaka, S., (2015). From genomics to gene therapy: induced pluripotent stem cells meet genome editing. *Annu. Rev. Genet*. **49**, 47–70

Hou, P., et al., (2013). Pluripotent stem cells induced from mouse somatic cells by small-molecule compounds. *Science* **341**, 651–654

Hough, S., et al., (2014). Single-cell gene expression profiles define self-renewing, pluripotent, and lineage primed states of human pluripotent stem cells. *Stem Cell Reports* **2**, 881–895

Howe, S., et al., (2008). Insertional mutagenesis combined with acquired somatic mutations causes leukemogenesis following gene therapy of SCID-X1 patients*. J Clin Invest*. **118**, 3143

Hsu, P., et al., (2013). DNA targeting specificity of RNA-guided Cas9 nucleases. *Nat. Biotechnol*. **31**, 827–832

Hsu, P., et al., (2014). Development and applications of CRISPR-Cas9 for genome engineering. *Cell* **157**, 1262–78

Hu, G., et al., (2011). Regulation of nucleosome landscape and transcription factor targeting at tissue-specific enhancers by BRG1. *Genome Res*. **21**, 1650–1658

Huang, P., et al., (2011). Induction of functional hepatocyte-like cells from mouse fibroblasts by defined factors. *Nature* **475**, 386–389

Huang, X. et al., (2014). *Extensive* ex vivo expansion of functional human erythroid precursors established from umbilical cord blood cells by defined factors. *Mol. Ther.* **22,** 451–463

Huber, K., et al., (2004). Haemangioblast commitment is initiated in the primitive streak of the mouse embryo. *Nature* **432**, 625–630

Hughes, J., et al. (2005). Annotation of cis-regulatory elements by identification, subclassification, and functional assessment of multispecies conserved sequences. *PNAS* **102**, 9830-9835

Hughes, J., et al., (2013). High-resolution analysis of cis-acting regulatory networks at the alpha-globin locus. *Philosophical transactions of the Royal Society of London Series B, Biological sciences* **368**, 20120361

Hughes, J., et al., (2014). Analysis of hundreds of cis-regulatory landscapes at high resolution in a single, high-throughput experiment. *Nature Genetics* **46**, 205-212

Hussein, S., et al., (2011). Copy number variation and selection during reprogramming to pluripotency. *Nature* **471**, 58–62

Hwang, W., et al., (2013). Heritable and precise zebrafish genome editing using a CRISPR-Cas system. *PLOS ONE* **8**, e68708

Inoue, H., et al., (2014). iPS cells: a game changer for future medicine. *EMBO Journal* **33**, 409–417

Inoue, M., et al., (2003). Non-transmissible virus-like particle formation by F-deficient Sendai virus is temperature sensitive and reduced by mutations in M and HN proteins. *J. Virol.* **77**, 3238–3246

Ishino, Y., et al., (1987). Nucleotide sequence of the iap gene, responsible for alkaline phosphatase isozyme conversion in Escherichia coli, and identification of the gene product. *J. Bacteriol.* **169**, 5429–5433

Itsara, A. et al., (2009). Population analysis of large copy number variants and hotspots of human genetic disease. *Am. J. Hum. Genet.* **84**, 148–161

Iwafuchi-Doi, M. and Zaret, K., (2014). Pioneer transcription factors in cell reprogramming. *Genes and Development* **28**, 2679-2692

Jia, F. et al., (2010). A nonviral minicircle vector for deriving human iPS cells. *Nat. Methods* **7**, 197–199

Jinek, M., et al., (2012). A programmable dualRNA-guided DNA endonuclease in adaptive bacterial immunity. *Science* **337**, 816–821

Jing, H., et al., (2008). Exchange of GATA factors mediates transitions in looped chromatin organization at a developmentally regulated gene locus. *Molecular Cell* **29**, 232–242

Kadam, S., et al., (2000). Functional selectivity of recombinant mammalian SWI/SNF subunits. *Genes and Development* **14**, 2441–2451

Kadonaga, J., (2012). Perspectives on the RNA polymerase II core promoter. *WIREs Dev. Biol.* **1**, 40–51

Kaji, K. et al., (2009). Virus-free induction of pluripotency and subsequent excision of reprogramming factors. *Nature* **458**, 771–775

Kang, L., et al., (2009). iPS cells can support full-term development of tetraploid blastocyst-complemented embryos. *Cell Stem Cell* **5**, 135–138

Kang, Y., et al., (2015). KLF1 stabilizes GATA-1 and TALI occupancy in the human Pglobin locus. *BBA - Gene Regulatory Mechanisms* **1849**, 282-289

Kapusta, A., et al., (2013). Transposable elements are major contributors to the origin, diversification, and regulation of vertebrate long noncoding RNAs. *PLoS Genet* **9**, e1003470

Kassouf, M., et al., (2010). Genome-wide identification of TAL1's functional targets: insights into its mechanisms of action in primary erythroid cells. *Genome Res*. **20**, 1064-1083

Kaufman, D., et al. (2001). Hematopoietic colony-forming cells derived from human embryonic stem cells. *PNAS* **98**, 10716–10721

Kearney, H., et al., (2011). Diagnostic implications of excessive homozygosity detected by SNP-based microarrays: consanguinity, uniparental disomy, and recessive single-gene mutations. *Clin. Lab. Med.* **31**, 595–613

Kelley, D., and Rinn, J., (2012) Transposable elements reveal a stem cell specific class of long noncoding RNAs. *Genome Biol* **13**, R107

Kennedy, M., et al, (2007). Development of the hemangioblast defines the onset of hematopoiesis in human es cell differentiation cultures. *Blood* **109**, 2679–2687

Kessel, K., et al., (2017). Emergence of CD43-Expressing Hematopoietic Progenitors from Human Induced Pluripotent Stem Cells. *Transfus Med Hemother.* **44**, 43–150

Khare, S., et al. (2012). HIstome--a relational knowledgebase of human histone proteins and histone modifying enzymes. *Nucleic acids research* **40**, 337-342

Kim, D., et al., (2009c). Generation of human induced pluripotent stem cells by direct delivery of reprogramming proteins. *Cell Stem Cell* **4**, 472–476

Kim, H., and Kim, J., (2014). A guide to genome engineering with programmable nucleases. *Nature reviews Genetics* **15**, 321-334

Kim, J., et al., (2008). Pluripotent stem cells induced from adult neural stem cells by reprogramming with two factors. *Nature* **454**, 646-650

Kim, K., et al., (2011). Donor cell type can influence the epigenome and differentiation potential of human induced pluripotent stem cells. *Nat. Biotechnol.* **29**, 1117–1119

Kim, S. et al., (2007). Dissecting molecular steps in chromatin domain activation during hematopoietic differentiation. *Mol. Cell. Biol.* **27**,4551–4565

Kim, S., et al., (2009a). BRG1 requirement for long-range interaction of a locus control region with a downstream promoter. *PNAS* **106**, 2259–2264

Kim, S., et al., (2009b) BRG1 directly regulates nucleosome structure and chromatin looping at the alpha globin locus to activate transcription. *Nucleic Acids Res*. **37**, 6019–6027

Kim, T., et al., (2005). A high-resolution map of active promoters in the human genome. *Nature* **436**, 876–880

Kissa, K., and Herbomel, P., (2010). Blood stem cells emerge from aortic endothelium by a novel type of cell transition. *Nature* **464**, 112–115

Knezetic, J., and Luse, D. (1986). The presence of nucleosomes on a DNA template prevents initiation by RNA polymerase II in vitro. *Cell* **45**, 95–104

Knight, J., (2014). Approaches for establishing the function of regulatory genetic variants involved in disease. *Genome Medicine* **6**, 92

Kobari, L., et al., (2012). Human induced pluripotent stem cells can reach complete terminal maturation: In vivo and in vitro evidence in the erythropoietic differentiation model. *Haematologica* **97**, 1795–1803

Koche, R., et al., (2011). Reprogramming factor expression initiates widespread targeted chromatin remodeling. *Cell Stem Cell* **8**, 96–105

Koonin, E., and Krupovic, M., (2015). Evolution of adaptive immunity from transposable elements combined with innate immune systems. *Nat. Rev. Genet*. **16**, 184–192

Kornberg, R. and Lorch, Y. (1999). Twenty-five years of the nucleosome, fundamental particle of the eukaryote chromosome. *Cell* **98**, 285-294

Koyanagi-Aoi, M., et al., (2013). Differentiation-defective phenotypes revealed by large-scale analyses of human pluripotent stem cells. *PNAS* **110**, 20569–20574

Krivega, I., et al., (2014). Role of LDB1 in the transition from chromatin looping to transcription activation. *Genes and Development* **28**, 1278-1290

Kulessa, H., et al., (1995). GATA-1 reprograms avian myelomonocytic cell lines into eosinophils, thromboblasts, and erythroblasts. *Genes and Development* **9**, 1250–1262

Kunarso G, et al., (2010). Transposable elements have rewired the core regulatory network of human embryonic stem cells. *Nat Genet* **42**, 631 – 634

Kurita, R. et al., (2013). Establishment of immortalized human erythroid progenitor cell lines able to produce enucleated red blood cells. *PLoS ONE* **8**, e59890

Kwiatkowski, N., et al., (2014). Targeting transcription regulation in cancer with a covalent CDK7 inhibitor. *Nature* **511**, 616-620

Lancrin, C., et al., (2009). The haemangioblast generates haematopoietic cells through a haemogenic endothelium stage. *Nature* **457**, 892–895

Lander, E., et al., (2001). Initial sequencing and analysis of the human genome. Nature 409, 860-921

Lapillonne, L., et al., (2010). Red blood cell generation from human induced pluripotent stem cells: perspectives for transfusion medicine. *Haematologica* **95**, 1651-1659

Latchman, D., (1997). Transcription factors: an overview. *The international journal of biochemistry & cell biology* **29**, 1305-1312

Laurent, L., et al., (2011). Dynamic changes in the copy number of pluripotency and cell proliferation genes in human ESCs and iPSCs during reprogramming and time in culture. *Cell Stem Cell* **8**, 106–118

Leder, A., et al., (1997). Mouse z- and a-Globin Genes: Embryonic Survival, a-Thalassemia, and Genetic Background Effects. *Blood* **90**, 1275 -1282

Lee, C., et al., (1999). Targeting a SWI/SNF-related chromatin remodelling complex to the beta-globin promoter in erythroid cells. *PNAS* **96**, 12311-12315

Lee, D., et al., (2005) Functional characterization of core promoter elements: the downstream core element is recognized by TAF1. *Mol Cell Biol* **25**, 9674–9686

Lee, T. and Young, R., (2000). Transcription of eukaryotic protein-coding genes. *Annual review of genetics* **34**, 77-137

Lelli, K., et al., (2012). Disentangling the many layers of eukaryotic transcriptional regulation. *Annual Rev. Genet*. **46**, 43–68

Lengerke, C., et al., (2008). BMP and Wnt specify hematopoietic fate by activation of the Cdx-Hox pathway. *Cell Stem Cell* **2**, 72–82

Lenhard, B., et al. (2012). Metazoan promoters: emerging characteristics and insights into transcriptional regulation. *Nature reviews. Genetics* **13**, 233-245

Letting, D., et al., 2003. Formation of a Tissue-Specific Histone Acetylation Pattern by the Hematopoietic Transcription Factor GATA-1. *Molecular and Cellular Biology* **23**, 1334-1340

Levine, M. and Tjian, R., (2003). Transcription regulation and animal diversity. *Nature* **424**, 147-151

Levine, M. (2010). Transcriptional Enhancers in Animal Development and Evolution. *Current Biology* **20**, 754–763

Ley, T., et al., (1989). Globingene expression in erythroid human fetal liver cells. *Journal of Clinical Investigation* **83**, 1032–1038

Li, B., et al. (2007). The role of chromatin during transcription. *Cell* **128**, 707-719

Li, C., et al., (2016). Novel HDAd/EBV reprogramming vector and highly efficient Ad/CRISPR–Cas sickle cell disease gene correction. *Sci. Rep.* **6**, 30422

Li, H., et al., (2000). A cytoplasmic RNA vector derived from non-transmissible Sendai virus with efficient gene transfer and expression. *J. Virol.* **74**, 6564–6569

Li, H., et al., (2015a). Efficient genomic correction methods in human iPS cells using CRISPR-Cas9 system. *Methods* **pii**, S1046–2023

Li, H., et al. (2015b). Precise correction of the dystrophin gene in Duchenne muscular dystrophy patient induced pluripotent stem cells by TALEN and CRISPRCas9. *Stem Cell Rep.* **4**, 143–54

Li, K., et al., (2014). Optimization of genome engineering approaches with the CRISPR/Cas9 system. *PLoS One* **9**, e105779

Li, L., et al., (2013). Ldb1-nucleated transcription complexes function as primary mediators of global erythroid gene activation. *Blood* **121**, 4575—4585

Li, Q., et al., (2002). Locus control regions. *Blood* **100**, 3077-3086

Lieber, M., (2010). The mechanism of double-strand DNA break repair by the nonhomologous DNA end-joining pathway. *Annu. Rev. Biochem*. **79**, 181–211

Lim, C., et al., (2004). The MTE, a new core promoter element for transcription by RNA polymerase II. *Genes Dev* **18**, 1606–1617

Lin, Y., et al., (2014). CRISPR/Cas9 systems have off-target activity with insertions or deletions between target DNA and guide RNA sequences. *Nucleic Acids Research* **42**, 7473–7485

Loh, Y., et al. (2009). Generation of induced pluripotent stem cells from human blood. *Blood* **113**, 5476–5479

Loken, M., et al., (1987). Flow Cytometric Analysis of Human Bone Marrow: I. Normal Erythroid Development. *Blood* **69**, 255-263

Lorch, Y., et al. (1987). Nucleosomes inhibit the initiation of transcription but allow chain elongation with the displacement of histones. *Cell* **49**, 203–210

Love, P., et al., (2014). Ldbl complexes: the new master regulators of erythroid gene transcription. *Trends in Genetics* **30**, 1-9

Lovén, J. et al., (2013). Selective inhibition of tumor oncogenes by disruption of superenhancers. *Cell* **153**, 320–334

Lovsin, N., et al., (2001). Evolutionary dynamics in a novel L2 clade of non-LTR retrotransposons in Deuterostomia. *Mol Biol Evol.* **18**, 2213-2224

Lowry, W. et al., (2008). Generation of human induced pluripotent stem cells from dermal fibroblasts. *PNAS* **105**, 2883-2888

Lozzio, C., and Lozzio, B., (1975). Human chronic myelogenous leukemia cell-line with positive Philadelphia chromosome. *Blood* **45**, 321–34

Lu, S., et al., (2008). Biologic properties and enucleation of red blood cells from human embryonic stem cells. *Blood* **112**, 4475–4484

Luck, L., et al., (2004). Human CD34(+) and CD34(+) CD38(-) hematopoietic progenitors in sickle cell disease differ phenotypically and functionally from normal and suggest distinct subpopulations that generate F cells. *Exp Hematol*. **32**, 483-493

Luckett, W., (1978). Origin and differentiation of the yolk sac and extraembryonic mesoderm in presomite human and rhesus monkey embryos. *Am. J. Anat.* **152**, 59–97

Lund, R., et al., (2012). Genetic and epigenetic stability of human pluripotent stem cells. *Nat. Rev. Genet.* **13**, 732–744

Lynch, M., et al., (2012). An interspecies analysis reveals a key role for unmethylated CpG dinucleotides in vertebrate Polycomb complex recruitment. *EMBO Journal* **31**, 317–329

Lynch, V., et al., (2011). Transposon-mediated rewiring of gene regulatory networks contributed to the evolution of pregnancy in mammals. *Nat Genet* **43**, 1154–1159

Lynch V., et al., (2015). Ancient transposable elements transformed the uterine regulatory landscape and transcriptome during the evolution of mammalian pregnancy. *Cell Rep* **10**, 551–561

Ma, H., et al., (2014). Pol III promoters to express small RNAs: delineation of transcription initiation. *Mol. Ther. Nucleic Acids* **3**, e161

MacDonald, J., et al., (2014). The Database of Genomic Variants: a curated collection of structural variation in the human genome. *Nucleic Acids Res.* **42**, 986–992

Maherali, N., et al., (2007). Directly reprogrammed fibroblasts show global epigenetic reprogramming and widespread tissue contribution. *Cell Stem Cell* **1**, 55-70

Maherali, N., et al., (2008). A high-efficiency system for the generation and study of human induced pluripotent stem cells. *Cell Stem Cell* **3**, 340-345

Makishima, H., and Maciejewski, J., (2011). Pathogenesis and consequences of uniparental disomy in cancer. *Clin. Cancer Res.* **17**, 3913–3923

Mali, P., et al., (2013). CAS9 transcriptional activators for target specificity screening and paired nickases for cooperative genome engineering. *Nat. Biotechnol.* **31**, 833–838

Malik, S. and Roeder, R., (2010). The metazoan Mediator co-activator complex as an integrative hub for transcriptional regulation. *Nature Review Genetics* **11**, 761–772

Mansour, M., et al., (2014). An oncogenic super-enhancer formed through somatic mutation of a noncoding intergenic element. *Science* **346**, 1373–1377

Marraffini, L., and Sontheimer, E., (2008). CRISPR interference limits horizontal gene transfer in staphylococci by targeting DNA. *Science* **322** 1843–1845

Martin, G., (1981). Isolation of a pluripotent cell line from early mouse embryos cultured in medium conditioned by teratocarcinoma stem cells. *PNAS* **78**, 7634-7638

Mateizel, I., et al., (2006). Derivation of human embryonic stem cell lines from embryos obtained after IVF and after PGD for monogenic disorders. *Hum. Reprod*. **21**, 503–511

Mathelier, A., et al., (2015). Identification of altered cis-regulatory elements in human disease. *Trends in Genetics* **31**, 67-76

Mathias, L., et al., (2000). Ineffective erythropoiesis in beta-thalassemia major is due to apoptosis at the polychromatophilic normoblast stage. *Exp Hematol.* **28**, 1343–1353

Mas, C., et al., (2011). Structural and functional characterization of an atypical activation domain in erythroid Kruppel-like factor (EKLF). *PNAS* **108**, 10484 – 10489

Massari, M., and Murre, C., (2000). Helix-loop-helix proteins: regulators of transcription in eucaryotic organisms. *Mol. Cell Biol.* **20**, 429-440

Maston, G., et al., (2006). Transcriptional regulatory elements in the human genome. *Annual review of genomics and human genetics* **7**, 29-59

Mayshar, Y., (2010). Identification and classification of chromosomal aberrations in human induced pluripotent stem cells. *Cell Stem Cell*. **7**, 521–531

McGrath, K., and Palis, J., (2008). Ontogeny of erythropoiesis in themammalian embryo. *Current Topics in Developmental Biology* **82**, 1–22

McGrath, K., et al. (2011). A transient definitive erythroid lineage with unique regulation of the beta-globin locus in the mammalian embryo. *Blood* **117**, 4600–4608

McVey, M., and Lee, S., (2008). MMEJ repair of double-strand breaks (director's cut): deleted sequences and alternative endings. *Trends in Genetics* **24**, 529–538

Meissner, A., et al., (2007). Direct reprogramming of genetically unmodified fibroblasts into pluripotent stem cells. *Nat. Biotechnol*. **25**, 1177-1181

Melnikov A, et al., (2012). Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. *Nat Biotechnol.* **30**, 271-277

Merika, M. and Orkin, S., (1995). Functional synergy and physical interactions of the erythroid transcription factor GATA-1 with the Kruppel family proteins Sp1 and EKLF. *Mol. Cell. Biol.* **15**, 2437–2447

Mettananda, S., et al., (2017). Editing an α-globin enhancer in primary human hematopoietic stem cells as a treatment for β-thalassemia. *Nature Communications* **8**, e424

Merryweather-Clarke, A., et al., (2011). Global gene expression analysis of human erythroid progenitors. *Blood* **117**, e96-e108

Miccio, A., et al., (2010). NuRD mediates activating and repressive functions of GATA-1 and FOG-1 during blood development. *EMBO Journal* **29**, 442–456

Migliaccio, G., et al., (1986). Human embryonic hemopoiesis. Kinetics of progenitors and precursors underlying the yolk sac—liver transition. *J. Clin. Invest.* **78**, 51–60

Mikkelsen, T., et al., (2007) Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature* **448**, 553–560

Mikkelsen, T., et al., (2008). Dissecting direct reprogramming through integrative genomic analysis. *Nature* **454**, 49-55

Mikkola, H.K. a et al., (2003). Haematopoietic stem cells retain long-term repopulating activity and multipotency in the absence of stem-cell leukaemia SCL/tal-1 gene. *Nature* **421**, 547-551

Miller, I., and Bieker, J., (1993). A Novel, Erythroid Cell-Specific Murine Transcription Factor That Binds to the CACCC Element and Is Related to the Kriippel Family of Nuclear Proteins. *Molecular and Cell Biology* **13**, 2776-2786

Miller, R., and Ruddle, F., (1976). Pluripotent teratocarcinoma-thymus somatic cell hybrids. *Cell* **9**, 45-55

Mitzelfelt, K., et al., (2017). Efficient Precision Genome Editing in iPSCs via Genetic Co-targeting with Selection. *Stem Cell Reports* **8**, 491–499

Miura, K., et al., (2009). Variation in the safety of induced pluripotent stem cell lines. *Nat. Biotechnol*. **27**, 743–745

Mojica, F., et al., (2005). Intervening sequences of regularly spaced prokaryotic repeats derive from foreign genetic elements. *J. Mol. Evol*. **60**, 174–182

Moreau, P., et al., (1981). The SV40 72 base repair repeat has a striking effect on gene expression both in SV40 and other chimeric recombinants. *Nucleic Acids Res*. **9**, 6047–6068

Mouse Genome Sequencing Consortium et al. (2002). Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**, 520–562

Mulin, J., et al., (2014). Exploring the function of genetic variants in the non-coding genomic regions: approaches for identifying human regulatory variants affecting gene expression
Briefings in Bioinformatics **16**, 393-412

Muller, F., et al., (2007). New problems in RNA polymerase II transcription initiation: matching the diversity of core promoters with a variety of promoter recognition factors. *J Biol Chem* **282**, 14685–14689

Murai, K., et al., (2016). The TLX-miR‐219 cascade regulates neural stem cell proliferation in neurodevelopment and schizophrenia iPSC model. *Nat. Commun*. **7**, 10965

Nagano, K., et al., (2008) Cell surface biomarkers of embryonic stem cells. *Proteomics* **8**, 4025–4035

Nakanishi, M., and Otsu, M., (2012). Development of Sendai Virus Vectors and their Potential Applications in Gene Therapy and Regenerative Medicine. *Curr Gene Ther.* **12**, 410–416

Narsinh, K., et al., (2011). Single cell transcriptional profiling reveals heterogeneity of human induced pluripotent stem cells. *J. Clin. Invest.* **121**, 1217–1221

Nasmyth, K., (2001) Disseminating the genome: joining, resolving, and separating sister chromatids during mitosis and meiosis. *Annu Rev Genet* **35**, 673–745

Neph, S., et al., (2012). An expansive human regulatory lexicon encoded in transcription factor footprints. *Nature* **489**, 83–90

Nichols, J., and Smith, A., (2009). Naive and primed pluripotent states. *Cell Stem Cell* **4**, 487–492

Nielsen, S., et al., (2013). Mechanism of eukaryotic RNA polymerase III transcription termination. *Science* **340**, 1577–1580

Nishimura, K. et al., (2010). Development of defective and persistent Sendai virus vector: a unique gene delivery/ expression system ideal for cell reprogramming. *J. Biol. Chem.* **286**, 4760–4771

O'Geen, H., et al., (2015). How specific is CRISPR/Cas9 really? *Curr. Opin. Chem. Biol.* **29**, 72–78

Ohi, Y., et al., (2011). Incomplete DNA methylation underlies a transcriptional memory of somatic cells in human iPS cells. *Nat. Cell Biol.* **13**, 541–549

Ohler, U., et al., (2002). Computational analysis of core promoters in the Drosophila genome. *Genome Biol* **3**, RESEARCH0087

Okumura, N., et al., (1992). Changes in cell surface antigen expressions during proliferation and differentiation of human erythroid progenitors. *Blood* **80**, 642-650

Okita, K., et al., (2007). Generation of germline-competent induced pluripotent stem cells. *Nature* **448**, 313-317

Okita, K., et al., (2008). Generation of mouse induced pluripotent stem cells without viral vectors. *Science* **322**, 949–953

Okita, K. et al., (2011). A more efficient method to generate integration-free human iPS cells. *Nat. Methods* **8**, 409–412

Olivier, E., et al., (2006). Large-scale production of embryonic red blood cells from human embryonic stem cells. *Exp Hematol*. **34**, 1635–1642

Olivier, E., et al., (2012). Novel, high-yield red blood cell production methods from CD34-positive cells derived from human embryonic stem, yolk sac, fetal liver, cord blood, and peripheral blood. *Stem Cells Translational Medicine* **1**, 604–614

Olivier, E., et al., (2016). High efficiency serum free feeder free erythroid differentiation of human pluripotent stem cells using small molecules. *Stem Cells Translational Medicine* **5**, 1394-1405

Ong, C., and Corces, V., (2011). Enhancer function: new insights into the regulation of tissue-specific gene expression. *Nature Reviews Genetics* **12**, 283-293

Ong, C., and Corces, V., (2014). CTCF: an architectural protein bridging genome topology and function. *Nature reviews. Genetics* **15**, 234-246

Org, T., et al., (2015). Scl binds to primed enhancers in mesoderm to regulate hematopoietic and cardiac fate divergence. *EMBO Journal* **34**, 759-777

Osafune, K., et al., (2008). Marked differences in differentiation propensity among human embryonic stem cell lines. *Nat. Biotechnol*. **26**, 313–315

Ott, M., et al., (2006). Correction of X-linked chronic granulomatous disease by gene therapy, augmented by insertional activation of MDS1-EVI1, PRDM16 Or SETBP1. *Nat Med*. **12**, 401–409

Palii, C., et al., (2011). Differential genomic targeting of the transcription factor TAL1 in alternate haematopoietic lineages. *EMBO Journal* **30**, 494-509

Palis, J., (2008). Ontogeny of erythropoiesis. *Current opinion in hematology* **15**, 155-161

Palis, J., (2014). Primitive and definitive erythropoiesis in mammals. *Frontiers in physiology*, **5**, 3

Palis, J., et al., (1999). Development of erythroid and myeloid progenitors in the yolk sac and embryo proper of the mouse. *Development* **126**, 5073–5084

Papayannopoulou, T., et al., (1980). Fetal Hb production during acute erythroid expansion. I. Observations in patients with transient erythroblastopenia and post-phlebotomy. *British Journal of Haematology* **44**, 535–546

Paquet, D., et al., (2016). Efficient introduction of specific homozygous and heterozygous mutations using CRISPR/Cas9. *Nature* **533**, 125-129

Park, C., et al., (2015). Functional correction of large factor VIII gene chromosomal inversions in hemophilia A patientderived iPSCs using CRISPR–Cas9. *Cell Stem Cell* **17**, 213–220

Park, I. (2008). Reprogramming of human somatic cells to pluripotency with defined factors. *Nature* **451**, 141-146

Patwardhan, R., et al., (2012). Massively parallel functional dissection of mammalian enhancers in vivo. *Nat Biotechnol.* **30**, 265-270

Paulson, R., et al., (2011). Stress erythropoiesis: new signals and new stress progenitor cells. *Curr Opin Hematol*. **18**, 139–145

Pearson, S., et al., (2008). The stepwise specification of embryonic stem cells to hematopoietic fate is driven by sequential exposure to Bmp4, activin A, bFGF and VEGF. *Development* **135**, 1525–1535.

Perkins, A., et al., (1995). Lethal beta-thalassaemia in mice lacking the erythroid CACCC-transcription factor EKLF. *Nature* **375**, 318-322

Perkins, A., et al., (1996). Silencing of human fetal globin expression is impaired in the absence of the adult beta-globin gene activator protein EKLF. *PNAS* **93**, 12267-12271

Perlingeiro, R., et al., (2003). A role for thrombopoietin in hemangioblast development. *Stem Cells* **21**, 272–280

Perry, J., et al., (2007). BMP4, SCF, and hypoxia cooperatively regulate the expansion of murine stress erythroid progenitors. *Blood* **109**, 4494-4502

Peschle, C., et al. (1985). Haemoglobin switching in human embryos: asynchrony of the ζ -> α and ε -> γ-globin switches in primitive and definitve erythropoietic lineage. *Nature* **313**, 235–238

Pevny, L., et al., (1991). Erythroid differentiation in chimeric mice blocked by a targeted mutation in the gene for transcription factor GATA-1. *Nature* **349**, 257-260

Phillips-Cremins, J., et al., (2013). Architectural protein subclasses shape 3D organization of genomes during lineage commitment. *Cell* **153**, 1281-1295

Phylipsen, M., et al., (2010). Thalassemia in Western Australia: 11 novel deletions characterized by Multiplex Ligation-dependent Probe Amplification. *Blood Cells Mol. Dis.* **44**, 146–151

Pick, M., et al., (2007). Differentiation of human embryonic stem cells in serum-free medium reveals distinct roles for bone morphogenetic protein 4, vascular endothelial growth factor, stem cell factor, and fibroblast growth factor 2 in haematopoiesis. *STEM CELLS* **25**, 2206–2214

Pimkin, M., et al., (2014). Divergent functions of hematopoietic transcription factors in lineage priming and differentiation during erythro-megakaryopoiesis. *Genome Res.* **24**, 1932-1944

Plessis, A., et al., (1992). Site-specific recombination determined by I-SceI, a mitochondrial group I intron-encoded endonuclease expressed in the yeast nucleus. *Genetics* **130**, 451–460

Plessy, C. et al., (2010). Linking promoters to functional transcripts in small samples with nanoCAGE and CAGEscan. *Nature Methods* **7**, 528–534

Poodt, J., et al., (2006). A newly identified deletion of 970 bp at the alpha-globin locus that removes the promoter region of the alpha1 gene. *Hemoglobin* **30**, 471–477

Porcher, C. et al., (1996). The T cell leukemia oncoprotein SCL/tal-1 is essential for development of all hematopoietic lineages. *Cell* **86**, 47-57

Porteus, M., (2016). Genome editing: a new approach to human therapeutics. *Annu. Rev. Pharmacol. Toxicol.* **56**, 163–190

Pourcel, C., et al., (2005). CRISPR elements in Yersinia pestis acquire new repeats by preferential uptake of bacteriophage DNA, and provide additional tools for evolutionary studies. *Microbiology* **151**, 653–663

Ptashne, M. (1986). Gene regulation by proteins acting nearby and at a distance. *Nature* **322**, 697–701

Putney, S., et al., (1981). A DNA fragment with an alpha-phosphorothioate nucleotide at one end is asymmetrically blocked from digestion by exonuclease III and can be replicated in vivo. *PNAS* **78**, 7350-7354

Qiu, C, et al., (2008) Globin switches in yolk sac-like primitive and fetal-like definitive red blood cells produced from human embryonic stem cells. *Blood* **111**,2400–2408

Quilez, J., et al., (2016). Polymorphic tandem repeats within gene promoters act as modifiers of gene expression and DNA methylation in humans. Nucleic Acids Research 44, 3750–3762

Ran, F., et al., (2013). Genome engineering using the CRISPR-Cas9 system. *Nature protocols* **8**, 2281-2308

Rath, D., et al., (2015). The CRISPR-Cas immune system: biology, mechanisms and applications. *Biochimie* **117**, 119–28

Redon, R. et al., (2006). Global variation in copy number in the human genome. *Nature* **444**, 444–454

Renaud, JB., et al., (2016). Improved genome editing efficiency and flexibility using modified oligonucleotides with TALEN and CRISPR-Cas9 nucleases. *Cell Rep*. **14**, 2263-

Richardson C., et al., (2016). Enhancing homology-directed genome editing by catalytically active and inactive CRISPR-Cas9 using asymmetric donor DNA. *Nature Biotechnology* **34**, 339–344

Rodriguez, P., et al., (2005). GATA-1 forms distinct activating and repressive complexes in erythroid cells. *The EMBO journal* **24**, 2354-2366

Rouet, P., et al., (1994a). Introduction of double-strand breaks into the genome of mouse cells by expression of a rare-cutting endonuclease. *Mol. Cell. Biol.* **14**, 8096–8106

Rouet, P., et al., (1994b). Expression of a site-specific endonuclease stimulates homologous recombination in mammalian cells. *PNAS* **91**, 6064–6068

Rudin, N., et al., (1989). Genetic and physical analysis of double-strand break repair and recombination in Saccharomyces cerevisiae. *Genetics* **122**, 519–534

Sawaya, S., et al., (2013). Microsatellite Tandem Repeats Are Abundant in Human Promoters and Are Associated with Regulatory Elements. *PLOS One* **8**, e54710

Saha, A., et al., (2006). Chromatin remodelling: the industrial revolution of DNA around histones. *Nature reviews. Molecular Cell Biology* **7**, 437-447

Sanborn, A., et al., (2015). Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes. *PNAS* **112**, E6456–E6465

Sanda, T., et al., (2012). Core transcriptional regulatory circuit controlled by the TAL1 complex in human T cell acute lymphoblastic leukemia. *Cancer Cell*. **22**, 209-221

Sandelin, A. et al., (2007). Mammalian RNA polymerase II core promoters: insights from genome-wide studies. *Nature Reviews Genetics* **8**, 424–436

Schaub, M., et al., (2012). Linking disease associations with regulatory information in the human genome. *Genome Research* **22**, 1748–1759

Schoenfelder, S., et al., (2010). The transcriptional interactome: gene expression in 3D. *Curr. Opin. Genet. Dev.* **20**, 127–133

Schwessinger, R., et al., (2017). Sasquatch: predicting the impact of regulatory SNPs on transcription factor binding from cell- and tissue-specific DNase footprints. *Genome Res*. **27**, 1730-1742

Seah, Y. et al., (2015). Induced pluripotency and gene editing in disease modelling: perspectives and challenges. *Int. J. Mol. Sci.* **16**, 28614–28634

Seki, T., et al., (2010). Generation of induced pluripotent stem cells from human terminally differentiated circulating T cells. *Stem Cell* **7**, 11–14

Sengupta, T., et al., (2009). Distinct modes of gene regulation by a cell-specific transcriptional activator. *PNAS* **106**, 4213–4218

Sergeant, A., et al., (1984). A transcription enhancer acts in vitro over distances of hundreds of base-pairs on both circular and linear templates but not on chromatin reconstituted DNA. *J. Mol. Biol.* **180**, 577–600

Shalem, O., et al., (2014). Genome-scale CRISPR-Cas9 knockout screening in human cells. *Science* **343**, 84–87

Sharpe, J., et al., (1992). Analysis of the human alpha globin upstream regulatory element (HS-40) in transgenic mice. *EMBO Journal* **11**, 4565-4572

Sharpe, J., et al., (1993) Analysis of the human alpha-globin gene cluster in transgenic mice. *PNAS* **90**, 11262-11266

Sharpe, J., et al., (1993b). Role of upstream DNase I hypersensitive sites in the regulation of human alpha globin gene expression. *Blood* **82**, 1666–1671

Shi, Y., et al., (2008). A combined chemical and genetic approach for the generation of induced pluripotent stem cells. *Cell Stem Cell* **2**, 525-528

Shi Y., et al., (2017). Induced pluripotent stem cell technology: a decade of progress. *Nat. Rev. Drug Discov.* **16**, 115-130

Shioda, T., et al., (1986). Determination of the complete nucleotide sequence of the Sendai virus genome RNA and the predicted amino acid sequences of the F, HN and L proteins. *Nucleic Acids Res.* **14**, 1545–1563

Shlyueva, D., et al., (2014). Transcriptional enhancers: from properties to genome-wide predictions. *Nature Reviews Genetics* **15**, 272-286

Simon, J., and Kingston, R., (2013). Occupying chromatin: Polycomb mechanisms for getting to genomic targets, stopping transcriptional traffic, and staying put. *Molecular cell* **49**, 808-824

Skarnes, W., et al., (2011). A conditional knockout resource for the genome-wide study of mouse gene function. *Nature* **474**, 337–342

Sloutskin, A., et al., (2015). ElemeNT: A Computational Tool for Detecting Core Promoter Elements. *Transcription* **6**, 41-50

Smale, S., and Kadonaga, J., (2003). The RNA polymerase II core promoter. *Annu Rev Biochem* **72**, 449–479

Smith, B., et al., (2013). The aryl hydrocarbon receptor directs hematopoietic progenitor cell expansion and differentiation. *Blood* **122**, 376-385

Smith, C., et al., (2014). Whole-genome sequencing analysis reveals high specificity of CRISPR–Cas9 and TALEN based genome editing in human iPSCs. *Cell Stem Cell* **15**, 12–13

Smith, C., et al., (2015). Efficient and allele-specific genome editing of disease loci in human iPSCs. *Mol. Ther.* **23**, 570–577

Soldner, F. et al., (2016). Parkinson-associated risk variant in distal enhancer of α-synuclein modulates target gene expression. *Nature* **533**, 95–99

Soler, E., et al., (2010). The genome-wide dynamics of the binding of Ldb1 complexes during erythroid differentiation. *Genes and development* **24**, 277-289

Song, B., et al., (2015). Improved hematopoietic differentiation efficiency of gene-corrected β-thalassemia induced pluripotent stem cells by CRISPR/Cas9 system. *Stem Cells Dev.* **24**, 1053–65

Song, L., and Crawford, G., (2010). DNase-seq: A High-Resolution Technique for Mapping Active Gene Regulatory Elements across the Genome from Mammalian Cells. *Cold Spring Harbor Protocols* **2**, e5384

Spitz, F. and Furlong, E., (2012). Transcription factors: from enhancer binding to developmental control. Nature reviews. *Genetics* **13**, 613-626

Stadtfeld, M., et al., (2008). Reprogramming of pancreatic beta cells into induced pluripotent stem cells. *Curr. Biol*. **18**, 890-894

Stadtfeld, M., et al., (2008b). Induced pluripotent stem cells generated without viral integration. *Science* **322**, 945–949

Stamatoyannopoulos, G. (2005) Control of globin gene expressionduring development and erythroid differentiation. *Experimental Hematology* **33**, 259–271

Stamatoyannopoulos, G., et al., (1985). HbF production in stressed erythropoiesis: observations and kinetic models. *Ann N Y Acad Sci*. **445**, 188–197

Strachan, T., et al. (eds) Molecular Genetics of Early Human Development (Academic Press, 1997)

Studer, L., et al., (2015). Programming and reprogramming cellular age in the era of induced pluripotency. *Cell Stem Cell* **16**, 591–600

Suzuki, K., et al., (2014). Targeted gene correction minimally impacts whole-genome mutational load in human-disease-specific induced pluripotent stem cell clones. *Cell Stem Cell* **15**, 31–36

Suzuki, Y., et al., (2001). Identification and characterization of the potential promoter regions of 1031 kinds of human genes. *Genome Res* **11**, 677–684

Symington, L., and Gautier, J., (2011). Double-strand break end resection and repair pathway choice. *Annual Review of Genetics* **45**, 247–271

Taapken, S., (2011). Karotypic abnormalities in human induced pluripotent stem cells and embryonic stem cells. *Nat. Biotechnol*. **29**, 313–314

Tada, M., et al., (1997). Embryonic germ cells induce epigenetic reprogramming of somatic nucleus in hybrid cells. *EMBO Journal* **16**, 6510-6520

Tada, M., et al., (2001). Nuclear reprogramming of somatic cells by in vitro hybridization with ES cells. *Curr. Biol*. **11**, 1553-1558

Takahashi, K., and Yamanaka, S. (2006). Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell* **126**, 663-676

Takahashi, K., et al., (2007). Induction of pluripotent stem cells from adult human fibroblasts by defined factors. *Cell* **131**, 861-872

Tallack, M., et al., (2010). A global role for KLF1 in erythropoiesis revealed by ChIP-seq in primary erythroid cells. *Genome Research* **20**, 1052-1063

Tallack, M., et al., (2012). Novel roles for KLF1 in erythropoiesis revealed by mRNAseq. *Genome research* **22**, 2385-98

Tan, M., et al., (2011). Identification of 67 histone marks and histone lysine crotonylation as a new type of histone modification. *Cell* **146**, 1016–1028

Thomas, M. and Chiang, C., (2006). The general transcription machinery and general cofactors. *Critical reviews in biochemistry and molecular biology* **41**, 105-178

Thomson, J. et al., (1998). Embryonic stem cell lines derived from human blastocysts. *Science* **282**, 1145-1147

Tokusumi, Y., et al., (2007). The new core promoter element XCPE1 (X Core Promoter Element 1) directs activator-, mediator-, and TATA-binding protein-dependent but TFIID-independent RNA polymerase II transcription from TATA-less promoters. *Mol Cell Biol* **27**, 1844–1858

Tolhuis, B., et al., (2002). Looping and interaction between hypersensitive sites in the active beta-globin locus. *Molecular cell* **10**, 1453-1465

Trakarnsanga, K., et al. (2017). An immortalized adult human erythroid line facilitates sustainable and scalable generation of functional red cells. *Nature Communications* **8**, 14750

Treisman, R., and Maniatis, T., (1985). Simian virus 40 enhancer increases number of RNA polymerase II molecules on linked DNA. *Nature* **315**, 73–75

Tripic, T., et al., (2009). SCL and associated proteins distinguish active from repressive GATA transcription factor complexes. *Blood* **113**, 2191-2201

Tsang, A., et al., (1998). Failure of megakaryopoiesis and arrested erythropoiesis in mice lacking the GATA-1 transcriptional cofactor FOG. *Genes & Development* **12**, 1176-1188

Tuan, D., et al., (1992). Transcription of the hypersensitive site HS2 enhancer in erythroid cells. *PNAS* **89**, 11219-11223

Urbach, A., and Benvenisty, N. (2009). Studying early lethality of 45, XO (Turner's syndrome) embryos using human embryonic stem cells. *PLoS ONE* **4**, e4175

Urbach, A., et al., (2004). Modeling for Lesch-Nyhan disease by gene targeting in human embryonic stem cells. *Stem Cells* **22**, 635–641

Vakoc, C., et al., (2005). Proximity among distant regulatory elements at the beta globin locus requires GATA-1 and FOG-1. *Molecular Cell* **17**, 453–462

Valen, E. and Sandelin, A., (2011). Genomic and chromatin signals underlying transcription start-site selection. *Trends Genet*. **27**, 475–485

van den Akker, E., et al., (2010). The majority of the in vitro erythroid expansion potential resides in CD34(-) cells, outweighing the contribution of CD34(+) cells and significantly increasing the erythroblast yield from peripheral blood samples. *Haematol-Hematol J* **95**, 1594-1598

van der Harst, P., et al., (2012). Seventy-five genetic loci influencing the human red blood cell. *Nature* **492**, 369–375

Veres, A., et al., (2014). Low incidence of off-target mutations in individual CRISPR–Cas9 and TALEN targeted human stem cell clones detected by whole-genome sequencing. *Cell Stem Cell* **15**, 27–30

Vernimmen, D., et al., (2007). Long-range chromosomal interactions regulate the timing of the transition between poised and active gene expression. *EMBO Journal* **26**, 2041–2051

Vernimmen, D., et al., (2009). Chromosome looping at the human alpha-globin locus is mediated via the major upstream regulatory element (HS -40). *Blood* **114**, 4253-4260

Vernimmen, D., et al. (2011). Polycomb eviction as a new distant enhancer function. *Genes and development* **25**, 1583-1588

Vernimmen, D., (2014). Uncovering Enhancer Functions Using the alpha-Globin Locus. *PLoS genetics* **10**, e1004668

Viprakasit, V., et al., (2003). De novo deletion within the telomeric region flanking the human alpha globin locus as a cause of alpha thalassaemia. *Br. J. Haematol.* **120**, 867–875

Viprakasit, V., et al., (2006). A novel deletion causing alpha thalassemia clarifies the importance of the major human alpha globin regulatory element. *Blood* **107**, 3811–3812

Visvader, J., et al., (1992). GATA-1 but not SCL induces megakaryocytic differentiation in an early myeloid line. *EMBO J* **11**, 4557–4564

Vo Ngoc L., et al., (2017). The punctilious RNA polymerase II core promoter. *Genes Dev*. **31**, 1289-1301

Vodyanik M. et al., (2005). Human embryonic stem cell-derived CD34+ cells: Efficient production in the coculture with OP9 stromal cells and analysis of lymphohematopoietic potential. *Blood* **105**, 617–626

Vodyanik M., et al., (2006). Leukosialin (CD43) defines hematopoietic progenitors in human embryonic stem cell differentiation cultures. *Blood* **108**, 2095-2105

Vyas, P., et al., (1992). Regulation of human embryonic globin genes zeta 2 and epsilon in stably transformed mouse erythroleukemia cells. *Blood* **80**, 1832-1837

Wadman, I., et al., (1997). The LIM-only protein Lmo2 is a bridging molecule assembling an erythroid, DNA-binding complex which includes the TALI, E47, GATA-1 and Ldbl/NLI proteins. *The EMBO Journal* **16**, 3145-3157

Wallace H. et al., (2007). Manipulating the mouse genome to engineer precise functional syntenic replacements with human sequence. *Cell* **128**, 197-209

Wang, G., et al., (2017). Efficient, footprint-free human iPSC genome editing by consolidation of Cas9/CRISPR and piggyBac technologies. *Nature Protocols* **12**, 88–103

Wang, Q., et al., (2005). Spatial and temporal recruitment of androgen receptor and its coactivators involves chromosomal looping and polymerase tracking. *Mol. Cell* **19**, 631–642

Wang, T., et al., (2014). Genetic screens in human cells using the CRISPR-Cas9 system. *Science* **343**, 80–84

Warren, L., et al., (2010). Highly efficient reprogramming to pluripotency and directed differentiation of human cells with synthetic modified mRNA. *Cell Stem Cell* **7**, 618–630

Weatherall, D., et al., (1975). A form of hereditary persistence of fetal haemoglobin characterized by uneven cellular distribution of haemoglobin F and the production of haemoglobins A and A2 in homozygotes. *Br J Haematol.* **29**, 205-220

Weissbein, U., et al., (2014). Quality control: genome maintenance in pluripotent stem cells, *J. Cell Biol.* **204**, 153–163

Welch, J., et al., (2004) Global regulation of erythroid gene expression by transcription factor GATA-1. *Blood* **104**, 3136–3147

Wen, Z. et al., (2014). Synaptic dysregulation in a human iPS cell model of mental disorders. *Nature* **515**, 414–418

Wernig, M., et al., (2007). In vitro reprogramming of fibroblasts into a pluripotent ES-cell-like state. *Nature* **448**, 318-324

Wernig, M., et al., (2008). A drug-inducible transgenic system for direct reprogramming of multiple somatic cell types. *Nat. Biotechnol*. **26**, 916-924

Whyte, W., et al., (2013). Master Transcription Factors and Mediator Establish Super-Enhancers at Key Cell Identity Genes. Cell **153,** 307–319

Wienert, B., et al., (2017). KLF1 drives the expression of fetal hemoglobin in British HPFH. *Blood* **130**, 803- 807

Wijgerde, M., et al., (1996). The role of EKLF in human beta-globin gene competition. *Genes and Development* **10**, 2894–2902

Wilmut, I., et al., (1997). Viable offspring derived from fetal and adult mammalian cells. *Nature* **385**, 810-813

Winichagoon, P., et al., (1982). Multiple arrangements of the human embryonic zeta globin genes. *Nucleic Acids Res*. **10**, 5853–5868

Woltjen, K. et al., (2009). piggyBac transposition reprograms fibroblasts to induced pluripotent stem cells. *Nature* **458**, 766–770

Wong, P., et al., (1986). Hemoglobin switching during murine embryonic development: evidence for two populations of embryonic erythropoietic progenitor cells. *Blood* **67**, 716–721

Wontakal, S., et al., (2012). A core erythroid transcriptional network is repressed by a master regulator of myelolymphoid differentiation. *PNAS* **109**, 3832–3837

Wood, W., et al., (1982). The British type of non-deletion HPFH: characterization of developmental changes in vivo and erythroid growth in vitro. *Br J Haematol.* **50**, 401-414

Woon Kim, Y., et al., 2011. The distinctive roles of erythroid specific activator GATA-1 and NF-E2 in transcription of the human fetal gamma-globin genes. *Nucleic Acids Research* **39**, 6944-6955

Wu, W., et al., (2014). Dynamic shifts in occupancy by TAL1 are guided by GATA factors and drive large-scale reprogramming of gene expression during hematopoiesis. *Genome Res.* **24**, 1945-1962

Xiang, J., et al., (2015). In vitro culture of stress erythroid progenitors identifies distinct progenitor populations and analogous human progenitors. *Blood* **125**, 1803–1812

Xie, F., et al. (2014). Seamless gene correction of β-thalassemia mutations in patient-specific iPSCs using CRISPR/Cas9 and piggyBac. *Genome Res.* **24**, 1526–1533

Xie, H., et al., (2004). Stepwise reprogramming of B cells into macrophages. *Cell* **117**, 663–676

Xu, P., et al., (2015). Both TALENs and CRISPR/Cas9 directly target the HBB IVS2-654 (C > T) mutation in β-thalassemia-derived iPSCs. *Sci. Rep*. **5**, 12065

Yamanaka, S. (2012). Induced pluripotent stem cells: past, present, and future. *Cell Stem Cell* **10**, 678–684

Yang L, et al. (2013) Optimization of scarless human stem cell genome editing. *Nucleic Acids Res* **41**, 9049–9061

Yang, L., et al. (2014). Targeted and genome-wide sequencing reveal single nucleotide variations impacting specificity of Cas9 in human stem cells. *Nat Commun* **5**, 5507

Ye, L., et al., (2014). Seamless modification of wild-type induced pluripotent stem cells to the natural CCR5Δ32 mutation confers resistance to HIV infection *PNAS* **111**, 9591-9596

Yu, J., et al. (2007). Induced pluripotent stem cell lines derived from human somatic cells. *Science* **318**, 1917-1920

Yu, J., et al., (2009). Human induced pluripotent stem cells free of vector and transgene sequences. *Science* **324**, 797–801

Yu, M. et al., (2009b). Insights into GATA-1-mediated gene activation versus repression via genome-wide chromatin occupancy analysis. *Molecular Cell* **36**, 682-695

Yusa, K., etal., (2009). Generation of transgene-free induced pluripotent mouse stem cells by the piggyBac transposon. *Nat. Methods* **6**, 363–369

Zambidis, E., et al., (2005). Hematopoietic differentiation of human embryonic stem cells progresses through sequential hematoendothelial, primitive, and definitive stages resembling human yolk sac development. *Blood* **106**, 860–870

Zarrei, M., et al., (2015). A copy number variation map of the human genome. *Nat. Rev. Genet.* **16**, 172–183

Zaret, K., and Carroll, J., (2011). Pioneer transcription factors: establishing competence for gene expression. *Genes & development* **25**, 2227-2241

Zaret, K., and Mango, S., (2016). Pioneer Transcription Factors, Chromatin Dynamics, and Cell Fate Control. *Curr Opin Genet Dev*. **37**, 76–81

Zhang, C., et al., (2006). Angiopoietinlike proteins stimulate ex vivo expansion of hematopoietic stem cells. *Nat Med* **12**, 240–245

Zhang, J., et al., (2004). E protein silencing by the leukemogenic AML1-ETO fusion protein. *Science* **305**, 1286-1289

Zhang, W. and Bieker, J., (1998). Acetylation and modulation of erythroid Kruppel-like factor (EKLF) activity by interaction with histone acetyltransferases. *PNAS* **95**, 9855–9860

Zhao, X., et al. (2009). iPS cells produce viable mice through tetraploid complementation. *Nature* **461**, 86–90

Zwaka, T., and Thomson, J., (2003). Homologous recombination in human embryonic stem cells. *Nat. Biotechnol*. **21**, 319–321