Acquired Alpha Thalassaemia in Myelodysplastic Syndrome



Pak Leng Cheong

Green Templeton College

A thesis submitted for the degree of

Doctor of Philosophy

Trinity 2018

Abstract

Alpha thalassaemia occurs when there is reduced production of the α-globin compared to β-globin chains of haemoglobin (α_2/β_2). Although this almost always results from inherited defects in the α-globin genes, over the past 50 years, a rare acquired form of α-thalassaemia (AT) has been reported in patients with myelodysplastic syndrome (MDS); the so-called ATMDS syndrome. With one exception, in these individuals there appear to be no mutations in the α-globin genes or their regulatory elements. However, the myeloid cells in patients with ATMDS frequently harbor mutations in the chromatin remodeling factor ATRX: an X-encoded *trans*-acting factor known to regulate α-globin expression. Defining how *trans*-acting mutations down regulate the well characterized α-globin genes in patients with ATMDS should increase our general understanding of how epigenetic modifiers may regulate mammalian gene expression in health and disease.

The first part of this project extended our analysis of the spectrum of mutations in patients with ATMDS. In addition to the expected mutations in *ATRX* we also found frequent mutations in other epigenetic modifiers particularly *ASXL1*, *U2AF1* and *PHF6*. The recurrent observation of mutations in all four of these genes suggested that they represent a mutational signature of ATMDS and indeed one case was retrospectively diagnosed based of this signature. Single cell analysis demonstrated that

mutations in all these four genes were present within the single $CD34^+$ stem/progenitor cells.

$CD34^+$ stem/progenitor cells from three newly identified individuals with ATMDS were expanded and differentiated along the erythroid pathway. In ATMDS most $CD34^+$ cells did not show erythroid potential, measured by expression of the erythroid cell surface markers CD71 and CD235a. This reflects the well known block in differentiation seen in patients with myelodysplasia. Of the cells that did express erythroid markers, the ratio of α to non-α globin gene expression was reduced by 50% recapitulating, to some degree, the α-thalassaemia observed in the peripheral blood of patients with ATMDS. Chromatin accessibility in differentiating erythroid cells, assayed by ATAC-seq, showed a generalised reduction in accessibility at the α-globin locus.

Constitutive *ATRX* mutations lead to mild α-thalassaemia in ATR-X syndrome. By contrast α-globin expression in some patients with ATMDS may be almost completely silenced with α/β globin RNA and protein synthesis ratios < 0.1. This cannot be explained by *ATRX* mutation alone. To analyse the role of the newly identified epigenetic modifiers mutated in ATMDS, we developed a CRISPR/Cas9 editing protocol, combined with single cell FACS sorting, to model ATMDS in the human umbilical cord blood derived erythroid cell line (HUDEP-2). *ATRX*, *PHF6*, *ASXL1* and *U2AF1* mutations were sequentially introduced into cells derived from a single clone of HUDEP-2. This clone with the ATRX long isoform knocked out did

not show any significant decrease in α-globin gene expression. However when all four mutations had been introduced sequentially into the same clone there was a both a block in erythroid differentiation, and α-globin gene expression was decreased. However, this was also accompanied by reduction in β-globin expression, at the same time, expression of γ globin was strongly increased. Therefore HUDEP-2 cells do not fully recapitulate the ATMDS phenotype and this may be an indication of their limitations as a model system, nevertheless they may provide an opportunity to study the effect of mutations that affect erythroid differentiation. It seems likely that the combination of mutations in the epigenetic modifiers *ATRX*, *ASXL1*, *PHF6* and *U2AF1* discovered here combine to produce a severe reduction in α–globin expression, and a more faithful cell model of this disease will be required to investigate this further.

Acknowledgement

The last four years had been an adventurous journey at the Weatherall Institute of Molecular Medicine, and it has been a privilege to experience the fascinating medical research here. I would like to acknowledge my supervisors Professor Doug Higgs and Professor Richard Gibbons for providing this environment. I have never been a compliant student and I thank them for being accommodating and supportive. I am also grateful to have received funding from the Nuffield Dominion Trust, Governo da Região Administrativa Especial de Macau, and from the Institute for this period.

The talented team at the unit has provided countless support throughout the research. I would like to thank Assistant Professor Petter Woll for his assistance in single cell assay, and Dr Joey Riepsaame and Dr Philip Hublitz for teaching me everything about CRISPR. I am grateful for Dr Jelena Telenius for her bioinformatic skills and teaching. This project would have been impossible without the help on the single cell sorting (and anything car-related) from Mr Kevin Clark at the Flow Cytometry Facility.

I would like to express my gratitude to the generosity of patients involved in this project. I am also indebted to colleagues who have helped to obtain samples for this project – Mr Terry Castanha, Dr Rachel Wells, Dr Charlotte Bomken, Dr Beatriz Arrizabalaga and Dr Lynne Quek, and Ms Jade Farrell who has travelled all the way to hand deliver the precious cargo. I am

thankful for the help from Dr Angela Hamblin on sequencing and analysis of patient samples throughout this project.

The Higgs, Gibbons and Buckle labs have been a wonderful group. Dr Yavor Bozhilov is more like the post-doc than a fellow DPhil student to me, being full of knowledge and as the go-to person for, literally, everything. I thank Dr Andrew King, Dr Robert Beagrie and Dr Mira Kassouf for providing numerous suggestions on the project. I would be lost without the help and guidance from Dr Chris Fisher, in and outside of the laboratory. I would also like to thank Dr David Clyne and Mr Tom Kent for all the discussion in and outside of science around the dining table. Ms Julia Truch has been a wonderful colleague to work with and her expertise on data analysis is much appreciated. A lot of effort from Dr Chris Babbs, Ms Gemma Moir-Meyer and Dr Caroline Scott has contributed to the manuscript and I also thank them for sharing a lot of tips and tricks on genome editing and erythroid differentiation.

I would like to thank Professor Adam Mead and Professor Jacqueline Boultwood for their guidance in the thesis committee, and Dr Andrea Pellagatti from the Boultwood lab for sharing the MDS samples.

There have been a lot of changes in life outside medicine. The ongoing support from my mentor, Professor Ron Trent, has been vital. This thesis is dedicated to the memory of my grandfather, my grandmother and my

father-in-law. I thank my family and my wife Lily for their unconditional support throughout the rough period.

To my daughter, I hope a thesis would show you about resilience in life.

Abbreviations

3C	Chromosome Capture Conformation
ADD	ATRX-DNMT3-DNMT3L
AML	Acute Myeloid Leukaemia
AML/MRC	Acute Myeloid Leukaemia with Myelodysplasia-Related Changes
ATAC-seq	Assay for Transposase-Accessible Chromatin with high-throughput Sequencing
ATMDS	Alpha Thalassaemia Myelodysplastic Syndrome
ATR-X	Alpha Thalassaemia mental Retardation X-linked
BAF	Brg/Brahma-Associated Factor
BAM	Binary Alignment/Map
BCB	Brilliant Cresyl Blue
BEL-A	Bristol Erythroid Line-Adult
BET	Bromodomain and Extra Terminal
BFLS	Börjeson-Forssman-Lehmann syndrome
BLAST	Basic Local Alignment Search Tool
BOPS	Bohring-Opitz Syndrome
bp	Base Pair
BPS	Branch-Point Sequence
BSA	Bovine Serum Albumin
ChIP-seq	Chromatin Immunoprecipitation with high-throughput Sequencing
CMML	Chronic Myelomonocytic Leukaemia
CRISPR	Clustered Regularly Interspaced Short Palindromic Repeats
CRISPR-Cas9	CRISPR-associated protein-9 nuclease
COSMIC	Catalogue of Somatic Mutations in Cancer
CSS	Coffin-Siris syndrome
CTD	Carboxyl-Terminal Domain
DAPI	4',6-diamidino-2-phenylindole
DMSO	Dimethyl Sulfoxide
E. coli	*Escherichia coli*
EPO	Erythropoietin
ESE	Exon Splicing Enhancer
FACS	Fluorescence Activated Cell Sorting
FBS	Fetal Bovine Serum
FRET	Förster Resonance Energy Transfer
G4	G-quadruplex
GATK	Genome Analysis Tool Kit
GFP	Green Fluorescent Protein
GM-CSF	Granulocyte-Macrophage Colony-Stimulating Factor
GMP	Granulocyte-Macrophage Progenitor
HARE-HTH	HB1/Asxl Restriction endonuclease-Helix-Turn-Helix
Hb	Haemoglobin
Hct	Haematocrit
HPFH	Hereditary Persistence of Fetal Haemoglobin
HPLC	High Performance Liquid Chromatography
HPV	Human Papillomavirus
HS	Hypersensitivity Site
HSC	Haematopoietic Stem Cell
HUDEP	Human Umbilical cord Derived Erythropoietic Progenitor cell
IEF	Isoelectric Focusing
IL	Interleukin

IMDM	Iscove's Modified Dulbecco's Medium
IPSS-R	Revised International Prognostic Scoring System
LCR	Locus Control Region
LOH	Loss of Heterozygosity
MACS	Magnetic Activated Cell Sorting
MACS2	Model-based Analysis of ChIP-Seq 2
MCH	Mean Cell Haemoglobin
MCS	Multispecies Conserved Sequence
MCV	Mean Corpuscular Volume
MDS	Myelodysplastic Syndrome
MEP	Megakaryocyte-Erythroid Progenitor
MEL	Murine Erythroleukaemia
MPN	Myeloproliferative Neoplasm
NCBI	National Center for Biotechnology Information
NoLS	Nucleolus Localisation Sequence
NuRD	Nucleosome Remodeling and Deacetylase
OMIM	Online Mendelian Inheritance in Man
PBS	Phosphate Buffered Saline
PCA	Principle Component Analysis
PCH	Pericentromeric heterochromatin
PHD	Plant Homeodomain
PML-NBs	Promyelocytic Leukaemia Nuclear Bodies
PRC	Polycomb Repressive Complex
PIC	Pre-Initiation Complex
PolII	Polymerase II
Poly(A)	Polyadenylation
P-TEFb	Positive Transcription Elongation Factor b
RAEB	Refractory Anaemia with Excess Blasts
RASL-seq	RNA-mediated oligonucleotide annealing, selection, and ligation with next-generation sequencing
RCMD	Refractory Cytopenia with Multilineage Dysplasia
RCMD-RS	Refractory Cytopenia with Multilineage Dysplasia – Ring Sideroblast
rDNA	Ribosomal DNA
rRNA	Ribosomal RNA
RRM1	RNA Recognition Motif 1 (of U2AF65)
RS-1	RAD51-Stimulatory compound 1
SCF	Stem Cell Factor
SFEM	Serum Free Expansion Media
sgRNA	Single Guide RNA
shRNA	short hairpin RNA
snRNP	small nuclear Ribonuclear Protein
SNP	Single Nucleotide Polymorphism
SR	arginine/Serine-Rich
ssDNA	Single Stranded Deoxyribonucleic Acid
SWI/SNF	SWItch/Sucrose Non-Fermentable
T-ALL	T-cell Acute Lymphoblastic Leukaemia
TBS	Tris Buffered Saline
TF	Transcription Factor
TPO	Thrombopoietin
TSCA	TruSeq Custom Amplicon
VAF	Variant Allele Frequency
VNTR	Variable Number Tandem Repeat
WGA	Whole Genome Amplification
WHO	World Health Organisation

Table of Contents

Abstract ii

Acknowledgement v

Abbreviations viii

1 Introduction 1

1.1 Alpha Thalassaemia 1

1.1.1 Haemoglobin imbalance and Haemoglobin H disease 4

1.1.2 Haemoglobin switching 5

1.1.3 Epigenetic regulation of alpha globin gene expression 6

1.2 Alpha Thalassaemia Mental Retardation X-linked syndrome 9

1.2.1 ATRX – structure and localisation 10

1.2.2 ATRX – domains and their interactions 11

1.2.3 ATRX – role in H3.3 deposition and G-quadruplex resolution 12

1.3 Acquired alpha thalassaemia in myelodysplastic syndrome 16

1.3.1 Myelodysplastic Syndrome 16

1.3.2 Erythropoiesis in normal haematopoiesis and myelodysplasia 18

1.3.3 Genetic basis of MDS 19

1.3.4 Acquired haemoglobin H disease 23

1.3.5 ATRX mutations in ATMDS 24

2 Materials and Methods 26

2.1 General Protocol 26

2.1.1 DNA extraction 26

2.1.2 General PCR and Sanger sequencing 26

2.1.3 RNA extraction 27

2.1.4 cDNA synthesis 27

2.1.5 Quantitative PCR 28

2.1.6 Western Blot 29

2.2 Assay for Transposase-Accessible Chromatin with high-throughput sequencing (ATAC-seq) 30

2.2.1 ATAC-seq data analysis 32

2.2.2 ATAC-seq differential peak analysis 32

2.3 Ion Torrent sequencing 33

2.4 Haematological staining 34

2.5 Identification of mutations in ATMDS using a myeloid panel 35

2.6 SNP array analysis 36

2.7 Peripheral blood mononuclear cell and CD34$^+$ haemopoietic stem cell separation 36

2.8 CD34$^+$ haemopoietic stem cell (HSC) isolation by MACS: 38

2.9 CD34$^+$ single cell expansion 38

2.10 CD34$^+$ HSC erythroid expansion and differentiation 40

2.11 CD34$^+$ HSC FACS analysis and FACS sorting 41

2.12 HUDEP cell expansion 42

2.13 CRISPR-Cas9 genome editing of HUDEP 42

2.13.1 Plasmid and oligonucleotide design 43

2.13.2 Surveyor assay 46

2.13.3 Homology-directed repair using single-stranded oligodeoxynucleotide 48

2.13.4 Nucleofection, single cell FACS sorting and expansion 48

2.14 Erythroid differentiation of HUDEP cells 50

3 Mutation Spectrum in ATMDS 55
3.1 Patient demographics 55
3.2 Findings from myeloid panel sequencing 56
3.2.1 Mutation load and enrichment in ATMDS 57
3.2.2 Correlation between enriched genes in ATMDS 61
3.2.3 Spliceosomal mutations in ATMDS 62
3.3 Temporal order of mutations in ATMDS 62
3.4 *ATRX* mutations in ATMDS 64
3.5 *ASXL1* truncating mutations in ATMDS 65
3.5.1 ASXL1 mutation in erythropoiesis 67
3.6 The role of U2AF1 in acceptor splice site recognition 69
3.6.1 Non-canonical role of U2AF1 71
3.6.2 U2AF1 hotspot mutations in MDS 72
3.6.3 U2AF1 in erythropoiesis 74
3.7 Constitutive and somatic mutations in *PHF6* 76
3.7.1 Somatic mutation of PHF6 in haematologic neoplasm 77
3.7.2 Function of PHF6 78
3.7.3 PHF6 mutation in ATMDS 80
3.8 Clinical investigations of ATMDS patients 83
3.9 ATMDS mutations at a single cell level 86
3.10 AML Transformation in ATMDS 89
3.11 Summary 92

4 Erythroid differentiation of ATMDS CD34⁺ progenitor cells 95
4.1 Globin gene expression analysis in peripheral blood 95
4.2 Block in erythroid differentiation in ATMDS 96
4.3 Separation of erythroid and non-erythroid populations during differentiation by FACS sorting 100
4.4 Low haemoglobin expression and alpha/non-alpha globin gene expression ratio in ATMDS erythroid population 102
4.5 Chromosome accessibility during erythroid differentiation in ATMDS 104
4.6 Differential ATAC-seq peak analysis 108
4.7 Summary 113

5 Generation of HUDEP-2 cell line with ATMDS mutations 117
5.1 Generation of ATRX knockout HUDEP-2 cell 119
5.2 Double knockout of ATRX and PHF6 120
5.3 Protection of wildtype allele in generation of heterozygous ASXL1 truncation 121
5.4 Two-step approach for generation of heterozygous p.Gln157Pro mutation in U2AF1 123
5.5 Repair of ATRX knockout in the quadruple mutant B10 E5 124
5.6 Erythroid differentiation of HUDEP cell 125
5.7 Chromosome accessibility in differentiated HUDEP 133
5.8 HUDEP clone with only *U2AF1* mutation 136
5.9 Summary 139

6 Discussion 144
6.1 Signature and significance of mutations in ATMDS 144
6.2 Block in differentiation in ATMDS CD34⁺ and HUDEP-2 erythroid differentiation 146

6.3 The potential role of spliceosomal and *ATRX* mutations in alpha globin gene expression **148**
6.4 *ASXL1* and *PHF6* in RNA polymerase II C-Terminal Domain phosphorylation **152**
6.5 Links of ATMDS mutations to PRC2 ..... **154**
6.6 HUDEP-2 as a model for erythroid differentiation **155**
7 Appendix **159**
7.1 Spectrum of mutations in ATMDS **160**
7.2 Characterisation of ATMDS patients **163**
7.3 Erythroid differentiation of ATMDS patients **166**
7.4 HUDEP-2 genome editing **168**
7.4.1 sgRNAs and ssODN used for introducing *ATRX*, *PHF6* and *ASXL1* mutations 168
7.4.2 Introduction of U2AF1 heterozygous p.Gln157Pro mutation in the triple mutant APA G7 168
7.4.3 Repairing *ATRX* mutation in quadruple mutant clone B10 E5 171
7.4.4 Characteristics of genome edited HUDEP-2 clones during erythroid differentiation 173
7.4.5 Introducing U2AF1 p.Gln157Pro mutation in wildtype HUDEP-2 176
7.5 Predicted G4 sequence in alpha and beta globin genes **178**
Publication resulted from this work **180**
Invited Presentation **180**
References **181**

1 Introduction

Alpha thalassaemia and myelodysplastic syndrome are two apparently unrelated diseases. The former is an inherited disorder, whereas myelodysplastic syndrome is an acquired bone marrow disorder. In the rare cases of acquired alpha thalassaemia myelodysplastic syndrome (ATMDS), severe alpha thalassaemia emerges in patients with an intact alpha globin locus. This presents an opportunity to study the regulation of alpha globin gene expression without perturbation of the locus. In the following sections, we will discuss the pathological basis of alpha thalassaemia, how alpha globin gene expression is controlled epigenetically, the role of ATRX in alpha thalassaemia and how this chromatin remodeler links alpha thalassaemia and myelodysplastic syndrome together. We will also touch on the molecular basis of myelodysplastic syndrome and show how ATRX alone cannot explain the severe alpha thalassaemia seen in these patients, which leads to this project with three aims. First, we will explore the mutation spectrum in ATMDS, followed by study on erythroid differentiation of haematopoietic progenitor cells in these patients to look for underlying cause for the alpha thalassaemia phenotype, and finally the attempt to establish a cellular model of ATMDS.

1.1 Alpha Thalassaemia

Alpha thalassaemia is one of the most common inherited genetic disorder worldwide, with 20.7% of the world's population being a carrier[1]. Alpha thalassaemia is an inherited disorder caused by reduction in alpha globin synthesis. Alpha globin is one of the two haem-containing subunits in adult

haemoglobin A, a tetrameric metalloprotein vital in oxygen transport in vertebrates (Figure 1-1). Haemoglobin A is formed by two alpha and two beta subunits ($\alpha_2\beta_2$). The two subunits are encoded by two *HBA* genes (*HBA1* and *HBA2*) on chromosome 16p13.3, and one *HBB* gene on chromosome 11p15.4, respectively. Imbalance in the globin output reduces the amount of haemoglobin A in red blood cells (RBCs). In addition, accumulation of the uncoupled subunit (see section on Haemoglobin H disease) leading to haemolysis, and to a lesser degree ineffective erythropoiesis, together result in hypochromic ("less red") microcytic ("smaller") anaemia.

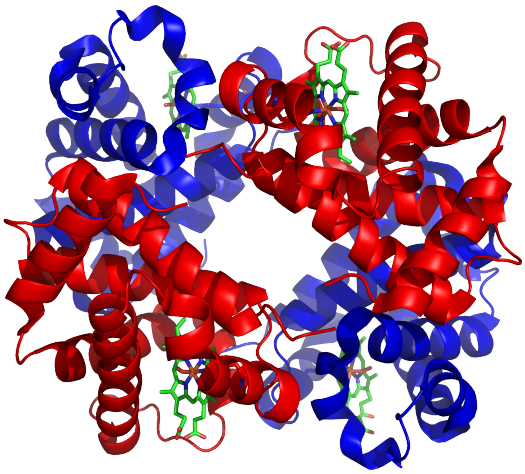


Figure 1-1. Three dimensional structure of adult haemoglobin A, showing the tetrameric structure formed by two alpha and two beta subunits (red and blue). Haem is shown in a ball-and-stick model within each subunit (PDB ID: 1GZX)[2].

In most cases, the underlying genetic cause for alpha thalassaemia is due to deletions or mutations involving *HBA1* and/or *HBA2*. These two genes are located in tandem along with one embryonic zeta globin gene (*HBZ*, ζ), two minimally expressed globin genes haemoglobin mu and theta (*HBM*, μ and *HBQ*, θ respectively), and two pseudogenes (ψζ and ψα1) within the alpha globin locus in a 5'-ζ-ψζ-μ-ψα1-α2-α1-θ-3' fashion (Figure 1-2).

The severity of alpha thalassaemia is dependent on the number of alpha globin genes affected. In normal individuals both copies of *HBA1* and *HBA2*, i.e. four *HBA* genes, contributes to the total alpha globin output. Numerous point mutations in alpha globin gene have been reported and they will not be discussed in details here (for review see Thom *et al*[3]). A heterozygous one-gene deletion such as the common 3.7kb deletion ($-\alpha^{3.7}$, Figure 1-2) removes one out of the four *HBA* genes. Such a deletion is usually clinically silent with slightly reduced haematologic indices (mean corpuscular volume, MCV and mean cell haemoglobin, MCH). An alpha thalassaemia trait with hypochromic microcytic anaemia is apparent clinically when two of the four genes are affected, such as those with South East Asian deletion (α^{-SEA}). Rare forms of deletional alpha thalassaemia have also been reported. ATR-16 syndrome refers to large (900-1700kb) deletions removing the whole alpha globin locus and beyond, resulting in alpha thalassaemia with mental retardation[4]. In another example, a deletion (named α^{-ZF}) that removes α1 and θ globin genes and juxtaposes the downstream *LUC7L* to the remaining intact α2 results in antisense RNA production and silencing of α_2 due to methylation[5].

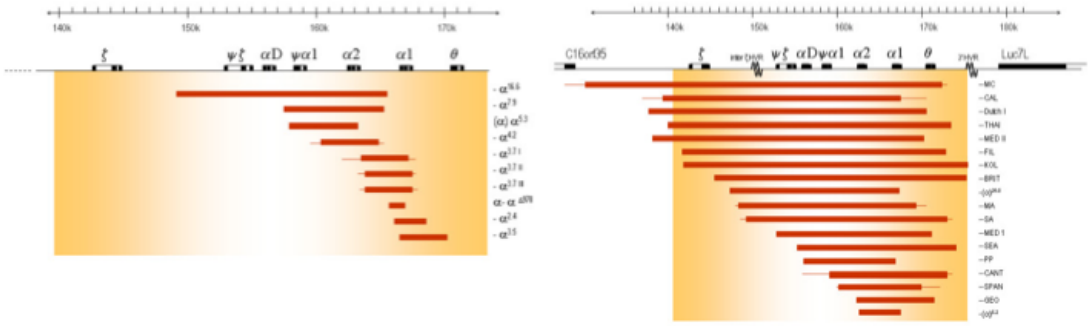


Figure 1-2. Spectrum of deletional alpha thalassaemia. Left and right panels show deletion patterns of one and two-gene deletions, respectively (red bars). The relative position of globin genes and pseudogenes is depicted above. Adapted from Harteveld and Higgs[6]. Haemoglobin mu is labeled as αD here as it is closely related to avian alpha-D globin. Previously haemoglobin mu was thought to be a pseudogene and was known as ψα2[7].

1.1.1 Haemoglobin imbalance and Haemoglobin H disease

Normal production of haemoglobin A requires precise balance between the synthesis of alpha and beta globin chains. Haemoglobin H disease occurs when three of the four *HBA* genes are affected. Haemoglobin H (Hb H) refers to tetramer of beta globin (β_4), formed due to the low alpha/beta globin chain synthesis. Hb H can be visualised in red blood cells using the supravital stain brilliant cresyl blue (BCB). Upon staining with this redox dye the unstable Hb H precipitates and form intracellular inclusions (Figure 1-3). It is readily detectable in Hb H disease (5-50% of RBCs). For patients with alpha thalassaemia trait, these inclusions can also be occasionally seen in a small proportion of cells (0.01-0.1%)[8]. The clinical presentation of Hb H disease is variable – ranging from a benign condition to transfusion-dependent anaemia[9, 10].

Loss of all four *HBA* genes leads to Hb Bart's fetalis hydrops syndrome, which is almost always lethal in utero or shortly after birth without intervention. Hb Barts refers to gamma globin tetramer (γ4), the fetal beta-like globin.

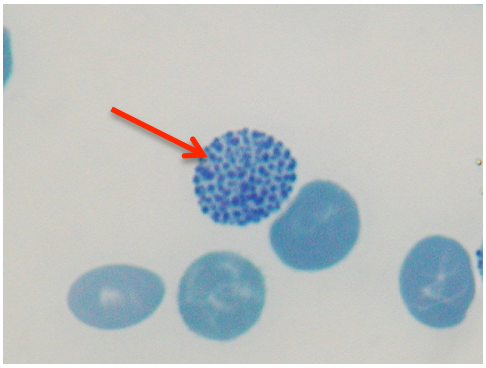


Figure 1-3. BCB stain showing multiple Hb H inclusions within a red blood cell (red arrow).

1.1.2 Haemoglobin switching

The reason for formation of Hb Barts rather than Hb H in Hb Bart's fetalis hydrops syndrome is due to the unique characteristics associated with haemoglobin switching. The beta globin locus resides on chromosome 11 and like the alpha globin locus, also contains several globin genes (5'-ε-γG-γA-δ-β-3', Figure 1-4). In human, the embryonic haemoglobin (Hb Gower, $\zeta_2\varepsilon_2$) is replaced by Hb F ($\alpha_2\gamma_2$), the major form of haemoglobin early in the developing fetus. Alpha globin continues to be expressed throughout the remaining prenatal and into the postnatal period. Subsequently fetal gamma globin is largely replaced by adult beta globin. This process only occurs at around 6-12 weeks after birth and therefore explains the detection of Hb Barts in Hb Bart's fetalis hydrops syndrome.

Hb F production is not completely switched off, and there is a large variation in the residual amount of Hb F in adults due to genetic factors[11, 12]. Deletions in the beta globin locus and point mutations in the gamma globin gene promoters cause hereditary persistence of fetal haemoglobin (HPFH). In addition, other trans-acting factors such as BCL11A, MYB, KLF1 and LRF have been identified via

genome wide association studies and functional studies to contribute to the variation in Hb F levels[13-17], highlighting the complexity in control of haemoglobin expression.

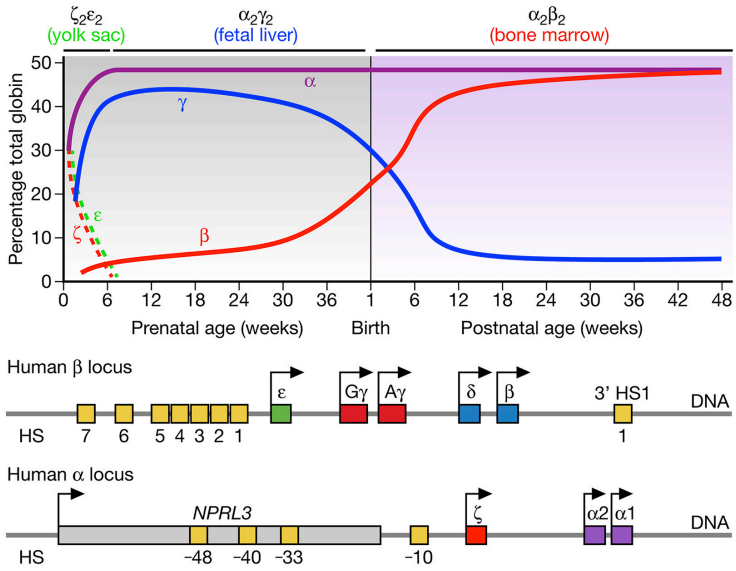


Figure 1-4. Timeline illustrating haemoglobin switching in human in pre and post-natal period. Embryonic haemoglobin ($\zeta_2\varepsilon_2$) is only expressed in yolk sac. In adults, small amount of Hb F can be detected. Yellow boxes in the map of beta and alpha loci are DNase I hypersensitivity sites (HSs), conserved noncoding sequences that constitute regulatory elements in globin expression. Three of the four HSs in the alpha globin locus reside within the gene body of the actively transcribed *NPRL3*. Adapted from Blobel *et al*[18].

1.1.3 Epigenetic regulation of alpha globin gene expression

Expression of haemoglobin is tightly controlled in a spatial (i.e. tissue-specific) and temporal (i.e. expresses only from fetus onwards) manner. Achieving appropriate alpha globin gene expression requires stringent epigenetic control given the fact that the locus is located in a gene-dense region, among which are genes (e.g. *NPRL3*) that are constitutively expressed in a variety of tissues. The locus is also relatively close to the telomere (150kb) where shortening of telomere is known to alter nearby gene expression (telomere position effect)[19, 20], although whether this effect also applies to alpha globin genes is not known.

The regulation of alpha globin gene expression has been studied in detail in mouse models. Expression is mediated by a panel of erythroid-specific transcription factors (TFs), such as GATA1, GATA2, NF-E2, SCL/TAL1 complex and KLF1. They are expressed at different stages during erythroid differentiation and bind to *cis*-regulatory elements upstream of the globin genes[21-24] (see Figure 1-5). In the human alpha globin locus, there are four *cis*-regulatory elements upstream named multispecies conserved sequence (MCS) R1 to R4 (these correspond to regions of accessible chromatin HS-48, HS-40, HS-33 and HS-10, respectively in Figure 1-5). These sites were identified based on cross species conservation and DNAse I hypersensitivity[25, 26]. Naturally occurring deletions that remove some of these elements[27-31], and various functional studies using transient expression or transgenic mice with a "humanized" alpha globin locus (where the mouse alpha globin domain was replaced by the human version) have confirmed that these elements are essential enhancers that drive alpha globin gene expression[32, 33].

Of the four enhancer elements, R2 has been shown to be the most conserved and possesses the strongest effect on alpha globin gene expression[25, 34, 35], but each element has an additive effect on expression[36]. R2 becomes bound by TFs early in erythroid differentiation. On the contrary, R3 is only involved at later stages in differentiation[21]. This is accompanied by recruitment of RNA polymerase II (PolII) and other components of pre-initiation complex (PIC) to the enhancer[37]. However, knockout of R3 and R4 in mouse had little effect on alpha globin expression and erythropoiesis[36].

Activation of alpha globin gene transcription is accompanied by looping of the enhancer elements, bound with transcription factors and PIC, to the HBA gene promoters. The PIC and promoters are brought together physically, as shown by chromosome capture conformation (3C) analysis in both mouse[37] and human[38], and coincides with transcription of alpha globin.

The activation of transcription is accompanied by changes in histone marks specifically in the alpha globin locus. In human, recruitment of polycomb repressive complex 2 (PRC2) is responsible for the silencing of alpha, but not beta globin gene transcription in non-erythroid tissues[39]. PRC2 recruitment leads to histone H3 tri-methylation at lysine 27 (H3K27me3). In erythroid tissue R2 is necessary to clear PRC2 from the alpha globin CpG island, via recruitment of demethylase JMJD3 to the enhancer. H3K27me3 is replaced by active histone marks such as histone H3 tri-methylation at lysine 4 (H3K4me3) and acetylation in histones H3 and H4[40-42].

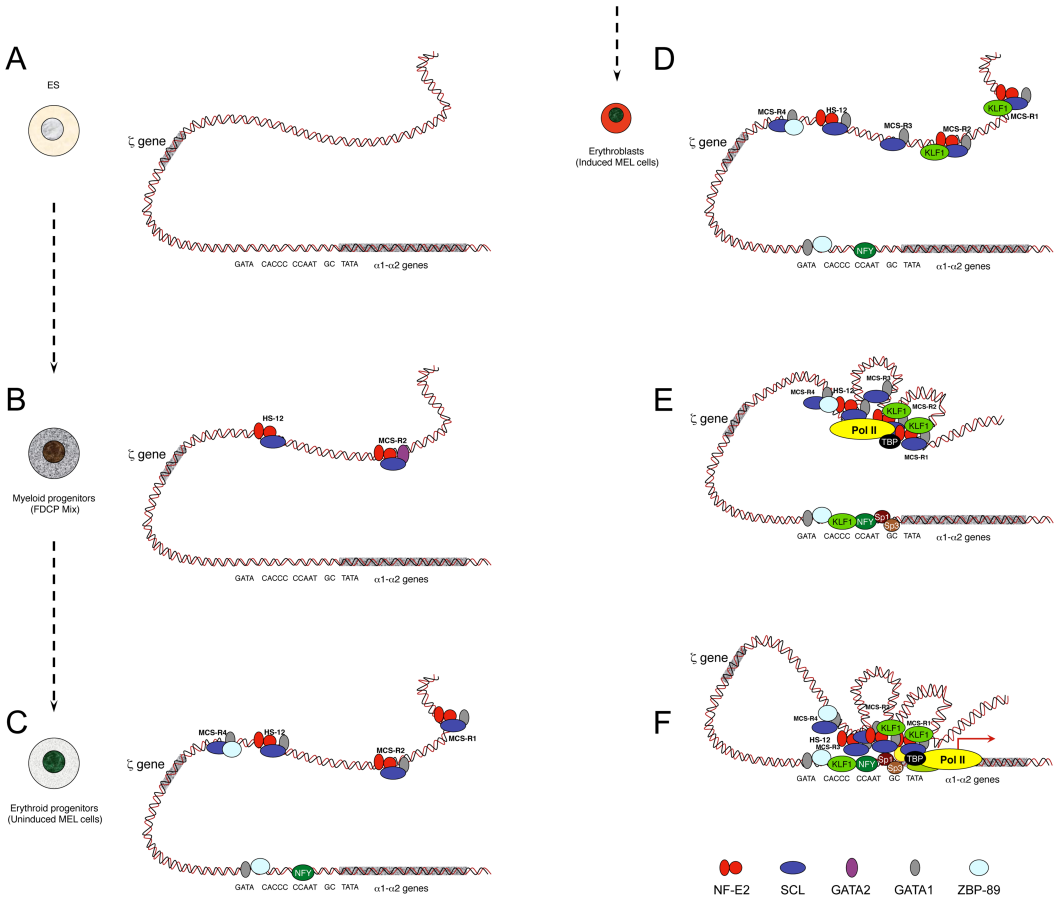


Figure 1-5. The model of transcriptional control in mouse alpha globin locus. As cells differentiate in the erythroid pathway from A to D, TFs are bound to DNAse I hypersensitive sites. In late stage of erythropoiesis PolII is recruited to the enhancer and the PolII-bound enhancers loop to contact the promoter region. N.B. 1: these studies were performed in different murine cell types to represent different stages of differentiation. HS-12 is unique to mouse and not found in human; 2: nuclear factor Y (NFY) binding is common to both human and mouse alpha globin gene promoter, whereas 89-kDa zinc finger protein (ZBP-89) binding is not found in human. Sp1 and Sp3 are TFs belonging to the specificity protein/Krüppel-like factor family. TBP: TATA binding protein. From Vernimmen[43].

1.2 Alpha Thalassaemia Mental Retardation X-linked syndrome

As mentioned above, a group of patients with alpha thalassaemia and intellectual disability exist who have large deletions involving the end of 16p13.3 (ATR-16 syndrome). However it was discovered that not all patients with alpha thalassaemia and intellectual disability had deletions in chromosome 16[44]. These patients have an intact alpha globin locus but develop alpha thalassaemia with variable degree of Hb H inclusion. The alpha thalassaemia is mild and sometimes undetectable without a careful examination using BCB stain[45, 46],

and in some cases not visible at all[47, 48]. Other clinical signs include severe mental retardation, specific facial characteristics and multi-system abnormalities (skeletal, genital, urorenal, gastrointestinal, reviewed in Gibbons[4]).

Family studies and linkage analysis mapped the defect to Xq12-q21.31[44, 49-51], hence the syndrome was named alpha thalassaemia mental retardation X-linked (ATR-X) syndrome. The gene responsible was later identified via cDNA hybridization from an ATR-X syndrome patient with deletion in the *XH2* gene in Xq13 (now known as *ATRX* gene)[52] and confirmed in other affected individuals.

1.2.1 ATRX – structure and localisation

The *ATRX* gene on Xq13.3 contains 35 exons, coding for 2,492 amino acids (NCBI reference sequence NG_008838.3 and NP_000480.3) producing a 280kD protein. An alternative transcript exists where exon 6 is spliced out, producing a smaller product at 265kD[53, 54]. There is also a truncated protein isoform (ATRXt), coded by a transcript that truncates at intron 11[55]. ATRXt is lowly expressed compared to the full-length protein. Sequence homology suggested *ATRX* contains two highly conserved domains. At the N-terminal is an ATRX-DNMT3-DNMT3L (ADD) domain (residues 161-292)[56] It is so named due to being a shared feature of ATRX and the methyltransferases DNMT3A, DNMT3B and DNMT3L. It contains a hybrid of cysteine-rich GATA-like and PHD-like zinc finger, followed by an alpha helix. The C-terminus contains an ATPase/helicase domain that belongs to the SWI2/SNF2 (SWItch/Sucrose Non-Fermentable) family (residues 1550-2226)[52]. The C-terminus also interacts with the methyl

CpG binding protein MeCP2[57, 58]. Between these two domains are regions where ATRX is reported to interact with HP1 alpha (residues 580-595, also known as CBX5), EZH2 (minimum residues 475-734) and death domain-associated protein DAXX (residues 1260-1289)[59-61].

In ATR-X syndrome, the majority of mutations are missense mutations within the ADD or the SWI2/SNF2 domain[62, 63]. Various studies have attempted to elucidate the function of ATRX and how it leads to alpha thalassaemia.

To gain insight on the function of ATRX, it is essential to explore its cellular and genomic localisation. At a cellular level, ATRX is found in promyelocytic leukaemia nuclear bodies (PML-NBs) at S and G2 phases[64, 65]. It is detected in short arms of acrocentric chromosomes at metaphase containing ribosomal DNA tandem repeats[66]. It is also found in the inactive X chromosome and both arms of the largely heterochromatic Y chromosome[67, 68]. Throughout the genome ATRX is found in pericentromeric and telomeric heterochromatin where PML-NBs co-localise. It is also enriched at retrotransposons, repetitive zinc finger motifs in the 3' exons of zinc finger genes, in G and CpG-rich regions as well as tandem repeats[66, 69-71]. The common theme from these studies points to the preference of ATRX to G-rich repetitive DNA sequences in heterochromatin.

1.2.2 ATRX – domains and their interactions

The specificity of genomic localisation of ATRX is generally thought to be conferred by its ADD domain, which shows higher affinity for chromatin signature associated with heterochromatin and repeats, rich in modified

histones H3K9me2/3 and H4K20me2/3 (histone 3 and 4 with di/tri-methylated lysine 9 and 20, respectively)[72]. The ADD domain binds unmethylated H3K4[73], but not the modified histone H3K4me2/3 which is associated with gene activation[74, 75]. The binding of ATRX to these regions is enhanced by interaction with HP1 alpha, whose chromodomain also associates with H3K9me3[73, 76-78].

The C-terminal SWI2/SNF2 domain of ATRX is considered the functional part of ATRX. The SWI/SNF group is a subfamily of chromatin remodelers. Chromatin is built from an average of 147 base pairs of DNA wrapped around an octamer of histones comprising of histones H2A, H2B, H3 and H4. Chromatin remodelers modify chromatin architecture by sliding nucleosomes, ejecting or inserting nucleosomes or histone components in an ATP-dependent manner to allow or restrict access to DNA from transcription factors and transcriptional machinery. Xue *et al*[64] showed that the ATRX SWI2/SNF2 domain in ATRX is atypical. While, like the SWI/SNF complex, it is able to disrupt DNA-histone interaction and possesses DNA translocase activity, it does not alter nucleosome phasing that is visible after on DNAse I digestion.

1.2.3 ATRX – role in H3.3 deposition and G-quadruplex resolution

Studies on ATRX domain functions suggest it is involved in remodeling by interaction with histone variant H3.3. In earlier studies ATRX was found to co-localise with DAXX and the histone variant H3.3 in pericentromeric and telomeric heterochromatin[79, 80]. H3.3 differs from canonical H3 by 4-5 amino acids, one of which (glycine 90) is crucial for its interaction with DAXX[81, 82].

Multiple studies revealed that it is actually DAXX that is required to deposit H3.3 in these regions, and in telomere ATRX recruits DAXX and provides processivity for this process[80, 83-86]. It is suggested that H3.3 is recruited as a dimer with H4 to PML-NB by DAXX prior to its deposition[85]. A later report also detected H3.3 deposition by ATRX/DAXX in heterochromatin throughout the genome in mouse, in particular endogenous retroviral repeats, intragenic CpG islands, and the methylated allele in imprinted, differentially methylated regions[87]. The general consensus is that DAXX performs the action of histone incorporation, while ATRX is responsible for the specificity and processivity of this action. In contrast, the deposition of H3.3 in euchromatin in facilitated by HIRA (histone regulator A)[80, 83, 88].

Compared to canonical histones H3.1 and H3.2, which are only expressed during S phase, H3.3 is expressed throughout the cell cycle[88]. This property is important in transcribed genes, in heterochromatic regions as well as following DNA damages where histones need to be replaced for “re-chromatinisation”. H3.3 marks promoters of actively transcribed genes[89]. Therefore although in general H3.3 deposition by ATRX/DAXX occurs in heterochromatic regions (i.e. not in expressed genes) it is possible that ATRX indirectly regulates gene expression in GC-rich regions, alpha globin genes being an example, by participating in H3.3 deposition. This was tested in Atrx knockout experiments. In a mouse model downregulation of ancestral pseudoautosomal region genes such as Csf2ra and Dhrsx is observed in Atrx knockout[90]. These genes are located in subtelomeric region, akin to the human alpha globin locus. Atrx and H3.3 co-localise to GC-rich regions within the gene bodies of these genes. When

Atrx was knocked out, H3.3 enrichment in these regions was reduced. Conversely, PolII binding was enriched in these genes, suggestive of stalling of PolII[90]. It is proposed that in GC-rich regions, the loss of Atrx and H3.3 deposition leads to formation of G-quadruplexes, which results in PolII stalling transcription suppression. The following paragraph will discuss the role of ATRX in G-quadruplex resolution, and how this could form the basis of alpha thalassaemia in ATR-X syndrome.

G-quadruplex (G4) is a non-B form, four-stranded DNA or RNA structure. Four interspersed guanine tracts are linked by Hoogsteen base-pairing, adopting a planar arrangement termed G-quartet[91] (Figure 1-6). The consensus sequence for G4 is G_x-N_{1-7}-G_x-N_{1-7}-G_x-N_{1-7}-G_x, where x≥3 and N is any nucleotide, although variations to this consensus do exists as shown by next-generation sequencing combined with G4-dependent DNA polymerase stalling (G4-seq)[92]. Stacking of G-quartets forms the three-dimensional G4[93]. G4 can be formed within the same DNA/RNA molecule, or between two or more strands of DNA/RNA. DNA and RNA can form hybrid G4[94].

The existence of G4 *in vitro* and *in vivo* is now generally accepted, with evidence from indirect detection using synthetic single chain antibody fragments (Sty 3 and Sty 49) targeting intermolecular G4 in protozoan[95], direct visualization of "G-loop" structure under electron microscopy in an G4 inducible system in *E. coli*[96], radiolabelled G-quadruplex ligand 3H-360A binding G4 in human telomeric regions[97], and two independent G4-specific antibodies

demonstrating presence of G4 in human cells by immunofluorescence microscopy[98-100].

At the genome level, given the guanine-rich nature of G4-forming sequences G4 is bioinformatically predicted to be enriched in telomeric repeats (TTAGGG in vertebrates), variable number tandem repeats (VNTRs), gene promoters, immunoglobulin switch regions and ribosomal DNA (rDNA). In addition, the non-coding strand at the 5' end of the first intron in almost half of all human genes, and both 5' and 3' untranslated regions are predicted to form G4[101, 102]. Some of these predictions have been verified by G4-seq[92].

Evidence of ATRX interacting with G4, and specifically G4 at the alpha globin locus comes from Law *et al*[71]. The authors showed that 50% of putative G4 sequences in the genome were bound by ATRX. At the alpha globin locus, the ψζVNTR with G4 forming potential 1kb upstream of HBM promoter is bound by ATRX. This VNTR, containing sequence $CGC(GGGGCGGGG)_n$, indeed formed G4 in vitro as shown by circular dichroism spectroscopy and atomic force microscopy[103]. The VNTR size is variable between individuals and the longer the VNTR, the more downregulation is observed in alpha globin expression in patients with ATRX syndrome[71]. ATRX was also shown to interact with G4 structure in vitro by circular dichroism spectroscopy and gel-shift assay[103]. Combining these data, ATRX was suggested to regulate alpha globin transcription by resolution of G4 structure.

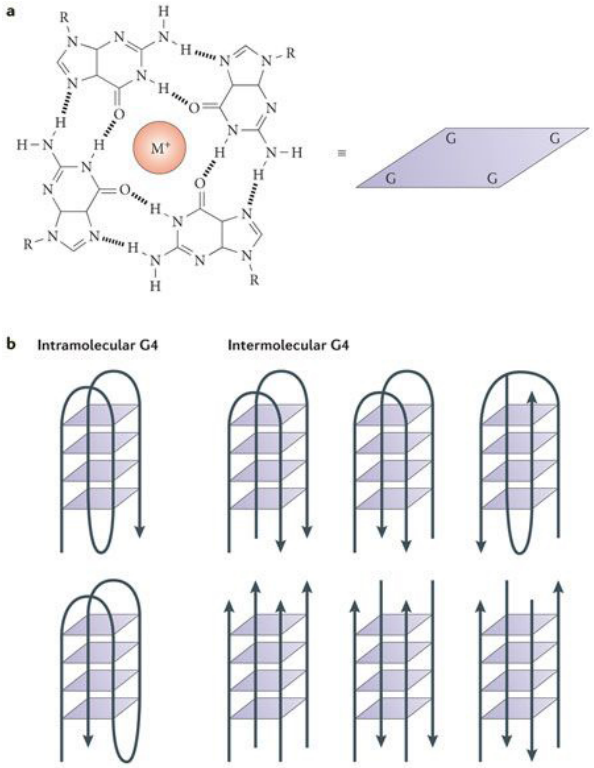


Figure 1-6. a: Illustration of G-quartet. b: G4 formed by stacking of G-quartets. On the left are two G4 configurations formed by the same DNA/RNA strand. G4s formed by different DNA/RNA strands are shown on the right. Adapted from Bochman *et al*[104].

1.3 Acquired alpha thalassaemia in myelodysplastic syndrome

The link between acquired alpha thalassaemia and myelodysplastic syndrome becomes more apparent with the discovery of somatic ATRX mutations in cases of acquired alpha thalassaemia myelodysplastic syndrome (ATMDS). The next section provides background on MDS and ATMDS, forming the basis of the current research.

1.3.1 Myelodysplastic Syndrome

Myelodysplastic syndrome (MDS) refers to a heterogeneous group of haematologic neoplasms due to clonal expansion of haematopoietic stem cells (HSCs) within the bone marrow. It is a disease of aged populations, with an incidence of up to 30 cases per 100,000 population per year in patients above 70

years old[105]. Clinically, it is manifested by refractory peripheral cytopenia, morphological dysplasia in haematopoietic cells, normo/hyper-cellular bone marrow indicating ineffective haematopoiesis, and risk of progression to acute myeloid leukaemia (AML). The cytopenia can involve one or more myeloid lineages of haematopoiesis.

Conventionally, World Health Organization (WHO) classifies MDS based on clinical, morphological and cytogenetic criteria (Table 1-1)[106]. In relation to this is the revised International Prognostic Scoring System (IPSS-R), which stratifies patients into risk groups based on the number of cytopenic lineages, bone marrow blast percentage and karyotype[107]. However, with the availability of large scale sequencing, information at the individual gene level has the potential to provide a more rational classification system for MDS.

Name	Dysplastic lineages	Cytopenias*	Ring sideroblasts as % of marrow erythroid elements	BM and PB blasts	Cytogenetics by conventional karyotype analysis
MDS with single lineage dysplasia (MDS-SLD)	1	1 or 2	<15%/<5%†	BM <5%, PB <1%, no Auer rods	Any, unless fulfills all criteria for MDS with isolated del(5q)
MDS with multilineage dysplasia (MDS-MLD)	2 or 3	1-3	<15%/<5%†	BM <5%, PB <1%, no Auer rods	Any, unless fulfills all criteria for MDS with isolated del(5q)
MDS with ring sideroblasts (MDS-RS)					
MDS-RS with single lineage dysplasia (MDS-RS-SLD)	1	1 or 2	≥15%/≥5%†	BM <5%, PB <1%, no Auer rods	Any, unless fulfills all criteria for MDS with isolated del(5q)
MDS-RS with multilineage dysplasia (MDS-RS-MLD)	2 or 3	1-3	≥15%/≥5%†	BM <5%, PB <1%, no Auer rods	Any, unless fulfills all criteria for MDS with isolated del(5q)
MDS with isolated del(5q)	1-3	1-2	None or any	BM <5%, PB <1%, no Auer rods	del(5q) alone or with 1 additional abnormality except −7 or del (7q)
MDS with excess blasts (MDS-EB)					
MDS-EB-1	0-3	1-3	None or any	BM 5%-9% or PB 2%-4%, no Auer rods	Any
MDS-EB-2	0-3	1-3	None or any	BM 10%-19% or PB 5%-19% or Auer rods	Any
MDS, unclassifiable (MDS-U)					
with 1% blood blasts	1-3	1-3	None or any	BM <5%, PB = 1%,‡ no Auer rods	Any
with single lineage dysplasia and pancytopenia	1	3	None or any	BM <5%, PB <1%, no Auer rods	Any
based on defining cytogenetic abnormality	0	1-3	<15%§	BM <5%, PB <1%, no Auer rods	MDS-defining abnormality
Refractory cytopenia of childhood	1-3	1-3	None	BM <5%, PB <2%	Any

*Cytopenias defined as: hemoglobin, <10 g/dL; platelet count, $<100 \times 10^9$/L; and absolute neutrophil count, $<1.8 \times 10^9$/L. Rarely, MDS may present with mild anemia or thrombocytopenia above these levels. PB monocytes must be $<1 \times 10^9$/L

†If *SF3B1* mutation is present.

‡One percent PB blasts must be recorded on at least 2 separate occasions.

§Cases with ≥15% ring sideroblasts by definition have significant erythroid dysplasia, and are classified as MDS-RS-SLD.

Table 1-1. The current 2016 revision of WHO classification on MDS. Note the term "MDS" replaces "refractory cytopenia" or "refractory anaemia" used in the previous 2008 WHO classification. From Arber *et al*[106].

1.3.2 Erythropoiesis in normal haematopoiesis and myelodysplasia

During erythropoiesis, bipotent progenitor cells (megakaryocyte-erythroid progenitors, MEPs) either differentiate into erythroid cells or megakaryocytes. Once committed to the erythroid lineage the progenitor cell undergoes several stages of erythroid differentiation that can be defined by morphology, cell surface markers and gene expression. MEPs can be enriched using various surface markers ($Lin^-CD34^+CD38^{mid}CD45RA^-FLT3^-MPL^+CD36^-CD41^-$[108]). The earliest morphologically recognizable erythroid cell is the proerythroblast, which undergoes successive mitoses to produce morphologically distinctive populations (basophilic, polychromatic and orthochromatic erythroblasts). The terms here describe the progressive change in colour in the stained cells as haemoglobin is produced (top pane, Figure 1-7). Finally the nuclei are expelled from orthochromatic erythroblasts to produce reticulocytes. During the process, the haematopoietic stem cell marker CD34 is lost from the cell surface, while erythroid markers such as CD71 and CD235a (Glycophorin A, GPA) become positive[109]. In terminal stages of erythropoiesis two markers, CD233 (Band 3) and CD49d (α4-integrin) become positive and negative, respectively[110](bottom pane, Figure 1-7). In patients with various forms of MDS the pattern of these terminal markers are altered, demonstrating an abnormality in terminal erythropoiesis[110].

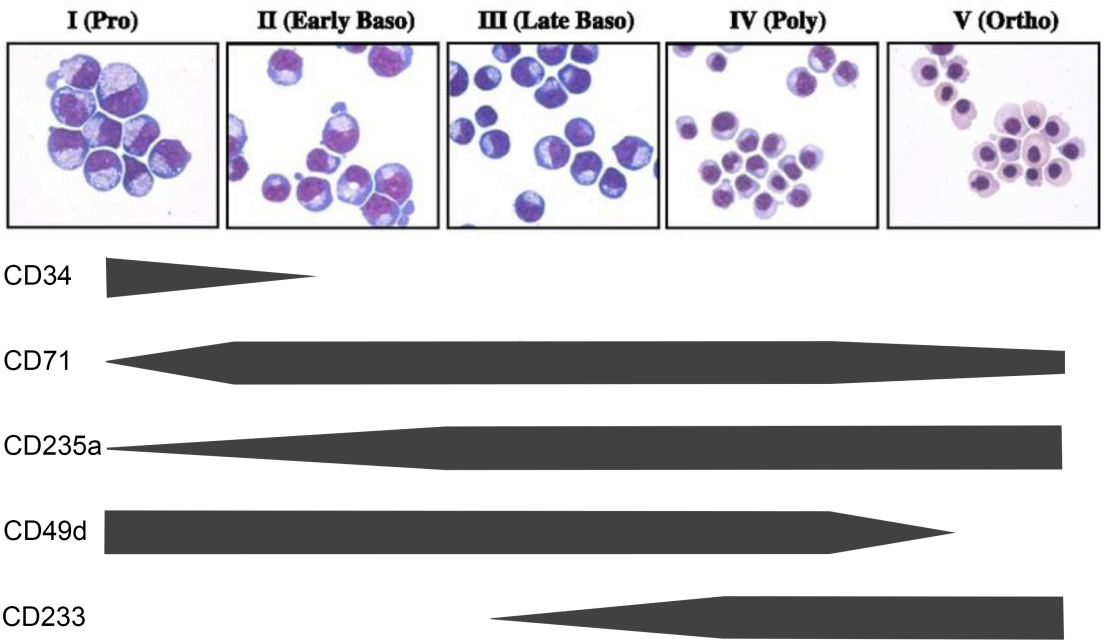


Figure 1-7. Top: Cytospin images showing morphological changes in human erythroblasts undergoing erythropoiesis in vitro. From left to right: I: proerythroblast; II: early basophilic erythroblast; III: late basophilic erythroblast; IV: polychromatic erythroblast; V: orthochromatic erythroblast. From Hu *et al*[110]. Bottom: illustration of cell surface marker expression during erythropoiesis.

Data from RNA-seq at each stage of human and murine erythropoiesis showed increasing level of alternative splicing and transcripts with intron retention at the whole transcriptomic level, particularly at the last four cell divisions before enucleation[111-113]. The changes in intron retention peaks at the orthochromatic erythroblast stage, and gene expression in general is inversely proportional to intron retention[113]. This highlights the importance of splicing in terminal erythropoiesis, which is commonly perturbed in MDS as discussed in the next section.

1.3.3 Genetic basis of MDS

The first such effort was achieved by a combination of mass spectrometry, Single Nucleotide Polymorphism (SNP) array, pyrosequencing and Sanger sequencing[114]. The most frequently mutated genes found in this report were

TET2, *ASXL1*, *RUNX1*, *TP53* and *EZH2*. These genes were involved in DNA methylation, chromatin modification, transcription regulation, DNA damage response (p53 pathway) and histone modification, respectively.

Subsequently two large MDS cohorts were characterized using focused gene panels. It became apparent that MDS is predominantly a disease of splicing and the epigenetic control of transcription (Figure 1-8), and the spectrum of mutations in MDS is broad and variable. Some interesting characteristics on the mutation spectrum of MDS emerged from these studies. Using a 111-gene panel to study 738 MDS patients, Papaemmanuil *et al*[115] detected at least one oncogenic mutation or MDS-related copy number variation in 78% of patients. More than 50% of patients had two or more mutations or copy number variations. Oncogenic mutations were found in 43 genes and only four of them (*SF3B1*, *TET2*, *SRSF2* and *ASXL1*) were represented in more than 10% of the study population. Similar observations were found in another study using a different panel in 944 patients[116].

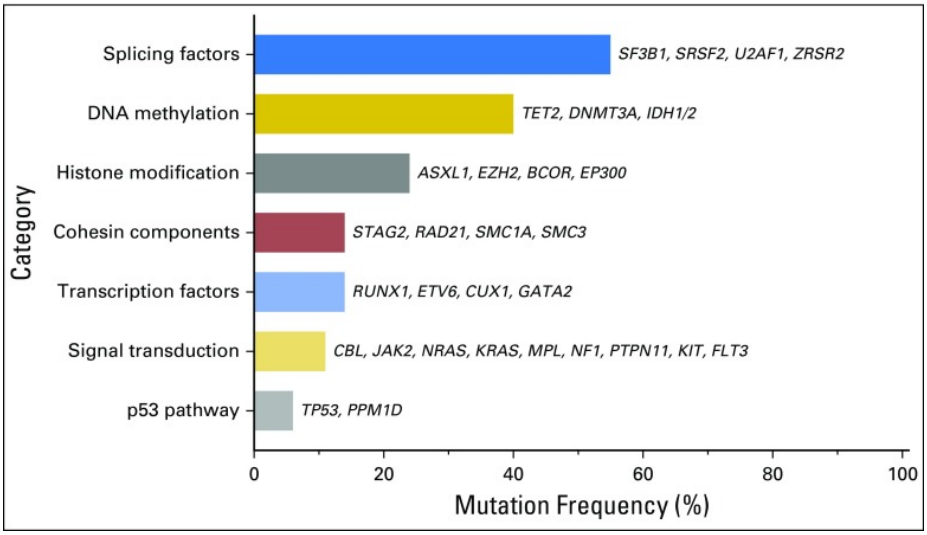


Figure 1-8. Commonly affected cellular processes in MDS ranked in order of mutation frequency. The most commonly affected process is splicing, with more than 50% of MDS cases containing a mutation. Genes mutated in each process are listed to the right of the respective bar. Data here include the two large cohorts by Papaemmanuil *et al*[115] and Haferlach *et al*[116]. Adapted from Kennedy and Ebert[117].

The most common cellular process affected in MDS is splicing, with more than 50% of MDS patients containing a mutation in *SF3B1*, *SRSF2*, *U2AF1* or *ZRSR2* (see Figure 1-8).

SF3B1 mutations are specifically enriched in MDS patients with ring sideroblast, while *SRSF2* mutations are more common in chronic myelomonocytic leukaemia (CMML), a rare chronic leukaemia with features of MDS (Figure 1-9). In the majority of cases, recognition of the 3' splice site in pre-mRNA processing is affected. Importantly, mutations found in *SF3B1*, *SRSF2* and *U2AF1* are all hotspot mutations. This suggests a change or gain-in-function rather than a loss of function in the mechanism of pathogenesis.

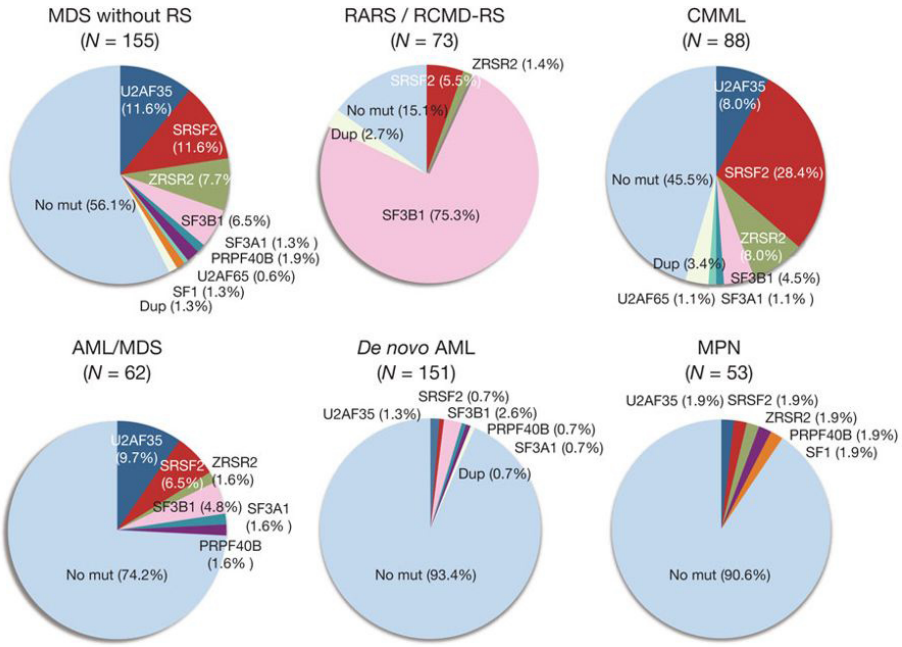


Figure 1-9. Distribution of mutations related to splicing in various haematologic malignancies. Note the significant enrichment of *SF3B1* mutation in RCMD-RS (Refractory Cytopenia Multilineage Dysplasia-Ring Sideroblast) and *SRSF2* in CMML. From Yoshida *et al*[118].

On closer examination of the mutation data from these studies, some interesting patterns have emerged. Mutations of genes in splicing pathway are mutually exclusive, suggesting that one mutation in this pathway is sufficient for pathogenesis[115, 118, 119]. Differentially expressed genes however showed little overlap among patients with different spliceosomal mutations, suggesting that more than one mutation could rather result in a functional disadvantage and becomes actively selected against[120]. Mutually exclusive mutations are also found across different cellular processes, such as between *ASXL1* and *DNMT3A* or *SF3B1*, and between *SRSF2* and *EZH2*. On the other hand, some mutations are more likely to co-occur, such as that of *ASXL1* and *U2AF1*[116].

As HSCs become dysplastic and ultimately transform into AML, it is thought that successive accrual of mutations provide the necessary survival advantage. The

genetic evolution in MDS was demonstrated using modeling based on variant allele frequency (VAF) from multiple cross-sectional samples, and from sequential sampling of the same individual patient with MDS[115, 121, 122]. These studies revealed that mutations in the splicing and DNA methylation pathways tend to occur earlier, while those in chromatin modification and signal transduction are often later mutations. Modeling suggested that both linear (where the same clone accumulates mutation sequentially) and branching (where several subclones with different spectrum of mutations co-exist) pattern exist in MDS, and the pattern is changed upon treatment. However these data have not been verified at the single cell level.

1.3.4 Acquired haemoglobin H disease

The first report on acquired haemoglobin H disease in the context of leukemia with no previous history of alpha thalassaemia was published in 1960[123]. In the majority of cases the haemoglobin H disease developed in the context of MDS, hence the term α-thalassaemia myelodysplastic syndrome (ATMDS, OMIM #300448)[124]. ATMDS is a rare subtype of MDS, with just over 120 cases in our in-house database since its first description.

Consistent with an alpha thalassaemia phenotype, peripheral blood from ATMDS patients showed microcytic hypochromic anaemia (average MCV 75.2fL versus 97.3 for MDS patients, and average MCH 22.4pg versus 32.3 for MDS patients, Figure 1-10)[124]. In contrast, MDS without acquired alpha thalassaemia is usually associated with macrocytic or normocytic anaemia. The percentage of MDS patients with macrocytosis (MCV>100fL) ranges from 39 to 45%[125-127],

while microcytosis (MCV<80fL) is only found in 4.7% of MDS patients in the Groupe Francophone des Myelodysplasies registry (n=2980)[128].

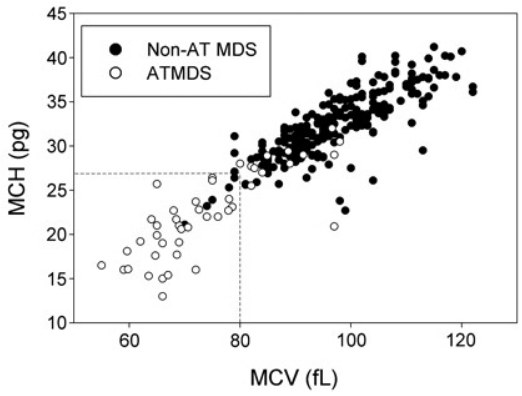


Figure 1-10. Scatter plot of MCV and MCH values in ATMDS patients versus MDS patients without alpha thalassaemia (Non-AT MDS). Adapted from Steensma *et al*[124].

In depth analyses of patients with acquired haemoglobin H disease showed that the alpha globin locus was intact, and translation was normal as the low globin chain synthesis matched the low α/β mRNA ratio[129, 130]. This suggests the underlying defect lies on decreased transcription of all four alpha globin genes, likely resulting from trans-acting factors. This notion was supported by experiment showing normalization of alpha globin synthesis when chromosome 16 from an ATMDS patient was transferred to mouse erythroleukemia cells[131].

1.3.5 ATRX mutations in ATMDS

Given the role of ATRX in alpha thalassaemia in ATR-X syndrome, it has been a strong candidate for causing acquired alpha thalassaemia in ATMDS. The first proof of somatic ATRX mutation involvement in ATMDS was discovered by cDNA microarray from an ATMDS patient. The expression of ATRX was downregulated in RNA isolated from neutrophils from this patient. It was discovered that the patient harboured a canonical donor splice site mutation in ATRX intron 1, which

was restricted to myeloid cells[132]. Since then further ATRX mutations have been described in ATMDS patients[133-137]. Overall, ATRX mutation is rare in MDS, with a mutation frequency of 0.2-0.8% in MDS[128].

The severity of alpha thalassaemia cannot be explained by ATRX mutation alone (Figure 1-11). An ATRX mutation (p.Leu192Phe) resulted in detectable Hb H in 50% of red blood cells (RBCs) in one ATMDS case, while the same mutation in a congenital ATR-X syndrome patient only saw 0.1% of RBCs with Hb H[133]. To put it in context, a similarly low percentage of Hb H (0.11%) was found in an atypical ATMDS case (ATMDS case #67, see Chapter 3) where the acquired alpha thalassaemia was caused by large heterozygous deletion involving both HBA genes (--/αα) rather than ATRX mutation[138]. Given the wide spectrum of mutations in MDS, other factors affecting alpha globin production may explain this discrepancy.

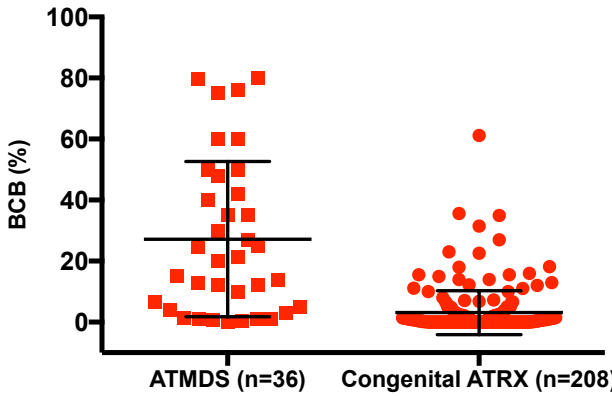


Figure 1-11. Comparison of Hb H inclusion between ATMDS and congenital ATR-X syndrome patients. The percentage of red blood cells with Hb H inclusions was determined by brilliant cresyl blue (BCB) stain. Average percentage of cells with Hb H inclusion was 27.2% in ATMDS, compared to 3.1% in congenital ATRX syndrome (P<0.0001, two tailed unpaired t test). Individual BCB% for ATMDS cases is available in Chapter 3

2 Materials and Methods

2.1 General Protocol

2.1.1 DNA extraction

DNA from patients or cell lines were extracted from a minimum of 150,000 cells using PureLink genomic DNA extraction kit (Invitrogen). DNA was quantified using Qubit DNA BR kit (Thermo Fisher Scientific). Quality of DNA was assessed using Nanodrop (Thermo Fisher Scientific).

2.1.2 General PCR and Sanger sequencing

PCR reactions were prepared using Platinium HiFi Supermix (Thermo Fisher Scientific):

Components	**Volume (µL)**
Forward primer (20µM)	2
Reverse primer (20µM)	2
Supermix	12
DNA (30ng/µL)	2

Thermocycling condition is listed as below. Touchdown method was used where the first cycle started with a higher annealing temperature (Ta) and reduced by 1°C per cycle for the first 5 cycles to increase the specificity of PCR product[139].

Step	Temperature	Time
1	94°C	2min
2	94°C	30s
3	Ta+4°C (-1°C/cycle)	30s
4	68°C	1kb/min
5	Repeat steps 2-4, 4x	
6	94°C	30s
7	Ta	30s
8	68°C	1kb/min
9	Repeat steps 6-7, 30x	
10	68°C	7min
11	4°C	indefinite

PCR products were visualised by gel electrophoresis. For Sanger sequencing, primers and unincorporated nucleotides were removed enzymatically using Exo-SAP IT express according to the manufacturer's protocol. Purified PCR products were sent for Sanger sequencing externally. Sequence traces were visualised using 4Peaks version 1.7.2.

2.1.3 RNA extraction

RNA was extracted using RNeasy Mini Kit (Qiagen) according to manufacturer's protocol. DNA removal was achieved by on-column DNase digestion using RNase-Free DNase set (Qiagen). Purified RNA was eluted in DPEC-treated water and quantified using Qubit RNA BR kit.

2.1.4 cDNA synthesis

First-strand cDNA synthesis was performed using SuperScript™ III First-Strand Synthesis SuperMix (ThermoFisher Scientific 18080400). For RNA extracted from FACS sorted samples, 50ng of input RNA (or 8µL if the RNA concentration was below detection limit of the Qubit RNA BR kit) was used. For other RNA samples up to 1µg input RNA was used.

2.1.5 Quantitative PCR

Expression levels of globin genes (*HBA*, *HBB*, *HBG*, *HBZ*, *HBE*) were assessed using TaqMan Real-Time PCR assays. TaqMan probes (Thermo Fisher Scientific) for the relevant genes are listed below. The ribosomal protein L13a (*RPL13A*) was used as a reference gene[140-142], as this gene had been shown empirically in our laboratory as one of the most stably expressed housekeeping genes during erythroid differentiation (Sachith Mettananda, unpublished data). It should however be noted that expression of most housekeeping genes including *RPL13A* do decrease as cells undergo erythroid differentiation (Table 2-1).

	Day 9	**Day 13**	**Day 16**
ATMDS patient	24.87	25.32	27.78
Normal control	25.10	26.50	29.98*
	Day 1	**Day 4**	**Day 10**
HUDEP wildtype	23.01	23.22	24.59

Table 2-1. Mean C_T values for RPL13A during erythroid differentiation of $CD34^+$ cells of an ATMDS patient and normal control (top) and wildtype HUDEP-2 cells (bottom). * the low C_T value is due to excessive cell death at terminal stage of differentiation.

TaqMan probes used:

Gene	**Assay ID**
HBZ	Hs00923579_m1
HBA2/HBA1	Hs00361191_g1
HBE	Hs00362216_m1
HBG2/HBG1	Hs00361131_g1
HBB	Hs00747223_g1
HBM	Hs01392876_g1
RPL13A	Hs04194366_g1

Reaction setup:

Component	**Volume (µL)**
TaqMan Universal Master Mix	5
Probe	0.5
ddH_2O	3.5
cDNA	1.0
Total volume	10.0

RT-PCR reactions were performed in StepOne PCR system (Applied Biosystems). Relative expression to RPL13A was calculated using the C_T values obtained from StepOne software 2.3. Results were average of triplicate reactions.

2.1.6 Western Blot

For Western blot to detect ATRX and PHF6 protein, $2x10^6$ HUDEP cells were washed in PBS and centrifuged to remove supernatant. Pellets were snap frozen and stored in -80°C. Pellets were lysed in 100µL of high salt RIPA lysis buffer with protease inhibitor (Roche). DNA was sheared by syringing the lysate through a 25G needle 20 times. The lysate was centrifuged at 20,000g for 10min at 4°C. Supernatant was recovered and stored at -80°C until electrophoresis.

For electrophoresis, equal volume of lysate and 2x loading buffer with 0.2M DTT were mixed and denatured at 95°C for 10min. Electrophoresis was performed in a NuPAGE 4-12% Bis-Tris gel (Thermo Fisher Scientific) in a Mini-PROTEAN system (Bio-Rad). The gel was transferred to a membrane overnight and blocked with TBS-T (Tris buffered saline+0.1% Tween-20) with 5% skimmed milk powder. Primary antibody staining with anti-ATRX (MABE1798, Merck Millipore or ab97508, Abcam), or anti-PHF6 (NB100-68262, Novus Biologicals) antibody was performed in TBS-T with 2.5% skim milk powder. Anti-H3 (ab1791, Abcam) antibody was used as loading control. After primary antibody staining (1hr)

membrane was washed in TBS-T three times before staining with secondary antibody for 1hr. After secondary staining the membrane was washed again in TBS-T before exposure on X-ray film.

2.2 Assay for Transposase-Accessible Chromatin with high-throughput sequencing (ATAC-seq)

For each ATAC-seq sample 75,000 cells were required. Two technical replicates were taken with each experiment. Cells were centrifuged at 400g for 5min at 4°C in a 1.5mL centrifuge tube (N.B. do not use Eppendorf LoBind tube as pellet formation would be poor). Remove supernatant.

Resuspend pellet gently in 50µL of cold ATAC lysis buffer. Centrifuge at 500g for 10min at 4°C. Remove supernatant and keep pellet on ice.

ATAC lysis buffer:

Component	**Concentration**
Tris-HCl	10mM
NaCl	10mM
$MgCl_2$	3mM
IGPAL CA-630	0.1% (v/v)

Resuspend pellet gently in 50µL of transposition reaction mix. Incubate at 37°C for 30min.

Transposition reaction mix (adapted from Illumina Nextera DNA Library Prep Kit FC-121-1030)

Component	**Volume (µL) per reaction**
2x Tagment DNA (TD) buffer	25.0
Tagment DNA enzyme (TDE1)	2.5
Nuclease-free water	22.5

Extract transposed DNA using MinElute PCR Purification kit (Qiagen). Elute transposed DNA in 23μL of Elution Buffer in DNA LoBind tube (Eppendorf). Sample can be stored in -20°C until PCR amplification and indexing.

For PCR amplification and indexing, combine the following:

PCR amplification and indexing

Component	**Volume (μL)**
NEBNext High-Fidelity 2x PCR Master Mix (New England BioLabs M0541)	25.0
PCR primer 1 (25μM)*	2.5
PCR primer 2 (25μM)*	2.5
Nuclease-free water	10.0
Transposed DNA	10.0

*PCR primer 1 and 2 are customised primers. Primer 2 contains the sequence for indexing.

Perform thermocycling reactions as below:

Step	**Temperature**	**time**
1	72°C	5min
2	98°C	30s
3	98°C	10s
4	63°C	30s
5	72°C	1min
6	Repeat steps 3-5, 9x	
7	4°C	indefinite

Purify the amplified and indexed ATAC library using QIAquick PCR purification kit (Qiagen). Elute library in 20μL elution buffer. Purified libraries were quantified using real-time PCR (KAPA Library Quantification Kit, Kapa Biosystems). Libraries were adjusted to 4nM and pooled for sequencing using a NextSeq 500/550 High Output v2 75 cycle kit (Illumina) in a NextSeq 500 sequencer (Illumina).

2.2.1 ATAC-seq data analysis

Data generated from sequencing (in FASTQ format) were processed using in-house pipeline (Jelena Telenius, unpublished data. Available at **http://userweb.molbiol.ox.ac.uk/public/telenius/PipeSite.html**). Briefly, the sequence reads were trimmed and quality assessed using FastQC. Reads were then mapped against hg19 reference genome using BOWTIE1. The mapped reads were filtered for PCR duplicates using SAMTOOL RMDUP. Data visualisation files were generated using BEDTOOLS and evaluated on UCSC genome browser.

To allow for comparison between samples, peaks were normalised against total aligned reads minus reads aligned to mitochondrial genome.

2.2.2 ATAC-seq differential peak analysis

Genome-wide peak calling for ATAC-seq was performed using MACS2 (Model-based Analysis of ChIP-Seq 2)[143]. Peaks between 2 to 50-fold change from background were called. For differential peak calling, the "(sample name)_peaks.narrowPeak" file and the corresponding BAM file from each sample were uploaded to the Galaxy web platform using the public server at usegalaxy.org. The former file contains peak locations, peaks summit, fold change, p value and q value generated from MACS2.

Differential peak calling was performed using the Bioconductor package DiffBind[144] on the Galaxy web platform (Galaxy version 2.6.6.4)[145]. For this project, peak calling data from three ATMDS patients (MK, ECT, NQ) and four normal controls (LC4, LC5, LC6 and LC8) were compared. Peaks were extended

250bp up and downstream of the summit during the analysis. The multiple testing corrected false discovery rate (FDR) p value threshold of 0.05 was used.

Transcription factor binding motif enrichment analysis for differential sites identified in DiffBind was performed using MEME-ChIP[146, 147]. Nucleotide sequences of differential sites in FASTA format were uploaded to MEME-ChIP web service (**http://meme-suite.org/tools/meme-chip**). Motifs with statistical significant enrichment relative to a random model based on frequencies of nucleotides in the differential sites were returned.

2.3 Ion Torrent sequencing

A custom AmpliSeq sequencing panel (IAD30921) that fully covers the coding sequences of *ATRX*, *ASXL1*, *U2AF1* and *PHF6* was previously designed (Andrew Campbell, unpublished data). The panel consists of two sets of primer pools for each sample. Library preparation was modified from manufacturer's protocol. The two sets were PCR amplified separately in 10μL reactions in the first step of library preparation, using 8ng of genomic DNA or whole genome amplified (WGA) DNA as input. The quantity of input DNA was measured by Qubit (dsDNA BR kit or for WGA DNA, ssDNA kit).

After initial PCR, the two sets of PCR products were pooled together (total volume 20μL) for downstream steps. The PCR products were partially digested and indexed by ligation using Ion Xpress barcode adaptors. The libraries were purified using Agencourt AMPure beads. The purified libraries were further amplified and underwent size selection using a two-round purification step with

Agencourt AMPure beads. Quantity of purified libraries were assessed using TapeStation D1000 High Sensitivity tape. Libraries were adjusted to 100pM and pooled for emulsion PCR using Ion PGM Template OT 200 Kit in the Ion OneTouch 2 System. Product was enriched using the OneTouch ES system.

After emulsion PCR, sequencing was performed on an Ion PGM system using the Ion PGM Sequencing 200 Kit v2 and Ion 316 or 318 Chip v2. Data analysis and variant calling were processed via the Variant Caller plugin in the Torrent Suite software version 4.2.1. VCF files were annotated using wANNOVAR[148]. Mutations detected on Ion Torrent were confirmed by Sanger sequencing.

2.4 Haematological staining

Microscope slides of cells during erythroid differentiation were prepared by centrifugation using a Cytospin (Thermo Fisher Scientific). In undifferentiated cells 80,000 cells were centrifuged at 400rpm for 4min. For differentiated cells 250,000 cells were centrifuged at 200rpm for 3min. Modified Wright stain were performed using a Hematek Slide Stainer (Bayer HealthCare) and imaged using an Infinity microscope camera (Lumenera).

Brilliant cresyl blue (BCB) stain was performed on whole blood by Dr. Chris Fisher. The percentage of red blood cells with Hb H inclusion bodies was counted after overnight staining.

2.5 Identification of mutations in ATMDS using a myeloid panel

Mutations in ATMDS patients were identified using a TruSeq Custom Amplicon (TSCA) myeloid sequencing panel (Illumina) on an Illumina MiSeq platform. The panel consists of 341 amplicons covering approximately 56kb over 32 genes. A list of genes and exons covered is available in Table 2-2. The minimum amount of DNA input was 50ng as measured by Qubit dsDNA BR assay kit (ThermoFisher Scientific). Library preparation and sequencing were performed at the Oxford Molecular Diagnostics Centre. Primary analysis was performed on the MiSeq Reporter software (Illumina) – custom banded Smith-Waterman algorithm was used for alignment and Genome Analysis Tool Kit (GATK) was used for variant calling. Variants were annotated using Variant Studio software (Illumina). Coverage was determined using in-house app CoverMi which calculated the coverage at every base within the design. Variants with quality score <100 (based on GATK) and variant allele frequency (VAF) <5% were filtered out. Variants without consequence (i.e. synonymous, non-coding or non-splice site variants) were not assessed. Variants previously reported in haematological malignancies in literature or COSMIC database, novel frameshift or nonsense mutations were highlighted as variants of predicted clinical significance.

Gene	Exons Covered	Gene	Exons Covered
ASXL1	12	**KRAS**	2-3
ATRX	8-10, 17-31	**MPL**	10
CBL	8-9	**NPM1**	11
CBLB	9-10	**NRAS**	2-3
CBLC	9-10	**PDGFRA**	12, 14, 18
CEBPA	1 (partially covered)	**PHF6**	2-10
CSF3R	14-17	**PTEN**	5-7
DNMT3a	23	**RUNX1**	3-8
ETV6	1-8	**SETBP1**	4
EZH2	2-20	**SF3B1**	14-15
FLT 3	14-15, 20	**SRSF2**	1
HRAS	2-3	**TET**	3-11
IDH1	4	**TP53**	4-9
IDH2	4	**U2AF1**	2 & 6
JAK2	12 & 14	**WT1**	7 & 9
KIT	2, 8-11, 13, 17	**ZRSR2**	1-11

Table 2-2. List of genes and exons covered by TSCA myeloid panel.

2.6 SNP array analysis

Single nucleotide polymorphism microarray (SNP array) was performed using Illumina HumanCytoSNP-12 BeadChip at the Wellcome Trust Centre for Human Genetics. Copy number variation calling was performed using KaryoStudio v1.4.

2.7 Peripheral blood mononuclear cell and $CD34^+$ haemopoietic stem cell separation

Blood samples were collected from ATMDS patients as close as possible prior to a due blood transfusion, or at least 2 weeks after a blood transfusion. This is to avoid contamination from donor blood. 30-100mL of blood were collected and processed within 48 hours.

Blood samples were first separated by gradient centrifugation in a LeucoSep tube to isolate peripheral blood mononuclear cells (PBMCs):

1. 16mL of Histopaque-1077 (Sigma) was added to each LeucoSep tube (Greiner) and centrifuged at 1,000g for 30s at room temperature.

2. Blood was diluted 1:2 with PBS and 35mL of diluted blood was added to each LeucoSep tube.

3. Leucosep tube was centrifuged at 1,000g for 10min at room temperature in a swinging-bucket rotor with no brakes.

4. After centrifugation, red cells and granulocytes lie below Histopaque, whereas the mononuclear cell layer lies above Histopaque. In ATMDS patients though, it is noted that the red cell separation was not completely clean and red cells can be seen in the Histopaque layer Figure 2-1.

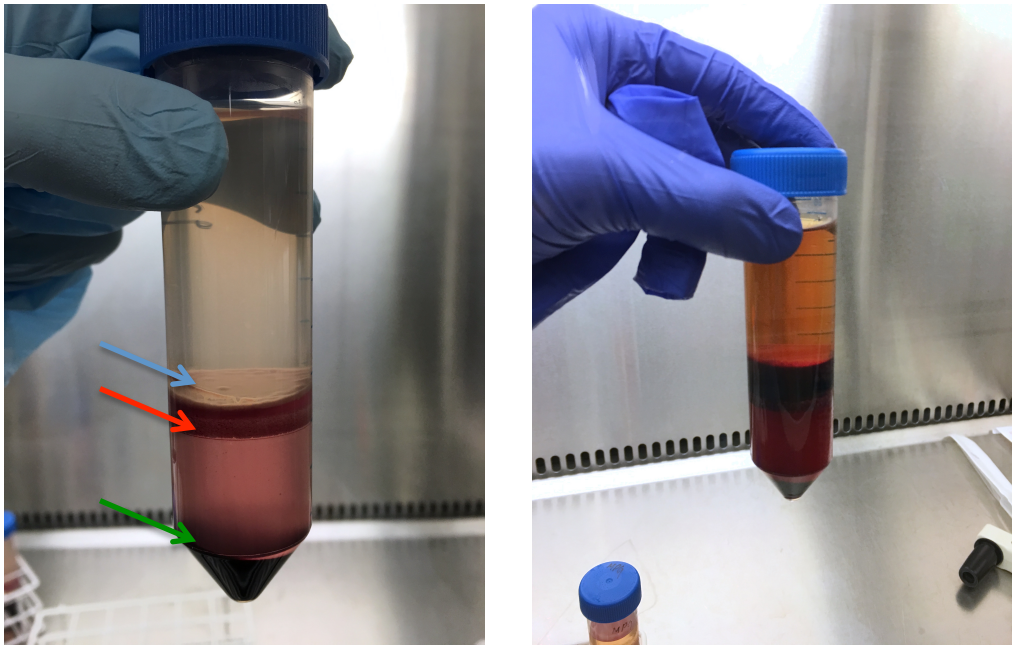


Figure 2-1. Gradient centrifugation using LeucoSep tube. Left – separation of PBMCs in blood from normal individuals. The layer of PBMCs (blue arrow) can be seen above the LeucoSep filter (red arrow), red blood cells and granulocytes (green arrow) lie below the layer of Histopaque-1077; Right – separation of PBMCs in ATMDS patients. Note the Histopaque layer is contaminated with red blood cells.

5. PBMCs were carefully removed using a Pasteur pipette. Cells were washed twice by adding 10mL of PBS and centrifuged at 300g for 10min at room temperature in the first wash, and 200g for 10min in the second wash. PBMCs be cryopreserved in 90% fetal bovine serum and 10% DMSO, or proceed directly to MACS.

2.8 CD34$^+$ haemopoietic stem cell (HSC) isolation by MACS:

6. Following PBMC separation, The washed pellet was resuspended in 900µL of MACS buffer (PBS with 0.5% bovine serum albumin and 2mM EDTA). 300µL of FcR blocking reagent (Miltenyi Biotec) and 300µL of CD34 MicroBeads UltraPure (Miltenyi Biotec) were added to the resuspended cells and incubate for 30min at 2-8°C.

7. After incubation, labeled cells were washed in 10mL of MACS buffer and centrifuged at 300g for 10min at room temperature. Cells were resuspended in 500µL of MACS buffer.

8. Magnetic separation of labeled cells were performed using LS column (Miltenyi Biotec) according to the manufacturer's protocol. After elultion, cells were counted and cryopreserved in 0.2 x 10^6 aliquots (in 500µL 90% fetal bovine serum and 10% DMSO).

2.9 CD34$^+$ single cell expansion

PBMCs were labeled using antibodies against CD34, CD3 (a T cell marker) and CD19 (a B cell marker). 15 x10^6 PBMCs were incubated in 150µL of antibody mix in PBS at 4°C for 30min. Unstained and single stain controls for each of the three antibodies were prepared for compensation.

Antibody mix	**Dilution (in PBS)**
CD34-APC (clone 8G12, BD Bioscience)	1:160
CD3-FITC (clone SK7, BD Bioscience)	1:80
CD19-PE (clone SJ25C1, BD Bioscience)	1:40
FcR block (Miltenyi Biotec)	1:10

After incubation, cells were washed in 10mL of PBS and centrifuged at 300g for 5min. Cells were resuspended in PBS with 5% FBS for single cell FACS. DAPI (Cell Signaling Technology) was added to the buffer to the final concentration of

1µg/mL before cell sorting for live/dead staining. FACS sorting was performed using FACSAria Fusion flow cytometer (BD Biosciences), with assistance from the WIMM FACS facility. Single $CD34^+$, $CD3^-$ and $CD19^-$ cells were sorted into 60-well sterile Terasaki plates (Thermo Fisher Scientific), i.e. each well contains a single cell in 20µL of expansion media.

Expansion media

Component	**Final Concentration**
IMDM	-
FBS	10%
L-Glutamine	1x
2-Mercaptoethanol	50µM
Penicillin/Streptomycin	1x
SCF	50ng/mL
Flt-3 ligand	50ng/mL
TPO	10ng/mL
IL-3	10ng/mL
IL-6	10ng/mL
GSM-CSF	5ng/mL
EPO	1IU/mL
G-CSF	10ng/mL

Due to the low volume, Terasaki plates were incubated at 37°C inside a humidity box with sterile water to prevent drying out. Plates were assessed on day 7 for clonal expansion. Wells with over 30 cells were retrieved for whole genome amplification.

Whole genome amplification (WGA) was performed directly from retrieved cells using Illustra Single Cell GenomiPhi DNA Amplification Kit (GE Healthcare). WGA DNA was quantified using Qubit ssDNA assay kit (Thermo Fisher).

2.10 CD34$^+$ HSC erythroid expansion and differentiation

CD34$^+$ HSCs were expanded and differentiated in the erythroid pathway using a three-phase system[149]. The media consisted of a base media supplemented with different combination of cytokines depending on the day of differentiation. Cells in cryovials were thawed in 37°C water bath until the last crystal had thawed. Cells were washed in 10mL of PBS and centrifuged at 300g for 5min. Supernatant was removed and cells resuspended in Phase I media at $0.5x10^6$/mL in a 24-well plate. Cell density was checked every second day and media added to maintain density at $0.5x10^6$/mL if needed. On day 7 and day 12 media was replaced with Phase II and III media, respectively.

Base media	**Final concentration**
IMDM	-
Iron saturated human holo-transferrin (Sigma T0665)	0.2mg/mL
Human insulin	0.01mg/mL
Heparin	3IU/mL
Human AB plasma	3%
FBS	2%
Penicillin/Streptomycin	1x
EPO	3IU/mL

Phase I supplement (Day 0-7)	**Final concentration**
IL-3	1ng/mL
SCF	10ng/mL

Phase II supplement (Day 8-11)	**Final concentration**
SCF	10ng/mL

Phase III supplement (Day 12-19)	**Final concentration**
Iron saturated human holo-transferrin	0.3mg/mL

2.11 CD34^{+} HSC FACS analysis and FACS sorting

Differentiation of CD34^{+} HSCs was assessed by FACS analysis, using CD34 and the two erythroid markers CD71 and CD235a. 0.2x10^{6} cells were centrifuged at 300g for 5min and resuspended in 50µL of antibody mix. Unstained and single stain controls were prepared for each of the three antibodies. Cells were incubated with antibody mix for 30min at 4°C. Stained cells were washed in 5mL PBS and centrifuged at 300g for 5min, then resuspended in PBS + 5% FBS for FACS. FACS analysis was performed using Attune NxT flow cytometer (Thermo Fisher Scientific).

Antibody mix	**Dilution (in PBS)**
CD34-APC (Miltenyi Biotec 130-090-954)	1:50
CD71-FITC (BD Bioscience 555536)	1:50
CD235a-PE (BD Bioscience 340947)	1:50
FcR block (Miltenyi Biotec)	1:10

For FACS sorting, 20µL of the same antibody mix was used per 1x10^{6} cells. Stained cells were FACS sorted using a Sony SH800 cell sorter into 15mL centrifuge tube. After FACS sorting, cells were collected by centrifugation at 400g for 10min. The pellet was resuspended in 1mL of PBS and counted before proceeding to downstream procedures (e.g. preparation for ATAC-seq and RNA extraction).

2.12 HUDEP cell expansion

HUDEP-2 cells were obtained from Cell Engineering Division, RIKEN BioResource Center, Tsukuba, Japan[150]. HUDEP-2 cells were maintained in expansion media as below.

Component	Final concentration
StemSpan SFEM	-
GlutaMAX	1x
Penicillin/Streptomycin	1x
SCF	100ng/mL
EPO	3IU/mL
Dexamethasone	10^{-6}M
Doxycycline	2µg/mL*

*note the doxycycline concentration here is double the concentration of the original publication describing HUDEP cell lines. The higher concentration is essential during expansion after single cell FACS sorting.

A complete change of media is required every two days to maintain the doxycycline concentration. Cell density was maintained between 0.5-2x10^6/mL.

2.13 CRISPR-Cas9 genome editing of HUDEP

Genome editing of HUDEP cells was achieved via nucleofection and single cell FACS sort. Compared to other strategies such as limiting dilution and selection by antibiotic resistance, single cell FACS sorting strategy eliminated concerns of mixed populations in the edited clones.

The plasmid used here is the pX458 plasmid from the Zhang lab, Massachusetts Institute of Technology[151] (Figure 2-2). The plasmid co-expresses *Streptococcus pyogenes* Cas9 (SpCas9) nuclease and enhanced green fluorescence protein (EGFP) via a T2A self-cleaving peptide, along with a single guide RNA (sgRNA) sequence for targeting. The plasmid also contains ampicillin resistance gene for selection.



Figure 2-2. Plasmid map of pX458. sgRNA is cloned 3' to the human U6 promoter (cyan). Locations for Cas9 (dark red), T2A peptide (pink), EGFP (bright green) and ampicillin resistance (AmpR, light green) are shown above (http://www.addgene.org/48138/).

2.13.1 Plasmid and oligonucleotide design

sgRNAs were designed using the online CRISPR design tool (**http://crispr.mit.edu/**). Sequence surrounding the targeted region was submitted and sgRNAs returned with score above 50 were considered. The final choice on sgRNA depended on the location of sgRNA (e.g. for knockout the sgRNA will need to be within an exon) and the specificity of sgRNA, which was checked by aligning the sgRNA to the human genome using BLAST (Basic Local Alignment Search Tool). The sgRNA was modified to contain a guanine at the 5' end if not already present in the sequence, as U6 promoter requires guanine at

the start of a transcript[152]. Typically sgRNA contained 18-20bp of homology to the target sequence.

sgRNAs were ordered as a pair (top and bottom) of complementary oligonucleotides. Nucleotide sequences "CACC" and "AAAC" were added to the 5' end of the top and bottom sgRNA respectively for cloning using BbsI restriction enzyme. An example for sgRNA targeting *ATRX* exon 16 (NG_008838.2) is shown below in Figure 2-3. A full list of sgRNA used in this project is available in Appendix.

sgRNA sequence:

ATRX-sgRNA-Ex16-TOP	5' caccGTCCAATAACAACCAAGT 3'
ATRX-sgRNA-Ex16-BOTTOM	5' aaacACTTGGTTGTTATTGGAC 3'

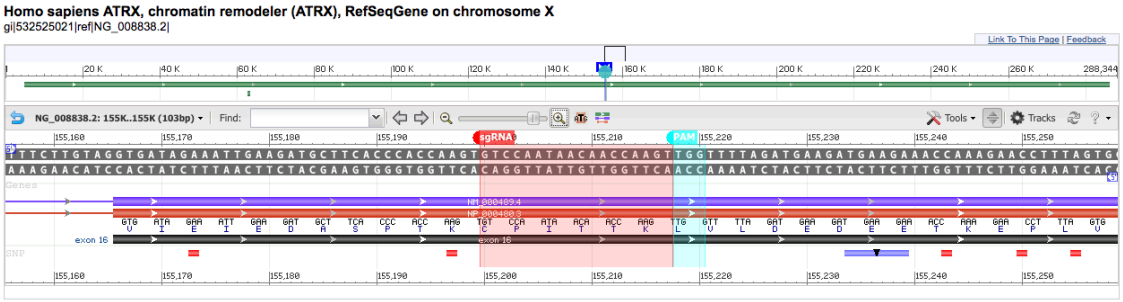


Figure 2-3. sgRNA targeting exon 16 of ATRX. The sgRNA sequence is highlighted in red, with the PAM (protospacer-adjacent motif) "NGG" sequence highlighted in cyan.

sgRNAs were cloned onto pX458 using the protocol provided by Ran *et al*[151]. sgRNAs were first phosphorylated and annealed in the following reaction.

Component	Volume (µL)
sgRNA top (100µM)	1.0
sgRNA bottom (100µM)	1.0
T4 ligation buffer (with ATP), 10x (New England BioLabs, B0202S)	1.0
T4 polynucleotide kinase (New England BioLabs, M0201S)	1.0
ddH_2O	6.0
Total	10.0

The reaction mixture was incubated in a thermocycler at 37°C for 30min, 95°C for 5min then ramp down to 25°C at 5°C/min. The phosphorylated and annealed oligos were diluted 1:200 with ddH_2O.

The diluted oligos were cloned onto pX458 in the following reaction:

Components	Volume (µL)
pX458 plasmid	add 100ng
Diluted oligo	2.0
Tango buffer, 10x (Fermentas, BY5)	2.0
DTT, 10 mM	1.0
ATP, 10 mM	1.0
FastDigest *Bbs*I (Fermentas, FD1014)	1.0
T7 DNA ligase (Enzymatics, L602L)	0.5
ddH_2O	to 20.0
Total	20.0

The reaction mixture was incubated in a thermocycler under the following condition:

Step	Temperature	Time
1	37°C	5min
2	21°C	5min
3	Repeat Step 1	5x

The ligation reaction product was treated with PlasmidSafe exonuclease (Lucigen) to digest residual linearalised DNA.

Component	Volume (µL)
Ligation reaction product	11.0
PlasmidSafe buffer, 10x	1.5
ATP, 10 mM	1.5
PlasmidSafe exonuclease	1.0
Total	15.0

The reaction mixture was incubated at 37°C for 30min, then 70°C for 30min. Exonuclease-treated plasmid was transformed into competent 5α Escherichia coli (New England BioLabs) following the manufacturer's protocol and grown overnight on a LB agar plate with ampicillin (100 µg/mL).

Colonies were picked and checked for presence of sgRNA by PCR. Colonies containing the correct sgRNA were expanded in overnight culture in the presence of ampicillin (100µg/mL). Plasmids were extracted using Plasmid Midi or Maxi kit (Qiagen). Plasmids were resuspended in sterile water in a sterile environment to 1-2µg/µL for use in nucleofection.

2.13.2 Surveyor assay

The efficacy of sgRNAs used for genome editing was assessed by surveyor assay. HEK293 cells were transfected using Lipofectamine 3000 (Thermo Fisher Scientific) according to the manufacturer's protocol. Transfection efficiency was assessed by direct observation of green fluorescence signal under microscope. A transfection efficiency of 60% or above was considered satisfactory. Transfected cells were extracted for DNA and proceeded to surveyor assay using Surveyor Mutation Detection Kit (IDT). In cases where sgRNA resulted in mutation in the targeted region, mixing PCR products containing the target region from transfected and untransfected cells would lead to formation of heteroduplex. The

heteroduplex would be cleaved by the surveyor nuclease and become visible as additional bands in gel electrophoresis (Figure 2-4).

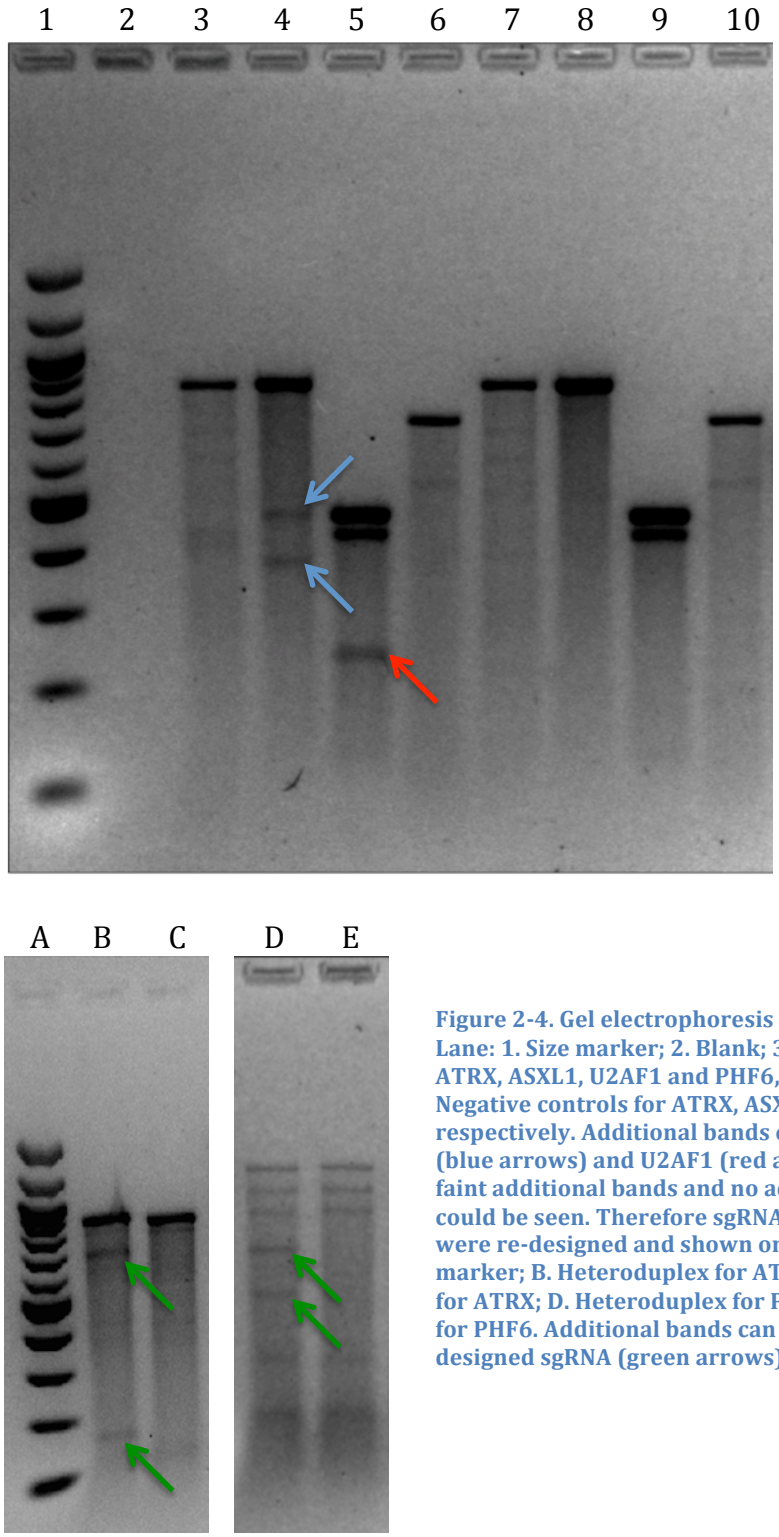


Figure 2-4. Gel electrophoresis of surveyor assay. Top: Lane: 1. Size marker; 2. Blank; 3-6. Heteroduplex for ATRX, ASXL1, U2AF1 and PHF6, respectively; 7-10. Negative controls for ATRX, ASXL1, U2AF1 and PHF6, respectively. Additional bands can be seen in ASXL1 (blue arrows) and U2AF1 (red arrow). ATRX had very faint additional bands and no additional bands on PHF6 could be seen. Therefore sgRNA for these two genes were re-designed and shown on the left. Lane: A. Size marker; B. Heteroduplex for ATRX; C. Negative control for ATRX; D. Heteroduplex for PHF6; E. Negative control for PHF6. Additional bands can be seen in the re-designed sgRNA (green arrows).

2.13.3 Homology-directed repair using single-stranded oligodeoxynucleotide

Homology-directed repair (HDR) is required to introduce a specific nucleotide change or to repair a previously introduced deletion/insertion. This is delivered via co-transfection with a single stranded oligodeoxynucleotide (ssODN) to act as a template. ssODNs were 100-120bp in length with the target region located between 30-70bp from the 5' end. A list of ssODNs used in this project can be found in Appendix. ssODNs were ordered as desalted oligonucleotides without further purification and reconstituted to 200μM in sterile water in a sterile environment. For each nucleofection 1μL of each ssODN was added to the nucleofector solution. Using less ssODN than specified here would result in lower HDR efficiency. Up to three different ssODNs could be used in the same nucleofection.

In agreement with published report, ssODNs complementary to the released strand (i.e. on the same strand as the "TOP" sgRNA used in this project) are more efficient in introducing HDR[153]. The same report also suggested that offsetting the location of target region so that the PAM site of the associated sgRNA is 36bp from the 5' end of ssODN also improved efficiency. In this project however, the three ssODNs used for ATRX repair (ATRX-repairWT-ssODN1 to -ssODN3) with different offsets achieved similar rate of HDR (4, 7 and 5 clones out of a total of 38 clones, respectively).

2.13.4 Nucleofection, single cell FACS sorting and expansion

Genome editing of HUDEP was based on nucleofection using Lonza human CD34 cell nucleofector kit on a Nucleofector 2b machine (Lonza). $1.5x10^6$ cells in

expansion phase were used per nucleofection. Cells were washed in 10mL of PBS and centrifuged at 300g for 5min. This PBS wash step is important to remove antibiotics present in the media that would negatively affect nucleofection. Cells were resuspended in 100µL of supplemented nucleofector solution containing 2µg of plasmid and, if required, ssODN (1µL at 200µM for each oligonucleotide, up to three ssODNs per transfection). The total volume of plasmid and ssODNs must be under 10µL (10% of total volume).

Resuspended cells were transferred to a nucleofection curvette and immediately proceeded to nucleofection on the device, using the programme U-08. After nucleofection the cells were immediately returned to a 24-well plate in 1mL of pre-warmed fresh media, using the Pasteur pipette provided in the kit. If ssODNs were used, the small molecule RAD51-Stimulatory compound 1 (RS-1, Sigma-Aldrich) was added to the media to a final concentration of 0.75µM. RS-1 promotes homology directed repair (HDR) by stimulating RAD51 activity [154-156]. Cells were incubated for two days before media was changed. It is normal to expect significant cell death as a result of nucleofection. No RS-1 was added to the media from this point onwards. The non-homologous end joining inhibitor SCR7[157, 158] was also trialed but SCR7 did not lead to increase HDR rate in HUDEP cells.

Single cell FACS sorting was performed on a FACSAria Fusion flow cytometer 72 hours post-transfection to select GFP-positive, DAPI-negative cells. It is important to wait for 72 hours before FACS sorting as this improves the survival rate in a Terasaki plate – the hypothesis being that the longer interval allows

those cells that would have otherwise not survived to apoptose and not be selected during FACS sorting. Single cells were sorted into 60-well sterile Terasaki plates containing 20µL of expansion media. The plates were placed inside a humidity box with sterile water for 5 days in the incubator.

On day 5, wells with more than 20 cells were transferred to a 96-well plate containing 90µL of expansion media. On day 7, 100µL of expansion media was added. From then on, 100µL of media was carefully removed from the top of the well and replaced with 100µL of fresh media every second day. When cells were 70% confluent in the well they were split into two wells in a 96-well plate and maintained as above by media change until both wells were 70% confluent. Cells from both wells were then pooled and transferred to a 24-well plate with 500µL media. From this point onwards cells were expanded routinely with full media change by centrifugation every second day. When the number of cells allowed sample was taken for DNA extraction for confirmation of genome editing. Remaining cells were cryopreserved. In general the duration from nucleofection to cryopreservation is 25-30 days.

2.14 Erythroid differentiation of HUDEP cells

HUDEP cells in expansion phase were already $CD34^-$ $CD71^+$ and $CD235a^+$[150]. Hence these markers could not be used to track erythroid differentiation. Two markers, α4-integrin (CD49d) and band 3 (CD233), have been shown to track terminal erythroid differentiation[110] (Figure 2-5) and were used to track HUDEP erythroid differentiation and FACS sort different populations of HUDEP cells.

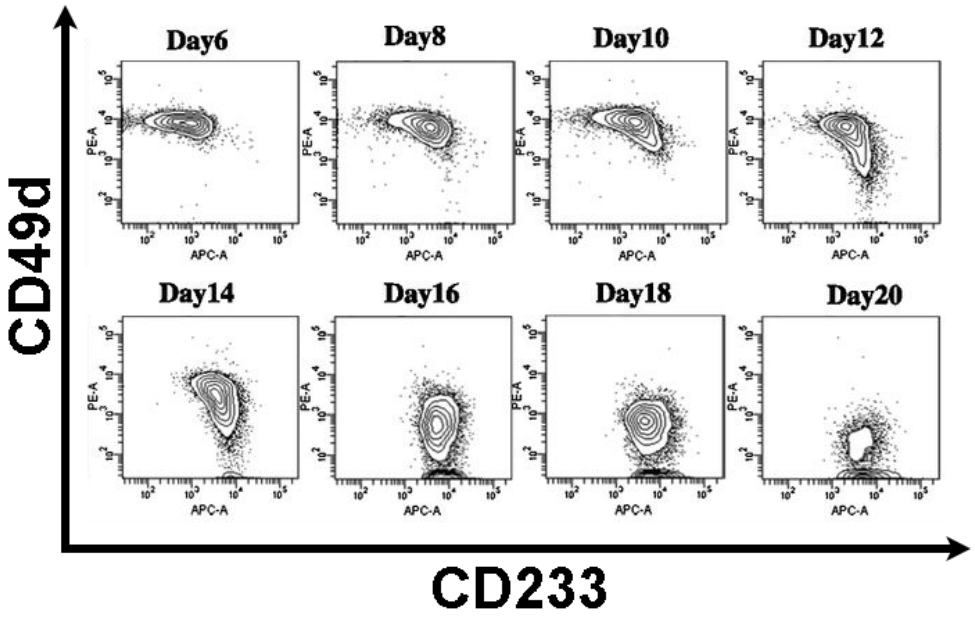


Figure 2-5. FACS analysis of α4-integrin (CD49d) and band 3 (CD233) during terminal erythroid differentiation of CD235a+ human erythroid cells. α4-integrin expression decreased gradually from Day 12 while expression of band 3 increased as cells matured. Adapted from Hu *et al*[110].

Clonal variability on the stage of differentiation was observed between HUDEP clones (Figure 2-6). In order to allow comparable erythroid differentiation between HUDEP clones, depletion of $CD233^{+}$ cells by MACS was performed.

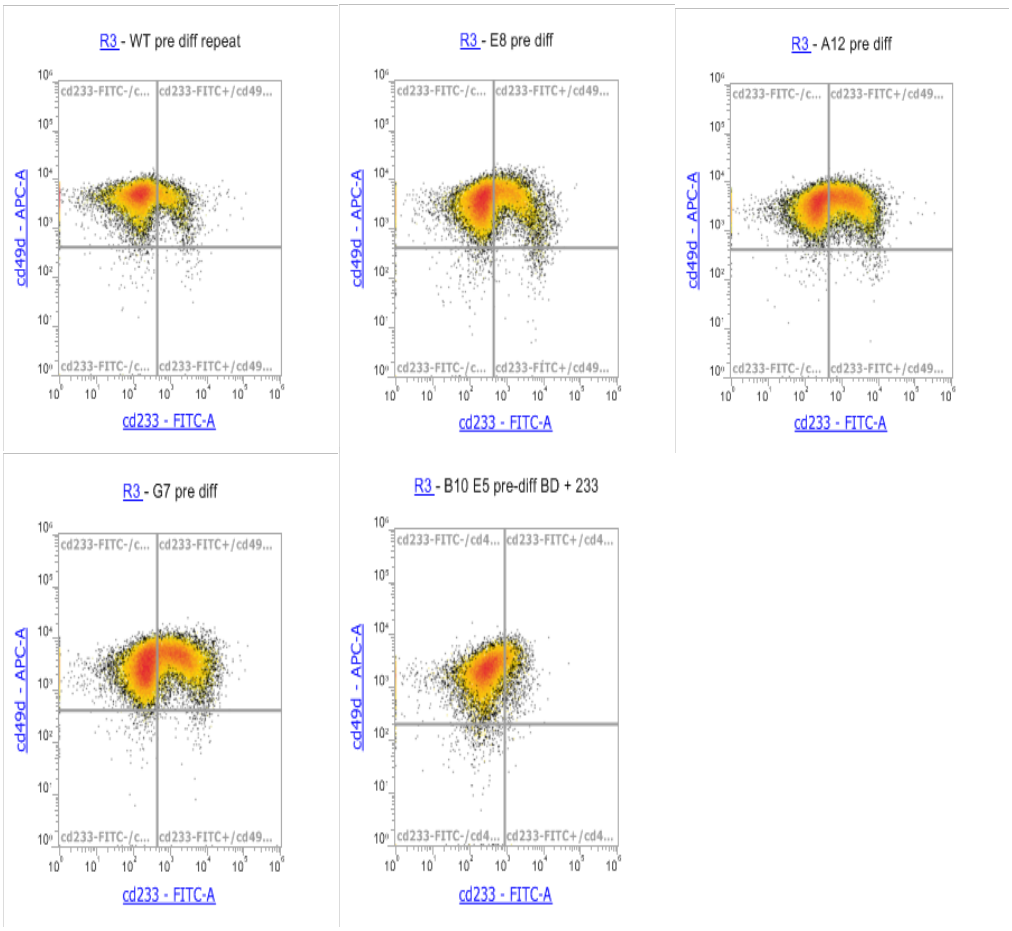


Figure 2-6. FACS analysis of CD49d and CD233 in five HUDEP clones in expansion phase. Top pane: from left to right – wildtype, ATRX E8, AP A12; Bottom pane: from left to right – APA G7 and B10 E5. ATRX E8, AP A12 and APA G7 clones had a significant proportion of $CD233^{+}$ cells, while most of the cells in B10 E5 clone were $CD233^{-}$.

HUDEP cells in expansion phase were stained with anti-CD233-FITC antibody (extracellular domain, clone BRIC6, International Blood Group Reference Laboratory) at 1:100 dilution in PBS at 4°C for 15min, using 20µL per 10^6 cells. Stained cells were washed in 5mL of PBS and centrifuged at 300g for 5min. Supernatant was removed and cells resuspended in 95µL of MACS buffer (for up to 107 cells). An aliquot of 5µL was removed and diluted in PBS for FACS analysis and lysed for RNA extraction. For every 90µL of cells in MACS buffer 10µL of anti-FITC microbeads (Miltenyi Biotec) were added and incubate at 4°C for 15min. The magnetically labeled cells were washed in 1mL of MACS buffer and

centrifuged. Pellet was resuspended in 500µL of MACS buffer and proceeded to negative selection in a LD column (Miltenyi Biotec). The LD column depleted $CD233^+$ cells. $CD233^-$ cells were eluted and an aliquot was analysed by FACS and lysed for RNA extraction. FACS analysis showed efficient removal of $CD233^+$ cells with this process – all HUDEP clones had less than 5% $CD233^+$ cells after depletion (Figure 2-7).

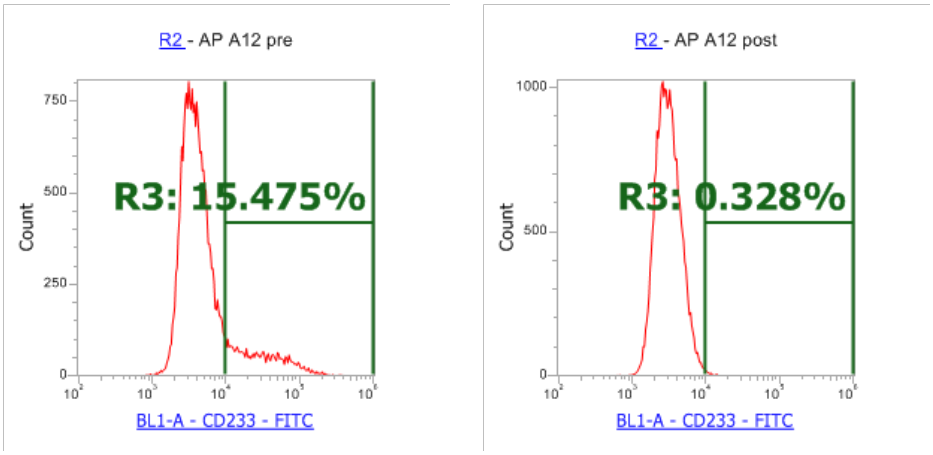


Figure 2-7. FACS analysis before and after CD233 MACS depletion. Left: Histogram of HUDEP cells in expansion phase prior to CD233 depletion showing a population of $CD233^+$ (R3: 15.475%) cells. Right: Histogram of HUDEP cells post-depletion.

Immediately after MACS depletion, CD233-depleted HUDEP cells were differentiated in the erythroid pathway using a two-stage differentiation protocol. Cells were resuspended in base media supplemented with stage I supplement. Cell density was adjusted to $0.5x10^6$/mL every second day with complete media change.

On day 8, media was changed to Phase II media without doxycycline. Cells were stained on Day 9 for CD233-FITC and CD49d-APC, using 20µL of antibody mix per 10^6 cells, at 4°C for 20min. Cells were washed in PBS and resuspended in MACS buffer containing DAPI at 1µg/mL. FACS sorting was performed on a Sony

SH800 cell sorter into 15mL centrifuge tube. FACS sorted cells were centrifuged at 400g for 10min and pellets collected for downstream processing (ATAC-seq and RNA extraction).

Base media component	**Final concentration**
IMDM	-
Penicillin/Streptomycin	1x
Human AB plasma	3%
FBS	2%
Heparin (Sigma)	3U/mL
Human insulin (Sigma)	10µg/mL
EPO	3IU/mL
Iron saturated human holo-transferrin	200µg/mL

Stage I supplement (Day 0-7)	**Final concentration**
SCF	10ng/mL
IL-3	1ng/mL
Doxycycline	2µg/mL

Stage II supplement (Day 8-10)	**Final concentration**
Iron saturated human holo-transferrin (additional)	300µg/mL

Antibody mix	**Dilution (in PBS)**
CD49d-APC (Clone 9F10, BD Biosciences)	1:25
CD233-FITC (BRIC6, International Blood Group Reference Laboratory)	1:100
FcR block (Miltenyi Biotec)	1:10

3 Mutation Spectrum in ATMDS

3.1 Patient demographics

DNA samples, prepared from peripheral blood, were available from a total of 38 ATMDS patients. Two samples failed library preparation twice due to poor DNA quality and were therefore not included. For four samples (#118, 122, 123 and 125) $CD34^+$ cells isolated from peripheral blood by MACS separation were used for DNA extraction and sequencing.

The demographics of ATMDS patients were compared to two large MDS cohorts[115, 116] in Table 3-1. Comparison on the MDS classification between ATMDS and MDS is difficult as most ATMDS cases were historical with no karyotyping, formal IPSS-R risk group or WHO MDS subtype classification. The average age at presentation was similar across the three groups. However the male to female ratio was significantly higher at 8:1 in ATMDS. Detailed haematologic indices were not available from these two studies.

	ATMDS (n=36)	**Papaemmauil *et al*[115] (n=738)**	**Haferlach *et al*[116] (n=944)**
Average age at presentation	70.6	68	72.8 (median)
Male:female ratio	8:1	1.28:1	1.6:1
Average Hb (g/dL)	9.7±1.9	9.0±2.0	N/A
Average MCV (80-100μm³)	73.9	N/A	N/A
Average MCH (27-32pg)	22.8	N/A	N/A
Karyotype			
- Normal	12 (33.3%)	425 (57.6%)	648 (68.6%)
- Abnormal	6 (16.7%)	213 (28.9%)	296 (31.4%)
- Not available	18 (50.0%)	100 (13.6%)	-

Table 3-1. Comparison of patient demographics and clinical features between ATMDS and two MDS cohorts. Hb: Haemoglobin; MCV: mean corpuscular volume; MCH: mean corpuscular haemoglobin.

3.2 Findings from myeloid panel sequencing

Some of the patients in the ATMDS cohort were previously investigated by targeted sequencing (Andrew Campbell, unpublished data), showing an enrichment of mutations in *ATRX*, *ASXL1* (Additional Sex Combs Like 1), *U2AF1* (U2 snRNP Auxiliary Factor 1) and *PHF6* (PHD Finger Protein 6) in ATMDS. We extended this dataset using targeted sequencing covered 32 genes available from the Oxford NIHR BRC Molecular Diagnostic Centre in 36 ATMDS cases.

In terms of quality control of the sequencing data and variant calling, a mean amplicon coverage of 5347x (3721-7821x, over 289 amplicons) was achieved. Uniformity of coverage (% of reads > 0.2 x mean amplicon coverage) was 93.0%. Variants with >5% minor allele frequency, read depth of 100x and quality score of >99 were called. Variants were assessed pathogenicity and classified based on the following criteria:

Variants of predicted clinical significance (highlighted in green in Table 3-3):

- Recurrent Variants: All mutations regardless of variant types (missense, nonsense, frameshift, in/out-of-frame indel, splice site mutation) that have been previously reported in haematologic malignancies in either the literature and/or COSMIC (Catalogue of Somatic Mutations in Cancer) database[159].
- Novel Frameshift or nonsense mutations: Mutations which cause premature stop codons or elongation of the protein are reported

under the assumption that the alteration to the mRNA will have an effect on protein function.

Variants of unknown clinical significance (highlighted in yellow in Table 3-3):

- Novel Splice Site Mutations: variants predicted to affect splicing by at least 3 of the 5 splice site prediction software used.
- Novel missense mutations occurring at known recurrently mutated codons.
- Novel missense mutations with *in vitro* functional data demonstrating that they are likely to alter the function of the protein.
- Recurrent variants with contradictory information available in the scientific literature: variants which are variably described as either germline polymorphisms or acquired somatic variants with no clear consensus as to their significance.

3.2.1 Mutation load and enrichment in ATMDS

Figure 3-1 shows the distribution on the number of mutations found in ATMDS patients, stratified by the severity of disease based on percentage of cells with Hb H inclusion bodies. Patients with 4-5 mutations comprise 67% of the cohort. All ATMDS patients have at least 2 mutations detected. This is in contrast to MDS without alpha thalassaemia where 10.5-26% of patients did not contain any oncogenic mutation[115, 116]. There is not a

clear correlation between the number of mutations and severity of disease – in fact patients with the most mutations (7-8 mutations) all have <40% Hb H inclusion bodies.

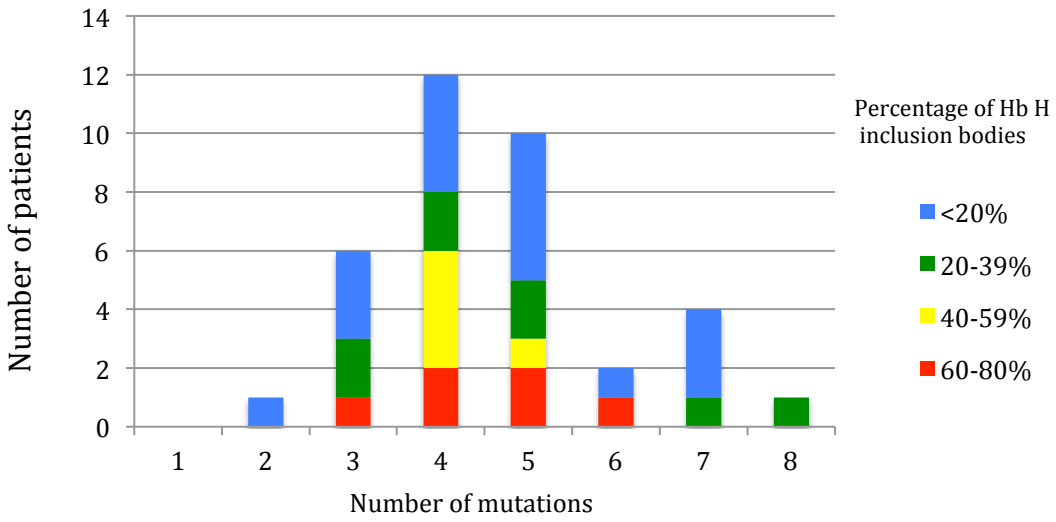


Figure 3-1. Distribution of number of mutations in ATMDS. Percentage of Hb H inclusion bodies is divided into four categories. Variants of unknown clinical significance were counted towards the total number of mutations.

Table 3-3 shows the mutational landscape in this ATMDS cohort. From top to bottom ATMDS cases were ranked by percentage of Hb H inclusion bodies as measured by BCB stain. Genes were listed from left to right in descending order of mutation frequency in this ATMDS cohort. When we interrogated specific genes, this table shows a significant enrichment of mutations in *ASXL1*, *U2AF1*, *PHF6* and *ATRX* in this cohort, with between 58-86% of patients containing mutation in these four genes (see Supplementary Table 1 in Appendix for detailed mutation incorporating data obtained from previous and current study).

To put the enrichment of these four genes in perspective, the highest percentages of gene mutation seen in MDS were *TET2* and *SF3B1*, both of

which were found in only 30-35% of patients[116]. The higher percentage of *SF3B1* mutation reported in the study was likely due to a significant proportion of patients (19.4%) with RCMD-RS (refractory cytopenia multilineage dysplasia-ring sideroblast) in the cohort, where *SF3B1* mutations are known to be highly enriched[118, 160]. Table 3-2 compares the prevalence of mutations in *ATRX*, *ASXL1*, *U2AF1* and *PHF6* in ATMDS versus the MDS cohort reported in Haferlach *et al*[116]. It is obvious that these four genes are highly enriched in the ATMDS cohort.

	ATRX	***ASXL1***	***U2AF1***	***PHF6***
ATMDS (n=36)	58% (21)	86% (31)	81% (29)	86% (31)
MDS (n=944)	1% (11)	24% (225)	8% (73)	3% (27)

Table 3-2. Comparison of mutation prevalence in ATRX, ASXL1, U2AF1 and PHF6 between the ATMDS cohort and published MDS cohort[116].

On the other hand, mutations in *SF3B1* and *DNMT3A* were under-represented in ATMDS compared to MDS cohorts. Particularly, *SF3B1* mutation is one of the most common mutations in MDS. But in ATMDS only one case was found to be mutated. This is in accordance with the negative correlation between *PHF6* and *SF3B1*[161]. Mutation frequencies in *TET2*, *RUNX1*, *SRSF2*, *IDH1*, *NRAS*, *TP53*, *ZRSR2*, *CEBPA*, *EZH2* and *KIT* are in concordance with the published MDS cohorts[115, 116]. No mutations were found in the following genes in ATMDS patients: *CBLB*, *CBLC*, *FLT3*, *HRAS*, *IDH1*, *MPL*, *NPM1*, *PDGFRA*, *PTEN* and *WT1*.

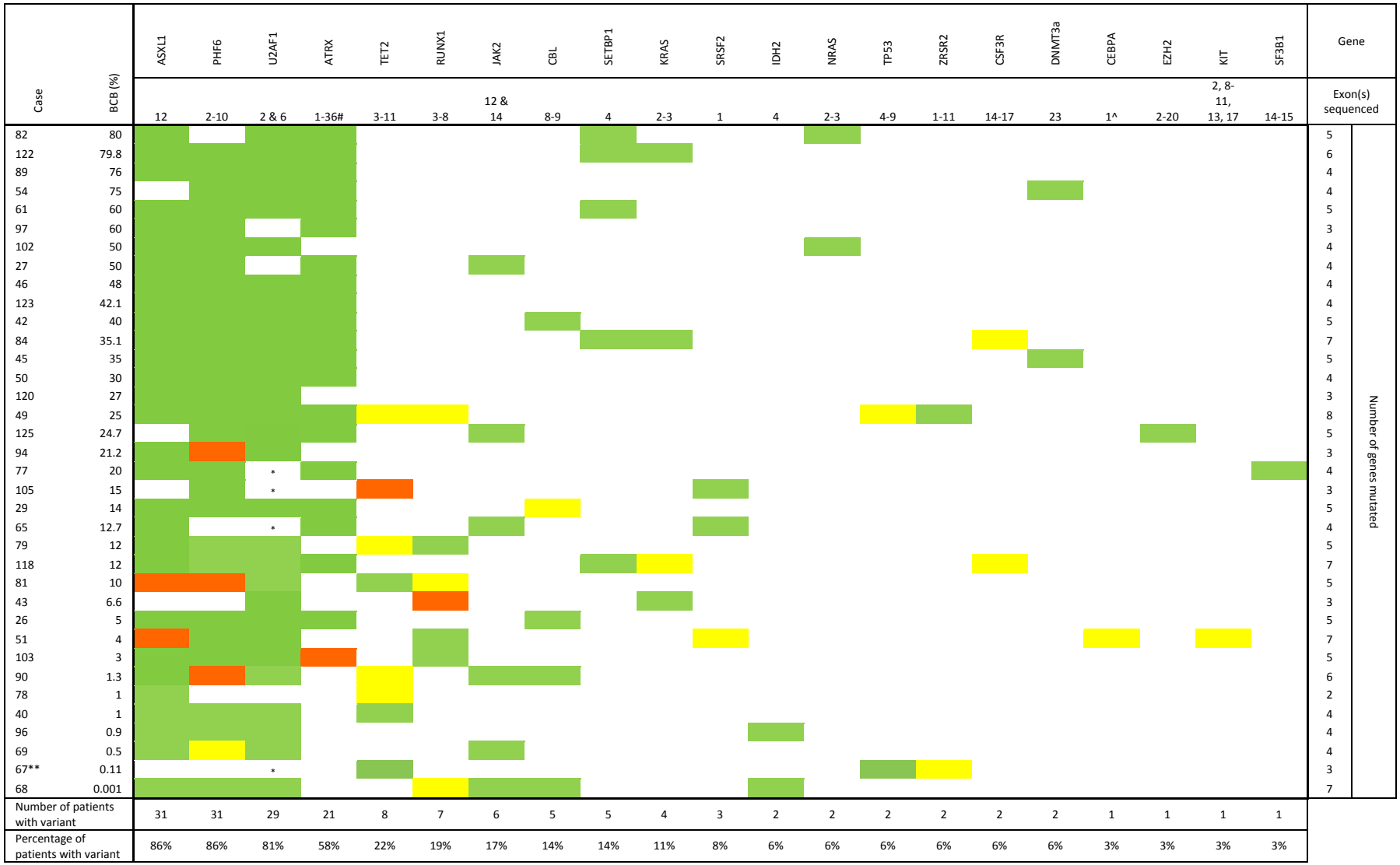


Table 3-3. List of mutations found in ATMDS patients using the Oxford myeloid panel. #: coverage is a combined from sequencing data obtained through Sanger sequencing, myeloid panel and Ion Torrent sequencing. Green and yellow highlight variants of predicted clinical significance and variants of unknown significance, respectively. Variants in orange denotes more than one mutation detected in the particular gene. A full table with variant position and variant allelic frequencies (VAFs) listed, and incorporating data from previous panel is available in Appendix (Supplementary Table 1). * indicates a spliceosomal mutation other than U2AF1 mutation was detected. **case #67 was found to have a heterozygous 1.9Mb deletion in chromosome 16 involving both HBA2 and HBA1 (i.e. equivalent to a two-gene deletion --/αα), reported in Steensma *et al*[138]. ^: only part of CEBPA exon 1 was sequenced.

3.2.2 Correlation between enriched genes in ATMDS

Within these four genes, the association between gene pairs of *ASXL1-PHF6*, *U2AF1-PHF6* and *ASXL1-U2AF1* had been described previously. *ASXL1* and *U2AF1* mutations are relatively more common in MDS, and they are frequently found to be co-mutated[116, 162]. In the ATMDS cohort, this co-mutation pattern is also observed in 77.1% of cases. In a study focusing on myeloid neoplasms with *PHF6* mutations, Mori *et al*[161] reported that concomitant mutations in *ASXL1* and *U2AF1* are found in 37.5% (18/48) and 20.8% (10/48) of *PHF6* mutation-positive cases, respectively. However, in ATMDS this association is even more significant, with these figures rising to 93.9% (31/33 for ASXL1-PHF6 pair) and 87.9% (29/33 for U2AF1-PHF6 pair) respectively. No data was available on the prevalence of *ASXL1*, *U2AF1* and *PHF6* triple mutant. Other genes significantly co-mutated with *PHF6* in the same report were *EZH2*, *RUNX1*, *SMC1A*, and *ZRSR2.* In ATMDS the prevalence of mutations in these genes were too low for meaningful comparison (n=1, 6 and 2 for *EZH2*, *RUNX1* and *ZRSR2*, respectively). *SMC1A* was not covered by the Oxford myeloid panel.

In Haferlach *et al*[116], *ATRX* mutations were detected in eight patients. While blood counts were not available in these patients, the mutation spectrum among these patients showed an interesting pattern. Only one was found to have all four enriched genes mutated (patient 6, see Supplementary Table 2), while another patient (patient 1) possessed all but *PHF6* mutation. One observation that stands out is all except one were

found to have a spliceosomal mutation (*U2AF1*, *ZRSR2* or *SRSF2*), suggesting an association between *ATRX* and spliceosomal mutations.

3.2.3 Spliceosomal mutations in ATMDS

Spliceosomal mutations tend to be mutually exclusive in MDS[115, 116]. If we look for other spliceosomal mutations in the seven ATMDS cases without a detectable *U2AF1* mutation, four were found to have spliceosomal mutations in *SF3B1*, *SRSF2* or *ZRSR2*. This translates to 92% of the cohort containing a spliceosomal mutation, a significant increase from just over 50% in all MDS patients (discussed in Introduction). In addition, spliceosomal mutations are not particularly enriched for *U2AF1* mutations in MDS. In MDS without ring sideroblasts, *SF3B1*, *U2AF1*, *SRSF2* and *ZRSR2* have mutation frequencies of 4.4-7%, 10.3-11.6%, 11.6% and 7.7% respectively[163]. This highlights the significance of specific enrichment for *U2AF1* mutation in ATMDS.

3.3 Temporal order of mutations in ATMDS

In MDS, modeling of clonal architecture using variant allele frequencies from sequencing data suggested that mutations in *U2AF1* and other spliceosomal genes occur early in pathogenesis, while *PHF6* and *ASXL1* mutations appear at a similar time point subsequently[115]. *PHF6* mutations have been reported to be subclonal in 54% of myeloid neoplasm[161]. We ranked VAFs of *ATRX*, *ASXL1*, *U2AF1* and *PHF6* for each individual patient in the ATMDS cohort (Figure 3-2). In agreement with

Papaemmanuil *et al*, in the majority of ATMDS cases the mutation with the highest VAF (i.e. suggesting it is an early event) is found in *U2AF1*. *PHF6* and *ATRX* mutations have lower VAFs in comparison.

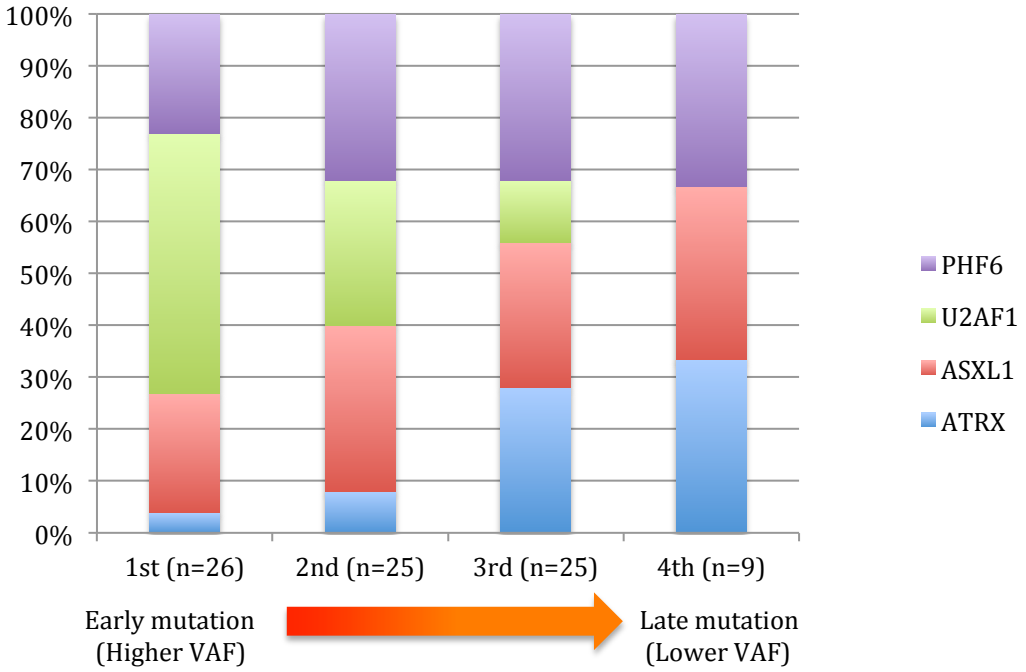


Figure 3-2. Modeling temporal sequence of mutation acquisition in ATMDS. For each patient mutations were ranked in descending order of VAFs (1st → 4th). The proportion of each gene in each rank is shown above. The VAFs for ATRX and PHF6 mutations were halved to compensate for gene dosage. ATRX was found to have the highest VAF in only one single case (case #82, who has the highest % of Hb H inclusion in RBCs in the cohort). U2AF1 mutation was never found to have a lower VAF than mutations in the other three genes. Only VAF data from sequencing of peripheral blood were used. Sequencing data from cases where $CD34^+$ cells were sequenced were excluded.

We then looked at the specific mutations in the four highly enriched genes to predict the functional consequences.

3.4 *ATRX* mutations in ATMDS

Within the four highly enriched genes, mutations in *ATRX* showed an interesting distribution pattern – the prevalence of mutation in those with Hb H% ≤10% was much lower compared to those with higher Hb H% (16.7% versus 79.2%, respectively. Fisher's exact test *P*=0.0008). Whereas *ASXL1*, *U2AF1* and *PHF6* mutations did not show any correlation with Hb H%.

In ATMDS, mutations in *ATRX* either cause 1. truncation or splicing error, or 2. missense mutations in the two functional domains (ADD or SWI2/SNF2 domain) (Table 3-4). There is no strong preference as to the location or type of mutations - the number of mutations found within the two domains is equal (6 cases each). There was no statistically significant correlation between location of ATRX mutation and Hb H inclusion bodies percentage (minimum adjusted *P* value 0.0851 between 5' end truncating mutation/splice site mutation and ADD domain mutations). Therefore while the presence of ATRX mutation aggravates the alpha thalassaemia phenotype, the mechanism seems to be related to loss of function.

Location of mutation (5' to 3')	ATRX Mutation	RBCs with Hb H inclusion bodies (%)	Average RBCs with Hb H inclusion bodies (%)
5' end truncating mutation/splice site mutation	c.20+1G>A	12.7	16.9
	p.Arg78*	3	
	p.Ser79*	35	
ADD domain (residues 161-292[75])	p.Cys174Tyr	76	58.5
	p.Leu192Phe	50	
	p.Cys200Tyr	60	
	p.Cys240Gly	48	
	p.Trp263Cys	42.1	
	p.Cys265Tyr	75	
SWI2/SNF2 domain (resides 1560-2259[63])	p.Cys1576Ser	12	35.2
	intron 18 splicing mutation	60	
	c.5567-3C>G (acceptor splice site of exon 23)	14	
	p.Arg2028Pro	30	
	p.Arg2131Gln	35.1	
	p.Leu2240Phe	60	
3' end truncating mutations	p.Asn2296Glnfs*26	40	37.5
	p.Asn2340Valfs*2	25	
	p.Glu2351Argfs*3	5	
	p.Arg2407*	80	

Table 3-4. Distribution of ATRX mutations in ATMDS. Mutations within the ADD and ATPase domains are highlighted in yellow and orange, respectively. Reference sequence: NM_000489.3, NP_000480.2.

3.5 *ASXL1* truncating mutations in ATMDS

ASXL1 is one of the most commonly mutated genes in MDS (see Table 3-2 above). It codes for a nuclear protein with an unusual plant homeodomain (PHD) at the C-terminus. In MDS, *ASXL1* mutations are exclusively heterozygous truncating mutations in the last exon (Figure 3-3). The truncation removes the PHD, leading to a gain-of-function mutation with detectable level of truncated protein[164-166].

The mechanism on how truncated ASXL1 promotes leukaemogenesis is not entirely clear. ASXL1 is co-immunoprecipitated with EZH2 and SUZ12, two

components of PRC2. Loss of ASXL1 results in reduction in H3K27me3 in specific oncogenic loci (e.g. *HOXA* genes) via inhibition of PRC2 recruitment[167]. ASXL1 also associates with the deubiquitinase BAP1. In a mouse model, a truncated form of ASXL1 enhances the ASXL1/BAP1 complex, resulting in erasure of H2A lysine 119 ubiquitination and upregulation of *Hoxa* genes[168, 169].

Recently, mass spectrometry and co-immunoprecipitation experiments showed that truncated ASXL1, but not wildtype ASXL1, recruits bromodomain protein 4 (BRD4) to genes[165]. BRD4 possesses pleiotropic functions in H3 lysine 122 acetylation, RNA polymerase II carboxy-terminal domain (CTD) phorphorylation, and telomere length regulation[170-172]. The recruitment of BRD4 by truncated ASXL1 led to increased transcription of genes related to stem cell maintenance and myeloid differentiation. Interestingly, in normal erythropoiesis BRD4 is recruited to GATA1-regulated genes. GATA1 is the major transcription factor involved in erythropoiesis, and depletion or inhibition of BRD4 leads to repression of GATA1-mediated erythroid gene, including alpha, beta and gamma globins[173, 174]. Whether this acquired ability to bind BRD4 in truncated ASXL1 could produce the same effect is unknown. No physical interaction was detectable between truncated ASXL1 and ATRX, U2AF1 or PHF6[165].

Consistent with the mutation pattern in MDS, in ATMDS all *ASXL1* mutations are also heterozygous truncations between codons 583-760 in the last exon. The most common truncating mutation was

NM_015338.5:c.1934dupG (NP_056153.2:p.Gly646TrpfsTer12), representing 36.7% of all *ASXL1* mutations in this cohort. The same mutation is also found to be the most common *ASXL1* mutation in primary MDS patients[175].

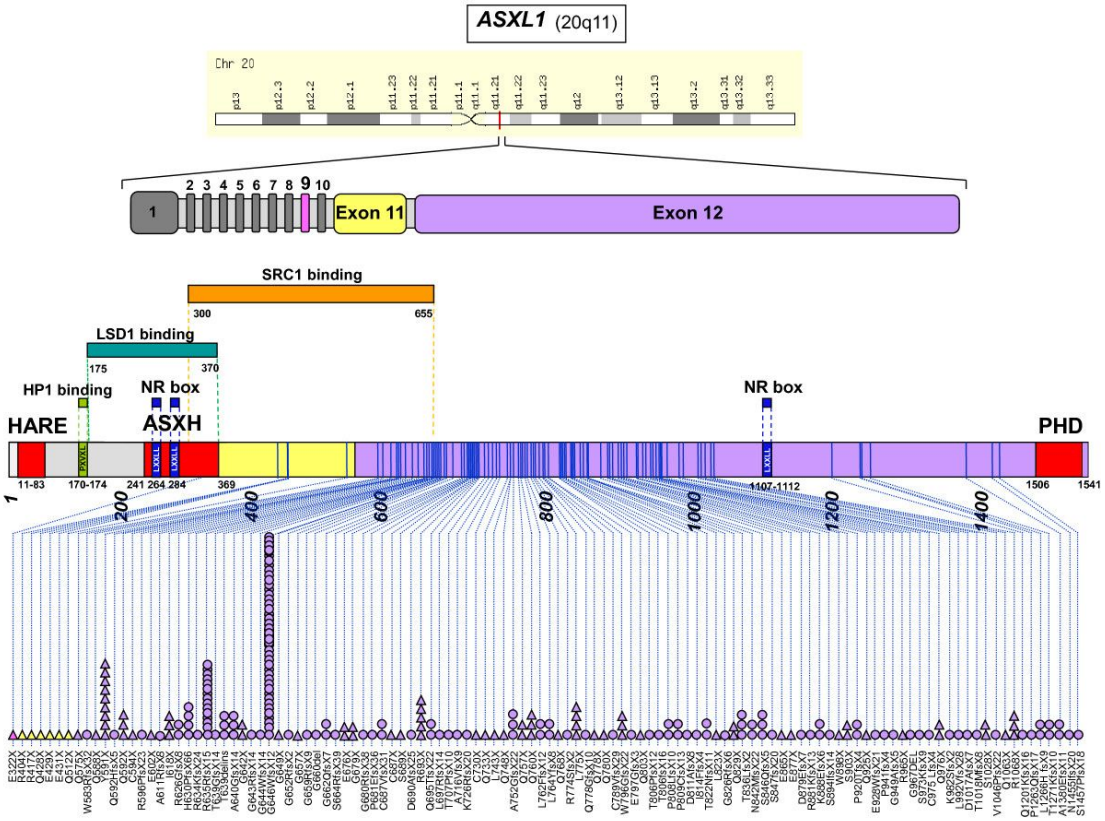


Figure 3-3. ASXL1 truncation in myeloid neoplasm. Human ASXL protein family (ASXL1, ASXL2 and ASXL3) share the same architecture with a N-terminal HB1/Asxl Restriction endonuclease helix-turn-helix (HARE-HTH) domain, followed by ASXH domain that associates with epigenetic regulators BRCA associated protein 1 (BAP1) and lysine demethylase 1A (LSD1), a steroid receptor coactivator 1 (SRC1) binding region and the C-terminal PHD domain. Most of the reported truncating mutations are within the last exon, all of which are 5' to the C-terminal PHD. Adapted from Gelsi-Boyer *et al*[176].

3.5.1 ASXL1 mutation in erythropoiesis

ASXL1 mutations play a role in erythropoiesis. MDS/MPN patients with *ASXL1* mutations have lower RBCs and haemoglobin levels compared to those with wildtype *ASXL1*[177]. The same is observed in a case report of the rare Bohring-Opitz syndrome (BOPS, OMIM#605039), caused by congenital *ASXL1* heterozygous truncating mutation[178]. The patient had

persistently low levels of RBCs and haemoglobin over a two-year period. Specifically, MCV and MCH were at low normal level and microcytic hypochromic RBCs were noted, although the iron status of this patient was not known (see Supplementary Table 3 for blood counts). Also, in chronic myelomonocytic leukemia where up to 45.3% of patients harbour *ASXL1* truncation, bone marrow examination showed evidence of block in erythroid differentiation with increased proportion of early stage erythroblasts and reduced enucleated erythrocytes[177, 179, 180]. However, in elderly individuals not diagnosed with haematologic disorders, *ASXL1* mutations were one of the more common somatic mutation detected (62 variants in 17,182 individuals), and individuals with only *ASXL1* somatic mutations did not show low haemoglobin or MCV[181].

Haematologic findings from patients with somatic *ASXL1* mutations are recapitulated in three independent Asxl1 truncation knock-in mouse models containing *Asxl1*^{E635RfsX15}, *Asxl1*^{G643fs} and *Asxl1*^{Y588X}[165, 182, 183]. Examination of bone marrow in these mice showed dysplastic myeloid cells and accumulation of cells in early stages of erythroblast maturation determined by Ter119/CD71 flow cytometry analysis in the former two models. In peripheral blood all three models showed thrombocytosis consistent with a MDS phenotype. Pseudo Pelger-Huët anomaly was also seen in neutrophils of Asxl1^{G643fs} mouse, as observed in the bone marrow of ATMDS patient MK (case #123). RBCs and haemoglobin levels were slightly reduced in both mouse models. However, the morphology of RBCs was normal. Importantly, MCV was increased in *Asxl1*^{E635RfsX15} mice, and

unchanged in *Asxl1*^{Y588X} mice (MCV was not measured in the *Asxl1*^{G643fs} model). RNA-seq from bone marrow HSCs of *Asxl1*^{E635RfsX15} knock-in mice showed upregulation in genes involved in mRNA processing and splicing, while targets involved in erythroid differentiation were downregulated. H3K4me3 ChIP-seq showed strong reduction in H3K4me3 in the murine *Hba* locus. Expression of *Hba-a1* was reduced but the level of reduction was not statistically significant[182]. These data suggest that while *ASXL1* truncation can affect erythropoiesis through blockage in erythroid differentiation, it does not seem to be able to induce an alpha thalassaemia phenotype alone.

3.6 The role of U2AF1 in acceptor splice site recognition

Transcripts from most genes in higher eukaryotes are transcribed as precursor mRNAs (pre-mRNAs). Pre-mRNAs contain introns that need to be excised prior to export from the nucleus for translation. Intron removal is performed over a two-step trans-esterification reaction. This reaction is catalyzed by the spliceosome (Figure 3-4). Spliceosome is a large complex containing five small nuclear ribonucleoproteins (snRNPs) (U1, U2, and the U4/U6.U5 tri-snRNP), their associated small nuclear RNAs and non-snRNP splicing factors. U2 snRNP Auxiliary Factor (U2AF) is one of such non-snRNP splicing factors to promote stable U2 snRNP association with pre-mRNAs[184]. U2AF is essential for 3' splice site recognition that occurs in the initial stage of spliceosome assembly[185].

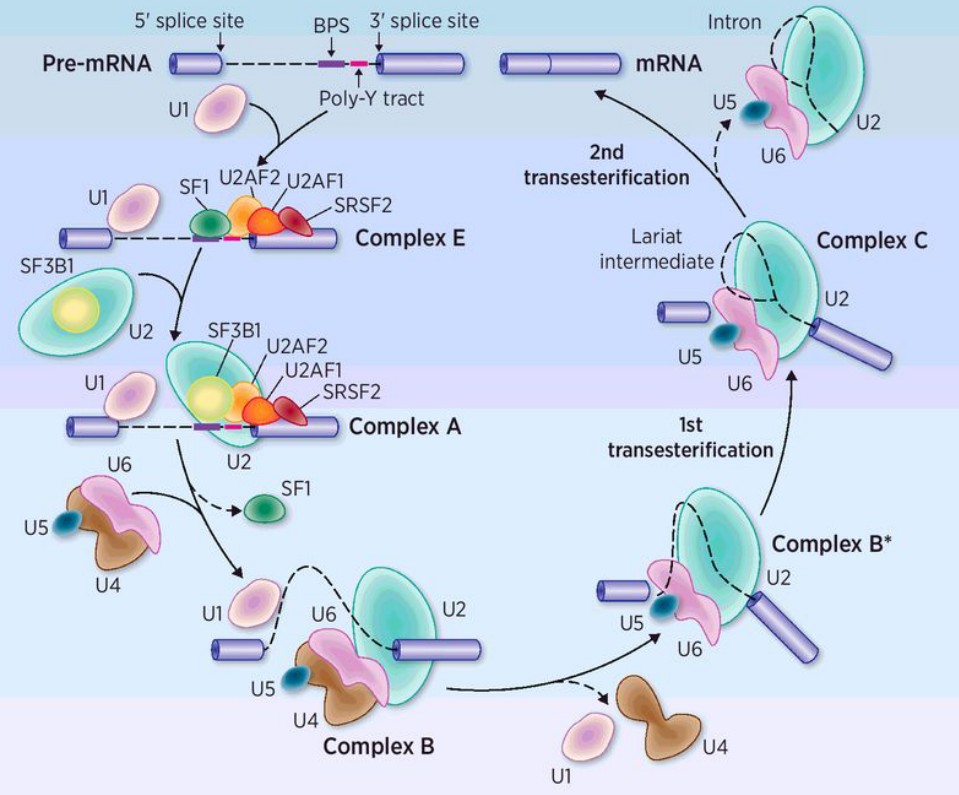


Figure 3-4. Spliceosome assembly and pre-mRNA splicing. Within the intron, there are four elements involved in splicing: the donor (5') splice site (GT dinucleotide), acceptor (3') splice site (AG dinucleotide), branch-point sequence (BPS) upstream of the acceptor splice site and a polypyrimidine tract between BPS and the acceptor splice site. In the first step of splicing of the major class of intron (called U2-type introns), U1 snRNP, Splicing Factor 1 (SF1), and U2AF2/U2AF1 bind to the donor splice site, BPS and polypyrimidine tract/acceptor splice site, respectively (Complex E). This in turns recruit the U2 snRNP containing SF3B1 to the BPS (Complex A). Following release of SF1, the U4/U6.U5 tri-snRNP is recruited and releases the U1 and U4 snRNPs to form the first catalytic Complex B*. This is followed by a two-step transesterification reaction. The 2'OH of BPS adenosine first carries out a nucleophilic attack on the donor splice site guanosine, yielding a 5' exon with a free 3'OH and a branched intron lariat attached to the 3' exon (Complex C). In the second step the 3'OH of the 5' exon attacks the first nucleotide downstream of the acceptor splice site to release the two ligated exons (spliced mRNA) and the excised intron lariat. For a small subset of introns (called U12-type introns), an alternative spliceosome complex involving ZRSR2 is responsible for splicing[186]. From Yoshimi and Abdel-Wahab[187].

U2AF is a heterodimer comprising the larger 65-kDa U2AF65 (official gene name *U2AF2*) and smaller 35-kDa U2AF35 (official gene name *U2AF1*) subunits. U2AF65 contacts the intronic polypyrimidine tract preceding the acceptor splice site. This polypyrimidine tract is important to identify the acceptor splice site during spliceosome assembly[188]. U2AF65 binds directly to phosphorylated CTD of RNA polymerase II, coupling splicing with transcription[189]. *U2AF2* is not frequently mutated in haematologic

malignancies, with six cases reported containing the most common mutation p.Leu187Val, located within the RNA recognition motif 1 (RRM1). The RRM1 motif is adjacent to the binding site to U2AF35[190].

The U2AF35 subunit, although not required for splicing *in vitro*, is essential for survival *in vivo*[191, 192]. It is responsible for recognition of the canonical acceptor splice site dinucleotide (AG) to enhance the affinity of U2AF heterodimer[193, 194]. U2AF35 contains two highly conserved zinc finger domains that bind cooperatively to the target RNA sequence as shown in the yeast homolog U2AF23[195, 196]. The two zinc finger domains are interspersed by the U2AF homology motif where binding to U2AF2 occurs. At the C-terminus there is an arginine/serine-rich (SR) domain that is responsible interaction with other SR proteins such as SRSF2 at the 3' exon splicing enhancer (ESE) site[197]. Thus, U2AF35 is thought to act as a bridge between SR proteins and U2AF65 at the acceptor splice site.

3.6.1 Non-canonical role of U2AF1

Apart from its role in splice site recognition, U2AF1 is known to shuttle between nucleus and cytoplasm[198]. Recently U2AF1 was shown to possess a non-canonical function on translational control in human cell line. Binding of hundreds of mature mRNA to cytoplasmic U2AF1 repressed translation, and this repression was lost in certain genes when the p.Ser34Phe mutation was introduced[199].

3.6.2 U2AF1 hotspot mutations in MDS

In MDS and other myeloid neoplasms (MPN, MDS/MPN, AML), *U2AF1* mutations are mostly found in two hotspots in codons 34 and 157. The mutation is always found in a heterozygous state, as the wildtype allele is required for cell survival[200]. The most common mutations in these two locations are p.Ser34Phe (S34F) and p.Gln157Pro (Q157P), accounting for more than 75% of all *U2AF1* mutations. Other mutations such as p.Ala26Val, p.Ser34Tyr, p.Arg35Leu, p.Arg156Gln, p.Gln157Arg and p.Gly213Ala are detected at a lower frequency[201, 202]. In the same study, *U2AF1* mutation was significantly more common in male than female MDS patients (83% vs 17%)[202]. These mutations are located at the expected RNA binding surfaces within the first and second zinc finger domain, respectively. In the yeast U2AF23 model, the p.Ser34Phe mutant loses RNA binding activity, while the p.Gln151Arg mutant (corresponding to p.Gln157 in human U2AF35) retains binding activity but loses sequence selectivity[196].

Expectedly, these two mutations differentially affect splicing via alternation in acceptor splice site preference. The p.Ser34 mutant has reduced affinity for uridine at the -3 position upstream of the acceptor splice site (Figure 3-5)[203, 204], while the p.Gln157 mutant has a higher preference for G and represses A at the +1 position (immediately following the canonical AG dinucleotide at the acceptor splice site)[205, 206]. This differential preference was shown in K562 cells where the two mutations showed little overlap in differentially spliced events[205].

In addition to alternative splicing, mutant U2AF1 was also shown to alter mRNA cleavage and polyadenylation (poly(A)) site usage[207]. RNA-seq on murine cell line expressing p.Ser34Phe mutant showed that increased usage of distal cleavage and poly(A) site comprised of 40.8% of alternative RNA processing events. Specifically, the alternative cleavage and poly(A) site led to a longer and inefficiently translated transcript in the gene *Atg7*. This longer transcript was detectable when U2AF1 p.Ser34Phe was expressed in human bronchial cell line, and in MDS patient with both p.Ser34Phe/Tyr and p.Gln157Pro mutations.

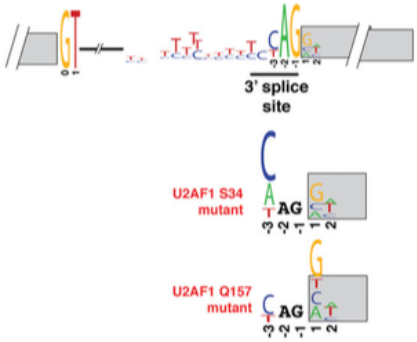


Figure 3-5. Differential effect on acceptor splice site recognition in the two U2AF1 hotspot mutations. Mutation at the p.Gln157 position means exons are more likely to be included when the +1 position is G instead of A. Adapted from Inoue et al[208].

The p.Ser34 mutant has been the focus for *in vitro* and *in vivo* studies as it was found to be the more common *U2AF1* mutation in MDS (see Table 3-5). Various RNA-seq datasets from primary tissues (bone marrow or blood of MDS/AML patients), cell lines and mouse models with U2AF1 p.Ser34 mutation showed differential exon usage pattern[202, 204, 209, 210]. However the pattern is different among these datasets, reflecting the difference in the source of input material or analysis pipeline.

3.6.3 U2AF1 in erythropoiesis

The specific effect of *U2AF1* mutation in erythropoiesis was investigated in Yip *et al*[210]. Transduced human $CD34^+$ progenitor cells overexpressing U2AF1 p.Ser34Phe mutation showed impaired erythroid differentiation as measured by surface markers (CD71 and CD235a) and colony forming assay. Apoptosis was increased in the mutant cells during erythroid differentiation. Comparison of RNA-seq data showed that genes with significant aberrant splicing in erythroid colonies with p.Ser34Phe mutation compared to wildtype U2AF1 were related to haem and mRNA processing.

In accordance to the *in vitro* findings, both U2AF1 mutations were associated with anaemia and thrombocytopenia in MDS[211]. However p.Gln157 mutation was more frequently associated with Hb <80g/L compared to the p.Ser34 counterpart, suggesting a more prominent effect on erythropoiesis.

In contrast to MDS, the majority (85.2%) of U2AF1 mutations in ATMDS were p.Gln157Pro (Table 3-5). The p.Ser34Phe mutation was only found in four cases (case #43, 94, 120 and 125, with BCB% between 6.6-27). No p.Ser34Tyr or p.Gln157Arg mutation was detected in the ATMDS cohort. Consistent with the observation in our cohort, p.Gln157 mutations are more likely to be co-mutated with *ASXL1* truncation than the p.Ser34 counterpart in both MDS and myelofibrosis[211, 212].

U2AF1 mutation	ATMDS (n=36)	Total MDS (n=3,258)	Individual MDS cohorts						
			Yoshida *et al*[118] (n=582)	Przychodzen *et al*[202] (n=524)	Walter *et al*[213] (n=150)	Haferlach *et al*[116] (n=944)	Wu *et al*[214] (n=304)	Li *et al*[211] (n=511)	Kim *et al*[215] (n=153)
Gln157Pro/ Gln157Arg	23	97	11	16	4	47	2	9	8
Ser34Phe/ Ser34Tyr	4	198	25	25	13	20	23	75	17

Table 3-5. Comparison of prevalence of the two hotspot *U2AF1* mutations in ATMDS versus MDS. Mutation in p.Gln157 is significantly enriched in ATMDS (P<0.0001, Fisher's exact test, two tailed). The actual numbers from each individual MDS cohort are shown on the right. Note some of these MDS studies include MDS, MDS/MPN and AML.

3.7 Constitutive and somatic mutations in *PHF6*

PHF6 is a small (365 amino acids), highly conserved nuclear protein containing two PHD-like zinc finger domains. Mutations of the *PHF6* gene, located at Xq26.2, were initially described in a congenital disorder called Börjeson-Forssman-Lehmann syndrome (BFLS, OMIM #301900), characterized by intellectual disability, epilepsy, hypogonadism, obesity with marked gynaecomastia and other dysmorphic features[216]. Haemoglobin was normal and there was no overt microcytic hypochromic anaemia in the few families where blood results were available, although one family showed borderline MCV in four individuals[217] (see Supplementary Table 4 for blood counts). The mutations described are either missense mutations within the two PHD domains of PHF6, or nonsense mutations leading to truncation, suggesting a loss of function effect.

Although *PHF6* is located on the X chromosome, multiple affected females with loss-of-function PHF6 mutations have been described[218-222]. Large deletions involving part or whole of *PHF6* were seen in females only (listed in[223]). The X chromosome inactivation in some of these cases was shown to be skewed[218, 221]. The phenotype of *PHF6* mutation in female is distinctive from the male patients. In fact, two of these patients were initially diagnosed as having Coffin-Siris syndrome (CSS, OMIM#135900), a development delay syndrome with overlapping clinical features. CSS is

caused by mutations in genes involving the Brg/Brahma-associated factor (BAF) complex, also called the mammalian SWI/SNF complex[220].

3.7.1 Somatic mutation of PHF6 in haematologic neoplasm

PHF6 is highly expressed in lymphoid progenitor populations, and less so in myeloid progenitors where the expression in megakaryocyte-erythroid progenitors (MEP) is higher than that of granulocyte-macrophage progenitors (GMP)[224]. The link between *PHF6* mutation and haematologic neoplasm was first described in T-cell acute lymphoblastic leukaemia (T-ALL), where *PHF6* is one of the most commonly mutated genes (between 5-40%, varying between pediatric and adult groups)[225-227]. *PHF6* mutation is common (20%) in T cell acute lymphoblastic lymphoma (T-ALL). When comparing between T-ALL patients with and without *PHF6* mutations, the haemoglobin of mutation positive patients was slightly but not significantly lower (median Hb 93 versus 102g/L)[228].

In myeloid neoplasm, a large study (n=1760) by Mori *et al*[161] (n=1760) showed that *PHF6* mutation is detected across the spectrum in MDS, MPN and AML (3.0%, 0.7% and 4.1% of cases, respectively). Mutation is however more frequent in RAEB (Refractory anaemia with excess blasts, 5.3%) and AML/MRC (Acute myeloid leukaemia with myelodysplasia-related changes, 15.4%), suggesting that it might be important in leukaemic transformation. Indeed, presence of *PHF6* mutation was significantly associated with higher bone marrow blast counts in multivariate analysis,

and a mouse model knockdown of Phf6 promotes growth of AML[229]. However, there was no association between presence of *PHF6* mutation and haemoglobin level (10.3 versus 10.2, between *PHF6* mutation positive and mutation negative group, *P*=0.83). *PHF6* mutations are usually truncating/splice site mutations (nonsense, frameshift and splice site mutations constitutes 72.6% of all detected mutations), and missense mutations are more common in the second PHD domain.

Whether *PHF6* mutations contribute to sex bias in these neoplasms is controversial. While Mori *et al* and other reports[228, 230] showed no sex bias, others have reported a strong male bias in PHF6 mutation positive cases[224, 225].

3.7.2 Function of PHF6

The potential cellular functions of PHF6 have only recently emerged from various functional and structural studies. PHF6 associates with proteins involved in multiple epigenetic processes, namely ribosomal RNA synthesis, chromatin modification, and transcriptional regulation.

The role of PHF6 in ribosomal RNA (rRNA) biogenesis has been the focus of investigation in PHF6 function. PHF6 localises to the nucleolus in a RNA-dependent manner via the first PHD domain[231, 232]. It is proposed that PHF6 is involved in the regulation of ribosomal RNA (rRNA) transcription and processing via its interaction with PAF1 (polymerase associated factor) transcription elongation complex. PHF6 physically interacts with PAF1 and

knockdown of PHF6 or PAF1 inhibition produced the same impairment in neuronal migration in mice[233]. Knockdown of PHF6 also induced DNA damage at the rDNA locus[231]. This was thought to be due to R-loop formation at the locus. R-loop is a three-stranded structure consisting of a DNA-RNA hybrid between the nascent RNA and its associated DNA template strand, along with the displaced single stranded DNA non-template strand[234]. Overexpression of RNAse H (which specifically removes the DNA-RNA hybrids) alleviated the DNA damage[235].

PHF6 co-precipitates with retinoblastoma-binding protein 4 (RBBP4, also known as retinoblastoma associated protein 48 RBAP48, or chromatin assembly factor-1 p48 subunit CAF-1 P48. Nur55 is the *Drosophila* homolog of RBBP4)[236]. RBBP4 is a component of both the Nucleosome Remodeling and Deacetylase (NuRD) complex and the Polycomb Repressive Complex 2 (PRC2). Structural studies showed that the PHF6 peptide (amino acid 162-170) binds the top surface of the seven bladed β propeller fold (called c-site) of RBBP4[237]. Arg-163 of PHF6 is directly inserted into the central channel of RBBP4, forming a bond with Glu-231 of RBBP4. The same Glu-231 also binds the arginine residue in Histone H3 (Arg-2) and FOG-1 (Arg-4 of Friend of GATA protein-1).

It has been shown that disruption of this Glu-Arg interaction between FOG-1 and NuRD reduces adult-type globin gene expression in murine definitive erythroid cells, highlighting a role of NuRD in activating adult haemoglobin expression[238]. In PRC2, the cofactor AEBP2 also binds to the c-site by

mimicking an unmodified H3 tail[239, 240]. Binding of AEBP2 to SUZ12, a core component of PRC2, is important in directing PRC2 to chromatin[241]. Whether PHF6 competes with AEBP2 in RBBP4 binding is unknown.

In the lymphoid lineage PHF6 acts as a lineage-specific regulator. Knockout of Phf6 promotes T-cell tumour growth and suppresses B-cell leukaemia growth. Conversely overexpression of Phf6 suppresses T-cell tumour growth[229]. The loss of Phf6 in B-cell leukaemia cells results in changes in chromosome accessibility, with regions for motifs of transcription factors associated with T-ALL and AML development becoming more accessible. Integrative analysis of ChIP-seq and RNA-seq showed that Phf6 localises to gene bodies and promoter/enhancer regions of many genes and Phf6 binding to transcription start site in highly expressed genes is enriched. Altogether these data suggest that Phf6 can regulate lineage-specific gene expression by modulating chromatin accessibility.

3.7.3 PHF6 mutation in ATMDS

In *PHF6*, the pattern of mutation in ATMDS is similar to published reports in that they are either truncating or splice site mutations, or missense mutations in the second PHD domain[161, 242] (Table 3-6). This is also suggestive of a loss-of-function effect in ATMDS. The amino acid positions of all missense mutations detected in this cohort (Glu271, Gly287, Gly291, Cys297 and His302) are all highly conserved across vertebrate species. Multiple PHF6 mutations within the same individual is noted in 13.2% of cases in myeloid neoplasm, suggestive of multiple subclones[161]. In the

ATMDS cohort, three male patients (case# 81, 90 and 94, Table 3-3) had more than one PHF6 mutations detected. Within each case there is at least one nonsense mutation.

Location of mutation (5' to 3')	PHF6 Mutation
	p.Cys20Leufs*13
	p.Asp25Glnfs*10
	p.Glu27Asnfs*6
	p.Leu32Thrfs*4
PHD domain 1 (residues 17-131)	c.241-1G>A
	p.His104Ilefs*9
	p.Arg116*
	c.374+1G>A
	c.418+1G>A
NoLS (residues 157-169) and RBBP4 binding motif (residues 162-170)	p.Arg163Profs*55
	p.Leu175Cysfs*43
	p.Gly186Glufs*32
	c.585+2T>C
PHD domain 2 (residues 208-333)	p.Arg225*
	c.730-2A>G
	p.Gln251*
	p.Glu271Ala
	c.834+1G>A
	p.Gly287Asp
	p.Gly291Arg
	p.Gly291Glu
	p.Cys297Ser
	p.His302Tyr
	p.His302Leu
	p.Ile314Thr
	p.Arg319*

Table 3-6. Distribution of PHF6 mutations in ATMDS. Location of the PHD domains are based on crystal structure by Liu *et al*[240]. The location of the PHD domain 2 here was extended from the predicted PHD domain in UniProKB (residues 277-329). Construct of this extended version was stable in solution, whereas the predicted domain was unstable. NoLS: Nucleolar localising sequence. In PHF6 NoLS overlaps with RBBP4 binding motif[241]. Reference sequence: NM_032458.2, NP_115834.1.

3.8 Clinical investigations of ATMDS patients

In this project, fresh blood samples were available from five male patients (NQ, case# 118; FW, case# 120; ECT, case# 122; MK, case# 123, also described in a report by Lannon and Wells[243] and KH, case# 125). These samples provided an opportunity to study the characteristics of ATMDS that was previously impossible. Table 3-7 shows the serial haematology results obtained in these five cases.

Case	Date	Hb (11-17 g/dL)	Hct (37-54%)	MCV (80-100fL)	MCH (27-32pg)	Platelet (150-500 x10^3/μL)	Leucocyte (4.0-10 x10^3/μL)	BCB (%)
NQ (#118)	12/13	**8.9 L**	**29.9 L**	**59 L**	**17.4 L**	476	4.7	-
	01/14	**8.8 L**	**29.6 L**	**62 L**	**18.4 L**	**612 H**	**3.7 L**	24.5
	10/14	**8.3 L**	**27.9 L**	**64 L**	**19.1 L**	**842 H**	**3.7 L**	23.5
	08/15	**10.1 L**	**32.3 L**	**74 L**	**23.2 L**	**140 L**	4.9	13.6
	02/16	**10.2 L**	**33.8 L**	**77 L**	**23.1 L**	200	4.4	6.6
FW (#120)	10/14	**9.1 L**	**28.6 L**	**75 L**	**24 L**	**106 L**	**1.9 L**	27.9
	08/15	**7.6 L**	**24.5 L**	**76 L**	**23.5 L**	**43 L**	**1.4 L**	20.2
	02/16*	**9.9 L**	**30.8 L**	84	**26.9 L**	**22 L**	**1.4 L**	6.1
ECT (#122)	02/11	16	46.3	97.1	**33.5 H**	163	7.7	-
	11/12	15.6	48	97.3	31.7	149	6.9	-
	04/15	**10.7 L**	45.7	**61 L**	**14.4 L**	**103 L**	**45 H**	-
	05/15	**9.2 L**	**36.7 L**	**61.7 L**	**15.5 L**	**42 L**	**14.6 H**	-
	06/15	**7.8 L**	**28.9 L**	**60 L**	**16.2 L**	-	**11.5 H**	79.8
MK (#123)	08/16	**8.5 L**	**26.8 L**	**76 L**	**24.3 L**	**69 L**	**16.8 H**	42.1
	12/16	**9.9 L**	**31.1 L**	**78 L**	**24.8 L**	**77 L**	**23.8 H**	40.5
	07/17*	**9.7 L**	**29.9 L**	87	28.3	**19 L**	**84.5 H**	
KH (#125)	01/17	12.8	42.9	**74 L**	**22.1 L**	417	**3.4 L**	29.7
	04/17	11.7	40.6	**77 L**	**22.1 L**	**865 H**	4.5	24.7

Table 3-7. Serial haematology for ATMDS patients. For case #122 two blood counts pre-MDS between 2011 and 2012 were shown. Unless indicated, blood samples were taken more than 1 month post-transfusion to avoid contamination by donor RBCs. Hb: haemoglobin, Hct: haematocrit, MCV: mean corpuscular volume, MCH: mean cell haemoglobin, BCB: brilliant cresyl blue. *blood sample was taken <1 week after transfusion. Abnormal values are in bold.

Case NQ (#118) and FW (#120): these two patients were referred from the same centre with an initial diagnosis of refractory anaemia in 2013. Haematology showed a hypochromic microcytic anaemia requiring frequent blood transfusion. Both had genetic testing for alpha thalassaemia which was negative. In NQ, SNP array of peripheral blood showed a region

of copy number-neutral loss of heterozygosity in chromosome 1p (see Supplementary Figure 1). Loss of heterozygosity in chromosome 1p36 is frequent in MDS (13%) and MDS transforming to AML (30%)[244-246]. Other chromosomes appeared normal. SNP array analysis for FW is detailed in the next section. Both patients transformed to AML and died in September and March 2016, respectively.

Case ECT (#122): This patient has one of the highest percentages of cells with Hb H inclusions in this cohort. The patient, who had normal blood count previously in 2011-2012 (Table 3-7), first presented in April 2015 at age 83 with hypochromic microcytic anaemia. Alpha thalassaemia screen was negative. Haemoglobin high performance liquid chromatography (HPLC) and isoelectric focusing (IEF) confirmed the presence of Hb H (Figure 3-6). Bone marrow biopsy confirmed RCMD with normal karyotype. Patient was managed expectedly and received regular transfusion. He died in April 2016, one year after initial diagnosis.

Case MK (#123): Patient was 68 years old when diagnosed with ATMDS in March 2014. He presented with ongoing malaise after suffering from an episode of pneumonia. Bone marrow showed hypercellular particles and trails, dysplastic changes in megakaryocytes were seen. Granulopoiesis was expanded and left-shifted, and many pseudo-Pelger-Huët anomaly cells were seen. Erythropoiesis was active with dysplastic changes and delayed nuclear maturation. Blasts <5%. The diagnosis was consistent with RCMD. His anaemia was managed by regular erythropoietin injection. Repeat bone

marrow in October 2015 showed stable disease. His anaemia however worsened. Erythropoietin was stopped and he was started on regular blood transfusion from December 2015 (2-3 units every 2-3 weeks). A third bone marrow biopsy in March 2016 showed secondary fibrosis with 5% blasts, normal karyotype. No mutations in *CSF3R* exons 14 and 17 were detectable by Sanger sequencing. Anti-glycophorin antibody demonstrated decreased and dispersed erythropoiesis. The last blood sample in July 2017 was taken after the patient transformed into AML (see AML transformation in ATMDS). Patient died in March 2018.

Case KH (#125): This patient was diagnosed with MDS/MPN overlap disease. Bone marrow showed normal karyotype. HPLC showed an elevated HbF of 14%, which could be related to use of hydroxyurea. A later bone marrow showed progression to RAEB2 (Refractory Anaemia Excess Blast) with 13% blast in bone marrow in October 2017.

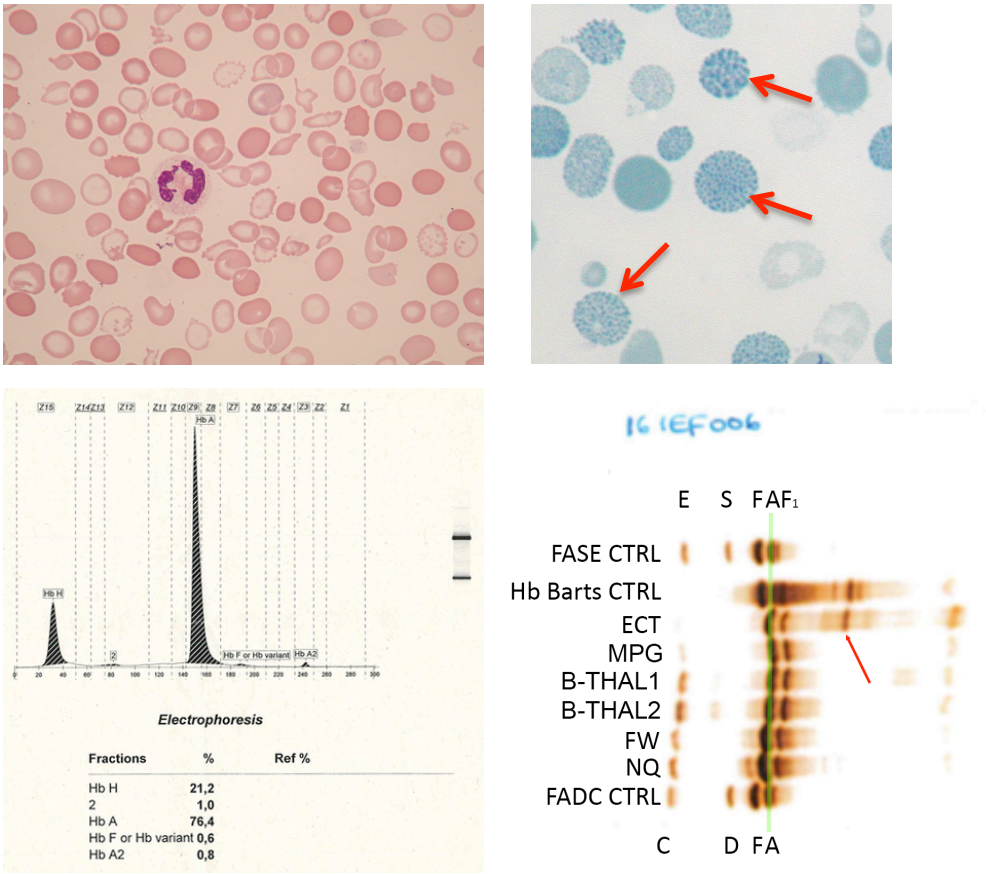


Figure 3-6. Top left and right – modified Wright's stain and BCB stain on peripheral blood of ECT (#122). Hypochromic, microcytic RBCs can be seen with significant anisocytosis, poikilocytes and red cell fragments. BCB stain showing numerous Hb H inclusions within RBCs (arrowed). Bottom left - Haemoglobin HPLC showing the distinctive peak of Hb H in ECT (blue arrow). Bottom right – Haemoglobin IEF from three ATMDS patients (FW case# 118, NQ case# 120 and ECT case# 122). Hb A band is highlighted in green. A strong band at the Hb Barts position (red arrow) can be seen in ECT (BCB% 79.8%) but not the other two patients with lower BCB% (12 and 27%, respectively). An Hb F1 (acetylated Hb F) band can be seen in all three ATMDS samples immediately to the right of Hb A band. The band running adjacent to the 'C' position in FW and NQ is Hb A2 (α2δ2). MPG is a MDS patient. B-THAL1 and 2 are two patients with beta thalassaemia.

3.9 ATMDS mutations at a single cell level

First we asked the question whether the four highly enriched genes are mutated in the same cell. To answer this we took DNA extracted directly from CD34⁺ for panel sequencing. The VAF in Table 3-8 shows that the four genes of interest had very high VAF – for the two genes in the X chromosome (*ATRX* and *PHF6*) it was close to 100%, whereas for the two autosomal genes (*ASXL1* and *U2AF1*) the VAF was close to 50%, with the

exception of p.Ser34Phe mutation where the VAF was lower suggesting a subclonal architecture for this particular mutation. The VAF data for *ASXL1* mutation are less reliable as they are typically insertions or deletions at a homopolymer region.

In two cases (NQ, case #118 and ECT, case #122) subclonal mutations in *SETBP1* and *KRAS* can be observed. *SETBP1* mutations are of interest as they are detected in the two cases with the highest percentages of Hb H inclusion in this cohort, one of which is ECT. In addition, in the two cases here the VAFs in *SETBP1* seem to correlate with the percentages of Hb H inclusion as well. In RCMD patients with *ASXL1* or *U2AF1* mutations serial sampling suggested that these gain of function mutations in *SETBP1* play a role in leukaemic transformation, especially in *ASXL1*-mutated MDS patients where *SETBP1* is co-mutated in 9.38% of cases[247, 248]. Indeed, in an *ASXL1*/*SETBP1* double knock-in mouse model the mutant SETBP1 augmented the differentiation block induced by *ASXL1* mutation, reduced apoptosis and increased myeloid colony output[248]. A significant decrease in haemoglobin level was observed in double mutant mice compared to single mutants, but in contrast MCV was increased significantly. Given the lack of a significant enrichment in ATMDS patients, subclonal mutations in *SETBP1* are unlikely to induce acquired alpha thalassaemia *per se* in the background of *ASXL1* mutation.

The high VAFs suggested mutations in *ATRX*, *ASXL1*, *U2AF1* and *PHF6* are in the same cell. To confirm this, $CD34^{+}$ $CD3^{-}$ $CD19^{-}$ cells were FACS sorted as

single cell into Terasaki plates and expanded for 7 days. Whole genome amplification (WGA) was performed directly from clonal populations and presence of mutations in these four genes was verified by Sanger sequencing (see Supplementary Figures 2 and 3). In NQ (case #118) and ECT (case#122), 15 out of 15 clonal populations sequenced from each case were shown to contain all four mutations.

In both cases with U2AF1 p.Ser34Phe mutation (case#120 and 125), the four enriched genes were not found to be all mutated. In FW (case#120), initially there was no *ATRX*, *ASXL1* or *PHF6* mutation detectable on both Ion Torrent and myeloid panel sequencing. For *ATRX*, one variant p.Ala914Thr was detected but the VAF was below the reportable threshold at 3%. Upon comparing raw sequencing data between peripheral blood and $CD34^+$ cells we found a somatic deletion covering the whole *PHF6*. We then investigated further via SNP array on paired tumour ($CD34^+$ cells) and normal (peripheral blood) samples (Figure 3-7). The SNP array confirmed a deletion in *PHF6*. In addition a region of loss of heterozygosity involving 3' end of *ASXL1* was suggested. No deletion in *ATRX* was detected.

In KH (case#125), no *ASXL1* truncating mutation was detectable. However a nonsense mutation in *EZH2*, also involved in chromatin modification, was detected along with the *JAK2* activating mutation. These two genes are frequently co-mutated in patients with myelofibrosis[249]. The loss of Ezh2 in $Jak2^{V617F}$ knock-in mice has been shown to inhibit erythropoiesis, increase megakaryopoiesis, and accelerated the development of

myelofibrosis[250]. However, clinically the only mutation associated with anaemia in myelofibrosis was *U2AF1*. There was no association between *ASXL1*, *SF3B1*, *SRSF2*, *TET2*, *EZH2*, *ZRSR2*, *IDH1/2* mutations and anaemia[251].

3.10 AML Transformation in ATMDS

We were able to obtain a second sample from one case (MK, case#123, sample obtained in July 2017) after the disease transformed into AML. The patient became progressively more transfusion-dependent, requiring three units of RBC per week. Myeloid panel sequencing of $CD34^+$ cells obtained from a post-transformation sample showed that the four mutations remained present, and the VAFs remained comparable to the pre-transformation sample (Table 3-8). Additionally, a homozygous missense mutation (p.Leu380Pro) in CBL was detected. The apparent homozygosity of this mutation could be due to acquired uniparental disomy known in this region (chr 11q)[252], but this had not been confirmed by SNP array analysis.

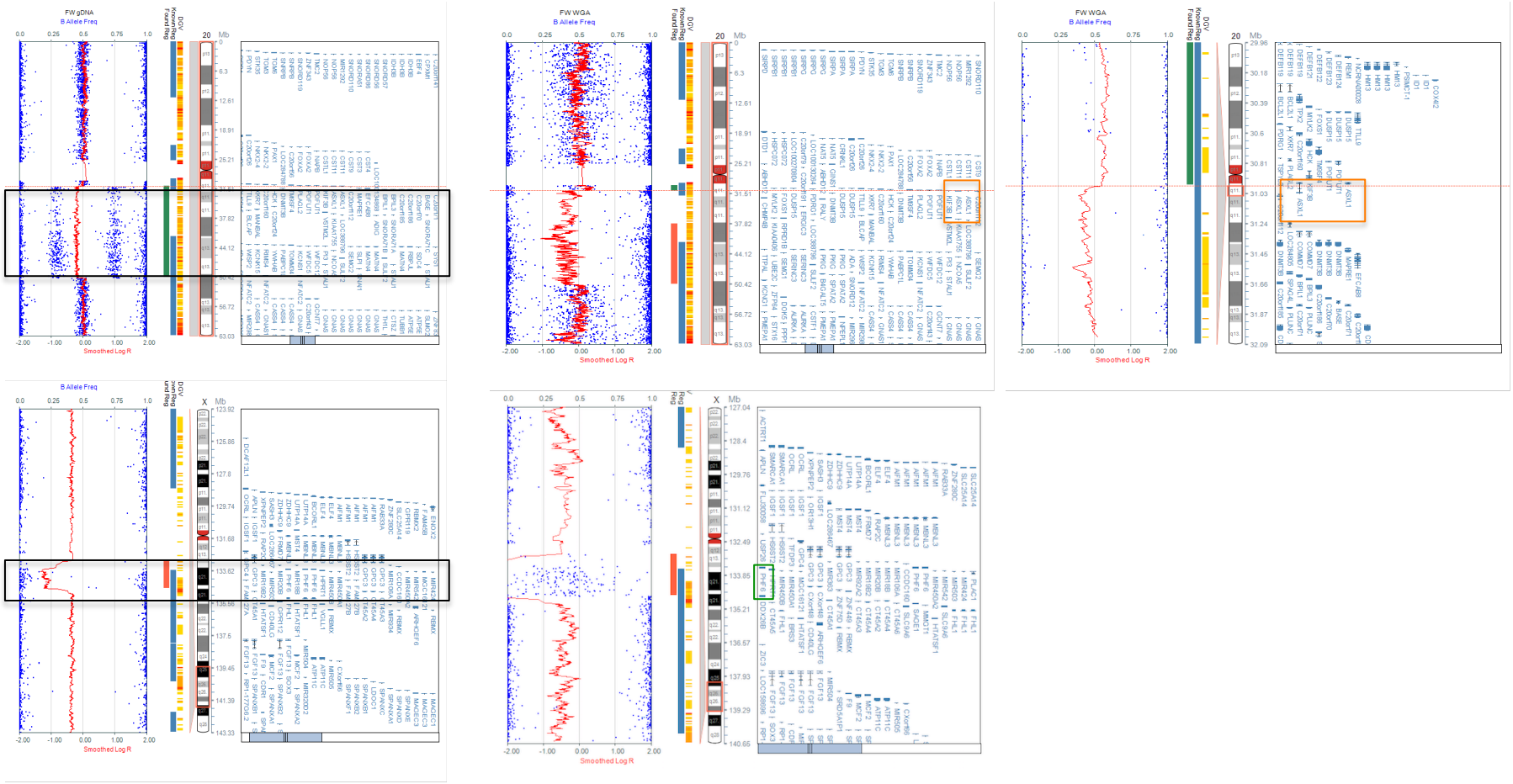


Figure 3-7. Tumour-normal paired SNParray in patient FW (case #120). Top left: whole chromosome view of chromosome 20 in genomic DNA showing a long stretch of CNV in a mosaic pattern in 20q11.21-q13.2 (black box); Top middle: the same whole chromosome view in whole genome amplified (WGA) DNA from CD3- CD19- CD34+ cells showing loss of heterozygosity (LOH) in the same region. Top right: close up view on the centromeric boundary of LOH, showing ASXL1 involvement. Bottom left and right: deletion in Xq26.2-26.3 (black box) involving the entire PHF6 gene detectable in both the genomic and WGA sample, respectively.

Sample	Initial	BCB (%)	ATRX	ASXL1	U2AF1	PHF6	KRAS	JAK2	CBL	SETBP1	EZH2
118	NQ	12	p.Cys1576Ser (99.41%)	p.Gly646TrpfsTer12 (42.67%)	p.Gln157Pro (48.13%)	p.His104IlefsTer9 (96.38%)	p.Tyr71Asp (15.43%) VUS	ND	ND	p.Gly870Arg (5.59%)	ND
120	FW	27	ND	del(20)(q11.21-13.2)	p.Ser34Phe (18.78%)	del(X)(q26.2-26.3)	ND	ND	ND	ND	ND
122	ECT	79.8	p.N2340Vfs*2 (100%)	p.Gly646TrpfsTer12 (39.97%)	p.Gln157Pro (48.28%)	c.585+2T>C (98.14%)	p.Gly60Asp (6.57%)	ND	ND	p.Asp868Asn (48.37%)	ND
123 (pre-AML)	MK	40.5	p.Trp263Cys (99.23%)	p.Arg634LysfsTer24 (50.59%)	p.Gln157Pro (47.2%)	p.Arg319Ter (99.41%)	ND	ND	ND	ND	ND
123 (post-AML)	MK	42.1	p.Trp263Cys (99.75%)	p.Arg634LysfsTer24 (48.78%)	p.Gln157Pro (50.4%)	p.Arg319Ter (99.66%)	ND	ND	p.Leu380Pro (99.69%)	ND	ND
125	KH	24.7	p.Asp2035Val (97.44%)	ND	p.Ser34Phe (5.46%)	p.Gly291Glu (97.21%)	ND	p.Val617Phe (25.18%)	ND	ND	p.Gln735Ter (45.83%)

Table 3-8. Somatic mutation in ATMDS obtained by sequencing of CD34⁺ cells. VAFs of each mutation is shown in bracket. The pre and post-AML sample in patient MK showed an additional CBL mutation. U2AF1 p.Ser34Phe mutation in both patient FW and KH showed a lower VAF compared to those with p.Gln157Pro mutation. ND: no mutation detected. VUS: Variant of unknown clinical significance.

3.11 Summary

In this chapter, we have confirmed a previous observation that mutations in *ATRX*, *ASXL1*, *U2AF1* and *PHF6* are highly enriched in ATMDS. The prevalence of mutations in *U2AF1*, *ASXL1* and *PHF6* in ATMDS is extraordinary high and well above the prevalence of the most common mutations detected in MDS. In agreement with its role in alpha thalassaemia, the presence of *ATRX* mutation in addition to mutations in the other three genes tend to involve patients with more severe acquired alpha thalassaemia as measured by the percentage of red blood cells with haemoglobin H inclusions. However, it could also be possible that in those patients with low percentages of Hb H inclusions the clone containing *ATRX* mutation was below detection limit. We did not find significant enrichment of other genes commonly involved in myeloid neoplasm. Using variant allele frequency as an indication of subclonal architecture it shows that *U2AF1* tends to be the earlier mutation acquired in pathogenesis, while the two X chromosome genes *ATRX* and *PHF6* tend to be later mutations.

The pattern of mutations in *ATRX* and *PHF6*, both of which contain zinc finger domains, is consistent with loss of function. In contrary, truncations in the last exon of *ASXL1* and hotspot mutations in *U2AF1* are gain of function mutations. These patterns are similar to what is observed in MDS. However the high prevalence of p.Gln157Pro mutation in U2AF1 contrasts with other myeloid neoplasms, where p.Ser34Phe mutation is more common. This is also reflected in the two ATMDS patients (FW and KH) where the p.Ser34Phe mutation is subclonal by VAF, but the cases with p.Gln157Pro are always close to 50% in

$CD34^+$ samples. This has potential implication in the pathogenesis of ATMDS as the two mutations adopt different splice site sequence preference.

The presence of two of these four genes on the X chromosome may explain the strong male bias observed in ATMDS. If the wildtype *ATRX* is re-activated in female due to age-related reactivation of the X chromosome, this could provide evidence in support of this male bias. This was however shown not to be the case by the study in Haas *et al*[253], who showed no change in methylation pattern at the *ATRX* promoter across different age groups in normal female and in two female ATMDS patients. For *PHF6* however, it is subjected to X-inactivation and preferential inactivation of the mutant allele was observed in female carriers of BFLS[216, 218, 219, 225, 254, 255]. Therefore, it is quite possible that *PHF6* is a contributing factor to the male bias in ATMDS.

Exploration into the temporal sequence of mutations and subclonal architecture with corresponding haematology data would provide useful information to delineate the role of each mutation in the pathogenesis of acquired alpha thalassaemia and disease progression. The acquisition of *CBL* mutation upon AML transformation in MK is a good example. Expansion of single cell clones did not detect any subclonal architecture within the four genes in the two samples (NQ, case #118 and ECT, case #122) examined, in agreement with the high VAFs detected in sequencing. A likely explanation is that by the time the acquired alpha thalassaemia became clinically significant, this dominant clone was already well established. It is worth noting that even within the three patients with all four genes mutated (NQ, ECT and MK), there is a vast variation in percentages of

cells with haemoglobin H (12, 79.8 and 40.5%, respectively). Other detected mutations (*SETBP1*, *KRAS*) does not seem to explain this variation. However, we have not accounted for cytogenetic abnormalities in every sample.

In the next chapter, we will perform erythroid differentiation on $CD34^+$ cells isolated from ATMDS patients to see if the alpha thalassaemia phenotype can be recapitulated *ex vivo*.

4 Erythroid differentiation of ATMDS CD34$^+$ progenitor cells

We were able to compare erythroid differentiation of MACS separated CD34$^+$ cells between ATMDS (n=3), MDS without alpha thalassaemia (n=2) and normal controls (leucocyte reduction system cones obtained from blood donors. n=5). In ATMDS, CD34$^+$ cells from both patients with the U2AF1 p.Ser34Phe (FW, case#120 and KH, case#125) failed to expand.

4.1 Globin gene expression analysis in peripheral blood

Before we can compare haemoglobin expression by RT-PCR in cultured cells, we first looked at haemoglobin expression in unfractionated peripheral blood of ATMDS patients to see if the phenotype can be demonstrated by RT-PCR (Figure 4-1). The presence of white blood cells in the sample is not expected to affect the readout as they do not express haemoglobin. Normal controls and patients with constitutional alpha thalassaemia were used for comparison. We calculated the alpha/beta globin gene expression ratio (expression of *HBA* divided by expression of *HBB*) as a marker for haemoglobin chain imbalance. Expression ratio of alpha/beta globin gene showed a significant reduction in ATMDS versus normal control. Expression of gamma globin was minimal in all cases and did not affect the overall ratio (<5% difference between alpha/beta ratio and alpha/non-alpha ratio. Non-alpha component is the sum of beta and gamma globin gene expression.

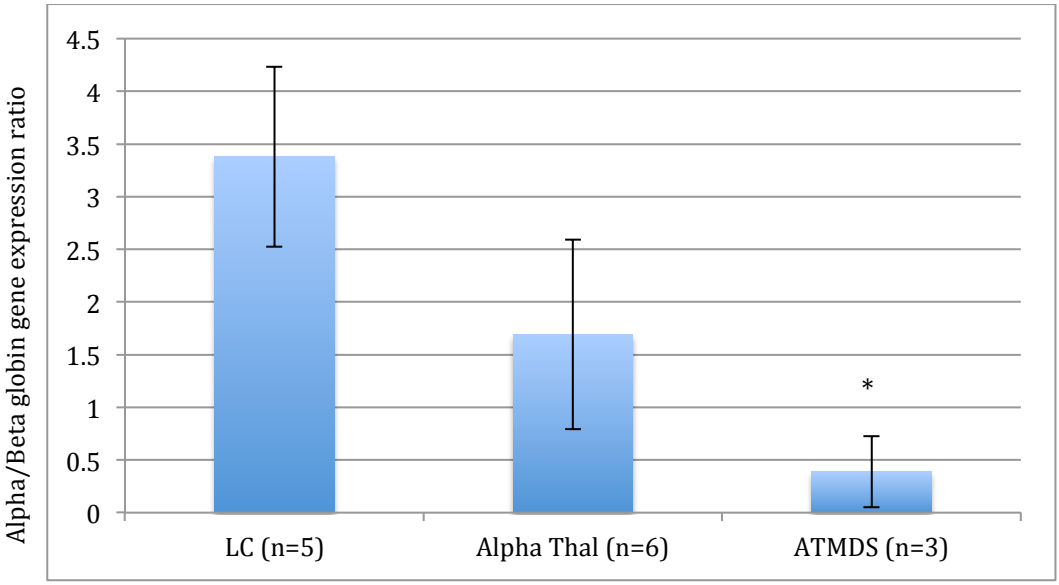


Figure 4-1. Alpha/Beta globin gene expression ratio in peripheral blood of normal leucocyte cone control (LC), ATMDS patients and patients with constitutional alpha thalassaemia (Alpha Thal). *:p<0.005 versus LC, Student's T-test, two-tailed.

4.2 Block in erythroid differentiation in ATMDS

CD34$^+$ cells of normal control and an ATMDS patient MK (case#123) were differentiated for up to 16 days. Serial samples were obtained for FACS, morphological and globin gene expression analyses. Expansion of culture during erythroid differentiation was less efficient in ATMDS as shown in Figure 4-2.

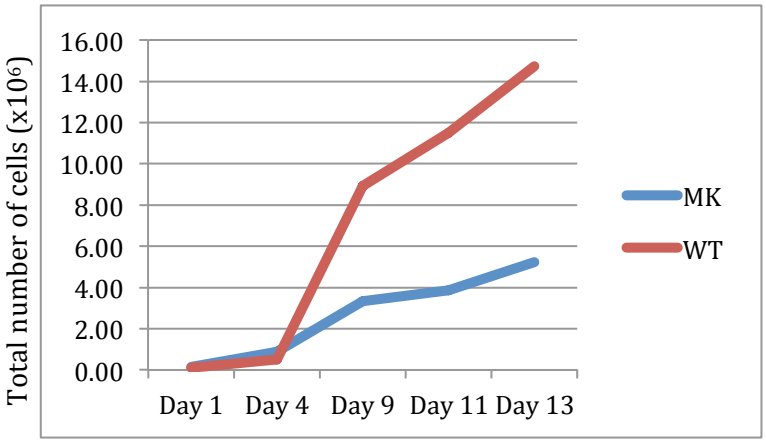


Figure 4-2. Growth curve during erythroid differentiation in ATMDS patient MK (blue) and wildtype control (red). Total cell count at start of culture was 0.15x10^6 and 0.12x10^6 for ATMDS patient MK and wildtype, respectively.

Serial FACS analysis showed that in both the ATMDS patient and normal control, CD34 marker became mostly negative by day 10 of differentiation (Figure 4-3).

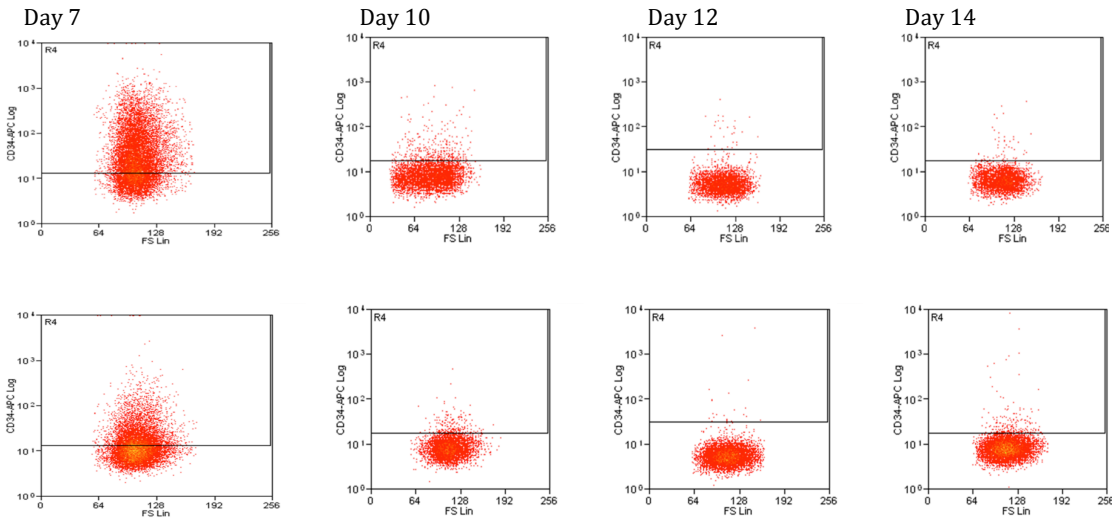


Figure 4-3. Serial FACS analysis showing the loss of CD34 positivity during erythroid differentiation. Top panel – normal controls. Bottom panel – ATMDS patient (MK, case #123). In both groups CD34 marker became negative between Day 7 and 10.

As for the erythroid markers, in wildtype controls both erythroid markers CD71 and CD235a became mostly positive by day 9, and CD71 appropriately progressed towards negative again as differentiation proceeded further to day 16[256]. In ATMDS patient there are clearly two populations – a double positive and a double negative population and this observation persisted throughout differentiation (Figure 4-4). Morphology shows a block in differentiation phenotypically in ATMDS patient. Most cells remain as late basophilic erythroblasts at day 16, with no orthochromatic erythroblast visible (Figure 4-5). In the normal control preparation of cytospin slides was less satisfactory due to cell loss towards the end of differentiation.

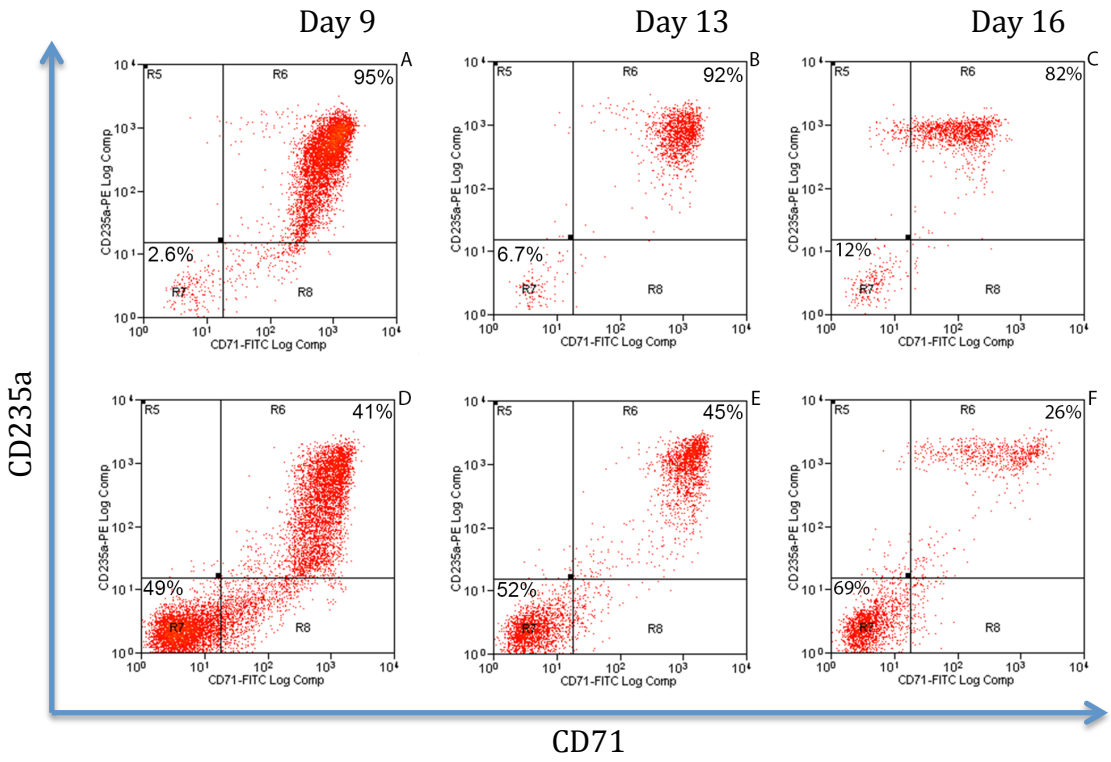


Figure 4-4. Serial FACS analysis on erythroid markers CD235a and CD71. Top pane: normal control (A to C). Bottom pane: ATMDS patient MK (D to F). Percentages of gated population (R7: $CD71^{-}$ $CD235a^{-}$; R6: $CD71^{+}$ $CD235a^{+}$) are shown in the plot.

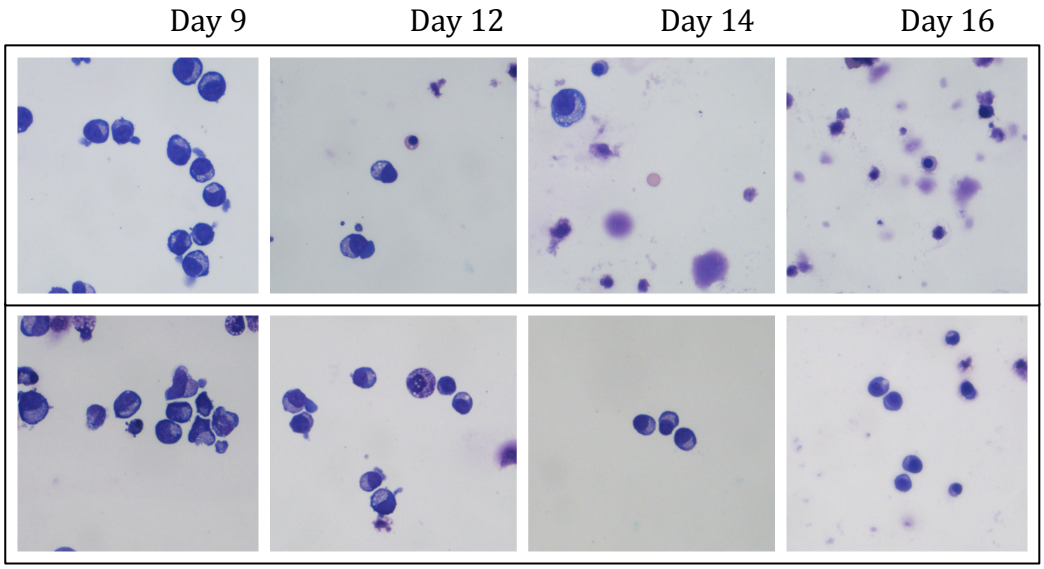


Figure 4-5. Cytospin slides of normal control (top) and ATMDS patient (bottom) during differentiation. There was substantial cell loss in the normal control towards the end of differentiation.

Globin gene expression was also measured in the bulk population during differentiation (Figure 4-6). Alpha globin gene was repressed throughout differentiation in ATMDS. A drop in the alpha to both beta and non-alpha (i.e. sum of beta and gamma) globin gene expression ratio was observed in all three time points (Figure 4-6B and C). An increase in gamma globin gene expression in ATMDS was particularly pronounced towards the end of differentiation, contributing to the larger difference in the alpha/non-alpha globin gene expression ratio between normal control and ATMDS.

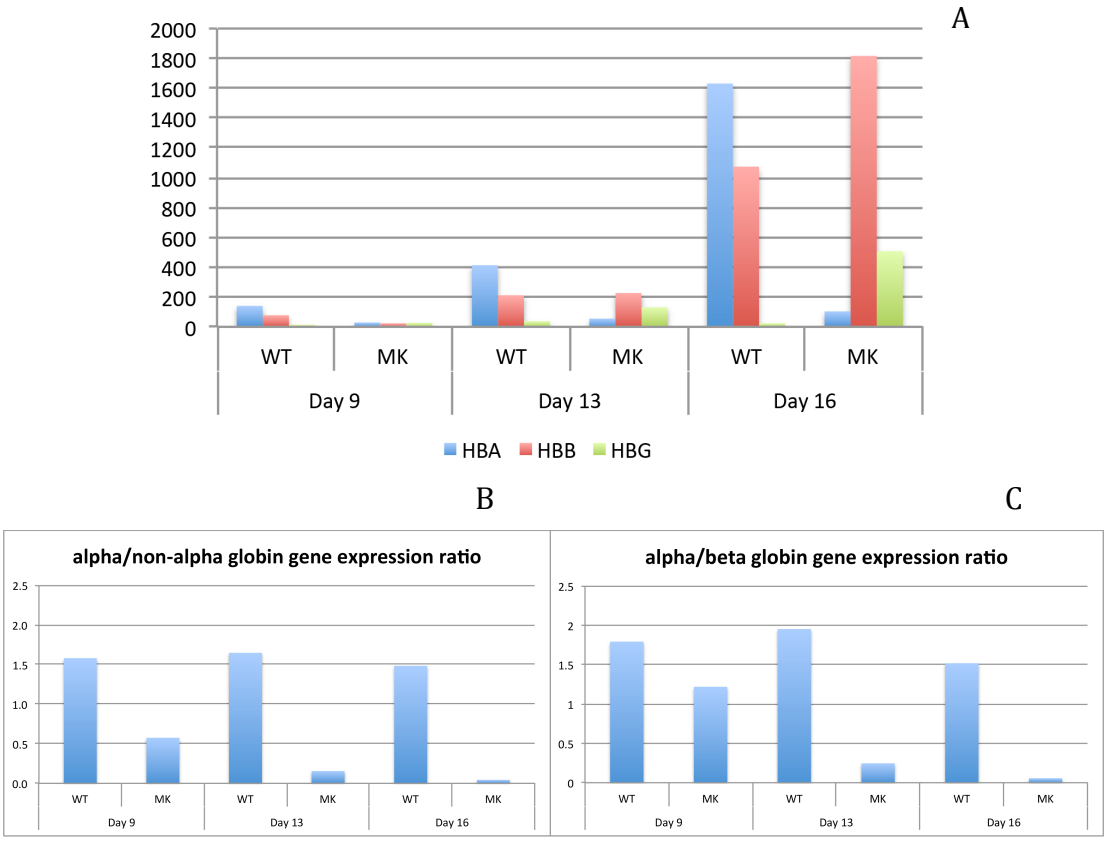


Figure 4-6. A: globin gene expression during erythroid differentiation. Expression of individual globin genes (*HBA*, *HBB* and *HBG*, respectively) was measured on days 9, 13 and 16 during differentiation in wildtype normal control (WT) and one ATMDS patient (MK). Expression is normalised as fold change over the housekeeping gene *RPL13A*; B: alpha/non-alpha globin gene expression ratio between normal control and ATMDS patient; C: alpha/beta globin gene expression ratio between normal control and ATMDS patient.

4.3 Separation of erythroid and non-erythroid populations during differentiation by FACS sorting

The large proportion of undifferentiated cells in ATMDS prevents comparable assessment on the erythroid population. We FACS sorted the double positive and double negative populations and processed the sample for globin gene expression analysis and ATAC-seq. Day 10 was used as the cutoff as further differentiation resulted in low viable cell yield upon FACS sort. For wildtype controls there was insufficient number of cells in the double negative populations for downstream assays (Figure 4-7).

To determine if MDS without mutations in the four enriched genes could itself produce an alpha thalassaemia phenotype, two $CD34^+$ samples from MDS cases without alpha thalassaemia were also differentiated (samples provided at the courtesy of Professor Jacqueline Boultwood and Dr Andrea Pellagatti, Nuffield Division of Clinical Laboratory Science, University of Oxford). The diagnosis for both cases (MDS591 and 629) was refractory cytopenia with multilineage dysplasia with normal karyotype (46,XY). Blast count was between 0-2%. Myeloid panel sequencing did not detect any mutation in MDS629, while the only mutation detected in MDS591 was a TET2 p.Gly1361Arg mutation (VAF 25.81%). While a small double negative population was visible in MDS591, there was not enough material for downstream assays. MDS629 behaved similarly to wildtype controls (Figure 4-7).

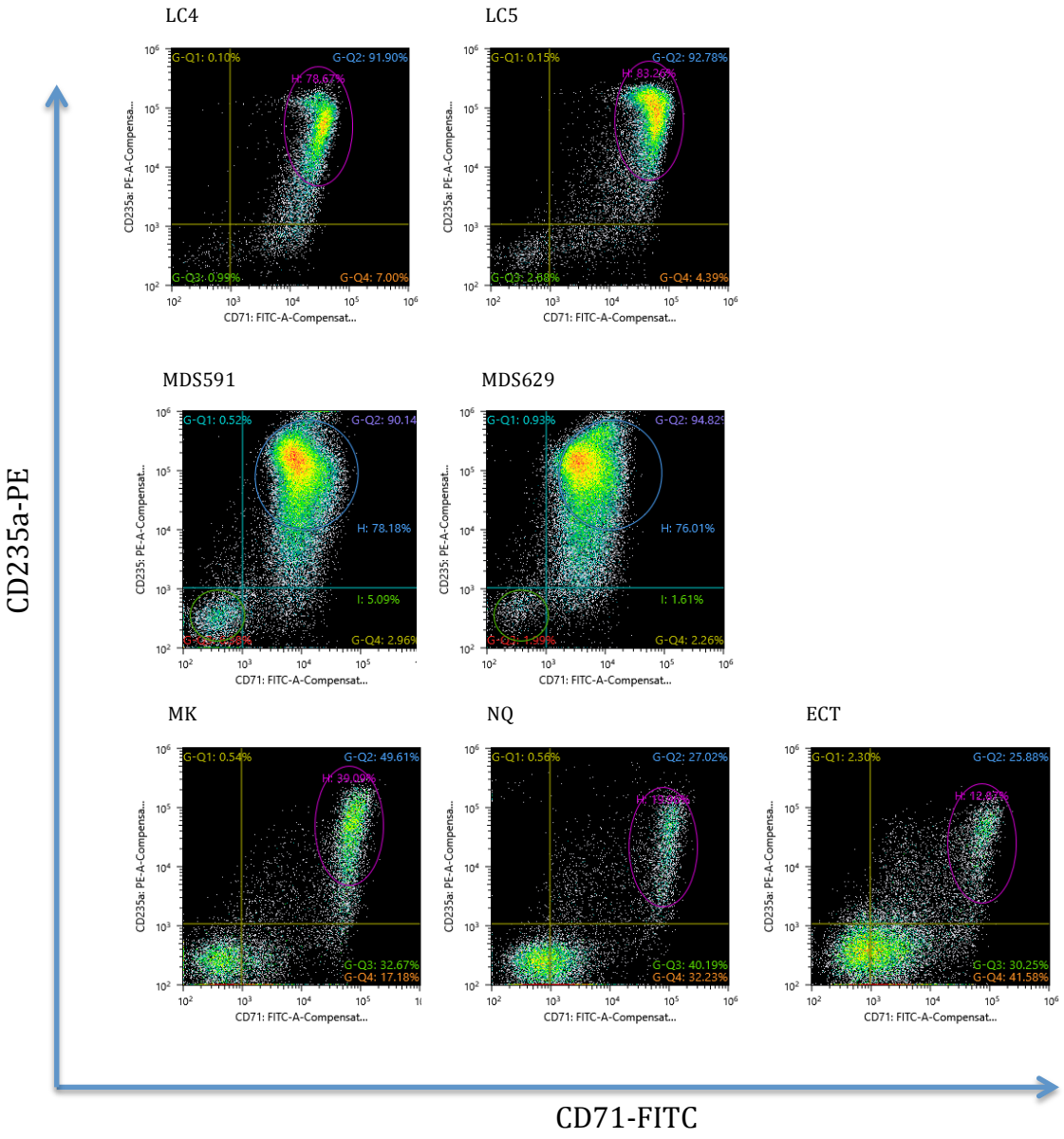


Figure 4-7. FACS analysis of normal controls (LC4, LC5), MDS without acquired alpha thalassaemia (MDS591 and MDS 629) and three ATMDS patients (MK, NQ and ECT) on Day 10 of erythroid differentiation. ATMDS patients all have a significant double negative population, whereas in the two MDS cases here the majority of cells do differentiate and become double positive.

4.4 Low haemoglobin expression and alpha/non-alpha globin gene expression ratio in ATMDS erythroid population

In ATMDS, measurement of individual globin gene expression on Day 10 in the $CD71^+$ $CD235a^+$ double positive population showed that expression of alpha globin was indeed lower in ATMDS compared to normal controls or MDS samples, whereas expression of gamma globin was increased in the MDS and ATMDS groups but the difference was not statistically significant (Figure 4-8A). Unexpectedly, expression of beta globin was also reduced in ATMDS. Expression of all globin genes in the double negative population was lower than that of the housekeeping gene *RPL13A* (data not shown).

A reduction in alpha/non-alpha ratio can be seen in ATMDS patients (Figure 4-8B). However, if gamma globin expression was excluded from analysis, alpha/beta globin expression ratio was not significantly different between the three groups (Figure 4-8C). We also looked at expression of embryonic haemoglobins (*HBZ* and *HBE*) and neither were expressed (data not shown). Therefore the low expression of *HBB* and the conversely increased expression of *HBG* was not due to reactivation of embryonic globin gene programme.

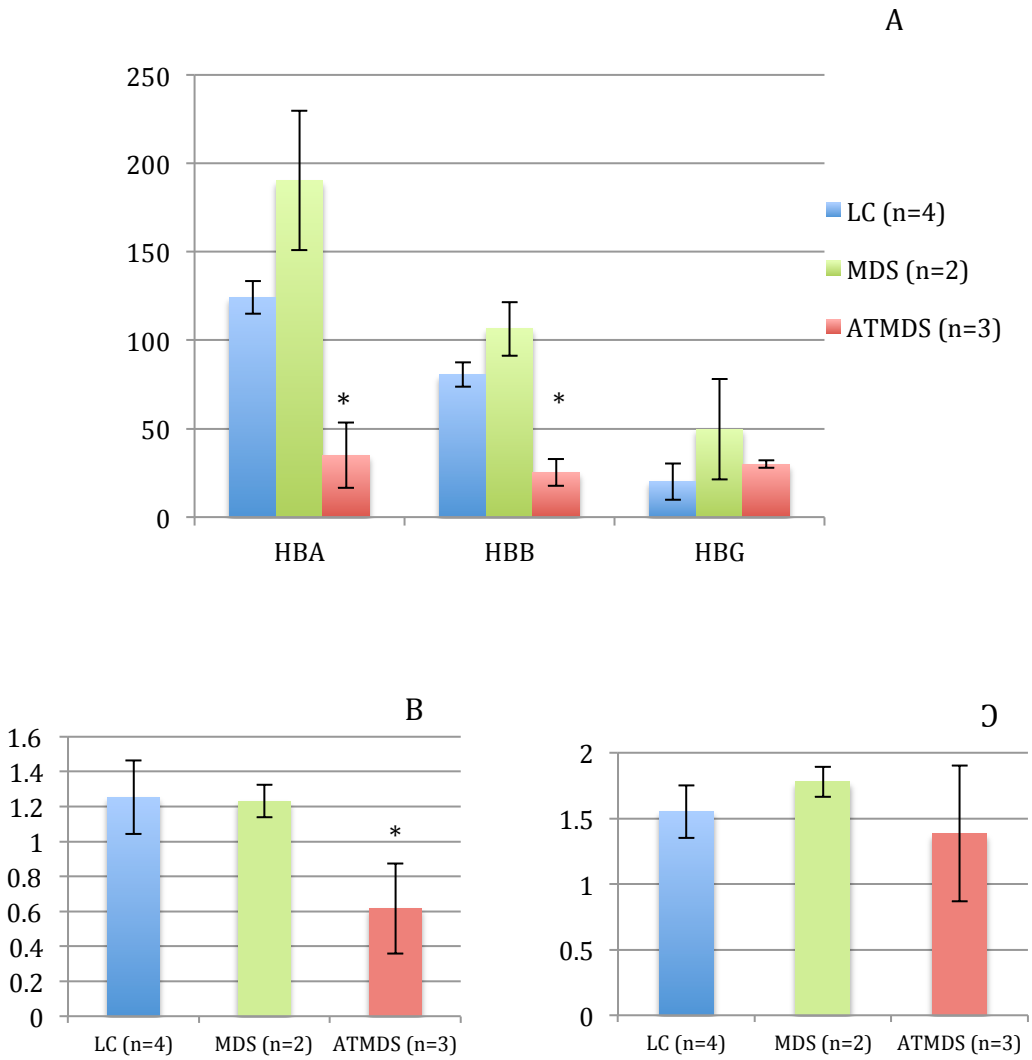


Figure 4-8. A: Individual globin gene expression on day 10 of erythroid differentiation in CD71⁺ CD235a⁺ cells of normal leucocyte cone controls (LC, blue), MDS without alpha thalassaemia (MDS, green) and ATMDS patients (red). Expression is normalised as fold change over *RPL13A*; B: Alpha/non-alpha globin gene expression ratio between the three groups (LC: 1.25, MDS: 1.23 and ATMDS: 0.62); C: Alpha/beta globin gene expression ratio between the three groups (LC: 1.55, MDS: 1.78 and ATMDS: 1.38). There was no significant difference in alpha/beta globin gene expression ratio between the three groups. *:$p<0.05$ versus LC, Student's T-test, two tailed.

4.5 Chromosome accessibility during erythroid differentiation in ATMDS

To see if the changes in globin gene expression were due to changes in chromosome accessibility, we performed ATAC-seq on both double positive and double negative populations.

At the alpha globin locus (Figure 4-9), the pattern of chromosome accessibility in the double negative populations in ATMDS resembled that of HSC, showing no accessibility in the four enhancer elements. In the double positive populations, only R2 was clearly present in all three samples. The other three elements, particularly R3 and R4, and the *HBA* gene body showed low or lack of peaks.

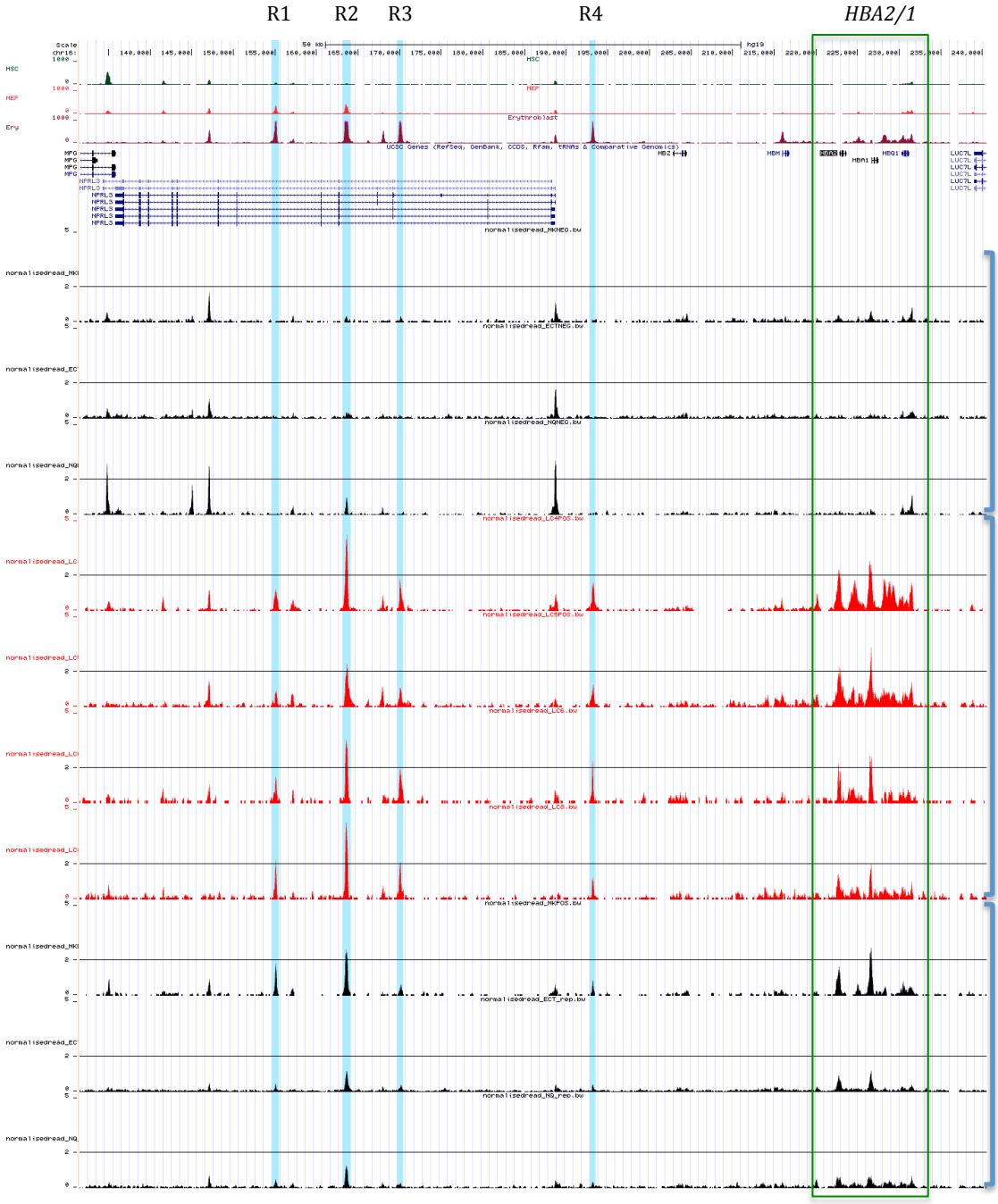


Figure 4-9. Normalised ATAC-seq tracks at the alpha globin locus. From top to bottom: tracks from FACS separated primary HSC (green), megakaryocyte-erythroid progenitor (MEP; red) and erythroblasts (dark red), produced by Corces *et al*[257]; normalised tracks from double negative populations from three ATMDS patients (MK, ECT and NQ) (black); normalised tracks from double positive populations from four independent normal LC controls (LC4, LC5, LC6 and LC8; red); normalised tracks from double positive populations of the same ATMDS patients (MK, ECT and NQ) (black). The four enhancer elements (R1-R4) are highlighted in cyan. Gene bodies of *HBA2*, *HBA1* and *HBQ1* (theta globin) are boxed (green).

In the beta/gamma globin locus (Figure 4-10), peaks were visible in the locus control region (LCR) as well as the gene body of *HBB* in the double positive populations in both ATMDS and controls. The gamma globin (*HBG1* and *HBG2*) region was not significantly different between ATMDS and controls.

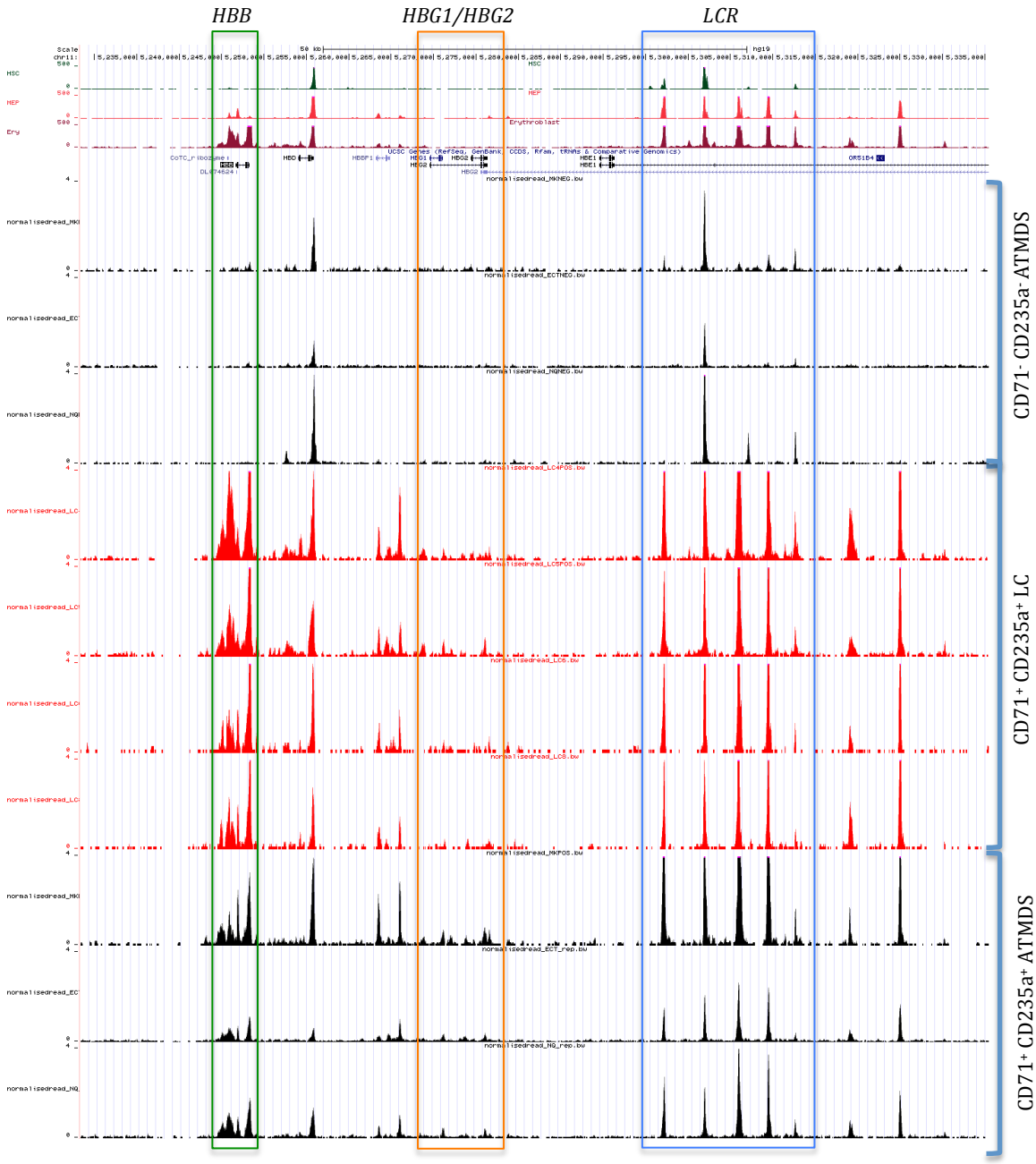


Figure 4-10. Normalised ATAC-seq peaks at the beta globin locus. LCR is boxed in blue on the right, the two gamma globin genes (*HBG1* and *HBG2*) are boxed in orange, and beta globin is boxed in green.

4.6 Differential ATAC-seq peak analysis

We performed genome-wide peak calling and compared peak calls between CD71⁺ CD235a⁺ cells in ATMDS patients and normal controls. DiffBind analysis showed that independent samples within each group clustered well together, particularly the four normal controls (Figure 4-11). The two groups were segregated on principle component analysis (PCA). There were 1,289 differential sites (i.e. genomic regions with significantly different ATAC-seq peak call) identified between the two groups (Figure 4-12). The majority of these differential sites (1,111 out of 1,289) indicated a reduction/absence of peaks in ATMDS patients.

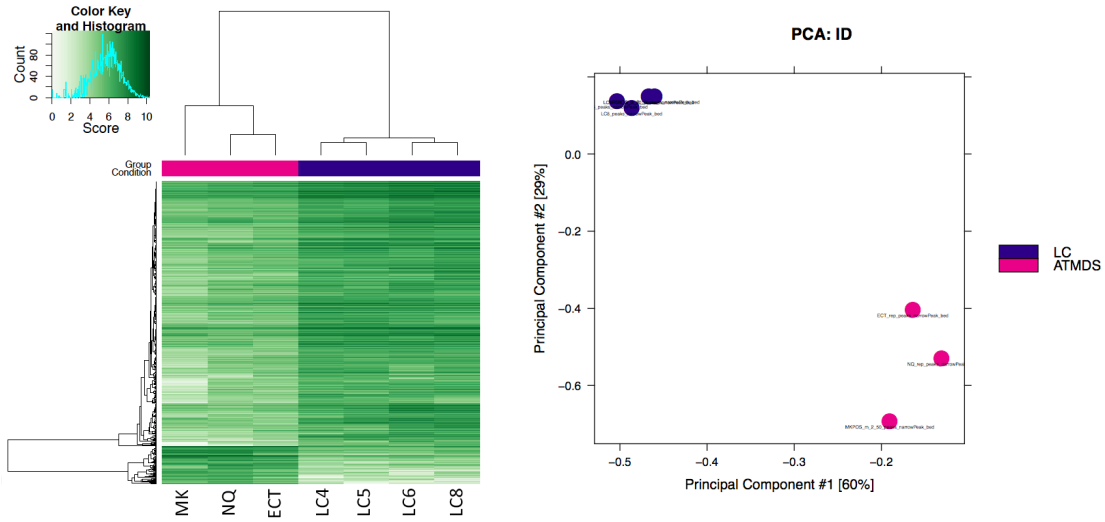


Figure 4-11. Correlation heatmap (left) and PCA plot (right) based on differential sites. The three ATMDS patients are shown in red, normal leucocyte cone controls (LC, n=4) are shown in blue.

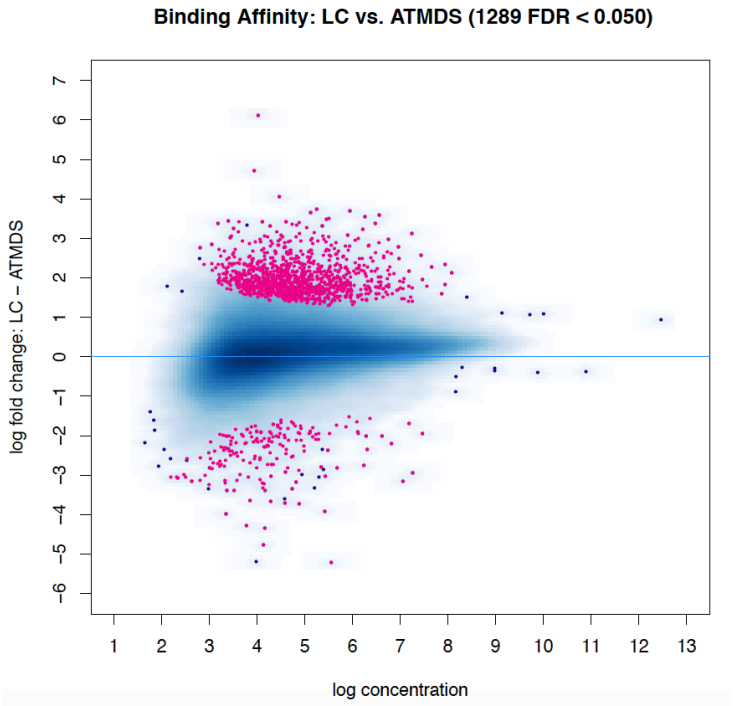


Figure 4-12. MA plot comparing ATAC-seq peak calls between normal leucocyte cone controls (LC) and ATMDS patients. Each differentially called site is shown as a red dot. The majority of sites showed higher peaks in LC group (i.e. above 1 in Y-axis).

We first focused on the differential sites in the alpha and beta globin loci. Although some of the peaks were lower in ATMDS, no differential sites were called in the beta globin locus. For 16p13.3 where the alpha globin locus is located, only four differential sites were called (see Table 4-1 for a list of differential sites in chromosome 16, and Supplementary Figure 4 for ATAC-seq tracks showing peak calls). Both R3 and R4 were identified as having lower peaks in the ATMDS group. One site within the gene body of *NME4* was also lower in ATMDS patients. ATRX is enriched in an imperfect VNTR within intron 1 of *NME4*[71]. However here the differential site was centered around intron 3, which contains GATA and TAL1 (also known as SCL) transcription factor binding sites within an internal promoter coding for an erythroid specific *NME4* transcript[258].

Start	End	Conc	Conc_LC	Conc_ATMDS	Fold	p.value	FDR	Genomic Location	Cytoband
169834	170335	5.15	5.66	3.95	1.72	2.01E-05	0.0029	R3 in alpha globin locus	16p13.3
192971	193472	4.83	5.28	3.83	1.45	0.000311	0.0214	R4 in alpha globin locus	16p13.3
438077	438578	4.59	5.13	3.24	1.89	9.42E-06	0.00163	Uncharacterised LOC100134368 (ncRNA)	16p13.3
449315	449816	5.1	5.62	3.85	1.77	3.12E-05	0.00396	*NME4*	16p13.3
24147498	24147999	4.21	4.79	2.64	2.15	6.10E-06	0.0012	*PRKCB*	16p12.2
29019293	29019794	4.01	4.62	2.29	2.33	2.02E-05	0.0029	*NPIPL1*	16p11.2
56072587	56073088	6.06	6.73	3.78	2.95	2.24E-09	2.71E-06	No known gene	16q12.2
58848994	58849495	4.65	5.29	2.67	2.62	2.36E-08	1.59E-05	*AK057513*	16q21
71498102	71498603	3.87	4.4	2.59	1.81	0.000749	0.0395	*ZNF23*	16q22.2
72078137	72078638	4.61	5.09	3.52	1.57	0.000866	0.0431	*TXNL4B*	16q22.2
83171013	83171514	4.43	3.43	5.17	-1.73	0.00107	0.0494	*CDH13*	16q23.3
84641262	84641763	3.72	4.25	2.42	1.83	0.00106	0.0492	*COTL1*	16q24.1
87339346	87339847	5.04	5.52	3.93	1.6	0.000189	0.0152	*C16ORF95*	16q24.2
87441218	87441719	3.42	1.93	4.32	-2.39	4.74E-05	0.00552	*ZCCHC14*	16q24.2
87840737	87841238	4.18	4.65	3.13	1.51	0.00105	0.0489	No known gene	16q24.2

Table 4-1. Differential sites between ATMDS and normal controls in chromosome 16. In the alpha globin locus, both R3 and R4 peaks were lower in ATMDS patients versus normal controls. Start: start position of site; End: end position of site; Conc: mean read concentration over all 7 samples (LC and ATMDS); Conc_LC: mean read concentration over leucocyte cone controls; Conc_ATMDS: mean read concentration over ATMDS patients; Fold: difference in mean read concentrations between the two groups. Positive values represent a lower mean read concentration (i.e. ATAC-seq peak) in ATMDS patients and vice versa; p.value: p-value for differentially bound sites; FDR: multiple testing corrected FDR p-value. Genomic Location indicates the locus or gene where the differential site lies within.

Expanding the differential site analysis to a genome-wide level, we used the motif enrichment analysis tool MEME-ChIP to look for enriched transcription factor binding motifs in the 1,111 differential sites where ATAC-seq peaks are reduced/absent in ATMDS patients.

The three most significantly enriched motifs discovered from the differential sites included motifs for 1. GATA; 2. KLF; and 3. TAL1-TCF3 heterodimer (TCF3 is also known as E2A) (Table 4-2). In other words, in ATMDS patients chromosome accessibility at various sites containing these motifs is reduced compared to normal controls. GATA, KLF and TAL1 are all well characterized transcription factors essential in normal erythropoiesis (reviewed in Kerenyi and Orkin[259]). These results suggest that in ATMDS the effect of differentiation block extends to later stages of erythroid differentiation in cells already expressing erythroid cell surface markers.

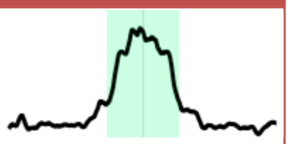
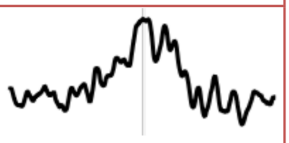
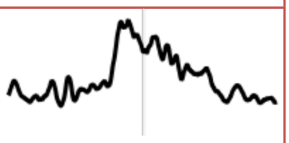
Motif found	E-value	Sites	Known or similar motifs	Distribution
	$2.3x10^{-189}$	672	GATA2 GATA5 Gata1	
	$2.2x10^{-83}$	211	Klf1 KLF9 KLF5	
	$9.4x10^{-39}$	581	TAL1-TCF3 ZBTB18	

Table 4-2. Top three motifs enriched in differential sites from ATAC-seq DiffBind analysis that are reduced/absent in ATMDS patients. E-value represents the statistical significance of enrichment. Sites: number of differential sites contributing to the construction of the enriched motif. Distribution is a graphical representation of the best matches to the motif within differential sites. This is highly dependent on the location of the extracted sequence from ATAC-seq. In the case of GATA motif, most of the differential sites contain the motif close to the centre (area shaded green). The distributions for KLF and TAL1-TCF3 motifs are less apparent, as reflected in the higher (less statistically significant) E-values.

4.7 Summary

In this chapter, we characterize erythroid differentiation and globin gene expression of haematopoietic progenitor cells in ATMDS patients *ex vivo*. FACS analysis shows failure to commit to erythroid differentiation in the majority of cells, and this failure in differentiation is more prominent compared to MDS patients without mutations in *ATRX*, *ASXL1*, *U2AF1* and *PHF6*. Of the cells that do express erythroid cell surface markers CD71 and CD235a in ATMDS patients, the expression of both alpha and beta globin genes were downregulated. The alpha/beta globin gene expression ratio, which was reduced in peripheral blood and bulk cell culture of ATMDS patients, was not significantly lower in $CD71^{+}$ $CD235a^{+}$ cells. However, when we took into account gamma globin expression there was a small but significant alpha to non-alpha globin gene expression imbalance.

Chromosome accessibility analysis on the two globin loci showed generalised reduction in accessibility, particularly the alpha globin locus. Differential peak analysis indicated that the two alpha globin enhancer elements R3 and R4 had lower accessibility in ATMDS, pointing to a defect in later stages of differentiation[37]. In agreement, at the genome-wide level, sites with reduced accessibility in ATMDS patients were enriched with motifs of erythroid-specific transcription factor GATA, KLF1 and TAL1 binding motifs.

GATA1, the finding member of the GATA transcription factor family, and TAL1 are part of a complex that regulates erythroid-specific genes[260].

Defects in GATA1 and TAL1 have been implicated to affect late stages of erythroid differentiation. In mouse, erythroid differentiation is arrested at early proerythroblast-like stage in Gata1 knockouts[261, 262]. Relating to MDS, GATA1 targets were also repressed in RNA-seq of knock-in mice expressing truncated ASXL1[182].

In mice, conditional knockout and knock-in of a mutant defective in DNA binding in Tal1 showed defect in proliferation of early erythroid progenitor with anaemia and perturbed processes during terminal erythroid maturation[263, 264]. *TAL1* transcript binds to enhancers of both alpha and beta globin, but globin gene expression levels were not altered by expression of Tal1 mutant defective in DNA binding[265]. However, alternative transcripts exist in *TAL1*[266]. The shorter transcript is more efficient in promoting erythroid differentiation and beta globin gene expression. Hotspot mutation in the spliceosomal gene *SF3B1* (p.Lys700Glu), frequently found in MDS with ring sideroblasts, has been shown to lower the ratio of the short/full-length *TAL1* transcript[267]. This highlights a relationship between aberrant splicing and regulation of transcription factor. Whether U2AF1 p.Gln157Pro mutation could also alter the short/full-length *TAL1* transcript ratio remains to be determined.

KLF1 is involved in both early and late stages of erythropoiesis. At early stage of erythropoiesis, KLF1 promotes erythropoiesis while suppressing megakaryopoiesis at the megakaryocyte-erythroid progenitor (MEP) stage[268]. KLF1 remains highly expressed during erythroid

differentiation. In later stages of differentiation Klf1-null erythroid cells ectopically expressed megakaryocyte markers[269, 270].

Another relevant role of KLF1 is its involvement in gamma to beta globin gene switching. KLF1 mutations are detected in individuals with hereditary persistence of fetal haemoglobin (HPFH)[16], and gamma to beta globin gene switching was delayed in longitudinal study in these individuals[271]. KLF1 binds to the locus control region of the beta globin locus and also to the beta globin gene promoter to drive beta globin gene transcription. This is shown in differentiated umbilical cord blood-derived $CD34^+$ cells where expression of beta globin is strongly correlated with KLF1 expression, and knockdown of KLF1 increased gamma globin mRNA[272]. KLF1 also promotes globin gene switching via regulation of the gamma globin repressors BCL11A and ZBTB7A[16, 273, 274]. Chromosome accessibility in *BCL11A* gene body was generally lower in ATMDS patients, and a differential peak containing KLF1 motif was identified upstream of *BCL11A* (see Supplementary Figure 5). Whether the low accessibility in *BCL11A* could explain the low beta globin gene expression in ATMDS remains to be investigated. No differential peaks were identified within 0.5Mb of *ZBTB7A*.

Overall, the results from this chapter suggest that defective erythroid differentiation is evident in ATMDS, but the imbalance in globin gene expression in differentiated $CD71^+$ $CD235^+$ cells *ex vivo* is not as pronounced as in peripheral blood. Western blot of individual globins may be performed to provide further evidence on the degree of globin

imbalance, as detection at the protein level accounts for post-translational effect. Differential peak analysis on chromosome accessibility assay showed that reduced erythroid transcription factor occupancy might be the underlying cause for such block in differentiation, and the block may exert differential effect between alpha and beta globin loci.

ATMDS is rare and samples are very limited. To study how mutations in ATMDS, individually and cooperatively, affect erythroid differentiation and the alpha globin locus, a cellular model is required. The next chapter will describe introduction of ATMDS mutations in an erythroid progenitor cell line HUDEP-2.

5 Generation of HUDEP-2 cell line with ATMDS mutations

The Human Umbilical cord Derived Erythroid Progenitor (HUDEP) cell line was established in 2013 to enable mass production of human red blood cells for transfusion purpose[150]. This cell line was created by lentiviral transduction of a tetracycline-inducible human papillomavirus E6/E7 (HPV E6/E7) expression system in a mixed population of $CD34^+$ cells obtained from umbilical cord blood. HPV E6 protein promotes degradation of p53 tumour suppressor, whereas E7 binds to another tumour suppressor retinoblastoma protein (pRB)[275]. These cells express precursor cell and erythroid markers, including CD71, CD235a, CD36 and c-kit, and are negative for markers of leucocytes and megakaryocytes (CD11b and CD41a, respectively). These cells can undergo differentiation to late stage erythroblasts and produce haemoglobin, therefore they serve as a potential model to study the effect of mutations in *ATRX*, *ASXL1*, *U2AF1* and *PHF6* on erythroid differentiation and globin gene expression.

Originally three HUDEP lines were created (HUDEP-1, -2 and -3). While all three lines showed expression of alpha globin upon differentiation, only HUDEP-2 expressed beta globin and hence was chosen for genome editing in this project. The other two lines expressed the fetal gamma globin only. Concordantly *BCL11A*, a known regulator in fetal haemoglobin silencing[17, 276], was only expressed in HUDEP-2.

The requirements for introducing mutations are different among the four genes of interest – the two X chromosome genes require knockout, *ASXL1* requires a heterozygous truncation and *U2AF1* requires a heterozygous missense mutation. To avoid introducing a large deletion in chromosome X, *ATRX* and *PHF6* had to be knocked out sequentially. Therefore a step-wise approach was taken to introduce these mutations.

Figure 5-1 shows the strategy in generation of ATMDS HUDEP cell line. In the first two steps, ATRX and PHF6 were sequentially knocked out and confirmed by Western Blot (Figure 5-2). In the previous chapter, *U2AF1* and *ASXL1* were shown to be earlier mutations in MDS. Here the two X chromosome genes were edited first as initially it was technically difficult to introduce heterozygous mutation into autosomal region. This was resolved later in the project and explained in the following section. It is also worth noting that no clonal selection was performed on the wildtype HUDEP-2 cells before genome editing, in another word the wildtype HUDEP-2 used here for comparison was a mixed population of cells that could contain different genomic variations.

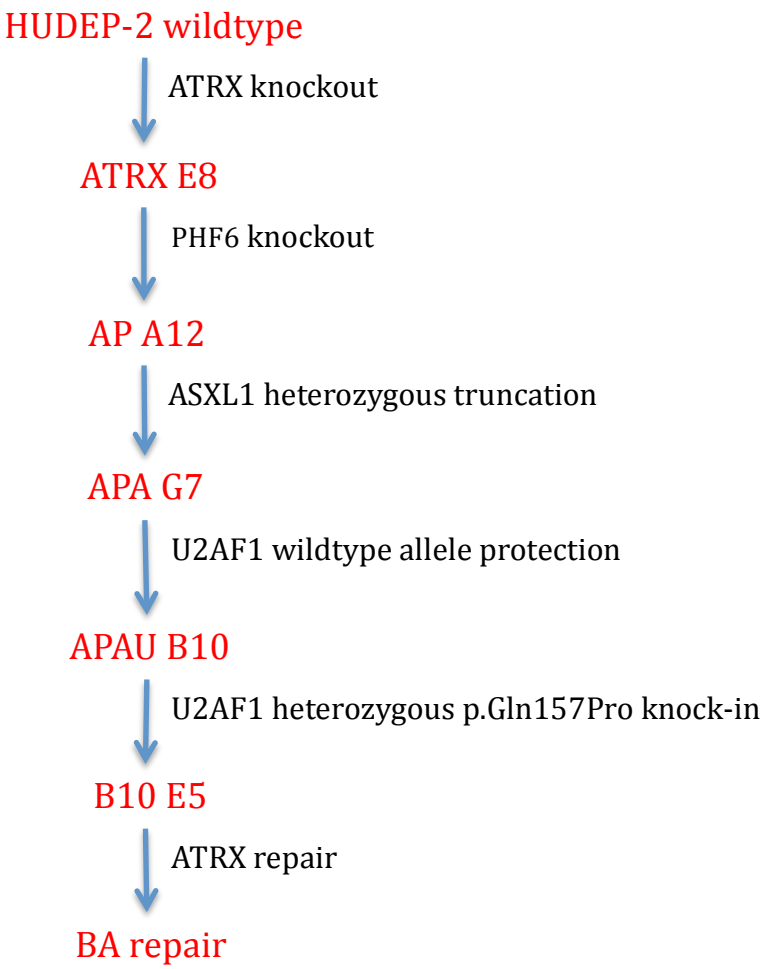


Figure 5-1. Flowchart on introducing mutations in ATMDS in HUDEP-2. Mutations in ATRX, PHF6, ASXL1 and U2AF1 were introduced sequentially. The clone APAU B10 was an intermediate clone made during the two-step process of introducing the heterozygous U2AF1 p.Gln157Pro mutation.

5.1 Generation of ATRX knockout HUDEP-2 cell

Mutations in each ATRX knockout clone are listed in Table 5-1. It is noted that in-frame deletions removing one or two amino acids (clones D2 and C7) did not affect the protein level, while in-frame deletion of three amino acids in similar position greatly reduced the level of protein (clone E3) (see Figure 5-2). Clone E8 was used for further editing, as with only a single base insertion it contains the least sequence manipulation.

Clone	Mutation	Predicted consequence
B5	Complex indel	Not predicted
D2	c.4606_4609delinsG	p.Lys1536_Leu1537delinsVal
C7	c.4601_4606delCAACCA	p.Thr1534_Thr1535del
E8	c.4607_4608insA	p.Leu1537Valfs*5
E3	c.4603_4612delACCAAGTTGGinsC	p.Thr1535_Val1538delinsLeu
D6	c.4600_4606delACAACCAinsT	p.Thr1534*
G4	c.4604del31	p.Leu1537Asnfs*3

Table 5-1. List of HUDEP-2 clones generated from a single ATRX genome editing experiment. The complex insertion/deletion in clone B5 involved a 60bp deletion, and a 53bp insertion of an inverted fragment from the non-coding strand. ATRX reference coding sequence: NM_000489.3; reference protein sequence NP_000480.2.

5.2 Double knockout of ATRX and PHF6

PHF6 knockout on ATRX E8 clone returned two clones with complete knock out at the protein level (clone AP A12 and AP G7, Table 5-2). Again, clone AP A12 with the least sequence manipulation was used for further editing.

Clone	Mutation	Predicted consequence
AP A7	c.48_65del18	p.Cys17_Ser22del
AP A12	c.48_49insA	p.Cys17Metfs*5
AP G5	c.44_52del9	p.Lys16_Gly18del
AP G7	c.48_49insAAAA	p.Cys17Lysfs*6

Table 5-2. List of ATRX-PHF6 double knockout clones. The two clones with in-frame deletion (AP A7 and AP G5) had low level of protein remaining on Western Blot (see Figure 5-2). PHF6 reference mRNA sequence: NM_001015877.1; reference protein sequence: NP_001015877.1.

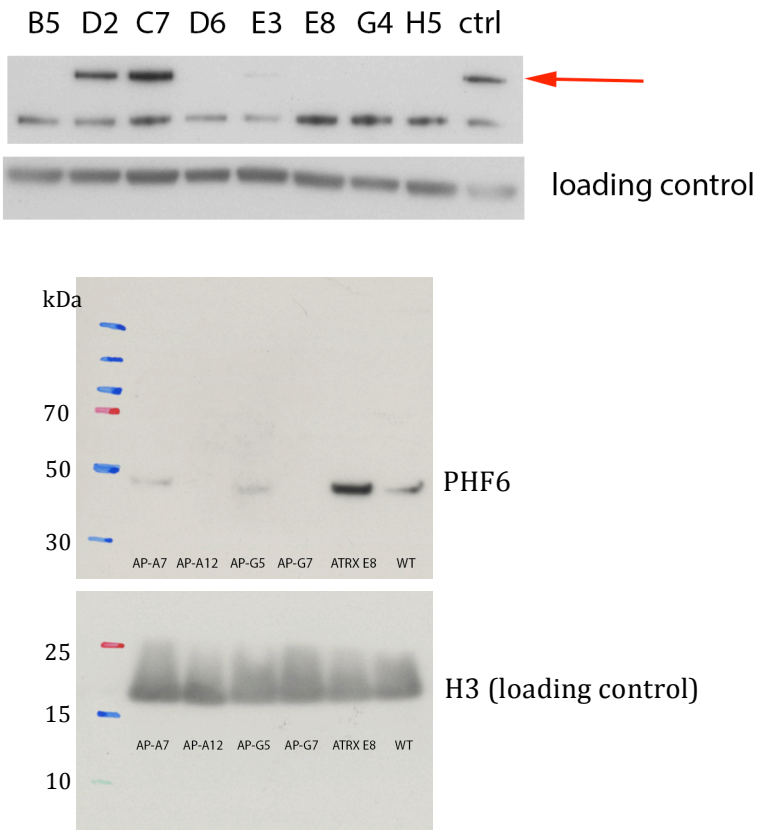


Figure 5-2. Western Blots for ATRX (top panel) and PHF6 (bottom panel). In ATRX, the primary antibody (39F) recognised both the long (top band, red arrow) and short (bottom band) ATRX isoforms. The sgRNA that targets ATRX exon 16 in this project only knocks out the long isoform, as seen in clones B5, D6, E8, G4 and H5. Wildtype HUDEP-2 was used as control. For PHF6 Western blot, AP A12 and AP G7 were complete knockouts. The parental clone ATRX E8 and wildtype HUDEP-2 were used as controls. Anti-H3 antibody was used as a loading control.

5.3 Protection of wildtype allele in generation of heterozygous ASXL1 truncation

An issue appeared during the attempt to introduce heterozygous truncation in ASXL1 – the CRISPR/Cas9 always cut both alleles leading to homozygous or compound heterozygous mutations. To protect one allele from being edited, we designed a single strand oligonucleotide (ssODN) containing three synonymous mutations, one of which within the PAM site (ASXL1-ssODN-WT, see Figure 5-3). We transfected cells from AP A12 clone with plasmid targeting the last exon of *ASXL1* along with the ssODN. To enhance the chance of knock-in by homology-directed repair (HDR), we

used the small molecule RS-1 (see Materials and Methods). Without RS-1, 7 clones were screened and no HDR event was observed. Using RS-1 four clones were screened and one clone (APA G7) contained a heterozygous 31bp deletion, leading to a truncation of the protein (p.Tyr700*) (Table 5-3 and Figure 5-3).

ASXL1-ssODN-WT:
ACAGCGAACACAACTACTGCCGCCTTATCCTCTAAATGGaGAGCATACaCAaGCCGGAACTGCCATGTC
CAGAG CTAGGAGAGAGGACCTGCCTTCT

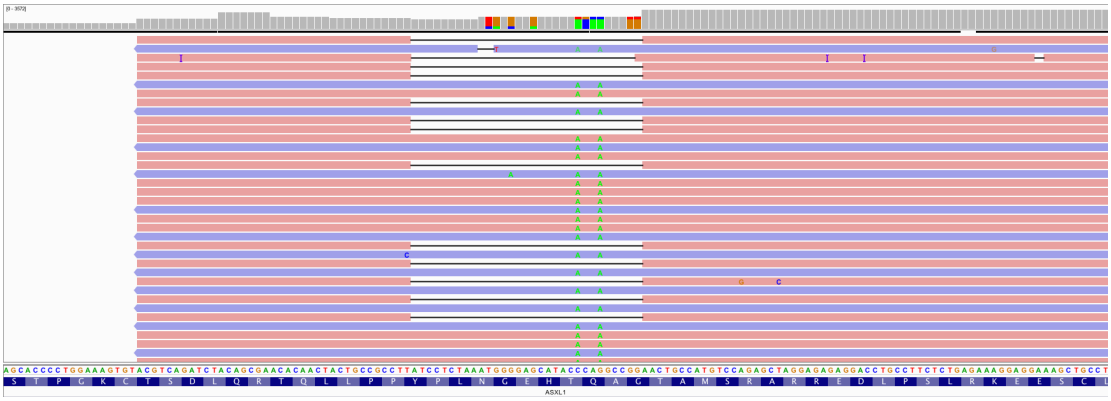


Figure 5-3. Top: sequence of ASXL1-WT-ssODN to create a protected wildtype allele in ASXL1. sgRNA sequence is highlighted in green. PAM sequence is highlighted in yellow. Deliberate synonymous mutations in small letters (red). Sequence reads of clone APA G7 at exon 14 of ASXL1. A heterozygous 31bp deletion (NM_015338.5:c.2100_2130del) can be seen. The protected wildtype allele only contained two of the three deliberate synonymous mutations - the deliberate synonymous mutation furthest from the PAM site was not incorporated. This is not unexpected given SpCas9 could tolerate mismatches distal to PAM site[152], and therefore such mismatch was not essential to prevent re-editing.

	Clone	*ASXL1* mutation
Without RS-1	A1	homozygous c.2118del
	A3	compound heterozygous deletion c.[2116_2121del];[2120_2121del]
	A10	homozygous c.2467_2551del
	C3	compound heterozygous deletion c.[2119_2130del]; [2115_2124del]
	C4	compound heterozygous (indel undefined)
	C11	compound heterozygous c.[2111_2127del];[2117_2118del]
	E6	homozygous c.2262_2559del
With RS-1	APA G5	homozygous c.2116_2121del in-frame deletion
	APA B9	large deletion (>300bp)
	APA G6	compound heterozygous deletion (indel undefined)
	APA G7	heterozygous c.2100_2130del (p.Tyr700*) with HDR on second allele

Table 5-3. Genotyping results for ASXL1 mutation generated with and without RS-1. Clones were screened by Sanger sequencing.

5.4 Two-step approach for generation of heterozygous p.Gln157Pro mutation in U2AF1

Introduction of the heterozygous U2AF1 p.Gln157Pro mutation was more challenging. The missense mutation was knocked in over a two-step process (Figure 5-4). In the first step, using the same strategy as for *ASXL1*, an intermediate clone (APAU B10) was created using ssODN (U2AF1-ssODN-WT, sequence shown in Supplementary Figure 6) to "protect" one wildtype *U2AF1* allele, whereas the other allele contained a 1bp insertion. The second step involved designing a sgRNA specifically targeting the allele with 1bp insertion, and introduction of a different ssODN (U2_Q157P_ssODN5) that 1. Repairs the 1bp insertion; 2. Introduce the p.Gln157Pro mutation and 3. Introduce synonymous mutation at the new PAM site to prevent re-editing. This step proved to be difficult and five different iterations of ssODN design were attempted before the mutation was successfully knocked in.

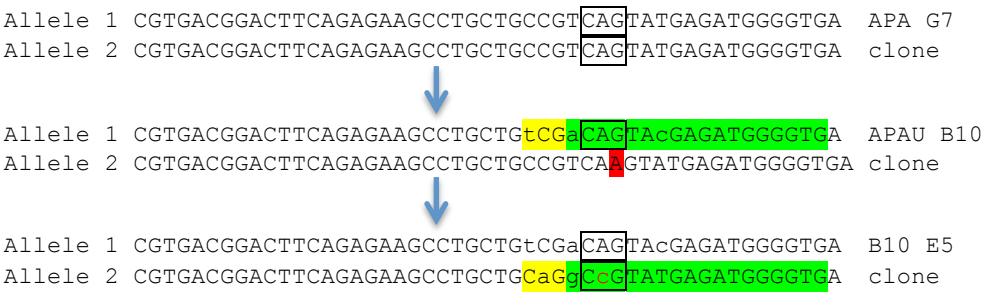


Figure 5-4. Schematic for introducing U2AF1 p.Gln157Pro mutation in the triple mutant APA G7 clone. Codon 157 is boxed. After the first round of editing, three synonymous mutations were introduced in Allele 1 to prevent re-editing (small letters), whereas in Allele 2 a single base insertion (highlighted in red) was introduced. In the second round, Allele 2 was specifically targeted to repair the insertion and to introduce the missense mutation leading to the p.Gln157Pro (CAG → CcG). Two deliberate synonymous mutations were also introduced (small letters). See Appendix (Supplementary Figures 6 and 7) for Sanger sequencing traces for APAU B10 and B10 E5 clones.

During this process it was realised that there are three *U2AF1* alleles in HUDEP-2 cell line (Figure 5-5). This was also confirmed in SNP array where copy number gain was detected in chr21q11.2-q22.3[277]. The edited quadruple mutant HUDEP-2 line with all four genes mutated, clone B10 E5, had one completely wildtype allele, one protected wildtype allele and one allele with the p.Gln157Pro mutation.

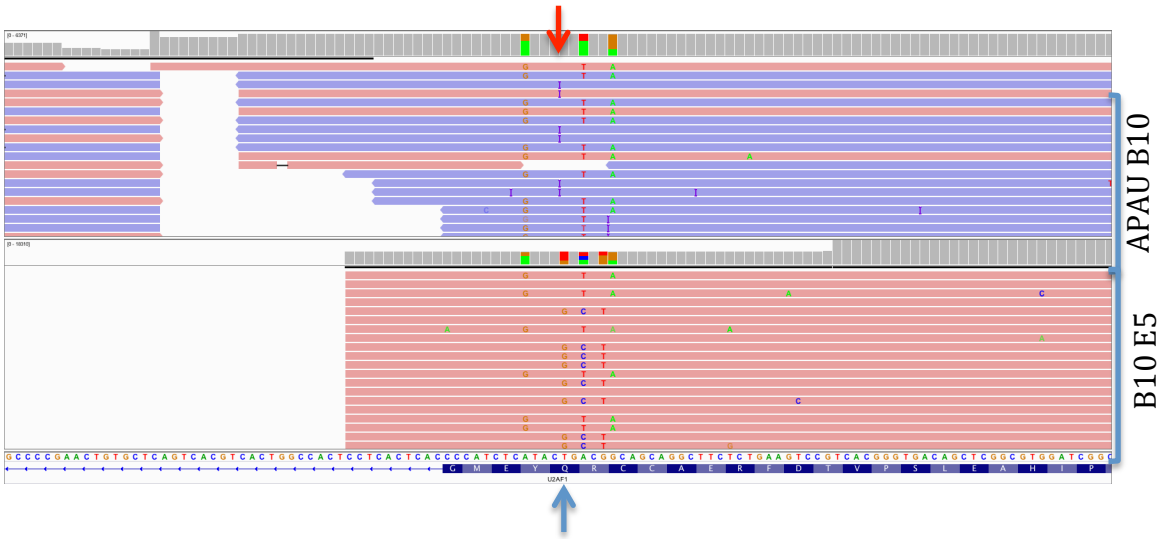


Figure 5-5. IGV view of sequence reads demonstrating the two-step editing process and presence of three alleles in U2AF1. Top panel – sequence reads from the intermediate clone APAU B10 showing the protected wildtype allele (with three deliberate synonymous mutations) and the 1bp insertion (red arrow, VAF 67.3% or 3692/5488 reads). Bottom panel – sequence reads from clone B10 E5. The allele with 1bp insertion was re-edited to knock in the p.Gln157Pro mutation (T>G change, blue arrow). Note the presence of completely wildtype allele. The presence of three alleles was also visible in Sanger sequencing trace (see Supplementary Figure 7). One possible explanation for the presence of this wildtype allele was that one of the two alleles containing the 1bp insertion was repaired back to complete wildtype in the second step.

5.5 Repair of ATRX knockout in the quadruple mutant B10 E5

Reversal of ATRX knockout in the B10 E5 clone would create a triple mutant line to mimick that of ATMDS patients with low percentages of Hb H. To maximize the chance of a successful repair, three independent ssODNs (see Appendix for ssODN sequences) with varying number of deliberate synonymous mutation were used simultaneously in the same

nucleofection. A total of 24 clones were retrieved from this experiment, out of which 10 clones were successfully repaired (2 clones repaired by ssODN1, and 4 clones each repaired by ssODN2 and ssODN3, see Supplementary Figure 8 for Sanger sequencing traces). This equates to an overall HDR success rate of 41.67%. The repair was also confirmed at the protein level by Western Blot on two clones (BA12 and BA2, repaired by ssODN2 and ssODN3, respectively; Figure 5-6). The previously introduced mutations in the other three genes (*ASXL1*, *U2AF1* and *PHF6*) were confirmed again by Sanger sequencing in these clones.

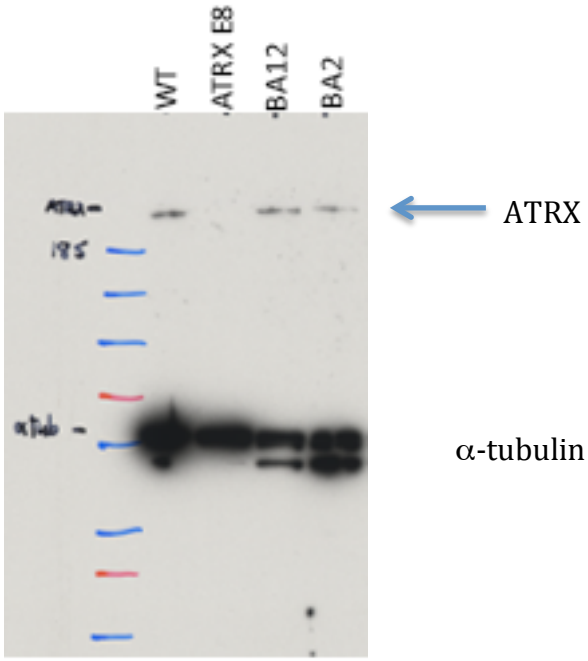


Figure 5-6. Western Blot to confirm ATRX repair in B10 E5 clone. A different anti-ATRX antibody (Abcam ab97508, targeting only the C-terminus of ATRX) was used compared to the previous ATRX Western Blot. The re-appearance of ATRX band can be seen in both BA12 and BA2 clones (blue arrow). Anti-alpha tubulin was used as input control (Abcam ab7291, bottom bands). Western Blot performed by Thomas Kent. WT: wildtype positive control. ATRX E8: ATRX knockout negative control.

5.6 Erythroid differentiation of HUDEP cell

HUDEP cells were expanded and depleted for CD233 marker to remove differentiating cells before the start of differentiation. All cell lines had <5%

CD233^{+} cells at the start of differentiation. Globin gene expression before and after CD233 depletion showed that the depletion process did not affect alpha/non-alpha globin gene expression ratio (Table 5-4). In fact, the alpha/non-alpha globin expression ratio was highest in the wildtype (WT, 1.22 in depleted cells) and dropped to less than half (0.47) upon introduction of PHF6 mutation. This low expression ratio was a result of lower alpha and beta globin gene expression, with an inversely increased gamma globin expression. The low expression ratio remained in subsequent clones (APA G7, B10 E5 and BA12), despite an increase in expression of all three globin genes in the B10 E5 and BA12 clones where U2AF1 p.Gln157Pro mutation was introduced. This variation in basal expression of globin genes could be a result of the clonal nature of genome editing strategy.

	Gene(s) mutated	*HBA*/*RPL13A*		*HBB*/*RPL13A*		*HBG*/*RPL13A*		Alpha/non-alpha	
		Pre-depletion	Post-depletion	Pre-depletion	Post-depletion	Pre-depletion	Post-depletion	Pre-depletion	Post-depletion
WT	N/A	20.6	11.2	15.6	9.2	0.0	0.0	1.32	1.22
ATRX E8	*ATRX*	22.6	11.0	21.5	10.8	0.3	0.1	1.04	1.00
AP A12	*ATRX, PHF6*	10.2	4.7	7.3	3.5	13.8	6.6	0.49	0.47
APA G7	*ATRX, PHF6, ASXL1*	13.6	7.1	7.4	3.9	15.2	8.9	0.60	0.55
B10 E5	*ATRX, PHF6, ASXL1, U2AF1*	47.9	30.0	27.0	17.1	62.6	38.3	0.53	0.54
BA12	*PHF6, ASXL1, U2AF1*	58.3	36.6	51.3	31.9	56.4	35.8	0.54	0.54

Table 5-4. Globin gene expression before and after CD233 depletion in HUDEP cells prior to differentiation. Shown are expression levels of individual globin genes (*HBA*, *HBB* or *HBG* over *RPL13A*) and the alpha/non-alpha ratio in 6 different HUDEP cell lines from wildtype to quadruple mutant (B10 E5) and quadruple mutant with ATRX repaired (BA12).

HUDEP cells were differentiated for 9 days before FACS sorting on day 10 (see Supplementary Figures 9-11 for growth curve, viability measurement, serial FACS analysis and morphology during erythroid differentiation). Figure 5-7 shows the FACS analysis on day 10 of differentiation. Two

populations were FACS sorted – populations in Gate M and N represent the less and more differentiated population, respectively (see Supplementary Figure 12 for morphology of various gated populations). A block in differentiation becomes apparent from the double knockout AP A12 clone onwards – the percentage of cells in Gate N was below 3% and hence there was insufficient material for analysis. The morphology of cells before and after FACS sorting can been seen in Figure 5-8 and Figure 5-9. Interestingly, the ATRX only knockout (ATRX E8) seems to differentiate further compared to the wildtype (Gate N population 22.32% in ATRX E8, versus 13.52% in wildtype), and rescue of ATRX from the quadruple mutant background did not restore the block in differentiation (clone BA12).

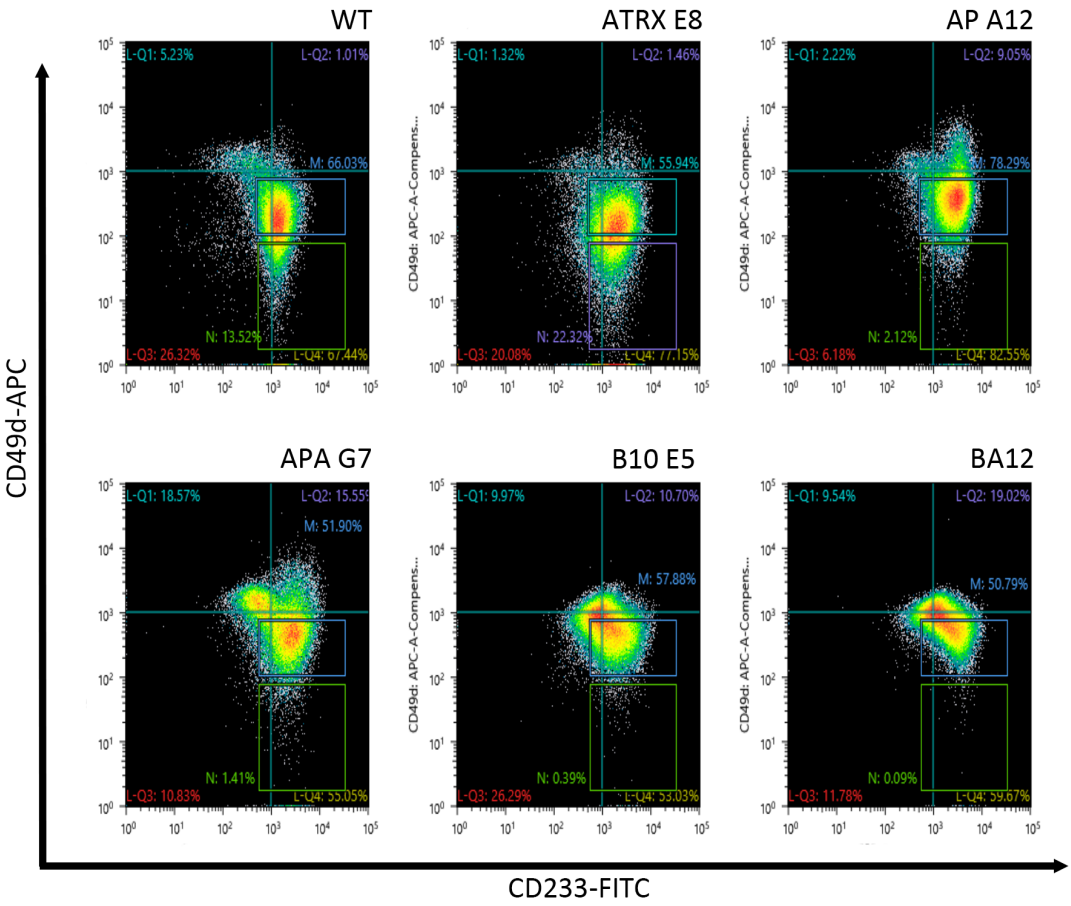


Figure 5-7. FACS analysis of HUDEP clones at day 10 of erythroid differentiation. As cells undergo differentiation CD233 signal increases and conversely the CD49d signal is reduced. Two gates (M and N) were used for FACS sorting. From top left to bottom right – WT: wildtype, ATRX E8: ATRX knockout, AP A12: ATRX PHF6 double knockout, APA G7: ATRX, PHF6 and ASXL1 mutant, B10 E5: ATRX, PHF6, ASXL1 and U2AF1 mutant, BA12: B10 E5 clone with ATRX repaired. Cells progressively lost the potential to differentiate upon introduction of ASXL1 and U2AF1 mutations (in APA G7 and B10 E5 clone, respectively). Repair of ATRX did not restore the block in differentiation in clone BA12.

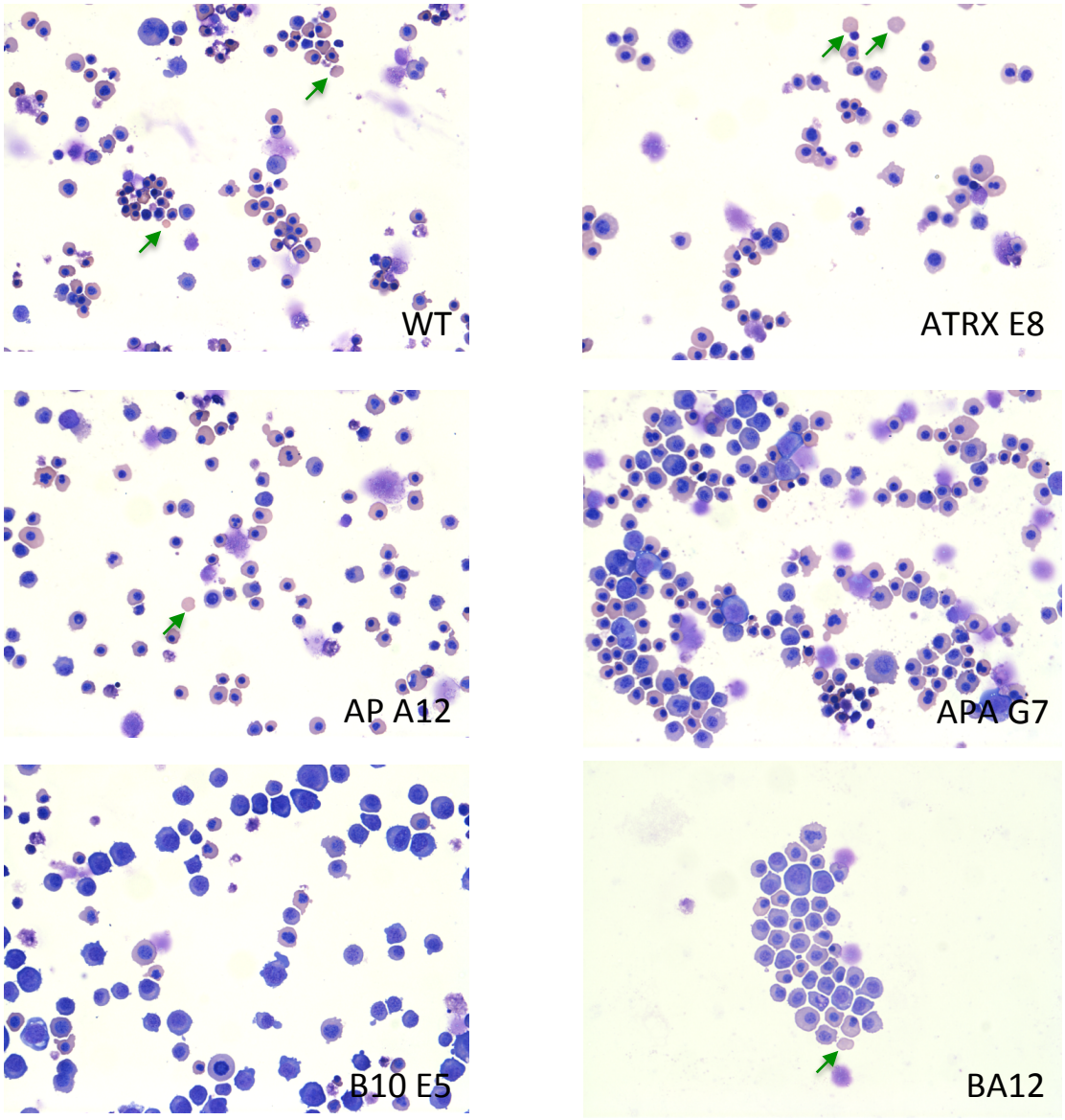


Figure 5-8. Cytospin pictures of differentiated HUDEP cells on Day 10 before FACS sorting. An increase in less differentiated cells (basophilic and polychromatic erythroblasts) can be seen in clones APA G7, B10 E5 and BA12. WT: wildtype, ATRX E8: ATRX knockout only, AP A12: ATRX, PHF6 double knockout, APA G7: ATRX, PHF6 and ASXL1 mutant, B10 E5: ATRX, PHF6, ASXL1 and U2AF1 mutant, BA12: B10 E5 clone with ATRX repaired. Green arrows: enucleated red blood cells.

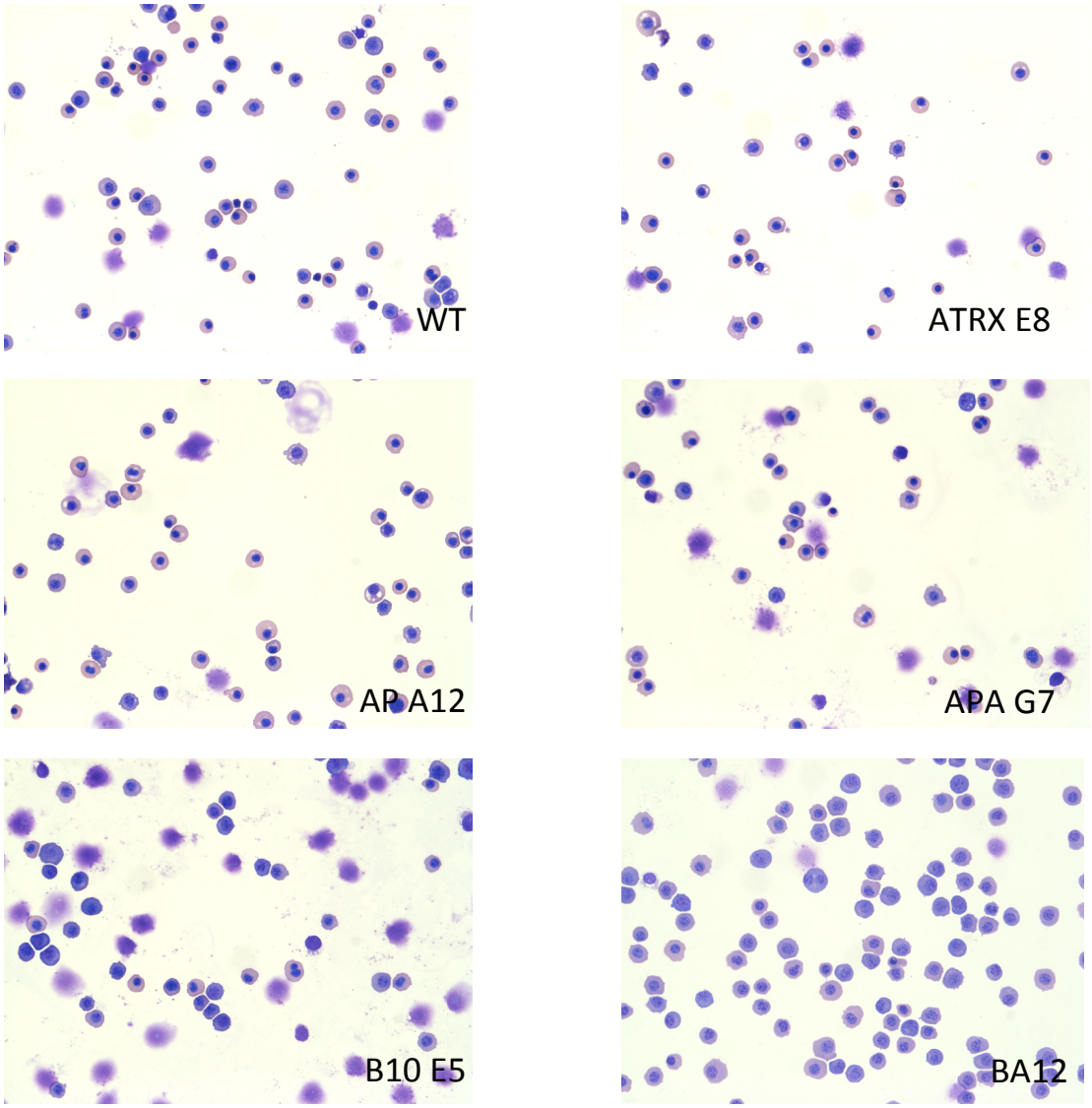


Figure 5-9. Cytospin pictures of FACS sorted cells (population "M") on Day 10 of differentiation. Basophilic erythroblasts were largely removed from the triple and quadruple mutants APA G7 and B10 E5 clones.

Comparison of globin gene expression between the less differentiated "M" and the more differentiated "N" populations in wildtype and three ATRX knockout clones (E8, D6 and G4) confirmed that globin gene expression increased as cells became more differentiated (Figure 5-10). In both populations the expression of *HBA* was lower in the ATRX knockout clones.

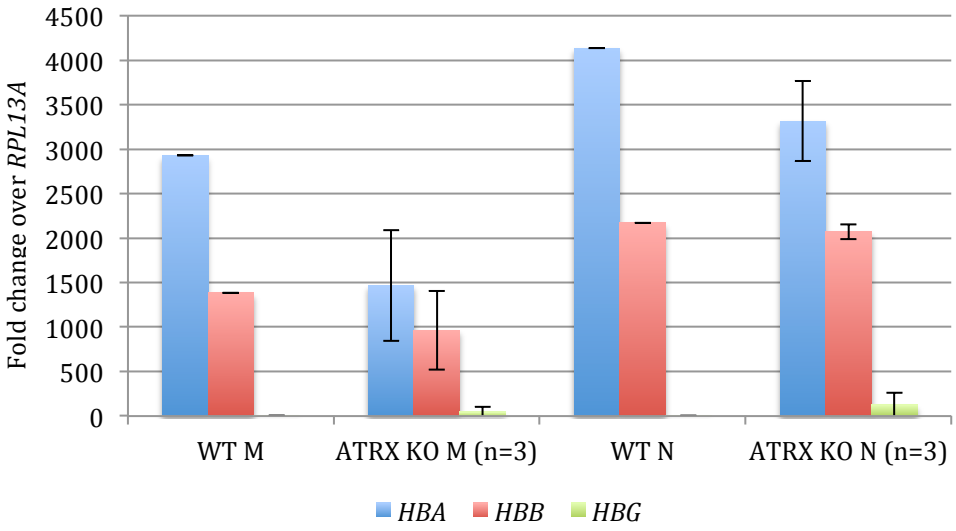


Figure 5-10. Individual globin gene expression in "M" (less differentiated) and "N" (more differentiated) populations in wildtype and ATRX knockout clones. Both alpha and beta globin showed increased expression in the more differentiated "N" population. Expression of gamma globin was very low in all populations (<20 fold over housekeeping gene *RPL13A*) except for ATRX G4 clone where it was 114 and 126 fold over *RPL13A* in the "M" and "N" populations, respectively.

We then compared globin gene expression of FACS sorted HUDEP cells in the "M" population across all clones (Figure 5-11). Between wildtype and the single (ATRX E8), double (AP A12) and triple (APA G7) mutant clones, there were no significant differences in alpha and beta globin expression. A significant reduction in alpha globin gene expression was observed between the wildtype and the quadruple mutant clone B10 E5. The same is

seen in the alpha/non-alpha expression ratio (Figure 5-12). Gamma globin expression was significantly increased in all genome-edited clones compared to wildtype. Expression of all globin genes increased when ATRX was repaired.

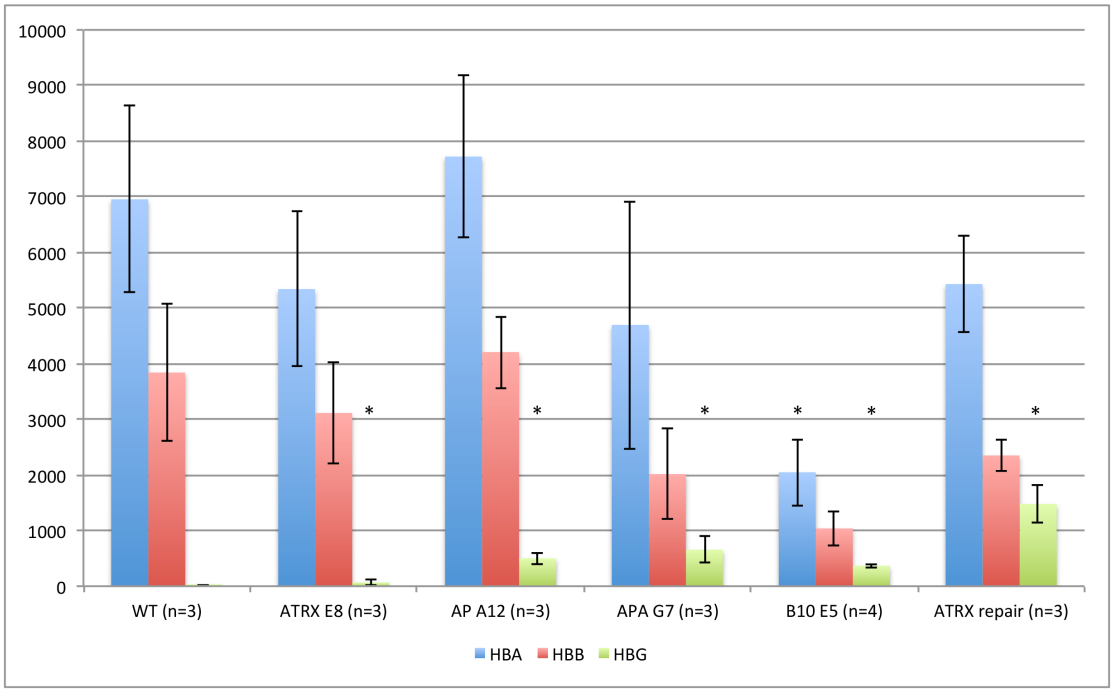


Figure 5-11. Globin gene expression of FACS sorted HUDEP clones at day 10 of erythroid differentiation (population "M"). Expression levels are normalised to RPL13A. Results are shown as the average of three (or four in the case of B10 E5) independent differentiation experiments ± standard deviation. *:p<0.05 compared to wildtype (Student's T-test, two tailed).

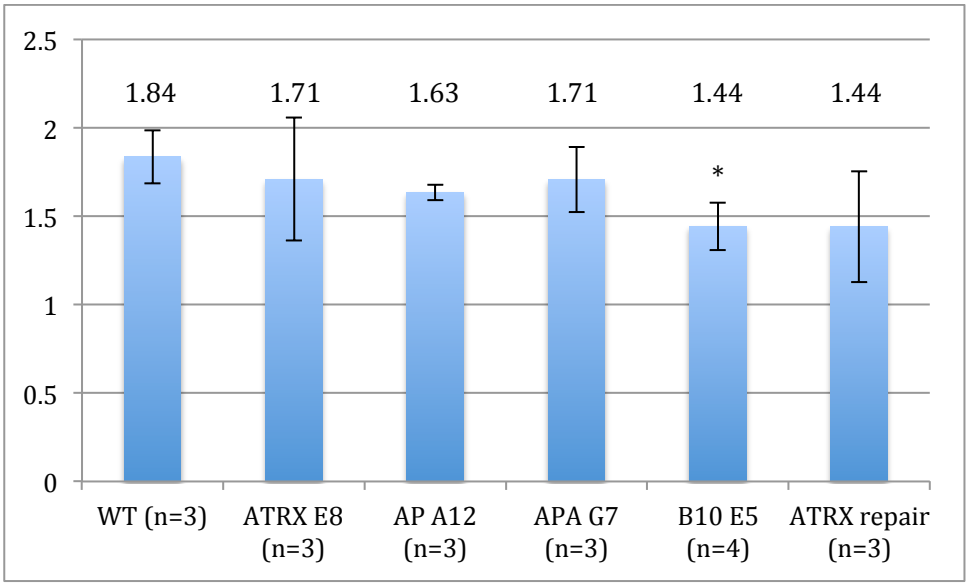


Figure 5-12. Alpha/non-alpha globin gene expression ratio calculated from expression levels in Figure 5-11. Actual ratio is shown above the bar. *:p=0.022 (Student's T-test, two tailed)

compared to wildtype. The differences between wildtype and clones other than B10 E5 are not statistically significant.

At the protein level, haemoglobin F ($\alpha_2\gamma_2$) was visible in all clones except wildtype and ATRX knockout (ATRX E8) (see Figure 5-17 in the following section for haemoglobin isoelectric focusing). This is consistent with the globin gene expression data. However, no Hb H (β_4) or Hb Barts (γ_4) were detected in these clones.

5.7 Chromosome accessibility in differentiated HUDEP

ATAC-seq data from FACS-sorted (population "M"), differentiated HUDEP cells saw R3 and R4 having the largest reduction in peak heights out of the four regulatory elements in the alpha globin locus (Figure 5-13). There was a generalised reduction in peak heights in the *HBA* genes themselves. Repairing *ATRX* did not restore chromosome accessibility back to the level of wildtype HUDEP.

Similar generalised peak height reduction was observed in the beta globin locus, with all of the hypersensitivity sites at the locus control region (LCR) visible (Figure 5-14). When compared to ATAC-seq track of differentiated $CD34^+$ cells (see previous Chapter), the peaks for *HBD* and *HBBP1* were lower across all clones. This was also observed in two of the three ATMDS patients. The BA12 clone with ATRX repaired showed partial restoration of ATAC peaks in the LCR and beta globin gene, which correlated with an increase in beta and gamma globin gene expression.

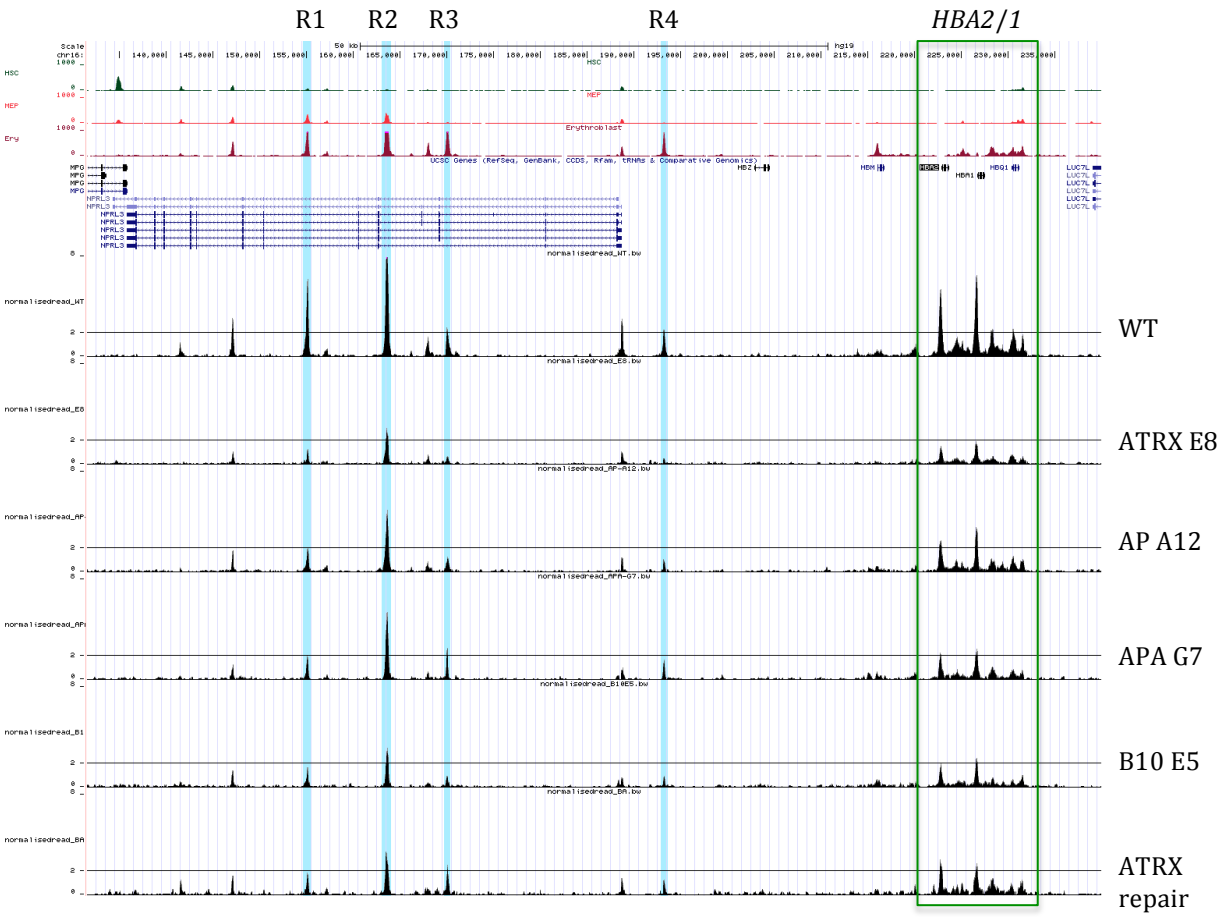


Figure 5-13. Normalised ATAC-seq tracks for differentiated, FACS-sorted HUDEP cells (population "M") on Day 10 of erythroid differentiation at the alpha globin locus. Tracks in black from top to bottom: wildtype (WT), ATRX knockout (ATRX E8), ATRX and PHF6 double knockout (AP A12), ATRX, PHF6 and ASXL1 mutant (APA G7), ATRX, PHF6, ASXL1 and U2AF1 mutant (B10 E5) and B10 E5 clone with ATRX repaired. The four regulatory elements are highlighted in blue (from left to right: R1-R4). The two alpha globin genes (*HBA2* and *HBA1*) and theta globin (*HBQ1*) are boxed in green. Y-axis represents number of reads per million sequencing reads.

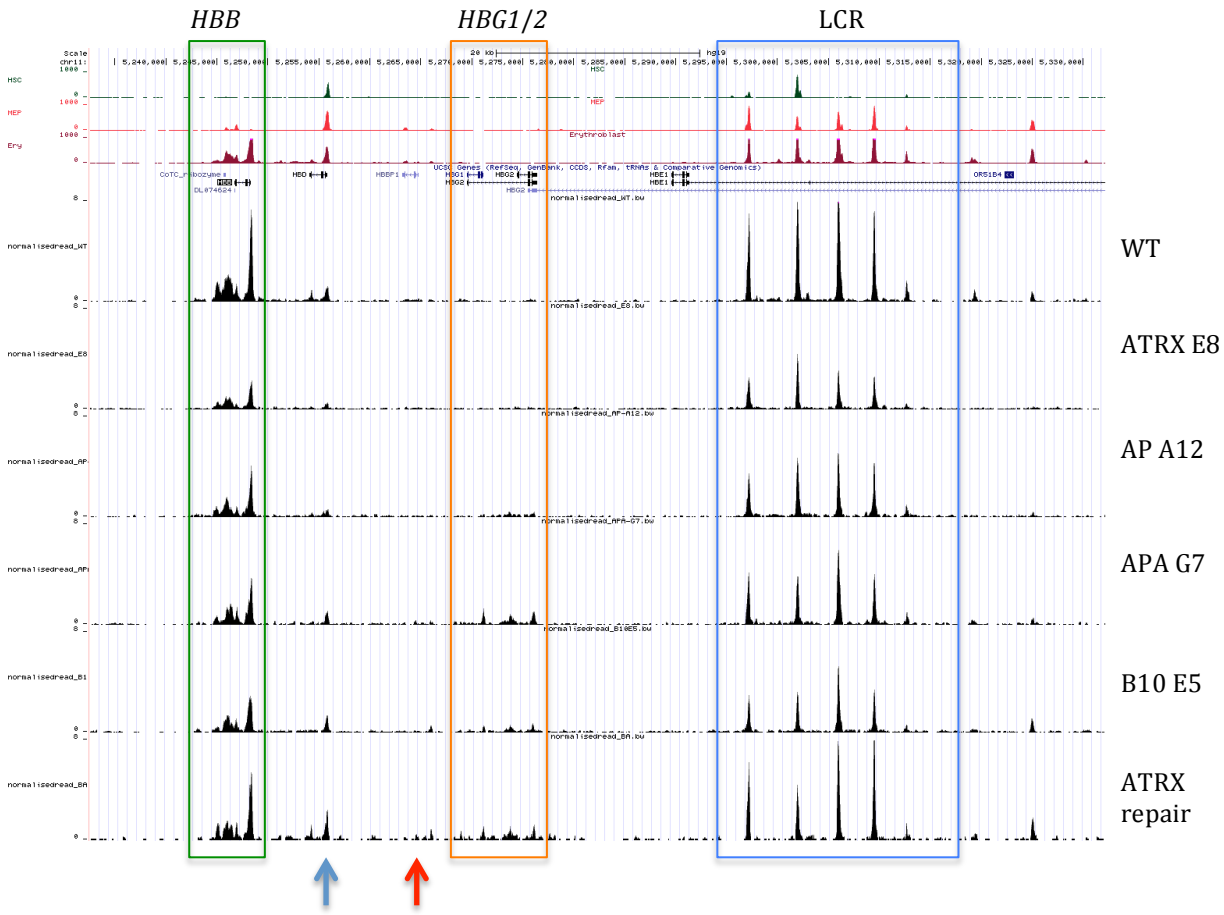


Figure 5-14. Normalised ATAC-seq tracks at the beta globin locus. From left to right beta globin gene is boxed in green, gamma globin genes (*HBG1* and *HBG2*) are boxed in orange. LCR is boxed in blue. Blue and red arrows point to the *HBD* and *HBBP1* regions, respectively. Y-axis represents number of reads per million sequencing reads.

5.8 HUDEP clone with only *U2AF1* mutation

Data from the globin gene expression analysis showed that *HBA* expression dropped further when *U2AF1* mutation was introduced in the triple mutant background. As mentioned before, modeling of MDS pathogensis suggested that *U2AF1* mutation is one of the earlier mutations, present before the appearance of *ASXL1* and *PHF6* mutations[115]. Therefore we sought to create a clone with U2AF1 p.Gln157Pro only in a wildtype background to delineate the role of U2AF1 in alpha globin expression. This clone would also form the base of an ATMDS HUDEP model that mimics the natural history of MDS in terms of sequence of mutation acquisition.

We used a similar two-step strategy as above to knock in the missense mutation (Figure 5-15). We first created an intermediate clone (UA1) with a heterozygous 5bp deletion (NG_029455.1:g.17917_17922delAGATG). The wildtype allele was unmodified despite inclusion of a ssODN to protect the wildtype allele. In the second step, a sgRNA specifically targeting the 5bp deletion allele was used to cut the mutated allele. The same ssODN (U2_Q157P_ssODN5) used to introduce the p.Gln157Pro mutation in the quadruple mutant B10 E5 clone was also used here to introduce the missense mutation.

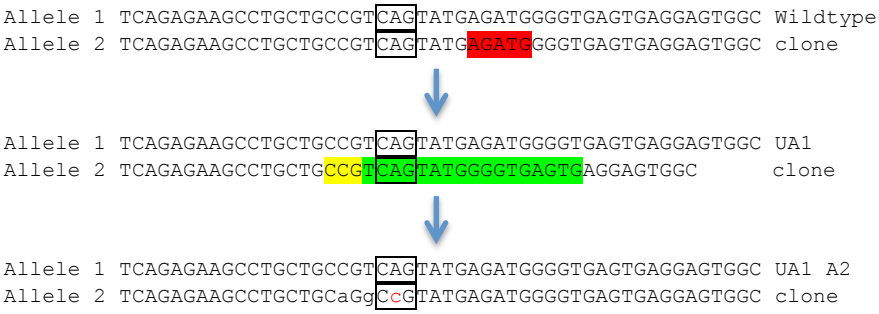


Figure 5-15. Schematic of introducing U2AF1 p.Gln157Pro in wildtype HUDEP cells. In the first step, one allele (Allele 2) was edited and a 5bp deletion (highlighted in red) was introduced, while Allele 1 remained completely wildtype. The p.Gln157 codon is boxed. The sgRNA sequence targeting specifically the 5bp deletion is highlighted in green with the PAM site in yellow (in reverse complement). The resultant UA1 A2 clone contains the missense mutation (small letter in red), along with two other deliberately introduced synonymous mutations (shown as small letters) in the preceding codon (CGT → aGg) to prevent re-editing in Allele 2. See Supplementary Figure 13 for Sanger sequencing traces.

The UA1 A2 clone differentiated similarly to other HUDEP clones (see Supplementary Figures 14 and 15 for serial FACS analysis and morphology comparison to wildtype). However, globin gene expression analysis showed a strong reduction in *HBA* and *HBB* expression, while a high *HBG* expression was noted (Figure 5-16). Alpha/non-alpha globin expression ratio was reduced to 0.59. This ratio was lower than all HUDEP clones we have differentiated and haemoglobin isoelectric focusing showed the presence of Hb Barts (γ_4) and Hb Portland ($\zeta_2\gamma_2$) (Figure 5-17).

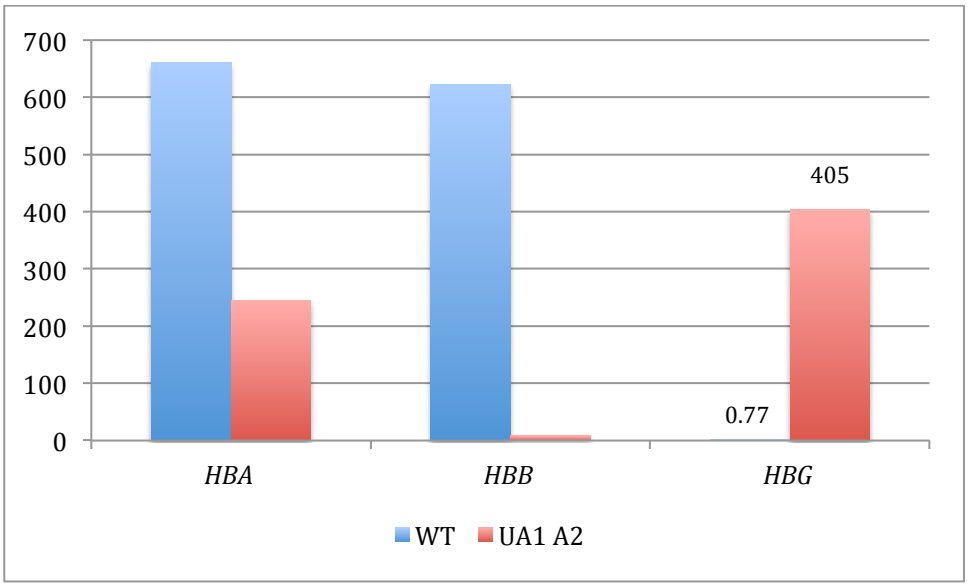


Figure5-16. Globin gene expression in differentiated UA1 A2 clone. Expression was normalised to *RPL13A*. Expression of gamma globin (*HBG*) was 405 fold over *RPL13A*, versus 0.77 fold in the wildtype.

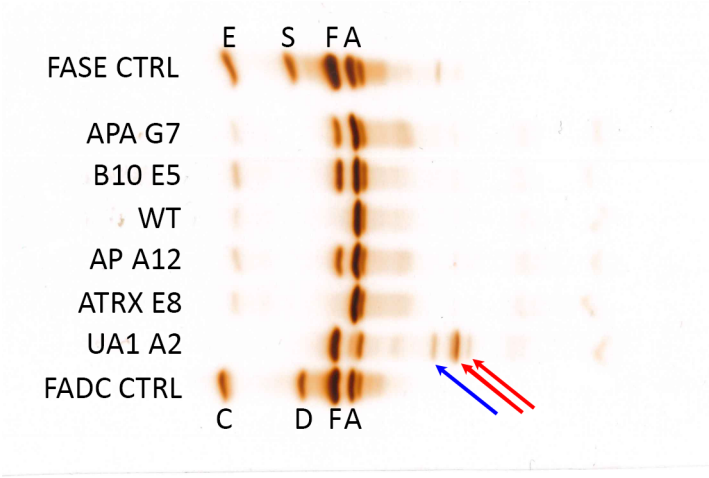


Figure 5-17. Haemoglobin isoelectric focusing of differentiated HUDEP cells. Hb F was not detectable in WT and ATRX E8 clone, consistent with the expression analysis by RT-PCR (Figure 5-11). In UA1 A2 clone, the two bands representing Hb Portland are arrowed in red. The band representing Hb Barts is arrowed in blue.

5.9 Summary

In this chapter, we have developed a protocol combining CRISPR/Cas9 genome editing and single cell FACS sorting to allow generation of pure genome-edited HUDEP-2 clone. This protocol is highly effective, with virtually 100% mutagenesis rate and requires only 25-30 days from transfection to being ready for cryopreservation. We have also managed to achieve a satisfactory rate of HDR via the use of HDR enhancer RS-1. This has made allele-specific knock-in and insertion of protein tag sequence possible (see publication resulted from this project)

The introduction of deliberate synonymous mutations effectively "protects" the allele from re-editing. Similar concept was published during the course of this project in a report by Paquet *et al*[278]. While the introduction of synonymous mutations may not be an issue for protein coding regions, it could affect protein-DNA interactions (transcription factors, spliceosomal components) in non-coding region. An extra round of editing to remove these synonymous changes may be necessary in these instances. In the future these issues may be solved by recent development in base editors. Base editors are chimeric proteins consisting of nuclease-inactive CRISPR/Cas9 fused to either activation-induced cytidine deaminase (for C→T change)[279, 280], or adenosine deaminase (for A→G change)[281]. They are able to introduce base changes without DNA cleavage. However currently there are several limitations, such as the lack of precision of editing and inability to edit a certain base in stretch of homopolymers.

With this protocol we have introduced mutations to knock out ATRX and PHF6 sequentially, followed by a heterozygous truncation in ASXL1 and finally a heterozygous p.Gln157Pro mutation in U2AF1. Due to technical difficulty initially, this order of mutagenesis was performed with the understanding that this may not reflect the sequence of events *in vivo* in ATMDS. Nevertheless, differentiation of these clones with increasing mutation load showed interesting features that may provide cues on the mechanism of acquired alpha thalassaemia in ATMDS.

ATRX knockout by itself did not lead to a block in differentiation according to the FACS profile, and only showed a modest decrease in alpha globin gene expression. This would be consistent with the mild phenotype seen in constitutive ATRX syndrome. We have not knocked out both isoforms of ATRX and therefore cannot exclude the possibility that the short isoform could contribute to alpha globin expression, although from the spectrum of *ATRX* mutation seen in ATMDS patients there was no indication that this is the case as half of the patients (3 out of 6) with ≥60% RBCs containing Hb H inclusions have mutations affecting the C-terminal ATPase domain.

The differentiation block seen in ATMDS patients became more apparent when *PHF6* and *ASXL1* were mutated, with loss of the more mature population by cell surface marker and appearance of immature erythroblasts in differentiated population. It appears that ASXL1 contributes to the differentiation block, in line with observations in MDS. However, PHF6 knockout and truncation of ASXL1 did not lead to

significant changes in alpha globin gene expression. Instead, gamma globin gene expression was significantly increased.

Introduction of the last mutation (U2AF1 p.Gln157Pro) led to a pronounced effect on globin gene expression. In fact, it was only then when alpha globin gene expression became significantly reduced, as well as the imbalance in globin gene expression. However, the drop in alpha/non-alpha globin gene expression ratio was smaller compared to what was observed in differentiated $CD34^+$ cells (from 1.84 in wildtype HUDEP to 1.44 in the quadruple mutant B10 E5, as opposed to 1.25 versus 0.62 in differentiated $CD34^+$ cells of normal control and ATMDS patients, respectively). Similar to ATMDS patients, increased gamma globin gene expression contributed to the lower alpha/non-alpha globin gene expression ratio. No Hb H or Hb Barts was detectable at the protein level by isoelectric focusing in these HUDEP clones, although it should be noted that even in ATMDS patients Hb H was only demonstrable in one case with a very high percentage of cells with Hb H inclusions. We were unable to obtain adequate number of enucleated RBCs for BCB staining.

As shown in the ATMDS cohort in this project, patients without *ATRX* mutation tend to have lower percentages of RBCs with Hb H inclusions. Therefore we repaired *ATRX* in the quadruple mutant B10 E5 clone to see if the globin imbalance could be improved. No significant change in the alpha/non-alpha globin expression ratio was demonstrated upon ATRX repair. However we could demonstrate a significant increase in expression

of all three globin genes upon the repair, which remained below the expression level of wildtype cells.

We looked at whether the chromosome accessibility pattern reflected the observations seen in differentiated $CD34^+$ cells and how each individual mutation altered chromosome accessibility. In the alpha globin locus, a generalised decrease in chromosome accessibility that was seen in ATMDS patients was also visible in FACS-sorted HUDEP mutants. More importantly, repairing ATRX did not increase chromosome accessibility despite the increase in expression, suggesting that the effect of ATRX on alpha globin gene downregulation was not due to altered accessibility *per se* at the enhancers or the two alpha globin genes themselves.

The significant drop in globin expression upon introduction of U2AF1 p.Gln157Pro mutation prompted us to examine this U2AF1 mutation in isolation. Given this mutation is not entirely unique to ATMDS, and it has been observed in MDS and other myeloid disorders, we did not expect U2AF1 p.Gln157Pro mutation to induce significant globin imbalance by itself. However, when this mutation was introduced into wildtype HUDEP, alpha globin gene expression was again reduced. More intriguing was the fact that beta globin gene expression was almost completely abolished, while the gamma globin gene expression was strongly induced. This led to a significant globin imbalance (alpha/non-alpha ratio of 0.59) that was observable at the protein level with the presence of the gamma globin

tetramer Hb Barts and the gamma globin containing Hb Portland on isoelectric focusing.

While it is interesting that introduction of U2AF1 p.Gln157Pro mutation in both the wildtype and triple mutant background led to suppression of alpha globin gene, the unexpected switch of beta to gamma globin in UA1 A2 clone casts doubt on whether HUDEP could faithfully recapitulate an acquired alpha thalassaemia phenotype. We have not confirmed the mutant/wildtype allele ratio in UA1 A2 clone. This will need to be addressed by massively parallel sequencing of *U2AF1* in the UA1 A2 clone. As mentioned, in the quadruple mutant B10 E5 clone only one of the three alleles contained the p.Gln157Pro mutation. It is known that splicing abnormalities induced by the other U2AF1 hotspot mutation (p.Ser34Phe) is proportional to the ratio of mutant/wildtype gene product[200]. If *U2AF1* mutation is implicated in globin gene expression then the mutant allele dosage may confound the observations seen in this clone.

It is recognised that only one clone from each combination of mutations was differentiated in the experiments above. Before HUDEP can be considered as a faithful ATMDS model, multiple clones containing mutations in these four genes, in isolation and in combination, will need to be produced and differentiated. Genomic variations that might have been introduced by genome editing need to be accounted for by copy number variation analysis and whole genome sequencing.

6 Discussion

6.1 Signature and significance of mutations in ATMDS

ATMDS presents as an ideal model to study epigenetic control of alpha globin gene expression as patients demonstrate a severe alpha thalassaemia phenotype with an intact alpha globin locus.

We have confirmed the extraordinary enrichment of *ATRX*, *ASXL1*, *U2AF1* and *PHF6* mutations in ATMDS, and these mutations are within the same cell. Mutations in spliceosomal genes, particularly *U2AF1*, are the most enriched among ATMDS patients. The mutation profiles suggest that in *ATRX* and *PHF6* they are loss of function mutations, while in *ASXL1* and *U2AF1* these are gain of function changes. Mutations in *ASXL1*, *U2AF1* and *PHF6* are independently associated in pairs in MDS[115, 116, 161], but the combination of these three genes, as well as the association with *ATRX* mutation, has not been described in other subtypes of MDS.

One aspect we have not fully investigated in ATMDS is the prevalence of cytogenetic changes. While chromosomes 16 and 11, where the two globin loci are located, are not common targets for cytogenetic changes in MDS, cytogenetic changes are found in 33% of MDS patients and *U2AF1* mutation is associated with del(7q)[115]. Cytogenetic changes may explain the gaps in mutation profiles in some ATMDS patients, as we have shown in the case of FW (case# 120).

Apart from *ATRX*, none of these genes have previously been associated with alpha thalassaemia. However, when these mutations are present the alpha thalassaemic phenotype is much more severe than that of constitutive *ATRX* mutation alone. Haemoglobin H inclusions were also detectable in patients with *ASXL1*, *U2AF1* and *PHF6* mutations only, supporting the involvement of these three genes in alpha globin gene expression. *ATRX* mutations were not as highly enriched in the ATMDS cohort as the other three genes and allelic frequency data from sequencing suggested it is a later mutation. However they are detected more commonly in ATMDS patients with higher percentages of cells with Hb H. This raises the possibility that *ATRX* mutations aggravate, rather than cause the acquired alpha thalassaemia phenotype. It must be noted that intronic mutations affecting *ATRX* splicing have not been excluded these patients. Alternative splicing event in *ATRX* is another possible explanation and this was detected by RASL-seq in MDS samples with *U2AF1* mutations but not in those with mutations in other splicing factors (*SRSF2*, *SF3B1*)[282], although this association was not reproduced in another study using RNA-seq[283]. It is also possible that in patients with low Hb H percentages the *ATRX* mutation is confined to a small subclone that is below the detection limit of the sequencing panel, highlighting the value of isolating $CD34^+$ cells for sequencing rather than relying on peripheral blood as input, as demonstrated in selected cases in this cohort.

Results from the sequencing panel presents a central question to this project – how does this particular mutation signature lead to the acquired

alpha thalassaemia phenotype? We attempted to address this via erythroid differentiation of $CD34^+$ cells from ATMDS patients and genome-edited HUDEP-2 cells.

6.2 Block in differentiation in ATMDS $CD34^+$ and HUDEP-2 erythroid differentiation

In the $CD34^+$ system, a block in differentiation was clearly visible as shown by cell surface markers and differential peak analysis of chromosome accessibility. This was accompanied by a reduction in alpha globin gene expression. Intriguingly, expression of beta globin was also repressed in FACS-sorted ATMDS $CD71^+$ $CD235a^+$ cells. Expression of beta globin gene was not affected in the normal controls and two MDS patients used in this study, as well as in normal control $CD34^+$ cells where alpha globin gene expression was downregulated via removal of the major enhancer element MCS-R2 in the alpha globin locus[284], although obviously these cells did not contain mutations affecting differentiation. The concomitant reduction in beta globin gene expression in this cell culture system therefore seems to be unique to ATMDS patients.

When we measured the alpha/beta globin gene expression ratio in peripheral blood in ATMDS patients, representing the end of erythroid differentiation, it was indeed significantly reduced. Therefore the cell culture system, while demonstrating the repression in alpha globin gene expression, failed to recapitulate the haemoglobin imbalance. Expression of beta globin gene initiates at least 12 hours behind alpha globin genes in

differentiating murine erythroleukaemia (MEL) cell line and heterokaryons formed by fusion of MEL cells and human non-erythroid cells[285]. In induced pluripotent stem cells with a GFP-based beta globin reporter system, increasing expression of beta globin correlates with maturation during erythroid differentiation[286]. Low viability from FACS sorting in cells undergoing terminal differentiation restricted the window for sampling. It is likely that in ATMDS, the block in differentiation means the events leading to haemoglobin imbalance in later stage of erythroid differentiation were not captured in the day 10 cell culture.

To overcome the issue of cell availability and to explore the contribution of each mutation to the phenotype, we introduced these mutations sequentially in HUDEP-2 cells. There was a noticeable but not statistically significant reduction in alpha globin gene expression when ATRX was knocked out and differentiation was not affected. A block in differentiation by cell surface markers was visible when *PHF6* and *ASXL1* mutations were subsequently introduced. The role of PHF6 in erythroid differentiation has not been examined in MDS previously, but the observation here is at least consistent with the reported role of ASXL1 in erythroid differentiation in MDS (see Chapter 3). Comparison of FACS profiles suggested mutations in these two genes led to additive effect on block of differentiation. In these clones the reduction in alpha globin gene expression was also modest, but there was an increase in gamma globin expression. It was only when U2AF1 p.Gln157Pro mutation was knocked in that led to a significant reduction in alpha globin expression, despite the higher basal expression in

the quadruple mutant clone. As in $CD34^+$ cell culture, beta globin expression was also downregulated, but in HUDEP-2 gamma globin expression was significantly induced. Repairing *ATRX* in the quadruple mutant did not resolve the block in differentiation. In terms of globin gene expression, *ATRX* repair only partially restored alpha and beta globin gene expression, but gamma globin gene expression was not repressed.

We examined the contribution of U2AF1 p.Gln157Pro to globin gene expression in isolation by introducing the mutation on a wildtype background. Expression of both alpha and beta globin gene was repressed upon differentiation, with beta globin gene expression reduced to almost background level. However there was a substantial increase in gamma globin gene expression and intriguingly haemoglobin imbalance was visible at the protein level with appearance of Hb Barts. These observations lead to two questions – could U2AF1 p.Gln157Pro mutation play a role in alpha globin gene expression, and what is the mechanism for this haemoglobin switch in HUDEP-2 cells?

6.3 The potential role of spliceosomal and *ATRX* mutations in alpha globin gene expression

As mentioned in the Introduction, an alternative splicing programme with intron-retention occurs during terminal erythroid differentiation, involving genes encoding other splicing factors such as *SF3B1*, mitochondrial iron importers and major cytoskeletal proteins[112, 113, 287]. Haemoglobin genes (*HBA1* and *HBB*) are efficiently spliced in cultured erythroblasts

derived from $CD34^+$ cells of normal individuals[112]. So how could spliceosomal mutation affect alpha globin gene expression? A clue comes from examination of the second intron of alpha globin gene. This intron is short, and has an unusual base composition with high guanine content (42%). These two properties are associated with differential splicing in MDS patients with *SF3B1* mutations[288]. The alpha globin gene intron contrasts with intron 2 of beta globin gene that is longer and AT-rich (see Appendix for comparison between *HBA2* and *HBB* sequences). The guanine in *HBA2* intron 2 presents as multiple triplets. G triplets are important for 5' and 3' intron boundary definition as mutations in these triplets reduced spliceosome assembly, and the G triplets are thought to compensate for the suboptimal 3' pyrimidine tract in this intron[289]. In introns with suboptimal pyrimidine tract, it was shown that interaction between U2AF1 and the 3' splice site is critical for efficient U2AF heterodimer binding[193, 290-292]. It is therefore attractive to speculate that mutations in *U2AF1* may affect alpha globin gene pre-mRNA splicing.

Splicing is coupled with transcription. The C-terminal domain (CTD) of RBP1, the largest subunit of RNA polymerase II, is required for efficient splicing and U2AF65 binds directly to this domain[189]. In a live cell imaging model using fluorescent reporter within intron 2 of beta globin gene, Coulon *et al*[293] showed that 15% of beta globin reporter transcripts were spliced prior to release from the chromatin. However, when U2AF1 p.Ser34Phe mutation was introduced, this pre-release splicing was abolished. In addition splicing in freely diffusing transcripts was also

slowed down. This observation was validated in the endogenous *FXR1* gene. As alpha globin gene expression increased during erythroid differentiation, defective splicing could therefore lead to accumulation of unspliced alpha globin gene RNA.

The high guanine content in the nascent alpha globin RNA containing intron 2 can promote formation of R-loop in the process of transcription. Mutations or pharmacological inhibition of spliceosomal genes, including *SRSF2*, *U2AF1* and *SF3B1*, have been shown to induce R-loop[294-296]. ChIP experiments specifically targeting R-loops in HEK293T cells carrying *U2AF1* mutations in the two hotspot locations showed that the mutations induced R-loop formation genome-wide. In particular, p.Gln157Pro mutation was more capable in inducing R-loops compared to p.Ser34Phe mutation[295]. This would fit in with the observation that p.Gln157Pro mutation is the predominant U2AF1 mutation in ATMDS.

What would R-loops do during alpha globin gene transcription? The formation of R-loop leaves a single stranded G-rich template DNA, which could form G-quadruplexes. Indeed, the G triplets in intron 2 of alpha globin gene are predicted to form G-quadruplexes (see Appendix). Using a real-time FRET-based (Förster Resonance Energy Transfer) technique, intramolecular formation of G-quadruplexes was shown to be proportional to transcriptional activity, and formation of R-loop stabilized G-quadruplexes[297]. R-loops and G-quadruplexes can both block the passage of RNA polymerase complex[298-301], and R-loop formation and

resolution occurs dynamically with transcription[302]. Timely resolution of R-loop and G-quadruplex is therefore important for normal transcription.

This brings ATRX into the picture. ATRX is recruited to telomere repeats that can form R-loops, and the recruitment is dependent on transcription of these telomeric repeats. Knockout of ATRX increases R-loop formation[303]. ATRX binds to gene bodies of alpha globin genes (Julia Truch, unpublished data). In ATMDS, it is possible that loss of ATRX aggravates the transcription block caused by R-loop formation secondary to *U2AF1* mutation, due to the inability to resolve R-loop structure in alpha globin pre-mRNA.

Studies on U2AF1 mutations to date did not examine specifically cells with p.Gln157Pro mutation during terminal erythroid differentiation. To answer the questions raised here, RNA-seq data especially on terminal erythroid cells from ATMDS patients would provide valuable information on whether alpha globin mRNA is processed normally and shed a light on other differentially expressed genes that could contribute to this phenotype. We will need to establish whether R-loop is formed within the alpha globin genes in wildtype and U2AF1 mutant, and whether R-loop is augmented in the absence of ATRX in differentiating cells.

6.4 *ASXL1* and *PHF6* in RNA polymerase II C-Terminal Domain phosphorylation

The role of the other two enriched genes, *PHF6* and *ASXL1*, in alpha globin gene expression is less clear from the expression data in HUDEP-2. However, it must be noted that in this project these mutations were introduced prior to U2AF1 mutation, and the effect of mutations in these genes may not be realized in the context of normal splicing. An example is that *ASXL1* itself is differentially spliced in K562 cells and AML patients with U2AF1 p.Ser34Phe mutation[205].

In the case of ASXL1, an additional layer of complication comes from reports that ASXL1 truncation may have loss or gain of function effects. Early studies on ASXL1 function were performed by knockdown or knockout in the assumption that it was a loss of function mutation. However, it was later shown that the truncated ASXL1 is detectable at a protein level[166]. While both loss of ASXL1 and overexpression of truncated ASXL1 promoted myeloid dysplasia and leukaemic transformation[304-306], the truncated protein gained functions in 1. enhancing activity of the deubiquitinase BAP1[168, 169, 307]; 2. inhibition of polycomb repressive complex 2 function[306]; and 3. acquired interaction with the BET (bromodomain and extra terminal domain) protein BRD4[165]. The latter interaction may be implicated in ATMDS as we will discuss below.

A possible role of *PHF6* and *ASXL1* in ATMDS lies on their potential involvement in RNA polymerase II CTD phosphorylation. Phosphorylation of serine 2 in the CTD has been shown to be important in splicing, as replacing serine 2 with alanine resulted in loss of U2AF65 recruitment[308]. Conversely, inhibition of splicing specifically decreased the level of serine 2 phosphorylation[309]. Serine 2 phosphorylation is achieved by recruitment by the positive transcription elongation factor P-TEFb. Activity of P-TEFb is sequestered in an inactive, multimeric 7SK snRNP complex[310-312], and release of P-TEFb from 7SK complex is required for its activity. Interestingly, SRSF2 is part of the 7SK complex at promoter and R-loop in SRSF2 mutant was formed due to impair transcription pause release secondary to inability to release P-TEFb from the 7SK complex[295, 313].

Activation of P-TEFb is achieved by bromodomain-containing protein 4 (BRD4), a member of the bromodomain and extra terminal (BET) family[314-316] and the polymerase associated factor PAF1[317, 318]. BRD4 also directly phosphorylates serine 2 *in vitro* and *in vivo*[319]. Whether the acquired interaction between truncated ASXL1 and BRD4 affects its ability to phosphorylate serine 2 remains to be investigated. The link between PHF6 and PAF1 has been mentioned in Chapter 3. PAF1 is also enriched particularly at 3' ends of R-loop positive genes[302]. Mutations *ASXL1* and *PHF6* may therefore contribute to the block in transcription via inhibition of CTD serine 2 phosphorylation, which in turn aggravate co-transcriptional splicing in the alpha globin gene in the context

of a co-mutated *U2AF1*. We will need to establish whether these components, namely PHF6, truncated ASXL1, BRD4 and PAF1, localise to alpha globin locus in wildtype and ATMDS cells. Examination on the level of RNA PolII CTD phosphorylated serine 2 in the alpha globin genes in differentiating cells would provide evidence to this hypothesis.

6.5 Links of ATMDS mutations to PRC2

Another complex that links these four genes together is the PRC2 complex. PRC2 complex composes of four core subunits (EZH2, EED, SUZ12 and RBBP4) and two cofactors (AEBP2 and JARID2)[239, 320]. The physical interaction between PHF6 and RBBP4 has been well described (see Chapter 3). EZH2 is involved with ATRX, ASXL1 and U2AF1. For *ATRX*, an EZH2 binding domain was identified by yeast two-hybrid analysis[59], although this association was not observed in co-immunoprecipitation experiments (Clare Jelinska, unpublished data). As mentioned above, ASXL1 also interacts with PRC2 via EZH2[167]. Interestingly, increased usage of a cryptic exon that leads to premature termination in EZH2 is detected in patients harbouring U2AF1 p.Ser34Phe mutation, but not in those with p.Gln157Pro mutation[288]. In the ATMDS cohort, only one patient has an EZH2 mutation (p.Gln735Ter, case #125) and sequencing of $CD34^+$ cells in this patient showed a subclonal U2AF1 p.Ser34Phe mutation with a VAF of 5.5%. Together with the observation of strong enrichment of p.Gln157Pro mutation in ATMDS, it may suggest that an intact EZH2 is required for the acquired alpha thalassaemia phenotype. Given PRC2 clearance is required

for alpha globin gene expression, it would be worthwhile investigating changes in PRC2 function or localization in ATMDS.

6.6 HUDEP-2 as a model for erythroid differentiation

Globin gene expression data from this project indicated that HUDEP-2 may not be the ideal model for studying haemoglobin imbalance. While a block in differentiation could be seen and alpha globin gene expression was indeed significantly repressed upon introduction of all four mutations in ATMDS, in the beta globin locus these genetic manipulations seemed to have induced haemoglobin switching from adult to fetal globin expression. Elevated fetal haemoglobin was noted in the two of the three original cases of acquired Hb H disease[123], as well as patients with acquired Hb H disease on the background of aplastic anaemia and primary myelofibrosis[321, 322].

The level of fetal haemoglobin in adult is variable and is considered a quantitative trait. Multiple genetic factors have been identified as modifiers, including *Xmn-HGB2* variant within the beta globin locus, variants in *HBS1L-MYB* intergenic region and microRNAs (miR-15a and miR-16) that control MYB level (see below), and *KLF1* mutations or hypermethylation of its enhancer[323] (for review, see [324, 325]). There are also two known gamma globin repressors, namely BCL11A and ZBRB7A (also called LRF), that bind directly to independent sites at the gamma globin promoter and mutations at these sites induce gamma globin

transcription in HUDEP-2 cells[17, 326]. The switch may also be indirectly due to changes in the kinetic of erythroid differentiation. In stress erythropoiesis it is known that erythroid cells express fetal haemoglobin[327, 328]. Changes in MYB level alters this kinetic and the level of fetal haemoglobin-producing red cells[329-331].

The reason for this haemoglobin switching in HUDEP-2 is not clear. HUDEP-2 originated from cord blood $CD34^+$ cells, which normally expresses fetal haemoglobin. In the original report, HUDEP-2 was in fact the only outlier cell line that expressed beta globin and *BCL11A*, while all other cell lines expressed gamma globin only[150]. This switch has also been observed in various genome-edited HUDEP-2 clones by other groups, even between two clones with the same genome modification.

In this project, the induction of gamma globin expression became apparent when PHF6 was knocked out. As mentioned, PHF6 binds to RBBP4. Using an nuclease-deficient Cas9-based capture technique, Liu *et al*[174] confirmed that components of NuRD complex (RBBP4, HDAC1 and CHD4) are located in the beta globin locus, and knockdown of these components by RNA interference in human primary erythroid cells was shown to increase gamma globin expression. The PHF6 binding site on RBBP4 is shared with FOG-1 (Friend of GATA protein 1), histone H3 lysine 4[240] and more recently BCL11A[332]. We will need to assess *BCL11A* expression in the genome-edited clones. It would also be interesting to examine the effect on gamma globin expression by disrupting RBBP4 binding motifs in

PHF6 without knocking the protein out. The level of the transcription factors (MYB, KLF1) influencing Hb F level may also provide clues to the underlying cause of this switch.

There are two possible solutions to this issue of undesired gamma globin expression. We can either immortalize $CD34^+$ cells from ATMDS patients. However given the underlying genetic aberrations in these patients they may not have a suitable genetic background for comparison. In fact, we attempted to reprogramme $CD34^+$ cells from one ATMDS patient (ECT, case #122) to produce induced pluripotent stem cells. Only one clone from ECT (out of a total of eight between the two patients) contained the four mutations, and this clone was difficult to remain pluripotent (with assistance from Yavor Bozhilov, data not shown). Recent comparison of transcriptomes along erythroid differentiation between cord blood and peripheral blood derived $CD34^+$ cells showed differences in gene expression pattern in various clusters, especially at the very early and very late stages of erythropoiesis[333]. It is therefore important to use an adult-derived cell line as the basis for genome editing. An example would be the immortalized progenitor cell line derived from adult $CD34^+$ cells (BEL-A, Bristol Erythroid Line-Adult) that is known to express exclusively beta globin[334].

A major shortcoming in these experiments is that we have not accounted for the potential effect of clonal variability, and each gene was only targeted by one sgRNA. Reproducing the same phenotype using an independent

sgRNA would be essential to validate these findings. Although HUDEP-2 may not be the ideal model for ATMDS, the genome editing protocol developed in this project with high efficiency and high HDR rate paves way for studying various diseases related to erythropoiesis, and the results here have provided insights for future attempts in producing a faithful cell line model of ATMDS.

7 Appendix

Supplementary Table 1. List of ATMDS patients with clinical and mutation data

Sample	Initial	Sex	Hb (g/dL)	MCV	MCH	BCB (%)	Age at presentation	Clinical diagnosis	Karyotype	ASXL1 NP_056153.2 12	PHF6 NM_032458.2; NP_115834.1 2-10	U2AF1 NP_001020374.1 2 & 6	ATRX NM_000489.3; NP_000480.2 combined results from myeloid panel sequencing (exons 8-10, 17-31), Sanger sequencing and DHPLC	TET2 NM_001127208.2; NP_001120680.1 3-11	RUNX1 NP_001745.2 3-8	JAK2 NP_004963.1 12 & 14	CBL NP_005179.2 8-9	SETBP1 NP_056374.2 4	KRAS NP_203524.1 2-3	SRSF2 NP_001182356.1 1	IDH2 NP_002159.2 4	NRAS NP_002515.1 2-3	TP53 NP_000537.3 4-9	ZRSR2 NP_005080.1 1-11	CSF3R NP_724781.1 14-17	DNMT3a NP_072046.2 23	CEBPA NP_004355.2 1 (partial)	EZH2 NP_004447.2 2-20	KIT NP_000213.1 2, 8-11, 13, 17	SF3B1 NP_036565.2 14-15	CBLB NP_001308715.1 9-10	CBLC NP_036248.3 9-10	ETV6 NP_001978.1 1-8	Additional mutations detected in previous panel in genes that are not covered in the current myeloid panel sequenicng at Oxford			Gene NCBI Reference sequence Exon(s) sequenced	Number of genes mutated
82	BT	M	7.5	59	14.2	80	N/A	N/A	N/A			p.Gln157Pro (35.34%)	p.Arg2407Ter (75%)					p.Gly870Ser (37.81%)			p.Ala59Asp (38.56%)											SUZ12 p.Thr370Met (52%)			6	
122 (CD34+)	ECT	M	7.8	60	16.2	79.8	N/A	RCMD	Normal	p.Gly646TrpfsTer12 (39.97%)		p.Gln157Pro (48.28%)	c.585+2T>C (98.14%)					p.Asp868Asn (48.37%)	p.Gly60Asp (6.57%)																6	
89	BW	F	9.4	64	16	76	N/A	RARS	Normal	p.Gly646TrpfsTer12 (31.39%)		p.Gln157Pro (39.92%)	c.374+1G>A (33.64%)	p.Cys174Tyr (40%)																						4	
54*	AK	M	11	69.3	18.9	75	83	RAEB	N/A		p.Arg225Ter (86.41%)	p.Gln157Pro (45.75%)	p.Cys265Tyr (72%)													p.Arg882Ser (41.21%)										4	
61*,*	JE	M	9.4	68.6	17.7	60	67	RAEB	Normal	p.Tyr591Ter (41.58%)	p.His302Leu (83.6%)	p.Gln157Pro (42.12%)	p.Leu2240Phe (78.67%); splicing mutation (intron 18) (% N/A)					p.Asp868Tyr (40.35%)		p.Pro95His (46%)																6	
97	MM	F	12.5	88.2	30.4	60	N/A	MDS NOS	N/A	p.Tyr591Ter (44.84%)	c.418+1G>A (32.9%)		p.Cys200Tyr (34%)																								3	
102	JB	M	12	80	28.1	50	59	RCMD	Normal	p.Tyr591Ter (48.1%)	p.Arg118Ter (98.75%)	p.Gln157Pro (48.37%)										p.Gly12Asp (19.37%)														4	
27#,*	JP	M	10.5	65	19.9	50	76	CMD NOS	46,XY[13]/47,XXY[1]	p.Glu635ArgfsTer15 (22.8%)	p.Glu271Ala (93.91%)		p.Leu192Phe (13%)			p.Val617Phe (38.57%)																						4	
46*	JS	M	9.2	83	24.5	48	70	RA	N/A	p.Glu635ArgfsTer15 (12.79%)	p.Gln251Ter (51.85%)	p.Gln157Pro (23.68%)	p.Cys240Gly (34.77%)																								4	
123 (CD34+)	MK	M	8.5	76	24.3	42.1	68	RCMD	Normal	p.Arg634LysfsTer24 (50.59%)	p.Arg319Ter (99.41%)	p.Gln157Pro (47.2%)	p.Trp263Cys (99.23%)																								4	
42*	DW	M	12.3	59.8	16.1	40	71	RA	Normal	p.Glu635ArgfsTer15 (20.18%)	p.Cys20LeufsTer13 (74.33%)	p.Gln157Pro (44.45%)	c.6884_6885insT;p.Asn2296Glnfs*26 (72%)				p.His398Pro																				5	
84	EW	F	8.1	76	22.4	35.1	N/A	RARS	Normal	p.Gly646TrpfsTer12 (36.39%)	c.834+1G>A (20.56%)	p.Gln157Pro (48.4%)	p.Asp2234Gly (9%);p.Arg2131Gln (7%)					p.Asp868Asn (2.04%)	p.Ala59Thr (9.98%)						p.Tyr814Ter (2.7%)											7	
45*,*	JP	M	10.2	67	15.4	35	60	RA	Normal	p.Arg693Ter (39.11%)	p.Ile314Thr (63.81%)	p.Gln157Pro (35.14%)	p.Ser79* (77%)	p.GLu178Lys (18%)												p.Arg882His (35.53%)										6	
50*	AS	M	13.2	88.4	29.5	30	79	RA	N/A	p.Glu635ArgfsTer15 (43.8%)	p.Gly291Arg (75.23%)	p.Gln157Pro (37.09%)	p.Arg2028Pro (14.88%)																								4	
120	FW	M	9.1	75	24	27	61	N/A	N/A	deletion in SNP array	deletion in SNP array	p.Ser34Phe (18.78%)																									3	
49*	GL	M	8.6	69.4	20.6	25	60	RA	N/A	p.Tyr591Ter (45.41%)	p.Cys297Ser (81.33%)	p.Gln157Pro (41.59%)	c.7014_7017delAACA; p.Asn2340Valfs*2 (% N/A)	p.Ala1283Ser (5.66%)	p.Arg250Ser (13.79%)									p.Val73Ala (5.15%)	p.Gln32Ter (17.88%)											8	
125 (CD34+)	KH	M	11.7	77	22.1	24.7	N/A	RAEB	Normal		p.Gly291Glu (97.21%)	p.Ser34Phe (5.46%)	p.Asp203Val (97.44%)			p.Val617Phe (25.18%)												p.Gln735Ter (45.83%)								5	
94	S	M	7	66	N/A	21.2	40	N/A	46,XY,-19,+mar[2]/46,XY[19]	p.Gly646TrpfsTer12 (18.1%)	p.Arg183ProfsTer55 (44.36%); p.Lys164Gln (46.75%)	p.Ser34Phe (22.74%)		p.Gln599Arg (50%)																					p.Arg369Gln (36%)			5	
77	PN	M	7.7	73.7	N/A	20	73	MDS/AML	N/A	p.Glu635ArgfsTer15 (24.53%)	p.Gly186GlufsTer32 (87.17%)		splicing mutation affecting exon 28 (% N/A)																	p.His662Gln (49.15%)					4	
105	MG	M	12.7	69	20.1	15	73	Therapy related MDS	N/A		p.Ile314Thr (15.37%)			p.Pro333HisfsTer14 (9.28%); p.Ser424ArgfsTer3 (10.44%); c.4183-2A>G (46.17%)						p.Pro95Leu (38.46%)																3	
29#,*	RC	M	8.8	82	27.7	14	86	?RARS	N/A	p.Gly646TrpfsTer12 (29.02%)	p.Arg116Ter (76.47%)	p.Gln157Pro (37.45%)	c.5567-3C>G (71%)				p.Lys382Thr (38.18%)																		KDM6HB p.Arg666Gln (39%)	NOTCH1 p.Asp809Asn (46%)	TLR2 p.Leu154Pro (6%)	8	
65*,*	JC	M	11.3	62	19.2	12.7	70	CMD NOS	46XY,del(7)(q11.2p15)[28]/46,XY[2]	p.Trp583Ter (48.94%)			c.20+1G>A			p.Val617Phe (53.01%)				p.Pro95His (56.41%)														NF1 p.Trp223Arg (20%)			5	
79	LD	M	7.7	93	24.1	12	83	RCMD	N/A	p.Gly646TrpfsTer12 (25.25%)	p.Asp25GlnfsTer10 (59.38%)	p.Gln157Pro (29.34%)		p.Val1225Gly (14.81%)	p.Arg201Ter (29.93%)																				TERT p.Glu441del (15%)			6	
118 (CD34+)	NQ	M	8.9	59	17.4	12	74	N/A	N/A	p.Gly646TrpfsTer12 (42.67%)	p.His104GlnfsTer9 (96.38%)	p.Gln157Pro (48.13%)	p.Cys1576Ser (99.41%)						p.Gly870Arg (5.59%)	p.Tyr71Asp (15.43%)					p.Asn816ThrfsTer13 (18.14%); p.Ser810AlafsTer19 (8.28%); p.Thr615Ala (5.23%)											7	
81	RT	M	7.1	74	26.1	10	71	N/A	N/A	p.Gly646TrpfsTer12 (33.91%); p.Asp1127Asn (49.79%)	p.Arg319Ter (11.08%); p.Glu274AsnfsTer6 (14.42%); p.Leu175CysfsTer43 (60.31%)	p.Gln157Pro (41.85%)		p.Pro1594GlnfsTer21 (43.84%)	p.Leu565er (49.34%)												p.Thr691Lys (12%)								GATA2 p.Ala411Val (59%)			7	
43*	CW	M	10.2	91	29	6.6	68	RA	47, XY, del(20)(q11),+8			p.Ser34Phe (37.78%)			p.Arg1669ProfsTer47 (8.04%); p.Ala149_Ala150InsThrAla (25.18%)				p.Gly12Asp (7.55%)																	3	
26#,*	EJ	M	5.9	97	20.9	5	61	MMM	N/A	p.Arg693Ter (47.29%)	c.730-2A>G (89.76%)	p.Gln157Pro (47.43%)	c.7049delT;p.Glu2351Argfs*3 (% N/A)			p.Val617Phe (5%)	p.Cys419Arg (28.97%)																				6	
51*	LF	M	10.3	69.8	22.7	4	60	RA	46, XY[inv 9]	p.Gly646TrpfsTer12 (33.01%); p.Gln760Ter (5.83%)	c.241-1G>A (83.37%)	p.Gln157Pro (47.41%)			p.Cys995ArgfsTer24 (39.03%)					p.Pro6His (8.56%)						p.Ser190Pro (10.9%); p.Pro188_Pro189ArgSer (5.07%); p.Ser85Gly (21.28%)		p.Ser590Asn (6.42%)								7	
103	HW	M	10.2	76	28.1	3	84	RCMD	Normal	p.Glu635ArgfsTer15 (19.84%)	p.His3027Tyr (76.41%)	p.Gln157Pro (38.91%)	p.Arg2028Pro (7.41%); p.Cys2405ser(6.87%)		p.Gly165Arg (41.24%)																						5	
90	AG	M	8.9	80.4	26.9	1.3	N/A	N/A	N/A	p.Leu739SerfsTer5 (25.85%)	p.Leu327ThrfsTer4 (7.76%); p.Arg319Ter (16.43%)	p.Gln157Pro (28.37%)		p.His1868Tyr (25.71%)		p.Val617Phe (14.1%)	p.Cys396Tyr (15.1%)																		NOTCH2 p.Phe1209Val (57%)			7	
78	AK	F	10	63	20	1	72	RCMD	N/A	p.Pro701SerfsTer16 (43.42%)				c.4183-1delG (43.68%)																							2	
40	CS	M	9.3	84.9	28.9	1	75	?RCMD	N/A	p.Tyr591Ter (41.53%)	p.Arg319Ter (60.94%)	p.Gln157Pro (38.35%)		p.Gln481Ter (42.71%)																							4	
96	RA	M	9.7	72	24.4	0.9	73	RA	Normal	p.Gly646TrpfsTer12 (34.13%)	p.Ile314Thr (40.94%)	p.Gln157Pro (34.81%)									p.Arg140Gln (36.22%)															4	
69*	FR	M	7.5	72	23.7	0.5	N/A	RA	Normal	p.Gln760Ter (47.56%)	p.Gly287Asp (98.89%)	p.Gln157Pro (49.52%)				p.Val617Phe (51.25%)																					4	
67^^	AW	M	10.4	64	21.7	0.11	72	RARS	Complex change including chr16 deletion: 45,X,-Y,add(3)(q25),der(3)add(3)(?q25),del(4)(q?15),der(5)add(5)(p?31)del(5)(p?34),del(6)(p?21)),-7,der(11)add(p?11)add(q?21),der(13)t(13;21)(p11;q11),der(16)add(p?)(?13)add(q?22)),-18,-20,-21,+mar [cp10], 46,XY[1]				p.Arg1359Cys (38.59%)								p.Arg213Ter (43.18%); p.Gly154AlafsTer16 (38.66%)	p.Phe525Ser (99.79%)												3	
68*	RD	M	12.6	75	26.1	0.1	68	MDS NOS	N/A	p.Gly646TrpfsTer12 (29.82%)	p.Arg319Ter (11.11%)	p.Gln157Pro (34.88%)			p.Leu985ArgfsTer24 (35.7%)	p.Val617Phe (57.24%)	p.Lys382Asn (7.65%)				p.Arg140Gln (6.23%)															7	
									Number of patients with mutation	31	31	29	20	10	7	7	5	4	2	2	2	2	3	1	1	1	1	0	0	1					
									Percentage of patients with mutation	86%	86%	81%	56%	28%	19%	19%	14%	11%	6%	6%	6%	6%	8%	3%	3%	3%	3%	0%	0%	3%					
									Total number of patients	36																										

Patients are ranked in descending order of BCB % from top to bottom. Genes are listed from left to right in descending order of enrichment in this cohort. The total number of gene(s) mutated is shown on the right. The total number of mutations in a particular gene in this cohort is shown at the bottom.

No mutations were detected in the following genes: CBLB, CBLC, FLT3, HRAS, IDH1, MPL, NPM1, PDGFRA, PTEN and WT1

*These individuals were described in Steensma *et al* 2004. *Blood* 103(6):2019-26 (PubMed ID: 14592816).

^These individuals were described in Gibbons *et al* 2003. *Nat Genet* 34,446-9 (PubMed ID: 12858175)

^^This individual with chromosome 16 deletion was described in Steensma et al 2004. *Blood* 103:1518-20 (PubMed ID: 14576055)

#These individuals were described in Higgs et al 1983. *Am J Med* 75(2):181-91 (PubMed ID: 6881169)

Abbreviations

CMD Chronic Myeloid Dysplasia
MDS Myelodysplastic Syndrome
MMM Myelofibrosis with Myeloid Metaplasia
NOS Not Otherwise Specified
RA Refractory Anaemia
RARS Refractory Anaemia Ring Sideroblast
RAEB Refractory Anaemia Excess Blasts
RCMD Refractory Cytopaenia multilineage dysplasia

Colour scheme

Known mutation
ATRX mutation detected by Sanger sequencing/DHPLC but not covered in myeloid panel sequencing
Variant of unknown clinical significance
More than one mutation detected
Cases #26, 27, 29, 42, 45, 54, 61, 65, 67, 68, 69, 77, 78, 79, 81, 82, 84, 89, 90, 94, 96, 97 and 103 were investigated previously using an independent panel (Andrew Campbell, unpublished data) in addition to the myeloid panel at Oxford. Additional mutations not detected by the myeloid panel used in this project are highlighted.
Cases #40, 43, 46, 49, 50, 51, 90, 102, 105, 120, 118, 122, 123 and 125 were investigated only by the myeloid panel at Oxford for mutations in genes other than ATRX

Patient	Sex	MDS diagnosis	Karyotype	Gene	Mutation	Protein change	VAF (%)	FISH ISCN
1	M	RCMD	46, XY	*ATRX*	c.A1250G	p.K417R	3.03	nuc ish 4q24(RP11-351K6x2,RP11-16G16x2),12p13(ETV6 upstreamx2,ETV6 downstreamx2)(ETV6 upstream sep ETV6 downstreamx0),17q11(NF1x2),17q22(MPOx2),20q11.21(5`ASXL1x2,3`ASXL1x2),20q11.21(DNMT3Bx1),20q12(3`PTPRTx1,5`PTPRTx1),20q13.32(GNASx2)
				U2AF1	c.C101T	p.S34F	37.40	
				ASXL1	CNV			
2	M	RAEB-1	46, XY	*ATRX*	c.A2357G	p.E786G	97.93	nuc ish 8cen(D8Z2x2)
				TET2	c.G2527T	p.E843*	43.15	
				TET2	c.3803+2T>G		21.74	
				TET2	c.5033_5034del	p.1678_1678del	18.64	
				ZRSR2	c.827+1G>A		35.80	
				ZRSR2	c.C952T	p.Q318*	49.81	
3	F	RAEB-1	45X, -X	*ATRX*	c.C1122A	p.D374E	4.05	N/A
				KRAS	c.G34A	p.G12S	7.76	
				TET2	c.1841_1842insA	p.G614fs	81.67	
				ZRSR2	c.1252delC	p.H418fs	39.10	
4	M	RAEB-1	Complex, XY	*ATRX*	c.G6218C	p.R2073P	62.37	nuc ish 3p25.3(FANCD2x1),3q26.2(D3S1243x3,hTERCx3)(D3S1243 sep hTERCx0),5p15.2(D5S721,D5S23x2-3),5q31(EGR1x1),5q33-q34(DTDST,CSF1R,PDGFRBx1),7cen(D7Z1x2),7q31(D7S486x2),17p13(p53x2)
				TP53	c.1054-1G>C		27.50	
				TP53	c.G422A	p.C141Y	28.79	
5	F	RAEB-2	46, XX	*ATRX*	c.G2940C	p.K980N	2.6	N/A
				SRSF2	c.C284T	p.P95L	47.85	
				TET2	c.2247_2253del	p.749_751del	29.72	
				TET2	c.5638_5645del	p.1880_1882del	25.46	
6	M	RAEB-2	46, XY	*ATRX*	c.6780_6781insGTAA	p.P2260_P2261delinsP*	74.86	N/A
				ASXL1	c.1926_1927insG	p.G642fs	25.97	
				NRAS	c.G35T	p.G12V	43.99	
				PHF6	c.T941C	p.I314T	92.01	
				U2AF1	c.A470C	p.Q157P	44.63	
7	M	RCMD	46, XY	*ATRX*	c.G6476A	p.R2159H	11.12	N/A
				ASXL1	c.C2197T	p.Q733*	24.90	
				RUNX1	c.1423delG	p.A475fs	18.67	
				RUNX1	c.292delC	p.L98fs	5.25	
				TET2	c.2259_2260del	p.753_754del	16.14	
				TET2	c.3998delT	p.M1333fs	16.21	
				U2AF1	c.A470C	p.Q157P	24.06	
8	M	RCMD-RS	47, XY, +21	*ATRX*	c.4263_4265del	p.1421_1422del	2.02	nuc ish Xq25(RP11-477A22x1,RP1-159G19x1),4q12(SCFD2x2),4q24(TET2x2),5p15(hTERTx2),5q31(CDC25C,EGR1x2),5q33(RPS14x2),5q35(5'NPM1x2)(3'NPM1x2)(5'NPM1 sep 3'NPM1x0),7q31.1(RP11-46B4x2,RP11-54P11x2),7q36.1(RP11-140E16x2,RP11-81I21x2),12p13(5'ETV6x2,3'ETV6x2)(5'ETV6 sep 3'ETV6x0),17p13(TP53x2),17q11(NF1x2),20q11.21(RP1-316I5x2,RP11-358N2x2),21q22(3'AML1x3,5'AML1x3)(3'AML1 sep 5'AML1x0),21q22.2(RP11-476D17x3,RP11-95I21x3)
				IDH1	c.G395A	p.R132H	2.03	
				IDH2	c.G515A	p.R172K	16.59	
				RUNX1	c.C319T	p.R107C	66.56	
				SRSF2	c.284_307del	p.95_103del	23.18	

Supplementary Table 2. Mutation profile of MDS patients with *ATRX* mutation detected (n=8 out of 944) in Haferlach *et al*[116]. Only patient 4 with a complex karyotype does not possess a spliceosomal mutation. Data here are obtained via personal communication with Prof. Torsten Haferlach (21/06/2015).

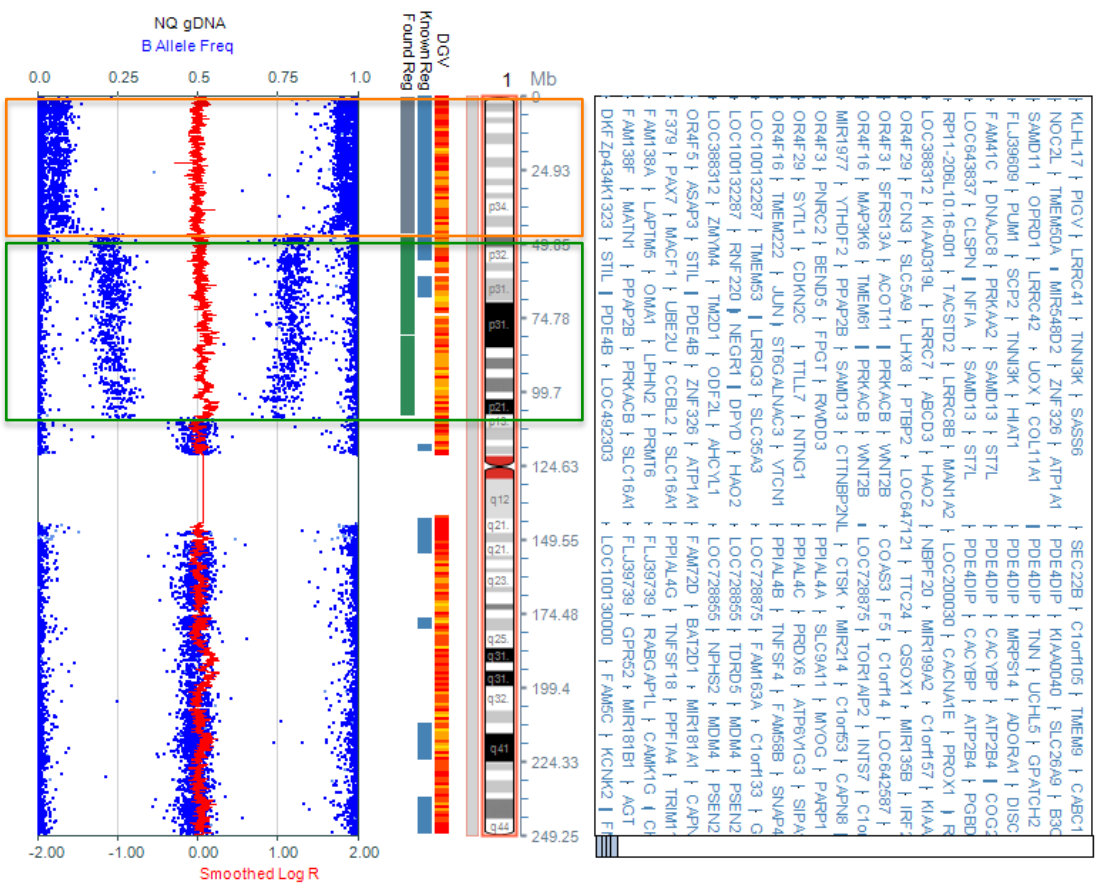
Age	Hb	MCV	MCH	Platelet
4 months	122	78.4	24.0	450
5 months	115	76.6	24.3	506
6 months	131	77.5	24.4	267
2 years	**95 L**	77.7	24.4	308
2 years*	**76 L**	77.0	24.3	212

Supplementary Table 3. Blood counts of a female patient with Bohring-Opitz syndrome. This patient was described in Visayaragawan *et al*{Visayaragawan, 2017, 28889139}. Blood counts were obtained via personal communication to Dr Nithiya Visayaragawan. * blood film review showed moderate anaemia with poor reticulocyte response. Hypochromic, microcytic red cells, few pencil cells, target cells and tear drop cells seen. Iron deficiency anaemia was suggested but no iron studies were available. Patient passed away 4 months after this final blood count secondary to sepsis. Note: the diagnosis of Bohring-Opitz syndrome was made clinically and ASXL1 mutation was not confirmed in this patient.

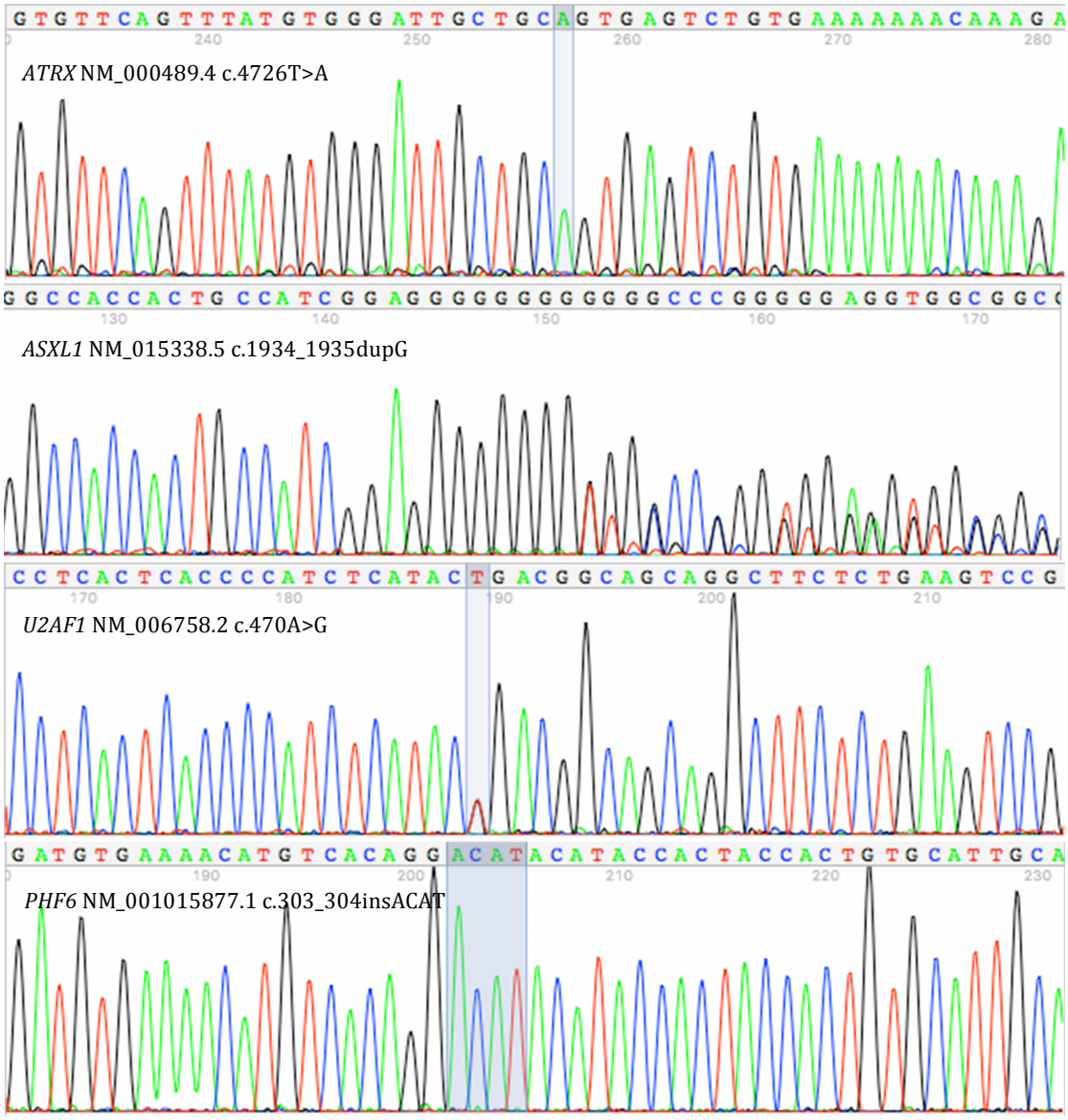
Patient	Sex	Age (years)	Mutation	Hb	MCV	MCH	Platelet	Ferritin
1	M	5 months	p.Val268Ile	130	76.1	25.6	337	N/A
		3		134	80	26.4	284	90
2	F	14	p.Leu282Pro	126	91	31.0	**432 H**	N/A
3	F	2	Deletion exon 6-7	**90 L**	75	24.9	352	N/A
4	M	1	p.Met1Val#	122	**68 L**	N/A	347	31
		7		127	74	N/A	275	N/A
5	M	1		118	**66 L**	N/A	366	**29 L**
		7		121	77	N/A	216	N/A
6	M	11		126	77	N/A	367	**28 L**
		16		148	**78 L**	N/A	251	34
7	M	1		131	76	N/A	401	66
		5		132	79	N/A	380	N/A
8	M	31	Clinically diagnosed, mutation not confirmed	153	83	28.5	276	N/A

Supplementary Table 4. List of patients with Börjeson-Forssman-Lehmann syndrome with blood counts. #Patients 4-7 are from the same family and have been described in Ernst et al{Ernst, 2015, 26648834}. Patients 4 and 5 are monozygotic twins. Patients 6 and 7 are siblings. Patient 6 was treated for iron-deficiency anaemia. Blood counts here were obtained via personal communications to Dr Patrick Yap (patients 1-3), Dr Tine Høg Sørensen (patients 4-7) and Dr Felicity Collins. Abnormal values are in bold (L: low, H: high)

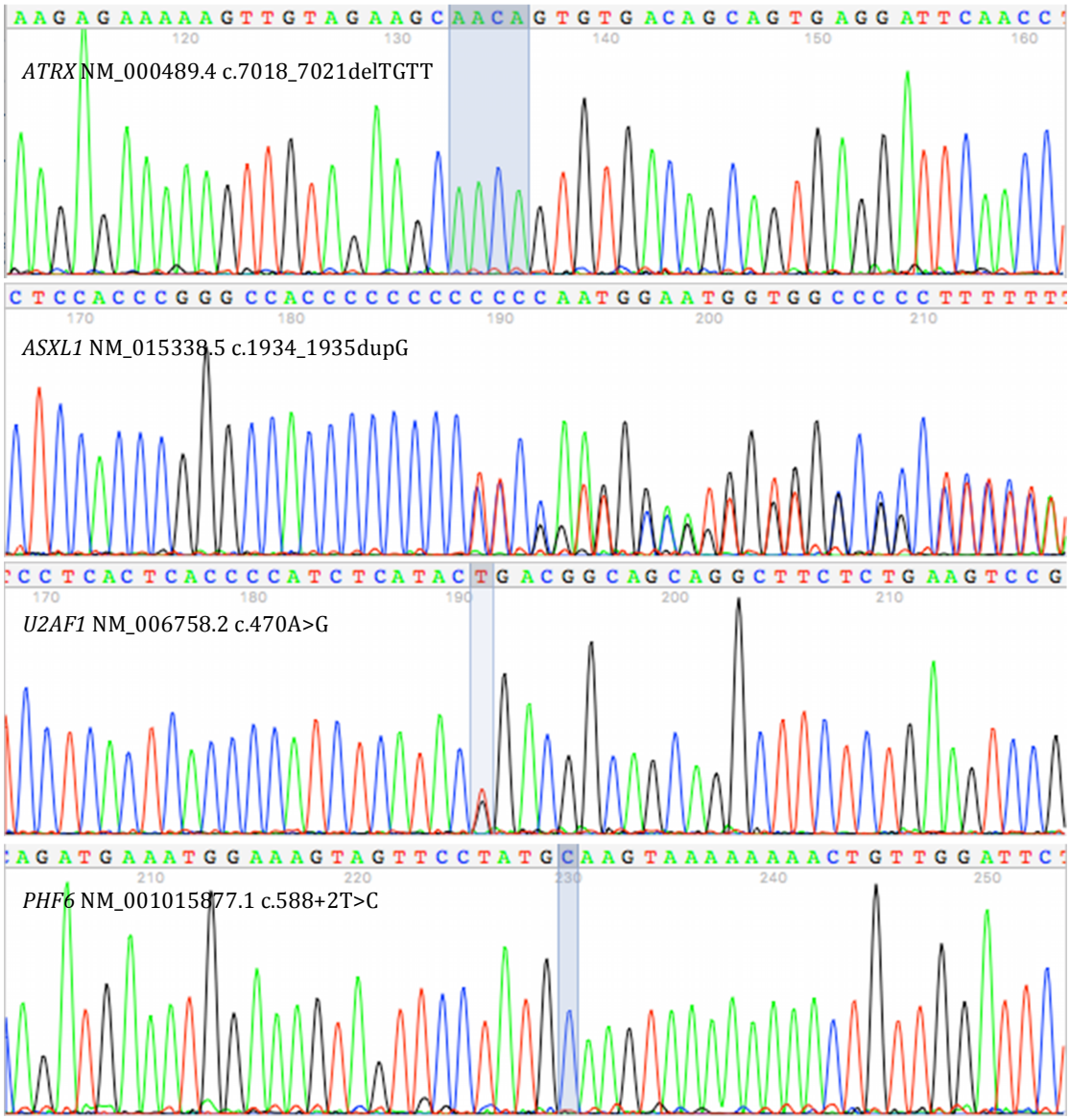
7.2 Characterisation of ATMDS patients



Supplementary Figure 1. SNP array analysis showing copy number-neutral loss of heterozygosity (LOH) in chromosome 1p. Two regions of LOH with different B allele frequency exist, suggestive of a subclonal structure with a parental clone with high percentage mosaicism in 1p33-36.33 (orange box), and a subclone with centromeric extension of LOH, shown as a lower percentage mosaicism (subclone) between 1p13.3-33 (green box).

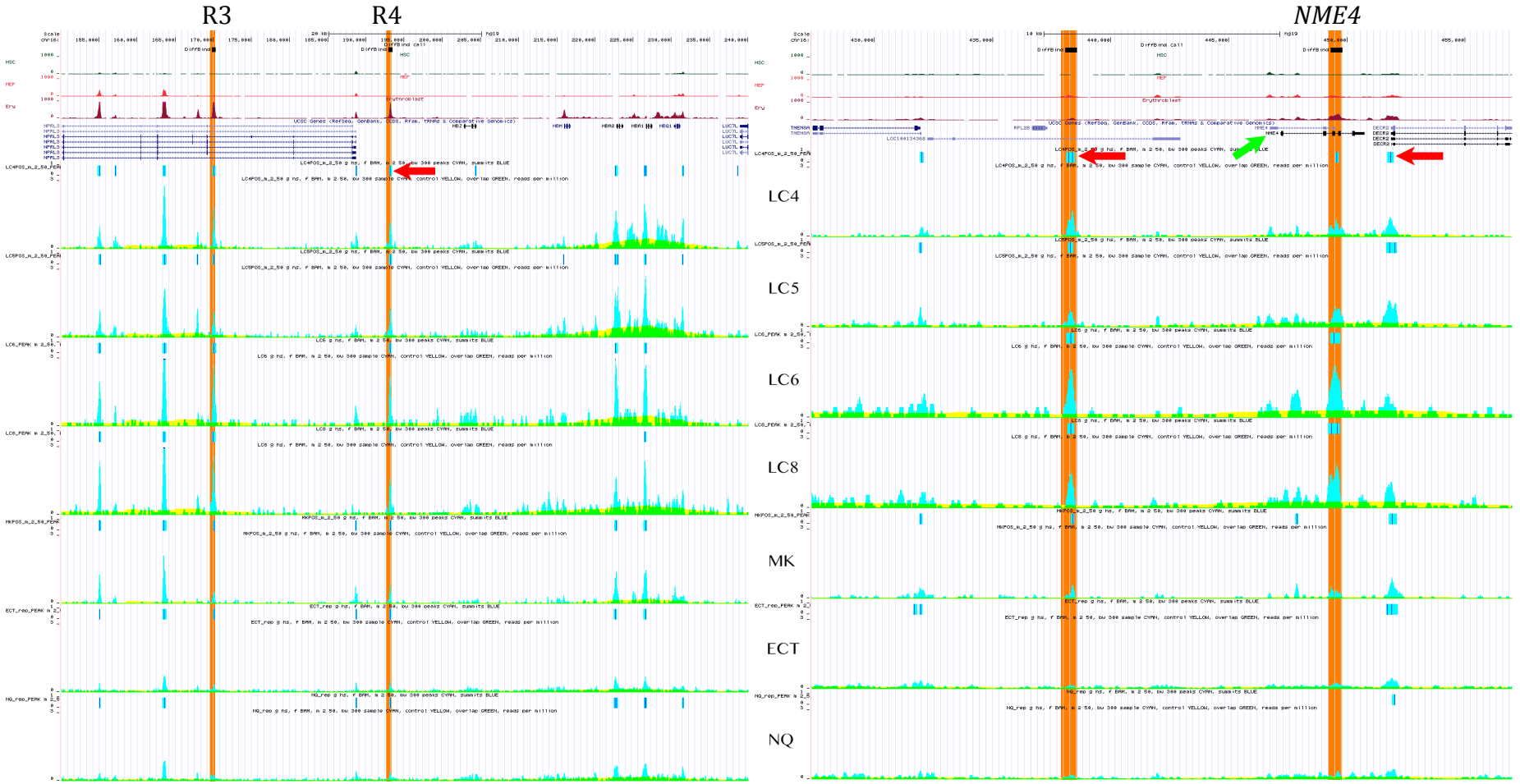


Supplementary Figure 2. Sanger sequencing traces of *ATRX*, *ASXL1*, *U2AF1* and *PHF6* (from top to bottom) from single cell culture in patient NQ (case #118). Locations of mutation are highlighted in light blue. Sequence trace for *U2AF1* is in reverse complement to coding strand.

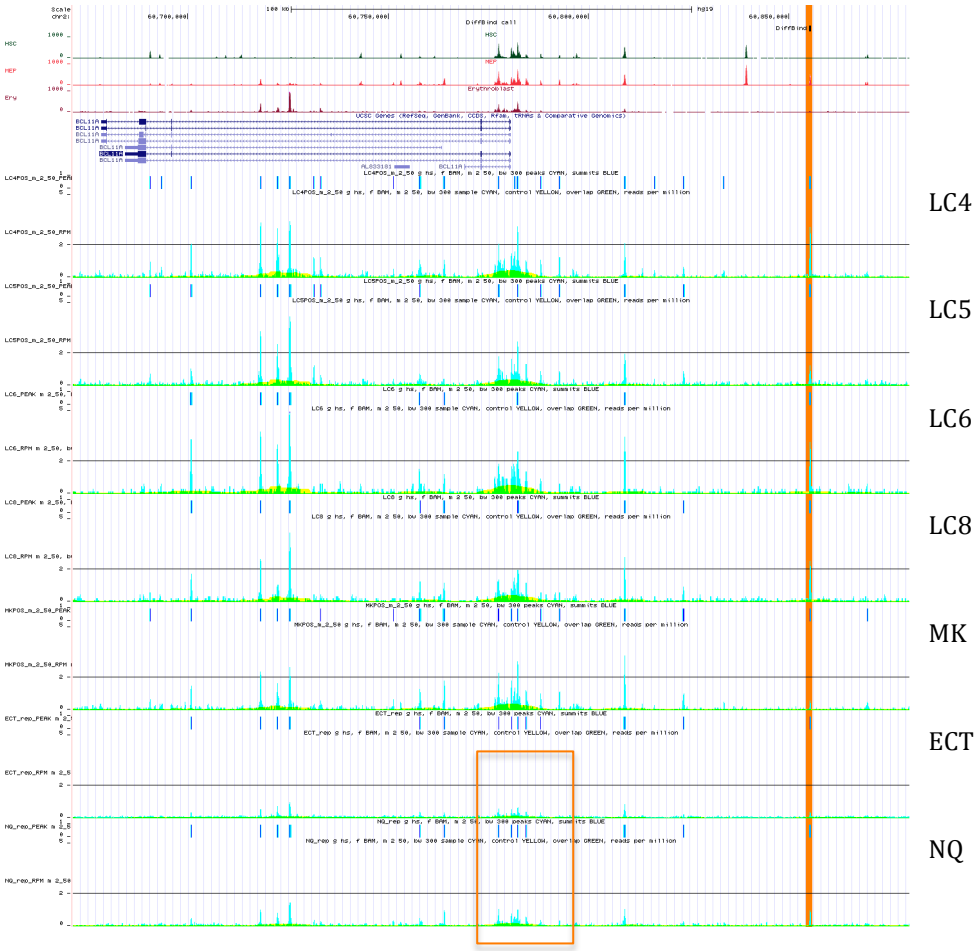


Supplementary Figure 3. Sanger sequencing traces of *ATRX*, *ASXL1*, *U2AF1* and *PHF6* (from top to bottom) from single cell culture in patient ECT (case #122). Locations of mutation are highlighted in light blue. Sequence traces for *ATRX*, *ASXL1* and *U2AF1* are in reverse complement to coding strand.

7.3 Erythroid differentiation of ATMDS patients



Supplementary Figure 4. ATAC-seq tracks with peak calling in chromosome 16p13.3 at the alpha globin locus (left) and *NME4* (right). From top to bottom the ATAC-seq tracks for four normal leucocyte cone control (LC4, LC5, LC6 and LC8) and three ATMDS patients (MK, ECT and NQ) are shown in cyan, with the background in yellow. Above each track peaks called by MACS2 are highlighted with the summits in blue (examples pointed out by red arrows). Differential sites are highlighted in orange, where peaks are clearly visible in LCs but reduced/absent in ATMDS patients. Green arrow: *NME4*.



Supplementary Figure 5. ATAC-seq tracks with peak calling at *BCL11A*. The differential site containing KLF1 motifs upstream of *BCL11A* (chr2:60855094-60855594) is highlighted in orange. Note that in the reference ATAC-seq tracks by Corces *et al*[257] this peak is only present in the MEP population during differentiation. ATAC-seq peaks in BCL11A gene body are generally lower in ATMDS patients, particularly ECT and NQ (orange box).

7.4 HUDEP-2 genome editing

7.4.1 sgRNAs and ssODN used for introducing *ATRX*, *PHF6* and *ASXL1* mutations

sgRNA used for ATRX knockout:

ATRX-sgRNA-Ex16-TOP	caccGTCCAATAACAACCAAGT
ATRX-sgRNA-Ex16-BOTTOM	aaacACTTGGTTGTTATTGGAC

sgRNA used for PHF6 knockout:

PHF6-sgRNA-Ex2-TOP	caccGTACAAGACAGCGCAAATG
PHF6-sgRNA-Ex2-BOTTOM	aaacCATTTGCGCTGTCTTGTAC

sgRNA used for ASXL1 heterozygous truncation:

ASXL1-sgRNA-Ex14-TOP	caccGTAAATGGGGAGCATACCC
ASXL1-sgRNA-Ex14-BOTTOM	aaacGGGTATGCTCCCCATTTAC

ssODN used for protecting *ASXL1* allele:

ASXL1-ssODN-WT
ACAGCGAACACAACTACTGCCGCCTTATCCTCTAAATGGaGAGCATACaCAaGCCGGA
ACTGCCATGTCCAGAGCTAGGAGAGAGGACCTGCCTTCT

ssODN designed to introduce deliberate synonymous mutations in one allele to protect it from further editing. sgRNA sequence is highlighted in green, while the PAM sequence is highlighted in yellow. Deliberately introduced synonymous mutations are in small letters in red.

7.4.2 Introduction of U2AF1 heterozygous p.Gln157Pro mutation in the triple mutant APA G7

Step 1: to introduce synonymous mutation in one allele to prevent editing

sgRNA used to target region near codon 157 in *U2AF1*.

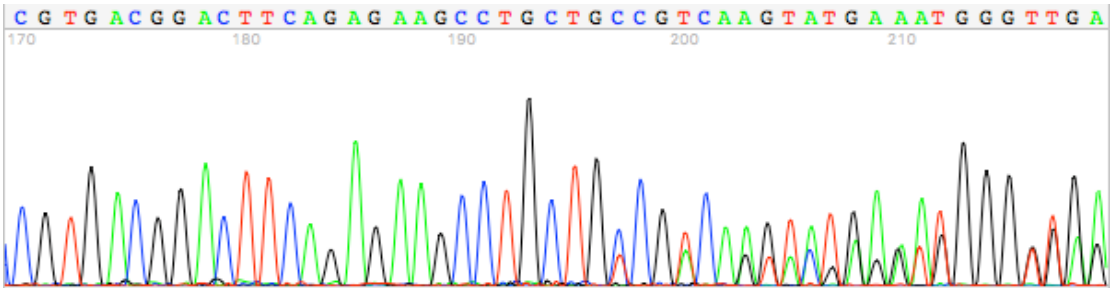
U2AF1-sgRNA-TOP	caccGCACCCCATCTCATACTGA
U2AF1-sgRNA-BOTTOM	aaacTCAGTATGAGATGGGGTGG

ssODN used for protecting one *U2AF1* allele in the first step:

U2AF1-ssODN-WT
CACGTCACTGGCCACTCCTCACTCACCCCATCTCgTACTGtCGaCAGCAGGCTTC
TCTGAAGTCCGTCACGGGTGACAGCTCGGCGTGGATCGGC

ssODN designed to protect the U2AF1 allele. sgRNA sequence is highlighted in green, while the PAM sequence in highlighted in

yellow. The codon for p.Gln157 is boxed. Deliberately introduced synonymous mutations are in small letters in red.



Supplementary Figure 6. Sanger sequencing trace of U2AF1 in the intermediate clone APAU B10, showing the 1bp insertion.

Step 2: to repair the 1bp insertion introduced in the intermediate clone APAU B10 and knock in p.Gln157Pro mutation.

sgRNA used to target specifically the 1bp insertion in APAU B10:

U2AF1-B10-TOP	caccGACCCCATCTCATACtTGA
U2AF1-B10-BOTTOM	aaacTCAaGTATGAGATGGGGTC

The 1bp insertion "t" is in small letter.

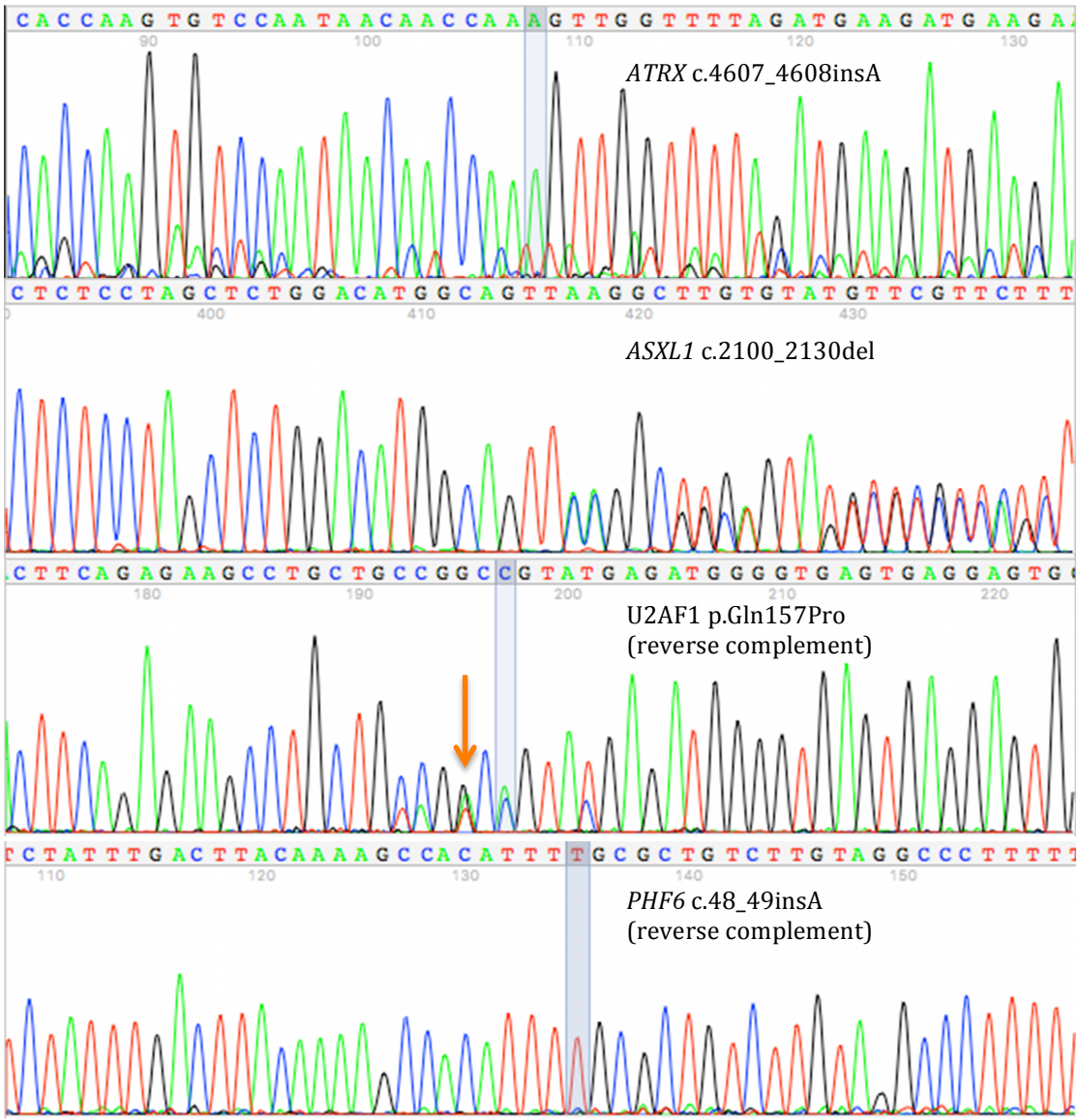
A list of ssODNs designed to repair the single base insertion in HUDEP clone APAU B10 and at the same time introduce the p.Gln157Pro mutation is shown in the table below. The codon for Gln157 is boxed. The mutation to introduce p.Gln157Pro change is in small letter in red. Other deliberate synonymous mutations are shown in small letters. Only U2_Q157P_ssODN5 was successful in introduction of the p.Gln157Pro mutation.

ssODN	**Sequence (5' to 3')**	**Note on ssODN design**
U2_Q157P_ssODN1	CGCCGAGCTGTCACCCGTGACGGACTTCAGAGAAGCCTGCTGtCGTCctTATGAGATGGGGTGAGTGAGGAGTGGCCAGTGACGTGACTGAGCAC	ssODN in reverse complement to "top" sgRNA oligo
U2_Q157P_ssODN2	GACAGCCGATCCACGCCGAGCTGTCACCCGTGACGGACTTCAGAGAAGCCTGCTGtCGTCctTATGAGATGGGGTGAGTGAGGAGTGGCCAGTGACG	different offset to U2_Q157P_ssODN1
U2_Q157P_ssODN3	CGTCACTGGCCACTCCTCACTCACCCCATCTCATAagGACGaCAGCAGGCTTCTCTGAAGTCCGTCACGGGTGACAGCTCGGCGTGGATCGGCTGTC	Same offset as U2_Q157P_ssODN2 but on the same strand as the "top" sgRNA oligo
U2_Q157P_ssODN4	CACGTCACTGGCCACTCCTCACTCACCCCATCTCATACgGACGaCAGCAGGCTTCTCTGAAGTCCGTCACGGGTGACAGCTCGGCGTGGATCGGC	Different offset to ssODN3 and one less deliberate mismatch
U2_Q157P_ssODN5	CACGTCACTGGCCACTCCTCACTCACCCCATCTCATACgGcCtGCAGCAGGCTTCTCTGAAGTCCGTCACGGGTGACAGCTCGGCGTGGATCGGC	Different deliberate mismatches to U2_Q157P_ssODN4

PCR primers used for screening of genome-edited clones by Sanger sequencing:

Gene	Primer name	Sequence (5' to 3')	Tm (°C)	PCR product size (bp)
ATRX	ATRX-CRISPR-Ex16F2	GATTTGTGGTGTGGGAGTAAATGTG	60.56	728
	ATRX-CRISPR-Ex16R2	AGCAAGAAAGGGGACTAGCAAC	60.55	
ASXL1	ASXL1-CRISPR-F	GGGTGGTTAAAGGTCAGCCC	60.6	836
	ASXL1-CRISPR-R	TGCTCTGTTCTGCAGGCAAT	60.3	
U2AF1	U2AF1-CRISPR-F	GCCACTGAGCTTAACAACGG	59.5	466
	U2AF1-CRISPR-R	CACCTTACTATACCTGAGTGTGT	57.1	
PHF6	PHF6-CRISPR-Ex2F2	ACATTAATTATGCCAGCAGGGGA	60.12	991*
	PHF6-CRISPR-Ex2R2	GACATCTTCAATAGAAAATCCACCA	57.37	

*For Sanger sequencing of PHF6 PCR product, an internal sequencing primer was used (PHF6-CRISPR-Ex2R4 5'-ATTCAGGAACATTCATGTG-3') to avoid homopolymer region.



Supplementary Figure 7. Sanger sequencing traces for mutations in the HUDEP B10 E5 clone. From top to bottom: ATRX, ASXL1, U2AF1 and PHF6. The signal in ATRX trace was too high and therefore background noise was visible. Mutations are highlighted in light blue. Note the three signal peaks in U2AF1 (orange arrow) suggesting .the presence of three alleles.

7.4.3 Repairing *ATRX* mutation in quadruple mutant clone B10 E5

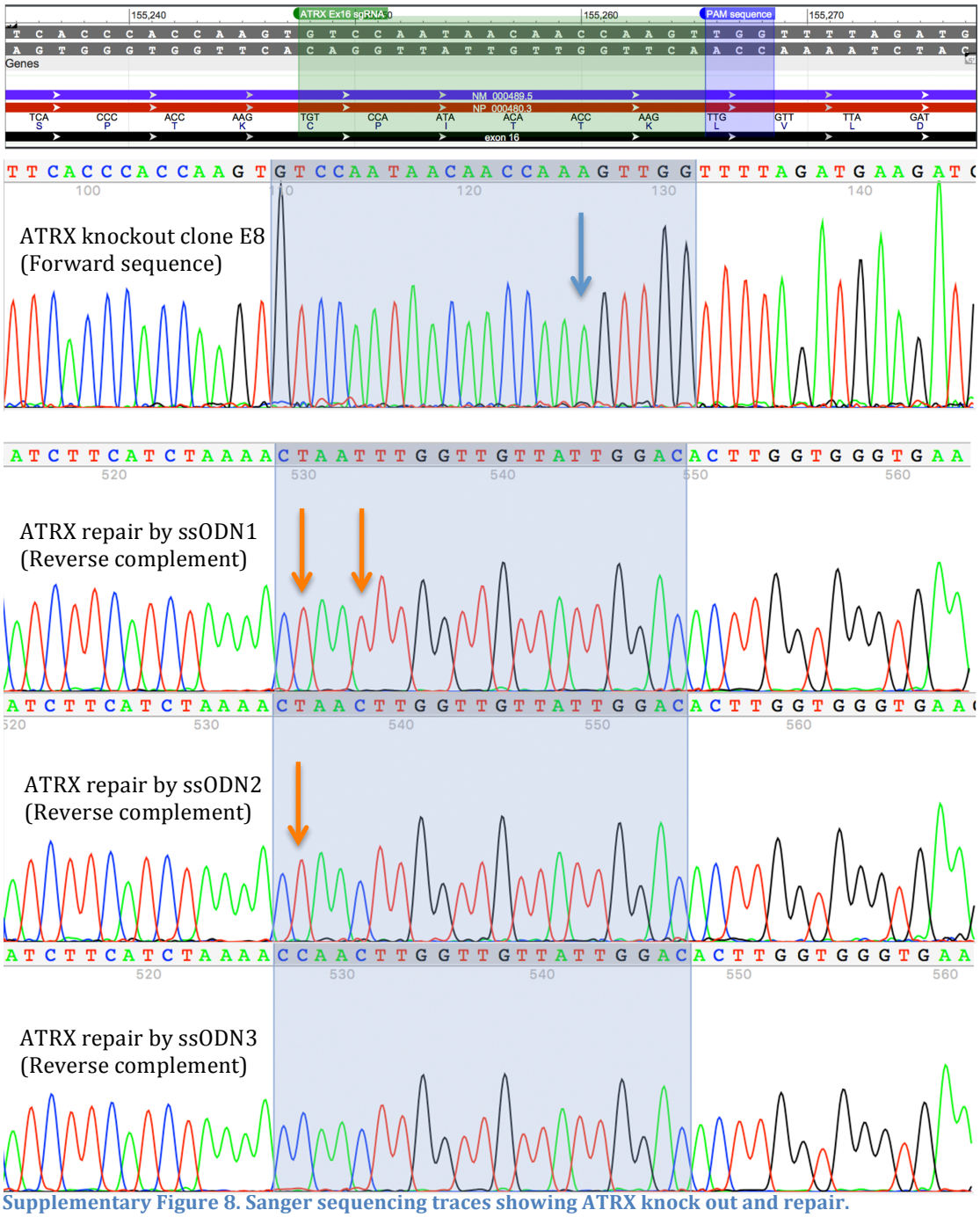
sgRNA used for ATRX knockout clone E8 repair

ATRX-E8-TOP	caccGTCCAATAACAACCAAAGT
ATRX-E8-BOTTOM	aaacACTTTGGTTGTTATTGGAC

The single base insertion in ATRX E8 clone is highlighted red.

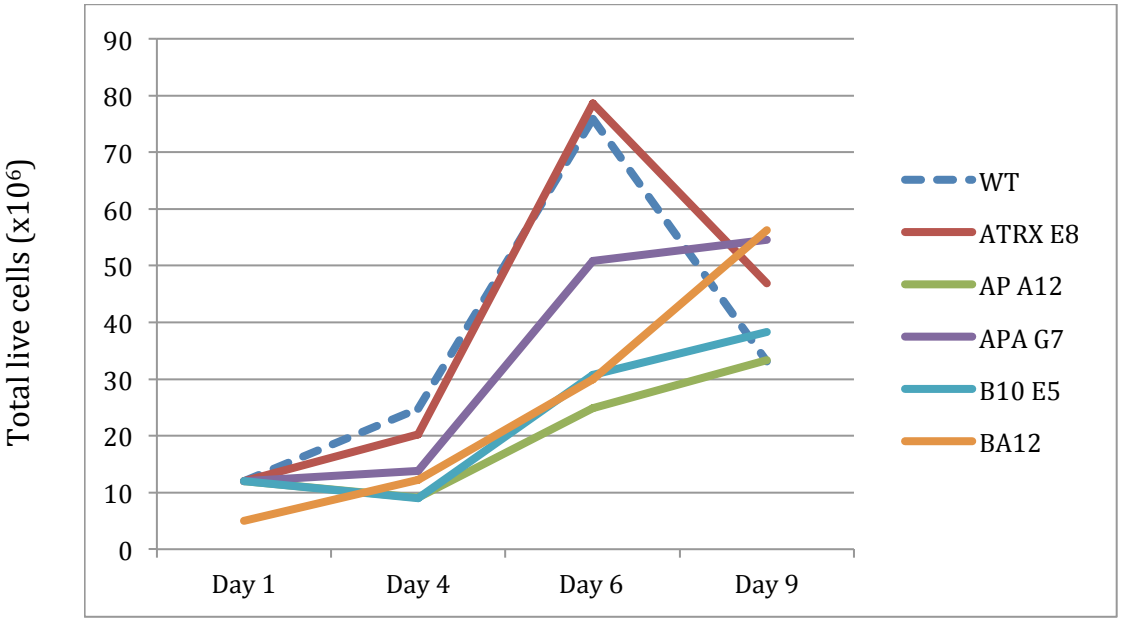
ssODNs for HDR repair of ATRX knockout clone E8:
Three pairs of ssODNs were used in the repair. Each ssODN vary slightly in offset and number of deliberate synonymous mutations (ssODN1 has two, shown in small letters in red, whereas ssODN2 has one only and ssODN3 has none). The sgRNA sequence of ATRX-E8-TOP is highlighted in green – note in the ssODN the sequence does not contain the single base insertion. PAM sequence is highlighted in yellow.

ssODN	Sequence (5' to 3')
ATRX-repairWT-ssODN1	CTTGTAGGTGATAGAAATTGAAGATGCTTCACCCACCAAGTGTCCAATAACAACCAAaTTaGTTTTAGATGAAGATGAAGAAACCAAAGAACCTTTAGTGCAGGTTCATAG
ATRX-repairWT-ssODN2	GAAATTGAAGATGCTTCACCCACCAAGTGTCCAATAACAACCAAGTTaGTTTTAGATGAAGATGAAGAAACCAAAGAACCTTTAGTGCAGGTTCATAGAAATATGGTTAT
ATRX-repairWT-ssODN3	CTTTGAGCTTAACTATCAACAACTCCTTTTCTTGTAGGTGATAGAAATTGAAGATGCTTCACCCACCAAGTGTCCAATAACAACCAAGTTGGTTTTAGATG



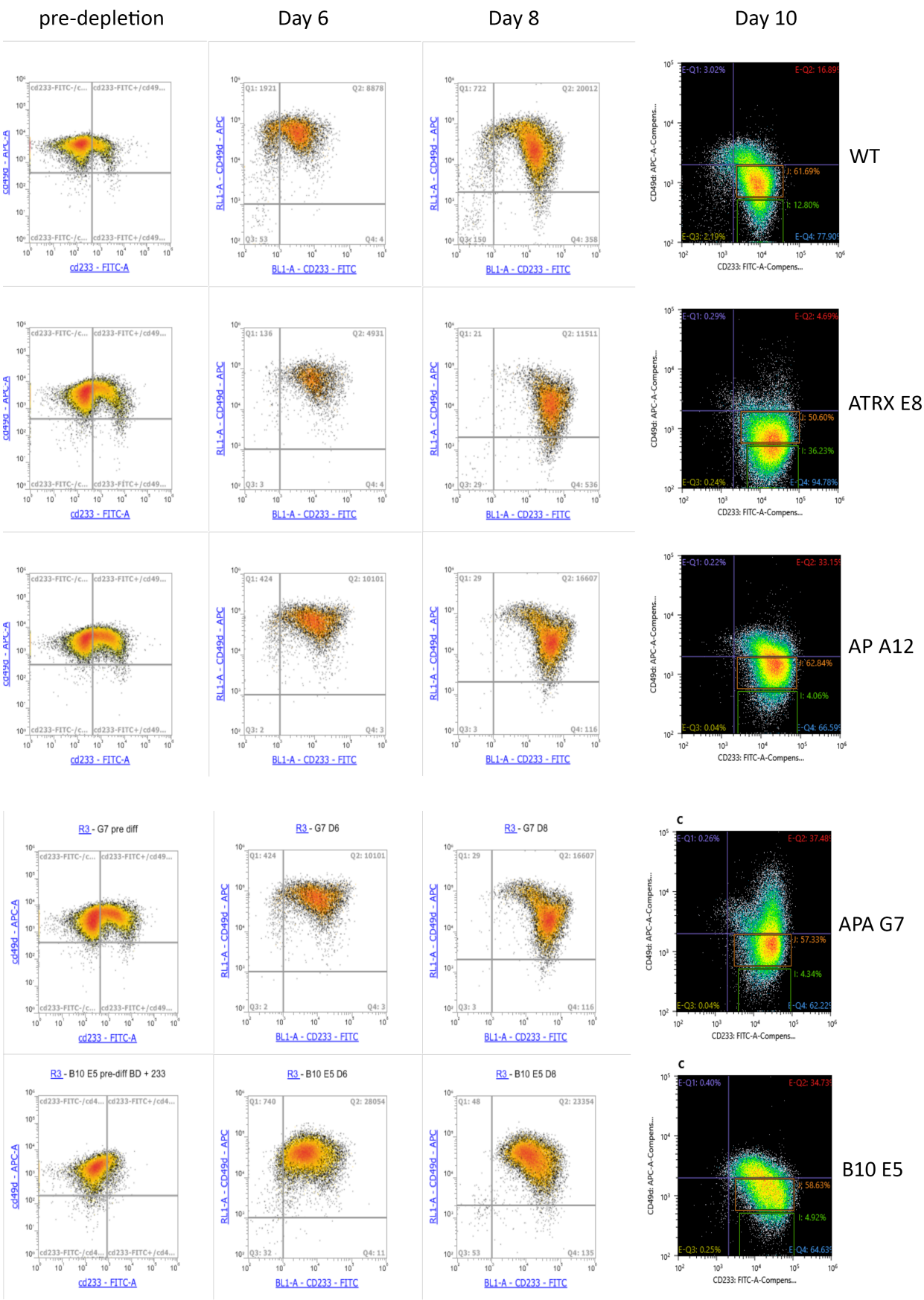
Supplementary Figure 8. Sanger sequencing traces showing ATRX knock out and repair. Reference sequence is shown at the top with sgRNA and PAM sequence highlighted. The 1bp insertion in ATRX knockout clone E8 is shown below (blue arrow). The next three traces (in reverse strand) show repair of the 1bp insertion by three ssODNs (ATRX-repairWT-ssODN 1 to 3). Deliberate synonymous mutations are indicated by orange arrows. ssODN3 repaired the sequence scarlessly to wildtype.

7.4.4 Characteristics of genome edited HUDEP-2 clones during erythroid differentiation

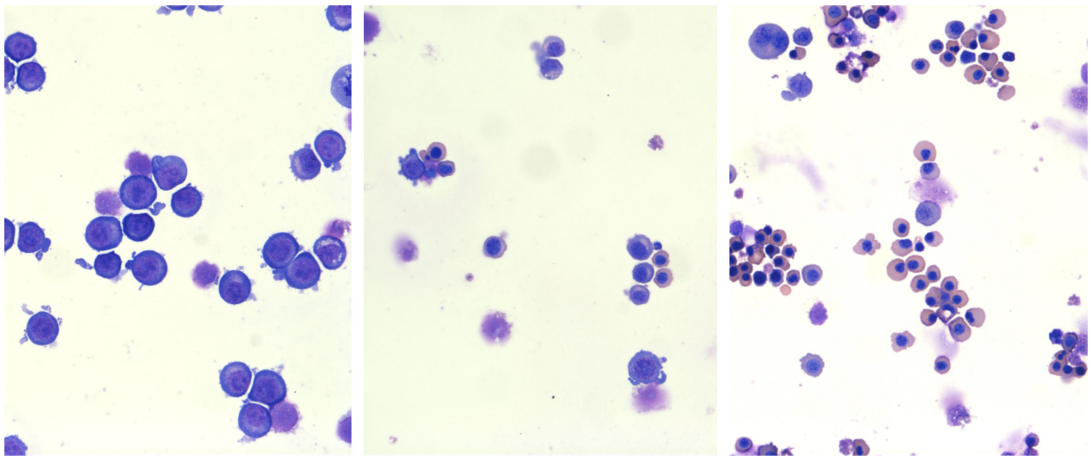


Viability (%)	Day 1	Day 4	Day 6	Day 9
WT	92.7	91.3	88.4	90.5
ATRX E8	93.6	92.7	96.2	93.3
AP A12	91.4	84.5	93.8	92.8
APA G7	88.4	94.2	94.9	95.6
B10 E5	94.8	90.1	86.5	95.3
BA12	94.5	94.9	98.4	97.1

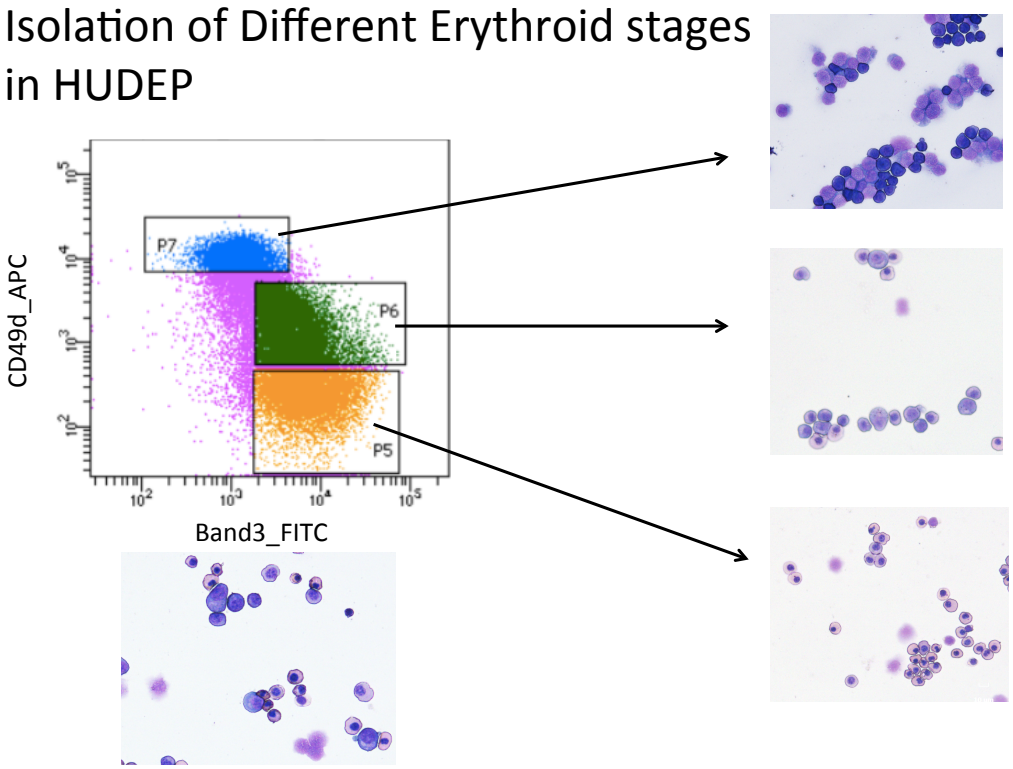
Supplementary Figure 9. Top: Growth curve of HUDEP clones during erythroid differentiation. The starting cell number is $12x10^6$ for all but BA12 clone, where the starting cell number was lowered to $5x10^6$ due to its rapid expansion. Bottom: cell viability at each time point for HUDEP clones during erythroid differentiation. Viability measurement was automated using a NucleoCounter NC-3000 (chemometec).



Supplementary Figure 10. Serial FACS analysis of HUDEP cells tracking erythroid differentiation. From left to right: analyses performed before CD233 depletion, on Day 6, Day 8 and Day 10 (at the time of FACS sorting). Five different HUDEP clones were shown here. As the cells differentiate they become CD233+ (X-axis) and CD49d- (Y-axis).



Supplementary Figure 11. Modified Wright stain of HUDEP wildtype during erythroid differentiation on Day 1, 6 and 9 (from left to right), demonstrating change in morphology from basophilic erythroblast to polychromatic erythroblast to orthochromatic erythroblast.



Supplementary Figure 12. FACS analysis of differentiated wildtype HUDEP cells. Modified Wright stains showing heterogeneity in differentiation in the bulk population (bottom). FACS-sorted populations (P7, P6 and P5) shows progressive differentiation as cells become CD49d negative and Band 3 (CD233) positive. Figure produced by Dr. Andrew King.

7.4.5 Introducing U2AF1 p.Gln157Pro mutation in wildtype HUDEP-2

Refer to sgRNA sequences above (U2AF1-sgRNA-TOP and –BOTTOM) for introduction of 5bp deletion in the intermediate UA1 clone.

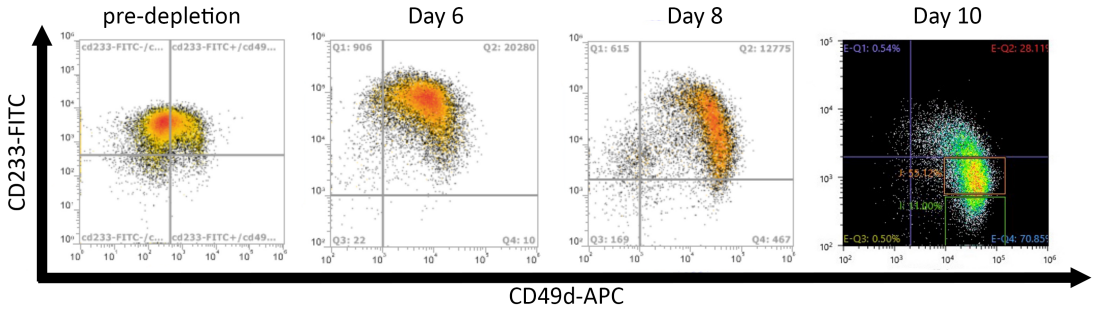
sgRNA used to repair 5bp deletion in UA1 intermediate clone

U2AF1-5bpdel-TOP	caccGCCTGCTGCCGTCAGTAT
U2AF1-5bpdel-BOTTOM	aaacATACTGACGGCAGCAGGC

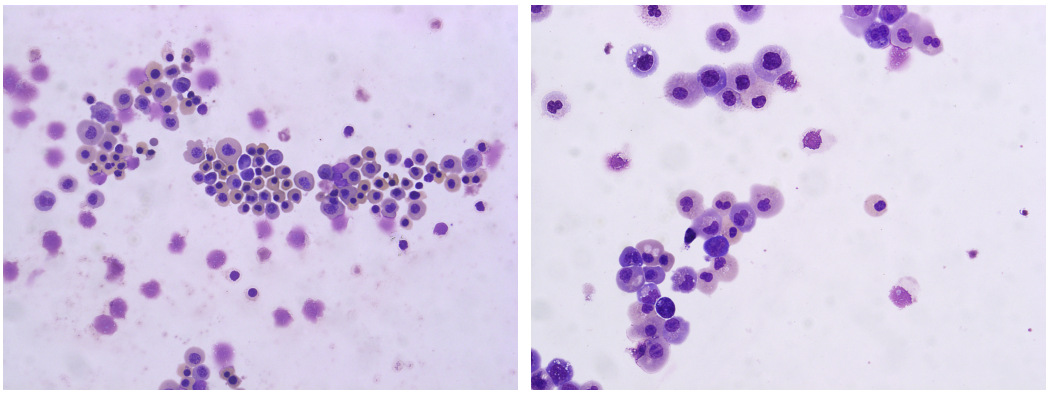
Refer to U2_Q157P_ssODN5 above for the ssODN used to repair and introduce the p.Gln157Pro mutation in the allele with 5bp deletion.



Supplementary Figure 13. Top: U2AF1 reference sequence showing the location of 5bp deletion (highlighted in green). Middle: Sanger sequencing trace of the intermediate UA1 clone containing a 5bp deletion 3' to codon Gln157. The base required to be mutated for p.Gln157Pro mutation (CAG → CCG) is highlighted in blue. Bottom: Sanger sequencing trace of the resultant UA1 A2 clone containing the p.Gln157Pro mutation. Two deliberate mutations in codon 156 leading to a synonymous change (CGT→AGG) were introduced by a ssODN.



Supplementary Figure 14. Serial FACS analysis of HUDEP UA1 A2 clone tracking erythroid differentiation. From left to right: analyses before differentiation, Day 6, Day 8 and Day 10 (at the time of FACS sorting).



Supplementary Figure 15. Modified Wright stain of differentiated cells on day 10 in UA1 A2 (right). The morphology of differentiated wildtype HUDEP from the same experiment is shown on the left for comparison.

7.5 Predicted G4 sequence in alpha and beta globin genes

Shown below are sequences of *HBA2* and *HBB*. In *HBA2*, intron 2 is short and GC-rich, with four potential sets of G4 sequences. This contrasts with *HBB* where its intron 2 is long and AT-rich.

Exons are highlighted in teal. Predicted G4 sequences are in red, with the G duplets or triplets in bold.

G4 sequences prediction is obtained from G-Rich Sequences DataBase v2 (GRSDB2).

HBA2

ACTCTTCTGG TCCCCACAGA CTCAGAGAGA ACCCACCATG
GTGCTGTCTC CTGCCGACAA GACCAACGTC AAGGCCGCCT
GGGGTAA**GG**T C**GG**CGCGCAC GCTGGCGAGT ATGGTGCGGA
GGCCCT**GG**AG A**GG**TGA**GG**CT CCCTCCCCTG CTCCGACCCG
GGCTCCTCGC CCGCCCGGAC CCACAGGCCA CCCTCAACCG
TCCTGGCCCC GGACCCAAAC CCCACCCCTC ACTCTGCTTC
TCCCCGCAGG ATGTTCCTGT CCTTCCCCAC CACCAAGACC
TACTTCCCGC ACTTCGACCT GAGCCACGGC TCTGCCCA**GG**
TTAA**GG**GCCA C**GG**CAAGAA**G** **G**TGGCCGACG CGCTGACCAA
CGCCGTGGCG CACGTGGACG ACATGCCCAA CGCGCTGTCC
GCCCTGAGCG ACCTGCACGC GCACAAGCTT CG**GG**TGGACC
C**GG**TCAACTT CAA**GG**TGAGC **GG**C**GGG**CC**GG** **G**AGCGATCT**G**
GGTCGAG**GGG** CGAGATGGCG CCTTCCTCTC A**GG**GCAGA**GG**
ATCACGC**GG**G TTGC**GG**GAGG TGTAGCGCA**G** **G**CGGC**GG**CTG
C**GG**GCCT**GG**G CCGCACTGAC CCTCTTCTCT GCACAGCTCC
TAAGCCACTG CCTGCTGGTG ACCCTGGCCG CCCACCTCCC
CGCCGAGTTC ACCCCTGCGG TGCACGCCTC CCTGGACAAG
TTCCTGGCTT CTGTGAGCAC CGTGCTGACC TCCAAATACC
GTTAAGCTGG AGCCTCGGTA GCCGTTCCTC CTGCCCGCTG
GGCCTCCCAA CGGGCCCTCC TCCCCTCCTT GCACCGGCCC
TTCCTGGTCT TTGAATAAAG TCTGAGTGGG CAGC

HBB

ACATTTGCTT CTGACACAAC TGTGTTCACT AGCAACCTCA
AACAGACACC ATGGTGCATC TGACTCCTGA GGAGAAGTCT
GCCGTTACTG CCCTGTGGGG CAAGGTGAAC GTGGATGAAG
TTGGTGGTGA GGCCCTGGGC AGGTTGGTAT CAAGGTTACA
AGACAGGTTT AAGGAGACCA ATAGAAACTG GGCATGTGGA
GACAGAGAAG ACTCTTGGGT TTCTGATAGG CACTGACTCT
CTCTGCCTAT TGGTCTATTT TCCCACCCTT AGGCTGCTGG
TGGTCTACCC TTGGACCCAG AGGTTCTTTG AGTCCTTTGG
GGATCTGTCC ACTCCTGATG CTGTTATGGG CAACCCTAAG
GTGAAGGCTC ATGGCAAGAA AGTGCTCGGT GCCTTTAGTG
ATGGCCTGGC TCACCTGGAC AACCTCAAGG GCACCTTTGC
CACACTGAGT GAGCTGCACT GTGACAAGCT GCACGTGGAT
CCTGAGAACT TCAGGGTGAG TCTATGGGAC GCTTGATGTT
TTCTTTCCCC TTCTTTTCTA TGGTTAAGTT CATGTCATAG
GAAGGGGATA AGTAACAGGG TACAGTTTAG AATGGGAAAC
AGACGAATGA TTGCATCAGT GTGGAAGTCT CAGGATCGTT
TTAGTTTCTT TTATTTGCTG TTCATAACAA TTGTTTTCTT
TTGTTTAATT CTTGCTTTCT TTTTTTTTCT TCTCCGCAAT
TTTTACTATT ATACTTAATG CCTTAACATT GTGTATAACA
AAAGGAAATA TCTCTGAGAT ACATTAAGTA ACTTAAAAAA
AAACTTTACA CAGTCTGCCT AGTACATTAC TATTTGGAAT
ATATGTGTGC TTATTTGCAT ATTCATAATC TCCCTACTTT
ATTTTCTTTT ATTTTTAATT GATACATAAT CATTATACAT
ATTTATGGGT TAAAGTGTAA TGTTTTAATA TGTGTACACA
TATTGACCAA ATCAGGGTAA TTTTGCATTT GTAATTTTAA
AAAATGCTTT CTTCTTTTAA TATACTTTTT TGTTTATCTT
ATTTCTAATA CTTTCCCTAA TCTCTTTCTT TCAGGGCAAT
AATGATACAA TGTATCATGC CTCTTTGCAC CATTCTAAAG
AATAACAGTG ATAATTTCTG GGTTAAGGCA ATAGCAATAT
CTCTGCATAT AAATATTTCT GCATATAAAT TGTAACTGAT
GTAAGAGGTT TCATATTGCT AATAGCAGCT ACAATCCAGC
TACCATTCTG CTTTTATTTT ATGGTTGGGA TAAGGCTGGA
TTATTCTGAG TCCAAGCTAG GCCCTTTTGC TAATCATGTT
CATACCTCTT ATCTTCCTCC CACAGCTCCT GGGCAACGTG
CTGGTCTGTG TGCTGGCCCA TCACTTTGGC AAAGAATTCA
CCCCACCAGT GCAGGCTGCC TATCAGAAAG TGGTGGCTGG
TGTGGCTAAT GCCCTGGCCC ACAAGTATCA CTAAGCTCGC
TTTCTTGCTG TCCAATTTCT ATTAAAGGTT CCTTTGTTCC
CTAAGTCCAA CTACTAAACT GGGGGATATT ATGAAGGGCC
TTGAGCATCT GGATTCTGCC TAATAAAAAA CATTTATTTT
CATTGC

Reference:

Kikin, O., et al., *GRSDB2 and GRS_UTRdb: Databases of Quadruplex Forming G-rich Sequences in pre-mRNAs and mRNAs*. Nucleic Acids Res, 2008. **36**(suppl_1): p.D141-D148.

Publication resulted from this work

Moir-Meyer, G., et al. *Robust CRISPR/Cas9 Genome Editing of the HUDEP-2 Erythroid Precursor Line Using Plasmids and Single-Stranded Oligonucleotide Donors*. Methods Protoc, 2018. **1**(3): 28.

Invited Presentation

"Achieving HDR in HUDEP cells", 6th Genome Engineering in Oxford Meeting. November, 2016. Weatherall Institute of Molecular Medicine, Oxford, United Kingdom.

References

1. Modell, B., *Global epidemiology of haemoglobin disorders and derived service indicators.* Bulletin of the World Health Organization, 2008. **2008**(6): p. 480-487.
2. Paoli, M., et al., *Crystal structure of T state haemoglobin with oxygen bound at all four haems.* J Mol Biol, 1996. **256**(4): p. 775-92.
3. Thom, C.S., et al., *Hemoglobin variants: biochemical properties and clinical correlates.* Cold Spring Harb Perspect Med, 2013. **3**(3): p. a011858.
4. Gibbons, R.J., *alpha-Thalassemia, mental retardation, and myelodysplastic syndrome.* Cold Spring Harb Perspect Med, 2012. **2**(10).
5. Higgs, D.R., *The molecular basis of alpha-thalassemia.* Cold Spring Harb Perspect Med, 2013. **3**(1): p. a011718.
6. Harteveld, C.L. and D.R. Higgs, *Alpha-thalassaemia.* Orphanet J Rare Dis, 2010. **5**: p. 13.
7. Goh, S.H., et al., *A newly discovered human alpha-globin gene.* Blood, 2005. **106**(4): p. 1466-72.
8. Old, J., et al., in *Prevention of Thalassaemias and Other Haemoglobin Disorders: Volume 2: Laboratory Protocols*, nd, Editor. 2012: Nicosia, Cyprus.
9. Galanello, R., et al., *Phenotype-genotype correlation in haemoglobin H disease in childhood.* J Med Genet, 1983. **20**(6): p. 425-9.
10. Laosombat, V., et al., *Clinical features and molecular analysis in Thai patients with HbH disease.* Ann Hematol, 2009. **88**(12): p. 1185-92.
11. Thein, S.L. and J.E. Craig, *Genetics of Hb F/F cell variance in adults and heterocellular hereditary persistence of fetal hemoglobin.* Hemoglobin, 1998. **22**(5-6): p. 401-14.
12. Garner, C., et al., *Genetic influences on F cells and other hematologic variables: a twin heritability study.* Blood, 2000. **95**(1): p. 342-6.
13. Garner, C., et al., *Haplotype Mapping of a Major Quantitative-Trait Locus for Fetal Hemoglobin Production, on Chromosome 6q23.* The American Journal of Human Genetics, 1998. **62**(6): p. 1468-1474.
14. Sankaran, V.G., et al., *Human fetal hemoglobin expression is regulated by the developmental stage-specific repressor BCL11A.* Science, 2008. **322**(5909): p. 1839-42.
15. Thein, S.L., et al., *Intergenic variants of HBS1L-MYB are responsible for a major quantitative trait locus on chromosome 6q23 influencing fetal hemoglobin levels in adults.* Proc Natl Acad Sci U S A, 2007. **104**(27): p. 11346-51.
16. Borg, J., et al., *Haploinsufficiency for the erythroid transcription factor KLF1 causes hereditary persistence of fetal hemoglobin.* Nat Genet, 2010. **42**(9): p. 801-5.
17. Masuda, T., et al., *Transcription factors LRF and BCL11A independently repress expression of fetal hemoglobin.* Science, 2016. **351**(6270): p. 285-9.
18. Blobel, G.A., et al., *An international effort to cure a global health problem: A report on the 19th Hemoglobin Switching Conference.* Exp Hematol, 2015. **43**(10): p. 821-37.

19. Gottschling, D.E., et al., *Position effect at S. cerevisiae telomeres: reversible repression of Pol II transcription.* Cell, 1990. **63**(4): p. 751-62.
20. Robin, J.D., et al., *Telomere position effect: regulation of gene expression with progressive telomere shortening over long distances.* Genes Dev, 2014. **28**(22): p. 2464-76.
21. Anguita, E., et al., *Globin gene activation during haemopoiesis is driven by protein complexes nucleated by GATA-1 and GATA-2.* EMBO J, 2004. **23**(14): p. 2841-52.
22. Loyd, M.R., *Role of AP1/NFE2 binding sites in endogenous -globin gene transcription.* Blood, 2003. **102**(12): p. 4223-4228.
23. Drissen, R., et al., *The active spatial organization of the beta-globin locus requires the transcription factor EKLF.* Genes Dev, 2004. **18**(20): p. 2485-90.
24. Tallack, M.R., et al., *A global role for KLF1 in erythropoiesis revealed by ChIP-seq in primary erythroid cells.* Genome Res, 2010. **20**(8): p. 1052-63.
25. Hughes, J.R., et al., *Annotation of cis-regulatory elements by identification, subclassification, and functional assessment of multispecies conserved sequences.* Proc Natl Acad Sci U S A, 2005. **102**(28): p. 9830-5.
26. Vyas, P., et al., *Cis-acting sequences regulating expression of the human alpha-globin cluster lie within constitutively open chromatin.* Cell, 1992. **69**(5): p. 781-93.
27. Viprakasit, V., et al., *De novo deletion within the telomeric region flanking the human alpha globin locus as a cause of alpha thalassaemia.* Br J Haematol, 2003. **120**(5): p. 867-75.
28. Viprakasit, V., et al., *A novel deletion causing alpha thalassemia clarifies the importance of the major human alpha globin regulatory element.* Blood, 2006. **107**(9): p. 3811-2.
29. Sollaino, M.C., et al., *Homozygous deletion of the major alpha-globin regulatory element (MCS-R2) responsible for a severe case of hemoglobin H disease.* Blood, 2010. **116**(12): p. 2193-4.
30. Coelho, A., et al., *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, 2010. **45**(2): p. 147-53.
31. Wu, M.Y., et al., *A novel selective deletion of the major alpha-globin regulatory element (MCS-R2) causing alpha-thalassaemia.* Br J Haematol, 2017. **176**(6): p. 984-986.
32. Higgs, D.R., J.A. Sharpe, and W.G. Wood, *Understanding alpha globin gene expression: a step towards effective gene therapy.* Semin Hematol, 1998. **35**(2): p. 93-104.
33. Wallace, H.A., et al., *Manipulating the mouse genome to engineer precise functional syntenic replacements with human sequence.* Cell, 2007. **128**(1): p. 197-209.
34. Jarman, A.P., et al., *Characterization of the major regulatory element upstream of the human alpha-globin gene cluster.* Mol Cell Biol, 1991. **11**(9): p. 4679-89.
35. Chen, H., C.H. Lowrey, and G. Stamatoyannopoulos, *Analysis of enhancer function of the HS-40 core sequence of the human alpha-globin cluster.* Nucleic Acids Res, 1997. **25**(14): p. 2917-22.
36. Hay, D., et al., *Genetic dissection of the alpha-globin super-enhancer in vivo.* Nat Genet, 2016. **48**(8): p. 895-903.

37. Vernimmen, D., et al., *Long-range chromosomal interactions regulate the timing of the transition between poised and active gene expression.* EMBO J, 2007. **26**(8): p. 2041-51.
38. Vernimmen, D., et al., *Chromosome looping at the human alpha-globin locus is mediated via the major upstream regulatory element (HS -40).* Blood, 2009. **114**(19): p. 4253-60.
39. Garrick, D., et al., *The role of the polycomb complex in silencing alpha-globin gene expression in nonerythroid cells.* Blood, 2008. **112**(9): p. 3889-99.
40. Anguita, E., et al., *Identification of a conserved erythroid specific domain of histone acetylation across the alpha-globin gene cluster.* Proc Natl Acad Sci U S A, 2001. **98**(21): p. 12114-9.
41. De Gobbi, M., et al., *Tissue-specific histone modification and transcription factor binding in alpha globin gene expression.* Blood, 2007. **110**(13): p. 4503-10.
42. Vernimmen, D., et al., *Polycomb eviction as a new distant enhancer function.* Genes Dev, 2011. **25**(15): p. 1583-8.
43. Vernimmen, D., *Uncovering enhancer functions using the alpha-globin locus.* PLoS Genet, 2014. **10**(10): p. e1004668.
44. Wilkie, A.O., et al., *Clinical features and molecular analysis of the alpha thalassemia/mental retardation syndromes. II. Cases without detectable abnormality of the alpha globin complex.* Am J Hum Genet, 1990. **46**(6): p. 1127-40.
45. McPherson, E.W., et al., *X-linked alpha-thalassemia/mental retardation (ATR-X) syndrome: a new kindred with severe genital anomalies and mild hematologic expression.* Am J Med Genet, 1995. **55**(3): p. 302-6.
46. Logie, L.J., et al., *Alpha thalassaemia mental retardation (ATR-X): an atypical family.* Arch Dis Child, 1994. **70**(5): p. 439-40.
47. Villard, L., D. Lacombe, and M. Fontes, *A point mutation in the XNP gene, associated with an ATR-X phenotype without alpha-thalassemia.* Eur J Hum Genet, 1996. **4**(6): p. 316-20.
48. Villard, L., et al., *Splicing mutation in the ATR-X gene can lead to a dysmorphic mental retardation phenotype without alpha-thalassemia.* Am J Hum Genet, 1996. **58**(3): p. 499-505.
49. Cole, T.R., A. May, and H.E. Hughes, *Alpha thalassaemia/mental retardation syndrome (non-deletional type): report of a family supporting X linked inheritance.* J Med Genet, 1991. **28**(11): p. 734-7.
50. Donnai, D., et al., *The non-deletion alpha thalassaemia/mental retardation syndrome: further support for X linkage.* J Med Genet, 1991. **28**(11): p. 742-5.
51. Gibbons, R.J., et al., *X-linked alpha-thalassemia/mental retardation (ATR-X) syndrome: localization to Xq12-q21.31 by X inactivation and linkage analysis.* Am J Hum Genet, 1992. **51**(5): p. 1136-49.
52. Gibbons, R.J., et al., *Mutations in a putative global transcriptional regulator cause X-linked mental retardation with alpha-thalassemia (ATR-X syndrome).* Cell, 1995. **80**(6): p. 837-45.
53. Villard, L., et al., *Determination of the genomic structure of the XNP/ATRX gene encoding a potential zinc finger helicase.* Genomics, 1997. **43**(2): p. 149-55.

54. Picketts, D.J., et al., *ATRX encodes a novel member of the SNF2 family of proteins: mutations point to a common mechanism underlying the ATR-X syndrome.* Hum Mol Genet, 1996. **5**(12): p. 1899-907.
55. Garrick, D., et al., *A conserved truncated isoform of the ATR-X syndrome protein lacking the SWI/SNF-homology domain.* Gene, 2004. **326**: p. 23-34.
56. Argentaro, A., et al., *Structural consequences of disease-causing mutations in the ATRX-DNMT3-DNMT3L (ADD) domain of the chromatin-associated protein ATRX.* Proceedings of the National Academy of Sciences, 2007. **104**(29): p. 11939-11944.
57. Nan, X., et al., *Interaction between chromatin proteins MECP2 and ATRX is disrupted by mutations that cause inherited mental retardation.* Proceedings of the National Academy of Sciences, 2007. **104**(8): p. 2709-2714.
58. Kernohan, K.D., et al., *Analysis of neonatal brain lacking ATRX or MeCP2 reveals changes in nucleosome density, CTCF binding and chromatin looping.* Nucleic Acids Research, 2014. **42**(13): p. 8356-8368.
59. Cardoso, C., et al., *Specific interaction between the XNP/ATR-X gene product and the SET domain of the human EZH2 protein.* Hum Mol Genet, 1998. **7**(4): p. 679-84.
60. Wang, X., et al., *Structural basis for DAXX interaction with ATRX.* Protein Cell, 2017. **8**(10): p. 767-771.
61. Hoelper, D., et al., *Structural and mechanistic insights into ATRX-dependent and -independent functions of the histone chaperone DAXX.* Nat Commun, 2017. **8**(1): p. 1193.
62. Gibbons, R.J., et al., *Mutations in the chromatin-associated protein ATRX.* Hum Mutat, 2008. **29**(6): p. 796-802.
63. Mitson, M., et al., *Functional significance of mutations in the Snf2 domain of ATRX.* Human Molecular Genetics, 2011. **20**(13): p. 2603-2610.
64. Xue, Y., et al., *The ATRX syndrome protein forms a chromatin-remodeling complex with Daxx and localizes in promyelocytic leukemia nuclear bodies.* Proc Natl Acad Sci U S A, 2003. **100**(19): p. 10635-40.
65. Luciani, J.J., *PML nuclear bodies are highly organised DNA-protein structures with a function in heterochromatin remodelling at the G2 phase.* Journal of Cell Science, 2006. **119**(12): p. 2518-2531.
66. McDowell, T.L., et al., *Localization of a putative transcriptional regulator (ATRX) at pericentromeric heterochromatin and the short arms of acrocentric chromosomes.* Proc Natl Acad Sci U S A, 1999. **96**(24): p. 13983-8.
67. Baumann, C., et al., *Association of ATRX with pericentric heterochromatin and the Y chromosome of neonatal mouse spermatogonia.* BMC Mol Biol, 2008. **9**: p. 29.
68. Baumann, C. and R. De La Fuente, *ATRX marks the inactive X chromosome (Xi) in somatic cells and during imprinted X chromosome inactivation in trophoblast stem cells.* Chromosoma, 2009. **118**(2): p. 209-22.
69. Sadic, D., et al., *Atrx promotes heterochromatin formation at retrotransposons.* EMBO reports, 2015. **16**(7): p. 836-850.
70. Valle-García, D., et al., *ATRX binds to atypical chromatin domains at the 3′ exons of zinc finger genes to preserve H3K9me3 enrichment.* Epigenetics, 2016. **11**(6): p. 398-414.

71. Law, M.J., et al., *ATR-X syndrome protein targets tandem repeats and influences allele-specific expression in a size-dependent manner.* Cell, 2010. **143**(3): p. 367-78.
72. Su, Z., et al., *ChIP-less analysis of chromatin states.* Epigenetics Chromatin, 2014. **7**: p. 7.
73. Eustermann, S., et al., *Combinatorial readout of histone H3 modifications specifies localization of ATRX to heterochromatin.* Nature Structural & Molecular Biology, 2011. **18**(7): p. 777-782.
74. Dhayalan, A., et al., *The ATRX-ADD domain binds to H3 tail peptides and reads the combined methylation state of K4 and K9.* Hum Mol Genet, 2011. **20**(11): p. 2195-203.
75. Iwase, S., et al., *ATRX ADD domain links an atypical histone methylation recognition mechanism to human mental-retardation syndrome.* Nat Struct Mol Biol, 2011. **18**(7): p. 769-76.
76. Jacobs, S.A., et al., *Specificity of the HP1 chromo domain for the methylated N-terminus of histone H3.* EMBO J, 2001. **20**(18): p. 5232-41.
77. Lechner, M.S., et al., *The mammalian heterochromatin protein 1 binds diverse nuclear proteins through a common motif that targets the chromoshadow domain.* Biochem Biophys Res Commun, 2005. **331**(4): p. 929-37.
78. Kourmouli, N., et al., *Epigenetic regulation of mammalian pericentric heterochromatin in vivo by HP1.* Biochem Biophys Res Commun, 2005. **337**(3): p. 901-7.
79. Wong, L.H., et al., *ATRX interacts with H3.3 in maintaining telomere structural integrity in pluripotent embryonic stem cells.* Genome Res, 2010. **20**(3): p. 351-60.
80. Goldberg, A.D., et al., *Distinct factors control histone variant H3.3 localization at specific genomic regions.* Cell, 2010. **140**(5): p. 678-91.
81. Elsasser, S.J., et al., *DAXX envelops a histone H3.3-H4 dimer for H3.3-specific recognition.* Nature, 2012. **491**(7425): p. 560-5.
82. Liu, C.P., et al., *Structure of the variant histone H3.3-H4 heterodimer in complex with its chaperone DAXX.* Nat Struct Mol Biol, 2012. **19**(12): p. 1287-92.
83. Drane, P., et al., *The death-associated protein DAXX is a novel histone chaperone involved in the replication-independent deposition of H3.3.* Genes Dev, 2010. **24**(12): p. 1253-65.
84. Lewis, P.W., et al., *Daxx is an H3.3-specific histone chaperone and cooperates with ATRX in replication-independent chromatin assembly at telomeres.* Proc Natl Acad Sci U S A, 2010. **107**(32): p. 14075-80.
85. Delbarre, E., et al., *DAXX-dependent supply of soluble (H3.3-H4) dimers to PML bodies pending deposition into chromatin.* Genome Res, 2013. **23**(3): p. 440-51.
86. Delbarre, E., et al., *PML protein organizes heterochromatin domains where it regulates histone H3.3 deposition by ATRX/DAXX.* Genome Research, 2017. **27**(6): p. 913-921.
87. Voon, Hsiao P.J., et al., *ATRX Plays a Key Role in Maintaining Silencing at Interstitial Heterochromatic Loci and Imprinted Genes.* Cell Reports, 2015. **11**(3): p. 405-418.

88. Tagami, H., et al., *Histone H3.1 and H3.3 complexes mediate nucleosome assembly pathways dependent or independent of DNA synthesis.* Cell, 2004. **116**(1): p. 51-61.
89. Chow, C.M., et al., *Variant histone H3.3 marks promoters of transcriptionally active genes during mammalian cell division.* EMBO Rep, 2005. **6**(4): p. 354-60.
90. Levy, M.A., et al., *ATRX promotes gene expression by facilitating transcriptional elongation through guanine-rich coding regions.* Hum Mol Genet, 2015. **24**(7): p. 1824-35.
91. Hoogsteen, K., *The crystal and molecular structure of a hydrogen-bonded complex between 1-methylthymine and 9-methyladenine.* Acta Crystallographica, 1963. **16**(9): p. 907-916.
92. Chambers, V.S., et al., *High-throughput sequencing of DNA G-quadruplex structures in the human genome.* Nat Biotechnol, 2015. **33**(8): p. 877-81.
93. Burge, S., et al., *Quadruplex DNA: sequence, topology and structure.* Nucleic Acids Research, 2006. **34**(19): p. 5402-5415.
94. Xu, Y., et al., *Finding a human telomere DNA–RNA hybrid G-quadruplex formed by human telomeric 6-mer RNA and 16-mer DNA using click chemistry: A protective structure for telomere end.* Bioorganic & Medicinal Chemistry, 2014. **22**(16): p. 4419-4421.
95. Schaffitzel, C., et al., *In vitro generated antibodies specific for telomeric guanine-quadruplex DNA react with Stylonychia lemnae macronuclei.* Proc Natl Acad Sci U S A, 2001. **98**(15): p. 8572-7.
96. Duquette, M.L., et al., *Intracellular transcription of G-rich DNAs induces formation of G-loops, novel structures containing G4 DNA.* Genes Dev, 2004. **18**(13): p. 1618-29.
97. Granotier, C., et al., *Preferential binding of a G-quadruplex ligand to human chromosome ends.* Nucleic Acids Res, 2005. **33**(13): p. 4182-90.
98. Biffi, G., et al., *Quantitative visualization of DNA G-quadruplex structures in human cells.* Nat Chem, 2013. **5**(3): p. 182-6.
99. Henderson, A., et al., *Detection of G-quadruplex DNA in mammalian cells.* Nucleic Acids Res, 2014. **42**(2): p. 860-9.
100. Hänsel-Hertsch, R., et al., *G-quadruplex structures mark human regulatory chromatin.* Nature Genetics, 2016. **48**(10): p. 1267-1272.
101. Eddy, J. and N. Maizels, *Conserved elements with potential to form polymorphic G-quadruplex structures in the first intron of human genes.* Nucleic Acids Res, 2008. **36**(4): p. 1321-33.
102. Huppert, J.L., et al., *G-quadruplexes: the beginning and end of UTRs.* Nucleic Acids Res, 2008. **36**(19): p. 6260-8.
103. Li, Y., et al., *Effect of ATRX and G-Quadruplex Formation by the VNTR Sequence on α-Globin Gene Expression.* ChemBioChem, 2016. **17**(10): p. 928-935.
104. Bochman, M.L., K. Paeschke, and V.A. Zakian, *DNA secondary structures: stability and function of G-quadruplex structures.* Nature Reviews Genetics, 2012. **13**(11): p. 770-780.
105. Sekeres, M.A., *The epidemiology of myelodysplastic syndromes.* Hematol Oncol Clin North Am, 2010. **24**(2): p. 287-94.

106. Arber, D.A., et al., *The 2016 revision to the World Health Organization classification of myeloid neoplasms and acute leukemia.* Blood, 2016. **127**(20): p. 2391-2405.
107. Greenberg, P.L., et al., *Revised International Prognostic Scoring System for Myelodysplastic Syndromes.* Blood, 2012. **120**(12): p. 2454-2465.
108. Sanada, C., et al., *Adult human megakaryocyte-erythroid progenitors are in the CD34+CD38mid fraction.* Blood, 2016. **128**(7): p. 923-33.
109. Chen, K., et al., *Resolving the distinct stages in erythroid differentiation based on dynamic changes in membrane protein expression during erythropoiesis.* Proceedings of the National Academy of Sciences, 2009. **106**(41): p. 17413-17418.
110. Hu, J., et al., *Isolation and functional characterization of human erythroblasts at distinct stages: implications for understanding of normal and disordered erythropoiesis in vivo.* Blood, 2013. **121**(16): p. 3246-53.
111. Pimentel, H., et al., *A dynamic alternative splicing program regulates gene expression during terminal erythropoiesis.* Nucleic Acids Res, 2014. **42**(6): p. 4031-42.
112. Pimentel, H., et al., *A dynamic intron retention program enriched in RNA processing genes regulates gene expression during terminal erythropoiesis.* Nucleic Acids Res, 2016. **44**(2): p. 838-51.
113. Edwards, C.R., et al., *A dynamic intron retention program in the mammalian megakaryocyte and erythrocyte lineages.* Blood, 2016. **127**(17): p. e24-e34.
114. Bejar, R., et al., *Clinical effect of point mutations in myelodysplastic syndromes.* N Engl J Med, 2011. **364**(26): p. 2496-506.
115. Papaemmanuil, E., et al., *Clinical and biological implications of driver mutations in myelodysplastic syndromes.* Blood, 2013. **122**(22): p. 3616-27; quiz 3699.
116. Haferlach, T., et al., *Landscape of genetic lesions in 944 patients with myelodysplastic syndromes.* Leukemia, 2014. **28**(2): p. 241-7.
117. Kennedy, J.A. and B.L. Ebert, *Clinical Implications of Genetic Mutations in Myelodysplastic Syndrome.* J Clin Oncol, 2017. **35**(9): p. 968-974.
118. Yoshida, K., et al., *Frequent pathway mutations of splicing machinery in myelodysplasia.* Nature, 2011. **478**(7367): p. 64-9.
119. Thol, F., et al., *Frequency and prognostic impact of mutations in SRSF2, U2AF1, and ZRSR2 in patients with myelodysplastic syndromes.* Blood, 2012. **119**(15): p. 3578-84.
120. Gerstung, M., et al., *Combining gene mutation with gene expression data improves outcome prediction in myelodysplastic syndromes.* Nature Communications, 2015. **6**(1).
121. Woll, P.S., et al., *Myelodysplastic syndromes are propagated by rare and distinct human cancer stem cells in vivo.* Cancer Cell, 2014. **25**(6): p. 794-808.
122. da Silva-Coelho, P., et al., *Clonal evolution in myelodysplastic syndromes.* Nat Commun, 2017. **8**: p. 15099.
123. White, J.C., et al., *An unstable haemoglobin associated with some cases of leukaemia.* Br J Haematol, 1960. **6**: p. 171-7.
124. Steensma, D.P., R.J. Gibbons, and D.R. Higgs, *Acquired alpha-thalassemia in association with myelodysplastic syndrome and other hematologic malignancies.* Blood, 2005. **105**(2): p. 443-52.

125. Tennant, G.B., A.I. Al-Sabah, and A.K. Burnett, *Prognosis of myelodysplasic patients: non-parametric multiple regression analysis of populations stratified by mean corpuscular volume and marrow myeloblast number.* Br J Haematol, 2002. **119**(1): p. 87-96.
126. May, S.J., et al., *The myelodysplastic syndrome: analysis of laboratory characteristics in relation to the FAB classification.* Br J Haematol, 1985. **59**(2): p. 311-9.
127. Wang, H., et al., *Mean corpuscular volume predicts prognosis in MDS patients with abnormal karyotypes.* Ann Hematol, 2010. **89**(7): p. 671-9.
128. Herbaux, C., et al., *Incidence of ATRX mutations in myelodysplastic syndromes, the value of microcytosis.* Am J Hematol, 2015. **90**(8): p. 737-8.
129. Higgs, D.R., et al., *Clinical features and molecular analysis of acquired hemoglobin H disease.* Am J Med, 1983. **75**(2): p. 181-91.
130. Anagnou, N.P., et al., *Acquired alpha-thalassemia in preleukemia is due to decreased expression of all four alpha-globin genes.* Proc Natl Acad Sci U S A, 1983. **80**(19): p. 6051-5.
131. Helder, J. and A. Deisseroth, *S1 nuclease analysis of alpha-globin gene expression in preleukemic patients with acquired hemoglobin H disease after transfer to mouse erythroleukemia cells.* Proceedings of the National Academy of Sciences, 1987. **84**(8): p. 2387-2390.
132. Gibbons, R.J., et al., *Identification of acquired somatic mutations in the gene encoding chromatin-remodeling factor ATRX in the alpha-thalassemia myelodysplasia syndrome (ATMDS).* Nat Genet, 2003. **34**(4): p. 446-9.
133. Steensma, D.P., et al., *Acquired somatic ATRX mutations in myelodysplastic syndrome associated with alpha thalassemia (ATMDS) convey a more severe hematologic phenotype than germline ATRX mutations.* Blood, 2004. **103**(6): p. 2019-26.
134. Nelson, M.E., et al., *A novel 5' ATRX mutation with splicing consequences in acquired alpha thalassemia-myelodysplastic syndrome.* Haematologica, 2005. **90**(11): p. 1463-70.
135. Costa, D.B., et al., *A novel mutation in the last exon of ATRX in a patient with alpha-thalassemia myelodysplastic syndrome.* Eur J Haematol, 2006. **76**(5): p. 432-5, 453.
136. Rose, C., et al., *Acquired alpha thalassemia myelodyslastic/myeloproliferative syndrome (ATMDS): evolution on hypomethylating agent therapy.* Leuk Res, 2011. **35**(11): p. e203-5.
137. Herbaux, C., et al., *A new ATRX mutation in a patient with acquired alpha-thalassemia myelodysplastic syndrome.* Hemoglobin, 2012. **36**(6): p. 581-5.
138. Steensma, D.P., et al., *Deletion of the alpha-globin gene cluster as a cause of acquired alpha-thalassemia in myelodysplastic syndrome.* Blood, 2004. **103**(4): p. 1518-20.
139. Korbie, D.J. and J.S. Mattick, *Touchdown PCR for increased specificity and sensitivity in PCR amplification.* Nat Protoc, 2008. **3**(9): p. 1452-6.
140. Studer, D., et al., *Ribosomal protein l13a as a reference gene for human bone marrow-derived mesenchymal stromal cells during expansion, adipo-, chondro-, and osteogenesis.* Tissue Eng Part C Methods, 2012. **18**(10): p. 761-71.
141. Gentile, A.M., et al., *RPL13A and EEF1A1 Are Suitable Reference Genes for qPCR during Adipocyte Differentiation of Vascular Stromal Cells from*

Patients with Different BMI and HOMA-IR. PLoS One, 2016. **11**(6): p. e0157002.
142. Jesnowski, R., et al., *Ribosomal Highly Basic 23-kDa Protein as a Reliable Standard for Gene Expression Analysis.* Pancreatology, 2002. **2**(4): p. 421-424.
143. Zhang, Y., et al., *Model-based analysis of ChIP-Seq (MACS).* Genome Biol, 2008. **9**(9): p. R137.
144. Ross-Innes, C.S., et al., *Differential oestrogen receptor binding is associated with clinical outcome in breast cancer.* Nature, 2012. **481**(7381): p. 389-393.
145. Afgan, E., et al., *The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update.* Nucleic Acids Res, 2018. **46**(W1): p. W537-W544.
146. Machanick, P. and T.L. Bailey, *MEME-ChIP: motif analysis of large DNA datasets.* Bioinformatics, 2011. **27**(12): p. 1696-1697.
147. Ma, W., W.S. Noble, and T.L. Bailey, *Motif-based analysis of large nucleotide data sets using MEME-ChIP.* Nature Protocols, 2014. **9**(6): p. 1428-1450.
148. Yang, H. and K. Wang, *Genomic variant annotation and prioritization with ANNOVAR and wANNOVAR.* Nature Protocols, 2015. **10**(10): p. 1556-1566.
149. Griffiths, R.E., et al., *Maturing reticulocytes internalize plasma membrane in glycophorin A-containing vesicles that fuse with autophagosomes before exocytosis.* Blood, 2012. **119**(26): p. 6296-306.
150. Kurita, R., et al., *Establishment of immortalized human erythroid progenitor cell lines able to produce enucleated red blood cells.* PLoS One, 2013. **8**(3): p. e59890.
151. Ran, F.A., et al., *Genome engineering using the CRISPR-Cas9 system.* Nat Protoc, 2013. **8**(11): p. 2281-2308.
152. Hsu, P.D., et al., *DNA targeting specificity of RNA-guided Cas9 nucleases.* Nature Biotechnology, 2013. **31**(9): p. 827-832.
153. Richardson, C.D., et al., *Enhancing homology-directed genome editing by catalytically active and inactive CRISPR-Cas9 using asymmetric donor DNA.* Nat Biotechnol, 2016. **34**(3): p. 339-44.
154. Jayathilaka, K., et al., *A chemical compound that stimulates the human homologous recombination protein RAD51.* Proceedings of the National Academy of Sciences, 2008. **105**(41): p. 15848-15853.
155. Song, J., et al., *RS-1 enhances CRISPR/Cas9- and TALEN-mediated knock-in efficiency.* Nature Communications, 2016. **7**: p. 10548.
156. Pinder, J., J. Salsman, and G. Dellaire, *Nuclear domain 'knock-in' screen for the evaluation and identification of small molecule enhancers of CRISPR-based genome editing.* Nucleic Acids Research, 2015. **43**(19): p. 9379-9392.
157. Chu, V.T., et al., *Increasing the efficiency of homology-directed repair for CRISPR-Cas9-induced precise gene editing in mammalian cells.* Nat Biotechnol, 2015. **33**(5): p. 543-8.
158. Li, G., et al., *Small molecules enhance CRISPR/Cas9-mediated homology-directed genome editing in primary cells.* Sci Rep, 2017. **7**(1): p. 8943.
159. Forbes, S.A., et al., *COSMIC: somatic cancer genetics at high-resolution.* Nucleic Acids Research, 2017. **45**(D1): p. D777-D783.
160. Papaemmanuil, E., et al., *Somatic SF3B1 mutation in myelodysplasia with ring sideroblasts.* N Engl J Med, 2011. **365**(15): p. 1384-95.

161. Mori, T., et al., *Somatic PHF6 mutations in 1760 cases with various myeloid neoplasms.* Leukemia, 2016. **30**(11): p. 2270-2273.
162. Damm, F., et al., *Mutations affecting mRNA splicing define distinct clinical phenotypes and correlate with patient outcome in myelodysplastic syndromes.* Blood, 2012. **119**(14): p. 3211-8.
163. Dvinge, H., et al., *RNA splicing factors as oncoproteins and tumour suppressors.* Nature Reviews Cancer, 2016. **16**(7): p. 413-430.
164. Boultwood, J., et al., *Frequent mutation of the polycomb-associated gene ASXL1 in the myelodysplastic syndromes and in acute myeloid leukemia.* Leukemia, 2010. **24**(5): p. 1062-1065.
165. Yang, H., et al., *Gain of function of ASXL1 truncating protein in the pathogenesis of myeloid malignancies.* Blood, 2018. **131**(3): p. 328-341.
166. Inoue, D., et al., *Truncation mutants of ASXL1 observed in myeloid malignancies are expressed at detectable protein levels.* Exp Hematol, 2016. **44**(3): p. 172-6 e1.
167. Abdel-Wahab, O., et al., *ASXL1 mutations promote myeloid transformation through loss of PRC2-mediated gene repression.* Cancer Cell, 2012. **22**(2): p. 180-93.
168. Balasubramani, A., et al., *Cancer-associated ASXL1 mutations may act as gain-of-function mutations of the ASXL1-BAP1 complex.* Nat Commun, 2015. **6**: p. 7307.
169. Asada, S., et al., *Mutant ASXL1 cooperates with BAP1 to promote myeloid leukaemogenesis.* Nat Commun, 2018. **9**(1): p. 2733.
170. Devaiah, B.N., et al., *BRD4 is a histone acetyltransferase that evicts nucleosomes from chromatin.* Nat Struct Mol Biol, 2016. **23**(6): p. 540-8.
171. Jang, M.K., et al., *The bromodomain protein Brd4 is a positive regulatory component of P-TEFb and stimulates RNA polymerase II-dependent transcription.* Mol Cell, 2005. **19**(4): p. 523-34.
172. Wang, S., et al., *BRD4 inhibitors block telomere elongation.* Nucleic Acids Res, 2017. **45**(14): p. 8403-8410.
173. Stonestrom, A.J., et al., *Functions of BET proteins in erythroid gene expression.* Blood, 2015. **125**(18): p. 2825-34.
174. Liu, X., et al., *In Situ Capture of Chromatin Interactions by Biotinylated dCas9.* Cell, 2017. **170**(5): p. 1028-1043 e19.
175. Chen, T.C., et al., *Dynamics of ASXL1 mutation and other associated genetic alterations during disease progression in patients with primary myelodysplastic syndrome.* Blood Cancer Journal, 2014. **4**(1): p. e177-e177.
176. Gelsi-Boyer, V., et al., *Mutations in ASXL1 are associated with poor prognosis across the spectrum of malignant myeloid diseases.* Journal of Hematology & Oncology, 2012. **5**(1): p. 12.
177. Shi, H., et al., *ASXL1 plays an important role in erythropoiesis.* Sci Rep, 2016. **6**: p. 28789.
178. Visayaragawan, N., et al., *Bohring-opitz syndrome - A case of a rare genetic disorder.* Med J Malaysia, 2017. **72**(4): p. 248-249.
179. Katoh, M., *Functional and cancer genomics of ASXL family members.* Br J Cancer, 2013. **109**(2): p. 299-306.
180. Patnaik, M.M. and A. Tefferi, *Chronic myelomonocytic leukemia: 2018 update on diagnosis, risk stratification and management.* Am J Hematol, 2018. **93**(6): p. 824-840.

181. Jaiswal, S., et al., *Age-Related Clonal Hematopoiesis Associated with Adverse Outcomes.* New England Journal of Medicine, 2014. **371**(26): p. 2488-2498.
182. Nagase, R., et al., *Expression of mutant Asxl1 perturbs hematopoiesis and promotes susceptibility to leukemic transformation.* J Exp Med, 2018. **215**(6): p. 1729-1747.
183. Uni, M., et al., *Modeling ASXL1 mutation revealed impaired hematopoiesis caused by derepression of p16Ink4a through aberrant PRC1-mediated histone modification.* Leukemia, 2018.
184. Valcarcel, J., et al., *Interaction of U2AF65 RS region with pre-mRNA branch point and promotion of base pairing with U2 snRNA [corrected].* Science, 1996. **273**(5282): p. 1706-9.
185. Zorio, D.A.R. and T. Blumenthal, *Both subunits of U2AF recognize the 3′ splice site in Caenorhabditis elegans.* Nature, 1999. **402**(6763): p. 835-838.
186. Madan, V., et al., *Aberrant splicing of U12-type introns is the hallmark of ZRSR2 mutant myelodysplastic syndrome.* Nat Commun, 2015. **6**: p. 6042.
187. Yoshimi, A. and O. Abdel-Wahab, *Molecular Pathways: Understanding and Targeting Mutant Spliceosomal Proteins.* Clinical Cancer Research, 2017. **23**(2): p. 336-341.
188. Wahl, M.C., C.L. Will, and R. Luhrmann, *The spliceosome: design principles of a dynamic RNP machine.* Cell, 2009. **136**(4): p. 701-18.
189. David, C.J., et al., *The RNA polymerase II C-terminal domain promotes splicing activation through recruitment of a U2AF65-Prp19 complex.* Genes Dev, 2011. **25**(9): p. 972-83.
190. Glasser, E., et al., *Cancer-Associated Mutations Mapped on High-Resolution Structures of the U2AF2 RNA Recognition Motifs.* Biochemistry, 2017. **56**(36): p. 4757-4761.
191. Zamore, P.D. and M.R. Green, *Biochemical characterization of U2 snRNP auxiliary factor: an essential pre-mRNA splicing factor with a novel intranuclear distribution.* The EMBO Journal, 1991. **10**(1): p. 207-214.
192. Rudner, D.Z., et al., *Mutations in the small subunit of the Drosophila U2AF splicing factor cause lethality and developmental defects.* Proc Natl Acad Sci U S A, 1996. **93**(19): p. 10333-7.
193. Wu, S., et al., *Functional recognition of the 3′ splice site AG by the splicing factor U2AF35.* Nature, 1999. **402**(6763): p. 832-5.
194. Merendino, L., et al., *Inhibition of msl-2 splicing by Sex-lethal reveals interaction between U2AF35 and the 3′ splice site AG.* Nature, 1999. **402**(6763): p. 838-41.
195. Webb, C.J. and J.A. Wise, *The splicing factor U2AF small subunit is functionally conserved between fission yeast and humans.* Mol Cell Biol, 2004. **24**(10): p. 4229-40.
196. Yoshida, H., et al., *A novel 3′ splice site recognition by the two zinc fingers in the U2AF small subunit.* Genes Dev, 2015. **29**(15): p. 1649-60.
197. Zuo, P. and T. Maniatis, *The splicing factor U2AF35 mediates critical protein-protein interactions in constitutive and enhancer-dependent splicing.* Genes Dev, 1996. **10**(11): p. 1356-68.
198. Gama-Carvalho, M., et al., *Nucleocytoplasmic shuttling of heterodimeric splicing factor U2AF.* J Biol Chem, 2001. **276**(16): p. 13104-12.

199. Palangat, M., et al., *The splicing factor U2AF1 contributes to cancer progression through a non-canonical role in translation regulation.* In press, 2018.
200. Fei, D.L., et al., *Wild-Type U2AF1 Antagonizes the Splicing Program Characteristic of U2AF1-Mutant Tumors and Is Required for Cell Survival.* PLoS Genet, 2016. **12**(10): p. e1006384.
201. Graubert, T.A., et al., *Recurrent mutations in the U2AF1 splicing factor in myelodysplastic syndromes.* Nat Genet, 2011. **44**(1): p. 53-7.
202. Przychodzen, B., et al., *Patterns of missplicing due to somatic U2AF1 mutations in myeloid neoplasms.* Blood, 2013. **122**(6): p. 999-1006.
203. Okeyo-Owuor, T., et al., *U2AF1 mutations alter sequence specificity of pre-mRNA binding and splicing.* Leukemia, 2015. **29**(4): p. 909-17.
204. Shirai, C.L., et al., *Mutant U2AF1 Expression Alters Hematopoiesis and Pre-mRNA Splicing In Vivo.* Cancer Cell, 2015. **27**(5): p. 631-43.
205. Ilagan, J.O., et al., *U2AF1 mutations alter splice site recognition in hematological malignancies.* Genome Res, 2015. **25**(1): p. 14-26.
206. Herdt, O., et al., *The cancer-associated U2AF35 470A>G (Q157R) mutation creates an in-frame alternative 5' splice site that impacts splicing regulation in Q157R patients.* RNA, 2017. **23**(12): p. 1796-1806.
207. Park, Sung M., et al., *U2AF35(S34F) Promotes Transformation by Directing Aberrant ATG7 Pre-mRNA 3′ End Formation.* Molecular Cell, 2016. **62**(4): p. 479-490.
208. Inoue, D., R.K. Bradley, and O. Abdel-Wahab, *Spliceosomal gene mutations in myelodysplasia: molecular links to clonal abnormalities of hematopoiesis.* Genes Dev, 2016. **30**(9): p. 989-1001.
209. Brooks, A.N., et al., *A pan-cancer analysis of transcriptome changes associated with somatic mutations in U2AF1 reveals commonly altered splicing events.* PLoS One, 2014. **9**(1): p. e87361.
210. Yip, B.H., et al., *The U2AF1S34F mutation induces lineage-specific splicing alterations in myelodysplastic syndromes.* J Clin Invest, 2017. **127**(6): p. 2206-2221.
211. Li, B., et al., *Clinical features and biological implications of different U2AF1 mutation types in myelodysplastic syndromes.* Genes, Chromosomes and Cancer, 2018. **57**(2): p. 80-88.
212. Tefferi, A., et al., *U2AF1 mutation types in primary myelofibrosis: phenotypic and prognostic distinctions.* Leukemia, 2018.
213. Walter, M.J., et al., *Clonal diversity of recurrently mutated genes in myelodysplastic syndromes.* Leukemia, 2013. **27**(6): p. 1275-82.
214. Wu, L., et al., *Genetic landscape of recurrent ASXL1, U2AF1, SF3B1, SRSF2, and EZH2 mutations in 304 Chinese patients with myelodysplastic syndromes.* Tumor Biology, 2015. **37**(4): p. 4633-4640.
215. Kim, S.Y., et al., *The high frequency of the U2AF1 S34Y mutation and its association with isolated trisomy 8 in myelodysplastic syndrome in Asians, but not in Caucasians.* Leuk Res, 2017. **61**: p. 96-103.
216. Lower, K.M., et al., *Mutations in PHF6 are associated with Borjeson-Forssman-Lehmann syndrome.* Nat Genet, 2002. **32**(4): p. 661-5.
217. Ernst, A., et al., *The PHF6 Mutation c.1A>G; pM1V Causes Borjeson-Forsman-Lehmann Syndrome in a Family with Four Affected Young Boys.* Mol Syndromol, 2015. **6**(4): p. 181-6.

218. Crawford, J., et al., *Mutation screening in Borjeson-Forssman-Lehmann syndrome: identification of a novel de novo PHF6 mutation in a female patient.* J Med Genet, 2006. **43**(3): p. 238-43.
219. Berland, S., et al., *PHF6 Deletions May Cause Borjeson-Forssman-Lehmann Syndrome in Females.* Mol Syndromol, 2011. **1**(6): p. 294-300.
220. Wieczorek, D., et al., *A comprehensive molecular study on Coffin-Siris and Nicolaides-Baraitser syndromes identifies a broad molecular and clinical spectrum converging on altered chromatin remodeling.* Hum Mol Genet, 2013. **22**(25): p. 5121-35.
221. Zweier, C., et al., *A new face of Borjeson-Forssman-Lehmann syndrome? De novo mutations in PHF6 in seven females with a distinct phenotype.* J Med Genet, 2013. **50**(12): p. 838-47.
222. Di Donato, N., et al., *Distinct phenotype of PHF6 deletions in females.* Eur J Med Genet, 2014. **57**(2-3): p. 85-9.
223. Todd, M.A., D. Ivanochko, and D.J. Picketts, *PHF6 Degrees of Separation: The Multifaceted Roles of a Chromatin Adaptor Protein.* Genes (Basel), 2015. **6**(2): p. 325-52.
224. Van Vlierberghe, P., et al., *PHF6 mutations in adult acute myeloid leukemia.* Leukemia, 2011. **25**(1): p. 130-4.
225. Van Vlierberghe, P., et al., *PHF6 mutations in T-cell acute lymphoblastic leukemia.* Nat Genet, 2010. **42**(4): p. 338-42.
226. Chao, M.M., et al., *T-cell acute lymphoblastic leukemia in association with Borjeson-Forssman-Lehmann syndrome due to a mutation in PHF6.* Pediatr Blood Cancer, 2010. **55**(4): p. 722-4.
227. Yoo, N.J., Y.R. Kim, and S.H. Lee, *Somatic mutation of PHF6 gene in T-cell acute lymphoblatic leukemia, acute myelogenous leukemia and hepatocellular carcinoma.* Acta Oncol, 2012. **51**(1): p. 107-11.
228. Wang, Q., et al., *Mutations of PHF6 are associated with mutations of NOTCH1, JAK1 and rearrangement of SET-NUP214 in T-cell acute lymphoblastic leukemia.* Haematologica, 2011. **96**(12): p. 1808-14.
229. Meacham, C.E., et al., *A genome-scale in vivo loss-of-function screen identifiesPhf6as a lineage-specific regulator of leukemia cell growth.* Genes & Development, 2015. **29**(5): p. 483-488.
230. Spinella, J.F., et al., *Genomic characterization of pediatric T-cell acute lymphoblastic leukemia reveals novel recurrent driver mutations.* Oncotarget, 2016. **7**(40): p. 65485-65503.
231. Wang, J., et al., *PHF6 regulates cell cycle progression by suppressing ribosomal RNA synthesis.* J Biol Chem, 2013. **288**(5): p. 3174-83.
232. Todd, M.A.M., M.S. Huh, and D.J. Picketts, *The sub-nucleolar localization of PHF6 defines its role in rDNA transcription and early processing events.* European Journal of Human Genetics, 2016. **24**(10): p. 1453-1459.
233. Zhang, C., et al., *The X-linked intellectual disability protein PHF6 associates with the PAF1 complex and regulates neuronal migration in the mammalian brain.* Neuron, 2013. **78**(6): p. 986-93.
234. Belotserkovskii, B.P. and P.C. Hanawalt, *Anchoring nascent RNA to the DNA template could interfere with transcription.* Biophys J, 2011. **100**(3): p. 675-684.

235. Wahba, L., et al., *RNase H and multiple RNA biogenesis factors cooperate to prevent RNA:DNA hybrids from generating genome instability.* Mol Cell, 2011. **44**(6): p. 978-88.
236. Todd, M.A. and D.J. Picketts, *PHF6 interacts with the nucleosome remodeling and deacetylation (NuRD) complex.* J Proteome Res, 2012. **11**(8): p. 4326-37.
237. Liu, Z., et al., *Structural basis of plant homeodomain finger 6 (PHF6) recognition by the retinoblastoma binding protein 4 (RBBP4) component of the nucleosome remodeling and deacetylase (NuRD) complex.* J Biol Chem, 2015. **290**(10): p. 6630-8.
238. Miccio, A. and G.A. Blobel, *Role of the GATA-1/FOG-1/NuRD pathway in the expression of human beta-like globin genes.* Mol Cell Biol, 2010. **30**(14): p. 3460-70.
239. Kasinath, V., et al., *Structures of human PRC2 with its cofactors AEBP2 and JARID2.* Science, 2018. **359**(6378): p. 940-944.
240. Chen, S., et al., *Unique Structural Platforms of Suz12 Dictate Distinct Classes of PRC2 for Chromatin Binding.* Mol Cell, 2018. **69**(5): p. 840-852 e5.
241. Youmans, D.T., J.C. Schmidt, and T.R. Cech, *Live-cell imaging reveals the dynamics of PRC2 and recruitment to chromatin by SUZ12-associated subunits.* Genes Dev, 2018. **32**(11-12): p. 794-805.
242. Liu, Z., et al., *Structural and functional insights into the human Borjeson-Forssman-Lehmann syndrome-associated protein PHF6.* J Biol Chem, 2014. **289**(14): p. 10069-83.
243. Lannon, M. and R. Wells, *Acquiring an H in MDS.* British Journal of Haematology, 2016. **174**(2): p. 173-173.
244. Hofmann, W.K., et al., *Frequent loss of heterozygosity in the region of D1S450 at 1p36.2 in myelodysplastic syndromes.* Leuk Res, 2001. **25**(10): p. 855-8.
245. Mori, N., et al., *Progression of myelodysplastic syndrome: allelic loss on chromosomal arm 1p.* Br J Haematol, 2003. **122**(2): p. 226-30.
246. Mori, N., et al., *Chromosome band 1p36 contains a putative tumor suppressor gene important in the evolution of chronic myelocytic leukemia.* Blood, 1998. **92**(9): p. 3405-9.
247. Makishima, H., et al., *Somatic SETBP1 mutations in myeloid malignancies.* Nat Genet, 2013. **45**(8): p. 942-6.
248. Inoue, D., et al., *SETBP1 mutations drive leukemic transformation in ASXL1-mutated MDS.* Leukemia, 2015. **29**(4): p. 847-57.
249. Guglielmelli, P., et al., *EZH2 mutational status predicts poor survival in myelofibrosis.* Blood, 2011. **118**(19): p. 5227-34.
250. Yang, Y., et al., *Loss of Ezh2 cooperates with Jak2V617F in the development of myelofibrosis in a mouse model of myeloproliferative neoplasm.* Blood, 2016. **127**(26): p. 3410-23.
251. Barraco, D., et al., *Molecular correlates of anemia in primary myelofibrosis: a significant and independent association with U2AF1 mutations.* Blood Cancer Journal, 2016. **6**(4): p. e415-e415.
252. Grand, F.H., et al., *Frequent CBL mutations associated with 11q acquired uniparental disomy in myeloproliferative neoplasms.* Blood, 2009. **113**(24): p. 6182-6192.

253. Haas, P.S., et al., *The role of X-inactivation in the gender bias of patients with acquired alpha-thalassaemia and myelodysplastic syndrome (ATMDS).* Br J Haematol, 2009. **144**(4): p. 538-45.
254. Peeters, S.B., et al., *Human cis-acting elements regulating escape from X-chromosome inactivation function in mouse.* Human Molecular Genetics, 2018. **27**(7): p. 1252-1262.
255. Turner, G., et al., *The clinical picture of the Borjeson-Forssman-Lehmann syndrome in males and heterozygous females with PHF6 mutations.* Clin Genet, 2004. **65**(3): p. 226-32.
256. Loken, M.R., et al., *Flow cytometric analysis of human bone marrow: I. Normal erythroid development.* Blood, 1987. **69**(1): p. 255-63.
257. Corces, M.R., et al., *Lineage-specific and single-cell chromatin accessibility charts human hematopoiesis and leukemia evolution.* Nat Genet, 2016. **48**(10): p. 1193-203.
258. Lower, K.M., et al., *Adventitious changes in long-range gene expression caused by polymorphic structural variation and promoter competition.* Proc Natl Acad Sci U S A, 2009. **106**(51): p. 21771-6.
259. Kerenyi, M.A. and S.H. Orkin, *Networking erythropoiesis.* J Exp Med, 2010. **207**(12): p. 2537-41.
260. Wadman, I.A., et al., *The LIM-only protein Lmo2 is a bridging molecule assembling an erythroid, DNA-binding complex which includes the TAL1, E47, GATA-1 and Ldb1/NLI proteins.* EMBO J, 1997. **16**(11): p. 3145-57.
261. Weiss, M.J. and S.H. Orkin, *Transcription factor GATA-1 permits survival and maturation of erythroid precursors by preventing apoptosis.* Proceedings of the National Academy of Sciences, 1995. **92**(21): p. 9623-9627.
262. De Maria, R., et al., Nature, 1999. **401**(6752): p. 489-493.
263. Hall, M.A., et al., *Functional but Abnormal Adult Erythropoiesis in the Absence of the Stem Cell Leukemia Gene.* Molecular and Cellular Biology, 2005. **25**(15): p. 6355-6362.
264. Kassouf, M.T., et al., *Differential use of SCL/TAL-1 DNA-binding domain in developmental hematopoiesis.* Blood, 2008. **112**(4): p. 1056-67.
265. Kassouf, M.T., et al., *Genome-wide identification of TAL1's functional targets: insights into its mechanisms of action in primary erythroid cells.* Genome Res, 2010. **20**(8): p. 1064-83.
266. Calkhoven, C.F., *Translational control of SCL-isoform expression in hematopoietic lineage choice.* Genes & Development, 2003. **17**(8): p. 959-964.
267. Jin, S., et al., *Splicing factor SF3B1K700E mutant dysregulates erythroid differentiation via aberrant alternative splicing of transcription factor TAL1.* PLoS One, 2017. **12**(5): p. e0175523.
268. Frontelo, P., et al., *Novel role for EKLF in megakaryocyte lineage commitment.* Blood, 2007. **110**(12): p. 3871-80.
269. Tallack, M.R. and A.C. Perkins, *Megakaryocyte-erythroid lineage promiscuity in EKLF null mouse blood.* Haematologica, 2010. **95**(1): p. 144-7.
270. Isern, J., et al., *Dose-dependent regulation of primitive erythroid maturation and identity by the transcription factor Eklf.* Blood, 2010. **116**(19): p. 3972-80.
271. Satta, S., et al., *Delayed fetal hemoglobin switching in subjects with KLF1 gene mutation.* Blood Cells Mol Dis, 2012. **48**(1): p. 22-4.

272. Vinjamur, D.S., et al., *Kruppel-Like Transcription Factor KLF1 Is Required for Optimal gamma- and beta-Globin Expression in Human Fetal Erythroblasts.* PLoS One, 2016. **11**(2): p. e0146802.
273. Zhou, D., et al., *KLF1 regulates BCL11A expression and gamma- to beta-globin gene switching.* Nat Genet, 2010. **42**(9): p. 742-4.
274. Norton, L.J., et al., *KLF1 directly activates expression of the novel fetal globin repressor ZBTB7A/LRF in erythroid cells.* Blood Advances, 2017. **1**(11): p. 685-692.
275. Narisawa-Saito, M. and T. Kiyono, *Basic mechanisms of high-risk human papillomavirus-induced carcinogenesis: roles of E6 and E7 proteins.* Cancer Sci, 2007. **98**(10): p. 1505-11.
276. Liu, N., et al., *Direct Promoter Repression by BCL11A Controls the Fetal to Adult Hemoglobin Switch.* Cell, 2018. **173**(2): p. 430-442 e17.
277. Moir-Meyer, G., et al., *Robust CRISPR/Cas9 Genome Editing of the HUDEP-2 Erythroid Precursor Line Using Plasmids and Single-Stranded Oligonucleotide Donors.* Methods and Protocols, 2018. **1**(3): p. 28.
278. Paquet, D., et al., *Efficient introduction of specific homozygous and heterozygous mutations using CRISPR/Cas9.* Nature, 2016. **533**(7601): p. 125-9.
279. Komor, A.C., et al., *Programmable editing of a target base in genomic DNA without double-stranded DNA cleavage.* Nature, 2016. **533**(7603): p. 420-4.
280. Nishida, K., et al., *Targeted nucleotide editing using hybrid prokaryotic and vertebrate adaptive immune systems.* Science, 2016. **353**(6305): p. aaf8729-aaf8729.
281. Gaudelli, N.M., et al., *Programmable base editing of A*T to G*C in genomic DNA without DNA cleavage.* Nature, 2017. **551**(7681): p. 464-471.
282. Qiu, J., et al., *Distinct splicing signatures affect converged pathways in myelodysplastic syndrome patients carrying mutations in different splicing regulators.* RNA, 2016. **22**(10): p. 1535-49.
283. Pellagatti, A., et al., *Impact of spliceosome mutations on RNA splicing in myelodysplasia: dysregulated genes/pathways and clinical associations.* Blood, 2018. **132**(12): p. 1225-1240.
284. Mettananda, S., et al., *Editing an alpha-globin enhancer in primary human hematopoietic stem cells as a treatment for beta-thalassemia.* Nat Commun, 2017. **8**(1): p. 424.
285. Baron, M.H. and T. Maniatis, *Regulated expression of human alpha- and beta-globin genes in transient heterokaryons.* Molecular and Cellular Biology, 1991. **11**(3): p. 1239-1247.
286. Vanuytsel, K., et al., *Induced pluripotent stem cell-based mapping of beta-globin expression throughout human erythropoietic development.* Blood Adv, 2018. **2**(15): p. 1998-2011.
287. Parra, M., et al., *An important class of intron retention events in human erythroblasts is regulated by cryptic exons proposed to function as splicing decoys.* RNA, 2018. **24**(9): p. 1255-1265.
288. Shiozawa, Y., et al., *Aberrant splicing and defective mRNA production induced by somatic spliceosome mutations in myelodysplasia.* Nature Communications, 2018. **9**(1).

289. McCullough, A.J. and S.M. Berget, *G triplets located throughout a class of small vertebrate introns enforce intron borders and regulate splice site selection.* Mol Cell Biol, 1997. **17**(8): p. 4562-71.
290. Zhu, J. and A.R. Krainer, *Pre-mRNA splicing in the absence of an SR protein RS domain.* Genes Dev, 2000. **14**(24): p. 3166-78.
291. Graveley, B.R., K.J. Hertel, and T. Maniatis, *The role of U2AF35 and U2AF65 in enhancer-dependent splicing.* RNA, 2001. **7**(6): p. 806-18.
292. Voith von Voithenberg, L., et al., *Recognition of the 3' splice site RNA by the U2AF heterodimer involves a dynamic population shift.* Proc Natl Acad Sci U S A, 2016. **113**(46): p. E7169-E7175.
293. Coulon, A., et al., *Kinetic competition during the transcription cycle results in stochastic RNA processing.* eLife, 2014. **3**.
294. Tanikawa, M., et al., *The spliceosome U2 snRNP factors promote genome stability through distinct mechanisms; transcription of repair factors and R-loop processing.* Oncogenesis, 2016. **5**(12): p. e280-e280.
295. Chen, L., et al., *The Augmented R-Loop Is a Unifying Mechanism for Myelodysplastic Syndromes Induced by High-Risk Splicing Factor Mutations.* Mol Cell, 2018. **69**(3): p. 412-425 e6.
296. Nguyen, H.D., et al., *Spliceosome Mutations Induce R Loop-Associated Sensitivity to ATR Inhibition in Myelodysplastic Syndromes.* Cancer Res, 2018. **78**(18): p. 5363-5374.
297. Zhao, Y., et al., *Real-Time Detection Reveals Responsive Cotranscriptional Formation of Persistent Intramolecular DNA and Intermolecular DNA:RNA Hybrid G-Quadruplexes Stabilized by R-Loop.* Anal Chem, 2017. **89**(11): p. 6036-6042.
298. Tous, C. and A. Aguilera, *Impairment of transcription elongation by R-loops in vitro.* Biochem Biophys Res Commun, 2007. **360**(2): p. 428-32.
299. Belotserkovskii, B.P., et al., *Mechanisms and implications of transcription blockage by guanine-rich DNA sequences.* Proc Natl Acad Sci U S A, 2010. **107**(29): p. 12816-21.
300. Broxson, C., J. Beckett, and S. Tornaletti, *Transcription Arrest by a G Quadruplex Forming-Trinucleotide Repeat Sequence from the Human c-myb Gene.* Biochemistry, 2011. **50**(19): p. 4162-4172.
301. Belotserkovskii, B.P., J.H. Soo Shin, and P.C. Hanawalt, *Strong transcription blockage mediated by R-loop formation within a G-rich homopurine-homopyrimidine sequence localized in the vicinity of the promoter.* Nucleic Acids Res, 2017. **45**(11): p. 6589-6599.
302. Sanz, L.A., et al., *Prevalent, Dynamic, and Conserved R-Loop Structures Associate with Specific Epigenomic Signatures in Mammals.* Mol Cell, 2016. **63**(1): p. 167-78.
303. Nguyen, D.T., et al., *The chromatin remodelling factor ATRX suppresses R-loops in transcribed telomeric repeats.* EMBO Rep, 2017. **18**(6): p. 914-928.
304. Abdel-Wahab, O., et al., *Deletion of Asxl1 results in myelodysplasia and severe developmental defects in vivo.* J Exp Med, 2013. **210**(12): p. 2641-59.
305. Wang, J., et al., *Loss of Asxl1 leads to myelodysplastic syndrome-like disease in mice.* Blood, 2014. **123**(4): p. 541-53.
306. Inoue, D., et al., *Myelodysplastic syndromes are induced by histone methylation-altering ASXL1 mutations.* J Clin Invest, 2013. **123**(11): p. 4627-40.

307. Guo, Y., et al., *Reduced BAP1 activity prevents ASXL1 truncation-driven myeloid malignancy in vivo.* Leukemia, 2018. **32**(8): p. 1834-1837.
308. Gu, B., D. Eick, and O. Bensaude, *CTD serine-2 plays a critical role in splicing and termination factor recruitment to RNA polymerase II in vivo.* Nucleic Acids Res, 2013. **41**(3): p. 1591-603.
309. Koga, M., M. Hayashi, and D. Kaida, *Splicing inhibition decreases phosphorylation level of Ser2 in Pol II CTD.* Nucleic Acids Res, 2015. **43**(17): p. 8258-67.
310. Li, Q., et al., *Analysis of the large inactive P-TEFb complex indicates that it contains one 7SK molecule, a dimer of HEXIM1 or HEXIM2, and two P-TEFb molecules containing Cdk9 phosphorylated at threonine 186.* J Biol Chem, 2005. **280**(31): p. 28819-26.
311. Dulac, C., et al., *Transcription-dependent association of multiple positive transcription elongation factor units to a HEXIM multimer.* J Biol Chem, 2005. **280**(34): p. 30619-29.
312. Prasanth, K.V., et al., *Nuclear organization and dynamics of 7SK RNA in regulating gene expression.* Mol Biol Cell, 2010. **21**(23): p. 4184-96.
313. Ji, X., et al., *SR proteins collaborate with 7SK and promoter-associated nascent RNA to release paused polymerase.* Cell, 2013. **153**(4): p. 855-68.
314. Yang, Z., et al., *Recruitment of P-TEFb for stimulation of transcriptional elongation by the bromodomain protein Brd4.* Mol Cell, 2005. **19**(4): p. 535-45.
315. Itzen, F., et al., *Brd4 activates P-TEFb for RNA polymerase II CTD phosphorylation.* Nucleic Acids Res, 2014. **42**(12): p. 7577-90.
316. Fujinaga, K., et al., *Visualization of positive transcription elongation factor b (P-TEFb) activation in living cells.* J Biol Chem, 2015. **290**(3): p. 1829-36.
317. Yu, M., et al., *RNA polymerase II-associated factor 1 regulates the release and phosphorylation of paused RNA polymerase II.* Science, 2015. **350**(6266): p. 1383-6.
318. Vos, S.M., et al., *Structure of activated transcription complex Pol II-DSIF-PAF-SPT6.* Nature, 2018. **560**(7720): p. 607-612.
319. Devaiah, B.N., et al., *BRD4 is an atypical kinase that phosphorylates serine2 of the RNA polymerase II carboxy-terminal domain.* Proc Natl Acad Sci U S A, 2012. **109**(18): p. 6927-32.
320. Ciferri, C., et al., *Molecular architecture of human polycomb repressive complex 2.* eLife, 2012. **1**.
321. Figueiredo, M.S., et al., *Acquired hemoglobin H disease in a patient with aplastic anemia evolving into acute myeloid leukemia.* Sao Paulo Med J, 2004. **122**(6): p. 273-5.
322. Rinaldi, C.R., et al., *Acquired Hb H disease associated with elevated Hb F level in patient affected by primary myelofibrosis.* Ann Hematol, 2010. **89**(8): p. 827-8.
323. Fluhr, S., et al., *Epigenetic dysregulation of the erythropoietic transcription factor KLF1 and the β-like globin locus in juvenile myelomonocytic leukemia.* Epigenetics, 2017. **12**(8): p. 715-723.
324. Sankaran, V.G., J. Xu, and S.H. Orkin, *Advances in the understanding of haemoglobin switching.* Br J Haematol, 2010. **149**(2): p. 181-94.
325. Thein, S.L., *Molecular basis of beta thalassemia and potential therapeutic targets.* Blood Cells Mol Dis, 2018. **70**: p. 54-65.

326. Martyn, G.E., et al., *Natural regulatory mutations elevate the fetal globin gene via disruption of BCL11A or ZBTB7A binding.* Nature Genetics, 2018. **50**(4): p. 498-503.
327. Alter, B.P., *Fetal erythropoiesis in stress hematopoiesis.* Exp Hematol, 1979. **7 Suppl 5**: p. 200-9.
328. Stamatoyannopoulos, G., et al., *Hb F production in stressed erythropoiesis: observations and kinetic models.* Ann N Y Acad Sci, 1985. **445**: p. 188-97.
329. Emambokus, N., et al., *Progression through key stages of haemopoiesis is dependent on distinct threshold levels of c-Myb.* EMBO J, 2003. **22**(17): p. 4478-88.
330. Sandberg, M.L., et al., *c-Myb and p300 regulate hematopoietic stem cell proliferation and differentiation.* Dev Cell, 2005. **8**(2): p. 153-66.
331. Menzel, S., et al., *The HBS1L-MYB intergenic region on chromosome 6q23.3 influences erythrocyte, platelet, and monocyte counts in humans.* Blood, 2007. **110**(10): p. 3624-6.
332. Moody, R.R., et al., *Probing the interaction between the histone methyltransferase/deacetylase subunit RBBP4/7 and the transcription factor BCL11A in epigenetic complexes.* J Biol Chem, 2018. **293**(6): p. 2125-2136.
333. Yan, H., et al., *Developmental differences between neonatal and adult human erythropoiesis.* American Journal of Hematology, 2018. **93**(4): p. 494-503.
334. Trakarnsanga, K., et al., *An immortalized adult human erythroid line facilitates sustainable and scalable generation of functional red cells.* Nature Communications, 2017. **8**: p. 14750.