Unravelling the Globin Gene Switch Mechanism in Patients with Hereditary Persistence of Foetal Haemoglobin

Nikita Camilleri

July 2023

A dissertation submitted to the Faculty of Health Sciences in fulfilment of the requirements for the degree of Master of Science in Applied Biomedical Science at the University of Malta

Supervisor: Prof. Joseph Borg BSc, MSc, PhD

L-Università
ta' Malta

FACULTY/INSTITUTE/CENTRE/SCHOOL Faculty of Health Sciences

DECLARATIONS BY POSTGRADUATE STUDENTS

(a) Authenticity of Dissertation

I hereby declare that I am the legitimate author of this Dissertation and that it is my original work.

No portion of this work has been submitted in support of an application for another degree or qualification of this or any other university or institution of higher education.

I hold the University of Malta harmless against any third party claims with regard to copyright violation, breach of confidentiality, defamation and any other third party right infringement.

(b) Research Code of Practice and Ethics Review Procedures

I declare that I have abided by the University's Research Ethics Review Procedures. Research Ethics & Data Protection form code 7560-11012021.

As a Master's student, as per Regulation 77 of the General Regulations for University Postgraduate Awards 2021, I accept that should my dissertation be awarded a Grade A, it will be made publicly available on the University of Malta Institutional Repository.

To Elise Camilleri and Jake Bonnici for their unwavering support and belief in my abilities, and for the invaluable encouragement and love throughout this academic journey.

Acknowledgements

First and foremost, I would like to express my sincere gratitude to my supervisor Professor Joseph Borg PhD for his constant guidance, patience, and support. Furthermore, I extend my heartfelt gratitude to Dr. Josef Borg PhD for his invaluable assistance in data analysis and bioinformatics data processing. A special thank you goes to Professor Marieke von Lindern PhD, Dr. Emile van den Akker PhD, Ms. Carmen van der Zwaan and Ms. Kerly Fu from the Department of Haematopoiesis, Sanquin Research, Amsterdam, who warmly welcomed me into their department and provided essential support in conducting flow cytometry and mass spectrometry experiments.

I would like to thank Professor Alex E. Felice MD PhD of the Thalassaemia Clinic and Professor Richard Muscat Director of the Centre for Molecular Medicine and Biobanking at the University of Malta for granting me permission to conduct my research at the Laboratory of Molecular Genetics at the Department of Applied Biomedical Science, University of Malta. I am also grateful to Professor Godfrey Grech PhD for giving me permission to use the Luminex equipment. Additionally, I extend my appreciation to Dr. Melissa Formosa PhD and Mr. Tony Carbonaro for their assistance in blood collection, and to Ms. Elaine Fenech for her invaluable assistance in the laboratory practice. A special mention goes to all the families who participated in this study. Without their generous cooperation, this research would not have been possible. Finally, I am deeply grateful to my family, friends, and loved ones for their unwavering support and motivation throughout the entirety of this project.

This study was funded by the Tertiary Education Scholarships Scheme, financed by the Ministry for Education and Employment in Malta.

Abstract

Haemoglobinopathies are widely recognised as one of the most common monogenic diseases globally, representing a significant global health issue. Hereditary persistence of foetal hemoglobin (HPFH) is a benign genetic condition characterised by continuous production of high levels of foetal hemoglobin (HbF) throughout adulthood, resulting from disrupted globin gene switching. Clinical investigations and molecular findings have demonstrated that the presence of HPFH in conjunction with other haemoglobinopathies reduces the severity of associated symptoms, attributed to elevated levels of HbF. This study focused on three Maltese families, encompassing 11 individuals affected by HPFH due to a truncation mutation (p.K288X) in the *KLF1* gene and 11 healthy relatives serving as controls. The primary objective was to gain insights into the underlying genetic and molecular mechanisms involved in the regulation of globin gene switching. Whole genome sequencing (WGS) was performed using DNA extracted from 11 affected individuals and 9 healthy controls. A total of 205 unique variants following a dominant inheritance pattern were identified and found to be present in all affected individuals. All these variants were discovered to be located on chromosome 19 in close proximity to the *KLF1* gene. Furthermore, novel variants were discovered in the *LMO2* and *KLF1* genes, which potentially contribute to the onset of HPFH. A subset analysis focusing on four subjects from Fam F1, exhibiting the highest HbF levels (>3%), revealed variants in the *NLRP3* gene and in the *RPS9* gene when following autosomal dominant and recessive inheritance patterns, respectively. These variants likely account for the sustained elevation of HbF levels in Fam F1, despite carrying the same *KLF1* mutation as other families. Flow cytometry data analysis confirmed the role of KLF1 in the regulation of antigen expression, as individuals with HPFH exhibited reduced levels of BCAM, CD44, and P1 antigens on erythrocytes. Furthermore, analysis of globin gene mRNA expression, revealed that healthy controls had elevated mRNA levels of adult α-, β-globin genes, while individuals with HPFH exhibited higher mRNA levels of the fetal $^{A}\gamma$-globin gene. Proteomic analysis using mass spectrometry (MS) supported these findings and additionally identified 53 proteins significantly correlated with HbF levels in HPFH-affected subjects, suggesting their involvement in globin gene regulation. In conclusion, this study highlights the importance of adapting an integrative approach to understand the molecular mechanisms underlying KLF1 deficiency. The identification of potentially causal variants associated with HPFH, provides valuable insights into the onset of this condition. Further investigations involving functional work to confirm the precise impact of these variants, as well as the role of the identified proteins in the upregulation of HbF levels, can provide a more comprehensive understanding of the underlying genetic architecture of HPFH.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

α	Alpha
β	Beta
δ	Delta
ε	Epsilon
γ	Gamma
μ	Mu
θ	Theta
ζ	Zeta
AAF	Alternative allele frequency
ACH	Active chromatin hub
ACHE	Acetylcholinesterase
AF-488	Alexa Fluor™ 488
AFR	African population
AMR	American population
APC	Anti-human allophycocyanin
BAM file	Binary alignment and map file
BCAM	Basal cell adhesion molecule
BCL11A	B-cell lymphoma/leukemia 11A
bDNA	Branched DNA
BFU-E	Erythroid burst forming unit
bp	Base pairs
BPGM	Bisphosphoglycerate mutase
BSA	Bovine serum albumin
BWA	Burrows-Wheeler alignment
CBC	Complete blood count
CDA	Congenital dyserythropoietic anaemia
CO_2	Carbon dioxide
COUP-TFII	Chicken ovalbumin upstream promoter transcription factor II
CFU-E	Erythroid colony forming unit
CMP	Common myeloid progenitor cell
CRISPR/Cas 9	Clustered regularly interspaced short palindromic repeats/CRISPR-associated protein 9
CSNK1A1	Casein kinase I isoform alpha
DBA	Diamond-Blackfan anaemia

DIA	Data independent acquisition
DNase I	Deoxyribonuclease I
DNMT	DNA methyltransferase enzyme
DRED	Direct repeat erythroid-definitive complex
EAS	East Asian population
EPO	Erythropoietin
FACS	Fluorescence-activated cell sorting
Fe^{2+}	Ferrous iron atom
FOG1	Friend of GATA1
FOP	Friend of protein arginine methyltransferase I
FSC	Forward scatter
FSC-A	Forward scatter area
FSC-H	Forward scatter height
g	G-force or relative centrifugal force
g/dL	Grams per decilitres
GAPDH	Glyceraldehyde-3-phosphate dehydrogenase
GATA1	GATA binding protein 1
GATK	Genome analysis toolkit
gDNA	Genomic DNA
gMFI	Geometric mean fluorescence intensity
gnomAD	Genome aggregation database
GRCh37	Genome research consortium human build 37
GRCh38	Genome research consortium human build 38
GRN	Gene regulatory networks
GVCF	Genomic VCF
GVHD	Graft versus host disease
Hb	Haemoglobin
HbF	Foetal haemoglobin
HDAC	Histone deacetylase
HEP	Human erythroid progenitor cell
HLA	Human leukocyte antigen
HMBS	Porphobilinogen deaminase
HMG	High mobility group
HMIP	*HBS1L-MYB* intergenic polymorphism
HPFH	Hereditary persistence of foetal haemoglobin
HPLC	High performance liquid chromatography

HS	Hypersensitive site
HSC	Haematopoietic stem cell
HSCT	Haematopoietic stem cell transplantation
HU	Hydroxyurea
Indels	Small insertions/deletions
In(Lu)	Inhibitor of Lutheran blood group antigen
IP	Immunoprecipitation
IPB	Illumina® purification beads
K_2-EDTA	Dipotassium ethylenediaminetetraacetic acid
kb	Kilobases
KCNN4	Intermediate conductance calcium-activated potassium channel protein 4
kDA	Kilodalton
KLF1	Krüppel-like factor
LC	Liquid chromatography
LCR	Locus control region
LD	Linkage disequilibrium
LMO2	LIM domain only 2
LRLD	Long-range linkage disequilibrium
Lu	Lutheran blood group
mA	Milliamp
MAPK	Mitogen-activated protein kinase
MCH	Mean corpuscular haemoglobin
MCV	Mean corpuscular volume
MDS	Myelodysplastic syndrome
MEP	Megakaryocyte-erythroid progenitor cell
MFI	Median fluorescence intensity
mg/dl	Milligrams per decilitre
mRNA	Messenger RNA
MS	Mass spectrometry
m/z	Mass to charge ratio
Nan	Neonatal anaemia
NaOH	Sodium hydroxide
NF-E4	Nuclear factor erythroid 4
NFE	Non-Finnish European population
ng/µl	Nanograms per microlitre
NMD	Nonsense-mediated mRNA decay

NTC	No template control
NuRD	Nucleosome remodeling deacetylase
O_2	Oxygen
PBS	Phosphate-buffered saline
PCA	Principal component analysis
PCR	Polymerase chain reaction
PTM	Post-translational modifications
Q-score	Quality score
QTL	Quantitative trait loci
r	Correlation coefficient
RDW	Red cell distribution width
rEPO	Recombinant human erythropoietin hormone
RNAseq	RNA sequencing
ROS	Reactive oxygen species
rpm	Revolutions per minute
RUNX1	Runt-related transcription factor 1
SAPE	Streptavidin-conjugated R-Phycoerythrin
SCA	Sickle cell anaemia
SCD	Sickle cell disease
SDC	Sodium deoxycholate
SIRT2	NAD-dependent protein deacetylase sirtuin-2
SNP	Single nucleotide polymorphism
SOX6	Sex determining region Y-box 6
SSC	Side scatter
SWI/SNF	Switch/sucrose nonfermenting complex
TAE buffer	Tris-acetate-EDTA buffer
TCEP	Tris(2-carboxyethyl) phosphin
UROS	Uroporphyrinogen-III synthase
UTR	Untranslated region
UV	Ultraviolet light
V	Volts
VCF	Variant call format
VEP	Varriant effet predictor
WGR	Whole genome resequencing
WGS	Whole genome sequencing
ZFN	Zinc finger nucleases

1. Literature Review

1.1 Haemoglobin Structure and Function

Haemoglobin (Hb) is a specialised oxygen (O_2) transport protein found inside erythrocytes that has been described by some as a two-way respiratory carrier due to its ability to transport O_2 from the lungs to cells throughout the body, while also transporting carbon dioxide (CO_2) from the cells back to the lungs so it can be disposed of efficiently. Hb is a 65 kilodalton (kDA) tetramer composed of four polypeptide chains, two alpha (α) globin chains and two beta (β) globin chains, which are similar in size and structure, but are not identical (Figure 1.1) (Marengo-Rowe 2006; Storz, Opazo et al. 2013).

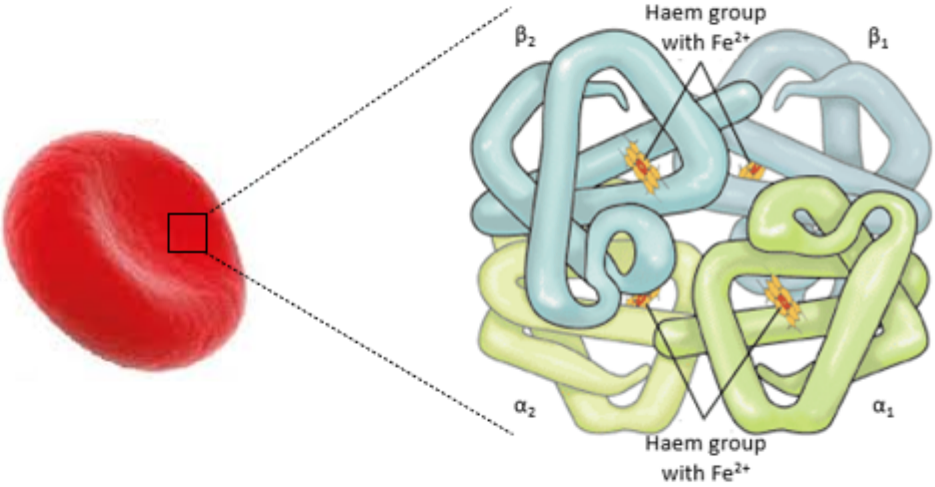


Figure 1.1: Erythrocyte and Haemoglobin structure, showing the characteristic biconcave structure of erythrocytes packed with Hb. Hb consists of 4 polypeptide chains (2 α- and 2 β- subunits) with a haem prosthetic group in the centre of each subunit (adapted from Encyclopaedia Britannica Inc. 2023).

Each of the four globin subunits has a large central space where a haem prosthetic moiety is bound. The haem group consists of an iron atom in the ferrous state (Fe^{2+}) held in the centre of a porphyrin ring by non-covalent bonds (Hsia 1998). The ferrous iron atom plays an important role in binding reversibly to CO_2 or O_2 ligands by covalent bonds, hence allowing the transport of four molecules of O_2 per red blood cell (Longeville, Stingaciu 2017; Schechter 2008).

The efficiency of Hb as an O_2 transport protein can be attributed to the ability of each of the four subunits of haem to communicate and interact directly during the process of oxygenation. This allows for cooperative binding, whereby the binding of an O_2 molecule to one haem subunit will induce tertiary conformational changes in other subunits, which facilitate the binding of further O_2 molecules to the remaining oxygen binding sites. Cooperativity between the α and β subunits of haem will also allow for the rapid and efficient unloading of O_2 to the tissues (Ahmed, Ghatge et al. 2020; Marengo-Rowe 2006; Storz, Opazo et al. 2013).

1.2 Globin Genes

The human globin genes are organised into two gene clusters which are located on separate chromosomes. The α-globin chains are coded for by genes grouped together in a segment of approximately 20 kilobases (kb), known as the α-globin gene cluster (Figure 1.2). This cluster of genes, located at the tip of the short arm of chromosome 16 (16p13.3), consists of three functional genes that are expressed during different stages of development (Liebhaber, Russel 1998).

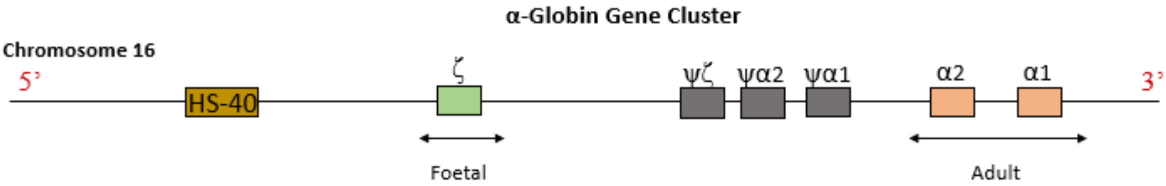


Figure 1.2: The α-Globin Gene Cluster located on chromosome 16, consists of three functional genes (ζ, α1 and α2), three pseudogenes (ψζ, ψα1 and ψα2) and the HS-40 regulatory element.

The zeta (ζ) globin gene, located nearest to the 5' region of the α-globin gene cluster, is the primary α-like globin gene that is expressed during the embryonic stage and is synthesised exclusively in the yolk sac. Expression of the ζ gene is silenced during the foetal and adult stages and is instead replaced by the expression of the α_1 and α_2 genes, which are paralogous genes having a high degree of homology but differing in abundance (Hardison 2012). The α-globin gene cluster also consists of three pseudogenes; ψζ, $\psi\alpha_1$ and $\psi\alpha_2$, which are homologous to functional genes but are not expressed due to inactivating mutations. The major regulatory element of the α-globin gene locus is known as the HS-40, which is an erythroid-specific deoxyribonuclease (DNase) I hypersensitive site (HS) located 40 kb upstream of the ζ-globin gene (Hua-bing, De-Pei et al. 2002).

The β-globin gene cluster is located along a segment of 60 kb on the short arm of chromosome 11 (11p) (Figure 1.3). It consists of five functional genes that are arranged in order of expression during development, having the embryogenic globin genes at the 5' end and the adult globin genes at the 3' end of the segment. The epsilon (ε) globin gene is the β-like globin gene expressed during embryogenesis. Two foetal gamma (γ) globin genes, $^{A}\gamma$ (alanine) and $^{G}\gamma$ (glycine), which differ by a single amino acid at position 136, are expressed during the foetal stages. Expression of the foetal γ-globin genes is replaced by expression of the delta (δ) and β-globin genes in the erythrocytes of adults. A single pseudogene, ψβ is also present in the β-globin gene cluster, and expression of this locus is controlled by the locus control region (LCR) (Hardison 2012, Ohls 2017).

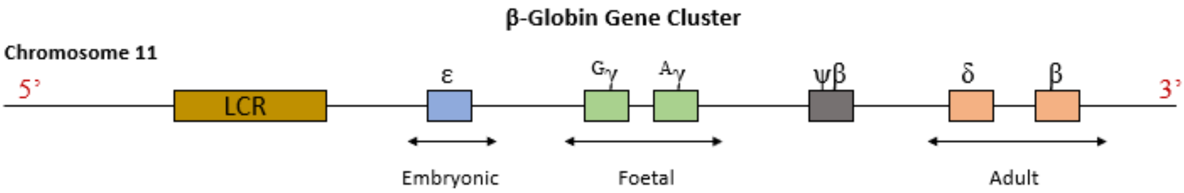


Figure 1.3: The β-Globin Gene Cluster located on chromosome 11, consists of five functional genes (ε, Aγ, Gγ, δ and β) arranged in order of developmental expression, one pseudogene (ψβ) and expression is under control of the LCR.

The expression of genes in both the α-globin and β-globin gene clusters is under strict regulation to ensure a balance in the production of α-globin and β-globin chains in the erythrocytes, which would otherwise lead to pathogenic phenotypes (Hardison 2012).

1.3 Haemoglobin Phenotypes

During distinct stages of human life, there is developmental timing of expression of the different globin genes which results in the production of seven different Hb isoforms having different O_2 affinities. The production of the different Hb isoforms is characterised by two developmental class switches during the embryonic and foetal stages, accompanied by a change in the site of erythropoiesis (Figure 1.4) (Storz, Opazo et al. 2013).

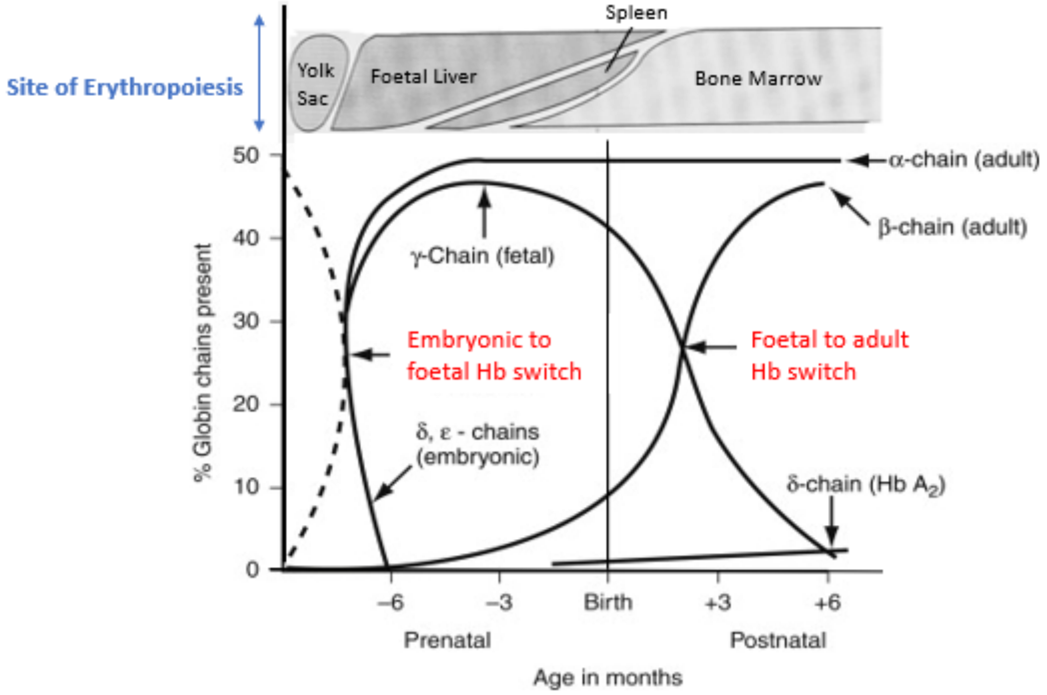


Figure 1.4: Developmental expression of different Hb genes and the site of erythropoiesis, showing the two main Hb class switching events occurring during embryoing and foetal life (adapted from Old 2013).

In the first few weeks following conception, three different forms of embryonic Hb will start to be expressed in the primitive erythrocytes developed in the yolk sac (Perry, Soreq 2002). The embryonic Hb include Hb Gower-1 ($\zeta_2\varepsilon_2$), Hb Gower-2 ($\alpha_2\varepsilon_2$) and Hb Portland

($\zeta_2\gamma_2$). The first major Hb switching event occurs around 6 weeks post-conception, during which the expression of ζ- and ε-globin genes is silenced, while there is transcriptional activation of both γ- and α-globin genes, leading to the production of foetal Hb (HbF : $\alpha_2\gamma_2$) (Peschle, Mavilio et al. 1985). The site of erythropoiesis also migrates from the yolk sac to the foetal liver and spleen. HbF is the major Hb isoform expressed during the course of gestation and postnatal life, and is coded for by both foetal $^{A}\gamma$- and $^{G}\gamma$- globin genes (Liebhaber, Russel 1998; Wilber, Nienhuis et al. 2011).

The second Hb class switch occurs following birth and is completed throughout infancy. It involves a decline in the synthesis of HbF, coupled with a complementary rise in the expression of the previously silent δ- and β-globin genes. This leads to an increased synthesis of the major adult haemoglobin HbA ($\alpha_2\beta_2$), and HbA$_2$ ($\alpha_2\delta_2$) in minor quantities. The second developmental class switch causes a shift in the site of erythropoiesis to the bone marrow which becomes the predominant haematopoietic organ (Perry, Soreq 2002). At around two years of age, HbA comprises approximately 95% of the total Hb in adults, HbA$_2$ is present in minor quantities of 2-3%, and almost all of the HbF synthesis ceases such that only 1% or less persists in the mature erythrocytes of healthy individuals (Manning, Russell et al. 2007; Sankaran, Xu et al. 2010).

1.4 Transcriptional Regulation of Globin Gene Expression

The process of developmental Hb gene switching and stage specific expression is regulated by transcription factors as well as other *cis*-acting elements, which interact with regulatory regions of the α- and β-globin gene clusters to ensure a balance in the production of α-globin and β-globin chains. Regulatory regions, including promoters, enhancers and silencers, are characterised by the presence of several DNase I HS, known as the LCR.

Regions located in proximity to the globin genes are important for the correct initiation of transcription, while more distant regions are required for maximal expression of the genes (Cao, Moi 2002).

The promoter regions of all globin genes have three major regulatory elements; a TATA box, a CAAT box and a CACCC box, with minor sequence variations for each globin promoter contributing to uniqueness. Several different transcription factors interact with the DNA sequences present in the promoter regions, resulting in the formation of DNA protein complexes which activate or repress stage specific gene expression (Hardison 2012). Important transcription factors include the Runt-related Transcription Factor 1 (RUNX1), GATA binding protein 1 (GATA1), Krüppel-Like Factor (KLF1), Sex Determining Region Y-box 6 (SOX6), Nuclear Factor Erythroid 4 (NF-E4) and B-cell lymphoma/leukemia 11A (BCL11A).

1.4.1 B-cell lymphoma/leukemia 11A (BCL11A)

The BCL11A transcription factor is a master regulator of γ-globin gene expression in adult erythroid progenitor cells. It participates in the second developmental switch by silencing the γ-globin gene and repressing the transcription of HbF, thus favouring the increased expression of HbA (Sankaran, Menne et al. 2008; Sankaran, Xu et al. 2010).

The BCL11A transcription factor occupies the β-globin gene locus in primary adult human erythroid progenitor (HEP) cells. It binds to distal regulatory elements in the LCR, including the third DNase I HS (HS3) site and to the intergenic region between the γ- and δ-globin genes (Sankaran, Menne et al. 2008). Together with the transcription factor SOX6 which binds to proximal γ-globin gene promoters, it reconfigures the β-globin gene locus and prevents interactions between the γ-globin genes and the LCR by a process known as chromatin looping, hence resulting in repression of the γ-globin genes. SOX6 is a member of SRY-related high-mobility group (HMG) box transcription factors that is co-expressed with BCL11A during erythroid development. It also co-occupies the β-globin gene cluster *in vivo.* In the absence of BCL11A, such as in foetal erythroid development, reconfiguration of β-globin gene occurs favouring interactions between the γ-globin genes and the LCR (Xu, Sankaran et al. 2010).

Therefore, BCL11A will mediate γ-globin gene silencing by long range interactions within the β-globin gene locus causing chromosomal loop formation, and by local interactions with SOX6 proteins at the proximal promotors of the γ-globin genes (Figure 1.5). The role of both BCL11A and SOX6 in silencing of the γ-globin genes was demonstrated in combined knockdown experiments whereby developing adult human erythroblasts presented with a robust increase in HbF levels (Sankaran, Xu et al. 2009; Xu, Sankaran et al. 2010). The activity of BCL11A as a transcriptional repressor of γ-globin expression in human erythroid progenitors is also dependant on the erythroid transcription factors GATA1 and friend of GATA1 (FOG1), and the nucleosome remodeling deacetylase (NuRD) repressor complex (Figure 1.5). This association was confirmed by immunoprecipitation (IP) (Sankaran, Menne et al. 2008).

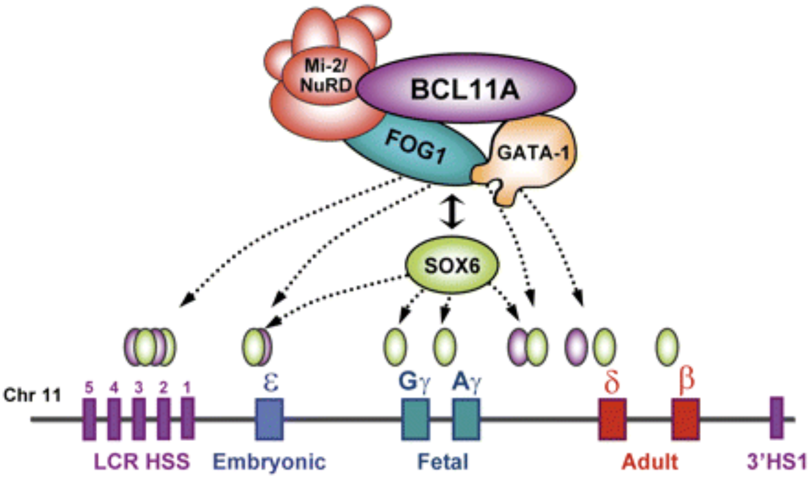


Figure 1.5: BCL11A mediates silencing of γ-globin genes by physically interacting with GATA1, FOG1, Mi-2/NuRD complexes and SOX6, causing long range interaction within the β-globin gene through chromosomal looping and also by local interactions with promoters of γ-globin genes (adapted from Xu, Sankaran et al. 2010).

1.4.2 Erythroid Krüppel-Like Factor (KLF1)

KLF1 is a zinc finger protein and one of the most specific erythroid transcription factors discovered. It plays an important role in the second Hb developmental switch and causes a decrease in γ-globin gene expression by a dual mechanism. KLF1 interacts directly with the CACCC box of the β-globin gene promoter, resulting in activation of the β-globin gene (Donze, Townes et al. 1995). KLF1 also indirectly represses the expression of the γ-globin gene by binding to the BCL11A promoter and activating the expression of the *BCL11A* gene.

During foetal development when KLF1 levels are low, protein complexes will bind to the CACCC region in the γ-globin gene promoter allowing for interactions between the LCR and γ-globin genes. KLF1 levels are insufficient to activate BCL11A expression and to repress γ-globin gene expression. The levels of KLF1 increase during adult erythroid development, resulting in the formation of complexes that preferentially bind to the CACCC box of the β-globin gene promoter. This allows the β-globin genes to interact with the LCR, and hence their expression increases. The elevated KLF1 levels will also upregulate the expression of the BCL11A gene which leads to silencing of the γ-globin genes (Figure 1.6) (Zhou, Liu et al. 2010).

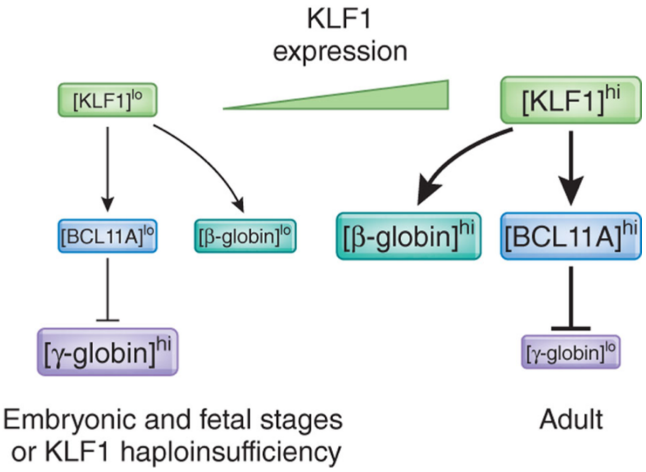


Figure 1.6: KLF1 regulates globin gene switching by directly acting on the β-globin gene and indirectly by activating BCL11A expression. Low levels of KLF1 (left), result in low levels of BCL11A and β-globins and elevated levels of γ-globin. Expression of *KLF1* increases during adult stage (right), causing repression of γ-globin genes due to upregulation of BCL11A and β-globin genes (adapted from Bieker 2010).

The role of KLF1 in Hb switching and its effect on BCL11A was discovered after performing genome-wide single nucleotide polymorphism (SNP) scan, followed by linkage analysis on a Maltese family that presented with hereditary persistence of foetal haemoglobin (HPFH). A nonsense mutation was identified in the *KLF1* gene which ablated the DNA binding domain resulting in *KLF1* haploinsufficiency and HPFH (Borg, Papadopoulos et al. 2010).

Functional work involving the knockdown of *KLF1* in human and mice erythroid progenitor cells highlight the significant role that KLF1 has in the foetal to adult globin switch. Knockdown of *KLF1* severely inhibits expression of BCL11A, consequently resulting in upregulation of the γ-globin gene expression (Wilber, Nienhuis et al. 2011; Zhou, Liu et al. 2010).

1.4.3 Other Transcription Factors

Other modifier transcription factors that play a role in globin gene switching include the TR2/TR4 direct repeat erythroid-definitive (DRED) complex and the chicken ovalbumin upstream promoter transcription factor II (COUP-TFII). Both TR2/TR4 DRED and COUP-TFII will bind to direct repeats in the γ-globin gene promoters, such as the CAAT element, resulting in silencing of the γ-globin genes (Filipe, Li et al. 1999; Omori, Tanabe et al. 2005). The transcription factor NF-E4 interacts with the ubiquitous transcription factor CP2 to activate transcription of γ-globin genes (Jane, Nienhuis et al. 1995), while the friend of protein arginine methyltransferase I (FOP) is suggested to silence the foetal globin genes, since knockdown of FOP in human erythroblasts results in increased production of HbF (Van Dijk, Gillemans et al. 2010; Wilber, Nienhuis et al. 2011).

1.4.4 Locus Control Region and Chromatin Architecture

The LCR is comprised of DNA sequences located approximately 40-60 kb upstream of the globin genes and is responsible for regulation of globin gene expression. The β-globin gene LCR consists of four erythroid specific DNase I HS and a fifth DNase I HS-5 located further upstream. DNase I HS refers to a combination of several DNA motifs that allow for DNA-protein interactions and facilitate access to several transcription factors and gene regulators by modelling the chromatin structure. The HSs in the LCR will act as a holocomplex that interacts directly with individual globin genes during erythroid development. The dynamic competition between the different globin genes and the LCR will give rise to specific gene expression during the different stages of human development. The importance of the LCR in regulation of globin gene expression was displayed in cases of γδβ-thalassemias, where naturally occurring deletions in the LCR results in silencing of all β-globin genes (Stamatoyannopoulos, Grosveld 2001; Wilber, Nienhuis et al. 2011).

Transcriptional regulation of globin gene expression is nowadays being regarded as a three-dimensional spatial organisation of the genome mediated by transcription factors, rather than a mechanism that involves regulation of single genes in isolation (Hattangadi, Wong et al. 2011; Schoenfelder, Sexton et al. 2009). A combination of several transcription factors like KLF1, GATA1 and FOG1 are necessary to mediate long range gene expression by looping of chromosome sequences outwards of their chromosomal territory. Chromosomal looping will therefore bring distal regulatory elements, like the LCR, into physical proximity to specific globin promoters, resulting in activation or silencing of gene expression (Figure 1.7) (Drissen, Palstra et al. 2004; Kadauke, Blobel 2009).

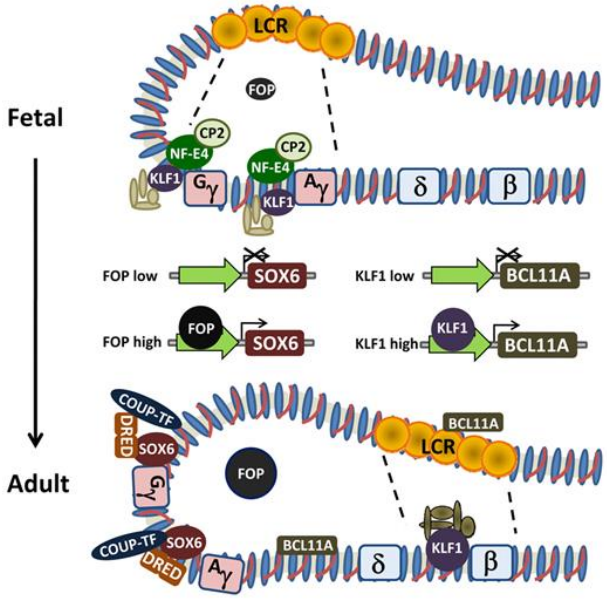


Figure 1.7: Chromosomal looping model and interaction of transcription factors with the LCR of globin genes during foetal stages (top) and during adult stages (bottom) (adapted from Wilber, Nienhuis et al. 2011).

Chromosome modification and remodelling is also mediated by protein complexes known as ATP-dependent nucleosome remodelers, which include the switch/sucrose nonfermenting (SWI/SNF) complex and the NuRD complex. The SWI/SNF protein complexes consist of large protein subunits that utilise ATP energy to generate an open chromatin conformation at promoters and enhancers. This promotes transcriptional activation of genes by making the DNA more accessible to binding with transcription factors (Bracken, Brien et al. 2019; Kwon, Imbalzano et al. 1994). The NuRD remodelers will oppose the activity of SWI/SNF complexes by mediating the formation of repressive chromatin through histone deacetylation and placement of nucleosomes at regulatory elements (Bracken, Brien et al. 2019).

The zinc finger transcription factor Ikaros is also crucial for efficient globin gene transcription by specifically binding to globin gene promoters and by assembling the β-globin genes into an active chromatin hub (ACH). The ACH is formed by long distance DNA looping between the LCR and the individual genes of the β-globin locus, hence facilitating interaction between the genes and their regulatory elements (Keys, Tallack et al. 2008; O’Neill, Schoetz et al. 2000).

1.5 Erythropoiesis

Erythropoiesis is the commitment of haematopoietic stem cells (HSC) to the erythroid lineage and their differentiation into mature end-point stage erythrocytes. This complex multistep process, which occurs over a span of approximately two weeks, ensures the replacement of senescent erythrocytes by the production of sufficient numbers of terminally differentiated red blood cells. The process of erythropoiesis changes temporally and spatially during the different developmental stages of life (Section 1.3). Early erythropoiesis, known as primitive erythropoiesis, involves the production of large, nucleated erythroblasts formed from mesoderm cells within the yolk sac. Such early erythroblasts express the three different embryonic haemoglobins; Hb Gower-I ($\zeta_2\varepsilon_2$), Hb Gower-II ($\alpha_2\varepsilon_2$) and Hb Portland ($\zeta_2\gamma_2$) (Perry, Soreq 2002).

During late foetal life, the early progenitor cells will migrate through the blood circulation to the foetal liver and eventually seed the bone marrow, where definitive erythropoiesis is established (Figure 1.8). The immature erythroid progenitor cells in definitive erythropoiesis are lineage committed and will result in the generation of well-defined enucleated erythrocytes that have a smaller size than the primitive erythroblasts (Palis 2014; Perry, Soreq 2002).

During the first steps of erythropoiesis, the pluripotent HSCs will divide and differentiate into the common myeloid progenitor (CMP) cells which are committed to the myeloid lineage. The CMPs will themselves differentiate into the megakaryocyte-erythroid progenitor (MEP) cells and eventually into the erythroid burst forming units (BFU-E) and erythroid colony forming units (CFU-E) (Figure 1.8). Both the BFU-E and CFU-E are immature progenitor cells committed to the erythroid lineage (Zivot, Lipton et al. 2018).

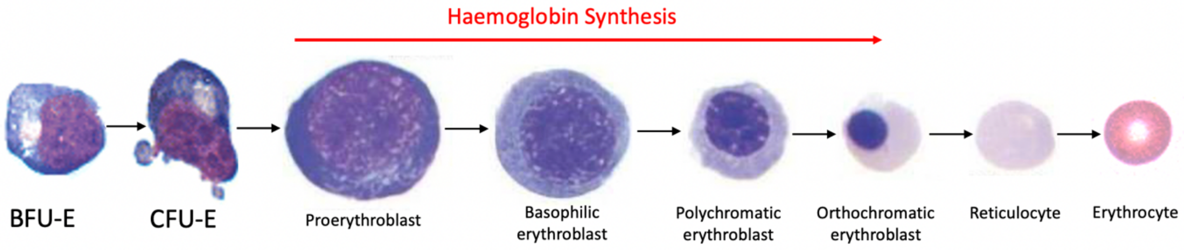


Figure 1.8: The Stages of Definitive Erythropoiesis, showing the series of cellular differentiations that each cell undergoes in the bone marrow, until the final mature erythrocytes are released into the blood stream (adapted from Song, Huang et al. 2021).

The second phase of erythropoiesis is known as the precursor stage and involves five rapid cell divisions from the CFU-E to the proerythroblasts, basophilic erythroblasts, polychromatic erythroblasts, orthochromatic erythroblasts, until the final reticulocytes are produced (Figure 1.8). Each step of cell division is accompanied by a progressive reduction in cellular size and in RNA content, and the gradual accumulation of Hb. Furthermore, erythroblasts will exhibit nuclear condensation and eventually enucleation which is the end result of precursor cell maturation (Dzierzak, Philipsen 2013).

Maturation of the reticulocytes to the terminally differentiated erythrocytes is the final stage of erythropoiesis, during which the reticulocytes will lose any residual cytoplasmic organelles, decrease the mean corpuscular volume (MCV) and acquire the characteristic biconcave shape through cytoskeletal remodelling. The definitive and mature erythrocytes are then released into circulation within 24 hours where they remain for approximately 120 days, after which they are broken down in the liver and spleen by resident macrophages (Palis 2014).

1.5.1 Regulation of Erythropoiesis

Both primitive and definitive erythropoiesis are controlled by a number of factors to ensure a balance between red blood cell synthesis and breakdown. One such factor is the hormone erythropoietin (EPO) which is synthesised by the kidneys and is essential for the survival and proliferation of BFU-E and CFU-E progenitor cells. EPO inhibits the apoptosis of the erythroid progenitor cells and upregulates the expression of globin genes (Zivot, Lipton et al. 2018). The precursor stage of erythropoiesis, which involves the production of Hb, is highly dependent on the supply of iron. Iron regulates the synthesis of globin chains at the transcriptional and translational level. Folic acid and Vitamin B12 are also necessary factors for erythropoiesis.

Furthermore, transcription factors play a critical role in driving lineage specific cell maturation. GATA1 transcription factor activates erythroid specific genes in both primitive and definitive erythropoiesis leading to erythroid differentiation. Its key role in erythroid commitment is demonstrated by the failure of erythroid precursors to reach maturation in *GATA*1 null mice (Pevny, Simon et al. 1991). The transcription factor MYB is abundantly expressed in the immature erythroblasts during primitive erythropoiesis. Its expression must then be downregulated to allow for terminal differentiation of the immature progenitor cells (Gonda, Metcalf 1984). Knockout experiments of *MYB* in mice exhibited severely impaired definitive erythropoiesis but normal early cellular expansion in homozygous null mice. This resulted in their early death due to failure to transition to adult stage erythropoiesis (Lieu, Reddy 2009). KLF1 transcription factor is important in both primitive and definitive erythropoiesis by controlling the expression of cytoskeletal proteins found in erythroid cells. It also regulates the expression of adult globin genes in late erythropoiesis. Therefore, an inadequate supply of any one of the transcription or growth factors will result in anaemia (Beckman, Silberstein et al. 2010; Perry, Soreq 2002).

1.6 Haemoglobinopathies

Haemoglobinopathies is an umbrella term used to refer to an inherited group of Hb disorders caused by mutations in genes coding for the globin component of Hb. They are considered as one of the most common monogenic diseases worldwide and one of the world's major health problems. Approximately 5% of the world's population are carriers for abnormal Hb and thalassaemia (Modell, Darlison 2008). These genetic disorders are classified into two main groups, depending on whether the globin gene mutations result in an abnormal Hb synthesis or an abnormal Hb structure. Thalassaemia syndromes are caused by mutations or deletions in the globin genes that result in the reduced or absent synthesis of α-globin or β-globin chains. Structural Hb variants refer to the group of abnormal Hb produced due to mutations in the globin genes. Such mutations result in an altered amino acid sequence of the α-globin and/or β-globin chains. The most clinically important structural Hb variants are the HbS, HbC and HbE. Disorders of Hb were originally identified in Mediterranean countries, as well as large parts of Africa and Asia, however due to international migration they have spread all over the world (Kohne 2011).

1.6.1 Sickle Cell Disease

Sickle cell disease (SCD) is a monogenic blood disorder that is caused by a single base pair point mutation (GAG to GTG) in the β-globin gene. This mutation results in the substitution of the sixth amino acid, from a glutamic acid to a valine, in the β-globin chain of Hb, forming the structural Hb variant referred to as HbS (Hoban, Orkin et al. 2016).

SCD was first reported in 1910 (Herrick 1910), and predominantly affects people of sub-Saharan African descent. Inheritance of homozygous HbS is referred to as sickle cell anaemia (SCA), which is the most common form of SCD. Co-inheritance of HbS and HbC can also occur and result in the structural variant HbSC. The pathophysiology of SCD is that erythrocytes with the variant HbS will undergo polymerisation and become rigid in environments with low O_2. The half life of these erythrocytes decreases from around 120 days in circulation to 10-20 days in circulation since they become more prone to haemolysis (Sebastiani, Nolan et al. 2007). The dense sickled erythrocytes will obstruct the microcirculation causing a vaso-occlusive crisis, while also restricting the blood flow to the organs (Figure 1.9). This leads to tissue ischaemia, oedema and infarction. SCD is hence a multi-organ multisystem disorder that results in acute and chronic complications (De Montalembert 2002; Manwani, Frenette 2013).

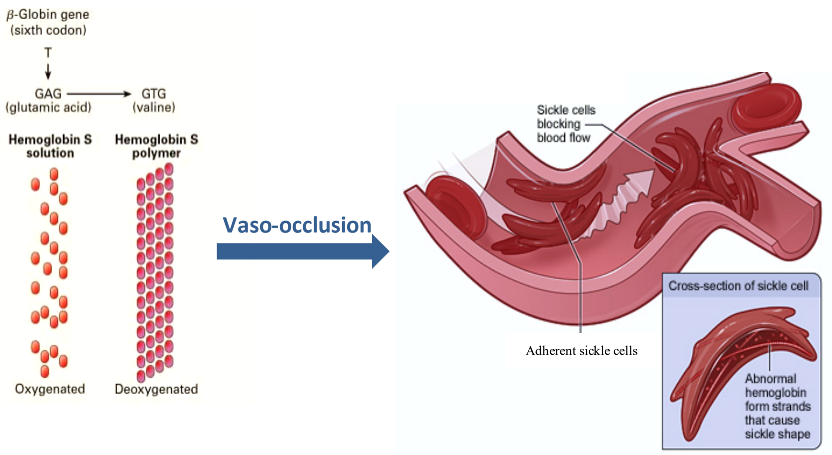

β-Globin gene (sixth codon)
T
GAG (glutamic acid)
GTG (valine)
Hemoglobin S solution
Hemoglobin S polymer
Oxygenated
Deoxygenated
Vaso-occlusion
Sickle cells blocking blood flow
Adherent sickle cells
Cross-section of sickle cell
Abnormal hemoglobin form strands that cause sickle shape

Figure 1.9: The Pathophysiology of sickle cell disease, where a single gene mutation results in the defective HbS that polymerisations when exposed to low oxygen levels forming sickle cells. The adherent sickle cells will occlude microvasculature causing vaso-occlusive crisis (adapted from National Heart Lung and Blood Institute 2022).

Individuals with SCD will start manifesting clinical features of haemolytic anaemia and vaso-occlusive episodes accompanied by intense pain at around 6 months of age, after the second Hb switch to adult HbA. Other complications include increased susceptibility to infections and enlarged spleen due to increased haemolysis. Hence, a multidisciplinary approach is required to manage patients with SCD (Inusa, Hsu et al. 2019).

1.6.2 Thalassaemia

Thalassaemia is one of the most common inherited autosomal recessive haemoglobinopathy worldwide, having the highest incidence among Mediterranean populations, as well as sub-Saharan Africa and South Asian countries (Colah, Gorakshakar et al. 2010). This group of blood disorders is characterised by anomalies in the synthesis of one or more human globin chains, leading to a reduced or an abnormal production of Hb. The two main types of thalassaemia are α-thalassaemia and β-thalassaemia caused by a reduced or absent synthesis of the α- and β-globin chains respectively (Mehdi, Al Dahmash 2011). Approximately 70,000 births per year are reported to be affected with thalassaemia disorders, with 30,000 babies being born with β-thalassaemia (Mettananda, Higgs 2018). This amounts to a global number of 80 million individuals being carriers of β-thalassaemia, as suggested by scientific studies (De Sanctis, Kattamis et al. 2017).

The α-thalassaemia disorders are caused by one or more deletions of the α-globin genes found on chromosome 16. There are four different clinical scenarios of α-thalassaemia with varying degrees of severity, depending on the number of α-globin genes that are affected by loss of function, as shown in Table 1.1 (Harteveld, Higgs 2010). The first two clinical scenarios are the α-thalassaemia minima and the α-thalassaemia minor, which are known as the silent carrier and the α-thalassaemia trait respectively. The third scenario is the HbH disease caused by three inactive α-globin genes (Chui, Fucharoen et al. 2003), while the final and most severe clinical scenario is the Hb Bart's Hydrops fetalis caused by complete deletion of the four α-globin genes. Individuals with Hb Bart's thalassaemia have very severe anaemia that is fatal in utero or at birth if left untreated, due to the absence of any α-globin chains (Lorey, Charoenkwan et al. 2001).

Table 1.1: The four different clinical scenarios of α-Thalassaemia and associated phenotypes.

Common Genotype	Common Name	Phenotypic Expression
- α / α α	α-Thalassaemia minima	Silent carrier, asymptomatic
- - / α α or - α / - α	α-Thalassaemia minor	Trait, mild anaemia
- - / - α	HbH disease	Moderate to severe anaemia
- - / - -	Hb Barts	Hydrops fetalis, incompatible with life

β-thalassaemia is a group of Hb disorders caused by point mutations or deletions in the β-globin gene cluster found on chromosome 11, leading to a reduced or absent synthesis of the β-globin chains. Around 534 recessive mutations have been identified to date and are known to cause β-thalassaemia (https://globin.bx.psu.edu/hbvar/). The nature of the mutation and its effect on various stages of the gene expression will effect the degree of reduction in the globin chain synthesis. The reduced amount of β-globin chains or their absence, will result in the excess synthesis of unbound α-globin chains as a compensation. These unbound chains will precipitate in the erythroid precursors found in the bone marrow leading to their premature death, and hence ineffective erythropoiesis and anaemia (Sankaran, Orkin 2012). This is therefore the main pathophysiological mechanism in β-thalassaemia (Figure 1.10).

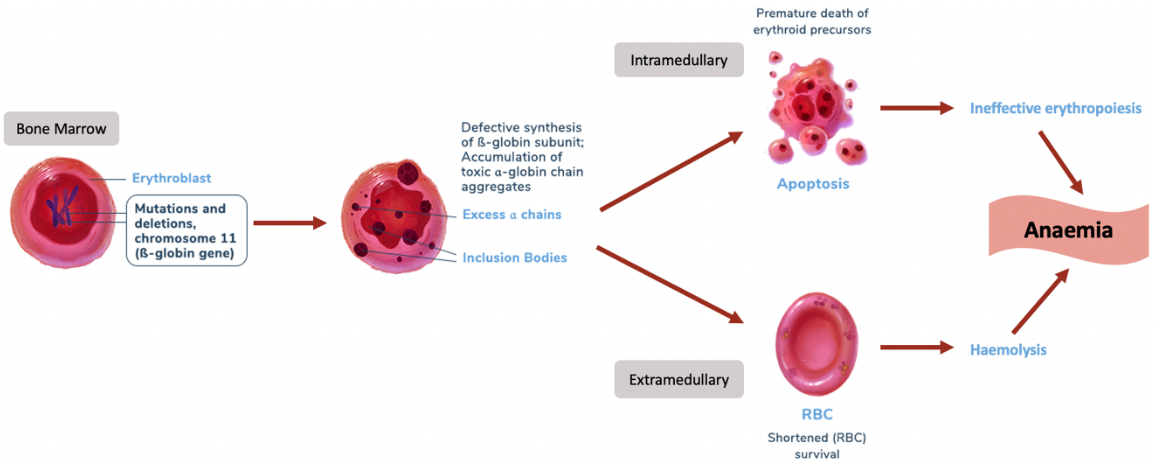

Bone Marrow
Erythroblast
Mutations and deletions, chromosome 11 (ß-globin gene)
Defective synthesis of ß-globin subunit; Accumulation of toxic α-globin chain aggregates
Excess α chains
Inclusion Bodies
Intramedullary
Premature death of erythroid precursors
Apoptosis
Ineffective erythropoiesis
Anaemia
Extramedullary
RBC
Shortened (RBC) survival
Haemolysis

Figure 1.10: The Pathophysiological Mechanism of β-thalassaemia, where the excess unbound α-globin chains form toxic aggregates inside the erythroid progenitor cells causing ineffective erythropoiesis and eventually anaemia (adapted from Rachmilewitz, Giardina 2011).

There are three main forms of β-thalassaemia; β-thalassaemia major (β^0) which is caused by complete absence of β-globin chains, β-thalassaemia intermedia (β^+) and β-thalassaemia minor (β^{++}), both caused by a mild to moderate reduction in the β-globin chains produced (Table 1.2). All forms of β-thalassaemia always present with hypochromia and microcytosis at the microscopic level. Both β-thalassaemia intermedia and β-thalassaemia minor are known as the carrier states. Patients will typically present with mild anaemia, commonly referred to a nontransfusion-dependent anaemia, and can even be completely asymptomatic (Fibach, Rachmilewitz 2017; Galanello, Origa 2010).

Table 1.2: The three different clinical scenarios of β-Thalassaemia and associated phenotypes.

Common Genotype	Common Name	Phenotypic Expression
β^0/β or β^+/β	β-Thalassaemia minor	Silent carrier, asymptomatic
β^0/β or β^+/β^+	β-Thalassaemia intermedia	Trait, mild anaemia, occasional transfusions
β^0/β^0	β-Thalassaemia major	Severe anaemia, transfusion dependent

Individuals suffering from β-thalassaemia major usually present with severe chronic anaemia (Hb levels <7g/dl) in the first six months to two years of age, following the second Hb class switch that causes silencing of the γ-globin genes (Thein 2012). The bone marrow compensates for the lack of β-globin chains by upregulating the synthesis of HbF even beyond the foetal stage. However, the HbF produced is unevenly distributed in the erythrocytes resulting in ineffective erythropoiesis with increased peripheral red cell haemolysis. Severe symptoms and complications will arise in patients with β-thalassaemia major if the anaemia is left untreated. This includes growth retardation in children, jaundice, hepatosplenomegaly, skeletal deformities due to expansion of the bone marrow, and may even be lethal in severe cases (Kohne 2011).

1.6.3 Hereditary Persistence of Foetal Haemoglobin (HPFH)

HPFH is a benign genetic condition which is characterised by a significant production of HbF that persists at higher levels (>1%) than normal throughout adulthood, without any associated morphological changes to the erythrocytes (Forget 1998). Mutations in the α-globin or β-globin gene clusters or in the promoter region of the γ-globin genes result in alteration of the normal Hb switching process, whereby silencing of γ-globin gene expression is inhibited (Sharma, Singhal et al. 2020). HPFH was first discovered in Nigeria in the 1960s but has nowadays also been encountered in people of African descent, as well as in Greece (Edington, Lehmann 1955).

Patients with HPFH may be found in heterozygous, homozygous, or compound heterozygous states; with heterozygotes having elevated HbF levels up to 30%, while homozygote or compound heterozygote individuals can have high levels of HbF that approach 100% of the total Hb (Sharma, Singhal et al. 2020; Thein, Craig 1998). Few cases of homozygote HPFH have been identified to date (Wheeler, Krevans 1961).

HPFH can be classified based on the cellular distribution pattern of Hb, based on the type of globin chain produced, and based on the type of molecular defect present. The distribution of HbF in the erythrocytes may be pancellular or heterocellular. Pancellular distribution involves a homogenous distribution of HbF among the erythrocytes and is typically associated with a higher expression of HbF. It is usually caused by large deletions in the β-globin gene cluster or by point mutations in promoters of the γ-globin genes which are inherited in a Mendelian fashion. Heterocellular HPFH involves the uneven distribution of HbF in red blood cells and is associated with a lower increase in HbF levels. Unlike pancellular HPFH, heterocellular HPFH is regarded as a multifactorial quantitative trait and its inheritance is polygenic (Shaukat, Paudel et al. 2018; Thein, Craig 1998). Patients having HPFH may also be classified according to the type of globin chain making up HbF, whether it is $^{A}\gamma$-globin, $^{G}\gamma$- globin or both. Furthermore, the proportion of $^{A}\gamma$-globin, and $^{G}\gamma$- globin chains produced varies between different HPFH patients.

Molecular basis of this disorder can have both deletional and non-deletional mutations. Deletional HPFH is a rare Mendelian form of HPFH that is characterised by large deletions (13-106kbp) in the β-globin gene cluster causing partial or full deletions of the β-globin genes and hence a reduced or absent synthesis of β-globin chains. Such large deletions will deregulate the normal developmental pattern of γ-globin expression leading to a compensatory increase of γ-globin synthesis and HbF. Seven types of deletional HPFH have been identified with the majority having a pancellular distribution of HbF among all red blood cells (Forget 1998; Old 2013).

An example of a deletional HPFH is the delta beta (δβ) thalassaemia, which is a rare Hb disorder characterised by reduced or absent synthesis of adult Hb due to a deletion in both δ- and β-globin genes found on chromosome 11p15. Such patients will present with mild microcytic hypochromic anaemia and elevated levels of HbF up to 90% of the total Hb due to persistent γ-chain synthesis as a compensation (Bollekens, Forget 1991; Ottolenghi, Comi et al. 1976). Because of the increased synthesis of HbF, homozygote patients for δβ-thalassaemia will present with mild phenotypes, similar to β-thalassaemia trait but with heterocellular distribution of HbF in erythrocytes (Verma, Bhargava et al. 2013).

Non-deletional HPFH is the more common form of HPFH that is caused by point mutations in the proximal promoter of the $^{A}\gamma$-globin or $^{G}\gamma$-globin genes, resulting in persistence of HbF in adult erythrocytes. Levels of HbF vary from 1% to 35% in heterozygotes and are less when compared to those observed in patients with deletional HPFH. Furthermore, heterocellular distribution of HbF is usually observed in non-deletional HPFH (Forget 1998). Several non-deletional HPFH-causing mutations have been identified to date, most of which alter the nucleotide sequence of regulatory elements, and in turn affect the binding of erythroid-specific or ubiquitous transcription factors (Ottolenghi, Comi et al. 1989). Non-deletional HPFH can also be caused by complex inheritance of three major quantitative trait loci (QTL), which are the *Xmn*I site upstream of the $^{G}\gamma$- globin genes on chromosome 11p15, *BCL11A* on chromosome 2p15, and *HBS1L-MYB* intergenic polymorphism (HMIP) on chromosome 6q23. Studies suggest that these QTL increase HbF levels by direct activation of the expression of γ- globin genes, as well as by indirectly altering the kinetics of erythroid maturation. They are responsible for 20-50% of the HbF variability (Braghini, Costa et al. 2016; Thein, Menzel 2009).

1.6.3.1 HPFH in Malta

In 2010, Borg, Papadopoulos et al. discovered HPFH in 10 members from a large Maltese family of 27 members, with elevated HbF levels ranging from 3.3% to 19.5%. Genome wide SNP followed by linkage analysis identified a candidate region on chromosome 19p13.12-13, and DNA sequencing revealed a nonsense mutation p.K288X in the *KLF1* gene. The p.K288X mutation is a truncation mutation that involves the replacement of alanine by thymine, resulting in a premature stop codon which completely ablates the DNA binding domain of the key erythroid transcription factor KLF1. The family members with HPFH were heterozygous carriers of this mutation, but it was absent from the general Maltese population (Borg, Papadopoulos et al. 2010).

Gene expression profiles conducted on cultured primary HEP cells from the family members revealed that in members with HPFH the γ-globin genes were upregulated, while the expression of *BCL11A* was downregulated. Similar results were obtained from knockdown experiments of *KLF1* in cultured HEPs and the BCL11A levels increased upon restoration of the KLF1 activity. This study hence showed the role of KLF1 as a dual regulator in the foetal-to-adult globin gene switching by directly activating the β-globin gene locus, and indirectly by activating the expression of BCL11A, which in turn represses the γ-globin genes (Figure 1.11) (Borg, Papadopoulos et al. 2010). Five additional Maltese families were identified, having the same p.K288X truncation mutation in the *KLF1* gene but who present with normal or slightly increased HbF levels and borderline HbA_2 (Grech, Borg et al. 2020).

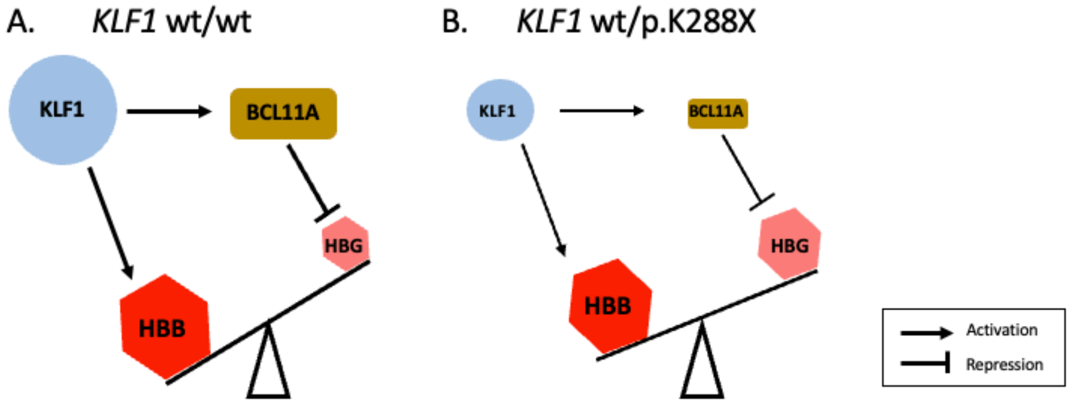


Figure 1.11: The effect of *KLF1* p.K288X mutation on globin expression; Figure A: In normal adults KLF1 preferentially activates β-globin genes and silences γ-globin gene expression by activating BCL11A. Figure B: Individuals with the *KLF1* p.K288X mutation have reduced KLF1 activity, which downregulates the expression of *BCL11A* and alleviates the repression of γ-globin genes (adapted from Borg, Papadopoulos et al. 2010).

1.6.3.2 *KLF1* Variants and Haematological Conditions

Due to the key role that KLF1 plays in erythropoiesis and other physiological processes, *KLF1* variants have been associated with haematological conditions of varying severity and altered red blood cell phenotype (Figure 1.12) (Huang, Zhang et al. 2015; Perkins, Xu et al. 2016).

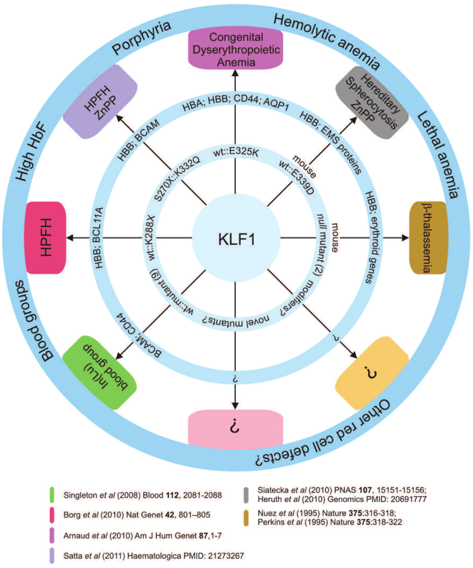

Porphyria
Hemolytic anemia
Lethal anemia
Other red cell defects?
Blood groups
High HbF
Congenital Dyserythropoietic Anemia
Hereditary Spherocytosis ZnPP
β-thalassemia
?
?
In(Lu) blood group
HPFH
HPFH ZnPP
HBA; HBB; CD44; AQP1
HBB; EMS proteins
HBB; erythroid genes
?
?
BCAM; CD44
HBB; BCL11A
HBB; BCAM
wt::E325K
mouse wt::E339D
mouse null mutant (2)
modifiers?
novel mutants?
wt::mutant (9)
wt::K288X
S270X::K332Q
KLF1
Singleton et al (2008) Blood 112, 2081-2088
Borg et al (2010) Nat Genet 42, 801–805
Arnaud et al (2010) Am J Hum Genet 87,1-7
Satta et al (2011) Haematologica PMID: 21273267
Siatecka et al (2010) PNAS 107, 15151-15156;
Heruth et al (2010) Genomics PMID: 20691777
Nuez et al (1995) Nature 375:316-318;
Perkins et al (1995) Nature 375:318-322

Figure 1.12: Phenotypes caused by *KLF1* mutations. *KLF1* mutations and potential modifiers are displayed in the inner ring, critical KLF1 target genes/loci are displayed in the middle ring, while the phenotypes are displayed in the outer ring (adapted from Borg, Patrinos et al. 2011).

The first reported mutation in the *KLF1* gene was associated with low or absent expression of the Lutheran (Lu) blood group antigens, which results in either the rare phenotype Lu(a-b-) with complete absence of the Lu antigens, or can result in the inhibitor of Lu (In(Lu)) phenotype where the expression of the Lu antigens is reduced. This phenotype is not associated with any pathology but may increase the risk of alloimmunization if transfusions are not properly matched (Eernstman, Veldhuisen et al. 2021). Mutations in the *KLF1* gene may result in very high levels of zinc protoporphyrin, since KLF1 regulates several enzymes in erythropoiesis leading to haem biosynthesis. This was demonstrated in a Sardinian family where two brothers having HPFH and elevated zinc protoporphyrin were found to have two mutations in the *KLF1* gene (Satta, Perseu et al. 2011). Furthermore, mice with *Nan* (neonatal anaemia) mutation that alters the DNA binding specificity of KLF1 displayed elevated levels of zinc protoporphyrin (Siatecka, Sahr et al. 2010).

KLF1 variants can also give rise to mild phenotypes such as elevated HbF levels, as described in a large Maltese family with HPFH (Borg, Papadopoulos et al. 2010), and raised levels of HbA_2, as shown by a study on 145 Italian subjects who had borderline levels of HbA_2 (3.3-4.1%) as a result of mutations in the *KLF1* gene (Perseu, Satta et al. 2011).

Congenital dyserythropoietic anaemia (CDA), which refers to a group of rare congenital anaemias caused by dyserythropoiesis in the bone marrow, may also result from mutations in the *KLF1* gene *(*Renella, Wood 2009). A study by Arnaud et al. in 2010, describes a missense mutation (c.973G>A) in the *KLF1* gene, which was associated with CDA and profound dysregulation of globin gene expression. This further emphasises the role of KLF1 in transcriptional regulation of human globin genes (Arnaud, Saison et al. 2010).

1.7 Management of Patients with Haemoglobinopathies

Clinical management of patients with haemoglobinopathies depends on the severity of the complications. Patients having β-thalassaemia major are dependent on lifelong blood transfusions every four to six weeks to correct their Hb concentration and maintain the levels between 9-10.5g/dL. Transfusion treatments usually commence before two years of age to enable normal growth and correct for the ineffective erythropoiesis. It is estimated that 2 million packed red blood cell units are required per year to treat patients suffering from β-thalassaemia worldwide (Yogalakshmi, Hemamalini et al. 2020). Complications related to iron overload are also present in such patients including irreversible organ damage, endocrine complications, cardiac toxicity and cirrhosis, leading to significant morbidity and mortality if left untreated (Mishra, Tiwari 2013; Olivieri, Brittenham 2013). Effective treatment of β-thalassaemia major patients must also include iron chelation therapy.

Patients with SCA must be properly diagnosed early in life to allow for early initiation of preventative treatment. If SCA is left untreated, patients will have a poor quality of life and early mortality rates due to haemolytic anaemia, acute vaso-occlusive events and chronic end-organ damage. The acute vaso-occlusive events will cause these individuals to suffer from sudden onset of painful episodes and increases their risk of respiratory distress, acute chest syndrome and stroke. Transfusion treatment is required to treat the anaemia, improve the oxygen carrying capacity and must be administered indefinitely following an initial stroke to prevent recurrent events (McGann, Nero et al. 2013). Furthermore, blood transfusions come with the risk of transfusion transmitted infections and adverse reactions. Management and treatment of patients with β-thalassaemia and SCA is therefore both expensive and inconvenient for the patients themselves.

The only available curative treatments of individuals with β-thalassaemia and SCA are haematopoietic stem cell transplantation (HSCT) or bone marrow transplantation from human leukocyte antigen (HLA) matched donors, with only a limited number of patients who benefit from this procedure. Patients must undergo chemotherapy sessions to prevent graft versus host disease (GVHD) by inhibiting any T-cells, and can even suffer from post-transplant infections and graft rejection. Bone marrow transplants also pose a financial and psychological burden on the recipients, and despite the efforts, 10% of the transplants remain unsuccessful and hence such individuals cannot benefit from a disease-free survival (Kohne 2011, Marengo-Rowe 2007).

1.7.1 Pharmacological induction of HbF

Both molecular studies and clinical findings confirm that elevated levels of HbF help ameliorate the clinical severity of the condition in patients with haemoglobinopathies by reducing the imbalance between the α-globin and β-globin chains, while also increasing Hb synthesis (Carrocini, Zamaro et al. 2011; Musallam, Taher et al. 2013). Novel therapeutic approaches involve reactivation of the expression of the foetal γ-globin chains through pharmacological means, with aims to act as substitutes for the defective β-globin genes. This can be done by preventing or reversing the second Hb gene switching event, hence increasing the overall levels of HbF that correct the severe chronic anaemia (Marengo-Rowe 2007; Panja, Basu 2015). Several studies have identified different classes of pharmacological agents (Table 1.3) that augment HbF levels both *in vivo* and *in vitro*.

Table 1.3: Pharmacological inducers of HbF in human erythroid cells.

Pharmacological Agent	Mechanism of Action	Limitations
Hydroxyurea (HU) (Charache, Terrin et al. 1995; Musallam, Taher et al. 2013; Torrents 2014)	• Inhibition of DNA synthesis by inactivating the enzyme ribonucleotide reductase	• Toxicity associated with long term use
DNA Methyltransferase Inhibitors Example: 5-azacytidine and 5-aza-2'-deoxycytidine (Carr, Rahbar et al. 1988; Christman 2002; Ley, Nienhuis 1985)	• Inhibition of DNA methyltransferases (DNMT) resulting in hypomethylation of DNA	• Potential carcinogenic effect by subtle toxicity on the bone marrow with repeated administration
Butyrates and butyrate derivatives (Faller, Perrine 1995; Fard, Hosseini et al. 2013; Weinberg, Ji et al. 2005)	• Histone deacetylase (HDAC) inhibitors • Increase efficiency of translation of γ-globin mRNA	• Very short half-life in plasma • Intravenous administration • Develop tolerance following long term exposure
Immunomodulators: Thalidomide and its derivatives (Pomalidomide) (Aerbajinai, Zhu et al. 2007; Yang, Hu et al. 2020)	• Activation of p38 mitogen-activated protein kinase (MAPK) pathway • Histone acetylation at the γ-globin gene promoter through generation of reactive oxygen species (ROS)	• Teratogenic effects • Peripheral neuropathy

Several other pharmacological agents have also been tested and found to influence HbF, including sirolimus, nicotinic acid and recombinant human erythropoietin (rEPO) (Fard, Hosseini et al. 2013). Some studies are also taking combinatorial approaches of drug testing which have proved to be more superior in elevating HbF levels and ameliorating clinical symptoms when compared to single drug treatments. However, there is no one single drug that can fully eliminate the complications associated with β-thalassaemia and SCD and so genetic strategies to reactivate HbF expression seem to be the only potential solution for a disease-free survival (Demirci, Leonard et al. 2020).

1.7.2 Gene Therapy

Gene therapy is a novel experimental therapeutic approach that involves the transplant of autologous HSC that have been genetically modified *ex vivo* to express the gene of interest. It can be grouped into two categories: gene addition and gene editing. Gene addition involves the *in vitro* insertion of β-globin producing genes into HSC, using lentiviral or retroviral vectors which integrate with the host cell's genome. The modified HSC are subsequently transfused into the patient following myeloablation, where they proliferate and repopulate the bone marrow (Soni 2020). In gene editing the genes are altered and corrected *in situ* by specific nucleases, such as zinc finger nucleases (ZFNs) and the CRISPR/Cas 9 system (clustered regularly interspaced short palindromic repeats (CRISPR)/CRISPR-associated protein 9) (Demirci, Leonard et al. 2020). Such designer nucleases will create double stranded breaks at precise areas of the genome, which are then repaired by replacement of the host cell sequence with that of the donor template provided (Goodman, Malik 2016). The goal of both strategies is to correct the ineffective erythropoiesis and haemolysis associated with haemoglobinopathies, reduce transfusion needs, while also avoiding GVHD associated complications and immunosuppression that accompany HSC transplants (Karponi, Zogas 2019).

Although gene therapy offers the possibility of cure, it still has several limitations, including expenses and appropriate infrastructure required to perform such advanced methods, which may not be easily accessed by most patients. Furthermore, patients must undergo myeloablative drug conditioning regimens to ensure a successful transplant of the modified HSC and long-term follow-ups are required to assess the safety and efficacy of such therapy. Long-term complications that may arise due to persistence of HbF are also still unknown (Rivers, Molokie et al. 2019).

1.7.3 Significance of the Problem

The clinical management of patients with haemoglobinopathies is hence a challenging and expensive process, often requiring long-term care and presenting significant inconvenience to patients. While bone marrow transplant and HSCT are the only curative treatments available, they are limited to a small subset of patients. Pharmacological induction of HbF can help alleviate symptoms, but no single drug can fully eliminate the complications of the disease. Gene therapy represents a novel experimental therapeutic option, but its high costs and unknown long-term complications make it inaccessible to most patients.

Studies have demonstrated that the coexistence of HPFH with other haemoglobinopathies, such as sickle cell disease and thalassaemia helps in decreasing the severity of clinical symptoms owing to the elevated HbF levels. This is because the increased HbF levels help reduce the concentration of sickling HbS in sickle cell disease, as well as decrease the level of unbound α-globin chains found in thalassaemia (Sharma, Singhal et al. 2020; Shaukat, Paudel et al. 2018). Therefore, understanding the molecular defects that cause HPFH may help provide new insights on the molecular mechanisms that control γ-globin gene expression, and hence identify novel points of therapeutic intervention (Sharma, Singhal et al. 2020). Application of both genomic and proteomic techniques to study patients with HPFH may provide valuable insights into the underlying molecular mechanisms and will help in identification of candidate modifier genes.

1.8 Genomic Profiling by Whole Genome Sequencing (WGS)

Whole genome sequencing (WGS) is a genetic technique used to determine the complete DNA sequence of an individual's genome. It involves the generation of millions to billions of short DNA sequences, called reads, from the DNA sample and then assembling these reads into the full genome sequence. The process of WGS starts off by fragmenting the DNA extracted into small pieces and amplifying these fragments by polymerase chain reaction (PCR). A library of DNA fragments is generated by attaching short sequences of DNA, known as adapters, to both ends of each DNA fragment. The library of DNA fragments is sequenced by using high-throughput sequencing platforms, which read the sequence of the DNA fragments by detecting the incorporation of fluorescently labelled nucleotides. The sequencing process produces millions to billions of short reads, each typically 100 to 300 base pairs in length, which are then used to generate the complete genome sequence (Fuentes-Pardo, Ruzzante 2017; Ng, Kirkness 2010).

WGS can be classified into two categories; whole genome *de novo* sequencing which involves assembling the resulting reads to produce a complete genome sequence for the first time, and whole genome resequencing (WGR) which involves aligning the resulting reads to a reference genome allowing for analysis of genomic variability between individuals or populations (Figure 1.13) (Fuentes-Pardo, Ruzzante 2017).

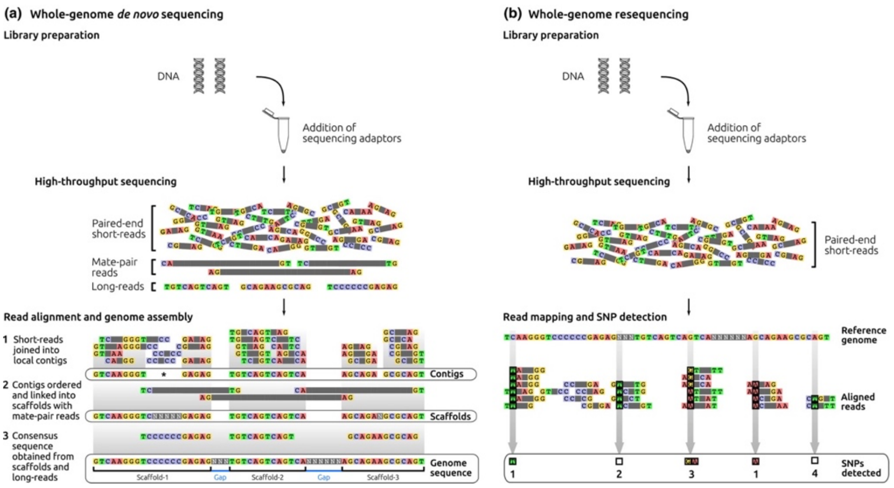

(a) Whole-genome de novo sequencing
Library preparation
DNA
Addition of sequencing adaptors
High-throughput sequencing
Paired-end short-reads
Mate-pair reads
Long-reads
Read alignment and genome assembly
1 Short-reads joined into local contigs
Contigs
2 Contigs ordered and linked into scaffolds with mate-pair reads
Scaffolds
3 Consensus sequence obtained from scaffolds and long-reads
Genome sequence
Scaffold-1
Gap
Scaffold-2
Gap
Scaffold-3
(b) Whole-genome resequencing
Library preparation
DNA
Addition of sequencing adaptors
High-throughput sequencing
Paired-end short-reads
Read mapping and SNP detection
Reference genome
Aligned reads
SNPs detected
1
2
3
1
4

Figure 1.13: Whole Genome de novo Sequencing vs Whole Genome Resequencing (adapted from Fuentes-Pardo, Ruzzante 2017).

One of the main advantages of WGS is that it allows for the comprehensive analysis of an individual's entire genome, including non-coding regions with high resolution (Gonzaga-Jauregui, Lupski et al. 2012). The ability to deliver large volumes of data in a short amount of time makes WGS a powerful tool for genomics. WGS also offers high accuracy and sensitivity, enabling the detection of large and small rare genetic variants that might be missed by targeted genomic techniques. It can also identify potential causative variants for studies of gene expression and regulation mechanisms including detection of SNPs, insertions/deletions, copy number changes, as well as large structural variants (Gong, Pan et al. 2018; Wrzeszczynski, Felice et al. 2018). This comprehensive analysis can provide a wealth of information about an individual's genetic predisposition to various diseases and can aid in the development of more personalized treatments (Brown, Aisner et al. 2018; Katsanis, Katsanis 2013). In addition, the use of high-throughput sequencing platforms has dramatically reduced the cost and time required for WGS, making it more accessible and feasible for large-scale genetic studies (Christensen, Dukhovny et al. 2015; Meienberg, Bruggmann et al. 2016).

1.9 Proteomic Profiling

The relationship between transcription factors and *cis*-regulatory elements that control globin gene expression and processes like erythropoiesis, is best understood in the context of dynamic gene regulatory networks (GRN). However, since GRN are mostly derived from transcriptomic data, their usefulness is limited. This is because it has recently been proven that there are major discrepancies between messenger RNA (mRNA) and protein abundance for master regulators of erythropoiesis (Brand, Ranish 2021). Proteomic studies and quantitative protein measurements are necessary to provide essential information to fully understand these processes. Flow cytometry and mass spectrometry (MS) are two powerful tools that are widely used in the study of blood proteomics to gain insight in the regulation of erythropoiesis, haematopoiesis, and globin gene expression (Maes, Cools et al. 2020).

1.9.1 Flow Cytometry

Flow cytometry is a high-throughput technique that analyses cells based on their physical and biochemical properties. It can also be used to isolate and analyse specific cell populations based on the protein expression on the cell surface. The basic principle of flow cytometry involves placing a heterogenous cell population in suspension and measuring the properties of individual cells as the cell suspension enters the flow cell and passes by a set of lasers. The scattered light and fluorescence signal emitted by each cell is collected by photodetectors and the data generated is then analysed and used to distinguish between different cell populations and determine the expression of specific proteins (Figure 1.14) (Adan, Alizada et al. 2017; Maes, Cools et al. 2020). This technique is called fluorescence-activated cell sorting (FACS). Flow cytometry has several advantages, including its high-throughput capacity, its ability to analyse thousands of cells or particles per second, and its ability to measure multiple parameters simultaneously. It is also highly sensitive and can detect even small changes in the properties of particles or cells, making it a valuable tool in the study of various diseases and biological processes.

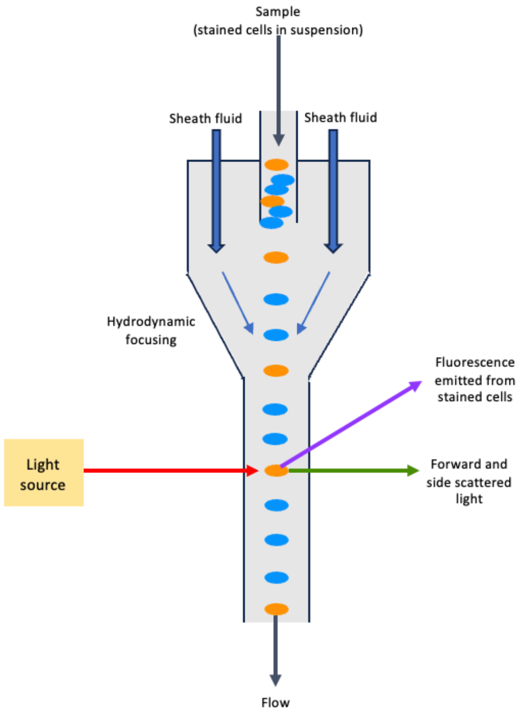

Sample
(stained cells in suspension)
Sheath fluid
Sheath fluid
Hydrodynamic focusing
Fluorescence emitted from stained cells
Light source
Forward and side scattered light
Flow

Figure 1.14: The basic principle of a Flow Cytometer; introduction of samples in suspension form where hydrodynamic focusing causes the cells to pass by the laser source one cell at a time. The forward and side scattered light as well as fluorescence signals emitted from the cells are detected.

1.9.2 Mass Spectrometry

MS is a powerful and robust analytic technique that can handle the complexities associated with the proteome. It allows for unbiased identification and quantification of thousands of proteins from the samples in an absolute or relative manner. The absolute quantification of proteins in samples will reveal information about the stoichiometry and dynamics of proteins involved in erythropoiesis, which is crucial in constructing quantitative models of GRN in erythropoiesis (Brand, Ranish 2021). The basic principle of MS is to ionise the proteins in a sample and measure the mass to charge (m/z) ratio of the resulting gas phase ions. The different proteins are then identified and quantified based on their unique mass and charge properties (Murayama, Kimura et al. 2009).

Every mass spectrometer has three main components; an ion source which converts the analyte molecules to gas phase ions, a mass analyser which separates the ionised analytes based on their m/z ratio, and a detector which records the number of ions at each m/z ratio (Figure 1.15) (Arevalo, Ni et al. 2019). The proteins are then identified by a top-down method or a bottom-up method. In the top-down method, whole protein analysis is carried out by separating the proteins in a complex mix before analysing with MS to obtain intact protein mass measurement. In the bottom-up method the proteins in a complex mixture are separated and digested into peptides by specific enzymes or chemicals. The shotgun method is the standard MS approach which involves digesting a mix of proteins into peptides and identifying the proteins based on the presence of at least one unique peptide (Han, Aslanian et al. 2008). MS has several advantages for the study of proteomics, including its high sensitivity and specificity, its ability to identify and quantify large numbers of proteins simultaneously, and its ability to detect low abundance proteins. It is also capable of determining the post-translational modifications (PTMs) of proteins, which can provide valuable information about protein function and the regulation of biological processes (Larsen, Trelle et al. 2006).

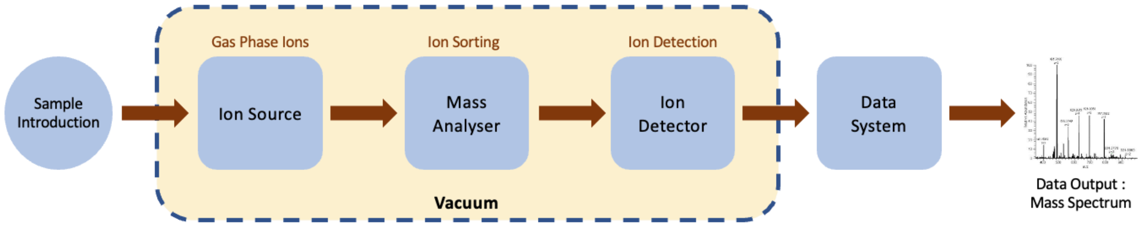


Figure 1.15: The basic components of a Mass Spectrometer; showing the conversion of the sample into gas phase ions by the ion source, the separation of the charged ions based on their m/z ratio by the mass analyser, detection of the ions by the ion detector and visualisation of the data collected as a mass spectrum following processing by the data system.

1.10 Aims & Objectives of the Study

The overall research objective of this study was to identify genetic and molecular pathways involved in the regulation of the globin gene switch. The aims of this research study were to:

- Identify and recruit three independent Maltese families with HPFH due to KLF1 deficiency for the study,
- Perform whole genome sequencing and subsequent bioinformatics analysis on the individuals with HPFH due to the KLF1 deficiency and their family members to identify candidate modifier genes,
- Compare the genetic profiles of HPFH affected individuals from the three distinct families to uncover variants contributing to the elevated levels of HbF, particularly in families exhibiting higher HbF levels despite carrying the same *KLF1* mutation,
- Analyse the unique proteomic data sets obtained by mass spectrometry from individuals with HPFH due to KLF1 deficiency and their family members,
- Integrate the data obtained from whole genome sequencing, mRNA gene expression assay, flow cytometry and mass spectrometry, to generate comprehensive multiomics profile of individuals with HPFH and gain insights into the regulation of globin gene expression.

2. Methodology

2.1 Patient Recruitment & Sample Collection

Patient recruitment involved the recall of 11 HPFH-KLF1 haploinsufficient family members from 3 separate families, as well as 11 healthy individuals from the same families to act as controls (Figure 2.1 and Figure 2.2). The extended original Maltese family with HPFH reported by Borg, Papadopoulos et al. (2010), consists of 12 individuals and is referred to as Fam F1 throughout this thesis. Two other families, referred to as Fam F2 and Fam F4, which consist of 4 and 6 individuals respectively, have the same p.K288X truncation mutation in the *KLF1* gene as the original Fam F1, but present with normal or slightly increased HbF levels and borderline HbA_2 (Grech, Borg et al. 2020). The original family Fam F1 was extended to include individual F1.12, while Fam F4 was extended with the parents (F4.1 and F4.2) of the proband (F4.4), together with her siblings F4.3, F4.5 and F4.6. The participants for this study had already provided informed consent to blood collection for previous studies (Approval number UREC 45/2014).

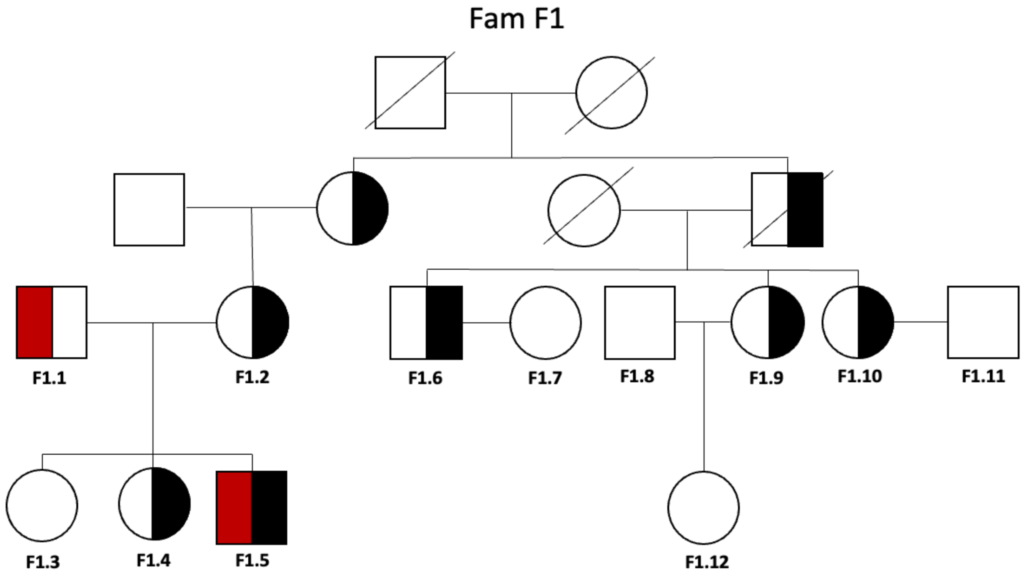


Figure 2.1: The family tree of the original Maltese HPFH family Fam F1, showing individuals affected with HPFH and carrying the *KLF1* mutation p.K288X as half-filled symbols (black), heterozygote members for Hb St. Luke's are shown by half-filled red symbols, while unaffected family members are shown as open figures.

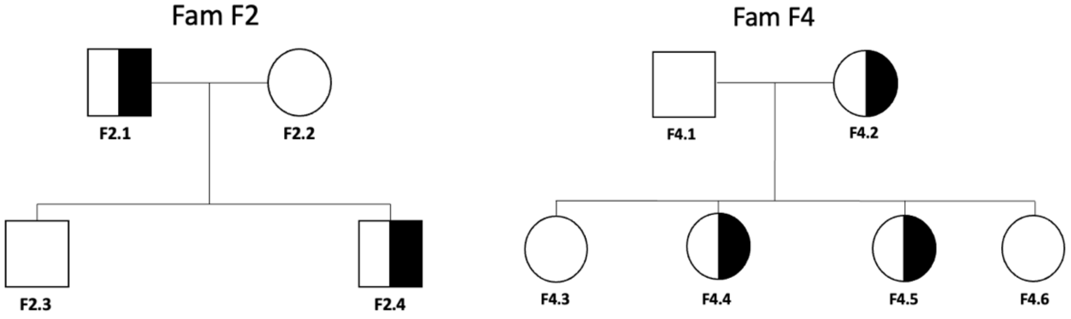


Figure 2.2: The family trees of two other Maltese HPFH families (Fam F2 and Fam F4), showing individuals affected with HPFH and carrying the *KLF1* mutation p.K288X as half-filled symbols (black), while unaffected family members are shown as open figures.

Blood was collected by trained phlebotomists at the Thalassaemia and Molecular Genetics clinic at Mater Dei Hospital in purple vacutainers containing K_2-EDTA (dipotassium ethylenediaminetetraacetic acid) as an anticoagulant. Four K_2-EDTA vacutainers were collected from each patient, which amounts to a total blood volume of 12 ml. One sample was used to measure the haematological parameters, as well as levels of HbA_2 and HbF, while the second sample was used for DNA and RNA work. The EDTA vacutainer for DNA and RNA work were stored at -20°C at the laboratories at the University of Malta until used, while the other blood sample was processed immediately at Mater Dei Hospital. The remaining two samples were transported to the Department of Haematopoiesis of the Sanquin Research Facility in Amsterdam (Netherlands) where proteomic lab work was performed.

2.2 Measurement of Haematological Parameters

Haematological parameters were measured using an automated haematology analyser (XT-20001, Sysmex, Canada) at the Haematology Laboratory at Mater Dei Hospital, while HbF and HbA_2 levels were determined by high performance liquid chromatography (HPLC) (VARIANT™ II System, Bio-Rad, USA) at the Molecular Genetics Laboratory, University of Malta.

The principle of the VARIANT™ II System is to separate and quantify the various haemoglobins present in a whole blood sample by ion exchange HPLC. The EDTA blood samples were loaded in sample racks and placed on the sampling station conveyer belt, where they were automatically mixed, diluted and subsequently injected into the analytical cartridge secured in a temperature-controlled chamber. Two pumps created a buffer gradient of increasing ionic strength across the cartridge, which was then used to separate the different haemoglobins based on their ionic charge. The separated haemoglobins were eluted at different times and the changes in absorbance at 415nm were measured as they passed through the flow cell of the filter photometer. A sample report and a chromatogram were generated for each sample. The various haemoglobins were identified based on their area percentages and their characteristic retention time, which refers to the duration between sample injection and the point at which a peak corresponding to Hb reaches its maximum height.

2.3 Flow Cytometry

Flow cytometry analysis was carried out at the Department of Haematopoiesis of the Sanquin Research Facility in Amsterdam (Netherlands). It was performed on the packed red cells from all the subjects to measure the protein expression of 3 different red cell surface antigens; BCAM (basal cell adhesion molecule), CD44 and P1 antigens. Previous studies have shown that the expression of these three antigens is affected in patients with *KLF1* variants, and hence a flow cytometry assay was used to determine their levels in the Maltese subjects with HPFH due to *KLF1* haploinsufficiency.

A volume of 1ml whole blood from each sample collected was transferred into labelled 1.5ml Eppendorf tubes. Packed red cell concentrates were prepared by centrifuging each

Eppendorf tube at 800 *g* (relative centrifugal force) for 10 minutes at room temperature. The plasma layer was discarded, and the remaining red cells were washed with 1ml of phosphate-buffered saline (PBS: 1X concentration; Sigma-Aldrich, USA) supplemented with 1% bovine serum albumin (BSA; Sigma-Aldrich, USA) for a total of 3 washes, to ensure any remaining antibodies present in the plasma are removed since these would otherwise interfere with the staining. The tubes were vortexed and centrifuged at 650 *g* for 2 minutes between each wash and the supernatant was discarded each time. A volume of 10µl of each sample of packed red cells was mixed with 490µl of PBS 1% BSA solution to dilute by a factor of 50. A batch of diluted antibodies were prepared using PBS 1% BSA, where BCAM was diluted 1:100, CD44 was diluted 1:200, and P1 was diluted 1:500. Since BCAM and P1 antibodies were unconjugated, secondary conjugated antibodies were used and dilutions were prepared. The secondary antibody rabbit anti-goat Alexa Fluor™ 488 (AF-488; Invitrogen) which binds to BCAM was diluted 1:200, while the secondary antibody goat anti-human allophycocyanin (APC; Invitrogen) which binds to P1 was diluted 1:100.

A volume of 5µl of the diluted red cells were pipetted into a 96-well plate, where 6 wells were used for each sample (BCAM, P1, CD44, Secondary antibody AF-488, Secondary antibody APC and unstained). The diluted red cells were then mixed with 50µl of P1, 50µl of CD44 and 25µl of BCAM diluted antibodies respectively and were incubated for 30 minutes at room temperature. After incubation the red cells were washed 3 times with 200µl of PBS 1% BSA and the plate was centrifuged at 360 *g* for 2 minutes. The supernatant was discarded by swift inversion of the plate and gentle blotting. A volume of 25µl diluted AF-488 secondary antibodies were added to each BCAM well, while a volume of 50µl diluted APC secondary antibodies were added to each P1 well. The plate was covered with an aluminium foil plate cover and incubated for 30 minutes at room temperature. Again, the red cells were washed with 200µl of PBS 1% BSA and centrifuged at 360 *g* for 2 minutes. This step was repeated 2 more times for a total of 3 washes. A final volume of 110µl of PBS 1% BSA was added to each well and the plate was loaded on the BD FACSCanto™ II Flow Cytometer (BD Biosciences). The fluorescence signal and the scattered light was measured for each well and the data was analysed using FlowJo™ software (BD Life Sciences).

2.4 DNA Extraction

Genomic DNA was extracted from peripheral blood leucocytes using the QIAamp® DNA Mini Kit (Qiagen, Germany) following the manufacturer's instructions. This kit makes use of guanidine hydrochloride and a high salt buffer to rapidly extract and purify DNA. The guanidine hydrochloride acts as a lysing buffer, while the high salt buffer removes any contaminating proteins, resulting in a good yield and high-quality DNA concentration. Figure 2.3 shows an overview of the steps to extract DNA using the QIAamp Spin Procedure. A volume of 20µl Proteinase K was pipetted into the bottom of sterile labelled 1.5ml microcentrifuge tubes, followed by addition of 200µl of each whole blood sample. Lysis of the samples was carried out by adding 200µl of Buffer AL and mixing by pulsed vortexing for 15 seconds. Care was taken to ensure that the samples and the Buffer AL were completely homogeneous to ensure efficient lysis. The samples were incubated at 56°C for 10 minutes, followed by brief centrifugation to remove any drops from the inside of the lid. Using sterile micropipettes, 200µl of absolute ethanol (Sigma-Aldrich, USA) was added to each sample and mixing was carried out by pulsed vortexing for 15 seconds. After mixing, the microcentrifuge tubes were briefly centrifuged to remove any drops from the inside of the lid. The lysates were carefully transferred to the QIAamp Mini spin column (Qiagen, Germany) assembled in a 2ml tube and centrifuged at 6,000 *g* for 1 minute at room temperature. The spin columns were placed in clean 2ml collection tubes and the tubes containing the flow-through were discarded.

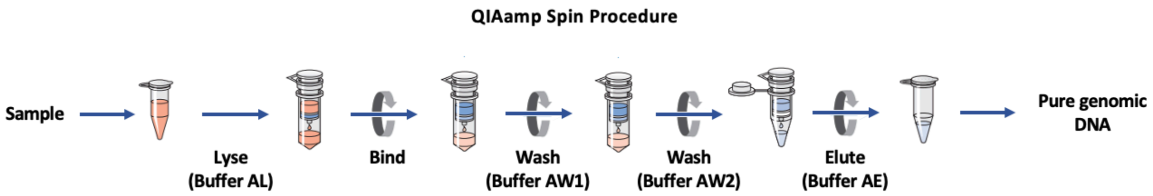


Figure 2.3: Extraction of Genomic DNA using the QIAamp® DNA Mini Kit (adapted from Qiagen, Germany).

The columns were treated with two consecutive washes using two different wash buffers; Buffer AW1 and Buffer AW2, to remove any residual contaminants, hence significantly improving the purity of the eluted DNA. Since both wash buffers AW1 and AW2 are supplied as concentrates, they were both prepared prior to their use by adding the appropriate amount of 100% ethanol as indicated on the bottle. A volume of 500µl of Buffer AW1 was added to each spin column, followed by centrifugation at 6,000 *g* for 1 minute at room temperature and finally transferring the columns to clean 2ml collection tubes. A second washing step was performed by carefully adding 500µl of Buffer AW2 to each spin column followed by centrifugation at full speed of 20,000 *g* for 3 minutes. The collection tubes with the flow-through were discarded and the spin columns were transferred to clean 2ml collection tubes. A centrifugation step at 20,000 *g* for 1 minute at room temperature was performed to dry the columns and eliminate any possible Buffer AW2 carryover. The extracted DNA was eluted by transferring the columns to new sterile 1.5ml microcentrifuge tubes, adding 150µl of DNA Elution Buffer AE, incubating for 5 minutes at room temperature and subsequently centrifuging the tubes at 6,000 *g* for 1 minute. The quality of the DNA was checked by agarose gel electrophoresis (Section

2.5.1) and the DNA yields were quantified using the NanoDrop^{TM} 2000 Spectrophotometer (Thermo Scientific, USA) as described in Section 2.5.2 below. The eluted DNA were stored at 4°C until analysed.

2.5 Measurements of DNA Quality and Quantity

2.5.1 Quality Check by Agarose Gel Electrophoresis

Agarose gel electrophoresis is a technique that is used to separate molecules according to size and electric charge by loading in an agarose gel with pores and creating an electric current across the gel. During the process of gel formation, the agarose polymers associate non-covalently to form a network of pores allowing the gel to act as a molecular sieve. The size of the pores is determined by the concentration of the agarose gel. When a current is applied through the gel matrix, the DNA fragments will start migrating towards the positively charged anodes. Smaller fragments travel fastest and furthest through the gel, as opposed to larger fragments which slow down as they travel through the gel matrix (Lee et al. 2012). The quality and integrity of the extracted genomic DNA was checked by loading the samples onto a 1% agarose gel and subsequently comparing the size of the fragments to that of a known DNA size marker. A tight dark band of high molecular weight was expected for good quality DNA, while a smear across the agarose gel would be observed in cases where the DNA was degraded.

A 1% agarose gel was prepared by weighing 1g of ultra pure agarose powder (AppliChem, Germany) in a weighing boat on an electronic balance. The measured agarose was transferred to a clean conical flask and a 100ml volume of 1X Tris-acetate-EDTA (TAE) buffer (121g Trizma base in 400ml of distilled water, 28.5ml of glacial (100%) acetic acid, 50ml of 0.5M EDTA pH 8.0) was added and mixed by swirling. The top of the conical flask was covered with a piece of plastic film that was pierced to allow release of any pressure that might have accumulated during heating, and the mixture was melted by heating for approximately 5 minutes until a dissolved and transparent homogenous solution was obtained. The dissolved agarose solution was swirled gently and allowed to cool down slightly. Using a micropipette, 5µl of ethidium bromide (10µg/L concentration; Sigma Aldrich, USA; 10µl per 100ml of 1X TAE buffer) was added to the agarose solution to allow for visualisation of the DNA fragments under ultraviolet (UV) light. Ethidium bromide is an intercalating dye that inserts itself between the base pairs in the DNA strand due to its resemblance to the nitrogenous bases. Upon excitation by UV light, ethidium bromide will fluoresce intensely and emit light in the visible spectrum which can be detected (Sigmon, Larcom 1996). The solution was mixed and poured into a clean casting tray. Any air bubbles were moved towards the edge of the casting tray, since these would otherwise interfere with the migration of the products in the gel. Combs were placed in the casting tray to create wells and the gel was left to set for approximately 30 minutes.

The casting tray with the solidified gel was placed in a levelled electrophoresis chamber with 1X TAE buffer, providing the ions and an adequate pH allowing for the separation of the DNA fragments, and the combs were carefully removed. Using a micropipette, 5µl of each DNA sample were mixed with 1µl of 6X loading buffer (0.25% bromophenol blue, 0.25% xylene cyanol FF, 15% Ficoll in water) and were carefully loaded in separate wells. A viral DNA size marker cut into known fragment sizes was also pipetted into the last well of the gel. The size marker used was φX 174 DNA digested with *Hae* III (New England BioLabs Inc, Ipswich, UK) which yields 11 fragments that have an optimal range of 100-1000bp (Figure 2.4). The electrophoresis chamber was closed with its lid, and the leads were connected to a power supply set at a voltage of 120 volts (V) with a current of 400 milliamps (mA) and was run for 20 minutes. The gel was viewed under a UV transilluminator (UVP Incorporation BioImaging systems, USA), and the DNA concentration and integrity were determined by the intensity and size of the bands.

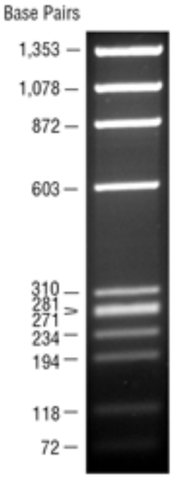


Figure 2.4: The DNA size marker φX 174 DNA-Hae III Digest showing 11 fragments visualised by ethidium bromide staining (adapted from New England BioLabs Inc, Ipswich, UK).

2.5.2 Quantification of the Extracted DNA

The concentration of the nucleic acids extracted from the peripheral whole blood was determined using the Thermo Scientific NanoDrop^{TM} 2000 Spectrophotometer (USA). Quantification of the DNA ensures a successful extraction procedure while checking that the extracted DNA products are of good quality and purity. The NanoDrop technology allows for measurements of microvolume samples without using cuvettes or capillaries to retain the samples. The surface tension of each sample is used to create a liquid column that forms a vertical optical path. The light that passes through the sample is then analysed and measured by a spectrophotometer. The Thermo Scientific NanoDrop^{TM} 2000 Spectrophotometer (USA) measured the absorbance at wavelengths of 230nm, 260nm and 280nm and displayed the results as a spectral image (Desjardins, Conklin 2010).

The upper and lower optical surfaces of the spectrophotometer were cleaned using lint free paper. The instrument was first blanked by pipetting 1µl of elution Buffer AE (Qiagen, Germany) onto the bottom pedestal, followed by the measurement of the eluted DNA samples, whereby 1µl of the purified products was pipetted on the pedestal. To avoid any carryover, a water sample was measured between each sample, and a lint free laboratory wipe was used to wipe each sample from the measurement pedestals. The concentration of the DNA in each sample was measured in nanograms per microlitre (ng/µl) based on the absorbance at the 260nm wavelength. The purity and quality of the nucleic acids in the sample was indicated by the ratio of absorbance at the 260nm and 280nm (260:280), which ranged between 1.8 to 2.0 for good quality DNA.

2.6 Whole Genome Sequencing (WGS)

The extracted genomic DNA from 20 samples was sent abroad to the facilities at Dante Genomics (Italy), where WGS was carried out. Upon arrival at the laboratories, the genomic DNA was prepared for processing using the Illumina® DNA Prep (Illumina Inc., San Diego, California, USA) library preparation kit. This kit makes use of bead-linked transposome complexes to tagment the DNA, by fragmenting and tagging the DNA with adapter sequences in a single step (Figure 2.5). Following tagmentation and post-tagmentation clean up, a limited cycle polymerase chain reaction (PCR) was used to amplify the DNA and to add adapter sequences to the ends of each fragment, hence generating libraries of DNA. A subsequent Illumina® Purification Beads (IPB) clean up step prepared the libraries for use on an Illumina® sequencing system by purifying the amplified libraries generated. The double stranded DNA libraries were normalised to the appropriate concentration using standard library quantification and quality control procedures, and were then pooled and denatured using freshly prepared sodium hydroxide (NaOH).

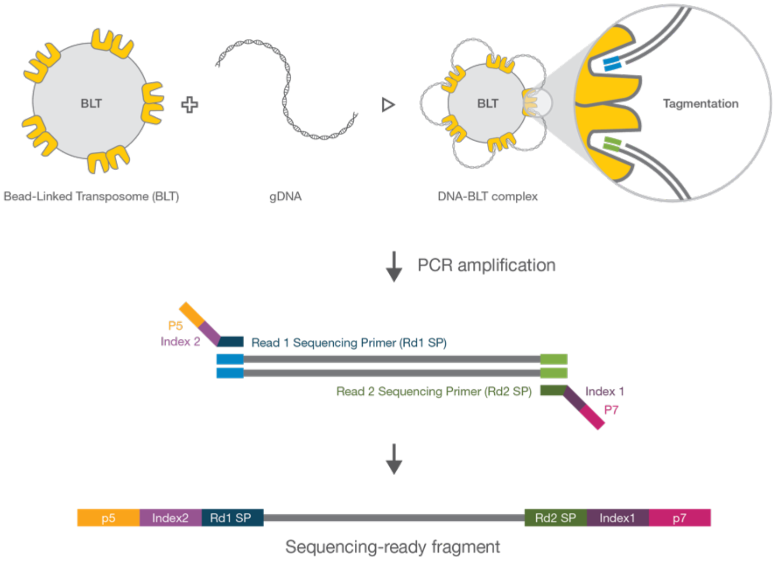

BLT
Bead-Linked Transposome (BLT)
gDNA
BLT
DNA-BLT complex
Tagmentation
PCR amplification
P5
Index 2
Read 1 Sequencing Primer (Rd1 SP)
Read 2 Sequencing Primer (Rd2 SP)
Index 1
P7
p5
Index2
Rd1 SP
Rd2 SP
Index1
p7
Sequencing-ready fragment

Figure 2.5: The library generation procedure using the Illumina® DNA Prep library preparation kit (adapted from Illumina Inc, San Diego, California, USA).

Sequencing was then carried out using the Illumina® NovaSeq™ 6000 System (Illumina, California, USA). The diluted and denatured library pools were first loaded into a cluster cartridge to generate clusters. The single DNA molecules were bound to the surface of a glass flow cell, containing billions of nanowells in an ordered arrangement, and were simultaneously amplified by extending by one base per cycle to form clusters. The clusters were imaged using bidirectional scanning and two-channel sequencing chemistry, that makes use of cameras with sensors that detect red and green light. The intensity values for each cluster in a given image were determined and base calling was performed to identify the base for every cluster in a given tile, based on the ratio of red to green signal detected. This process was repeated for each cycle of sequencing. Real time data analysis was carried out to monitor the progress of the run and to generate quality scores (Q-score) for each base call. Q-scores were used to predict the probability of an incorrect base call, with a high Q-score value representing a high-quality base call which is less likely to be incorrect. The raw data files were then analysed.

2.6.1 Data Analysis of WGS

The primary sequencing output files have a FASTQ format, which are text-based file formats that contain the DNA sequencing data from the clusters that pass the filters on a flow cell. Each record in a FASTQ file consists of four lines which give information about each cluster, including a unique sequence identifier for the sequence read, the actual DNA sequence for that read and the base call quality scores, which reflect the confidence level of the base calls in the sequence data. Data analysis of the FASTQ files involves generating a pipeline with a series of processing steps that identify variants of interest, while eliminating any errors or artifacts that can affect the accuracy and reliability of the variant calls. The raw sequencing files were analysed at Dante Genomics facilities (Italy) using the advanced analysis pipeline on the Illumina® DRAGEN™ Bio-IT platform. Although the specific steps involved in analysing raw sequencing files vary depending on the pipeline used, the ultimate goal of obtaining a final variant call format (VCF) file remains constant.

The GATK (Genome Analysis Toolkit) best practices workflow is a highly regarded benchmark for processing genomic data and detecting high quality variants and is frequently consulted as a reference when constructing a data analysis pipeline. It has been developed and optimised by a team of experts at the Broad Institute to ensure reliable and reproducible data processing, while maximising variant calling accuracy and sensitivity. The workflow incorporates steps that are specifically designed to address common sequencing challenges and to correct common sources of noise and bias. Extensive validation against large-scale datasets and gold standard reference datasets demonstrates the reliability of the GATK best practices workflow (DePristo, Banks et al. 2011; Van der Auwera, O'Connor 2020).

The data analysis process started off by conversion of the raw unmapped FASTQ files to an unaligned BAM (binary alignment and map) file that is in a binary format. The validity of the BAM file was checked to ensure it meets the necessary quality standards, and adapter trimming tags were added to mark the location of the adapter sequences. Quality control of the raw sequencing reads was carried out and the BAM file was converted back to a FASTQ file so that alignment could be carried out. GATK best practices workflow recommends using the Burrows-Wheeler Alignment (BWA) tool (Li, Durbin 2009) to map the reads to the latest build of the human reference genome, named GRCh38 (Genome Research Consortium human build 38), to get the aligned BAM file. However, the analysis pipeline of the Illumina® DRAGEN™ Bio-IT platform involved mapping the reads to the human reference genome GRCh37.

After mapping, PCR duplicates were marked and removed to eliminate any potential artifacts arising from library preparation. Subsequently, base quality scores were recalibrated to correct for systematic errors in the sequencing data. Additionally, local realignment was performed to address small indels that may have been missed during the initial mapping stage, thus enabling accurate mapping of the reads and improved precision in variant calling. Quality control tools were utilised to analyse the aligned BAM files and generate multiple metrics, including the number of aligned reads, mean coverage, coverage distribution, GC bias, and contamination rate. These metrics are crucial for assessing data quality and detecting potential issues in alignment and coverage, such as low coverage, high duplication rates, or GC bias. Statistics on the level of coverage achieved for different regions of the genome, such as exons and GC-rich regions, were also calculated to ensure uniform coverage across the genome. The resulting BAM files were then ready for downstream analysis.

After obtaining the analysis-ready BAM files, the GATK Best Practices Workflow recommends using a variant calling tool, such as HaplotypeCaller, to identify genomic variants in each sample. This tool performed local *de novo* assembly of haplotypes to improve sensitivity and specificity of variant calling, especially for complex variants such as indels and structural variants. The resulting output file was in GVCF (Genomic VCF) format and contained information about all possible variant sites.

Sample specific GVCFs were subsequently joint genotyped to obtain one, full cohort raw and unfiltered VCF. Such VCF file contained variant-specific information for all samples under study. This is vital for further downstream processing as it gives zygosity information for all samples for each variant, while also improving accuracy on low-frequency variant calls. This will allow direct comparison between different samples when eventually implementing zygosity filtering.

Subsequently, multiallelic sites were identified and each separate variant pertaining to the same site was separately considered as a different variant in the raw unfiltered VCF files. VCF files were split into two separate files, with one file containing only single nucleotide polymorphisms (SNPs) and another containing only small insertions/deletions (Indels). A hard filtration step ensued to filter variants in both files based on quality metrics, specific for both variant types according to recommended GATK practices. After the filtering step, the SNP and indel files were consolidated.

The variant effect predictor (VEP) functional annotation tool was then used to annotate the VCF file. This annotation step provides additional information on the functional impact of variants and their allele frequency in population databases, which enables the identification of potential functional consequences of the identified variants. Upon obtaining analysis-ready VCF file, different filtering techniques were applied to identify variants of interest as described in Section 3.4.1 .

2.7 Globin Gene mRNA Expression Assay

The mRNA expression of the α-globin (*HBA1*), β-globin (*HBB*) and γ-globin (*HBG1*) genes, as well as the expression of the housekeeping gene glyceraldehyde-3-phosphate dehydrogenase (*GAPDH*) in the whole blood samples of the family members was measured using the Invitrogen™ QuantiGene™ Plex Assay kit (Thermo Scientific, USA). This assay combines branched DNA (bDNA) technology, that captures and quantifies target-specific RNA using probe hybridisation methods that amplify signal rather than target, with the Luminex™ xMAP™ multi-analyte technology that allows for multiplex gene expression analysis. Colour coded fluorescent magnetic microspheres conjugated with a custom oligonucleotide probe set were used to capture specific target RNA sequences. The probe set consisted of three types of target specific probes; the capture extenders which help in differentiating between different capture beads, the label extenders which support bDNA signal amplification, and the blocking probes. Sequential hybridisation of bDNA oligonucleotides (pre-amplifier and amplifier) occured to each amplification unit, including the final hybridisation to the biotinylated label probe that binds Streptavidin-conjugated R-Phycoerythrin (SAPE) (Figure 2.6). The resulting fluorescence signal was measured by the Luminex™ instrument and was reported as a median fluorescence intensity (MFI). This signal was proportional to the amount of target RNA present in the samples.

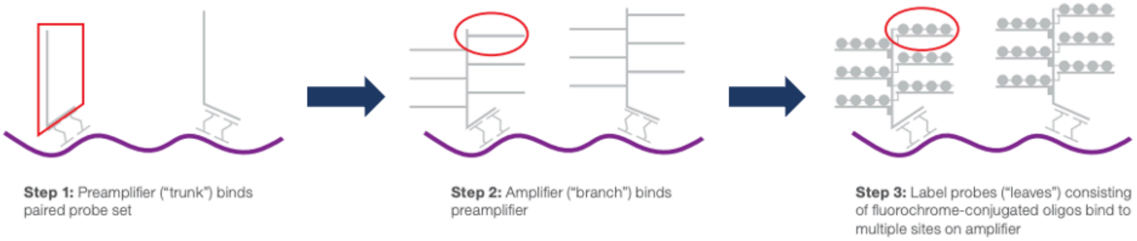

Step 1: Preamplifier ("trunk") binds paired probe set
Step 2: Amplifier ("branch") binds preamplifier
Step 3: Label probes ("leaves") consisting of fluorochrome-conjugated oligos bind to multiple sites on amplifier

Figure 2.6: The bDNA technology used in the Invitrogen™ QuantiGene™ Plex Assay kit, showing sequential hybridisation of pre-amplifier oligonucleotides, amplifier oligonucleotides and labelled probes to amplify the fluorescence signal from each amplification unit (adapted from Thermo Scientific, USA).

Whole blood lysates were first prepared for all the samples using the Invitrogen™ QuantiGene™ Sample Processing kit (Thermo Scientific, USA) starting from frozen blood. The Lysis Mixture was firstly pre-warmed at 37°C for 30 minutes, followed by gentle swirling, while the frozen peripheral whole bloods were allowed to thaw at room temperature. A Whole Blood Working Lysis Mixture was prepared by combining Lysis Mixture, sterile nuclease-free water and Proteinase K in the volumes shown in Table 2.1 below, and the mixture was vortexed.

Table 2.1: The volumes used to prepare the Whole Blood Working Lysis Mixture.

Component	**Volume added per reaction (µl)**
Lysis Mixture	32µl
Sterile nuclease-free water	50µl
Proteinase K	2µl
Final volume	**84 µl**

Since a total of two technical replicates were analysed for each sample, a volume of 168µl Whole Blood Working Lysis Mixture was added to 24µl of whole blood in a 1.5ml microcentrifuge tube. The mixture was vortexed immediately for 1 minute, followed by incubation at 60°C for 1 hour in a shaking incubator set at 275 rpm (revolutions per minute). The lysates were stored at -80°C until further use.

For the first day of the QuantiGene™ Plex Assay, the frozen lysates were removed from the freezer and thawed at room temperature, followed by an incubation at 37°C for 30 minutes and brief vortexing. The reagents of the Invitrogen™ QuantiGene™ Plex Assay kit (Thermo Scientific, USA) were prepared prior to use. The Lysis Mixture was prewarmed at 37°C for 30 minutes, after which it was gently swirled. The Probe Set and Blocking Reagent were thawed and mixed by brief vortexing, and the Probe Set was centrifuged briefly to collect the contents at the bottom of the tube. The Capture Beads were taken out of storage right before use and were sonicated for 3 minutes, followed by vortexing and brief centrifugation. Care was taken to keep the capture beads away from light. The Proteinase K was kept on ice. The samples were also diluted by a factor of 2 using a Diluted Lysis mixture made up of 1 volume of Lysis Mixture and 2 volumes of nuclease-free water.

The appropriate volume of Working Bead Mixture was prepared by combining nuclease-free water, prewarmed Lysis Mixture, Blocking Reagent, Proteinase K, Capture Beads and Probe Set, in the order as shown in Table 2.2. The resulting Working Bead Mix was protected from light by covering with foil. The Working Bead Mix was vortexed for 10 seconds and a volume of 20µl was pipetted into each well of the Hybridisation Plate. A micropipette was then used to transfer 80µl of each diluted lysate into separate wells containing the Working Bead Mix, such that the final volume in each well was 100µl. Three background controls were also included by adding 80µl of the diluted Lysis Mixture to 3 wells containing the Working Bead Mixture. The Hybridisation Plate was completely sealed using a pressure seal and was incubated at 54°C for 18-22 hours in a shaking incubator set at 600 rpm.

Table 2.2: The volumes used to prepare the Working Bead Mixture.

Reagent	1 Well (µl)	96 Well (µl)
Nuclease-free Water	5.2	624
Lysis Mixture	6.6	792
Blocking Reagent	2	240
Proteinase K	0.2	24
Capture Beads (vortexed 30 seconds before adding)	1	120
Probe Set	5	600
Total	**20**	**2,400**

On the second day of the QuantiGene™ Plex Assay, the Luminex™ protocol was set up on the Luminex™ instrument using the following parameters shown in Table 2.3.

Table 2.3: The parameters used to set up the Luminex™ protocol.

Sample Size	DD Gate	Timeout	Bead event/ Bead region
100µl	5,000-25,000	45 sec	100

The Pre-Amplifier Solution, Amplifier Solution and Label Probe Solution were warmed at 37°C for 30 minutes to dissolve any precipitates, were mixed well by inversion before use and were left at room temperature until used. The SAPE diluent was also brought to room temperature. A 1X Wash Buffer was prepared by mixing 0.6ml Wash Buffer Component 1, 10ml Wash Buffer Component 2 and 189ml nuclease-free water, which is sufficient for 1 plate (Table 2.4).

Table 2.4: The Volumes used to prepare the 1X Wash Buffer.

1X Wash Buffer	1 Well (ml)	96 Well (ml)
Wash Buffer Component 1	0.006	0.6
Wash Buffer Component 2	0.104	10
Nuclease-free water	1.97	189

Following overnight incubation, the Hybridisation Plate was removed from the shaking incubator and was centrifuged at 240 *g* for 1 minute at room temperature. The pressure seal was removed, and each sample was pipetted up and down for 5 times before completely transferred into a Magnetic Separation Plate. The Magnetic Separation Plate was placed into the handheld magnetic plate washer for 1 minute to allow the magnetic beads to accumulate on the bottom of each well, before removing the solution in the wells by quick inversion over a waste container. The plate was washed for a total of 3 times by adding 100µl of 1X Wash Buffer into each well. The magnetic beads were allowed to accumulate on the bottom of each well by waiting 15 seconds between each wash, after which the plate was inverted swiftly to remove the solution and blotted gently using paper towels.

Using a sterile micropipette, 100µl of Pre-Amplifier solution were pipetted into each well and the plate was sealed well. The Magnetic Separation Plate was removed from the handheld magnetic plate washer, shaken at 800 rpm for 1 minute at room temperature to resuspend the capture beads, and was subsequently incubated for 1 hour at 50^{o}C in a shaking incubator set at 600 rpm. After incubation with the Pre-Amplifier solution, the Magnetic Separation Plate was placed into the handheld magnetic plate washer for 1 minute before removing the solution in the wells by quick inversion over a waste container. The plate was washed for a total of 3 times by adding 100µl of 1X Wash Buffer into each well and waiting for 15 seconds between each wash before removing the solution in the wells by plate inversion.

A volume of 100µl of Amplifier solution was pipetted into each well, and again the plate was sealed, shaken at 800 rpm for 1 minute at room temperature and incubated for 1 hour at 50ºC in a shaking incubator set at 600 rpm. This was followed by transferring the plate onto the handheld magnetic plate washer, waiting 1 minute before removing the solution by inversion and performing 3 consecutive washes using 100µl 1X Wash Buffer. The beads were allowed to accumulate at the bottom of the well for 15 seconds after each wash, and the plate was inverted swiftly to remove the solution and blotted gently using paper towels.

Hybridisation with the label probes was carried out by transferring 100µl of Label Probe solution into each well, shaking at 800 rpm for 1 minute at room temperature and then incubating the plate in a shaking incubator for 1 hour at 50ºC with shaking at 600 rpm. After the 1 hour incubation, the hybridisation plate was transferred onto the handheld magnetic plate wash and the residual Label Probe solution was rinsed off by washing each well 3 times with 100µl 1X Wash Buffer, with 15 seconds waiting time between each wash to allow for bead accumulation.

The SAPE working reagent was prepared in a 15ml tube by adding 36µl of SAPE and 12ml SAPE Diluent and the tube was covered with foil to protect from light. A volume of 100µl of SAPE Working Reagent was transferred into each well. The Magnetic Separation Plate was sealed well, covered with foil to protect from any light, placed on a shaking platform at room temperature and shaken at 800 rpm for 1 minute, followed by shaking at 600 rpm for 30 minutes. The plate was transferred into the handheld magnetic plate washer and was washed 3 times using 100µl 1X Wash Buffer, with 15 seconds waiting on the handheld magnetic plate washer between each wash to allow accumulation of the magnetic beads.

Finally, a volume of 130µl SAPE Wash Buffer was added into each assay well and the plate was sealed well and covered with foil to protect from any light. The Magnetic Separation Plate was placed on the Microtiter Plate Shaker, shaken at 800 rpm for 3 minutes and analysed immediately on the Luminex™ instrument. The normalised mRNA expression levels were then calculated for each target gene in each sample.

2.7.1 Globin Gene mRNA Expression Data Analysis

The raw data obtained from the Invitrogen™ QuantiGene™ Plex Assay was analysed to quantify the expression of each target gene in each sample. First, the average background value for each gene was calculated by averaging the background signal values of the probes that corresponded to the gene. Next, the average background signal of each gene was subtracted from the sample values of that same gene, so that the resulting value represented the normalised signal value of the target gene in the sample. The expression of the *GAPDH* housekeeping gene was used as an internal control to normalise the samples by correcting for any variations in the amount of RNA present in each sample. The normalised gene expression for each target gene in the samples was calculated by dividing the signal value of the target gene in the sample by the *GAPDH* housekeeping gene. Finally, statistical analyses were performed on the normalised gene expression values using Statistical Package for Social Sciences (SPSS) version 26 (SPSS, Chicago, USA) to determine if there were any significant differences in mRNA gene expression between the samples.

2.8 Proteomics by Mass Spectrometry

Proteomics of the red cells by liquid chromatography-mass spectrometry (LC-MS)/mass spectrometry (MS) was also carried out at the Sanquin Research Facility in Amsterdam

(Netherlands). Packed red cell concentrate were prepared by transferring 1ml of whole blood into labelled 1.5ml Eppendorf tubes, followed by centrifugating the tubes at 800 *g* for 10 minutes at room temperature. A volume of 100µl packed red cells were transferred into labelled Eppendorf tubes and washed 5 times with 1ml PBS. The cells were vortexed and centrifuged at 650 *g* for 2 minutes between each wash and the supernatant was completely discarded each time. The red cell pellets were snap frozen in liquid nitrogen and stored at -20°C freezers until processed on a mass spectrometer.

Prior to starting MS, the frozen red cell pellets were thawed and 3x10^6 erythrocytes were isolated and subsequently lysed using a buffer comprising Tris(2-carboxyethyl) phosphine (TCEP), chloroacetamide, sodium deoxycholate (SDC), and Tris-HCl buffer (pH 8) and heated at 95^{o}C for 5 minutes. The protein concentration of the lysates was determined and 10µg of total protein was subjected to overnight trypsin digestion. After SDC precipitation, 500ng of the resulting peptides were loaded onto Evotips Pure™ (Evosep, Denmark), which are disposable trap columns with integrated purification and sample loading capabilities, that increase peptide identification and allow for a deep proteome coverage. The samples were then analysed on the Orbitrap Fusion™ Lumos™ Tribrid™ Mass Spectrometer (Thermo Scientific, USA) using the Evosep One liquid chromatography (LC) system (Evosep, Denmark) with a 15cm performance column. Data independent acquisition (DIA) was carried out during the run and the resulting raw data were analysed using the universal DIA-NN automated software for DIA proteomics data analysis (Demichev, Messner et al. 2019). Perseus software (Tufts University, Massachusetts) was used for missing value imputation and statistical analysis was carried out using Python software (Version 3.10).

3. Results

3.1 Families with *KLF1* p.K288X mutation

In this study, analysis was performed on a cohort consisting of 22 individuals from three distinct families affected with HPFH and known to harbour the *KLF1* p.K288X mutation (Figure 3.1 and Figure 3.2). All individuals who participated were known to be absent from mutations in the α- and β-globin genes from previous studies, excluding individuals F1.1 and F1.5 from Fam F1 who are heterozygote carriers of the Hb St.Luke's variant due to a mutation in the α-globin gene.

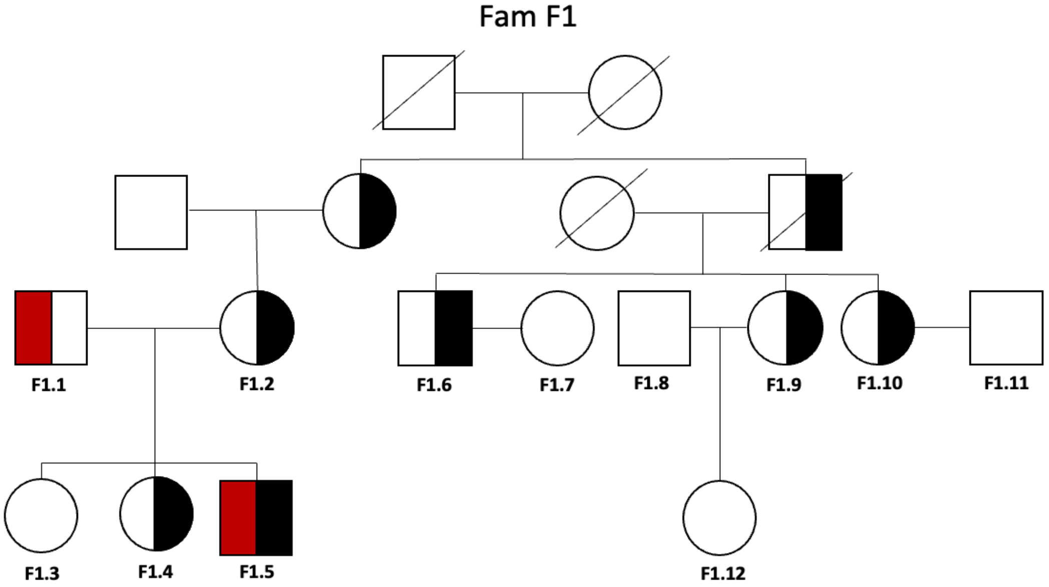


Figure 3.1: The family tree of Fam F1, showing the 12 recruited subjects where individuals affected with HPFH due to the *KLF1* p.K288X mutation are shown as half filled black symbols, heterozygote members for Hb St. Luke's are shown by half-filled red symbols, while unaffected family members are shown as open figures.

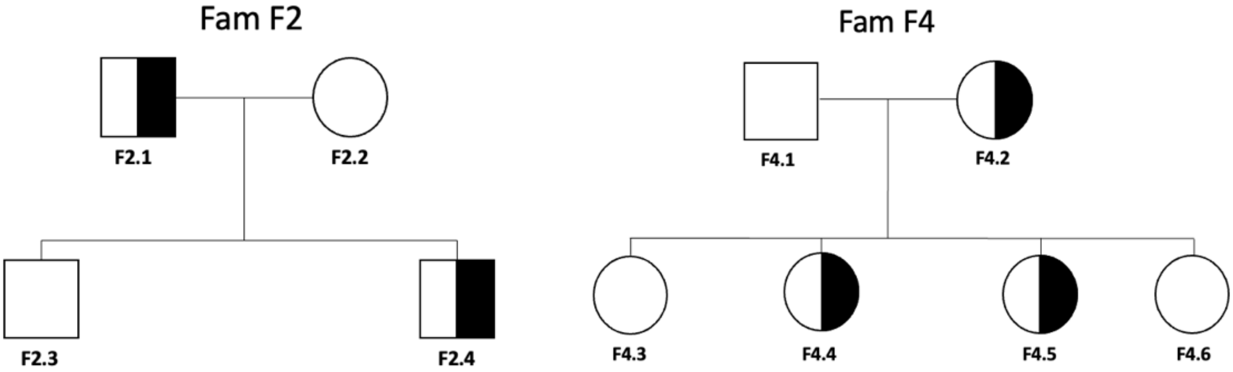


Figure 3.2: The family trees of Fam F2 and Fam F4, showing the recruited subjects where individuals affected with HPFH due to the *KLF1* p.K288X mutation are shown as half filled black symbols while unaffected family members are shown as open figures.

All 22 blood samples collected were subjected to a complete blood count (CBC) to measure various haematological parameters. The original Fam F1 was expanded to include subject F1.12 who presented with normal haematological parameters, with an HbF level of 0.4% and an HbA_2 level of 2.6%. WGS confirmed that F1.12 was homozygous wildtype for the p.K288X pathogenic mutation. In Fam F4, the proband's father (F4.1) and siblings (F4.3 and F4.6) displayed normal haematological parameters and exhibited no variations in the *KLF1* gene following WGS analysis. However, the proband's mother (F4.2) exhibited elevated levels of HbF (3.0%) and presented with HbA_2 levels of 2.8%, MCV of 83.9fL, and mean corpuscular haemoglobin (MCH) of 25.7pg. WGS revealed the presence of the pathogenic p.K288X mutation in the *KLF1* gene in subject F4.2. Consequently, this study comprised 11 individuals who classified as healthy controls and an additional 11 individuals who identified as carriers of the *KLF1* p.K288X variation.

The CBC results revealed distinct differences between individuals with the *KLF1* p.K288X mutation and the control group (Table 3.1). Specifically, individuals with the mutation exhibited higher levels of HbF and HbA_2. Individuals within Fam F1 who carried the *KLF1* variant demonstrated significantly elevated HbF levels when compared to the other two families. Among these individuals, F1.2 of Fam F1 displayed the highest concentration of HbF, measuring 6.4%. In contrast, the control group exhibited HbF levels ranging from 0.3% to 0.7%. Furthermore, individuals from Fam F2 and Fam F4 displayed the highest levels of HbA_2, with F4.5 from Fam F4 exhibiting the highest HbA_2 level of 3.6%. Based on the observed HbF levels, the individuals with the *KLF1* p.K288X truncation mutation were categorised into two groups: those with low HbF levels (≤3%) and those with moderate HbF levels (>3%).

Table 3.1: Table showing the CBC parameters of all individuals in families Fam F1, Fam F2 and Fam F4. Individuals marked in bold red were classified as having moderate levels of HbF (%). Individuals carrying the *KLF1* p.K288X mutation in a heterozygous state are shown as +/-, while the individuals who lacked the mutation are shown as -/-.

Family	**Member**	**Hb (g/dL)** (12.0-17.2g/dL)	**HbF (%)** (0.8-2.0%)	**HbA_2 (%)** (2.0-3.0%)	**MCV (fL)** (79.0-97.0fL)	**MCH (pg)** (27.0-32.0pg)	**MCHC (g/dL)** (32.0-36.0g/dL)	**RDW (%)** (11.9-14.6%)	***KLF1* p.K288X status**
Fam F1	F1.1	14.9	0.5	2.3	85.7	29.3	34.2	12.7	-/-
	F1.2	11.5	6.4	2.7	81.1	26.1	32.2	15.2	+/-
	F1.3	12.7	0.7	2.5	79.9	25.6	32.0	14.2	-/-
	F1.4	12.2	2.8	3.2	80.0	27.2	34.0	13.2	+/-
	F1.5	12.8	4.6	2.5	70.5	23.1	32.8	15.6	+/-
	F1.6	14.4	2.2	3.3	82.1	27.6	33.6	13.1	+/-
	F1.7	15.1	0.3	2.8	88.3	30.0	34.0	11.7	-/-
	F1.8	16.6	0.3	2.7	89.4	30.4	34.0	13.2	-/-
	F1.9	12.3	3.4	2.8	78.2	24.8	31.7	14.6	+/-
	F1.10	13.6	5.6	2.9	79.5	25.1	31.6	14.3	+/-
	F1.11	15.9	0.4	3.0	90.9	29.5	32.4	12.6	-/-
	F1.12	12.4	0.4	2.6	84.3	27.1	32.1	13.0	-/-
Fam F2	F2.1	14.9	1.2	3.4	86.7	27.9	32.1	13.2	+/-
	F2.2	12.8	0.3	2.8	91.9	30.3	33.0	11.9	-/-
	F2.3	15.7	0.6	3.1	93.2	29.8	32.0	12.0	-/-
	F2.4	15.5	1.4	3.4	85.3	27.1	31.8	13.5	+/-
Fam F4	F4.1	15.2	0.4	2.8	90.3	29.4	32.5	13.3	-/-
	F4.2	12.1	3.0	2.8	83.9	25.7	30.6	14.5	+/-
	F4.3	12.6	0.5	2.7	89.8	28.6	31.8	12.4	-/-
	F4.4	13.2	1.6	3.1	81.6	26.4	32.4	15.5	+/-
	F4.5	13.9	1.2	3.6	91.7	30.2	32.9	13.1	+/-
	F4.6	13.7	0.7	2.7	96.5	31.7	32.9	12.2	-/-

To explore potential linear correlations between haematological parameters, Pearson correlation coefficients (r) were computed. The results suggest a robust positive correlation between HbF levels and red cell distribution width (RDW; r = 0.727), demonstrating that the red cell distribution width increased as HbF levels increased (Figure 3.3). Conversely, the values suggest that strong negative correlations were present between HbF levels and MCV (r = -0.698) as well as MCH (r = -0.726), indicating that both parameters decreased with increasing HbF levels (Figure 3.4).

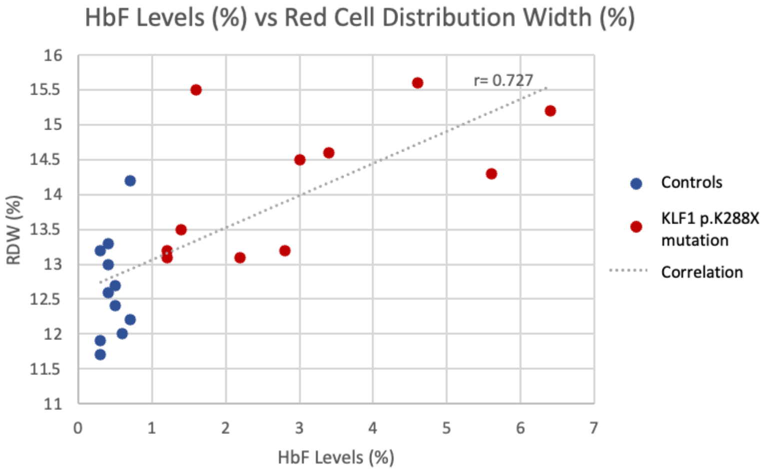


Figure 3.3: Scatter plot to show the strong positive correlation between HbF levels and the red cell distribution width for healthy controls (blue) and individuals with the *KLF1* p.K288X mutation (red).

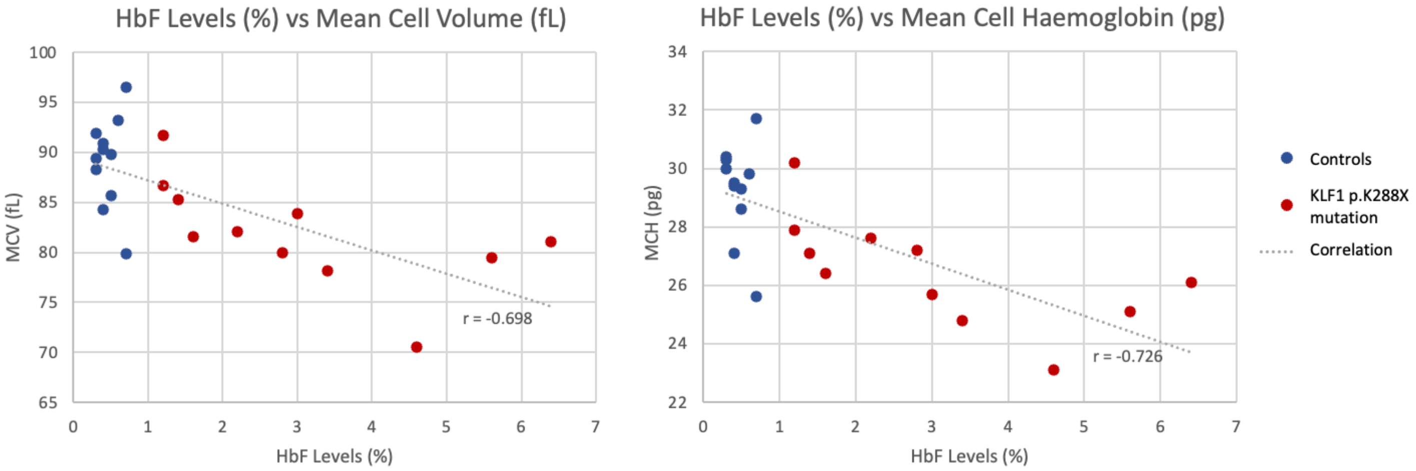


Figure 3.4: Scatter plots to show the strong negative correlation between HbF levels and the mean corpuscular volume (left) and the mean corpuscular haemoglobin (right) for healthy controls (blue) and individuals with the *KLF1* p.K288X mutation (red).

3.2 Flow Cytometry Data Analysis

Data analysis of the raw fluorescence signals and scattered light measured by the BD FACSCanto™ II Flow Cytometer (BD Biosciences) was carried out using the FlowJo™ software (BD Life Sciences). After importing the data files into the software, gating strategies were implemented based on forward scatter (FSC) and side scatter (SSC) parameters to enable accurate analysis of erythrocytes, single cells, and to remove any background signal. Initially, erythrocytes were gated by displaying the FSC and SSC parameters on a scatter plot and drawing a gate around the erythrocyte cluster to include only these events in subsequent analyses. Gating for single cells involved visualising the FSC height (FSC-H) and FSC area (FSC-A) parameters on a scatter plot and drawing a gate around the single cell cluster to exclude doublet cells and debris (Figure 3.5).

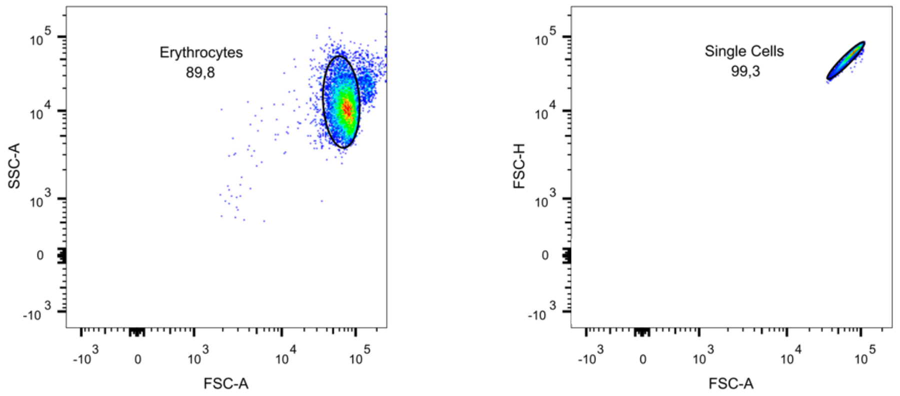


Figure 3.5: Gating strategy for erythrocytes (left) and for single cells (right) in flow cytometry analysis.

Unstained samples acted as negative controls for removing background signal. Fluorescence channels of interest (FITC and APC) were displayed on histograms, and a gate was defined to exclude events with background fluorescence by identifying the region of minimal or no fluorescence (Figure 3.6). These gating strategies were consistently applied to all samples, ensuring uniformity in the analysis. A comprehensive batch report summarising the gating results was generated for all the samples, providing an extensive overview of the gating strategy. This report facilitated efficient data interpretation and allowed for comparisons to be made across different antigens and sample groups.

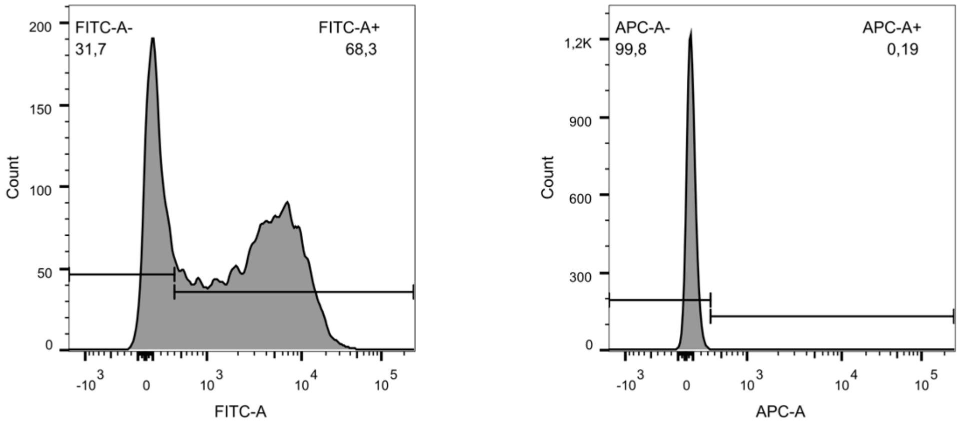


Figure 3.6: Gating strategy for background signal removal in the FITC channel (left) and APC channel (right) using unstained samples.

For each of the three antigens tested, the fluorescence intensity data from all the samples were plotted on the same graph. Two sets of overlay graphs were then created to analyse antigen expression patterns. The first set encompassed data from Fam F1, while the second set included the data from Fam F2 and Fam F4. By superimposing the individual histograms for each sample group, a clear distinction in the antigen expression patterns was observed between individuals that had the p.K288X mutation in the *KLF1* gene and the healthy controls. These discernible differences highlight the presence of unique expression profiles associated with the *KLF1* p.K288X mutation (Figure 3.7). Three distinct histograms were also generated for each participant involved in this study to demonstrate the unique antigen expression profiles of the three tested antigens (BCAM, CD44 and P1). These graphs are available in Appendix A.

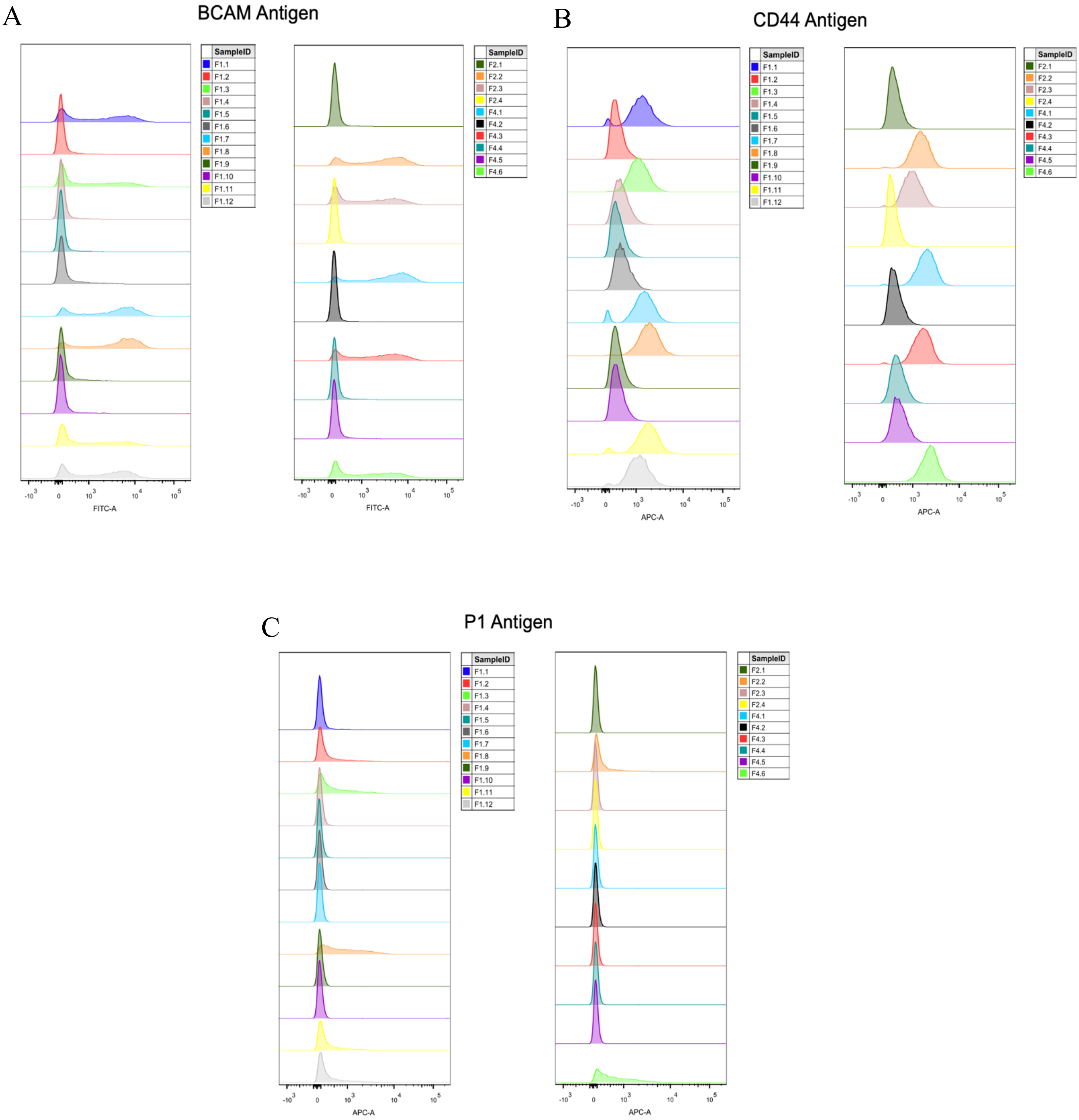


Figure 3.7: Overlay graphs to show antigen expression patterns in individuals having the *KLF1* mutation and healthy controls from Fam F1, Fam F2 and Fam F4; where figure A shows the overlay graphs for BCAM antigen, figure B shows the overlay graphs for CD44 antigen and figure C shows overlay graphs for P1 antigen.

Significant differences in antigen expression patterns were evident between the healthy control group and individuals with the *KLF1* mutation, particularly for the BCAM and CD44 antigens (Figure 3.7 A and B). Individuals carrying the *KLF1* mutation exhibited noticeably reduced levels of both BCAM and CD44 antigens, as indicated by a leftward shift in the overlay graphs. In contrast, the P1 antigen demonstrated the least variation in expression levels between affected individuals and healthy controls (Figure 3.7 C).

To further understand the prevalence of antigen levels within each group, the percentage of positive cells for each antigen of interest was subsequently quantified by assessing the proportion of cells within the gated populations (Appendix B). Bar charts were generated for each of the three antigens, providing a visual representation of the differences in the percentage of positive cells among individuals (BCAM: Figure 3.8; CD44: Figure 3.9; P1: Figure 3.10).

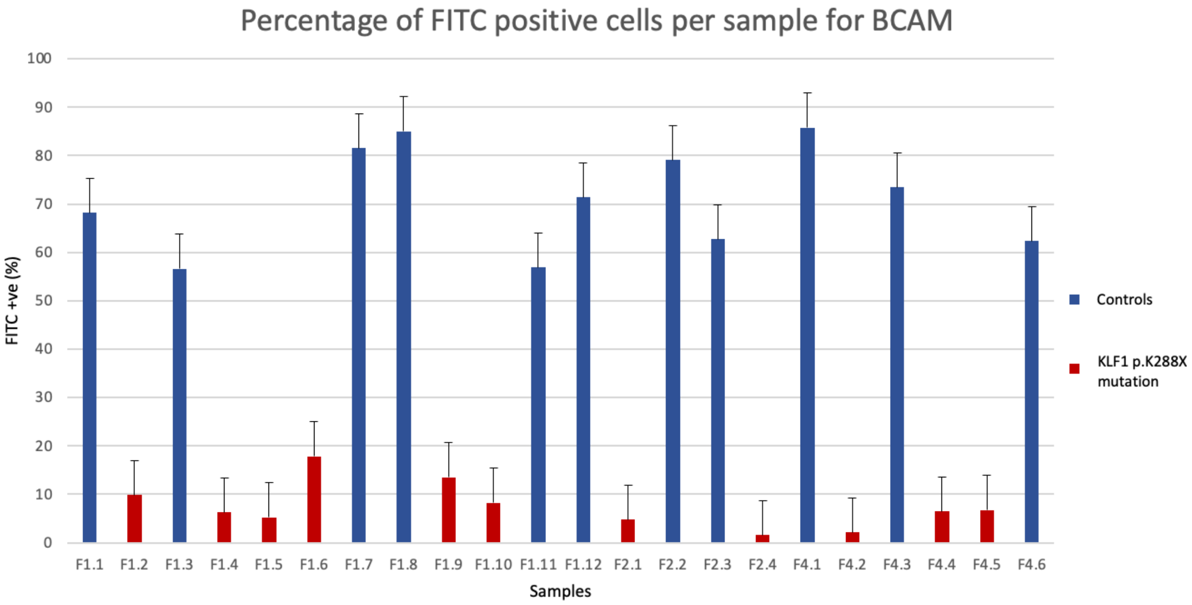


Figure 3.8: The percentage of FITC positive cells detected for each sample for the BCAM antigen, showing the healthy controls in blue and the individuals with the *KLF1* p.K288X mutation in red.

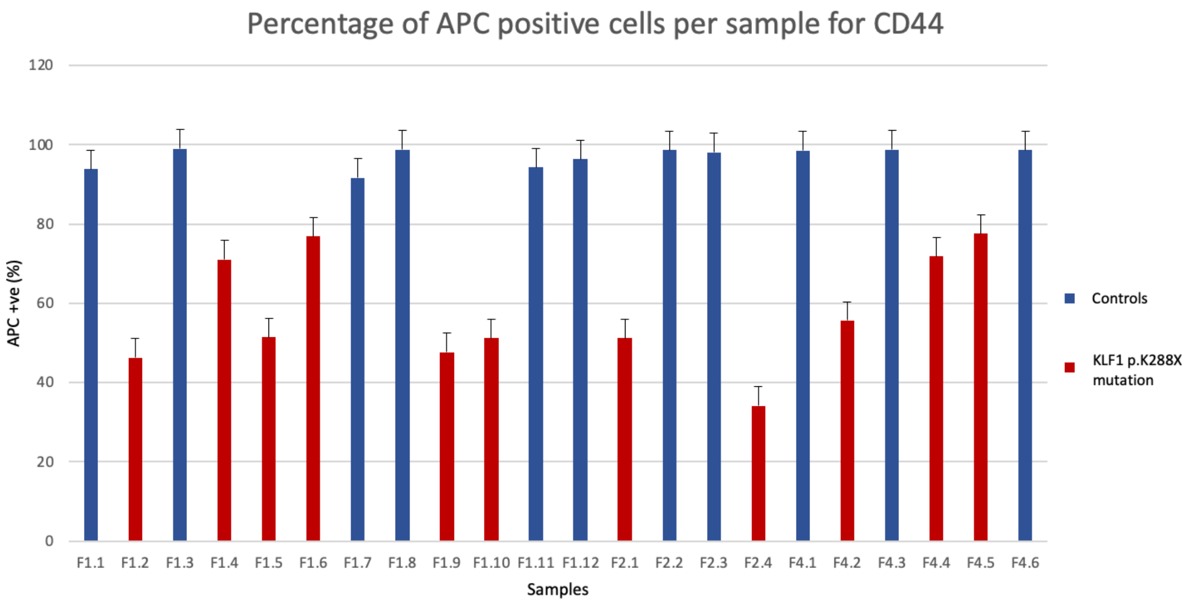


Figure 3.9: The percentage of APC positive cells detected for each sample for the CD44 antigen, showing the healthy controls in blue and the individuals with the *KLF1* p.K288X mutation in red.

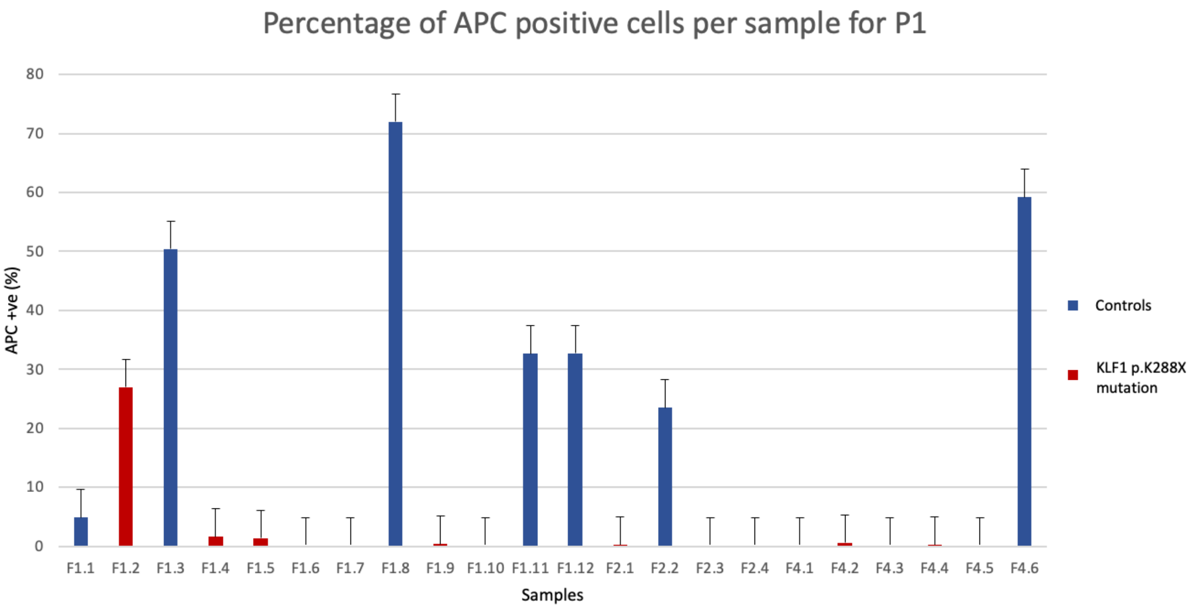


Figure 3.10: The percentage of APC positive cells detected for each sample for the P1 antigen, showing the healthy controls in blue and the individuals with the *KLF1* p.K288X mutation in red.

The most notable difference in the percentage of positive cells between the individuals having the *KLF1* variant and the healthy controls was observed for BCAM and CD44 antigens. For the BCAM antigen, healthy controls exhibited a higher levels of BCAM antigen as indicated by a higher percentage of positive cells, while individuals carrying the *KLF1* p.K288X mutation had a very low percentage of positive cells (Figure 3.8). CD44 antigen displayed a similar expression levels to that observed with the BCAM antigen, where KLF1 deficient individuals had a reduced number of percentage positive cells when compared to the healthy controls (Figure 3.9). However, the reduction in positive cells was not as significant as that observed with the BCAM antigen. P1 antigen, on the otherhand, was the least affected in KLF1 deficient individuals and the controls, since both groups had cases of very low percentage positive cells (Figure 3.10). This indicates that P1 antigen is the least influenced by the *KLF1* p.K288X mutation.

Moreover, the geometric mean fluorescence intensity (gMFI) values were calculated for each antigen in both the control and KLF1 deficient groups (Appendix B) and scatter graphs were generated to visualise the results. These intensity values served as quantitative indicators of antigen levels, enabling a precise comparison between the two groups. As observed in all three scatter plots, the gMFI of the controls was consistently higher than those of the individuals with the *KLF1* p.K288X mutation (BCAM: Figure 3.11; CD44: Figure 3.12; P1: Figure 3.13). This was also supported by a higher average gMFI observed in the controls when compared to the KLF1 deficient group. In Figure 3.13, it can be noted that the P1 antigen gMFI values of the controls were dispersed, contrasting with the more concentrated gMFI values below 100 observed in the KLF1 deficient group. Individual F1.2 from Fam F1 deviated from this pattern, displaying a high gMFI of approximately 300 for the P1 antigen, despite carrying the *KLF1* mutation.

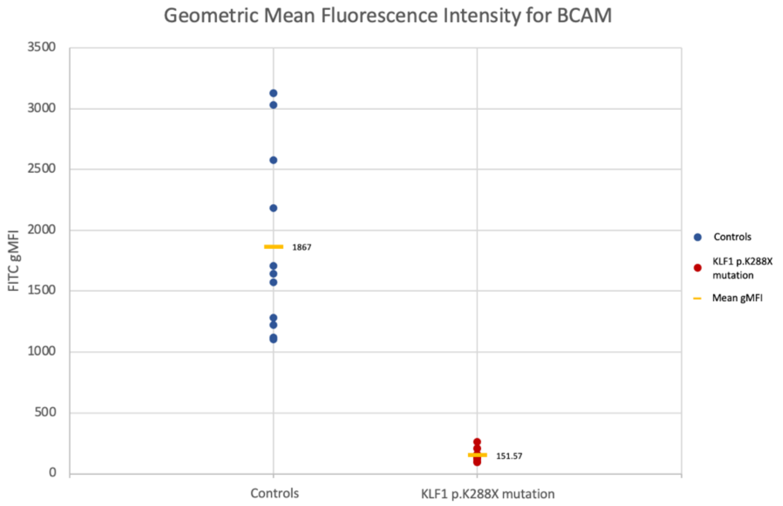


Figure 3.11: The gMFI for BCAM antigen of healthy controls (blue) and individuals with the *KLF1* p.K288X mutation (red), with the mean gMFI for both groups shown in orange.

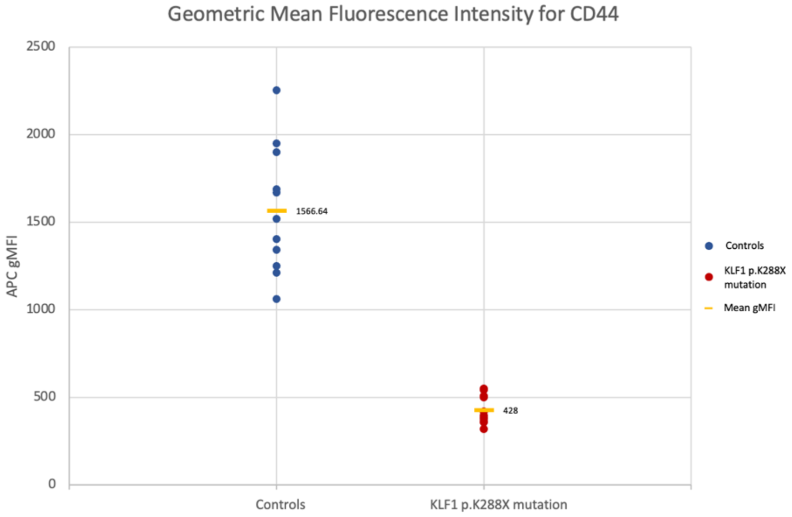


Figure 3.12: The gMFI for CD44 antigen of healthy controls (blue) and individuals with the *KLF1* p.K288X mutation (red), with the mean gMFI for both groups shown in orange.

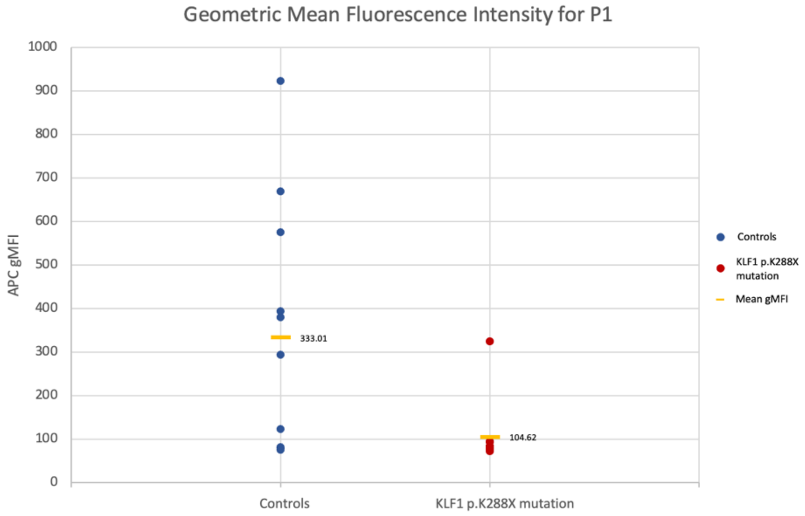


Figure 3.13: The gMFI for P1 antigen of healthy controls (blue) and individuals with the *KLF1* p.K288X mutation (red), with the mean gMFI for both groups shown in orange.

The statistical independent sample T-test was subsequently performed on the data to check if the presence of the *KLF1* mutation affects the antigen levels. A significance threshold of 0.05 (alpha value) was chosen for accepting or rejecting the null hypothesis, which stated that there was no difference in the mean antigen levels between the two groups. The T-test results for BCAM and CD44 antigens yielded p-values of 2.66 x10^{-7} and 2.45 x10^{-9} respectively. Both p-values were below the significance threshold of 0.05, indicating a statistically significant correlation between levels of BCAM and CD44 antigens in individuals with the *KLF1* p.K288X mutation and healthy controls. For the P1 antigen, the T-test resulted in a p-value of 0.019, confirming a significant correlation between the levels of P1 antigen in the two groups. However, this correlation was less statistically significant compared to BCAM and CD44 antigens.

The relationship between the percentage of positive cells and the levels of HbF (%) was explored by generating scatter plots for each antigen against HbF levels (BCAM: Figure 3.14; CD44: Figure 3.15; P1: Figure 3.16). A consistent pattern emerged for BCAM and CD44 antigens, which revealed an inverse association between the antigen levels and HbF levels, where antigen levels decreased with increasing HbF concentration. It was also observed that surpassing a threshold of 1% HbF level, caused a significant reduction in the proportion of positive cells detected. However, subsequent increase in the levels of HbF did not yield in a linear decrease in the percentage of positive cells, resulting in similar proportions among individuals with the *KLF1* mutation, irrespective of varying HbF levels. An exception was noted in P1 expression of individual F1.2 from Fam F1, who exhibited elevated levels of P1 positive cells, despite having a high HbF concentration of 6.4% (Figure 3.16).

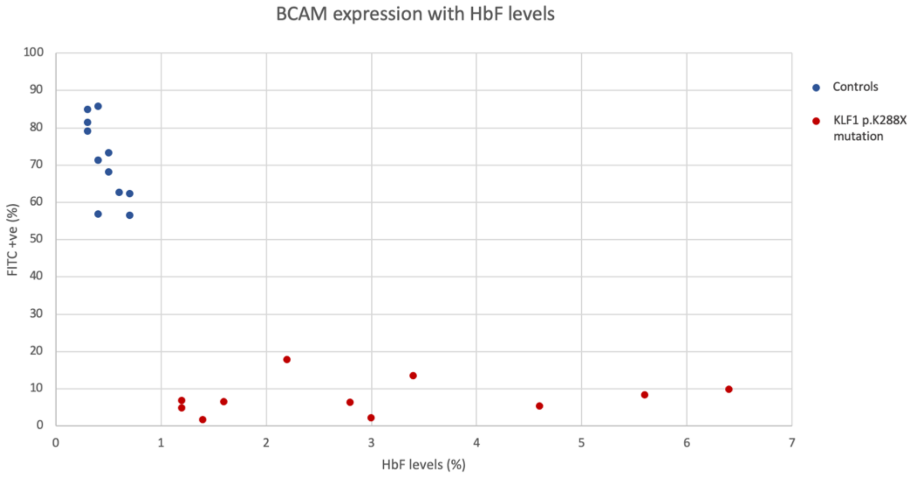


Figure 3.14: The relationship between BCAM positive cells (%) and HbF levels (%) for healthy controls (blue) and individuals with the *KLF1* p.K288X mutation (red).

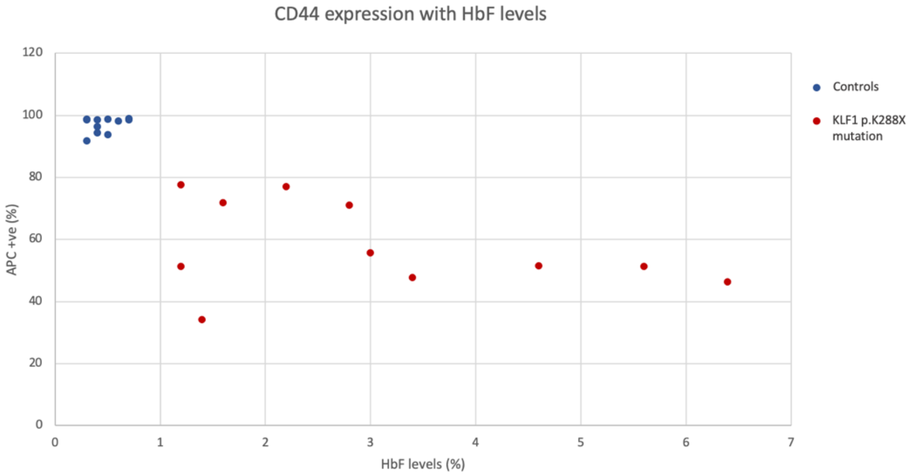


Figure 3.15: The relationship between CD44 positive cells (%) and HbF levels (%) for healthy controls (blue) and individuals with the *KLF1* p.K288X mutation (red).

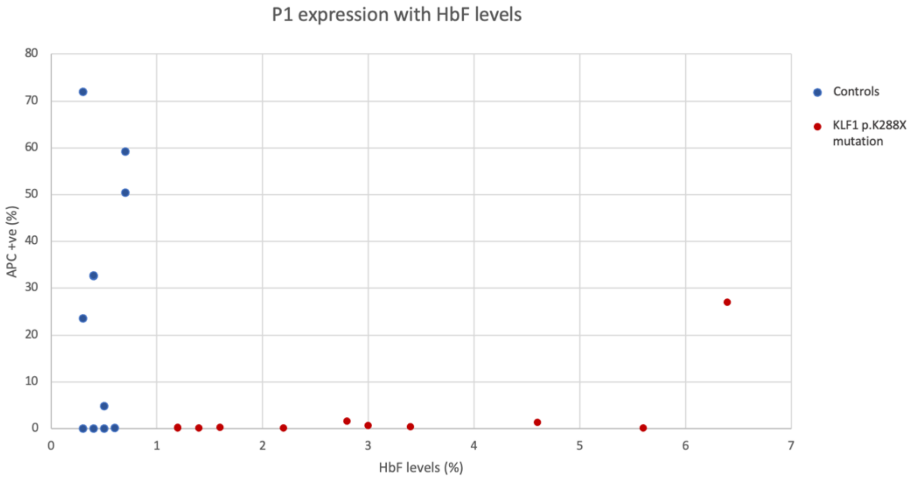


Figure 3.16: The relationship between P1 positive cells (%) and HbF levels (%) for healthy controls (blue) and individuals with the *KLF1* p.K288X mutation (red).

3.3 Checking the quality of the extracted DNA

The quality of the extracted genomic DNA (gDNA) was assessed by loading it onto a 1% agarose gel. The agarose gel was prepared following the instructions provided in Section 2.5.1. The gel was run at a voltage of 120V and a current of 400mA and for 20 minutes. To ensure the absence of contamination, a no template control (NTC) consisting of sterile nuclease free water was included in the run. Additionally, the size marker φX 174 DNA digested with Hae III, which yields 11 fragments ranging from 100 to1000 base pairs (bp) was used. A thick dark band having a high molecular weight was observed for all the samples, confirming a successful DNA extraction process while ascertaining the quality and integrity of the gDNA. The intensity of the band corresponded to the concentration of the DNA in each sample, where a thicker band was observed for samples having a higher concentration of nucleic acids (Figure 3.17).

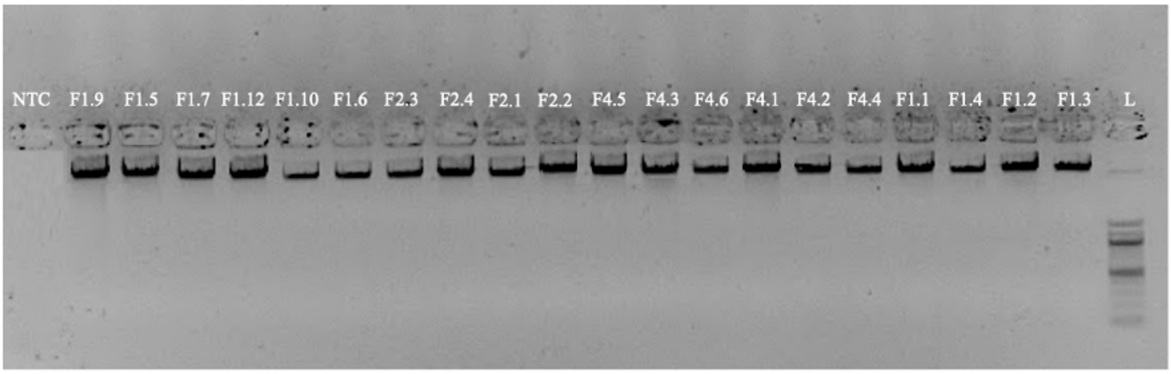


Figure 3.17: Agarose gel electrophoresis showing a thick dark band of high molecular weight for all 20 samples. A DNA ladder (L) is also shown in the last lane and a NTC is present in the first lane.

3.4 Analysis of WGS Data

3.4.1 WGS Data Filtering

Several filtering techniques were employed to refine the selection of variants in the analysis-ready combined and annotated VCF file. The first filtering step involved the application of an allele frequency filter, which excluded variants exhibiting an alternative allele frequency (AAF) equal to or exceeding 15% in any of the four primary populations of interest in the Genome Aggregation Database (gnomAD; Lek, Karczewski et al., 2016). These populations included non-Finnish Europeans (NFE), Africans (AFR), east Asians (EAS), and Americans (AMR). Variants present at a frequency of 15% or higher in any of these populations were considered common and unlikely to be associated with the relatively rare phenotype under investigation and consequently these variants were excluded from further analysis.

The second filtering step implemented a lenient version of the zygosity filter, which proved to be the most effective filter employed. This filter allowed no-calls to be retained to ensure that no valuable variants were erroneously discarded due to processing errors, such as inadequate coverage at specific genomic positions. Lenient zygosity filtering allows the possibility of rescuing potentially relevant variants found on genes of interest, which could have otherwise been missed by more stringent zygosity filters. Distinct zygosity filters were applied depending on the two different modes of inheritance, namely dominant and recessive, considering the entire cohort of individuals (11 affected vs. 9 unaffected). Additionally, separate dominant and recessive zygosity filters were applied in a subgroup analysis, which focused on the four individuals, namely F1.2, F1.5, F1.9 and F1.10, who exhibited the highest levels of HbF (>3%) (4 affected vs. 16 unaffected). These individuals belonged to Fam F1, which has a known history of significantly elevated HbF levels (Borg, Papadopoulos et al. 2010). Variant filters were specifically applied to this subgroup to identify potential causal variants present only in Fam F1 that contribute to the observed higher HbF levels among its affected individuals, despite carrying the same truncation mutation p.K288X in the *KLF1* gene as the two other families with HPFH.

The third filter applied was the coverage filter, which required at least one individual to have a coverage of 20X or higher at the specific variant position for the variant to be accepted. The fourth filter, known as the allele fraction filter, required the presence of an alternate allelic fraction greater than or equal to 10% in at least one of the affected individuals for the variant to satisfy this requirement. The allele fraction refers to the ratio of sequencing reads at a specific variant position that carry the alternative allele in comparison to the reference allele. Variants exhibiting an allelic fraction of less than 10% among the sequencing reads at that particular variant position were deemed inadequately represented and hence disregarded. This filter ensured that variants were not retained solely due to any processing artifacts.

Lastly, the final filter was the 'stringent' version of the zygosity filter, which eliminated all no-calls. This filter, consistent with the two modes of inheritance (dominant and recessive), was applied to both the entire cohort of individuals and also the cohort where the four individuals with the highest HbF levels were treated as affected. This stringent zygosity filter further reduced the list of variants, by retaining with certainty only those variants that were fully confirmed to be present in the cohort of interest.

3.4.1.1 Filtering of Variants following a Dominant Inheritance Pattern for the Full Cohort

A stepwise approach was employed to filter variants within the entire cohort consisting of 11 individuals affected by HPFH and 9 healthy relatives exhibiting a dominant inheritance pattern. The variant filtering process involved the application of zygosity filters designed to identify variants conforming to a dominant inheritance pattern. In the initial zygosity filter, variants exhibiting heterozygous, homozygous alternate, or no-call zygosities across all 11 affected individuals, and homozygous reference or no-call zygosities across all 9 unaffected individuals, were selectively retained. This initial zygosity filter together with the other filters yielded a total of 6,426 unique variants. Subsequently, the more stringent zygosity filter was implemented, where only variants with heterozygous or homozygous alternate zygosities across all 11 affected individuals, and homozygous reference zygosities across all 9 unaffected individuals, were retained. This more stringent filtering step together with the other filters resulted in a total of 205 unique variants that adhered to the dominant inheritance pattern, being present in all 11 affected individuals while being absent in all 9 unaffected individuals (Online Supplementary Data A). Details of the filtering steps performed, along with the number of variants remaining at each stage, are presented in Table 3.2.

Table 3.2: Sequential filtering steps for identifying potential causal variants in the full cohort with a dominant inheritance model.

Filtering steps for the full cohort following a dominant inheritance pattern	Unique Variant Counts
Total number of unique variants in the original annotated VCF file prior to filtering	10,627,339
Unique variants having an observed allele frequency of ≤15% in at least one of the four major gnomAD populations of interest (NFE, AFR, EAS and AMR)	10,597,744
Remaining number of unique variant counts after applying relaxed zygosity filters where no-calls were accepted	9,193
Number of unique variants with a ≥20X coverage in at least one of the affected individuals	6,448
Remaining unique variants after filtering by an alternate allelic fraction of ≥10% in at least one of the affected individuals	6,426
Remaining number of unique variant counts after applying stringent full cohort zygosity filters where no-calls were rejected	205

To assess the consequences of the identified variants, a comprehensive analysis was conducted by generating pie charts that visually depicted the distribution of variant consequence types (Figure 3.18). A threshold of 5% was established, where variant consequences accounting for less than 5% were consolidated into an "Others" category. Furthermore, an additional pie chart was constructed specifically to demonstrate the breakdown of consequences within this "Others" group (Figure 3.19). Variants with multiple consequences were allocated to multiple groups, resulting in a higher count of individual variant consequences (597) compared to the actual number of unique variants (205).

Figure 3.18 illustrates that the majority of the identified variants were categorised as intronic variants (25.5%). Similar proportions were observed for upstream variants

(14.7%), nonsense-mediated mRNA decay (NMD) transcript variants (14.4%), downstream variants (14.1%), and variants within regulatory regions (13.1%). Non-coding transcript variants accounted for 6.7% of the total variant consequences, while minor variant types constituted 11.6%.

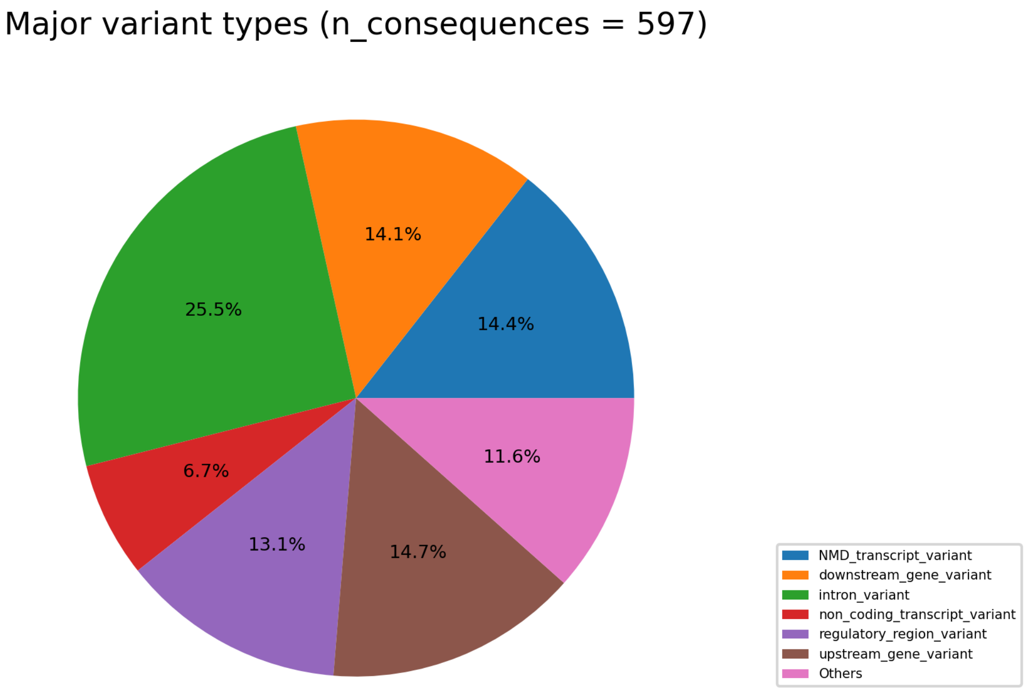


Figure 3.18: Distribution of the major variant consequence types in the full cohort filtered for a dominant inheritance pattern.

Furthermore, Figure 3.19 provides insights into the 69 variant consequences, each present in less than 5% of the variants. Notably, the most prevalent among these minor variant consequences was the transcription factor binding site variant, accounting for 26.1% of the consequences.

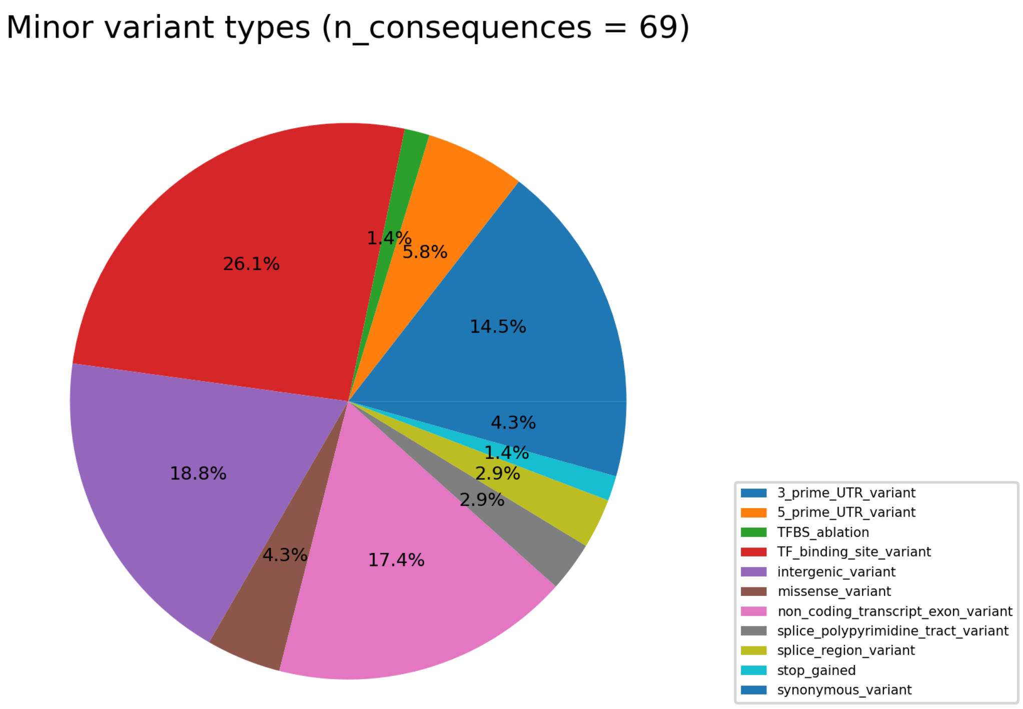


Figure 3.19: Distribution of the minor variant consequence types in the full cohort filtered for a dominant inheritance pattern.

3.4.1.2 Filtering of Variants following a Recessive Inheritance Pattern for the Full Cohort

To explore the possibility of variants conforming to a recessive mode of inheritance, different zygosity filters were applied to the entire cohort consisting of 11 individuals affected by HPFH and 9 healthy relatives. The outcomes of this filtering process are presented in Table 3.3. Initially, the lenient zygosity filter was employed, where variants were selectively retained if they had homozygous alternate or no-call zygosities across all 11 affected individuals, and heterozygous, homozygous reference, or no-call zygosities across all 9 unaffected individuals. The lenient zygosity filters together with other filters yielded a total of 5,779 variants. The stringent zygosity filter implemented afterwards required variants to exhibit homozygous alternate zygosities across all 11 affected individuals, and heterozygous or homozygous reference zygosities across all 9 unaffected individuals. As a result, no unique variants adhering to the recessive inheritance pattern were identified across all affected individuals.

Table 3.3: Sequential filtering steps for identifying potential causal variants in the full cohort with a recessive inheritance model.

Filtering steps for the full cohort following a recessive inheritance pattern	Unique Variant Counts
Total number of unique variants in the original annotated VCF file prior to filtering	10,627,339
Unique variants having an observed allele frequency of ≤15% in at least one of the four major gnomAD populations of interest (NFE, AFR, EAS and AMR)	10,597,744
Remaining number of unique variant counts after applying relaxed zygosity filters where no-calls were accepted	8,423
Number of unique variants with a ≥20X coverage in at least one of the affected individuals	5,800
Remaining unique variants after filtering by an alternate allelic fraction of ≥10% in at least one of the affected individuals	5,779
Remaining number of unique variant counts after applying stringent full cohort zygosity filters where no-calls were rejected	0

3.4.1.3 Filtering of Variants following a Dominant Inheritance Pattern for the Four Affected Individuals with the Highest HbF Levels Cohort

The combined and annotated VCF file was subjected to separate dominant and recessive zygosity filters, this time focusing on individuals F1.2, F1.5, F1.9 and F1.10, who exhibited the highest levels of HbF (>3%) and treating them as the affected group, while considering the remaining individuals as unaffected (4 affected vs. 16 unaffected). In the dominant inheritance lenient zygosity filter for the cohort of the four individuals with the highest HbF levels, only variants exhibiting heterozygous, homozygous alternate, or no-call zygosities for all four highest HbF individuals, and homozygous reference or no-call zygosities for the remaining individuals, were retained. Consequently, a total of 5,588 unique variants were identified when combining the lenient zygosity filters with the other filters.

For the subsequent stringent zygosity filtering, variants with no-call zygosities were excluded, and only those displaying heterozygous or homozygous alternate zygosities in individuals F1.2, F1.5, F1.9 and F1.10, along with homozygous reference zygosities in the other individuals, were retained. This filtering step yielded a total of 272 unique variants which were present only in all the four individuals with the highest HbF levels (>3%), but were absent in all other individuals (Online Supplementary Data B). The details of the filtering process are presented in Table 3.4.

Table 3.4: Sequential filtering steps for identifying potential causal variants in the subgroup with four highest HbF affected individuals (F1.2, F1.5, F1.9 and F1.10) following a dominant inheritance model.

Filtering steps for the subgroup analysis with 4 affected individuals following a dominant inheritance pattern	Unique Variant Counts
Total number of unique variants in the original annotated VCF file prior to filtering	10,627,339
Unique variants having an observed allele frequency of ≤15% in at least one of the four major gnomAD populations of interest (NFE, AFR, EAS and AMR)	10,597,744
Remaining number of unique variant counts after applying relaxed zygosity filters where no-calls were accepted	7,508
Number of unique variants with a ≥20X coverage in at least one of the affected individuals	5,600
Remaining unique variants after filtering by an alternate allelic fraction of ≥10% in at least one of the affected individuals	5,588
Remaining number of unique variant counts after applying stringent zygosity filters for the 4 individuals with the highest HbF levels where no-calls were rejected	272

The obtained list of unique variants was subjected to an analysis based on their consequence types to assess the distribution of each identified variant consequence (Figure 3.20). To facilitate this analysis, a threshold of 5% was again employed to group variant consequence types present in less than 5% into the "Others" category. A dedicated pie chart was then generated to illustrate the composition of consequences within this minor category (Figure 3.21). Since variants with multiple consequences were allocated to multiple groups, the total number of variant consequences (837) was higher than the number of unique variants identified (272) following a dominant mode of inheritance.

In the subgroup analysis considering only the four individuals with the highest HbF levels as affected (Figure 3.20), the majority of the identified variants were found to be intronic (24.7%). Comparable proportions were observed for noncoding transcript variants (16.5%), downstream variants (14.9%), NMD transcript variants (14.0%), and upstream variants (13.9%). Variants located in regulatory regions exhibited the lowest prevalence (7.4%), while minor variant consequences accounted for 8.6% of the total.

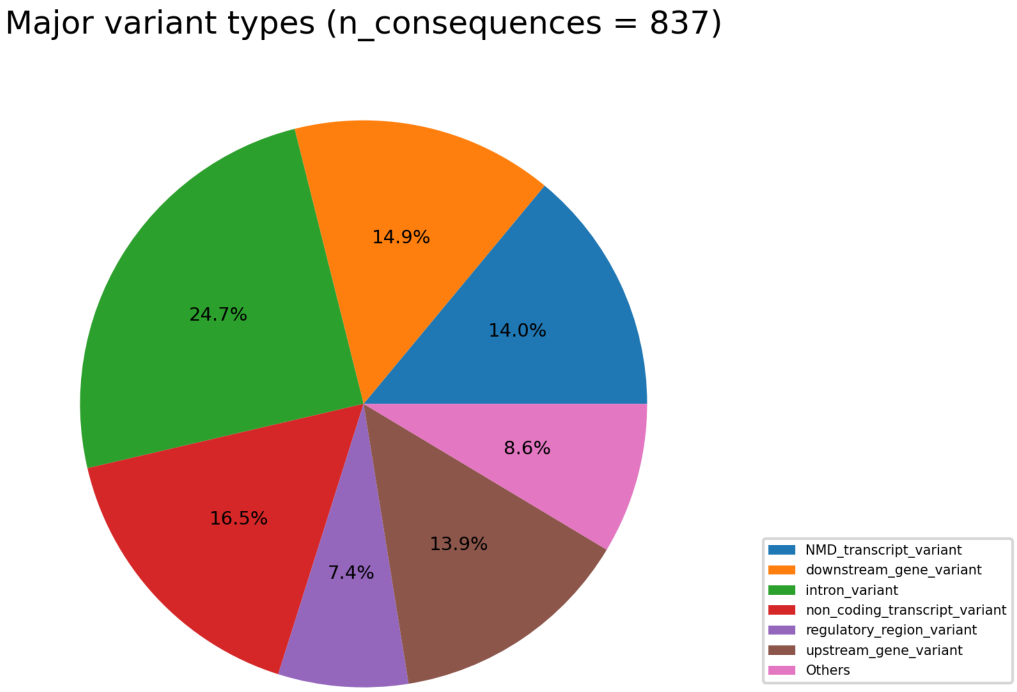


Figure 3.20: Distribution of the major variant consequence types in the subgroup consisting of 4 individuals with the highest HbF levels and 16 unaffected individuals filtered for a dominant inheritance pattern.

Amongst the minor variant consequences (Figure 3.21), the most prevalent were variants in the intergenic region, comprising 27.85% of the minor variant consequences. Additionally, variants in the 3' untranslated region (UTR) and non-coding transcript exons each accounted for 23.6% of the total minor variant consequences.

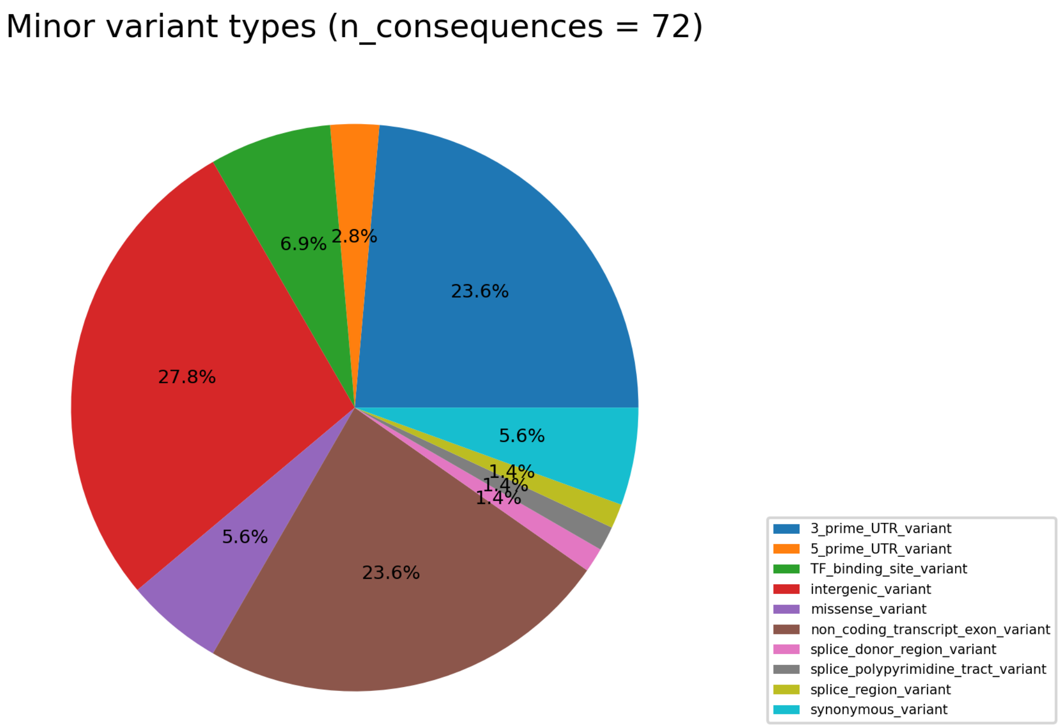


Figure 3.21: Distribution of the minor variant consequence types in the subgroup consisting of 4 individuals with the highest HbF levels and 16 unaffected individuals filtered for a dominant inheritance pattern.

3.4.1.4 Filtering of Variants following a Recessive Inheritance Pattern for the Four Affected Individuals with the Highest HbF Levels

To identify variants adhering to a recessive mode of inheritance within the four highest HbF cohort (>3%), a relaxed set of zygosity filters were first employed. Specifically, only variants with homozygous alternate or no-call zygosities for all four individuals with the highest HbF levels, while possessing heterozygous, homozygous reference, or no-call zygosities for all other individuals, were retained. Lenient zygosity filtering together with the other filters yielded a total of 27,585 unique variants. To further refine the selection, the more stringent zygosity filters were implemented to exclude all no-call variants. This filter retained only those variants with homozygous alternate zygosities for the four individuals with the highest HbF levels, while keeping variants with heterozygous or homozygous reference zygosities for all other individuals. As a result, a total of 200 unique variants were identified, exclusively present in the four affected individuals (F1.2, F1.5, F1.9, and F1.10) with the highest HbF levels (>3%), and entirely absent in all other individuals, including both healthy controls and affected individuals from other families (Online Supplementary Data C). The sequential filtering steps utilised are shown in Table 3.5.

Table 3.5: Sequential filtering steps for identifying potential causal variants in the subgroup with four highest HbF affected individuals (F1.2, F1.5, F1.9 and F1.10) following a recessive inheritance model.

Filtering steps for the subgroup analysis with 4 affected individuals following a recessive inheritance pattern	**Unique Variant Counts**
Total number of unique variants in the original annotated VCF file prior to filtering	10,627,339
Unique variants having an observed allele frequency of ≤15% in at least one of the four major gnomAD populations of interest (NFE, AFR, EAS and AMR)	10,597,744
Remaining number of unique variant counts after applying relaxed zygosity filters where no-calls were accepted	30,836
Number of unique variants with a ≥20X coverage in at least one of the affected individuals	27,625
Remaining unique variants after filtering by an alternate allelic fraction of ≥10% in at least one of the affected individuals	27,585
Remaining number of unique variant counts after applying stringent zygosity filters for the 4 individuals with the highest HbF levels where no-calls were rejected	200

During the analysis of consequence types within the list of 200 unique variants identified by filtering using a recessive inheritance pattern, a total of 367 individual variant consequences were discovered. Figure 3.22 visually represents the distribution of these consequences, revealing that the majority of the identified variants fell into the intronic category, comprising 27.8% of the total. Intergenic variants accounted for 22.6% of the total, indicating their significant presence as well. Among the variant consequences, non-coding transcript variants constituted 15.8%, while NMD transcript variants represented 13.6% of the total. Variants occurring in regulatory regions made up 9.8% of the consequence types. In contrast, downstream variants comprised only 5.7% of the total variant consequences identified.

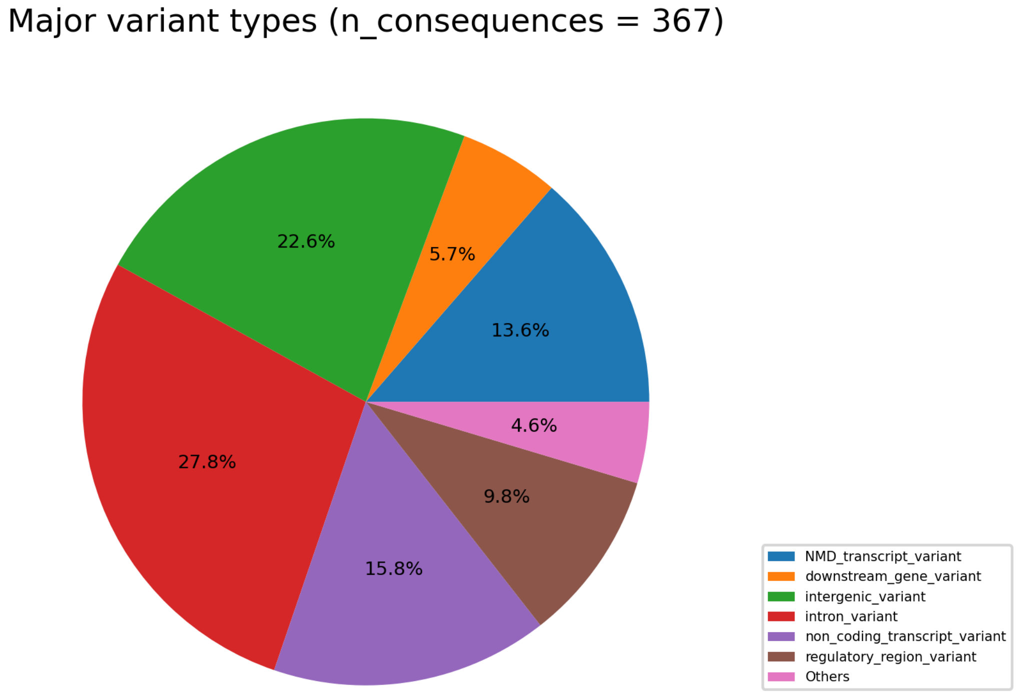


Figure 3.22: Distribution of the major variant consequence types in the subgroup consisting of 4 individuals with the highest HbF levels and 16 unaffected individuals filtered for a recessive inheritance pattern.

Further analysis of the 17 minor variant types, accounting for 4.6% of the total consequences, revealed that upstream genetic variants were the most abundant, constituting 76.5% of this subgroup. These findings are visually presented in Figure 3.23.

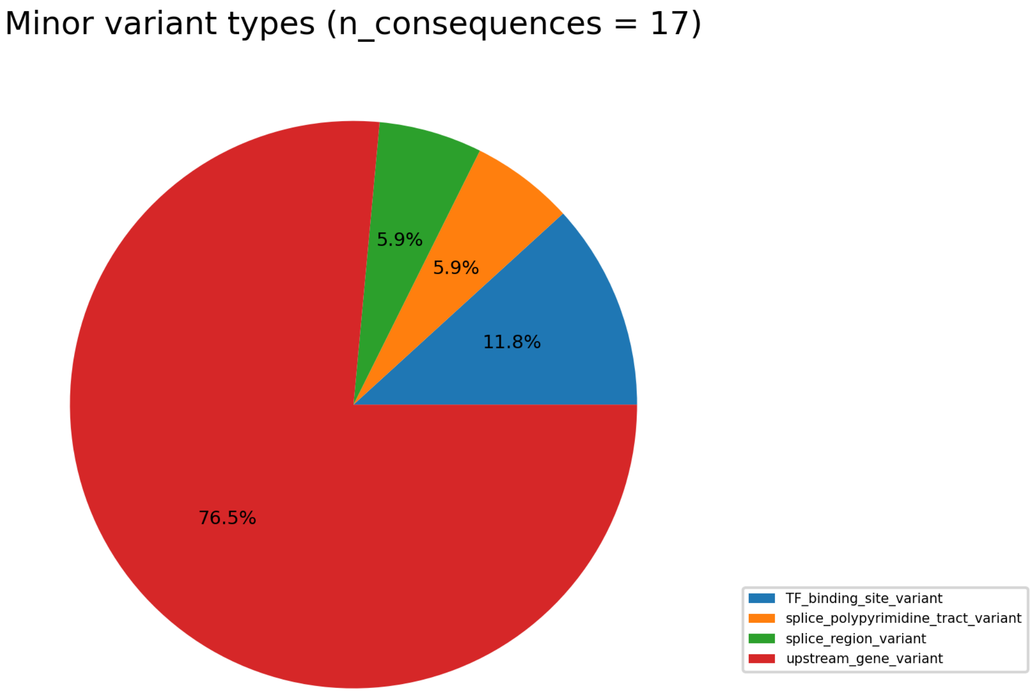


Figure 3.23: Distribution of the minor variant consequence types in the subgroup consisting of 4 individuals with the highest HbF levels and 16 unaffected individuals filtered for a recessive inheritance pattern.

3.4.2 Identifying Variants of Interest

After applying the necessary filtering steps to refine the variant lists as explained in Section 3.4.1, the different variant lists obtained, including those generated with the relaxed zygosity filters, were utilised to identify any potential variants of interest. Firstly, since the objective of this study was to uncover potential causative variants involved in the globin gene switching mechanism, a comprehensive analysis of the entire β-globin locus was conducted across the four scenarios (dominant inheritance and recessive inheritance in the full cohort, and dominant inheritance and recessive inheritance in the cohort with the four affected individuals having the highest HbF levels (>3%)). However, no notable variants were detected in this region.

Afterwards, attention was directed towards genes of interest associated with erythropoiesis, Hb regulation, globin gene switching mechanisms, and haemoglobinopathies, specifically aiming to identify variants potentially linked to the target condition, HPFH. After performing extensive background research and in-depth analysis of scientific research papers, I curated a list of genes of interest used in this study, which is outlined in Table 3.6. While a few variants were observed in some of the genes of interest when analysing the different modes of inheritance within both the full cohort and the cohort consisting of the four individuals with the highest HbF levels, these variants were disregarded due to a substantial number of no-call results, indicating insufficient representation of the variants.

Table 3.6: List of genes of interest known in scientific literature to be associated with globin gene switching, erythropoiesis and haemoglobinopathies, used to identify the presence of variants of interest in the three Maltese families being studied.

Genes of interest	Description
BCL11A	Known repressor of γ-globin gene (Borg, Papadopoulos et al. 2010; Wilber, Nienhuis et al. 2011; Zhou, Liu et al. 2010)
KLF1	Direct positive transcriptional regulator of BCL11A (Borg, Papadopoulos et al. 2010; Zhou, Liu et al. 2010)
BMI1 (PCGF4)	Polycomb group RING finger protein repressor of HbF (Qin, Lan et al. 2023)
IGF2BP1/3	Highly expressed in human fetal erythroblasts, switches transcription from β- to γ-globin (Chambers, Gross et al. 2020)
Lin28B	Increased expression in adult erythroblasts increases HbF levels (Lee, De Vasconcellos et al. 2013)
ZBTB7A	Trans-acting factor known to repress γ-globin gene expression (Martyn, Wienert et al. 2018; Masuda, Wang et al. 2016)
MAPK1	Synergistic activation by SCF and EPO crucial for enhanced erythropoiesis (Sui, Krantz et al. 1998)
SP1	Essential for proper erythrocyte maturation, coregulator of erythropoiesis with KLF1 and GATA1 (Gregory, Taxman et al. 1996; Kruger, Vollmer et al. 2007)
CHD4	Knockdown leads to higher expression of γ-globin gene, increases *KLF1* and *BCL11A* expression (Amaya, Desai et al. 2012)
HDAC2	Part of NuRD complex, silences HbF production (Bradner, Mak et al. 2010)
TFCP2	Interacts with NFE4, forms a complex activating the γ-globin promoter (Jane, Nienhuis et al. 1995; Zhou, Clouston et al. 2000; Zhou, Zhao et al. 2004)
RUNX1	Inhibits erythroid differentiation, represses *KLF1* expression during megakaryocytic differentiation (Kuvardina, Herglotz et al. 2015)
HBS1L	Intergenic region between *HBS1L* and *MYB* genes influences HbF levels (Lettre, Sankaran et al. 2008; Menzel, Jiang et al. 2007; Thein, Menzel et al. 2007)
NR2C1	Direct transcriptional repressor of embryonic and fetal globin genes (Tanabe, Katsuoka et al. 2002; Tanabe, McPhee et al. 2007; Tanimoto, Liu et al. 2000)
NR2C2	Direct transcriptional repressor of embryonic and fetal globin genes (Tanabe, Katsuoka et al. 2002; Tanabe, McPhee et al. 2007; Tanimoto, Liu et al. 2000)

Genes of interest	Description
NR2C2AP	Critical for erythroid maturation and globin gene expression (Gothwal, Wehrle et al. 2016)
GATA1	Forms chromatin complex with BCL11A and NuRD, represses γ-globin genes, involved in erythroid maturation (Pevny, Simon et al. 1991; Sankaran, Menne et al. 2008; Weiss, Keller et al. 1994)
LDB1	Recruits and activates erythroid genes by interacting with TAL1, GATA1, and KLF1 (Love, Warzecha et al. 2014)
SOX6	Coordinates with BCL11A in suppressing γ-globin gene expression (Xu, Sankaran et al. 2010)
NFE2	Plays a role in erythroid maturation, controls transcription of erythroid-specific genes along with KLF1 and GATA1 (Ingley, Tilbrook et al. 2004; Johnson, Grass et al. 2002)
LMO2	Essential in erythropoiesis and regulation of hematopoiesis (Wadman, Osada et al. 1997; Warren, Colledge et al. 1994)
TAL1	Essential in erythropoiesis and regulation of hematopoiesis (Robb, Lyons et al. 1995; Shivdasani, Mayer et al. 1995; Wadman, Osada et al. 1997)
MYB	Supports erythropoiesis through transactivation of *KLF1* and *LMO2* expression, intergenic region with HBS1L influences HbF levels (Lettre, Sankaran et al. 2008; Menzel, Jiang et al. 2007; Thein, Menzel et al. 2007)
MTA2	Works with GATA1 in a repressive FOG-1 mediated MeCP1 complex (Rodriguez, Bonte et al. 2005)
ICAM4	Directly regulated by KLF1 during erythroid differentiation (Xue, Galdass et al. 2014)
CD44	Direct target of KLF1 transcription factor (Arnaud, Saison et al. 2010)
BCAM	Direct target of KLF1 transcription factor (Arnaud, Saison et al. 2010)
AHSP	KLF1 directly regulates expression of *AHSP* gene (Magor, Tallack et al. 2015)
HEBP1	Known to be dependent on *FOG-1* gene (Tanimura, Miller et al. 2015)
EIF2AK1	Essential for regulation of globin gene translation, deletion in sickle erythroblasts increases HbF production (Adema, Ma et al. 2022; Liu, Bhattacharya et al. 2008)

In total, five well-represented variants were identified within two genes of interest when employing a dominant inheritance pattern in the full cohort. These included four variants located on the *KLF1* gene, and one variant which was located on the *LMO2* gene (Table 3.7).

Table 3.7: Five variants identified on *KLF1* and *LMO2* genes of interest in a dominant inheritance mode showing the zygosity status of each individual, where a no-call is represented as .|., a homozygous reference status is represented as a 0|0, and a heterozygous alternate status is represented as 0|1.

Family	Member	HPFH status	***LMO2*** 11:33912893 Intronic **AACAC>A** (rs57617009)	***KLF1*** 19:12991919 Downstream **G>C** (rs112348773)	***KLF1*** 19:12996182 Coding region **T>A** (rs267607202)	***KLF1*** 19:12999087 Upstream **G>T** (rs112943513)	***KLF1*** 19:13001454 Upstream **CACAG>C** (rs1395948183)
Fam F1	F1.1	Control	.|.	0|0	0|0	0|0	0|0
	F1.2	Affected	.|.	0|1	0|1	0|1	0|1
	F1.3	Control	.|.	0|0	0|0	0|0	0|0
	F1.4	Affected	.|.	0|1	0|1	0|1	0|1
	F1.5	Affected	.|.	0|1	0|1	0|1	0|1
	F1.6	Affected	0|1	0|1	0|1	0|1	0|1
	F1.7	Control	0|0	0|0	0|0	0|0	0|0
	F1.9	Affected	0|1	0|1	0|1	0|1	0|1
	F1.10	Affected	0|1	0|1	0|1	0|1	0|1
	F1.12	Control	0|0	0|0	0|0	0|0	0|0
Fam F2	F2.1	Affected	0|1	0|1	0|1	0|1	0|1
	F2.2	Control	0|0	0|0	0|0	0|0	0|0
	F2.3	Control	0|0	0|0	0|0	0|0	0|0
	F2.4	Affected	.|.	0|1	0|1	0|1	0|1
Fam F4	F4.1	Control	.|.	0|0	0|0	0|0	0|0
	F4.2	Affected	0|1	0|1	0|1	0|1	0|1
	F4.3	Control	.|.	0|0	0|0	0|0	0|0
	F4.4	Affected	.|.	0|1	0|1	0|1	0|1
	F4.5	Affected	0|1	0|1	0|1	0|1	0|1
	F4.6	Control	0|0	0|0	0|0	0|0	0|0

The identified variant within the *LMO2* gene was characterized as a deletion. It was observed to be present with a heterozygous alternate zygosity in six out of the eleven affected individuals and confirmed to be absent with a homozygous reference zygosity in five out of the nine unaffected individuals. This variant exhibited significant prevalence within the full cohort, thereby highlighting its substantial importance. Within the *KLF1* gene, a total of four variants were identified, comprising the previously reported p.K288X (rs267607202) stop-gained mutation and three novel variants found to be in *cis* to the p.K288X mutation. All four variants identified within the *KLF1* gene were consistently found with a heterozygous alternate zygosity in all eleven affected individuals with HPFH and confirmed to be absent with a homozygous reference zygosity in all nine unaffected individuals, thus indicating a strong likelihood of association with the target condition.

3.5 Quantification of Globin Gene mRNA Expression

Following the analysis of the raw data obtained from the Luminex™ instrument and subsequent quantification of normalised mRNA globin expression in each sample, statistical analysis was performed to examine variations between individuals who are KLF1 deficient and healthy controls. Mean and standard deviation measures were calculated to assess the data's dispersion (Table 3.8). The average relative mRNA expression of the *HBG1* globin gene was noted to be higher in the KLF1 deficient group, while the average relative mRNA expression of the *HBB* and *HBA1* genes were higher in the controls. The standard deviation was the highest for the mRNA expression of the *HBA1* gene indicating a significant amount of variability in the data set.

Table 3.8: Table showing the average mRNA expression and standard deviation of each globin gene in the controls and in individuals who are KLF1 deficient.

Target Genes	Analysis	Controls	KLF1 deficient
HBB	Mean	29.27	25.78
	Standard Deviation	34.73	31.13
HBG1	Mean	0.26	0.82
	Standard Deviation	0.29	1.08
HBA1	Mean	382.82	259.93
	Standard Deviation	411.48	285.18

Box plots were generated to visualise the mRNA expression levels of each target gene in both the control group and the KLF1 deficient group. The findings revealed notable differences in mRNA expression patterns between the control group and the KLF1 deficient group. Specifically, the mRNA expression of the *HBB* and *HBA1* genes were higher in the control group (Figure 3.24 and Figure 3.25), whereas the KLF1 deficient group displayed elevated mRNA expression of the *HBG1* gene (Figure 3.26).

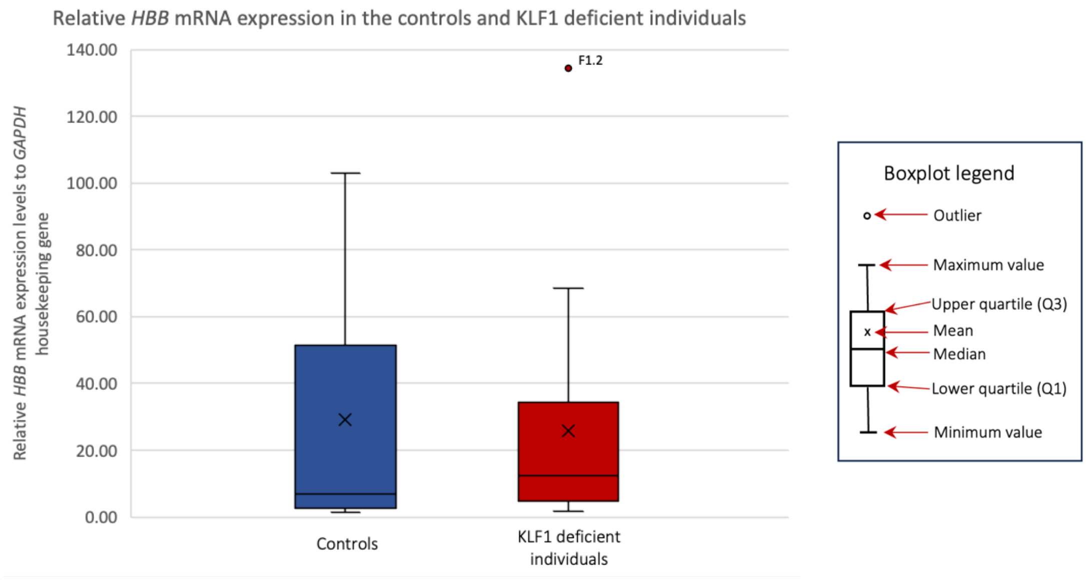


Figure 3.24: Comparative box plots showing the relative mRNA expression levels of the *HBB* gene in the control group and the KLF1 deficient group, normalised to the expression of *GAPDH* housekeeping gene.

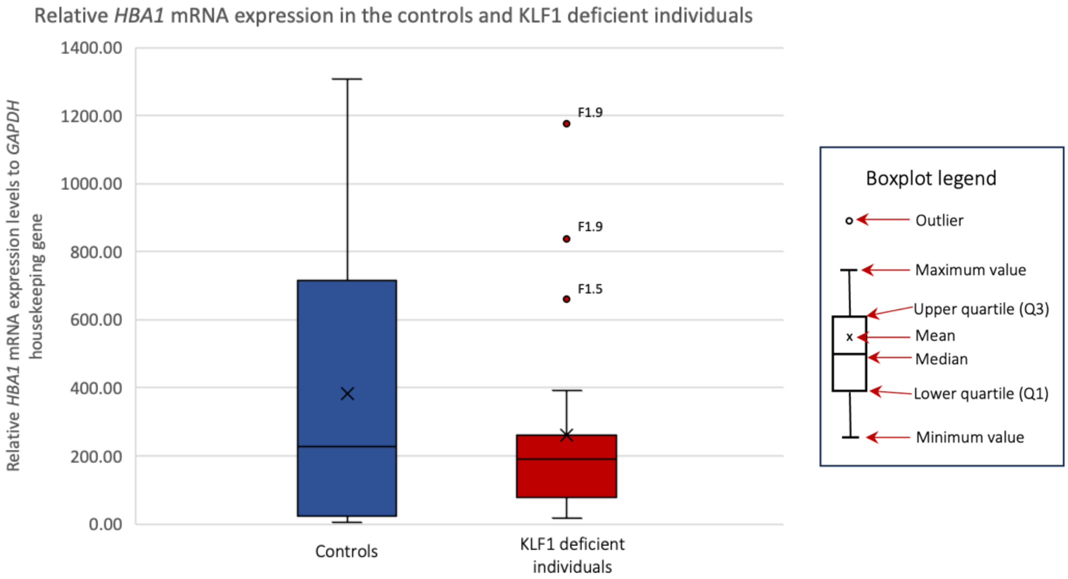


Figure 3.25: Comparative box plots showing the relative mRNA expression levels of the *HBA1* gene in the control group and the KLF1 deficient group, normalised to the expression of *GAPDH* housekeeping gene.

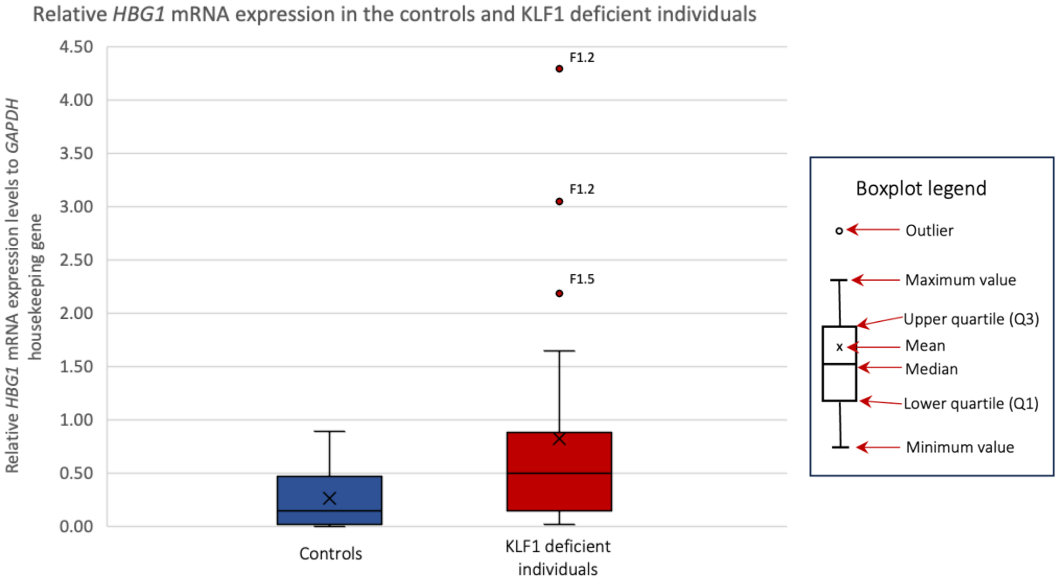
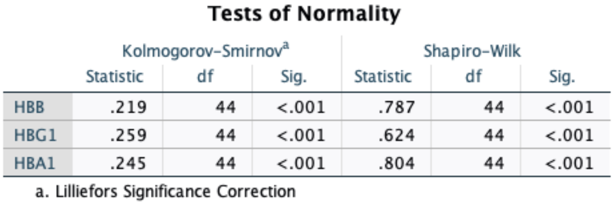


Figure 3.26: Comparative box plots showing the relative mRNA expression levels of the *HBG1* gene in the control group and the KLF1 deficient group, normalised to the expression of *GAPDH* housekeeping gene.

The relative mRNA expression of the *HBB* and *HBA1* gene was found to be higher in the control group due to a higher average expression, as well as a longer box plot indicating higher values. However, the median expression level of the *HBB* gene was actually lower in the control group compared to the KLF1 deficient group. Conversely, the relative mRNA expression of the *HBG1* gene was higher in the KLF1 deficient group, as evidenced by the larger box plot, as well as both higher mean and median values. Outliers within the dataset included the duplicate readings of individual F1.9 for the mRNA expression of the *HBA1* gene, and the mRNA expression levels of individual F1.2 for the *HBG1* gene. One of the duplicate measurements of individual F1.5 was an outlier for both the *HBA1* and *HBG1* mRNA expression levels.

The Shapiro-Wilk test and the Kolmogorov-Smirnov test were used to assess the normality distribution of the samples. The normality assumptions were violated, as shown by p-values smaller than 0.05 (Table 3.9). Consequently, the non-parametric Mann-Whitney U test was used to investigate potential significant differences in gene expression between the two sample groups. The analysis revealed no statistically significant differences in the average mRNA expression of the *HBB* and *HBA1* genes for the two groups under study, as indicated by p-values exceeding 0.05 (*HBB*: p-value = 0.664; *HBA1*: p-value = 0.742). However, a significant discrepancy was observed between the average mRNA expression of the *HBG1* gene of the two groups (p-value = 0.015), where the expression of *HBG1* was found to be higher in individuals who were KLF1 deficient compared to controls (Table 3.10).

Table 3.9: Table showing the test statistics for normality assumptions and significance levels in the dataset, where normality assumptions were rejected.

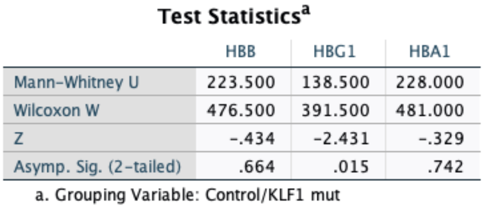


Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
HBB	.219	44	<.001	.787	44	<.001
HBG1	.259	44	<.001	.624	44	<.001
HBA1	.245	44	<.001	.804	44	<.001

a. Lilliefors Significance Correction

Table 3.10: Table showing the results of the Mann-Whitney U test for statistical comparison of mRNA gene expression levels between the control and the KLF1 deficient groups.



Test Statistics[a]

	HBB	HBG1	HBA1
Mann-Whitney U	223.500	138.500	228.000
Wilcoxon W	476.500	391.500	481.000
Z	-.434	-2.431	-.329
Asymp. Sig. (2-tailed)	.664	.015	.742

a. Grouping Variable: Control/KLF1 mut

3.6 Analysis of the Proteomic Data by Mass Spectrometry

After subjecting the peptides to LC-MS/MS analysis, the raw data files were analysed at the Department of Haematopoiesis of the Sanquin Research Facility in Amsterdam (Netherlands), using the DIA-NN automated software that identifies the peptides by associating them with protein entries in a database. The automated software performed label free protein quantification by utilising the peptide intensity values to measure the relative abundance of each protein in the samples and subsequently carried out log2 transformation. Log2 transformation helps in stabilising the variance of the peptide intensity data by compressing larger values while expanding smaller values, hence, reducing the influence of extreme values across the dataset. The raw data analysis using the DIA-NN software resulted in the identification of a total of 1600 proteins.

Missing value imputation was carried out using the Perseus software, following the assumption that missing values arise from low expression proteins. To accurately impute these values, a Gaussian distribution was employed, where the median was shifted towards lower expression levels compared to the measured data distribution. The mode parameter was utilised to determine the measured data distribution for calculating the random distribution. Specifically, a down-shift of 1.8 and a distribution width of 0.5 were applied to simulate the expression of low abundant proteins that fall below the detection limit (Tyanova, Temu et al. 2016).

Principal component analysis (PCA) was conducted to explore variations among individuals, including controls and those with the *KLF1* mutation. The aim of PCA was to reveal underlying patterns and identify similarities or dissimilarities in protein expression patterns by visually representing the distribution of samples based on the protein abundance levels. Each individual sample was represented by a point on the plot and the proximity of the points reflected the similarity of their protein expression profiles. The PCA plot depicted in Figure 3.27 showcases clear segregation of the 22 samples into two discrete clusters: individuals harbouring the p.K288X truncation mutation in the *KLF1* gene (shown in red) and the unaffected, healthy controls (shown in blue). The evident separation observed between these groups provides evidence of significant differences in the protein profiles between the two cohorts.

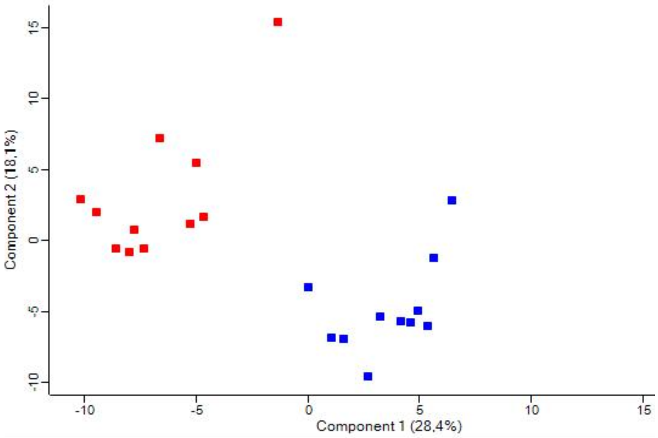


Figure 3.27: PCA plot of all the samples, showing distinct protein expression patterns between individuals having the *KLF1* p.K288X mutation shown in red, and the unaffected healthy controls shown in blue.

Furthermore, the PCA plot showing the protein expression patterns was dominated by differential globin expressions. The $^{A}\gamma$, $^{G}\gamma$ (*HBG1* and *HBG2*) and ζ (*HBZ*) globins exhibited the most pronounced dissimilarity among the various globin isoforms when comparing the two groups (Figure 3.28).

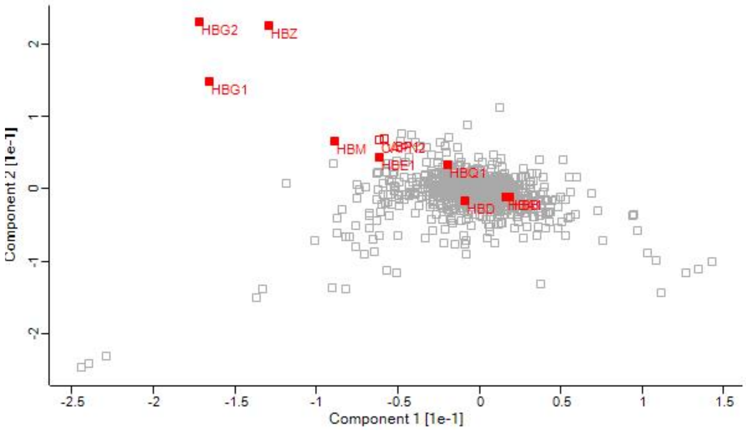


Figure 3.28: PCA analysis of protein expression levels, reveals distinct separation primarily driven by differential globin expression levels, where $^{A}\gamma$, $^{G}\gamma$ and ζ- globin levels are mostly varied between the two cohorts.

Subsequently, statistical analyses were conducted using Python to assess the functional significance of these proteins. Mean and standard deviation calculations were calculated for the control group and individuals with the *KLF1* mutation to assess protein variation. To assess the normality distribution of protein values, the Shapiro-Wilk test was separately conducted for the two groups under study for each identified protein. Proteins exhibiting a normal distribution were subjected to a independent sample T-test, while proteins with non-normal distributions underwent the Mann-Whitney U test. A significance threshold of 0.05 (alpha value) was utilised to determine acceptance or rejection of the null hypothesis, which stated that there is no statistically significant difference in the mean or median of the two groups. Consequently, a filtering process based on the alpha value was implemented to generate a list of proteins deemed 95% statistically significant. This filtering step resulted in the identification of 534 proteins that exhibited statistical significance with an p-value less than 0.05 (Online Supplementary Data D).

Z-scores were computed individually for each significant protein by subtracting the mean expression value and dividing it by the standard deviation. These calculated z-scores were subsequently utilised to generate a heatmap using Python software, in which the color gradient represents the varying expression levels of each protein across the samples. Proteins exhibiting high expression levels were represented by dark red shading, while proteins with low expressions were depicted in shades of blue.

The samples were classified as either belonging to the control group (labelled as 'C') or the group of individuals affected by the *KLF1* truncation mutation (labelled as 'M'). This labelling strategy was implemented to ensure samples from the same group were arranged in close proximity. Figure 3.29 presents a heatmap illustrating the relative protein expression patterns between samples derived from healthy controls and those derived from affected individuals. The upper section of the heatmap depicts proteins with increased expression in the samples that were KLF1 deficient, as indicated by the red coloration. Conversely, the lower section of the heatmap represents proteins with decreased expression (blue coloration) in the affected individuals.

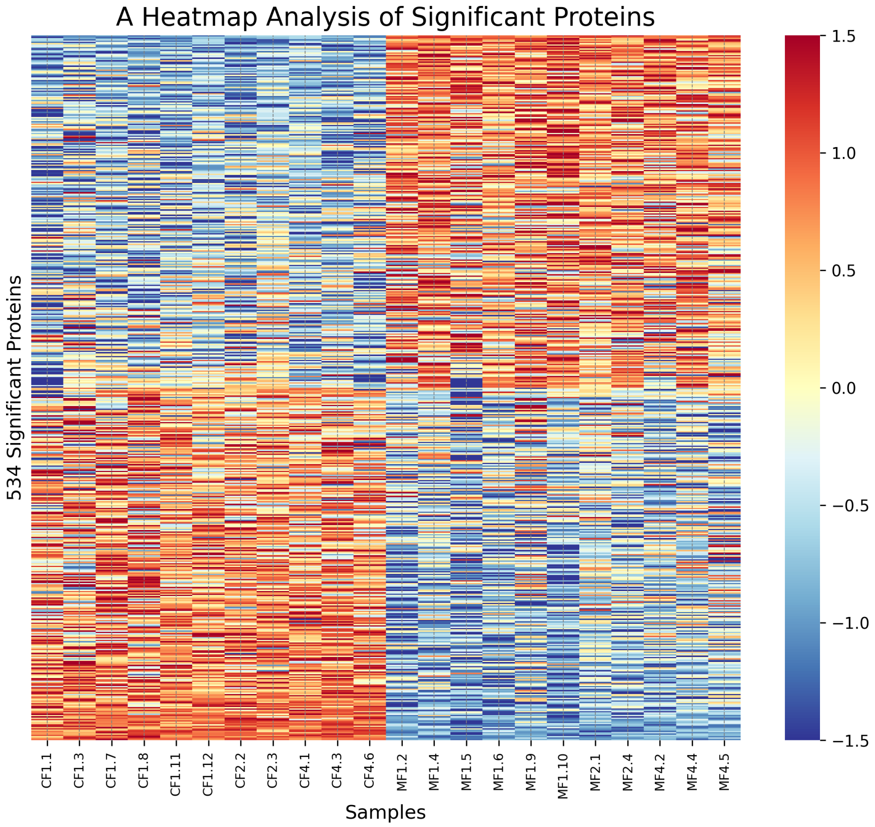


Figure 3.29: Heatmap analysis of 534 significant proteins showing comparison of the relative protein expression levels in KLF1 deficient samples and controls. The colour scale on the right shows how proteins with low expression are shown in shades of blue, while proteins with increased expression are shown in shades of red.

To identify proteins with the most pronounced differential expression, a separate heatmap was generated specifically for the top 40 proteins displaying the largest significant differences, as determined by their corresponding p-values (Figure 3.30). Notably, the five proteins with the highest upregulation in the group with KLF1 deficiency, were encoded by the genes *TBC1D24*, *CLIC1*, *HBG1*, *GOT1*, and *GAPDH*. On the other hand, the top five proteins demonstrating reduced expression in the presence of the *KLF1* p.K288X mutation were encoded by the genes *WDR91*, *BCAM*, *SH3BGRL3*, *S100A6*, and *CD44*.

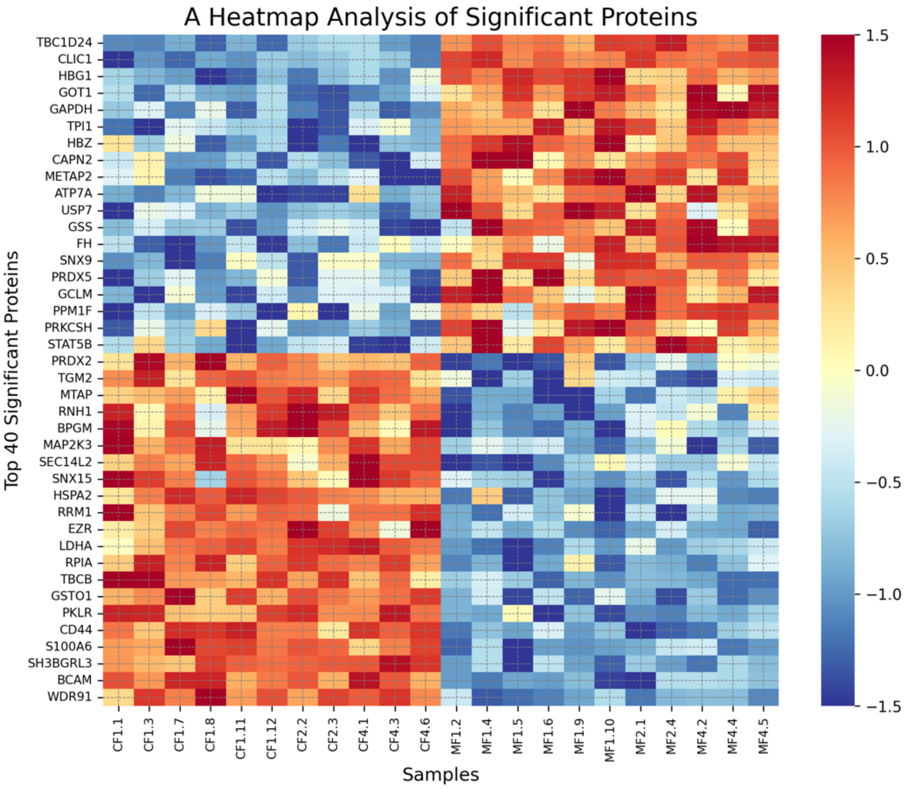


Figure 3.30: Heatmap analysis of the top 40 significant proteins showing comparison of the relative protein expression levels in samples with the *KLF1* p.K288X mutation and controls. The colour scale on the right shows how proteins with low expression are shown in shades of blue, while proteins with increased expression are shown in shades of red.

A distinct heatmap was generated to visually represent the variations in the different Hb subunit protein levels between the two groups. In Figure 3.31 it can be noted that in the KLF1 deficient group, augmented levels were observed in several globin subunits, namely the $^{A}\gamma$, $^{G}\gamma$, ζ, ε (HBE1), mu (*μ;* HBM), and theta-2 (θ; HBQ1) globin subunits. However, the control group demonstrated higher levels of the adult α- globin subunit (HBA1) and the β- globin subunit (HBB). Exceptions included individual F4.5 from Fam F4, who despite being KLF1 deficient, exhibited elevated levels of α- and β- globin subunits; and individual F1.3 from Fam F1 who displayed lower levels of adult Hb subunits despite their healthy status.

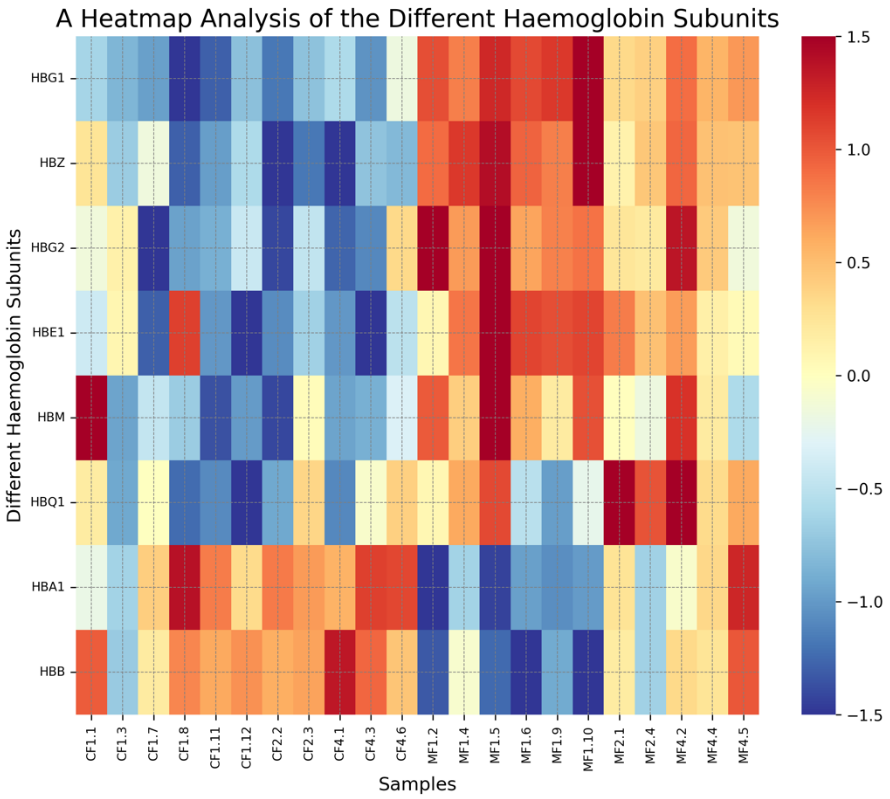


Figure 3.31: Heatmap analysis of the 6 different haemoglobin subunit protein levels detected by mass spectrometry showing comparison of the relative protein expression levels in KLF1 deficient samples and controls. The colour scale on the right shows how proteins with low expression are shown in shades of blue, while proteins with increased expression are shown in shades of red.

Subsequently, correlation analysis was performed to explore the linear relationship between the significant proteins identified in individuals affected by the *KLF1* truncation mutation and the levels of HbF to determine any significant associations. Pearson correlation analysis was carried out, employing a significance threshold of 0.05 to assess the significance of the correlations, where only protein correlations exhibiting a p-value below 0.05 were considered statistically significant. This correlation analysis identified a total of 53 proteins displaying significant correlations with HbF levels (Appendix C). Table 3.11 presents the top 20 proteins exhibiting values which suggest the strongest and most significant negative correlation with HbF levels, while Table 3.12 provides a list of the top 20 proteins demonstrating values which suggest the strongest positive correlations with HbF levels. Scatter plots were generated to display the correlation.

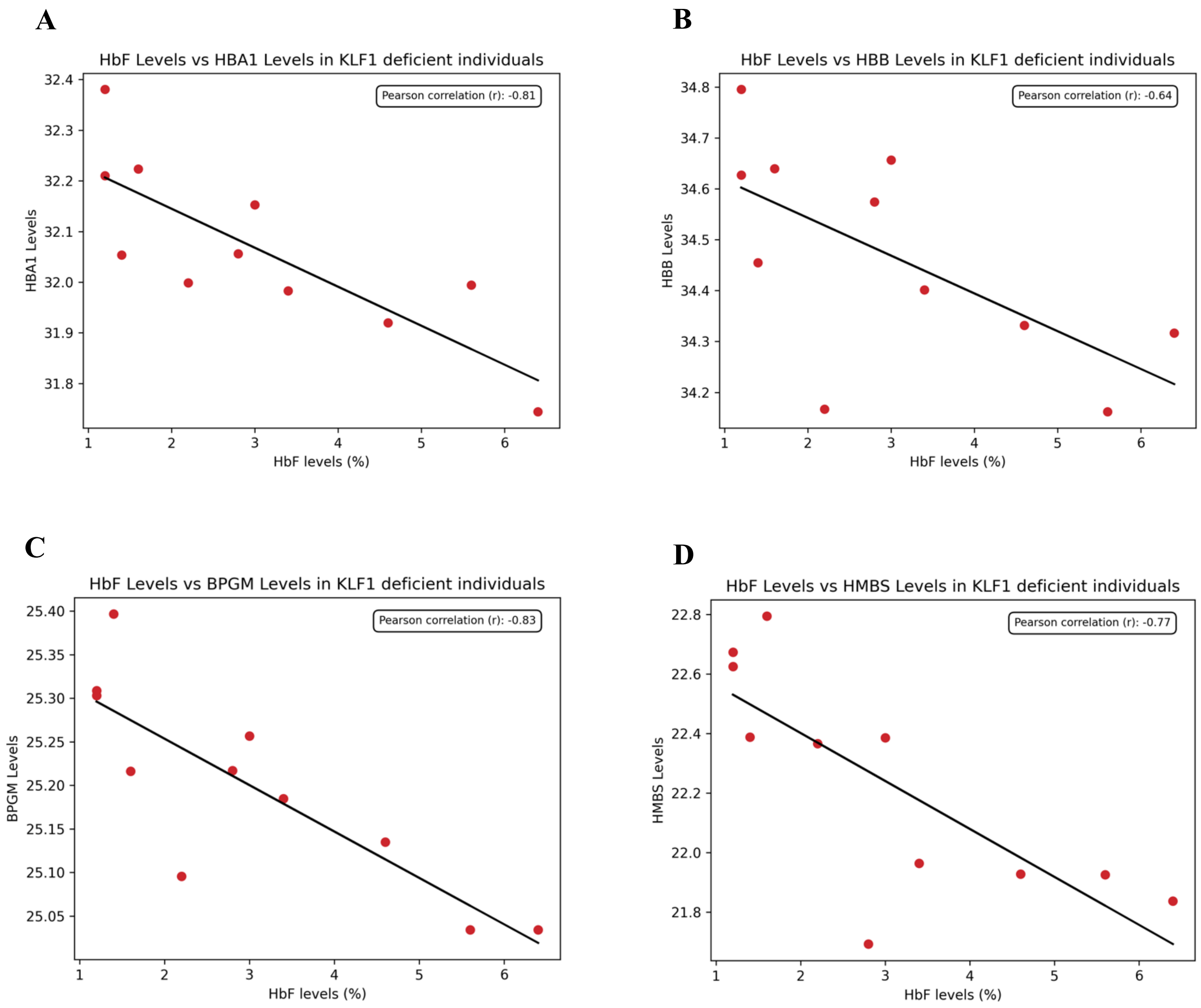
Table 3.11: Table showing the top 20 most significant negatively correlated proteins with HbF levels in individuals affected with the *KLF1* p.K288X truncation mutation, the gene name that codes for the protein and the Pearson correlation coefficient (r).

Negatively Correlated Proteins with HbF Levels	Gene Name	Pearson Correlation (r)
Rho GTPase-activating protein 1	ARHGAP1	-0.877802836
Bisphosphoglycerate mutase	BPGM	-0.828901392
Hemoglobin subunit alpha	HBA1	-0.812287181
Exportin-2	CSE1L	-0.80475275
14-3-3 protein beta/alpha	YWHAB	-0.789303564
Keratinocyte proline-rich protein	KPRP	-0.78633549
Transport and Golgi organization protein 2 homolog	TANGO2	-0.774967796
Porphobilinogen deaminase	HMBS	-0.769124214
Golgi-associated plant pathogenesis-related protein 1	GLIPR2	-0.757394781
HCLS1-binding protein 3	HS1BP3	-0.739579359
Ribose-phosphate pyrophosphokinase 1	PRPS1	-0.726665613
Intermediate conductance calcium-activated potassium channel protein 4	KCNN4	-0.698806207
Acetylcholinesterase	ACHE	-0.693079109
Stomatin	STOM	-0.691080936
Casein kinase I isoform alpha	CSNK1A1	-0.684434867
EH domain-binding protein 1-like protein 1	EHBP1L1	-0.681459511
Alpha-actinin-4	ACTN4	-0.673861227
Interleukin-18	IL18	-0.65705484
Hemoglobin subunit beta	HBB	-0.640863445
BRO1 domain-containing protein BROX	BROX	-0.636241879

Table 3.12: Table showing the top 20 most significant positively correlated proteins with HbF levels in individuals affected with the *KLF1* p.K288X truncation mutation, the gene name that codes for the protein and the Pearson correlation coefficient (r).

Positively Correlated Proteins with HbF Levels	Gene Name	Pearson Correlation (r)
NAD-dependent protein deacetylase sirtuin-2	SIRT2	0.886077033
Tensin-1	TNS1	0.847433568
Eukaryotic translation initiation factor 4E	EIF4E	0.816275016
Hemoglobin subunit gamma-2	HBG2	0.812475993
Hemoglobin subunit gamma-1	HBG1	0.798469871
Serine--tRNA ligase, cytoplasmic	SARS1	0.765328475
BRISC complex subunit Abraxas 2	ABRAXAS2	0.761433756
Hemoglobin subunit zeta	HBZ	0.757627054
Proteasome subunit alpha type-2	PSMA2	0.752631483
Hemoglobin subunit mu	HBM	0.720165558
Protein MEMO1	MEMO1	0.712408471
Serine/threonine-protein phosphatase 2A 56 kDa regulatory subunit delta isoform	PPP2R5D	0.70555279
Transitional endoplasmic reticulum ATPase	VCP	0.703223663
E3 ubiquitin-protein ligase UBR4	UBR4	0.672546401
Heat shock protein 105 kDa	HSPH1	0.666138946
Rap guanine nucleotide exchange factor 2	RAPGEF2	0.656798969
Ubiquitin-protein ligase E3A	UBE3A	0.645668939
Protein DPCD	DPCD	0.632616312
Charged multivesicular body protein 2a	CHMP2A	0.626403332
Uroporphyrinogen-III synthase	UROS	0.623499265

Prominent among the proteins exhibiting a strong negative correlation with HbF levels were the α-globin subunit (r = -0.812) and the β-globin subunit (r = -0.641) (Figure 3.32 A and B). Additional proteins demonstrating significant negative correlations, known to be associated with erythropoiesis or haemoglobinopathies, included bisphosphoglycerate mutase (BPGM, r = -0.829) which reduces Hb O_2 affinity, porphobilinogen deaminase (HMBS; r = -0.769) which is crucial in haem biosynthesis, intermediate conductance calcium-activated potassium channel protein 4 (KCNN4; r = -0.699) which promotes calcium influx, acetylcholinesterase (ACHE; r = -0.693) which forms part of the Yt blood group antigen, and casein kinase I isoform alpha (CSNK1A1; r = -0.684) which is important in haematopoiesis (Figure 3.32 C to G).



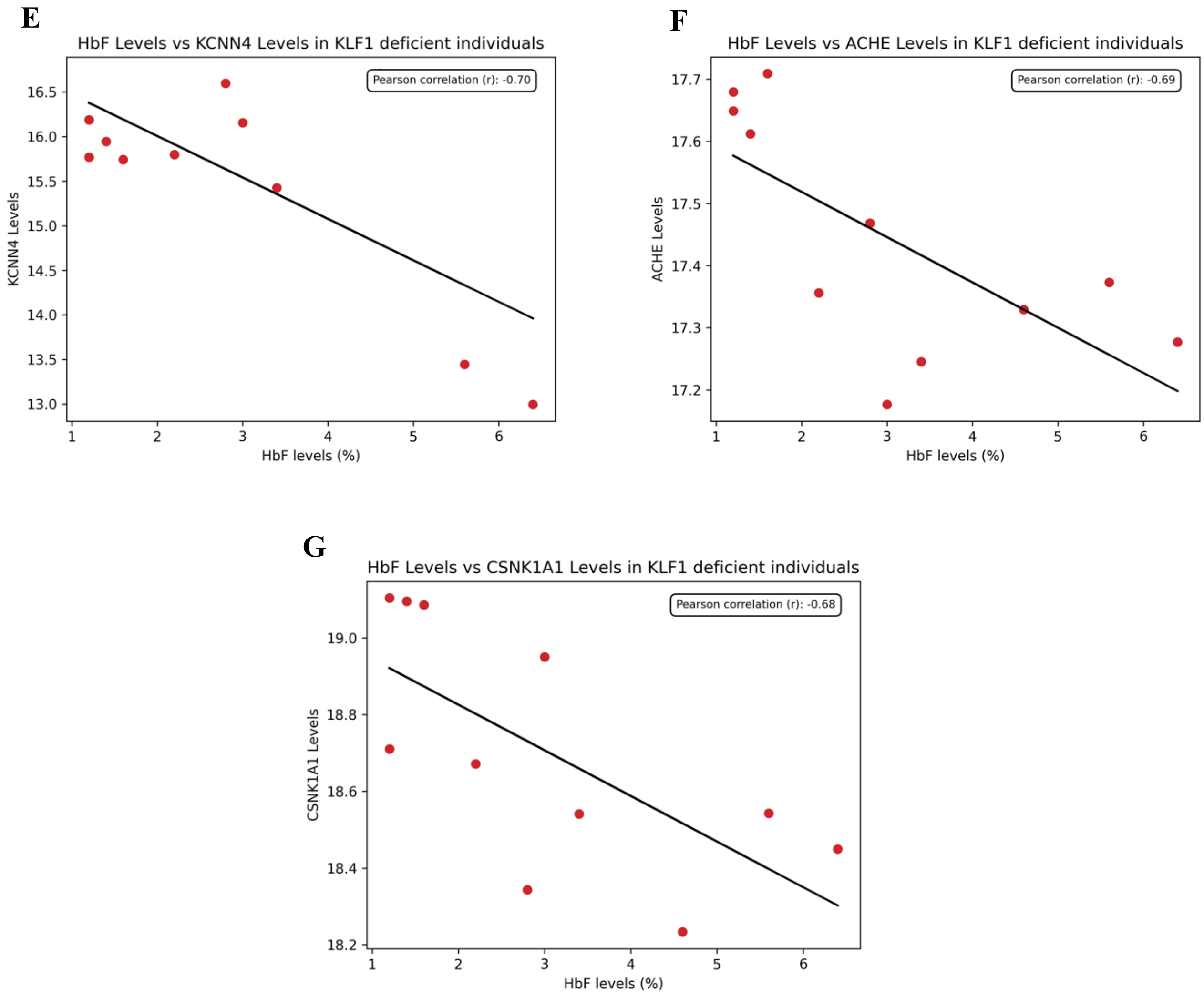


Figure 3.32: Scatter plots to represent the strong negative correlations between levels of HbF (%) in individuals affected with HPFH and their respective protein levels of proteins known to be associated with erythropoiesis and haemoglobinopathies (A: HBA1 levels, B: HBB levels, C: BPGM levels, D: HMBS levels, E: KCNN4 levels, F: ACHE levels, G: CSNK1A1).

Proteins displaying robust positive correlations with HbF levels included the the $^{A}\gamma$ (HBG1; r = 0.798) and $^{G}\gamma$ (HBG2; r = 0.812) foetal globin subunits, and uroporphyrinogen-III synthase (UROS; r = 0.623), which catalyses haem biosynthesis (Figure 3.33).

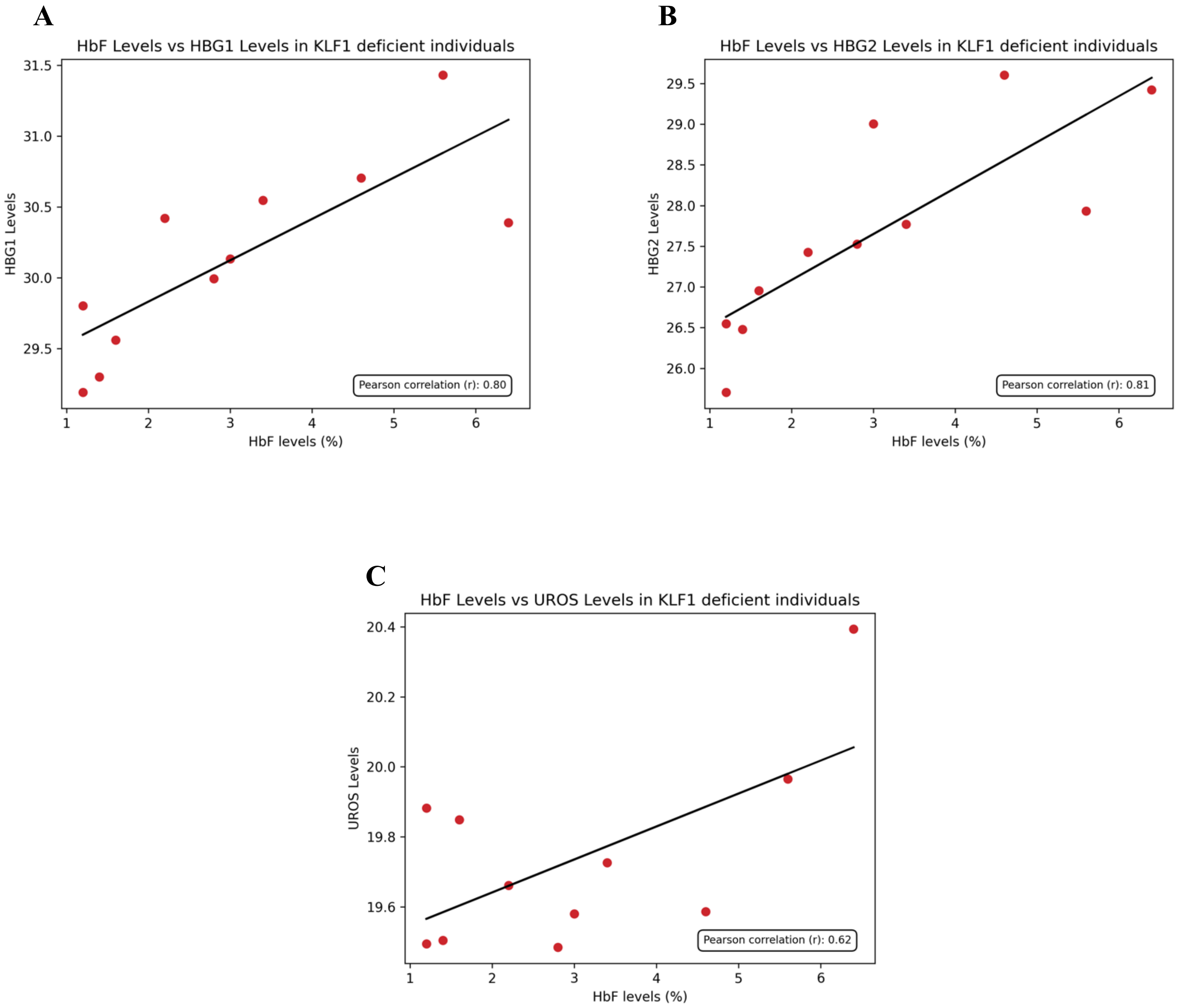


Figure 3.33: Scatter plots to represent the strong positive correlations between levels of HbF (%) in individuals affected with HPFH and their respective protein levels of proteins known to be associated with erythropoiesis and haemoglobinopathies (A: HBG1 levels, B: HBG2 levels, C: UROS levels).

4. Discussion

Haemoglobinopathies, such as thalassemia and SCD, are recognised as among the most prevalent monogenic diseases worldwide and constitute a significant global health concern (Harteveld, Achour et al. 2022). Severe forms of haemoglobinopathies, like β-thalassaemia major, can result in severe symptoms and complications if left untreated, posing substantial challenges and expenses in terms of patient management and long-term care.

HPFH is a benign genetic condition characterised by sustained production of high levels of HbF (>1%) throughout adulthood. This persistence arises from a disruption in the regular process of switching from foetal to adult Hb, where silencing of γ-globin gene expression is reduced (Sharma, Singhal et al. 2020). Research has indicated that the presence of HPFH alongside other haemoglobinopathies can ameliorate the severity of clinical symptoms by elevating HbF levels. Hence, investigating the molecular defects underlying HPFH may provide valuable insights into the mechanisms regulating γ-globin gene expression, thereby identifying potential therapeutic targets (Wilber, Nienhuis et al. 2011).

The initial discovery of the first Maltese family with HPFH due to a nonsense mutation p.K288X in the *KLF1* gene dates back to 2010 (Borg, Papadopoulos et al. 2010). Subsequently, five additional Maltese families carrying the same p.K288X truncation mutation in the *KLF1* gene were identified. However, these families exhibited normal or slightly increased HbF levels and borderline HbA_2 (Grech, Borg et al. 2020). This study focuses on three Maltese families (Fam F1, Fam F2, and Fam F4) affected by HPFH due to a truncation mutation in the *KLF1* gene. A total of 11 individuals with HPFH and 11 healthy relatives were recruited for this familial study. To gain valuable insights into the underlying molecular mechanisms involved in globin gene switching, genomic and proteomic techniques were employed with the aim of potentially identifying candidate modifier genes associated with this condition.

To assess the impact of KLF1 deficiency on antigen levels and mRNA globin gene expression in individuals with HPFH, flow cytometry and an mRNA globin gene expression assay were carried out. Individuals affected with HPFH displayed elevated mRNA expression levels of the $^{A}\gamma$-globin gene compared to their otherwise healthy relatives, as substantiated by existing literature (Wienert, Martyn et al. 2017). Conversely, the presence of HPFH was associated with decreased mRNA expression of the α- and β-globin genes. Additionally, flow cytometry data revealed that individuals carrying the *KLF1* p.K288X mutation exhibited prominently reduced levels of BCAM and CD44 antigens, with a lesser extent observed for P1 antigen. This was demonstrated by both a decrease in the percentage of positive cells detected, and by the reduced gMFI values in the HPFH affected group.

Mutations in the *KLF1* gene have been related to altered blood group phenotypes. Examples of this include CDA (Arnaud, Saison et al. 2010) and the rare In(Lu) phenotype, characterised by markedly diminished levels of BCAM, which carries the Lutheran blood group antigens, and CD44, which carries the Indian blood group antigens (Borg, Patrinos et al. 2011; Kawai, Obara et al. 2017; Singleton, Burton et al. 2008). Furthermore, individuals affected with the In(Lu) phenotype also exhibited reduced levels of the P1 antigen, which forms part of the P1PK blood group system (Bruce 2018). The finding of the reduced levels of these three antigens in individuals with HPFH due to KLF1 haploinsufficiency, emphasises the regulatory role of KLF1 not only in the globin gene switching mechanism but also in the modulation of specific red cell antigens (Eernstman, Veldhuisen et al. 2021; Kawai, Obara et al. 2017).

The outcomes derived from the flow cytometry data analysis and mRNA expression assays were further validated by the MS data. Consistent with the flow cytometry results, the MS analysis confirmed reduced protein levels of both BCAM and CD44 antigens in individuals affected with HPFH. Additionally, the measurement of different Hb protein subunits using MS concurred with the mRNA expression assay findings, where foetal globin subunits (HBG1 and HBG2) as well as the embryonic HbZ globin subunit were found at higher levels in affected individuals, while adult HBA1 and HBB globin protein subunits were more abundant in the otherwise healthy subjects. This was further confirmed by the MS correlation analysis between protein levels and levels of HbF in individuals affected by HPFH caused by the *KLF1* p.K288X mutation. It was revealed that the adult globin subunits HBA1 and HBB, were among the top 20 proteins displaying the most pronounced negative correlation with HbF levels, while the foetal globin subunits HBG1 and HBG2, and the embryonic subunit HBZ, were among the top 20 proteins exhibiting a strong positive correlation with HbF levels.

The MS correlation analysis also identified additional proteins known to be involved in haem biosynthesis or are associated with haemoglobinopathies, that exhibit a significant correlation with levels of HbF in HPFH-KLF1 haploinsufficient individuals. These proteins encompassed BPGM (Petousi, Copley et al. 2014), HMBS (Zheng, Xu et al. 2018), KCNN4 (Allegrini, Jedele et al. 2022), ACHE (Prall, Gambhir et al. 1998), CSNK1A1 (Schneider, Ademà et al. 2014), and UROS (Iolascon, De Falco et al. 2009). However, the precise role or contribution of these proteins in the persistence of elevated HbF levels remains unknown, necessitating further investigations.

Subsequently, WGS was performed on a subset of 20 individuals, including 11 affected individuals and 9 healthy relatives, to identify potential variants related to HPFH. The application of a series of filtering steps outlined in Section 3.4.1 facilitated the identification of candidate causal variants. Given the absence of prior knowledge regarding specific variants and the mode of inheritance, both dominant and recessive inheritance patterns were investigated in this study. Notably, no variants were detected when filtering for autosomal recessive inheritance pattern. Conversely, a total of 205 unique variants were identified when applying an autosomal dominant inheritance pattern to the entire cohort. These variants were confirmed to be present in all affected individuals with HPFH from the three families while being absent in all unaffected individuals. Interestingly, an analysis of the identified variant list revealed that all 205 unique variants resided on chromosome 19 within a concentrated region spanning approximately 5 million bp, in close proximity to the *KLF1* gene (Figure 4.1).

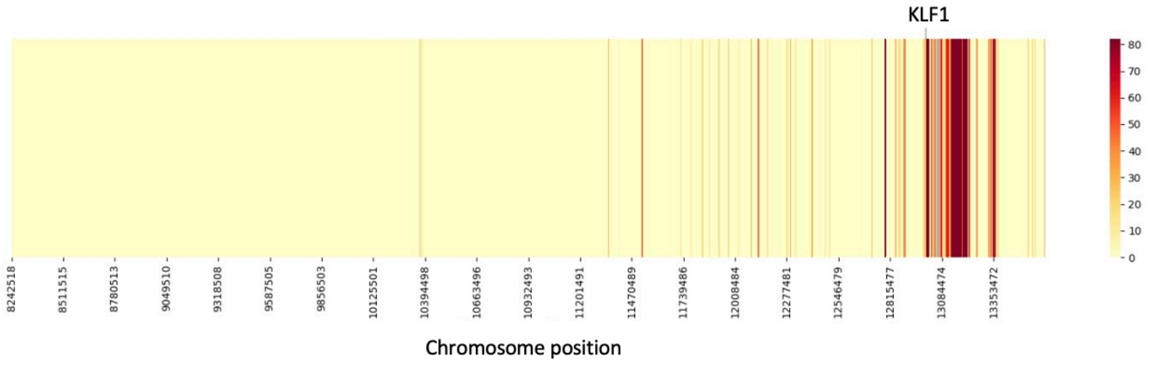


Figure 4.1: Localised heatmap of chromosome 19 showing the entire range of chromosome positions of the unique 205 variants following a dominant inheritance pattern identified in all individuals affected with HPFH. The scale on the right shows the number of variants that were called by the variant caller at the respective positions on chromosome 19.

The co-occurrence of 205 unique variants on chromosome 19 among all individuals affected by HPFH, indicates that these variants are likely genetically linked and hence inherited together as a group more frequently than would be expected by chance. This finding suggests the presence of long-range linkage disequilibrium (LRLD), where distantly located variants on a chromosome exhibit non-random associations, known as linkage disequilibrium (LD) (Pritchard, Przeworski 2001). The presence of LRLD further implies that these variants on chromosome 19 may collectively play a role in the regulation of genes or molecular pathways involved in the expression of HbF and in the onset of the HPFH condition. However, this does not establish a functional relationship between the variants, nor does it confirm that these variants are directly causing HPFH. Therefore, additional experimental and functional studies are required to determine the functional significance of these variants.

A curated list of genes of interest (Table 3.6) was generated as explained in Section 3.4.2 and was used for variant analysis. These genes were selected based on their significance in the foetal to adult globin switch, their pivotal role in erythropoiesis, or their association with haemoglobinopathies. In the examination of an autosomal dominant inheritance pattern within the entire cohort, a total of five variants were identified on two genes of interest: four variants were located on the *KLF1* gene, and one variant was identified on the *LMO2* gene (Table 3.7).

The variants discovered on the *KLF1* gene were observed to be present in all individuals affected by HPFH. Among these variants, was the rare stop gain mutation (rs267607202, p.K288X), initially reported by Borg, Papadopoulos et al. in 2010. This mutation was classified as a high impact variant as it completely abolishes the DNA binding domain of the key erythroid transcription factor KLF1. All affected family members were confirmed to be heterozygous carriers of this variant. Three additional novel mutations were detected in the *KLF1* gene, classified as having a modifier impact. The mutation rs112348773 represented a downstream SNP variant resulting in a G>C base pair change. The two other mutations were upstream variants, with rs112943513 being an SNP causing a G>T base pair substitution, and rs1395948183 denoting a deletion of CACAG>C. Considering that the *KLF1* gene resides on chromosome 19, the presence of these variants in all affected individuals from the three families suggests the existence of LD, due to their co-inheritance with the truncation variant rs267607202.

The LIM Domain Only 2 (LMO2) transcription factor has been recognized as a master regulator of erythroid differentiation, as proven by studies demonstrating that LMO2-deficient mice experience prenatal death due to impaired embryonic yolk sac erythropoiesis (Warren, Colledge et al. 1994). Studies have also established that *LMO2* and *KLF1* are co-activated by the transcription factor c-myb to promote erythropoiesis (Bianchi, Zini et al. 2010). In this study, an intronic variant (rs57617009) located on the *LMO2* gene situated on chromosome 11, was identified. This variant represents a deletion of AACAC>A and was categorised as having a modifier impact. It was verified to be present in six out of the eleven individuals affected by HPFH, while it was absent in five out of the nine unaffected subjects. Although not observed in all affected individuals, this variant was deemed representative within the studied cohort, suggesting its potential involvement in the onset of the condition.

An independent analysis was subsequently conducted to identify variants that potentially contribute to the onset of higher HbF levels within the affected cohort. This analysis specifically focused on four individuals, namely F1.2, F1.5, F1.9, and F1.10, who exhibited the highest HbF levels (>3%). These individuals belonged to Fam F1, which is known for having significantly elevated HbF levels when compared to other families affected with HPFH, despite carrying the same truncation mutation p.K288X in the *KLF1* gene (Borg, Papadopoulos et al. 2010). Variant filtering methods were specifically applied to this subgroup, aiming to identify potential causative variants unique to Fam F1 that contribute to the observed heightened HbF levels among its affected members.

Upon filtering for variants that follow an autosomal dominant inheritance pattern in the cohort exhibiting the highest HbF levels, a total of 272 distinct variants were identified. These variants were confirmed to be present in all four individuals with the most elevated

HbF levels (>3%), while being absent in all other subjects. Interestingly, three variants were detected within the *NLRP3* gene located on chromosome 1, which were found in either a homozygous alternate state or a heterozygous state in the four individuals exhibiting the highest HbF levels (Table 4.1). Two of the identified mutations in the *NLRP3* gene were SNPs, resulting in an A>G nucleotide change at two distinct positions in the gene (chr1:247593142 and chr1:247596321), while the third mutation entailed a deletion at chr1:247596971.

Table 4.1: Three variants identified on *NLRP3* gene in a dominant inheritance mode showing the zygosity status of each individual, where a homozygous reference status is represented as a 0|0, a heterozygous alternate status is represented as 0|1, and a homozygous alternate status is represented as 1|1. The four individuals with the highest HbF levels are marked in red.

Family	Member	HPFH status	*NLRP3* 1:247593142 A>G	*NLRP3* 1:247596321 A>G	*NLRP3* 1:247596971 AGAGCTCTCTGGTCA GGTGTGTCCTGATGC TTTCTCTATTCCG>A
Fam F1	F1.1	Control	0|0	0|0	0|0
	F1.2	Affected	0|1	0|1	0|1
	F1.3	Control	0|0	0|0	0|0
	F1.4	Affected	0|0	0|0	0|0
	F1.5	Affected	0|1	0|1	1|1
	F1.6	Affected	0|0	0|0	0|0
	F1.7	Control	0|0	0|0	0|0
	F1.9	Affected	0|1	0|1	0|1
	F1.10	Affected	1|1	1|1	0|1
	F1.12	Control	0|0	0|0	0|0
Fam F2	F2.1	Affected	0|0	0|0	0|0
	F2.2	Control	0|0	0|0	0|0
	F2.3	Control	0|0	0|0	0|0
	F2.4	Affected	0|0	0|0	0|0
Fam F4	F4.1	Control	0|0	0|0	0|0
	F4.2	Affected	0|0	0|0	0|0
	F4.3	Control	0|0	0|0	0|0
	F4.4	Affected	0|0	0|0	0|0
	F4.5	Affected	0|0	0|0	0|0
	F4.6	Control	0|0	0|0	0|0

The *NLRP3* gene can become activated to assemble the NLRP3 inflammasome, which contributes to anaemia upon activation in myeloid cells (Wang, Xu et al. 2017). Mutations within the *NLRP3* gene have been associated with hyperactivation of the NLRP3 inflammasome, leading to mild anaemia and impaired erythropoiesis in the bone marrow (Brill, Bourne et al. 2023). Disorders such as SCD and myelodysplastic syndromes (MDS) have also been linked to upregulated expression of the *NLRP3* gene and aberrant activation of the inflammasome (Basiorka, McGraw et al. 2016; Vogel, Kaminura et al. 2021). The specific functional consequences of the identified *NLRP3* gene variants in the four subjects exhibiting the highest HbF levels, are still unknown. Hence, further investigation into whether these mutations enhance or inhibit the activation of the NLRP3 inflammasome is required.

Interestingly, when performing the correlation analysis to examine the relationship between HbF levels and the significant proteins identified and quantified by MS, the NAD-dependent protein deacetylase sirtuin-2 (SIRT2), encoded by the *SIRT2* gene, exhibited the strongest positive correlation in individuals affected with HPFH. Extensive research has demonstrated that states of mitochondrial stress trigger aberrant activation of the NLRP3 inflammasome, and the SIRT2 protein is crucial for the maintenance and regeneration of HSCs by suppressing the activation of the NLRP3 inflammasome in an autonomous manner (Luo, Mu et al. 2019; Yang, Hu et al. 2021). The discovery of a robust positive correlation between SIRT2 protein levels and HbF levels in individuals with HPFH suggests the hypothesis that mutations in the *NLRP3* gene among the four individuals with the highest HbF levels (>3%) may lead to enhanced activation of the NLRP3 inflammasome. Consequently, elevated levels of SIRT2 proteins are required to inhibit this complex and maintain homeostasis. However, further experimental studies would be necessary to fully understand the functional implication of these proteins and to establish a direct causal relationship that validates the proposed hypothesis.

Additionally, another study revealed that SIRT2 protein interacts with the LMO2 transcription regulator and triggers its deacetylation, thereby inducing early stages of haematopoietic differentiation (Morishima, Bernhard et al. 2015). This interaction could also potentially account for the strong positive correlation observed with HbF levels, resulting in elevated levels of SIRT2 protein in individuals with HPFH, especially in the four individuals from Fam F1 with the highest HbF levels (>3%).

A total of 200 unique variants exhibiting an autosomal recessive inheritance pattern were identified within the cohort having the highest HbF levels. These variants were unequivocally present in a homozygous alternate state in the four individuals (F1.2, F1.5, F1.9 and F1.10) from Fam F1 who exhibited the highest levels of HbF (>3%), while they were either found in a heterozygous state or a homozygous reference state in all other subjects. The retention of 200 unique homozygous alternate variants within the restricted subset of the four individuals displaying the highest HbF levels is of significant interest, due to the rarity of such variants. Particularly noteworthy were the variants identified on the *RPS9* gene located on chromosome 19 (Table 4.2), which encodes ribosomal protein 9 that constitutes the 40S subunit of the ribosome and has been associated with Diamond-Blackfan anaemia (DBA) (Farrar, Vlachos et al. 2011).

Table 4.2: Two variants identified on *RPS9* gene in a recessive inheritance mode showing the zygosity status of each individual, where a homozygous reference status is represented as a 0|0, a heterozygous alternate status is represented as 0|1, and a homozygous alternate status is represented as 1|1. The four individuals with the highest HbF levels are marked in red.

Family	**Member**	**HPFH status**	***RPS9*** 19:54733630 **A>AT**	***RPS9*** 19:54733640 **G>C**
Fam F1	F1.1	Control	0|1	0|1
	F1.2	Affected	1|1	1|1
	F1.3	Control	0|1	0|1
	F1.4	Affected	0|1	0|1
	F1.5	Affected	1|1	1|1
	F1.6	Affected	0|1	0|1
	F1.7	Control	0|1	0|1
	F1.9	Affected	1|1	1|1
	F1.10	Affected	1|1	1|1
	F1.12	Control	0|1	0|1
Fam F2	F2.1	Affected	0|1	0|1
	F2.2	Control	0|1	0|1
	F2.3	Control	0|1	0|1
	F2.4	Affected	0|1	0|1
Fam F4	F4.1	Control	0|1	0|1
	F4.2	Affected	0|1	0|1
	F4.3	Control	0|0	0|0
	F4.4	Affected	0|1	0|1
	F4.5	Affected	0|0	0|0
	F4.6	Control	0|0	0|0

Individuals afflicted with DBA present with abnormal laboratory parameters, including macrocytic anaemia, deficiency of red cell precursors and elevated HbF levels (Halperin, Freedman 1989). Experimental investigations involving zebrafish, where mutations were induced in the *RPS9* gene, have provided compelling evidence regarding its essential role in the normal maturation of red blood cells. Disruption of this gene resulted in anaemia attributed to the inhibition of terminal erythrocyte maturation (Chen, Huang et al. 2019; Moetter, Kartal et al. 2011).

In this study, two mutations were detected within the *RPS9* gene, which included an insertion of a T nucleotide at position chr19:54733630, and a SNP involving a G to C substitution at position chr19:54733640 (Table 4.2). The presence of these two variants within the cohort characterised by the highest HbF levels (>3%) might explain the persistence of higher HbF levels observed individuals originating from Fam F1, distinguishing them from other families similarly affected by HPFH due to KLF1 deficiency. Nevertheless, further experimental investigations are required to assess the functional impact of these variants and their relation to HPFH and HbF levels.

The comprehensive genomic data obtained through WGS, examination of mRNA globin gene expression, and proteomic analysis using MS have yielded valuable insights into the distinctions between individuals affected by HPFH due to KLF1 deficiency and their otherwise healthy relatives. These findings provide a deeper understanding of the underlying molecular mechanisms and protein expression profiles associated with the *KLF1* mutation, elucidating its influence on globin gene expression.

Limitations of the study

One limitation of this study is the potential introduction of selection bias when recruiting the three Maltese families affected by HPFH due to *KLF1* mutations. Among the six families affected by KLF1 haploinsufficiency identified in Malta, Fam F1 was specifically chosen due to its notable differences in HbF levels when compared to the other families. However, the selection of the remaining two families was conducted randomly. This approach may have introduced a form of selection bias, as the study population may not be fully representative of all families affected by KLF1 haploinsufficiency in Malta. The study might also have been limited by a relatively small sample size, which could affect the generalisability of the findings. A larger sample size would yield more robust results and enable more comprehensive statistical analysis. Yet *KLF1* mutations and deficiency is a rare occurrence, making it challenging to access and recruit new families for the study. Partnering with international research groups may increase the number of families with KLF1 deficiency, allowing for more extensive investigations into the effects of KLF1 background.

WGS and subsequent data analysis to obtain the raw unfiltered VCF files was carried out at Dante Genomics facilities (Italy) using the Illumina® DRAGEN™ Bio-IT platform. The analysis involved mapping the sequencing reads to the human reference genome GRCh37 instead of the latest release GRCh38. Unfortunately, the choice to utilise the outdated GRCh37 reference genome in the data analysis process was beyond our control. Despite its outdated status, GRCh37 remains widely adopted by most research and clinical laboratories as the primary reference in various human genetic tools (Guo, Dai et al. 2017; Li, Dawood et al. 2021).

The WGS analysis revealed the presence of variants of interest on the *KLF1* and *LMO2* genes, suggesting their potential involvement in the onset of HPFH as evidenced by their presence in the affected individuals. Additionally, variants on the *NLRP3* and *RPS9* genes were identified within the cohort consisting of four individuals from Fam F1 exhibiting the highest HbF levels (>3%), which may contribute to the sustained elevation of HbF levels in Fam F1. However, the use of *in silico* tools to predict the variant effect and conducting any *in vitro* functional work were not possible due to time constraints. This hence, prevented a comprehensive assessment of the influence of the identified variants on the observed phenotype.

The study employed specific technologies and instruments, such as Luminex and LC-MS/MS, for data acquisition and analysis. These techniques have their own limitations and technical challenges and understanding these limitations is crucial to accurately interpret the results. One limitation when carrying out the mRNA gene expression assay, was that the measurement of mRNA expression was performed from frozen whole blood samples, which has resulted in reduced mRNA expression levels. Additionally, only the mRNA expression of three globin genes (*HBA1*, *HBB*, and *HBG1*) were included in this assay. Utilising fresh blood samples for the mRNA expression assay and carrying out immediate processing would have significantly improved the yield of the mRNA expression levels. Furthermore, the inclusion of additional target genes to assess their mRNA expression would have provided a more comprehensive analysis. However, due to time constraints and limited access to fresh blood samples, these improvements were not feasible in the current study.

The presence of high variability in the data, as indicated by the standard deviations and box plots, suggests the existence of additional factors, such as environmental influences contributing to the observed gene expression patterns and protein profiles, which were not accounted for in the study.

Future work

To validate the findings of WGS, future work should focus on conducting functional studies using cell or animal models to investigate the impact of the identified variants on gene expression and protein function. Techniques such as CRISPR/Cas9-mediated gene knockdown can be employed to establish a causal relationship between the mutations and the observed molecular changes, by targeting and disrupting the identified variants of interest within cell models (Shinmyo, Tanaka et al. 2016; Straub, Granger et al. 2014). The application of this technique enables direct observation of the effects resulting from genetic changes, facilitating the establishment of causal relationships between the identified variants and the observed molecular and phenotypic alterations.

Additional bioinformatics analyses can be carried out to support the identification of key genes. One approach could involve applying a gradient of zygosity filters, gradually eliminating the "no calls" from the WGS data. This stepwise filtering process would ensure that variants of interest are not lost due to overly stringent zygosity filtering. Implementing these gradient zygosity filters requires additional time and computational resources, which were limited in the current study.

To enhance the mRNA gene expression assay, the analysis should be repeated using freshly processed whole blood to ensure optimal mRNA yield. Additionally, the investigation of other target genes alongside the globin genes would be beneficial. Single-cell RNA sequencing techniques can be employed on HbF-producing erythroblasts to identify key transcripts having differential expression in cases of high and low HbF levels. However, it is crucial to remember that single-cell RNA sequencing may overlook genes exhibiting weak expression (Deng, Ramsköld et al. 2014; Islam, Zeisel et al. 2013). To obtain a more comprehensive understanding of differential gene expression, transcriptomic analysis by high throughput RNA sequencing (RNAseq) can be conducted in the future. RNAseq involves profiling gene expression across the entire transcriptome to reveal broader dysregulation patterns and identify potential molecular pathways involved (Sá, Sadee et al. 2017).

Further exploration of protein expression profiles could provide valuable insights into the downstream effects of KLF1 deficiency. Quantitative proteomics approaches, such as targeted protein quantification, can be employed to characterise protein alterations and their functional implications (Deracinois, Flahaut et al. 2013). Considering the observed correlations between protein levels and HbF levels, deeper investigations could be conducted to explore the functional relationship and underlying mechanisms. This could involve studying the effect of specific proteins on HbF expression and elucidating their regulatory roles in globin gene switching. In addition, the proteomic data obtained by MS can also be employed in future multivariate studies, which involve simultaneous analysis of multiple proteins to investigate their combined contributions to the elevated HbF levels observed in individuals affected with HPFH (Long, Veenstra 2013).

Conclusion

The integration of advanced molecular tools in this study, including flow cytometry, WGS, mRNA expression assay and MS, have provided crucial insights into the regulation of globin gene expression and identified potential modifier genes involved in HPFH. Through WGS analysis, a comprehensive set of 205 distinct variants co-occurring on chromosome 19, were identified among all individuals affected by HPFH due to KLF1 haploinsufficiency. These findings strongly suggest the presence of LRLD and implies the possible significance of these variants in the development of HPFH. Furthermore, novel variants were discovered within the *KLF1* gene and the *LMO2* gene, thereby suggesting their significant contributory roles in the onset of HPFH. An independent analysis focusing on a subset of individuals from Fam F1, characterised by significantly elevated HbF levels (>3%) despite carrying the same *KLF1* truncation mutation, revealed potential candidate causal variants on the *NLRP3* and *RPS9* genes. These variants may play a role in sustaining high HbF levels in this family. Although complete functional validation of these variants could not be achieved within the constraints of this study's timeframe, the identification of these potential genetic contributors lays the groundwork for future functional studies. Additionally, through the analysis of MS data, 53 proteins were identified displaying significant correlation with HbF levels, thus proposing their active involvement in the process of globin gene regulation. However, the specific role of the identified proteins in the upregulation of HbF levels remains unknown.

Validating these findings via functional studies using cell or animal models and CRISPR/Cas9-mediated gene knockdown, will be crucial to establish a causal relationship between the identified variants and observed molecular changes. Incorporating single-cell RNA sequencing on HbF-producing erythroblasts will identify key transcripts with differential expression in high and low HbF levels, offering deeper insights into genes influencing HbF production. The application of a multiomics approach and integration of cutting-edge molecular techniques, have advanced our understanding of the complex interplay between genetic variants, gene expression profiles, and protein levels in HPFH. This study, hence, establishes a strong foundation for future investigations aiming to identify potential therapeutic targets in individuals with haemoglobinopathies.

References

ADAN, A., ALIZADA, G., KIRAZ, Y., BARAN, Y., and NALBANT, A., 2017. Flow cytometry: Basic principles and applications. *Critical Reviews in Biotechnology*, 37(2), pp.163–176.

ADEMA, V., MA, F., KANAGAL-SHAMANNA, R., THONGON, N., MONTALBAN-BRAVO, G., YANG, H., PESLAK, S.A., WANG, F., ACHA, P., SOLE, F., LOCKYER, P., CASSARI, M., MACIEJEWSKI, J.P., VISCONTE, V., GAÑÁN-GÓMEZ, I., SONG, Y., BUESO-RAMOS, C., PELLEGRINI, M., TAN, T.M., BEJAR, R., CAREW, J.S., HALENE, S., SANTINI, V., AL-ATRASH, G., CLISE-DWYER, K., GARCIA-MANERO, G., BLOBEL, G.A., and COLLA, S., 2022. Targeting the EIF2AK1 signaling pathway rescues red blood cell production in *sf3b1*-mutant myelodysplastic syndromes with ringed sideroblasts. *Blood Cancer Discovery*, 3(6), pp.554–567.

AERBAJINAI, W., ZHU, J., GAO, Z., CHIN, K. and RODGERS, G., 2007. Thalidomide induces γ-globin gene expression through increased reactive oxygen species–mediated p38 MAPK signaling and histone H4 acetylation in adult erythropoiesis. *Blood*, 110(8), pp.2864-2871.

AHMED, M., GHATGE, M. and SAFO, M., 2020. Hemoglobin: Structure, Function and Allostery. *Subcellular Biochemistry*, 94, pp.345-382.

ALLEGRINI, B., JEDELE, S., DAVID NGUYEN, L., MIGNOTET, M., RAPETTI-MAUSS, R., ETCHEBEST, C., FENNETEAU, O., LOUBAT, A., BOUTET, A., THOMAS, C., DURIN, J., PETIT, A., BADENS, C., GARÇON, L., DA COSTA, L. and GUIZOUARN, H., 2022. New KCNN4 variants associated with anemia: Stomatocytosis without erythrocyte dehydration. *Frontiers in Physiology*, 13.

AMAYA, M., DESAI, M., GNANAPRAGASAM, M.N., WANG, S.Z., ZU ZHU, S., WILLIAMS, D.C., and GINDER, G.D., 2013. Mi2β-mediated silencing of the fetal γ-globin gene in adult erythroid cells. *Blood*, 121(17), pp.3493–3501.

AREVALO, R., NI, Z., and DANELL, R.M., 2019. Mass spectrometry and planetary exploration: A brief review and future projection. *Journal of Mass Spectrometry*, 55(1), p.e4454.

ARNAUD, L., SAISON, C., HELIAS, V., LUCIEN, N., STESCHENKO, D., GIARRATANA, M.-C., PREHU, C., FOLIGUET, B., MONTOUT, L., DE BREVERN, A.G., FRANCINA, A., RIPOCHE, P., FENNETEAU, O., DA COSTA, L., PEYRARD, T., COGHLAN, G., ILLUM, N., BIRGENS, H., TAMARY, H., IOLASCON, A., DELAUNAY, J., TCHERNIA, G., and CARTRON, J.-P., 2010. A dominant mutation in the gene encoding the erythroid transcription factor KLF1 causes a congenital dyserythropoietic anemia. *The American Journal of Human Genetics*, 87(5), pp.721–727.

BASIORKA, A.A., MCGRAW, K.L., EKSIOGLU, E.A., CHEN, X., JOHNSON, J., ZHANG, L., ZHANG, Q., IRVINE, B.A., CLUZEAU, T., SALLMAN, D.A., PADRON, E., KOMROKJI, R., SOKOL, L., COLL, R.C., ROBERTSON, A.A., COOPER, M.A., CLEVELAND, J.L., O'NEILL, L.A., WEI, S. and LIST, A.F., 2016. The NLRP3 inflammasome functions as a driver of the myelodysplastic syndrome phenotype. *Blood*, 128(25), pp.2960–2975.

BECKMAN, B., SILBERSTEIN, P. and ALDOSS, I., 2010. Erythropoiesis. In: *xPharm: The Comprehensive Pharmacology Reference*, 1st ed. Elsevier, pp.1-4.

BIANCHI, E., ZINI, R., SALATI, S., TENEDINI, E., NORFO, R., TAGLIAFICO, E., MANFREDINI, R. and FERRARI, S., 2010. C-myb supports erythropoiesis through the transactivation of KLF1 and LMO2 expression. *Blood*, 116(22), pp.e99–e110.

BIEKER, J.J., 2010. Putting a finger on the switch. *Nature Genetics*, 42(9), pp.733–734.

BOLLEKENS, J. and FORGET, B., 1991. Delta beta thalassemia and hereditary persistence of fetal hemoglobin. *Hematology/Oncology clinics of North America*, 5(3), pp.399-422.

BORG, J., PAPADOPOULOS, P., GEORGITSI, M., GUTIÉRREZ, L., GRECH, G., FANIS, P., PHYLACTIDES, M., VERKERK, A.J., VAN DER SPEK, P.J., SCERRI, C.A., CASSAR, W., GALDIES, R., VAN IJCKEN, W., ÖZGÜR, Z., GILLEMANS, N., HOU, J., BUGEJA, M., GROSVELD, F.G., VON LINDERN, M., FELICE, A.E., PATRINOS, G.P., and PHILIPSEN, S., 2010. Haploinsufficiency for the erythroid transcription factor KLF1 causes hereditary persistence of fetal hemoglobin. *Nature Genetics*, 42(9), pp.801–805.

BORG, J., PATRINOS, G.P., FELICE, A.E., and PHILIPSEN, S., 2011. Erythroid phenotypes associated with KLF1 mutations. *Haematologica*, 96(5), pp.635–638.

BRACKEN, A.P., BRIEN, G.L. and VERRIJZER, C.P., 2019. Dangerous liaisons: interplay between SWI/SNF, NuRD, and Polycomb in chromatin regulation and cancer. *Genes & Development*, 33(15-16), pp.936–959. doi:10.1101/gad.326066.119.

BRADNER, J.E., MAK, R., TANGUTURI, S.K., MAZITSCHEK, R., HAGGARTY, S.J., ROSS, K., CHANG, C.Y., BOSCO, J., WEST, N., MORSE, E., LIN, K., SHEN, J.P., KWIATKOWSKI, N.P., GHELDOF, N., DEKKER, J., DEANGELO, D.J., CARR, S.A., SCHREIBER, S.L., GOLUB, T.R., and EBERT, B.L., 2010. Chemical genetic strategy identifies histone deacetylase 1 (HDAC1) and HDAC2 as therapeutic targets in sickle cell disease. *Proceedings of the National Academy of Sciences*, 107(28), pp.12617–12622.

BRAGHINI, C., COSTA, F., FEDOSYUK, H., NEADES, R., NOVIKOVA, L., PARKER, M., WINEFIELD, R. and PETERSON, K., 2016. Original Research: Generation of non-deletional hereditary persistence of fetal hemoglobin β-globin locus yeast artificial chromosome transgenic mouse models: -175 Black HPFH and -195 Brazilian HPFH. *Experimental Biology and Medicine*, 241(7), pp.697-705.

BRAND, M., and RANISH, J.A., 2021. Proteomic/transcriptomic analysis of erythropoiesis. *Current Opinion in Hematology*, 28(3), pp.150–157.

BRILL, A., BOURNE, J., CAMPOS, J., HOPKIN, S., WHITWORTH, K., PALIS, J., SENIS, Y., RAYES, J. and IQBAL, A., 2023. Megakaryocyte NLRP3 hyperactivation induces anemia and potentiates inflammatory response in mice. *Research Square*.

BROWN, N.A., AISNER, D.L., and OXNARD, G.R., 2018. Precision medicine in non–small cell lung cancer: Current standards in pathology and Biomarker Interpretation. *American Society of Clinical Oncology Educational Book*, 38, pp.708–715.

BRUCE, L.J., 2018. Molecular mechanism of P1 antigen expression. *Blood*, 131(14), pp.1505–1506.

CANVER, M., SMITH, E., SHER, F., PINELLO, L., SANJANA, N., SHALEM, O., CHEN, D., SCHUPP, P., VINJAMUR, D., GARCIA, S., LUC, S., KURITA, R., NAKAMURA, Y., FUJIWARA, Y., MAEDA, T., YUAN, G., ZHANG, F., ORKIN, S. and BAUER, D., 2015. BCL11A enhancer dissection by Cas9-mediated in situ saturating mutagenesis. *Nature*, 527, pp.192-197.

CAO, A. and MOI, P., 2002. Regulation of the Globin Genes. *Pediatric Research*, 51, pp.415-420.

CARR, B., RAHBAR, S., ASMERON, Y., RIGGS, A. and WINBERG, C., 1988. Carcinogenicity and haemoglobin synthesis induction by cytidine analogues. *British Journal of Cancer*, 57(4), pp.395-402.

CARROCINI, G., ZAMARO, P. and BONINI-DOMINGOS, C., 2011. What influences Hb fetal production in adulthood?. *Revista Brasileira de Hematologia e Hemoterapia*, 33(3), pp.231-236.

CAVAZZANA-CALVO, M., PAYEN, E., NEGRE, O., WANG, G., HEHIR, K., FUSIL, F., DOWN, J., DENARO, M., BRADY, T., WESTERMAN, K., CAVALLESCO, R., GILLET-LEGRAND, B., CACCAVELLI, L., SGARRA, R., MAOUCHE-CHRÉTIEN, L., BERNAUDIN, F., GIROT, R., DORAZIO, R., MULDER, G., POLACK, A., BANK, A., SOULIER, J., LARGHERO, J., KABBARA, N., DALLE, B., GOURMEL, B., SOCIE, G., CHRÉTIEN, S., CARTIER, N., AUBOURG, P., FISCHER, A., CORNETTA, K., GALACTEROS, F., BEUZARD, Y., GLUCKMAN, E., BUSHMAN, F., HACEIN-BEY-ABINA, S. and LEBOULCH, P., 2010. Transfusion independence and HMGA2 activation after gene therapy of human β-thalassaemia. *Nature*, 467, pp.318-322.

CHAMBERS, C.B., GROSS, J., PRATT, K., GUO, X., BYRNES, C., LEE, Y.T., LAVELLE, D., DEAN, A., MILLER, J.L., and WILBER, A., 2020. The mrna-binding protein IGF2BP1 restores fetal hemoglobin in cultured erythroid cells from patients with β-hemoglobin disorders. *Molecular Therapy - Methods & Clinical Development*, 17, pp.429–440.

CHARACHE, S., TERRIN, M., MOORE, R., DOVER, G., BARTON, F., ECKERT, S., MCMAHON, R. and BONDS, D., 1995. Effect of Hydroxyurea on the Frequency of Painful Crises in Sickle Cell Anemia. *New England Journal of Medicine*, 332(20), pp.1317-1322.

CHEN, C., HUANG, H., YAN, R., LIN, S. and QIN, W., 2019. Loss of rps9 in Zebrafish Leads to p53-Dependent Anemia. *G3 Genes|Genomes|Genetics*, 9(12), pp.4149–4157.

CHRISTENSEN, K., DUKHOVNY, D., SIEBERT, U., and GREEN, R., 2015. Assessing the costs and cost-effectiveness of genomic sequencing. *Journal of Personalized Medicine*, 5(4), pp.470–486.

CHRISTMAN, J., 2002. 5-Azacytidine and 5-aza-2′-deoxycytidine as inhibitors of DNA methylation: mechanistic studies and their implications for cancer therapy. *Oncogene*, 21(35), pp.5483-5495.

CHUI, D., FUCHAROEN, S. and CHAN, V., 2003. Hemoglobin H disease: not necessarily a benign disorder. *Blood*, 101(3), pp.791-800.

COLAH, R., GORAKSHAKAR, A. and NADKARNI, A., 2010. Global burden, distribution and prevention of β-thalassemias and hemoglobin E disorders. *Expert Review of Hematology*, 3(1), pp.103-117.

DEMICHEV, V., MESSNER, C.B., VERNARDIS, S.I., LILLEY, K.S., and RALSER, M., 2019. Dia-NN: Neural Networks and interference correction enable deep proteome coverage in high throughput. *Nature Methods*, 17(1), pp.41–44.

DEMIRCI, S., LEONARD, A. and TISDALE, J., 2020. Genome editing strategies for fetal hemoglobin induction in beta-hemoglobinopathies. *Human Molecular Genetics*, 29(R1), pp.R100-R106.

DE MONTALEMBERT, M., 2002. Prise en charge des enfants drépanocytaires : Un travail d'équipe. *Archives de Pédiatrie*, 9(11), pp.1195–1201.

DENG, Q., RAMSKÖLD, D., REINIUS, B. and SANDBERG, R., 2014. Single-cell RNA-seq reveals dynamic, random monoallelic gene expression in mammalian cells. *Science*, 343(6167), pp.193–196.

DEPRISTO, M.A., BANKS, E., POPLIN, R., GARIMELLA, K.V., MAGUIRE, J.R., HARTL, C., PHILIPPAKIS, A.A., DEL ANGEL, G., RIVAS, M.A., HANNA, M., MCKENNA, A., FENNELL, T.J., KERNYTSKY, A.M., SIVACHENKO, A.Y., CIBULSKIS, K., GABRIEL, S.B., ALTSHULER, D. and DALY, M.J., 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics*, 43(5), pp.491–498.

DERACINOIS, B., FLAHAUT, C., DUBAN-DEWEER, S. and KARAMANOS, Y., 2013. Comparative and quantitative global proteomics approaches: An overview. *Proteomes*, 1(3), pp.180–218.

DE SANCTIS, V., KATTAMIS, C., CANATAN, D., SOLIMAN, A., ELSEDFY, H., KARIMI, M., DAAR, S., WALI, Y., YASSIN, M., SOLIMAN, N., SOBTI, P., AL JAOUNI, S., EL KHOLY, M., FISCINA, B. and ANGASTINIOTIS, M., 2017. β-thalassemia distribution in the old world: a historical standpoint of an ancient disease. *Mediterranean Journal of Hematology and Infectious Diseases*, 9(1), p.e2017018.

DESJARDINS, P., and CONKLIN, D., 2010. Nanodrop microvolume quantitation of Nucleic Acids. *Journal of Visualized Experiments*, (45), pp.2565.

DONZE, D., TOWNES, T.M., and BIEKER, J.J., 1995. Role of erythroid kruppel-like factor in human γ- to β-globin gene switching. *Journal of Biological Chemistry*, 270(4), pp.1955–1959.

DRISSEN, R., PALSTRA, R.J., GILLEMANS, N., SPLINTER, E., GROSVELD, F., PHILIPSEN, S. and DE LAAT, W., 2004. The active spatial organization of the -globin locus requires the transcription factor EKLF. *Genes & Development*, 18(20), pp.2485–2490. doi:10.1101/gad.317004.

DZIERZAK, E. and PHILIPSEN, S., 2013. Erythropoiesis: Development and Differentiation. *Cold Spring Harbor Perspectives in Medicine*, 3(4), pp.a011601-a011601.

EDINGTON, G. and LEHMANN, H., 1955. Expression of the Sickle-cell Gene in Africa. *BMJ*, 1(4925), pp.1308-1311.

EERNSTMAN, J., VELDHUISEN, B., LIGTHART, P., VON LINDERN, M., VAN DER SCHOOT, C.E. and VAN DEN AKKER, E., 2021. Novel variants in Krueppel like factor 1 that cause persistence of fetal hemoglobin in In(Lu) individuals. *Scientific Reports*, 11(1). doi:10.1038/s41598-021-97149-y.

Encyclopedia Britannica, 2023-last update. Hemoglobin [Encyclopedia Britannica], [Online]. Available: www.britannica.com/science/hemoglobin [Apr 1, 2023].

FALLER, D. and PERRINE, S., 1995. Butyrate in the treatment of sickle cell disease and β-thalassemia. *Current Opinion in Hematology*, 2(2), pp.109-117.

FARD, A., HOSSEINI, S., SHAHJAHANI, M., SALARI, F. and JASEB, K., 2013. Evaluation of Novel Fetal Hemoglobin Inducer Drugs in Treatment of β-Hemoglobinopathy Disorders. *International Journal of Hematology-Oncology and Stem Cell Research*, 7(3), pp.47-54.

FARRAR, J.E., VLACHOS, A., ATSIDAFTOS, E., CARLSON-DONOHOE, H., MARKELLO, T.C., ARCECI, R.J., ELLIS, S.R., LIPTON, J.M. and BODINE, D.M., 2011. Ribosomal protein gene deletions in diamond-blackfan anemia. *Blood*, 118(26), pp.6943–6951.

FIBACH, E. and RACHMILEWITZ, E., 2017. Pathophysiology and treatment of patients with beta-thalassemia – an update. *F1000Research*, 6, pp.2156.

FILIPE, A., LI, Q., DEVEAUX, S., GODIN, I., ROMÉO, P.-H., STAMATOYANNOPOULOS, G., and MIGNOTTE, V., 1999. Regulation of embryonic/fetal globin genes by nuclear hormone receptors: A novel perspective on hemoglobin switching. *The EMBO Journal*, 18(3), pp.687–697.

FORGET, B., 1998. Molecular Basis of Hereditary Persistence of Fetal Hemoglobin. *Annals of the New York Academy of Sciences*, 850(1), pp.38-44.

FUENTES-PARDO, A.P., and RUZZANTE, D.E., 2017. Whole-genome sequencing approaches for conservation biology: Advantages, limitations and practical recommendations. *Molecular Ecology*, 26(20), pp.5369–5406.

GALANELLO, R. and ORIGA, R., 2010. Beta-thalassemia. *Orphanet Journal of Rare Diseases*, 5(1).

GONDA, T. and METCALF, D., 1984. Expression of myb, myc and fos proto-oncogenes during the differentiation of a murine myeloid leukaemia. *Nature*, 310(5974), pp.249-251.

GONG, J., PAN, K., FAKIH, M., PAL, S., and SALGIA, R., 2018. Value-based genomics. *Oncotarget*, 9(21), pp.15792–15815.

GONZAGA-JAUREGUI, C., LUPSKI, J.R., and GIBBS, R.A., 2012. Human genome sequencing in health and disease. *Annual Review of Medicine*, 63(1), pp.35–61.

GOODMAN, M. and MALIK, P., 2016. The potential of gene therapy approaches for the treatment of hemoglobinopathies: achievements and challenges. *Therapeutic Advances in Hematology*, 7(5), pp.302-315.

GOTHWAL, M., WEHRLE, J., AUMANN, K., ZIMMERMANN, V., GRUNDER, A., and PAHL, H.L., 2016. A novel role for nuclear factor-erythroid 2 in erythroid maturation by modulation of mitochondrial autophagy. *Haematologica*, 101(9), pp.1054–1064.

GRECH, L., BORG, J., GALDIES, R., ATTARD, C., SCERRI, C.A., PHILIPSEN, S. and FELICE, A.E., 2020. Genetic Heterogeneity of KLF1, a Master Regulator of Erythropoiesis, Revealed an Autosomal Recessive Ψβ-Thalassemia and a Very Strong Promoter In Vivo. *Blood*, 136(1). doi:10.1182/blood-2020-143018.

GREGORY, R., TAXMAN, D., SESHASAYEE, D., KENSINGER, M., BIEKER, J., and WOJCHOWSKI, D., 1996. Functional interaction of GATA1 with erythroid kruppel-like factor and SP1 at defined erythroid promoters. *Blood*, 87(5), pp.1793–1801.

GUO, Y., DAI, Y., YU, H., ZHAO, S., SAMUELS, D.C. and SHYR, Y., 2017. Improvements and impacts of GRCH38 human reference on high throughput sequencing data analysis. *Genomics*, 109(2), pp.83–90.

HACEIN-BEY-ABINA, S., GARRIGUE, A., WANG, G., SOULIER, J., LIM, A., MORILLON, E., CLAPPIER, E., CACCAVELLI, L., DELABESSE, E., BELDJORD, K., ASNAFI, V., MACINTYRE, E., DAL CORTIVO, L., RADFORD, I., BROUSSE, N., SIGAUX, F., MOSHOUS, D., HAUER, J., BORKHARDT, A., BELOHRADSKY, B., WINTERGERST, U., VELEZ, M., LEIVA, L., SORENSEN, R., WULFFRAAT, N., BLANCHE, S., BUSHMAN, F., FISCHER, A. and CAVAZZANA-CALVO, M., 2008. Insertional oncogenesis in 4 patients after retrovirus-mediated gene therapy of SCID-X1. *Journal of Clinical Investigation*, 118(9), pp.3132-3142.

HALPERIN, D.S. and FREEDMAN, M.H., 1989. Diamond-blackfan anemia: etiology, pathophysiology, and treatment. *The American journal of pediatric hematology/oncology*, 11(4), pp.380–394.

HAN, X., ASLANIAN, A., and YATES, J.R., 2008. Mass spectrometry for proteomics. *Current Opinion in Chemical Biology*, 12(5), pp.483–490.

HARDISON, R., 2012. Evolution of Hemoglobin and Its Genes. *Cold Spring Harbor Perspectives in Medicine*, 2(12), pp.a011627-a011627.

HARTEVELD, C.L., ACHOUR, A., ARKESTEIJN, S.J., TER HUURNE, J., VERSCHUREN, M., BHAGWANDIEN-BISOEN, S., SCHAAP, R., VIJFHUIZEN, L., EL IDRISSI, H. and KOOPMANN, T.T., 2022. The hemoglobinopathies, molecular disease mechanisms and diagnostics. *International Journal of Laboratory Hematology*, 44(S1), pp.28–36.

HARTEVELD, C. and HIGGS, D., 2010. α-thalassaemia. *Orphanet Journal of Rare Diseases*, 5(13).

HATTANGADI, S.M., WONG, P., ZHANG, L., FLYGARE, J. and LODISH, H.F., 2011. From stem cell to red cell: regulation of erythropoiesis at multiple levels by multiple proteins, RNAs, and chromatin modifications. *Blood*, 118(24), pp.6258–6268. doi:10.1182/blood-2011-07-356006.

HERRICK, J.B., 1910. Peculiar elongated and sickle-shaped red blood corpuscles in a case of severe anemia. *Yale Journal of Biology and Medicine*, 74(3), pp.179–184.

HOBAN, M.D., ORKIN, S.H., and BAUER, D.E., 2016. Genetic treatment of a molecular disorder: Gene therapy approaches to sickle cell disease. *Blood*, 127(7), pp.839–848.

HSIA, C., 1998. Respiratory Function of Hemoglobin. *New England Journal of Medicine*, 338(4), pp.239-248.

HUA-BING, Z., DE-PEI, L. and CHIH-CHUAN, L., 2002. The Control of Expression of the α-Globin Gene Cluster. *International Journal of Hematology*, 76(5), pp.420-426.

HUANG, J., ZHANG, X., LIU, D., WEI, X., SHANG, X., XIONG, F., YU, L., YIN, X. and XU, X., 2015. Compound heterozygosity for KLF1 mutations is associated with microcytic hypochromic anemia and increased fetal hemoglobin. *European Journal of Human Genetics*, 23, pp.1341–1348. doi:10.1038/ejhg.2014.291.

INGLEY, E., TILBROOK, P., and KLINKEN, S.P., 2004. New insights into the regulation of erythroid cells. *IUBMB Life (International Union of Biochemistry and Molecular Biology: Life)*, 56(4), pp.177–184.

INUSA, B., HSU, L., KOHLI, N., PATEL, A., OMINU-EVBOTA, K., ANIE, K., and ATOYEBI, W., 2019. Sickle cell disease—genetics, pathophysiology, clinical presentation and treatment. *International Journal of Neonatal Screening*, 5(2), p.20.

IOLASCON, A., DE FALCO, L. and BEAUMONT, C., 2009. Molecular basis of inherited microcytic anemia due to defects in iron acquisition or heme synthesis. *Haematologica*, 94(3), pp.395–408.

ISLAM, S., ZEISEL, A., JOOST, S., LA MANNO, G., ZAJAC, P., KASPER, M., LÖNNERBERG, P. and LINNARSSON, S., 2013. Quantitative single-cell RNA-seq with unique molecular identifiers. *Nature Methods*, 11(2), pp.163–166.

JANE, S.M., NIENHUIS, A.W., and CUNNINGHAM, J.M., 1995. Hemoglobin switching in man and chicken is mediated by a heteromeric complex between the ubiquitous transcription factor CP2 and a developmentally specific protein. *The EMBO Journal*, 14(1), pp.97–105.

JOHNSON, K.D., GRASS, J.A., BOYER, M.E., KIEKHAEFER, C.M., BLOBEL, G.A., WEISS, M.J., and BRESNICK, E.H., 2002. Cooperative activities of hematopoietic regulators recruit RNA polymerase II to a tissue-specific chromatin domain. *Proceedings of the National Academy of Sciences*, 99(18), pp.11760–11765.

KADAUKE, S. and BLOBEL, G.A., 2009. Chromatin loops in gene regulation. *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms*, 1789(1), pp.17–25. doi:10.1016/j.bbagrm.2008.07.002.

KARPONI, G. and ZOGAS, N., 2019. Gene Therapy For Beta-Thalassemia: Updated Perspectives. *The Application of Clinical Genetics*, Volume 12, pp.167-180.

KATSANIS, S.H., and KATSANIS, N., 2013. Molecular genetic testing and the future of Clinical Genomics. *Nature Reviews Genetics*, 14(6), pp.415–426.

KAWAI, M., OBARA, K., ONODERA, T., ENOMOTO, T., OGASAWARA, K., TSUNEYAMA, H., UCHIKAWA, M. and INABA, S., 2017. Mutations of the KLF1 gene detected in Japanese with the In(Lu) phenotype. *Transfusion*, 57(4), pp.1072–1077.

KEYS, J.R., TALLACK, M.R., ZHAN, Y., PAPATHANASIOU, P., GOODNOW, C.C., GAENSLER, K.M., CROSSLEY, M., DEKKER, J. and PERKINS, A.C., 2008. A mechanism for Ikaros regulation of human globin gene switching. *British Journal of Haematology*, 141(3), pp.398–406. doi:10.1111/j.1365-2141.2008.07065.x.

KOHNE, E., 2011. Hemoglobinopathies - Clinical Manifestations, Diagnosis, and Treatment. *Deutsches Aerzteblatt International*, 108(31-32), pp.532–540.

KRÜGER, I., VOLLMER, M., SIMMONS, D., ELSÄSSER, H.-P., PHILIPSEN, S., and SUSKE, G., 2007. Sp1/Sp3 compound heterozygous mice are not viable: Impaired erythropoiesis and severe placental defects. *Developmental Dynamics*, 236(8), pp.2235–2244.

KUVARDINA, O.N., HERGLOTZ, J., KOLODZIEJ, S., KOHRS, N., HERKT, S., WOJCIK, B., OELLERICH, T., CORSO, J., BEHRENS, K., KUMAR, A., HUSSONG, H., URLAUB, H., KOCH, J., SERVE, H., BONIG, H., STOCKING, C., RIEGER, M.A., and LAUSEN, J., 2015. Runx1 represses the erythroid gene expression program during megakaryocytic differentiation. *Blood*, 125(23), pp.3570–3579.

KWON, H., IMBALZANO, A.N., KHAVARI, P.A., KINGSTON, R.E. and GREEN, M.R., 1994. Nucleosome disruption and enhancement of activator binding by a human SW1/SNF complex. *Nature*, 370(6489), pp.477–481. doi:10.1038/370477a0.

LARSEN, M.R., TRELLE, M.B., THINGHOLM, T.E., and JENSEN, O.N., 2006. Analysis of posttranslational modifications of proteins by tandem mass spectrometry. *BioTechniques*, 40(6), pp.790–798.

LEE, P., COSTUMBRADO, J., HSU, C. and KIM, Y., 2012. Agarose Gel Electrophoresis for the Separation of DNA Fragments. *Journal of Visualized Experiments*, (62).

LEE, Y.T., DE VASCONCELLOS, J.F., YUAN, J., BYRNES, C., NOH, S.-J., MEIER, E.R., KIM, K.S., RABEL, A., KAUSHAL, M., MULJO, S.A., and MILLER, J.L., 2013. LIN28B-mediated expression of fetal hemoglobin and production of fetal-like erythrocytes from adult human erythroblasts ex vivo. *Blood*, 122(6), pp.1034–1041.

LEK, M., KARCZEWSKI, K.J., MINIKEL, E.V., SAMOCHA, K.E., BANKS, E., FENNELL, T., O DONNELL-LURIA, A.H., WARE, J.S., HILL, A.J., CUMMINGS, B.B., TUKIAINEN, T., BIRNBAUM, D.P., KOSMICKI, J.A., DUNCAN, L.E., ESTRADA, K., ZHAO, F., ZOU, J., PIERCE-HOFFMAN, E., BERGHOUT, J., COOPER, D.N., DEFLAUX, N., DEPRISTO, M., DO, R., FLANNICK, J., FROMER, M., GAUTHIER, L., GOLDSTEIN, J., GUPTA, N., HOWRIGAN, D., KIEZUN, A., KURKI, M.I., MOONSHINE, A.L., NATARAJAN, P., OROZCO, L., PELOSO, G.M., POPLIN, R., RIVAS, M.A., RUANO-RUBIO, V., ROSE, S.A., RUDERFER, D.M., SHAKIR, K., STENSON, P.D., STEVENS, C., THOMAS, B.P., TIAO, G., TUSIE-LUNA, M.T., WEISBURD, B., WON, H., YU, D., ALTSHULER, D.M., ARDISSINO, D., BOEHNKE, M., DANESH, J., DONNELLY, S., ELOSUA, R., FLOREZ, J.C., GABRIEL, S.B., GETZ, G., GLATT, S.J., HULTMAN, C.M., KATHIRESAN, S., LAAKSO, M., MCCARROLL, S., MCCARTHY, M.I., MCGOVERN, D., MCPHERSON, R., NEALE, B.M., PALOTIE, A., PURCELL, S.M., SALEHEEN, D., SCHARF, J.M., SKLAR, P., SULLIVAN, P.F., TUOMILEHTO, J., TSUANG, M.T., WATKINS, H.C., WILSON, J.G., DALY, M.J. and MACARTHUR, D.G., 2016. Analysis of protein-coding genetic variation in 60,706 humans. *Nature,* **536**(7616), pp. 285-291.

LETTRE, G., SANKARAN, V.G., BEZERRA, M.A., ARAÚJO, A.S., UDA, M., SANNA, S., CAO, A., SCHLESSINGER, D., COSTA, F.F., HIRSCHHORN, J.N., and ORKIN, S.H., 2008. DNA polymorphisms at the BCL11A, HBS1L-MYB, and β-globin loci associate with fetal hemoglobin levels and pain crises in sickle cell disease. *Proceedings of the National Academy of Sciences*, 105(33), pp.11869–11874.

LEY, T. and NIENHUIS, A., 1985. Induction of Hemoglobin F Synthesis in Patients with β Thalassemia. *Annual Review of Medicine*, 36(1), pp.485-498.

LI, H., DAWOOD, M., KHAYAT, M.M., FAREK, J.R., JHANGIANI, S.N., KHAN, Z.M., MITANI, T., COBAN-AKDEMIR, Z., LUPSKI, J.R., VENNER, E., POSEY, J.E., SABO, A. and GIBBS, R.A., 2021. Exome variant discrepancies due to reference-genome differences. *The American Journal of Human Genetics*, 108(7), pp.1239–1250.

LI, H. and DURBIN, R., 2009. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics*, 25(14), pp.1754–1760.

LIEBHABER, S. and RUSSEL, J., 1998. Expression and Developmental Control of the Human alpha-Globin Gene Cluster. *Annals of the New York Academy of Sciences*, 850(1), pp.54-63.

LIEU, Y. and REDDY, E., 2009. Conditional c-myb knockout in adult hematopoietic stem cells leads to loss of self-renewal due to impaired proliferation and accelerated differentiation. *Proceedings of the National Academy of Sciences*, 106(51), pp.21689-21694.

LIU, S., BHATTACHARYA, S., HAN, A., SURAGANI, R.N., ZHAO, W., FRY, R.C., and CHEN, J.-J., 2008. Haem-regulated eIF2α kinase is necessary for adaptive gene expression in erythroid precursors under the stress of iron deficiency. *British Journal of Haematology*, 143(1), pp.129–137.

LONG, F.H. and VEENSTRA, T.D., 2013. Chapter 19 - Multivariate Analysis for Metabolomics and Proteomics Data,. In: H.J. Issaq, ed., *Proteomic and Metabolomic Approaches to Biomarker Discovery*. [online] Academic Press, pp.299–311. Available: https://doi.org/10.1016/B978-0-12-394446-7.00019-4 [Jul 9, 2023].

LONGEVILLE, S. and STINGACIU, L., 2017. Hemoglobin diffusion and the dynamics of oxygen capture by red blood cells. *Scientific Reports*, 7(1).

LOREY, F., CHAROENKWAN, P., WITKOWSKA, H., LAFFERTY, J., PATTERSON, M., ENG, B., WAYE, J., FINKLESTEIN, J. and CHUI, D., 2001. Hb H hydrops foetalis syndrome: a case report and review of literature. *British Journal of Haematology*, 115(1), pp.72-78.

LOVE, P.E., WARZECHA, C., and LI, L., 2014. LDB1 complexes: The new master regulators of erythroid gene transcription. *Trends in Genetics*, 30(1), pp.1–9.

LUO, H., MU, W.-C., KARKI, R., CHIANG, H.-H., MOHRIN, M., SHIN, J.J., OHKUBO, R., ITO, K., KANNEGANTI, T.-D. and CHEN, D., 2019. Mitochondrial stress-initiated aberrant activation of the NLRP3 inflammasome regulates the functional deterioration of hematopoietic stem cell aging. *Cell Reports*, 26(4), pp.945-954.e4.

MAES, E., COOLS, N., WILLEMS, H., and BAGGERMAN, G., 2020. FACS-based proteomics enables profiling of proteins in rare cell populations. *International Journal of Molecular Sciences*, 21(18), p.6557.

MAGOR, G.W., TALLACK, M.R., GILLINDER, K.R., BELL, C.C., MCCALLUM, N., WILLIAMS, B., and PERKINS, A.C., 2015. KLF1-null neonates display hydrops fetalis and a deranged erythroid transcriptome. *Blood*, 125(15), pp.2405–2417.

MANNING, L., RUSSELL, J., PADOVAN, J., CHAIT, B., POPOWICZ, A., MANNING, R. and MANNING, J., 2007. Human embryonic, fetal, and adult hemoglobins have different subunit interface strengths. Correlation with lifespan in the red cell. *Protein Science*, 16(8), pp.1641-1658.

MANWANI, D., and FRENETTE, P.S., 2013. Vaso-occlusion in sickle cell disease: Pathophysiology and novel targeted therapies. *Blood*, 122(24), pp.3892–3898.

MARENGO-ROWE, A., 2007. The Thalassemias and Related Disorders. *Baylor University Medical Center Proceedings*, 20(1), pp.27-31.

MARENGO-ROWE, A., 2006. Structure-Function Relations of Human Hemoglobins. *Baylor University Medical Center Proceedings*, 19(3), pp.239-245.

MARTYN, G.E., WIENERT, B., YANG, L., SHAH, M., NORTON, L.J., BURDACH, J., KURITA, R., NAKAMURA, Y., PEARSON, R.C., FUNNELL, A.P., QUINLAN, K.G., and CROSSLEY, M., 2018. Natural regulatory mutations elevate the fetal globin gene via disruption of BCL11A or ZBTB7A binding. *Nature Genetics*, 50(4), pp.498–503.

MASUDA, T., WANG, X., MAEDA, M., CANVER, M.C., SHER, F., FUNNELL, A.P., FISHER, C., SUCIU, M., MARTYN, G.E., NORTON, L.J., ZHU, C., KURITA, R., NAKAMURA, Y., XU, J., HIGGS, D.R., CROSSLEY, M., BAUER, D.E., ORKIN, S.H., KHARCHENKO, P.V., and MAEDA, T., 2016. Transcription factors LRF and BCL11A independently repress expression of fetal hemoglobin. *Science*, 351(6270), pp.285–289.

MCGANN, P.T., NERO, A.C., and WARE, R.E., 2013. Current management of sickle cell anemia. *Cold Spring Harbor Perspectives in Medicine*, 3(8).

MEHDI, S. and AL DAHMASH, B., 2011. A comparative study of hematological parameters of α and β thalassemias in a high prevalence zone: Saudi Arabia. *Indian Journal of Human Genetics*, 17(3), pp.207-211.

MEIENBERG, J., BRUGGMANN, R., OEXLE, K., and MATYAS, G., 2016. Clinical sequencing: Is WGS the better wes? *Human Genetics*, 135(3), pp.359–362.

MENZEL, S., JIANG, J., SILVER, N., GALLAGHER, J., CUNNINGHAM, J., SURDULESCU, G., LATHROP, M., FARRALL, M., SPECTOR, T.D., and THEIN, S.L., 2007. The HBS1L-MYB intergenic region on chromosome 6q23.3 influences erythrocyte, platelet, and monocyte counts in humans. *Blood*, 110(10), pp.3624–3626.

METTANANDA, S. and HIGGS, D., 2018. Molecular Basis and Genetic Modifiers of Thalassemia. *Hematology/Oncology Clinics of North America*, 32(2), pp.177-191.

MISHRA, A. and TIWARI, A., 2013. Iron Overload in Beta Thalassaemia Major and Intermedia Patients. *Maedica*, 8(4), pp.328-332.

MODELL, B. and DARLISON, M., 2008. Global epidemiology of haemoglobin disorders and derived service indicators. *Bulletin of the World Health Organization*, 86(6), pp.480-487.

MOETTER, J., KARTAL, M., MEERPOHL, J., FISCHER, A., SIMON, T., URBANIAK, S., HIRABAYASHI, S., KAPP, F., NOELLKE, P., KRATZ, C., NIEMEYER, C.M. and WLODARSKI, M.W., 2011. Analysis of ribosomal protein genes associated with Diamond Blackfan Anemia (DBA) in German DBA patients and their relatives. *Blood*, 118(21), p.729.

MORISHIMA, T., BERNHARD, R., ZAHABI, A., DANNENMANN, B., NASRI, M., SAMAREH, B., KANZ, L., WELTE, K. and SKOKOWA, J., 2015. SIRT2 plays essential role in early hematopoiesis through deacetylation of LMO2 protein. *Blood*, 126(23), pp.3574–3574.

MURAYAMA, C., KIMURA, Y., and SETOU, M., 2009. Imaging mass spectrometry: Principle and application. *Biophysical Reviews*, 1(3), pp.131–139.

MUSALLAM, K., TAHER, A., CAPPELLINI, M. and SANKARAN, V., 2013. Clinical experience with fetal hemoglobin induction therapy in patients with β-thalassemia. *Blood*, 121(12), pp.2199-2212.

National Heart Lung and Blood Institute, 2022-last update. Sickle cell disease [National Heart Lung and Blood Institute], [online]. Available: www.nhlbi.nih.gov/resources/sickle-cell-disease [Apr 1, 2023].

NG, P.C., and KIRKNESS, E.F., 2010. Whole genome sequencing. *Methods in Molecular Biology*, 628, pp.215–226.

OHLS, R., 2017. *Fetal and neonatal physiology- Chapter 116 Developmental Erythropoiesis*. 5th ed. Elsevier, pp.1112-1134.e4.

OLD, J., 2013. Chapter 71 - Hemoglobinopathies and Thalassemias. In: D. Rimoin, R. Pyeritz and B. Korf, ed., *Emery and Rimoin's Principles and Practice of Medical Genetics*, 6th ed. Academic Press, pp.1-44.

OLIVIERI, N. and BRITTENHAM, G., 2013. Management of the Thalassemias. *Cold Spring Harbor Perspectives in Medicine*, 3(6), pp.a011767-a011767.

OMORI, A., TANABE, O., ENGEL, J.D., FUKAMIZU, A., and TANIMOTO, K., 2005. Adult stage γ-globin silencing is mediated by a promoter direct repeat element. *Molecular and Cellular Biology*, 25(9), pp.3443–3451.

O'NEILL, D.W., SCHOETZ, S.S., LOPEZ, R.A., CASTLE, M., RABINOWITZ, L., SHOR, E., KRAWCHUK, D., GOLL, M.G., RENZ, M., SEELIG, H.-P., HAN, S., SEONG, R.H., PARK, S.D., AGALIOTI, T., MUNSHI, N., THANOS, D., ERDJUMENT-BROMAGE, H., TEMPST, P. and BANK, A., 2000. An Ikaros-Containing Chromatin-Remodeling Complex in Adult-Type Erythroid Cells. *Molecular and Cellular Biology*, 20(20), pp.7572–7582. doi:10.1128/mcb.20.20.7572-7582.2000.

OTTOLENGHI, S., COMI, P., GIGLIONI, B., TOLSTOSHEV, P., LANYON, W., MITCHELL, G., WILLIAMSON, R., RUSSO, G., MUSUMECI, S., SCHILIRO, G., TSISTRAKIS, G., CHARACHE, S., WOOD, W., CLEGG, J. and WEATHERALL, D., 1976. δβ-Thalassemia is due to a gene deletion. *Cell*, 9(1), pp.71-80.

OTTOLENGHI, S., MANTOVANI, R., NICOLIS, S., RONCHI, A. and GIGLIONI, B., 1989. DNA Sequences Regulating Human Globin Gene Transcription in Nondeletional Hereditary Persistence of Fetal Hemoglobin. *Hemoglobin*, 13(6), pp.523-541.

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

PANJA, A. and BASU, A., 2015. Pharmacogenomics of the Drugs used for the Treatment of Thalassemia. *Journal of Cytology & Histology*, 6(5), pp.1-3.

PERKINS, A., XU, X., HIGGS, D.R., PATRINOS, G.P., ARNAUD, L., BIEKER, J.J. and PHILIPSEN, S., 2016. Krüppeling erythropoiesis: an unexpected broad spectrum of human red blood cell disorders due to KLF1 variants. *Blood*, 127(15), pp.1856–1862. doi:10.1182/blood-2016-01-694331.

PERRY, C. and SOREQ, H., 2002. Transcriptional regulation of erythropoiesis. *European Journal of Biochemistry*, 269(15), pp.3607-3618.

PERSEU, L., SATTA, S., MOI, P., DEMARTIS, F.R., MANUNZA, L., SOLLAINO, M.C., BARELLA, S., CAO, A. and GALANELLO, R., 2011. KLF1 gene mutations cause borderline HbA2. *Blood*, 118(16), pp.4454–4458. doi:10.1182/blood-2011-04-345736.

PESCHLE, C., MAVILIO, F., CARÈ, A., MIGLIACCIO, G., MIGLIACCIO, A., SALVO, G., SAMOGGIA, P., PETTI, S., GUERRIERO, R., MARINUCCI, M., LAZZARO, D., RUSSO, G. and MASTROBERARDINO, G., 1985. Haemoglobin switching in human embryos: asynchrony of ζ → α and ε → γ-globin switches in primitive and definitive erythropoietic lineage. *Nature*, 313(5999), pp.235-238.

PETOUSI, N., COPLEY, R.R., LAPPIN, T.R., HAGGAN, S.E., BENTO, C.M., CARIO, H., PERCY, M.J., RATCLIFFE, P.J., ROBBINS, P.A., MCMULLIN, M.F., DONNELLY, P., BELL, J., BENTLEY, D., MCVEAN, G., RATCLIFFE, P., TAYLOR, J., WILKIE, A., DONELLY, P., BROXHOLME, J., BUCK, D., CAZIER, J.-B., CORNALL, R., GREGORY, L., KNIGHT, J., LUNTER, G., MCVEAN, G., TOMLINSON, I., WILKIE, A., BUCK, D., ALLAN, C., ATTAR, M., GREEN, A., GREGORY, L., HUMPHRAY, S., KINGSBURY, Z., LAMBLE, S., LONIE, L., PAGNAMENTA, A., PIAZZA, P., POLANCO, G., TREBES, A., MCVEAN, G., DONNELLY, P., CAZIER, J.-B., BROXHOLME, J., COPLEY, R., FIDDY, S., GROCOCK, R., HATTON, E., HOLMES, C., HUGHES, L., HUMBURG, P., KANAPIN, A., LISE, S., LUNTER, G., MARTIN, H., MURRAY, L., MCCARTHY, D., RIMMER, A., SAHGAL, N., WRIGHT, B. and YAU, C., 2014. Erythrocytosis associated with a novel missense mutation in the BPGM gene. *Haematologica*, 99(10), pp.e201–e204.

PEVNY, L., SIMON, M., ROBERTSON, E., KLEIN, W., TSAI, S., D'AGATI, V., ORKIN, S. and COSTANTINI, F., 1991. Erythroid differentiation in chimaeric mice blocked by a targeted mutation in the gene for transcription factor GATA-1. *Nature*, 349(6306), pp.257-260.

PRALL, Y.G., GAMBHIR, K.K. and AMPY, F.R., 1998. Acetylcholinesterase: An enzymatic marker of human red blood cell aging. *Life Sciences*, 63(3), pp.177–184.

PRITCHARD, J.K. and PRZEWORSKI, M., 2001. Linkage disequilibrium in humans: Models and Data. *The American Journal of Human Genetics*, 69(1), pp.1–14.

QIN, K., LAN, X., HUANG, P., SAARI, M.S., KHANDROS, E., KELLER, C.A., GIARDINE, B.M., ABDULMALIK, O., SHI, J., HARDISON, R.C., and BLOBEL, G.A., 2023. Molecular basis of Polycomb group protein-mediated fetal hemoglobin repression. *Blood*, 141(22), pp.2756–2770.

RACHMILEWITZ, E.A., and GIARDINA, P.J., 2011. How I treat thalassemia. *Blood*, 118(13), pp.3479–3488.

RENELLA, R. and WOOD, W.G., 2009. The Congenital Dyserythropoietic Anemias. *Hematology/Oncology Clinics of North America*, 23(2), pp.283–306. doi:10.1016/j.hoc.2009.01.010.

RIVERS, A., MOLOKIE, R. and LAVELLE, D., 2019. A new target for fetal hemoglobin reactivation. *Haematologica*, 104(12), pp.2325-2327.

ROBB, L., LYONS, I., LI, R., HARTLEY, L., KÖNTGEN, F., HARVEY, R.P., METCALF, D., and BEGLEY, C.G., 1995. Absence of yolk sac hematopoiesis from mice with a targeted disruption of the SCL gene. *Proceedings of the National Academy of Sciences*, 92(15), pp.7075–7079.

RODRIGUEZ, P., BONTE, E., KRIJGSVELD, J., KOLODZIEJ, K.E., GUYOT, B., HECK, A.J., VYAS, P., DE BOER, E., GROSVELD, F., and STROUBOULIS, J., 2005. GATA-1 forms distinct activating and repressive complexes in erythroid cells. *The EMBO Journal*, 24(13), pp.2354–2366.

SÁ, A.C., SADEE, W. and JOHNSON, J.A., 2017. Whole transcriptome profiling: An RNA-seq primer and Implications for Pharmacogenomics Research. *Clinical and Translational Science*, 11(2), pp.153–161.

SANKARAN, V., MENNE, T.F., XU, J., AKIE, T.E., LETTRE, G., VAN HANDEL, B., MIKKOLA, H.K., HIRSCHHORN, J.N., CANTOR, A.B., and ORKIN, S.H., 2008. Human fetal hemoglobin expression is regulated by the developmental stage-specific repressor *bcl11a*. *Science*, 322(5909), pp.1839–1842.

SANKARAN, V. and ORKIN, S., 2012. The Switch from Fetal to Adult Hemoglobin. *Cold Spring Harbor Perspectives in Medicine*, 3(1), pp.a011643-a011643.

SANKARAN, V., XU, J. and ORKIN, S., 2010. Advances in the understanding of haemoglobin switching. *British Journal of Haematology*, 149(2), pp.181-194.

SANKARAN, V., XU, J., RAGOCZY, T., IPPOLITO, G.C., WALKLEY, C.R., MAIKA, S.D., FUJIWARA, Y., ITO, M., GROUDINE, M., BENDER, M.A., TUCKER, P.W., and ORKIN, S.H., 2009. Developmental and species-divergent globin switching are driven by BCL11A. *Nature*, 460(7259), pp.1093–1097.

SATTA, S., PERSEU, L., MOI, P., ASUNIS, I., CABRIOLU, A., MACCIONI, L., DEMARTIS, F.R., MANUNZA, L., CAO, A. and GALANELLO, R., 2011. Compound heterozygosity for KLF1 mutations associated with remarkable increase of fetal hemoglobin and red cell protoporphyrin. *Haematologica*, 96(5), pp.767–770. doi:10.3324/haematol.2010.037333.

SCHECHTER, A., 2008. Hemoglobin research and the origins of molecular medicine. *Blood*, 112(10), pp.3927-3938.

SCHNEIDER, R.K., ADEMÀ, V., HECKL, D., JÄRÅS, M., MALLO, M., LORD, A.M., CHU, L.P., MCCONKEY, M.E., KRAMANN, R., MULLALLY, A., BEJAR, R., SOLÉ, F. and EBERT, B.L., 2014. Role of casein kinase 1A1 in the biology and targeted therapy of Del(5Q) MDS. *Cancer Cell*, 26(4), pp.509–520.

SCHOENFELDER, S., SEXTON, T., CHAKALOVA, L., COPE, N.F., HORTON, A., ANDREWS, S., KURUKUTI, S., MITCHELL, J.A., UMLAUF, D., DIMITROVA, D.S., ESKIW, C.H., LUO, Y., WEI, C.-L., RUAN, Y., BIEKER, J.J. and FRASER, P., 2009. Preferential associations between co-regulated genes reveal a transcriptional interactome in erythroid cells. *Nature Genetics*, 42(1), pp.53–61. doi:10.1038/ng.496.

SEBASTIANI, P., NOLAN, V.G., BALDWIN, C.T., ABAD-GRAU, M.M., WANG, L., ADEWOYE, A.H., MCMAHON, L.C., FARRER, L.A., TAYLOR, J.G., KATO, G.J., GLADWIN, M.T., and STEINBERG, M.H., 2007. A network model to predict the risk of death in sickle cell disease. *Blood*, 110(7), pp.2727–2735.

SHARMA, D., SINGHAL, S., WOIKE, P., RAI, S., YADAV, M. and GAUR, R., 2020. Hereditary persistence of fetal hemoglobin. *Asian Journal of Transfusion Science*, 14(2), pp.185-186.

SHAUKAT, I., PAUDEL, A., YASSIN, S., HÖTI, N. and MUSTAFA, S., 2018. Blessing in disguise; a case of Hereditary Persistence of Fetal Hemoglobin. *Journal of Community Hospital Internal Medicine Perspectives*, 8(6), pp.380-381.

SHINMYO, Y., TANAKA, S., TSUNODA, S., HOSOMICHI, K., TAJIMA, A. and KAWASAKI, H., 2016. CRISPR/Cas9-mediated gene knockout in the mouse brain using in utero electroporation. *Scientific Reports*, 6(1).

SHIVDASANI, R.A., MAYER, E.L., and ORKIN, S.H., 1995. Absence of blood formation in mice lacking the T-cell leukaemia oncoprotein tal-1/SCL. *Nature*, 373(6513), pp.432–434.

SIATECKA, M., SAHR, K.E., ANDERSEN, S.G., MEZEI, M., BIEKER, J.J. and PETERS, L.L., 2010. Severe anemia in the Nan mutant mouse caused by sequence-selective disruption of erythroid Krüppel-like factor. *Proceedings of the National Academy of Sciences*, 107(34), pp.15151–15156. doi:10.1073/pnas.1004996107.

SIGMON, J. and LARCOM, L., 1996. The effect of ethidium bromide on mobility of DNA fragments in agarose gel electrophoresis. *Electrophoresis*, 17(10), pp.1524-1527.

SINGLETON, B.K., BURTON, N.M., GREEN, C., BRADY, R.L. and ANSTEE, D.J., 2008. Mutations in EKLF/KLF1 form the molecular basis of the rare blood group in(lu) phenotype. *Blood*, 112(5), pp.2081–2088.

SONG, W., HUANG, P., WANG, J., SHEN, Y., ZHANG, J., LU, Z., LI, D. and LIU, D., 2021. Red Blood Cell Classification Based on Attention Residual Feature Pyramid Network. *Frontiers in Medicine*, 8.

SONI, S., 2020. Gene therapies for transfusion dependent β-thalassemia: Current status and critical criteria for success. *American Journal of Hematology*, 95(9), pp.1099-1112.

STAMATOYANNOPOULOS, G. and GROSVELD, F., 2001. Haemoglobin switching. In: *The Molecular Basis of Blood Diseases*. Philadelphia: Saunders, pp.135–165.

STAMATOYANNOPOULOS, G., NAVAS, P.A., LI, Q., STEINBERG, M.H., FORGET, B.G., HIGGS, D.R., and WEATHERALL, D.J., 2009. Molecular and Cellular Basis of Hemoglobin Switching. In: *Disorders of Hemoglobin: Genetics, Pathophysiology, and Clinical Management*, 2nd ed. Cambridge: Cambridge University Press, pp.86–100.

STORZ, J., OPAZO, J. and HOFFMANN, F., 2013. Gene duplication, genome duplication, and the functional diversification of vertebrate globins. *Molecular Phylogenetics and Evolution*, 66(2), pp.469-478.

STRAUB, C., GRANGER, A.J., SAULNIER, J.L. and SABATINI, B.L., 2014. CRISPR/Cas9-mediated gene knock-down in post-mitotic neurons. *PLoS ONE*, 9(8).

SUI, X., KRANTZ, S.B., YOU, M., and ZHAO, Z., 1998. Synergistic activation of Map Kinase (ERK1/2) by erythropoietin and stem cell factor is essential for expanded erythropoiesis. *Blood*, 92(4), pp.1142–1149.

TANABE, O., KATSUOKA, F., CAMPBELL, A.D., SONG, W., YAMAMOTO, M., TANIMOTO, K., and DOUGLAS ENGEL, J., 2002. An embryonic/fetal beta-type globin gene repressor contains a nuclear receptor TR2/TR4 heterodimer. *The EMBO Journal*, 21(13), pp.3434–3442.

TANABE, O., MCPHEE, D., KOBAYASHI, S., SHEN, Y., BRANDT, W., JIANG, X., CAMPBELL, A.D., CHEN, Y.-T., CHANG, C. SHANG, YAMAMOTO, M., TANIMOTO, K., and ENGEL, J.D., 2007. Embryonic and fetal β-globin gene repression by the orphan nuclear receptors, TR2 and TR4. *The EMBO Journal*, 26(9), pp.2295–2306.

TANIMOTO, K., LIU, Q., GROSVELD, F., BUNGERT, J., and ENGEL, J.D., 2000. Context-dependent EKLF responsiveness defines the developmental specificity of the human ε-globin gene in erythroid cells of YAC transgenic mice. *Genes & Development*, 14(21), pp.2778–2794.

TANIMURA, N., MILLER, E., IGARASHI, K., YANG, D., BURSTYN, J.N., DEWEY, C.N., and BRESNICK, E.H., 2015. Mechanism governing heme synthesis reveals a gata factor/heme circuit that controls differentiation. *EMBO reports*, 17(2), pp.249–265.

THEIN, S., 2012. The emerging role of fetal hemoglobin induction in non-transfusion-dependent thalassemia. *Blood Reviews*, 26, pp.S35-S39.

THEIN, S. and CRAIG, I., 1998. Genetics of Hb F/F Cell Variance in Adults and Heterocellular Hereditary Persistence of Fetal Hemoglobin. *Hemoglobin*, 22(5-6), pp.401-414.

THEIN, S. and MENZEL, S., 2009. Discovering the genetics underlying foetal haemoglobin production in adults. *British Journal of Haematology*, 145(4), pp.455-467.

THEIN, S.L., MENZEL, S., PENG, X., BEST, S., JIANG, J., CLOSE, J., SILVER, N., GEROVASILLI, A., PING, C., YAMAGUCHI, M., WAHLBERG, K., ULUG, P., SPECTOR, T.D., GARNER, C., MATSUDA, F., FARRALL, M., and LATHROP, M., 2007. Intergenic variants of HBS1L-myb are responsible for a major quantitative trait locus on chromosome 6q23 influencing fetal hemoglobin levels in adults. *Proceedings of the National Academy of Sciences*, 104(27), pp.11346–11351.

TORRENTS, E., 2014. Ribonucleotide reductases: essential enzymes for bacterial life. *Frontiers in Cellular and Infection Microbiology*, 4(52).

TYANOVA, S., TEMU, T., SINITCYN, P., CARLSON, A., HEIN, M.Y., GEIGER, T., MANN, M., and COX, J., 2016. The Perseus Computational Platform for comprehensive analysis of (prote)omics data. *Nature Methods*, 13(9), pp.731–740.

VAN DER AUWERA, G.A. and O'CONNOR, B.D., 2020. *Genomics in the cloud: Using Docker, GATK and WDL in Terra*. 1st ed. Beijing: O'Reilly Media.

VAN DIJK, T.B., GILLEMANS, N., STEIN, C., FANIS, P., DEMMERS, J., VAN DE CORPUT MARIËTTE, ESSERS, J., GROSVELD, F., BAUER, U.-M., and PHILIPSEN, S., 2010. Friend of PRMT1, a novel chromatin target of protein arginine methyltransferases. *Molecular and Cellular Biology*, 30(1), pp.260–272.

VERMA, S., BHARGAVA, M., MITTAL, S. and GUPTA, R., 2013. Homozygous delta-beta Thalassemia in a Child: a Rare Cause of Elevated Fetal Hemoglobin. *Iranian journal of pediatric hematology and oncology*, 3(1), pp.222-227.

VOGEL, S., KAMIMURA, S., ARORA, T., SMITH, M.L., ALMEIDA, L.E.F., COMBS, C.A., THEIN, S.L. and QUEZADO, Z.M.N., 2021. NLRP3 inflammasome and Bruton tyrosine kinase inhibition interferes with upregulated platelet aggregation and in vitro thrombus formation in sickle cell mice. *Biochemical and Biophysical Research Communications*, 555, pp.196–201.

WADMAN, I.A., OSADA, H., GRÜTZ, G.G., AGULNICK , A.D., WESTPHAL, H., FORSTER, A., and RABBITTS, T.H., 1997. 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. *The EMBO Journal*, 16(11), pp.3145–3157.

WANG, C., XU, C.-X., ALIPPE, Y., QU, C., XIAO, J., SCHIPANI, E., CIVITELLI, R., ABU-AMER, Y. and MBALAVIELE, G., 2017. Chronic inflammation triggered by the NLRP3 inflammasome in myeloid cells promotes growth plate dysplasia by mesenchymal cells. *Scientific Reports*, 7(1).

WARREN, A.J., COLLEDGE, W.H., CARLTON, M.B.L., EVANS, M.J., SMITH, A.J.H., and RABBITTS, T.H., 1994. The oncogenic cysteine-rich LIM domain protein RBTN2 is essential for erythroid development. *Cell*, 78(1), pp.45–57.

WEINBERG, R., JI, X., SUTTON, M., PERRINE, S., GALPERIN, Y., LI, Q., LIEBHABER, S., STAMATOYANNOPOULOS, G. and ATWEH, G., 2005. Butyrate increases the efficiency of translation of γ-globin mRNA. *Blood*, 105(4), pp.1807-1809.

WEISS, M.J., KELLER, G., and ORKIN, S.H., 1994. Novel insights into erythroid development revealed through in vitro differentiation of GATA-1 embryonic stem cells. *Genes & Development*, 8(10), pp.1184–1197.

WHEELER, J. and KREVANS, J., 1961. The homozygous state of persistent fetal hemoglobin and the interaction of persistent fetal hemoglobin with thalassemia. *Bull Johns Hopkins Hosp.*, (109), pp.217-233.

WIENERT, B., MARTYN, G.E., KURITA, R., NAKAMURA, Y., QUINLAN, K.G. and CROSSLEY, M., 2017. KLF1 drives the expression of fetal hemoglobin in British hpfh. *Blood*, 130(6), pp.803–807.

WILBER, A., NIENHUIS, A. and PERSONS, D., 2011. Transcriptional regulation of fetal to adult hemoglobin switching: new therapeutic opportunities. *Blood*, 117(15), pp.3945-3953.

WRZESZCZYNSKI, K.O., FELICE, V., SHAH, M., RAHMAN, S., EMDE, A.-K., JOBANPUTRA, V., O. FRANK, M., and DARNELL, R.B., 2018. Whole genome sequencing-based discovery of structural variants in glioblastoma. *Methods in Molecular Biology*, 1741, pp.1–29.

XU, J., SANKARAN, V.G., NI, M., MENNE, T.F., PURAM, R.V., KIM, W., and ORKIN, S.H., 2010. Transcriptional silencing of γ-globin by BCL11A involves long-range interactions and cooperation with SOX6. *Genes & Development*, 24(8), pp.783–798.

XUE, L., GALDASS, M., GNANAPRAGASAM, M.N., MANWANI, D., and BIEKER, J.J., 2014. Extrinsic and intrinsic control by EKLF (KLF1) within a specialized erythroid niche. *Development*, 141(11), pp.2245–2254.

YANG, K., WU, Y., ZHOU, Y., LONG, B., LU, Q., ZHOU, T., WANG, L., GENG, Z. and YIN, X., 2020. Thalidomide for Patients with β-Thalassemia: A Multicenter Experience. *Mediterranean Journal of Hematology and Infectious Diseases*, 12(1), p.e2020021.

YANG, L., HU, M., LU, Y., HAN, S. and WANG, J., 2021. Inflammasomes and the maintenance of hematopoietic homeostasis: New Perspectives and Opportunities. *Molecules*, 26(2), p.309.

YOGALAKSHMI, E., HEMAMALINI, N. and CHITRA, S., 2020. Screening for Haemoglobinopathies- Better prevent than regret - A Tip of the iceberg experience in a Tertiary Care Hospital in Chennai. *International Journal of Research in Pharmaceutical Sciences*, 11(SPL2), pp.251-259.

ZHENG, Y., XU, J., LIANG, S., LIN, D. and BANERJEE, S., 2018. Whole exome sequencing identified a novel heterozygous mutation in HMBS gene in a Chinese patient with acute intermittent porphyria with rare type of mild anemia. *Frontiers in Genetics*, 9.

ZHOU, D., LIU, K., SUN, C.-W., PAWLIK, K.M., and TOWNES, T.M., 2010. KLF1 regulates BCL11A expression and γ- to β-globin gene switching. *Nature Genetics*, 42(9), pp.742–744.

ZHOU, W., CLOUSTON, D.R., WANG, X., CERRUTI, L., CUNNINGHAM, J.M., and JANE, S.M., 2000. Induction of human fetal globin gene expression by a novel erythroid factor, NF-E4. *Molecular and Cellular Biology*, 20(20), pp.7662–7672.

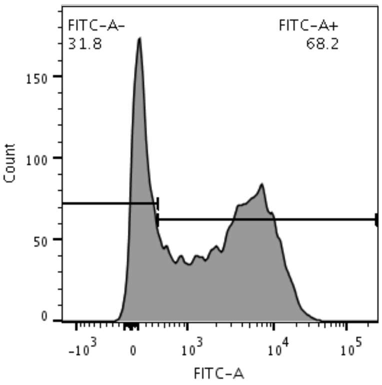
ZHOU, W., ZHAO, Q., SUTTON, R., CUMMING, H., WANG, X., CERRUTI, L., HALL, M., WU, R., CUNNINGHAM, J.M., and JANE, S.M., 2004. The role of p22 NF-E4 in human globin gene switching. *Journal of Biological Chemistry*, 279(25), pp.26227–26232.

ZIVOT, A., LIPTON, J., NARLA, A. and BLANC, L., 2018. Erythropoiesis: insights into pathophysiology and treatments in 2017. *Molecular Medicine*, 24(1).

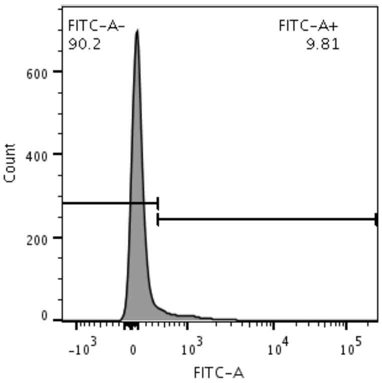
Appendix A

Flow cytometry antigen expression profiles of each subject

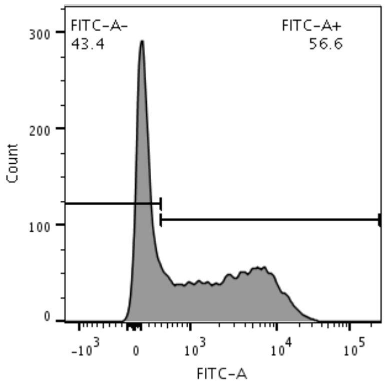
BCAM Antigen Expression Profiles



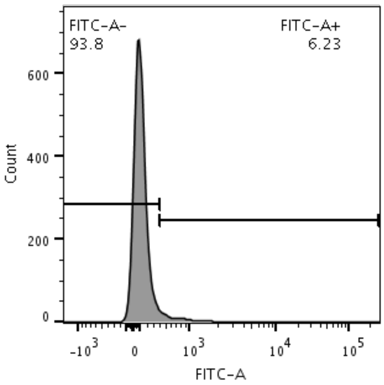
SampleID	Antibody	Type
F1.1	BCAM	Control



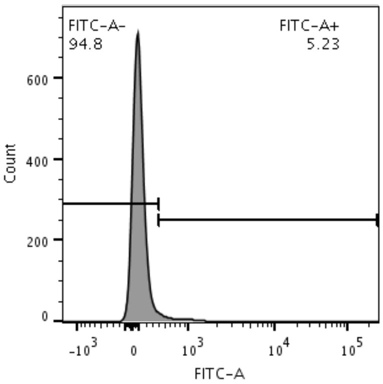
SampleID	Antibody	Type
F1.2	BCAM	HPFH



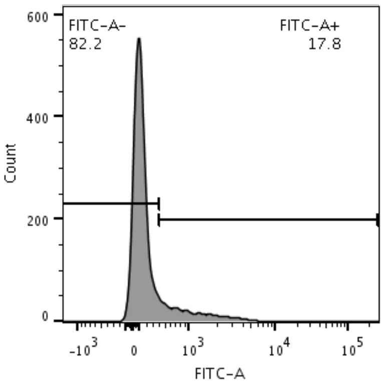
SampleID	Antibody	Type
F1.3	BCAM	Control



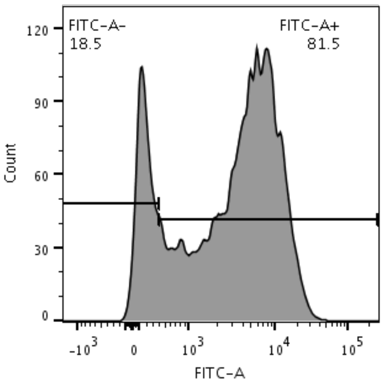
SampleID	Antibody	Type
F1.4	BCAM	HPFH



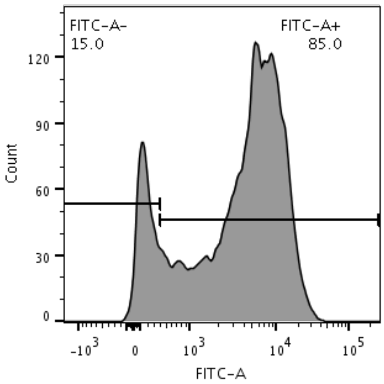
SampleID	Antibody	Type
F1.5	BCAM	HPFH



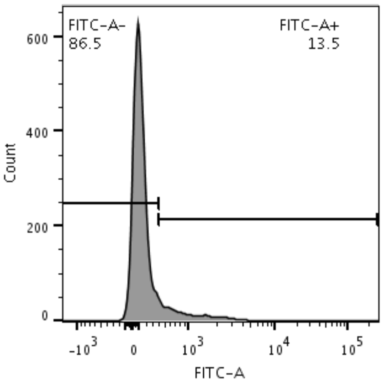
SampleID	Antibody	Type
F1.6	BCAM	HPFH



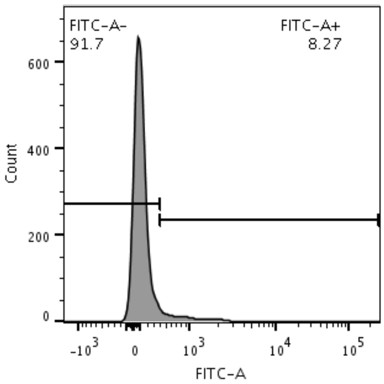
SampleID	Antibody	Type
F1.7	BCAM	Control



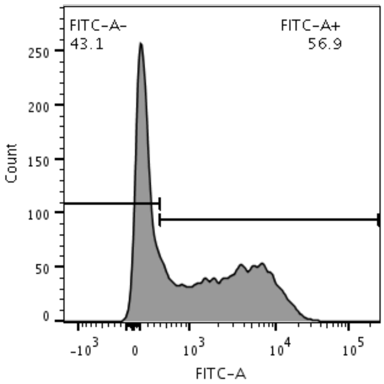
SampleID	Antibody	Type
F1.8	BCAM	Control



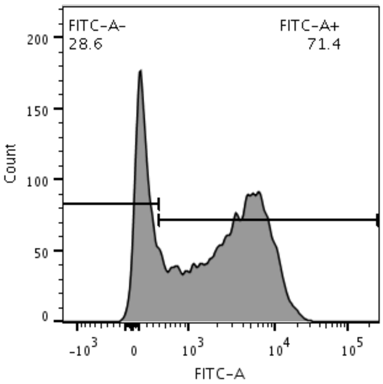
SampleID	Antibody	Type
F1.9	BCAM	HPFH



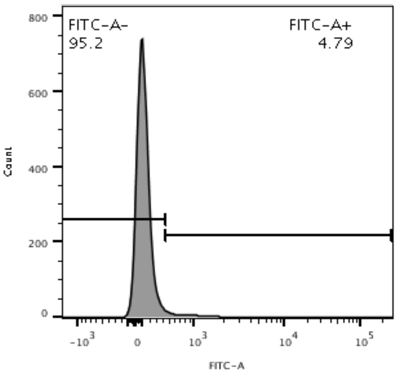
SampleID	Antibody	Type
F1.10	BCAM	HPFH



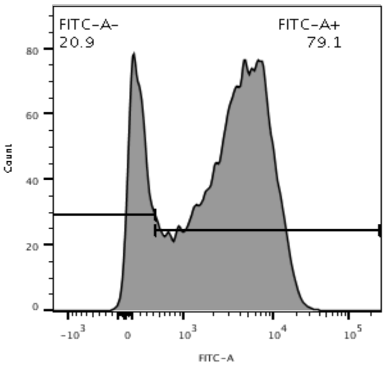
SampleID	Antibody	Type
F1.11	BCAM	Control



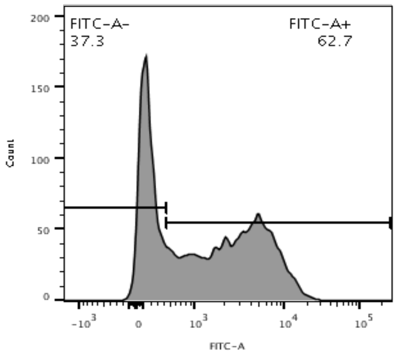
SampleID	Antibody	Type
F1.12	BCAM	Control



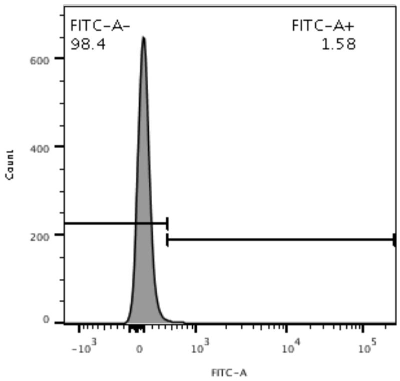
SampleID	Antibody	Type
F2.1	BCAM	HPFH



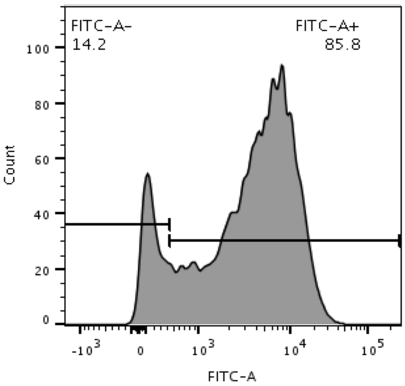
SampleID	Antibody	Type
F2.2	BCAM	Control



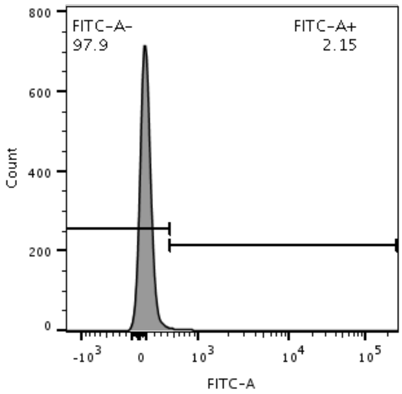
SampleID	Antibody	Type
F2.3	BCAM	Control



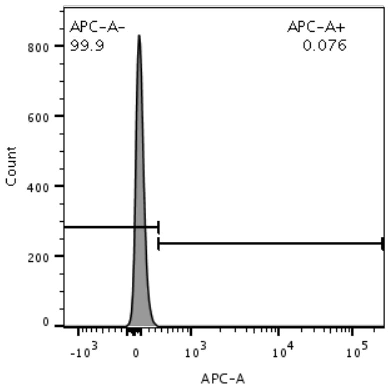
SampleID	Antibody	Type
F2.4	BCAM	HPFH



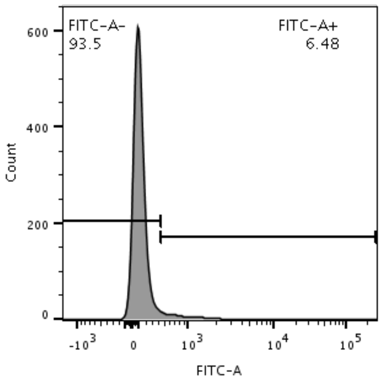
SampleID	Antibody	Type
F4.1	BCAM	Control



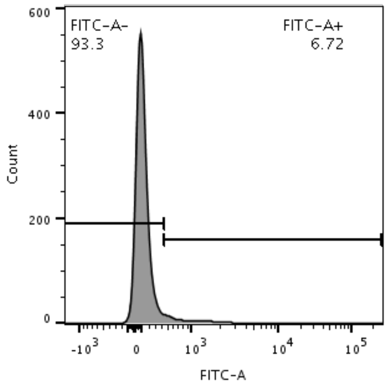
SampleID	Antibody	Type
F4.2	BCAM	HPFH



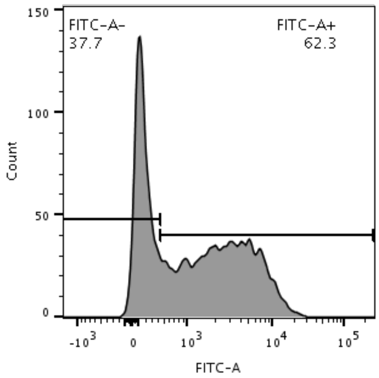
SampleID	Antibody	Type
F4.3	BCAM	Control



SampleID	Antibody	Type
F4.4	BCAM	HPFH

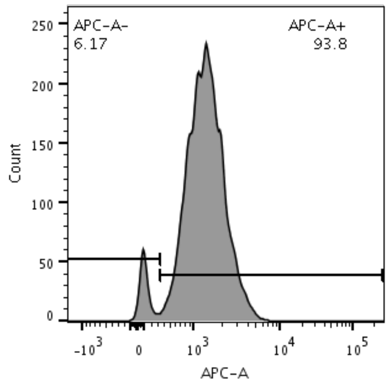


SampleID	Antibody	Type
F4.5	BCAM	HPFH

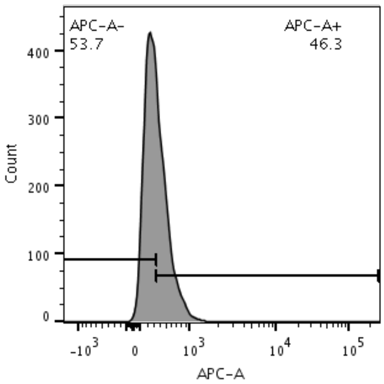


SampleID	Antibody	Type
F4.6	BCAM	Control

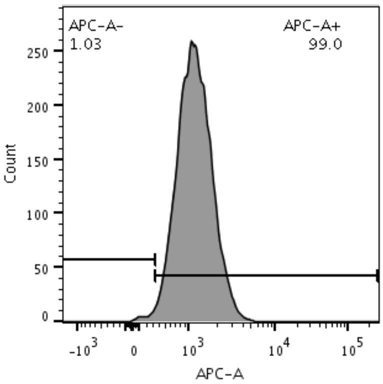
CD44 Antigen Expression Profiles



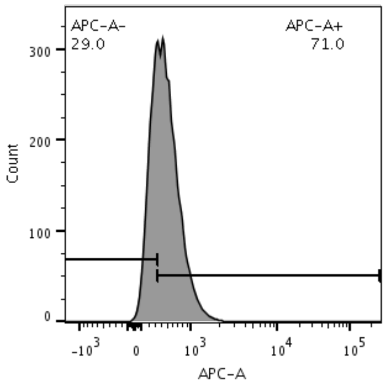
SampleID	Antibody	Type
F1.1	CD44	Control



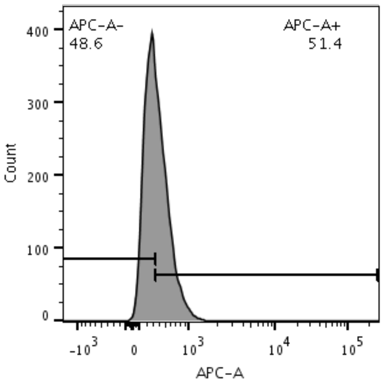
SampleID	Antibody	Type
F1.2	CD44	HPFH



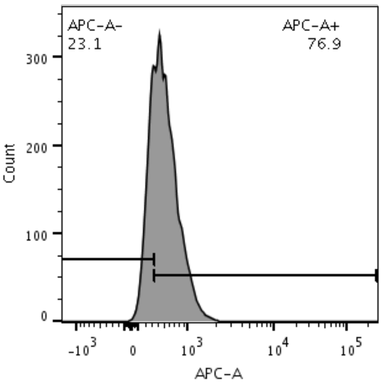
SampleID	Antibody	Type
F1.3	CD44	Control



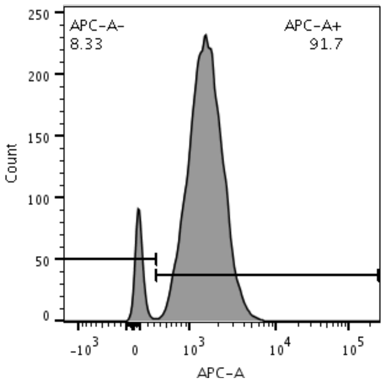
SampleID	Antibody	Type
F1.4	CD44	HPFH



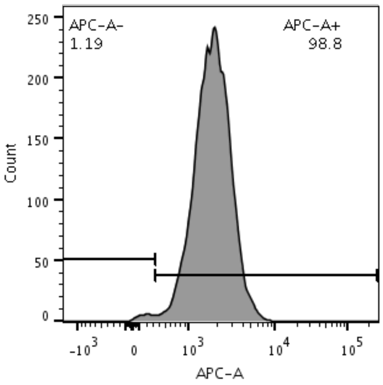
SampleID	Antibody	Type
F1.5	CD44	HPFH



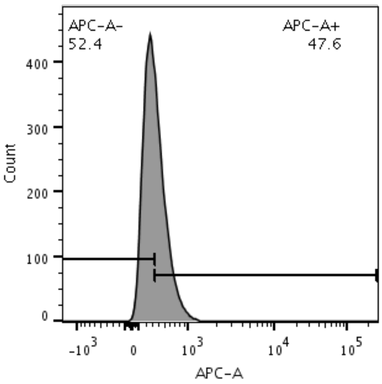
SampleID	Antibody	Type
F1.6	CD44	HPFH



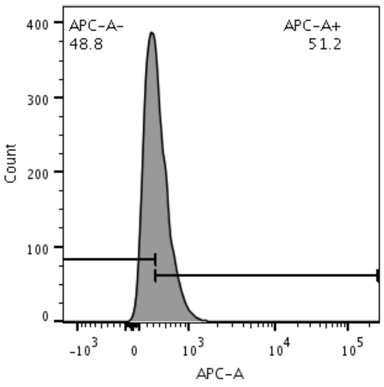
SampleID	Antibody	Type
F1.7	CD44	Control



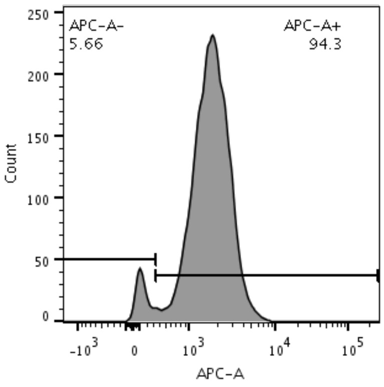
SampleID	Antibody	Type
F1.8	CD44	Control



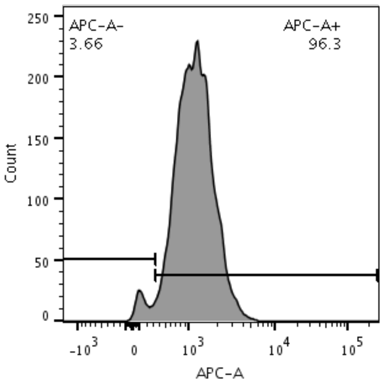
SampleID	Antibody	Type
F1.9	CD44	HPFH



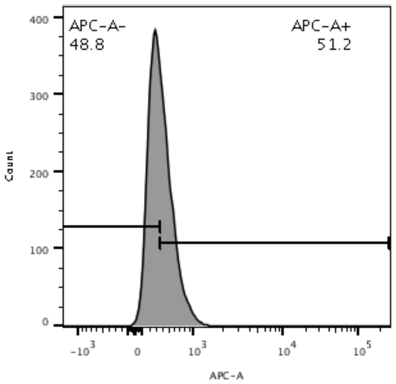
SampleID	Antibody	Type
F1.10	CD44	HPFH



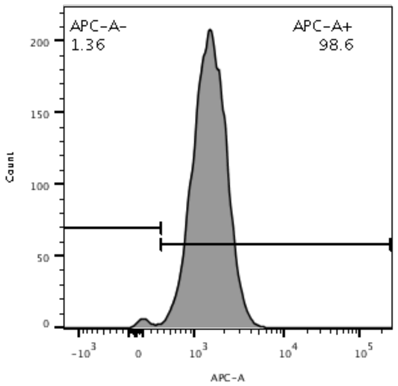
SampleID	Antibody	Type
F1.11	CD44	Control



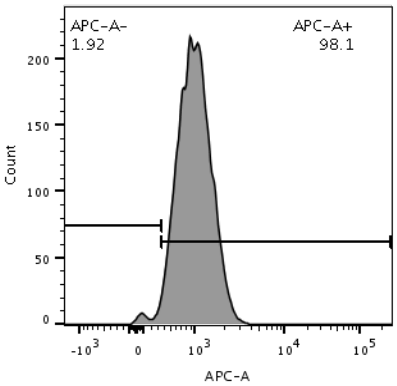
SampleID	Antibody	Type
F1.12	CD44	Control



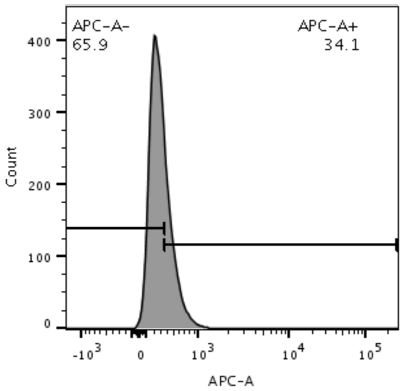
SampleID	Antibody	Type
F2.1	CD44	HPFH



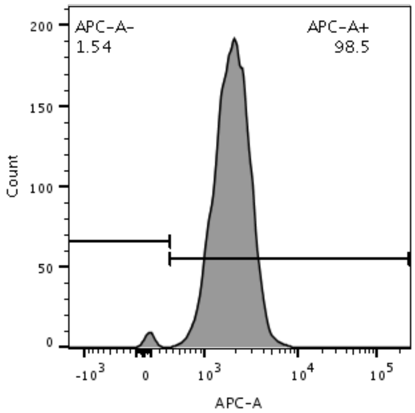
SampleID	Antibody	Type
F2.2	CD44	Control



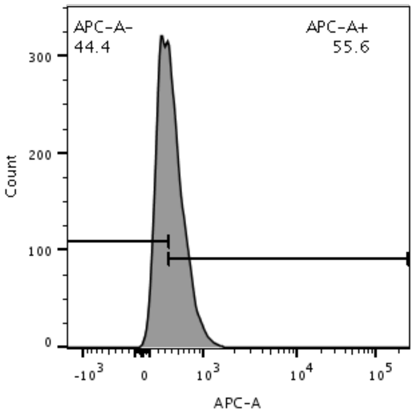
SampleID	Antibody	Type
F2.3	CD44	Control



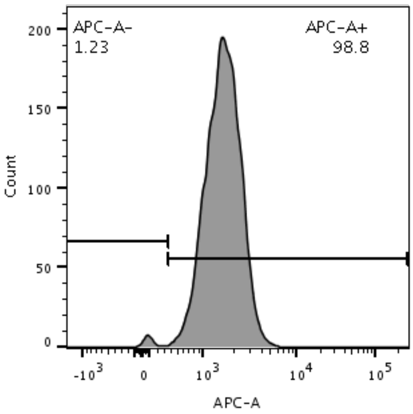
SampleID	Antibody	Type
F2.4	CD44	HPFH



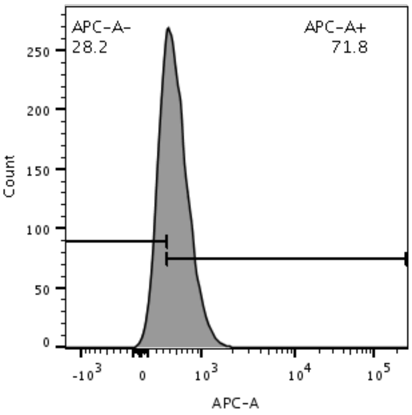
SampleID	Antibody	Type
F4.1	CD44	Control



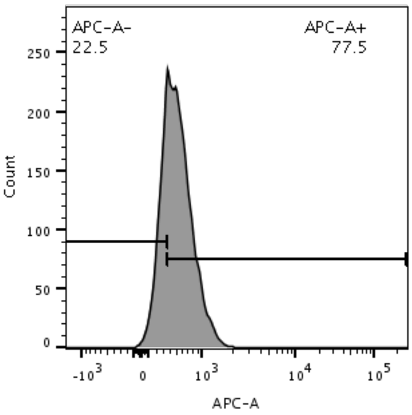
SampleID	Antibody	Type
F4.2	CD44	HPFH



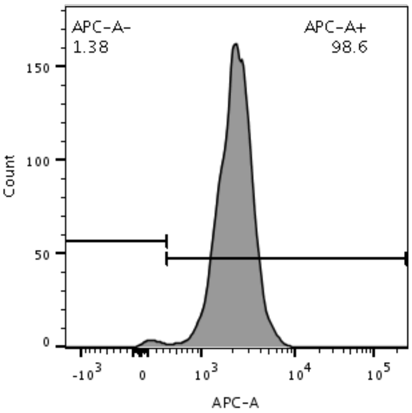
SampleID	Antibody	Type
F4.3	CD44	Control



SampleID	Antibody	Type
F4.4	CD44	HPFH

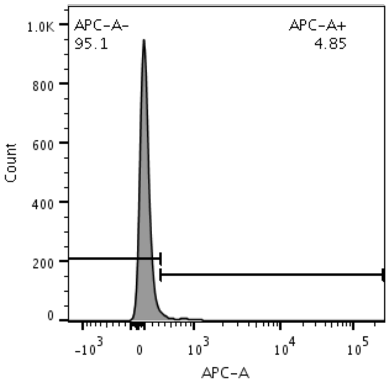


SampleID	Antibody	Type
F4.5	CD44	HPFH

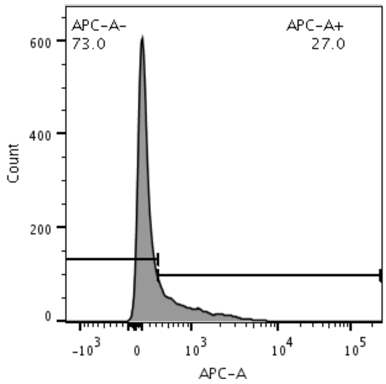


SampleID	Antibody	Type
F4.6	CD44	Control

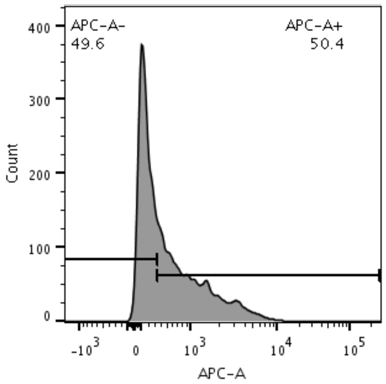
P1 Antigen Expression Profiles



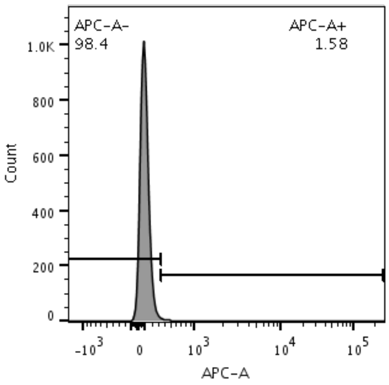
SampleID	Antibody	Type
F1.1	P1	Control



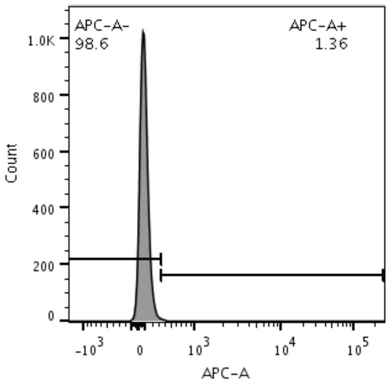
SampleID	Antibody	Type
F1.2	P1	HPFH



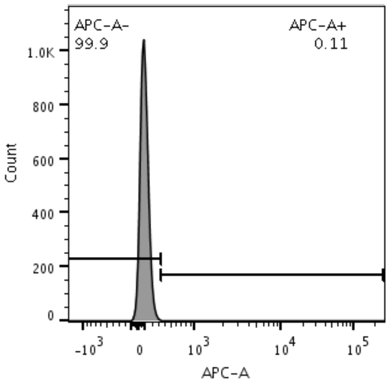
SampleID	Antibody	Type
F1.3	P1	Control



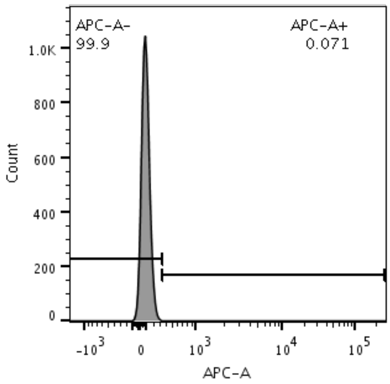
SampleID	Antibody	Type
F1.4	P1	HPFH



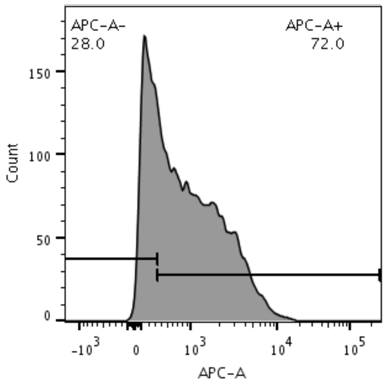
SampleID	Antibody	Type
F1.5	P1	HPFH



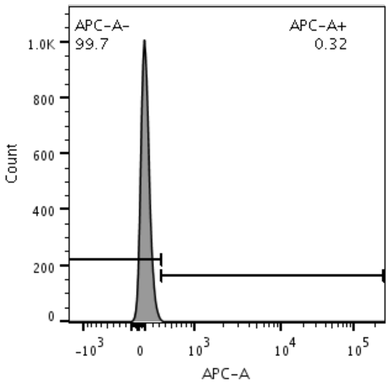
SampleID	Antibody	Type
F1.6	P1	HPFH



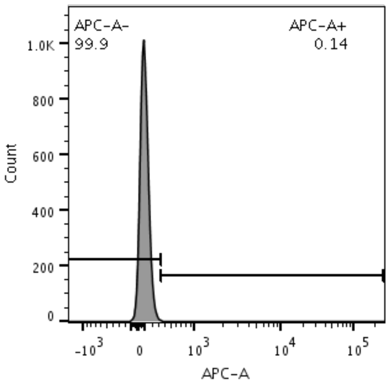
SampleID	Antibody	Type
F1.7	P1	Control



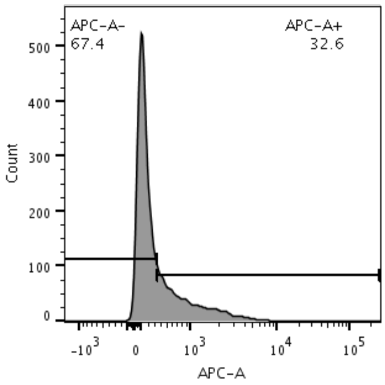
SampleID	Antibody	Type
F1.8	P1	Control



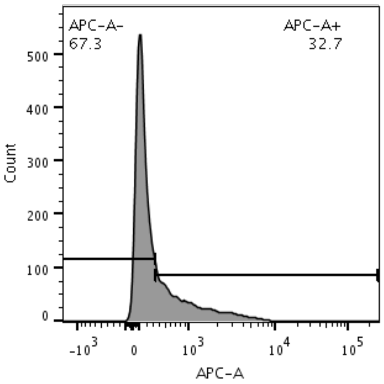
SampleID	Antibody	Type
F1.9	P1	HPFH



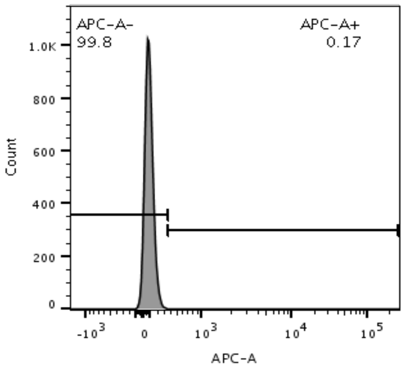
SampleID	Antibody	Type
F1.10	P1	HPFH



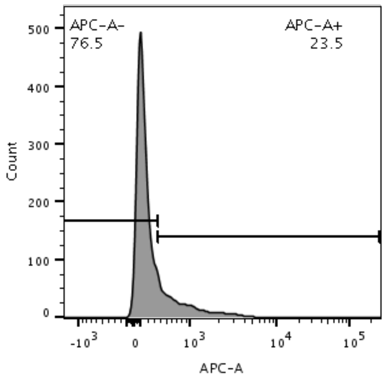
SampleID	Antibody	Type
F1.11	P1	Control



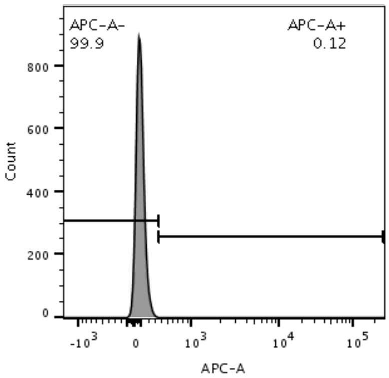
SampleID	Antibody	Type
F1.12	P1	Control



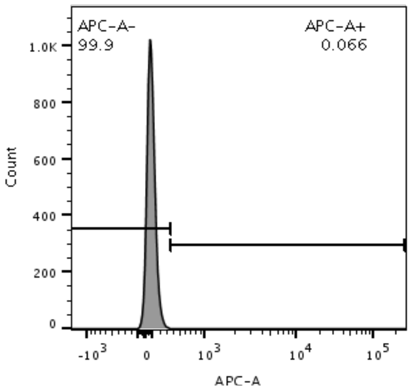
SampleID	Antibody	Type
F2.1	P1	HPFH



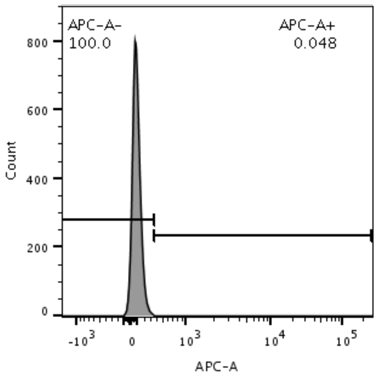
SampleID	Antibody	Type
F2.2	P1	Control



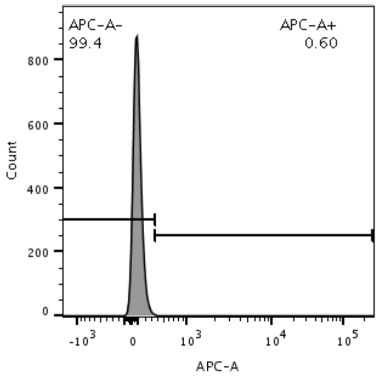
SampleID	Antibody	Type
F2.3	P1	Control



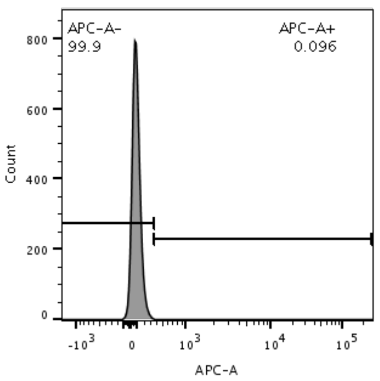
SampleID	Antibody	Type
F2.4	P1	HPFH



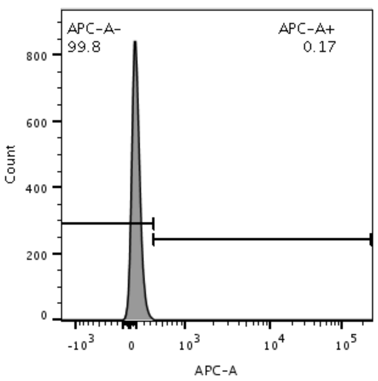
SampleID	Antibody	Type
F4.1	P1	Control



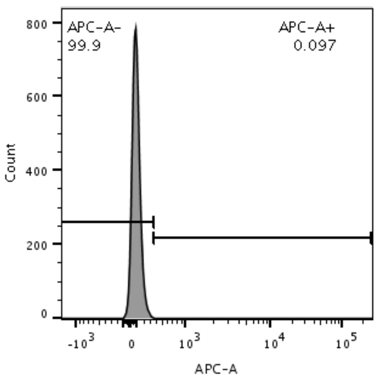
SampleID	Antibody	Type
F4.2	P1	HPFH



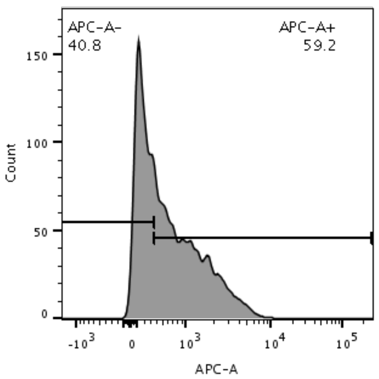
SampleID	Antibody	Type
F4.3	P1	Control



SampleID	Antibody	Type
F4.4	P1	HPFH



SampleID	Antibody	Type
F4.5	P1	HPFH



SampleID	Antibody	Type
F4.6	P1	Control

Appendix B

Antigen expression data obtained by flow cytometry

Sample ID	Type	BCAM Antigen			CD44 Antigen			P1 Antigen		
		FITC gMFI	FITC +ve (%)	FITC -ve (%)	APC gMFI	APC +ve (%)	APC -ve (%)	APC gMFI	APC +ve (%)	APC -ve (%)
F1.1	Control	1568	68.2	31.8	1340	93.8	6.17	122	4.85	95.1
F1.2	HPFH	159	9.81	90.2	358	46.3	53.7	325	27	73
F1.3	Control	1103	56.6	43.4	1249	99	1.03	575	50.4	49.6
F1.4	HPFH	136	6.23	93.8	498	71	29	93.7	1.58	98.4
F1.5	HPFH	119	5.23	94.8	382	51.4	48.6	81.8	1.36	98.6
F1.6	HPFH	258	17.8	82.2	541	76.9	23.1	80.8	0.11	99.9
F1.7	Control	2573	81.5	18.5	1403	91.7	8.33	80.7	0.071	99.9
F1.8	Control	3129	85	15	1899	98.8	1.19	922	72	28
F1.9	HPFH	207	13.5	86.5	361	47.6	52.4	92.8	0.32	99.7
F1.10	HPFH	153	8.27	91.7	380	51.2	48.8	83.2	0.14	99.9
F1.11	Control	1115	56.9	43.1	1688	94.3	5.66	379	32.6	67.4
F1.12	Control	1639	71.4	28.6	1210	96.3	3.66	393	32.7	67.3
F2.1	HPFH	143	4.79	95.2	394	51.2	48.8	76.1	0.17	99.8
F2.2	Control	2180	79.1	20.9	1518	98.6	1.36	293	23.5	76.5
F2.3	Control	1279	62.7	37.3	1058	98.1	1.92	74.6	0.12	99.9
F2.4	HPFH	96.5	1.58	98.4	318	34.1	65.9	72.2	0.066	99.9
F4.1	Control	3026	85.8	14.2	1949	98.5	1.54	78.3	0.048	100
F4.2	HPFH	94.8	2.15	97.9	417	55.6	44.4	83	0.6	99.4
F4.3	Control	1706	73.4	26.6	1668	98.8	1.23	77.5	0.096	99.9
F4.4	HPFH	151	6.48	93.5	508	71.8	28.2	82.1	0.17	99.8
F4.5	HPFH	150	6.72	93.3	551	77.5	22.5	80.1	0.097	99.9
F4.6	Control	1219	62.3	37.7	2251	98.6	1.38	668	59.2	40.8

Appendix C

Correlation analysis of mass spectrometry data identified 53 proteins significantly correlated with levels of HbF

Gene name	Correlation Coefficient (r)	P-value
ARHGAP1	-0.8778028	0.00038143
BPGM	-0.8289014	0.00161099
HBA1	-0.8122872	0.00238308
CSE1L	-0.8047528	0.00281194
YWHAB	-0.7893036	0.00386762
KPRP	-0.7863355	0.00409996
TANGO2	-0.7749678	0.00508619
HMBS	-0.7691242	0.00565622
GLIPR2	-0.7573948	0.00694035
HS1BP3	-0.7395794	0.00928235
PRPS1	-0.7266656	0.01130672
KCNN4	-0.6988062	0.01673579
ACHE	-0.6930791	0.0180488
STOM	-0.6910809	0.01852377
CSNK1A1	-0.6844349	0.02016788
EHBP1L1	-0.6814595	0.02093661
ACTN4	-0.6738612	0.02299413
IL18	-0.6570548	0.02804809
HBB	-0.6408634	0.03361282
BROX	-0.6362419	0.03533307
LMAN2	-0.6342095	0.03610868
IDH2	-0.6328831	0.03662121
PITPNA	-0.6307733	0.03744685
PAFAH1B3	-0.630587	0.0375204
SEMA7A	-0.6168545	0.04322115
PPIA	-0.6144208	0.04429042
DAZAP1	-0.6136284	0.04464247
DDAH2	-0.6118103	0.04545755

Gene name	Correlation Coefficient (r)	P-value
DNAJC9	-0.6111972	0.04573469
CA3	-0.6024362	0.04982326
RANBP6	0.60479113	0.04870054
OXSR1	0.60971741	0.04640842
REXO2	0.61458436	0.044218
UROS	0.62349927	0.04039286
CHMP2A	0.62640333	0.0391981
DPCD	0.63261631	0.03672491
UBE3A	0.64566894	0.03188716
RAPGEF2	0.65679897	0.0281306
HSPH1	0.66613895	0.0252283
UBR4	0.6725464	0.0233642
VCP	0.70322366	0.01577093
PPP2R5D	0.70555279	0.01527862
MEMO1	0.71240847	0.01389369
HBM	0.72016556	0.01243847
PSMA2	0.75263148	0.00751837
HBZ	0.75762705	0.00691302
ABRAXAS2	0.76143376	0.00647631
SARS1	0.76532848	0.0060508
HBG1	0.79846987	0.00321135
HBG2	0.81247599	0.00237301
EIF4E	0.81627502	0.00217689
TNS1	0.84743357	0.00098926
SIRT2	0.88607703	0.00028169

Online Supplementary Data

https://github.com/borgjoseph/HPFH_Project

File A: The 205 unique variants identified in the full cohort following a dominant inheritance pattern after stringent filtering steps. These variants were found to be present in all 11 affected individuals while being absent in all 9 unaffected individuals. The file shows the chromosome position, the reference and alternate alleles, the associated consequence, the gene name and any annotations of the identified variants.

File B: The 272 unique variants identified in the cohort consisting of the four individuals having the highest HbF levels (F1.2, F1.5, F1.9 and F1.10) from Fam F1, following a dominant inheritance pattern after stringent filtering steps. These variants were present only in all the four individuals with the highest HbF levels and were absent in all other individuals. The file shows the chromosome position, the reference and alternate alleles, the associated consequence, the gene name and any annotations of the identified variants.

File C: The 200 unique variants identified in the cohort consisting of the four individuals having the highest HbF levels (F1.2, F1.5, F1.9 and F1.10) from Fam F1, following a recessive inheritance pattern after stringent filtering steps. These variants were present only in all the four individuals with the highest HbF levels and were absent in all other individuals. The file shows the chromosome position, the reference and alternate alleles, the associated consequence, the gene name and any annotations of the identified variants.

File D: The 534 statistically significant proteins identified by mass spectrometry, having a p-value less than 0.05.