

Genomics of sickle cell disease and fetal hemoglobin in African populations

By: Kevin Esoh Kum

ESHKEV001



Submitted to the University of Cape Town
In fulfilment of the requirements of the PhD degree in Human Genetics
University of Cape Town, Faculty of Health Sciences, Department of Pathology,
Division of Human Genetics.

Supervisor: Professor Ambroise Wonkam

Co-supervisor: Dr Jack Morrice

Date of submission: 3 October 2023

The copyright of this thesis vests in the author. No quotation from it or information derived from it is to be published without full acknowledgement of the source. The thesis is to be used for private study or non-commercial research purposes only.

Published by the University of Cape Town (UCT) in terms of the non-exclusive license granted to UCT by the author.

The copyright of this thesis vests in the author. No quotation from it or information derived from it is to be published without full acknowledgement of the source. The thesis is to be used for private study or noncommercial research purposes only.

Published by the University of Cape Town (UCT) in terms of the non-exclusive license granted to UCT by the author.

Declaration

I, Kevin Esoh Kum, hereby declare that this dissertation/thesis is based on my original work (except where acknowledgments indicate otherwise) and that neither the whole work nor any part of it has been or is being submitted for another degree in this or any other university.

I give permission to the university to reproduce, for the purpose of research, either the whole or portion of the content, in any manner whatsoever.

Signature:

Date: 3 October 2023

Student Name: Kevin Esoh Kum

Student Number: ESHKEV001

Dedication

In loving memory of my mother Monica Beteh Kang, my great-grand mother Lydia Mameh, and my sister Rachel Bih Esoh. Your hearts of gold and spirits of a lioness paved the way for me.

Acknowledgments

I would like to give thanks with a grateful heart to the holy one, God, for giving me this purpose in life, and for making me see it.

Sincere gratitude to my supervisor, Professor Ambroise Wonkam, for giving me the opportunity and space to express my potential. His supervision, guidance, encouragement and mentorship were simply genius.

Many thanks to my co-supervisor, Dr Jack Morrice, for the guidance, advice and encouragement.

I am grateful to all the study participants who contributed their priceless samples to the success of this project, and immense thanks to Dr Valentina Josiane Ngo Bitoungui from Cameroon who, together with Professor Ambroise Wonkam, recruited the study participants.

Much thanks to the funders of my project, the National Heart, Lung, and Blood Institute of the National Institutes of Health (NHLBI-NIH, USA), through grant [U24HL135600-02] to Professor Ambroise Wonkam for the Sickle in Africa Data Coordinating Center (SADaCC).

Thanks to Dr Victoria Nembaware, SADaCC project manager, and the late Dr Gaston Manzandu, for their immense support.

Sincere thanks to Dr Nchangwi Syntia Munung for her mentorship, advice, and encouragement throughout the journey.

Thanks to all my SADaCC team members for their support, with special thanks to Mario Jonas for assisting with data access and allocation of computing resources and the day-to-day troubleshooting of technical issues.

Thanks to the Centre for High Performance Computing (CHPC), South Africa, for providing computational resources to this research project. Computations were also performed using facilities provided by the University of Cape Town's ICTS High Performance Computing team to whom I am very grateful.

Thanks to Dr Emile Chimusa for the expert advice on genomics and bioinformatics.

Lastly, I would like to thank my father, Mr Jacob Kang Esoh, for his constant words of wisdom and encouragement, my siblings, Mr Richard Ketchem Esoh, Reverend Dr Felix Kang Esoh, Mr Samson Achmi Esoh, Ms Maxine Mbong Esoh, and Reverend Randuph Akue Esoh, and my niece, Tracy Mameh Esoh, for their fervent support.

List of publications included in this thesis

I confirm that I have been granted permission by the University of Cape Town's Doctoral Degrees Board to include the following publications in my PhD thesis, and where co-authorships are involved, my co-authors have agreed that I may include these publications:

1. **Esoh, K.**, Wonkam-Tingang, E., & Wonkam, A. (2021). Sickle cell disease in sub-Saharan Africa: transferable strategies for prevention and care. *The Lancet. Haematology*, 8(10), e744–e755. [https://doi.org/10.1016/S2352-3026\(21\)00191-5](https://doi.org/10.1016/S2352-3026(21)00191-5)
2. **Esoh, K.**, & Wonkam, A. (2021). Evolutionary history of sickle-cell mutation: implications for global genetic medicine. *Human molecular genetics*, 30(R1), R119–R128. <https://doi.org/10.1093/hmg/ddab004>

Signature:

Signed by candidate

Date: 3 October 2023

Student Name: Kevin Esoh Kum

Student Number: ESHKEV001

Publications from collaborative efforts not included in the thesis

Publications I made significant contributions to during the time of my registration as a PhD student.

1. **Esoh KK**, Apinjoh TO, Amambua-Ngwa A, Nyanjom SG, Chimusa ER, Amenga-Etego L, Wonkam A, Achidi EA. Genome-wide association study identifies novel candidate malaria resistance genes in Cameroon. *Hum Mol Genet.* **2023** Jun 5;32(12):1946-1958. <https://doi.org/10.1093/hmg/ddad026>. PMID: 36752565; PMCID: PMC10244224.
2. Wonkam A, Munung NS, Dandara C, **Esoh KK**, Hanchard NA, Landoure G. Five Priorities of African Genomics Research: The Next Frontier. *Annu Rev Genomics Hum Genet.* 2022 Aug 31;23:499-521. <https://doi.org/10.1146/annurev-genom-111521-102452>. Epub 2022 May 16. PMID: 35576571.
3. Nembaware V, Mazandu GK, Hotchkiss J, Safari Serufuri JM, Kent J, Kengne AP, Anie K, Munung NS, Bukini D, Bitoungui VJN, Munube D, Chirwa U, Chunda-Liyoka C, Jonathan A, Flor-Park MV, **Esoh KK**, Jonas M, Mnika K, Oosterwyk C, Masamu U, Morrice J, Uwineza A, Nguweneza A, Banda K, Nyanor I, Adjei DN, Siebu NE, Nkanyemka M, Kuona P, Tayo BO, Campbell A, Oron AP, Nnodu OE, Painstil V, Makani J, Mulder N, Wonkam A. The Sickle Cell Disease Ontology: Enabling Collaborative Research and Co-Designing of New Planetary Health Applications. *OMICS.* **2020** Oct;24(10):559-567. <https://doi.org/10.1089/omi.2020.0153>. PMID: 33021900; PMCID: PMC7549008.
4. Nembaware V, Bennett D, Chimusa ER, Chikowore T, Daodu R, Bitoungui VN, Williams SM, Fatumo S, Healy S, Seoighe C, Wonkam A; African Society of Human Genetics. The African Society of Human Genetics successfully launches global data science workshops. *Trends Genet.* 2023 Sep 13:S0168-9525(23)00135-X. doi: <https://doi.org/10.1016/j.tig.2023.06.004>. Epub ahead of print. PMID: 37714735.

Preface

Over 300,000 babies are born with sickle cell disease (SCD) each year, about 79% of these births occurring in sub-Saharan Africa where the sickle cell mutation is known to originate. Although the variant is highly recessively lethal, the protection it confers against severe malaria in its heterozygous form has resulted in its persistence in sub-Saharan Africa. Sickle cell anemia (SCA) caused by inheritance of two copies of the sickle cell mutation is therefore prevalent in sub-Saharan Africa where malaria is endemic. Without intervention, 50–90% of affected children in many sub-Saharan African countries are estimated to die before their fifth birthday. The search for a definitive cure or effective disease-modifying therapy is therefore an imperative. Fetal hemoglobin (HbF) has long been recognized to ameliorate SCD severity and patients harboring genetic variations that lead to the persistence of high levels of HbF in their blood tend to live longer with fewer complications. The persistence of high HbF levels in adult blood is a highly heritable quantitative trait (heritability of 89%), and the strongest modifier of SCD severity known to date. It is therefore the focus of enormous research as a promising therapeutic approach. Genetic variants that influence HbF level have been identified in three major loci, and associated with about 50% of fetal hemoglobin heritability in African American SCA patients and non-anemic European individuals. Two of the three loci that have been observed in SCA patients in Africa contribute less than 30% in HbF heritability, indicating the existence of other loci that remain to be uncovered. A better understanding of the genetic architecture of HbF in African SCA patients will foster research into potentially novel SCD therapeutic targets.

This thesis project was my conception with guidance and supervision from Professor Ambroise Wonkam (supervisor and project principal investigator). The study involved solely genomic data analysis using primary data that was generated from samples collected for cross-sectional epidemiological and genetic studies between 2016 and 2018 in Cameroon by Prof Wonkam and Dr Valentina Josiane Ngo Bitoungui, a former PhD student in Prof Wonkam's group at the University of Cape Town. Through a collaboration with Tanzanian investigators, we also obtained and utilized genotype data from Tanzanian SCD patients generated between 2011 and 2014. The data contributed to the first genome-wide association study (GWAS) of HbF in Africa-based SCD patients in 2014 lead by Dr Siana Nkya. Summary statistics of HbF GWAS involving seven cohorts of African ancestry SCD patients based in the United States of America (USA) were also obtained through collaborators and utilized for meta-analysis. A proposal for the study developed with supervision from Prof Wonkam and Dr. Jack Morrice (co-supervisor, who gave simply remarkable writing tips) was approved by the Department of Pathology Departmental Research Committee and the University of Cape Town, Faculty of Health Sciences Human Research Ethics Committee under the reference [HREC REF: 606/2021].

A substantial amount of 'data' was generated from the project and was published or being prepared for publication during the time of my registration. This is therefore a Doctoral thesis that includes publications, not merely a compilation of publications. The thesis is presented in

the traditional way, and according to the general guidelines of the Doctoral Degrees Board of the University of Cape Town for the inclusion of publications in a doctoral thesis. The thesis is presented in five cohesive chapters circling around SCD management. Chapter 1 presents a general background including historical aspects of SCD, from discovery, burden, and evolving management options, to the use of genomics to find better alternatives for management and cure. Chapter 2 looks at the evolutionary history of the sickle cell mutation, focusing on its origin in Africa, spread around the world through migration with implication on global genetic medicine, genetic modifiers of SCD severity, and co-evolution or co-occurrence with other adaptively selected variants which affect or might potentially affect SCD severity and consequently management. Chapter 3 presents a comprehensive overview of SCD management options that have been successful in reducing morbidity and mortality in high-resource settings. It recommends transferable strategies for low-resource settings, and areas for further research. Chapter 4 presents an exploration of genetic loci that might influence HbF level in African sickle cell patients, and that could serve as targets for gene therapy. This was achieved through GWAS and meta-analyses. It first assessed the application of multiple genotype imputation strategies for GWAS, and then used dataset imputed using multiple panels to investigate HbF-associated loci. It also assessed the heritability of HbF in African sickle cell anemia patients, and went further to perform *in silico* and *in vitro* functional analyses for significant loci identified in the association and meta-analyses. Chapter 5 presents a general discussion of the study findings, and concludes with future directions (perspectives).

The inclusion of published work in this thesis served the following purposes: i) to demonstrate my aptitude to synthesize available information in the field, analyze complex bibliographic and large-scale genomic data, and generate new knowledge, all fundamental skill sets necessary to attaining high research quality; ii) to fulfill a requirement for continued financial support from the Sickle in Africa Data Coordinating Center (SADaCC) project; iii) to enable participation in academic conferences and workshops including: the Sickle in Africa Consortium meetings, the Human Heredity and Health in Africa (H3Africa) 18th and 19th meetings in 2021 and 2022 respectively, the African Society of Human Genetics (AfSHG) 13th meeting in 2021, the International Congress of Human Genetics in 2023, and the Great Lakes Bioinformatics Conference in 2023. I was also invited through the AfSHG to present a case study of my work on the impact of African genetic diversity on the genomics of SCD and malaria for a Computational Biology course that was taught to graduate students in Ireland; iv) to contribute to the University's body of high impact work and to fulfill the Faculty of Health Science's policies that promote the publication of students' research work for the dissemination of knowledge, and contributing towards maintaining the institution's status as the best in Africa and one of the best in the world, while also improve my research profile.

In submitting this thesis, I have met all requirements and approval of UCT's Doctoral Degrees Board, under Rules GP6.7 stating:

- a. The candidate's proposal to include publications in the current thesis was

approved by the UCT Faculty of Health Sciences Doctoral Degree Board.

- b. The thesis contains an adequate summary, introduction, a chapter on the aims and objectives, a comprehensive academic discussion of the results, forming the basis of the conclusions and perspectives drawn from this research.
- c. Each results chapter with publications included is preceded by a synopsis of how the publications directly tie to the aims and objectives of the project, as well as to the thesis.
- d. All included publications were written and published during the candidate's tenure as a PhD student since 2020.

The candidate,

Signed by candidate

Kevin Esoh Kum

Table of Contents

Declaration	iii
Dedication.....	iv
Acknowledgments.....	v
List of publications included in this thesis.....	vi
Publications from collaborative efforts not included in the thesis	vii
Preface	viii
Table of Contents	xi
List of Figures.....	xvi
List of Tables	xvii
List of Abbreviations	xviii
Abstract	xix
Chapter 1: General Introduction.....	1
1.1. Sick cell disease burden: a historical background.....	1
1.2. Strategies for prevention and patient care.....	2
1.3. Gene therapy for sickle cell disease.....	3
1.4. Fetal hemoglobin as a potent sickle cell disease curative agent through gene therapy	3
1.5. Strategies employed in the study of HbF-influencing loci	6
1.6. This study: a genome-wide association approach	7
1.7. Rationale	9
1.8. Aim	10
1.9. Objectives.....	10
Chapter 2: Evolutionary history of sickle cell mutation: implications for global genetic medicine	12
<i>Synopsis</i>	12
2.1. Abstract.....	12
2.2. Introduction	13
2.3. Method	14
2.3.1. Database	14
2.3.2. Search strategy.....	14
2.3.3. Exclusion criteria:	15
2.3.4. Inclusion criteria:.....	15
2.3.5. Data extraction.....	15
2.4. Results.....	16

2.4.1.	Common Variants in Hemoglobin Genes	16
2.4.2.	Global distribution of the HBB- β S variant.....	17
2.4.3.	Global distribution of the HBB-like genes cluster haplotypes	18
2.4.4.	The African origin of the <i>HBB</i> - β^S variant.....	20
2.4.5.	SCA and genetic modifiers	22
2.4.6.	Co-evolution of HBB- β S and Other Malaria-Resistant Variants: Epistatic and Non-epistatic Interactions	23
2.4.7.	Co-evolution of HBB- β S and Other Variants Selected under Specific Non-malaria Environmental Pressures	25
2.4.8.	No HbF protection against malaria	25
2.5.	Conclusion and Perspectives.....	26
2.6.	Limitation	26
2.7.	Funding	27
Chapter 3: Sickle cell disease in sub-Saharan Africa: transferable strategies for prevention and care ..		28
	<i>Synopsis</i>	28
3.1.	Abstract.....	28
3.2.	Introduction	29
	Search strategy and selection criteria.....	30
3.3.	Epidemiology.....	30
3.3.1.	Sickle cell disease and malaria endemicity	30
3.3.2.	Co-inheritance of sickle cell disease and other genetic determinants of outcomes in sub-Saharan Africa	32
3.4.	Prevalence of sickle cell disease in sub-Saharan Africa	35
3.5.	Diagnosis of sickle cell disease and neonate screening in sub-Saharan Africa.....	36
	PANEL 1	38
	Challenges and recommendations for sickle cell disease research in sub-Saharan Africa.....	38
	PANEL 2	39
	Proposed guidelines for the care of patients with sickle cell disease in sub-Saharan Africa	39
3.6.	Mortality and life expectancy of patients with sickle cell disease in sub-Saharan Africa	40
3.7.	Pathophysiology.....	40
3.8.	Important blood components measured in sickle cell disease.....	41
3.9.	Sub-phenotypes of sickle cell disease in sub-Saharan Africa.....	42
3.10.	Specific environmental determinants of sickle cell disease clinical outcome in sub-Saharan Africa	43

3.11.	Management of sickle cell disease in sub-Saharan Africa	44
3.12.	Psychosocial effects of sickle cell disease	46
3.13.	Conclusion and perspectives.....	46
3.14.	Contributors	47
3.15.	Acknowledgments.....	47
Chapter 4: Genome-wide association study of fetal hemoglobin in sickle cell disease patients of African Ancestry		48
	<i>Synopsis</i>	48
4.1.	Abstract.....	49
4.2.	Introduction	50
4.3.	Results	51
4.3.1.	The dataset	51
4.3.2.	Genotype data quality assessment and population structure.....	54
4.3.3.	Comparative performance across different imputation panels	57
4.3.4.	Association testing supports complementary use of multiple imputation panels	61
4.3.5.	Replication of the major HbF-influencing loci: <i>BCL11A</i> and <i>HBS1L-MYB</i>	65
4.3.6.	Replication in other known and recent genetically nominated loci	66
4.3.7.	Identification of novel HbF-associated loci.....	67
4.3.8.	Functional mapping of the <i>FLT1</i> signal	71
4.3.9.	Functional effects of <i>FLT1</i> fine-mapped variants.....	73
4.3.10.	Association of <i>FLT1</i> fine-mapped variants with other blood traits.....	74
4.3.11.	Haplotype structure reveals reason for absence of <i>FLT1</i> association in the Tanzanian cohort. 76	
4.3.12.	Haplotype association.....	79
4.3.13.	Gene-based, gene set, and heritability analyses further support the association results .80	
4.3.14.	Heritability of fetal hemoglobin.....	84
4.3.15.	GWAS of combined CAM and TZN genotyped samples mirrors meta-analysis results and confirms robustness of heritability analysis	86
4.3.16.	Involvement of <i>FLT1</i> in erythropoiesis is hypoxia-dependent and likely necessary for hypoxia-induced HbF induction	87
4.3.17.	Expression patterns of <i>FLT1</i> and other erythropoiesis and hypoxia related genes in HUDEP-2 and K562 cells under hypoxia.....	91
4.4.	Discussion.....	94
4.4.1.	Summary of key points	94

4.4.2.	Heterogeneity in imputation panels	94
4.4.3.	Heterogeneity in association signals.....	95
4.4.4.	Heritability supported by association results, pathway enrichment, and potentially, selection pressure (at least in Cameroonians) evidenced by high LD and haplotype conservation in the <i>FLT1</i> 40kb regulatory region.....	95
4.4.5.	Functional relevance of <i>FLT1</i> associations.....	96
4.4.6.	Methodological considerations/limitations/strengths	97
4.5.	Conclusions	99
4.6.	Methods.....	99
4.6.1.	Ethical approvals	99
4.6.2.	Patient participants.....	100
4.6.3.	Measurements of hematological indices	100
4.6.4.	Molecular methods.....	100
4.6.5.	Cohorts description.....	101
4.6.6.	Quality control (QC)	102
4.6.7.	Haplotype estimation (phasing) and genotype imputation.....	102
4.6.8.	Creation of custom imputation panel.....	103
4.6.9.	Association analysis	104
4.6.10.	Meta-analysis	104
4.6.11.	Statistical and functional fine mapping	105
4.6.12.	Gene set analysis and functional mapping	106
4.6.13.	<i>FLT1</i> haplotype structure and association analysis.....	106
4.6.14.	Assessment of transferability/replication of signals.....	107
4.6.15.	Estimation of HbF heritability	107
4.6.16.	HUDEP-2 and K562 cell culture, differentiation, and hypoxia treatment.....	107
Chapter 5: General Discussion and Conclusion		109
5.1.	Introduction	109
5.2.	Evolutionary history of the sickle cell mutation and opportunities for further research.....	109
5.3.	Co-evolution of sickle cell disease and other disease-related adaptive signatures	110
5.4.	SCD management: newborn screening as the main pivot.....	110
5.5.	Genomics of fetal hemoglobin.....	111
5.6.	Limitations.....	113
5.7.	Conclusion and perspectives.....	113
References		116

Appendix I: Distribution of βS haplotypes in various World populations	140
Appendix II: Birth and population prevalence of SCA in Africa	143
Appendix III: Sickle cell anemia pathophysiology	144
Appendix IV: Expression of hypoxia inducible factor (HIF) associated genes in HUDEP-2 and K562 cells after 24-hour and 10-Day hypoxia	145
Appendix V: Expression of genes connecting hypoxia and hemoglobin production pathways in HUDEP-2 and K562 cells after 24-hour and 10-Day hypoxia	147
Appendix VI: Expression of transcription factors involved in hemoglobin production in HUDEP-2 and K562 cells after 24-hour hypoxia	149
Appendix VII: Expression of transcription factors whose binding motifs are disrupted by <i>FLT1</i> regulatory region variants in HUDEP-2 and K562 cells after 10-Day hypoxia	150
Appendix VIII: Research Material and other web resources.....	151
Appendix IX: Associations for the merged, non-imputed, genotype dataset of Cameroon and Tanzania with FDR < 0.20	153
Appendix X: Associations for the meta-analysis of Cameroon and Tanzania non-imputed datasets with FDR < 0.20	155

List of Figures

Figure 1.1 Hemoglobin switching	5
Figure 2.1 HbS and HBB haplotypes Search Strategy	16
Figure 2.2 HBB-like genes cluster haplotypes.....	19
Figure 2.3 HBB- β^S origin in Africa	21
Figure 2.4 Co-evolution of the HBB- β^S variant and other genes variant in Africa.....	24
Figure 3.1 Frequencies of the sickle hemoglobin allele and other gene variants in Africa.....	32
Figure 3.2 Distribution of mutated HBB gene cluster haplotypes in regional variants	34
Figure 3.3 Sickle cell disease treatment centres and HSCT centres in Africa	36
Figure 4.1 Description of basic demographic and clinical characteristics	53
Figure 4.2 Genotype data quality control	55
Figure 4.3 Data quality assessment and population structure.....	56
Figure 4.4 Number of variants imputed per imputation panel	58
Figure 4.5 Comparative analysis of imputation panels.....	60
Figure 4.6 Data flow in the major analysis stages	61
Figure 4.7 Evidence of association.....	63
Figure 4.8 Significant and suggestive loci in the TOPMed panel and Tanzania reanalysis	69
Figure 4.9 Regional association and boxplots of Cameroon and Tanzania exclusive signals	70
Figure 4.10 Statistical and functional fine mapping of the FLT1 signal	72
Figure 4.11 Functional effects of FLT1 fine-mapped variants.....	75
Figure 4.12 Boxplots of all FLT1 fine-mapped variants.....	76
Figure 4.13 Genetic heterogeneity and haplotype substructure within the FLT1 –40kb regulatory region	78
Figure 4.14 Heterogeneity plot of sentinel variants of all the significant loci.....	79
Figure 4.15 Significant loci gene-based association tests.....	80
Figure 4.16 Gene-based and gene set enrichment analysis.	82
Figure 4.17 GTEx version 8 differentially expressed genes in 30 general tissue types.....	83
Figure 4.18 heritability of fetal hemoglobin	85
Figure 4.19 Association signals in merged genotype and meta-analysis non-imputed datasets.	87
Figure 4.20 Proposed model for FLT1 inactivation, reactivation, and involvement in erythropoiesis and HbF induction.....	89
Figure 4.21 Protein-protein interaction network of key erythropoietic and hypoxia response pathway proteins	90
Figure 4.22 Expression pattern of FLT1 and erythropoiesis transcription factors in HUDEP-2 and K562 cells under short-term hypoxia (3-24 hours)	93
Figure 4.23 Expression pattern of FLT1 and erythropoiesis transcription factors in HUDEP-2 and K562 cells under long-term hypoxia (4-10 days)	94

List of Tables

Table 4.1 Clinical and demographic features of the Cameroonian cohort	52
Table 4.2 Basic characteristics of Cameroonian and Tanzanian cohorts	52
Table 4.3 General statistics of the imputation panels utilized	58
Table 4.4 Number of variants analyzed per panel for each dataset	62
Table 4.5 Summary of significant signal in independent association tests.....	64
Table 4.6 Summary of significant signals in meta-analyses	65

List of Abbreviations

AGR	African genome resource
BMDAC	Bone marrow-derived angiogenic cell
CAAPA	Consortium on Asthma among African-ancestry Populations in the Americas
CRISPR	Clustered Regularly Interspaced Short Palindromic Repeats
DNA	Deoxyribonucleic acid
EHT	Endothelial to hematopoietic transition
GWAS	Genome-wide association study
H3Africa	Human Heredity and Health in Africa
Hb	Hemoglobin
HbA	Normal adult hemoglobin
HbF	Fetal hemoglobin
HbS	Sickle hemoglobin
HEC	Hemogenic endothelial cell
hESC	Human embryonic stem cell
HPFH	Hereditary persistence of fetal hemoglobin
HRE	Hypoxia response element
HSC	Hematopoietic stem cell
HSCT	Hematopoietic stem cell transplantation
HSPC	Hematopoietic stem and progenitor cell
HU	Hydroxyurea
HUDEP	Human umbilical cord blood-derived erythroid progenitor
HUVEC	Human umbilical vein endothelial cell
KGP	The 1000 Genomes Project
LD	Linkage disequilibrium
NHLBI	National Heart, Lung, and Blood Institute
NIH	National Institutes of Health
QTL	Quantitative trait locus/loci
RFLP	Restriction fragment length polymorphism
RNA	Ribonucleic acid
SCA	Sickle cell anemia
SCD	Sickle cell disease
SCT	Sickle cell trait
SNP	Single nucleotide polymorphism
SNV	Single nucleotide variation
TF	Transcription factor
TFBS	Transcription factor binding site
TOPMed	Trans-Omics for Precision Medicine

Abstract

Background

More than 300,000 babies are born with sickle cell disease (SCD) each year. About 79% of these births occurs in sub-Saharan Africa where the sickle variant is known to have originated in the genetic background of the ancestors of Agriculturalist populations. Although the variant is highly lethal, the protection it confers against severe malaria in its heterozygous form has resulted in its persistence in sub-Saharan Africa where malaria is endemic. Without intervention, 50–90% of affected children in many sub-Saharan African countries die before their fifth birthday. The search for a definitive cure or effective disease-modifying therapy is therefore an imperative. Fetal hemoglobin (HbF) has long been recognized to ameliorate SCD severity whereby patients harboring natural genetic variations that lead to the persistence of high HbF levels in their blood tend to live longer with fewer complications. The HbF quantitative trait is highly heritable (89%), and is therefore the focus of enormous research for therapeutic purposes. Genetic polymorphisms influencing HbF level have been identified in three major loci; *BCL11A*, *HBS1L-MYB*, and *HBG2*. However, these loci jointly explain less than 30% of HbF variability in African sickle cell anemia (SCA) patients as compared to ~50% in African Americans and non-anemic Europeans. Genome-wide association studies have been employed to replicate two of the three major loci (*BCL11A* and *HBS1L-MYB*) in African SCA patients, as well as to identify new loci including *BACH2* involving a meta-analysis of African Americans and Tanzanians, *SLC28A3*, *TICRR*, and *PIEZO2* in Nigerians, and *FRMPD4* in Tanzanians. While *BCL11A* and *HBS1L-MYB* have been replicated in Cameroonian SCD patients through candidate genotyping studies, there has yet to be a genome-wide investigation. Moreover, the long survival of SCA patients in Africa despite a higher disease severity and mortality in the region hints at the enrichment of African genomes with ‘protective’ polymorphisms potentially impacting the level of HbF, the strongest modifier of SCD severity. A better understanding of the genetic architecture of HbF in African SCA patients is therefore needed to foster research into potentially novel sickle cell disease therapeutic targets.

Aims and Methods

This thesis project aimed to: **1)** present an in-depth description of the evolutionary history of the sickle cell mutation and its implication for global genetic medicine through a synthesis of publicly available data. This included updating and summarising the global georeferenced databases of the sickle gene and the *HBB* gene cluster haplotypes, systematically and critically evaluating the age and place of origin of the sickle cell mutation by incorporating information about malaria with which its evolution is highly intertwined, summarising bibliographic information on the genetic modifiers of SCD severity, and, importantly, examining other gene variants that co-occur with the sickle gene in sub-Saharan Africa that might hint at possible co-evolution and effect on SCD severity; **2)** provide a comprehensive overview of SCD in Sub-Saharan Africa, bringing out

transferable strategies and recommendations for prevention and care, through bibliographic searches; and **3)** to investigate the missing heritability of HbF in SCA patients of African ancestry from Cameroon, Tanzanian, and the USA through a multiple imputation panel and genome-wide association approach, with fine-mapping and functional analysis. Briefly, the performance of different reference haplotype panels on genotype imputation accuracy for African SCA populations from Cameroon and Tanzania was assessed. Genome-wide association analyses for the two populations using all the imputation panels was performed, and then meta-analyses of the two populations, as well as with summary statistics from the USA-based cohorts. Statistical fine mapping and extensive *in silico* functional analyses were next performed to determine the functional relevance of significant associations, while extensive haplotype structure analysis was performed to illuminate the reason for substantial heterogeneity in association signals across different populations of the same ancestry. Finally, gene-based and gene set enrichment analyses were undertaken to identify additional significant loci and significantly enriched pathways, while heritability analysis was performed to further appreciate the observations of multiple significant signals and to better understand the genetic architecture of HbF in African SCA patients.

Results

Evolutionary history of sickle cell mutation

In this first part, we successfully updated the global georeferenced databases of the sickle gene and the *HBB* gene cluster haplotypes. We showed that changes in population dynamics, contributed by migration and gene flow, might be introducing some *HBB* haplotypes in regions where they were previously absent, reflecting changes in regional SCD severity. Through our systematic and critical evaluation of the origin of the sickle cell mutation, we identified limitations to the models that have been used to estimate the age of the mutation, and we determined that the mutation is likely older than its currently held age of 22,000 years. Using data on the emergence of malaria, we determined that the mutation is most likely to have originated somewhere in West-Central Africa. Importantly, we showed overlapping distribution of the sickle gene and other gene variants that are under natural selection in sub-Saharan Africa. Data suggest that some of these gene variants impact SCD severity, while others have known modifying effects on the disease severity.

Overview of sickle cell disease in sub-Saharan Africa

We provided an overview of SCD in sub-Saharan Africa with transferable strategies for prevention and care as part of the *Lancet Haematology's* 2021 series on hematological care priorities in sub-Saharan Africa. We touched on aspects of SCD such as epidemiology, burden, mortality and life expectancy, hematological parameters, diagnosis, pathophysiology, biological and genetic modifiers of severity, management, environmental determinants, and psychosocial

effects. We also presented challenges and proposed recommendations for SCD research in sub-Saharan Africa.

Impact of reference haplotype panels on genotype imputation in African sickle cell anemia populations

To use the genomics of HbF to search for other, perhaps more effective, targets of gene editing for better management and perhaps cure of SCD, we first assessed the impact of imputing missing genotypes into genome-wide single nucleotide polymorphism (SNP) data using different reference haplotype panels. We used six different panels including one which we created ourselves from whole genome sequencing data of fifty Cameroonians. The key observations included: i) different imputation performance for different African populations; ii) different imputation performance among different imputation panels within each African population, indicating that one variant can be imputed with vastly different accuracies across different panels, reflecting differences in haplotype structures across the panels resulting from different tagging schemes. This underscores the complementary use of the panels, with an expectation of different patterns of association (panel-specific significant signals).

Genome-wide association analysis, statistical and functional fine mapping

As expected from our assessment of imputation performance, we observed multiple panel-specific significant signals. We replicated the major known loci including *BCL11A* and *HBS1L-MYB*, and uncovered fourteen novel loci. The most significant of the novel loci, *FLT1*, which was observed in Cameroonians and replicated in a meta-analysis with Tanzanians, has known role in hematopoiesis. Fine-mapping and *in silico* functional analyses suggest an important role in HbF induction.

Gene-based, gene set enrichment, and heritability analyses

Gene-based analysis confirmed significant signals in the three major HbF-influencing loci, *BCL11A*, *HBS1L-MYB*, and *HBB*. Gene set enrichment analysis revealed an overwhelming enrichment of hematopoiesis related pathways, as well as hemoglobin as the major enriched biological component, while blood traits were the most enriched phenotypes. Consistent with these, we estimated HbF heritability in a joint cohort of Cameroonian and Tanzanian SCA patients at 94% suggesting an enrichment of these populations with HbF-influencing loci in a way that has probably been underappreciated, and that might be better revealed through whole-genome sequencing.

Conclusion and perspectives

Our study presented data with overwhelmingly support for a single African origin of the sickle cell mutation, with opportunity for further research on determining the true age of the mutation.

Tracking the movement of the mutation through the distribution of the *HBB* gene cluster haplotypes highlighted changing population dynamics that are important for public health. Calling attention to gene variants whose distribution in sub-Saharan Africa overlaps with the sickle gene is important because the co-occurrence could mean co-evolution which might suggest an impact on SCD severity. Therefore, further evolutionary genetic studies are warranted to understand the interactions of these gene variants and the sickle gene. We equally showed that progress in implementing newborn screening and comprehensive care in some sub-Saharan African countries has been encouraging. Early diagnosis and family education on management can reduce morbidity, while immunisation and hydroxyurea therapy initiated in affected children as young as 9 months old can greatly improve quality and quantity of life. Research in sub-Saharan Africa is urgently needed to establish the exact prevalence, mortality, and morbidity, environmental and genetic factors affecting clinical complications, and life expectancy of patients with sickle cell disease, to facilitate the design of future risk models and to investigate novel routes for therapeutic options with the ultimate aim of improving the clinical outcomes of patients with SCD in all parts of the world. Finally, our discovery of new genes that are associated with blood levels of HbF which is the strongest modifier of SCD severity and the target of gene editing means that our work has tremendous potential towards discovery of novel SCD therapeutic targets, but also in improving our understanding of the genetic architecture of HbF in understudied African populations.

Chapter 1: General Introduction

“either it [sickle cell anemia] kills directly or so weakens the patient that it renders him liable to succumb to some intercurrent infection”

Dr E. A. Beet (1947)

1.1. Sickle cell disease burden: a historical background

Recent estimates published by the Global Burden of Disease (GBD) 2021 Sickle Cell Disease Collaborators reveal that the number of people living with sickle cell disease (SCD) globally rose from 5.46 million in 2000 to 7.74 million in 2021, a 41.4% increase that was largely driven by population growth in sub-Saharan Africa and an overall increase in survival of patients ¹. In 2021 alone: 515,000 babies were born with sickle cell disease—already exceeding a forecast of 404,190 by 2050 made by Piel et al., in 2013 ²—and 78.6% of these births (i.e. 405,000) occurred in sub-Saharan Africa; an estimated 376,000 individuals with SCD died globally, of which 70% (i.e. 265,000) occurred in sub-Saharan Africa. Certainly, SCD is most commonly seen in individuals of African origin including admixed populations with large proportions of African ancestry, and the highest disease prevalence and mortality rates are seen in sub-Saharan Africa where sickle cell anemia (SCA), the most severe and highly lethal form of the disease, contributes 70% of all cases. Globally, SCA accounted for 76.5% of all SCD births in 2021 (i.e. 394,000 of 515,000) ¹.

The first formal report of SCD involving a male patient of African origin occurred in 1910 in Chicago by Dr James B. Herrick ³. Two other reports of female patients of African ancestry in the United States of America (USA) followed, and the fourth case of a male patient of African origin reported from the Johns Hopkins Hospital by Dr Verne R. Mason in 1922 saw the first use of the term ‘sickle cell anemia’ ⁴. Based on the observation of family trees of individuals whose blood samples showed sickling (so called sickle cell trait), Drs William H. Taliaferro and J. G. Huck from the Johns Hopkins University and Hospital concluded in 1923 that *“the disease is definitely inherited”* ⁵. Between 1947 and 1949, the independent works of Dr E. A. Beet, a British Medical Officer who was based in Northern Rhodesia (present day Zambia), and Dr James V. Neel from the University of Michigan, clarified the genetic basis using family trees ^{6,7}. In 1949, the American chemist, Linus Pauling, and his students showed through hemoglobin electrophoresis that sickle-cell hemoglobin (HbS) carried more positive electric charges than normal hemoglobin (HbA), thus tracing the cause of a disease to a molecule for the first time—he coined the term ‘molecular disease’ ⁸. The molecular and genetic basis of SCD were eventually confirmed in 1958 following a series of experiments conducted by the German-American scientist, Dr Vernon M. Ingram, who was working at the British Medical Research Council in Cambridge, UK, at the time ^{9,10}. He

identified a single abnormal peptide out of twenty-six from a trypsin digest of HbA and HbS and showed that the abnormality resulted from the substitution of valine for glutamic acid at position 6 of the beta chain of hemoglobin. Remarkably, his experiments also led to the discovery of the hemoglobin C (HbC), a substitution on the same codon as the S allele in which the glutamic acid is replaced by lysine.

The rarity of sickle cell anemia cases in individuals over the age of five despite a high proportion of individuals whose blood demonstrated the sickling trait (heterozygotes) prompted Dr E. A. Beet to postulate in 1947 that *“either it [sickle cell anemia] kills directly or so weakens the patient that it renders him liable to succumb to some intercurrent infection”*⁶. As a matter of fact, it has been estimated that a majority (historically 50-90%) of affected children die before they turn five years old (compared to unaffected children)¹¹ as a result of one or many complications (discussed in Chapter 3). The prevalence and mortality rates presented at the opening of this chapter thus reflect the highly deleterious nature of the HbS mutation—almost as many affected individuals die as are born with SCD, the classic negative selection phenomenon in evolutionary genetics (discussed further in Chapter 2). The increasing progress in reducing global childhood mortality from infectious diseases has now seen non-communicable diseases, including SCD, become prominent causes of childhood mortality. In 2021, SCD was the 12th leading cause of global under-5 mortality, and it was in the top three causes of under-5 mortality in Portugal, Jamaica, Libya, Oman, and San Marino. In countries with the highest incidence rates, SCD contributes on average 3.8% of all deaths in children younger than 5 years¹.

1.2. Strategies for prevention and patient care

In most high-income countries where SCD is prevalent such as in the USA, childhood mortality due to SCD has been virtually eliminated¹² following interventions implemented as a result of research and increased awareness that followed the characterisation of the disease, and the adoption of policies for research funding such as by the USA National Institutes of Health and the British Medical Research Council in the 70s. Notable studies include the Jamaican Cohort Study of Sickle Cell Disease initiated in 1973¹³ and the Cooperative Study of Sickle Cell Disease initiated in the USA in 1979¹⁴ with key objectives to understand the clinical course (“natural history”) of the disease from childhood to death, and to provide information to improve management. The improved survival of patients in high-income countries thus led to an epidemiological transition whereby patients die from chronic complications later in life such as organ damage (particularly kidney damage), stroke and other cardiovascular complications rather than from complications in early childhood. Further research on the management of chronic complications has led to other disease-modifying and potentially curative strategies like the use of hydroxyurea (HU, also called hydroxycarbamide – which is capable of reversing the effect of the sickle hemoglobin, thus considerably reducing complications), transcranial doppler test to detect the risk of stroke,

chronic blood transfusion, bone marrow transplantation, hematopoietic stem cell transplantation (HSCT) (discussed in Chapter 3) and, more recently, gene therapy.

1.3. Gene therapy for sickle cell disease

Gene therapy holds tremendous promise as a more effective curative or highly disease-modifying option over HSCT which is primarily of the allogenic type ¹⁵. Allogenic HSCT relies on the availability of HLA-matched healthy donors who are often scarce to find, and also has multiple limitations that make it unavailable to all SCD patients. HSCT works simply by replacing HbS-containing blood cells with HbA-containing ones. Gene therapy, on the other hand, can be considered as an autologous HSCT in which the patient's hematopoietic stem cell (HSC) is harvested, the defected gene is "corrected", and then re-implanted into the patient's system to generate and repopulate the blood. Four main types of gene therapy that differ by the way they replace or reduce the proportion of HbS exist: i) gene addition in which a viral vector is used to deliver a new normal gene that encodes HbA to the stem cell of the patient; ii) gene correction in which highly sophisticated techniques are used, such as guide RNAs, to directly target and change the defective S allele to a normal A allele; iii) gene editing in which other genes that influence the proportion of HbS are targeted and directly changed; and iv) gene silencing in which the regulation of expression of other genes that influence the proportion of HbS is targeted such as by reducing their messenger RNA levels through RNA interference. Multiple clinical trials of gene therapy have been initiated including: the LentiGlobin gene addition trial ([NCT02140554](#)) which transduces the patient's hematopoietic stem and progenitor cells (HSPC) with a lentiviral vector that encodes a modified beta-globin gene ¹⁶; the MOMENTUM trial ([NCT02186418](#)) that transduces the patient's hematopoietic stem cells with a lentiviral vector containing the γ -globin gene which produces fetal hemoglobin (HbF), a potent anti-sickling molecule; the CEDAR gene correction trial ([NCT04819841](#)) that uses the CRISPR-Cas9 technique to directly target and reverse the sickle cell mutation in the patient's HSPCs; the CLIMB ([NCT03745287](#)) and PRECIZN-1 ([NCT03653247](#)) gene editing trials that use enzymes that target and disrupt the production of the master repressor of the γ -globin gene, *BCL11A*, in red blood cells only, thereby restoring high levels of HbF in patients; and the BCH-BB694 gene silencing trial ([NCT03282656](#)) which uses a lentiviral vector encoding a small hairpin (sh) RNA that targets and neutralizes the messenger RNA of *BCL11A*. The CEDAR trial was terminated in the early half of 2023 due to reports of severe adverse effects ¹⁷. Currently, the majority of SCD gene therapy strategies aim at increasing the level of HbF to a therapeutic level of >20% in the peripheral blood.

1.4. Fetal hemoglobin as a potent sickle cell disease curative agent through gene therapy

To understand the enormous interest in targeting HbF as a potent SCD gene therapeutic strategy, an understanding of the production, regulation, and composition of various hemoglobin forms at different human developmental stages is important. The human hemoglobin (Hb) molecule is a

composite of four proteins called globin chains: two alpha-like globin chains (α_2) encoded by the alpha-like globin gene cluster (HBA, containing *HBA1* and *HBA2* and located on chromosome 16), and two beta-like globin chains (β_2) encoded by the beta-like globin gene cluster (HBB, containing five active and one inactive genes and located on chromosome 11) ¹⁸. The six genes of the *HBB* gene cluster include: epsilon (ϵ or *HBE1*); gamma A ($A\gamma$ or *HBG1*); gamma G ($G\gamma$ or *HBG2*); delta (δ or *HBD*); beta pseudogene ($\psi\beta$ or *HBBP1*); and beta (β or *HBB*). Their genetic arrangement, 5'—*HBE1*—*HBG2*—*HBG1*—*HBBP1*—*HBD*—*HBB*—3', allows for a robust and highly sophisticated stage-specific regulation of their expression at different developmental stages: *HBE1* is expressed in the embryonic stage (in the yolk sac), *HBG1* and *HBG2* at the fetal stage (in fetal liver), while *HBD* and *HBB* are the most predominant after birth (produced in the bone marrow) ¹⁹ (Figure 1.1). A distal locus control region (LCR) approximately 6kb – 18kb upstream of *HBE1* is responsible for the regulation of the entire locus while gene-specific promoters that are selectively activated or inactivated at each developmental stage by a stage selector element are responsible for the regulation of each gene ¹⁹. Expression of the alpha-globin chains is maintained throughout life, and their pairing with the different beta-globin chains give rise to different compositions of hemoglobin in blood at the different developmental stages: i) embryonic stage: mainly $\alpha_2\epsilon_2$ and only traces of $\alpha_2\gamma_2$ (HbF); ii) fetal stage: mainly $\alpha_2\gamma_2$ (HbF – encoded by the *HBG2* gene); iii) after birth (greater than one year of age): mainly $\alpha_2\beta_2$ (HbA; at least 97%) and $\alpha_2\delta_2$ (HbA₂; 1 - 3%). After birth, the level of HbF generally drops to less than 15% and less than 1% in adults in the general population ^{19,20}.

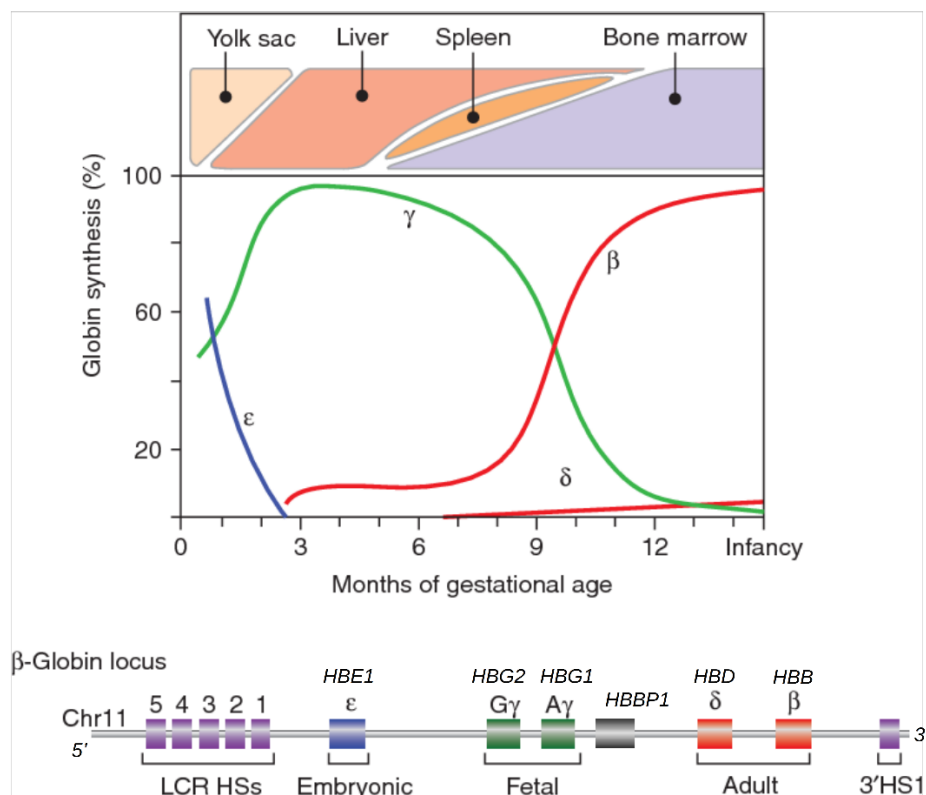


Figure 1.1 Hemoglobin switching

Expression of the human beta-like globin genes at different developmental stages. The top panel shows the different developmental stages (colored differently) along which different human hemoglobin types are produced. The plot shows the expression of the different hemoglobin types against their respective developmental stages, while the bottom panel shows the beta-globin gene cluster on chromosome 11 in which the different genes are colored according the developmental stages in which they are expressed. DNase I hypersensitivity sites of the locus control region (LCR) upstream of the *HBE1* gene and another downstream site are indicated in purple color. The figure was adapted from ²⁰.

Interestingly, it was discovered that in a subset of individuals in the general population (anemic and non-anemic), HbF concentration in blood remains relative high well into adulthood (up to 60%), and this was apparently inherited—a condition termed hereditary persistence of fetal hemoglobin (HPFH) ^{21,22}. In 1961, researchers from the University of Mississippi confirmed that elevated amounts of HbF in the blood of SCA patients was the reason for the amelioration of the disease ²³. These observations raised questions as to how the switch from fetal to adult hemoglobin occurs, and whether it can be reversed to reactivate fetal hemoglobin in adults. Therefore, when it was observed in 1984 that hydroxyurea (HU - a cytotoxic agent used to treat

certain cancers) increases HbF level in anemic monkeys ²⁴, research involving SCA patients immediately followed, and showed the effectiveness of the drug to treating sickle cell disease ^{25,26}. HU became the first approved drug by the USA Food and Drug Administration (FDA) to treat SCD in adults in 1998 and children in 2017 ^{27,28}. However, certain factors such as toxicity, fear of infertility, poor adherence, variable individual response (about 20% of recipients do not respond to the treatment), and the cost associated with its long-term use, limit its wide application and effectiveness, particularly in under-resourced settings. In Africa, HU is available only to a few populations, mostly through clinical trials assessing its effectiveness ²⁹.

The recognition that high HbF levels ameliorate SCD spurred research into the genetic basis of the HPFH. This led to the discovery and characterisation of large genomic deletions involving the removal of the *HBB* gene responsible for the deletional HPFH ^{30–32}, and single nucleotide variations in the regulatory region of the *HBG1/2* genes responsible for the non-deletional HPFH ^{33–35}. In the year 2000, a study of twins in the United Kingdom (UK) estimated that high amounts of HbF in the blood can indeed be explained by human genetic factors up to 89% of the times—so called heritability of HbF ³⁶. Three major genetic loci harboring single nucleotide variations (SNVs) have since been discovered to explain the most variability in HbF level in blood. These include *BCL11A* on chromosome 2, intergenic variants between *HBS1L* and *MYB* on chromosome 6 (*HBS1L-MYB*), and the *HBG2* regulatory region variants on chromosome 11 (particularly at a restriction site for *XmnI*, referred to as *XmnI-HBG2* and represented by the reference SNP ID rs1427407) ³⁷. The discovery of *BCL11A* as the master repressor of HbF production ³⁸, as well as *BCL11A* regulatory elements (enhancers) that are specific to erythroid cell lines ³⁹, was the major impetus for genetic manipulation to cure SCD through HbF reactivation: the disruption of erythroid-specific *BCL11A* enhancers led to an almost complete reactivation of HbF ³⁹. It was estimated in the UK twin population that the HbF heritability associated with the three major loci sum to ~50% ⁴⁰. The loci were later confirmed in SCA patients based in the United States and were found to account for ~50% of HbF heritability in the population ⁴¹. Two studies have attempted to estimate the total heritability of HbF in SCD populations, including in France (~32%) ⁴², and a combination of populations of sickle cell patients from the USA, Brazil, and Tanzania (~31 – 36%) ⁴³.

1.5. Strategies employed in the study of HbF-influencing loci

The discovery of restriction enzymes in the 1970s revolutionized molecular biology and aided considerably in the understanding of molecular diseases, including SCD. A combination of the restriction fragment length polymorphism (RFLP) technique and gel electrophoresis led to the discovery of the deletional and non-deletional HPFH ^{32,34,35}. The integration of statistical approaches such as linkage analysis and regression saw the confirmation of the *HBG2* regulatory region genetic variants and the discovery of the chromosome 6 locus ^{44,45} which was later fine-mapped to the *HBS1L-MYB* intergenic region through targeted resequencing association analyses

⁴⁶. Advancements in association mapping that saw the development of the powerful genome-wide association approach allowed the discovery of the chromosome 2 locus, *BCL11A* ^{40,47}, in which the amount of F-cells (red blood cells containing HbF) was analysed as a quantitative trait. Much of the earlier studies on HbF-influencing loci involved non-African normal and thalassemic populations. In 2010, a fine-mapping using resequencing and genotyping confirmed the presence of the three major HbF quantitative trait loci (QTL) in African American sickle cell anemia patients ⁴¹, while targeted genotyping and genome-wide association studies confirmed *BCL11A* and *HBS1L-MYB* in a Tanzanian sickle cell anemia cohort ^{48,49}. It became apparent that sickle cell anemia patients in Africa lack the *HBB* association when a targeted genotyping association study involving Cameroonian sickle cell anemia patients confirmed all but the locus ⁵⁰. It is now known that the *HBB* signal occurs on a specific haplotype background which is either very uncommon or absent in patients living in Africa. In Tanzanians, the *BCL11A* and *HBS1L-MYB* loci explain 20% of HbF variability, while the two loci were observed to explain less than 15% of HbF variability in Cameroonians. In 1992, researchers from the Johns Hopkins University used linkage analysis to identify a new locus on the short arm of the X chromosome (Xp22.2) that apparently partially explained the difference in HbF level among males and females ⁵¹. Using targeted genotyping, a genetic polymorphism associated with levels of F-cells was recently fine-mapped to Xp22.2, in *FRMPD4*, in Tanzanian SCA patients ⁵² but is yet to be replicated in other studies/populations. More recently, a GWAS of HbF in a Nigerian sickle cell anemia cohort reported significant associations in three new loci, *SLC28A3*, *TICRR*, and *PIEZO2*, in addition to the *BCL11A* and *HBS1L-MYB* loci which explained 24% of HbF variability in the population ⁵³, while a GWAS involving SCD patients from the USA, Brazil, Tanzania, and Thailand described a novel locus on chromosome 6, *BACH2* ⁴³.

Other more specialized techniques have seen use recently. For instance the identification of *BCL2L1* on chromosome 20 through gene expression and pathway correlation ⁵⁴, *ZNF410* through transcription factor (TF) CRISPR screening ⁵⁵, *ZNF802* (*JAZF1*) through RNA interference experiments ⁵⁶, and *HIF1A* through a combination of CRISPR screening and CUT&RUN analysis ⁵⁷. In all, the limitations associated with hydroxyurea use, the high heritability of HbF with only part explained thus far, the successes achieved in the discovery and manipulation of genetic variants for HbF induction, the increased interest in and funding for research on SCD, as well as the increasing advocacy for equitable access to therapy, mean that “the future of sickle cell disease therapeutics rests in genomics” ⁵⁸.

1.6. This study: a genome-wide association approach

Aside from genome-wide association study (GWAS) and global gene expression and pathway correlation analysis, which are hypothesis-free strategies for studying disease-gene associations, the rest of the strategies mentioned above are hypothesis-dependent, relying on knowledge of potential susceptibility or trait-modifying genes. GWAS has been used with significant success to

uncover thousands of disease-associated genes ⁵⁹, requiring large sample sizes of unrelated individuals genotyped across hundreds of thousands to millions of single nucleotide polymorphisms (SNPs) ⁶⁰, although whole-genome sequencing can be used, but is however expensive given the associated sample size constraint. The most powerful whole-genome genotyping assays include more than 2 million SNPs that have been ascertained to be in linkage disequilibrium with multiple other SNPs. The most popular include the illumina^(TM) [Infinium Omni2.5-8 array](#) consisting of ~2.4 million SNPs and designed from diverse world populations in the 1000 Genomes Project (KGP), and the Infinium [Human Heredity and Health in Africa \(H3Africa\) array](#) that includes ~2.3 million SNPs ascertained in African populations, with about 10,000 ascertained in specific diseases, including SCD. The H3Africa array is thus optimized for African populations and was used to genotype samples from SCA patients in Cameroon between 2018 and 2019 in furtherance of genetic studies of HbF-influencing loci at a genome-wide scale.

Genotype imputation is a widely used method for improving power in GWAS involving whole-genome genotyping by statistically inferring missing genotypes from reference haplotype panels of whole-genome sequenced data using patterns of linkage disequilibrium ^{61,62}. Until recently, the 1000 Genomes Project's (KGP) reference haplotype panel of 2504 individuals representative of global ancestries was the most popular. The first GWAS of HbF in SCA patients conducted in 2014 by Tanzanian investigators used the KGP panel for imputation of missing genotypes ⁴⁹. Over the years, multiple, more powerful, panels have been developed. These panels typically include a large proportion of individuals of a particular ancestry and small proportions of individuals of other ancestries. These include: the Trans-Omics for Precision Medicine (TOPMed), the Haplotype Reference Consortium (HRC), the Genome Asia Pilot (GAsP), the Consortium on Asthma among African-ancestry Populations in the Americas (CAAPA), the UK10K, the African Genome Resource (AGR), the H3Africa (H3A), and the panel for imputing African Americans (AFAM) (discussed further in Chapter 4). In addition, advancements in genome-sequencing and approaches for variant calling and haplotype estimation have seen substantial improvements in the reference haplotype panels ⁶³. For instance, the 2014 Tanzanian GWAS study was based on Phase I of the KGP data, whereas the most recent versions of the panel are generated from the more improved and updated dataset (KGP Phase III version 5).

Most studies that involve imputation choose one of these panels—usually the one with the largest representation of individuals of a similar ancestry as the study samples. There have been multiple studies comparing the relative performance of the panels for different ancestral populations, and there is a consensus that the TOPMed panel is superior in terms of imputation accuracy, including for African populations ^{64–66}. Comparing the performance of the panels has helped researchers to make informed decisions as to which panel to use. However, only one study has attempted to use data imputed from multiple panels (three in total) to perform GWAS ⁶⁶. This lack of multi-panel imputation and association studies reflects the significant challenges that the

strategy poses. Recently, a strategy for combining datasets imputed with multiple panels (meta-imputation) was developed to address the challenges of the multi-panel imputation and association strategy ⁶⁷. However, the meta-imputation method requires users to have direct access to all the reference haplotype panels which is unrealistic.

Against this background, we employed in this study the multi-panel genotype imputation and association strategy to perform a genome-wide association analysis of the whole-genome genotype data that was generated from samples of Cameroonian SCA patients in furtherance of HbF genomics studies in the population. We integrated and analysed genome-wide SNP data from Tanzanian SCA patients that was genotyped using the Illumina Infinium Omni2.5-8 array, as well as summary results from previous HbF GWASs involving seven cohorts of SCA patients based in the USA. These were undertaken under the premises that: i) GWAS using an Africa-specific genotyping array would be more powerful in identifying significant associations in an African population than a more diverse array; ii) association analysis with data imputed using multiple genotype imputation panels is more powerful than with a single panel; and iii) genotype imputation with more updated versions of reference haplotype panels is more likely to identify significant loci that would be otherwise missed with earlier versions of the panels.

1.7. Rationale

Until recently, research on SCD in sub-Saharan Africa has been negligible due to limited funding. The growing recognition of SCD as a major public health challenge and the increase in funding such as by the USA National Institutes of Health (NIH) and the UK-based Wellcome Trust are seeing an uptick in research on the condition in sub-Saharan Africa by African investigators aimed at providing effective standards of care, management and treatment options to African patients. This means that a comprehensive and up-to-date understanding of what has worked in high-income settings, as well as newly developed affordable strategies, is necessary to inform priority areas that can be effectively transferred to the low- and middle-income African setting. An important area with significant implication on SCD management in Africa is the disease-modifying effect of co-morbidities, especially those associated with infectious diseases. A classic example is the African trypanosomiasis (sleeping sickness) for which protective genetic variants of the *APOL1* gene rather predispose individuals of African ancestry to kidney dysfunction, including SCD patients who are already prone to the latter ⁶⁸. The distribution of these variants remarkably resembles that of the sickle cell gene, indicating possible co-evolution and genetic interactions. Shining the spotlight on this and other such variants and co-morbidities in the context of SCD is not only important to guide the development of standards of care and management, but also in informing the formulation of questions for future research.

The ultimate goal of research on SCD is to develop a cure or a highly-disease modifying therapy such as through increasing the level of fetal hemoglobin (HbF) in the blood. The enormous focus that has been placed on genomics to achieve this goal is reflected in the recent

partnership between the NIH and the Gates Foundation to each commit at least 100 million US Dollars toward developing affordable gene-based cures for SCD and HIV, and make these available to Africans⁶⁹. The recent successes of gene therapy for SCD indicate that developing a successful and effective gene-based cure is only a matter of time. One of the aims of current HbF genomic studies is to increase the target pool beyond just *BCL11A* and *HBG2* and ultimately finding a target that is much easier to manipulate genetically, effective in inducing HbF (beyond the therapeutic limit of 20-30%), and can be clinically evaluated. The fact that only part (~50%) of the heritability of HbF has been accounted for means that there is opportunity to find additional targets, especially in Africa where the proportion of HbF heritability that has been explained is less than 30%. Moreover, surviving SCA patients in Africa (older than five years of age) have traversed childhood mortality largely without the benefits of modern healthcare strategies such as newborn screening followed by comprehensive life-saving care with penicillin prophylaxis and hydroxyurea treatment. Considering the historically high under-five mortality rate of SCA in Africa, therefore, it is likely that this group of patients represents a naturally selected population enriched with genetic variants that favor “long survival” (such as has been previously shown in patients from Cameroon through whole-exome sequencing analysis⁷⁰). It is therefore conceivable that genetic variants enriched in, and associated with pathways leading to elevated HbF levels and favoring long survival are more likely to be inherited in this population than others, translating to a relatively higher heritability of HbF. Thus, a better understanding of the genetic architecture of HbF in African SCA patients, particularly Cameroonian patients for whom no genome-wide investigation has been undertaken, is important to foster research into potentially novel SCD therapeutic targets.

1.8. Aim

This thesis project aimed generally at finding better strategies for improving the health of sickle cell disease patients, including potential targets for gene-based cures.

1.9. Objectives

1. To determine adaptively selected gene variants that co-occur and/or co-evolve with the sickle cell gene in Africa that might impact sickle cell disease severity through a synthesis of publicly available data.
2. To catalog effective and transferable strategies that have been used to care for sickle cell patients in high-income settings, and make recommendations for patient care in sub-Saharan Africa.
3. To assess the impact of different reference haplotype panels on genotype imputation in genome-wide association studies of African sickle cell anemia populations.
4. To investigate genetic variants associated with fetal hemoglobin level in sickle cell anemia patients of African ancestry at a genome-wide scale using a multi-genotype

imputation panel approach.

5. To determine the functional relevance of identified significant fetal hemoglobin associations.
6. To estimate the heritability of fetal hemoglobin in African sickle cell anemia patients.

Chapter 2: Evolutionary history of sickle cell mutation: implications for global genetic medicine

Synopsis

This chapter takes a deep dive into the evolution of the sickle cell gene variant in a way that has not been previously appreciated. It includes extensive synthesis of publicly available data, with a systematic analysis of haplotype data and knowledge of the emergence of malaria and its co-evolution with sickle cell disease, to track the movement of the sickle cell mutation in Africa and globally, and to determine the mutation origin. The chapter reviews critical data revealing that the currently held age of the sickle cell mutation is likely still an underestimation. Importantly, the chapter reviews and collates information on other adaptively selected gene variants related to other disease, and demonstrates their overlapping distribution with that of the sickle cell mutation. Drawing from the fact that some of these gene variants have modifying effects on sickle cell disease severity, the chapter concludes by recommending evolutionary genetic analysis to elucidate any co-evolution with the sickle cell mutation, and thus reveal information that might be important to consider when developing sickle cell disease standards of care and management. This chapter has been published in the Human Molecular Genetics journal with the details below, and has been included here as was published. In the publication, the method section is included as supplementary information. In this thesis, the method has been included between the introduction and the results of the chapter for clarity.

Kevin Esoh, Ambroise Wonkam, Evolutionary history of sickle-cell mutation: implications for global genetic medicine, **Human Molecular Genetics**, Volume 30, Issue R1, 1 March 2021, Pages R119–R128, <https://doi.org/10.1093/hmg/ddab004>

Author contribution: KE did the literature search, data extraction and analysis, and drafted and finalized the manuscript. AW conceived the manuscript, supervised revisions, reviewed the literature search, and approved the final version of the manuscript.

2.1. Abstract

Resistance afforded by the sickle-cell trait against severe malaria has led to high frequencies of the sickle-cell mutation [*HBB*; c.20T>A, p.Glu6Val; OMIM: 141900 (*HBB*- β^S)] in most parts of Africa. For over the years, the RFLP-based classical *HBB* gene cluster haplotypes have been vital

in studying the evolutionary history of *HBB-β^S*. Although it is now apparent that they are not informative enough regarding genetic variability in the gene cluster, they remain a rich source of information regarding human migration patterns that have led to the distribution of *HBB-β^S*. Based on whole-genome sequencing data, researchers have now confirmed the single African origin of *HBB-β^S*. The overlapping distribution of *HBB-β^S* and other disease-associated variants means that their evolutionary genetics must be investigated concurrently. In this review: 1) we explore the evolutionary history of *HBB-β^S* and its implications in understanding human migration within and out of Africa: e.g. *HBB* haplotypes and recent migration paths of the Bantu expansion, occurrence of ~7% of the Senegal haplotype in Angola reflecting changes in population/SCD dynamics, and existence of all five classical *HBB* haplotype in Cameroon and Egypt suggesting a much longer presence of *HBB-β^S* in these regions; 2) we discuss the time estimates of the emergence of *HBB-β^S* in Africa; and finally, 3) we discuss implications for genetic medicine in understanding complex epistatic interactions between *HBB-β^S* and other gene variants selected under environmental pressure in Africa e.g. variants in *HBB*, *HBA*, *G6PD*, *APOL*, *APOE*, *OSBPL10*, *RXRA*.

2.2. Introduction

Sickle cell disease (SCD) is a group of blood disorders caused by mutations in *HBB* that promote hemoglobin (Hb) polymerization and sickling of red blood cells. The most common and most clinically severe form of SCD is sickle cell anemia (SCA, MIM: 603903), caused by homozygosity of the sickle-cell gene variant [*HBB*; c.20T>A, p.Glu6Val; OMIM: 141900 (*HBB-β^S*)], or the co-inheritance of *HBB-β^S* and β⁰-thalassemia. Although the *HBB-β^S* is estimated to have originated more than 7,000 years ago ⁷¹, with a high recessive lethality, and excess mortality of 50%-90%, it has persisted in appreciable frequencies because of the protection that heterozygotes have against severe *Plasmodium falciparum* malaria. Nevertheless, the classical *HBB*-like gene cluster haplotypes remain a rich source of information for variable sickle cell disease (SCD) clinical expressions, and for interpreting regional and global human migrations. An estimated 300,000 babies are born each year with SCD worldwide, with nearly 75% of the births occurring in sub-Saharan Africa (SSA) ⁷². Therefore, as global population expansion and migration rates are predicted to double in the next three decades ⁷³, it is expected that the *HBB-β^S* frequency will increase and reach places previously naïve to it. Moreover, increased awareness, newborn screening programs, and comprehensive care including emerging therapeutics and curative interventions will also contribute to increased SCD survival, and the *HBB-β^S* variant globally, urging for the need to understand the historical distribution of the gene in an evolutionary context.

There is considerable overlap in the geographical distribution between *HBB-β^S* and other gene variants that are known to have been adaptively selected in African genomes to protect against malaria, such as the thalassemias (alpha- and beta-), and G6PD deficiency variants. Similar

overlap exists between *HBB-β^S* and gene variants adaptively selected against other infectious comorbidities such as the African trypanosomiasis, i.e., variants in *APLO1*. Therefore, concurrent co-inheritance of *HBB-β^S* and other gene variants in African individuals could result in complex gene-gene (epistatic/non-allelic and allelic), and gene-environment interactions that may pose specific challenges or hold specific benefits that are yet to be fully investigated. In this review, we 1) explored the evolutionary genetic history of *HBB-β^S* and its implications in understanding human migration within and out of Africa, 2) reviewed and discussed the time estimates of emergence of the *HBB-β^S* variants in Africa, and finally, 3) discussed perspectives in understanding complex epistatic interactions between the *HBB-β^S* and other gene variants adaptively selected under environmental pressure in Africa (e.g. variants in *HBB*, *HBA* (OMIM: 141800), *G6PD* [OMIM: 305900] and *APOL1* [OMIM: 603743]), and their implications for genetic medicine for African populations, and beyond.

2.3. Method

The global georeferenced database of HbS data (1950 – 2015) by Piel et al.,⁷⁴ was retrieved and complemented by electronic database searches between 2015 – 2020 using the search strategies below and Figure 2.1.

2.3.1. Database

- **PubMed:** <https://pubmed.ncbi.nlm.nih.gov>
- **Scopus:** <https://www.scopus.com>
- **Web of Science:** <http://clarivate.libguides.com/webofscienceplatform/alldb>

2.3.2. Search strategy

HbS allele frequency

("sickle cell trait" OR "sickle-cell mutation" OR "sickle mutation" OR "sickle cell mutation" OR "Hb S" OR "hemoglobin S" OR "hemoglobin S" OR "hemoglobin mutation" OR "hemoglobin mutation") AND ("sickle cell anemia"[keyword] OR "hemoglobin S"[keyword] OR "sickle cell"[keyword]) NOT ("case report"[title] OR "with sickle cell disease"[title])

HBB haplotypes

("sickle cell trait" OR "sickle-cell mutation" OR "sickle mutation" OR "sickle cell mutation" OR "Hb S" OR "hemoglobin S" OR "hemoglobin S" OR "hemoglobin mutation" OR "hemoglobin mutation" OR "sickle cell") AND haplotype; limited to the last 10 years

Filters: Original **journal articles** with available **abstract** indexed in **MEDLINE, EMBASE**, or present in the **Web of Science core collection** and published in **English** language between **2010** and **2020** were retrieved.

last search: June 28th 2020, 03:09

Initial screening of the records was based on title and abstract, involving utilizing the following exclusion criteria.

2.3.3. Exclusion criteria:

- Case reports and reviews.
- Studies on the prevalence of SCD sub-phenotypes that did not directly measure SCD prevalence
- Studies based on evaluating diagnostic methods without directly measuring the prevalence of HbS genotypes in specific settings.
- Studies with samples unrepresentative of the population e.g. studies on women only, or athletes, or US army soldiers, samples ascertained for hemoglobinopathies etc.
- Studies of HbAS frequency in cases without any matched controls
- Family-based study

The 227 records retained after initial screening were further screened by applying the following inclusion criteria and after manual removal of further duplicates.

2.3.4. Inclusion criteria:

- Studies with direct measure of HbAS

2.3.5. Data extraction

- Demographic: Country, province, age group, sample size
- Genetic: HbAS, HbSS
- Study information: authors and affiliation etc.

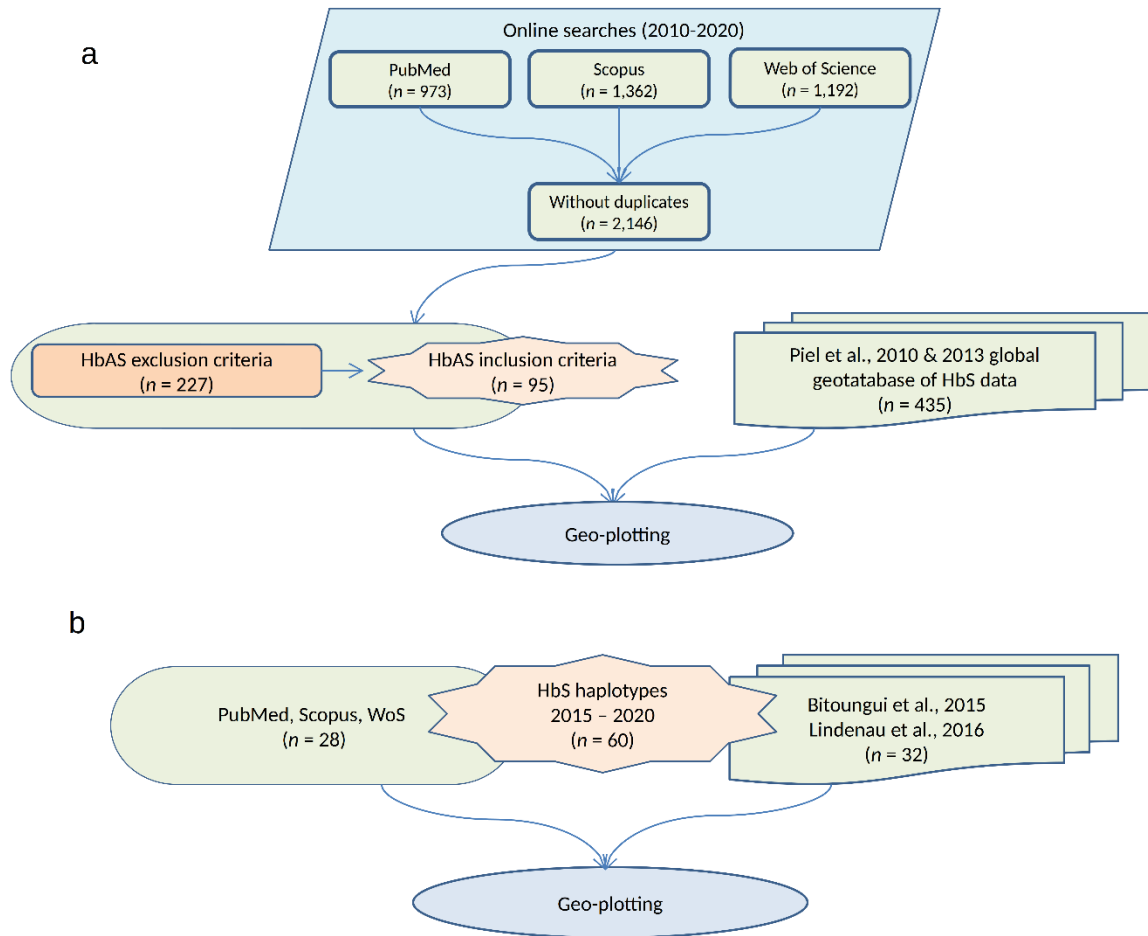


Figure 2.1 HbS and *HBB* haplotypes Search Strategy

a) Search protocol for HbS global allele frequency in three online electronic bibliographic databases. New searches were integrated with Piel et al. global geotatabase of HbS. b) Search protocol for global HBB distribution. New searches were integrated with Bitoungui et al., 2015⁷⁵ database of global HBB distribution, and Lindenau et al., 2016⁷⁶ database of HBB distribution in Brazil.

2.4. Results

2.4.1. Common Variants in Hemoglobin Genes

Over 3000 hemoglobin variants have been reported (<https://www.ithanet.eu/db/ithagenes>), several of which either affect qualitatively hemoglobin structure (hemoglobinopathies) or quantitatively hemoglobin expression level (thalassemias). A plethora of *HBB* genes hemoglobinopathies exist, of which the *HBB*- β^S (*HBB*; c.20T>A, p.Glu6Val; OMIM: 141900), HbC (*HBB*; c.19G>A, p.Glu6Lys; OMIM: 141900) and HbE (*HBB*; c.79G>A, p.Glu26Lys; OMIM: 142100) are the most clinically significant^{77,78}. Alpha thalassemias (OMIM: 604131) result mostly from the

deletions at least one of the four alpha-globin genes (*HBA*, OMIM: 141800) and rarely from point mutations, while beta-thalassemias (OMIM: 613985) are mostly caused by single nucleotide polymorphisms (SNPs) in the *HBB* gene ⁷⁹. Thalassemias major (α^0/β^0) involve the complete absence of the alpha-/beta-globin chains, and are usually clinically severe, with associated public health implications, while thalassemia intermedia (α^+/β^+), minor or carrier (trait), involving decreased production of the globin chains, are usually benign ⁸⁰. The severity of thalassemias largely depends on the degree of imbalance of alpha and beta hemoglobin chains ⁸¹. For instance, in β^0 -thalassemia, there is a marked imbalance. In β -thalassemia intermedia the ratio of α - to non- α globin is about 3-4:1, while a twofold excess in α -globin is seen in the β -thalassemia trait ⁸¹.

2.4.2. Global distribution of the *HBB*- β^S variant

The evolutionary interaction between *HBB*- β^S and malaria, that has persisted for at least 5,000 years, is perhaps the most characterized selection force in the human genome. The evolutionary link between *HBB* gene cluster variants and malaria—known as the ‘malaria hypothesis’—was postulated in 1949 by J.B.S. Haldane ⁸² and confirmed in 1954 by A. C. Allison based on β^S data from Uganda and Kenya ⁸³. Linkage analysis has shown that human genetic factors contribute up to 25% in severe malaria phenotypic variations, and that the sickle cell trait (SCT), the heterozygous state of *HBB*- β^S variant, exerts the strongest monogenic effect, albeit only 2% ⁸⁴. Indeed, as Haldane noted, only a small gain in fitness would be required for a variant to attain its equilibrium frequency ⁸². The strongest SCT protective effect has now been recorded in West African (Gambia), while the weakest effect was recorded in Cameroon, although it is unclear whether fine-scale genetic sub-structure in Cameroon may have confounded these analyses ^{85,86}.

The gain in fitness afforded by the SCT has resulted in the overlapping geographic distribution of the *HBB*- β^S variant with the historical malaria endemicity regions ⁸⁷. Generally, equatorial regions, particularly the stretch from western Ghana, through West-Central Africa and down to the north of the Zambezi, with historical malaria holo- and hyper-endemicity (year out malaria exposure and reinfections) harbour near-equilibrium frequencies of *HBB*- β^S suggesting that the mutation has been around this region longer than anywhere else ^{87,88}. The low prevalence of the *HBB*- β^S in northern Africa, its virtual absence from indigenous populations in southern Africa, as well as in the horn of Africa (Ethiopia, Eritrea, Somalia and Djibouti), reflects a low *Plasmodium falciparum* malaria selective pressure in these areas. The distribution of the *HBB*- β^S outside of Africa greatly reflects historical regional migrations from Africa to the Mediterranean, the Middle East, and the Indian sub-continent regions, and more recently trade routes such as the trans-Saharan slave trade (involving northern Africa and the Persian Gulf – the Arab World), the Islamic and European slave trades across the Nile and the Indian Ocean (involving the Arabian peninsula, India, and South East Asia), and the trans-Atlantic slave trade (involving West-Central Africa, the Americas, and Europe) ⁸⁹. The East African slave trade dates, back to the 7th century and involved mostly women taken as domestic and sex slaves, hence the high prevalence of the *HBB*- β^S gene

in Middle Eastern and South Asian indigenous populations. Although the trans-Atlantic slave trade only effectively took off in the 16th century and involved mostly men taken as plantation workers, it is believed to have involved, by far, more slaves than the East African slave trade. Malaria selective environment in these regions further augmented the gene frequencies. Recent migrations have also contributed to the prevalence of the gene in regions with low or no malaria prevalence, for instance in southern Africa ⁹⁰, North America ⁷³, and Europe ⁹¹.

Interestingly, patches of regions within Central and West Africa, notably south of Liberia, surrounded by regions with high malaria endemicity, display much lower frequencies of *HBB-β^S* ⁹². In some of these regions, the *HBB-β^S* prevalence is offset by the opposing *HBB-HbC* variant such as in Burkina Faso, southern Mali, and Northern Ghana where it is most prevalent, reaching frequencies >15% ⁹³. The homozygous HbCC affords a stronger protection against severe *P. falciparum* malaria than the HbAC and HbAS ⁹⁴. Therefore, a positive directional selection is favouring the *HBB-HbC* variant to eventually replace the *HBB-β^S* in malaria endemic regions. However, it has been observed that the rate of increase of the *HBB-HbC* variant is slower than that of the *HBB-β^S* due to the superiority in excess of average fitness associated with the *HBB-β^S* allele ⁹⁵. See ref ⁹⁶ for more information on the natural history of sickle cell disease and the environmental aspects that contributed to the increase in *HBB-β^S* frequency.

2.4.3. Global distribution of the HBB-like genes cluster haplotypes

The association of combinations of restriction fragment length polymorphic (RFLP) sites, and the *HBB-β^S* variant results in five 'classical' or 'typical' HBB haplotypes named after the region from where they were first identified (not necessarily where they originated); Senegal (SEN), Benin (BEN), Cameroon (CAM), Central African Republic (CAR) or Bantu and the Arab-India (AI) (Figure 2.2a). These have served as powerful tools for understanding the evolutionary history of *HBB-β^S* for over the years. There is considerable genetic diversity within the *HBB-like* gene cluster because of the balancing selection, and this is preserved by a recombination hotspot ~500bp upstream of *HBB-β^S* variant. Meanwhile, long-range haplotypes extending across the recombination hotspot in some West African populations ^{97,98} is consistent with a relatively recent acquisition of the *HBB-β^S* variant in these populations without enough time for recombination to break them. Generally, the AI, SEN, and BEN haplotypes are associated favourable SCD clinical outcomes whereas the CAM and CAR are associated with severe outcomes. However, a recent whole-genome sequencing study that reclassified the classical haplotypes based on linkage disequilibrium, found sub-structuring of the haplotypes which may have confounded previous associations of the haplotypes with SCD clinical severity ⁷¹.

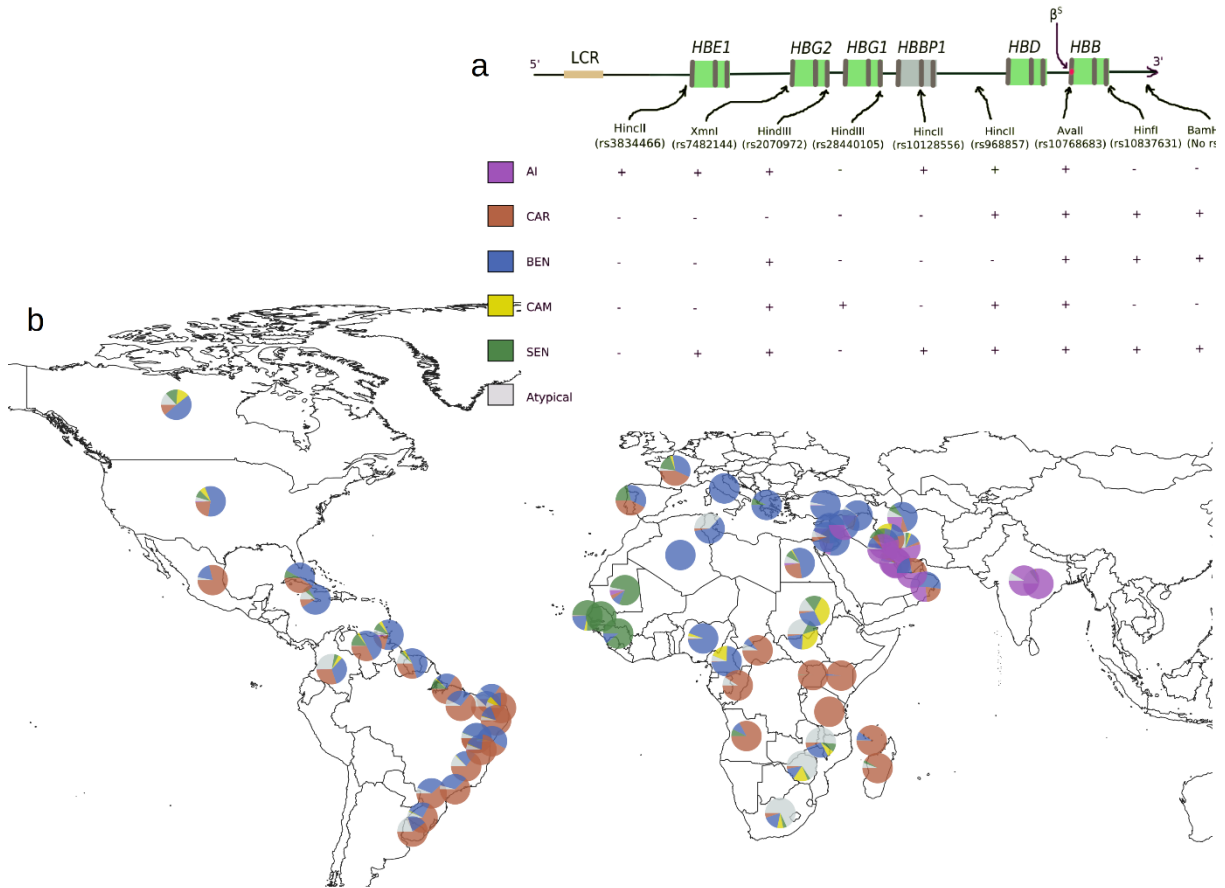


Figure 2.2 HBB-like genes cluster haplotypes

a) Haplotypes defined by presence (+) or absence (-) of specific restriction sites that are associated with the HbS mutation; LCR = locus control region. b) Global distribution of HBB haplotypes. CAR = Central African Republic, BEN = Benin, CAM = Cameroon, SEN = Senegal.

Using 60 carefully screened electronic bibliographic records; 28 from recent searches spanning 2015–2020 (Figure 2.1) and 32 from the reference lists of ⁷⁵ and ⁷⁶, we made the following observations (Figure 2.2b): First, the distribution of the Bantu and BEN haplotypes in the coasts of South America, the Caribbean and North America is consistent with the trans-Atlantic slave trade routes. Second, the prevalence of the Bantu haplotype in Eastern Africa and into Madagascar and the archipelago of Mayotte reflects ancient and recent migrant paths of the Bantu expansion, and more recent migrations. Third, the occurrence of ~7% of the SEN haplotype in the Bengo population of Angola where it was previously absent reflects changing population and disease dynamics that should be important to public health and genetic medicine ⁹⁹. Fourth, the excess of ‘atypical’ haplotypes (RFPL recombinants that differ from the five major types) in indigenous southern African populations is consistent with low malaria pressure necessary to

maintain the conservation of the *HBB* haplotypes in these populations ⁹⁰. Fifth, Cameroon and Egypt are the only African countries to report all five classical haplotype backgrounds thus far, albeit with low prevalence of the Arab-India haplotype (Appendix I). Although gene flow may be responsible for this observation, it is likely that *HBB*- β^S has been in these populations longer, especially in Cameroon (see section The African origin of the *HBB*- β^S variant).

2.4.4. The African origin of the *HBB*- β^S variant

Until recently, there has been no consensus regarding the origin of the *HBB*- β^S variant. Some studies argued a single origin (unicentric model) ^{100–102} while others posited multiple independent origins via recurrent mutations (multicentric model) coinciding with the adoption of agriculture ~4,000 – 5000 years ago ^{103,104}. It was previously thought that the classical *HBB* haplotypes represented independent origins. However, it is clear that recurrent mutations are less likely to give rise to independent origins of *HBB*- β^S variant ^{82,105}. Besides, the rate of mutation in hemoglobin genes is too slow to account for multiple independent events of the *HBB*- β^S variant ¹⁰⁶. In addition, gene conversions appear to play no role in the evolution of *HBB*- β^S ¹⁰⁷. Moreover, the absence of the *HBB*- β^S variant in most indigenous populations outside of Africa means that recurrent mutations are not generating novel *HBB*- β^S variant. It is now apparent that the different RFLPs used to classify the *HBB*- β^S haplotypes are not dense or informative enough to capture the extent of the genetic complexity necessary to explain its evolutionary history and clinical significance. Moreover, the specific number and combination of RFLPs used lacks consensus. Genomics techniques have now been used to classify the typical haplotypes based on phased SNP data, meaning that haplotypes can now be predicted in non-homozygous *HBB*- β^S individuals, thus significantly increasing the power to study the evolutionary history of the *HBB*- β^S variant ^{71,108,109}

On this basis, one study estimated *HBB*- β^S to have emerged ~7,300 years ago ⁷¹ while another report estimated *HBB*- β^S to have originated ~22,000 years ago ⁸⁸. The mutation was found to have originated in the ancestors of agriculturalists (AGR) from present-day Cameroon ⁷¹ (Figure 2.3), while only recently (~3,000 years ago) acquired by rain forest hunter gatherers (RHG) following increased gene flow among the two populations in the last ~6,000 years ⁸⁸. There is indeed evidence of increased gene flow among AGR and RHG in the past ~10,000 – 20,000 years ¹¹⁰. The observation that the CAM haplotype was the closest to the *HBB*- β^S ancestral haplotype of the five classical haplotypes offers another line of support for a Cameroonian origin ⁷¹. Interestingly, data points to the emergence of *Plasmodium falciparum* from gorillas some 40,000 – 60,000 years ago in the western African rainforests around Cameroon ¹¹¹. More so, the first known close relative of *P. falciparum*—the chimpanzee parasite *P. reichenowi* which diverged even earlier than the human-gorilla parasites split—was first discovered in Cameroon ¹¹². Thus, there appear to have been subtertian malaria around this region earlier than anywhere else, and predating the adoption of agriculture. More virulent strains that succeeded to establish repeated infections later emerged following a bottleneck in the parasite population around 5,000 – 6,000 years ago

and rapid population expansions leading to the emergence of the *P. falciparum*-specific erythrocyte invasion protein EBA-175 around 4,000 years ago ¹¹¹. Coinciding with the adoption of agriculture, these events likely ramped up pressure on the human genome, possibly contributing to the Bantu expansion ^{111,113}.

Despite the notable and important recent progresses to estimate the age of *HBB*- β^S , there have been some shortcomings in the methods used thus far. For instance, ⁷¹ assumed a complete recessive lethality of the *HBB*- β^S (relative fitness = 0) whereas it is known that homozygotes (HbSS) have a relative fitness of 0.2 (recessive lethality = 0.8) ¹¹⁴. While ⁸⁸ made use of empirical relative fitnesses, they utilized the average *HBB*- β^S frequency to mean the equilibrium frequency, which is clearly not the case. Given that the estimated *HBB*- β^S equilibrium frequency (11%) is higher than the average (8.3%) used by ⁸⁸, it implies that the *HBB*- β^S variant is older than the estimated 22,000 years. Therefore, more research is needed to determine the true age of *HBB*- β^S .

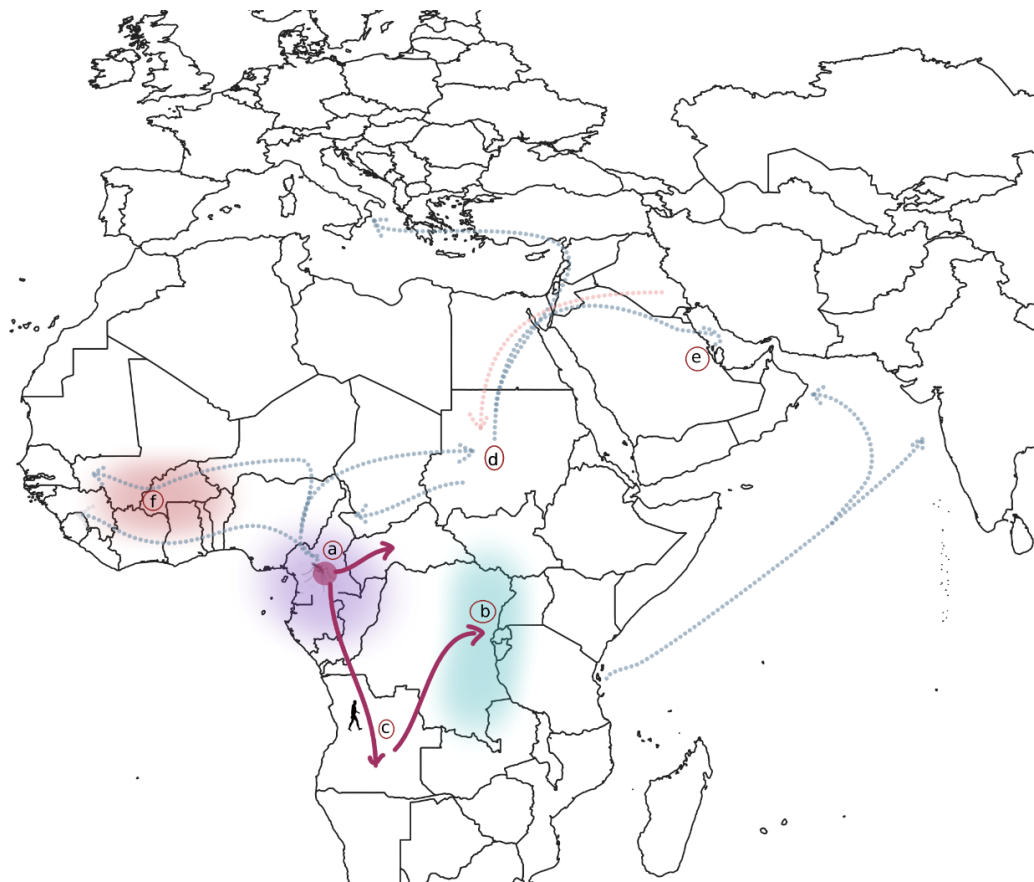


Figure 2.3 *HBB*- β^S origin in Africa

Current data support a single origin of the *HBB*- β^S variant in central-west Africa in the vicinity of present day Cameroon around 7,300 years ago ⁷¹, or ~22,000 years ago ⁸⁸ (**a**) (the precise date of occurrence still remains to be determined). This is supported by the following lines of evidence;

First, *P. falciparum* diverged from its common ancestor with *P. praefalciparum* 40,000 – 60,000 years ago in gorillas found around Cameroon ¹¹⁵, supported by the absence of *Plasmodium* infection in eastern gorillas **(b)**. Subtertian malaria pressure probably led to the emergence of *HBB-β^S* on a CAM *HBB* haplotype background **(a)** in line with ⁷¹ (It is therefore possible that *HBB-β^S* is much older than the current estimates). A generic bottleneck around 5,000 – 6,000 years ago and selection of more virulent strains of *P. falciparum* around 4,000 years ago with a rapid parasite expansion then occurred ^{111,113}, possibly triggering at least one of the waves of the Bantu expansion **(c)**. With Bantu expansion, and populations settling in various parts of Africa, it is likely that genomic recombination events within the *HBB* gene cluster have generated the other classical *HBB* haplotypes (Figure 2.2b), and the regional distributions of these haplotypes have been subsequently shaped by intra African back and forth migratory events, whose sequences and dates are still to be determined **(d)**. The distribution of the *HBB-β^S* from Africa into the Mediterranean, the Middle East, and the Indian sub-continent, where the Benin and Indian-Arab haplotype are the most prevalent, could reflect a much recent historical regional migration out of Africa as well as population admixtures, which still need to be properly investigated **(e)**. The high prevalence of HbC in West Africa even though HbS associated genotypes are known to demonstrate excess of average fitness higher than HbC associated genotypes indicate that HbS is recent in this West African region **(f)**. This needs to be properly investigation from an evolutionary genetic point of view **(f)**. Lastly, from the time *HBB-β^S* attained equilibrium in populations [currently estimated at 5000 years ago ⁷¹], it is likely that the pressure due to SCA on the human genome could have contributed to the enrichment of additional adaptive signatures as demonstrated by the ‘phenotype’ of “long-survival” in SCA patients in Africa. Therefore, genome scans for such signatures are needed to shed more light.

2.4.5. SCA and genetic modifiers

Individuals homozygous for *HBB-β^S* (HbSS) have a relative fitness of 0.2 in malarial environments. Indeed the decreased mortality due to malaria reported in HbSS patients, as compared to homozygous normal individuals (HbAA) ¹¹⁶, is likely to be explained by the by lower parasitaemia in HbSS individuals ¹¹⁷. In addition, a proportion of SCA patients usually live well into their adulthood with few complications and hospitalisations. This long-term survival, without medical interventions, has been associated with the strongest modifiers of SCD, the hereditary persistence of ‘non-deletional’ fetal Hb (HbF), a heritable quantitative trait subjected to genetic variations. Three major genetic loci have been associated with higher HbF levels in adults: *XmnI-HBG2* (OMIM: 142250, 11p15.4), *BCL11A* (OMIM: 606557, 2p16.1), and *HBS1L-MYB* (OMIM: 612450, 6q24) ¹¹⁸. The *XmnI-HBG2* (rs7482144) marks the Arab-India and SEN haplotypes and is strongly associated with favourable SCA outcomes ¹¹⁹. *BCL11A*, is an HbF-silencer and whose deactivation—such as with hydroxyurea use—leads to high HbF levels ^{118,120,121}. Targeted variants

in *BCL11A* are strongly associated with HbF levels in SCD, including among African patients ^{122–128}. Variants in the intergenic region of *HBSL1* and *MYB* showed even stronger association with HbF levels ¹²². Other loci recently associated with HbF level include *BCL2L1* (OMIM: 600039, 20q11.21) ¹²⁹ (which is yet to be replicated in Africa), and *FRMPD4* (OMIM: 300838, Xp22.2) found among patients living with SCD in Tanzania ¹³⁰. While, it is estimated that the HbF trait is highly heritable (87%) ¹³¹, the current known catalogue of variants explain just about 20-50% of the HbF phenotypic variance, indicating that many more variants remain to be uncovered, particularly in African populations, which harbour the highest genetic diversities with usually more extreme phenotypes ¹³². A particularly interesting new area of research would be to perform genome scans of SCA patients, particularly the long-term survivors, for signatures of adaptation, with specific interest to genomic loci involved in stage-specific expression of globin genes.

A recent whole-exome sequencing study uncovered novel gene modifiers (including *CLCN6* [OMIM: 602726], *OGDHL* [OMIM 617513], and *ATP2B4* [OMIM 108732]) in long-term SCD survivors among SCA patients in Africa who had neither significantly augmented HbF levels, nor alpha-thalassemia, nor favourable *HBB-β^S* haplotypes, and yet had significantly fewer complications and hospitalisation rates ¹³³. Therefore, although SCD is a monogenic disorder, it is expressed in a polygenic manner such that the tools that are commonly employed for complex disease analysis should be used to yield vital information, to unravel its pathophysiology, and in understanding the contribution of genetic factors to clinical variability in SCD.

2.4.6. Co-evolution of HBB-β^S and Other Malaria-Resistant Variants: Epistatic and Non-epistatic Interactions

Co-inheritance of HBB-β^S with alpha-thalassemia

The α-thalassemia trait (α+—thalassemia or —α/—α), is known to protect against severe malaria and like HBB-β^S (Figure 2.4a), is thus highly prevalent in malaria endemic areas, particularly the 3.7 kb alpha-globin gene (HBA1/HBA2) deletion in Africa (Figure 2.4b). In addition, α+—thalassemia is independently associated with favorable SCD clinical outcomes in Africa, whereby an increase in the frequency of the common 3.7 kb alpha-globin gene (HBA1/HBA2) deletion leads to low intracellular Hb concentration, reduced hemolytic anemia, and ultimately delayed onset of clinical manifestations and improved survival ^{128,134,135}. The prevalence of α+—thalassemia in SCA patients shows great age dependence, with increasing prevalence with age, for instance in individuals of African descent in the USA and Cuba ^{136,137}. Thus, there might be adaptive signatures in response to SCA in the HBA gene cluster. In Africa, the co-inheritance of α+—thalassemia and SCT, results in reduced severe malaria protection, marking a negative epistatic interaction ^{138,139}. The low prevalence of HBB-β^S in the Mediterranean and its almost complete absence in Oceania (Papua New Guinea) can therefore be explained by very high frequencies of α+—thalassemia (which rises to fixation levels in Papua New Guinea) ¹⁴⁰.

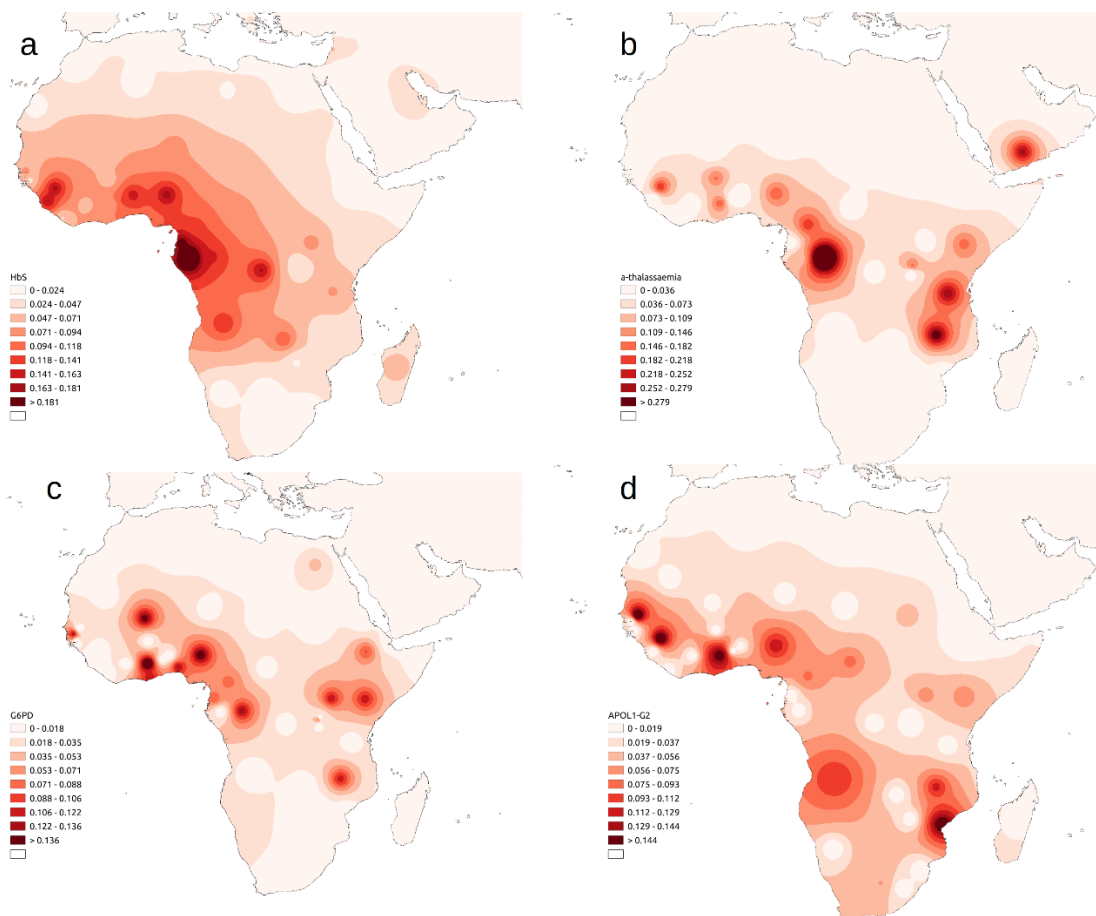


Figure 2.4 Co-evolution of the *HBB*-β^S variant and other genes variant in Africa

a) Distribution of the *HBB*-β^S variant [frequencies adapted from ⁸⁷]. b) Distribution of the 3.7 kb alpha-globin gene (*HBA1/HBA2*) deletion [frequencies adapted from ¹³⁴ and ⁸⁵]. c) Distribution of the G6PD deficiency. d) Distribution of the *APOL1* G2 [frequencies adapted from ¹⁴¹].

Co-inheritance of *HBB*-β^S with G6PD

Glucose-6-phosphate dehydrogenase (G6PD) deficiency that is heterogeneously associated with malaria protection and thus highly prevalent in malaria-endemic areas (Figure 2.4c), appears to be negatively correlated with the SCT in some populations; in India for instance ¹⁴². This is expected given that G6PD is required to protect the RBCs from oxidative stress and this anti-oxidant effect protects against haemolytic anemia in individuals with HbAA affected by malaria, although reports elsewhere are have been conflicting ^{143,144}. However, in some individuals with HbAA, G6PD deficiency is known to increase the risk of severe malaria anemia ¹⁴⁵. Increased risk of severe haemolytic anemia due to increased oxidative stress therefore exposes SCT carries to harsher clinical outcomes.

Co-inheritance of HBB- β S with APOE

APOE is encoded by the APOE (OMIM: 107741) gene on chromosome 19q13.2 and it is involved in the transfer of plasma cholesterol, an important component of Plasmodium parasite metabolism. The ancestral APOE ϵ 4 is associated with elevated plasma cholesterol levels and increased risk of Alzheimer's disease, and has therefore come under strong negative selection leading to low prevalence in many populations ¹⁴⁶. However, in sub-Saharan Africa, where the protein is reported to protect against malaria and hepatitis C, it is observed at high frequencies, particularly in hunter-gatherer populations without any association with Alzheimer's disease ¹⁴⁷. It has been observed from studies involving African children that SCT maintained its protective effect against malaria in children only in the presence of the ancestral APOE ϵ 4 ¹⁴⁸. However, the specific interaction between the SCT and APOE has not been extensively explored.

2.4.7. Co-evolution of HBB- β S and Other Variants Selected under Specific Non-malaria Environmental Pressures

Co-inheritance of HBB- β S with variants in APOL1 (trypanosomes pressure)

Association of targeted variants in *APOL1* (G1/G2; rs60910145, rs73885319, rs71785313) with increased risk of nephropathy and end stage renal disease (ESRD) in individuals of African descent has been linked to two coding variants [G1; two nonsynonymous coding variants S342G and I384M, and G2; deletion of two amino acids N388 and Y38] that have been adaptively selected for their protective role against the African trypanosomiasis ¹⁴⁹. These *APOL1* gene variants are therefore highly prevalent in sub-Saharan Africa (Figure 2.4d) and in African Americans in whom hypertension and kidney disease are highly prevalent. A gene editing study observed a significant exacerbation of nephropathy by the G2 variant under anaemic stress ¹⁵⁰. These observations have been largely confirmed by recent associations of the high-risk variants with exacerbated kidney dysfunction in SCD patients ^{151,152}, and not SCT carriers ^{153,154}.

Co-inheritance of HBB- β S with variants in OSBPL10 and RXRA (dengue pressure)

A similar increased risk of severe dengue has been associated with SCD ¹⁵⁵. Interestingly, the distribution of dengue greatly overlaps that of malaria and SCD, prompting a recent editorial to advocate the implementation of dengue vaccination for children with SCD ¹⁵⁶. However, any genetic interaction between variants in *OSBPL10* (OMIM: 606738) and *RXRA* (OMIM: 180245) that are protective against dengue fever ¹⁵⁷, and *HBB- β S* variant remains to be investigated.

2.4.8. No HbF protection against malaria

Previous studies had alluded to a possible protective role of HbF against malaria, particularly in neonates ^{158,159}. This spurred interest in a possible epistatic interaction between variants in the HbF-promoting loci and the *HBB- β S* variant, among individuals living in malaria-endemic regions in Tanzania ¹²³. However, a recent investigation by Archer *et al.*, which cited specific methodological deficiencies in previous studies, found no significant effect of HbF level on *Plasmodium* parasite growth ¹⁶⁰. This observation is indeed consistent with clinical trials in Africa,

which have consistently shown that hydroxyurea remains relatively safe in malaria-burdened regions ¹⁶¹.

2.5. Conclusion and Perspectives

There is now overwhelming evidence for the single African origin of *HBB*- β^S variant, likely in the rainforest in the vicinity of the present days Cameroon, at least 7000 years ago. However, the present review has revealed that there is still opportunity to refine the date and, perhaps the location of the emergence of *HBB*- β^S in Africa. This study emphasized that the classical *HBB*-like genes cluster haplotypes remain a rich source of information in understanding human migration within Africa and out of Africa, by highlighting the distribution of the *HBB* haplotypes and the understanding of the Bantu expansion, as well as the recent migrations, particularly exemplified by the occurrence of ~7% of the Senegal haplotype in Angola that reflects a recent changes in populations migratory dynamics within Africa, and the report of all five classical *HBB* haplotype in Cameroon and Egypt suggests a much longer presence of *HBB*- β^S in these countries/regions. Thus, it will be interesting to further the studies on the origin of the *HBB*- β^S variant in these populations.

This review also highlighted the overlapping distribution of *HBB*- β^S and other malaria resistant variants in *HBB*, *HBA*, *G6PD*, *APOE*, as well as genomic variants selected against trypanosome in *APOL1*, and dengue infections in *OSBPL10* and *RXRA*, for which evolutionary genetic investigations must be urgently performed concurrently, to reveal complex epistasis interactions. Lastly, the present review underlines the need to investigate the enormous pressure that SCA has exerted on the human genome and related adaptively selected genomic loci that are still to be fully uncovered. Therefore, future research should explore the evolutionary history of SCD in relation with other adaptive signatures within the exome, looking at variants of high functional impact, and in the genome, focusing on investigating the missing heritability of the strongest clinical modifiers of SCD i.e. HbF levels. It is also evident that methods used for complex trait and polygenic analyses would yield vital information *HBB*- β^S genomics, to decipher the panel of genomic information that could enhance genetic medicine practice in Africa, and globally.

2.6. Limitation

Initial screening of articles was based on article title and abstract only. Only a few articles that did not contain HbAS information in title or abstract, which were however sure to contain such information in the main text were retained. Hence some articles with relevant information in the main text were likely missed at this stage.

Conflict of Interest statement: None declared.

2.7. Funding

The National Institutes of Health (NIH, USA) (grants 5u24HL135600-02 and U01-HG-009716 to A.W.); the African Academy of Sciences/Wellcome Trust (grant H3A/18/001; grant 5u24HL135600-02 to K.E.).

Chapter 3: Sickle cell disease in sub-Saharan Africa: transferable strategies for prevention and care

Synopsis

This chapter provides an overview of sickle cell disease in sub-Saharan Africa with transferable strategies for prevention and care. This was part of The Lancet Haematology's 2021 series on hematological care priorities in sub-Saharan Africa. It touches on aspects of sickle cell disease such as epidemiology, burden, mortality and life expectancy, hematological parameters, diagnosis, pathophysiology, biological and genetic modifiers of severity, management, environmental determinants, and psychosocial effects. It also presents challenges and proposes recommendations for sickle cell disease management and research in sub-Saharan Africa. Results in this chapter were published in The Lancet Haematology journal with the details below. In the publication, the search strategy is added at the end, after conclusion. In this thesis, the search strategy has been brought forward just after the introduction and before the result of the chapter for clarity.

Esoh, K., Wonkam-Tingang, E., & Wonkam, A. (2021). Sickle cell disease in sub-Saharan Africa: transferable strategies for prevention and care. **The Lancet Haematology**, 8(10), e744–e755. [https://doi.org/10.1016/S2352-3026\(21\)00191-5](https://doi.org/10.1016/S2352-3026(21)00191-5)

Author contribution: KE did the literature search and drafted and finalized the manuscript. EW-T reviewed the selected references for quality and revised the manuscript. AW conceived the manuscript, supervised revisions, reviewed the literature search, and approved the final version of the manuscript.

3.1. Abstract

Sickle cell disease can be life-threatening or chronically debilitating for both children and adults. Worldwide, more than 300 000 children are born with sickle cell disease every year, over 75% of whom in sub-Saharan Africa. Increased awareness and early interventions, such as neonate screening and comprehensive care, have led to considerable reductions in mortality in children younger than 5 years in high-income countries. However, sickle cell disease prevention and care have largely been neglected in Africa. Without intervention, 50–90% of affected children in many sub-Saharan African countries die before their fifth birthday. Fortunately, increasing initiatives in sub-Saharan Africa are piloting interventions such as neonate screening and comprehensive care,

and as mortality declines, quality of life and increased life expectancy become major targets for interventions. Hydroxyurea (hydroxycarbamide) and hematopoietic stem-cell transplantation have already been shown to be effective therapies in high-income countries, but are either not widely accessible or too expensive for most African populations. These challenges are being alleviated by numerous networks evolving through international collaborations that are positively changing the outlook of sickle cell disease management in sub-Saharan Africa. In this Series paper, we describe the epidemiology, pathophysiology, clinicobiological profile, and psychosocial effects of sickle cell disease in sub-Saharan Africa. We highlight transferable strategies already used for the successful management of the condition and key strategies and recommendations for affordable and comprehensive care on the continent.

3.2. Introduction

Sickle cell disease is a group of blood disorders characterised by the presence of sickle hemoglobin in red blood cells, and it mostly affects individuals of African descent. Sickle cell anemia, the most severe and the commonest form of the disease in Africa, results from the inheritance of two mutated alleles encoding sickle hemoglobin. In the mutated hemoglobin subunit β gene (*HBB*), a single nucleotide change causes an amino acid change (Glu6Val) that leads to a defect in the structure of hemoglobin (hemoglobinopathy). Mutations that lead to a complete absence of hemoglobin subunit β synthesis (β^0 -thalassemia), when co-inherited with one sickle hemoglobin allele, also cause sickle cell anemia. The co-inheritance of the sickle hemoglobin allele with other *HBB* variants (eg, hemoglobin C) results in other forms of sickle cell disease. The sickle cell mutation is now known to have originated in Africa¹⁶², where over 75% of the estimated 300 000 livebirths of children with sickle cell anemia annually occur⁸⁷. As a result of the protection that carriers of one copy of the sickle hemoglobin allele have against severe *Plasmodium falciparum* malaria, the mutation has persisted at high frequencies in Africa, where malaria caused by *P falciparum* is endemic. In addition, because of insufficient public health interventions in most African countries, such as neonate screening for sickle cell disease followed by comprehensive preventive care, historically 50–90% of affected children in sub-Saharan Africa have been estimated to die before their fifth birthday¹⁶³.

Although sickle cell disease was first described more than 110 years ago, expenditure on care and research of this condition in sub-Saharan Africa has been negligible. WHO only recognised sickle cell disease as a major public health issue in the 59th World Health Assembly, in May, 2006. In the 63rd session of its General Assembly, in 2008, the UN also recognised sickle cell disease as a public health problem and June 19 was established as World Sickle Cell Awareness Day¹⁶⁴. Advocacy by the American Society of Hematology resulted, in 2018, in the US Congress approving bills to strengthen global awareness and research into the disorder, such as the bill by Senators

Tim Scott and Cory Booker that designated September as Sickle Cell Disease Awareness Month in the USA. The effect of these resolutions and advocacies is reflected in the momentous funding for research in sub-Saharan Africa that has led to the establishment of organisations such as the SickleGenAfrica Network, funded through the Human Heredity and Health in Africa consortium, and the [Sickle In Africa Consortium](#) (which comprises the Sickle Africa Data Coordinating Centre and the Sickle Pan-African Research Consortium), funded by the US National Heart, Lung, and Blood Institute. These initiatives are committed to translating research output into high-quality and affordable care, and to working towards a definitive cure for sickle cell disease. In this Series paper, we explore the epidemiology, clinicobiological profile, and psychosocial effects of sickle cell disease in sub-Saharan Africa, and explore key interventions, previously implemented with success, that have the potential to improve care and management of sickle cell disease in the region.

Search strategy and selection criteria

References for this Series paper were identified through searches of PubMed with the words “sickle cell” and one or more of the following words or phrases to build specific search strategies: “sickle mutation”, “origin”, “HBB haplotype”, “morbidity”, “mortality”, “haematological parameter”, “newborn screening”, “neonatal screening”, “prenatal diagnosis”, “psychosocial impact”, “genetic modifier”, “hydroxyurea”, “clinical trial”, “stem-cell transplant”, “pathophysiology”, “treatment”, “management”, “transfusion”, “complication”, “clinical profile”, “prevalence”, “burden”, “diagnosis”, “point-of-care test”, “standard-of-care”, “infection”, and “bacteraemia”. Searches were done between May 27, 2020, and Jan 31, 2021, and broadly covered the field with specific attention to studies on transferable interventions for sickle cell disease published within the past 36 months. The final references list was generated from original articles and expert reviews or commentaries in English exclusively on specific topics relevant to the broad scope of this Review.

3.3. Epidemiology

3.3.1. Sickle cell disease and malaria endemicity

More than 110 years have elapsed since the first description of sickle cell disease in a 20-year-old African American by James B Herrick in Chicago, USA ¹⁶⁵. The African origin of mutated hemoglobin subunit β has been confirmed by large-scale genomic analyses, although a review, published in 2021, indicates that the correct age of the mutation and the precise location of its origin are still undetermined ¹⁶². When Anthony C Allison and his colleagues showed, in 1954, that carrying a single allele of sickle hemoglobin (called the sickle cell trait) protects against severe *P. falciparum* malaria ⁸³, they confirmed the evolutionary link between hemoglobin

subunit β variants and malaria—the so-called malaria hypothesis that had been postulated by John BS Haldane on the basis of data on β -thalassemia, in 1949⁸². Studies have since shown that human genetic factors are responsible for up to 25% of the interindividual differences in the clinical expression of severe malaria, and that the sickle cell trait contributes the largest proportion due to a single gene (up to 2%)⁸⁴. The sickle cell trait has been observed to reduce the risk of severe malaria caused by *P. falciparum* by up to 90% in carriers, compared with individuals with a normal allele^{166,167}. As a result, the sickle hemoglobin allele is highly prevalent in regions with historical malaria endemicity (Figure 3.1a)—up to 20% in some regions; although statistical estimates have placed an upper limit at 18%¹⁶⁸. Outside of Africa, the sickle hemoglobin allele is most frequent in individuals of African descent. This observation reflects the effect of migration on the spread of pathogenic mutations, for instance to the Mediterranean region (such as Padova and Monza in Italy)⁹¹ and to the American continent (where it is absent in the native population), or in regions where malaria is endemic or was historically endemic, such as in the Indian subcontinent, the Middle East, and the Mediterranean, but where this mutation was not prevalent⁸⁷. The expression of hemoglobin C, the second most common cause of sickle cell disease in Africa, is almost entirely restricted to west Africa, with an epicentre around Burkina Faso,¹³ although changing population dynamics is resulting in a wider distribution of the variant (for instance, in the population surrounding Karachi, Pakistan)^{94,169}. A homozygous mutation in *HBB* is the most common cause of sickle cell disease in Africa (65–70% of cases), the compound heterozygote form with hemoglobin C is the second most common (about 30% of cases), and β^0 -thalassemia with one sickle hemoglobin allele accounts for the remaining cases of sickle cell disease in the continent¹⁷⁰.

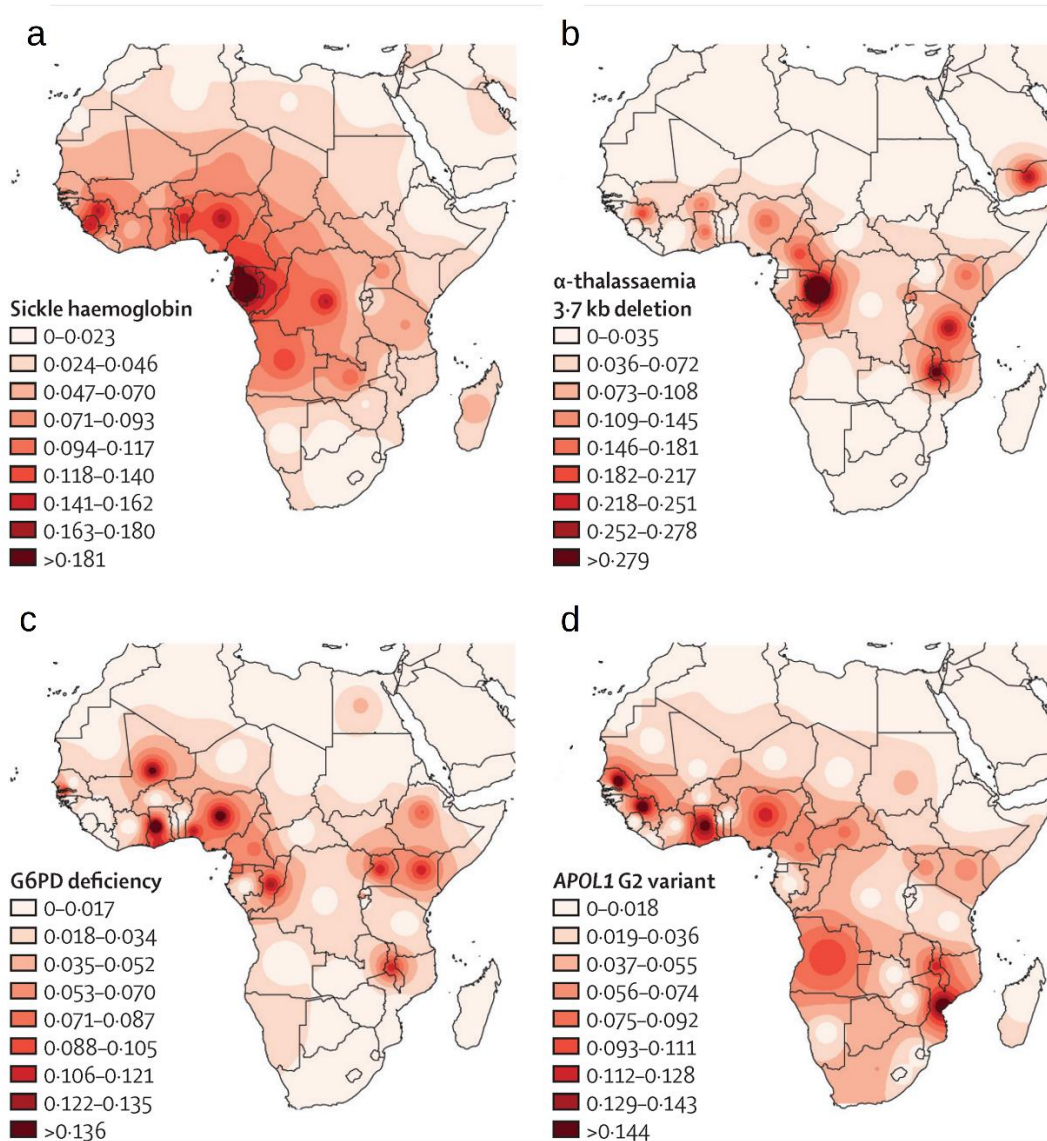


Figure 3.1 Frequencies of the sickle hemoglobin allele and other gene variants in Africa

Data shown are allele frequencies per 100 individuals in the general population. Adapted from Esoh and Wonkam¹⁶². (a) Distribution of the sickle hemoglobin allele. (b) Distribution of the 3.7 kb deletion in α -thalassemia variant. (c) Distribution of G6PD deficiency. (d) Distribution of the *APOL1* G2 variant.

3.3.2. Co-inheritance of sickle cell disease and other genetic determinants of outcomes in sub-Saharan Africa

The interaction between the sickle hemoglobin allele and other red blood cell genetic variants that also confer malaria resistance—and that are also prevalent in malaria-endemic regions—has resulted in complex genetic interactions that either increase or reduce the clinical severity of

sickle cell disease. For instance, individuals with the 3·7 kb deletion in α -thalassemia, (Figure 3.1b) which is particularly prevalent among patients with sickle cell disease in sub-Saharan Africa (compared with ethnolinguistically matched control populations), have better sickle cell disease haematological indices and lower hospitalisation rates than individuals with sickle cell disease without this deletion ¹³⁴. The G6PD deficiency variants that protect against some severe forms of malaria in sub-Saharan Africa (Figure 3.1c) are, however, associated with an increased risk of haemolytic anemia in sickle cell disease. The *APOL1* G1 and G2 variants that are prevalent in sub-Saharan Africa (Figure 3.1d) because of the protection that they provide against African trypanosomiasis (sleeping sickness) are also known to predispose individuals of African ancestry—including patients with sickle cell disease—to end-stage renal disease ¹⁴¹. The *HBB* gene cluster haplotypes (Figure 3.2) that are present in all chromosomes with the sickle cell mutation are also associated with variable sickle cell disease clinical severity: the Arab-India, Senegal, and Benin haplotypes are associated with less severe sickle cell disease clinical outcomes, whereas the Cameroon and Central African Republic haplotypes are associated with poorer outcomes ¹⁷¹. However, one study based on sequence data rather than restriction site data found sub-structuring among the haplotypes that might have confounded their association with sickle cell disease clinical severity in some populations ⁷¹, implying that further research is needed to understand the exact role of the different haplotypes in the clinical outcomes of sickle cell disease.

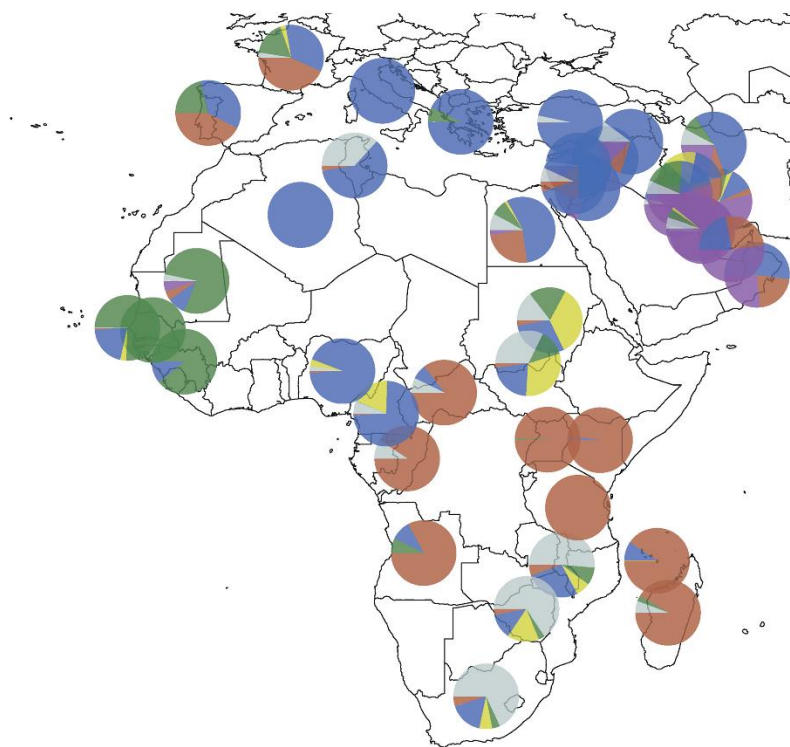
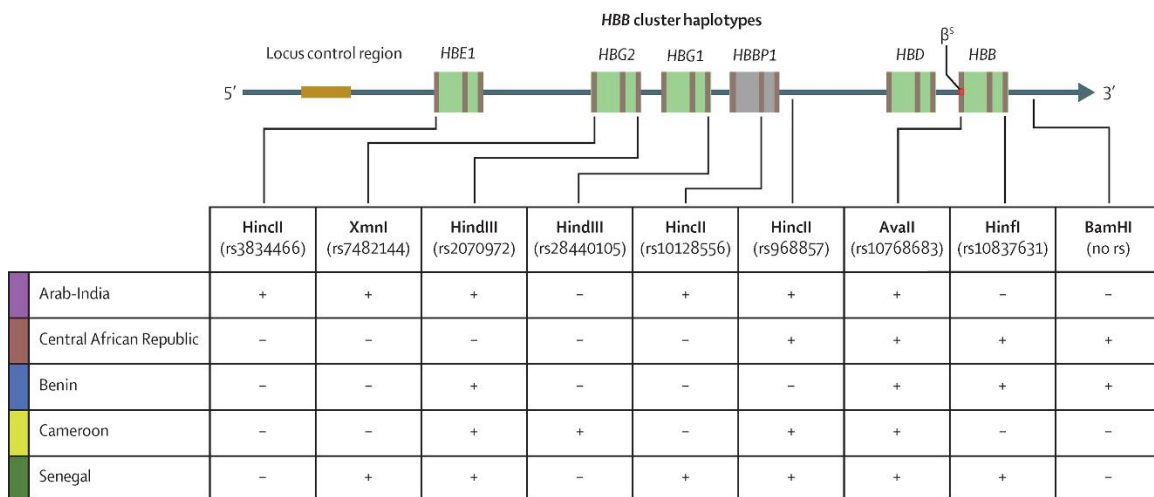


Figure 3.2 Distribution of mutated HBB gene cluster haplotypes in regional variants

The placement of pie charts across the map represents the differential proportions of *HBB* gene cluster haplotypes in countries for which data are available. Some countries have multiple pie charts representing different studies from which data were extracted, whereas for some other countries no data are publicly available.

The strongest known modifier of clinical outcome of sickle cell disease is, however, the concentration of fetal hemoglobin—a highly heritable trait—in adult blood. Three major loci have been identified to influence fetal hemoglobin concentrations: *Xmn* I- *HBG2* (rs7482144), located

at 11p15.4, *BCL11A* (rs1427407), located at 2p16.1, and *HBS1L-MYB* (rs66650371), located at 6q24¹¹⁸. Polymorphisms in these loci are associated with variable concentrations of fetal hemoglobin (1–30%) in adults. Functionally relevant variants have been proposed for each of the loci: the *Xmn*I restriction variant at *HBG2* (frequent in populations in the Middle East), the *BCL11A* variant (the most influential fetal hemoglobin modifying locus in Africa), and the 3 bp deletion at *HBS1L-MYB*. *BCL11A* is a transcriptional repressor of fetal hemoglobin, with a well characterised fetal hemoglobin-silencing role in adults¹¹⁸, and is therefore the target of hydroxyurea (hydroxycarbamide), the drug most widely used to treat sickle cell disease¹⁷². Other loci recently associated with fetal hemoglobin concentrations include *BCL2L1* (20q11.21)¹²⁹, which is yet to be found in Africa, and *FRMPD4* (Xp22.2) among Tanzanian patients with sickle cell disease¹³⁰. Most studies on fetal hemoglobin-promoting loci have been based on non-African populations, and explain up to 50% of variation in fetal hemoglobin concentrations, which means that much is yet to be discovered about other potential influencing loci in sub-Saharan African populations. In 2020, a whole-exome sequencing study of samples from Cameroonian patients with sickle cell disease uncovered variants in multiple genes, such as *CLCN6*, *SERPINC1*, *ATP2B4*, and *OGDHL*, that are linked to improved survival without complications requiring therapy¹³³. Variants in other genes (eg, *COL11A1*, *ABCC1*, *INSR*, and *ATP2B4*) have been associated with an increased susceptibility to stroke in patients with sickle cell disease¹³³. For comprehensive care to be achieved, the genetic factors that predispose some individuals to variable clinical outcomes should also be considered. African populations have the highest amounts of genetic diversity and often have more extreme disease phenotypes, so genetic studies on these populations could have a substantially greater sensitivity for disease-associated variants on the continent.

3.4. Prevalence of sickle cell disease in sub-Saharan Africa

The prevalence of sickle cell disease in sub-Saharan Africa is largely driven by the prevalence of the sickle cell trait, hemoglobin C, and β^0 -thalassemia. The best measure of the prevalence of sickle cell disease in a population would be the total number of affected births per 1000 livebirths (birth prevalence). Such estimates have been possible to obtain in high-income countries in which universal neonate screening has been implemented. Making such estimates in specific African countries, where neonate screening has been piloted as a sickle cell disease management strategy, has only become possible with the establishment of sickle cell disease treatment centres across the continent (Figure 3.3a). Generally, most national sickle cell disease prevalence rates do not exceed 2% (Appendix II). However, birth prevalence rates greater than 3% have been reported in places such as Shirati, Tanzania (3.9%)¹⁷³, suggesting that there might be more such cases in specific geographical regions in sub-Saharan Africa—especially in rural areas, where awareness of sickle cell disease is typically low. In 2019, in Nigeria, where up to 20 children with sickle cell disease are born per 1000 livebirths annually¹⁷⁴, a National Demographic and Health Survey showed that 10% of children with ages between 6 months and 5 years who had severe

anemia also had sickle cell anemia ¹⁷⁵. This finding stresses the need for countries to invest in epidemiological surveys of the prevalence of the important *HBB* variants in sub-Saharan Africa; sickle hemoglobin, hemoglobin C, and β -thalassemia. Increased investment in epidemiological surveys of all variants will enable the identification and prioritisation of regions with the highest need for crucial interventions. For example, demographic projections show that, by 2050, the prevalence of sickle cell disease will increase by about 100 000 neonates globally, and that Nigeria and the Democratic Republic of the Congo will be the countries most in need of interventions ⁷². Therefore, the prompt rollout of intervention strategies such as universal, or at least large-scale, neonate screening in these countries is particularly imperative. A survey of children aged 16–17 years in Bahrain showed that health education, carrier screening, and premarital genetic counselling are the best ways to reduce the prevalence of inherited blood disorders, including sickle cell disease ¹⁷⁶.

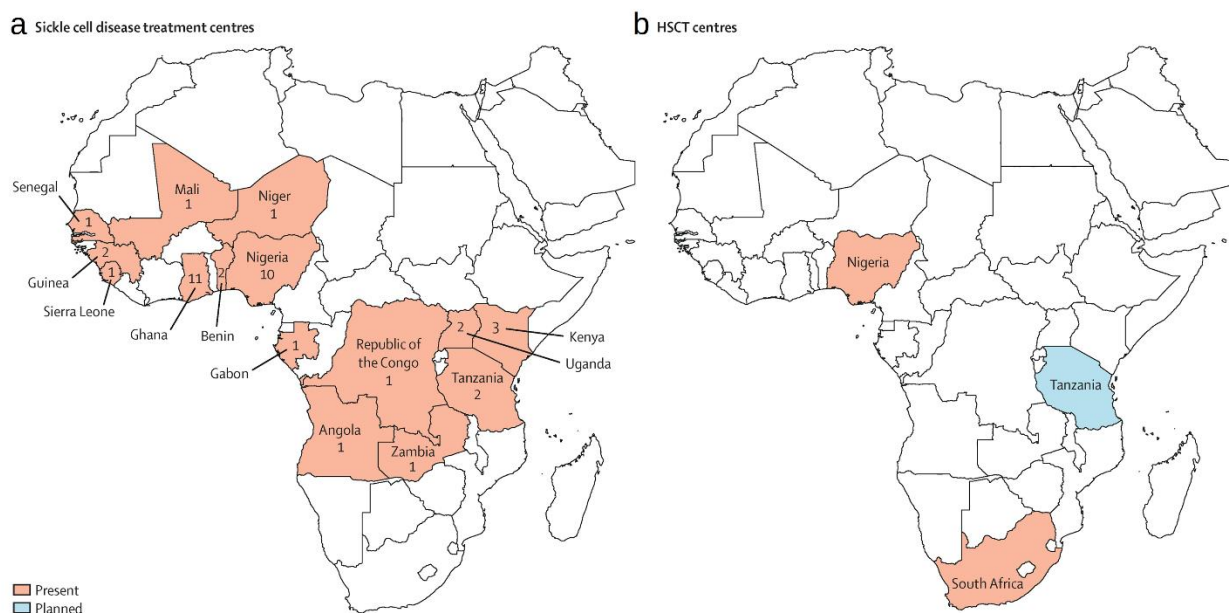


Figure 3.3 Sickle cell disease treatment centres and HSCT centres in Africa

The figures within countries show the number of reported sickle cell disease treatment centres. HSCT=hematopoietic stem-cell transplantation.

3.5. Diagnosis of sickle cell disease and neonate screening in sub-Saharan Africa

The three laboratory tests most commonly used for the diagnosis of sickle cell disease are hemoglobin electrophoresis, isoelectric focusing, and high-performance liquid chromatography. A screening test is typically done first in a clinical setting, and is followed by a confirmatory test with any of the three aforementioned techniques. A sickle cell solubility test, which involves

treating a thin blood film with metabisulfite or sodium dithionite under hypoxic conditions and observing for sickling under a light microscope, has been the major screening technique used over the years, and a positive result can suggest either sickle cell anemia or the sickle cell trait¹⁷⁷. Rapid, inexpensive, and sensitive immunoassay-based point-of-care tests, such as Sickie SCAN¹⁷⁸ and HemoTypeSC¹⁷⁹, have also been developed and found to have high sensitivity and specificity in sub-Saharan African settings^{179,180}. These tests could propel the implementation of neonate screening programmes, which have already been piloted in several sub-Saharan African countries, across the entire region. In Nigeria, for instance, a study that assessed the feasibility of implementing point-of-care tests as part of immunisation programmes found that the tests had specificity (100%) and sensitivity (100%) similar to the gold standard (high-performance liquid chromatography)¹⁸⁰. The currently available point-of-care tests are affordable—eg, the HemoTypeSC costs less than US\$ 2 per test—and can therefore be adopted by, and scaled up in, most sub-Saharan African countries. These low prices might curtail the challenges of financial cost and logistical burden of laboratory diagnostic tests in neonate screening programmes, and can result in substantial coverage of neonate screening, including in remote areas. Therefore, universal neonate screening might be achievable in sub-Saharan Africa, provided that other challenges (panel 1) are addressed—such as increasing the number of sickle cell disease clinics offering point-of-care tests, improving accessibility to such clinics, training personnel with the necessary skills to handle point-of-care tests and individual results, and offering genetic counselling to patients with positive results. Appropriate feedback of sickle cell disease neonate screening test results is essential to parental wellbeing¹⁸¹. Genetic counselling targeting individuals with the sickle cell trait will be particularly helpful in raising awareness of reproductive options, the potential implications of having a child with sickle cell disease, and the benefits of early intervention. In addition, systems for follow-up of screen-detected babies should be put in place; for example, through the integration of neonate screening programmes in immunisation campaigns, or via prospective cohort studies, such as previously done in Kenya (panel 2)¹³⁵.

PANEL 1

Challenges and recommendations for sickle cell disease research in sub-Saharan Africa

- Challenges to care of sickle cell disease in sub-Saharan Africa
- Malaria endemicity, which increases the risk of anemia and splenic dysfunction
- Absence of a universal health-care scheme in many countries, limiting access to proper sickle cell disease management
- Scarcity of penicillin prophylaxis and comprehensive immunisation programmes against pneumococcal infection and other preventable infectious diseases
- Inadequate blood supply and poor management of complications related to blood transfusion
- Inaccessibility of hydroxyurea (hydroxycarbamide) and hematopoietic stem-cell transplantation in most sub-Saharan African countries

Strategies to improve care

- Implement a neonate screening programme for an early diagnosis and management of sickle cell disease
- Establish premarital testing and genetic counselling clinics to reduce the prevalence of sickle cell disease and assist families in making decisions related to reproductive options • Ensure the continuous professional development of clinicians on the diagnosis and management of sickle cell disease
- Improve infrastructures in health-care facilities, create hematology services, and train more qualified specialists
- Increase public awareness of sickle cell disease to mitigate stigma and negative thoughts surrounding the condition, and thus reduce the psychological effects of sickle cell disease

Research recommendations

- Assess the prevalence and incidence of sickle cell disease and other hemoglobinopathies in sub-Saharan Africa
- Use whole-genome sequencing analysis to identify genetic modifiers of sickle cell disease
- Identify environmental factors and other comorbidities affecting the clinical outcomes of sickle cell disease
- Assess the co-morbidity of nephropathy caused by mutated APOL1 in patients with sickle cell disease in trypanosomiasis-endemic regions
- Explore the use of artemisinin combination therapy regimens as prophylaxis for patients with sickle cell disease, and its potential benefit on complications of the disease

PANEL 2

Proposed guidelines for the care of patients with sickle cell disease in sub-Saharan Africa

- Newborn infants with a positive screening result should be seen by a pediatrician to confirm the result, and families should be referred to a genetic counsellor if available
- Penicillin prophylaxis should be prescribed by the age of 90 days and continued throughout childhood (until age 5 years).
- Ensure immunisation against pneumococcal, meningococcal, salmonella, hepatitis A, and hepatitis B infections
- Ensure malaria prophylaxis, folic acid supply, and adequate hydration for all patients with sickle cell disease
- Patients with acute fever should be offered a malaria screening test and cultures of blood, urine, and other possible site of infections (eg, joints) if applicable
- In case of suspected bacterial infection, propose an antibiotic regimen in collaboration with a microbiologist if possible, and ensure pneumococcus infection is covered
- An annual transcranial doppler scan should be done from the age of 2 years where feasible, and hydroxyurea (hydroxycarbamide) should be prescribed in case of abnormal transcranial doppler velocity (mean velocities >200 cm/s) to detect stroke risk
- Patients with neurological signs and symptoms should be referred to a specialist (pediatrician or neurologist) and be offered a brain MRI or CT scan, if possible, to detect stroke
- Ensure a regular psychological assessment of patients and parents or legal guardians, and consider referral to a specialist (psychologist or psychiatrist) in case of depression, anxiety, or both
- Patients with recurrent episodes of pain should be prescribed hydroxyurea and have a specialist consultation (with a pediatrician or hematologist) if available
- Acute splenic sequestration and transient red cell aplasia should be suspected in case of acute fall in hemoglobin concentration (>3 g/dL below the steady state), and patients should be cared for by a specialist if available
- Acute blood transfusion should be proposed in case of acute or severe anemia
- The presence of acute chest syndrome, characterised by chest pain, fever, abnormal chest examination, and new pulmonary infiltrates on the chest x-ray, should be assessed
- The treatment of acute chest syndrome should include oxygen delivery, intravenous fluids, analgesics, antibiotics (which should cover pneumococcus), physiotherapy, and acute blood transfusion
- Consider regular transfusions, phlebotomy or exchange transfusions for the secondary prevention of stroke and in case of recurrent acute chest syndrome episodes not prevented by hydroxyurea, and prescribe iron-chelating agents to manage iron overload.
- Iron deficiency should be treated by iron supplementation under close medical supervision
- Avascular necrosis should be considered in patients with sickle cell disease presenting with either sudden onset or progressive joint pain, and a plain x-ray should be done
- Patients should be monitored at least annually for symptoms or signs of renal disease (e.g., hematuria), hypertension, and presence or progression of albuminuria, proteinuria, and

3.6. Mortality and life expectancy of patients with sickle cell disease in sub-Saharan Africa

The recessive sickle hemoglobin allele is estimated to have a lethality of 80% in malaria-endemic regions, meaning that, without therapeutic interventions, 80% of all children born with sickle cell anemia will die of the disease before the age of 5 years if the disease is not diagnosed and managed immediately ¹¹⁴. In high-income settings, such as the USA and Europe, universal neonate screening, prophylactic penicillin, and pneumococcal vaccination have drastically reduced childhood mortality rates for sickle cell disease. As sickle cell disease becomes more of a chronic disease in adults in these countries, other interventions such as hydroxyurea, blood transfusion, hematopoietic stem-cell transplantation (HSCT), and transcranial doppler screening to detect risk of stroke, among others, have considerably increased the life expectancy of patients with sickle cell disease to more than 50 years (although the life expectancy of people with the disease is still about 20 years shorter than that of unaffected individuals). In sub-Saharan Africa, where early detection of sickle cell disease and initiation of prophylactic penicillin and pneumococcal vaccination, among other interventions, have been scarce, and where the high burden of infectious diseases (such as malaria) and respiratory diseases contributes to childhood mortality, 150 000–300 000 children with SCD die before the age of 5 years every year ¹⁶³. Because these numbers are mostly estimates, studies to establish the exact mortality rates of sickle cell disease should be done across sub-Saharan Africa. The few reports of sickle cell disease mortality in sub-Saharan Africa have found mortality rates in children younger than 5 years to be between 2% ¹⁸² and 16% ¹⁸³.

An overall mortality in children younger than 5 years of 16% reported in Benin was reduced by 10 times after neonate screening and follow-up programmes were piloted in the country ¹⁸³. Mortality rates of 9.8% (Angola) ¹⁸⁴, 5.8% (Kenya) ¹³⁵, and 1.9% (Tanzania) ¹⁸² have also been reported. The implication of these high childhood mortality rates due to sickle cell disease in sub-Saharan Africa is that most affected neonates are at high risk of death in settings where disease awareness is low and access to health care is limited or non-existent. These high rates are also the rationale for the establishment of more specialised SCD clinics across the continent and for the expansion of accessibility to such clinics to patients with sickle cell disease.

3.7. Pathophysiology

Hemoglobin polymerization triggers the downstream complications of SCD ¹⁸⁵, of which vascular occlusion and haemolysis are the two most important pathophysiological events (Appendix III). Hemoglobin polymerization is mediated by a high concentration of sickle hemoglobin and low oxygen tension in red blood cells, and causes red blood cell sickling and premature breakdown ¹⁸⁶. This breakdown results in the release of hemoglobin, which depletes nitric oxide, leading to

vasoconstriction, increased expression of endothelial cell adhesion molecules, and increased platelet activation. In addition, arginase 1 is released as a result of red blood cell destruction, leading to the depletion of L-arginine (the substrate of nitric oxide synthesis). These events lead to repeated tissue ischaemia (ischaemia–reperfusion) and hypoxia, which causes oxidative stress. Increased platelet and white blood cell (neutrophils and monocytes) counts incite inflammatory responses. The major factors affecting hemoglobin polymerization include a high intracellular concentration of sickle hemoglobin relative to other hemoglobins (particularly fetal hemoglobin, which helps to prevent hemoglobin polymerization), duration of low oxygen tension, and pH ¹⁸⁵. The vaso-occlusive events and haemolysis that follow hemoglobin polymerization are the hallmarks of sickle cell disease ¹⁷⁰. Notable processes triggered by these events include increased activity of pro-inflammatory pathways, increased expression of endothelial cell adhesion molecules (eg, P-selectin and E-selectin), reduced nitric oxide bioavailability, and increased neutrophil, monocyte, and platelet counts. The understanding and targeting of these major pathophysiological pathways have led to the discovery of several therapeutics targets ¹⁷⁰. For instance, in November, 2019, the drug crizanlizumab, an anti-P-selectin molecule, was approved by the US Food and Drug Administration (FDA) to treat vaso-occlusive crisis in patients with sickle cell disease aged 16 years or older ¹⁸⁷. The drug was shown to be safe and effective in the USA, Brazil, and Jamaica ¹⁸⁷, and has now been planned for trials in Ghana and Kenya ¹⁸⁸. L-glutamine, an agent that reduces the redox potential and oxidative stress in sickle red blood cells, was also approved by the FDA in July, 2017, for the treatment of vaso-occlusive crises, acute chest syndrome, and hospitalisation of patients with sickle cell disease ¹⁸⁹. Before these drugs were approved, hydroxyurea, an anti-sickling agent, was the only FDA-approved therapy for sickle cell disease, for its ability to prevent hemoglobin polymerization through a mechanism of action not yet known ¹⁷². Several other strategies targeting various pathophysiological pathways, such as agents targeting inflammatory and hemoglobin polymerization pathways, are currently being tested in clinical trials ¹⁷⁰. Continued research is therefore important to delineate more pathophysiological pathways that might uncover novel therapeutic targets.

3.8. Important blood components measured in sickle cell disease

The components of red blood cells and their breakdown products (such as hemoglobin, haem, and lactate dehydrogenase) released after haemolysis constitute some of the haematological indices routinely measured in patients with sickle cell disease. Substances released from the liver are equally important indicators of disease severity. The liver plays a major role in the metabolism of bilirubin, an important breakdown product of haem, converting free serum bilirubin to a soluble form (conjugated bilirubin) that can be easily excreted through urine. Serum and urine bilirubin concentrations, as well as some important liver enzymes such as aspartate aminotransferase and alanine aminotransferase, are therefore routinely measured to assess the health of the liver under anaemic and oxidative stress. Haemolytic anemia is especially severe in

cases of homozygous sickle cell disease and leads to substantial reductions in hemoglobin concentrations (<7 g/dL), particularly in children younger than 5 years ¹⁹⁰. Hemoglobin concentrations in patients with sickle cell disease older than 5 years are typically stable ¹⁹¹. In a multinational study, Inusa and colleagues ¹⁹² observed that baseline (before initiation of therapy) hemoglobin concentrations are generally lower in patients with sickle cell disease in sub-Saharan Africa than in patients living in the USA and Europe, who usually have access to better care. However, the fact that patients with sickle cell disease in sub-Saharan Africa have lower baseline hemoglobin concentrations does not necessarily mean that they need higher doses of hydroxyurea to have a proportional increase in fetal hemoglobin concentrations in comparison with patients in the USA and Europe. Furthermore, geographical heterogeneity in haematological profiles within Africa has been observed (e.g., higher baseline haemolysis and total bilirubin in east African than in west African patients) ¹⁹². Low hemoglobin and high total bilirubin concentrations are typical predictors of risk of death ¹⁸², which implies that sickle cell disease clinical outcomes might be more severe in east Africa than in west Africa. This finding would be consistent with the prevalence of the Central African Republic haplotype in east Africa, and of the Senegal and Benin haplotypes in west Africa. Regional differences in fetal hemoglobin concentrations typically disappear after hydroxyurea therapy ^{193,194}. In general, monitoring of blood components in sickle cell disease is crucial to determine the risk of complications (such as nephropathy) and when specific life-saving interventions, such as initiation of hydroxyurea or regular blood transfusion for stroke prevention, are needed.

3.9. Sub-phenotypes of sickle cell disease in sub-Saharan Africa

Sickle cell disease clinical phenotypes are highly heterogeneous; some individuals can be completely asymptomatic, whereas others living in the same geographical region and under similar environmental conditions can have more severe outcomes. The clinical course of sickle cell disease typically includes a steady-state period characterised by no history of painful episodes requiring emergency procedures or hospital admissions, blood transfusions within the previous 6 months (except in case of sickle hemoglobin co-inheritance with a transfusion-dependent β^+ -thalassemia), or infections requiring medications such as antibiotics ¹⁹⁵. By implementing universal neonate screening and comprehensive care programmes in high-income countries, reduction in mortality of children younger than 5 years has transformed sickle cell disease from a life-threatening paediatric disease to a debilitating condition of adults, with chronic complications that might ultimately culminate in end-organ dysfunction ¹⁹⁶. Patients in sub-Saharan Africa face the particular challenges of infectious comorbidities and poor access to care, associated with poorer baseline haematological indices, and severe sickle cell disease clinical events with multiorgan damages. These challenges specific to sub-Saharan Africa mean that early detection and monitoring via follow-up programmes and early initiation of comprehensive care are essential to prevent the progression of chronic complications.

3.10. Specific environmental determinants of sickle cell disease clinical outcome in sub-Saharan Africa

Splenic dysfunctions such as autosplenectomy or hyposplenism can substantially predispose patients with sickle cell disease to infections. Bacteraemia is a major cause of pneumococcal disease in these patients, and the causative organisms in sub-Saharan African children with sickle cell anemia are not different from those in high-income countries: *Streptococcus pneumoniae*, non-typhoidal *Salmonella* spp, *Haemophilus influenzae* type b, *Acinetobacter* spp, and *Escherichia coli* ^{197,198}. In addition, patients with sickle cell disease have increased susceptibility to transfusion-transmissible infections, such as HIV or *Yersinia enterocolitica* infections. Malaria also presents a particular challenge to the management of sickle cell disease in sub-Saharan Africa because it is independently associated with severe anemia and splenic dysfunctions that considerably worsen the clinical outcomes of patients with sickle cell disease, particularly in children younger than 5 years ¹⁹⁹. However, studies have reported that current antimalaria chemoprophylaxis (i.e., chloroquine, mefloquine, artesunate–mefloquine, proguanil, pyrimethamine, sulfadoxine–pyrimethamine, and sulfadoxine–pyrimethamine–amodiaquine) have very little or no effect in preventing sickle cell disease complications ²⁰⁰; further research is needed to assess other antimalarial drugs in this setting. Acute chest syndrome is an important cause of morbidity and mortality in children, who are particularly susceptible to bacterial and viral pneumonias. The emergence of COVID-19 was a cause for alarm because of its potential to rapidly exacerbate acute chest syndrome in patients with sickle cell disease. Data from these patients in Europe and the USA, however, indicate that COVID-19 has not been as upsetting to this group as was anticipated ²⁰¹. Although vaso-occlusive crisis was highly prevalent in patients with sickle cell disease infected by SARS-CoV-2, no proportional increase in acute chest syndrome was seen, and acute chest syndrome was not the major cause of fatality in such patients ²⁰². There is a paucity of information on COVID-19 and sickle cell disease in sub-Saharan Africa. Other infectious comorbidities in sub-Saharan Africa that pose a challenge to patients with sickle cell disease in the region include trypanosomiasis, for which protective genetic variants in *APOL1* have been shown to increase susceptibility to end-stage renal disease ¹⁵⁰. Adherence to penicillin prophylaxis has virtually eliminated pneumococcal bacteraemia in patients with sickle cell disease and has led to reductions in morbidity and mortality in high-income countries, and should be urgently and widely implemented in sub-Saharan Africa, alongside vaccination programmes against pneumococcal infections ¹⁹⁶. Additional considerations should be given to research on dengue fever, the prevalence of which largely overlaps with sickle cell disease, and to eventually developing and implementing policies for the incorporation of dengue vaccination in the comprehensive programmes for prevention of infectious diseases and care of patients with sickle cell disease ¹⁵⁶. In addition, further research is needed to assess the efficacy of other antimalarial drugs, and there is an urgent need to

develop laboratories and infrastructure capacity for appropriate infection detection to improve transfusion services for patients with sickle cell disease in sub-Saharan Africa.

3.11. Management of sickle cell disease in sub-Saharan Africa

Effective management of sickle cell disease involves early detection with neonate screening programmes, followed by comprehensive preventive care. There are relatively few reported specialised sickle cell disease treatment centres across sub-Saharan Africa (Figure 3.3a) ²⁰³ that offer standard-of-care treatment for sickle cell disease, including home-based care and family education on primary preventive measures, such as the recognition of when to seek medical care, how to detect an enlarged spleen, and prevention of dehydration. Hydroxycarbamide remains the most extensively tested therapeutic (with over 100 ongoing or completed trials listed in [ClinicalTrials.gov](https://clinicaltrials.gov)), and the best known disease-modifying therapy for sickle cell disease, recommended for adults and for all children with sickle cell disease from the age of 9 months ²⁰⁴. Clinical trials—notably the REACH ²⁰⁵ and NOHARM ¹⁶¹ dose-escalation studies—have shown the safety and efficacy of hydroxyurea in treating sickle cell disease in sub-Saharan Africa. A major challenge with hydroxyurea has been establishing the optimal dosing and monitoring regimens, especially given the high interpatient variability in hydroxyurea response ²⁰⁶. Previous clinical trials with a fixed dose of 15–20 mg/kg per day have shown the effectiveness of hydroxyurea and that it causes only a few reversible toxicities, such as mild to moderate neutropenia followed by reticulocytopenia and then by thrombocytopenia ²⁰⁶. The REACH and NOHARM trials in sub-Saharan Africa have now shown that even higher concentrations of fetal hemoglobin (>30% increase) can be observed by gradually increasing the hydroxyurea dose every 2 months until a maximum tolerated dose of about 35 mg/kg per day is reached ¹⁶¹. Researchers have also developed a new pharmacokinetic-guided precision dosing method for children, which reduces the time to establish a child's maximum tolerated dose and improves fetal hemoglobin concentrations to more than 40% ²⁰⁷. The potential benefit of this method is that it will substantially reduce the logistical challenges of the trial-and-error dose-escalation dosing method and of laboratory monitoring and the need for provider expertise. However, this method is yet to be tested in adult patients with sickle cell anemia in particular, and in sub-Saharan Africa in general. Other potential challenges to hydroxyurea therapy in sub-Saharan Africa are hesitancy and low adherence to treatment outside experimental conditions. In addition, there is ongoing discussion on whether long-term hydroxyurea use increases the risk of adverse effects such as infertility ²⁰⁸. However, information on this subject is scarce, ²⁰⁹ thus emphasising the need for longitudinal cohort studies with long-term follow-up to accompany the implementation of hydroxyurea therapy in sub-Saharan Africa.

Blood transfusion is commonly used in sub-Saharan Africa to manage acute complications of sickle cell disease such as splenic sequestration, haemolysis, and malaria-induced haemolytic anemia. However, major challenges that hinder blood transfusion for patients with sickle cell

disease in sub-Saharan Africa include inadequate blood supply, high risk of transfusion-transmitted infections such as bacterial and viral infections, and malaria, that are usually high in the general population, and haemolytic transfusion reactions due to the so-called innocent bystander mechanism, whereby antibodies against transfused red blood cells (alloantibodies) destroy the host's own red blood cells. In addition, repeated blood transfusions can lead to iron overload. Therefore, improving the transfusion system in sub-Saharan Africa through increasing blood supply by encouraging blood donation, improving blood screening tests and donor-recipient matching, and increasing accessibility to iron-chelating agents in the case of chronic blood transfusions will greatly reduce sickle cell disease morbidity and mortality ²¹⁰. The topic of safe blood transfusion sub-Saharan Africa is covered in further detail in the fourth paper of this Series.

Allogeneic HSCT is currently the only curative measure for sickle cell disease, with a cure rate of more than 95% in HLA-matched sibling donors ¹⁷⁰. When done before the age of 2 years, allogeneic HSCT can completely restore marrow function in patients with sickle cell disease. Immunological reactions such as graft-versus-host disease limit the use of unrelated matched donors, but this complication could be alleviated with gene editing and autologous HSCT. In Africa, where large families are common, allogeneic HSCT might be facilitated by the availability of matched sibling donors. However, access to and the costs of HSCT are challenges in the continent. The process and techniques required for these interventions, such as harvesting hematopoietic stem cells (CD34⁺), require specialist training that is scarce in Africa. Only two countries in sub-Saharan Africa have HSCT centres: Nigeria and South Africa ²¹¹; Tanzania has also put forward an agenda to establish a centre ²¹² (Figure 3.3b). Increasing funding by sub-Saharan African governments and international organisations is required to invest in HSCT centres and for further clinical trials in sub-Saharan Africa, which are particularly important because genetic studies on the continent are progressively identifying Africa-specific genetic factors in sickle cell disease.

The prospect of potential curative options for sickle cell disease has recently improved by a joint initiative from Novartis and the Bill and Melinda Gates Foundation to develop a single-shot gene therapy without the need to engineer cells outside the body. Considering the small number of specialised treatment centres for sickle cell disease in Africa (Figure 3.3), the success and equability of gene therapy implementation in the continent is questionable. However, the recent, faster-than-expected development of COVID-19 vaccines by use of genomic approaches is encouraging, and accompanying equity mechanisms through international agencies such as WHO might be expected. The responsibility of implementing future gene editing and therapy for sickle cell disease in Africa will also lie with African governments and funding bodies, such as the Cure Sickle Cell Disease Initiative of the US National Institutes of Health. The time might be right to set

an ambitious global genomic research programme to uncover curative options for sickle cell disease.

3.12. Psychosocial effects of sickle cell disease

The morbidity and mortality associated with sickle cell disease impose substantial psychosocial weight on affected patients and families. The labelling of affected individuals as ‘sicklers’ in settings where awareness about sickle cell disease is low often stigmatises patients, and is frequently accompanied by suicidal thoughts. Sickle cell disease in sub-Saharan Africa is also frequently associated with witchcraft and the mother is often blamed ²¹³. Therefore, education on sickle cell disease should not be limited to patients and their families, but be extended to the general community. Results from screening tests could be particularly burdensome to families with affected individuals. Premarital screening and genetic counselling, or prenatal screening of carrier mothers, with genetic counselling on reproductive options, could also contribute to the reduction of the morbidity and mortality due to sickle cell disease.

3.13. Conclusion and perspectives

Sickle cell disease preventive measures, including primary prevention through neonate screening and genetic counselling focused on individuals with the sickle cell trait, can greatly reduce the burden and prevalence of sickle cell disease in sub-Saharan Africa and globally. Neonate screening is currently one of the most powerful tools available for sickle cell disease management; however, its application can only be effective if associated with patient follow-up, comprehensive care, and genetic counselling of sickle cell trait carriers, as previously shown in Bahrain ²¹⁴. In addition, public awareness campaigns are important to reduce the stigma associated with the disease and to promote knowledge of warning signs that should prompt medical attention. In this Series paper, we have shown that progress in implementing neonate screening and comprehensive care in some sub-Saharan African countries has been encouraging. When the disease is detected early, family education on the management of sickle cell disease can reduce morbidity; hospital attendance for immunisation campaigns should be encouraged, and hydroxyurea therapy should be initiated in affected children as young as 9 months old. Therefore, African governments must embrace implementation of universal health care and invest in making hydroxyurea more accessible and affordable to all individuals affected by sickle cell disease. There is also an immediate need to improve Africa's blood transfusion services and to understand the socioeconomic and psychosocial effects of sickle cell disease, including when providing sickle cell disease screening test results to families. Such research projects will be best done in longitudinal cohort studies ²¹⁵, requiring the concerted effort of governments, the general public, and international collaborators. Increased research in pharmacogenetics is needed to optimise long-term treatment of chronic complications and to manage toxicities, in view of the anticipated increase in the number of patients with sickle cell disease taking hydroxyurea. As hydroxyurea dose-escalation algorithms promise to further reduce sickle cell

disease morbidity and mortality, more sickle cell disease clinics in semi-urban and rural settings need to be established to reach all patients with sickle cell disease who do not have access to existing specialised treatment centres in sub-Saharan Africa. Finally, research in sub-Saharan Africa is urgently needed to establish the exact prevalence, mortality, and morbidity, environmental and genetic factors affecting clinical complications, and life expectancy of patients with sickle cell disease, to facilitate the design of future risk models and to investigate novel routes for therapeutic options with the ultimate aim of improving the clinical outcomes of patients with sickle cell disease in all parts of the world.

3.14. Contributors

KE did the literature search and drafted the manuscript. EW-T reviewed the selected references for quality and revised the manuscript. AW conceived the manuscript, supervised revisions, reviewed the literature search, and approved the final version of the manuscript.

Declaration of interests: We declare no competing interests.

3.15. Acknowledgments

AW is director of the Sickle in Africa Data Coordinating Centre, based at the University of Cape Town, South Africa, and funded by the US National Institutes of Health under grant number U24HL135600–02; and is director of the Hearing Impairment Genetic Studies in Africa, funded by the US National Institutes of Health under grant number U01-HG-009716, by the African Academy of Sciences, and by the Wellcome Trust grant number H3A/18/001. KE is a student at the Sickle in Africa Data Coordinating Centre, funded by grant U24HL135600–02. EW-T is a student at the Hearing Impairment Genetic Studies in Africa, funded by grant U01-HG-009716.

Chapter 4: Genome-wide association study of fetal hemoglobin in sickle cell disease patients of African Ancestry

Synopsis

This chapter involves a genome-wide study of genetic variants associated with fetal hemoglobin level in Cameroonian, Tanzanian, and USA-based sickle cell anemia populations. It first assessed the impact of reference haplotype panels on genotype imputation in Cameroonian and Tanzanian sickle cell anemia populations, then used datasets imputed using multiple reference panels to perform genome-wide association and meta-analyses, discovering multiple novel significant loci, including the FLT1 locus in Cameroonians which has a known role in the proliferation of hematopoietic stem cells. Functional genome-wide association and heritability analysis of fetal hemoglobin performed in sickle cell anemia patients from Cameroon and Tanzania revealed that the genomes of African sickle cell anemia patients might be enriched with fetal hemoglobin-influencing variants in a way that has probably been previously underappreciated. At the time of this thesis write up, the results in this chapter had been incorporated in a manuscript, reviewed extensively internally by senior co-investigators, and in the final stages of preparation to be submitted for review and publication with the details below. In the original manuscript and in this thesis, the method section has been deliberately placed after the results, discussions, and conclusion sections for conciseness, on the basis of complexity of the methodology. However, the supplementary information of the original manuscript has been incorporated in the main text in this thesis write up to allow a seamless flow for clarity.

FLT1 and other Novel Fetal Hemoglobin Modifying Loci in Sickle Cell Disease from Cameroon and other African Ancestries

Ambroise Wonkam^{1,2,#,*}, Kevin K. Esoh^{2,#}, Valentina Josiane Ngo Bitoungui³, Khuthala Mnika², Siana Nkya⁴, Raphael Sangeda⁵, Victoria Nembaware², Jack Morrice², Fajr Osman¹, Mike Beer¹, Julie Makani⁴, Nicola Mulder⁶, Guillaume Lettre⁷, Martin Steinberg⁸, Johnathan S. Yen⁹, Gregory A. Newby¹, Rachel Latanich¹, Rachel Levine⁷, James F Cassela¹⁰, Dan E. Arking¹, Emile Chimusa¹¹, and Stylianos E. Antonarakis¹²

¹McKusick-Nathans Institute and Department of Genetic Medicine, Johns Hopkins University School of Medicine, Baltimore, MD 21205, USA; ²Division of Human Genetics, Faculty of Health Sciences, University of Cape Town, Cape Town 7925, South Africa; ³Department of Microbiology, Haematology and Immunology, University of Dschang, Cameroon; ⁴Department of Haematology

and Blood Transfusion, Muhimbili University of Health and Allied Sciences, Dar es salaam, Tanzania; ⁵Department of Pharmaceutical Microbiology, Muhimbili University of Health and Allied Sciences, Dar es Salaam, Tanzania; ⁶Computational Biology Division, Department of Integrative Biomedical Sciences, Institute of Infectious Disease and Molecular Medicine, CIDRI-Africa Wellcome Trust Centre, Faculty of Health Sciences, University of Cape Town, Cape Town, South Africa; ⁷Institut de Cardiologie de Montréal, Université de Montréal, Montreal, Quebec, Canada; ⁸Pediatrics, Pathology and Laboratory Medicine, Boston University School of Medicine, Boston, MA, USA; ⁹Department of Hematology, St. Jude Children's Research Hospital, Memphis, TN, USA; ¹⁰Department of Pediatrics, Division of Hematology, The Johns Hopkins University School of Medicine, Baltimore, Maryland; ¹¹Department of Applied Sciences, Faculty of Health and Life Sciences, Northumbria University, Newcastle, Tyne and Wear, NE1 8ST, UK; ¹²Department of Genetic Medicine, Faculty of Medicine, University of Geneva, Switzerland.

* Correspondence: awonkam1@jhmi.edu

Joint first authors

Author Contribution: Conceived and designed the experiments: AW, **Kevin Esoh (KKE)**, EC, SEA. Performed the experiments: AW, **KKE**, KN, VJNB, SN, RS, VN, JM, FO, MB, JM, NM, GL, MS, JY, GN, RL, RL, JFC, DA, EC, SEA. Patient recruitment, samples and clinical data collection and processing: AW, VJNB, KM, SN, RS, JM. Analysed the data: AW, **KKE**, EC, DA, MS, GL. Contributed reagents/materials/analysis tools: AW, **KKE**, EC, SN, NM. Wrote the paper: AW, KKE, EC. Revised and approved the manuscript: AW, **KKE**, KN, VJNB, SN, RS, VN, JM, FO, MB, JM, NM, GL, MS, JY, GN, RL, RL, JFC, DA, EC, SEA. We thank the St. Jude Children's Research Hospital Center for Advanced Genome Engineering for designing and validating nuclease guide RNAs as well as measuring editing outcomes. The study was funded by the National Institutes of Health, USA grants 1U01HG007459-01 and U24-HL-135600 to AW. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. The corresponding author confirms that he has full access to all the data in the study and has final responsibility for the decision to submit for publication.

4.1. Abstract

Known fetal hemoglobin (HbF)-modulating genes/loci explain 10-20% variation of HbF levels in African with Sickle Cell Disease (SCD), compared to 50% among Europeans. We used the H3Africa SNP genotyping array to perform genome wide association study (GWAS) of HbF in patients with SCD from Cameroon (n = 827), followed by metanalysis with data from Tanzania (n = 884) and the United States of America (USA) (n = 2040). Six GWAS imputation panels showed improved accuracy with their size, low proportion of shared variants (28%), and wide range of exclusive variants (2.3-16%) across panels. Known HbF-associated loci e.g., *BCL11A*, and *HBS1L-MYB* were

replicated. Remarkably, fourteen new loci were identified (*FLT1*, *ZNF804A*, *MRPS22*, *PRLR*, *FRMD1*, *DMRT1*, *OPCML*, *SERPINA6*, *PLCB1*, *SLC4A4*, *GFRA1*, *LOC574080*, *LOC107984174*, *LINC01898*). The heritability of HbF levels explained by these GWAS signals is 94% (SE = 0.01). African-specific variants were found significant across panels in *FLT1* ($P = 3.3 \times 10^{-8}$), and fine-mapped within three haploblocks. Published data on mice, and cellular models suggest that *FLT1* is directly associated to (hypoxia-) induced erythropoiesis, and indirectly with HbF induction, possibly mediated by GFI1, and HIF1A which are known to derepress HbF production via the NuRD complex. In conclusion, GWAS array enriched with variants from African genomes enhanced discovering much more HbF-associated loci, including *FLT1*, a potential target for SCD therapy.

4.2. Introduction

Sickle-cell disease (SCD) is caused by a biallelic single nucleotide substitution in the beta-globin gene resulting in an amino acid substitution, *HBB* (Glu7Val, formerly known as Glu6Val)²¹⁶. As a result of the partial protection conferred by the heterozygosity of the sickle variant against severe malaria, SCD has become prevalent in areas of the world where malaria is endemic²¹⁷. It is estimated that about 300 000 babies are born worldwide each year with SCD, with nearly 75% of these births being in sub-Saharan Africa²¹⁸. In Africa, at least 30-50% of children with untreated SCD die before the age of 5 years^{11,219}. Therefore, accelerating the path for novel therapies for SCD through genomics research on fetal hemoglobin (HbF; $\alpha_2\gamma_2$) is critical.

During fetal life, HbF is the dominant form of hemoglobin. After birth, the level of HbF decreases progressively to approximately 1%, in about 8 to 12 weeks, and it is replaced by adult hemoglobin (HbA; $\alpha_2\beta_2$)²²⁰. The regulation of Hb production is controlled by repressive transcription factors including *BCL11A* and *ZBTB7A* that bind to the *HBG1* and *HBG2* gene promoters²²¹. Genetic variations in HbF-modulating genes allow some individuals the capacity to continue producing HbF in adult life. SCD patients that produce higher levels of HbF (>8%) after birth have longer life expectancy²²², because the presence of HbF in sickle RBCs delays deoxy-HbS polymerization, and thus reduces clinical complications. A highly successful gene-editing strategy for treating individuals with the most common and severe subtype of SCD, is the induction of HbF expression through downregulation of the transcription factor *BCL11A*^{223,224}.

Variants in the currently known HbF-modulating genes/loci, i.e., *BCL11A*, *HBS1L-MYB*, and *XmnI-HBG2*, explain only 10-20% of the variation of HbF levels in African individuals with SCD^{48,50}, compared with nearly 50% of the variation in HbF levels among Europeans⁴⁰. Expanding genomic research in populations of African ancestry could uncover the missing heritability of HbF-promoting loci⁵⁸.

In this study, we used the Human Heredity and Health (H3Africa) consortium SNP genotyping array developed from whole genome enriched for common variants in sub-Saharan Africans with

3280 individuals from 17 African countries (<https://www.h3abionet.org/h3africa-chip>), to identify genomic variations associated with HbF levels in a Cameroonian cohort of 827 patients living with Sickle Cell Anemia from Cameroon. This was followed by a meta-analysis with previously published data from 884 SCD samples from Tanzania ⁴⁹ and summary statistics from four African American SCD cohorts (2040 samples) ²²⁵, reaching a combined sample size of 3751. We identified new loci, in both the Cameroonian cohort, and the meta-analysis, that could be future targets of therapeutic interventions of SCD.

4.3. Results

4.3.1. The dataset

Our study included 3751 Sickle cell anemia (SCA; HbSS) patients of African ancestry from Cameroon, Tanzania, and the United States of America (USA). A total of 1250 participants from Cameroon were recruited between 2016 and 2018 as previously described ^{226,227}. Genome-wide SNP typing was performed between 2018 and 2019 using the African ancestry-specific H3Africa Illumina Omni 2.3M SNP array. Clinical and SNP genotype data for 1022 Tanzanian SCA patients generated using the Illumina Infinum Omni2.5-8 array, and that contributed to the first GWAS of HbF in Africa published nearly a decade ago ⁴⁹ was obtained from our Tanzanian collaborators and co-authors. Meta-analysed summary statistics of HbF GWAS for 2040 individuals of African ancestry recruited as part of several multi-site studies in the USA covering a cumulative period of over four decades was also obtained from our collaborators and co-authors for meta-analysis ²²⁵ (see Methods section). HbF measurement in the Cameroonian and Tanzanian cohorts was by high-performance liquid chromatography (HPLC). In the USA-based cohorts, HbF measurement was achieved through either HPLC or alkali denaturation. In the previous meta-analysis of the USA-based cohorts, Bae et al. Standardized HbF level by cubic root transformation to control for potential confounding resulting from difference in method of HbF measurement ²²⁵.

The basic demographic characteristics, clinical and hematological features, as well as the alpha-thalassemia genotypes and the *HBB* gene cluster haplotypes of participants in the Cameroonian cohort (Cameroonian patients) are presented in Table 4.1. Table 4.2 presents the basic characteristics common to the Cameroonian and Tanzanian cohorts, while Figure 4.1 presents the ratio of males and females, the distribution of HbF before and after normalization in each cohort, as well as the definition of the five classical *HBB* gene cluster haplotypes. None of the patients were treated with hydroxyurea at the time HbF was measure. We restricted our analyses to participants age five years and older and we normalized HbF level in both cohorts by cubic root transformation to match the age distributions and transformations in the USA-based studies.

Table 4.1 Clinical and demographic features of the Cameroonian cohort

Demographic characteristics			
Variable	Median [IQR]	Range	Observations (<i>n</i>)
Age (years)	16 [IQR: 10 – 24]	5 - 66	1250
Sex (female/male)	657/593	-	1250
Hematological indices			
HbF (%)	8.50 [4.7 - 13.6]	0.70 - 37.4	1250
RBC (×10 ¹² /l)	2.70 [2.29 - 3.12]	0.84 - 7.70	1235
Hb (g/l)	7.60 [6.80 - 8.50]	2.90 - 14.5	1237
MCV (fl)	87.0 [81.0 - 94.0]	26.0 - 132	1236
MCHC (g/l)	31.0 [25.9 - 33.0]	3.80 - 54.2	1233
WBC (×10 ⁹ /l)	11.2 [8.61 - 14.2]	0.40 - 49.8	1236
Lymphocytes (×10 ⁹ /l)	4.70 [3.57 - 6.21]	0.10 - 22.6	1237
Monocytes (×10 ⁹ /l)	1.00 [0.70 - 1.40]	0.00 – 9.40	1236
Platelet count (×10 ⁹ /l)	401 [310 – 498]	22 - 1185	1237
HbA2 (%)	3.20 [2.70 - 3.80]	0.00 - 18.2	1247
Clinical events			
VOC (<i>n</i> /year)	2 [1 – 4]	0 - 48	1221
Consultations (<i>n</i> /year)	1 [0 – 2]	0 - 15	445
Hospitalisation (<i>n</i> /year)	1 [1 – 2]	0 - 24	1191
Stroke #	46/1195 (3.85%)	-	1241
3·7 HBA1/HBA2 deletion genotypes (n = 405)			
αα/αα	286		
αα/α3·7	101		
α 3·7/α 3·7	18		
HBB haplotype (n = 604)			
Benin/Benin	331		
Benin/Cameroon	99		
Cameroon/Cameroon	174		

Number of patients with at least one overt stroke, IQR = interquartile range, HbF = fetal hemoglobin, RBC = red blood cell, Hb = hemoglobin, MCV = mean corpuscular volume, MCHC = mean corpuscular hemoglobin concentration, WBC = white blood cell, VOC = vaso-occlusive crisis

Table 4.2 Basic characteristics of Cameroonian and Tanzanian cohorts

	Cameroon	Tanzania
Sample size	1250 [IQR]	1022 [IQR]
Age (years)	16 [10 – 24]	11 [IQR: 7.9 - 16.3]
Sex (female/male)	657/593	539/480*
HbF (%)	8.50 [4.7 - 13.6]	4.8 [2.6 - 7.8]

IQR: Interquartile range; * Three individuals had ambiguous sex information

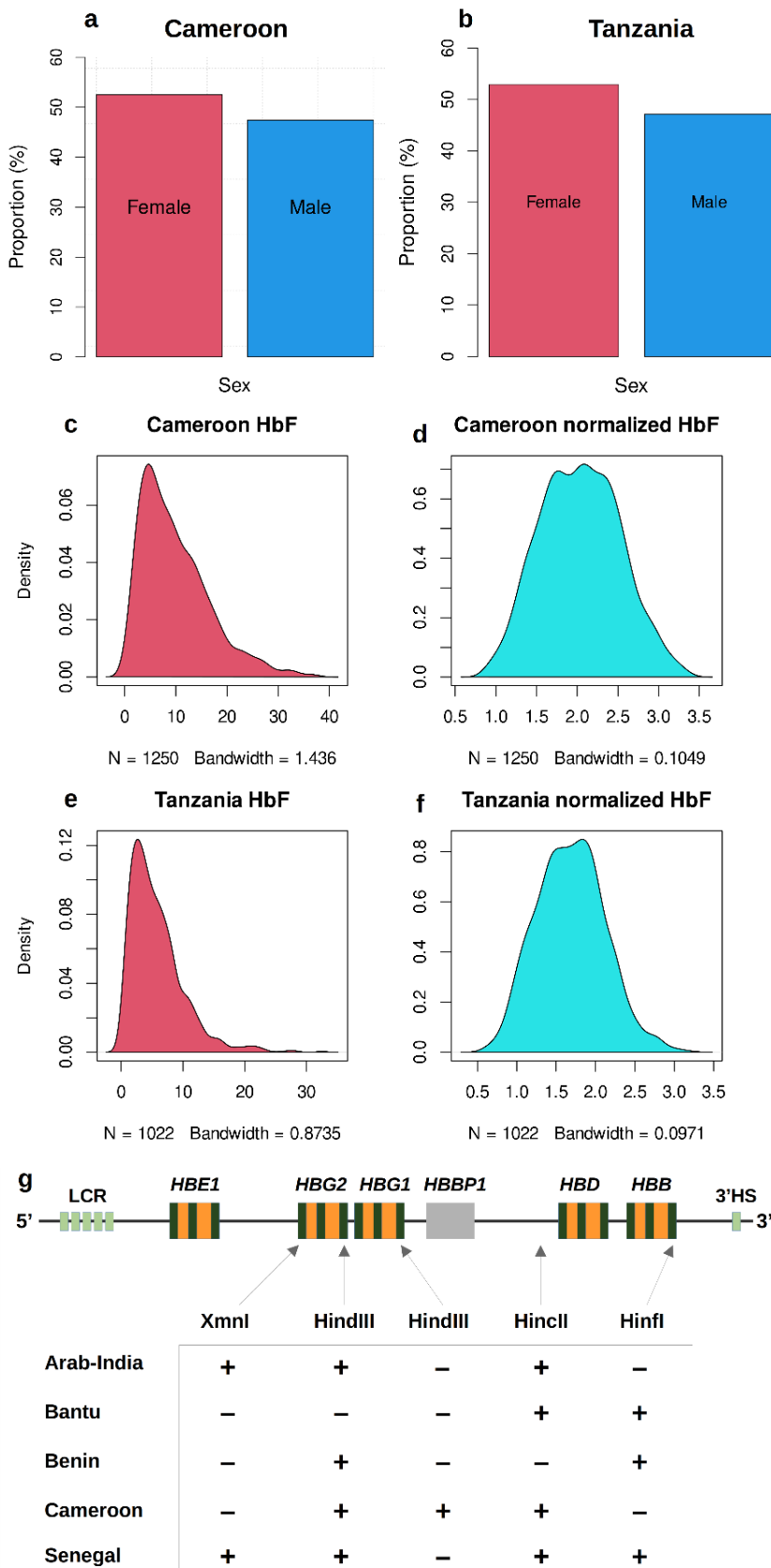


Figure 4.1 Description of basic demographic and clinical characteristics

(a&b) Sex ratio and (c-f) distribution of fetal hemoglobin before and after normalization in Cameroonian and Tanzanian sickle cell anemia cohorts. Fetal hemoglobin was normalized by a cubic root transformation. (g) Definition of five classical *HBB* gene cluster haplotypes based on the five restriction fragment length polymorphic sites used to genotype Cameroonian sickle cell anemia patients as presented in ²²⁸. LCR = locus control region, HS = DNaseI hypersensitive site.

4.3.2. Genotype data quality assessment and population structure

Accurate inferences from genome-wide association studies (GWAS) rely on accurate phenotyping, but also on rigorous quality control (QC) of the corresponding genotype data. We therefore conducted careful QC for the Cameroonian discovery and Tanzanian cohorts separately (Figure 4.2a) (See Methods section). Of the 1,250 samples in the Cameroonian cohort that passed all filters, 836 had complete HbF, sex, and age information. In the Tanzanian cohort, 942 of the 1,022 samples passed all filters and had complete HbF, sex, and age information. When we plotted the first two principal components for each of the cohorts to assess possible batch effects and population stratification, we observed expected sub-structure in Cameroon as previously described²²⁹ (Figure 4.2b). Tanzanian samples demonstrated a considerable degree of spread (heterogeneity) consistent with previously published data on the cohort⁴⁹ (Figure 4.2c).

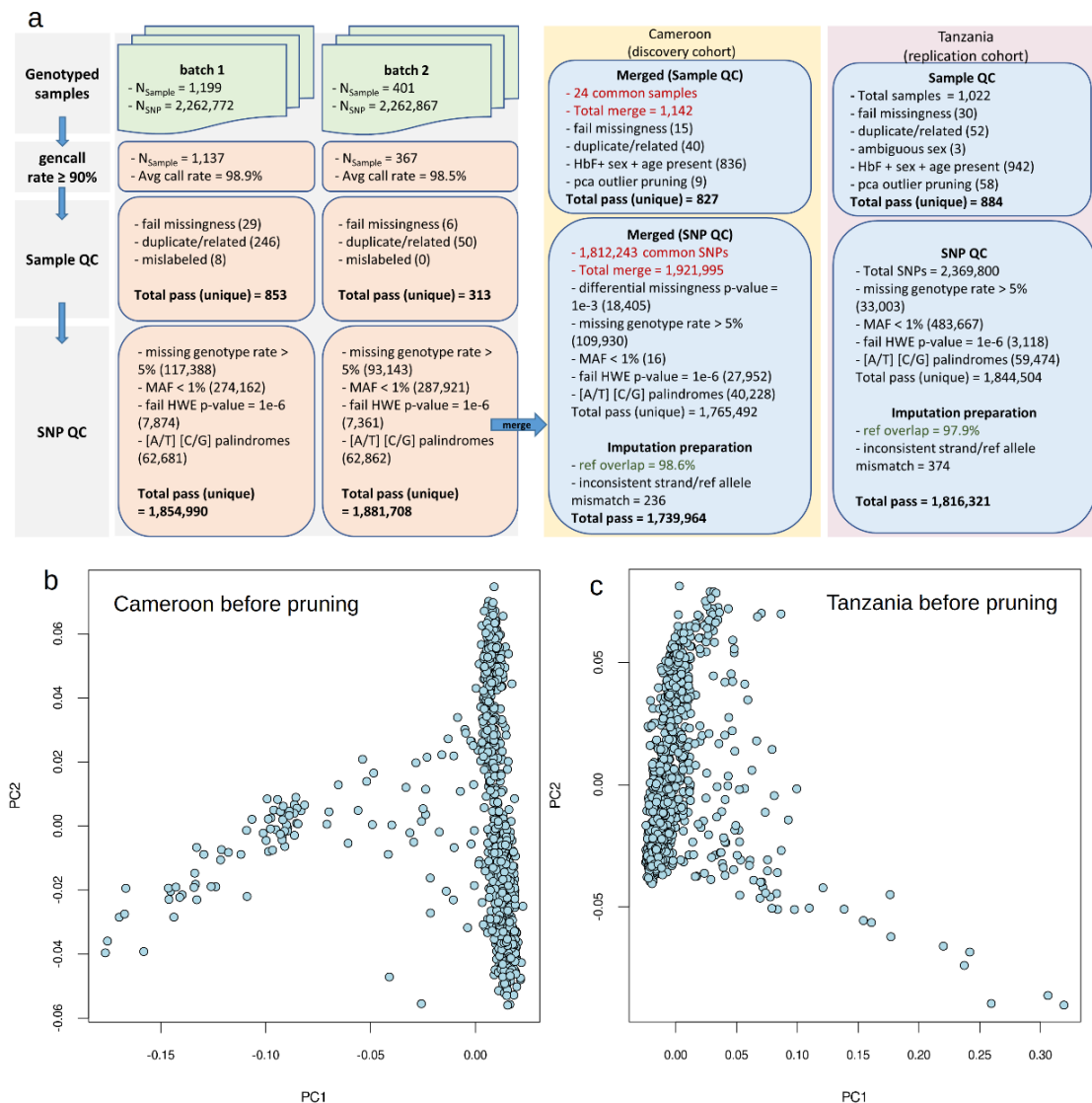


Figure 4.2 Genotype data quality control

(a) Two batches of Cameroonian samples (green boxes) that were involved, including the total number of samples and variants that were genotyped. The three stages that were involved in processing the samples are indicated as left-most grey boxes: (i) genotype calling and filtering out of samples that had genotype call rate < 90%. The total number of samples retained is indicated in the corresponding orange boxes as well as the total call rate after the first filter; (ii) per-sample quality control was next performed, and (iii) then per-SNP quality control (see “Methods” for a detailed description of the QC procedure). The two batches of Cameroonian samples were then merged after QC, and then QC was performed again for the merged set (yellow panel), as well as for the Tanzanian cohort (pink panel). The final datasets were each aligned to the reference haplotype panel (African populations in the 1000 Genomes project ²³⁰) to ascertain reference/alternate allele assignment, strand, SNP IDs, and chromosomes, in preparation for imputation. Principal component analysis before outlier pruning in the Cameroonian discovery (b) and Tanzanian (c) cohorts.

To reduce potential confounding by this variability, we performed genetic outlier pruning for the two cohorts which resulted in the exclusion of nine (9) samples from the Cameroonian cohort and fifty-eight (58) samples from the Tanzanian cohort (see Methods section). Outlier pruning had little effect in the Cameroonian cohort consistent with the observation that this is the characteristic clustering pattern regarding the three major ethnolinguistic groups in Cameroon (albeit we lacked complete ethnolinguistic information to highlight the clusters accordingly) ²²⁹ (Figure 4.3a). However, we noted a substantial “homogenizing” effect of outlier pruning in the Tanzanian cohort (Figure 4.3b). A total of 827 Cameroonian and 884 Tanzanian samples passed all QC procedures respectively.

We next performed principal component analysis (PCA) of these two populations with other African ancestry populations from the 1000 Genomes project (KGP; Figure 4.3a-b) which revealed expected clustering: the west-central clustering of the Cameroonian cohort close to Nigerian populations reflects the geographic location of Cameroon (West-Central Africa), while the Tanzanian cohort clustered with Kenyan individuals reflecting the geographic location of Tanzania (East Africa) (Figure 4.3c). Previously described genetic differentiation of the Fulani of Cameroon from the Bantu and Semi-Bantu ethnolinguistic groups was apparent ⁶⁶, while a cline from the Tanzanian population towards Cameroonians most likely reflects the well-characterized Bantu expansion. In general, these observations, along with the expected clustering of the two cohorts with other African populations on a PCA with global ancestries in the 1000 Genomes reference panel (Figure 4.3d) confirmed the integrity and quality of our data. Moreover, when we highlighted PCA of the Cameroonian cohort by the genotyping batch, we observed identical

clustering of the two batches of samples (Figure 4.3e), ascertaining the robustness of our QC procedures to effectively eliminate batch effect in downstream analyses.

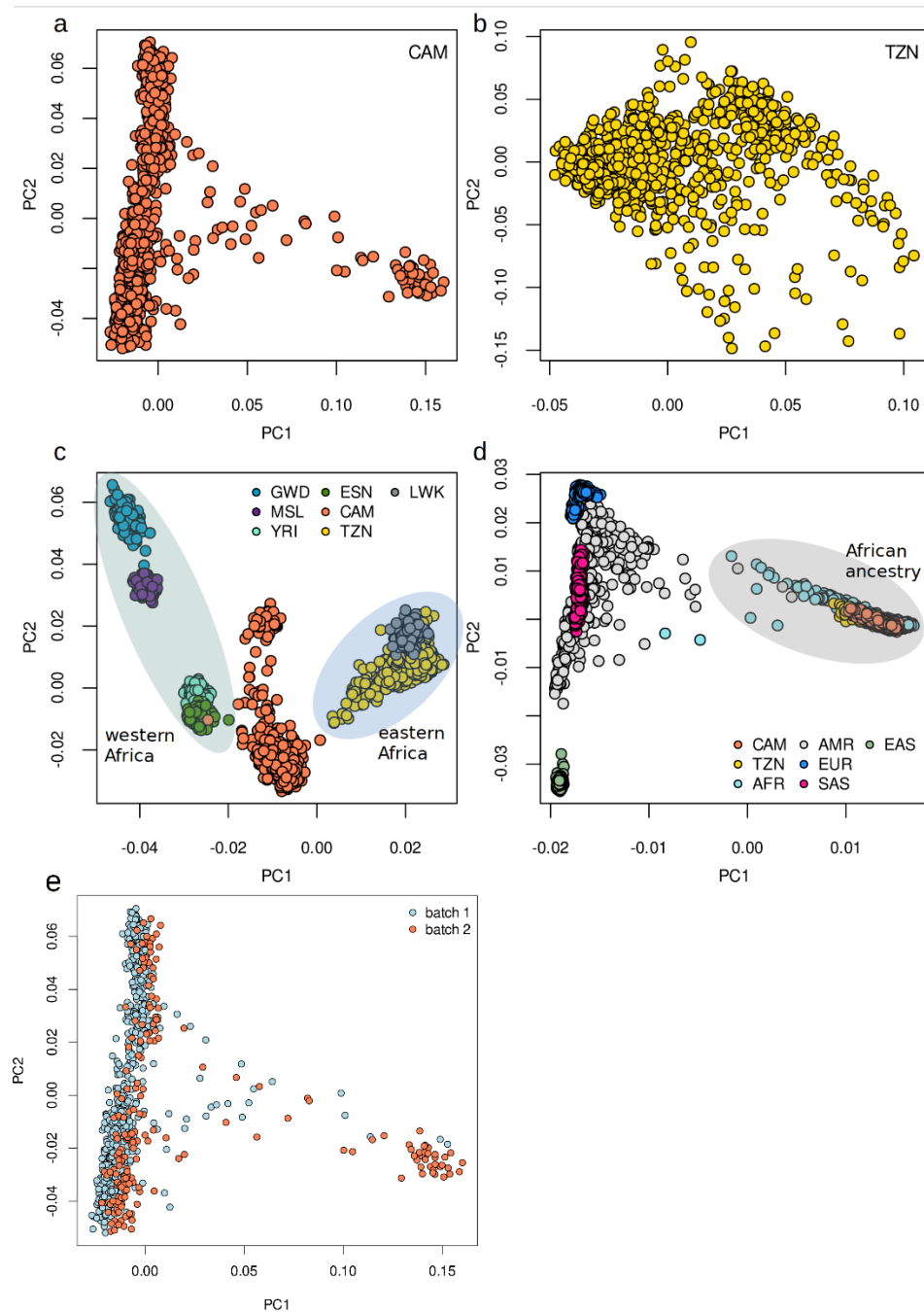


Figure 4.3 Data quality assessment and population structure

(a) The Cameroonian cohort after outlier pruning shows expected clustering, (b) Tanzanian re-analysis cohort shows “homogenizing” effect after outlier pruning, (c)

Cameroonians (CAM) clustered west-centrally towards Nigerian Esan (ESN) and Yoruba (YRI) populations, which in turn clustered with other western African populations (light green shaded area) including the Mende from Sierra Leone (MSL) and the Gambian Mandinka in the western division (GWD). Tanzanian (TZN) populations clustered eastwards with the Luhya populations from Webuye in Kenya (LWK; light blue shaded area). A cline from Tanzania towards Cameroon is consistent with the Bantu expansion. (d) CAM and TZN populations clustered with other African ancestry (AFR) populations (grey shaded area) on a PCA with global ancestries: American (AMR), European (EUR), South Asian (SAS), and East Asian (EAS). (e) Principal component assessment of batch effect in Cameroonian samples after quality control and outlier pruning.

To prepare our data for imputation, we next aligned the SNPs that passed all filters to the African populations in the 1000 Genomes reference panel in GRCh37 coordinates. This revealed 1.4% of SNPs in the Cameroonian cohort and 2.1% of SNPs in the Tanzanian cohort that were absent in the reference panel. The difference in reference allele overlap among the cohorts might be explained by the difference in genotyping chips utilized; the Cameroonian cohort was genotyped with an African ancestry-specific chip while the Tanzanian cohort was genotyped with an Illumina Omni2.3 array that is biased towards the European ancestry. Moreover, more than 90% of African populations in the 1000 Genomes reference panel are of West African origin who are more closely related to Cameroonians than to Tanzanians (Figure 4.3c).

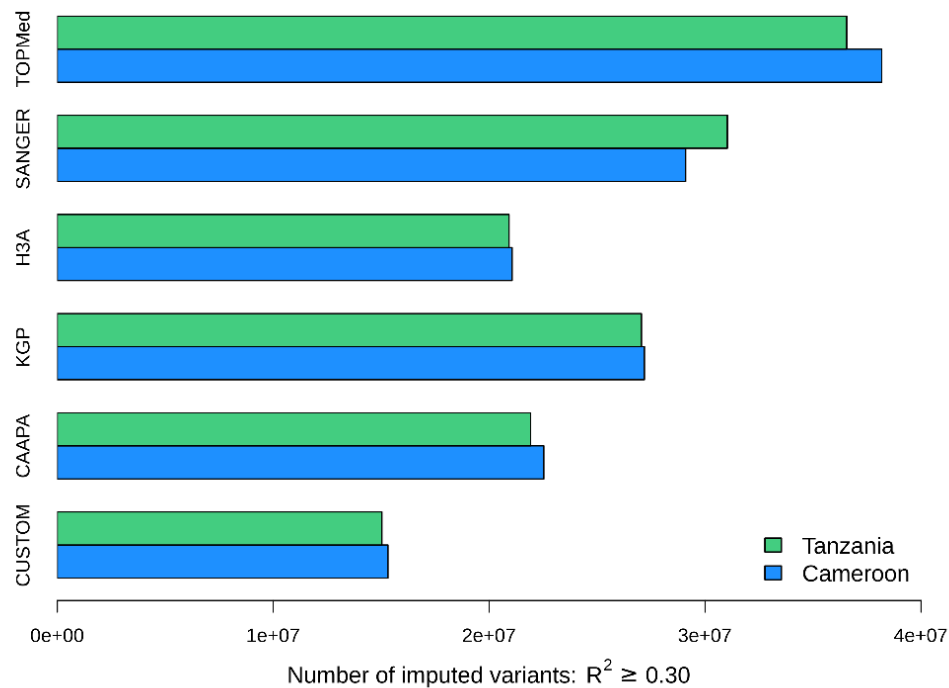
4.3.3. Comparative performance across different imputation panels

We next imputed genotypes into each cohort separately using six different imputation panels (see Methods section). Details of the panels, including size and chromosomes supported, are presented in Table 4.3. To assess imputation performance, we applied a baseline imputation accuracy threshold (R^2) of 0.3 below which variants were excluded from all the panels. Imputation from the TOPMed panel was performed in the GRCh38 coordinates. The rest of the panels were in the GRCh37 coordinates (we specifically retained our data in GRCh37 coordinates where possible to avoid any loss of variants that would have ensued from migrating between reference builds). Therefore, certain imputation statistics such as shared variants could only be compared across datasets imputed in GRCh37 coordinates. Figure 4.4 presents the number of variants imputed per panel with $R^2 \geq 0.3$.

Table 4.3 General statistics of the imputation panels utilized

<i>Panel</i>	<i>Sample Size (#AFR)</i>	<i>% African ancestry</i>	<i>Number of variants</i>	<i>Chromosomes supported</i>	<i>Imputation Service</i>
<i>Custom</i>	50	100	20,769,182	1-22, X	This study
<i>H3Africa</i>	~3,793	-	114,306,738	1-22	H3Africa
<i>TOPMed</i>	97,256 (51,110)	29	308,107,085	1-22, X	TOPMed
<i>KGP</i>	2504 (661)	26.4	49,143,605	1-22, X	Michigan
<i>CAAPA</i>	~883	100	31,163,897	1-22	Michigan
<i>AGR</i>	4,956 (~3,061)	62	93,421,145	1-22, X	Sanger

#AFR = number of samples of African ancestry; H3Africa = Human Heredity and Health in Africa; TOPMed = Trans-Omics for Precision Medicine; KGP = The 1000 Genomes Project; CAAPA = The Consortium on Asthma among African-ancestry Populations in the Americas; AGR = African Genome Resources. Links to the web servers are provided in Appendix VIII.

**Figure 4.4** Number of variants imputed per imputation panel

There was a general and expected positive correlation between the number of imputed variants and panel size (Figure 4.4). However, this trend broke down when comparing the larger H3A panel with the smaller CAAPA and KGP panels, implying that a substantial proportion of variants from

the H3A panel were imputed with very low accuracy ($R^2 < 0.3$; we note that the H3A panel is still under development and the version used in this study may represent a beta version with limited capabilities). We noted that while our custom panel and the KGP and TOPMed panels imputed both SNPs and INDELs, the H3A, AGR, and CAAPA panels imputed only SNPs. In addition, the H3A and CAAPA panels supported only autosomes (chromosomes 1-22; Table 4.3), most likely contributing to the breakdown in correlation between panel size and number of imputed variants. Panel size was also positively correlated with imputation accuracy, with the TOPMed panel outperforming the rest (Figure 4.5a-d). However, and surprisingly, our custom and the KGP panels performed better than the H3A and AGR conceivably due to the close genetic and phenotypic proximity of our custom panel to our study population, the use of trio information in the creation of the KGP panel, which is known to boost imputation performance⁶¹. Moreover, there was a surprising exclusion of related individuals in the AGR panel and this is expected to impose a cost on performance²³¹.

Comparing the mean R^2 per chromosome, our custom panel performed better than the H3A, CAAPA, and AGR in the Cameroonian cohort (Figure 4.5a). In the Tanzanian SCD cohort, the performance of the AGR was second only to the TOPMed panel, reflecting its enrichment with eastern and southern African haplotypes that more closely resemble those from Tanzania (Figure 4.5b). Interestingly, all the panels performed relatively poorly with chromosome 19; the H3A panel performing the least with mean $R^2 < 0.80$. Zooming into each chromosome by considering mean R^2 for different minor allele frequency (MAF) bins, the H3A panel demonstrated an overall better performance over our custom panel for many chromosomes, and particularly at lower MAFs (< 0.1 ; Figure 4.5c&d). Imputation accuracy was generally slightly higher for our Cameroonian than the Tanzanian cohort. This could be explained by the difference in genotyping chips utilized in the two populations, the H3A chip being more Afrocentric with tag SNPs that more accurately match African haplotypes.

When we considered shared variants across the panels at the defined R^2 threshold of 0.3 (those imputed in GRCh37 coordinates), we observed that each panel had a proportion of variants that was not present in the others (Figure 4.5e&f). In fact, the number of variants that overlapped across the panels was less than 30% in each cohort. In addition, up to 16% of variants imputed with the AGR panel in the Cameroonian cohort (19% in the Tanzanian cohort) were absent from other panels at the defined R^2 threshold. This means that the same variant can be imputed with vastly different accuracies across different imputation panels, reflecting differences in haplotype structures across the panels resulting from different tagging schemes, and underscoring the complementary use of the panels. But more importantly, it means that we can expect different patterns of association when different panels are utilized, such that when a signal is observed in one panel, its failure to be detected in another panel does not rule out its significance.

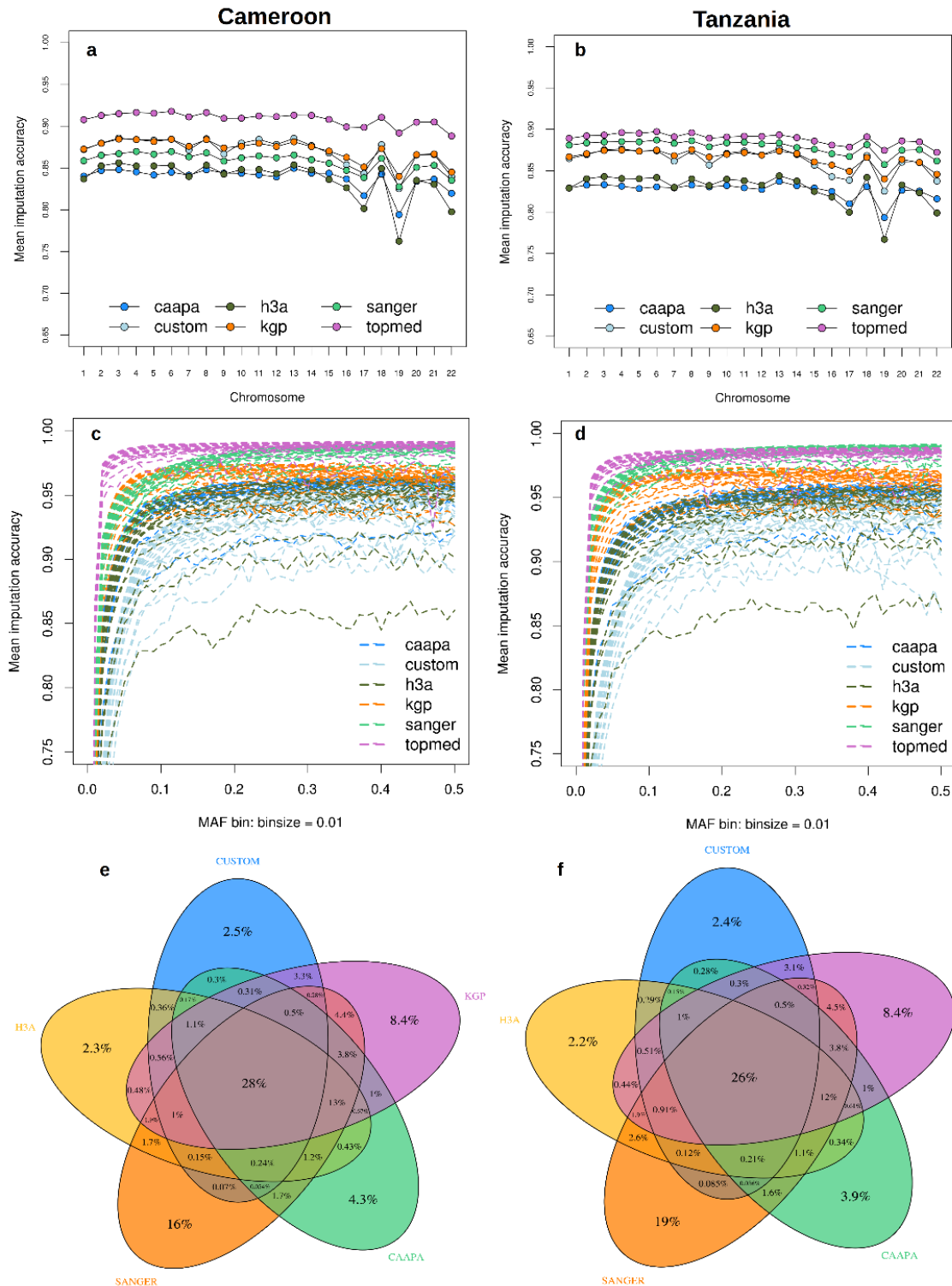


Figure 4.5 Comparative analysis of imputation panels

(a&b) Dot and line plots of R^2 per chromosome. Only autosomes (chromosome 1 - 22) were used since some of the panels (H3A and CAAPA) did not support the sex chromosomes. (c&d) Dashed-

line plots of R^2 within different minor allele frequency (MAF) bins per autosome. A bin size of 0.01 was used to bin the variants into fifty (50) bins from 0 to 0.5. (e&f) Five overlapping Venn diagrams showing proportions of shared and panel-specific variants. Only panels in GRCh37 coordinates are shown.

4.3.4. Association testing supports complementary use of multiple imputation panels

Following the above observations, we utilized datasets from all six panels for downstream association analyses in three stages (each panel dataset was analyzed separately): In **stage 1**, we performed single variant association tests for Cameroonian discovery and Tanzanian re-analysis cohort imputed datasets filtered to include only biallelic SNVs (SNPs and INDELs) with $R^2 \geq 0.6$ (see Methods section); **stage 2** involved a meta-analysis of two African-based cohorts (hereafter CAM-TZN meta-analysis), while in **stage 3**, we included summary statistics from HbF GWAS involving sickle cell anemia cohorts of African ancestry based in the United States of America to perform an overall or a global meta-analysis (Figure 4.6). The total number of variants analyzed per panel for each dataset (hereafter referred to as *analysis unit*) is presented in Table 4.4. Genome-wide significance was defined by $P < 5E-08$, while variants for which the Benjamini-Hochberg false discovery rate (FDR) was less than 0.05 and $P > 5E-08$ were considered moderately significant, and strongly suggestive associations were considered at FDR between 0.05 and 0.10 or $P < 5E-06$. A strong evidence of association was also inferred when a locus had moderately significant signals in an association test and in a meta-analysis.

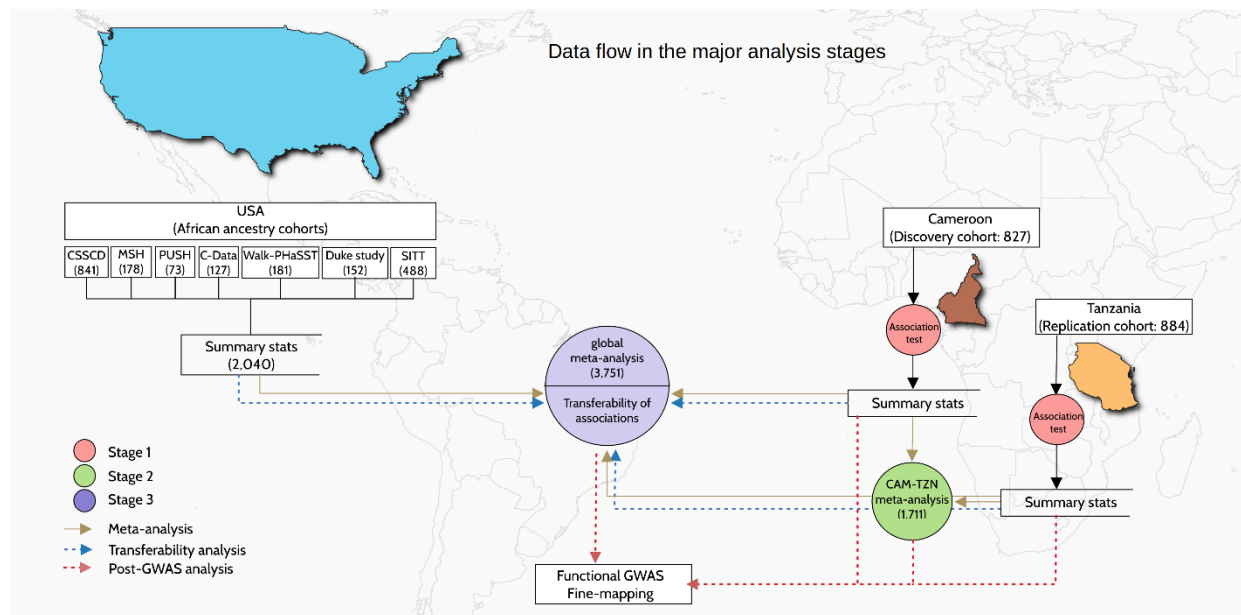


Figure 4.6 Data flow in the major analysis stages

Single variant association tests for each imputed dataset of the Cameroonian and Tanzanian cohorts separately was performed in Stage 1 (red circle), Summary statistics from the two cohorts were then used to perform meta-analyses in Stage 2 (green circle), while Summary statistics from the two cohorts were combined with summary statistics from sickle cell anemia cohorts based in the United States of America (USA) to perform an overall/global meta-analysis in Stage 3 (purple circle). The summary statistics from all three stages of association analysis were then used to perform functional GWAS and fine-mapping.

Table 4.4 Number of variants analyzed per panel for each dataset

<i>panel</i>	dataset			
	Cameroon (cm)	Tanzania (tz)	cm_tz_metal	global_metal
<i>custom</i>	15,397,530	15,298,435	14,639,103	1,031,534
<i>h3a</i>	13,421,490	13,657,802	12,393,754	1,050,271
<i>agr</i>	13,675,045	14,033,241	12,703,630	1,053,829
<i>caapa</i>	13,009,851	11,526,627	10,393,439	740,999
<i>kgp</i>	16,420,379	13,078,253	11,778,220	756,985
<i>topmed</i>	15,433,801	15,800,308	14,285,219	1,069,243
	Stage 1		Stage 2	Stage 3

h3a = H3Africa, *agr* = African Genome Resource, *caapa* = , *kgp* = 1000 Genomes Project, *topmed* = Trans-Omics for Precision Medicine

For meta-analysis, the number of variants represents only sites that were common to the Cameroonian and Tanzanian data sets (cm_tz_metal) or common to the Cameroonian, Tanzanian and USA-based cohorts (global meta_metal). The stage at which each variants analyzed per panel per dataset or *analysis unit* was analyzed is indicated in the last row of the table as colored cells. The colors correspond to the color codes in Figure 4.6 and Figure 4.7a. The numbers represent the total variants that were successfully tested for association and this was determined from the resulting summary statistics.

Figure 4.7a presents an overview of the loci for which we observed significant associations, with comparative performance of the panels with respect to number of loci identified. As anticipated, different patterns of association were observed across the imputation panels with multiple panel-specific signals. In the Cameroonian cohort, the Custom, Sanger (AGR) and TOPMed panels proved the most powerful, while the AGR panel was the most powerful for the Tanzanian cohort. Significantly more loci were observed in the Tanzanian cohort with eight (8) potentially novel loci., suggesting a marked improvement of recently developed imputation panels to capture genetic variations in African populations. By our definitions above, two strongly significant (*FLT1* and

OPCML) novel loci and twelve moderately significant (*ZNF804A*, *MRPS22*, *PRLR*, *FRMD1*, *DMRT1*, *SERPINA6*, *PLCB1*, *SLC4A4*, *LOC574080*, *LOC107984174*, *GFRA1*, *LINC01898*) potentially novel loci were observed. Taking the results of the two cohorts cumulatively, the H3A panel performed the least. However, when the meta-analysis results were also considered, it was the KGP panel that demonstrated the least performance, followed by the CAAPA panel. The AGR and Custom panels had the best overall performances respectively in terms of number of novel signals identified. Notably, two of the sixteen loci observed were the well-characterized *BCL11A* and *HBS1L-MYB*.

The quantile-quantile (Q-Q) plots of the expected against the observed p-values, as well as the genomic control inflation factors are presented in Figure 4.7b. Manhattan plots showing the significant and suggestive signals are presented in Figure 4.7c and Figure 4.8. Tables 4.5 and 4.6 present a summary of the most significant variants in each locus for each analysis unit after stepwise conditional analyses.

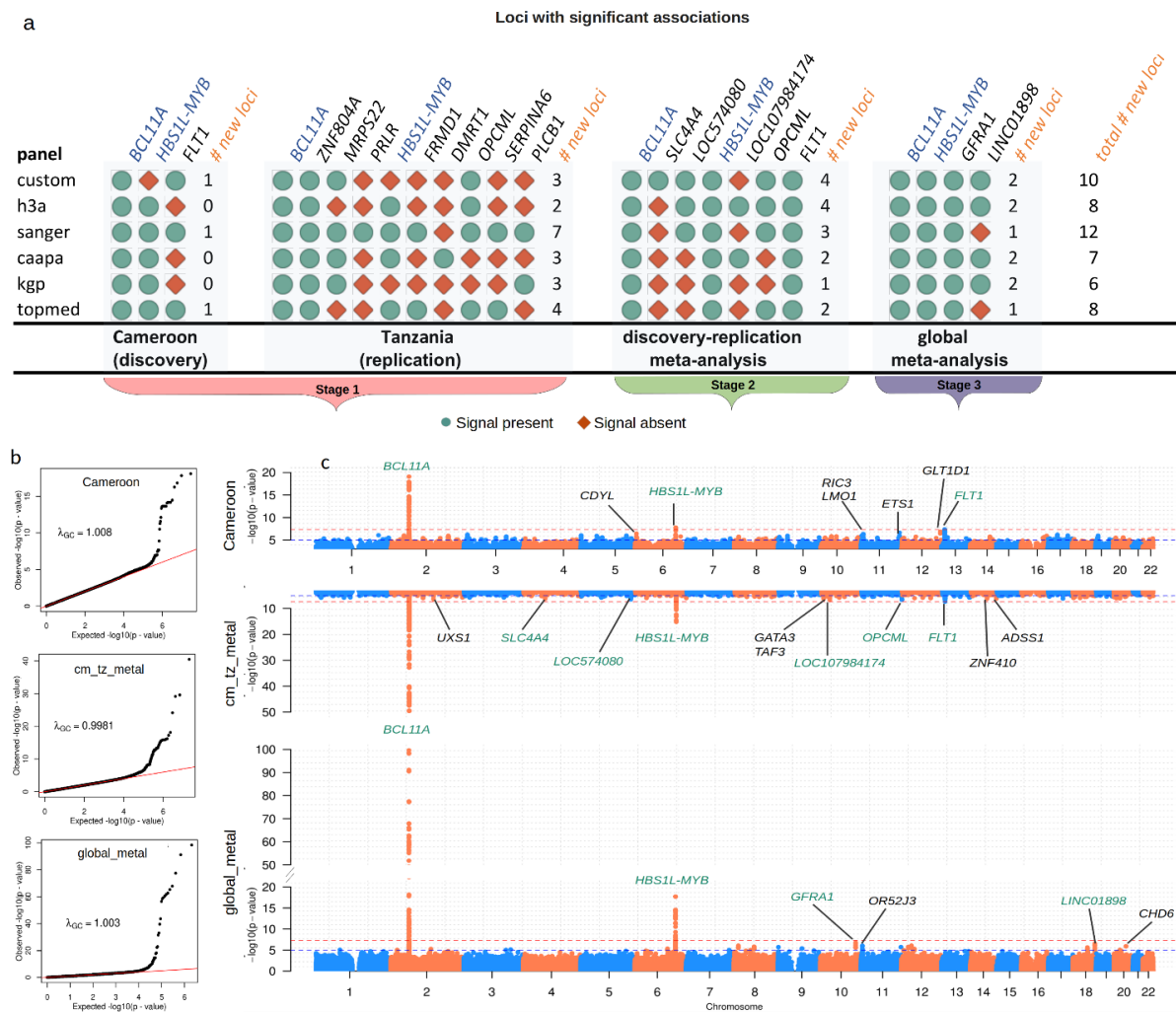


Figure 4.7 Evidence of association

(a) Overview of significant loci detected per analysis unit. Green circles indicate presence while red diamond indicates absence of significant signal in the corresponding locus. Blue colored loci represent the major known HbF-influencing loci, while the rest in black are novel loci. (b) The quantile-quantile (Q-Q) plots of the expected against the observed p-values, as well as the genomic control inflation factors demonstrate the robustness of our association results and show that our test statistics were not inflated. (c) Manhattan plots showing the significant signals (green) and suggestive signals (black) in the Cameroonian discovery, Cameroon-Tanzania meta-analysis, and global meta-analyses. The plots were generated using a combination of summary statistics in GRCh37 genomic coordinate (CUSTOM, H3A, AGR, CAAPA, and KGP panels). Manhattan plots generated from TOPMed results are shown in Figure 4.8 below.

Table 4.5 Summary of significant signal in independent association tests

The list is sorted by coordinate in ascending order.

Rsid	Coordinate (hg19)	Alleles	MAF	Beta	SE	P	FDR	PVE	Nearest gene	Panel	R ²
<i>Cameroon discovery</i>											
<i>rs7606173</i>	chr2:60725451	G/C	0.45	-0.22	0.024	8.25E-20	1.13E-12	0.091	<i>BCL11A</i>	AGR	0.98
<i>rs9399137</i>	chr6:135419018	T/C	0.03	0.38	0.067	1.76E-08	1.24E-02	0.037	<i>HBS1L-MYB</i>	H3A	0.97
<i>rs115695442</i>	chr13:29080292	C/T	0.1	0.21	0.038	3.35E-08	1.99E-02	0.036	<i>FLT1</i>	TOPMED	0.99
<i>Tanzania reanalysis</i>											
<i>rs1427407</i>	chr2:60718043	T/G	0.22	0.3	0.024	4.26E-36	2.91E-29	0.151	<i>BCL11A</i>	H3A	1
<i>rs193286999</i>	chr2:185797285	C/T	0.01	0.46	0.086	8.97E-08	1.56E-02	0.031	<i>ZNF804A</i>	KGP	0.94
<i>rs2307030</i>	chr3:139066061	A/G	0.14	-0.15	0.028	1.86E-07	2.19E-02	0.03	<i>MRPS22</i>	CUSTOM	0.82
<i>rs111464043</i>	chr5:35252204	A/G	0.02	-0.39	0.077	3.14E-07	3.32E-02	0.029	<i>PRLR</i>	AGR	0.91
<i>rs55634702</i>	chr6:135418633	T/TTAC	0.03	0.35	0.057	1.13E-09	4.00E-04	0.04	<i>HBS1L-MYB</i>	TOPMED	0.94
<i>rs6914083</i>	chr6:168579466	C/T	0.23	0.12	0.023	2.64E-07	2.96E-02	0.029	<i>FRMD1</i>	AGR	0.93
<i>rs202189241</i>	chr9:956578	C/A	0.03	0.27	0.054	3.37E-07	3.81E-02	0.029	<i>DMRT1</i>	CAAPA	0.78
<i>rs6590706</i>	chr11:133334906	G/A	0.18	-0.13	0.025	1.79E-07	2.35E-02	0.03	<i>OPCML</i>	H3A	0.97
<i>rs143333989</i>	chr14:94783372	T/A	0.02	0.37	0.067	5.69E-08	8.60E-03	0.032	<i>SERPINA6</i>	TOPMED	0.95
<i>rs189534069</i>	chr20:8870372	C/T	0.03	-0.29	0.058	4.97E-07	4.77E-02	0.028	<i>PLCB1</i>	AGR	0.95

Rsid = reference SNP identifier, *MAF* = minor allele frequency, *Beta* = effect size, *SE* = standard error, *P* = unadjusted p-value, *FDR* = Benjamini-Hochberg false discovery rate, *PVE* = proportion of variance explained, *R²* = imputation accuracy

Table 4.6 Summary of significant signals in meta-analyses

The list is sorted by coordinate in ascending order

Rsid	Coordinate (hg19)	Alleles	MAF	Beta	SE	P	FDR	PVE	Nearest gene	Panel	Direction	HetPval	HetISq
<i>Cameroon-Tanzania meta-analysis</i>													
rs1427407	chr2:60718043	T/G	0.24	0.27	0.018	2.48E-50	3.64E-43	0.116	<i>BCL11A</i>	H3A	++	0.05	73.9
rs116237841	chr4:72226303	T/G	0.03	-0.25	0.051	5.75E-07	4.96E-02	0.014	<i>SLC4A4</i>	CUSTOM	--	0.893	0
rs1373981	chr5:165266584	A/G	0.1	-0.13	0.025	3.09E-07	3.15E-02	0.015	<i>LOC574080</i>	H3A	--	0.293	9.4
rs55634702	chr6:135418633	T/TTAC	0.03	0.36	0.044	2.32E-16	1.01E-10	0.038	<i>HBS1L-MYB</i>	TOPMED	++	0.761	0
rs12240332	chr10:29314402	A/G	0.08	0.14	0.027	2.11E-07	2.68E-02	0.015	<i>LOC107984174</i>	CAAPA	++	0.647	0
rs6590705	chr11:133334522	A/C	0.19	0.1	0.019	2.35E-07	2.26E-02	0.015	<i>OPCML</i>	CUSTOM	++	0.069	69.8
rs74617914	chr13:29084891	T/C	0.05	-0.2	0.036	4.38E-08	6.50E-03	0.017	<i>FLT1</i>	CAAPA	--	0.303	5.6
<i>Global meta-analysis</i>													
rs766432	chr2:60719970	A/C	0.28	-0.24	0.011	2.42E-100	3.57E-93	0.107	<i>BCL11A</i>	H3A	---	0.112	54.2
rs9399137	chr6:135419018	T/C	0.05	-0.27	0.025	4.37E-29	3.36E-23	0.032	<i>HBS1L-MYB</i>	TOPMED	---	0.058	64.9
rs3781532	chr10:117895604	A/G	0.27	0.06	0.012	1.87E-07	1.70E-02	0.007	<i>GFRA1</i>	CUSTOM	+++	0.975	0
rs7237840	chr18:73485764	T/C	0.1	0.09	0.018	2.98E-07	2.52E-02	0.007	<i>LINC01898</i>	CUSTOM	+++	0.234	31.1

Rsid = reference SNP identifier, *MAF* = minor allele frequency, *Beta* = effect size, *SE* = standard error, *P* = unadjusted p-value, *FDR* = Benjamini-Hochberg false discovery rate, *PVE* = proportion of variance explained, *HetPval* = Heterogeneity p-value, *HetISq* = Heterogeneity I^2 statistic indicating the amount of effect size difference that is due to differences between the populations involved in the meta-analysis.

4.3.5. Replication of the major HbF-influencing loci: *BCL11A* and *HBS1L-MYB*

Across all analyses, *BCL11A* and *HBS1L-MYB* were the most significant loci (Tables 4.5&4.6), and establishing that they are the largest contributors to HbF variability in these cohorts. rs1427407 and rs9399137 are the most widely and frequently reported sentinel variants in *BCL11A* and *HBS1L-MYB* respectively, and in fact multi-ancestry fine-mapping has suggested rs1427407 as the likely functionally relevant variant within the *BCL11A* locus⁴³. Yet, the sentinel variants in these loci sometimes differ from one cohort to another. We observed in our study that the strongest associations within the loci differed among the cohorts. For instance, in the Cameroonian cohort, it was rs7606173 that emerged as the *BCL11A* sentinel variant ($P = 8.25E-20$) apparently because it was almost twice as frequent as rs1427407 ($MAF_{rs7606173} = 0.45$; $MAF_{rs1427407} = 0.26$) since the two variants had similar effects ($\beta_{rs7606173} = -0.22$; $\beta_{rs1427407} = -0.23$). rs7606173 therefore contributed the largest proportion (9.1%) in HbF variability, and together, the two variants in *BCL11A* accounted for 8.85% in HbF variance in the Cameroonian cohort. In the Tanzanian cohort, rs1896294 and rs1427407 were the *BCL11A* sentinel variants with similar significance ($P = 4.26E-$

36), and rs1896294 was even slightly more significant due to a moderately higher frequency ($MAF_{rs1896294} = 0.28$; $MAF_{rs1427407} = 0.22$) although rs1427407 had a rather larger effect ($\beta_{rs1896294} = 0.28$; $\beta_{rs1427407} = 0.30$). Individually and together, the variants contributed ~15.2% in HbF variability in the Tanzanian cohort similar to previous reports⁴⁰, and substantially higher than the variance explained by *BCL11A* sentinel variants in the Cameroonian cohort. In a meta-analysis of the two cohorts, rs1427407 was indeed the sentinel *BCL11A* variant with the largest effect on HbF level ($P = 2.48E-50$, $\beta = 0.27$), and it contributed ~11.6% in HbF variability in the joint cohort. In the global meta-analysis, however, rs766432, another widely reported association emerged as the *BCL11A* sentinel variant ($P = 2.42E-100$), contributing 10.7% in HbF variability in the combined cohorts. Yet, rs1427407 still had the largest effect ($\beta_{rs766432} = -0.24$; $\beta_{rs1427407} = 0.26$), supporting the attribution of functional relevance to it within the *BCL11A* locus⁴³, although it was slightly less frequent ($MAF_{rs766432} = 0.28$; $MAF_{rs1427407} = 0.24$). rs1427407 therefore contributed a smaller proportion (7.6%) in HbF variability.

In the *HBS1L-MYB* intergenic region, rs9399137 and rs35786788 were the sentinel variants in the Cameroonian cohort (cm) having identical test statistics ($P = 1.76E-08$, $\beta = 0.38$). In the Tanzanian cohort (tz), as well as in the Cameroon-Tanzania meta-analysis, the rs55634702 INDEL was the sentinel variant ($P_{tz} = 1.13E-09$, $\beta_{tz} = 0.35$; $P_{cm_tz_metal} = 2.32E-16$, $\beta_{cm_tz_metal} = 0.36$). Generally, these *HBS1L-MYB* sentinel variants explained ~4% of HbF variance, consistent with previous findings. rs9399137 was the sentinel variant in the global meta-analysis and contributed 3.2% in HbF variability. The relatively low proportion of HbF variance explained by the *HBS1L-MYB* variants notwithstanding their relatively large effects is owed to their low frequencies in African ancestries ($MAF \leq 0.03$) compared with other ancestries where they generally occur at high frequencies; greater than 10%.

4.3.6. Replication in other known and recent genetically nominated loci

A notable absence from our list of significant loci was the chromosome 11 *HBG2* signal which completes the triad of major and well-characterized HbF-influencing loci. Marked by the rs7482144 SNP (Xmnl-*HBG2*)⁴⁵, the signal is mostly observed in non-African ancestries and admixed populations carrying the Senegal and Arab-India *HBB* gene cluster haplotype, which are absent in the two cohorts herein analyzed. In the global meta-analysis of our study, we only observed replication signals ($P < 5E-05$) in a region 2kb upstream of the *HBG2* promoter that is only active in erythroid-specific K562 cells (as a strong enhancer) and human embryonic stem cells (hESCs; as a weak enhancer). These included suggestive associations such as rs5010979 and rs2855122 that occurred in multiple transcription factor binding sites. Multiple replication signals with $P < 5E-04$ were also observed in the *HBG2* –2kb enhancer region in the Cameroonian cohort and CAM-TZN meta-analysis. Interestingly, this region was recently confirmed to interact with the *HBG2* promoter both in wild type and *BCL11A* knockout erythroid-specific HUDEP-2 cells²²¹,

suggesting a potentially different and underappreciated mechanism of HbF regulation at the *HBG2* locus in African ancestries. An additional replication signal (rs3898917) that occurred in GATA1 peaks upstream of the *HBB* locus control region (LCR) exclusively in K562 erythroid lineage was observed in the Cameroonian Cameroonian cohort only, and remarkably in all the panels. No *HBB* region signals were observed in the Tanzanian cohort.

Other genetic regulators of HbF level have recently been identified through CRISPR screening, CUT&RUN analysis, and RNA interference. Using transcription factor (TF) CRISPR screening, for instance, Divya et al. identified *ZNF410* as an HbF repressor that acts by promoting the expression of CHD4, a key member of the nucleosome remodeling and deacetylating complex (NuRD) which is required for HbF repression⁵⁵. We looked up variants at nominal significance ($P < 0.05$) within 100kb downstream and 100kb upstream of *ZNF410* given that some cis-regulatory elements could be tens of thousands of bases away. We observed multiple suggestive variants at $P < 5e-06$ in the CAM-TZN meta-analysis including rs145213285 at $P = 6e-07$ ($\beta = -0.20$, MAF = 0.035). Multiple replication signals at $P < 5e-04$ were observed in the global meta-analysis, as well as in the separate discovery and Tanzanian cohorts with, interestingly, similar effect sizes as *BCL11A* variants (β 0.22 - 0.26), but at however lower frequencies (MAF = 0.04), hence contributing a small fraction in HbF variance (1.5 - 2.1%). More recently in their HbF GWAS study, Cato et al have described a novel genetically nominated HbF-influencing locus on chromosome 6, *BACH2*, whose product might potentially compete with prominent erythropoiesis regulators like *NFE2*⁴³. Similar to *ZNF410*, we observed replication variants ($P < 5e-04$) that might potentially tag *BACH2* in all but the Cameroonian cohort. In an RNA interference experiment, Chokdee et al., demonstrated that downregulation of *ZNF802* (*JAZF1*) in thalassemia patients significantly restored HbF levels to greater than therapeutic limit (up to 35%)⁵⁶. We investigated potential replication of this locus in our analysis and also observed some support for replication in the Cameroonian discovery and Tanzanian Tanzanian cohorts ($P < 5e-04$) with also appreciable effects (generally greater than 0.13), especially in the Cameroonian cohort (β 0.18 - 0.23; MAF = 0.06; pve 1.5 - 1.9%). In general, some of the replicated variants are relatively common, have appreciable effect sizes, and some are intronic, and may thus be further explored for functional relevance.

4.3.7. Identification of novel HbF-associated loci

In the Cameroonian Cameroonian cohort, a third significant signal that mapped to a novel locus upstream of the FMS related receptor tyrosine kinase 1 gene (*FLT1*, also known as vascular endothelial growth factor receptor 1 - *VEGFR1*) on chromosome 13 (13q12.3) was observed. The sentinel variant, rs115695442 ($P = 4.18E-08$, $\beta = 0.21$) (Figure 4.9a), was relatively common in the cohort (MAF = 0.10), and it contributed 3.5% in HbF variance. Notably, in the Cameroonian cohort, *FLT1* significant variants were relatively common than *HBS1L-MYB* ones (*FLT1*; MAF = 0.076 - 0.105 Vs *HBS1L-MYB*; MAF < 0.04) and had similar effects as *BCL11A* associations ($\beta = 0.20 - 0.23$).

Although no significant *FLT1* associations were observed in the Tanzanian cohort, we observed multiple variants at p-value < 9E-03 and with appreciable effects ($\beta = 0.14 - 0.16$), and generally low frequencies (MAF < 0.05). Similarly, variants within the genomic region of *FLT1* were observed in the Cooperative Study of Sickle Cell Disease (CSSCD) cohort at $P = 6.8E-03$ (rs61763174, intronic variant, $\beta = -0.24$, MAF = 0.06)²³², and in the Silent Cerebral Infarct Transfusion Trial (SITT) cohort at $P = 2E-04$ (rs9578046, 94kb upstream, $\beta = 0.14$, MAF = 0.12)²³³.

Upon meta-analysis of the Cameroonian and Tanzanian cohorts, five novel loci that included *FLT1* and *OPCML* were identified, of which *FLT1* had the strongest associations with HbF levels (Table 4.6). Interestingly, rs74617914 emerged as the *FLT1* meta-analysis sentinel variant ($P = 4.38E-08$, $\beta = 0.20$) although it did not reach genome-wide significance in the independent association tests of the Cameroonian and Tanzanian cohorts. Two additional novel loci were uncovered upon a global meta-analysis, including *GFRA1* and *LINC01898*. Importantly, we uncovered seven potentially novel locus in the Tanzanian re-analysis cohort (Figure 4.8) of which only *OPCLM* was previously reported almost a decade ago in this same cohort albeit it did not reach genome-wide significance⁴⁹, hence classified herein as novel. Each of the novel loci contributed ~3% in HbF variance. *ZNF804A* variants demonstrated the largest effect on HbF level across our entire analysis ($P = 8.97E-08$, $\beta = 0.46$) and the relatively small proportion in HbF variance (3.1%) that they contributed was due to their very low frequencies in the population (MAF < 0.02). Indeed, the variants are very rare in Africans generally, whilst absent in other ancestries. In the Tanzanian cohort, the derived alleles were only observed in heterozygotes (as would be expected for rare variants) and were associated with higher HbF levels (Figure 4.9).

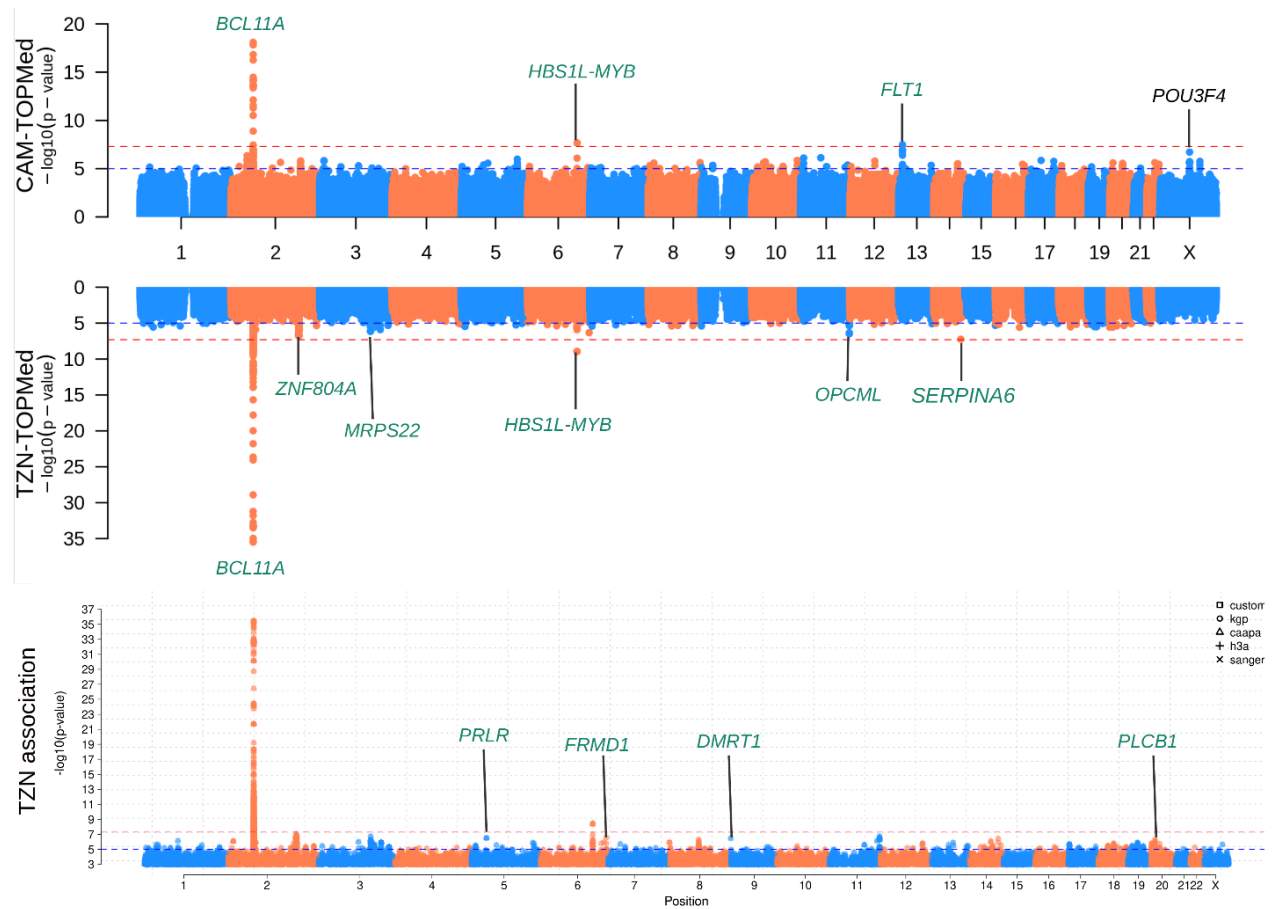


Figure 4.8 Significant and suggestive loci in the TOPMed panel and Tanzania reanalysis
Green loci are significant while black is suggestive

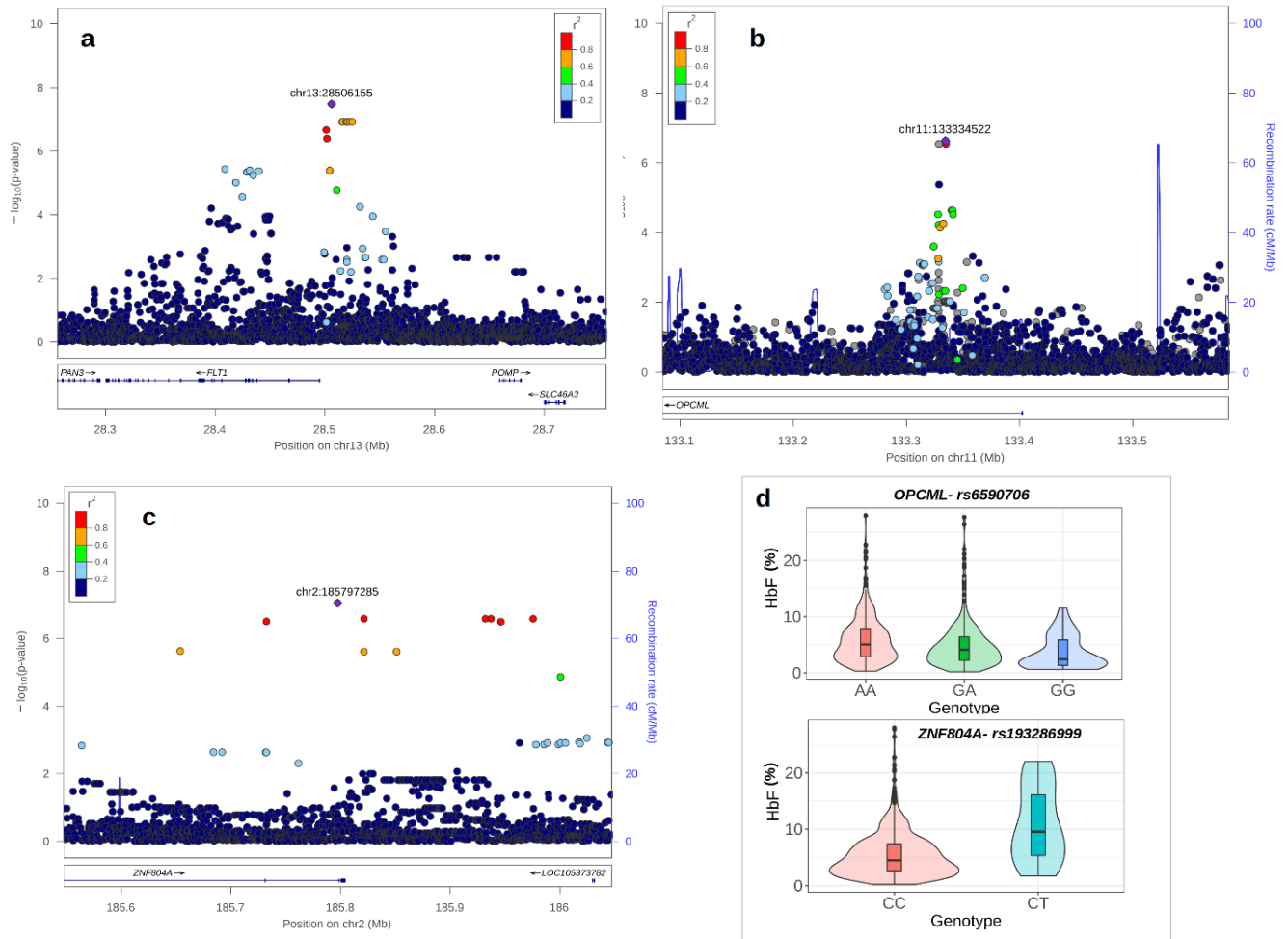


Figure 4.9 Regional association and boxplots of Cameroon and Tanzania exclusive signals

(a) Cameroon *FLT1* signal regional association plot showing partial to strong linkage disequilibrium of the sentinel variant, rs115695442, with haploblocks 5 and 7 variants respectively (see Fig.5 in main text). (b) *OPCML* and (c) *ZNF804A* regional association plots in the Tanzania cohort. (d) Boxplots of sentinel variants in *OPCML* and *ZNF804A* showing fetal hemoglobin (HbF) distribution among individuals carrying the different genotypes of the variants. The derived allele of *OPCML*-rs6590706-G was associated with lower HbF levels, while the derived allele of *ZNF80A*-rs193286999-T was associated with higher HbF levels in an additive manner. The rare nature of rs193286999 means that the derived alleles are only seen in heterozygotes. Generally, the *ZNF804* signal extended over a large genomic distance (~2Mb) downstream covering many other genes (*ZC3H15*, *ITGAV*, *FAM171B*, *ZSWIM2*, *FSIP2*) with very strong LD (greater than 0.6).

Of the novel loci, *FLT1* was particularly interesting in that it was identified in a population that has not been previously investigated genome-wide, was replicated in meta-analysis with

consistent signals across all the imputation panels, and harbored the third strongest signal after *BCL11A* and *HBS1L-MYB*. We therefore focused on the *FLT1* signal for further functional evidence.

4.3.8. Functional mapping of the *FLT1* signal

Fine mapping of the functionally relevant *FLT1* variants in the Cameroonian cohort revealed a single 95% credible set that included nine variants of which rs115695442 had the highest posterior inclusion (causal) probability (PIP = 0.36) (Figure 4.10a). In the meta-analysis, a single 95% credible set that included only rs74617914 with causal probability of 0.99 was revealed. Interestingly, six of the variants with identical test statistics (effect size, standard error, p-values, MAFs) and in perfect linkage disequilibrium (LD = 1) with each other (as well as with rs74617914) also had identical causal probabilities (PIP = 0.089) which summed to greater than 50% (Figure 4.10b). These variants additionally had larger effect sizes ($\beta = 0.23$) than the *FLT1* sentinel variants rs115695442 and rs74617914, suggesting that the most probable causal variant(s) might be among the six. In general, *FLT1* associations were mapped to a 40kb interval (chr13:29069272–29110372; GRCh37) spanning the *FLT1* promoter and a candidate enhancer region of ~30kb upstream of the *FLT1* transcription start site (TSS) (Figure 4.10c). Most of the variants occurred within multiple transcription factor (TF) binding sites (TFBSs), including five of the six variants that showed perfect LD with each other (Figure 4.10b&c). We noted that none of the variants was reported in GTEx upon a search for expression or splicing quantitative trait loci (eQTL/sQTL), consistent with the fact that they are virtually absent in non-African ancestries that make up the bulk of data in this database (Figure 4.10d).

transcription factor binding site (TFBS) predicted by the JASPAR algorithm, as well as the distance of the variants from the *FLT1* transcription start site (dTSS). (c) Mapping the variants to their genomic locations upstream of *FLT1* (using the UCSC Genome Browser, Human hg19 reference sequence) revealed their occurrence in potentially functional regions in different cell lineages as predicted by the ChromHMM algorithm based on chromatin immuno-precipitation sequencing (ChIP-seq) data ²³⁴. Two regions with DNase I hypersensitivity sites (HSs) are indicated. Hypoxia response elements (HREs) bound by the hypoxia-inducible factors (HIFs), HIF1A and HIF2A, are notable in the HS1 and HS2 and consistent with the fact *FLT1* is a target gene of the HIFs. Three of the TFs whose motifs were disrupted, and which play critical roles in erythropoiesis are indicated in bold. Importantly, K562-specific CTCF peaks in both HSs suggests interaction of the two regions via a loop formation in erythroid cells, while K562-HS2-specific MYC, IKZF1, and GATA2 peaks strongly support an erythroid-specific activity of the HS2 region. (d) Minor allele frequency (MAF) distribution of the *FLT1* fine-mapped variants and other variants looked up in the Tanzanian, CSSCD, and SITT cohorts. The MAFs are displayed for Cameroonian and Tanzanian sickle cell anemia populations, as well as unascertained global ancestries. One of the six fine-mapped variants that had identical statistics is used to represent the group of variants.

4.3.9. Functional effects of *FLT1* fine-mapped variants

The binding motifs of six TFs whose transcriptional activity is in the same direction as *FLT1* transcription were disrupted. Interestingly, three of the TFs implicated—STAT5A, GFI1, and MXI1—appear to play crucial roles in hematopoiesis/erythropoiesis ^{235–238} and their binding motifs were disrupted by three of the six variants in perfect LD (rs11840478, rs75294023, and rs11843606 respectively) (Figure 4.11a-c), thus supporting the attribution of functional relevance to these. To exert their effect, the TFs must be able to access these binding sites. Chromatin immunoprecipitation sequencing (ChIP-seq) analysis of chromatin accessibility revealed two DNase I hypersensitive sites (HS) within the 40kb region (Figure 4.10c). HS1 corresponded to the *FLT1* promoter which showed activity only in human umbilical vein endothelial cells (HUVECs; strong) and human embryonic cells (hESCs; weak), while HS2 had methylation and acetylation patterns marking an active promoter or strong enhancer exclusively in erythroid-specific K562 cells, HUVECs, and lymphoid-specific GM12878 cells. Remarkably, strong TF ChIP-seq peaks (for GATA2, MYC, IKZF1, and CTCF) were present at HS2 in K562 cells only, suggesting a restricted activity of this region in the erythro-lymphopoietic system, potentially involving loop formation (evidenced by CTCF binding sites in HS1 and HS2) ²³⁹. There were strong HDAC2 peaks at HS1 and HS2 in hESCs, as well as polycomb-repressive complex (PRC) marks in most cell lines including K562, consistent with predicted polycomb repression of the promoter (HS1) in K562 cells. These HS1 chromatin marks are consistent with bivalent promoters which are associated with developmentally regulated genes ²⁴⁰, and the general chromatin accessibility pattern in the 40kb

region suggests a tight cell-type and stage-specific regulation of *FLT1* along the developmental axis ²⁴¹.

Interestingly, there is not much experimental data on HS2 as an enhancer evidenced by its absence in the ENCODE project, the VISTA and FANTOM5 enhancer browsers, hence no predicted interaction by GeneHancer. However, data in the ENSEMBL resource suggest some experimental evidence, while ENCODE classifies the region as “distal enhancer-like”, hence our classification as “candidate enhancer”. These data supports a hypothesis that HS2 is only transiently active, leading to a brief upregulation of *FLT1*, another key feature of developmentally regulated genes. Our fine-mapped variants were sandwiched between the HS1 and HS2 region, flanking the promoter and the candidate enhancer (Figure 4.10c), and indeed, ATAC-seq peaks from three datasets of erythropoietic lineages revealed the variants to be enhancer-associated (Figure 4.11d). The variants generally exhibited additive effects on HbF level with the genotypes carrying the minor alleles associated with higher HbF levels (Figure 4.12).

4.3.10. Association of *FLT1* fine-mapped variants with other blood traits

We further tested association of the fine-mapped *FLT1* variants with other blood traits listed in Table 1 in our Cameroonian cohort. We observed significant associations of rs74617914, as well as the six other variants that were in perfect LD, with mean corpuscular volume (MCV; $P = 0.03$) (see Methods section). The derived alleles of the variants were associated with a slight increase in red blood cell (RBCs) size (i.e. MCV; Figure 4.11a-c). In addition, rs74617914 alone was significantly associated with RBC count ($P = 0.03$), although no marked difference was apparent upon observing the distribution of RBC count amongst the different genotypes.

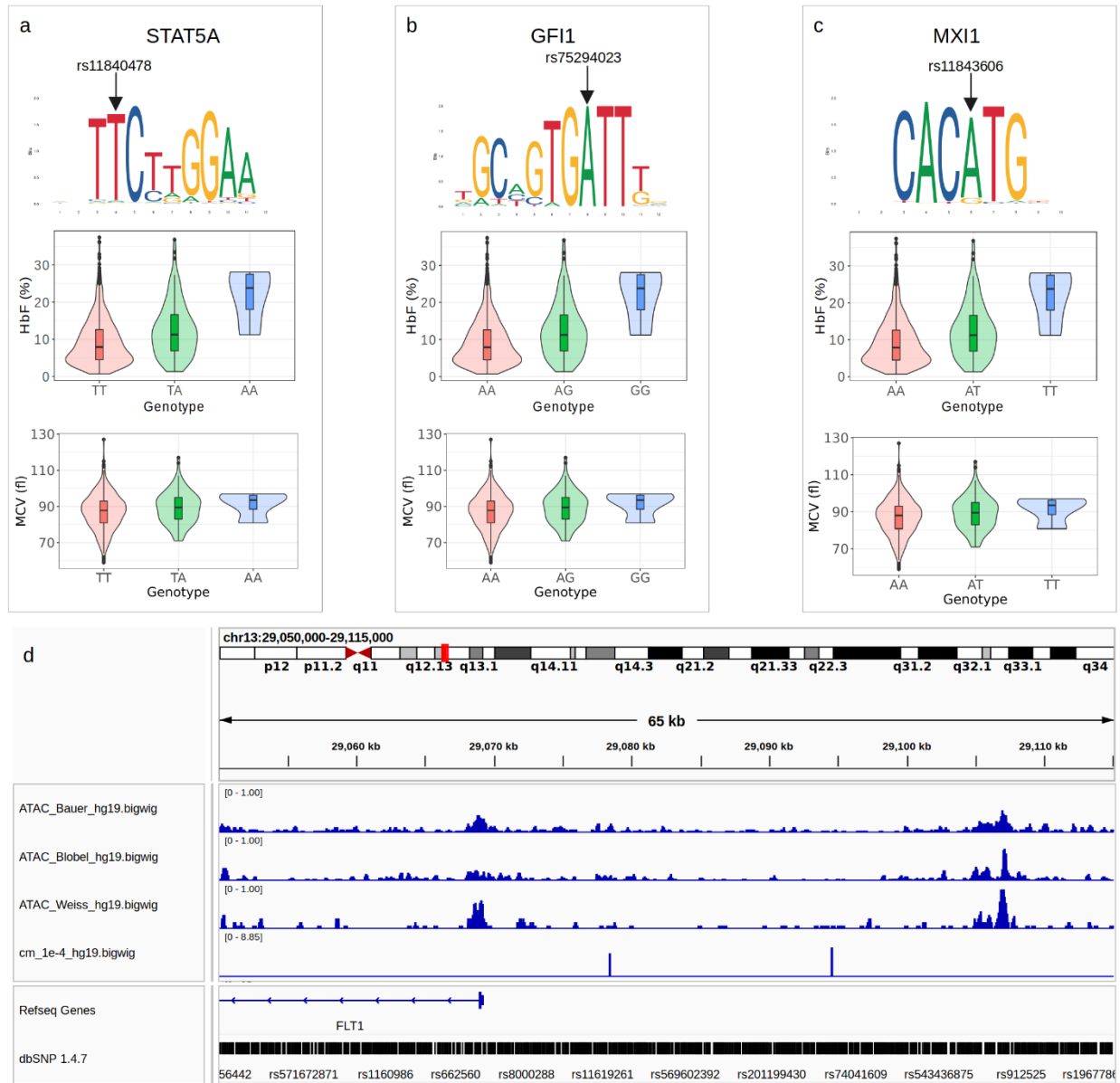


Figure 4.11 Functional effects of *FLT1* fine-mapped variants

In (a-c), the sequence logo of the binding motifs of the three TFs that were disrupted, as well as box plots showing the effects of the variants on fetal hemoglobin (HbF) level and mean corpuscular volume (MCV) are shown. The sequence logos show very high conservation of the bases affected, and the box plots show additive effects on HbF and MCV. (h) represents ATAC-seq data from 3 datasets of erythropoietic cell lines provided visual overlap, showing the *FLT1* signal to be enhancer-associated. *BCL11A* signal is used here as control for both Cameroon and Tanzania GWAS.

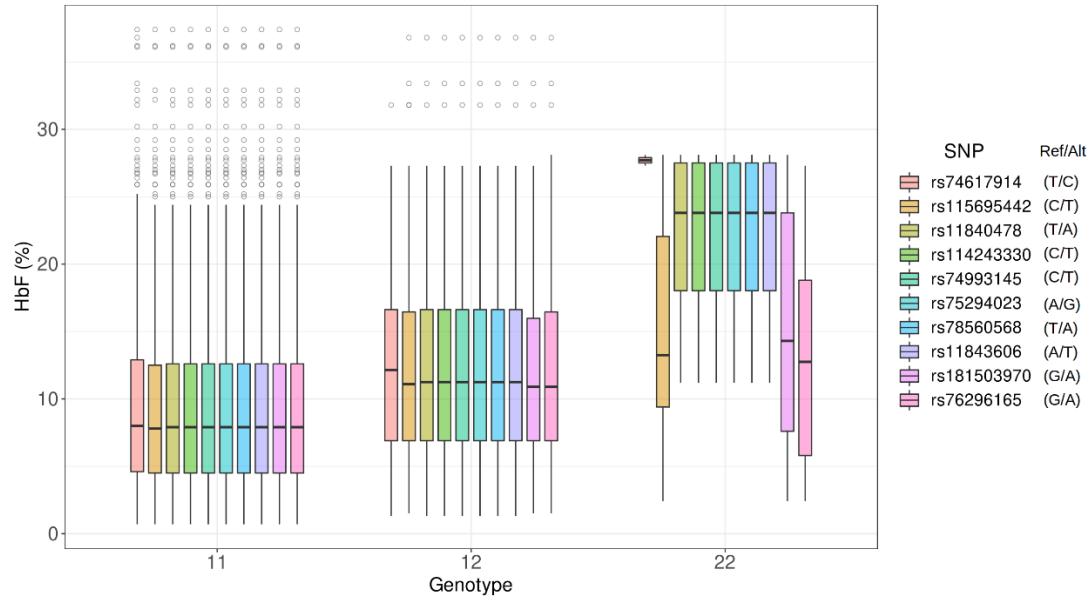


Figure 4.12 Boxplots of all *FLT1* fine-mapped variants

1 = reference (Ref) allele (major), 2 = alternate (Alt) allele (minor)

4.3.11. Haplotype structure reveals reason for absence of *FLT1* association in the Tanzanian cohort.

The difference in *FLT1* sentinel variants observed in the Cameroonian association and Cameroon-Tanzania meta-analysis (Figure 4.13a), as well as significant heterogeneity in effect sizes observed at all the fine-mapped variants ($P < 0.01$) with the exception of rs74617914 ($P > 0.29$) (Figure 4.13b), suggest different haplotype structures within the *FLT1* 40kb region between Cameroonians and Tanzanians. To investigate this, we analyzed haplotype blocks (haploblocks) within the region, including 25kb upstream and downstream (see Methods). At similar SNP densities, we observed higher and longer range LD, as well as lower haplotype diversity in Cameroonians than in Tanzanians. As suspected, there were substantial differences in haplotype structure amongst the populations. In Cameroonians, the fine-mapped variants were distributed across three haploblocks (blocks 5, 6, and 7) (Figure 4.13c); rs181503970 and rs76296165 that flanked the *FLT1* promoter occupied block 5, rs115695442 had no haploblock participation (it occurred between blocks 6 and 7), while the rest of the perfect-LD SNPs occupied block 7 which flanked the candidate enhancer. Intriguingly, all the haploblocks were in strong LD, evidenced by high D' values ($D' > 0.96$), indicative of little historical recombination; although rs115695442 had no haploblock participation, it demonstrated strong LD (> 0.8) with block 5 variants and partial to strong LD (> 0.6) with block 7 variants (see Figure 4.10a), making it an excellent tag for this locus in Cameroonians. In Tanzanians, the haplotype structure was quite different: (i) the fine-mapped variants were distributed within five haploblocks (blocks 8-12), and (ii) low D' values were

observed among the haploblocks ($D' < 0.90$), indicating evidence of high historical recombination (consistent with a recombination hotspot within the large haploblock), and meaning that the variants are largely evolving independently in this cohort. Indeed, LD between rs115695442 and all the haploblocks was less than 0.2 in Tanzanians, even though the D' between haploblocks 11 and 12 (where rs115695442 occurred) was quite high (0.95). There was high intra-block (short range) LD between rs181503970 and rs76296165 ($LD = 1$) and between rs11840478 and rs114243330 ($LD = 0.99$), all pairs of variants that are very close to each other.

To gain a better understanding of the evolution of this genomic region, we computed haploblocks for 25 non-SCD Cameroonian individuals, as well as for HbS-negative genomes from populations in the 1000 Genomes Project. We observed substantially lower LD for all the populations with multiple small haploblocks in African populations. Against a minor allele frequency threshold of 5%, we noted that the fine-mapped variants were present only in African ancestries, albeit with no consistent pattern in their haploblock participation. In fact, many of the variants had no haploblock participation at all. In addition, there was generally high historical recombination among the haploblocks as would be expected under neutral evolution. These suggest that some evolutionary force, such as natural selection, might be preserving haplotypes in the *FLT1* 40kb region in Cameroonian SCA populations.

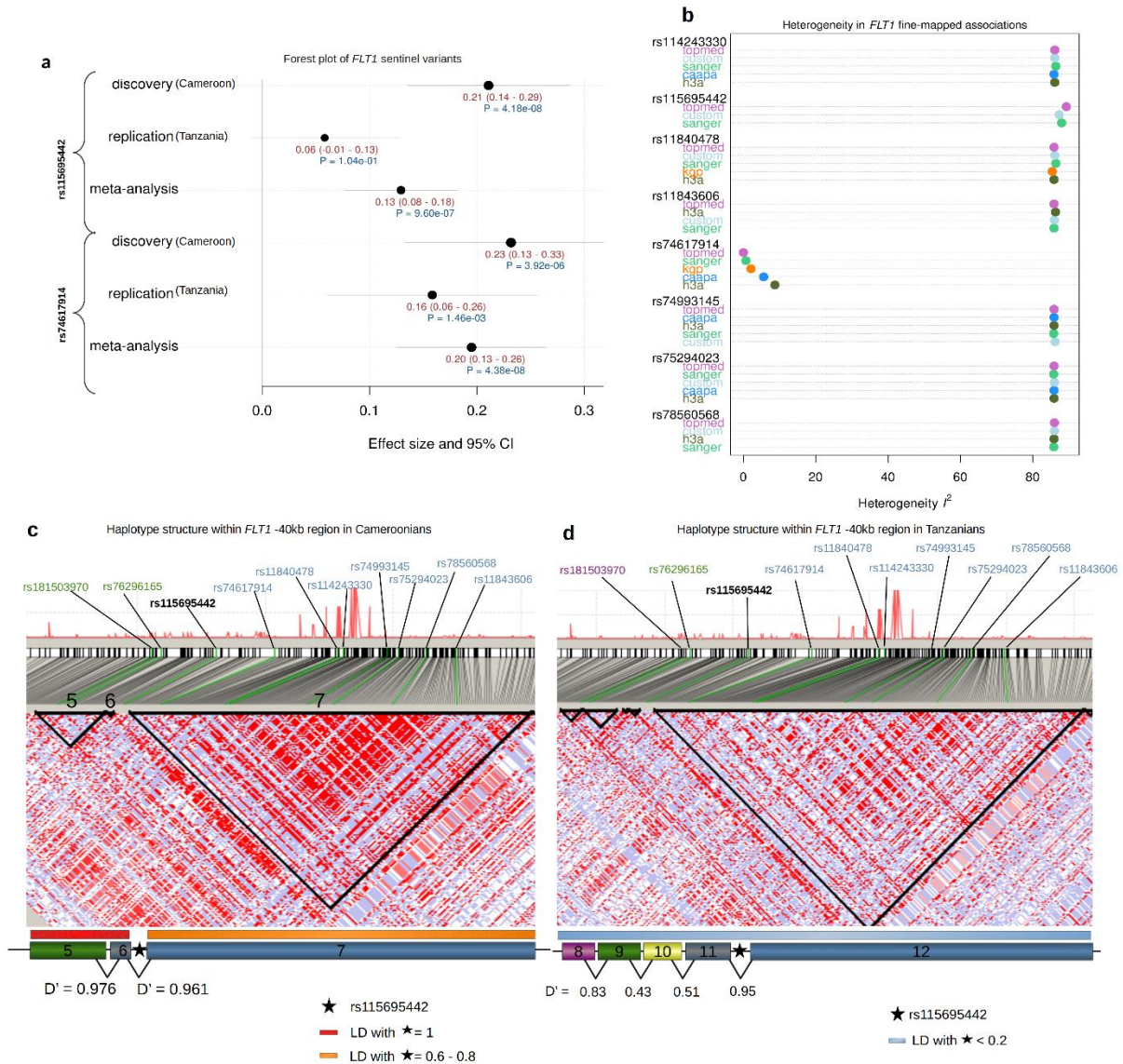


Figure 4.13 Genetic heterogeneity and haplotype substructure within the *FLT1* -40kb regulatory region

(a) Forest plot of *FLT1* lead associations in the Cameroonian cohort (rs115695442) and Cameroon-Tanzania meta-analysis (rs74617914). rs115695442 was only significant in Cameroonians and suggestive in the meta-analysis. rs74617914 was only significant in the meta-analysis and insignificant in the respective cohorts. (b) Heterogeneity plot of *FLT1* fine-mapped variants shows significant heterogeneity (greater than 80%) at all the variants except rs74617914, indicating complex haplotype structure within the region, that was confirmed by haplotype analysis (c & d): three haploblocks (black triangles) in Cameroonians (c) harboring the fine-mapped variants were in strong linkage disequilibrium (LD) evidenced by high D' values, explaining the strong LD (> 0.6) between rs115695442 and the blocks 5 and 7 variants (red and orange bars); Low LD among

haploblocks harboring the fine-mapped variants in Tanzanians (d) explained low LD (< 0.2) between rs115695442 and all other fine-mapped variants. There was generally higher LD in Cameroonians than in Tanzanians (shown by intensity of red color within the region). The probability of historical recombination between blocks is shown as D' (the higher the value, the lower the probability). A recombination map of the region, generated by haploview from the hapmap recombination map ²⁴² in GRCh37 coordinate, is shown as red line-plot above the haplotype map. A recombination hotspot is evidenced by the tallest peak.

4.3.12. Haplotype association

To see if any specific haplotypes were associated with HbF level, we performed a haplotype association test using haploview (see Methods). While no significant haplotypes were observed at a nominal p-value of 0.05 after 100,000 permutations, a strongly suggestive association was observed for a haplotype 11 carrying the derived alleles of two SNPs, rs7989474-A and rs1967786-T in Cameroonians ($P = 0.053$). The two SNPs also flanked the candidate enhancer upstream and occurred within GATA1 peaks in a region that is associated with weak enhancer and weak transcriptional activity. There were no significant haplotype associations in the Tanzanian cohort.

Similar substantial heterogeneity in effect sizes was observed within the major HbF-influencing loci, as well as some of the novel loci (Figure 4.14), thus explaining the difference in sentinel variants observed in different populations.

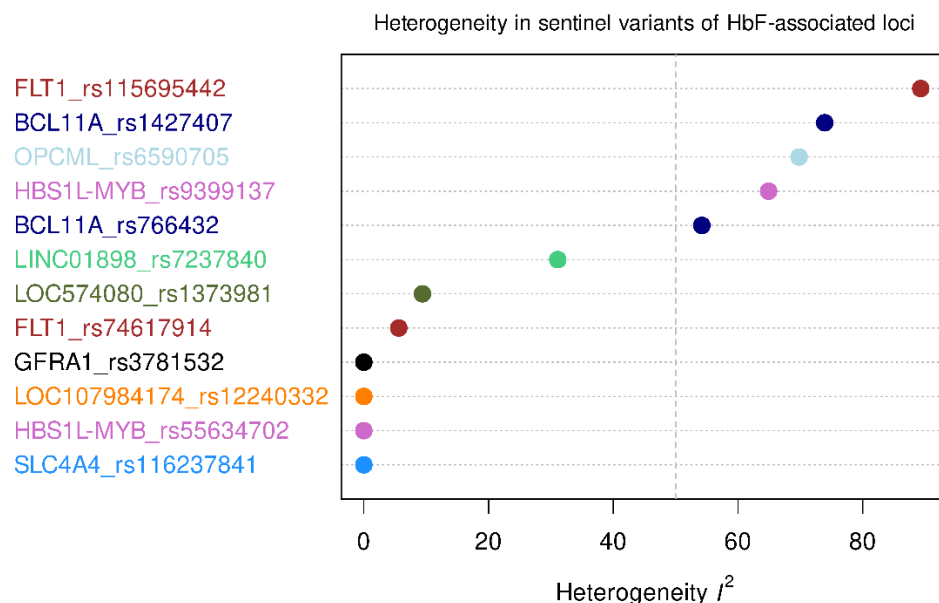


Figure 4.14 Heterogeneity plot of sentinel variants of all the significant loci

Substantial heterogeneity for some of the major known loci (*BCL11A* and *HBS1L-MYB*; greater than 50%) which could be explained by differences in haplotype structure similar to *FLT1*.

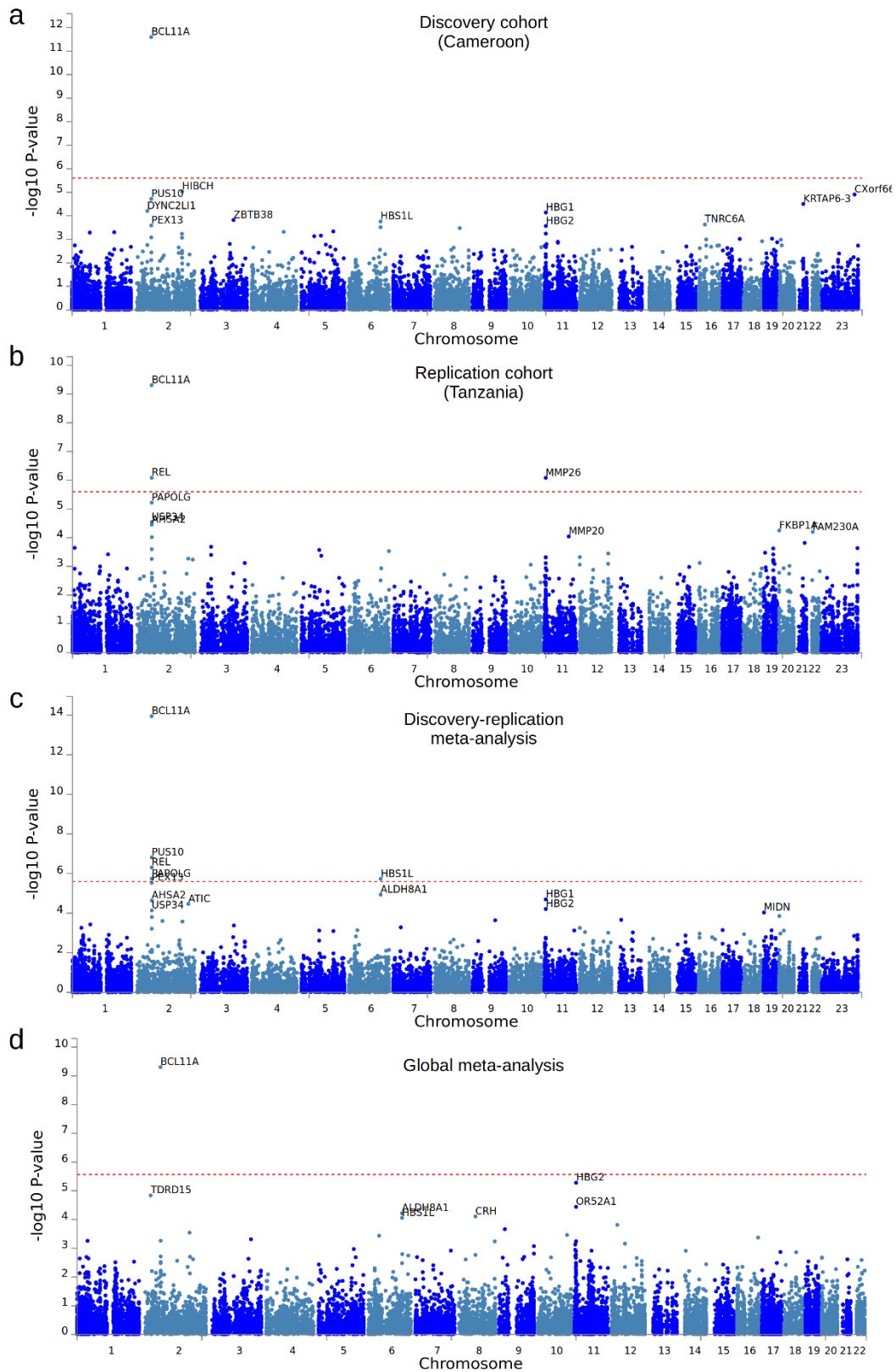
4.3.13. Gene-based, gene set, and heritability analyses further support the association results

Gene-based analysis revealed multiple genomic regions with significant ($P < 2e-06$) or strongly suggestive ($P < 2e-04$) evidence of association across all the datasets (p-value threshold $2.5e-06$; see Methods section; Figure 4.15). *BCL11A* and *HBS1L-MYB* were the most significant loci, while evidence of *HBG2* association was more apparent in this gene-based test (with strong signals in all but the Tanzanian dataset), The signal spanned *HBG1*, *HBG2*, *HBE1*, and the *HBB* gene cluster locus control region, and includes a large number of variants (greater than 3000). In the Tanzanian cohort, a novel significant locus, *MMP26*, that mapped just downstream of the *HBB*-3'HS1 on chromosome 11 was observed ($P = 8.32e-07$; Figure 4.15b). In addition, the *FLT1* locus demonstrated signs of replication in the Cameroonian cohort only ($P = 0.005$).

Figure 4.15 Significant loci gene-based association tests

The red horizontal line indicates genome-wide significance (p-value = $2.5e-06$) determined by dividing 0.05 times the total number of genes analyzed (i.e. $0.05/19867$) (**Next Page**)

Gene-base association analysis



While only a few loci overlapped between the datasets at a p-value threshold of $2e-04$, there were multiple overlaps at less stringent p-values (< 0.002), including all three major HbF-influencing loci (*BCL11A*, *HBS1L-MYB*, and *HBG2*). It was therefore unsurprising the almost identical gene set enrichment pattern across the datasets; the most enriched pathway being hematopoietic stem cell differentiation (Figure 4.16a). Indeed, there was an overwhelming enrichment of factors in the hematopoietic system (particularly erythropoietic), with some of the most significantly enriched pathways involving GATA1, KLF1, cMYB, RUNX1, STAT5A, HIF1A, and HDAC ($P_{\text{adjusted}} < 0.05$). The hypoxia pathway was also significantly enriched in all the datasets. Myeloid cell differentiation and gas (oxygen) transport were the most significantly enriched biologically processes ($P_{\text{adjusted}} < 2e-3$), while hemoglobin complex was the most significantly enriched cellular component ($P_{\text{adjusted}} < 9.60e-6$; Figure 4.16b&c). Unsurprisingly, RBC traits, including MCV and RBC count, were among the most significantly enriched phenotypes (adjusted $P < 6e-08$) (Figure 4.16d). Erectile dysfunction (priapism), an important sub-phenotype of SCD, was the most significantly enriched trait in the Tanzanian cohort and the CAM-TZN meta-analysis. Consistent with these observations, the blood and spleen were the sites with the most significant differentially upregulated genes in all the datasets, and mostly erythropoiesis related genes including *GATA1*, *KLF1*, and *HBG1* (Figure 4.17).

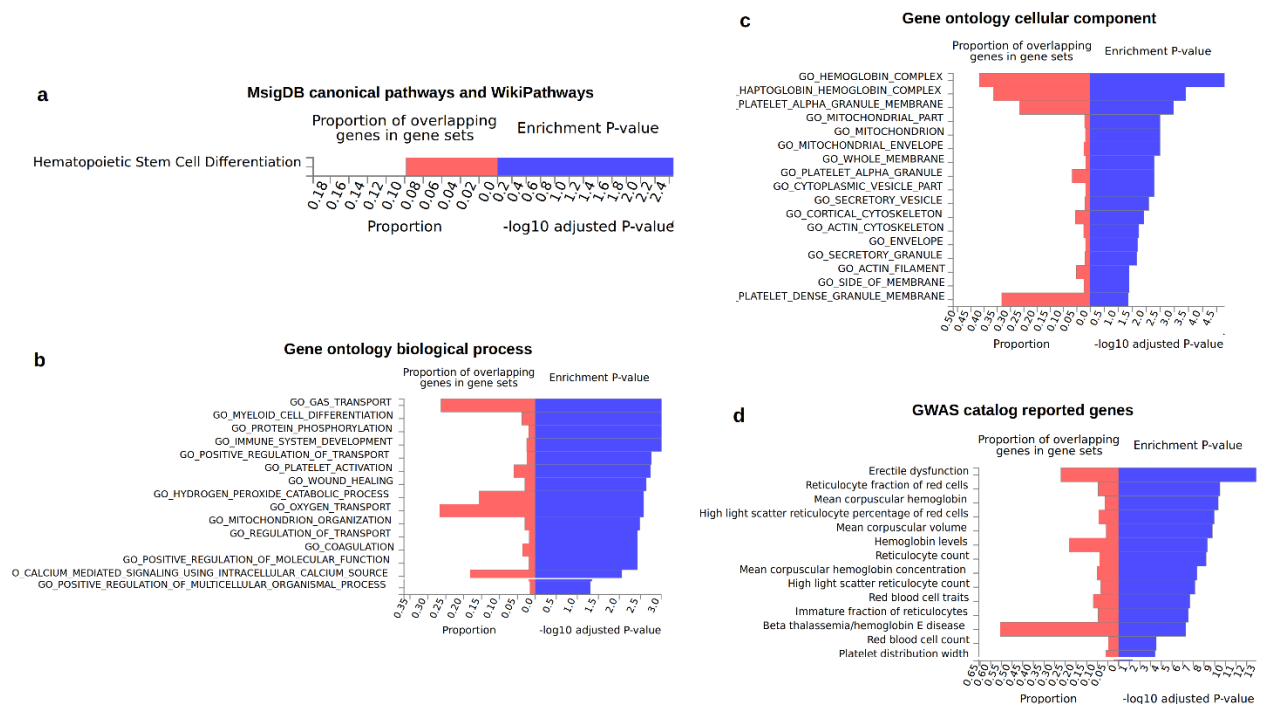


Figure 4.16 Gene-based and gene set enrichment analysis.

(a) Most significantly enriched pathway based on Molecular Signatures Database (MSigDB) and WikiPathways. Gene ontology terms for MSigDB most significantly enriched: biological process (b),

cellular component (c), GWAS catalog reported genes (d). Gene sets used in the analyses were furnished by the SNP2GENE MAGMA analysis of gene-based tests.

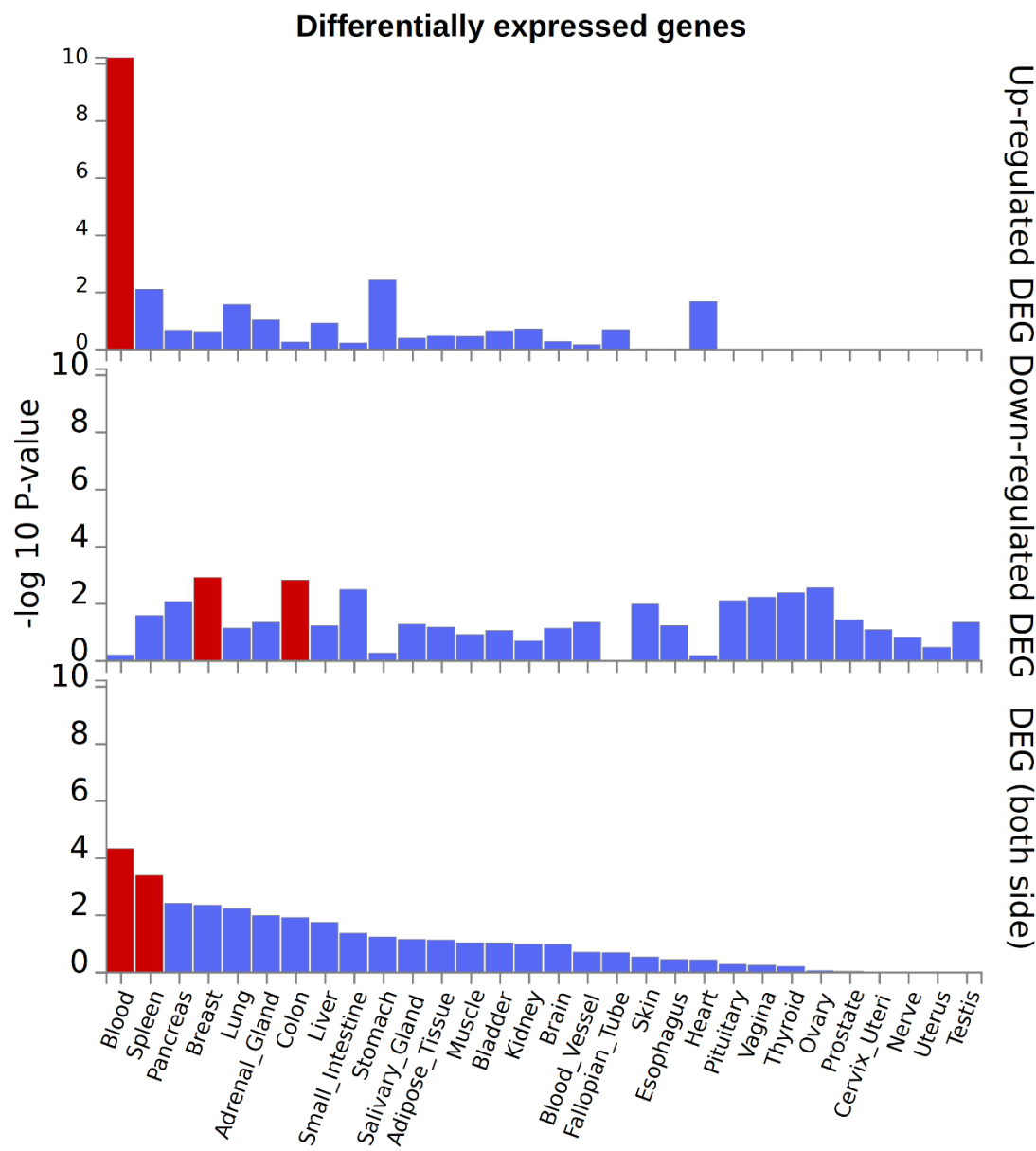


Figure 4.17 GTEx version 8 differentially expressed genes in 30 general tissue types

There was significant upregulation and downregulation of genes in the blood and spleen, while significant downregulation was observed in the breast and colon.

4.3.14. Heritability of fetal hemoglobin

The results presented thus far suggest a robust capture of HbF-modifying loci, with likely many small-effect-size variants that did not reach genome-wide significance. Indeed, furthering this observation, we estimated HbF SNP heritability in a combined, high quality cohort of 1682 Cameroonian and Tanzanian SCA populations at 0.94 (SE 0.01; 95% confidence interval [CI] 0.92 - 0.96; Figure 4.18a), slightly higher than for unselected Europeans (i.e. 0.89)³⁶, and substantially higher than previous estimates for SCA populations of African ancestry living in Europe and America (i.e. up to 0.50)⁴¹. No significant difference was observed when age squared was added as a covariate, while only a moderate improvement was observed by increasing the number of principal components from 20 to 100; a stable decrease, which therefore capped the estimate at ~96%. We note that our approach jointly estimated the additive and dominance genetic variance components²⁴³, as opposed to only the additive genetic variance component (narrow-sense heritability) estimated in previous studies. Attempts at estimating narrow-sense heritability produced highly variable outcomes (mean = 0.70; 95%CI = 0.37 - 1.04; SE = 0.24) (Figure 4.18b), probably reflecting the low power associated with estimating the additive variance component only.

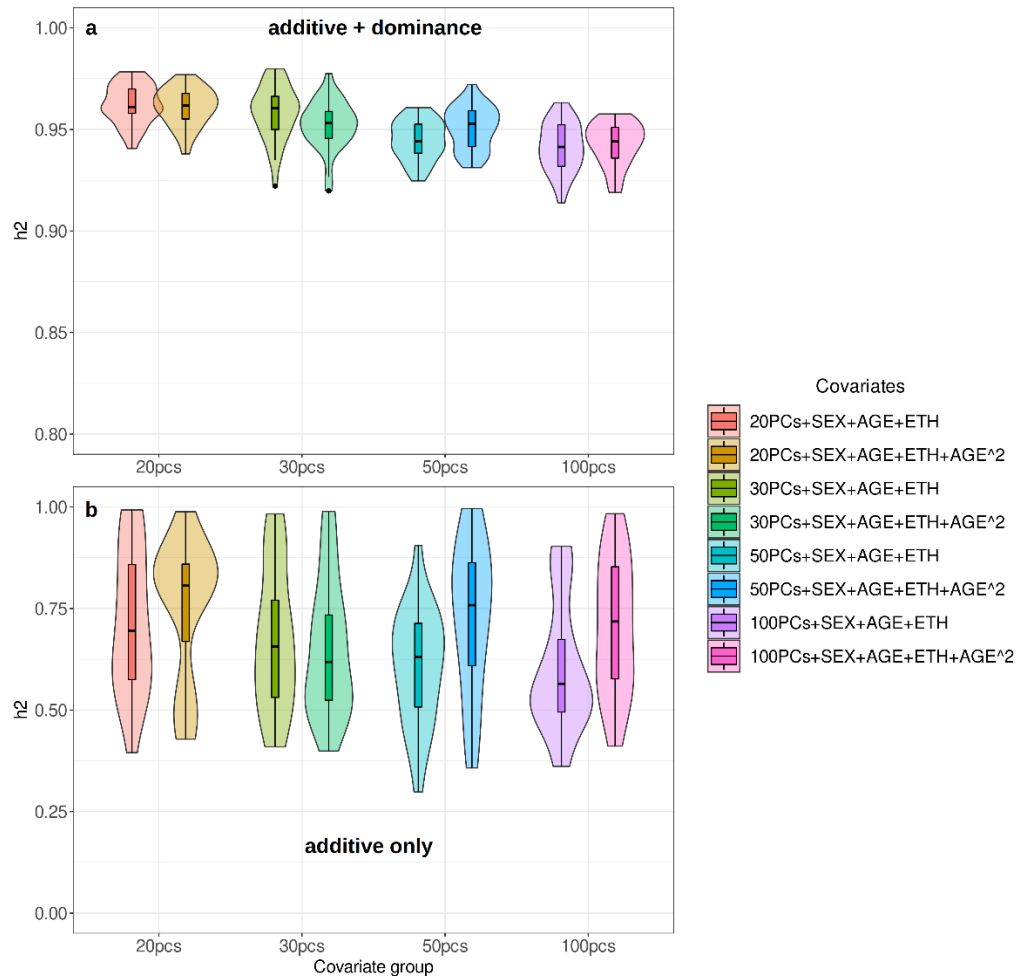


Figure 4.18 heritability of fetal hemoglobin

(a) Estimate of heritability of fetal hemoglobin (HbF) in a merged cohort of Cameroonian and Tanzanian sickle cell anemia patients ($n = 1862$). Additive and dominance genetic variance components were jointly estimated. Thirty (30) replicates of the estimates were generated using four categories of covariates. The violin plots show the density of the distribution of the estimates while the insert boxplots show the spread of the estimates; the lower and upper whiskers are the minimum and maximum estimates, the middle line of the box represents the median estimate, while the lower and upper parts of the box represent the lower and upper quartiles. (b) Additive (narrow-sense) heritability of fetal hemoglobin in Cameroonian and Tanzanian sickle cell anemia populations. Heritability was estimated using the `mc` function of the RHE tool, with four categories of covariates in thirty (30) replicates each. The estimates were plotted using the `ggplot2` R package with trimming of outliers²⁴⁴ and all scripts and workflows are provided in Appendix VIII.

4.3.15. GWAS of combined CAM and TZN genotyped samples mirrors meta-analysis results and confirms robustness of heritability analysis

To ensure that our heritability analysis was not based on spurious associations, we performed a separate association test in the merged, non-imputed, dataset of 1682 Cameroonian and Tanzanian samples that was used in heritability analysis. If all confounders were adequately controlled, we would expect the association results to mirror those from meta-analysis; the same significant signals should be detected with similar strengths of association (p-values, effect sizes and standard errors). Any difference between the two analyses should reflect slightly more power in the merged genotype dataset as was recently suggested ²⁴⁵, and this would mean that the merged dataset is even more powerful for heritability analysis.

Indeed, results of the two analyses were quite similar (Figure 4.19), with substantial overlaps at an adjusted p-value (FDR) of < 0.20 (Appendices IX and X). These included the significant associations in *BCL11A*, *HBS1L-MYB*, *OPCML*, *FLT1*, and *PLCB1*. In addition, the strengths of the associations, particularly the most significant *BCL11A* association, rs1427407, was all but identical in the two analyses. We controlled for confounding by adding the top 10 principal components (PCs), as well as age and country (ethnicity) as covariates. Increasing the number of PCs to 20 only resulted in a slight loss of power without any difference in the number of significant loci: the genomic control inflation factor (lambda) increased slightly from 1.007 to 1.008, the strength of the rs1427407 association reduced slightly from $P = 6.58E-50$ to $P = 2.42E-49$, and the adjusted p-values (FDR) generally increased slightly, although remained significant ($FDR < 0.05$) for the significant loci. This suggests that the stable decrease in heritability estimates observed when the PCs were increased from 20 to 100 could be an undesirable effect of over-correction. Therefore, the upper limit of the estimate at 96% (SE 0.01) might reflect the true heritability in the joint cohort.

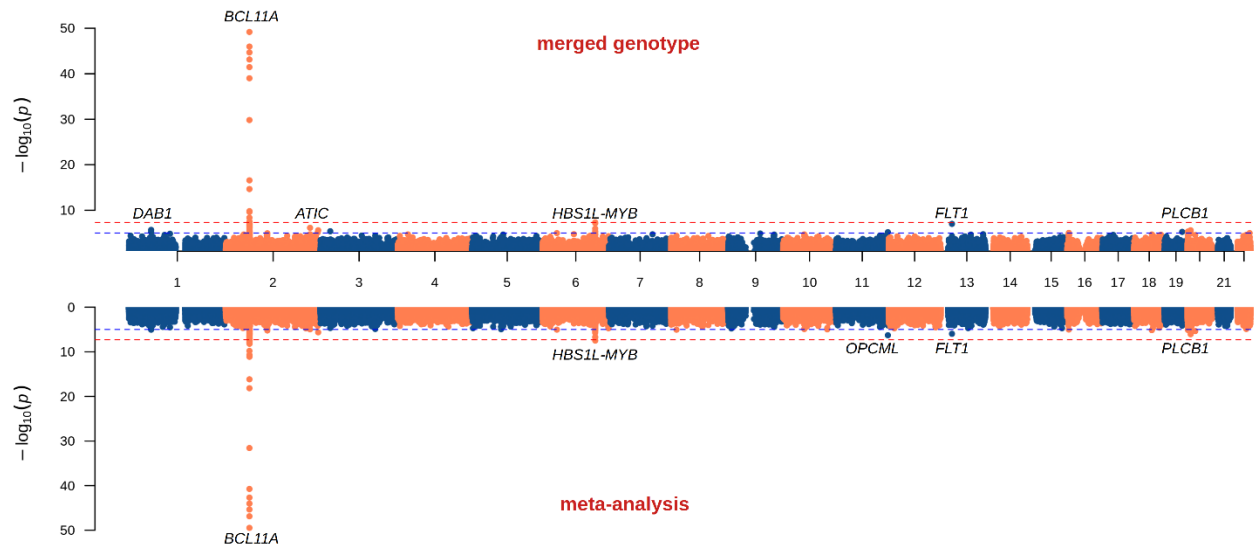


Figure 4.19 Association signals in merged genotype and meta-analysis non-imputed datasets
The plots form a remarkable mirror image of each other.

4.3.16. Involvement of *FLT1* in erythropoiesis is hypoxia-dependent and likely necessary for hypoxia-induced HbF induction

There is growing evidence for the involvement of *FLT1* in hematopoiesis, for instance in the reconstitution of hematopoietic stem cells (HSCs) through recruitment of bone marrow-derived angiogenic cells (BMDACs) and proliferation of HSPCs in a hypoxia-induced manner ^{246,247}. A recent study that investigated the mechanism of HbF induction under hypoxia-induced stress erythropoiesis implicated HIF1A as a key mediator, directly recruiting transcriptional activators (GATA1 and P300) to the γ -globin regulatory elements, targeting chromatin accessibility, and preferentially directing the powerful long-range HBB cluster enhancers of the locus control region to the γ -globin genes ⁵⁷. Consistent with previous data ²⁴⁸, we observed hypoxia response elements (HREs—binding sites for HIFs) in the *FLT1* HS1 (promoter) and HS2, which together with chromatin state data, suggest that a full reactivation of *FLT1* in the erythroid lineage might only be achieved via occupancy of both HSs by HIFs (Figure 4.20a, S1 and S4; see Figure 4.20c). The could result in the commitment of hemogenic endothelial cells (HECs) or BMDACs to HSPCs characterized by downregulation of KDR (a key HEC and BMDAC marker) via the known VEGFA negative signaling role of *FLT1* and its soluble isoform sFLT1 (in a RUNX1-/GFI1-independent manner) ²⁴⁹, thus mimicking the endothelial-to-hematopoietic transition (EHT; Figure 4.20b, (i) and (ii); see Discussion). Importantly, it has been shown that a hypoxia-driven loop between VEGFA, *FLT1*, and phosphorylated extracellular-signal regulated kinase 1/2 (ERK1/2; two mitogen-activated protein kinases – MAPKs) in a neuroblastoma model is capable of trans-activating (by phosphorylation) HIF1A, favoring its nuclear localisation, accumulation and transcriptional

activity (Figure 4.20b) ²⁵⁰. This therefore implicates *FLT1* in the HIF1A-HbF induction nexus (Figure 4.20b). The combination of hypoxia and the disruption of the highly conserved GFI1 binding motif (Figure 4.11b) provides a reasonable model for a full *FLT1* reactivation, and possible recapitulation of embryonic/fetal erythropoiesis, which is further supported by the association of *FLT1* variants with slightly larger erythrocytes (MCV).

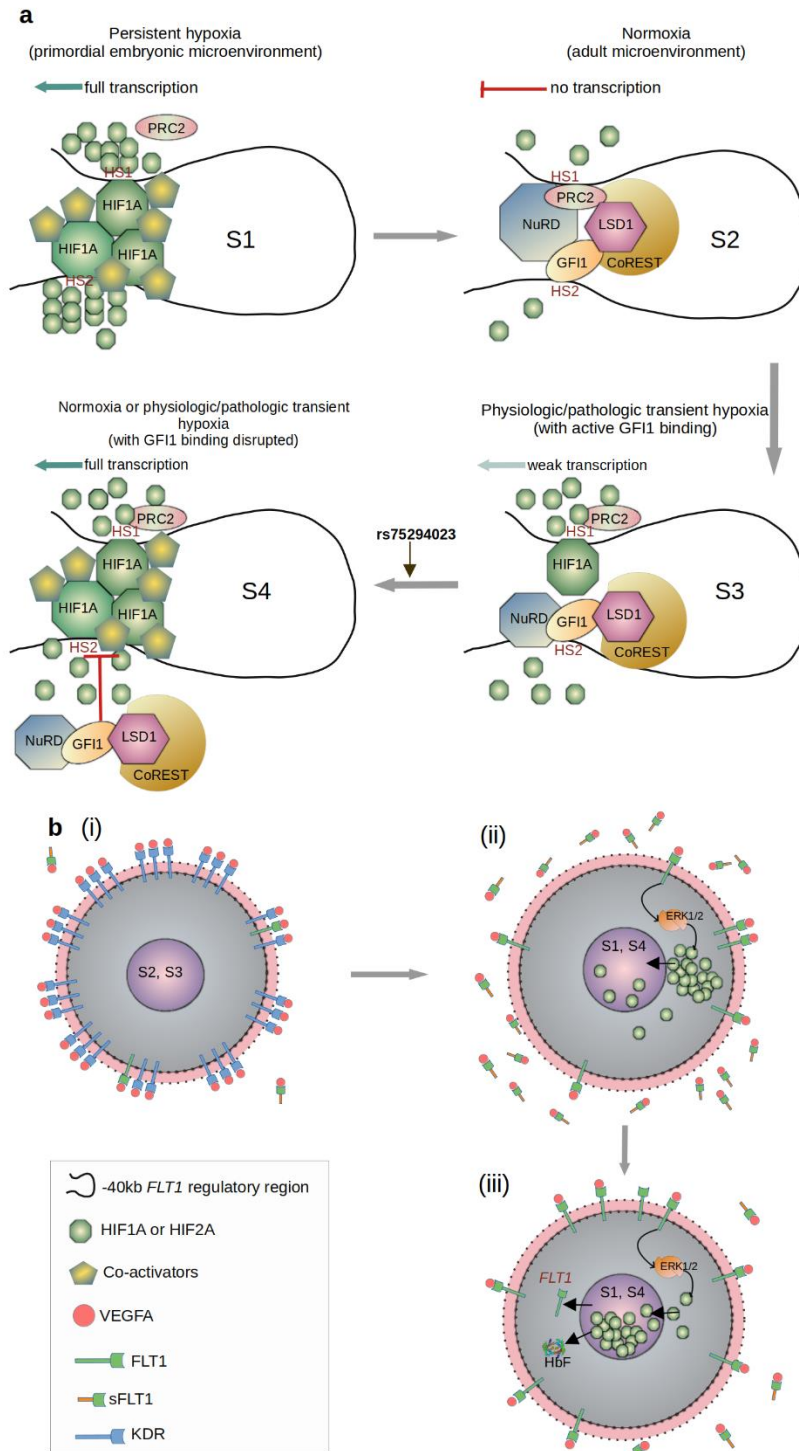


Figure 4.20 Proposed model for *FLT1* inactivation, reactivation, and involvement in erythropoiesis and HbF induction

(a) We inferred four nuclear states (**S1**, **S2**, **S3**, and **S4**) from available data (see Figure 4.10c and Figure 4.11d, and Discussion) that can differentially impact *FLT1* transcription in the endothelial and erythroid lineages. **S1** represents the original embryonic microenvironment which is known to be hypoxic, implying high levels of HIF1A and HIF2A and hence their target gene products including *FLT1*. At increasing oxygen (O₂) concentrations, as the transcriptional program shifts (i.e. degradation of HIFs mediated by VHL ²⁴⁸ and upregulation of the definitive erythropoietic program in the erythroid lineage), RUNX1 and its target gene products (i.e. GFI1—a key repressor in the erythropoietic pathway) are upregulated ²⁵¹, GFI1 tethers other co-repressors (i.e. NuRD and LSD1-CoREST complexes ²⁵²) to *FLT1* HS1 and HS2 facilitated by loop formation evidenced by CTCF peaks in the HS1 and HS2 ²³⁹ (see Figure 4.10c), NuRD (i.e. CHD4 and HDAC) removes transcriptionally active H3K27ac mark (deacetylation) ²⁵³, PRC2 adds transcriptionally inactive H3K27me3 mark (methylation) ²⁴⁰, while LSD1 enforces silencing at HS2 by mediating removal of the transcriptionally active mark H3K4me2 ²⁵⁴. O₂ therefore forms a critical gradient within which there is rapid shift in transcriptional programs, accounting for the endothelial-to-hematopoietic transition (EHT) and *FLT1* inactivation, which, remarkably, resembles hemoglobin switching (i.e. *HBG2* inactivation) ²⁵⁵. These observations are supported by data that suggest the placenta is the site for highest *FLT1* expression (<https://www.ncbi.nlm.nih.gov/gene/2321>) which coincides with a transition from low to normal O₂ tension in the developmental axis ²⁵⁶ and it is thus unclear whether *FLT1* might itself be a critical ‘switch factor’. Ultimately at normoxia (**S2**), such as later in fetal development and adult life, **S2** becomes the default silenced state in the erythroid lineage. No confirmed interaction of the HIFs with GFI1 (or the NuRD and CoREST complexes) ensures that the HS2 silenced state remains intact under physiologic/pathologic hypoxia, such that there is only a weak activation of *FLT1* mediated by HIF1A through reversal of the HS1 polycomb-repressed state (**S3**). A disruption of GFI1 binding might therefore allow a full *FLT1* reactivation under hypoxia (**S4**). (b) represents cellular changes that follow the different nuclear states: in (i) a hemogenic endothelial cell (HEC) or bone marrow-derived androgenic cell (BMDAC) is marked by KDR (VEGFR2). Under hypoxia (**S1**, **S3** or **S4**), *FLT1* is upregulated by HIF1A. Activation of *FLT1* (VEGFR1) by VEGFA closes the VEGFA-*FLT1*-ERK1/2 circuitry in HECs or BMDACs, further activating HIF1A and mediating its nuclear accumulation and transcriptional activity (ii). The superior VEGFA binding affinity of (s)*FLT1* ²⁴⁶ negatively regulates and drastically downregulates KDR, marking the RUNX1-/GFI1-independent transition to hematopoietic stem and progenitor cells (HSPCs). Erythrocytes resulting from these microenvironmental cues might thus express high levels of fetal hemoglobin (HbF) via, at least, the known HIF1A nexus ⁵⁷.

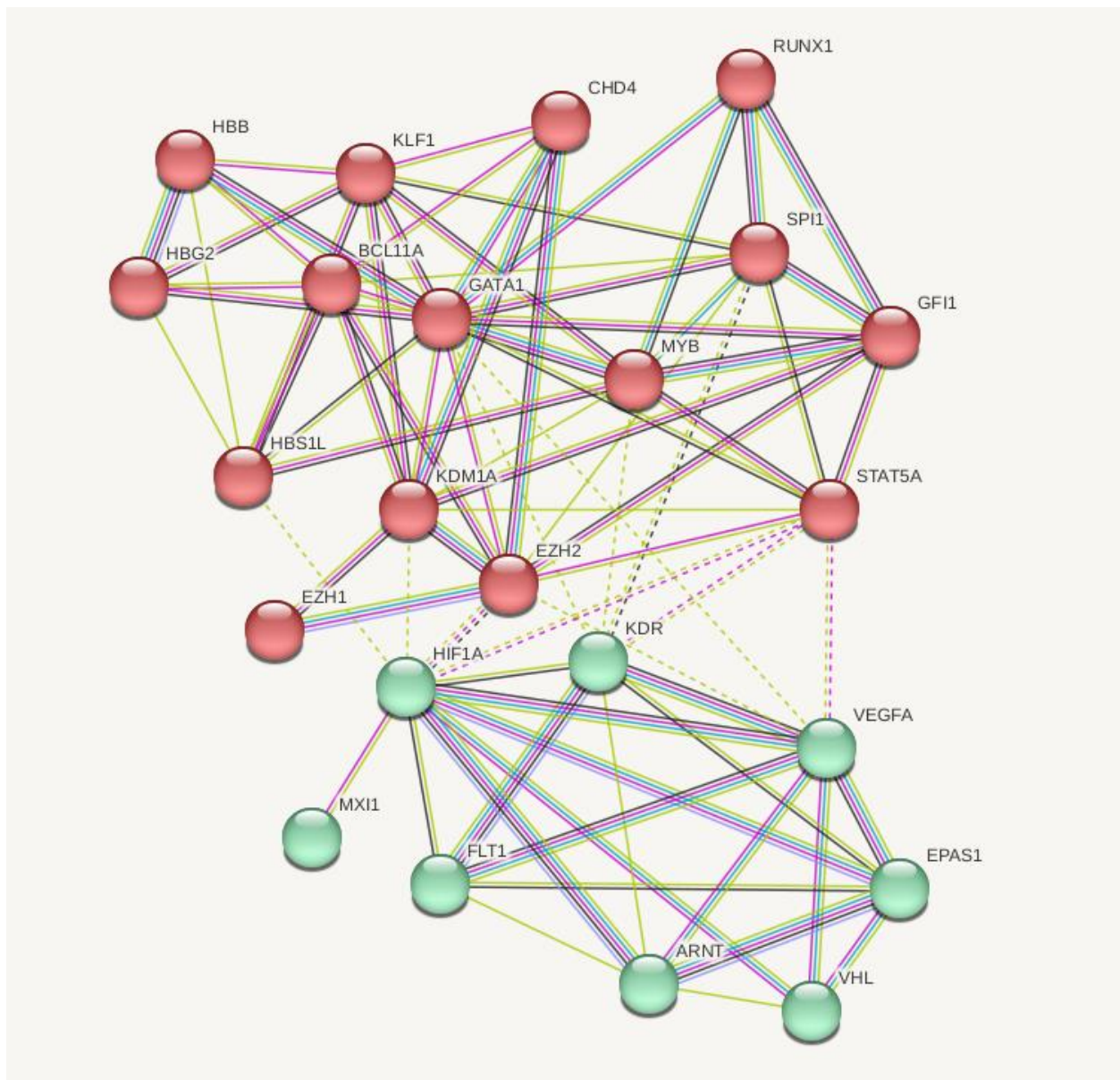


Figure 4.21 Protein-protein interaction network of key erythropoietic and hypoxia response pathway proteins

This supports a main role of HIF1A as mediating the de-repression of polycomb-repressed states via inactivation of EZH2, the major catalytic subunit of PRC2, whose inactivation has also been linked to the de-repression of the embryonic ϵ -globin gene ²⁵⁵. KDR also appears to be a major point of contact between the two pathways.

4.3.17. Expression patterns of *FLT1* and other erythropoiesis and hypoxia related genes in HUDEP-2 and K562 cells under hypoxia

Notably, HUDEP-2 cells display adult erythropoiesis in which the beta-globin gene (*HBB*) is the main expressed, while K562 cells display embryonic erythropoiesis with the epsilon-globin gene (*HBE1*) being the main expressed and only trace amounts of the gamma-globin genes (*HBG1/2*) expressed. However, both cells are capable of expressing high amounts of HbF (*HBG1/2*) in response to certain stress conditions such as hypoxia²⁵⁷ or, as in the case of K562 cells, in response to certain inducers such as rapamycin²⁵⁸. We assessed the expression patterns of *FLT1* and other erythropoiesis and hypoxia related genes in two erythroid cell lines, HUDEP-2 and K562 cells, under short-term (3-24 hours) and long-term (4-10 days) hypoxia (2% O₂) using quantitative polymerase chain reaction (qPCR). There was an apparent rapid induction of *FLT1* under short-term hypoxia in K562 cells (Figure 4.22a&b). The expression of *FLT1* increased with increase in the duration of hypoxia, which was characterized by increase in the expression of the hypoxia biomarkers, 5'-aminolevulinate synthase 2 (*ALAS2*) and carbonic anhydrase 9 (*CA9*), *ALAS2* of which is erythroid-specific (Appendix IV). *FLT1* was not expressed in HUDEP-2 cells consistent with our *in silico* functional data and supporting our hypothesis that *FLT1* is likely only involved in early (embryonic) erythropoiesis (Figure 4.22a). As predicted in our *in silico* models, *FLT1* expression in K562 cells was likely contributed by downregulation of *EZH2*, the major active subunit of the polycomb-repressive complex 2 (*PRC2*) (Appendix V).

After 24-hour hypoxia, *BCL11A* expression decreased in HUDEP-2 cells (consistent with increase HbF level under hypoxia-induced stress erythropoiesis²⁵⁷) but increased in K562 (Appendix VI). However, while *BCL11A* expression increased in HUDEP-2 cells after 10-day hypoxia, it was not expressed in K562 cells, possibly explaining the elevated expression of *HBG1/2* observed in K562 cells after 10-day hypoxia, almost six times the expression in HUDEP-2 cells (Figure 4.23a&b). No expression of *KDR* was observed under short-term hypoxia in both cells (Appendix V), consistent with the fact that these are already committed erythroid cells. However, and remarkably, *KDR* was apparently reactivated under long-term exposure to hypoxia: moderately in HUDEP-2 and considerably in K562. Recall that *KDR* is a key marker defining hematopoietic stem cells²⁵⁹, and therefore its reactivation in the erythroid lineage might indicate a critical step in the reconstitution and proliferation of hematopoietic stem and progenitors cells (HSPCs).

The expression of erythropoiesis regulators after 10-day hypoxia was largely consistent with transcriptional activation in K562 cells and repression of *HBG1/2* in HUDEP-2 cells (Figure 4.23a&b). When considering the transcription factors whose binding motifs were disrupted by *FLT1* regulatory region variants in our study, there was an apparent rapid increase in the expression of *GFI1* on the first day of hypoxia exposure in both cells, and a rapid decrease thereafter in HUDEP-2 cells (Appendix VII). There was only a slight decrease in K562 cells and the

expression appeared to plateau from day 4. *STAT5A* expression in both cells appeared to be generally restricted (downregulated), while *MXI1* appeared to be appreciably induced only after 24-hour hypoxia expression, before and after which the expression was either minimal or downregulated.

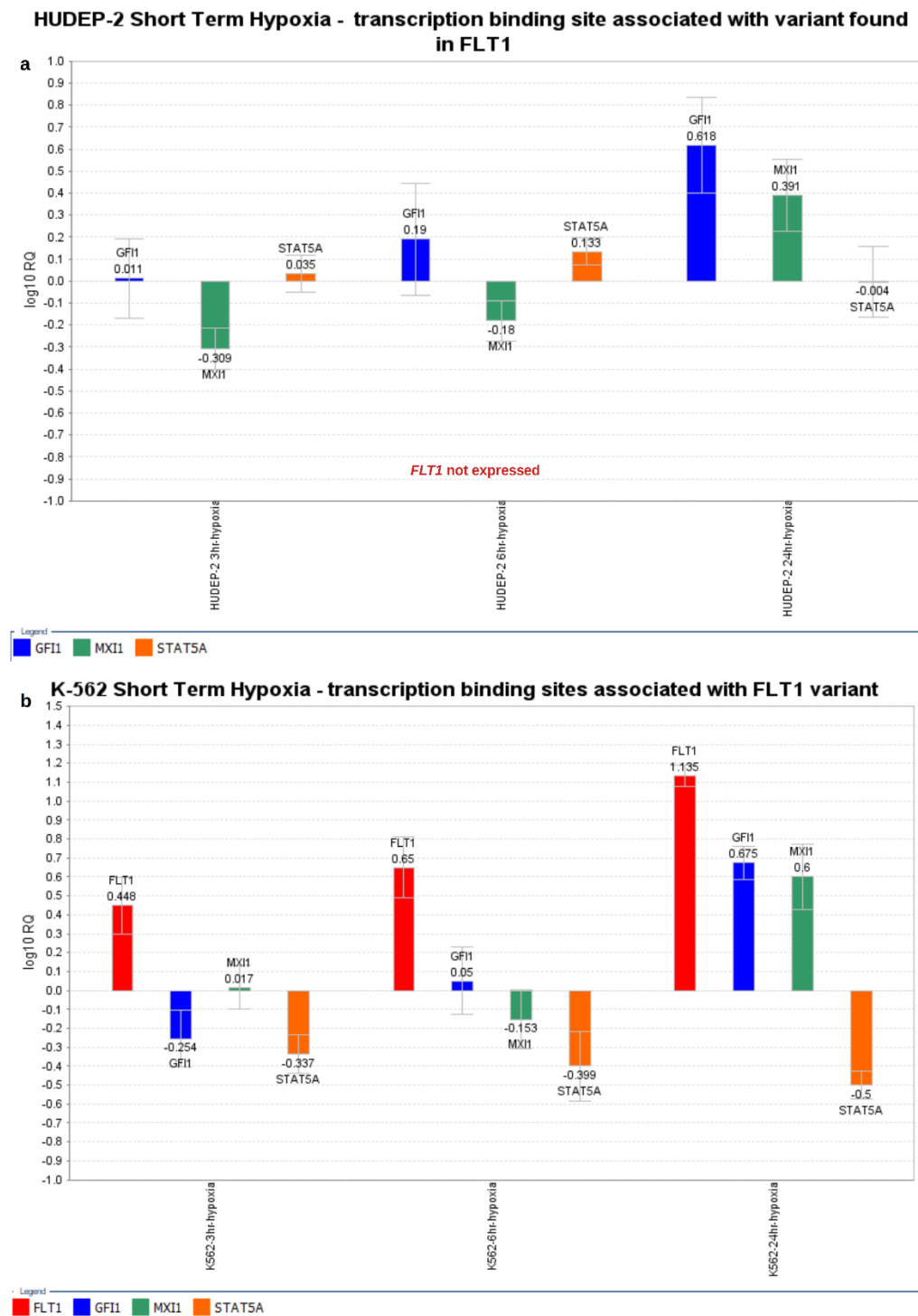


Figure 4.22 Expression pattern of FLT1 and erythropoiesis transcription factors in HUDEP-2 and K562 cells under short-term hypoxia (3-24 hours)

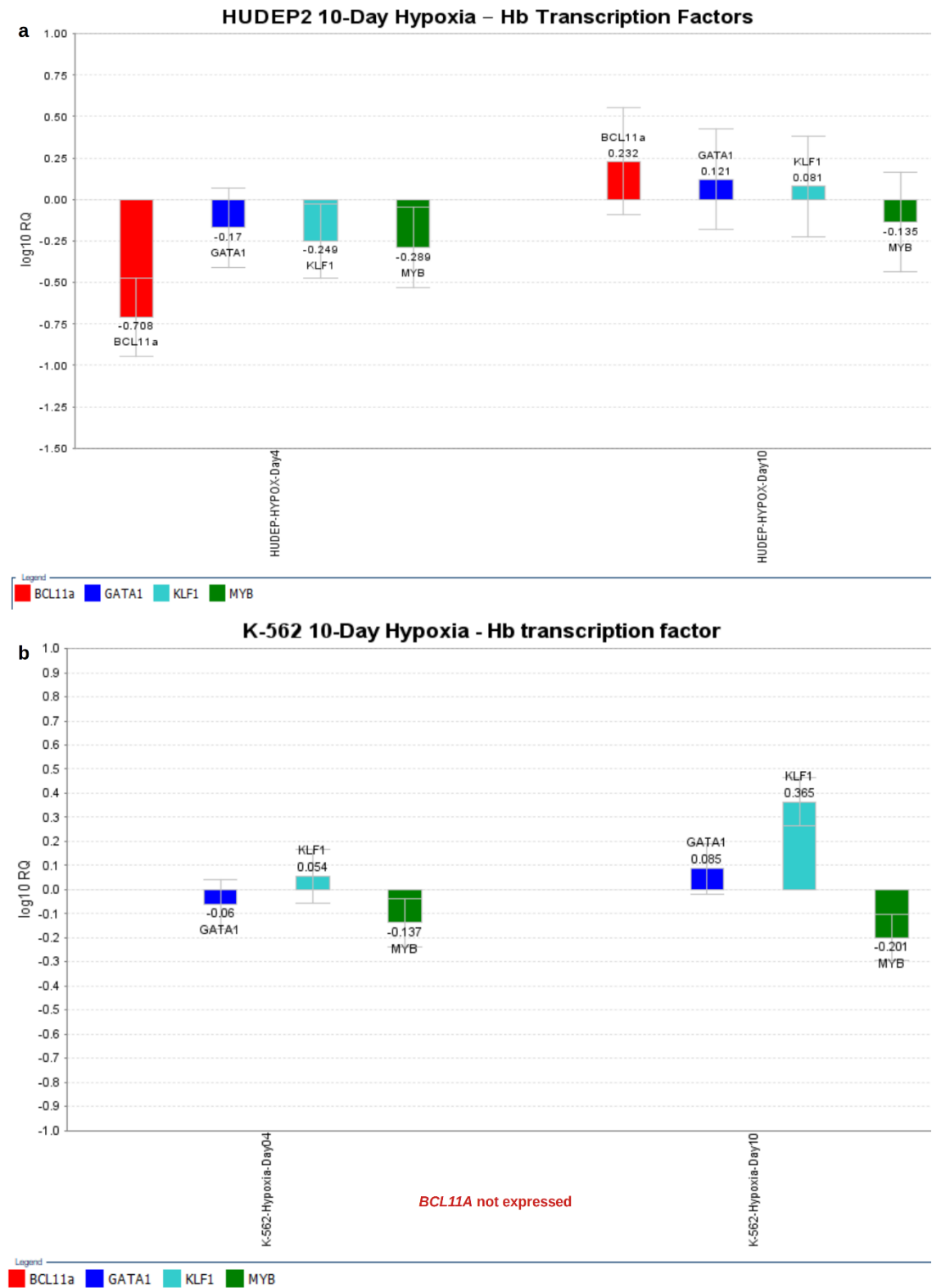


Figure 4.23 Expression pattern of FLT1 and erythropoiesis transcription factors in HUDEP-2 and K562 cells under long-term hypoxia (4-10 days)

4.4. Discussion

4.4.1. Summary of key points

In this study we provide a comprehensive attempt to investigate genetic variations influencing HbF level in an African sickle cell disease cohort with the use of a multi-panel imputation and association strategy against two distinct genotyping arrays tailored to capture a broad spectrum of Africa-specific and non-African genetic variations. This is expected to contribute to the much-needed new data from populations with African ancestry^{260–262}. Specific advantages of our study include: a) a Cameroonian cohort made of hydroxyurea-naïve patients living with SCD from Cameroon which provides a proxy for the natural disease history; b) the identification of fourteen potentially novel loci enhanced by the reanalysis of previously reported data from Tanzania, and global meta-analysis including data from individuals of African ancestry living with SCD in the United States of America; c) in particular, the description of variants in *FLT1* which are largely specific to African populations and look promising as future therapeutic targets, as well as our elucidation of the complex haplotype architecture of *FLT1* which provides evidence of substantial genomic variability that can be extended to other known and novel loci to explain the difference in sentinel variants observed in different populations; and d) a detailed *in silico* and *in vitro* cell-based functional exploration of the potential mechanism for *FLT1* involvement in erythropoiesis and HbF induction.

4.4.2. Heterogeneity in imputation panels

The use of multiple imputation panels for association analysis in our study is innovative, as other studies typically select a best-performing panel for association testing. Although the dissection of the comparative performance of imputation panels for sub-Saharan African populations has been performed in previous studies^{64,66}, our study presents comparisons for a larger variety of panels and variants that are relevant to populations of African ancestry, including a custom panel that was highly similar to the study population. Our observations were largely similar to previous reports; differential imputation performance, substantial panel-specific variants, and relatively low proportion of shared variants (less than 30%) that implied that a given variant can be imputed with vastly different accuracies across different panels, reflecting differences in haplotype structures resulting from different tagging schemes in the panels⁶¹. The TOPMed panel is without doubt the single best. However, its current version (freeze 8) has a known limitation for African ancestry populations as we revealed in our malaria GWAS study published in February of 2023⁶⁶; the panel fails to impute critical functionally relevant Africa-specific variants, such as the sickle cell mutation (rs334), that are imputed with very high accuracies (> 90%) using other panels. Importantly, our findings underscore the complementarity of the panels, particular in highly diverse populations, reflected in panel-specific signals. This supports the utilisation of all the

panels for downstream association as an optimal approach rather than selecting a ‘best-performing’ panel. Better still, the recently developed meta-imputation procedure for combining multi-panel imputed datasets (which was not available at the time of our study) ⁶⁷ would be a more computationally tractable way of handling such datasets given the enormous challenges accompanying separate analyses. Alternatively, future association studies involving highly diverse populations should consider whole-genome sequencing (WGS), as much as possible, to alleviate the large inconsistencies and complexities that come with utilizing multiple imputation panels.

4.4.3. Heterogeneity in association signals

Disparities in haplotype structure were the major reason for differences in imputation performance and association signals, but also differences in sentinel variants of significant loci amongst the cohorts in our study, as demonstrated in the *FLT1* -40kb region. While genetic admixture can account for these haplotype differences, it could also mean that genetic loci influencing HbF level, and other modifiers of sickle cell disease, have been through different evolutionary trajectories, especially in sub-Saharan African populations that have been exposed to vastly different ecologies which have shaped their genetic material differently over the roughly 300,000 years of modern human existence on the continent ²⁶⁰. The continued revelation of extensive uncaptured genomic variations within African populations ^{263,264}, some of which are population-specific, such as the *FLT1* variants reported here, reflect the enormous selective pressures that the populations have had to contend. The lack of these, sometimes functionally relevant, genetic variants in notable databases such as the GTEx portal reflects a current limitation for global genetic medicine. For instance, we recently reported malaria protective associations in the enhancer region of *CHST15* which tag strong eQTLs in tissues relevant to the disease biology, but are absent in the GTEx portal ⁶⁶. In cases where none of the tag variants is present in such databases, such as in the current study, a critical piece of functional information would be lost. The importance of increasing the representation of understudied populations in global omics databases could therefore not be overstated.

4.4.4. Heritability supported by association results, pathway enrichment, and potentially, selection pressure (at least in Cameroonians) evidenced by high LD and haplotype conservation in the *FLT1* 40kb regulatory region.

Over two decades ago, the heritability of HbF level was estimated in an unselected (not ascertained for any disease) twin population in the United Kingdom at 89% ³⁶, and subsequently in sickle cell anemia patients of African ancestry based in the United States of America at ~50% ⁴¹, and ~32% in SCD patients of African ancestry older than 15 years of age and living in France ⁴². Even lower estimates have recently been suggested for SCD patients of African ancestry ⁴³. In the European unselected population, half of the total HbF heritability is explained by just the three major loci i.e. *BCL11A*, *HBS1L-MYB*, and *HBG2* ⁴⁰. Our estimate of 94% is unsurprisingly higher than previous estimates for several reasons: (i) our approach jointly estimated additive and

dominance genetic variance components, whereas previous approaches estimated only the additive variance component (narrow-sense heritability), suggesting that a substantial portion of HbF heritability in selected patients from Africa could be explained by a dominance genetic variance component ²⁶⁵, (ii) our cohorts are fundamentally different from the other cohorts in that our samples represent individuals with the most severe form of sickle cell disease who have traversed childhood mortality largely without the benefits of current healthcare strategies such as newborn screening followed by comprehensive life-saving care with penicillin prophylaxis and hydroxyurea treatment. Considering the historically high under-five mortality rate (50-90%) of sickle cell anemia in Africa ¹¹, therefore this group of patients is likely to represent a naturally selected population enriched with genetic variants that favor “long survival” such as has been previously shown in patients from Cameroon ⁷⁰. It is therefore reasonable to imagine that HbF-induction is among the most enriched pathways given that it is the most potent modifier of SCD severity known to date. Moreover, the high number of potentially novel loci reported here, the overwhelming enrichment of variants in the hematopoietic pathway with the sheer number of replication level associations in the *HBG2* gene, as well as replication of other recently observed loci, support the effectiveness of our strategy to capture a broad spectrum of HbF-associated loci. However, larger sample sizes of patients living in Africa with SCD, and standardized measurements of HbF, would be needed to confirm the true heritability of HbF in SCD in Africa.

4.4.5. Functional relevance of *FLT1* associations

Our study reports with unambiguous significance associations around the *FLT1* locus. Transcription factors GFI1, STAT5A, and MXI1, whose binding motifs were disrupted by variants observed within the *FLT1* regulatory region have known roles in erythropoiesis/hematopoiesis; GFI1 (for which rs75294023 disrupted an indispensable nucleotide in its binding site) regulates chromatin state essential in human endothelial-to-hematopoietic transition (EHT), STAT5A is essential for erythropoietin response in stress erythropoiesis in the absence of erythropoietin receptor, while MXI1 competes for MAX with MYC, protecting against MYC-dependent sensitisation to hypoxia-induced apoptosis and is also essential for erythrocyte enucleation in terminal erythropoiesis ^{235–238}. Together with protein-protein interactions of major erythropoiesis and hypoxia pathway proteins, these suggest that EHT and hypoxia-induced stress erythropoiesis are two biological processes in which *FLT1* might play a crucial role. FLT1 (also known as VEGFR1) is a member of the vascular endothelial growth factor (VEGF) receptor family along with VEGFR2 [also known as kinase insert domain receptor (KDR) or fetal liver kinase-1 (FLK1)], responsible for regulating angiogenesis and vascular permeability mediated by VEGFA, exclusively through the mitogen-activated protein kinase (MAPK) pathway ²⁴⁶. While FLT1 has a higher binding affinity for VEGFA than KDR, its signal transducing (angiogenic) potential is about 10-folds lower ²⁴⁶. FLT1 is additionally activated by the placental growth factor (PIGF) and this is known to reconstitute hematopoiesis. In addition, soluble isoforms of FLT1 (sFLT1) which lack the signal transducing

domain are potent antagonists of VEGFA²⁶⁶, eliminating VEGFA induced mitogenesis of HUVECs²⁶⁷.

It is now well-established that hematopoietic stem and progenitor cells (HSPCs) derive from a small population of vascular endothelial cells called hemogenic endothelial cells (HECs) via the transient process of EHT during embryonic development²⁶⁸. Interestingly, the EHT, controlled by the RUNX1 transcription factor, is marked by a drastic downregulation of KDR²⁵¹. GFI1, a target of RUNX1, has been shown to be responsible for KDR downregulation in the transition²⁶⁹. Yet, the survival of *Runx1*^{-/-} primitive erythrocytes and embryos suggest a redundancy in the activation of the EHT during primitive erythropoiesis²⁷⁰. Therefore, the VEGFA negative signaling role of (s)FLT1 could mediate the downregulation of KDR given that KDR upregulation depends on a VEGFA-KDR positive signaling. Indeed, the elevated expression of *FLT1* and the reactivation of *KDR* in K562 cells (which display embryonic erythropoiesis) under hypoxic conditions support these premises. Besides, the reactivation of KDR in K562 cells under hypoxia all but confirms its role in linking the hypoxia response and erythropoiesis pathways displayed in the protein-protein interaction network in Figure 4.21. Hence, understanding the exact microenvironmental cues that lead to KDR reactivation in the erythroid lineage, and its role thereafter seems like a great opportunity to further understand the drivers of HSC proliferation, and potential targeting to supplement gene therapy strategies aimed at increasing HbF level for F-cell augmentation. Therefore, the erythroid-specific induction of *FLT1* by hypoxia inducible factors as a result of the disruption of GFI1 binding motif in *FLT1* regulatory region, the downregulation of KDR mediated by (s)FLT1, as well as the VEGFA-FLT1-ERK1/2 loop involved in activating HIF1A²⁵⁰, provide a reasonable model for the involvement of *FLT1* in hypoxia-induced proliferation of HSPCs and HbF induction. *FLT1* might thus be an essential component of early/primitive erythropoietic microenvironmental cues.

The functional relevance of *FLT1* in HbF induction is further supported by previous studies involving SCD patients in the USA²⁷¹ and beta-thalassemia patients from Greece²⁷², showing *FLT1* to be associated with improved hydroxyurea-induced HbF level and clinical outcome, which correlated with various measures of haemolysis, and erythropoietin level (an inducer of erythropoiesis)²⁷³. In addition, data from *Flt-1* and *Flk-1* knock-out mice show disruption of erythropoiesis^{274,275}. Detailed phenotyping of surviving, *Flt1*^{-/-} as well *Flt1*^{+/-} and *Flk1*^{+/-} and other knock-in mice model, coupled with additional functional analysis in appropriate cell lines such as HUDEP-1 will assist in investigating the role of *FLT1* in erythropoiesis, F-cells, and HbF production.

4.4.6. Methodological considerations/limitations/strengths

There is ongoing discussion regarding the optimal p-value cut-off for defining genome-wide significance in associations studies such as in those involving populations with relatively low

linkage disequilibrium or studies that used imputed datasets such as ours ²⁴⁵. Suggestions include considering more stringent thresholds owing to the extensive genetic diversity of African populations (e.g. $P < 5E-09$) ²³¹. Our definition of moderate significance at $FDR < 0.05$ was motivated by the reasoning that, while a high genetic diversity among African populations would translate to a high false positive rate, the false negative rate would also be expected to be high, owing to the largely contribution of historical environmental (climate, diet, disease) pressures. This would imply more complex disease biology with potentially multiple small-effect associations not strong enough to be detected at the currently accepted genome-wide significance threshold ($P < 5E-08$). While a $FDR < 0.05$ (meaning that less than 5% of all variants called significant could be false positives) slightly increases the likelihood of false positivity, it also increases the power to detect true associations by decreasing the false negative rate compared to other more conservative methods. Indeed, (i) the enrichment of variants in our study in pathways that are relevant to SCD, and more importantly to HbF production, (ii) the suggestive variants we observed in loci that were recently detected through unconventional association techniques, e.g. in *ZNF410* and *JAZF1* through CRISPR screening and RNA interference respectively ^{56,276}, and (iii) the very high heritability of HbF and robustness of our strategy suggest that the loci presented here are most likely true associations.

Additional support comes from the recent detection of a putative novel erythropoietin QTL on chromosome 15 with evidence of association at $P = 1.05E-07$ ²⁷⁷, which is just below widely accepted genome-wide significance threshold. Indeed, *SLC4A4* which was detected in the Tanzanian cohort at $P = 5.75E-07$ is a bicarbonate cotransporter that is involved in regulating intracellular pH, a major factor that determines HbS polymerization and red blood cell sickling, and may therefore be involved in HbF production in some rather convoluted manner. In addition, *OPCML* which was detected in the Tanzanian cohort in 2014 at $P < 1E-06$ ⁴⁹ was replicated in this reanalysis at $P < 3E-07$ in both Tanzania and CAM-TZN meta-analysis, suggesting some functional relevance in HbF production although the evidence of association falls short of the conventional significance threshold. Our strategy therefore appears robust to detecting true signals, although a combination of statistical significance with extensive functional analyses, is strongly encouraged to confirm true biological significance.

The absence of large-effect novel associations in the meta-analysis suggests that we are approaching saturation in the discovery of major HbF-associated loci with variants down to $MAF > 1\%$. It could also mean the saturation of loci that contribute to HbF variability additively in these cohorts. Ascertainment bias may also have limited our discovery of other novel loci in the global meta-analysis given that only about 1.2 million variants that were present in the SITT cohort were analyzed across the cohorts. However, it could be expected that increasing sample size may uncover additional loci as supported by a recent study which described HbF-associated variants in *BACH2* ⁴³, for which suggestive associations were observed in the present study. Nevertheless,

the observation of substantial genetic heterogeneity, supported by novel African-specific sentinel SNPs including potentially causal variants in *FLT1*, and relatively smaller credible sets following fine-mapping, despite a modest Cameroonian cohort sample size, highlights the importance of investigating larger African populations, from multiple countries.

4.5. Conclusions

In this study, differential analysis of six imputation panels showed an improved accuracy with the number of genomes used, and genetic relatedness to study population, relatively low proportion of shared variants (28%), and wide range of exclusive variants (2.3 - 16%) per panel, and the need for local/regional imputation panel illustrated by the surprisingly higher than expected performance of the *custom* panel. Known HbF-associated loci such as *BCL11A* were replicated, and fourteen potentially new loci were identified (*FLT1*, *ZNF804A*, *MRPS22*, *PRLR*, *FRMD1*, *DMRT1*, *OPCML*, *SERPINA6*, *PLCB1*, *SLC4A4*, *GFRA1*, *LOC574080*, *LOC107984174*, *LINC01898*). Estimated HbF heritability explained was slightly higher (94%) than previously reported (89%) in an unselected population in a British twin study. Variants in known and novel HbF-associated loci showed high levels of genetic variability, reflected in differences in haplotype structures in *FLT1* regulatory region in which Africa-specific variants were observed. Published data on *Flt-1* and *Flk-1* knock-out mice, as well as *in silico* and *in vitro* functional explorations suggest that *FLT1* is directly associated to hypoxia-induced erythropoiesis, and potentially to HbF-induction through a nexus involving VEGFA and HIF1A, and driven by HIF1A. The disruption of the GFI1 repressor binding motif within *FLT1* regulatory region in erythroid-specific cells therefore suggests a possible mechanism for the involvement of *FLT1* in HbF induction. The data emphasize the importance of African populations in discovering variants for quantitative traits, and further the understanding of SCD pathobiology. Future studies in mice models, coupled with additional functional analysis in appropriate cell lines will assist in further elucidating the role of *FLT1* in erythropoiesis and HbF production, as well as the involvement of other novel loci that may present alternative targets for therapeutic interventions in SCD.

4.6. Methods

4.6.1. Ethical approvals

The research was performed in accordance with the Declaration of Helsinki. Approval was obtained from the University of Cape Town, Faculty of Health Sciences Human Research Ethics Committee, Cape Town, South Africa (HREC/REF: R015/2018), and National Ethical Committee of the Ministry of Public Health of Cameroon (No 193/CNE/SE/15). All patients older than 18 years signed consent forms, while informed consent was given by the parents or guardians for participants younger than 18 years old. Written and signed informed consent forms were obtained from adult participants and parents/guardians of minor patients. An assent was also

obtained from the participants of more than 7 years old. The present study involved a secondary analysis of existing data, and was reviewed and approved by the University of Cape Town, Faculty of Health Sciences Human Research Ethics Committee, Cape Town, South Africa (HREC REF: 606/2021).

4.6.2. Patient participants

The data were collected from nine hospitals from five cities in Cameroon, including Yaoundé, Douala, Bafoussam, Bertoua, and Maroua from May 2016 to July 2018. Socio-demographic and clinical events were collected by means of a structured questionnaire administered to parents/guardians and adult SCD patients. Patients' medical records were reviewed, to delineate their clinical features over the past 3 years. Only patients older than 5 years of age (to avoid age-related changes in the complete blood count and HbF level), who had not received a blood transfusion or hospitalisation in the past 6 weeks were included. None was currently treated with hydroxycarbamide or opioids. The sampling strategy was not restricted to hospital-based patients to avoid the bias that might result from including only the sickest patients. To accomplish this goal two SCA patients' associations in Cameroon were engaged for collaboration and additional patients were recruited during their monthly meetings. No incentive was provided for participation in the study.

4.6.3. Measurements of hematological indices

Hemoglobin electrophoresis and complete routine blood count of the SCA patients were conducted upon arrival at the hospital. High performance liquid chromatography (HPLC) was used for the measurement of HbF levels at the hematological laboratory of the Centre Pasteur in Yaoundé, as previously described ^{226,227}.

4.6.4. Molecular methods

Genotyping of the sickle cell anemia mutation, HBB cluster haplotypes, and 3.7 kb HBA1/HBA2 deletion

DNA was extracted from peripheral blood following the manufacturer's instructions (Puregene Blood Kit; Qiagen, Hilden, Germany). Molecular analysis to determine the presence of the sickle mutation was carried out on 200 ng DNA by polymerase chain reaction (PCR) to amplify a 770 bp segment of the *HBB*, followed by DdeI restriction analysis of the PCR product ²⁷⁸. The present analysis was restricted to sickle cell anemia (homozygous HbS) due to the well-known differences in laboratory parameters ^{279,280}, and to allow single sickle genotype (HbSS) for genetic associations. Using published primers and methods, five restriction fragment length polymorphism (RFLP) sites in the *HBB* cluster were amplified to analyse the XmnI (5'G γ), HindIII (G γ), HindIII (A γ), HincII (3 $\psi\beta'$) and HinfI (5' β) for the *HBB* haplotype background (see Figure 4.1g) ²²⁸. The 3.7 kb *HBA1/HBA2* deletion was successfully screened, using the expand-long template PCR (Roche Diagnostics, Basel, Switzerland), as previously published ²⁸¹.

4.6.5. Cohorts description

Cameroonian Sickle Cell Anemia patient participants: Genotyping and genotype calling

Two batches of samples of sickle cell anemia patients from Cameroon (batch 1: $n = 1199$, batch 2: $n = 403$) were genotyped on the 2.3M H3Africa SNP array at Illumina® FastTrack™ Microarray services (Illumina, San Diego, USA) between 2018 and 2019. Genotype calling was performed for each batch using the Illumina gencall algorithm from the Illumina Array Analysis Platform Genotyping Command Line Interface (IAAP-CLI) version 1.1 (Appendix VIII). Briefly, gencall was used to process intensity data in IDAT format to GTC formats, utilizing manifest and cluster files specific to the H3Africa chip retrieved from <https://chipinfo.h3abionet.org/downloads> (Accessed: December 5, 2021). Thereafter, the per-sample GTC files were converted to a single VCF file for the separate batches of samples using the gtc2vcf plugin of bcftools version 1.15.1, while aligning to the human reference sequence in build 37 (hg19) coordinates.

Tanzanian Sickle Cell Anemia patient participants

The data consisted of genotypes for 1213 Tanzanian sickle cell anemia (SCA) patients generated using the Illumina Human Omnichip 2.5 platform (Illumina Inc., San Diego, CA, USA), and available at the European Genome Phenome Archive (EGA) under the study accession number EGAS00001000990⁴⁹. The genotype data mapped to the human reference in build 37 coordinates, in PLINK binary format, as well as clinical data were obtained from the Tanzanian investigators.

USA-based Sickle cell disease cohorts with patients of African ancestry

We obtained meta-analysed summary statistics of fetal hemoglobin GWAS involving seven cohorts of sickle cell anemia (HbSS) patients based in the United States of America (USA) from the study by Harold T. Bae et al., totaling 2040 samples and 1,198,700 variants²²⁵. The cohorts included: Cooperative Study of Sickle Cell Disease (CSSCD: $n = 841$), Multicenter Study of Hydroxyurea (MSH: $n = 178$), Pulmonary Hypertension and the Hypoxic Response in Sickle Cell Disease (PUSH) study ($n = 73$), Comprehensive Sickle Cell Centers Collaborative Data (C-data) project ($n = 127$), Treatment of Pulmonary Hypertension and Sickle Cell Disease with Sildenafil Treatment (Walk-PHaSST) trial ($n = 181$), Duke University Outcome Modifying Genes study ($n = 152$), and Silent Infarct Transfusion (SIT) trial (SITT: $n = 488$). The meta-analysis was performed using the inverse variance method of the METAL software. Apart from SNP coordinates (chromosome and position), the summary statistics included all information necessary to perform meta-analysis, including dbSNP and Illumina SNP identifiers (1,198,700 SNPs in total). We also obtained complete GWAS summary statistics from the SITT cohort in which HbF was cubic root normalized. The summary statistics included SNP coordinates in the human reference build 36 (hg18), as well as dbSNP and Illumina SNP identifiers (1,138,137 SNPs in total).

4.6.6. Quality control (QC)

Genotype quality control was performed for batch 1 and 2 of our stage 1 GWAS data set separately. First, each batch of samples with gencall call rate $\geq 90\%$ (batch1 $n = 1137$, batch2 $n = 367$) was converted to plink binary file sets using PLINK2²⁸² while excluding duplicate SNPs. Samples that failed missingness criteria (outlying heterozygosity and missing genotype rate $> 10\%$) were excluded. Duplicate and related individuals (up to 2nd degree relationships) were identified using the Kinship-based **I**nference for **G**enome-wide association studies (KING v2.2.4) software²⁸³, and one individual from each pair of duplicate or related individuals was excluded. Apparently mislabeled samples were also identified using the KING software, and all samples that failed QC were excluded. SNP QC was performed by excluding SNPs with missing genotype rate $> 5\%$, minor allele frequency (MAF) $< 1\%$, and SNPs that failed the Hardy-Weinger equilibrium (HWE) test at a p-value threshold of $1e-6$, as well as palindromic [A/T] and [C/G] SNPs were also excluded. The two batches of genotype data were then merged using PLINK v1.9 *-bmerge*²⁸². Additional quality control on the merged data set was performed to exclude samples that failed missingness criteria, duplicate and/or related samples, SNPs with MAF $< 1\%$, SNPs with missing genotype rate $> 5\%$, and SNPs that failed the HWE test at $P = 1e-6$. To control for potential batch effects, SNPs with significant (p-value < 0.001) allele frequency difference (differential missingness) between the batch 1 and batch 2 data sets were excluded. In addition, principal component analysis (PCA) was performed on a set of high-quality independent SNPs using *smartpca* of the EIGENSOFT package (version 7.2.1)²⁸⁴ to investigate batch effects and to remove population and ancestry outliers. The independent set of SNPs were obtained by linkage disequilibrium (LD) pruning using the following parameters: linkage disequilibrium < 0.2 , window size of 50 bp, and step size of 10 bp. Population outliers were pruned with *smartpca* using the following pruning parameters: 10 principal components along which to remove outliers with 5 outlier removal iterations and specifying 6.0 standard deviations which an individual must exceed along one of the top 10 principal components to be excluded as an outlier. Only the merged genotype data was considered for subsequent analyses. We applied the same quality control procedure to the Tanzanian cohort genotype data.

4.6.7. Haplotype estimation (phasing) and genotype imputation

Haplotypes were estimated for the stage 1 and 2 genotype data sets separately using the 1000 Genomes reference panel^{230,285} for all autosomes and the X chromosome. First, the genotype data were aligned to the 1000 Genomes haplotype reference panel (phase 3, version 5) to ensure allele overlap with the reference panel using the *conform-gt* program from the BEAGLE utils (see Appendix VIII). SNPs that were absent in African populations in the reference panel, as well as SNPs with inconsistent strand and allele mismatch as compared to the reference panel were excluded. We then used the EAGLE v2.4.2 software²⁸⁶ to phase the data sets with the combined hapmap recombination map used to provide genetic distance, and set 20,000 conditioning haplotypes (-Kpbwt, default 10,000) to improve phasing accuracy. Genotypes were imputed from

six different panels: a custom panel created from whole genome sequence data of 50 individuals of Cameroonian origin (see *Creation of custom imputation panel*), the H3Africa panel consisting of ~3280 individuals from 17 African countries, the TOPMed panel consisting of ~180,000 individuals pooled from the NHLBI's studies of which 29% are of African ancestry ⁷⁸, the 1000 Genomes reference panel (KGP, phase 3 version 5) consisting of 661 individuals from West and East Africa ²⁸⁵, the Consortium on Asthma among African-ancestry populations in America (CAAPA) panel consisting of 883 individuals of African ancestry ⁷⁹, and the African Genome Resource (AGR) consisting of 4,956 individuals, 62% (~3061) of whom are of African ancestry mostly from eastern and southern Africa (~2501, 82% of all the African samples) ²³¹. The TOPMed panel was accessed via the TOPMed imputation web service, the KGP and CAAPA panels were accessed via the Michigan imputation web service ⁶⁵, the AGR was accessed via the Sanger imputation web service ²⁸⁷, while access to the H3Africa panel via the H3Africa imputation web service was granted upon request. In our in-house procedure, we used BEAGLE v5 ²⁸⁸ for imputation of each chromosome separately, leaving all default parameters and using the single chromosome hapmap recombination map. For the TOPMed and Michigan imputation web services ²⁸⁹, we selected the MINIMAC4 software ²⁹⁰ for imputation and retrieved only variants with imputation accuracy, $R^2 \geq 0.3$. For the Sanger imputation web service, we selected the Positional Burrows-Wheeler Transform (PBWT) package for imputation ²⁹¹. Imputed data from each panel and for each analysis stage were processed separately. Quality control of the imputed data included removal of variants with imputation accuracy (R^2) < 0.60 and genotype call rate < 95%. Only biallelic SNPs and INDELs were retained for subsequent analysis.

4.6.8. Creation of custom imputation panel

We used whole genome sequencing (WGS) data of 24 Cameroonian SCD patients, as well as WGS data of 26 individuals of Cameroonian origin that contributed to the H3Africa Trypanogen project for the custom panel creation. First, the quality of the FASTQ reads were checked using FastQC, and then mapped to the human reference genome in build 37 coordinate using BWA-MEM ²⁹². The resulting SAM files were converted to BAM and sorted by coordinate (chromosome and base pair position) using samtools ²⁹³. Duplicate reads were marked, and base quality scores recalibrated using GATK version 4.2.5.0 ²⁹⁴. For variant calling, we used a recently optimized pipeline involving DeepVariant version 1.3.0 for single-sample variant calling and GLNexus version 1.4.3 for joint variant calling (DV-GLN-OPT) ⁶³. In their optimisation and benchmarking study, Yun et al., developed a variant filtration scheme based on four tunable parameters which gave the DV-GLN-OPT an edge over the popular GATK-VQSR Best Practices pipeline for all data types analyzed (including whole-exome). These parameters have now been coded into the GLNexus package as the default settings and they were therefore utilized in our study. In addition, Yun et al., showed that reference imputation panel created using variant calls from the optimized pipeline outperformed that created using call sets from GATK best practices pipeline. To create our custom panel, we applied additional filters on our DV-GLN-OPT joint call set: we excluded

variants with read depth (DP) < 10, genotype quality (GQ) < 20, as well as monoallelic and singleton sites. We then phased each chromosome separately without reference using EAGLE v2.4.2 as previously described.

4.6.9. Association analysis

Association testing was performed using the Scalable and Accurate Implementation of GEneralized mixed model (SAIGE) software, version 0.38²⁹⁵. First, we extracted independent SNPs for each non-imputed dataset of the discovery and Tanzanian cohort through linkage disequilibrium pruning in PLINK2 according to the following parameters: window size of 500,000 base pairs (bp), step-size of 50 markers, and pairwise LD (r^2) < 0.2. Next, 20 principal components (PC) were computed for each of the datasets using the high-quality independent SNPs. Thereafter, a null generalized linear mixed model (GLMM) was fitted for each of the full non-imputed datasets including only SNPs with minor allele count (MAC) ≥ 20 as recommended²⁹⁵. A full genetic relationship matrix (GRM) calculated on the fly from the plink binary file sets was used to fit the null GLMM on the cubic root transformed HbF quantitative trait while including the top 10 PCs, as well as age and sex as covariates. Using the fitted null GLMMs for each cohort, single variant association tests were next performed for each imputed dataset filtered to include only biallelic SNPs and INDELs with minor allele frequency (MAF) ≥ 0.01 , imputation accuracy ≥ 0.6 , genotype call rate $\geq 95\%$, as well as SNPs that passed the HWE test at $P = 1e-06$. Association analysis in the Cameroonian cohort involved 827 samples, 52% of whom were females, and the average age of the participants was 17.61. In the Tanzanian cohort, 884 samples were analyzed, 53% of whom were females, and the average age of the participants was 13.19. The Benjamini-Hockberg false discovery rate (FDR) method implemented in the *p.adjust* function of the R statistical package²⁹⁶ was then used to correct for multiple testing.

4.6.10. Meta-analysis

We performed fixed effects meta-analysis using the METAL software²⁹⁷ on the basis that all the populations were of the same ethnic background. We used a two-step approach which constituted *Stage two* and *Stage three* of our GWAS analysis; i) GWAS *Stage two*: involved a meta-analysis of Cameroonian and Tanzanian cohorts using summary statistics from the *Stage one* GWAS, ii) GWAS *Stage three*: involved meta-analysis of Cameroonian, Tanzanian, and the USA-based cohorts. For accurately matching of markers across the studies, we standardized the variant IDs ("MarkerName") using "chromosome:position:SNV" (e.g., 2:60718043:SNV). Considering that summary statistics for the USA-based cohorts lacked coordinates (that is chromosome and position), and that the SITT cohort with coordinate information was mapped to build 36, we first updated the coordinates in the SITT cohort to build 37 (as well as to build 38 for meta-analysis with the TOPMed panel) using the UCSC liftOver tool. Briefly, we created a bed file from the build 36 coordinates and used it as input for liftOver. A total of 1,137,886 and 1,137,522 SNPs were successfully updated to build 37 and build 38, respectively. We then used the updated SITT cohort

coordinates to update the USA-based cohorts summary statistics using the variant ID column as the primary key. Therefore, only the 1,137,886 of 1,137,522 SNPs that were successfully updated in the SITT cohort were retained in the updated USA-based cohorts for meta-analysis. We noted similar effect size estimates and standard errors across the studies indicating similar phenotyping and normalization, and we therefore used the inverse variance method of METAL for meta-analysis. We used genomic control to account for population stratification, as well as allele frequency tracking to help identify allele flips. We also enabled heterogeneity analysis in which METAL computes the I^2 statistic (and corresponding p-values) which measures the amount of effect size variation across the studies that is due to heterogeneity rather than chance.

4.6.11. Statistical and functional fine mapping

We used the ‘sum of single effects’ (SuSiE) model implemented in the SusieR package ²⁹⁸ to fine-map functionally relevant variants in each region that showed significant association(s) in our analyses. SusieR employs an iterative Bayesian stepwise selection (IBSS) procedure that affords it the advantage of capturing uncertainty in which variable to select in its variable selection scheme and is thus well suited for highly correlated data. That is, the estimate of uncertainty provides a framework for determining which variant is most probably ‘causal’ in a scenario of completely (highly) correlated variants. For significant associations in the stage 1 GWAS results, we computed 95% credible sets using in-sample correlation (LD) matrices for Cameroon and Tanzania respectively. We used out-sample correlation matrices—computed from African samples in the 1000 Genomes reference panel—to compute 95% credible sets for significant associations from the meta-analysis results. All correlation matrices (specifically r as recommended) were calculated using Plink1.9. Regional association plots were then generated for each signal, highlighting the fine-mapped variants using LocusZoom ²⁹⁹. For loci that were significant in the independent cohort association tests and meta-analysis, fine mapping was performed for both results and the credible sets were merged together for functional mapping. Functional fine-mapping involved: (i) searching in the Genotype Tissue Expression (GTEx) portal whether the fine-mapped variants were expression quantitative or splicing quantitative trait loci (eQTLs and sQTLs respectively), (ii) searching in the ENSEMBL database for functional classifications, (iii) mapping their locations relative to the nearest gene, taking into account any evidence of recombination hotspots within the genomic area as represented in the regional association plots, (iv) investigating their occurrence in functionally relevant regions using the University of California Santa Cruz (UCSC) genome browser tracts, including chromatin state segmentation by the ChromHMM algorithm, transcription factor binding site (TFBS) by the JASPAR algorithm, gene-enhancer interaction by the GeneHancer algorithm, transcription factor (TF) chromatin immunoprecipitation sequencing (ChIP-seq) peaks from the ENCODE project, etc, and (v) investigation of enhancer classifications in the ENCODE and VISTA Enhancer databases. Sequence logos of binding motifs for transcription factors whose motifs were affected by the fine-mapped variants were obtained from the JASPAR website (Appendix VIII).

4.6.12. Gene set analysis and functional mapping

The FUMA v1.5.4 online platform³⁰⁰ was used for functional annotation of the GWAS results as follows: all summary statistics in GRCh37 coordinate were uploaded to the SNP2GENE algorithm using the following parameters: p-value threshold for lead SNP = $5e-7$ and minimum LD for defining lead SNP = 0.4. We selected the African populations (AFR) of the 100 Genomes projects (KGP) as reference, while all other default parameters were used (Appendix VIII). SNP2GENE uses the ANNOVAR tool to functionally annotate independent lead SNPs and their LD tags and map them to their corresponding genes. Prioritized genes based on positional, eQTL, and chromatin interaction mapping, are then processed with the GENE2FUNCTION algorithm to obtain insight into putative biological mechanisms and pathways. In addition to functional mappings, FUMA also performs gene-based test and gene set analysis as implemented in the MAGMA v1.08 tool³⁰¹. Specifically, for gene-based tests, MAGMA uses the SNP p-values from the summary statistics to compute Chi-Square statistics for a gene with LD generated from a reference panel, and the mean or the top Chi-Square statistic is taken as the gene test statistic. FUMA implements the mean model in which a gene p-value is obtained by using a known approximation of the sampling distribution, and the significance threshold is calculated by $0.05/\text{number of mapped genes}$ (Bonferroni correction). In our study, the threshold was $2.517e-6$ ($P = 0.05/19867$). For gene-set analysis, the gene p-value for each gene from the gene-based analysis is converted to a Z-value that reflects the strength of the association of each gene with the phenotype.

4.6.13. *FLT1* haplotype structure and association analysis

Haplotype analysis was performed for the Cameroonian and Tanzanian cohorts using haploview v4.2³⁰². First, the *FLT1* 40kb region and 25kb upstream and downstream was extracted for each of the imputed datasets using PLINK1.9 according to the following coordinates: GRCh38 28476218 – 28549906, GRCh37 29050355 – 29124043. At the same time, the HbF phenotype was transformed into case-control whereby cases were defined as HbF level $\geq 8.6\%$ and controls as HbF level $\leq 3.1\%$ as previously described³⁰³. A total of 408 variants and 520 samples were retained in the Cameroonian cohort, of which 413 were cases and 107 were controls (240 males and 280 females). A total of 464 variants and 448 samples were retained in the Tanzanian cohort, of which 152 were cases and 296 were controls (228 males and 220 females). PLINK1.9 was used to recode the data into haploview format. The default quality filters of haploview we used; no sample or SNP failed any of the filters. Haplotype blocks were computed using the Gabriel block definition³⁰⁴, i.e. 95% confidence bounds on D prime (D') are generated and each comparison is called "strong LD", "inconclusive" or "strong recombination". A block is created if 95% of informative (i.e. non-inconclusive) comparisons are "strong LD". Variants with MAF < 0.05 were not included in block calculations, and blocks were non-overlapping. LD plots showing haplotype blocks were generated and the *FLT1* fine-mapped variants were generated by haploview. The chromosome 13 hapmap recombination map was used as a track file to generate a recombination plot along side the LD plot, as well as to highlight the *FLT1* fine-mapped variants. Finally, single variant and haplotype

association tests were performed using 100,000 permutation. Haplotype structure analysis was also performed for the non-SCD Cameroonians (n=25) that contributed to the custom panel creation, as well as for genomes of individuals that were negative for the sickle mutation from the 1000 Genomes Project that.

4.6.14. Assessment of transferability/replication of signals

Genomic regions associated with HbF level were identified from the respective summary statistics. Recent genetically nominated HbF-influencing loci were also identified through literature search. From each cohort in which the significant signal was absent, we extracted the loci (genes) including 100kb downstream and 100kb upstream given that some cis-regulatory elements could be tens of thousands of bases away. We next looked up variants within the extracted region, and replication or transferability of the signal was defined as occurrence of variants at $P < 0.05$.

4.6.15. Estimation of HbF heritability

We first combined the VCF files from both cohorts that were aligned to the 1000 Genomes panel during preparation for imputation. There were 1711 samples all together. Next, we extracted only biallelic SNPs across autosomes that passed the following filtering criteria: MAF $> 1\%$, missing genotype rate $< 5\%$, individual missingness $< 10\%$, HWE p-value of $1e-06$, and individuals with average heterozygosity within three standard deviations of the mean heterozygosity. Twenty nine individuals with outlying heterozygosity were excluded, while no SNPs were excluded based on missingness criteria. To assess potential batch effects, we calculated differential missingness among the two cohorts. No SNPs were excluded due to differential missingness test at p-value < 0.01 . We then estimated HbF heritability in the resulting cohort of 1682 high quality samples of Cameroonian and Tanzanian SCA patients using the Randomized Haseman–Elston regression for Multi-variance Components (RHE-mc) software ²⁴³, which jointly estimates additive and dominance genetic variance components. We set the number of random vectors (K) to 10, and the number of block Jackknifes (B) for standard error estimation to 1000 as recommended. In addition, we included four sets of principal components (20PCs, 30PCs, 50PCs, and 100PCs), as well as age, sex, and country as covariates. To accurately capture the effect of age, we performed another set of analyses in which we used the square of age as an additional covariate.

4.6.16. HUDEP-2 and K562 cell culture, differentiation, and hypoxia treatment

HUDEP-2 and K562 cells were expanded in SFEM media (Stem Cell Technologies, 09650) supplemented with 50 ng/mL recombinant human SCF (Peprotech, 300-07) 3 units/mL EPO, 1 µg/mL doxycycline (Sigma Aldrich, D9891), 0.4 µg/mL dexamethasone (Sigma Aldrich, D4902), and 1% Penicillin-Streptomycin solution. HUDEP2 cells were differentiated using a 2-phase protocol. During phase 1 (days 0-3), cells were cultured at 0.5E6 cells/mL-1.5E6 cells/mL in IMDM with 2% fetal bovine serum, 2% human blood type AB plasma (Seracare, 1810-0001), 1% penicillin/streptomycin, 3 units/mL heparin, 10 µg/mL insulin (Sigma, I9278), 3 units/mL EPO,

1mg/mL holo-transferrin (Millipore Sigma, T0665), 50 ng/mL SCF and 1 µg/mL doxycycline. After 3 days of culture in phase 1, the media was replaced with fresh media containing the same ingredients but without SCF, and cultured at 1E6 cells/mL-2E6 cells/mL for 7 additional days (10 days total). For hypoxic treatment, cells were differentiated, and sample collections and media changes were performed within a Whitley H35 HEPA Hypoxystation incubator at 2% O₂.

Chapter 5: General Discussion and Conclusion

5.1. Introduction

Progress made in the fight against infectious diseases in sub-Saharan Africa has exposed the significant deficiencies in the recognition of the public health burden of rare monogenic conditions, particularly sickle cell disease. The absence of any comprehensive newborn screening for sickle cell disease in many parts of sub-Saharan Africa and consequent paucity of data implies that current SCD prevalence and mortality burden estimates are underrated. The extent of SCD burden is reflected in the high carrier frequency of the sickle gene in sub-Saharan Africa in spite of its high recessive lethality—so eloquently captured by the statement of Dr E. A. Beet. In this thesis project, we set out generally: (i) to examine the evolutionary history of the sickle cell mutation including its time and place of origin, its interaction with other gene variants in sub-Saharan Africa that have either been established to be under adaptive evolution or show signs of adaptive evolution with consequent impact on sickle cell disease severity; (ii) to describe the burden of sickle cell disease in sub-Saharan including strategies that have been successful for management and care in high-resource settings and that can be effectively transferred to sub-Saharan Africa; and (iii) to use the genome-wide association approach to investigate new HbF-influencing loci that could serve as therapeutic targets for genetic manipulation.

5.2. Evolutionary history of the sickle cell mutation and opportunities for further research

That the sickle cell mutation arose more than 22,000 years ago means that SCD has potentially been affecting humans for more than 20,000 years. Its emergence in sub-Saharan Africa in a region where *Plasmodium falciparum* malaria is believed to have been subtertian (since about 40,000 - 60,000 years ago) certainly fits a model in which the gene is directly introduced into a selective environment that prevents its immediate purging by negative selection. However, it is possible that the highest disease burden (prevalence and mortality) has only been experienced in the last 5,000 years; the dawn of the Holocene ~12,000 years ago saw a wetter and greener Sahara, the formation and expansion of water bodies such as the Lake Chad and Lake Victoria, the glaciation of mountains like the Mount Kilimanjaro, and the adoption of agriculture around 8,000 - 10,000 years ago that eventually transformed small mobile hunter-gatherer groups into larger settlements of agriculturalist³⁰⁵, and promoting rapid population expansions, interactions with infectious agents such as the malaria parasite, increasing deleterious mutational loads, and consequent evolutionary pressure¹¹⁰. Further evidential support for the introduction of the sickle cell mutation into a malaria selective environment around West-Central Africa comes from a recent study involving ancient human DNA from sub-Saharan Africa which showed the remarkable presence of the protective Duffy-null variants against the *Plasmodium vivax* malaria parasite exclusively in the genomes of ~3,000 - 8,000 years old Cameroonian samples^{306,307}, thus demonstrating the strong adaptive evolution that was well underway in this region in response to malaria. Data presented in *chapter two* of this thesis show that there is great opportunity to further understand the evolution of SCD in Africa, and general patterns of human migration

through the lens of the sickle cell mutation and the *HBB* gene cluster haplotypes. For instance, the presence of the Arab-India haplotype in some African genomic backgrounds raises the question whether it actually arose in Africa and spread to South Asia, or whether it was introduced in Africa through back migration. While there is substantial evidence for an origin of the sickle cell mutation in West-Central Africa, there is still opportunity to determine when this occurred, and to fine-map its exact location. This would offer a better appreciation of the extent of the pressure that the mutation has exerted on the human genome.

5.3. Co-evolution of sickle cell disease and other disease-related adaptive signatures

The systemic nature of sickle cell disease means that it has a vast array of molecular connections, thereby exposing SCD patients to a great many pathological conditions with a concomitant increased risk of death. The interactions of SCD with malaria and other diseases presented in this thesis are likely just a subset of the possibly numerous interactions accrued from the over 300,000 years of modern human evolution in Africa. Indeed, infectious disease and changes in dietary lifestyle have been the preponderant drivers of genetic diversity in Africa, and this is exemplified by the enormous diversity in the immune system (human leukocyte antigen - HLA) and red blood cell receptor gene variants. For instance, another diseases-related adaptive evolutionary signature in the human genome that was not presented here and that should be explored for a possible interaction with SCD is the interferon lambda 4 (*IFNL4*) gene variants which protect European and Asian populations against viral infections but predispose Africans to a less effective response to the latter, with an increased susceptibility to several other common infections including malaria and gastrointestinal infections^{308–310}. While the variant has been negatively selected and almost completely purged outside of Africa, it remains at intermediate frequencies in Africa^{311,312}. This and other unknown adaptive signatures that might potentially interact with SCD could be explored using evolutionary genetic techniques. For instance, a study looking at the evolutionary history of human disease genes found that disease pairs connected by similar evolutionary rates showed a nearly 2-fold increase in comorbidity in comparison with unconnected disease pairs³¹³. Interestingly, genes involved in hematological conditions were the third fastest evolving after immunological and respiratory disease classes, consistent with immense evolutionary pressure on these systems. The discovery of the genetic basis of pathology in these connections, if they modify SCD outcome, would be vital in guiding management and developing therapeutics. In addition, as genetic testing becomes more readily available and affordable, it can be forecasted that patients scanning for pathogenic genes will likely be screened for epistatic (interacting) genes as well³¹⁴ for better tailoring of management options, much like pharmacogenomics.

5.4. SCD management: newborn screening as the main pivot

The reduction of childhood mortality largely through a combination of newborn screening and the implementation of simple interventions to prevent against infections is one of the major successes of SCD research. Newborn screening is pivotal to the management of SCD, and with the development of inexpensive point-of-care tests, universal newborn screening for sickle cell

disease is feasible in Africa. Therefore, new collaborative initiatives like the Sickle In Africa Consortium and The Consortium on Newborn Screening in Africa for Sickle Cell Disease (CONSA; which aims to implement standardized newborn hemoglobinopathy screening and early intervention for children with SCD in sub-Saharan Africa ³¹⁵) are much needed to raise awareness of the disease in all sectors through education and advocacy campaigns, to champion the implementation of newborn screening at scale, and to eventually close the knowledge gap on the disease burden, especially on the exact contribution of the other less common genotypes, i.e. HbS/C and HbS/beta-thalassemia. These initiatives will also provide a massive opportunity to study the natural disease history in its original ecosystem, as well as the response to different interventions through integration with multicenter longitudinal studies. The major challenge of handling of positive test results that universal newborn screening will present should be mitigated by investment, from African governments, in expanding facilities providing confirmatory tests and equipping them with the necessary tools, providing robust genetic counselling services, and adequate personnel training.

While current approaches for management and care of SCD, explored in *chapter three* of this thesis, significantly improve survival, the cost of management becomes astronomical over time. Currently, the cost of managing SCD throughout a patient's life (to age 50) is estimated to exceed 8 million US Dollars (\$), far out of the reach of the majority of patients, especially those in Africa. Some of the most effective management options like hydroxyurea are also not without limitations. A once-off treatment like hematopoietic stem cell transplantation (HSCT), though relatively expensive (costing between \$344 029–\$799 219) ³¹⁶, would be much cost-effective than long-term management. However, the limitations of HSCT that make it unavailable to more than 80% of patients means that better alternatives are needed. The promise of therapy for SCD by reactivating fetal hemoglobin through genetic manipulation, manifest in the multiple clinical trials that are underway, calls for the search for more, perhaps better, targets.

5.5. Genomics of fetal hemoglobin

In *chapter four*, we explored novel genetic loci that might explain additional variability in fetal hemoglobin level in African SCD patients, and potentially serve as therapeutic targets. Our identification of fourteen novel loci supports the notion that African genomes hold the key to sickle cell disease gene therapy. The high HbF heritability estimate in African SCD patients, with less than 30% accounted for by all the loci described so far, means that there is still opportunity for novel discoveries. The loci described here, particularly *FLT1*, hold tremendous promise as potential SCD therapeutic targets. Generally, insight from our *in silico* and *in vitro* functional analyses has already shed significant light on pathophysiological pathways that are relevant to erythropoiesis, and this could reveal other therapeutic targets. For instance, a central but previously unrecognized role for *KDR* in stress erythropoiesis is becoming increasingly apparent from our functional data. These functional studies are proceeding and are expected to further clarify the role of *KDR* and some of the other members of the protein interaction network linking hypoxia response and erythropoiesis. Furthermore, a close examination of several of the loci that

only reached suggestive level of significance revealed a clear enrichment in pathways relevant to erythropoiesis and HbF production (as was also apparent in the functional genome-wide analysis). For example, *ADSS1*, detected in our custom panel in the meta-analysis of Cameroonian and Tanzanian cohorts at p-value 6.05E-07, and *ATIC*, detected in the association analysis of the merged genotype dataset of the two cohorts at p-value 7.38E-07, are enzymes involved in the de novo biosynthesis of purine nucleotides which are required for DNA synthesis ³¹⁷. Notably hydroxyurea acts by inhibiting DNA synthesis through inhibition of ribonucleotide reductase which is responsible for the biosynthesis of deoxynucleotides from nucleotides ³¹⁸. Therefore, variants that interfere with the function of *ADSS1* and *ATIC* may impair DNA synthesis and exhibit a hydroxyurea-like response. With rigorous consideration of the loci presented here, both significant and suggestive, our study therefore holds tremendous potential, not only in identifying potential targets for SCD gene therapy, but also in unravelling important SCD pathophysiological mechanisms.

Core to our analysis was the previously unexplored genome-wide SNP data from Cameroonian patients generated using the H3Africa chip that is proving to be immensely important for African GWAS. The chip was recently used to discover three novel HbF-associated loci in a Nigerian sickle cell cohort, namely *SLC28A3*, *TICRR*, and *PIEZO2* ⁵³. The generally higher average imputation accuracy in the Cameroonian cohort than in the Tanzanian cohort also reflected the power of the H3Africa chip over the illumina Omni2.5-8 array for African GWAS. Therefore, the *FLT1* novel locus detected in Cameroonians may not have been seen using the illumina Omni2.5-8 chip given that the significant *FLT1* associations are completely absent or very rare in non-African populations who are the basis, for the most part, of the illumina Omni2.5-8 chip. There is still, however, a great opportunity for improvement of the H3Africa array. For instance, the array currently has similar SNP density as the illumina Omni2.5-8 chip (~2.4 million SNPs) although it is well established that the higher genetic diversity of African genomes is generally characterized by smaller haplotype blocks, meaning shorter-ranged linkage disequilibrium (LD). Therefore, certain regions of European genomes requiring a single tag SNP, for instance, might require many times more tag SNPs in African genomes—the exact extent of which could be an interesting subject for further investigation, such as by using some robust mathematical or statistical models for quantitative comparisons of genome-wide haplotype blocks in the different ancestries. The point is that an Africa-specific genotyping array should generally have a higher SNP density than the illumina Omni2.5-8 array.

Another major novelty of our study is the multi-imputation panel and association strategy which, to the best of our knowledge, makes our study the most comprehensive GWAS of HbF in sickle cell patients, if not all GWAS involving African populations, thus far. We note that the KGP panel has been the most popular and the basis of many studies involving African populations for over the years, including the GWAS previously conducted with Tanzanian samples. Our failure to detect the *FLT1* novel locus in the panel exposes a critical limitation that might have hindered discovery in previous studies. These observations suggest that a reappraisal of GWAS studies that have been based on the KGP panel, this time utilizing more diverse and larger panels might reveal

additional (previously missed) loci. Besides, creating a custom imputation panel can be challenging and considering the limitation of the KGP panel, as well as the fact that the H3A panel is not yet publicly available (access was granted upon request), it means that the Sanger, TOPMed, and CAAPA panels are currently the most valuable genotype imputation resources publicly available for African ancestry populations. However, accessing the Sanger panel requires the use of globus (a secure file transfer application) which can be particularly challenging to users, thus placing a constraint on utilizing the panel. It must also be mentioned that the relatively new imputation panel suitable for imputing Africans, AFAM, could not be utilized here because although it is apparently publicly available, it is not free accessible.

5.6. Limitations

A key limitation of our genomics study was the small sample size for each cohort which hindered per cohort heritability analysis. Even with the merged genotype dataset, partitioning of heritability by chromosome or functional enrichment category could not be archived. Therefore, larger samples sizes across multiple African populations are needed to confirm the true heritability of HbF among sickle cell patients. Another factor that apparently limited discovery of new loci was ascertainment bias associated with the summary statistics that were obtained for the USA-based cohorts. The summary statistics contained just over 1.3 million SNPs with only about 1 million that were effectively mapped across all the studies (Cameroon, Tanzania, and USA) and used in meta-analysis as compared to over 10 million variants (SNPs and INDELs) between Cameroon and Tanzania. Therefore, many potential associations that would have been apparent after global meta-analysis were likely missed. Besides, the complete merged summary statistics representing a combined sample size of more than 3,700 was still of insufficient power to determine heritability using summary statistics-based algorithms.

5.7. Conclusion and perspectives

This project presented data with overwhelmingly support for a single African origin of the sickle cell mutation, with opportunity for further research on determining the true age of the mutation, the place of origin, and historical disease burden. One way to approach these would be to investigate the population demographic history of SCD populations across Africa from whole genome sequencing data, including the estimation of the effective population size at different epochs from the estimated time of emergence of the mutation, the historical population genetic structure, and migration patterns. Luckily, the rapid evolution of the field of bioinformatics has seen the development of algorithms that can measure these variables to unprecedented resolutions. It is also evident that ancient genomes are important tools for better understanding the evolutionary history (historical burden) of human diseases and should thus be incorporated in SCD evolutionary genetic studies.

Tracking the movement of the sickle cell mutation through the distribution of the *HBB* gene cluster haplotypes highlighted changing population dynamics that are important for public health. With the increasing availability of whole-genome sequencing data for African sickle cell disease patients, the *HBB* gene cluster haplotypes should be better classified and examined for a

much clearer association with sickle cell disease severity. In addition, focus should also be placed on genetic determinants of co-morbidities affecting SCD severity. Progress in implementing newborn screening and comprehensive care in some sub-Saharan African countries has been encouraging. Early diagnosis and family education on management can reduce morbidity, while immunisation and hydroxyurea therapy initiated in affected children as young as 9 months old can greatly improve quality and quantity of life. Research is therefore urgently needed in sub-Saharan Africa to establish the exact prevalence, mortality, and morbidity, environmental and genetic factors affecting clinical complications, and life expectancy of patients with sickle cell disease, to facilitate the design of future risk models and to investigate novel routes for therapeutic options with the ultimate aim of improving the clinical outcomes of patients with sickle cell disease in all parts of the world.

Gene therapy for SCD is promising, and our study adds significantly to the growing list of potential targets for genetic manipulation. Our study approach was comprehensive but not exhaustive in terms of identifying HbF-influencing loci. Only SNPs and small INDELs with prevalence greater than 1% were examined. But rare variants might also contribute significantly to HbF variability, and large structural variants do contribute to the deletional HPFH. There also could be other uncharacterised variants (large insertions, inversions, duplications, and gene conversions for instance) that contribute significantly to HbF variability, but are poorly or not tagged at all by current genotyping chips. These premises are supported by the large missing heritability of HbF, and should be best investigated using long-read whole-genome sequencing. This will provide a more comprehensive understanding of the genetic architecture of HbF in SCD and may reveal additional, perhaps more potent, therapeutic targets.

An important area of SCD research that has surprisingly eluded the attention of researchers is to investigate genetic determinants of SCD modifiers from an evolutionary perspective, such as through natural selection mapping. One would expect that the long evolutionary history of the sickle cell mutation with an associated high lethality would mean that sickle cell anemia has exerted enormous pressure, in its own right, on the human genome. While the human genome enlisted the sickle cell mutation for malaria protection, thus maintaining sickle cell disease in the population, it might have also, over time, enlisted protective variants through natural selection to protect against sickle cell disease. It is therefore worthwhile to scan the genomes of sickle cell patients for any adaptive signatures of long-term survival.

Lastly, and perhaps innovatively, a creative way of investigating novel HbF-influencing loci could be to scan the genome for erythroid-specific enhancers or regulatory elements which could be genetically manipulated. One of the impetuses to use *BCL11A* as a therapeutic target was the discovery of erythroid-specific enhancers that allow for targeted disruption. The growing body of chromatin state data from chromatin immunoprecipitation sequencing (ChIP-Seq) and ATAC-Seq experiments have allowed the development of algorithms which can predict regulatory genomic regions like enhancers with remarkable accuracy. An example is the ChromHMM algorithm²³⁴ whose predictions are publicly available and form tracks in the UCSC genome browser or can be

accessed with the LocusZoom tool ²⁹⁹. One could, for instance, scan all high-scoring erythroid-specific enhancers to generate a list of loci that could be further prioritized based on other variables such as co-expression and co-localisation with known HbF regulators.

References

1. Thomson, A. M. *et al.* Global, regional, and national prevalence and mortality burden of sickle cell disease, 2000–2021: a systematic analysis from the Global Burden of Disease Study 2021. *Lancet Haematol.* **10**, e585–e599 (2023).
2. Piel, F. B., Hay, S. I., Gupta, S., Weatherall, D. J. & Williams, T. N. Global Burden of Sickle Cell Anaemia in Children under Five, 2010–2050: Modelling Based on Demographics, Excess Mortality, and Interventions. *PLoS Med.* **10**, e1001484 (2013).
3. Herrick, J. B. Peculiar elongated and sickle-shaped red blood corpuscles in a case of severe anemia. *Arch. Intern. Med.* **VI**, 517–521 (1910).
4. Mason, V. R. Sickle cell anemia. *J. Am. Med. Assoc.* **79**, 1318–1320 (1922).
5. Taliaferro, W. H. & Huck, J. G. THE INHERITANCE OF SICKLE-CELL ANAEMIA IN MAN. *Genetics* **8**, 594–598 (1923).
6. Beet, E. A. The Genetics of the Sickle-Cell Trait in a Bantu Tribe. *Ann. Eugen.* **14**, 279–284 (1947).
7. Neel, J. V. The Inheritance of Sickle Cell Anemia. *Science* **110**, 64–66 (1949).
8. Pauling, L. *et al.* ‘Sickle cell anemia, a molecular disease’. *Science* **286**, 1488–1490 (1949).
9. Ingram, V. M. A Specific Chemical Difference Between the Globins of Normal Human and Sickle-Cell Anæmia Hæmoglobin. *Nature* **178**, 792–794 (1956).
10. Hunt, J. A. & Ingram, V. M. Allelomorphism and the Chemical Differences of the Human Hæmoglobins A, S and C. *Nature* **181**, 1062–1063 (1958).
11. Grosse, S. D. *et al.* Sickle Cell Disease in Africa: A Neglected Cause of Early Childhood Mortality. *Am. J. Prev. Med.* **41**, S398–S405 (2011).
12. Payne, A. B. *et al.* Trends in Sickle Cell Disease–Related Mortality in the United States, 1979 to 2017. *Ann. Emerg. Med.* **76**, S28–S36 (2020).
13. Serjeant, G. R. & Serjeant, B. E. Management of Sickle Cell Disease; Lessons from the Jamaican Cohort Study. *Blood Rev.* **7**, 137–145 (1993).
14. Farber, M. D., Koshy, M., Kinney, T. R., & the Cooperative Study Of Sickle Cell Disease. Cooperative Study of Sickle Cell Disease: Demographic and socioeconomic characteristics of patients and families with Sickle Cell Disease. *J. Chronic Dis.* **38**, 495–505 (1985).
15. Kanter, J. & Falcon, C. Gene therapy for sickle cell disease: where we are now? *Hematology* **2021**, 174–180 (2021).
16. Kanter, J. *et al.* Biologic and Clinical Efficacy of LentiGlobin for Sickle Cell Disease. *N. Engl. J. Med.* **386**, 617–628 (2022).
17. Larkin, H. D. Sickle Cell Disease Gene Therapy Trial Paused. *JAMA* **329**, 364 (2023).
18. Weatherall, D. J. & Clegg, J. B. Inherited haemoglobin disorders: An increasing global health problem. *Bull. World Health Organ.* **79**, 704–712 (2001).
19. Levings, P. P. & Bungert, J. The human β -globin locus control region. A center of attraction. *Eur. J. Biochem.* **269**, 1589–1599 (2002).

20. Sankaran, V. G. & Orkin, S. H. The switch from fetal to adult hemoglobin. *Cold Spring Harb. Perspect. Med.* **3**, a011643 (2013).
21. Chernoff, A. I. The Human Hemoglobins in Health and Disease. *N. Engl. J. Med.* **253**, 365–374 (1955).
22. Jacob, G. F. & Raper, A. B. Hereditary Persistence of Foetal Haemoglobin Production, and its Interaction with the Sickle-Cell Trait. *Br. J. Haematol.* **4**, 138–149 (1958).
23. Jackson, J. F., Odom, J. L. & Bell, W. N. Amelioration of Sickle Cell Disease by Persistent Fetal Hemoglobin. *JAMA* **177**, 867–869 (1961).
24. Letvin, N. L., Linch, D. C., Beardsley, G. P., McIntyre, K. W. & Nathan, D. G. Augmentation of Fetal-Hemoglobin Production in Anemic Monkeys by Hydroxyurea. *N. Engl. J. Med.* **310**, 869–873 (1984).
25. Platt, O. S. *et al.* Hydroxyurea enhances fetal hemoglobin production in sickle cell anemia. *J. Clin. Invest.* **74**, 652–656 (1984).
26. Charache, S. *et al.* Effect of Hydroxyurea on the Frequency of Painful Crises in Sickle Cell Anemia. *N. Engl. J. Med.* **332**, 1317–1322 (1995).
27. Research, C. for D. E. and. FDA approves hydroxyurea for treatment of pediatric patients with sickle cell anemia. *FDA* (2019).
28. Ault, A. US FDA approves first drug for sickle-cell anaemia. *The Lancet* **351**, 809 (1998).
29. Esoh, K., Wonkam-Tingang, E. & Wonkam, A. Sickle cell disease in sub-Saharan Africa: transferable strategies for prevention and care. *Lancet Haematol.* **8**, e744–e755 (2021).
30. Kan, Y. W., Holland, J. P., Dozy, A. M., Charache, S. & Kazazian, H. H. Deletion of the β -globin structure gene in hereditary persistence of foetal haemoglobin. *Nature* **258**, 162–163 (1975).
31. Fritsch, E. F., Lawn, R. M. & Maniatis, T. Characterisation of deletions which affect the expression of fetal globin genes in man. *Nature* **279**, 598–603 (1979).
32. Collins, F. S., Cole, J. L., Lockwood, W. K. & Iannuzzi, M. C. The Deletion in Both Common Types of Hereditary Persistence of Fetal Hemoglobin is Approximately 105 Kilobases. *Blood* **70**, 1797–1803 (1987).
33. Antonarakis, S. E., Boehm, C. D., Giardina, P. J. V. & Kazazian, H. H. Nonrandom association of polymorphic restriction sites in the β -globin gene cluster. *Proc. Natl. Acad. Sci.* **79**, 137–141 (1982).
34. Gilman, J. G. & Huisman, T. H. J. DNA Sequence Variation Associated With Elevated Fetal Gy Globin Production. *Blood* **66**, 783–787 (1985).
35. Collins, F. S., Stoeckert, C. J., Serjeant, G. R., Forget, B. G. & Weissman, S. M. G gamma beta+ hereditary persistence of fetal hemoglobin: cosmid cloning and identification of a specific mutation 5' to the G gamma gene. *Proc. Natl. Acad. Sci.* **81**, 4894–4898 (1984).
36. Garner, C. *et al.* Genetic influences on F cells and other hematologic variables: A twin heritability study. *Blood* **95**, 342–346 (2000).

37. Menzel, S. & Thein, S. L. Genetic Modifiers of Fetal Haemoglobin in Sickle Cell Disease. *Mol. Diagn. Ther.* **23**, 235–244 (2019).
38. Sankaran, V. G. *et al.* Human Fetal Hemoglobin Expression Is Regulated by the Developmental Stage-Specific Repressor BCL11A. *Science* **322**, 1839–1842 (2008).
39. Bauer, D. E. *et al.* An Erythroid Enhancer of BCL11A Subject to Genetic Variation Determines Fetal Hemoglobin Level. *SCIENCE* **342**, 253–257 (2013).
40. Menzel, S. *et al.* A QTL influencing F cell production maps to a gene encoding a zinc-finger protein on chromosome 2p15. *Nat. Genet.* **39**, 1197–1199 (2007).
41. Galarneau, G. *et al.* Fine-mapping at three loci known to affect fetal hemoglobin levels explains additional genetic variation. *Nat. Genet.* **42**, 1049–1051 (2010).
42. Bao, E. L. *et al.* Heritability of fetal hemoglobin, white cell count, and other clinical traits from a sickle cell disease family cohort. *Am. J. Hematol.* **94**, 522–527 (2019).
43. Cato, L. D. *et al.* Genetic regulation of fetal hemoglobin across global populations. 2023.03.24.23287659 Preprint at <https://doi.org/10.1101/2023.03.24.23287659> (2023).
44. Sl, T. & Dj, W. A non-deletion hereditary persistence of fetal hemoglobin (HPFH) determinant not linked to the beta-globin gene complex. *Prog. Clin. Biol. Res.* **316B**, 97–111 (1989).
45. Craig, J. E. *et al.* Dissecting the loci controlling fetal haemoglobin production on chromosomes 11p and 6q by the regressive approach. *Nat. Genet.* **12**, 58–64 (1996).
46. Thein, S. L. *et al.* Intergenic variants of HBS1L-MYB are responsible for a major quantitative trait locus on chromosome 6q23 influencing fetal hemoglobin levels in adults. *Proc. Natl. Acad. Sci. U. S. A.* **104**, 11346–11351 (2007).
47. Uda, M. *et al.* Genome-wide association study shows BCL11A associated with persistent fetal hemoglobin and amelioration of the phenotype of β -thalassemia. *Proc. Natl. Acad. Sci. U. S. A.* **105**, 1620–1625 (2008).
48. Makani, J. *et al.* Genetics of fetal hemoglobin in Tanzanian and British patients with sickle cell anemia. *Blood* **117**, 1390–1392 (2011).
49. Mtatiro, S. N. *et al.* Genome wide association study of fetal hemoglobin in sickle cell anemia in Tanzania. *PloS One* **9**, e111464 (2014).
50. Wonkam, A. *et al.* Association of variants at BCL11A and HBS1L-MYB with hemoglobin F and hospitalization rates among sickle cell patients in Cameroon. *PloS One* **9**, e92506 (2014).
51. Dover, G. J. *et al.* Fetal Hemoglobin Levels in Sickle Cell Disease and Normal Individuals Are Partially Controlled by an X-Linked Gene Located at Xp22.2. *Blood* **80**, 816–824 (1992).
52. Urio, F. *et al.* F cell numbers are associated with an X-linked genetic polymorphism and correlate with haematological parameters in patients with sickle cell disease. *Br. J. Haematol.* **bjh.17102** (2020) doi:10.1111/bjh.17102.

53. Ojewunmi, O. O. *et al.* The genetic dissection of fetal haemoglobin persistence in sickle cell disease in Nigeria. 2023.05.16.23289851 Preprint at <https://doi.org/10.1101/2023.05.16.23289851> (2023).
54. Dai, Y. *et al.* BCL2L1 is associated with γ -globin gene expression. *Blood Adv.* **3**, 2995–3001 (2019).
55. Vinjamur, D. S. *et al.* ZNF410 represses fetal globin by singular control of CHD4. *Nat. Genet.* **53**, 719–728 (2021).
56. Wongborisuth, C. *et al.* Down-regulation of the transcriptional repressor ZNF802 (JAZF1) reactivates fetal hemoglobin in β 0-thalassemia/HbE. *Sci. Rep.* **12**, 4952 (2022).
57. Feng, R. *et al.* Regulation of Fetal Hemoglobin Expression By the VHL-HIF1 α Oxygen Sensing System. *Blood* **138**, 574 (2021).
58. Wonkam, A. The future of sickle cell disease therapeutics rests in genomics. *Dis. Model. Mech.* **16**, dmm049765 (2023).
59. Sollis, E. *et al.* The NHGRI-EBI GWAS Catalog: knowledgebase and deposition resource. *Nucleic Acids Res.* **51**, D977–D985 (2023).
60. Spencer, C. C. A., Su, Z., Donnelly, P. & Marchini, J. Designing genome-wide association studies: Sample size, power, imputation, and the choice of genotyping chip. *PLoS Genet.* **5**, (2009).
61. Marchini, J. & Howie, B. Genotype imputation for genome-wide association studies. *Nat. Rev. Genet.* **11**, 499–511 (2010).
62. Das, S., Abecasis, G. R. & Browning, B. L. Genotype Imputation from Large Reference Panels. *Annu. Rev. Genomics Hum. Genet.* **19**, 73–96 (2018).
63. Yun, T. *et al.* Accurate, scalable cohort variant calls using DeepVariant and GLnexus. *Bioinformatics* **36**, 5582–5589 (2020).
64. Sengupta, D. *et al.* Performance and accuracy evaluation of reference panels for genotype imputation in sub-Saharan African populations. *Cell Genomics* **3**, 100332 (2023).
65. Das, S. *et al.* Next-generation genotype imputation service and methods. *Nat. Genet.* **48**, 1284–1287 (2016).
66. Esoh, K. K. *et al.* Genome-wide association study identifies novel candidate malaria resistance genes in Cameroon. *Hum. Mol. Genet.* **32**, 1946–1958 (2023).
67. Yu, K. *et al.* Meta-imputation: An efficient method to combine genotype data after imputation with multiple reference panels. *Am. J. Hum. Genet.* **109**, 1007–1015 (2022).
68. Cooper, A. *et al.* APOL1 renal risk variants have contrasting resistance and susceptibility associations with African trypanosomiasis. *eLife* **6**, (2017).
69. NIH launches new collaboration to develop gene-based cures for sickle cell disease and HIV on global scale. *National Institutes of Health (NIH)* <https://www.nih.gov/news-events/news-releases/nih-launches-new-collaboration-develop-gene-based-cures-sickle-cell-disease-hiv-global-scale> (2019).

70. Wonkam, A. *et al.* Genetic modifiers of long-term survival in sickle cell anemia. *Clin. Transl. Med.* **10**, e152 (2020).
71. Shriner, D. & Rotimi, C. N. Whole-Genome-Sequence-Based Haplotypes Reveal Single Origin of the Sickle Allele during the Holocene Wet Phase. *Am. J. Hum. Genet.* **102**, 547–556 (2018).
72. Piel, F. B., Hay, S. I., Gupta, S., Weatherall, D. J. & Williams, T. N. Global Burden of Sickle Cell Anaemia in Children under Five, 2010–2050: Modelling Based on Demographics, Excess Mortality, and Interventions. *PLoS Med.* **10**, e1001484 (2013).
73. Piel, F. B. *et al.* Global migration and the changing distribution of sickle haemoglobin: A quantitative study of temporal trends between 1960 and 2000. *Lancet Glob. Health* **2**, e80–9 (2014).
74. Piel, F. B. *et al.* Global distribution of the sickle cell gene and geographical confirmation of the malaria hypothesis. *Nat. Commun.* **1**, 104 (2010).
75. Bitoungui, V. J. N. *et al.* Beta-globin gene haplotypes among cameroonians and review of the global distribution: is there a case for a single sickle mutation origin in Africa? *Omic J. Integr. Biol.* **19**, 171–179 (2015).
76. Lindenau, J. D. *et al.* The effects of old and recent migration waves in the distribution of HBB**S* globin gene haplotypes. *Genet. Mol. Biol.* **39**, 515–523 (2016).
77. Ha, J., Martinson, R., Iwamoto, S. K. & Nishi, A. Hemoglobin E, malaria and natural selection. *Evol. Med. Public Health* **2019**, 232–241 (2019).
78. Williams, T. N. & Weatherall, D. J. World distribution, population genetics, and health burden of the hemoglobinopathies. *Cold Spring Harb. Perspect. Med.* **2**, a011692–a011692 (2012).
79. Muncie, H. L. & Campbell, J. S. Alpha and beta thalassemia. *Am. Fam. Physician* **80**, (2009).
80. Taher, A. T., Weatherall, D. J. & Cappellini, M. D. Thalassaemia. *Lancet Lond. Engl.* **391**, 155–167 (2018).
81. Nienhuis, A. W. & Nathan, D. G. Pathophysiology and Clinical Manifestations of the β -Thalassemias. *Cold Spring Harb. Perspect. Med.* **2**, a011726 (2012).
82. HALDANE, J. B. S. THE RATE OF MUTATION OF HUMAN GENES. *Hereditas* **35**, 267–273 (1949).
83. Allison, A. C. Protection afforded by sickle-cell trait against subtertian malarial infection. *Br. Med. J.* **4857**, 290–294 (1954).
84. Mackinnon, M. J., Mwangi, T. W., Snow, R. W., Marsh, K. & Williams, T. N. Heritability of malaria in Africa. *PLoS Med.* **2**, 1253–1259 (2005).
85. Malaria Genomic Epidemiology Network. Insights into malaria susceptibility using genome-wide data on 17,000 individuals from Africa, Asia and Oceania. *Nat. Commun.* **10**, 5732 (2019).

86. Chaichoompu, K. *et al.* A different view on fine-scale population structure in Western African populations. *Hum. Genet.* **139**, 45–59 (2020).
87. Piel, F. B. *et al.* Global epidemiology of Sick cell haemoglobin in neonates: A contemporary geostatistical model-based map and population estimates. *The Lancet* **381**, 142–151 (2013).
88. Laval, G. *et al.* Recent Adaptive Acquisition by African Rainforest Hunter-Gatherers of the Late Pleistocene Sick cell-Mutation Suggests Past Differences in Malaria Exposure. *Am. J. Hum. Genet.* **104**, 553–561 (2019).
89. Hunwick, J. O. Black slaves in the mediterranean world: Introduction to a neglected aspect of the african diaspora. *Slavery Abolit.* **13**, 5–38 (1992).
90. Pule, G. D. *et al.* Beta-globin gene haplotypes and selected Malaria-associated variants among black Southern African populations. *Glob. Health Epidemiol. Genomics* **2**, e17 (2017).
91. Colombatti, R. *et al.* Results of a multicenter universal newborn screening program for sick cell disease in Italy: A call to action. *Pediatr. Blood Cancer* **66**, e27657 (2019).
92. Livingstone, F. B. Malaria and Human Polymorphisms. *Annu. Rev. Genet.* **5**, 33–64 (1971).
93. Piel, F. B. *et al.* The distribution of haemoglobin C and its prevalence in newborns in Africa. *Sci. Rep.* **3**, 1671 (2013).
94. Modiano, D. *et al.* Haemoglobin C protects against clinical Plasmodium falciparum malaria. *Nature* **414**, 305–308 (2001).
95. Stephens, C. NATURAL SELECTION. in *Philosophy of Biology* (eds. Matthen, M. & Stephens, C.) 111–127 (North-Holland, 2007). doi:10.1016/B978-044451543-8/50008-3.
96. Serjeant, G. R. The natural history of sick cell disease. *Cold Spring Harb. Perspect. Med.* **3**, a011783 (2013).
97. Hanchard, N. *et al.* Classical sick cell beta-globin haplotypes exhibit a high degree of long-range haplotype similarity in African and Afro-Caribbean populations. *BMC Genet.* **8**, 52 (2007).
98. Ghansah, A. *et al.* Haplotype analyses of haemoglobin C and haemoglobin S and the dynamics of the evolutionary response to malaria in Kassena-Nankana district of Ghana. *PLoS ONE* **7**, e34565 (2012).
99. Borges, E. *et al.* Unusual β -Globin Haplotype Distribution in Newborns from Bengo, Angola. *Hemoglobin* **43**, 149–154 (2019).
100. Trabuchet, G. *et al.* Nucleotide sequence evidence of the unicentric origin of the β C mutation in Africa. *Hum. Genet.* **87**, 597–601 (1991).
101. Mears, J. G. *et al.* Sick cell gene. Its origin and diffusion from West Africa. *J. Clin. Invest.* **68**, 606–610 (1981).

102. Labie, D. *et al.* Haplotypes in tribal Indians bearing the sickle gene: evidence for the unicentric origin of the beta S mutation and the unicentric origin of the tribal populations of India. *Hum. Biol.* **61**, 479–491 (1989).
103. Pagnier, J. *et al.* Evidence for the multicentric origin of the sickle cell hemoglobin gene in Africa. *Proc. Natl. Acad. Sci. U. S. A.* **81**, 1771–1773 (1984).
104. Chebloune, Y. *et al.* Structural analysis of the 5' flanking region of the beta-globin gene in African sickle cell anemia patients: further evidence for three origins of the sickle cell mutation in Africa. *Proc. Natl. Acad. Sci. U. S. A.* **85**, 4431–4435 (1988).
105. Livingstone, F. B. Simulation of the diffusion of the beta-globin variants in the Old World. *Hum. Biol.* **61**, 297–309 (1989).
106. Veytsman, B. A. Environment change, geographic migration and sickle cell anaemia. *Evol. Ecol.* **11**, 519–529 (1997).
107. Hockham, C., Piel, F. B., Gupta, S. & Penman, B. S. Understanding the contrasting spatial haplotype patterns of malaria-protective β -globin polymorphisms. *Infect. Genet. Evol. J. Mol. Epidemiol. Evol. Genet. Infect. Dis.* **36**, 174–183 (2015).
108. Liu, L. *et al.* High-density SNP genotyping to define beta-globin locus haplotypes. *Blood Cells. Mol. Dis.* **42**, 16–24 (2009).
109. Shaikho, E. M. *et al.* A phased SNP-based classification of sickle cell anemia HBB haplotypes. *BMC Genomics* **18**, 608 (2017).
110. Lopez, M. *et al.* The demographic history and mutational load of African hunter-gatherers and farmers. *Nat. Ecol. Evol.* **2**, 721–730 (2018).
111. Otto, T. D. *et al.* Genomes of all known members of a Plasmodium subgenus reveal paths to virulent human malaria. *Nat. Microbiol.* **3**, 687–697 (2018).
112. Rayner, J. C., Liu, W., Peeters, M., Sharp, P. M. & Hahn, B. H. A plethora of Plasmodium species in wild apes: A source of human infection? *Trends in Parasitology* vol. 27 222–229 Preprint at <https://doi.org/10.1016/j.pt.2011.01.006> (2011).
113. Bostoen, K. *et al.* Middle to late holocene paleoclimatic change and the early bantu expansion in the rain forests of Western Central Africa. *Curr. Anthropol.* **56**, 354–384 (2015).
114. R. Templeton, A. Natural selection. in *Philosophy of Biology* 111–127 (Elsevier, 2007). doi:10.1016/B978-044451543-8/50008-3.
115. Liu, W. *et al.* Origin of the human malaria parasite Plasmodium falciparum in gorillas. *Nature* **467**, 420–425 (2010).
116. Eleonore, N. L. E. *et al.* Malaria in patients with sickle cell anaemia: burden, risk factors and outcome at the Laquintinie hospital, Cameroon. *BMC Infect. Dis.* **20**, 40 (2020).
117. Komba, A. N. *et al.* Malaria as a cause of morbidity and mortality in children with homozygous sickle cell disease on the coast of Kenya. *Clin. Infect. Dis.* **49**, 216–222 (2009).

118. Menzel, S. & Thein, S. L. Genetic Modifiers of Fetal Haemoglobin in Sickle Cell Disease. *Mol. Diagn. Ther.* **23**, 235–244 (2019).
119. Galarneau, G. *et al.* Fine-mapping at three loci known to affect fetal hemoglobin levels explains additional genetic variation. *Nat. Genet.* **42**, 1049–1051 (2010).
120. Shah, F. & Dwivedi, M. Pathophysiology and recent therapeutic insights of sickle cell disease. *Ann. Hematol.* **99**, 925–935 (2020).
121. Nandakumar, S. K., Liao, X. & Sankaran, V. G. In The Blood: Connecting Variant to Function In Human Hematopoiesis. *Trends Genet.* 1–14 (2020) doi:10.1016/j.tig.2020.05.006.
122. Wonkam, A. *et al.* Association of variants at BCL11A and HBS1L-MYB with hemoglobin F and hospitalization rates among sickle cell patients in Cameroon. *PloS One* **9**, e92506 (2014).
123. Mmbando, B. P. *et al.* Negative Epistasis between Sickle and Foetal Haemoglobin Suggests a Reduction in Protection against Malaria. *PloS One* **10**, e0125929 (2015).
124. Lettre, G. *et al.* DNA polymorphisms at the BCL11A, HBS1L-MYB, and β -globin loci associate with fetal hemoglobin levels and pain crises in sickle cell disease. *Proc. Natl. Acad. Sci. U. S. A.* **105**, 11869–11874 (2008).
125. Uda, M. *et al.* Genome-wide association study shows BCL11A associated with persistent fetal hemoglobin and amelioration of the phenotype of β -thalassemia. *Proc. Natl. Acad. Sci. U. S. A.* **105**, 1620–1625 (2008).
126. Menzel, S. *et al.* A QTL influencing F cell production maps to a gene encoding a zinc-finger protein on chromosome 2p15. *Nat. Genet.* **39**, 1197–1199 (2007).
127. Sales, R. R. *et al.* Functional polymorphisms of BCL11A and HBS1L-MYB genes affect both fetal hemoglobin level and clinical outcomes in a cohort of children with sickle cell anemia. *Ann. Hematol.* **99**, 1453–1463 (2020).
128. Bernaudin, F. *et al.* Biological impact of α genes, β haplotypes, and G6PD activity in sickle cell anemia at baseline and with hydroxyurea. *Blood Adv.* **2**, 626–637 (2018).
129. Dai, Y. *et al.* BCL2L1 is associated with γ -globin gene expression. *Blood Adv.* **3**, 2995–3001 (2019).
130. Urio, F. *et al.* F cell numbers are associated with an X-linked genetic polymorphism and correlate with haematological parameters in patients with sickle cell disease. *Br. J. Haematol.* bjh.17102 (2020) doi:10.1111/bjh.17102.
131. Garner, C. *et al.* Genetic influences on F cells and other hematologic variables: A twin heritability study. *Blood* **95**, 342–346 (2000).
132. McGuire, A. L. *et al.* The road ahead in genetics and genomics. *Nat. Rev. Genet.* **21**, 581–596 (2020).
133. Wonkam, A. *et al.* Genetic modifiers of long-term survival in sickle cell anemia. *Clin. Transl. Med.* **10**, 1–15 (2020).

134. Rumaney, M. B. *et al.* The co-inheritance of alpha-thalassemia and sickle cell anemia is associated with better hematological indices and lower consultations rate in Cameroonian patients and could improve their survival. *PLoS ONE* **9**, e100516 (2014).
135. Uyoga, S. *et al.* The epidemiology of sickle cell disease in children recruited in infancy in Kilifi, Kenya: a prospective cohort study. *Lancet Glob. Health* **7**, e1458–e1466 (2019).
136. Fabry, M. E. *et al.* *Dense Cells in Sickle Cell Anemia: The Effects of Gene Interaction*. *Blood* vol. 64 (1984).
137. Martinez, G., Muniz, A., Svarch, E., Espinosa, E. & Nagel, R. L. *Age dependence of the gene frequency of alpha-thalassemia in sickle cell anemia in Cuba*. *Blood* vol. 88 1898–1899 (1996).
138. Williams, T. N. *et al.* Negative epistasis between the malaria-protective effects of α^+ -thalassemia and the sickle cell trait. *Nat. Genet.* **37**, 1253–1257 (2005).
139. Mpimbaza, A. *et al.* Associations between erythrocyte polymorphisms and risks of uncomplicated and severe malaria in Ugandan children: A case control study. *PloS One* **13**, e0203229 (2018).
140. Penman, B. S., Pybus, O. G., Weatherall, D. J. & Gupta, S. Epistatic interactions between genetic disorders of hemoglobin can explain why the sickle-cell gene is uncommon in the Mediterranean. *Proc. Natl. Acad. Sci.* **106**, 21242–21246 (2009).
141. Cooper, A. *et al.* APOL1 renal risk variants have contrasting resistance and susceptibility associations with African trypanosomiasis. *eLife* **6**, (2017).
142. Gautam, N. *et al.* Glucose 6 phosphate dehydrogenase deficiency and hemoglobinopathy in South Western Region Nepal: A boon or burden. *BMC Res. Notes* **12**, 1–6 (2019).
143. Fasola, F. A., Fowodu, F. O., Shokunbi, W. A. & Kotila, T. R. The effect of the coinheritance of Glucose-6-phosphate dehydrogenase deficiency on the severity of sickle cell disease. *Niger. Postgrad. Med. J.* **26**, 118–122 (2019).
144. Okafor, I. M., Okoroiwu, H. U. & Ekechi, C. A. Hemoglobin S and Glucose-6-Phosphate Dehydrogenase Deficiency Coinheritance in AS and SS Individuals in Malaria-Endemic Region: A Study in Calabar, Nigeria. *J. Glob. Infect. Dis.* **11**, 118–122 (2019).
145. Gordeuk, V. R. *et al.* CYB5R3 c . 350C >G and G6PD A Alleles Modify Severity of Anemia in Malaria and Sickle Cell Disease. *Am. J. Hematol.* ajh.25941 (2020) doi:10.1002/ajh.25941.
146. Corbo, R. M. & Scacchp, R. Apolipoprotein E (APOE) allele distribution in the world. Is APOE*4 a 'thrifty' allele? *Ann. Hum. Genet.* **63**, 301–310 (1999).
147. Smith, C. J., Ashford, J. W. & Perfetti, T. A. Putative Survival Advantages in Young Apolipoprotein ϵ 4 Carriers are Associated with Increased Neural Stress. *J. Alzheimers Dis.* **68**, 885–923 (2019).
148. Rougeron, V. *et al.* Epistatic Interactions between Apolipoprotein E and Hemoglobin S Genes in Regulation of Malaria Parasitemia. *PLoS ONE* **8**, e76924 (2013).

149. Ekulu, P. M. *et al.* A focus on the association of Apol1 with kidney disease in children. *Pediatr. Nephrol.* **1**, (2020).
150. Anderson, B. R. *et al.* In vivo Modeling Implicates APOL1 in Nephropathy: Evidence for Dominant Negative Effects and Epistasis under Anemic Stress. *PLoS Genet.* **11**, e1005349 (2015).
151. Geard, A. *et al.* Clinical and genetic predictors of renal dysfunctions in sickle cell anaemia in Cameroon. *Br. J. Haematol.* **178**, 629–639 (2017).
152. Wonkam, A. *et al.* Clinical and genetic factors are associated with pain and hospitalisation rates in sickle cell anaemia in Cameroon. *Br. J. Haematol.* **180**, 134–146 (2018).
153. Alladagbin, D. J. *et al.* The sickle cell trait and end stage renal disease in Salvador, Brazil. *PLoS ONE* **13**, e0209036 (2018).
154. Sood, R. *et al.* Sickle cell trait, estimated glomerular filtration rate, and risk of adverse outcomes in chronic kidney disease. *American Journal of Hematology* vol. 94 E275–E278 Preprint at <https://doi.org/10.1002/ajh.25588> (2019).
155. Iversen, P. O. *et al.* Sickle cell disease, malaria and dengue fever: A case of triple jeopardy. *Journal of Travel Medicine* vol. 26 1–2 Preprint at <https://doi.org/10.1093/jtm/taz070> (2019).
156. Elenga, N., Nacher, M. & Trehan, I. The Imperative of Dengue Vaccination for Children with Sickle Cell Disease. *J. Trop. Pediatr.* **66**, 245–247 (2020).
157. Sierra, B. *et al.* OSBPL10, RXRA and lipid metabolism confer African-ancestry protection against dengue haemorrhagic fever in admixed CUBANS. *PLoS Pathog.* **13**, e1006220 (2017).
158. Billig, E. M. W., McQueen, P. G. & McKenzie, F. E. Foetal haemoglobin and the dynamics of paediatric malaria. *Malar. J.* **11**, 396 (2012).
159. Kangoye, D. T. *et al.* Plasmodium falciparum malaria in children aged 0-2 years: the role of foetal haemoglobin and maternal antibodies to two asexual malaria vaccine candidates (MSP3 and GLURP). *PloS One* **9**, e107965 (2014).
160. Archer, N. M., Petersen, N. & Duraisingh, M. T. Fetal hemoglobin does not inhibit Plasmodium falciparum growth. *Blood Adv.* **3**, 2149–2152 (2019).
161. John, C. C. *et al.* Hydroxyurea Dose Escalation for Sickle Cell Anemia in Sub-Saharan Africa. *N. Engl. J. Med.* **382**, 2524–2533 (2020).
162. Esoh, K. & Wonkam, A. Evolutionary history of sickle cell mutation: implications for global genetic medicine. *Hum. Mol. Genet.* **00**, (2021).
163. Grosse, S. D. *et al.* Sickle Cell Disease in Africa. *Am. J. Prev. Med.* **41**, S398–S405 (2011).
164. UN. United Nations General Assembly. Recognition of sickle cell anaemia as a public health problem. **63**, 22–27 (2008).
165. Herrick, J. B. Peculiar elongated and sickle-shaped red blood corpuscles in a case of severe anemia. *Arch. Intern. Med.* **VI**, 517–521 (1910).

166. Williams, T. N. *et al.* Sick cell trait and the risk of Plasmodium falciparum malaria and other childhood diseases. *J. Infect. Dis.* **192**, 178–186 (2005).
167. Malaria Genomic Epidemiology Network. Reappraisal of known malaria resistance loci in a large multicenter study. *Nat. Genet.* **46**, 1197–1204 (2014).
168. Piel, F. B. *et al.* Global distribution of the sickle cell gene and geographical confirmation of the malaria hypothesis. *Nat. Commun.* **1**, 104 (2010).
169. Ghani, R., Manji, M. A. & Ahmed, N. Hemoglobinopathies among five major ethnic groups in Karachi, Pakistan. *Southeast Asian J. Trop. Med. Public Health* **33**, 855–861 (2002).
170. Salinas Cisneros, G. & Thein, S. L. Recent Advances in the Treatment of Sickle Cell Disease. *Frontiers in Physiology* vol. 11 Preprint at <https://doi.org/10.3389/fphys.2020.00435> (2020).
171. Kazadi, A. L., Ngiyulu, R. M., Gini-Ehungu, J. L., Mbuyi-Muamba, J. M. & Aloni, M. N. The clinical characteristics of Congolese children and adolescents suffering from sickle-cell anemia are marked by the high frequencies of epistaxis compared to Western series. *Pediatr. Hematol. Oncol.* **36**, 267–276 (2019).
172. McGann, P. T., Ware, R. E. & Children, C. Hydroxyurea therapy for sickle cell anemia. *Expert Opin Drug Saf* **14**, 1749–1758 (2015).
173. Eastburg, L. *et al.* Extremely high birth prevalence of sickle cell disease in rural Tanzania. *Pediatr. Blood Cancer* **67**, (2020).
174. Ojewunmi, O. O., Adeyemo, T. A., Ayinde, O. C., Iwalokun, B. & Adekile, A. Current perspectives of sickle cell disease in Nigeria: changing the narratives. *Expert Rev. Hematol.* **12**, 609–620 (2019).
175. National Population Commission - NPC/Nigeria and ICF. *Nigeria Demographic and Health Survey 2018*. (2019).
176. Al-Arrayed, S., Hafadh, N., Amin, S., Al-Mukhareq, H. & Sanad, H. Student screening for inherited blood disorders in Bahrain. *East. Mediterr. Health J.* **9**, 344–352 (2003).
177. Adekile, A. *et al.* The Sickle Cell Disease Ontology: Enabling universal sickle cell-based knowledge representation. *Database* **2019**, 1–12 (2019).
178. Kanter, J. *et al.* Validation of a novel point of care testing device for sickle cell disease. *BMC Med.* **13**, 225 (2015).
179. Steele, C. *et al.* Point-of-care screening for sickle cell disease in low-resource settings: A multi-center evaluation of HemoTypeSC, a novel rapid test. *Am. J. Hematol.* **94**, 39–45 (2019).
180. Nnodu, O. E. *et al.* Implementing newborn screening for sickle cell disease as part of immunisation programmes in Nigeria: a feasibility study. *The Lancet Haematology* vol. 7 e534–e540 Preprint at [https://doi.org/10.1016/S2352-3026\(20\)30143-5](https://doi.org/10.1016/S2352-3026(20)30143-5) (2020).
181. Jaja, C., Gibson, R. W. & Edem-Hotah, J. Communicating sickle cell disease point-of-care testing results. (2020) doi:10.1016/j.jpeds.2020.03.027.

182. Makani, J. *et al.* Mortality in sickle cell anemia in Africa: a prospective cohort study in Tanzania. *PloS One* **6**, e14699 (2011).
183. Rahimy, M. C., Gangbo, A., Ahouignan, G. & Alihonou, E. Newborn screening for sickle cell disease in the Republic of Benin. *J. Clin. Pathol.* **62**, 46–48 (2009).
184. McGann, P. T. *et al.* A Prospective Pilot Newborn Screening and Treatment Program for Sickle Cell Anemia in the Republic of Angola. *Blood* **120**, 480–480 (2012).
185. Sundd, P., Gladwin, M. T. & Novelli, E. M. Pathophysiology of Sickle Cell Disease. *Annu. Rev. Pathol. Mech. Dis.* **14**, 263–292 (2019).
186. Steinberg, M. H. Overview of sickle cell anemia pathophysiology. *Sickle Cell Anemia: From Basic Science to Clinical Practice* 49–74 Preprint at https://doi.org/10.1007/978-3-319-06713-1_3 (2016).
187. Ataga, K. I. *et al.* Crizanlizumab for the Prevention of Pain Crises in Sickle Cell Disease. *N. Engl. J. Med.* **376**, 429–439 (2017).
188. Adepoju, P. Ghana takes on sickle-cell disease. *The Lancet* vol. 395 402 Preprint at [https://doi.org/10.1016/S0140-6736\(20\)30293-2](https://doi.org/10.1016/S0140-6736(20)30293-2) (2020).
189. Niihara, Y. *et al.* A Phase 3 Trial of L-Glutamine in Sickle Cell Disease. *N. Engl. J. Med.* **379**, 226–235 (2018).
190. WHO. Haemoglobin concentrations for the diagnosis of anaemia and assessment of severity. Vitamin and Mineral Nutrition Information System. *Geneva Switz. World Health Organ.* 1–6 (2011).
191. Hoppe, C. C. *et al.* Design of the DOVE (Determining Effects of Platelet Inhibition on Vaso-Occlusive Events) trial: A global Phase 3 double-blind, randomized, placebo-controlled, multicenter study of the efficacy and safety of prasugrel in pediatric patients with sickle cell. *Pediatr. Blood Cancer* **63**, 299–305 (2016).
192. Inusa *et al.* Geographic Differences in Phenotype and Treatment of Children with Sickle Cell Anemia from the Multinational DOVE Study. *J. Clin. Med.* **8**, 2009 (2019).
193. Anyanwu, J. N. *et al.* Novel Use of Hydroxyurea in an African Region With Malaria: Protocol for a Randomized Controlled Clinical Trial. *JMIR Res. Protoc.* **5**, e110 (2016).
194. Tshilolo, L. *et al.* Hydroxyurea for Children with Sickle Cell Anemia in Sub-Saharan Africa. *N. Engl. J. Med.* **380**, 121–131 (2019).
195. Ballas, S. K. More definitions in sickle cell disease: Steady state v base line data. *American Journal of Hematology* vol. 87 338–338 Preprint at <https://doi.org/10.1002/ajh.22259> (2012).
196. Chaturvedi, S. & DeBaun, M. R. Evolution of sickle cell disease from a life-threatening disease of children to a chronic disease of adults: The last 40 years. *Am. J. Hematol.* **91**, 5–14 (2016).
197. Williams, T. N. *et al.* Bacteraemia in Kenyan children with sickle-cell anaemia: a retrospective cohort and case-control study. *The Lancet* **374**, 1364–1370 (2009).

198. Kizito, M. E., Mworozzi, E., Ndugwa, C. & Serjeant, G. R. Bacteraemia in homozygous sickle cell disease in Africa: Is pneumococcal prophylaxis justified? *Arch. Dis. Child.* **92**, 21–23 (2007).
199. Williams, T. N. & Obaro, S. K. Sickle cell disease and malaria morbidity: a tale with two tails. *Trends Parasitol.* **27**, 315–320 (2011).
200. Frimpong, A., Thiam, L. G., Arko-Boham, B., Owusu, E. D. A. & Adjei, G. O. Safety and effectiveness of antimalarial therapy in sickle cell disease: a systematic review and network meta-analysis. *BMC Infect. Dis.* **18**, 650 (2018).
201. Arlet, J. B. *et al.* Prognosis of patients with sickle cell disease and COVID-19: a French experience. *The Lancet Haematology* vol. 7 e632–e634 Preprint at [https://doi.org/10.1016/S2352-3026\(20\)30204-0](https://doi.org/10.1016/S2352-3026(20)30204-0) (2020).
202. Menapace, L. A. & Thein, S. L. COVID-19 and sickle cell disease. *Haematologica* **105**, 2501–2504 (2020).
203. Global Sickle Cell Disease Network- About. <https://www.globalsicklecelldisease.org/treat-centre-maps>.
204. Yawn, B. P. *et al.* Management of sickle cell disease: summary of the 2014 evidence-based report by expert panel members. *JAMA* **312**, 1033–1048 (2014).
205. McGann, P. T. *et al.* Realizing effectiveness across continents with hydroxyurea: Enrollment and baseline characteristics of the multicenter REACH study in Sub-Saharan Africa. *Am. J. Hematol.* **93**, 537–545 (2018).
206. McGann, P. T. & Ware, R. E. Hydroxyurea therapy for sickle cell anemia. *Expert Opin. Drug Saf.* **14**, 1749–1758 (2015).
207. Dong, M., McGann, P. T., Mizuno, T., Ware, R. E. & Vinks, A. A. Development of a pharmacokinetic-guided dose individualization strategy for hydroxyurea treatment in children with sickle cell anaemia. *Br. J. Clin. Pharmacol.* **81**, 742–752 (2016).
208. Haywood, C. *et al.* Examining the characteristics and beliefs of hydroxyurea users and nonusers among adults with sickle cell disease. *Am. J. Hematol.* **86**, 85–87 (2011).
209. Nevitt, S. J., Jones, A. P. & Howard, J. Hydroxyurea (hydroxycarbamide) for sickle cell disease. *Cochrane Database of Systematic Reviews* vol. 2017 Preprint at <https://doi.org/10.1002/14651858.CD002202.pub2> (2017).
210. Diop, S. & Pirenne, F. Transfusion and Sickle cell anemia in Africa. *Transfus. Clin. Biol.* (2021) doi:10.1016/j.tracli.2021.01.013.
211. Harif, M. *et al.* Special report: Summary of the first meeting of African Blood and Marrow Transplantation (AfBMT) group, Casablanca, Morocco, April 19–21, 2018 held under the auspices of the Worldwide Network for Blood and Marrow Transplantation (WBMT). *Hematol. Oncol. Stem Cell Ther.* (2019) doi:10.1016/j.hemonc.2019.05.003.
212. Mtenga, J. *et al.* Haematopoietic stem cell transplantation in Tanzania. *British Journal of Haematology* bjh.17106 Preprint at <https://doi.org/10.1111/bjh.17106> (2020).

213. Marsh, V. M., Kamuya, D. M. & Molyneux, S. S. 'All her children are born that way': gendered experiences of stigma in families affected by sickle cell disorder in rural Kenya. *Ethn. Health* **16**, 343–359 (2011).
214. Al-Arrayed, S., Hafadh, N., Amin, S., Al-Mukhareq, H. & Sanad, H. Student screening for inherited blood disorders in Bahrain. *East. Mediterr. Health J. Rev. Sante Mediterr. Orient. Al-Majallah Al-Sihhiyah Li-Sharq Al-Mutawassit* **9**, 344–352 (2003).
215. Wonkam, A. & Makani, J. Sickle cell disease in Africa: an urgent need for longitudinal cohort studies. *The Lancet Global Health* vol. 7 e1310–e1311 Preprint at [https://doi.org/10.1016/S2214-109X\(19\)30364-X](https://doi.org/10.1016/S2214-109X(19)30364-X) (2019).
216. Antonarakis, S. E. *et al.* Origin of the beta S-globin gene in blacks: the contribution of recurrent mutation or gene conversion or both. *Proc. Natl. Acad. Sci.* **81**, 853–856 (1984).
217. Allison, A. C. Protection afforded by sickle-cell trait against subtertian malarial infection. *Br. Med. J.* **4857**, 290–294 (1954).
218. Piel, F. B. *et al.* Global epidemiology of Sickle haemoglobin in neonates: A contemporary geostatistical model-based map and population estimates. *The Lancet* **381**, 142–151 (2013).
219. Ranque, B. *et al.* Estimating the risk of child mortality attributable to sickle cell anaemia in sub-Saharan Africa: a retrospective, multicentre, case-control study. *Lancet Haematol.* **9**, e208–e216 (2022).
220. Steinberg, M. H. & Nagel, R. L. Hemoglobins of the Embryo, Fetus, and Adult. in *Disorders of Hemoglobin: Genetics, Pathophysiology, and Clinical Management* (eds. Forget, B. G., Weatherall, D. J., Higgs, D. R. & Steinberg, M. H.) 119–136 (Cambridge University Press, 2009). doi:10.1017/CBO9780511596582.011.
221. Shen, Y. *et al.* A unified model of human hemoglobin switching through single-cell genome editing. *Nat. Commun.* **12**, 4991 (2021).
222. Platt, O. S. *et al.* Mortality In Sickle Cell Disease -- Life Expectancy and Risk Factors for Early Death. *N. Engl. J. Med.* **330**, 1639–1644 (1994).
223. Esrick, E. B. *et al.* Post-Transcriptional Genetic Silencing of BCL11A to Treat Sickle Cell Disease. *N. Engl. J. Med.* **384**, 205–215 (2021).
224. Frangoul, H. *et al.* CRISPR-Cas9 Gene Editing for Sickle Cell Disease and β -Thalassemia. *N. Engl. J. Med.* **384**, 252–260 (2021).
225. Bae, H. T. *et al.* Meta-analysis of 2040 sickle cell anemia patients: BCL11A and HBS1L-MYB are the major modifiers of HbF in African Americans. *Blood* vol. 120 (2012).
226. Wonkam, A. *et al.* Clinical and genetic factors are associated with pain and hospitalisation rates in sickle cell anaemia in Cameroon. *Br. J. Haematol.* **180**, 134–146 (2018).
227. Nguweneza, A. *et al.* Clinical characteristics and risk factors of relative systemic hypertension and hypertension among sickle cell patients in Cameroon. *Front. Med.* **9**, (2022).

228. Bitoungui, V. J. N. *et al.* Beta-globin gene haplotypes among cameroonians and review of the global distribution: is there a case for a single sickle mutation origin in Africa? *Omics J. Integr. Biol.* **19**, 171–179 (2015).
229. Esoh, K. K. *et al.* Fine scale human genetic structure in three regions of Cameroon reveals episodic diversifying selection. *Sci. Rep.* **11**, 1039 (2021).
230. Gibbs, R. A. *et al.* *A global reference for human genetic variation.* *Nature* vol. 526 (Nature Publishing Group, 2015).
231. Gurdasani, D. *et al.* Uganda Genome Resource Enables Insights into Population History and Genomic Discovery in Africa. *Cell* **179**, 984–1002.e36 (2019).
232. Solovieff, N. *et al.* Fetal hemoglobin in sickle cell anemia: Genome-wide association studies suggest a regulatory region in the 5' olfactory receptor gene cluster. *Blood* **115**, 1815–1822 (2010).
233. Bhatnagar, P. *et al.* Genome-wide association study identifies genetic variants influencing F-cell levels in sickle-cell patients. *J. Hum. Genet.* **56**, 316–323 (2011).
234. Ernst, J. & Kellis, M. Chromatin-state discovery and genome annotation with ChromHMM. *Nat. Protoc.* **12**, 2478–2492 (2017).
235. Grebien, F. *et al.* Stat5 activation enables erythropoiesis in the absence of EpoR and Jak2. *Blood* **111**, 4511–4522 (2008).
236. Thambyrajah, R. *et al.* GFI1 proteins orchestrate the emergence of haematopoietic stem cells through recruitment of LSD1. *Nat. Cell Biol.* **18**, 21–32 (2016).
237. Zhang, L., Flygare, J., Wong, P., Lim, B. & Lodish, H. F. miR-191 regulates mouse erythroblast enucleation by down-regulating Riok3 and Mxi1. *Genes Dev.* **25**, 119–124 (2011).
238. Corn, P. G. *et al.* Mxi1 is induced by hypoxia in a HIF-1–dependent manner and protects cells from c-Myc-induced apoptosis. *Cancer Biol. Ther.* **4**, 1285–1294 (2005).
239. Xi, W. & Beer, M. A. Loop competition and extrusion model predicts CTCF interaction specificity. *Nat. Commun.* **12**, 1046 (2021).
240. Blanco, E., González-Ramírez, M., Alcaine-Colet, A., Aranda, S. & Croce, L. D. The Bivalent Genome: Characterization, Structure, and Regulation. *Trends Genet.* **36**, 118–131 (2020).
241. Yu, Y. *et al.* H3K27me3-H3K4me1 transition at bivalent promoters instructs lineage specification in development. *Cell Biosci.* **13**, 66 (2023).
242. Frazer, K. A. *et al.* A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449**, 851–861 (2007).
243. Pazokitoroudi, A. *et al.* Efficient variance components analysis across millions of genomes. *Nat. Commun.* **11**, 4020 (2020).
244. Wickham, H. *ggplot2: Elegant Graphics for Data Analysis.* (Springer, 2009). doi:10.1007/978-0-387-98141-3.

245. Wojcik, G. L. *et al.* Genetic analyses of diverse populations improves discovery for complex traits. *Nature* **570**, 514–518 (2019).
246. Ferrara, N., Gerber, H.-P. & LeCouter, J. The biology of VEGF and its receptors. *Nat. Med.* **9**, 669–676 (2003).
247. Florentin, J. *et al.* VEGF Receptor 1 Promotes Hypoxia-Induced Hematopoietic Progenitor Proliferation and Differentiation. *Front. Immunol.* **13**, (2022).
248. Semenza, G. L. The Genomics and Genetics of Oxygen Homeostasis. *Annu. Rev. Genomics Hum. Genet.* **21**, 183–204 (2020).
249. Lie-A-Ling, M. *et al.* Regulation of RUNX1 dosage is crucial for efficient blood formation from hemogenic endothelium. *Development* **145**, dev149419 (2018).
250. Das, B. *et al.* A Hypoxia-Driven Vascular Endothelial Growth Factor/Flt1 Autocrine Loop Interacts with Hypoxia-Inducible Factor-1 α through Mitogen-Activated Protein Kinase/Extracellular Signal-Regulated Kinase 1/2 Pathway in Neuroblastoma. *Cancer Res.* **65**, 7267–7275 (2005).
251. Swiers, G. *et al.* Early dynamic fate changes in haemogenic endothelium characterized at the single-cell level. *Nat. Commun.* **4**, 2924 (2013).
252. Helness, A. *et al.* GFI1 tethers the NuRD complex to open and transcriptionally active chromatin in myeloid progenitors. *Commun. Biol.* **4**, 1–16 (2021).
253. Lai, A. Y. & Wade, P. A. Cancer biology and NuRD: a multifaceted chromatin remodelling complex. *Nat. Rev. Cancer* **11**, 588–596 (2011).
254. Kim, D., Kim, K. I. & Baek, S. H. Roles of lysine-specific demethylase 1 (LSD1) in homeostasis and diseases. *J. Biomed. Sci.* **28**, 41 (2021).
255. Xu, J. *et al.* Corepressor-dependent silencing of fetal hemoglobin expression by BCL11A. *Proc. Natl. Acad. Sci.* **110**, 6518–6523 (2013).
256. Tuuli, M. G., Longtine, M. S. & Nelson, D. M. Review: Oxygen and trophoblast biology – A source of controversy. *Placenta* **32**, S109–S118 (2011).
257. Feng, R. *et al.* Activation of γ -globin expression by hypoxia-inducible factor 1 α . *Nature* **610**, 783–790 (2022).
258. Lechauve, C. *et al.* The autophagy-activating kinase ULK1 mediates clearance of free α -globin in β -thalassemia. *Sci. Transl. Med.* **11**, eaav4881 (2019).
259. Ziegler, B. L. *et al.* KDR Receptor: A Key Marker Defining Hematopoietic Stem Cells. *Science* **285**, 1553–1558 (1999).
260. Wonkam, A. *et al.* Five Priorities of African Genomics Research: The Next Frontier. *Annu. Rev. Genomics Hum. Genet.* **23**, 499–521 (2022).
261. Ju, D., Hui, D., Hammond, D. A., Wonkam, A. & Tishkoff, S. A. Importance of Including Non-European Populations in Large Human Genetic Studies to Enhance Precision Medicine. *Annu. Rev. Biomed. Data Sci.* **5**, 321–339 (2022).

262. Sirugo, G., Williams, S. M. & Tishkoff, S. A. The Missing Diversity in Human Genetic Studies. *Cell* **177**, 26–31 (2019).
263. Choudhury, A. *et al.* High-depth African genomes inform human migration and health. *Nature* **586**, 741–748 (2020).
264. Fan, S. *et al.* Whole-genome sequencing reveals a complex African population demographic history and signatures of local adaptation. *Cell* **186**, 923–939.e14 (2023).
265. Milner, P. F. *et al.* Increased HbF in Sickle Cell Anemia is Determined by a Factor Linked to the β S Gene From One Parent. *Blood* **63**, 64–72 (1984).
266. Palmer, K. R., Tong, S. & Kaitu'u-Lino, T. J. Placental-specific sFLT-1: role in pre-eclamptic pathophysiology and its translational possibilities for clinical prediction and diagnosis. *Mol. Hum. Reprod.* **23**, 69–78 (2017).
267. Zhai, Y. *et al.* The soluble VEGF receptor sFlt-1 contributes to endothelial dysfunction in IgA nephropathy. *PLOS ONE* **15**, e0234492 (2020).
268. Wu, Y. & Hirschi, K. K. Regulation of Hemogenic Endothelial Cell Development and Function. *Annu. Rev. Physiol.* **83**, 17–37 (2021).
269. Lancrin, C. *et al.* GFI1 and GFI1B control the loss of endothelial identity of hemogenic endothelium during hematopoietic commitment. *Blood* **120**, 314–322 (2012).
270. Yzaguirre, A. D., de Bruijn, M. F. T. R. & Speck, N. A. The Role of Runx1 in Embryonic Blood Cell Formation. in *RUNX Proteins in Development and Cancer* (eds. Groner, Y. *et al.*) 47–64 (Springer, 2017). doi:10.1007/978-981-10-3233-2_4.
271. Ma, Q. *et al.* Fetal hemoglobin in sickle cell anemia: genetic determinants of response to hydroxyurea. *Pharmacogenomics J.* **7**, 386–394 (2007).
272. Kolliopoulou, A. *et al.* Role of Genomic Biomarkers in Increasing Fetal Hemoglobin Levels Upon Hydroxyurea Therapy and in β -Thalassemia Intermedia: A Validation Cohort Study. *Hemoglobin* **43**, 27–33 (2019).
273. Ataga, K. I. *et al.* Association of soluble fms-like tyrosine kinase-1 with pulmonary hypertension and haemolysis in sickle cell disease. *Br. J. Haematol.* **152**, 485–491 (2011).
274. Fong, G.-H., Rossant, J., Gertsenstein, M. & Breitman, M. L. Role of the Flt-1 receptor tyrosine kinase in regulating the assembly of vascular endothelium. *Nature* **376**, 66–70 (1995).
275. Fong, G.-H., Zhang, L., Bryce, D.-M. & Peng, J. Increased hemangioblast commitment, not vascular disorganization, is the primary defect in flt-1 knock-out mice. *Development* **126**, 3015–3025 (1999).
276. Lan, X. *et al.* ZNF410 Uniquely Activates the NuRD Component CHD4 to Silence Fetal Hemoglobin Expression. *Mol. Cell* **81**, 239–254.e8 (2021).
277. Corre, T. *et al.* Heritability and association with distinct genetic loci of erythropoietin levels in the general population. *Haematologica* **106**, 2499–2501 (2021).

278. Saiki, R. K. *et al.* Enzymatic Amplification of β -Globin Genomic Sequences and Restriction Site Analysis for Diagnosis of Sickle Cell Anemia. *Science* **230**, 1350–1354 (1985).
279. Platt, O. S. *et al.* Pain in Sickle Cell Disease. *N. Engl. J. Med.* **325**, 11–16 (1991).
280. Darbari, D. S. *et al.* Severe painful vaso-occlusive crises and mortality in a contemporary adult sickle cell anemia cohort study. *PloS One* **8**, e79923 (2013).
281. Rumaney, M. B. *et al.* The co-inheritance of alpha-thalassemia and sickle cell anemia is associated with better hematological indices and lower consultations rate in Cameroonian patients and could improve their survival. *PLoS ONE* **9**, e100516 (2014).
282. Chang, C. C. *et al.* Second-generation PLINK: Rising to the challenge of larger and richer datasets. *GigaScience* **4**, 7 (2015).
283. Manichaikul, A. *et al.* Robust relationship inference in genome-wide association studies. *Bioinformatics* **26**, 2867–2873 (2010).
284. Patterson, N., Price, A. L. & Reich, D. Population structure and eigenanalysis. *PLoS Genet.* **2**, 2074–2093 (2006).
285. Sudmant, P. H. *et al.* An integrated map of structural variation in 2,504 human genomes. *Nature* **526**, 75–81 (2015).
286. Loh, P. R., Palamara, P. F. & Price, A. L. Fast and accurate long-range phasing in a UK Biobank cohort. *Nat. Genet.* **48**, 811–816 (2016).
287. McCarthy, S. *et al.* A reference panel of 64,976 haplotypes for genotype imputation. *Nat. Genet.* **48**, 1279–1283 (2016).
288. Browning, B. L. & Browning, S. R. Genotype Imputation with Millions of Reference Samples. *Am. J. Hum. Genet.* **98**, 116–126 (2016).
289. Taliun, D. *et al.* Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *bioRxiv* **2**, 563866 (2019).
290. Fuchsberger, C., Abecasis, G. R. & Hinds, D. A. Minimac2: Faster genotype imputation. *Bioinformatics* **31**, 782–784 (2015).
291. Durbin, R. Efficient haplotype matching and storage using the positional Burrows–Wheeler transform (PBWT). *Bioinformatics* **30**, 1266–1272 (2014).
292. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows–Wheeler transform. *Bioinformatics* **26**, 589–595 (2010).
293. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
294. McKenna, A. *et al.* The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* (2010) doi:10.1101/gr.107524.110.
295. Zhou, W. *et al.* Scalable generalized linear mixed model for region-based association tests in large biobanks and cohorts. *Nat. Genet.* **52**, 634–639 (2020).
296. R Core Team. R: A Language and Environment for Statistical Computing. *R Found. Stat. Comput. Vienna Austria* (2023) doi:https://www.r-project.org.

297. Willer, C. J., Li, Y. & Abecasis, G. R. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* **26**, 2190–2191 (2010).
298. Zou, Y., Carbonetto, P., Wang, G. & Stephens, M. Fine-mapping from summary data with the “Sum of Single Effects” model. *PLOS Genet.* **18**, e1010299 (2022).
299. Pruim, R. J. *et al.* LocusZoom: regional visualization of genome-wide association scan results. *Bioinformatics* **26**, 2336–2337 (2010).
300. Watanabe, K., Taskesen, E., van Bochoven, A. & Posthuma, D. Functional mapping and annotation of genetic associations with FUMA. *Nat. Commun.* **8**, 1826 (2017).
301. Leeuw, C. A. de, Mooij, J. M., Heskes, T. & Posthuma, D. MAGMA: Generalized Gene-Set Analysis of GWAS Data. *PLOS Comput. Biol.* **11**, e1004219 (2015).
302. Barrett, J. C., Fry, B., Maller, J. & Daly, M. J. Haploview: Analysis and visualization of LD and haplotype maps. *Bioinformatics* (2005) doi:10.1093/bioinformatics/bth457.
303. Liu, L. *et al.* Original Research: A case-control genome-wide association study identifies genetic modifiers of fetal hemoglobin in sickle cell disease. *Exp. Biol. Med. Maywood NJ* **241**, 706–718 (2016).
304. Gabriel, S. B. *et al.* The structure of haplotype blocks in the human genome. *Science* **296**, 2225–2229 (2002).
305. Richerson, P. J., Boyd, R. & Bettinger, R. L. Was Agriculture Impossible during the Pleistocene but Mandatory during the Holocene ? A Climate Change Hypothesis. *Am. Antiq.* **66**, 387–411 (2001).
306. Lipson, M. *et al.* Ancient West African foragers in the context of African population history. *Nature* **577**, 665–670 (2020).
307. Lipson, M. *et al.* Ancient DNA and deep population structure in sub-Saharan African foragers. *Nature* **603**, 290–296 (2022).
308. Samayoa-Reyes, G. *et al.* IFN- λ 4 genetic variants influence clinical malaria episodes in a cohort of Kenyan children. *Malar. J.* **20**, 196 (2021).
309. Prokunina-Olsson, L. *et al.* IFN- λ 4 is associated with increased risk and earlier occurrence of several common infections in African children. *Genes Immun.* **22**, 44–55 (2021).
310. Waldenström, J. *et al.* Absence of interferon- λ 4 enhances spontaneous clearance of acute hepatitis C virus genotypes 1-3 infection. *Scand. J. Gastroenterol.* **56**, 855–861 (2021).
311. Manry, J. *et al.* Evolutionary genetic dissection of human interferons. *J. Exp. Med.* **208**, 2747–2759 (2011).
312. Key, F. M. *et al.* Selection on a Variant Associated with Improved Viral Clearance Drives Local, Adaptive Pseudogenization of Interferon Lambda 4 (IFNL4). *PLoS Genet.* **10**, e1004681 (2014).
313. Park, S. *et al.* Evolutionary history of human disease genes reveals phenotypic connections and comorbidity among genetic diseases. *Sci. Rep.* **2**, 757 (2012).
314. Nagel, R. L. Epistasis and the genetics of human diseases. *C. R. Biol.* **328**, 606–615 (2005).

315. The Consortium on Newborn Screening in Africa for sickle cell disease: study rationale and methodology | Blood Advances | American Society of Hematology.
<https://ashpublications.org/bloodadvances/article/6/24/6187/486808/The-Consortium-on-Newborn-Screening-in-Africa-for>.
316. Leonard, A., Tisdale, J. & Abraham, A. Curative options for sickle cell disease: haploidentical stem cell transplantation or gene therapy? *Br. J. Haematol.* **189**, 408–423 (2020).
317. Pareek, V., Pedley, A. M. & Benkovic, S. J. Human de novo purine biosynthesis. *Crit. Rev. Biochem. Mol. Biol.* **56**, 1–16 (2021).
318. Pizzorno, G., Diasio, R. B. & Cheng, Y.-C. Purine Analogs. in *Holland-Frei Cancer Medicine. 6th edition* (BC Decker, 2003).
319. Lavinha, J. *et al.* Importation route of the sickle cell trait into Portugal: contribution of molecular epidemiology. *Hum. Biol.* **64**, 891–901 (1992).
320. Mouélé, R., Pambou, O., Feingold, J. & Galactéros, F. alpha-thalassemia in Bantu population from Congo-Brazzaville: its interaction with sickle cell anemia. *Hum. Hered.* **50**, 118–125 (2000).
321. Abou-Elew, H. H. *et al.* β S globin gene haplotype and the stroke risk among Egyptian children with sickle cell disease. *Hematology* **23**, 362–367 (2018).
322. Sow, A. *et al.* Linkage-disequilibrium of the senegal haplotype with the beta S gene in the republic of guinea. *Am. J. Hematol.* **50**, 301–303 (1995).
323. Ojwang, P. J. *et al.* Haplotypes and alpha globin gene analyses in sickle cell anaemia patients from Kenya. *Br. J. Haematol.* **65**, 211–215 (1987).
324. Hewitt, R., Krause, A., Goldman, A., Campbell, G. & Jenkins, T. Beta-globin haplotype analysis suggests that a major source of Malagasy ancestry is derived from Bantu-speaking Negroids. *Am. J. Hum. Genet.* **58**, 1303–1308 (1996).
325. Pule, G. D. *et al.* Burden, genotype and phenotype profiles of adult patients with sickle cell disease in Cape Town, South Africa. *South Afr. Med. J. Suid-Afr. Tydskr. Vir Geneesk.* **107**, 149–155 (2017).
326. Veten, F. M. *et al.* Hb S [β 6(A3)Glu→Val, GAG>GTG] and β -globin gene cluster haplotype distribution in Mauritania. *Hemoglobin* **36**, 311–315 (2012).
327. Adekile, A. D. *et al.* Haplotypes in SS patients from Nigeria; characterization of one atypical beta S haplotype no. 19 (Benin) associated with elevated HB F and high G gamma levels. *Ann. Hematol.* **65**, 41–45 (1992).
328. Currat, M. *et al.* Molecular analysis of the beta-globin gene cluster in the Niokholo Mandenka population reveals a recent origin of the beta(S) Senegal mutation. *Am. J. Hum. Genet.* **70**, 207–223 (2002).

329. Tall, F. G. *et al.* Genetic Background of the Sickle Cell Disease Pediatric Population of Dakar, Senegal, and Characterization of a Novel Frameshift -Thalassemia Mutation {[HBB: c.265_266del; p.Leu89Glufs{*}2]}. *HEMOGLOBIN* **41**, 89–95 (2017).
330. ELDERDERY, A. Y. *et al.* Molecular analysis of the β -globin gene cluster haplotypes in a Sudanese population with sickle cell anaemia. *Int. J. Lab. Hematol.* **34**, 262–266 (2012).
331. Daak, A. A. *et al.* Sickle cell disease in western Sudan: genetic epidemiology and predictors of knowledge attitude and practices. *Trop. Med. Int. Health TM IH* **21**, 642–653 (2016).
332. Oner, C. *et al.* Beta S haplotypes in various world populations. *Hum. Genet.* **89**, 99–104 (1992).
333. Imen, M. *et al.* Restriction mapping of β S locus among Tunisian sickle-cell patients. *Am. J. Hum. Biol. Off. J. Hum. Biol. Counc.* **23**, 815–819 (2011).
334. Ndugwa, C. *et al.* Homozygous sickle cell disease in Uganda and Jamaica a comparison of Bantu and Benin haplotypes. *West Indian Med. J.* **61**, 684–691 (2012).
335. Muszlak, M. *et al.* Genetic Modifiers of Sickle Cell Disease: A Genotype-Phenotype Relationship Study in a Cohort of 82 Children on Mayotte Island. *Hemoglobin* **39**, 156–161 (2015).
336. Mukherjee, M. B. *et al.* Beta-globin gene cluster haplotypes linked to the betaS gene in western India. *Hemoglobin* **28**, 157–161 (2004).
337. Nongbri, S. R. L., Verma, H. K., Lakkakula, B. V. K. S. & Patra, P. K. Presence of atypical beta globin (HBB) gene cluster haplotypes in sickle cell anemia patients of India. *Rev. Bras. Hematol. E Hemoter.* **39**, 180–182 (2017).
338. Al Arrayed, S. & Haite, N. Features of sickle-cell disease in Bahrain. *East. Mediterr. Health J.* **1**, (1995).
339. Rahimi, Z., Karimi, M., Haghshenass, M. & Merat, A. Beta-globin gene cluster haplotypes in sickle cell patients from southwest Iran. *Am. J. Hematol.* **74**, 156–160 (2003).
340. Aghajani, F. *et al.* Identification of β -globin haplotypes linked to sickle hemoglobin (Hb S) alleles in Mazandaran province, Iran. *Genes Genet. Syst.* **91**, 311–313 (2016).
341. Keikhaei, B. *et al.* Beta-Globin Gene Cluster Haplotypes in Iranian Sickle Cell Patients : Relation to Some Hematologic Parameters. *IJBC* **4**, 105–110 (2012).
342. Al-Allawi, N. A. S. *et al.* Sickle cell disease in the Kurdish population of northern Iraq. *Hemoglobin* **36**, 333–342 (2012).
343. Yaseen, N., Al-Mamoori, H. & Hassan, M. Sickle β -globin haplotypes among patients with sickle cell anemia in Basra, Iraq: A cross-sectional study. *Iraqi J. Hematol.* **9**, 23 (2020).
344. el-Hazmi, M. A. *et al.* Haplotypes of the beta-globin gene as prognostic factors in sickle-cell disease. *East. Mediterr. Health J. Rev. Sante Mediterr. Orient. Al-Majallah Al-Sihhiyah Li-Sharq Al-Mutawassit* **5**, 1154–1158 (1999).
345. Adekile, A. D. & Haider, M. Z. Morbidity, beta S haplotype and alpha-globin gene patterns among sickle cell anemia patients in Kuwait. *Acta Haematol.* **96**, 150–154 (1996).

346. Inati, A. *et al.* Beta-globin gene cluster haplotypes and HbF levels are not the only modulators of sickle cell disease in Lebanon. *Eur. J. Haematol.* **70**, 79–83 (2003).
347. Daar, S., Hussain, H. M., Gravel, D., Nagel, R. L. & Krishnamoorthy, R. Genetic epidemiology of HbS in Oman: multicentric origin for the betaS gene. *Am. J. Hematol.* **64**, 39–46 (2000).
348. Samarah, F., Ayesh, S., Athanasiou, M., Christakis, J. & Vavatsi, N. beta(S)-Globin gene cluster haplotypes in the West Bank of Palestine. *Hemoglobin* **33**, 143–149 (2009).
349. Al-Ali, A. K. *et al.* Prevalence and Diversity of Haplotypes of Sickle Cell Disease in the Eastern Province of Saudi Arabia. *Hemoglobin* **44**, 78–81 (2020).
350. el-Kalla, S. & Baysal, E. Genotype-phenotype correlation of sickle cell disease in the United Arab Emirates. *Pediatr. Hematol. Oncol.* **15**, 237–242 (1998).
351. Muniz, A. *et al.* Sickle cell anemia and beta-gene cluster haplotypes in Cuba. *Am. J. Hematol.* **49**, 163–164 (1995).
352. Magaña, M. T. *et al.* Analysis of betaS and betaA genes in a Mexican population with African roots. *Blood Cells. Mol. Dis.* **28**, 121–126 (2002).
353. Jones-Lecointe, A. *et al.* beta-globin gene cluster Haplotypes and alpha-thalassemia in sickle cell disease patients from Trinidad. *Am. J. Hum. Biol.* **20**, 342–344 (2008).
354. Crawford, D. C. *et al.* Characterization of beta-globin haplotypes using blood spots from a population-based cohort of newborns with homozygous HbS. *Genet. Med. Off. J. Am. Coll. Med. Genet.* **4**, 328–335 (2002).
355. da Luz, J. A. *et al.* α -thalassemia, HbS, and β -globin gene cluster haplotypes in two Afro-Uruguayan sub-populations from northern and southern Uruguay. *Genet. Mol. Biol.* **29**, 595–600 (2006).
356. Adorno, E. V. *et al.* Clinical and molecular characteristics of sickle cell anemia in the northeast of Brazil. *Genet. Mol. Biol.* **31**, 621–625 (2008).
357. Nascimento, A. F. F. *et al.* Haplotypes and polymorphism in the CCR5 gene in sickle cell disease. *Genet. Mol. Res.* **16**, (2017).
358. Okumura, J. V *et al.* Atypical β -S haplotypes: classification and genetic modulation in patients with sickle cell anemia. *J. Hum. Genet.* **64**, 239–248 (2019).
359. Alves, A. C. *et al.* Sickle cell anemia in the state of Maranhão: a haplotype study. *Ann. Hematol.* **99**, 1225–1230 (2020).
360. Cardoso, G. L. *et al.* African gene flow to North Brazil as revealed by HBB*S gene haplotype analysis. *Am. J. Hum. Biol.* **18**, 93–98 (2006).
361. Da Silva, L. B., Gonçalves, R. P. & Rabenhorst, S. H. B. Analysis of sickle cell anemia haplotypes in Fortaleza reveals the ethnic origins of Ceará state population. *J. Bras. Patol. E Med. Lab.* **45**, 115–118 (2009).
362. Goncalves, M. S. *et al.* Sickle cell disease in a Brazilian population from Sao Paulo: a study of the beta s haplotypes. *Hum. Hered.* **44**, 322–327 (1994).

363. Leal, A. S., Martins, P. R. J. & Balarin, M. A. S. Haplotype of the β S-globin cluster in patients with sickle cell anemia at a University Hospital in the Triangulo Mineiro, Minas Gerais. *Rev. Bras. Hematol. E Hemoter.* **37**, 140–141 (2015).
364. Gonçalves, M. S. *et al.* β S-Haplotypes in sickle cell anemia patients from Salvador, Bahia, Northeastern Brazil. *Braz. J. Med. Biol. Res.* **36**, 1283–1288 (2003).
365. Bezerra, M. A. C. *et al.* Molecular variations linked to the grouping of beta- and alpha-globin genes in neonatal patients with sickle cell disease in the State of Pernambuco, Brazil. *Hemoglobin* **31**, 83–88 (2007).
366. Cabral, C. H. K. Determinação de haplótipos do gene beta S em pacientes com anemia falciforme. *Rev. Bras. Hematol. E Hemoter.* **32**, 491–492 (2010).
367. LitsukoTomimatsu Shimauti, E. *et al.* Prevalence of beta(S)-globin gene haplotypes, alpha-thalassemia (3.7 kb deletion) and redox status in patients with sickle cell anemia in the state of Parana, Brazil. *Genet. Mol. Biol.* **38**, 316–323 (2015).
368. Fong, C. *et al.* Frequency and origin of haplotypes associated with the beta-globin gene cluster in individuals with trait and sickle cell anemia in the Atlantic and Pacific coastal regions of Colombia. *Genet. Mol. Biol.* **36**, 494–497 (2013).
369. Arends, A. *et al.* Determination of beta-globin gene cluster haplotypes and prevalence of alpha-thalassemia in sickle cell anemia patients in Venezuela. *Am. J. Hematol.* **64**, 87–90 (2000).
370. Schiliro, G. *et al.* Clinical, hematological, and molecular features in Sicilians with Hb S-beta-thalassemia. *Am. J. Hematol.* **41**, 264–269 (1992).
371. Tubman, V. N. *et al.* Newborn Screening for Sickle Cell Disease in Liberia: A Pilot Study. *Pediatr. Blood Cancer* **63**, 671–676 (2016).
372. Kafando, E., Sawadogo, M., Cotton, F., Vertongen, F. & Gulbis, B. Neonatal screening for sickle cell disorders in Ouagadougou, Burkina Faso: a pilot study. *J. Med. Screen.* **12**, 112–114 (2005).
373. Zohoun, A., Baglo Agbodande, T., Zohoun, L. & Anani, L. Prevalence of hemoglobin abnormalities in an apparently healthy population in Benin. *Hematol. Transfus. Cell Ther.* **42**, 145–149 (2020).
374. Couque, N., Trawinski, E. & Elion, J. Génétique des maladies de l'hémoglobine. *Rev. Francoph. Lab.* **2016**, 49–60 (2016).
375. Inusa, B. P., Juliana Olufunke, Y. D. & John Dada, L. Sickle Cell Disease Screening in Northern Nigeria: The Co-Existence of B- Thalassemia Inheritance. *Pediatr. Ther.* **05**, (2015).
376. Françoise Ngo Sack Senior, F., Françoise Ngo Sack, F., Moria Njangtang, D., Chetcha Chemegni, B. & de Paul Djientcheu, V. Prevalence of sickle cell disease in newborns in the Yaounde Central Hospital. *J. Med. Res.* **3**, 277–279 (2017).

377. Agasa, B. *et al.* Prevalence of sickle cell disease in a northeastern region of the Democratic Republic of Congo: What impact on transfusion policy? *Transfus. Med.* **20**, 62–65 (2010).
378. Tshilolo, L. *et al.* Neonatal screening for sickle cell anaemia in the Democratic Republic of the Congo: Experience from a pioneer project on 31 204 newborns. *J. Clin. Pathol.* **62**, 35–38 (2009).
379. McGann, P. T. *et al.* A prospective newborn screening and treatment program for sickle cell anemia in Luanda, Angola. *Am. J. Hematol.* **88**, 984–989 (2013).
380. Kiyaga, C. *et al.* Sickle cell screening in Uganda: High burden, human immunodeficiency virus comorbidity, and genetic modifiers. *Pediatr. Blood Cancer* **66**, e27807 (2019).
381. Ndeezi, G. *et al.* Burden of sickle cell trait and disease in the Uganda Sickle Surveillance Study (US3): a cross-sectional study. *Lancet Glob. Health* **4**, e195–e200 (2016).
382. Nkya, S. *et al.* Newborn screening for sickle cell disease: An innovative pilot program to improve child survival in Dares Salaam, Tanzania. *Int. Health* **11**, 589–595 (2019).
383. Mutesa, L. *et al.* Neonatal screening for sickle cell disease in Central Africa: A study of 1825 newborns with a new enzyme-linked immunosorbent assay test. *J. Med. Screen.* **14**, 113–116 (2007).
384. Adam, M. A., Adam, N. K. & Mohamed, B. A. Prevalence of sickle cell disease and sickle cell trait among children admitted to Al Fashir Teaching Hospital North Darfur State, Sudan. *BMC Res. Notes* **12**, 659 (2019).
385. Steinberg, M. H. Sickle cell anemia, the first molecular disease: Overview of molecular etiology, pathophysiology, and therapeutic approaches. *TheScientificWorldJournal* vol. 8 1295–1324 Preprint at <https://doi.org/10.1100/tsw.2008.157> (2008).

Appendix I: Distribution of β S haplotypes in various World populations

Haplotypes (%)

Continent	Country/Region	Arab-India	CAR/Bantu	BEN	CAM	SEN	Atypical	Reference
<i>Africa</i>	Algeria	0	0	100	0	0	0	103
	Angola	0	95.5	4.5	0	0	0	319
	Angola, Bengo	0	82.93	9.76	0	7.31	0	99
	Cameroon	0.3	0.5	73.8	19.1	0.2	6.1	75
	Republic of Congo	0	91	0	0	0	9.1	320
	Egypt	2	18.27	41.34	1	5.8	5.8	321
	Guinea	0	0	11	0	89	0	322
	Kenya	0	98.2	1.8	0	0	0	323
	Madagascar	0	91.4	0	0	2.9	5.7	324
	Malawi	0	5.7	27.1	7.1	8.7	51.4	325
	Mauritania	5.6	4.4	8.9	0	77.8	3.3	326
	Nigeria	0	1	93.2	3.4	0	2.4	327
	Senegal	0	0	0	0	100	0	328
	Senegal, Dakar	0	0.9	20.64	2.9	75.4	0.18	329
	South Africa	0	5.1	16.6	6.4	3.9	68	325
	Sudan, White Nile	0	2.8	29.4	35	18.2	14.6	330
	Sudan, Kordofan	0	2.2	21.8	30.8	12.8	32.2	331
	Tanzania	0	100	0	0	0	0	332
	Tunisia	0	2.7	60.5	0	0	36.7	333
	Uganda	0	99.5	0	0	0.5	0	334
	Mayotte Island	0	89.8	9.3	0.085	0	0	335
	Zimbabwe	0	3.1	12.5	15.6	3.1	65.7	325
<i>Asia</i>	India	91.4	0	0	0	0	8.6	336
	India, Chhattisgarh	78	1	2	0	4	15	337
	Bahrain	89.2	5.4	2.7	0	0	2.7	338
	Iran, Shiraz	53.7	3.1	11.7	2.5	3.7	25.3	339
	Iran, Mazandaran	23	7.7	53.9	0	7.7	7.7	340
	Iran, Khuzestan	38	16	18	16	12	0	341
	Iraq, Kurdish	12.5	7.8	69.5	0	0	10.2	342
	Iraq, Basra	54.8	0	19.4	0	19.4	6.5	343
	Jordan	20	0	80	0	0	0	344

America	Kuwait	80.8	5.6	11.2	0	0	2.4	345
	Lebanon	10	15	73	0	0	2	346
	Oman	26.5	21.4	52.1	0	0	0	347
	West Bank	0	5.1	88.1	0	0	6.8	348
	Saudi Arabia, south eastern	1.6	0	98.4	0	0	0	344
	Saudi Arabia, eastern province	82.3	0.07	0.07	1.2	4.02	4.76	349
	Syria	33.3	0	66.7	0	0	0	344
	United Arab Emirates	52	26	22	0	0	0	350
	Canada	0	11.5	49.2	13.1	13.1	13.1	332
	Cuba	0	40.9	51	0	8.1	0	351
	Jamaica	0	8.3	76	0	5.2	10.5	334
	Mexico	0	78.8	18.2	0	0	3	352
	Trinidad	3.2	17.3	61.8	3.5	8.5	5.7	353
	USA	0.8	21.6	58.9	6.1	8	4.6	354
	Uruguay	0	60	20	0	0	20	355
	Brazil, Salvador	0.4	41.6	55.2	1.2	0.4	1.2	356
	Brazil, Bahia	0	11.5	28.6	2.8	2.8	54.3	357
	Brazil, Rio De Janerio	0.3	73.2	19.9	0.6	0.3	5.7	358
	Brazil, Maranhão	0	64.28	28.57	0	0	7.15	359
	Brazil, Belém	0	66	21.8	1.3	10.9	0	360
	Brazil, Ceará	0	66.2	22.1	0	0	11.8	361
	Brazil, São Paulo	0	62.2	33.8	0	0	4.1	362
	Brazil, Rio Grande do Sul	0	67.3	25	0.9	0.5	6.4	76
	Brazil, Minas Gerais	0	64.8	22.1	0.8	0	12.3	363
	Brazil, Bahia	0	48.1	45.6	0	0.63	5.63	364
	Brazil, Pernambuco	0	81.1	14.2	0.8	0	3.9	365
	Brazil, Rio Grande do Norte	0	75.5	11.3	6.6	0	6.6	366
	Brazil, Fortaleza	0	66.2	22	0	0	11.8	361
	Brazil, Paraná	0	62	32	0	0	6	367
Europe	Colombia	0.4	29.7	33.2	4.4	4.4	27.9	368
	Suriname	0	29.9	53.2	2.6	2.6	11.7	332
	Venezuela	0	32.2	50.8	2.3	14.1	0.6	369
	Portugal	0	42.4	36.4	0	21.2	0	319
	Italy	0	0	100	0	0	0	370

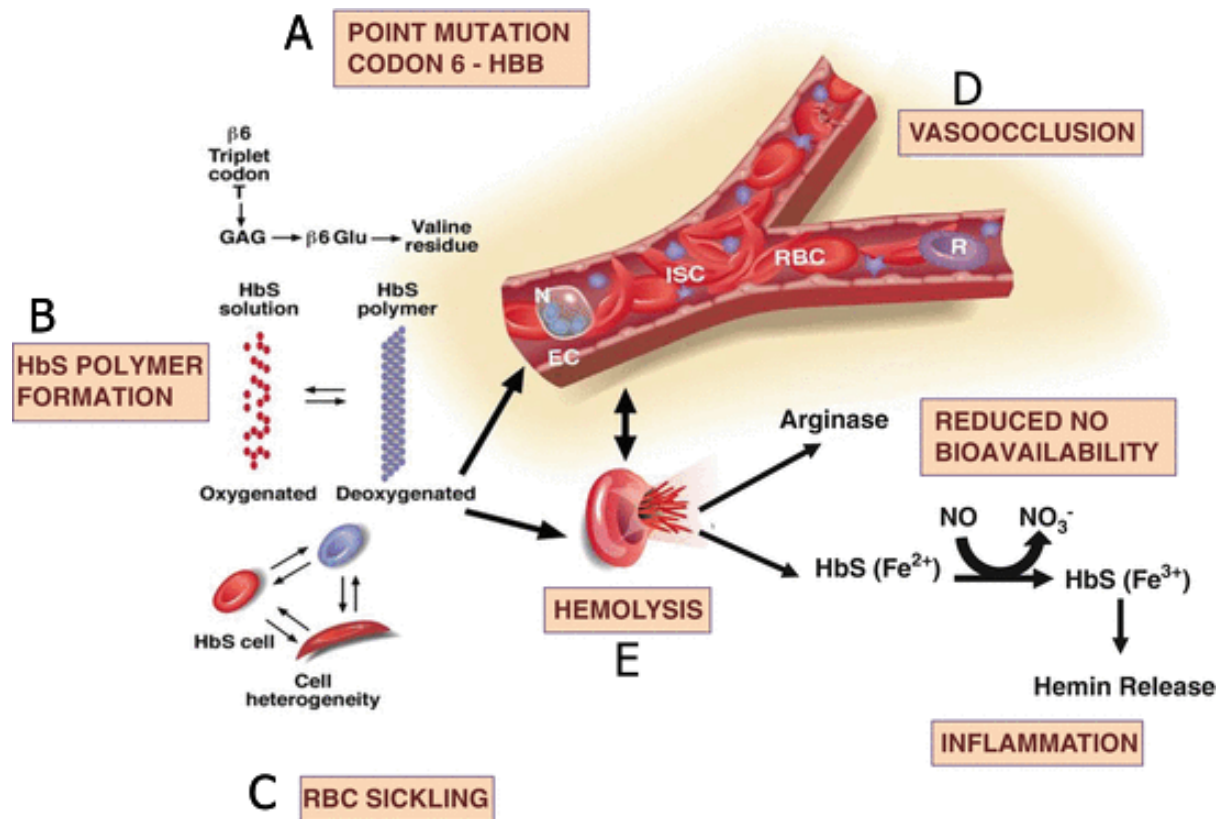
Turkey	0.5	0	96.3	0	0	3.3	332
Greece	0	0	92.9	0	7.1	0	332
France	0	43.2	34.6	2.7	17	2.5	128

Appendix II: Birth and population prevalence of SCA in Africa

Country, Region	Birth prevalence (%)	Population prevalence (%)	SCT prevalence (%)	References
<i>Liberia</i>	1.19	-	10.31	371
<i>Burkina Faso, Ouagadougou</i>	1.75	-	7.3	372
<i>Benin</i>	-	0.2-4	22.3	373,374
<i>Nigeria, Abuja</i>	1-1.4	-	21	180
<i>Nigeria, northern</i>	1.72	-	22.36	375
<i>Nigeria, national</i>	-	2.69-5	-	174
<i>Cameroon, Yaounde</i>		1.79	18.2	376
<i>Democratic Republic of Congo (DRC), Kisangani</i>	-	0.96	23.3	377
<i>DRC, national</i>	1.4	-	16.9	378
<i>Angola, Luanda</i>	1.51	-	21	379
<i>Angola, national</i>	1.55	-	21	379
<i>Uganda, national</i>	0.7	-	>18	380
<i>Uganda, East Central and Mid-Northern districts</i>	-	>1	>18	380
<i>Uganda, Eastern</i>	-	1.5	19.8	381
<i>Tanzania</i>	0.8	-	12.6	382
<i>Tanzania, Shirati</i>	3.9	-	31.6	173
<i>Kenya, Kilifi</i>	-	0.8	7.8	135
<i>Burundi</i>	0.16	-	4.40	383
<i>Rwanda</i>	0.09	-	2.54	383
<i>Sudan, Darfur</i>	3.5	-	11.3	384

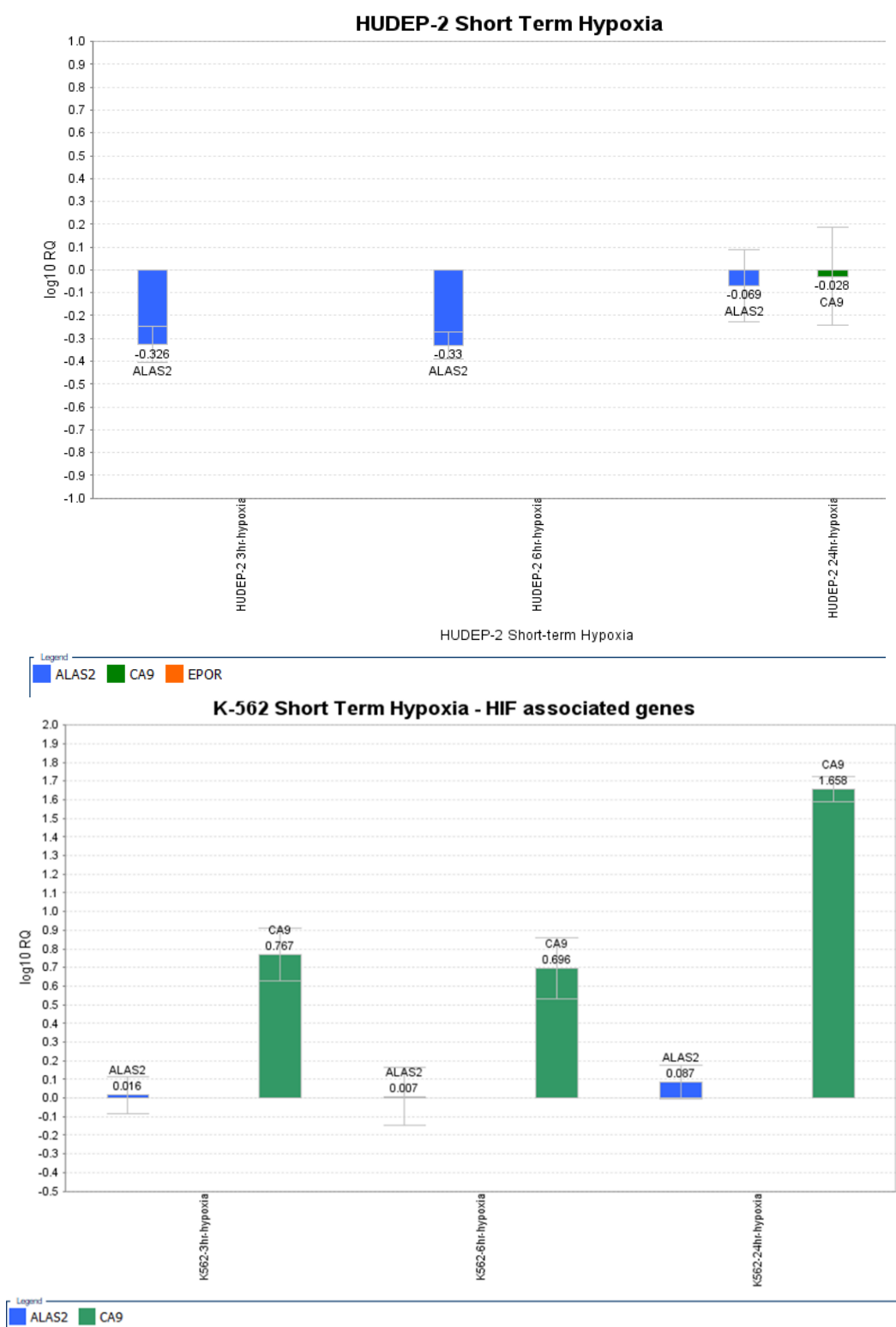
SCT = sickle-cell trait

Appendix III: Sick cell anemia pathophysiology

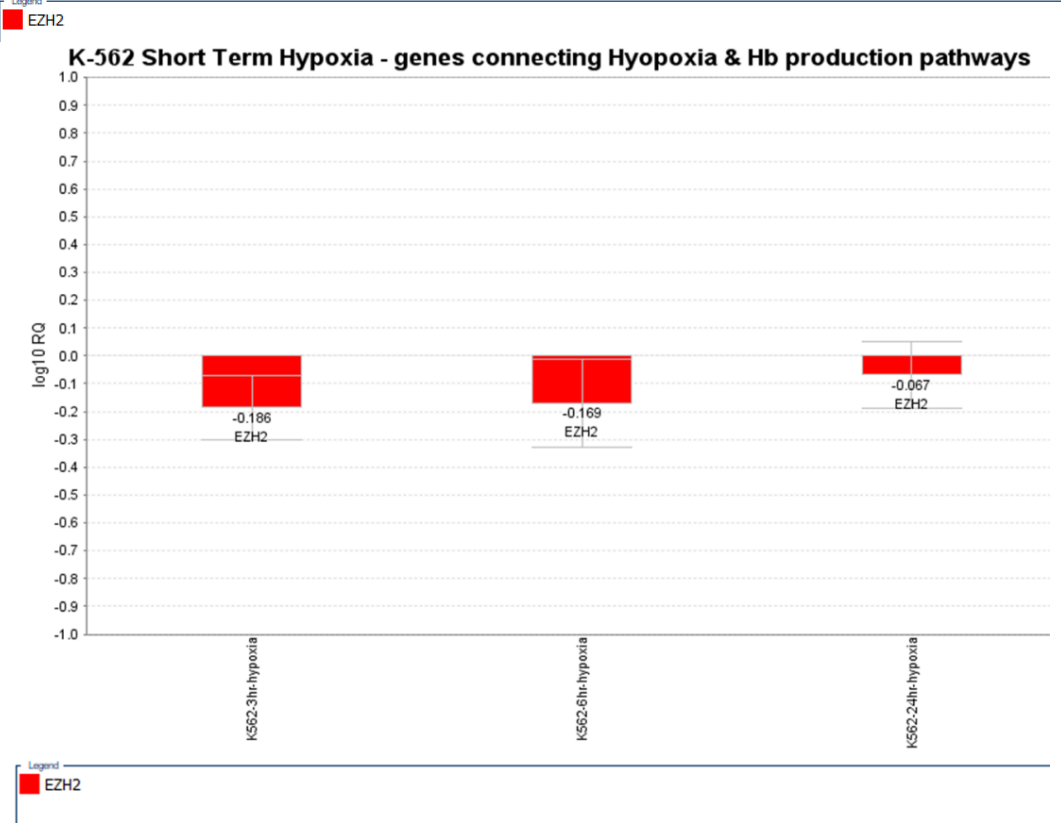
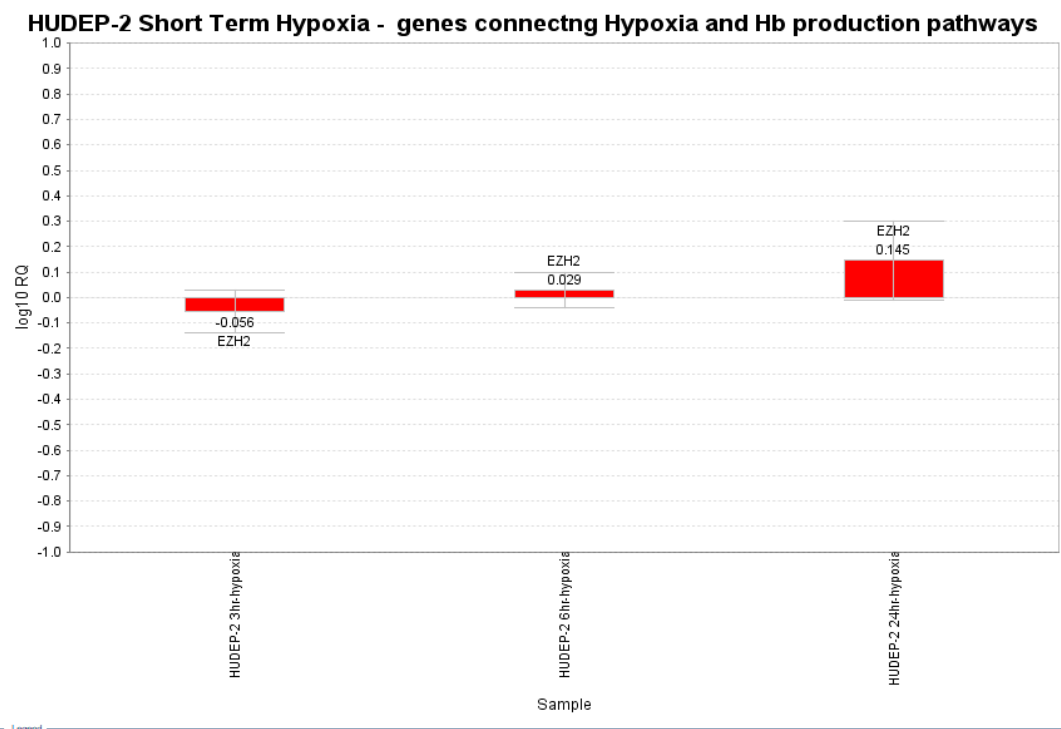


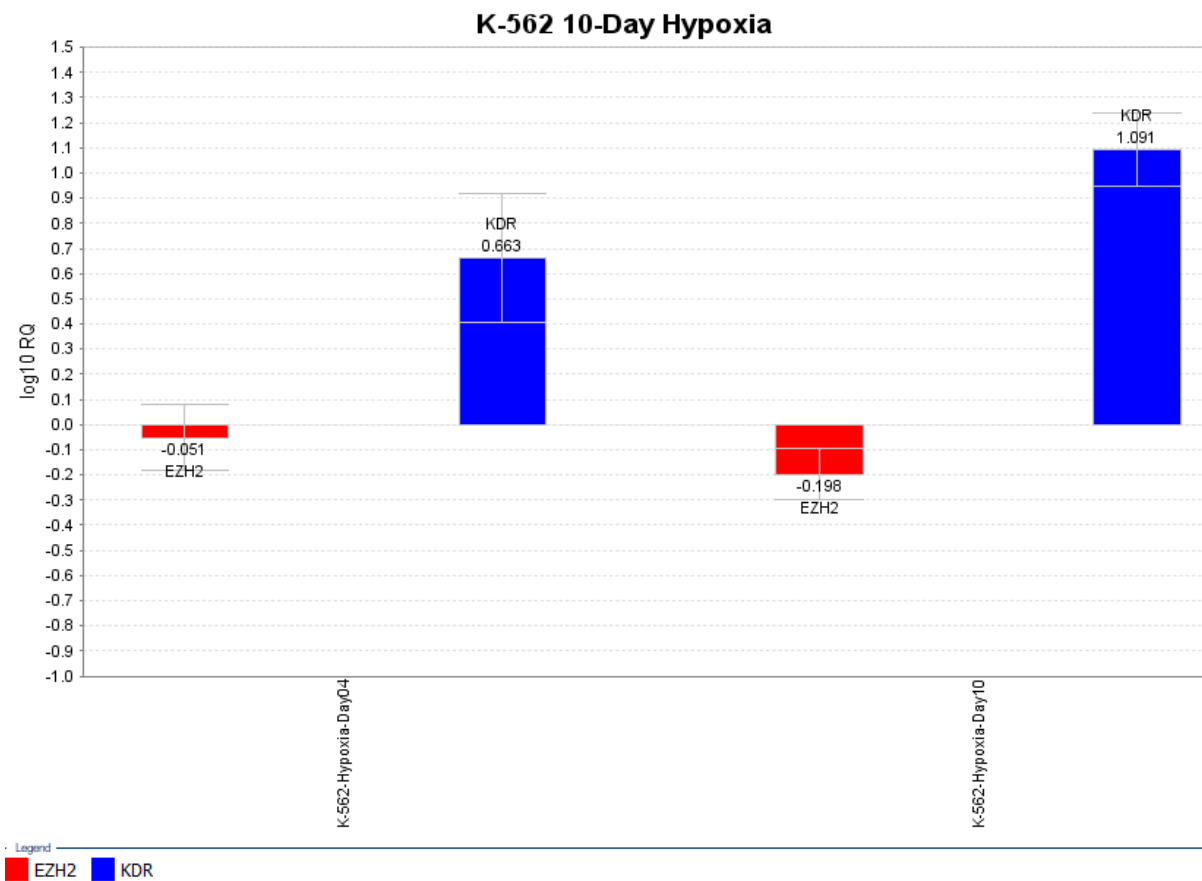
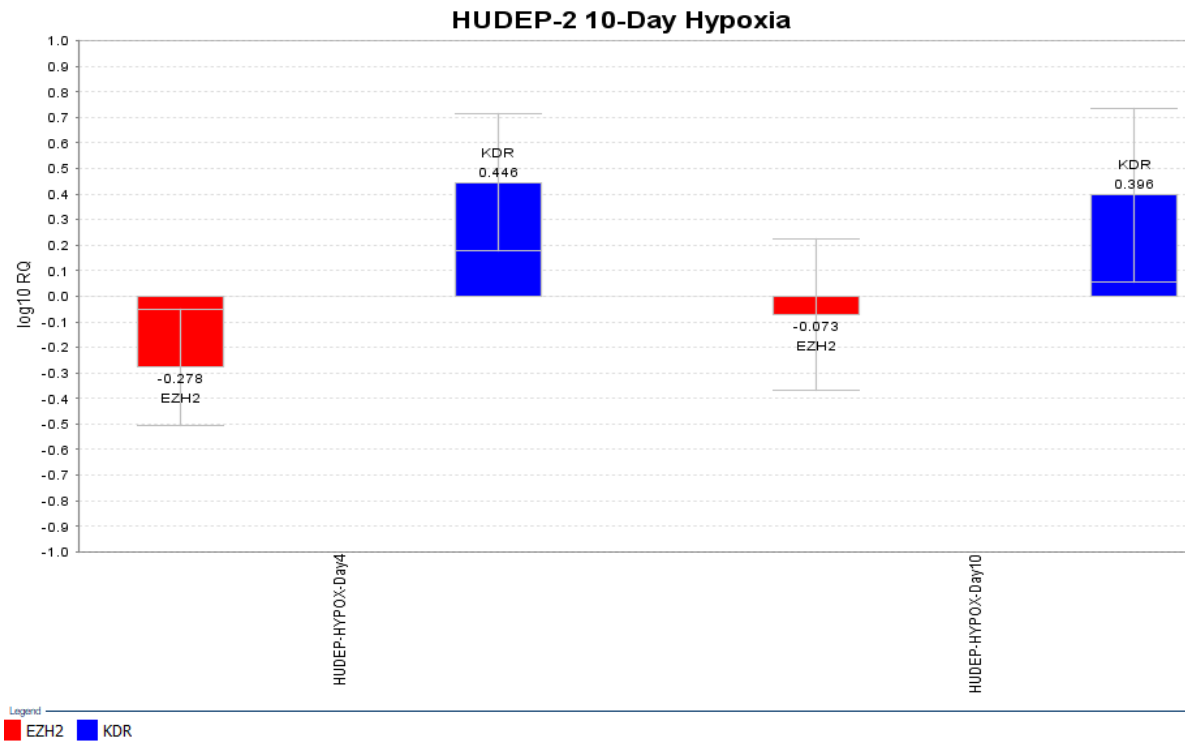
The sickle-cell mutation at position 20 of codon six on the beta-globin gene (A) triggers hemoglobin polymerization (B). The red blood cell (RBC) sickling and changes in haematological indices that ensue (C) cause changes to RBC rheology and blockage of blood vessels (vaso-occlusion) (D). The sickled and rigid RBCs breakdown (haemolysis) (E), releasing hemoglobin and its breakdown product (haem), as well as enzymes including arginase. Haem depletes nitric oxide (NO), the regulator of endothelial cell (EC) function. Arginase breaks down L-arginine, the precursor for NO synthesis. The cascade of reactions leads to endothelial cell dysfunction and vascular occlusion leading to ischaemia (ISC) (Adapted from ³⁸⁵, available via license CC BY 3.0).

Appendix IV: Expression of hypoxia inducible factor (HIF) associated genes in HUDEP-2 and K562 cells after 24-hour and 10-Day hypoxia

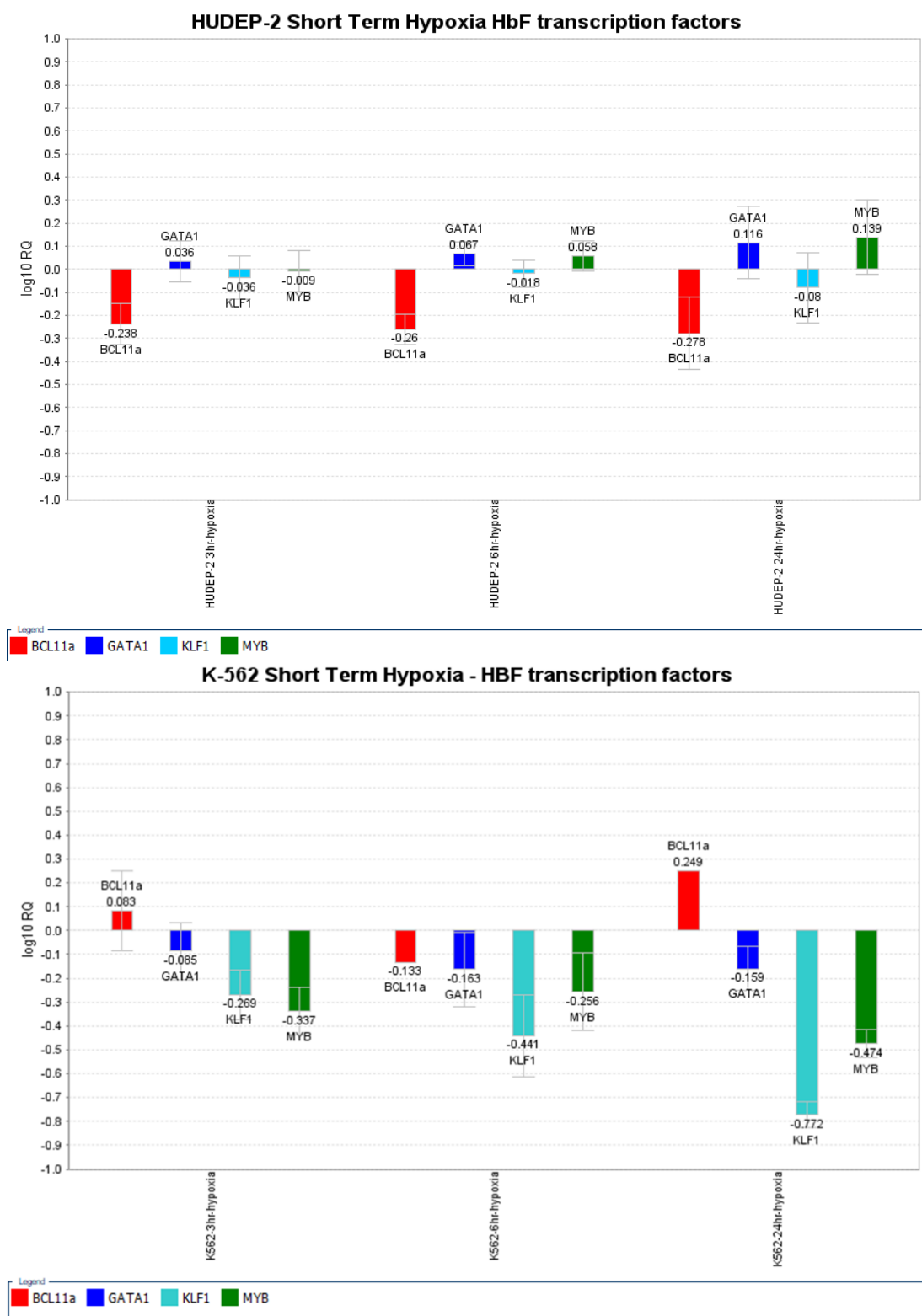


Appendix V: Expression of genes connecting hypoxia and hemoglobin production pathways in HUDEP-2 and K562 cells after 24-hour and 10-Day hypoxia

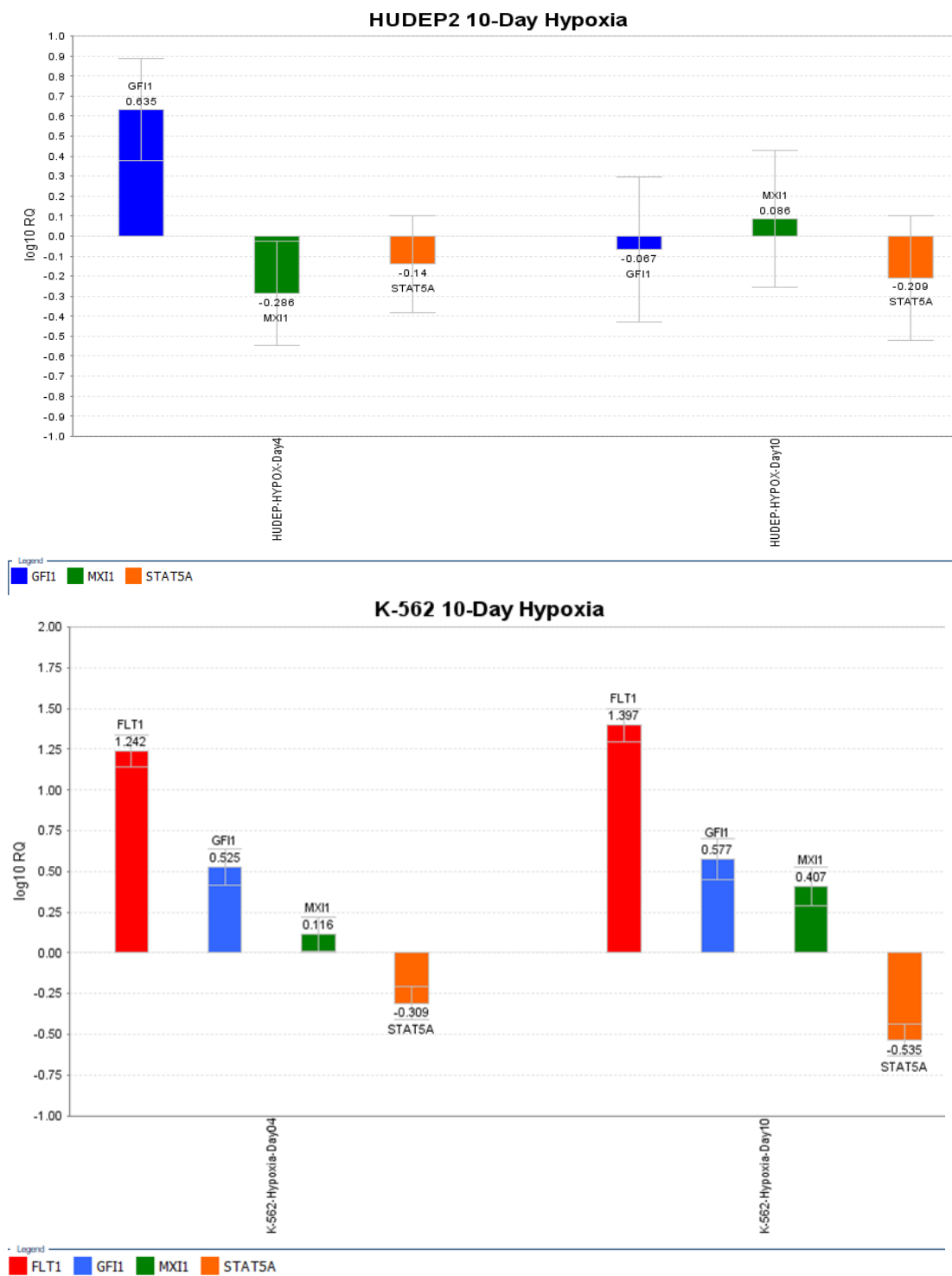




Appendix VI: Expression of transcription factors involved in hemoglobin production in HUDEP-2 and K562 cells after 24-hour hypoxia



Appendix VII: Expression of transcription factors whose binding motifs are disrupted by *FLT1* regulatory region variants in HUDEP-2 and K562 cells after 10-Day hypoxia



Appendix VIII: Research Material and other web resources

During the time of my registration as a PhD student, I assisted in multiple group projects and also worked part-time as data analyst for the Genetic Medicine of African Populations (GeneMAP) research group. In these capacities, and as part of my research project, I developed multiple robust workflow for data analysis that contributed to multiple publications. These demonstrate the skills I amassed during my registration, and my ability to work collaboratively. Some of the workflows which continue to be used within the group to process large-scale genomic data, and also available for public use, include:

1. **Genome-wide association study (GWAS):** While the scripts used in our GWAS study are made publicly available on github (<https://github.com/GeneMAP-Research/genemapgwas>) as a journal requirement prior to publication of our results, we intend to finalize development of the workflow and release it as a web-based platform to enhance such analyses in other African populations. In particular, we were able to incorporate a critical piece of the pipeline in our workflow that has been rather elusive to other research groups; the processing of raw data files in Illumina intensity (IDAT) format to the more popular VCF and PLINK binary formats. The workflow was used for our GWAS analysis.
2. **Next-generation sequencing (NGS) analysis** (<https://github.com/GeneMAP-Research/genemapngs>): We developed a robust workflow for whole-genome and whole-exome sequencing analysis that is capable of cutting the processing time for thousands of samples by several folds. This has already been utilized in many of our group projects. The workflow was used in the generation of the custom genotype imputation panel in our study.
3. **Genotype imputation service** (<https://github.com/GeneMAP-Research/genemapimputationservice>): We developed a workflow for haplotype estimation (phasing), genotype imputation, and custom panel creation. We plan to release this as a web-based tool. It implements a variety of tools for phasing and imputation. An important innovation is that it will allow meta-imputation in which imputed data from different reference imputation panels will be merged together, generating a single score that can be used for filtering. This will significantly alleviate the challenges that come with handling such datasets separately as revealed in our GWAS study. Together with the NGS workflow, this contributed in the generation of our custom panel and in-house genotype imputation.
4. **Mutation age estimation and haplotype analysis** (<https://github.com/GeneMAP-Research/mutationage>): We developed a workflow for estimation of mutation age given

select markers, and haplotype analysis. This contributed in at least two publications as of 2023.

These workflows remain under the custody of the GeneMAP Research Center lead by Professor Ambroise Wonkam in the Division of Human Genetics and are available for public use.

Other web resources

TOPMed Imputation Service: <https://imputation.biodatacatalyst.nhlbi.nih.gov/>

Michigan Imputation Service: <https://imputationserver.sph.umich.edu/>

Sanger Imputation Service: <https://imputation.sanger.ac.uk/>

H3Africa genotyping array: <https://www.illumina.com/content/dam/illumina/gcs/assembled-assets/marketing-literature/infinium-h3africa-consortium-array-data-sheet-370-2020-001/infinium-h3africa-consortium-array-data-sheet-370-2020-001.pdf>

Human reference genome, GRCh37:

<http://hgdownload.cse.ucsc.edu/goldenpath/hg19/bigZips/latest/hg19.fa.gz>

IAAP Genotyping command line interface: <https://emea.support.illumina.com/downloads/iaap-genotyping-cli.html>

conform-gt: <https://faculty.washington.edu/browning/conform-gt.html>

LiftOver: <https://genome.ucsc.edu/cgi-bin/hgLiftOver>

FUMA: <https://fuma.ctglab.nl/>

FUMA Job parameters:

<https://github.com/esohkevin/hbfgwas/blob/main/functionalmapping/params.config>

UCSC Browser: <https://genome.ucsc.edu/index.html>

GTEx Portal: <https://www.gtexportal.org/home/>

ENCODE Project: <https://www.encodeproject.org/>

VISTA Enhancer: <https://enhancer.lbl.gov/>

JASPAR: <https://jaspar.genereg.net/>

Appendix IX: Associations for the merged, non-imputed, genotype dataset of Cameroon and Tanzania with FDR < 0.20

CHROM	BP	RSID	REF	ALT	N	BETA	SE	P	FDR
chr1	57789439	rs267645	G	A	1675	-0.08	0.018	2.11E-06	0.084988
chr1	57789928	rs76054128	G	T	1679	-0.08	0.018	6.22E-06	0.172604
chr2	60291204	rs11890420	C	T	1680	0.07	0.016	5.65E-06	0.172604
chr2	60698397	rs12477097	A	C	1677	-0.08	0.016	2.12E-07	0.011531
chr2	60710738	rs11692396	G	A	1680	0.20	0.017	1.50E-30	2.67E-25
chr2	60718043	rs1427407	T	G	1682	0.26	0.017	6.58E-50	8.21E-44
chr2	60719074	rs1896294	C	T	1682	0.24	0.016	1.06E-46	6.63E-41
chr2	60719970	rs766432	C	A	1681	0.23	0.016	3.46E-42	8.65E-37
chr2	60720246	rs11886868	C	T	1679	0.24	0.016	2.04E-45	8.50E-40
chr2	60720589	rs10195871	A	G	1680	0.23	0.016	7.37E-44	2.30E-38
chr2	60722040	rs6706648	C	T	1682	-0.21	0.015	9.76E-40	2.03E-34
chr2	60729294	rs10184550	G	A	1680	0.11	0.018	4.50E-09	0.000468
chr2	60736529	rs7340264	G	A	1680	-0.16	0.019	2.67E-17	4.17E-12
chr2	60736837	rs6724431	C	T	1681	0.09	0.017	1.60E-07	0.009528
chr2	60740323	rs6740172	G	A	1681	0.08	0.017	6.09E-06	0.172604
chr2	60743605	rs72962602	A	G	1682	-0.16	0.019	2.22E-15	3.07E-10
chr2	60797896	rs2901066	A	C	1681	0.10	0.016	2.45E-10	2.78E-05
chr2	60799121	rs356982	G	A	1682	0.08	0.016	6.10E-07	0.029294
chr2	60800641	rs971742	T	C	1680	0.08	0.016	2.04E-07	0.011531
chr2	60830276	rs61632082	T	G	1682	-0.20	0.031	1.60E-10	1.99E-05
chr2	60851349	rs842773	A	C	1682	0.08	0.017	1.43E-06	0.05936
chr2	60910033	rs10198826	C	T	1682	0.08	0.017	3.06E-06	0.112436
chr2	60914577	rs6732794	C	A	1676	0.09	0.017	4.15E-07	0.020711
chr2	60952810	rs113481561	A	G	1682	-0.15	0.027	8.12E-08	0.006337
chr2	60967758	rs72880218	G	A	1682	-0.16	0.028	3.15E-08	0.003023
chr2	61077693	rs1177210	A	G	1649	-0.09	0.021	5.82E-06	0.172604
chr2	61079440	rs1177214	C	A	1681	-0.10	0.021	6.75E-06	0.179316
chr2	61095245	rs842639	G	A	1680	0.09	0.016	1.09E-07	0.007594
chr2	61119471	rs842647	G	A	1682	0.09	0.016	1.40E-07	0.00921
chr2	61142317	rs72887242	C	T	1681	-0.14	0.027	1.59E-07	0.009528
chr2	61204856	rs7608910	A	G	1679	-0.07	0.015	6.13E-06	0.172604
chr2	61213261	rs13398512	C	A	1677	-0.07	0.016	3.66E-06	0.130413
chr2	61224259	rs10181042	C	T	1681	0.08	0.015	3.85E-07	0.020014
chr2	61225330	rs10207958	A	C	1668	0.08	0.016	6.93E-07	0.03204
chr2	2.16E+08	rs12105093	C	T	1681	-0.11	0.023	7.38E-07	0.032908
chr2	2.36E+08	rs7569222	G	A	1682	0.09	0.019	7.21E-06	0.18756
chr2	2.37E+08	rs76847022	C	T	1675	0.12	0.024	2.58E-06	0.097776
chr3	25014458	rs113927714	T	G	1680	-0.17	0.036	4.17E-06	0.140796
chr6	1.35E+08	rs2064101	G	A	1679	-0.11	0.021	4.79E-08	0.004269

<i>chr6</i>	1.35E+08	rs2072825	A	G	1681	-0.08	0.017	4.05E-06	0.14056
<i>chr6</i>	1.35E+08	rs4289677	T	C	1680	-0.10	0.020	1.12E-06	0.048405
<i>chr6</i>	1.35E+08	rs9389269	T	C	1679	0.21	0.046	6.13E-06	0.172604
<i>chr6</i>	1.35E+08	rs9494145	T	C	1679	0.20	0.036	7.23E-08	0.006013
<i>chr11</i>	1.33E+08	rs6590705	C	A	1680	-0.09	0.020	6.68E-06	0.179316
<i>chr13</i>	29094555	rs75294023	A	G	1680	0.16	0.030	1.00E-07	0.007378
<i>chr19</i>	45596073	rs8111966	C	A	1681	-0.19	0.041	5.72E-06	0.172604
<i>chr20</i>	1339076	rs6033534	G	A	1678	-0.15	0.033	4.95E-06	0.162576
<i>chr20</i>	8298480	rs6086403	C	A	1666	0.12	0.025	2.58E-06	0.097776

Appendix X: Associations for the meta-analysis of Cameroon and Tanzania non-imputed datasets with FDR < 0.20

CHROM	BP	RSID	REF	ALT	N	BETA	SE	P	FDR
chr2	60291204	2:60291204:SNP	T	C	1711	-0.08	0.015	2.19E-07	0.015654
chr2	60698397	2:60698397:SNP	A	C	1711	0.09	0.016	1.06E-08	0.001275
chr2	60710738	2:60710738:SNP	A	G	1711	-0.20	0.017	2.66E-32	7.60E-27
chr2	60718043	2:60718043:SNP	T	G	1711	0.27	0.018	3.39E-50	7.76E-44
chr2	60719074	2:60719074:SNP	T	C	1711	-0.24	0.017	1.35E-47	1.54E-41
chr2	60719918	2:60719918:SNP	A	G	1711	-0.13	0.026	4.34E-07	0.029231
chr2	60719970	2:60719970:SNP	A	C	1711	-0.23	0.017	2.10E-43	9.61E-38
chr2	60720246	2:60720246:SNP	T	C	1711	-0.24	0.017	4.47E-46	3.41E-40
chr2	60720589	2:60720589:SNP	A	G	1711	0.24	0.017	9.57E-45	5.47E-39
chr2	60722040	2:60722040:SNP	T	C	1711	-0.21	0.016	1.80E-41	6.88E-36
chr2	60726427	2:60726427:SNP	T	C	1711	-0.10	0.022	2.35E-06	0.107555
chr2	60729294	2:60729294:SNP	A	G	1711	-0.11	0.018	1.55E-10	2.22E-05
chr2	60736529	2:60736529:SNP	A	G	1711	-0.17	0.019	6.80E-19	1.56E-13
chr2	60736837	2:60736837:SNP	T	C	1711	-0.09	0.016	1.32E-08	0.00144
chr2	60740323	2:60740323:SNP	A	G	1711	-0.08	0.017	5.78E-07	0.036729
chr2	60743605	2:60743605:SNP	A	G	1711	0.16	0.019	6.83E-17	1.42E-11
chr2	60744331	2:60744331:SNP	A	G	1711	0.07	0.015	2.33E-06	0.107555
chr2	60797896	2:60797896:SNP	A	C	1711	0.10	0.015	1.97E-11	3.47E-06
chr2	60799121	2:60799121:SNP	A	G	1711	0.08	0.016	8.27E-08	0.007278
chr2	60800641	2:60800641:SNP	T	C	1711	0.09	0.016	3.11E-08	0.003089
chr2	60830276	2:60830276:SNP	T	G	1711	0.21	0.030	7.54E-12	1.44E-06
chr2	60851349	2:60851349:SNP	A	C	1711	-0.08	0.017	2.82E-06	0.126535
chr2	60914577	2:60914577:SNP	A	C	1711	-0.08	0.017	9.08E-07	0.05197
chr2	60952810	2:60952810:SNP	A	G	1711	0.15	0.026	1.15E-08	0.001317
chr2	60967758	2:60967758:SNP	A	G	1711	-0.16	0.028	6.16E-09	0.00083
chr2	60971274	2:60971274:SNP	A	G	1711	-0.10	0.021	1.42E-06	0.073593
chr2	61073814	2:61073814:SNP	T	C	1711	-0.10	0.021	4.10E-07	0.028446
chr2	61077693	2:61077693:SNP	A	G	1711	0.09	0.020	2.95E-06	0.129691
chr2	61079440	2:61079440:SNP	A	C	1711	-0.10	0.021	1.07E-06	0.058954
chr2	61095245	2:61095245:SNP	A	G	1711	-0.08	0.016	1.19E-07	0.00975
chr2	61116481	2:61116481:SNP	A	G	1711	0.09	0.019	4.03E-06	0.167678
chr2	61119471	2:61119471:SNP	A	G	1711	-0.08	0.016	8.60E-08	0.007292
chr2	61142317	2:61142317:SNP	T	C	1711	-0.14	0.026	2.72E-08	0.002829
chr2	61213261	2:61213261:SNP	A	C	1711	-0.07	0.016	1.77E-06	0.089756
chr2	61224259	2:61224259:SNP	T	C	1711	-0.07	0.015	7.95E-07	0.047864
chr2	61225330	2:61225330:SNP	A	C	1711	0.07	0.015	1.28E-06	0.068333
chr2	61235855	2:61235855:SNP	A	C	1711	-0.07	0.016	4.85E-06	0.191397
chr2	2.37E+08	2:237099162:SNP	T	C	1711	0.11	0.024	2.34E-06	0.107555
chr6	1.35E+08	6:135250530:SNP	A	G	1711	0.10	0.020	1.75E-07	0.013817

<i>chr6</i>	1.35E+08	6:135253451:SNP	A	G	1711	-0.08	0.017	3.80E-06	0.160909
<i>chr6</i>	1.35E+08	6:135275485:SNP	T	C	1711	0.09	0.019	1.91E-06	0.095168
<i>chr6</i>	1.35E+08	6:135287533:SNP	A	G	1711	0.08	0.018	4.49E-06	0.180382
<i>chr6</i>	1.35E+08	6:135427159:SNP	T	C	1711	-0.23	0.045	2.01E-07	0.014993
<i>chr6</i>	1.35E+08	6:135432552:SNP	T	C	1711	-0.19	0.035	3.25E-08	0.003101
<i>chr11</i>	1.33E+08	11:133334522:SNP	A	C	1711	0.10	0.019	5.06E-07	0.033084
<i>chr13</i>	29094555	13:29094555:SNP	A	G	1711	-0.14	0.029	1.08E-06	0.058954
<i>chr20</i>	8298480	20:8298480:SNP	A	C	1711	0.12	0.025	8.66E-07	0.05082
<i>chr20</i>	20169399	20:20169399:SNP	A	G	1711	0.21	0.045	4.28E-06	0.175022