Identifying genetic variants and pathways associated with extreme levels of fetal hemoglobin in Sickle Cell Disease in Tanzania.

Siana Nkya[1,2*], Liberata Mwita[2], Josephine Mgaya[2], Happiness Kumburu[3], Marco van Zwetselaar [3], Stephan Menzel[4], Gaston K. Mazandu[5,6,7], Raphael Z. Sangeda[2,8], Emile R. Chimusa[5], Julie Makani[2].

[1]Department of Biological Sciences, Dar es Salaam University College of Education, Dar es Salaam, Tanzania.

[2]Sickle Cell Program, Department of Hematology and Blood Transfusion, Muhimbili University of Health and Allied Sciences, Dar es Salaam, Tanzania.

[3]Department of Biotechnology Laboratory, Kilimanjaro Clinical Research Institute, Kilimanjaro, Tanzania.

[4]Department of Molecular Hematology, King's College of London, London, UK.

[5]Department of Pathology, Division of Human Genetics, University of Cape Town, IDM, Cape Town, South Africa.

[6]Department of Biomedical Sciences, Computational Biology Division, University of Cape Town, South Africa.

[7]African Institute for Mathematical Sciences, Muizenberg 7945, Cape Town, South Africa

[8]Department of Pharmaceutical Microbiology, Muhimbili University of Health and Allied Sciences, Dar es Salaam, Tanzania.

*Correspondence: Siana Nkya, E-mail: snkyamtatiro@gmail.com

Abstract

Sickle cell disease (SCD) is a blood disorder caused by a point mutation on the beta globin gene resulting in the synthesis of abnormal hemoglobin. Fetal hemoglobin (HbF) reduces disease severity, but the levels vary from one individual to another. Most research has focused on common variants which differ across populations and hence do not fully account for HbF variation. To investigate rare and common variants influencing HbF levels in SCD, we performed targeted next generation sequencing covering exonic and other significant fetal hemoglobin-associated loci, including ***BCL11A***, ***MYB***, ***HOXA9***, ***HBB***, ***HBG1***, ***HBG2***, ***CHD4***, ***KLF1***, ***MBD3***, ***ZBTB7A*** and ***PGLYRP1***. Results revealed a range of functionally relevant genetic variants. Notably, there were significantly more deletions in individuals with high HbF levels (11% vs 0.9%). We identified frameshift deletion in individuals with high HbF levels and frameshift insertions in individuals with low HbF. ***CHD4*** and ***MBD3*** genes, interacting in the same sub-network, were identified to have a significant number of pathogenic or non-synonymous mutations in individuals with low HbF levels, suggesting an important role of epigenetic pathways in the regulation of HbF synthesis. This study provides new insights in selecting essential variants associated with extreme HbF levels in SCD.

Keywords: Sickle cell disease; Genetic disorder; Fetal hemoglobin; Hemoglobinopathy; Tanzania

Introduction

Sickle cell disease (SCD) and thalassemia are the most common hemoglobinopathies worldwide, with 270 million carriers and 300,000 to 500,000 annual births [1]. Up to 70% of global SCD annual births occur in sub-Saharan Africa. Reports show that 50% to 80% of affected children in these countries die annually [2]. Tanzania ranks fifth worldwide regarding the number of children born with SCD, estimated at 8,000-11,000 births annually. 15-20% of the population are SCD carriers (HbAS) and therefore potential parents of future babies with SCD [3, 4]. Without intervention, it is estimated that up to 50% of children with SCD will die before the age of 5 years [1]. Thus, SCD intervention at early stages of life may prevent premature deaths and reduce under-five mortality.

SCD is a monogenic condition resulting from a single mutation in the β-globin gene (*HBB*), on chromosome 11, leading to the production of an abnormal β-hemoglobin chain (HbS). SCD affects multiple organs and hence is a multisystem disease. Clinical manifestations vary immensely, with some individuals being entirely asymptomatic while others suffer from severe forms of the disease. The marked phenotypic heterogeneity of SCD is due to both, genetic and environmental determinants [5]. A major disease modifier is the presence of fetal hemoglobin (HbF): high HbF levels are associated with reduced morbidity and mortality [6, 7].

Hemoglobin is a tetrametric molecule composed by 2-alpha-globin and 2 gamma globin molecules in HbF and 2 alpha-globin and two 2 beta-globin molecules in HbA [8]. HbF is normally expressed during the development of the fetus and starts to decline just before birth, when it is replaced by adult hemoglobin (HbA) in normal individuals and hemoglobin S (HbS) in individuals with SCD [9]. Red blood cells of normal adults (HbAA) contain mainly hemoglobin A (HbA), with 2.5-3.5% Hemoglobin A_2 (HbA_2), and < 1% HbF [10]. However, 10% to 15% of adults possess higher HbF levels (up to 5.0%). Although this has no significant consequences in healthy individuals, HbF background variability in SCD can reach levels with clinical benefit to patients [11]. Consequently, efforts to understand and control the production of HbF in SCD patients may result in interventions of significant clinical benefit to individuals with SCD.

The levels of both HbF and F cells (erythrocytes with measurable amounts of HbF) are highly heritable traits [12] with up to 89% of variation being influenced by genetic factors. The remaining proportion is accounted for by age, sex and environmental factors. It is now clear that HbF is a quantitative trait which is shaped by genetic factors both linked and unlinked to the β-globin gene. Three main loci, namely *BCL11A* on chromosome 2, *HMIP* on chromosome 6, and *HBG* on chromosome 11, have been identified across populations as associated with HbF levels [13, 14, 15]. The variants in these loci have been reported to contribute 20-50% of HbF variation in non-African populations, however the impact of these variants is different from one population to another. An example is a strong variant at *HMIP*, which is rare in the Tanzanian population and hence has a smaller impact on HbF levels there [16, 17]. HbF levels in SCD, as a quantitative trait, is expected to be influenced by other polymorphisms, including insertions/deletions, rare mutations or copy number variations [15].

New genetic and proteomic techniques have led to the identification of several HbF expression regulators have been identified. *Kruppel like factor* (*KLF*1) has been reported as one of the key regulators of HbF expression. Reports indicate that *KLF1* has dual functions: direct activation of HbF expression through activation of β-globin [18] and an indirect silencing of γ-globin gene through *BCL11A1* [19]. Other players within the HbF regulation network that have been reported include *GATA1, FOG1* and *SOX6*, which are erythroid transcription factors and are believed to interact with *BCL11A* in HbF regulation [20]. In addition, nuclear receptors *TR2/TR4* which are associated with *corepressors of DNA methyltransferase 1* (*DNMT1*) and *lysine-specific demethylase 1* (*LSD1*) have also been implicated. *DNMT1* and *LSD1* are a part of the DRED complex, a known repressor of embryonic and fetal globin genes in adults [21]. Recently, studies of epigenetic pathways of HbF regulation have elucidated the involvement of the *nucleosome remodeling and deacetylase* (NuRD) complex [22, 23].

Despite the high prevalence of SCD in Africa, African patient populations remain understudied. Unique insight can be obtained from these patients, considering the substantial African genetic diversity and exceptional mapping resolution. The high burden of SCD in sub Saharan Africa makes it important that genetic studies, ultimately aimed at improved therapeutic intervention, are carried out in African countries. To address this, we conducted a Genome Wide Association Study (GWAS) [16, 17, 24] and candidate genotyping for HbF in Tanzanian individuals with SCD, which led to validation of known HbF variants and identification of novel ones. This report documents a follow-up study aimed at performing in-depth targeted sequencing around previously identified loci to descriptively compare, in detail, discovered polymorphisms between individuals with extreme HbF levels. For the first time, we have conducted targeted next-generation sequencing to investigate known and novel genetic variants and pathways associated with extreme HbF levels in individuals with Sickle cell disease (SCD) in Tanzania. From these selected individuals, we have identified different types of polymorphisms, including SNPs, INDELS, suggesting potential modifier effect. Interestingly, key discovered variants, together with previously identified variants, are enriched in biological pathways that underlie the HbF regulation.

Materials and Methods

Study design and population:

We performed a cross-sectional study involving the Muhimbili SCD cohort, which has been described previously [3]. Written informed consent was obtained for each patient and ethical approval given by the Muhimbili University Research and Publications Committee (MU/RP/AEC/VOLX1/33 and 2017-03-06/AEC/Vol X11/65). The study involved 14 individuals confirmed to have SCD (HbSS or S-β°thalassaemia), over five years old, with extreme HbF levels. Excluded were individuals confirmed to be AS or AA following Hb electrophoresis and HPLC, those with HbF measured at an age of less than 5 years, with inconclusive SCD laboratory diagnosis where a repeat test for confirmation could not be performed, and individuals who were on hydroxyurea therapy.

Phenotyping

Individuals were selected using previously collected HbF data. In this population, the median HbF was 4.6 [Interquartile range (IQR): 2.5-7.7)] [17] and therefore 0-2.5% was considered a

low HbF level (HbF≤2.5%) while 7.7% and above was considered a high HbF level (HbF≥7.7%).

Sequencing

DNA was extracted from archived buffy coat samples using the Nucleon BACC II system (GE Healthcare, Little Chalfont, UK). The sequencing panel was adopted from a research panel at King's College London and customized using Illumina DesignStudio (https://designstudio.illumina.com/). Targeted sequencing covered exons and non-coding regions around validated and candidate fetal hemoglobin-influencing loci, including *B-cell lymphoma/leukemia 11A* (*BCL11A*), *proto-oncogene, transcription factor* (*MYB*), *homeobox A9* (*HOXA9*), *hemoglobin subunit beta* (*HBB), hemoglobin subunit gamma 1* (*HBG1*), *hemoglobin subunit gamma 2* (*HBG2*), *chromodomain helicase DNA binding protein 4* (*CHD4*), *Kruppel like factor 1* (*KLF1*), *methyl-CpG binding domain protein 3* (*MBD3*), *zinc finger and BTB domain containing 7A* (*ZBTB7A*), *peptidoglycan recognition protein 1* (*PGLYRP1*) on chromosomes 2, 6, 7, 11, 12 and 19, respectively (see **Table 2**). Selection of target regions was based on previous associated known and novel loci in the studied population and those reported recently in other populations. Sequencing was performed on the Illumina MiSeq platform at the Kilimanjaro Clinical Research Institute (*KCRI*), Tanzania, following TruSeq Custom Amplicon Low Input Kit protocol.

Reads Mapping, Alignment, Variant Calling and Variant Calling Quality Control

Figure 1 illustrates and summarizes the pipeline used from alignment to prioritization of mutation. We reconstructed the reads by realigning them to the complete reference genome build hg38 using BWA [25]. The Picard tool kit [26] was used to sort and mark reads duplication, after alignment. We used an ensemble approach implemented in VariantMetaCaller [27] that may find a call consensus in detecting SNPs and short indels [28]. The best practice specific to each caller were adopted [29]. We combined information generated from two independent variant caller pipelines: (1) An incremental joint variant discovery implemented in GATK 3.0 HaplotypeCaller [26], which calls samples independently to produce gVCF files and leverages the information from the independent gVCF file to produce a final call-set at the genotyping step; (2) bcftools via mpileup [30, 31] variant callers (**Figure 1**). The final call-set from each subject group, were produced from VariantMetaCaller [27].

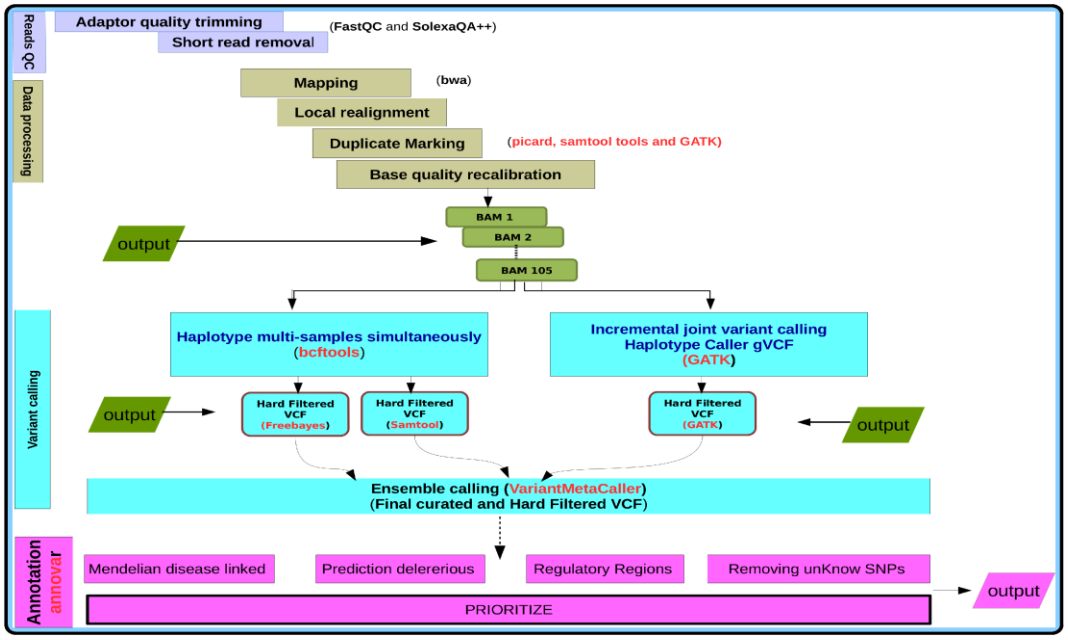


Figure 1. Workflow of the data analysis. Describes the bioinformatics pipelines from alignment of DNA reads, variants calling to in silico mutation prioritization.

Annotation, In Silico Prediction of Mutation and Prioritization

High confidence variants were called using VariantMetaCaller [27] from the dataset including 14 Tanzanian SCD patients (nine with high HbF and five with low HbF levels). We used ANNOVAR [32] to perform gene-based annotation to detect whether SNPs cause protein coding changes and to produce a list of the amino acids that are affected. ANNOVAR contains up to 21 different functional scores including SIFT [33, 34], LRT [35], MutationTaster, MutationAssessor [36], FATHMM [37], fathmm-MKL [38], RadialSVM, LR, PROVEAN, MetaSVM, MetaLR, DANN, M-CAP, Eigen, GenoCanyon [39], CADD [40], GERP++ [41], Polyphen2 HVAR, Polyphen2 HDIV [42] and PhyloP, SiPhy [43].

From the resulting functional annotated dataset, we first filtered variants for rarity, exonic variants, non-synonymous, stop codons; predicted functional significance and deleteriousness [33, 34]. First, the resulting functional annotated data set was independently filtered for predicted functional status (of which each predicted functional status is "deleterious" (D), "probably damaging" (D), "disease_causing_automatic" (A) or "disease_causing" (D) [44, 45, 46] from these 21 in silico prediction mutation tools. Recent evaluation of in silico prediction tools for mutation effects suggested these tools are quite similar [47]. However, the evaluation of these tools was conducted mostly in non-African populations. Here we opted for an extreme casting vote approach to retain only a variant if it had at least 17 predicted functional status “D” or “A” out of 21, as one can expect a true in silico mutant variant to similarly be reported from most of these tools. Second, the retained variants were further filtered for rarity, exonic variants, nonsynonymous mutations, yielding a final candidate list of predicted mutant and genetic modifier variants.

Network and Enrichment Analysis

To find out how predicted in silico mutant and modifier genes interact with others at the systems level, we analyzed how the set of all interactive genes from knowledge-based Protein-Protein Interaction (PPI) interacted with our identified in silico mutant genes and the rest of targeted genes, respectively. This to identify potential biological pathways in which these genes participate. To achieve this, we first mapped the identify mutant SNPs to their closest genes. We mapped genes to a comprehensive human Protein-Protein Interaction (PPI) network [48, 49] to identify sub-networks containing mutant and genetics modifier variant genes and their interactions. Using the Enrichr software [50], we examined how closely these genes within the extracted sub-networks are associated with human phenotypes and elucidate biological processes and pathways in which these genes participate, molecular functions and association with potential human phenotypes. The most significant pathway enriched for genes in the networks were selected from KEGG [51], Panther [52], Biocarta [53] and Reactome [54]. Gene ontologies, including molecular functions and biological process, from the Gene Ontology database [55].

Results

Sample Characterization

This study involved 14 SCD individuals with extreme (9 with high and 5 with low) HbF levels. **Table 1**, describes the age and HbF ranges of the included individuals.

Table 1. Characteristics of Tanzanian individuals sickle cell disease (SCD) with extreme fetal hemoglobin levels.

	High HbF ≥7.7%	Low HbF≤2.5%
N	9	5
Age range (Years)	5 - 19	8 - 21
HbF (%)	15-32	0.3-2.2

Summary of variants found in individuals with high and low HbF levels

A total of 873 and 1196 highly confident variants were determined in SCD patients with high and low HbF levels, respectively, on chromosomes 2, 6, 7, 11, 12 and 19. Surprisingly, this shows a difference in the overall variation between the two groups of individuals with SCD.

The identified variants are comprised of 77% and 82 % biallelic SNPs, 0.15% and 0.11% multi-allelic SNPs, 11% and 0.9% deletions and 0.9% and 0.7% insertions in patients with high and low HbF levels (adjusted $\chi 2$ p-values = 1.16e-03 and 2.96e-06, as compared to uniform distribution), respectively. From these discovered variants, we detect 1 and 0 frameshift-deletions, 2 and 4 frameshift-insertions, 1 and 1 non-frameshift-insertions, 34 and 41 nonsynonymous, 3 and 3 stop-gain, 49 and 60 synonymous variants in SCD individuals with high/low HbF level, respectively. Based on our targeted chromosomal sequencing, we found significant difference in coverage of variants in the molecular structure (**Figure 2**) between

SCD patients with high and low HbF level (adjusted Fisher exact p-value = 6.1e-04), at 3'untranslated region (3'UTR) (2.9% versus 4.2%), 5' untranslated region (5'UTR) (4.2% versus 0.68%), upstream (0.2%, 2.9%). Critically, we observed that patients with high HbF have 0% variants in splicing regions, while patients with low HbF level have 1.48% (**Figure 2**).

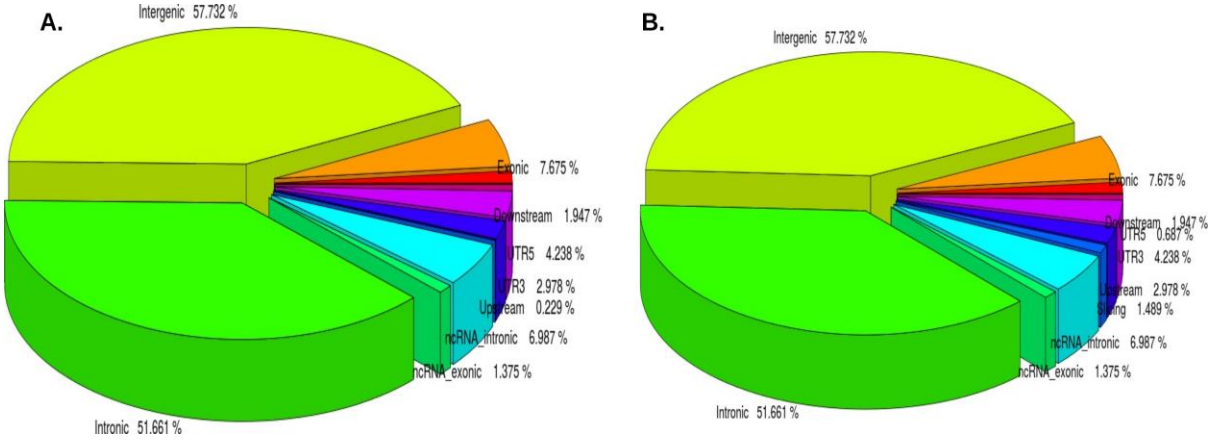


Figure 2. Characterization of SCD gene function and exome map from the targeted next generation sequencing: This included the exon and full regions for validated and novel fetal hemoglobin-associated loci, including *B-cell lymphoma/leukemia 11A* (*BCL11A*), *proto-oncogene, transcription factor* (*MYB*), *Homeobox A9* (*HOXA9*), *Hemoglobin subunit beta* (*HBB), hemoglobin subunit gamma 1* (*HBG1*), *hemoglobin subunit gamma 2* (*HBG2*), *chromodomain helicase DNA binding protein 4* (*CHD4*), *Kruppel like factor 1* (*KLF1*), *methyl-CpG binding domain protein 3* (*MBD3*), *zinc finger and BTB domain containing 7A* (*ZBTB7A*), *Peptidoglycan recognition protein 1* (*PGLYRP1*) on chromosomes 2, 6, 7, 11, 12 and 19, respectively . (A) Gene functions from patients with high HbF level.

Table 2. Characterization of polymorphisms within mutant and modifiers genes in SCD patients from Tanzania. Details of gene variants can be found in Supplementary File: **Table S1**.

(High; low HbF level)

GENE	#Polymorphisms	#MNP	#SNPs	#Deletion	#Insertion	#Pathogenic	#Benign	#USig*
PGLYRP1	**4; 4**	**0; 0**	**4; 4**	**0; 0**	**0; 0**	**0; 0**	**1; 0**	**3;4**
ZBTB7A	**13; 11**	**0; 1**	**10; 9**	**0; 0**	**1; 1**	**0; 0**	**2; 2**	**11;9**
CHD4	**25; 32**	**3; 3**	**19; 27**	**1; 0**	**1; 3**	**2; 5**	**3; 4**	**20;23**
MBD3	**14; 19**	**0; 2**	**12; 14**	**1; 1**	**1; 4**	**1; 2**	**0; 1**	**12;17**
KLF1	**11; 4**	**1; 0**	**10; 4**	**0; 0**	**0; 0**	**0; 0**	**1; 1**	**10;3**
MYB	**24; 27**	**1; 1**	**20; 23**	**0; 2**	**3; 1**	**0; 0**	**3; 1**	**21;26**
BCL11A	**27; 27**	**1; 2**	**25; 21**	**1; 2**	**0; 2**	**0; 0**	**4; 1**	**23; 26**
HBG2	**5; 17**	**1; 1**	**3; 12**	**0; 2**	**1; 2**	**0; 0**	**0; 0**	**5; 17**
HOXA9	**2; 2**	**0; 0**	**1; 1**	**0; 0**	**1; 1**	**0; 0**	**0; 0**	**2; 2**
HBB	**9; 10**	**0; 0**	**9; 10**	**0; 0**	**0; 0**	**0; 0**	**1; 1**	**8; 9**

Abbreviations: USig* is the number variant with uncertain significance of pathogenicity.

Potential pathogenic variants

Although African-specific reported pathogenic variants are underrepresented in current databases of pathogenic variants [56], here we aimed at descriptively characterizing possible pathogenic variants from the set of polymorphisms in the retained candidate in silico mutant genes and our initial target genes discovery variants between the two patient groups. Following our pipeline and mutation prioritization, we identified six SNPs in genes (*ZBTB7A, CHD*4, *HBB, PGLYRP1, MBD3* and *MYB*) with functional impact (**Table 2** and Supplementary File: **Table S2**) in both data generated from the SCD patients with high and low HbF levels. Two genes, *CHD4* and the *MBD3,* were found with a difference in the number of pathogenic variants (**Table 3** and Supplementary File: **Table S2**): individuals with SCD with low HbF levels were found to have more pathogenic, benign or uncertain significant pathogenic variants. Individuals with SCD with lower HbF levels had a significantly higher number of variants with insertions at both *CHD4* and *MBD3* than patients with high HbF levels (**Table 2** and Supplementary File: **Table S1**). While both groups have small numbers of deletion variants, individuals with low HbF level had fewer deletions than those with high HbF level.

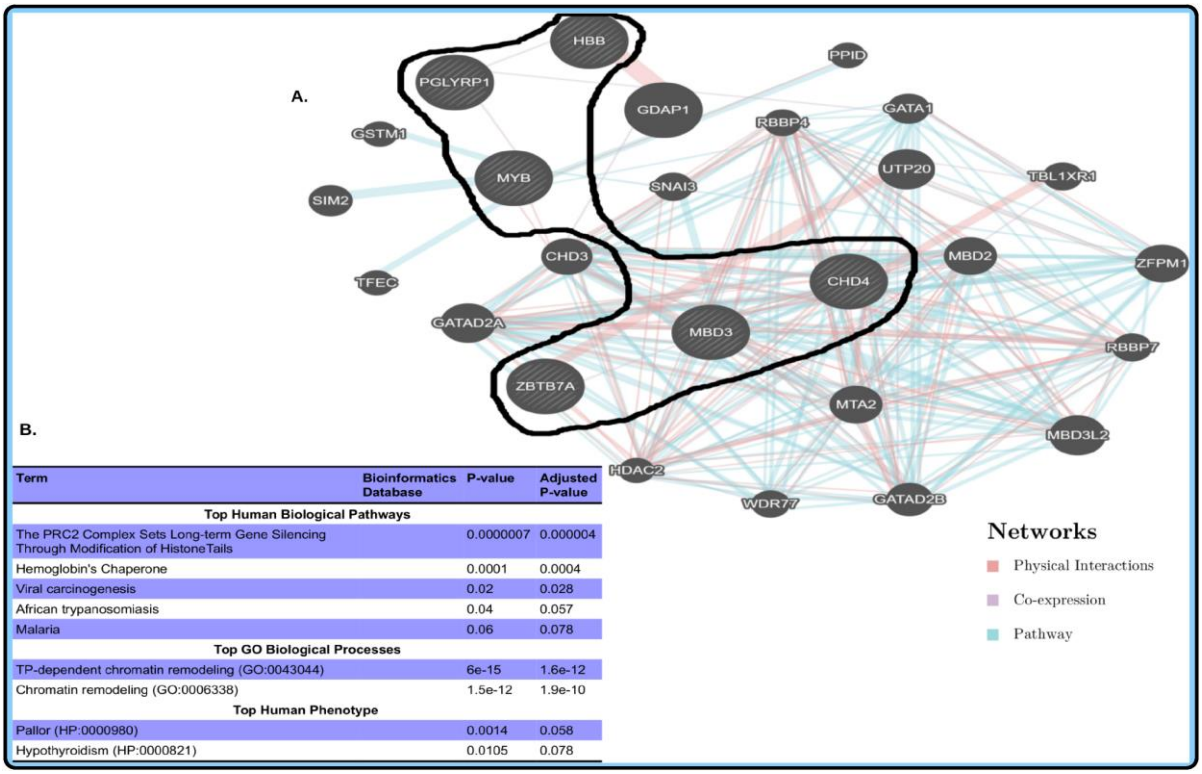
Based on Exome Aggregation Consortium (ExAC) database of pathogenic mutation [56], we found no significant difference in the number of pathogenic variants in both SCD patients with high or low HbF levels in genes (*BCL11A*), proto-oncogene, *transcription factor* (*MYB*), *Homeobox A9* (*HOXA9*), *hemoglobin subunit gamma 2* (*HBG2*), *Kruppel like factor 1* (*KLF1*), *zinc finger and BTB domain containing 7A* (*ZBTB7A*) in chromosomes 2, 6, 7, 11, 12 and 19, respectively. Overall, our targeted next generation sequencing of HbF associated genetic loci identified a disproportional number of loci with a few, variants, particularly deletions, present in patients with high levels of HbF.

Table 3. Genes with high deleterious and loss-of-function mutations in SCD patients from Tanzania. Details of mutation on SNPs below can be found in Supplementary File: **Table S2**.

CHR	Gene	#SNPs (High;low HbF level)	Exonic Function	# SP[1]
chr19	*ZBTB7A*	2; 1	Nonsynonymous	MutationTaster,FATHMM, fathmm-MKL, RadialSVM, LR, PROVEAN, MetaSVM, MetaLR, CADD, GERP++, DANN, M-CAP, Eigen, GenoCanyon, Polyphen2 HVAR, Polyphen2 HDIV, PhyloP and SiPhy
chr12	*CHD4*	11; 4	Nonsynonymous	SIFT, LRT, MutationTaster, MutationAssessor, FATHMM, fathmm-MKL, RadialSVM, LR, PROVEAN, MetaSVM, MetaLR, CADD, GERP++, DANN, M-CAP, GenoCanyon, Polyphen2 HVAR, Polyphen2 HDIV
chr11	*HBB*	3; 2	Nonsynonymous	SIFT, LRT, MutationAssessor, FATHMM, fathmm-MKL, RadialSVM, LR, ROVEAN, MetaSVM, MetaLR, CADD, DANN, Polyphen2 HVAR, Polyphen2 HDIV, PhyloP and

				SiPhy
chr19	*PGLYRP1*	4; 4	Nonsynonymous	SIFT, LRT, MutationAssessor, FATHMM, fathmm-MKL, RadialSVM, LR, PROVEAN, MetaSVM, DANN, M-CAP, GenoCanyon, Polyphen2 HVAR, Polyphen2 HDIV, PhyloP and SiPhy
chr19	*MBD3*	1; 2	Stop-gain	SIFT, LRT, MutationTaster, MutationAssessor, LR, PROVEAN, MetaSVM, MetaLR, CADD, GERP++, DANN, M-CAP, Eigen, GenoCanyon, Polyphen2 HVAR, Polyphen2 HDIV, PhyloP and SiPhy
chr6	*MYB*	1; 1	Nonsynonymous	SIFT, LRT, MutationTaster, MutationAssessor, FATHMM, fathmm-MKL, RadialSVM, LR, PROVEAN, MetaSVM, MetaLR, CADD, GERP++, DANN, M-CAP, GenoCanyon, Polyphen2 HVAR, Polyphen2 HDIV, PhyloP

Abbreviations: # SP[1] is the number of in silico mutation tools predicted and considered damaging.

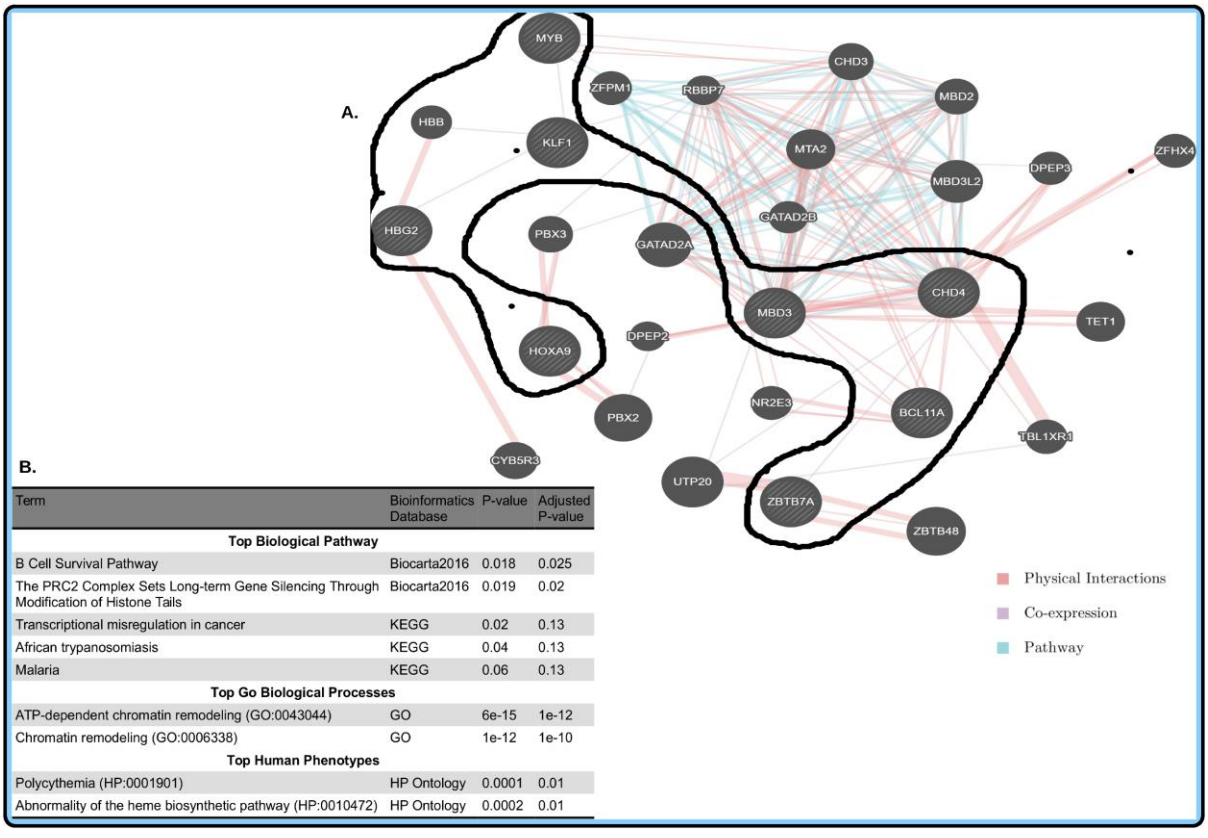


Term	Bioinformatics Database	P-value	Adjusted P-value
Top Human Biological Pathways			
The PRC2 Complex Sets Long-term Gene Silencing Through Modification of Histone Tails		0.0000007	0.000004
Hemoglobin's Chaperone		0.0001	0.0004
Viral carcinogenesis		0.02	0.028
African trypanosomiasis		0.04	0.057
Malaria		0.06	0.078
Top GO Biological Processes			
TP-dependent chromatin remodeling (GO:0043044)		6e-15	1.6e-12
Chromatin remodeling (GO:0006338)		1.5e-12	1.9e-10
Top Human Phenotype			
Pallor (HP:0000980)		0.0014	0.058
Hypothyroidism (HP:0000821)		0.0105	0.078

Figure 3. Biological sub-network of the candidate mutant gene and identified genetics modifier genes in 14 SCD patients from Tanzania. A) sub-networks of the mutant gene and identified candidate genetic modifiers include *CHD4 and MBD3*. B) description of the top most significant pathways, GO biological process, and Human Phenotypes associated with the identified variants.

Biological pathways and processes associated with genes with high mutational burdens

Independent roles of the identified candidate in silico mutant genes (Supplementary File: **Table S1, Figure 2**) or our initial targeted nine genes are known in Sickle Cell disease. However, how these genes interact with others at the systems level is currently unknown, particularly in African population of Sickle cell patients. As described in method section, using the set of all interactive genes including our identified mutant genes and the rest of targeted genes may contribute in identifying potential Sickle Cell-specific pathways in which modifier and mutant genes participate together in conferring variation in Sickle Cell Disease severity. The identified Protein-Protein Interaction (PPI) sub-network formed from 2 genes (**Figure 3A**) showed an enrichment of rare variants with deleterious effects was enriched for the *PRC2 complex* which influence long-term gene silencing through modification of histone tails (P = 0.000004; **Figure 3B**), and is highly associated with or involved in the *TP-dependent chromatin remodeling* (P = 1.6e-12, **Figure 3B**) biological process, nominally associated with *pallor* (P = 0.0014, **Figure 3B**). *CHD4* and *MBD3* were found to be the most important genes (hubs) of sub-network (**Figure 4A**), which are nominally associated with the B cell survival pathway (P = 0.018, **Figure 4B**), known to be implicated in the *ATP-dependent chromatin re-modeling* biological process (P = 6e-15, **Figure 4B**) and associated with *polycythemia disorder* (P = 0.0001, **Figure 4B**).



B.

Term	Bioinformatics Database	P-value	Adjusted P-value
Top Biological Pathway			
B Cell Survival Pathway	Biocarta2016	0.018	0.025
The PRC2 Complex Sets Long-term Gene Silencing Through Modification of Histone Tails	Biocarta2016	0.019	0.02
Transcriptional misregulation in cancer	KEGG	0.02	0.13
African trypanosomiasis	KEGG	0.04	0.13
Malaria	KEGG	0.06	0.13
Top Go Biological Processes			
ATP-dependent chromatin remodeling (GO:0043044)	GO	6e-15	1e-12
Chromatin remodeling (GO:0006338)	GO	1e-12	1e-10
Top Human Phenotypes			
Polycythemia (HP:0001901)	HP Ontology	0.0001	0.01
Abnormality of the heme biosynthetic pathway (HP:0010472)	HP Ontology	0.0002	0.01

Figure 4. Biological sub-network of the candidate mutant gene and identified genetics modifier genes in 14 SCD patients from Tanzania. A) sub-networks of our target sequencing variants include *ZBTB7A, BCL11A, MYB, HBB, HOXA9, HBG2, CHD4, KLF1, MBD3.* B) description of the top most significant pathways, GO biological process, and Human Phenotypes associated with the identified variants.

Discussion

This is the first study in Africa to conduct targeted next generation sequencing to investigate genetic modifiers and pathways associated with extreme fetal hemoglobin (HbF) in individuals with SCD. Most of the loci (SNPs) that have been found to associate with HbF by GWAS only show possible associations with variants covered by the array chip GWAS. The approach taken in this study was to perform in-depth sequencing around previously identified loci to descriptively compare, in detail, discovered polymorphisms between individuals with extreme HbF levels. We have identified single nucleotide polymorphisms (SNPs), insertions (IN) and deletions (DEL) across 8 targeted regions in chromosomes 2, 6, 7, 11, 12 and 19. We found differing types of polymorphisms, including SNPs and INDELS between individuals with low HbF versus those with high HbF, suggesting potential modifier effect. Interestingly, key discovered variants, together with previously identified variants, are enriched in biological pathways that underlie the HbF regulation.

It is worth also noting that possible structural variants in these patient groups may make the sequencing off between the two groups. Furthermore, current challenges include (1) limitation of variant calling tools in African data [57], (2) sequencing errors and structural variants in African data [58] and (3) under-representation of African samples in the current reference genome [58, 59] and may contribute to the observed difference in variants discovered in both high and low HbF level in individuals with SCD. We found more deletions in individuals with high HbF than those with low HbF levels indicating their role in HbF synthesis pathways. A number of significant deletions have been reported before, particularly in the globin cluster [60, 61, 62]. In this study, we have identified additional potential deletions across the targeted regions (**Table 2**). We observed more insertions in individuals with high HbF than in those with low HbF. However, frameshift deletions were found more in individuals with high HbF, while frameshift insertions were more prevalent in individuals with low HbF. Frameshift deletions and insertion may lead to abnormal proteins due to shorter or longer sequences, respectively.

We also looked at variants located at untranslated regions (UTR) both at 3' and 5' ends which are involved differently in regulation of gene expression. Interestingly, there were more variants in the 5'UTR in individuals with high HbF levels, whereas there were more variants in the 3'UTR in individuals with low HbF levels. Molecular mechanisms of the 5'UTR include regulating translation of main coding sequences while the 3'UTR contain binding sites for micro RNA (miRNA) which takes part in the timing and rate of translation of the corresponding mRNA. Hence the difference in variants in these two regions between individuals with high HbF versus those with low HbF is notably and may participate in differential in the regulation of HbF synthesis.

We looked specifically at non-synonymous mutations and found that out of the eight targets, six were found to have mutations with functional impact. Of interest, the genes *CHD4* and *MBD3*, functionally interacting in the same sub-network (see **Figure 3**), had more pathogenic mutations in individuals with low HbF levels than those with high HbF. *CHD4* is a chromatin organization modifier which confers the chromatin remodeling function of the NuRD complex. *CHD4* has been reported to repress *γ-globin* gene expression in mouse [22, 23]. Similarly, *MBD3* operates as a NuRD complex and is associated with the transcription factors GATA-1 and FOG-1 to directly regulate genes within the β-globin locus.

The human protein-protein interaction (PPI) (**Figure 3A**) for CHD4 and MBD3 proteins indicates that they are essential to system survival and hence their biological functions tend to be evolutionary conserved [63]. Thus, in presence of non-synonymous mutations, it is expected that individual components (proteins and interactions) in the system must adapt to a changing environment while maintaining the system's primary function. In this study, we observed that, to maintain its robustness while sustaining its function under fluctuating environmental conditions, the system possibly triggers different mechanisms. This ensures that the network retains the modularity degree in order to provide a selective advantage for the host system by conserving and/or gaining useful functional interactions within the network to ensure an increase of HbF levels. As an illustration, *CHD4*, as well as *MBD3*, indirectly interact with *KLF1* and *MYB*, which are potent activators of *BCL11A*. *CHD4* is believed to exert its gamma globin silencing effect by positively regulating the *BCL11A* and *KLF1* genes. In addition, *BCL11A* and *MYB* are known to be involved in γ-globin gene regulation, leading to either elevation or reduction of HbF levels [64]. The difference in frequency of non-synonymous mutations in the individuals with high HbF levels versus those with low levels reflect different interactions within this network and the resulting levels of HbF.

Given our study design, we did not perform genetics differentiation test or statistical tests of differences in minor allele frequencies or genotype counts. Instead, we have aimed at descriptively characterizing the proportion of variants between the low/high HbF from high confident variants calling, compare the count of pathogenic variants between the groups and identify potential Sickle Cell-specific pathways in which modifier and mutant genes participate in conferring variation in Sickle Cell severity. Importantly, our current study suggests (1) a difference in the overall variation between Sickle Cell patients with high and low HbF level and, (2) biological pathways include the *PRC2 complex* which sets long-term gene silencing through modification of histone tails (P = 0.000004; **Figure 3B**), hemoglobin's Chaperone (P = 0.001, **Figure 4B**) and B-cell Survival (P = 0.018, **Figure 4B**). These identified pathways may harbour potential interactive Sickle Cell-specific genes including modifier, mutant and other genes (**Figures 3-4**) in conferring variation in severity among individuals with SCD. This work has focused on the importance of studying both genetic and epigenetic pathways in HbF regulation. Our findings suggest an in-depth whole genome sequencing study to to fully characterize modifier genes implicated in the variation of SCD severity. This approach may contribute to future development of interventions for SCD, including drugs and gene therapy.

Acknowledgments: The authors thank the patients and staff of Muhimbili National Hospital, Muhimbili University of Health and Allied Sciences, Tanzania and the Sickle Cell Program. This work was supported by Wellcome Trust (Grant no: 095009, 093727, 080025 & 084538) and Fogarty Global Health Fellowship sponsored by the National Institutes of Health (NIH). The authors extend special gratitude to Dr Barnaby Clark (PhD) who was Principal Clinical Scientist at King's College Hospital. Dr Clark shared the next generation sequencing panel that was customized and adopted for this study.

Author Contributions: SN, HK, SM and JBM designed the study, JAM collected data, MZ, LM, RS, GM and EC processed and analysed the data. All authors contributed to the drafts of the manuscript.

Conflict of Interest: The authors declare that they have no conflict of interests.

References

1. Weatherall, D., Akinyanju, O., Fucharoen, S., Olivieri, N., and Musgrove, P. (2006). Chapter 34 Inherited Disorders of Hemoglobin. Disease Control Priorities in Developing Countries, 663–680.
2. World Health Organization (2006). Management of birth defects and haemoglobin disorders: report of a joint WHO-March of Dimes meeting, Geneva, Switzerland, 17-19 may 2006. World Health Organization: Regional office for South-East-Asia, 5–17.
3. Makani, J., Cox, S. E., Soka, D., Komba, A. N., Oruo, J., Mwamtemi, H., et al. (2011a). Mortality in sickle cell anemia in africa: A prospective cohort study in Tanzania. PLoS ONE, 6(2), e14699. doi:10.1371/journal.pone.0014699.
4. Piel, F. B., Patil, A. P., Howes, R. E., Nyangiri, O. A., Gething, P. W., Dewi, M., et al. (2013). Global epidemiology of Sickle haemoglobin in neonates: A contemporary geostatistical model-based map and population estimates. The Lancet, 381, 142–151. doi:10.1016/S0140-6736(12)61229-X.
5. Weatherall, D. J. (2001). Phenotype-genotype relationships in monogenic disease: Lessons from the thalassaemias. Nature Reviews Genetics, 2, 245–255. doi:10.1038/35066048.
6. Dampier, C., Ely, E., Eggleston, B., Brodecki, D., and O'Neal, P. (2004). Physical and cognitive-behavioral activities used in the home management of sickle pain: A daily diary study in children and adolescents. Pediatric Blood and Cancer, 43, 674–678. doi:10.1002/pbc.20162.
7. Platt, O. S., Thorington, B. D., Brambilla, D. J., Milner, P. F., Rosse, W. F., Vichinsky, E., et al. (1991). Pain in sickle cell disease. Rates and risk factors. N. Engl. J. Med., 325, 11–16. doi:10.1056/NEJM199107043250103.
8. Manning, L. R., Russell, J. E., Padovan, J. C., Chait, B. T., Popowicz, A., Manning, R. S., et al. (2007). Human embryonic, fetal, and adult hemoglobins have different subunit interface strengths. Correlation with lifespan in the red cell. Protein Sci, 16, 1641–1658. doi:10.1110/ps.072891007.
9. Thein, S. L., and Menzel, S. (2009). Discovering the genetics underlying foetal haemoglobin production in adults. Br J Haematol., 145(4), 455-467. doi:10.1111/j.1365-2141.2009.07650.x.
10. Mosca, A., Paleari, R., Ivaldi, G., Galanello, R., and Giordano, P. C. (2009). The role of haemoglobin A2 testing in the diagnosis of thalassaemias and related haemoglobinopathies. Journal of Clinical Pathology, 62, 13–17. doi:10.1136/jcp.2008.056945.
11. Menzel, S., Garner, C., Gut, I., Matsuda, F., Yamaguchi, M., Heath, S., et al. (2007). A QTL influencing F cell production maps to a gene encoding a zinc-finger protein on chromosome 2p15. Nature Genetics, 39, 1197–1199. doi:10.1038/ng2108.
12. Garner, C., Tatu, T., Reittie, J. E., Littlewood, T., Darley, J., Cervino, S., et al. (2000). Genetic influences on F cells and other hematologic variables: a twin heritability

study. Blood, 95, 342–346.
13. Menzel, S., and Lay, S. (2009). Genetic architecture of hemoglobin F control. Curr Opin Hematol., 16(3), 179–186. doi:10.1097/MOH.0b013e328329d07a.
14. Thein, S. L., Menzel, S., Peng, X., Best, S., Jiang, J., Close, J., et al. (2007). Intergenic variants of HBS1L-MYB are responsible for a major quantitative trait locus on chromosome 6q23 influencing fetal hemoglobin levels in adults. Proceedings of the National Academy of Sciences, 104, 11346–11351. doi:10.1073/pnas.0611393104.
15. Thein, S. L., Menzel, S., Lathrop, M., and Garner, C. (2009). Control of fetal hemoglobin: New insights emerging from genomics and clinical implications. Human Molecular Genetics, 18, 216–223. doi:10.1093/hmg/ddp401.
16. Makani, J., Menzel, S., Nkya, S., Cox, S. E., Drasar, E., Soka, D., et al. (2011b). Genetics of fetal hemoglobin in Tanzanian and British patients with sickle cell anemia. Blood, 117, 1390–1392. doi:10.1182/blood-2010-08-302703.
17. Mtatiro, S. N., Singh, T., Rooks, H., Mgaya, J., Mariki, H., Soka, D., et al. (2014). Genome wide association study of fetal hemoglobin in sickle cell Anemia in Tanzania. PLoS ONE, 9(11), e111464. doi:10.1371/journal.pone.0111464.
18. Zhou, D., Liu, K., Sun, C. W., Pawlik, K. M., and Townes, T. M. (2010). KLF1 regulates BCL11A expression and gamma- to beta-globin gene switching. Nature Genetics, 42(9), 742–744. doi:10.1038/ng.637.
19. Siatecka, M., Bieker, J. J., and Dc, W. (2011). The multifunctional role of EKLF / KLF1 during erythropoiesis. Blood, 118, 2044–2054. doi:10.1182/blood-2011-03-331371.
20. Sankaran, V. G., and Orkin, S. H. (2013). The switch from fetal to adult hemoglobin. Cold Spring Harb Perspect Med., 3(1), a011643. doi:10.1101/cshperspect.a011643.
21. Thein, S. L. (2013). Genetic association studies in β-hemoglobinopathies. ASH Education Program Book 2013, 354–361. doi:10.1182/asheducation-2013.1.354.
22. Amaya, M., Desai, M., Gnanapragasam, M. N., Wang, S. Z., Zu Zhu, S., Williams, D. C., et al. (2013). Mi2β-mediated silencing of the fetal γ-globin gene in adult erythroid cells. Blood, 121, 3493–3501. doi:10.1182/blood-2012-11-466227.
23. Torrado, M., Low, J. K. K., Silva, A. P. G., Schmidberger, J. W., Sana, M., Tabar, M. S., et al. (2017). Refinement of the subunit interaction network within the nucleosome remodelling and deacetylase (NuRD) complex. The FEBS Journal, 284, 4216–4232. doi:10.1111/febs.14301.
24. Mtatiro, S. N., Mgaya, J., Singh, T., Mariki, H., Rooks, H., Soka, D., et al. (2015). Genetic association of fetal-hemoglobin levels in individuals with sickle cell disease in Tanzania maps to conserved regulatory elements within the MYB core enhancer. BMC Medical Genetics 16:4. doi:10.1186/s12881-015-0148-3.
25. Li, H., Ruan, J., and Durbin, R. (2008). Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome Res., 18, 1851–1858. doi:10.1101/gr.078212.108.
26. McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., et al. (2010). The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res., 20, 1297–1303. doi:10.1101/gr.107524.110.
27. Gézsi, A., Bolgár, B., Marx, P., Sarkozy, P., Szalai, C., and Antal, P. (2015). VariantMetaCaller: automated fusion of variant calling pipelines for quantitative, precision-based filtering. BMC Genomics, 16, 875. doi:10.1186/s12864-015-2050-y.

28. 1000 Genomes Project Consortium, Abecasis, G. R., Auton, A., Brooks, L. D., DePristo, M. A., Durbin, R. M., Handsaker, R. E., Kang, H. M., Marth, G. T., McVean, G. A. (2012). An integrated map of genetic variation from 1,092 human genomes. Nature, 491(7422), 56-65. doi: 10.1038/nature11632.
29. Cornish, A., and Guda, C. (2015). A Comparison of Variant Calling Pipelines Using Genome in a Bottle as a Reference. Biomed Res. Int., 2015, 1–11. doi:10.1155/2015/456479.
30. Quinlan, A. R., and Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics, 26, 841–842. doi:10.1093/bioinformatics/btq033.
31. Li, H. (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics, 27, 2987–2993. doi:10.1093/bioinformatics/btr509.
32. Wang, K., Li, M., and Hakonarson, H. (2010). ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res., 38, e164. doi:10.1093/nar/gkq603.
33. Ng, P. C., and Henikoff, S. (2003). SIFT: Predicting amino acid changes that affect protein function. Nucleic Acids Res., 31, 3812–3814
34. Ng, P. C., and Henikoff, S. (2006). Predicting the Effects of Amino Acid Substitutions on Protein Function. Annu. Rev. Genomics Hum. Genet., 7, 61–80. doi:10.1146/annurev.genom.7.080505.115630.
35. Fujita, A., Kojima, K., Patriota, A. G., Sato, J. R., Severino, P., and Miyano, S. (2010). A fast and robust statistical test based on likelihood ratio with Bartlett correction to identify Granger causality between gene sets. Bioinformatics, 26, 2349–2351. doi:10.1093/bioinformatics/btq427.
36. Reva, B., Antipin, Y., and Sander, C. (2007). Determinants of protein function revealed by combinatorial entropy optimization. Genome Biol., 8, R232. doi:10.1186/gb-2007-8-11-r232.
37. Reva, B., Antipin, Y., and Sander, C. (2011). Predicting the functional impact of protein mutations: application to cancer genomics. Nucleic Acids Res., 39, e118–e118. doi:10.1093/nar/gkr407.
38. Shihab, H. A., Gough, J., Mort, M., Cooper, D. N., Day, I. N., and Gaunt, T. R. (2014). Ranking non-synonymous single nucleotide polymorphisms based on disease concepts. Hum. Genomics, 8, 11. doi:10.1186/1479-7364-8-11.
39. Dong, C., Wei, P., Jian, X., Gibbs, R., Boerwinkle, E., Wang, K., et al. (2015). Comparison and integration of deleteriousness prediction methods for nonsynonymous SNVs in whole exome sequencing studies. Hum. Mol. Genet., 24, 2125–2137. doi:10.1093/hmg/ddu733.
40. Kircher, M., Witten, D. M., Jain, P., O'Roak, B. J., Cooper, G. M., and Shendure, J. (2014). A general framework for estimating the relative pathogenicity of human genetic variants. Nat. Genet., 46, 310–315. doi:10.1038/ng.2892.
41. Cooper, G. M., Stone, E. A., Asimenos, G., NISC Comparative Sequencing Program, Green, E. D., Batzoglou, S., et al. (2005). Distribution and intensity of constraint in mammalian genomic sequence. Genome Res., 15, 901–913. doi:10.1101/gr.3577405.
42. Adzhubei, I. A., Schmidt, S., Peshkin, L., Ramensky, V. E., Gerasimova, A., Bork, P., et al. (2010). A method and server for predicting damaging missense mutations. Nat.

Methods, 7, 248–249. doi:10.1038/nmeth0410-248.
43. Garber, M., Guttman, M., Clamp, M., Zody, M. C., Friedman, N., and Xie, X. (2009). Identifying novel constrained elements by exploiting biased substitution patterns. Bioinformatics, 25, i54-62. doi:10.1093/bioinformatics/btp190.
44. Li, M.-X., Gui, H.-S., Kwan, J. S. H., Bao, S.-Y., and Sham, P. C. (2012). A comprehensive framework for prioritizing variants in exome sequencing studies of Mendelian diseases. Nucleic Acids Res., 40, e53–e53. doi:10.1093/nar/gkr1257.
45. Li, M.-X., Kwan, J. S. H., Bao, S.-Y., Yang, W., Ho, S.-L., Song, Y.-Q., et al. (2013). Predicting Mendelian Disease-Causing Non-Synonymous Single Nucleotide Variants in Exome Sequencing Studies. PLoS Genet., 9, e1003143. doi:10.1371/journal.pgen.1003143.
46. Sayers, E. W., Agarwala, R., Bolton, E. E., Brister, J. R., Canese, K., Clark, K., Connor, R., Fiorini, N., Funk, K., Hefferon, T., et al. (2019). Database resources of the National Center for Biotechnology Information. Nucleic Acids Res., 47(D1), D23-D28.
47. Houdayer, C., Dehainault, C., Mattler, C., Michaux, D., Caux-Moncoutier, V., Pagès-Berhouet, S., et al. (2008). Evaluation of in silico splice tools for decision-making in molecular diagnosis. Human mutation, 29, 975-982. doi: 10.1002/humu.20765.
48. Chimusa, E. R., Mbiyavanga, M., Mazandu, G. K., and Mulder, N. J. (2016). ancGWAS: a post genome-wide association study method for interaction, pathway and ancestry analysis in homogeneous and admixed populations. Bioinformatics, 32, 549–556. doi:10.1093/bioinformatics/btv619.
49. Wu, J., Vallenius, T., Ovaska, K., Westermarck, J., Mäkelä, T. P., and Hautaniemi, S. (2009). Integrated network analysis platform for protein-protein interactions. Nat. Methods, 6, 75–77. doi:10.1038/nmeth.1282.
50. Kuleshov, M. V., Jones, M. R., Rouillard, A. D., Fernandez, N. F., Duan, Q., Wang, Z., et al. (2016). Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. Nucleic Acids Res., 44, W90–W97. doi:10.1093/nar/gkw377.
51. Kanehisa, M., Sato, Y., Furumichi, M., Morishima, K., and Tanabe, M. (2019). New approach for understanding genome variations in KEGG. Nucleic Acids Res., 47, D590-D595. doi: 10.1093/nar/gky962.
52. Mi, H., Muruganujan, A., Ebert, D., Huang, X., Thomas, P. D. (2019). PANTHER version 14: more genomes, a new PANTHER GO-slim and improvements in enrichment analysis tools. Nucleic Acids Res., 47(D1), D419–D426. doi: 10.1093/nar/gky1038
53. Nishimura, D. (2001). BioCarta. Biotech Softw. Internet Rep., 2, 117–120. doi: 10.1089/152791601750294344.
54. Fabregat, A., Jupe, S., Matthews, L., Sidiropoulos, K., Gillespie, M., Garapati, P., Haw, R., Jassal, B., Korninger, F., May, B., et al. (2018). The Reactome Pathway Knowledgebase. Nucleic Acids Res., 46(D1), D649-D655. doi: 10.1093/nar/gkx1132.
55. The Gene Ontology Consortium. (2019). The Gene Ontology Resource: 20 years and still GOing strong. Nucleic Acids Res., 47(D1), D330-D338. doi: 10.1093/nar/gky1055.
56. Karczewski, K. J., Weisburd, B., Thomas, B., Solomonson, M., Ruderfer, D. M., Kavanagh, D., et al. (2017). The ExAC browser: displaying reference data information from over 60 000 exomes. Nucleic Acids Res., 45, D840–D845. doi:10.1093/nar/gkw971.
57. Pirooznia, M., Kramer, M., Parla, J., Goes, F.S., Potash, J.B., McCombie, W.R. and

Zandi, P.P., 2014. Validation and assessment of variant calling pipelines for next-generation sequencing. Human genomics 8(1):14.

58. Teo, Y.Y., Small, K.S. and Kwiatkowski, D.P., 2010. Methodological challenges of genome-wide association analysis in Africa. Nature Reviews Genetics, 11(2), 149.
59. O'Leary, N. A., Wright, M. W., Brister, J. R., Ciufo, S., Haddad, D., McVeigh, R., et al. (2016). Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res., 44, D733-745. doi:10.1093/nar/gkv1189.
60. Chalaow, N., Thein, S. L., and Viprakasit, V. (2013). The 12.6 kb-deletion in the β-globin gene cluster is the known Thai/Vietnamese (δβ)0-thalassemia commonly found in Southeast Asia. Haematologica, 98, e117–e118. doi:10.3324/haematol.2013.090613.
61. Hamid, M., Nejad, L.D., Shariati, G., et al. (2017). The First Report of a 290-bp Deletion in β-Globin Gene in the South of Iran. Iranian Biomedical Journal, 21(2), 126–128. doi:10.18869/acadpub.ibj.21.2.126.
62. Thein, S. L., and Craig, J. E. (1998). Genetics of Hb F/F cell variance in adults and heterocellular hereditary persistence of fetal hemoglobin. Hemoglobin, 22, 401–414.
63. Akinola, R. O., Mazandu, G. K., and Mulder, N. J. (2016). A Quantitative Approach to Analyzing Genome Reductive Evolution Using Protein–Protein Interaction Networks: A Case Study of Mycobacterium leprae. Front Genet, 7, 39. doi:10.3389/fgene.2016.00039.
64. Jiang, J., Best, S., Menzel, S., Silver, N., Lai, M. I., Surdulescu, G. L., et al. (2006). cMYB is involved in the regulation of fetal hemoglobin production in adults. Blood, 108, 1077–1083. doi:10.1182/blood-2006-01-008912.