


RESEARCH ARTICLE

Open Access



# Identifying genetic variants and pathways associated with extreme levels of fetal hemoglobin in sickle cell disease in Tanzania

Siana Nkya<sup>1,2\*</sup>, Liberata Mwita<sup>2</sup>, Josephine Mgaya<sup>2</sup>, Happiness Kumburu<sup>3</sup>, Marco van Zwetselaar<sup>3</sup>, Stephan Menzel<sup>4</sup>, Gaston Kuzamunu Mazandu<sup>5,6,7\*</sup> , Raphael Sangeda<sup>2,8</sup>, Emile Chimusa<sup>5</sup> and Julie Makani<sup>2</sup>

## Abstract

**Background:** 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 genetic variants which differ across populations and hence do not fully account for HbF variation.

**Methods:** We investigated rare and common genetic variants that influence HbF levels in 14 SCD patients to elucidate variants and pathways in SCD patients with extreme HbF levels ( $\geq 7.7\%$  for high HbF) and ( $\leq 2.5\%$  for low HbF) in Tanzania. We performed targeted next generation sequencing (Illumina\_Miseq) covering exonic and other significant fetal hemoglobin-associated loci, including *BCL11A*, *MYB*, *HOXA9*, *HBB*, *HBG1*, *HBG2*, *CHD4*, *KLF1*, *MBD3*, *ZBTB7A* and *PGLYRP1*.

**Results:** Results revealed a range of genetic variants, including bi-allelic and multi-allelic SNPs, frameshift insertions and deletions, some of which have functional importance. Notably, there were significantly more deletions in individuals with high HbF levels (11% vs 0.9%). We identified frameshift deletions 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.

**Conclusions:** This study provides new insights in selecting essential variants and identifying potential biological pathways associated with extreme HbF levels in SCD interrogating multiple genomic variants associated with HbF in SCD.

**Keywords:** Sickle cell disease, Genetic disorder, Fetal hemoglobin, Hemoglobinopathy, Tanzania

\* Correspondence: [snkyamtiro@gmail.com](mailto:snkyamtiro@gmail.com); [gmazandu@gmail.com](mailto:gmazandu@gmail.com)

<sup>1</sup>Department of Biological Sciences, Dar es Salaam University College of Education, Dar es Salaam, Tanzania

<sup>5</sup>Department of Pathology, Division of Human Genetics, University of Cape Town, IDM, Cape Town, South Africa

Full list of author information is available at the end of the article



© The Author(s). 2020 **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

## Background

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 8000–11,000 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  $\beta$ -globin gene or hemoglobin subunit beta (*HBB*), on chromosome 11, leading to the production of an abnormal  $\beta$ -hemoglobin chain namely hemoglobin S (HbS). SCD is a complex hemoglobin disorder with multiple phenotypic expressions that manifest as both chronic and acute complications, affecting multiple organs. 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 tetrameric molecule composed of 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 namely hemoglobin A (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<sub>2</sub> (HbA<sub>2</sub>), 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  $\beta$ -globin gene. Three main loci, namely *BCL11A* on chromosome 2, *HMIP* on chromosome 6, and *HBB* on chromosome 11, have been identified across populations as associated with HbF levels [13–15]. 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. *Kruppel like factor (KLF1)* has been reported as one of the key regulators of HbF expression with dual functions: direct activation of HbF expression through activation of  $\beta$ -globin [18] and an indirect silencing of  $\gamma$ -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 single nucleotide polymorphisms (SNPs), structural variants such as insertions and deletions (INDELS), suggesting potential modifier effects. Interestingly, key discovered variants, together with previously identified variants, are enriched in biological pathways that underlie the HbF regulation.

## Methods

### Study design and population

We performed a cross-sectional study involving the Dar-es-Salaam (Tanzania) Muhimbili National Hospital SCD cohort, which consisted of 1725 SCD patients, recruited between 2004 and 2009, for prospective surveillance, with three monthly interval visits for routine check-up [3]. These patients were subjected to folic acid (5 mg/day) and penicillin. Different hematological factors, including complete blood counts and foetal haemoglobin (HbF) quantifications, were measured during hospital visits. Written informed consent was obtained for each adult patient (> 16 years) 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). Informed and written consent was obtained from parents or guardians for all minor patients ( $\leq 16$  years). The study involved 14 individuals confirmed to have SCD (HbSS or S- $\beta^0$ thalassaemia), over 5 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 while 7.7% and above was considered a high HbF level.

### 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 (Table 2). Selection of target regions was based on previously 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.

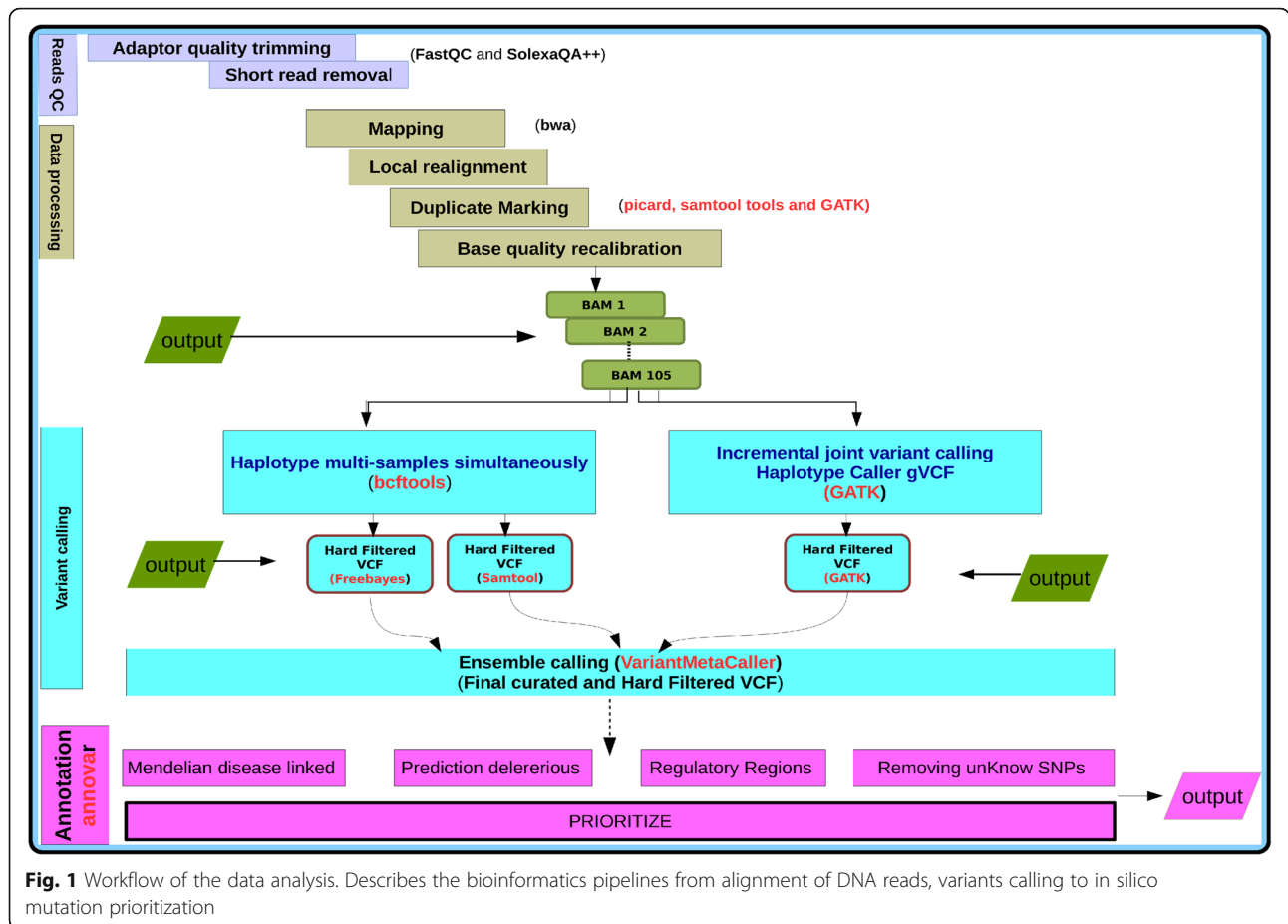
### Reads mapping, alignment, variant calling and variant calling quality control

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 (Fig. 1). The final call-set from each subject group was produced from VariantMetaCaller [27].

### 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) [44–46] 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 has enabled the identification of potential

biological pathways in which these genes participate. To achieve this, we first mapped the identified mutant SNPs to their closest genes. We mapped genes to a comprehensive human 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 processes, from the Gene Ontology database [55].

**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

**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.

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

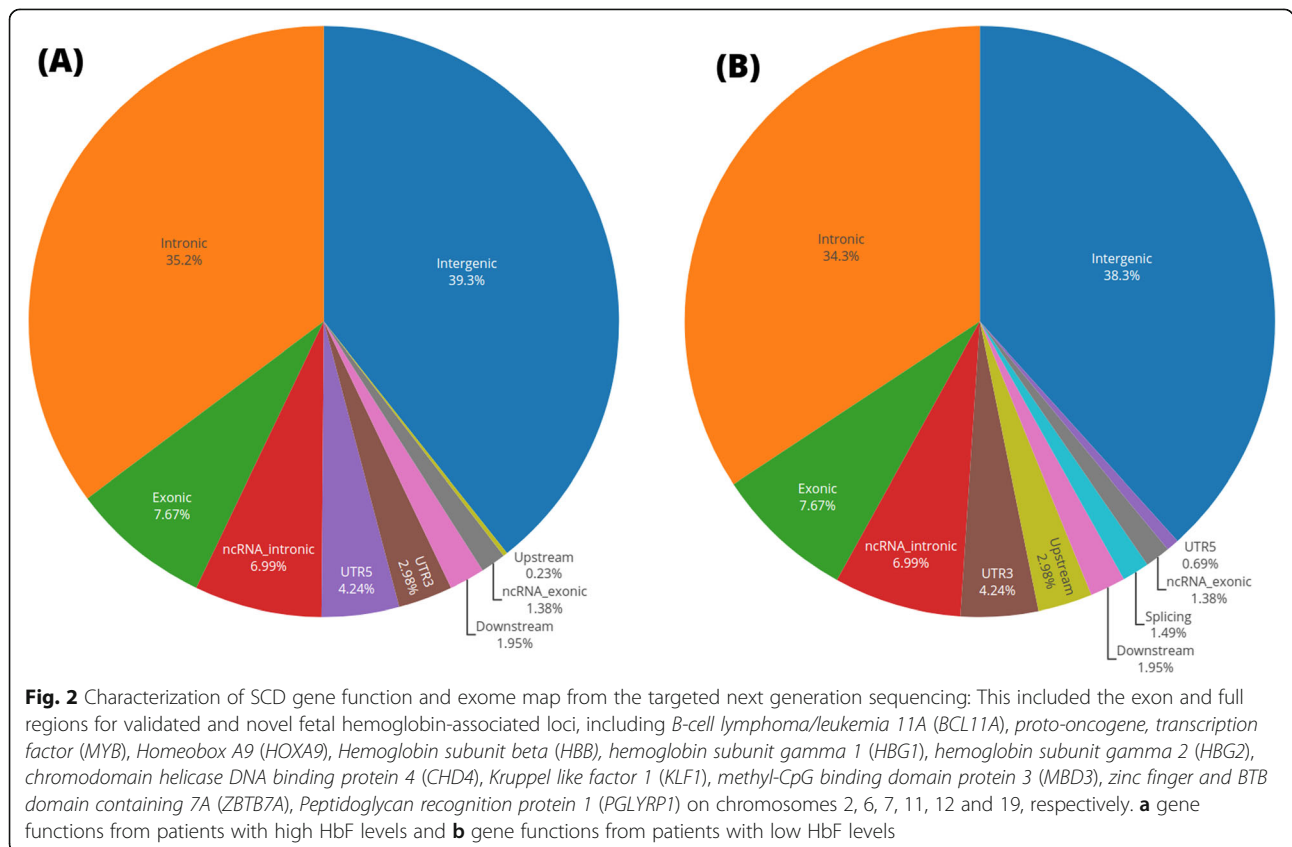
A total of 873 and 1196 highly confident variants were determined in SCD patients with high and low HbF levels, respectively, on chromosomes 2, 6, 7, 11, 12 and 19. Surprisingly, this shows a difference in the overall variation between the two groups of individuals with SCD.

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 (Fig. 2) between SCD

patients with high and low HbF level (adjusted Fisher exact  $p$ -value = 6.1e-04), at 3'untranslated region (3'UTR) (2.98% versus 4.24%), 5' untranslated region (5'UTR) (4.24% versus 0.69%), upstream (0.23, 2.98%). Critically, we observed that patients with high HbF have 0% variants in splicing regions, while patients with low HbF level have 1.49% (Fig. 2).

**Potential pathogenic variants**

Because 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*, *CHD4*, *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.





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

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

Abbreviation: USig\* is the number variant with uncertain significance of pathogenicity

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 [25], 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.

#### 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, Fig. 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 in various populations of African SCD patients. As described in the **Methods** 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

**Table 3** Genes with high deleterious and loss-of-function mutations in SCA 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 <sup>1</sup>
chr19	<i>ZBTB7A</i>	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	<i>CHD4</i>	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	<i>HBB</i>	3; 2	Nonsynonymous	SIFT, LRT, MutationAssessor, FATHMM, fathmm-MKL, RadialSVM, LR, ROVEAN, MetaSVM, MetaLR, CADD, DANN, Polyphen2 HVAR, Polyphen2 HDIV, PhyloP and SiPhy
chr19	<i>PGLYRP1</i>	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	<i>MBD3</i>	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	<i>MYB</i>	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

Abbreviation: # SP<sup>1</sup> is the number of in silico mutation tools predicted and considered damaging

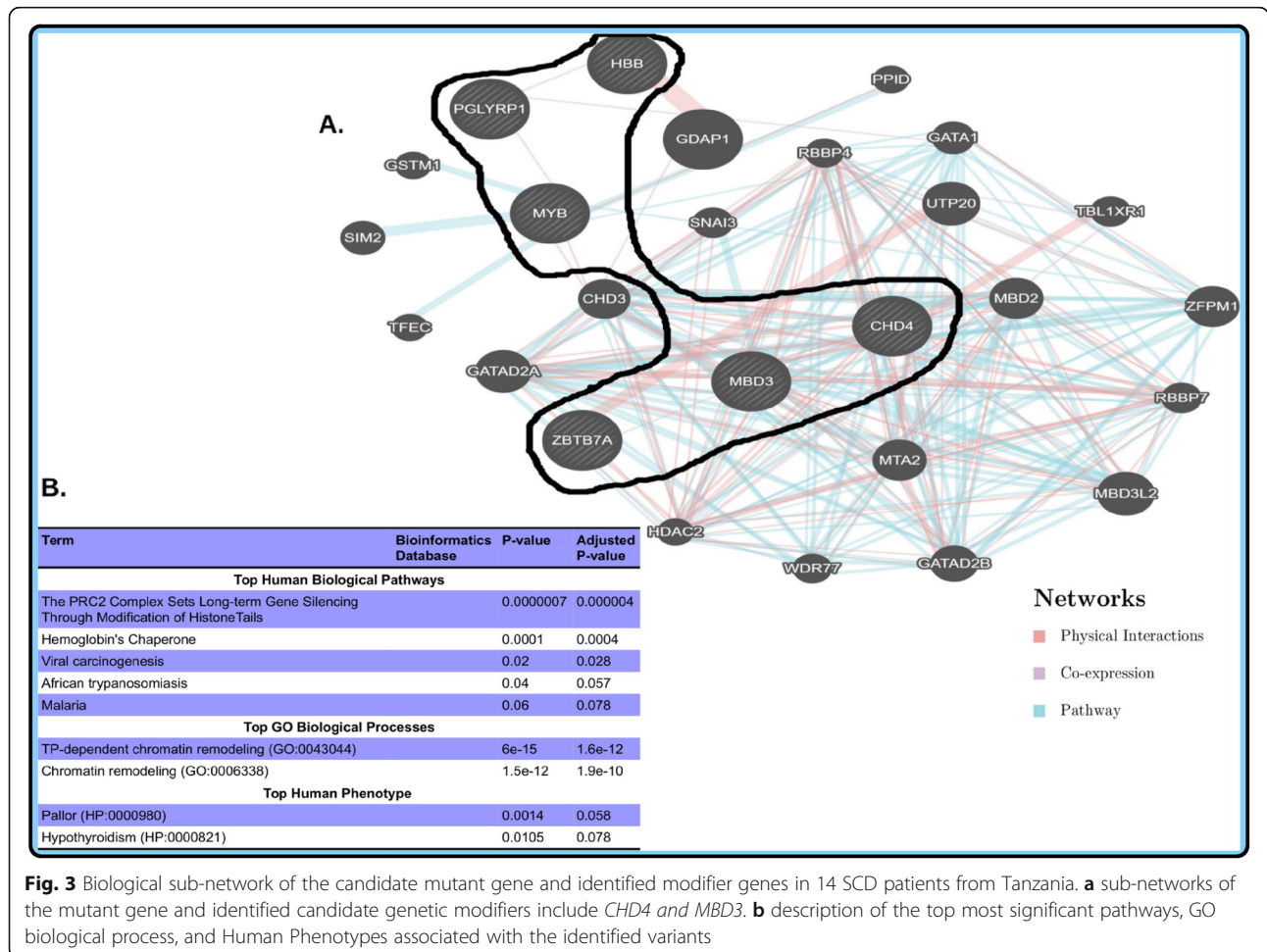
pathways in which modifier and mutant genes participate together in conferring variation in Sickle Cell Disease severity. The identified Protein-Protein Interaction (PPI) sub-network formed from 2 genes (Fig. 3a) showed an enrichment of rare variants with deleterious effects was enriched for the *PRC2 complex* which influence long-term gene silencing through modification of histone tails ( $P = 0.000004$ ; Fig. 3b), and is highly associated with or involved in the *TP-dependent chromatin remodeling* ( $P = 1.6e-12$ , Fig. 3b) biological process, nominally associated with *pallor* ( $P = 0.0014$ , Fig. 3b). *CHD4* and *MBD3* were found to be the most important genes (hubs) of sub-network (Fig. 4a), which are nominally associated with the B cell survival pathway ( $P = 0.018$ , Fig. 4b), known to be implicated in the *ATP-dependent chromatin re-modeling* biological process ( $P = 6e-15$ , Fig. 4b) and associated with *polycythemia disorder* ( $P = 0.0001$ , Fig. 4b).

**Discussion**

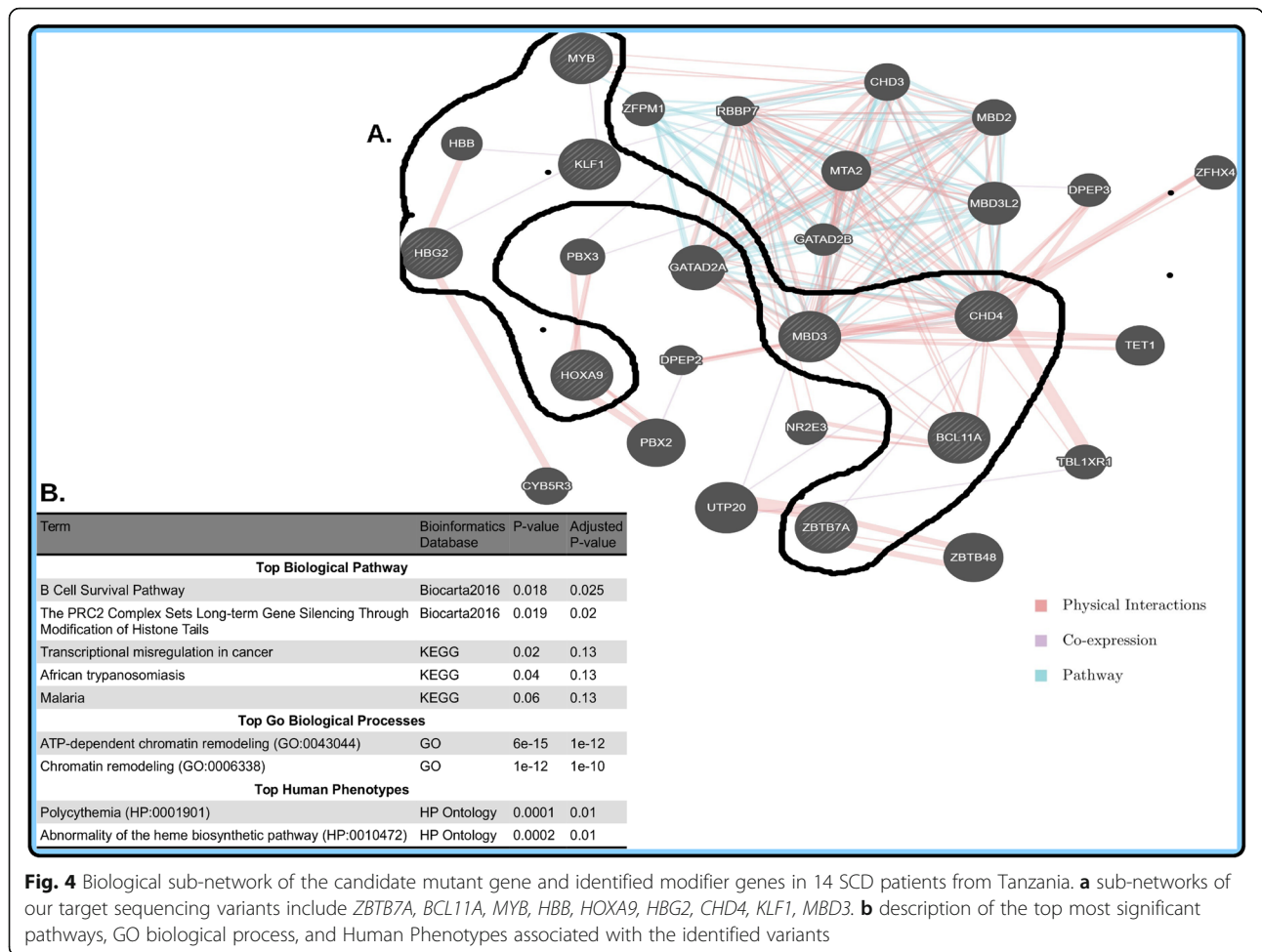
This is the first study in Africa to conduct targeted next generation sequencing to investigate genetic modifiers and pathways associated with extreme fetal hemoglobin

(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 used. 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, including (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



**Fig. 3** Biological sub-network of the candidate mutant gene and identified 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



**Fig. 4** Biological sub-network of the candidate mutant gene and identified 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

African samples in the current reference genome [58, 59], 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 [60–62]. 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 more prevalent 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, in individuals with high HbF levels, variants in the 5'UTR were more prevalent as opposed to 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 microRNA (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 notable and may contribute differently 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 Fig. 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 mice [63, 64]. Similarly, *MBD3* operates as a NuRD complex and is associated with the transcription factors GATA-1 and FOG-1, which directly regulate genes within the  $\beta$ -globin locus.



The human protein-protein interaction (PPI) (Fig. 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  $\gamma$ -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.

Though these post-analysis results are consistent with the literature and are biologically relevant, it is worth noting that, due to relatively high noise related to high-throughput data or experiments from which interactions are inferred, the protein-protein interaction network used may contain incorrectly classified interactions, i.e., failing to detect interactions (false negatives) or wrongly identifying some other interactions (false positives). This suggests these results still need to be validated experimentally. In this study, we minimized the likelihood of incorrectly classified interaction computationally by: (1) using a data integration model, combining information from multiple interacting data sources into one unified network, and (2) applying a strict interaction reliability or confidence score cutoff. These techniques are expected to significantly reduce the false negative and positive rate of the network produced, leading to a PPI network of high confidence interactions with an increased coverage [65].

Given our study design, we did not perform genetics differentiation tests 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 genetic variation between Sickle Cell patients with high and low HbF level and, (2) biological pathways, including the *PRC2 complex* which sets long-term gene silencing through modification of histone tails ( $P = 0.000004$ ; Fig. 3b), hemoglobin's Chaperone ( $P = 0.001$ , Fig. 4b) and B-cell Survival ( $P = 0.018$ , Fig. 4b). These identified pathways may harbour potential interactive Sickle Cell-specific genes including modifier, mutant and other genes (Figs. 3 and 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 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. Finally, to note is, the modest sample size limited the expected statistical power, which could yield false positive associations and missed others. However, different results obtained provide a strong hypothesis for future studies. With a larger sample size, it would be possible to perform genetics differentiation tests or statistical tests of differences in minor allele frequencies or genotype counts and possibly identify additional essential variants and biological pathways associated with extreme HbF levels in SCD using the model set by this study.

## Conclusions

This study has shown that the analysis of genetic modifiers associated with HbF in SCD patients can elucidate genetic factors underlying extreme (low or high) HbF levels in these patients. The study has identified frameshift deletion in SCD patients with high HbF levels and frameshift insertions in both *CHD4* and *MBD3* for those with low HbF, and some of these insertions are associated with the SCD pathogenesis.

## Supplementary information

**Supplementary information** accompanies this paper at <https://doi.org/10.1186/s12881-020-01059-1>.

**Additional file 1: Table S1.** Summary of chromosomal positions and sequenced regions of the targeted genes.

**Additional file 2: Table S2.** Details of gene variants identified in SCD patients from Tanzania.

## Abbreviations

ANNOVAR: ANNOtate VARIation; BCF: Binary variant call format; BCL11A: B-cell lymphoma/leukemia 11A; BWA: Burrows-Wheeler Alignment; CHD4: Chromodomain helicase DNA binding protein 4; DNMT1: DNA methyltransferase 1; ExAC: Exome Aggregation Consortium; FATHMM: Functional Analysis through Hidden Markov Models; FOG1: Friend of GATA1; GATA1: GATA Binding Protein 1; GVCf: Genomic variant call format; GWAS: Genome Wide Association Study; HbA: Hemoglobin A; HbAs: Hemoglobin AS; HBB: Hemoglobin Subunit beta; HbF: Fetal

hemoglobin; HBG2: Hemoglobin subunit gamma 2; HbS: Hemoglobin S; HOXA9: Homeobox A9; INDELS: Insertions and deletions; KCR1: Kilimanjaro Clinical Research Institute; KLF1: Kruppel like factor 1; LSD1: Lysine-specific demethylase 1; MBD3: Methyl-CpG binding domain protein 3; MYB: Myeloblastosis transcription factor; NuRD: Nucleosome remodeling and deacetylase; PPI: Protein-protein interaction; PRC2: Polycomb repressive complex 2; SCD: Sickle cell disease; SNP: Single nucleotide polymorphisms; TR2: Testicular nuclear receptor 2; TR4: Testicular nuclear receptor 4; UTR: Untranslated region; ZBTB7A: Zinc finger and BTB domain containing 7A

#### 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.

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.

#### Authors' contributions

SN, HK, SM and JBM designed the study, JAM, MZ, LM, RS, GKM and EC collected, processed and analyzed the data. All authors contributed to the drafts of the manuscript, GKM and SN finalized the manuscript. The author(s) read and approved the final manuscript.

#### Funding

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 funders had no role in study design, data collection, analysis and interpretation, and decision to publish or preparation of the manuscript. The content of this manuscript is solely the responsibility of the authors and does not necessarily represent the official views of the funders.

#### Availability of data and materials

Additional supporting information can be found in the supplementary file at the end of the article. The datasets used are available at the European Genome-phenome Archive (EGA), accession number EGAS00001000990, and accessible via the following link: <https://www.ebi.ac.uk/ega/studies/EGAS00001000990>

#### Ethics approval and consent to participate

This study was performed in accordance with the Declaration of Helsinki and with the approval of the Muhimbili University Research and Publications Committee (MU/RP/AEC/VOLX1/33 and 2017-03-06/AEC/Vol X11/65). Informed and written consent was obtained from all patients that were all adult participants (> 16 years). Informed and written consent was requested from parents or guardians for all minor participants (≤16 years).

#### Consent for publication

Not applicable.

#### Competing interests

The authors declare that they have no conflict of interests.

#### Author details

<sup>1</sup>Department of Biological Sciences, Dar es Salaam University College of Education, Dar es Salaam, Tanzania. <sup>2</sup>Sickle Cell Program, Department of Hematology and Blood Transfusion, Muhimbili University of Health and Allied Sciences, Dar es Salaam, Tanzania. <sup>3</sup>Department of Biotechnology Laboratory, Kilimanjaro Clinical Research Institute, Kilimanjaro, Tanzania. <sup>4</sup>Department of Molecular Hematology, King's College of London, London, UK. <sup>5</sup>Department of Pathology, Division of Human Genetics, University of Cape Town, ID, Cape Town, South Africa. <sup>6</sup>Department of Integrative Biomedical Sciences, Computational Biology Division, University of Cape Town, Observatory 7925, South Africa. <sup>7</sup>African Institute for Mathematical Sciences, Muizenberg, Cape Town 7945, South Africa. <sup>8</sup>Department of Pharmaceutical Microbiology, Muhimbili University of Health and Allied Sciences, Dar es Salaam, Tanzania.

Received: 27 November 2019 Accepted: 24 May 2020

Published online: 05 June 2020

#### References

- Weatherall D, Akinyanju O, Fucharoen S, Olivieri N, Musgrove P. Chapter 34 inherited disorders of hemoglobin. In: Disease control priorities in developing countries; 2006. p. 663–80.
- Joint WHO-March of Dimes Meeting on Management of Birth Defects and Haemoglobin Disorders (2nd: 2006: Geneva, Switzerland), World Health Organization & March of Dimes. Management of birth defects and haemoglobin disorders: report of a joint WHO-March of Dimes meeting. Geneva: World Health Organization; 2006. <https://apps.who.int/iris/handle/10665/43587>.
- Makani J, Cox SE, Soka D, Komba AN, Oruo J, Mwamtemi H, et al. Mortality in sickle cell anemia in Africa: a prospective cohort study in Tanzania. *PLoS One*. 2011;6(2):e14699. <https://doi.org/10.1371/journal.pone.0014699>.
- Piel FB, Patil AP, Howes RE, Nyangiri OA, Gething PW, Dewi M, et al. Global epidemiology of sickle haemoglobin in neonates: a contemporary geostatistical model-based map and population estimates. *Lancet*. 2013;381:142–51. [https://doi.org/10.1016/S0140-6736\(12\)61229-X](https://doi.org/10.1016/S0140-6736(12)61229-X).
- Weatherall DJ. Phenotype-genotype relationships in monogenic disease: lessons from the thalassaemias. *Nat Rev Genet*. 2001;2:245–55. <https://doi.org/10.1038/35066048>.
- Dampier C, Ely E, Eggleston B, Brodecki D, O'Neal P. Physical and cognitive-behavioral activities used in the home management of sickle pain: a daily diary study in children and adolescents. *Pediatr Blood Cancer*. 2004;43:674–8. <https://doi.org/10.1002/pbc.20162>.
- Platt OS, Thorington BD, Brambilla DJ, Milner PF, Rosse WF, Vichinsky E, et al. Pain in sickle cell disease. Rates and risk factors. *N Engl J Med*. 1991;325:11–6. <https://doi.org/10.1056/NEJM199107043250103>.
- Manning LR, Russell JE, Padovan JC, Chait BT, Popowicz A, Manning RS, et al. Human embryonic, fetal, and adult hemoglobins have different subunit interface strengths. Correlation with lifespan in the red cell. *Protein Sci*. 2007;16:1641–58. <https://doi.org/10.1110/ps.072891007>.
- Thein SL, Menzel S. Discovering the genetics underlying foetal haemoglobin production in adults. *Br J Haematol*. 2009;145(4):455–67. <https://doi.org/10.1111/j.1365-2141.2009.07650.x>.
- Mosca A, Paleari R, Ivaldi G, Galanello R, Giordano PC. The role of hemoglobin A2 testing in the diagnosis of thalassaemias and related hemoglobinopathies. *J Clin Pathol*. 2009;2:13–7. <https://doi.org/10.1136/jcp.2008.056945>.
- Menzel S, Garner C, Gut I, Matsuda F, Yamaguchi M, Heath S, et al. A QTL influencing F cell production maps to a gene encoding a zinc-finger protein on chromosome 2p15. *Nat Genet*. 2007;39:1197–9. <https://doi.org/10.1038/ng2108>.
- Garner C, Tatu T, Reittie JE, Littlewood T, Darley J, Cervino S, et al. Genetic influences on F cells and other hematologic variables: a twin heritability study. *Blood*. 2000;95:342–6.
- Menzel S, Lay S. Genetic architecture of hemoglobin F control. *Curr Opin Hematol*. 2009;16(3):179–86. <https://doi.org/10.1097/MOH.0b013e328329d07a>.
- Thein SL, Menzel S, Peng X, Best S, Jiang J, Close J, 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*. 2007;104:11346–51. <https://doi.org/10.1073/pnas.0611393104>.
- Thein SL, Menzel S, Lathrop M, Garner C. Control of fetal hemoglobin: new insights emerging from genomics and clinical implications. *Hum Mol Genet*. 2009;18:216–23. <https://doi.org/10.1093/hmg/ddp401>.
- Makani J, Menzel S, Nkya S, Cox SE, Drasar E, Soka D, et al. Genetics of fetal hemoglobin in Tanzanian and British patients with sickle cell anemia. *Blood*. 2011;117:1390–2. <https://doi.org/10.1182/blood-2010-08-302703>.
- Mtatiro SN, Singh T, Rooks H, Mgaya J, Mariki H, Soka D, et al. Genome wide association study of fetal hemoglobin in sickle cell anemia in Tanzania. *PLoS One*. 2014;9(11):e111464. <https://doi.org/10.1371/journal.pone.0111464>.
- Zhou D, Liu K, Sun CW, Pawlik KM, Townes TM. KLF1 regulates BCL11A expression and gamma- to beta-globin gene switching. *Nat Genet*. 2010;42(9):742–4. <https://doi.org/10.1038/ng.637>.
- Siatecka M, Bieker JJ, DC W. The multifunctional role of EKLF / KLF1 during erythropoiesis. *Blood*. 2011;118:2044–54. <https://doi.org/10.1182/blood-2011-03-331371>.

20. Sankaran VG, Orkin SH. The switch from fetal to adult hemoglobin. *Cold Spring Harb Perspect Med*. 2013;3(1):a011643. <https://doi.org/10.1101/cshperspect.a011643>.
21. Thein SL. Genetic association studies in  $\beta$ -hemoglobinopathies. In: *ASH education program book*; 2013. p. 354–61. <https://doi.org/10.1182/asheducation-2013.1.354>.
22. Amaya M, Desai M, Gnanapragasam MN, Wang SZ, Zhu SZ, Williams DC, et al. Mi2 $\beta$ -mediated silencing of the fetal  $\gamma$ -globin gene in adult erythroid cells. *Blood*. 2013;121:3493–501. <https://doi.org/10.1182/blood-2012-11-466227>.
23. Torrado M, Low JKK, Silva APG, Schmidberger JW, Sana M, Tabar MS, et al. Refinement of the subunit interaction network within the nucleosome remodelling and deacetylase (NuRD) complex. *FEBS J*. 2017;284:4216–32. <https://doi.org/10.1111/febs.14301>.
24. Mtairo SN, Mgaya J, Singh T, Mariki H, Rooks H, Soka D, et al. 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 Med Genet*. 2015;16:4. <https://doi.org/10.1186/s12881-015-0148-3>.
25. Li H, Ruan J, Durbin R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res*. 2008;18:1851–8. <https://doi.org/10.1101/gr.078212.108>.
26. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytzky A, et al. The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res*. 2010;20:1297–303. <https://doi.org/10.1101/gr.107524.110>.
27. Gézi A, Bolgár B, Marx P, Sarkozy P, Szalai C, Antal P. VariantMetaCaller: automated fusion of variant calling pipelines for quantitative, precision-based filtering. *BMC Genomics*. 2015;16:875. <https://doi.org/10.1186/s12864-015-2050-y>.
28. 1000 Genomes Project Consortium, Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, Handsaker RE, et al. An integrated map of genetic variation from 1,092 human genomes. *Nature*. 2012;491(7422):56–65. <https://doi.org/10.1038/nature11632>.
29. Cornish A, Guda CA. Comparison of variant calling pipelines using genome in a bottle as a reference. *Biomed Res Int*. 2015;1–11. <https://doi.org/10.1155/2015/456479>.
30. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*. 2010;26:841–2. <https://doi.org/10.1093/bioinformatics/btq033>.
31. Li H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*. 2011;27:2987–93. <https://doi.org/10.1093/bioinformatics/btr509>.
32. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res*. 2010;38:e164. <https://doi.org/10.1093/nar/gkq603>.
33. Ng PC, Henikoff S. SIFT: predicting amino acid changes that affect protein function. *Nucleic Acids Res*. 2003;31:3812–4.
34. Ng PC, Henikoff S. Predicting the effects of amino acid substitutions on protein function. *Annu Rev Genomics Hum Genet*. 2006;7:61–80. <https://doi.org/10.1146/annurev.genom.7.080505.115630>.
35. Fujita A, Kojima K, Patriota AG, Sato JR, Severino P, Miyano S. A fast and robust statistical test based on likelihood ratio with Bartlett correction to identify granger causality between gene sets. *Bioinformatics*. 2010;26:2349–51. <https://doi.org/10.1093/bioinformatics/btq427>.
36. Reva B, Antipin Y, Sander C. Determinants of protein function revealed by combinatorial entropy optimization. *Genome Biol*. 2007;8:R232. <https://doi.org/10.1186/gb-2007-8-11-r232>.
37. Reva B, Antipin Y, Sander C. Predicting the functional impact of protein mutations: application to cancer genomics. *Nucleic Acids Res*. 2011;39:e118. <https://doi.org/10.1093/nar/gkr407>.
38. Shihab HA, Gough J, Mort M, Cooper DN, Day IN, Gaunt TR. Ranking non-synonymous single nucleotide polymorphisms based on disease concepts. *Hum Genomics*. 2014;8:11. <https://doi.org/10.1186/1479-7364-8-11>.
39. Dong C, Wei P, Jian X, Gibbs R, Boerwinkle E, Wang K, et al. Comparison and integration of deleteriousness prediction methods for nonsynonymous SNVs in whole exome sequencing studies. *Hum Mol Genet*. 2015;24:2125–37. <https://doi.org/10.1093/hmg/ddu733>.
40. Kircher M, Witten DM, Jain P, O’Roak BJ, Cooper GM, Shendure J. A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet*. 2014;46:310–5. <https://doi.org/10.1038/ng.2892>.
41. Cooper GM, Stone EA, Asimenos G, NISC Comparative Sequencing Program, Green ED, Batzoglu S, et al. Distribution and intensity of constraint in mammalian genomic sequence. *Genome Res*. 2005;15:901–13. <https://doi.org/10.1101/gr.3577405>.
42. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, et al. A method and server for predicting damaging missense mutations. *Nat Methods*. 2010;7:248–9. <https://doi.org/10.1038/nmeth0410-248>.
43. Garber M, Guttman M, Clamp M, Zody MC, Friedman N, Xie X. Identifying novel constrained elements by exploiting biased substitution patterns. *Bioinformatics*. 2009;25:i54–62. <https://doi.org/10.1093/bioinformatics/btp190>.
44. Li MX, Gui HS, Kwan JSH, Bao SY, Sham PC. A comprehensive framework for prioritizing variants in exome sequencing studies of Mendelian diseases. *Nucleic Acids Res*. 2012;40:e53. <https://doi.org/10.1093/nar/gkr1257>.
45. Li MX, Kwan JSH, Bao SY, Yang W, Ho SL, Song YQ, et al. Predicting Mendelian disease-causing non-synonymous single nucleotide variants in exome sequencing studies. *PLoS Genet*. 2013;9:e1003143. <https://doi.org/10.1371/journal.pgen.1003143>.
46. Sayers EW, Agarwala R, Bolton EE, Brister JR, Canese K, Clark K, Connor R, Fiorini N, Funk K, Hefferon T, et al. Database resources of the national center for biotechnology information. *Nucleic Acids Res*. 2019;47(D1):D23–8.
47. Houdayer C, Dehainault C, Mattler C, Michaux D, Caux-Moncoutier V, Pagès-Berhouet S, et al. Evaluation of in silico splice tools for decision-making in molecular diagnosis. *Hum Mutat*. 2008;29:975–82. <https://doi.org/10.1002/humu.20765>.
48. Chimusa ER, Mbiyavanga M, Mazandu GK, Mulder NJ. ancGWAS: a post genome-wide association study method for interaction, pathway and ancestry analysis in homogeneous and admixed populations. *Bioinformatics*. 2016;32:549–56. <https://doi.org/10.1093/bioinformatics/btv619>.
49. Wu J, Vallenius T, Ovaska K, Westermarck J, Mäkelä TP, Hautaniemi S. Integrated network analysis platform for protein-protein interactions. *Nat Methods*. 2009;6:75–7. <https://doi.org/10.1038/nmeth.1282>.
50. Kuleshov MV, Jones MR, Rouillard AD, Fernandez NF, Duan Q, Wang Z, et al. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res*. 2016;44:W90–7. <https://doi.org/10.1093/nar/gkw377>.
51. Kanehisa M, Sato Y, Furumichi M, Morishima K, Tanabe M. New approach for understanding genome variations in KEGG. *Nucleic Acids Res*. 2019;47:D590–5. <https://doi.org/10.1093/nar/gky962>.
52. Mi H, Muruganujan A, Ebert D, Huang X, Thomas PD. PANTHER version 14: more genomes, a new PANTHER GO-slim and improvements in enrichment analysis tools. *Nucleic Acids Res*. 2019;47(D1):D419–26. <https://doi.org/10.1093/nar/gky1038>.
53. Nishimura D. BioCarta. *Biotech Softw Internet Rep*. 2001;2:117–20. <https://doi.org/10.1089/152791601750294344>.
54. Fargat A, Juge S, Matthews L, Sidiropoulos K, Gillespie M, Garapati P, Haw R, Jassal B, Korninger F, May B, et al. The Reactome pathway knowledgebase. *Nucleic Acids Res*. 2018;46(D1):D649–55. <https://doi.org/10.1093/nar/gkx1132>.
55. The Gene Ontology Consortium. The gene ontology resource: 20 years and still GOing strong. *Nucleic Acids Res*. 2019;47(D1):D330–8. <https://doi.org/10.1093/nar/gky1055>.
56. Karczewski KJ, Weisburd B, Thomas B, Solomonson M, Ruderfer DM, Kavanagh D, et al. The ExAC browser: displaying reference data information from over 60 000 exomes. *Nucleic Acids Res*. 2017;45:D840–5. <https://doi.org/10.1093/nar/gkw971>.
57. Pirooznia M, Kramer M, Parla J, Goes FS, Potash JB, McCombie WR, Zandi PP. Validation and assessment of variant calling pipelines for next-generation sequencing. *Hum Genomics*. 2014;8(1):14.
58. Teo YY, Small KS, Kwiatkowski DP. Methodological challenges of genome-wide association analysis in Africa. *Nat Rev Genet*. 2010;11(2):149.
59. O’Leary NA, Wright MW, Brister JR, Ciufu S, Haddad D, McVeigh R, et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res*. 2016;44:D733–45. <https://doi.org/10.1093/nar/gkv1189>.
60. Chalaow N, Thein SL, Viprakasit V. The 12.6 kb-deletion in the  $\beta$ -globin gene cluster is the known Thai/Vietnamese ( $\delta\beta$ )0-thalassemia commonly found in Southeast Asia. *Haematologica*. 2013;98:e117–8. <https://doi.org/10.3324/haematol.2013.090613>.
61. Hamid M, Nejad LD, Shariati G, Galehdari H, Saberi A, Mohammadi-Anaei M, et al. The first report of a 290-bp deletion in  $\beta$ -Globin gene in the South of

- Iran. *Iran Biomed J.* 2017;21(2):126–8. <https://doi.org/10.18869/acadpub.ijb.21.2.126>.
62. Thein SL, Craig JE. Genetics of Hb F/F cell variance in adults and heterocellular hereditary persistence of fetal hemoglobin. *Hemoglobin.* 1998;22:401–14.
63. Akinola RO, Mazandu GK, Mulder NJ. A quantitative approach to analyzing genome reductive evolution using protein–protein interaction networks: a case study of *Mycobacterium leprae*. *Front Genet.* 2016;7:39. <https://doi.org/10.3389/fgene.2016.00039>.
64. Jiang J, Best S, Menzel S, Silver N, Lai MI, Surdulescu GL, et al. cMYB is involved in the regulation of fetal hemoglobin production in adults. *Blood.* 2006;108:1077–83. <https://doi.org/10.1182/blood-2006-01-008912>.
65. Mazandu GK, Mulder NJ. Generation and analysis of large-scale data-driven *Mycobacterium tuberculosis* functional networks for drug target identification. *Adv Bioinform.* 2011;2011:801478, 14 pages. <https://doi.org/10.1155/2011/801478>.

### Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Ready to submit your research? Choose BMC and benefit from:**

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

**At BMC, research is always in progress.**

Learn more [biomedcentral.com/submissions](https://biomedcentral.com/submissions)

