OXFORD

Human Molecular Genetics, 2023, Vol. 32, 12, 1946–1958
https://doi.org/10.1093/hmg/ddad026
Advance access publication date 8 February 2023
Original Article

Genome-wide association study identifies novel candidate malaria resistance genes in Cameroon

Kevin K. Esoh[1,†], Tobias O. Apinjoh[2,†], Alfred Amambua-Ngwa[3], Steven G. Nyanjom[4], Emile R. Chimusa[5], Lucas Amenga-Etego[6], Ambroise Wonkam[1,7,‡,*] and Eric A. Achidi[2,‡]

[1]Division of Human Genetics, Department of Pathology, University of Cape Town, Health Sciences Campus, Anzio Rd, Observatory, Cape Town, 7925, South Africa
[2]Department of Biochemistry and Molecular Biology, University of Buea, PO Box 63, South West Region, Buea, Cameroon
[3]Medical Research Council Unit, The Gambia, at LSHTM, PO Box 273, Banjul, The Gambia
[4]Department of Biochemistry, Jomo Kenyatta University of Agriculture and Technology, PO Box 62000, City Square, Nairobi, Kenya
[5]Department of Applied Sciences, Faculty of Health and Life Sciences, Northumbria University, Newcastle, Tyne and Wear NE1 8ST, UK
[6]West African Centre for Cell Biology of Infectious Pathogens, University of Ghana, PO Box LG 25, Legon, Accra, Ghana
[7]McKusick-Nathans Institute of Genetic Medicine and Department of Genetic Medicine, Johns Hopkins University School of Medicine, 773 N. Broadway, MRB 439 , Baltimore, MD 21205, USA
*To whom correspondence should be addressed at: McKusick-Nathans Institute & Department of Genetic Medicine, Johns Hopkins University School of Medicine, 773 N. Broadway, MRB 439, Baltimore, MD 21205, USA. Tel: +1 4109553378; Email: awonkam1@jhmi.edu
†These authors contributed equally to this work.
‡Co-senior authors have equally contributed.

Abstract

Recent data suggest that only a small fraction of severe malaria heritability is explained by the totality of genetic markers discovered so far. The extensive genetic diversity within African populations means that significant associations are likely to be found in Africa. In their series of multi-site genome-wide association studies (GWAS) across sub-Saharan Africa, the Malaria Genomic Epidemiology Network (MalariaGEN) observed specific limitations and encouraged country-specific analyses. Here, we present findings of a GWAS of Cameroonian participants that contributed to MalariaGEN projects ($n = 1103$). We identified protective associations at polymorphisms within the enhancer region of *CHST15* [Benjamin–Hochberg false discovery rate (FDR) < 0.02] that are specific to populations of African ancestry, and that tag strong eQTLs of *CHST15* in hepatic cells. In-silico functional analysis revealed a signature of epigenetic regulation of *CHST15* that is preserved in populations in historically malaria endemic regions, with haplotype analysis revealing a haplotype that is specific to these populations. Association analysis by ethnolinguistic group identified protective associations within *SOD2* (FDR < 0.04), a gene previously shown to be significantly induced in pre-asymptomatic malaria patients from Cameroon. Haplotype analysis revealed substantial heterogeneity within the beta-like globin (*HBB*) gene cluster amongst the major ethnic groups in Cameroon confirming differential malaria pressure and underscoring age-old fine-scale genetic structure within the country. Our findings revealed novel insights in the evolutionary genetics of populations living in Cameroon under malaria pressure with new significant protective loci (*CHST15* and *SOD2*) and emphasized the significant attenuation of genetic association signals by fine-scale genetic structure.

Introduction

Host genetic factors play a major role in malaria phenotypic variance, contributing up to 25% of the differences in clinical expressions in severe malaria (SM) (Mendelian Inheritance in Man (MIM): 611162) amongst individuals (1,2). Although the sickle cell heterozygous variant (sickle cell trait; HbAS) affords the strongest protection against SM, it only explains ∼2% of the total SM variance (1,3,4). Recent data further demonstrate that only a small fraction of SM heritability is explained by the totality of genetic markers discovered so far (1), implying the existence of many unknown markers. The discovery of malaria susceptibility loci is important in elucidating disease pathophysiological pathways, which can serve as critical therapeutic targets. From the time of its establishment in 2005, the Malaria Genomic Epidemiology Network (MalariaGEN) (5) has led in this area, uncovering multiple malaria susceptibility loci in sub-Saharan Africa (*ATP2B4*, *FREM3-GYP* and *EPHA7* for example) (1,6–11), whilst confirming other previously characterized loci (*HBB*, *ABO* and *G6PD*). A majority of these loci are shared across sub-Saharan African populations (HbS, *G6PD* and *ABO* for instance) whilst some are restricted to specific geographic regions; for instance, the HbC in West Africa with an epicentre around Burkina Faso (12) and the Dantu (*GYPA-B*) in East Africa (1,9).

As a result of the high genetic diversity of African populations, a characteristic observation from MalariaGEN multi-site genome-wide association studies (GWAS) in sub-Saharan Africa is heterogeneity in association signals. For instance, the sickle cell trait was observed to exhibit its weakest effect in Cameroon as compared with other populations where the trait is similarly prevalent (such as in The Gambia where the strongest effect was observed) or less prevalent (such as in Ghana) (1). This weak protective effect of sickle cell trait against SM in Cameroon could not be explained by the opposing HbC variant, which is very rare in the population. Variants in other known malaria protective loci (e.g. *FREM3, ATP2B4, GYPA-B*) have also not been independently

Received: August 16, 2022. **Revised:** January 17, 2023. **Accepted:** February 7, 2023

replicated in Cameroon. In addition, the *G6PD* deficiency variants that afford protection against various malaria sub-phenotypes in some African populations exhibit opposing effects in Cameroon (13). In the face of these observations, the MalariaGEN recommended country-specific analyses (10) that have been particularly successful in countries like The Gambia, Ghana (14), Tanzania (15) and Kenya (9).

Considering that Cameroonian ancestral populations are amongst the oldest on the continent (16,17), and considering data that suggest the virulent *Plasmodium falciparum* malaria and the HbS mutation may have their origins in Cameroon (18–20), it is likely that the genomes of Cameroonians are enriched with population-specific variants that are nearly—if not equally or more strongly—as protective as sickle cell trait. We sought to investigate this by taking advantage of genome-wide genotype data for Cameroonian participants that contributed to MalariaGEN consortium studies (21) consisting of 2.3 million Single-nucleotide polymorphism (SNPs) genotyped across 1471 samples from individuals belonging to the Bantu (BA), Semi-Bantu (SB) and Fulani (FU) ethnolinguistic groups. We assessed the possibility of increasing the informativeness of our data by employing different publicly available genotype imputation panels (22–24). We also investigated the possibility that Genomes of Cameroon's major ethnolinguistic groups may have evolved differently under pressure from malaria. We report novel candidate protective loci that may shed further light on the contribution of Cameroonian genomes to malaria phenotypic variance, and we report substantial differences in the association pattern of the two ethnolinguistic groups that should be considered in future genetic studies in the country. Finally, we present evidence showing differential evolutionary paths of Cameroon's major ethnic populations under malaria pressure.

Results

Characteristics of the study participants

Quality control included identification and exclusion of related individuals (Supplementary Material, Fig. S1), outlying heterozygosity and missing genotype proportion (Supplementary Material, Fig. S2). Our analysis was on the basis of all clinical malaria cases versus controls as encoded in the sample process report retrieved from the MalariaGEN site, hence further stratifying the cases into the different sub-phenotypes yielded low samples, which underpowered the study. Moreover, the most prevalent sub-phenotype in the pooled data was SM (13), hence we did not expect to lose much power to detect potential resistance loci by pooling the sub-phenotypes together. The clinical and demographic characteristics of our cohort has been previously described by Achidi *et al.* (25) and MalariaGEN CP1 (Accessed September 21, 2020). A basic description of the characteristics of the participants retained for our analysis are presented in Figure 1.

Assessment of imputation performance and allele frequencies at known malaria loci

Following our previous analysis of fine-scale genetic structure in this cohort (26), we pruned population outliers by increasing the number of outlier removal iterations from the default 5–10 using *smartpca*. The remaining samples clustered with other African populations as expected (Fig. 2A). The spread observed within the Cameroonian population was expected and consistent with previous findings, confirming substantial genetic sub-structure within Cameroon as compared with other sSA populations (26,27). Upon alignment of our data to the 1000 Genomes reference panel (KGP), we observed that ∼2% of our SNPs were absent from the reference panel. A similar observation was made with the KGP panel from the Michigan Imputation (MI) server, whilst the TOPMed Imputation (TI) panel lacked ∼4% of SNPs we investigated. The Sanger Imputation server did not produce reference allele overlap information; however, we expect this percentage to be similar, if not smaller, to our in-house strategy (on the basis of the KGP panel) and the MI panels given the larger size of the African Genome Resource (AGR) in the Sanger Imputation server with respect to African representation as compared with the KGP panel. The disparate reference allele overlap was reflected in the low squared correlation (r^2) between reference allele frequencies for the TOPMed panel as compared with the KGP panel (Supplementary Material, Fig. S3a and b). Interestingly, the low reference allele overlap of the TOPMed panel did not appear to affect its accuracy as it outperformed all the other panels (Fig. 2B). We also noted a slightly better imputation accuracy for our in-house strategy than for both the Michigan and Sanger strategies that could be attributed to power gain by tuning the imputation parameters to increase accuracy (Supplementary Material, Fig. S3c). Unsurprisingly, the Sanger imputation service exhibited a better performance at ultra-rare variants [minor allele frequency (MAF) < 0.002] reflecting its enrichment with indigenous African samples from the AGR (Supplementary Material, Fig. S3d). Further assessment of the imputation performance at key malaria loci revealed an unexpected failure of the TI service to impute the HbS allele whilst the Michigan, Sanger and our in-house imputation strategies imputed the variant with $r^2 > 0.88$, $r^2 > 0.86$ and $r^2 > 0.90$, respectively. In a personal communication, the TI team acknowledged the deficiency in their 'Freeze 8' release and attributed it to an automatic quality check that was marked to be resolved in their 'Freeze 9' release.

Supplementary Material, Table S1 shows the allele frequencies and imputation accuracy of variants in key malaria-associated loci in our data from each of the imputation panels and for continental populations. The alleles were imputed with a generally high accuracy. Allele frequencies were largely similar amongst the BA and SB ethnic groups, consistent with our previous demonstration of similar genome-wide allele frequency spectra amongst the ethnic groups (26). However, there were noticeable differences at some loci (highlighted rows) such as higher HbS and *G6PD* variants in the BA, and a slightly higher prevalence of the *FREM3* variant rs184895969 in the SB. These allele frequency differences might indicate differences in the evolution of the two ethnic groups under malaria pressure. Interestingly, the loci showing some of the lowest imputation accuracy reside in known regions of balancing selection harbouring high haplotype diversities and low linkage disequilibrium (LD).

Association analysis reveals candidate loci and disparate patterns of association

Prior to association testing, we estimated narrow sense heritability/pseudo-heritability (heritability contributed by additive genetic variance component) in the pooled dataset at h^2 (Vg/Vp) ∼23% using EMMAX (22.2%), GCTA (21.9%) and BOLT-LMM (23.6%) in the imputed dataset, and at $h^2 \sim 22\%$ using EMMAX in the pre-imputation dataset. These values are similar to previous estimates (∼23%) of SM heritability (1,2), and reflect the fact that SM was the most prevalent sub-phenotype in our data. However, the large standard errors of estimates that ranged from 0.16 to 0.48 reflect the low sample size in the study. The total genetic variants were predicted to explain not >9% of the heritability.

Association analyses were performed for the merged set of 1103 samples that passed all filters (405 cases and 598 controls; 444 females, 659 males), and for the BA ($n = 490$; 209 cases

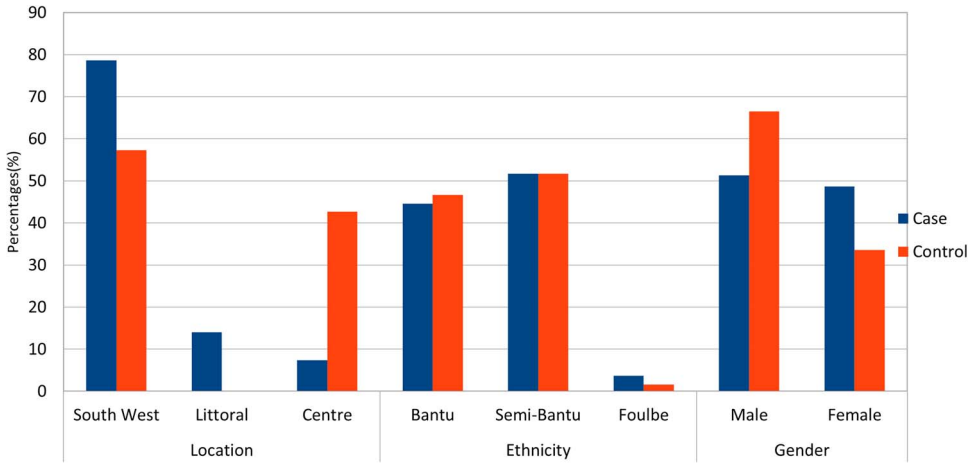

Figure 1. Basic demographic characteristics, and recruitment sites of cases and controls.

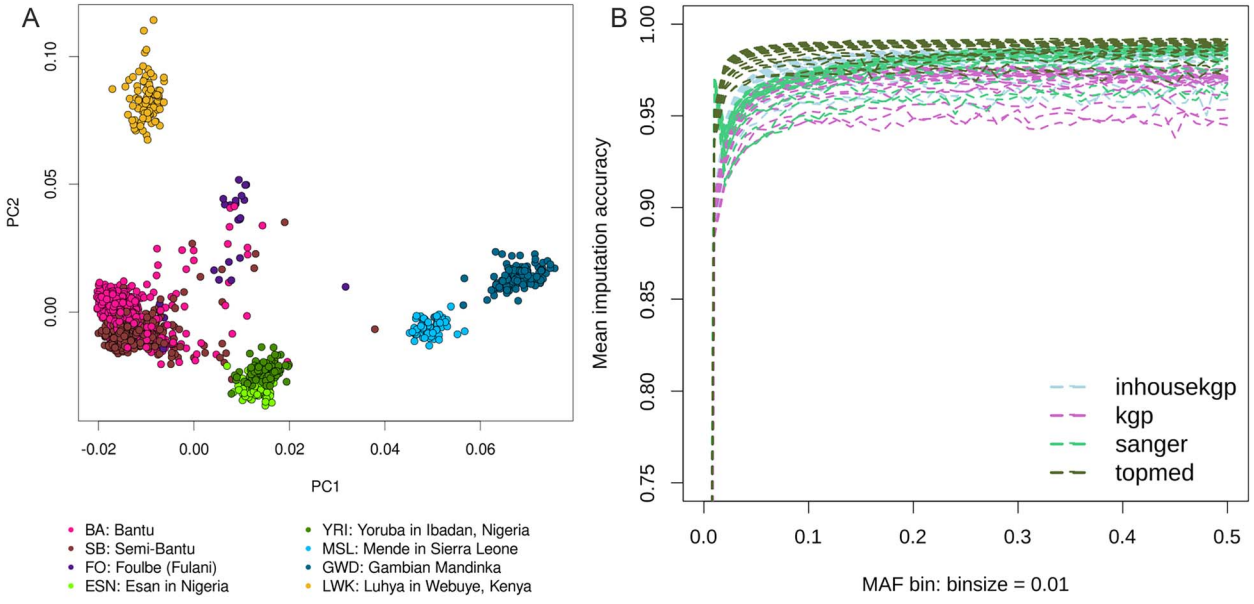

Figure 2. Principal Component Analysis (PCA) and Imputation Performance. (**A**) PCA plot of Cameroonians and other African populations in the 1000 Genomes Project. Cameroonians cluster centrally between East (LWK = Luhya in Webuye, Kenya) and West African (YRI = Yoruba in Ibadan, Nigeria, ESN = Esan in Nigeria, GWD = Gambian in Western Division, Mandinka and MSL = Mende in Sierra Leone) populations as expected. (**B**) Per-chromosome imputation accuracy of our in-house, TOPMed, Michigan and Sanger imputation strategies.

and 281 controls; 185 females and 305 males) as well as the SB ($n = 539$; 252 cases and 287 controls; 228 females and 311 males) populations separately. Multiple testing correction of the association test results was performed by the Benjamin–Hochberg false discovery rate (FDR). Table 1 shows candidate loci for that we observed significant associations when the imputed dataset was filtered to exclude variants with imputation accuracy, $r^2 < 0.60$, MAF < 1%, genotype call rate < 95% and variants that failed the Hardy–Weinberg equilibrium (HWE) test at $P = 1e - 04$. We observed significant protective variants (rs113508623 and rs80169640, FDR < 0.02) in the enhancer region of the carbohydrate-sulfotransferase 15 (*CHST15*, 10q26.13) in the pooled dataset under an additive model of inheritance (MOI) (Fig. 3A). Interestingly, the associations were only observed in data imputed with the AGR from the Sanger imputation server. Significant protective associations were also observed in the superoxide dismutase 2 (*SOD2*, 6q25.3) gene (FDR < 0.04) in the SB under a dominant MOI. In the BA, a significant variant (rs75944478) associated with increased risk of clinical malaria in *ZCCHC14* appeared to only be suggestive after multiple testing correction (FDR = 0.22). Notably, we did not observe significant associations in the well-characterized HbS locus. We noted that the variant was filtered out from all the panels that imputed it following the exclusion of variants with genotype call rate < 95%. However, we observed a suggestive association at rs1378749 ($P = 4.98e - 06$) in the SB, downstream of the HBB-3′HS1, a region known to be critical in the regulation of *HBB* gene cluster expression, implying that a larger sample size is likely to reveal significant SNPs in this region. The complete list of the top significant and suggestive signals observed from our analyses is presented as Supplementary Material, Table S2. Zooming into the regions of *CHST15* and *SOD2* showing significant associations (Table 2, Supplementary Material, Table S3), we observed moderate to strong LD localizing within and around the genes (Fig. 3B and C). Considering their biological roles, the genes seem appealing as plausible associations implicated in modulating malaria phenotypes. We therefore performed in-silico functional analyses to gain additional insight.

Table 1. Phased SNP-based beta-like globin (*HBB*) gene cluster haplotypes and other non-classical haplotypes in Cameroonians

Haplotype Name	rs968857 (*HBBP1*—HincII)	rs10128556 (*HBBP1*—HincII)	rs28440105 (*HBG1*—HindIII)	rs3834466 (*HBE1*—HincII)	Haplotype
SNP (aaf/R^2)	0.20/0.99	0.12/0.94	0.87/0.96	0.20/0.99	–
AI	T (0)	T (1)	C (1)	GT (1)	0111
SEN	T (0)	T (1)	C (1)	G (0)	0110
BEN	T (0)	C (0)	C (1)	G (0)	0010
CAR	C (1)	C (0)	C (1)	G (0)	1010
CAM	T (0)	C (0)	A (0)	G (0)	0000
Hap1	T (0)	C (0)	A (0)	GT (1)	0001
Hap2	T (0)	C (0)	C (1)	GT (1)	0011
Hap3	T (0)	T (1)	A (0)	G (0)	0100
Hap4	C (1)	C (0)	A (0)	G (0)	1000
Hap5	C (1)	C (0)	A (0)	GT (1)	1001
Hap6	C (1)	C (0)	C (1)	GT (1)	1011
Hap7	C (1)	T (1)	C (1)	G (0)	1110

aaf = Alternate allele frequency, R^2 = imputation accuracy, Hap = other haplotypes named chronologically, 0 = reference allele, 1 = alternate allele.

Table 2. Candidate associated loci with significant and suggestive association signals in the post-imputation dataset

rsid	chr:pos	ref/alt	maf	Unadjusted *P*-value	FDR	Odds ratio	95% confidence interval	gene(s)	MOI
Pooled									
rs80169640	10:125859979	G/A	0.17	2.07e-09	0.015	0.41	0.31–0.55	*CHST15*	additive
rs113508623	10:125859606	T/C	0.17	2.07e-09	0.015	0.41	0.31–0.55	*CHST15*	additive
SB									
rs2842958	6:160108425	T/G	0.32	4.62e-09	0.03	0.33	0.23–0.49	*SOD2*	dominant
BA									
rs75944478	16:87486312	G/A	0.22	3.30e-08	0.21	2.79	1.95–3.99	*ZCCHC14*	additive

rsid = reference SNP ID, chr:pos = chromosome and position in hg19 coordinate, ref/alt = reference/alternate allele.

An evolutionarily preserved signature of epigenetic regulation of *CHST15* in malarious environments suggests a role in clinical malaria protection

CHST15, also known as B-cell Rag-associated gene for its co-expression with recombination-activating gene 1 (*RAG1*) in B-cells is a type II trans-membrane glycoprotein that; (i) induces *RAG1* expressions in B-cell lines (28), (ii) serves as signalling receptor on the surface of unstimulated mature B-cells (29) and (iii) possesses sulfotransferase enzymatic activity whereby it catalyzes the transfer of sulphate residues to chondroitin sulphate A (CSA) and dermatan sulphate (DS) (30). When it transfers sulphates to the C-4 and C-6 hydroxyl groups of CSA, it produces chondroitin sulphate E (CSE). Recall that CSA is the receptor of choice in the placenta for the *P. falciparum* erythrocyte membrane protein-1 encoded by the *VAR2CSA* gene (31,32), hence is a risk factor for pregnancy-associated malaria (PAM). Therefore, given a high burden of PAM such as in West and Central Africa where the prevalence reaches 35% (33), one could imagine a scenario in which the expression of *CHST15* is augmented leading to an increase in sulfotransferase activity and a consequent reduction in CSA, thus effectively protecting against PAM. Furthermore, *CHST15* as a signalling receptor on the surface of unstimulated mature B-cells, and its ability to induce *RAG1* expression in B-cell lines may serve as a means to equip the immune system with the capacity to utilize the enormous repertoire of antibody specificities that come with V(D)J recombination (34) in order to mount effective defence mechanisms against the *Plasmodium* parasite. Moreover, gene expression data show that whole blood and the spleen (critical components of the immune system) are amongst the tissues in which *CHST15* is most expressed (Supplementary Material, Fig. S4), thus making the gene an attractive target for adaptive evolution in malarious environments.

Observing that rs113508623 and rs80169640 are regulatory region (specifically in the enhancer of *CHST15*) and intergenic variants, respectively, we sought to investigate any effect on *CHST15* expression in relevant tissues, which might explain its association with clinical malaria. Searching Genotype-Tissue Expression (GTEx) data for known *cis*-expression and splicing quantitative trait loci (*cis*-Expression quantitative trait loci (eQTL) and sQTL) via the GTEx Portal, (version 8, https://gtexportal.org), we found no report of the variants as either *cis*-eQTLs or sQTLs. Importantly, we noted that whilst these variants are common amongst populations of African ancestry, they are either very rare or completely absent in other ancestries and this might explain why they have not been reported as QTLs given the under-representation of African populations in genomics databases. Drawing from our regional association analysis (Fig. 3B), we next investigated whether the variants tag known QTLs. Three SNPs were found to be in LD with the two lead SNPs including; rs28657878, rs77797884 and rs4359147 (Supplementary Material, Fig. S5). Two of the SNPs, rs28657878 and rs4359147, were reported as *cis*-eQTLs occurring in open chromatin and transcription factor binding sites, respectively, in the liver (exhibiting strong upregulatory effects) and in the thyroid (exhibiting weak upregulatory effects) and were also strongly suggestive signals in our analysis (Supplementary Material, Table S2).

When we compared the distribution of all five SNPs in global populations, we observed an apparent restriction of rs80169640, rs113508623 and rs77797884 to regions of historical malaria endemicity; including amongst individuals of African ancestry, in

South America and in Vietnam (the Greater Mekong Subregion). Assuming that the minor alleles were the derived alleles, the distribution of the variants would suggest a positive selection in malarious environments. Intriguingly however, ENSEMBL's Enredo-Pecan-Ortheus (EPO) (35) multiple sequence alignment pipeline with nine primate species revealed that the minor alleles of the three SNPs, as well as rs4359147, are in fact the ancestral states (Fig. 4A). It also occurs that these ancestral alleles are the effect alleles of the variants (the alleles associated with increased *CHST15* expression) according to ENSEMBL and GTEx data, hence effectively positing that a negative selection has rather purged the alleles from most populations, whilst an opposing force has preserved the alleles in populations that have endured historical malaria pressure because of some selective advantage, such as protection against malaria. Whereas we noted extremely low LD in the genomic region harbouring the variants in all ancestries, LD structure around the lead SNPs was similar for populations that have endured historical malaria pressure (Supplementary Material, Fig. S5), and haplotype analysis revealed a shared haplotype (Hap8) carrying the ancestral alleles amongst these populations (Fig. 4C, see Supplementary Material, Table S4). In addition, other haplotypes carrying the lead SNPs were more specific to populations of African ancestry (Hap3, Hap6 and Hap7) (Supplementary Material, Table S4). These observations, together with extended haplotype homozygosity (EHH) associated with the ancestral alleles (Fig. 4B), suggest a recent balancing selection in the genomic region in malarious environments.

To gain further insight on the possible regulatory effect of the variants, we analysed publicly available ChiP-Seq peaks and chromatin state data via the University of California Santa Cruz (UCSC) Genome Browser (https://genome.ucsc.edu). These revealed at least three DNaseI hypersensitive sites 3′ to *CHST15* (hereafter HS1, HS2 and HS3) indicating open chromatin regions with transcriptional activity (Supplementary Material, Fig. S6). In addition, deoxyribonucleic acid (DNA) methylation, acetylation and enhancer data revealed HS1 as the core regulatory element of *CHST15* harbouring the transcription start site (TSS), whilst HS3 located ∼12 kb upstream of the TSS was predicted as an active enhancer. Interestingly, our SNPs were located within HS2 located ∼6 kb upstream of the TSS, with apparently high levels of H3K9me3 methylation, marking an inactive chromatin state. Indeed, the HS2 was predicted to be only active in human embryonic stem cell lines, whilst either polycomb-repressed or inactive (heterochromatin) in other cell lines. Transcription factor binding site data revealed that rs28657878, rs113508623, rs77797884 and rs4359147 are located within the binding sites of some key transcriptional activators (SP1, SP2, KLF14, ZKSCAN5, ZNF530) and in fact, rs28657878, rs77797884 and rs4359147 disrupt transcription factor binding motifs (Supplementary Material, Tables S5 and S6, and Fig. S7).

SOD2 is significantly induced in clinical malaria patients from Cameroon

The role of reactive oxygen species (ROS) and reactive nitrogen species (RNS), respectively, in malaria has been extensively documented and reviewed in Ref. (36). Host defence mechanisms mounted against *Plasmodium* parasites usually result in the generation of ROS and RNS via the stimulation of inflammatory responses and oxidative stress by pro-inflammatory cytokines (such as tumour necrosis factor alpha; TNF-α) to eliminate the parasites. Antimalarial drugs are also thought to act by eliciting oxidative stress. However, this mechanism is only active in the acute phase of infection as excess and prolonged oxidative stress is toxic to host cells and can exacerbate malaria pathology. Interestingly, significant and suggestive signals were observed in TLR9 and TNF in this same cohort, which apparently increased susceptibility to malaria (37), suggestive of inflammatory responses and prolonged oxidative stress. SOD2 (MIM: 147460), a mitochondrial matrix enzyme encoded in nuclear DNA (6q25.3) and highly expressed in many organs including the liver, is an effective scavenger of ROS, preventing excess oxidative stress.

In 2006, a study involving malaria naïve individuals from the United States and clinical malaria patients from Cameroon found genes involved in the Interferon gamma (IFN-γ) signalling pathway to be amongst the most significantly induced genes in Cameroonian pre-symptomatic patients (38). Of note, IFN-γ is a major pro-inflammatory cytokine with direct antiparasitic activity, but however favours host hyperresponsiveness to Toll-like receptor (TLR) agonists resulting in overproduction of pro-inflammatory cytokines (39). Amongst the genes significantly induced in the IFN-γ signalling pathway in Cameroonian patients were the transcription factors *STAT1* and *IRF-1*, as well as *SOD2*, which, interestingly is a target gene of *IRF-1*. The data reviewed therefore provides a basis and model for the protective role of *SOD2* in early infection in Cameroonian malaria patients; host hyperresponsiveness to *Plasmodium* parasites as a result of INF-γ-induced TLR-mediated inflammation elicits a negative feedback response via *IFR-1* induction of *SOD2* to protect against oxidative stress. This would be consistent with the predominance of protective signals we observed in the *SOD2* locus. The very strong LD within the locus extending over a 100 000 bp region with a remarkably low recombination rate made it challenging to pinpoint the functionally relevant variant(s). QTL analysis showed that some of the significant variants are weak *cis*-eQTLs with heterogenous effects in various tissues.

HBB gene cluster haplotypes reveal differences amongst Cameroon's major ethnic groups because of malaria pressure

To investigate the possibility that Cameroon's major ethnic groups might have come under differential malaria pressures and therefore harbour different haplotype structures, which may be driving some of the differences in genetic associations, we investigated haplotypes in the beta-like globin (*HBB*; 11p15) gene cluster where HbS resides. The *HBB* gene cluster consists of beta-like globin genes arranged in a 5′—*HBE1-HBG2-HBG1-HBBP1-HBD-HBB*—3′ fashion (40). Unique combinations of restriction fragment length polymorphic sites (RFLPs) in the cluster give rise to specific 'classical' or 'typical' haplotypes that are named according to the place in which they were first observed; Senegal (SEN), Benin (BEN), Cameroon (CAM), Central African Republic (CAR—also called BU) and the Arab-India (AI) (41). Reference SNPs for most of the RFLPs have been identified, and phased SNP-based (haplotype) approaches have been described to accurately classify HbS-carrying chromosomes into the different classes (Table 1) (42). Because of extremely low levels of LD within the gene cluster, *HBB* haplotypes are therefore routinely used to track the evolution of HbS in populations. We therefore generated haplotypes for all HbS-carrying samples and computed haplotype frequencies on the basis of four previously characterized SNPs (rs3834466, rs28440105, rs10128556, rs968857) (42) (Table 4). We also computed haplotype frequencies for HbS-negative samples (base population) to gain insight into haplotype diversity or conservation in this population.

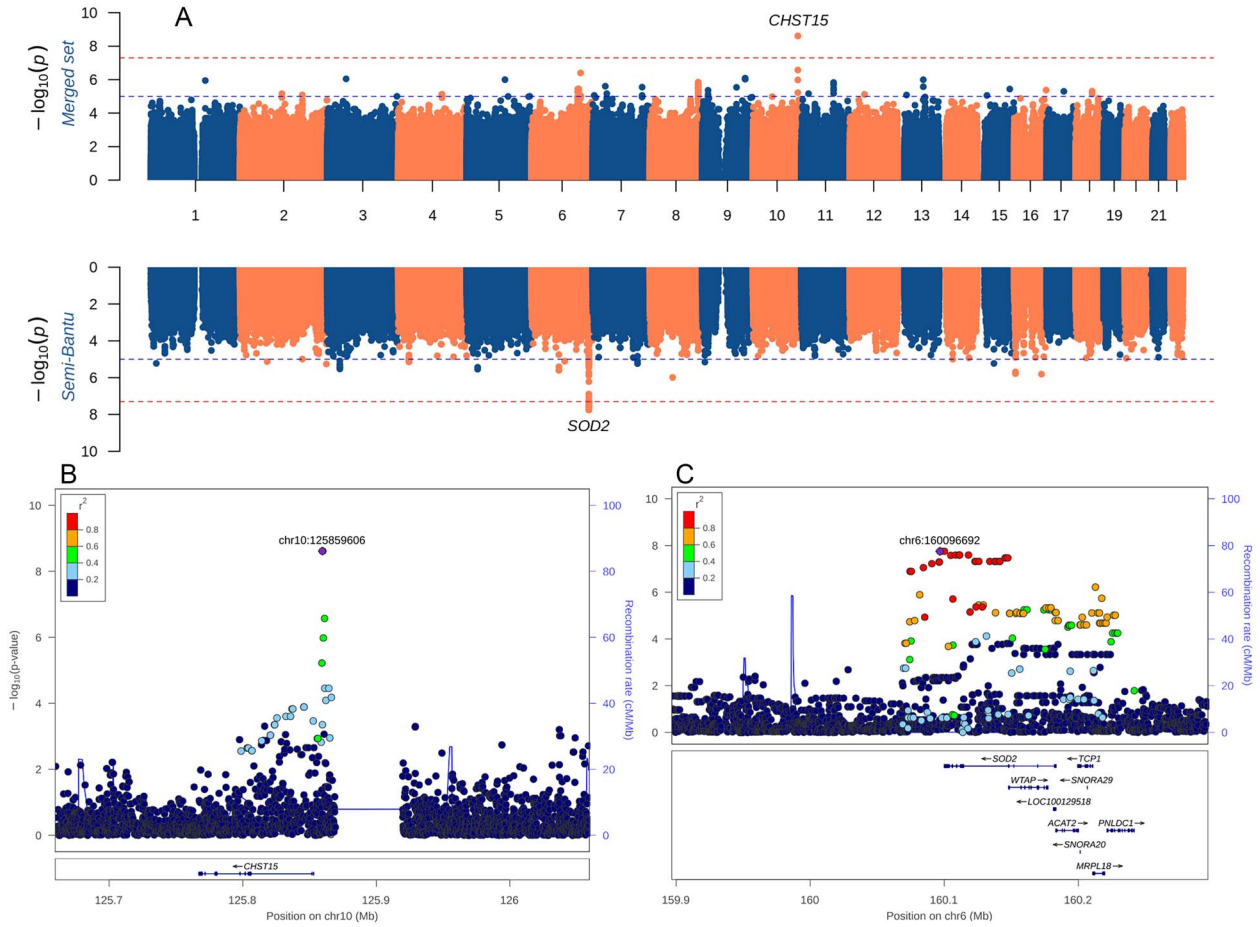

Figure 3. Manhattan plot of significant association signals. (**A**) Significant associations around *CHST15* in the merged set of BA, SB and FU individuals (top Manhattan plot) and in *SOD2* in SB individuals. Red line = genome-wide significance (5e – 08); blue line = suggestive line (1e – 05). (**B**) Regional association plot showing LD between the lead signal and other variants around *CHST15*. Moderate LD (0.4–0.6) is observed between the lead SNPs and other suggestive associations. (**C**) Regional association plot showing LD between the lead signal and other variants in *SOD2*. Strong LD extends over a 100 Kb region covering *SOD2* and other genes.

Notably, all the SNPs were imputed with high accuracy (average $r^2 = 0.97$) (Table 4). A total of 12 (12) haplotypes were generated across all the groups. The 'base' populations harboured a higher haplotypic diversity than the HbS-carrying populations as expected given the decreased need for haplotype conservation in the absence of HbS (Fig. 5). The BEN haplotype was the most prevalent in all population groups, and was more prevalent in the HbS-positive groups than in the base population groups consistent with its global distribution and association with favourable sickle cell disease outcomes (20). Interestingly, the CAM haplotype that is associated with unfavourable sickle cell disease outcomes was most common in the BA and FU, whilst the frequency was significantly lower in HbS-positive SB individuals. In addition, the Sem-Bantu had a higher prevalence of the SEN haplotype, and a lower prevalence of the CAR haplotype (Supplementary Material, Table S7).

Discussion

The study has uniquely characterized the genetic structure and malaria susceptibility of Cameroonians on the basis of available genome-wide SNP data that contributed to MalariaGEN studies and revealed the need for country-specific analysis with the description of novel significant loci. This study is likely to inform the design of subsequent genetic associations studies in other African Countries, as well as the prioritization of genes for targeted genotyping studies.

Importantly, our study has identified potentially novel malaria-associated loci in Cameroonians which, considering their biological role, seem appealing as plausible associations deserving further investigation. Studies of gene expression under stringent conditions show that *CHST15* is only expressed in human and baboon DNA highlighting its importance in the two species (28). We have shown that variants in LD with the significant variants we identified in *CHST15* are strong eQTLs, specifically in the liver, and we have shown the possible mechanisms by which the variants induce their effect.

The current information on chromatin state in the HS2 region of *CHST15* is largely driven by data from over-represented populations, which are enriched for the derived alleles of the variants we analysed, and therefore not adequately informative of the state in populations carrying the ancestral alleles but that are under-represented in omics databases. Considering the disruptive nature of the derived alleles on important transcription factors, we propose that HS2 was a previously active enhancer of *CHST15* that has now been adaptively inactivated (such as would be expected following its role in the progression of certain tumours (43,44)), whilst preserved in relevant cell lines in malarious environments to meet the challenge of malaria by upregulating the expression of *CHST15*.

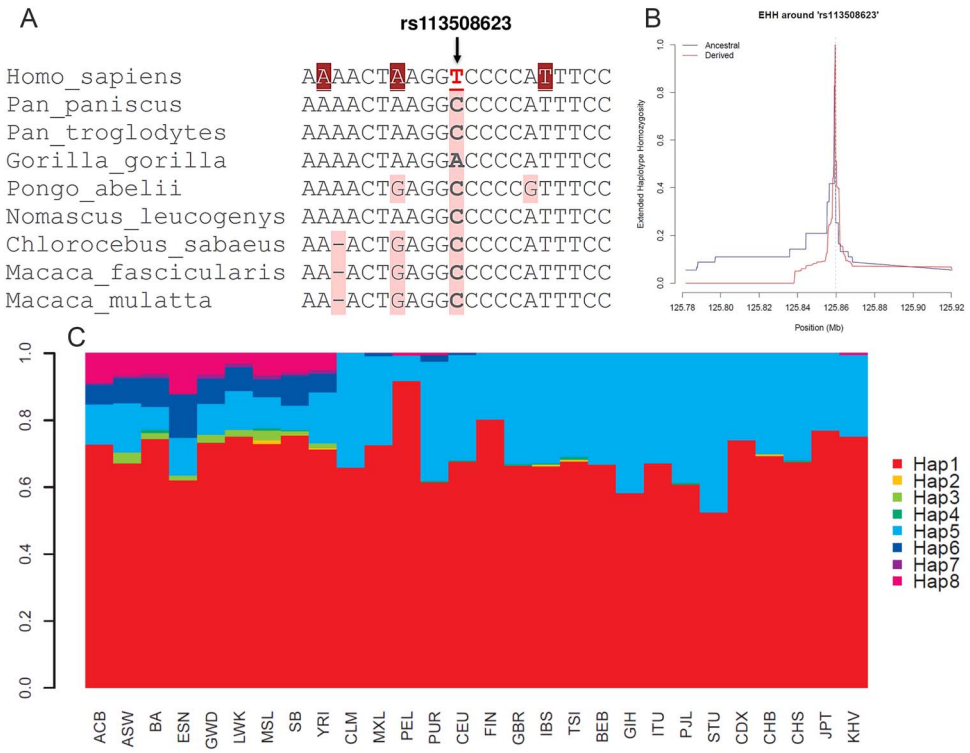

Figure 4. Genetic Diversity within and around CHST15. (**A**) Alignment of the rs113508623 significant variant with nine primate species from the ENSEMBL 10 EPO multiple sequence alignment. The bases highlighted in deep red represent regulatory region variants. The bases highlighted in light red represent positions with more than one allele across the different species. The underlined base shows the variant of interest, which is also a regulatory region variant. The variants occur in the enhancer region upstream of CHST15. (**B**) EHH plot showing haplotype decay around the rs113508623 significant variant. The ancestral allele (C) (blue line) shows an overall higher EHH score as compared with the derived allele (T) (red line). (**C**) Haplotype frequencies in the enhancer region upstream of CHST15 in global populations. The colours were generated using the rainbow colour function in R and numbered from bottom to top of the barplot, hence Hap1 is the bottom-most colour, whilst Hap8 is the top-most colour.

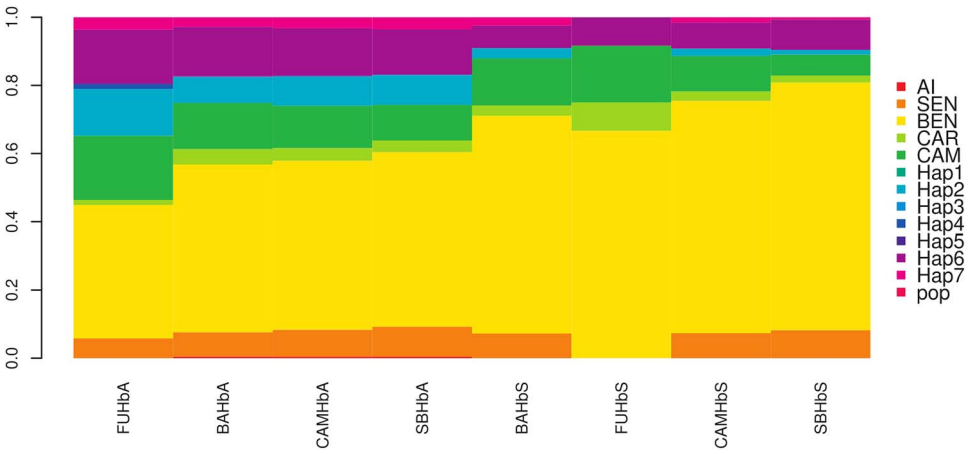

Figure 5. Haplotype frequencies in the *HBB* gene cluster in Cameroonians. CAM = merged set of all the ethnic groups. HbS = HbS-positive samples, HbA = HbS-negative samples. The colours were generated using the rainbow colour function in R and numbered from bottom to top of the barplot, hence AI is the bottom-most colour, whilst Hap7 is the top-most colour.

Moreover, polycomb repression is known to be an effective means of tissue–/cell type-specific gene expression regulation, and chromatin state data show that the HS2 is only polycomb-repressed in hepatic cell lines. This seems to explain why rs28657878 and rs4359147 are strong eQTLs in the liver although the liver was not initially identified as a site for high *CHST15* expression; the preservation of transcription factor binding sites and the ultimate tissue-/cell type-specific recruitment of the transcription machinery likely prevents polycomb repression in hepatic cell lines, leading to enhanced *CHST15* expression. Note that *Plasmodium* parasites spend much of the human phase of their life cycle in the liver, which would be consistent with a model of *CHST15* upregulation in the liver as a result of the parasite pressure.

Therefore, the question of whether *CHST15* is evolving under a recent balancing selection in malarious environments and what evolutionary forces may be driving the allele frequency distortions in human populations deserves further scrutiny, preferably in

whole-genome sequence data, and including a higher representation of populations in historical malaria endemic regions.

We also reviewed data that demonstrate strong support for the protective role of *SOD2* in Cameroonian malaria endemic populations, especially in the early stages of malaria infection. This should be important for future study designs in the country as most studies tend to focus heavily on extremes of phenotype such as SM, and might thus be missing signals that might only apparent or more strongly apparent during the early stages of infection.

Two key differences between our study and previous GWASs involving this same cohort afforded us the advantage to uncover the novel loci presented herein; the use of multiple reference imputation panels with varying representation of populations of African ancestry, and the stratification of our association analysis by ethnic group given the high genetic sub-structure within the country. The recent expansion and public availability of reference haplotype panels afforded us the opportunity to improve power for our analysis. The effect of this strategy proved to be very significant for our cohort given that significant associations in *CHST15* were only observed in data imputed with the AGR from the Sanger imputation panel. This however did not seem to be explained by the larger representation of indigenous African haplotypes (9912 haplotypes across 4956 individuals) in the Sanger panel as compared with other panels, given that the significant variants in *CHST15* were also imputed in the other panels with high accuracy and similar allele frequencies.

Generally, our imputation results confirmed the superiority of the TI panel over other panels for African populations (22). However, the higher performance of the Sanger panel at ultra-rare variants (MAF < 0.002) compared with the other panels makes it more suitable for rare variant analysis in African populations. The higher performance of our in-house imputation strategy over the Michigan (and Sanger at low frequency SNPs) demonstrates the advantage of tuning the imputation parameters in a study-specific manner to improve accuracy, and, in some cases, this might just be the difference between identifying significant associations or not.

The differences in allele frequency observed between the SB and BA at key malaria-associated loci may have no effect at all at highly frequent SNPs. However, this can lead to a loss of power at lower frequency sites as they may appear as rare variants when the ethnic groups are pooled together, especially in highly structured populations such as in the current data. These rare variants risk being filtered out during quality control (QC) procedures. The low imputation accuracy observed at key malaria loci in regions of ancient balancing selection reflect the need to increase the diversity of existing reference panels with populations that are closer genetically to the study population. The lower frequency of HbS and HbC in the SB is consistent with a previous report that found the frequency of HbS (8.5%,) in the SB to be significantly lower than in other ethnic groups in Cameroon (45).

We did not observe the well-characterized HbS (rs334) association in our analysis most likely caused by the exclusion of the variant in our post-imputation filtering. Subsequent Cameroon-specific GWASs that directly type the variant may be profitable. We, however, observed substantial differences in the haplotypic background of the major ethnic groups in Cameroon in the *HBB* gene cluster, and these are most likely the result of differential malaria pressures on the ethnic groups. The substantially low haplotype diversity in the HbS-positive backgrounds confirms longstanding knowledge that malaria has shaped the human genome for much of the last 5000 years (46). The haplotype frequencies we observed are also largely expected; the BEN is the most prevalent worldwide, the AI was only recently reported in Cameroon (41), and has only been reported again in Egypt (47) and Mauritania (48), and predicted in a single chromosome in Kenya (49). The lower prevalence of the unfavourable CAM and CAR haplotypes and higher prevalence of the favourable BEN and SEN haplotypes in the SB suggests that the SB have carried HbS for a longer period of time allowing for enough time to purge unfavourable haplotypes whilst preserving the more favourable ones. In addition, the presence of Hap3 in the SB only and Hap5 in the BA only further supports different extents of evolutionary pressure on the ethnic groups, consistent with age-old fine-scale population structure in Cameroon. Indeed, our previous results showed that the BA harboured strong signatures of positive selection at specific loci in the HLA region (*HLA-DPB1*) not present in the SB supporting differential evolutionary forces. A similar finding was previously reported in BA-speaking populations of West-Central Africa indicating a differential evolutionary signature under disease pressures as compared with other ethnic populations (50).

It is now widely accepted that the HbS mutation emerged once in the haplotype background of the ancestors of agriculturalists populations earlier than 20 000 years ago (ya) (20) and only relatively recently (∼3800 ya) acquired by rain forest hunter gatherer (RHG) populations via adaptive gene flow in the last 6000 years (46). As the variant frequency became augmented in malarious areas because of its protective role against SM, it saw an increase in the Sickle Cell Disease (SCD) burden, which in tend mounted pressure on HbS-carrying genomes. Thus, the *HBB* haplotypes are a result of these two co-evolutionary forces. Whilst the SEN, AI and BEN haplotypes are associated with favourable SCD clinical outcomes, the CAM and CAR haplotypes associate with poorer outcomes (41). The BA ethnic group in our analysis harboured a lower prevalence of the SEN and BEN haplotypes, twice as much of the CAM haplotype and a higher haplotype diversity in their HbS-positive haplotype background than the SB. This may reflect a relatively recent acquisition of HbS in the BA population, such that there has not been as much time to select the favourable haplotypes and attain more haplotype conservation as in the SB. This would also seem to explain the absence of association of HbS with SM in the BA as previously reported in this same cohort (37). Taking only the BA of the Central region in Cameroon, the previously described assumption seems likely as they are closer to the RHG BA populations of the South of Cameroon. Indeed, our previous analysis showed a substantial sub-structure within the BA ethnic group, which may indicate that the HbS mutation is indeed younger in some BA tribes of Cameroon.

Although the discussion of the specific prevalence and importance of the typical haplotypes in our population falls out of the scope of the current analysis, we note that they can be particular applicable to SCD research. For instance, the presence of the AI haplotype in all our 'base' populations and its absence from HbS-positive populations is particularly interesting given that the haplotype is associated with the most favourable SCD clinical outcomes and would be expected to be prevalent in HbS-positive chromosomes.

An important limitation of our study was the unavailability of the age information of the participants. Therefore, further studies on a larger sample size and more complete participant information are imperative. Also, it is crucial to increase the representation of the FU ethnic group in such studies as they have been shown to be particularly resistant to malaria in other African countries. Other possible limitation is the absence of a replication

cohort. Nevertheless, the Cameroonian population in this study has already been meta-analysed with other African populations in MalariaGEN studies, and no effect in the loci in multiple populations was detected. It is likely that these signals we report are population specific, and will require further sampling of a larger cohort of Cameroonians, and investigation is other unstudied populations. Owing to the huge variability and complex genetic architecture found in African populations, our result support that in addition to metanalysis, research should consistently and exhaustively analysis separately available data from individual African country, population and region. Taking collectively, our findings should further the specific understanding of the contribution of genomes of individual living in Cameroon under malaria evolutionary pressure, but also stimulate further questions for future research.

Our findings revealed novel insights in the evolutionary genetics of populations living in Cameroon under malaria pressure with novel significant protective loci (*CHST15* and *SOD2*), and emphasized the significant attenuation of genetic association signals by fine-scale genetic structure. The present study could inform the design of similar investigations in other African countries.

Materials and Methods

Ethical statement

This study was conducted in adherence to the principles of the Helsinki Declaration. Ethical and administrative clearance were obtained from the University of Buea Institutional Review Board and the South West Regional Delegation of Public Health, respectively. Authorization to conduct the surveys in primary schools was obtained from the Regional Delegation of Basic Education or the Catholic Education Secretariat of the South West Region. Only individuals who fulfilled inclusion criteria and volunteered to participate after adequate sensitization on the project objectives, methods and possible benefits/risks were enrolled into the study. A health facility or school was only investigated with the approval of its Director or Head Teacher and study participants were only enrolled if they or their caregivers/guardians gave written informed consent/assent.

Sampling and case definition

The characteristics of the study participants are extensively described in Table 1 of MalariaGEN CP1 in (21), and in (13,25,37). Briefly, cases consisted unrelated children with SM or uncomplicated malaria (UM). SM was defined according to the World Health Organization guidelines (51) as follows; presence of asexual *Plasmodium* parasitemia and at least one of the following conditions: cerebral malaria [impaired consciousness or unrousable coma (Blantyre coma score ≤ 2) and no record of recent severe head trauma, neurological disease or any other cause of coma]; severe malaria anaemia [haemoglobin < 5 g/dl or haematocrit $< 15\%$, no cases of severe bleeding or observed convulsions]; hyperpyrexia (axillary temperature ≥ 40°C); hyperparasite (>250 000 parasites/μl); convulsions before/during admission; respiratory distress (presence of alar flaring, intercostals or subcostal chest recession, use of accessory muscles of respiration, or abnormally deep respiration) and hypoglycaemia (blood glucose < 2.2 mmol/l/40 mg/dl). UM was defined by; axillary temperature ≥ 37.5°C associated with a *Plasmodium* positive blood film, haemoglobin ≥ 8 g/dl and full consciousness, in the absence of clinical signs and symptoms of SM and/or evidence of vital organ dysfunction. Controls were apparently healthy (afebrile) children (aged 1–14 years) and asymptomatic adults (aged 17–52 years) of the BA and SB ethnic groups. All cases were recruited from healthcare centres or hospital paediatric wards in the South West, Littorale and Centre regions of Cameroon. Children were recruited during malaria cross-sectional surveys from primary schools located in the South West region (Buea Metropolis) between 2004–2005 and 2007–2008. Adults were identified from a blood bank in the Centre region (Mother and Child Hospital—Yaounde) between July and August 2007.

Genotyping and data QC

Genotyping was performed on the Illumina Omni2.5M array at the MalariaGEN Oxford Resource Centre and alignment was performed against the human reference genome in GRCh37 (build 37) coordinates. Genotype calling was performed according to the MalariaGEN three-way genotype calling algorithm (10). Genotype data was retrieved from MalariaGEN site using secure file transfer protocols (sftp) on approval by site principal investors and according to MalariaGEN data access policies. The data consisted single chromosome Variant Call Format (VCF) files alongside a sample file with case-control information. Sample and SNPs process report files were also retrieved for initial quality information. The data consisted of ~2.3 million SNPs and 1471 samples (693 cases, 778 controls).

Standard GWAS quality control (52) was applied including; removal of all non-Cameroonian individuals and individuals with unreported ethnicity. Individuals with inconsistent sex information (discordance between self-reported and genotyped sex or sex information obtained by computing X chromosome inbreeding coefficient—F_{IS}) were identified and removed using Plink1.9's —*check-sex* function (53). The X chromosome pseudo-autosomal regions (PARs) were split prior to sex check. One individual from each pair of related individuals [2nd, FS, PO, MZ (Supplementary Fig. S1)] was excluded by computing an identity by descent report using the KING v2.2.4 software (54). Individuals with outlying heterozygosity (out of the range 0.180–0.230) and individuals with >10% missing genotype count were excluded (Supplementary Material, Fig. S2a and b).

Ancestral and population outliers were checked by running *smartpca* of the EIGENSOFT package (55) with 10 outlier removal iterations whilst projecting the dataset against the African populations from The 1000 Genomes Project (KGP) reference panel (phase 3 version 5) (56,57). A total of 368 unique individuals were removed following sample QC procedures leaving 1103 individuals. SNP QC involved removing SNPs with MAF $< 1\%$, genotype call rate $< 95\%$, SNPs with significant ($P < 1e-4$) differential call rate in cases and controls (differential missingness) that may indicate batch effect, and SNPs that failed the HWE test at $P < 1e-4$ in controls. A further differential missingness for the X chromosome was performed after merging the X chromosome PARs. A total of 1 863 254 SNPs of 2 261 351 were left following SNP QC.

Phasing and imputation

Prior to phasing, palindromic [A/T] and [C/G] SNPs (49468) were excluded. The remaining SNPs (1813786) were checked and validated against the KGP reference panel for strand, ref/alt SNP assignment, ID names, positions and alleles using the *conform-gt v24May2016* programme of the BEAGLE utils (58). Phasing and Imputation were performed using four strategies: an in-house imputation pipeline using EAGLE v2.4 (59) for phasing and IMPUTE2 (60) for imputation against the KGP, and public resources including the TOPMed, Michigan and Sanger Imputation services (TIS, MIS, SIS, respectively). In the TIS and MIS, EAGLE v2 was used for pre-phasing and Minimac v4 was used for imputation

(22–24). In the SIS, EAGLE v2 was used for pre-phasing and the PBWT algorithm was used for imputation (61,62). In our in-house phasing strategy, we set the number of conditioning haplotypes (—*Kpbwt*) to 50 000 (default 10 000) and the number of iterations (—*pbwtIters*) to 10 to improve phasing accuracy. For imputation, we used an effective population size (Ne) of 20 000, chunk sizes of 5 million with 1000 kb buffer size on each side of the chunks, whilst filtering out low frequency (MAF < 1%) variants in non-African populations in the KGP reference panel. Only biallelic SNPs and indels with imputation accuracy (R^2 or IMPUTE info score) ≥ 0.60, MAF ≥ 0.01, and genotype probability ≥95% were included in post-imputation association analyses.

Association analysis

Association analysis was performed using EMMAX (intel-binary-20120210) (63), BOLT-LMM v2.3.4 (64), GCTA v1.93.2 (65) and PLINK2 (53). First, 50 (50) genetic covariates (principal components—PCs) were computed on the non-imputed dataset using the PLINK2 —*pca* function. For BOLT-LMM, we obtained (∼500 000) SNPs in near linkage equilibrium for variance component estimation (*–modelSnps*) by LD-pruning using *–indep-pairwise* 50 5 0.2 in PLINK2. We then used the default *–lmm* function for association analysis. The same procedure was applied for the imputed dataset. For GCTA, we first generated a genetic relatedness matrix (*grm*). For the pre-imputation set, we directly ran the *–reml* function to estimate heritability, and the *–mlma* function for mixed linear model association analysis (whilst including the aforementioned PCs) at each step. For the imputed set, we first calculated segment-based LD scores, then stratified our SNPs on the basis of these scores, generated *grm*s, then performed *–reml* and *–mlma* using these *grm*s and PCs. Using EMMAX, we generated a covariates file including an intercept with all 1's and the PCs. A kinship matrix was also generated, whilst all other default parameters were used for association analysis. PLINK2 *–glm* function was then used to test different models (additive, genotypic, hethom, dominant and recessive), whilst adjusting for the PCs. Sex was included as a covariate in all association analyses. All association analyses were performed for the merged set of all samples and for the BA and SB populations separately. Association analysis could not be performed with the FU population only because of low sample size ($n < 30$).

Regional association

Regional association plots were generated for the lead SNPs in our analysis using the LocusZoom v1.4 command line package (66). When the lead SNP was found in the TOPMed panel, the human reference genome in build 38 (hg38) coordinates was used. Otherwise, the build 37 (hg19) was used. For LD calculations, we extracted a 1 Mb region around each lead SNP from the VCF file of the corresponding chromosome.

Functional analysis

Functional analysis of the significant variants observed was performed using various web-based tools. First we computed pairwise LD statistics for 1 Mb regions around the significant variants of *CHST15* and *SOD2* and we investigated the lead SNPs as well as all SNPs in LD with the lead SNPs for quantitative trait loci. Gene expression data, cis-eQTL and sQTL were determined using the GTEx Protal (https://gtexportal.org, accessed August 11, 2022). Chromatin state information was determined using the ChiP-Seq peak data, GeneHancer prediction of enhancers (67), JASPAR prediction of transcription factor binding sites (https://jaspar.uio.no/; accessed August 11, 2022), as well as ChromHMM prediction of chromatin state (68). These were visualized using the UCSC Genome Browser (https://genome.ucsc.edu; accessed August 11, 2022).

HBB gene cluster haplotype analysis

From our in-house imputed dataset, we identified HbS-carrying chromosomes (all samples that were homozygous or heterozygous for HbS) using the bcftools v1.9 *view* and *query* commands (69), whilst samples that were negative for HbS (hereafter referred to as 'base population') were extracted into another file. We then generated haplotypes on the basis of four (4) SNPs (rs3834466, rs28440105, rs10128556, rs968857) corresponding to different RFLPs that have previously been shown to accurately predict the classical *HBB* gene cluster haplotypes (SEN, BEN, CAM, CAR and AI) (42). Haplotypes were generated for samples carrying HbS for each ethnic group and the pooled dataset ($N_{Bantu} = 83$, $N_{Semi\text{-}Bantu} = 73$, $N_{Fulani} = 6$, $N_{Pooled} = 162$), as well as for the base population of each ethnic group and the pooled dataset ($N_{Bantu} = 407$, $N_{Semi\text{-}Bantu} = 466$, $N_{Fulani} = 69$, $N_{Pooled} = 942$). Haplotype frequencies were then computed using the haplotype analysis pipeline of a nextflow workflow available at https://github.com/esohkevin/mutationAge (accessed July 14, 2022).

Genetic diversity in the *CHST15* gene

We extracted a 2 Mb region around the *CHST15* gene (10q26.13) (hence 125 767 184 – 1 000 000 to 125 853 114 + 1 000 000 in build 37 coordinates) from our data imputed on the Sanger imputation server into a separate VCF file. We also extracted the same region from the KGP reference panel and merged with our data using the bcftools *isec* and *merge* commands so that only SNPs that were present in the two files were merged. A total of 50 samples from each of the populations in the KGP and from our data were randomly selected and extracted into separate VCF files for the calculation of genetic diversity statistics. We computed the Tajima's D statistic using vcftools (70) with a window size of 100 bp, and LD using PLINK1.9. We also computed the EHH score for each population using the REHH v3 R package (71). Allele frequency information was obtained from dbSNP and the ENSEMBL browser. Additional LD statistics around significant variants in *CHST15* in global populations were obtained from the ENSEMBL browser.

Web resources

MalariaGEN Consortial Project 1 (Cameroon): https://www.malariagen.net/about/where-we-work/human-genetic-determinants-severe-malaria-three-regions-cameroon

Codes used in this analysis can be found at: https://github.com/esohkevin/CamGWAS

Supplementary Material

Supplementary Material is available at *HMG* online.

Acknowledgements

We thank the participants and healthcare workers from the communities without whose sacrifice this work would not have been realized. We thank the MalariaGEN Oxford Resource Centre for their valuable work to avail the data, as well as the DELGEME and

H3ABioNet organizations for the training initiatives that led to the data analysis.

Conflict of Interest statement. None declared.

Data Availability

Data sources are available at the MalariaGEN Wed resources provided. Summary statistics and additional data from the analysis are available from the authors, upon reasonable request.

Funding

This research was funded by MIMPAC and MIM/TDR grant (A11034 to E.A.A.), and National Institute of Health (NIH), National Heart Lung and Blood Institute grant (5U24HL135600-04 to A.W). The views expressed in this publication are those of the author(s).

The authors acknowledge the Centre for High Performance Computing (CHPC), South Africa, for providing computational resources to this research project.

Computations were also performed using facilities provided by the University of Cape Town's ICTS High Performance Computing team: hpc.uct.ac.za.

References

1. Malaria Genomic Epidemiology Network (2019) Insights into malaria susceptibility using genome-wide data on 17,000 individuals from Africa, Asia and Oceania. *Nat. Commun.*, **10**, 5732.
2. Damena, D. and Chimusa, E.R. (2020) Genome-wide heritability analysis of severe malaria resistance reveals evidence of polygenic inheritance. *Hum. Mol. Genet.*, **29**, 168–176.
3. Mackinnon, M.J., Mwangi, T.W., Snow, R.W., Marsh, K. and Williams, T.N. (2005) Heritability of malaria in Africa. *PLoS Med.*, **2**, 1253–1259.
4. Kariuki, S.N. and Williams, T.N. (2020) Human genetics and malaria resistance. *Hum. Genet.*, **139**, 801–811.
5. Malaria Genomic Epidemiology Network (2008) A global network for investigating the genomic epidemiology of malaria. *Nature*, **456**, 732–737.
6. Band, G., Le, Q.S., Jostins, L., Pirinen, M., Kivinen, K., Jallow, M., Sisay-Joof, F., Bojang, K., Pinder, M., Sirugo, G. *et al.* (2013) Imputation-based meta-analysis of severe malaria in three African populations. *PLoS Genet.*, **9**, e1003509.
7. Leffler, E.M., Band, G., Busby, G.B., Kivinen, K., Si Le, Q., Clarke, G.M., Bojang, K.A., Conway, D.J., Jallow, M., Sisay-Joof, F. *et al.* (2017) Resistance to malaria through structural variation of red blood cell invasion receptors. *Science*, **356**, eaam6393.
8. Clarke, G.M., Rockett, K., Kivinen, K., Hubbart, C., Jeffreys, A.E., Rowlands, K., Jallow, M., Conway, D.J., Bojang, K.A., Pinder, M. *et al.* (2017) Characterisation of the opposing effects of G6PD deficiency on cerebral malaria and severe malarial anaemia. *elife*, **6**, e15085.
9. Ndila, C.M., Uyoga, S., Macharia, A.W., Nyutu, G., Peshu, N., Ojal, J., Shebe, M., Awuondo, K.O., Mturi, N., Tsofa, B. *et al.* (2018) Human candidate gene polymorphisms and risk of severe malaria in children in Kilifi, Kenya: a case-control association study. *Lancet Haematol.*, **5**, e333–e345.
10. Band, G., Rockett, K.A., Spencer, C.C.A. and Kwiatkowski, D.P. (2015) A novel locus of resistance to severe malaria in a region of ancient balancing selection. *Nature*, **526**, 253–257.
11. Malaria Genomic Epidemiology Network (2014) Reappraisal of known malaria resistance loci in a large multicenter study. *Nat. Genet.*, **46**, 1197–1204.
12. Modiano, D., Luoni, G., Sirima, B.S., Simporé, J., Verra, F., Konaté, A., Rastrelli, E., Olivieri, A., Calissano, C., Paganotti, G.M. *et al.* (2001) Haemoglobin C protects against clinical plasmodium falciparum malaria. *Nature*, **414**, 305–308.
13. Apinjoh, T.O., Anchang-Kimbi, J.K., Njua-Yafi, C., Ngwai, A.N., Mugri, R.N., Clark, T.G., Rockett, K.A., Kwiatkowski, D.P., Achidi, E.A. and The MalariaGEN Consortium (2014) Association of candidate gene polymorphisms and TGF-beta/IL-10 levels with malaria in three regions of Cameroon: a case-control study. *Malar. J.*, **13**, 236.
14. Timmann, C., Thye, T., Vens, M., Evans, J., May, J., Ehmen, C., Sievertsen, J., Muntau, B., Ruge, G., Loag, W. *et al.* (2012) Genome-wide association study indicates two novel resistance loci for severe malaria. *Nature*, **489**, 443–446.
15. Ravenhall, M., Campino, S., Sepúlveda, N., Manjurano, A., Nadjm, B., Mtove, G., Wangai, H., Maxwell, C., Olomi, R., Reyburn, H. *et al.* (2018) Novel genetic polymorphisms associated with severe malaria and under selective pressure in north-eastern Tanzania. *PLoS Genet.*, **14**, e1007172.
16. Mendez, F.L., Krahn, T., Schrack, B., Krahn, A.M., Veeramah, K.R., Woerner, A.E., Fomine, F.L.M., Bradman, N., Thomas, M.G., Karafet, T.M. and Hammer, M.F. (2013) An African American paternal lineage adds an extremely ancient root to the human y chromosome phylogenetic tree. *Am. J. Hum. Genet.*, **92**, 454–459.
17. Lipson, M., Ribot, I., Mallick, S., Rohland, N., Olalde, I., Adamski, N., Broomandkhoshbacht, N., Lawson, A.M., López, S., Oppenheimer, J. *et al.* (2020) Ancient west African foragers in the context of African population history. *Nature*, **577**, 665–670.
18. Liu, W., Li, Y., Learn, G.H., Rudicell, R.S., Robertson, J.D., Keele, B.F., Ndjango, J.B.N., Sanz, C.M., Morgan, D.B., Locatelli, S. *et al.* (2010) Origin of the human malaria parasite plasmodium falciparum in gorillas. *Nature*, **467**, 420–425.
19. Otto, T.D., Gilabert, A., Crellen, T., Böhme, U., Arnathau, C., Sanders, M., Oyola, S.O., Okouga, A.P., Boundenga, L., Willaume, E. *et al.* (2018) Genomes of all known members of a plasmodium subgenus reveal paths to virulent human malaria. *Nat. Microbiol.*, **3**, 687–697.
20. Esoh, K. and Wonkam, A. (2021) Evolutionary history of sickle cell mutation: implications for global genetic medicine. *Hum. Mol. Genet.*, **30**, R119–R128.
21. Apinjoh, T.O., Anchang-Kimbi, J.K., Njua-Yafi, C., Mugri, R.N., Ngwai, A.N., Rockett, K.A., Mbunwe, E., Besingi, R.N., Clark, T.G., Kwiatkowski, D.P., Achidi, E.A. (2013) MalariaGEN Consortium. Association of cytokine and Toll-like receptor gene polymorphisms with severe malaria in three regions of Cameroon. *PLoS One*, **8**, e81071.
22. Taliun, D., Harris, D., Kessler, M., Carlson, J., Szpiech, Z., Torres, R., Taliun, S., Corvelo, A., Gogarten, S., Kang, H.M. *et al.* (2021) Sequencing of 53,831 diverse genomes from the NHLBI TOPMed program. *Nature*, **590**, 290–299.
23. Fuchsberger, C., Abecasis, G.R. and Hinds, D.A. (2015) Minimac2: faster genotype imputation. *Bioinformatics*, **31**, 782–784.
24. Das, S., Forer, L., Schönherr, S., Sidore, C., Locke, A.E., Kwong, A., Vrieze, S.I., Chew, E.Y., Levy, S., McGue, M. *et al.* (2016) Next-generation genotype imputation service and methods. *Nat. Genet.*, **48**, 1284–1287.
25. Achidi, E.A., Apinjoh, T.O., Anchang-Kimbi, J.K., Mugri, R.N., Ngwai, A.N. and Yafi, C.N. (2012) Severe and uncomplicated falciparum malaria in children from three regions and three ethnic groups in Cameroon: prospective study. *Malar. J.*, **11**, 215.
26. Esoh, K.K., Apinjoh, T.O., Nyanjom, S.G., Wonkam, A., Chimusa, E.R., Amenga-Etego, L., Amambua-Ngwa, A. and Achidi, E.A. (2021) Fine scale human genetic structure in three regions of

Cameroon reveals episodic diversifying selection. *Sci. Rep.*, **11**, 1039.
27. Chaichoompu, K., Abegaz, F., Cavadas, B., Fernandes, V., Müller-Myhsok, B., Pereira, L. and Van Steen, K. (2020) A different view on fine-scale population structure in western African populations. *Hum. Genet.*, **139**, 45–59.
28. Verkoczy, L.K., Marsden, P.A. and Berinstein, N.L. (1998) HBRAG, a novel B cell lineage cDNA encoding a type II transmembrane glycoprotein potentially involved in the regulation of recombination activating gene 1 (RAG1). *Eur. J. Immunol.*, **28**, 2839–2853.
29. Verkoczy, L.K., Guinn, B.A. and Berinstein, N.L. (2000) Characterization of the human B cell RAG-associate gene, hBRAG, as a B, cell receptor signal-enhancing glycoprotein dimer that associates with phosphorylated proteins in resting B cells. *J. Biol. Chem.*, **275**, 20967–20979.
30. Ohtake, S., Ito, Y., Fukuta, M. and Habuchi, O. (2001) Human N-Acetylgalactosamine 4-sulfate 6-O-sulfotransferase cDNA is related to human B cell recombination activating gene-associated gene. *J. Biol. Chem.*, **276**, 43894–43900.
31. Salanti, A., Dahlbäck, M., Turner, L., Nielsen, M.A., Barfod, L., Magistrado, P., Jensen, A.T.R., Lavstsen, T., Ofori, M.F., Marsh, K., Hviid, L. and Theander, T.G. (2004) Evidence for the involvement of VAR2CSA in pregnancy-associated malaria. *J. Exp. Med.*, **200**, 1197–1203.
32. Fried, M. and Duffy, P.E. (1996) Adherence of plasmodium falciparum to chondroitin sulfate a in the human placenta. *Science*, **272**, 1502–1504.
33. WHO (2019) *World Malaria Report*. https://www.who.int/malaria. *World Malaria Report*. https://www.who.int/malaria; (2019).
34. Yu, W., Misulovin, Z., Suh, H., Hardy, R.R., Jankovic, M., Yannoutsos, N. and Nussenzweig, M.C. (1999) Coordinate regulation of RAG1 and RAG2 by cell type-specific DNA elements 5′ of RAG2. *Science*, **285**, 1080–1084.
35. Herrero, J., Muffato, M., Beal, K., Fitzgerald, S., Gordon, L., Pignatelli, M., Vilella, A.J., Searle, S.M.J., Amode, R., Brent, S. *et al.* (2016) Ensembl comparative genomics resources. *Database*, **2016**, bav096.
36. Kavishe, R.A., Koenderink, J.B. and Alifrangis, M. (2017) Oxidative stress in malaria and artemisinin combination therapy: pros and cons. *FEBS J.*, **284**, 2579–2591.
37. Apinjoh, T.O., Anchang-Kimbi, J.K., Njua-Yafi, C., Mugri, R.N., Ngwai, A.N., Rockett, K.A., Mbunwe, E., Besingi, R.N., Clark, T.G., Kwiatkowski, D.P., Achidi, E.A. and The MalariaGEN Consortium (2013) Association of cytokine and toll-like receptor gene polymorphisms with severe malaria in three regions of Cameroon. *PLoS One*, **8**, e81071.
38. Ockenhouse, C.F., Hu, W., Kester, K.E., Cummings, J.F., Stewart, A., Heppner, D.G., Jedlicka, A.E., Scott, A.L., Wolfe, N.D., Vahey, M. and Burke, D.S. (2006) Common and divergent immune response Signaling pathways discovered in peripheral blood mononuclear cell gene expression patterns in Presymptomatic and clinically apparent malaria. *Infect. Immun.*, **74**, 5561–5573.
39. Franklin, B.S., Parroche, P., Ataíde, M.A., Lauw, F., Ropert, C., de Oliveira, R.B., Pereira, D., Tada, M.S., Nogueira, P., da Silva, L.H.P. *et al.* (2009) Malaria primes the innate immune response due to interferon-γ induced enhancement of toll-like receptor expression and function. *Proc. Natl. Acad. Sci.*, **106**, 5789–5794.
40. Levings, P.P. and Bungert, J. (2002) The human β-globin locus control region. A center of attraction. *Eur. J. Biochem.*, **269**, 1589–1599.
41. Bitoungui, V.J.N., Pule, G.D., Hanchard, N., Ngogang, J. and Wonkam, A. (2015) Beta-globin gene haplotypes among cameroonians and review of the global distribution: is there a case for a single sickle mutation origin in Africa? *Omics. J. Integr. Biol.*, **19**, 171–179.
42. Shaikho, E.M., Farrell, J.J., Alsultan, A., Qutub, H., Al-Ali, A.K., Figueiredo, M.S., Chui, D.H.K., Farrer, L.A., Murphy, G.J., Mostoslavsky, G. *et al.* (2017) A phased SNP-based classification of sickle cell anemia HBB haplotypes. *BMC Genomics*, **18**, 608.
43. Wang, X., Cheng, G., Zhang, T., Deng, L., Xu, K., Xu, X., Wang, W., Zhou, Z., Feng, Q., Chen, D., Bi, N. and Wang, L. (2020) CHST15 promotes the proliferation of TE-1 cells via multiple pathways in esophageal cancer. *Oncol. Rep.*, **43**, 75–86.
44. Chen, Y., Zhang, Y., Wang, Z., Wang, Y., Luo, Y., Sun, N., Zheng, S., Yan, W., Xiao, X., Liu, S. *et al.* (2022) CHST15 gene germline mutation is associated with the development of familial myeloproliferative neoplasms and higher transformation risk. *Cell Death Dis.*, **13**, 1–11.
45. Engle-Stone, R., Williams, T.N., Nankap, M., Ndjebayi, A., Gimou, M.-M., Oyono, Y., Tarini, A., Brown, K.H. and Green, R. (2017) Prevalence of inherited Hemoglobin disorders and relationships with anemia and micronutrient status among children in Yaoundé and Douala, Cameroon. *Nutrients*, **9**, 693.
46. Laval, G., Peyrégne, S., Zidane, N., Harmant, C., Renaud, F., Patin, E., Prugnolle, F. and Quintana-Murci, L. (2019) Recent adaptive acquisition by African rainforest hunter-gatherers of the late Pleistocene sickle-cell mutation suggests past differences in malaria exposure. *Am. J. Hum. Genet.*, **104**, 553–561.
47. Abou-Elew, H.H., Youssry, I., Hefny, S., Hashem, R.H., Fouad, N. and Zayed, R.A. (2018) β S globin gene haplotype and the stroke risk among Egyptian children with sickle cell disease. *Hematology*, **23**, 362–367.
48. Veten, F.M., Abdelhamid, I.O., Meiloud, G.M., Ghaber, S.M., Salem, M.L., Abbes, S. and Houmeida, A.O. (2012) Hb S [β6(A3)Glu→Val, GAG\textgreaterGTG] and β-globin gene cluster haplotype distribution in Mauritania. *Hemoglobin*, **36**, 311–315.
49. Shriner, D. and Rotimi, C.N. (2018) Whole-genome-sequence-based haplotypes reveal single origin of the sickle allele during the Holocene wet phase. *Am. J. Hum. Genet.*, **102**, 547–556.
50. Patin, E., Lopez, M., Grollemund, R., Verdu, P., Harmant, C., Quach, H., Laval, G., Perry, G.H., Barreiro, L.B., Froment, A. *et al.* (2017) Dispersals and genetic adaptation of bantu-speaking populations in Africa and North America. *Science*, **356**, 543–546.
51. WHO (2014, 2014) Severe malaria. *Severe Malaria*, **19**, 7–131.
52. Anderson, C.A., Pettersson, F.H., Clarke, G.M., Cardon, L.R., Morris, A.P. and Zondervan, K.T. (2010) Data quality control in genetic case-control association studies. *Nat. Protoc.*, **5**, 1564–1573.
53. Chang, C.C., Chow, C.C., Tellier, L.C., Vattikuti, S., Purcell, S.M. and Lee, J.J. (2015) Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience*, **4**, 7.
54. Manichaikul, A., Mychaleckyj, J.C., Rich, S.S., Daly, K., Sale, M. and Chen, W.M. (2010) Robust relationship inference in genome-wide association studies. *Bioinformatics*, **26**, 2867–2873.
55. Patterson, N., Price, A.L. and Reich, D. (2006) Population structure and eigenanalysis. *PLoS Genet.*, **2**, 2074–2093.
56. Gibbs, R.A., Boerwinkle, E., Doddapaneni, H., Han, Y., Korchina, V., Kovar, C., Lee, S., Muzny, D., Reid, J.G., Zhu, Y. *et al.* (2015) A global reference for human genetic variation. *Nature*, **526**, 68–74.
57. Sudmant, P.H., Rausch, T., Gardner, E.J., Handsaker, R.E., Abyzov, A., Huddleston, J., Zhang, Y., Ye, K., Jun, G., Hsi-Yang Fritz, M. *et al.* (2015) An integrated map of structural variation in 2,504 human genomes. *Nature*, **526**, 75–81.
58. Browning, B. (2016) Conform-gt. *Conform-gt*. 2016.
59. Loh, P.R., Palamara, P.F. and Price, A.L. (2016) Fast and accurate long-range phasing in a UK biobank cohort. *Nat. Genet.*, **48**, 811–816.

60. Howie, B.N., Donnelly, P. and Marchini, J. (2009) A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet.*, **5**, e1000529.
61. McCarthy, S., Das, S., Kretzschmar, W., Delaneau, O., Wood, A.R., Teumer, A., Kang, H.M., Fuchsberger, C., Danecek, P., Sharp, K. *et al.* (2016) A reference panel of 64,976 haplotypes for genotype imputation. *Nat. Genet.*, **48**, 1279–1283.
62. Durbin, R. (2014) Efficient haplotype matching and storage using the positional burrows–wheeler transform (PBWT). *Bioinformatics*, **30**, 1266–1272.
63. Kang, H.M., Sul, J.H., Service, S.K., Zaitlen, N.A., Kong, S.-Y., Freimer, N.B., Sabatti, C. and Eskin, E. (2010) Variance component model to account for sample structure in genome-wide association studies. *Nat. Genet.*, **42**, 348–354.
64. Loh, P.R., Tucker, G., Bulik-Sullivan, B.K., Vilhjálmsson, B.J., Finucane, H.K., Salem, R.M., Chasman, D.I., Ridker, P.M., Neale, B.M., Berger, B., Patterson, N. and Price, A.L. (2015) Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nat. Genet.*, **47**, 284–290.
65. Yang, J., Lee, S.H., Goddard, M.E. and Visscher, P.M. (2011) GCTA: a tool for genome-wide complex trait analysis. *Am. J. Hum. Genet.*, **88**, 76–82.
66. Pruim, R.J., Welch, R.P., Sanna, S., Teslovich, T.M., Chines, P.S., Gliedt, T.P., Boehnke, M., Abecasis, G.R., Willer, C.J. LocusZoom: regional visualization of genome-wide association scan results. *Bioinformatics*, **26**, 2336–7.
67. Fishilevich, S., Nudel, R., Rappaport, N., Hadar, R., Plaschkes, I., Iny Stein, T., Rosen, N., Kohn, A., Twik, M., Safran, M. *et al.* (2017) GeneHancer: genome-wide integration of enhancers and target genes in GeneCards. *Database*, **2017**, bax028.
68. Ernst, J. and Kellis, M. (2017) Chromatin-state discovery and genome annotation with ChromHMM. *Nat. Protoc.*, **12**, 2478–2492.
69. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R. and 1000 Genome Project Data Processing Subgroup (2009) The sequence alignment/map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.
70. Danecek, P., Auton, A., Abecasis, G., Albers, C.A., Banks, E., DePristo, M.A., Handsaker, R.E., Lunter, G., Marth, G.T., Sherry, S.T. *et al.* (2011) The variant call format and VCFtools. *Bioinformatics*, **27**, 2156–2158.
71. Gautier, M., Klassmann, A. and Vitalis, R. (2017) Rehh 2.0: a reimplementation of the R package rehh to detect positive selection from haplotype structure. *Mol. Ecol. Resour.*, **17**, 78–90.