GENETIC DIVERSITY AND MARKERS OF SYMPTOMATIC MALARIA SUSCEPTIBILITY IN THREE MALARIA-ENDEMIC REGIONS OF CAMEROON

KEVIN ESOH KUM

MASTER OF SCIENCE

(Bioinformatics and Molecular Biology)

JOMO KENYATTA UNIVERSITY OF AGRICULTURE AND TECHNOLOGY

2021

Genetic Diversity and Markers of Symptomatic Malaria Susceptibility in Three Malaria-Endemic Regions of Cameroon

Kevin Esoh Kum

A Thesis Submitted in Partial Fulfilment of the Requirements for the Degree of Master of Science in Bioinformatics and Molecular Biology of the Jomo Kenyatta University of Agriculture and Technology

2021

DECLARATION

This thesis is my original work and has not been presented for a degree in any other University.

Signature…….Date: ……

Kevin Esoh Kum

This thesis has been submitted for examination with our approval as supervisors.

Signature…….Date: ……

Dr. Steven Ger Nyanjom, PhD

JKUAT, Kenya

Signature…….Date: ……

Dr. Tobias Apinjoh, PhD

UB, Cameroon

Signature…….Date: ……

Prof. Ambroise Wonkam, MD, PhD

UCT, South Africa

Signature…….Date: ……

Dr. Emile Chimusa, PhD

UCT, South Africa

DEDICATION

This work is dedicated to my parents, Mrs Monica Beteh Kang and Mr Jacob Kang Esoh, and also to my siblings for their enormous support.

ACKNOWLEDGEMENT

I am most grateful to God almighty for the strength he gave me to go through this process of learning.

I am indebted to the participants and healthcare workers from the communities who contributed to the studies that led to the generation of the data I used in my research.

This research could not be possible without funding and as a result, I am immensely thankful to my funders, the Developing Excellence in Leadership and Genomics Training for Malaria Elimination in Sub-Saharan Africa (DELGEME). DELGEME was supported through the DELTAS Africa Initiative [DELGEME Grant 107740/Z/15/Z]. The DELTAS Africa Initiative is an independent funding scheme of the African Academy of Sciences (AAS)'s Alliance for Accelerating Excellence in Science in Africa (AESA) and supported by the New Partnership for Africa's Development Planning and Coordinating Agency (NEPAD Agency) with funding from the Wellcome Trust [DELGEME Grant 107740/Z/15/Z] and the UK government. Funding for bioinformatics skill development and data analysis was provided by the H3ABioNet project; I am therefore grateful to Professor Nicola Mulder for the opportunity to gain skills through an internship with her group at the University of Cape Town. The H3ABioNet is supported by the National Institutes of Health (NIH) Common Fund [Grant Number U41HG006941]. The views expressed in this research are mine and not necessarily those of AAS, NEPAD Agency, Wellcome Trust, NIH or the UK government.

I wish to thank the Jomo Kenyatta University of Agriculture and technology for hosting such a great Bioinformatics course and the College of Health Sciences for admitting me into the Department of Biochemistry which hosted this course. I am heartily indebted to all the staff of the Department of Biochemistry who played one role or another in seeing the success of this program. And also importantly, many thanks goes to all my colleagues (course mates) with whom I rode the storms of the MSc program.

I am equally grateful to my supervisors Dr Steven Nyanjom, Dr Apinjoh Tobias, Prof Ambroise Wonkam, and Dr Emile Chimusa for the time and effort they dedicated to

seeing the success of my project, and particularly Dr Steven Ger Nyanjom for his earnest and fervent support throughout my research.

During my studies at JKUAT, I worked closely with my co-funded colleagues Ms Catherine Bakari, Dr Desire Ehouni, Ms Elikplim Amegashie, Mr Deriba Abera, and Ms Jeniffer Mumbe. Their support could not have been more important.

I also wish to thank Dr Caleb Kibet for his mentorship and the Kenya Education Network (KENET) for their generosity to use their internet resources for my data transfer.

I thank the NSC library staff of JKUAT for their input in putting the final copy of this work in good shape.

Finally, I wish to thank my family and my girlfriend for their amazing support.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF APPENDICES

ABBREVIATIONS AND ACRONYMS

ACT: Artemisinin-based Combination Therapy

AF: Allele frequency

AIMs: Ancestry informative markers

AMA: Apical Membrane Protein

BA: Bantu

BF: Bayes Factor

CD: Cluster of Differentiation

CDC: Center for Disease Control and Prevention

CNV: Copy number variation

CSA: Chondroitin sulphate A

CSE: Chondroitin sulphate E

CSP: Circomsporozoite Protein

DBR: Duffy-binding Receptors

EBL: Erythrocyte-binding Ligand

EHH: Extended Haplotype Homozygosity

FO: Foulbe

FLK: Lewontin-Krakauer Fst outlier test

Fst: Wright's F-statistic

GDP: Gross Domestic Product

G6PD: Glucose-6-phosphate dehydrogenase

GPI: Glycosylphosphatidylinositol

GWAS: Genome Wide Association Study

HBB: Hemoglobin beta gene

HSPG: Heparan Sulphate Proteoglycans

HWE: Hardy-Weinberg Equilibrium

ICAM: Intercellular Cell Adhesion Molecule

iHS: Integrated Haplotype Score

INDELs: Insertions and deletions

LD: Linkage Disequilibrium

LLINs: Long-lasting insecticide nets

MalariaGEN: Malaria Genomic Epidemiology Network

MAF: Minor allele frequency

MRCA: Most recent common ancestor" (

MDS: Multidimensional Scaling

MSP: Merozoite Surface Protein

OR: Odds Ratio

PAM: Pregnancy-associated malaria

PC: Principal Component

PCA: Principal Component Analysis

***Pf*EMP**: *Plasmodium falciparum* Erythrocyte Membrane Protein

RON: Rhoptry Neck Protein

SB: Semi-Bantu

SNP: Single Nucleotide Polymorphism

SV: Structural variation

TGF: Transforming growth factor

TRAP: Thrombospondin-related adhesive proteins

VCF: Variant Calling Format

WHO: World Health Organization

ABSTRACT

Malaria remains a global public health problem as the current lines of treatment and control are becoming increasingly less effective or under threat of decreased efficacy. While much progress has been made towards malaria vaccine development, the enormous diversity of the parasite antigens, and redundancy in their interaction with humans pose a great challenge to the current vaccine candidates, particularly in regions with high malaria transmission. In Cameroon, malaria contributes significantly to mortality reported at healthcare facilities; a large proportion of which are children below 5 years old. Evidence for the involvement of host (human) genetic factors in increased or decreased malaria susceptibility has mounted for over the years. Large scale multi-site genome-wide association studies (GWAS) have now revealed multiple human genetic factors associated with malaria susceptibility. Yet, due to the high genetic diversity (population structure) among African populations, these methods have been limited in multi-site studies. In this study, a country-specific GWAS for malaria case and control participants from the South West, Littorale, and Centre regions in Cameroon was performed in order to uncover genetic variants that may not have been found by previous studies. First, population structure analysis was performed on 1073 samples to ascertaining whether there was significant genetic differentiation within the regions that could reduce the power of the GWAS. The study confirmed significant genetic diversity among Cameroon's major ethnic groups; Bantu, Semi-Bantu, and Fulani. Association analysis confirmed attenuation of GWAS signals due to this population structure. In addition, markers in potentially novel malaria protective loci were uncovered; *SOD2* specific to Cameroon's Semi-Bantu ethnic group and *CHST15* in Cameroon's Bantu and Semi-Bantu ethnic groups. Furthermore, heterogeneity within the beta-like globin (*HBB*) gene cluster was revealed among the Bantu and Semi-Bantu ethnic groups that underscores age old fine-scale structure within the country. The findings of this study highlight population-specific variants and disparate genetic association patterns among Cameroon's major ethnic groups that should be important for future genetic association studies in the Country. The findings also further the understanding of the evolutionary course of Cameroon's major ethnic populations under malaria pressure

CHAPTER ONE

INTRODUCTION

1.1. Background Information

Host genetic factors play a major role in malaria phenotypic variance, contributing up to 25% of the inter-individual differences of severe malaria (SM) expression (Mackinnon et al., 2005). Although the sickle-cell mutation (HbS) affords the strongest protection (in its heterozygous state [HbAS]) against severe malaria across sub-Saharan Africa (sSA), it only explains 2% of the total SM variance (Mackinnon et al., 2005). Current data further show that only a small fraction of malaria phenotypic variance is explained by the totality of genetic markers discovered so far. Hence, many more variants remain to be discovered (Malaria Genomic Epidemiology Network, 2019).

The establishment of the Malaria Genomic Epidemiology Network (MalariaGEN) in 2005 brought together researchers from multiple countries with the goal of leveraging large-scale genomic variation studies to uncover malaria-associated variants (Malaria Genomic Epidemiology Network, 2008b). By pooling human genetic data in targeted genotyping and consortium-based genome-wide association studies (GWASs), multiple resistance/susceptibility loci in African genomes were replicated (*HBB*, *ABO*, and *G6PD* for example), while others, novel, were uncovered (*ATP2B4*, *FREM3-GYP*, and *EPHA7* for example) (Malaria Genomic Epidemiology Network, 2019). Although a majority of these loci are shared across sSA (HbS, *G6PD* deficiency, *ABO*), others have evolved disparately and restricted to specific geographic regions; for instance, the HbC in West Africa with an epicenter around Burkina Faso, and the Dantu (*GYPA-B*) in East Africa (Kariuki et al., 2020).

Meanwhile, the two targeted genotyping studies conducted in Cameroon between 2013 and 2014 in the early stages of the MalariaGEN collaboration found polymorphisms in several loci associated with malaria resistance (*HBB*, *IL10*, *IL17RE*, *NOS2*, and *ADCY9*) and malaria susceptibility (*G6PD*, *IL17RD*, *EMR1*, and *RTN3*) in two major ethnic groups in the country; Bantu and Semi-Bantu (Apinjoh et al., 2013, 2014). These remain the only Cameroon-specific large-scale human genetic studies of

malaria reported to date. The contribution of Cameroonian genomes to malaria phenotypic variance has been further elucidated by MalariaGEN's multi-site studies (Malaria Genomic Epidemiology Network, 2019). These studies, though largely successful, have often met specific challenges dealing with ethnically diverse populations that prompted the authors to recommend country-specific analyses which have thus far been conducted with success in Tanzania (Ravenhall et al., 2018), Kenya (Ndila et al., 2018), the Gambia (Jallow et al., 2009), and Ghana (Timmann et al., 2012), uncovering novel loci in the populations. This remains to be performed in Cameroon.

A peculiar observation was made with Cameroonian participants in a MalariaGEN's 2019 study where, while the HbAS was found to be most strongly protective in The Gambia, its weakest effect was observed in Cameroon (Malaria Genomic Epidemiology Network, 2019). Ironically, the HbS gene is believed to have originated from around Cameroon and is still highly prevalent in Cameroon, and would have been expected to be most protective in this region (Esoh & Wonkam, 2021). Also, the HbC variant, which tends to 'compete' with the HbS, is almost completely absent in Cameroon (Malaria Genomic Epidemiology Network, 2019). In addition, variants in other known malaria protective loci (*FREM3, ATP2B4, GYPA-B*) have not been independently replicated in Cameroon.

Considering that Cameroonian ancestral populations are among the earliest to have walked the continent (Lipson et al., 2020), and considering data that suggest the virulent *P. falciparum* malaria and the HbS mutation may have been around Cameroon perhaps earlier than anywhere else (Otto et al., 2018), it is therefore likely that the genomes of Cameroonians are enriched with population-specific variants that are nearly—if not equally or more strongly—as protective as the HbAS. Therefore, an analysis focusing on Cameroonian samples only could shed some light on these grey areas.

In this project, I used a genome-wide association approach (GWAS) to investigate the genomes of samples of malaria cases and controls belonging to three ethnic groups (Bantu, Semi-Bantu, and Foulbe or Fulani) from three regions in Cameroon (South

West, Littorale, and Centre) where malaria is endemic. First, I ascertaining whether there was significant population genetic structure that may negatively impact the association signals by computing several measures of genetic distance. I then Scan the genomes of the case-control participants for genetic markers of resistance/susceptibility to malaria. To the best of my knowledge, this study is the first of its kind in Cameroon. It has confirmed the presence of fine-scale population genetic structure within Cameroon and uncovered malaria protective loci that have not be reported by previous studies. This study is fundamentally a hypothesis generating study that adds to the global body of knowledge of malaria protection, furthers the understanding of the impact of age-old malaria pressure on Cameroonian genomes, and importantly, will direct future human genetic studies of malaria in the country.

1.2. Statement of Problem

Malaria continues to kill almost half a million people every year globally. To date, no approved vaccine against malaria exists, and resistance to the drugs at the first-line of action, the artemisinin derivatives such as in South East Asia continues to threaten the global fight against malaria give they spread to Africa. In addition, the mosquitoes that transmit malaria (female *Anopheles* mosquitoes) are becoming resistant to the insecticides that are used to control them (Etang et al., 2016; Hien et al., 2017). As a result, the world has seen a slowdown in the progress against the disease for the past three years (WHO, 2020). Without an effective vaccine or a more effective drug, malaria will persist in the most endemic areas particularly in Africa, and the world may see a roll back in the gains made in prevention and treatment of malaria. Should resistance to artemisinin-derived drugs spread throughout Africa, malaria would be expected to continue to claim many lives.

Severe malaria is the preponderant cause of malaria morbidity and mortality in sub-Saharan Africa (WHO, 2020). The magnitude of the burden of the disease is reflected in its impact on the long-term economic growth and development of countries affected including Cameroon (Malaney et al., 2004). The large economic cost of prevention, treatment and loss of productivity due to disease-related morbidity and mortality play significant roles in reducing the gross domestic product (GDP) of these highly burdened countries (Malaney et al., 2004).

In Cameroon, where the entire population of ~23 million individuals is at risk of malaria and the virulent *Plasmodium falciparum* is the most prevalent, intermittent preventive treatment in pregnancy (IPTp3), and the number of children under 5 years with fever who visit hospitals for treatment are usually low (≤30%). Cameroon has therefore contributed 3% of the global malaria deaths over the past 3 decades. According to the most recent estimates, this accounted for ~22% of all deaths reported in health care facilities in 2017 (Severe Malaria Observatory, 2018) in the country.

Although there has been progress in reducing the prevalence of malaria in Cameroon from 41% since the year 2000 to 24% in 2017, the disease is yet to be controlled in the country (Antonio-Nkondjio et al., 2019). In addition, a recent review found that there has been a 7% decline in the efficacy of artemisinin-based combination therapies in the country from 97% in 2006 to 90% in 2016, with *P. falciparum* chloroquine resistance (*Pfcrt)* and *P. falciparum* multidrug resistance 1 (*Pfmdr1*) gene resistant mutations predominantly circulating among the parasite populations in the country (Antonio-Nkondjio et al., 2019). Meanwhile, the circulation of *An. gambiae*, *An. coluzzii*, *An. arabiensis* and *An. funestus* (the predominant Anopheles species in the country) resistant mutations in the country threaten the use of long-lasting insecticide nets (LLINs) (Antonio-Nkondjio et al., 2019).

Therefore, if progress against malaria is to be maintained towards reduction in deaths, cases and consequently elimination and eradication, continued research into more effective strategies to combat the disease is imperative.

1.3. Rationale

Genomic and epidemiologic studies have shown that host genetic factors play a major role in the inter-individual expression of severe malaria. Over the past decade, genetic studies such as candidate gene association analyses have proven valuable in probing mechanisms by which the malaria parasite, *Plasmodium spp.*, evade drugs and vaccines. However, such studies have been less powerful in and narrow in scope as only a few genes and variants can be studied each time. The development of genome-wide association studies (GWASs) ushered in a new era of disease variant discovery with significant power (Spencer et al., 2009). Thousands of disease associated polymorphisms have been discovered; 5687 GWASs comprising 71,673 variant-trait associations as of September 2018 (Buniello et al., 2019). While population-specific

GWAS studies have been conducted with arguable success in some African populations, this has not been performed in Cameroon. The availability of genome-wide data of Cameroonian malaria case and control participants offers a unique opportunity to explore the power of the method and unravel novel—perhaps "medically actionable"—gene variants in the population.

Although GWA studies have led to the discovery of important molecules in human populations, heterogeneity in study populations—characterized by genetic diversity—have often hampered the association analyses, particularly in sub-Saharan Africa. Genetic diversity in Africa is the greatest in the world, posing a particular challenge to GWASs. The genetic diversity of specific African populations has been largely understudied; with the exception being the Southern African population where extensive fine-scale population structure has been studied (Uren et al., 2016). Cameroon, the world's most culturally diverse population (J. D. Fearon, 2003) may have significant genomic differences among its ethnic groups that may be biasing genetic studies. These studies usually involve sampling large numbers of unrelated individuals, usually from two groups in the general population; a group with the condition or disease (called the case group) and another without the condition (called the control group).

Despite attempts to address heterogeneity in multi-site cohort genetic studies on the continent, these studies continue to underperform in Africa due to genetic diversity that is usually not fully accounted for by existing models—this prompted the MalariaGEN to recommend country-specific analyses (Malaria Genomic Epidemiology Network, 2015, 2019). Even in a highly structure population, country-specific association studies would require careful analysis design. Therefore, understanding the specific genetic structure of Cameroonian populations will be essential to inform sampling and analysis designs that are necessary to increase the power of current methods of association study. This would be expected to reduce the false positive rate, and increase the signals of disease association.

1.4. Research Questions

i. Is there significant genetic population structure in Cameroon that may bias association analysis of severe malaria and other genomic studies in the country?

ii. Are there novel genetic polymorphisms in Cameroonian populations and specific ethnic groups that are significantly associated with malaria and malaria sub-phenotypes?

1.5. Hypothesis

i. There is no significant population structure in Cameroon that may bias genomic studies.

ii. There are no novel genetic variants that are associated with malaria and malaria sub-phenotypes in malaria patients from the South West, Littoral, and Centre regions of Cameroon.

1.6. Objectives

1.6.1. General Objective

To investigate genetic diversity and markers of symptomatic malaria susceptibility in three malaria-endemic regions of Cameroon .

1.6.2. Specific Objectives

i. To perform population structure analysis of Cameroonians living in the Southwest, Littorale, and Centre region using measures of genetic proximity.

ii. To investigate the genomes of Cameroonians from the Southwest, Littorale, and Centre regions for markers that may be associated with symptomatic malaria susceptibility via genome-wide association analyses.

CHAPTER TWO

LITERATURE REVIEW

2.1. Introduction

Malaria is caused by apicomplexan parasites of the genus *Plasmodium* of which four species predominantly infect humans: *Plasmodium falciparum* (*P. falciparum*), *Plasmodium vivax*, *Plasmodium malariae*, and *Plasmodium ovale*. Another type called *Plasmodium knowlesi* found in Southeast Asia is closely linked to human malaria and infect monkeys (CDC, 2020). *P. falciparum* is the most predominant species in Africa, making over 95% of *Plasmodium* species in most African countries. The next most prevalent species in Africa is usually *P. vivax* which accounts for less than 5% of malaria infections in most populations, but up to 58% of infections in some regions of Ethiopia (Deress & Girma, 2019). Mosquitoes of the genus *Anopheles* are responsible for malaria transmission.

Malaria's devastating toll on the humankind is best captured by the opening statement of Livingstone's 1971 review, "Malaria has probably killed more human beings than any single disease" (Livingstone, 1971). Malaria has afflicted humans for much of the last 5,000 years following the adoption and spread of agriculture almost 10,000 years ago from around the fertile crescent, the region spanning modern-day Iraq, Syria, Lebanon, Palestine, Israel, Jordan, and Egypt, together with the south-eastern region of Turkey and the western fringes of Iran (Phillipson, 2006). While malaria is believed to have existed up to 100,000 years ago (Kwiatkowski, 2005), the virulent Plasmodium falciparum malaria is known to have diverged from gorillas about 40,000 – 60,000 years ago (Otto et al., 2018). A population bottleneck in the parasite population almost 5,000 – 6,000 years ago followed by rapid population expansions leading to the emergence of the *P. falciparum*-specific erythrocyte invasion protein EBA-175 ~4000 years ago then led to even more virulent strains (Otto et al., 2018).

Today, malaria remains a global public health concern. It still kills nearly half a million people each year; 409,000 in 2019 according to recent World Health Organization (WHO) reports, despite the gains made in prevention, control, and treatment (WHO, 2020). Sub-Saharan Africa accounts for a majority of malaria cases: 215 million (94%) of the 229 million cases reported in 2019, and the most malaria

deaths: 95% of the 409,000 deaths reported worldwide in 2019. Children below five years of age are the most vulnerable, and they accounted for 67% of all malaria deaths in 2019 (WHO, 2020).

2.2. Malaria transmission and *Plasmodium* parasite life cycle

Generally, malaria transmission is seasonal and closely linked to climatic conditions: precipitation, temperature, and humidity. In areas where all these conditions are favorable for the breeding and growth of the vectors, and consequently transmission, malaria, is usually endemic—transmitted throughout the year. However, in regions where at least one of these conditions is absent, malaria transmission may be sporadic (epidemics) or endemic in a sub-set of vulnerable immunologically naïve population.

The life cycle of malaria parasites is complex, requiring two hosts; the *Anopheles* mosquito and the human host (**Figure 2.1**) (Meibalan & Marti, 2017). Human malaria infection is established after the bite of a female Anopheles mosquito carrying *Plasmodium* sporozoites. Within 30 minutes of injection, the parasite sporozoites migrate through the lymphatic system and eventually get presented to the host immune system in the lymph nodes where some of them get eliminated by host innate immunity (Acharya *et al.,* 2017), and some of them make their way to the liver through liver sinusoids (Langhorne & Duffy, 2016).

The sporozoites infect liver cells (hepatocytes) using specialized receptors, and remain shielded from the host immune system in parasitophorous vacuole (Cowman et al., 2016). Some of the sporozoites are however eliminated by cytotoxic T cells (Tc) and interferon gamma (INF-gamma) producing cells. The sporozoites that make it to the liver undergo exo-erythrocytic schizogony by differentiating into schizonts which divide into thousands of merozoites after about 2 – 10 days (Cowman et al., 2016; Soulard et al., 2015). Five to six days after the hepatic phase, schizonts rupture to release merozoites which invade red blood cells in a multi-step process that last only about 2 minutes (Cowman et al., 2016). This presents a short transit and exposure time of parasite antigens to the host immune system and this has been regarded as a mechanism of evading the host immune system.

In the erythrocytes, the merozoites differentiate into trophozoites that further differentiate into schizonts. The schizonts undergo repeated mitotic division into merozoites that rupture the erythrocyte after about 48 hours and infect other red blood cells. Each burst releases about 20 – 30 new merozoites. This erythrocytic schizogony of the parasite life cycle involves the repeated division of schizonts into ring stage trophozoites, which in turn mature into schizonts again that burst to release merozoites. Some trophozoites differentiate and commit to gametocytes (sexual gametes: male or micro-gametocyte and female or acro-gametocyte) which are important for the continuation of the parasite life cycle. During another blood meal, the female Anopheles mosquito ingests the gametes. The sexual stage of the parasite life cycle occurs in the mosquito, wherein, the gametes fuse to form a zygote which develops into a motile, elongated ookinete. The ookinete migrates to the midgut of the mosquito and develops into Oocysts which later raptures to release sporozoites, ready to be injected into another human host. In the human host, the parasite interacts with many cell surface ligands and receptors to make its way into host cells.

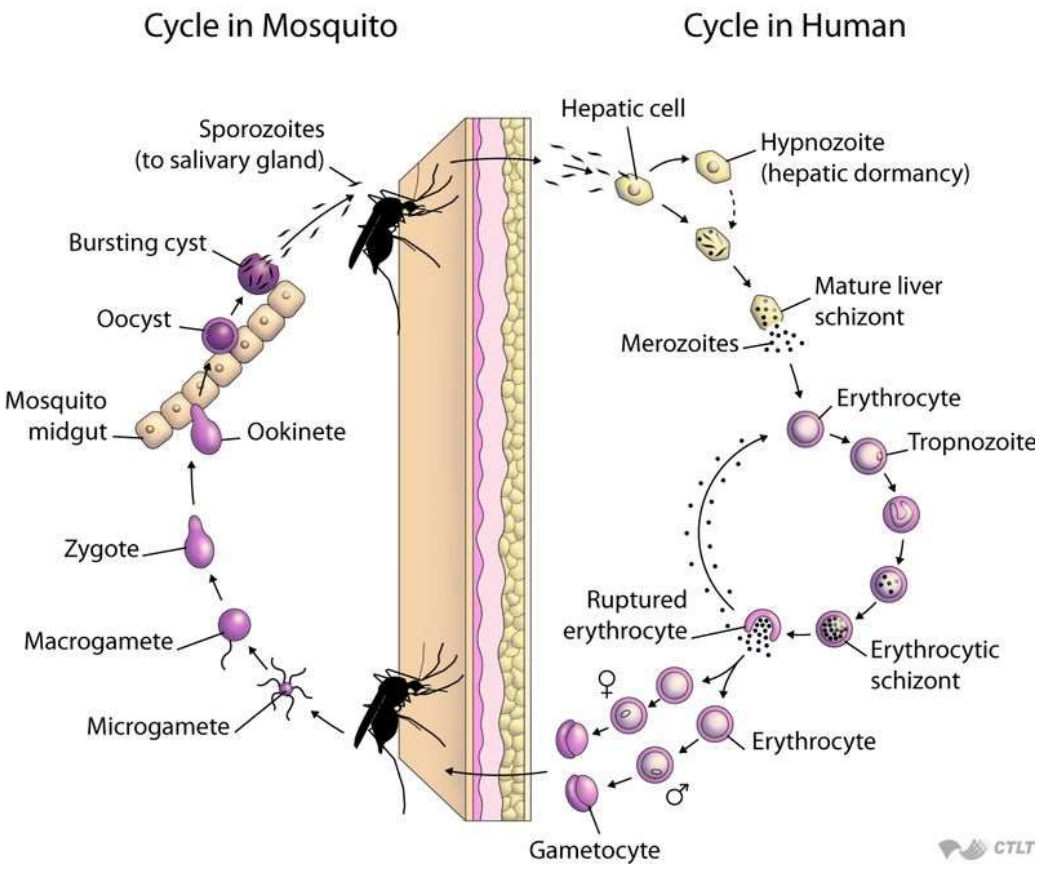


Figure 01: Life cycle of malaria parasite. (Source: Life cycle of the malaria parasite" from Epidemiology of Infectious Diseases. Retrieved from: http://ocw.jhsph.edu. Copyright © Johns Hopkins Bloomberg School of Public Health. Creative Commons BY-NC-SA)

2.3. Host-parasite interactions in the establishment of a malaria infection

After an infection is successfully established, parasites migrate in the blood vessels via interaction of their thrombospondin-related adhesive proteins (TRAP) with CD36 receptors on human endothelial cells (CD36 is receptor for several molecules like thrombospondin and short chain fatty acids) (Dundas et al., 2018). The sporozoites contain circomsporozoite proteins (CSP) on their surface which interact with heparan sulphate proteoglycans (HSPG - which are crucial in the uptake of chylomicron remnants by liver cells) on the surface of hepatocytes to penetrate liver cells (Langhorne & Duffy, 2016). The absence of HSPG on the skin cells is probably the

reason why the cells are not infected upon inoculation of sporozoites after a mosquito bite (Acharya et al., 2017).

After the hepatic cycle, the schizonts release merozoites in to the blood. Merozoites are characterized by small, polarized pear-shaped structures with multiple antigens on their surfaces (Wright & Rayner, 2014). They also contain specialized subcellular structures known as rhoptries and micronemes which are essential for erythrocyte invasion (Cowman et al., 2016). Merozoites use one of either two ways to invade erythrocytes: sialic acid (SA)-dependent pathway and SA-independent pathway (Acharya et al., 2017). The *P. falciparum* erythrocyte binding ligands (EBL) utilize the SA-dependent pathway which involves erythrocyte surface receptors bearing SA residues. Erythrocyte receptors implicated in this pathway include the glycophorins A (GYPA), B (GYPB) and C (GYPC) (Kwiatkowski, 2005). It has been shown that the *P. falciparum* EBL-175, EBL-1 and EBL-140 interact with glycophorins A, B and C respectively during parasite invasion (Wright & Rayner, 2014).

Merozoite surface proteins (MSP) are merozoite antigens that utilize the SA-independent pathway to invade the erythrocytes (Acharya et al., 2017). They possess an extracellular Duffy binding-like (MSPDBL) domain which binds Duffy binding-like receptors on erythrocyte surface (Jespersen et al., 2016; Kwiatkowski, 2005). Duffy-binding receptors (DBR) are erythrocyte surface receptors for binding of chemokines. They are also used as the principal pathway by *P. vivax* to invade RBCs (Langhorne & Duffy, 2016). An absence of these receptors on erythrocytes prevent *P. vivax* invasion, hence the reason for the almost complete absence of *P. vivax* infections in African populations which are largely Duffy negative (Kwiatkowski, 2005). MSPs are also very highly polymorphic and use several Duffy binding-like receptors on erythrocytes for invasion. This makes their invasive mechanism highly complex and redundant, hence effectively evading host immune system. Apical membrane protein-1 (AMA-1) is a merozoite surface antigen secreted from merozoite microneme that is also essential for red cell invasion. Its interaction with rhoptry neck protein (RON) forms a complex that triggers junction formation and hence invasion (Wright & Rayner, 2014).

Plasmodium vivax interacts with the Duffer-binding proteins on erythrocyte (preferably reticulocytes) surface of Duffy positive individuals and is therefore rare among Africans who are mostly Duffy negative (Howes et al., 2011). However, *P. falciparum*, which is the most virulent species, infects reticulocytes and mature erythrocytes via interaction with several Duffy binding-like receptors (Kwiatkowski, 2005). The reticulocyte binding-like homologous proteins of *P. falciparum* recognize other erythrocyte receptors different from Duffy binding-like receptors on erythrocytes and reticulocytes. This gives rise to several alternate and redundant pathways of parasite erythrocyte invasion (Wright & Rayner, 2014).

During the erythrocytic stage of its life cycle, the P. falciparum parasite expresses many proteins on the surface of erythrocytes. The var gene family of these proteins are one of the most characterized. They encode about 60 hypervariable antigens known as P. falciparum erythrocyte membrane proteins (PfEMP) which bind a variety of human host receptors (Dara et al., 2017). The N-terminal region consist of hypervariable Cysteine-rich interdomain region (CIDR) and extracellular Duffy binding-like (DBL) domain (Jespersen et al., 2016). PfEMP-1 binds CD36 on the surface of endothelial cells, dendritic cells, chondroitin sulfate A (CSA) in the placenta and other receptors to cause parasitized erythrocytes to sequester in deep vascular beds and placenta (Smith et al., 2013). This keeps infected cells away from general circulation hence promoting parasite growth and re-invasion, while shielding parasite from the immune system. PfEMP-1-mediated clustering of infected erythrocytes with uninfected erythrocytes (rosetting) promotes re-invasion (Saiwaew et al., 2017). VAR2CSA is the most conserved PfEMP-1 variant and is under investigation for possible malaria vaccine application (Kwiatkowski, 2005).

2.4. Leveraging host-parasite interactions for malaria vaccine development

Although the highly polymorphic nature of CSP limit its use as vaccine candidates, the most advanced pre-erythrocytic malaria vaccine candidate today, RTS,S/AS01E (brand name Mosquirix^{TM}), is based on the CSP protein (Duffy & Patrick Gorres, 2020). It comprises a central repeat (hence “R”) and C-terminal regions (containing T cell epitopes, hence “T”) which are fused to hepatitis B surface antigen (hence “S”), and expressed in yeast cells that also carry the hepatitis B surface antigen expression

cassette, hence the name “RTS,S” (Duffy & Patrick Gorres, 2020). Its formulation with the AS01 adjuvant completed Phase III clinical trials in sub-Sharan Africa by 2012 sponsored by the biotech company GlaxoSmithKline (GSK, hence RTS,S/AS01) (RTS, 2012). The vaccine candidate received favorable safety and efficacy measures in African children in 2015 (RTS, 2015). A larger 3-year trial in a real clinical setting was launched in 2019 in Malawi, Kenya, and Ghana and it was slated to involve over 1 million children. However, a major question surrounding the vaccine candidate and currently investigated by some researchers is how its efficacy might potentially change over time in regions of high malaria transmission that may drive genetic diversity of in the parasite (Amegashie et al., 2020).

PfSPZ on the other hand is a whole sporozoite vaccine that confers sterilizing immunity against Plasmodium sporozoites. In the PfSPZ-based vaccine, the attenuation of the sporozoites is either radiation attenuation, chemoattenuation (PfSPZ-CVac), or genetic (PfSPZ-GA) (Lyke et al., 2017; Mordmüller et al., 2017; Mueller et al., 2005). However, logistical challenges of this vaccine candidate include liquid nitrogen cold chain and intravenous inoculation (Seder et al., 2013). The efficacy of PfSPZ in humans is dose-dependent, while trials are underway to assess its safety (Duffy & Patrick Gorres, 2020).

Another category of malaria vaccines are the blood stage vaccine candidates which target the asexual parasite forms. Recall that merozoites are the blood stage form of the parasite that are transiently exposed to the immunity after their release from the liver and infection of RBCs. The transient nature of merozoites in blood therefore makes it challenging to develop anti-merozoite vaccines. In addition, the proteins expressed on merozoite surfaces (MSP and AMA) are highly polymorphic, hence serving as an effective vaccine escape measure. Furthermore, the pathway of erythrocyte invasion by merozoites is redundant, meaning that they can get into red blood cells using more than one means, and targeting only one route will make virtually no difference. Although blood stage vaccines have been developed mainly targeting MSP1and AMA1, and MSP3 and EBA-175 to some extent, their responses have not been encouraging; for instance, although the AMA-1 based vaccine candidates FMP2.1/AS01, AMA-1/AS01B, and AMA-1/AS02A were shown to

induce strong protective responses in vitro, they failed to elicit a similar response in controlled human infections (Payne et al., 2016; Spring et al., 2009). Other blood stage vaccine candidates include PfRH5 and AMA1-RON2. However, their efficacy demonstrated in monkeys is yet to be replicated in human infections (Duffy & Patrick Gorres, 2020).

Perhaps, one of them most exciting developments in the malaria vaccine pursuit is the recognition of the PfEMP1 as an immunodominant protein with vaccine potentials. Although the enormous polymorphic nature of the protein has largely impeded vaccine design, the VAR2CSA provides an exception with specific conserved domains and is thus an attractive target for placental malaria vaccines. Over the past five years, VAR2CSA-based vaccine trials have been conducted based on specific domains or combinations of domains (Duffy & Patrick Gorres, 2020). The transmission-blocking vaccine candidates Pfs230, Pfs48/45, Pfs25, and Pfs28 have also been tested, but however show poor immunogenicity as monomers. Therefore, protein-protein conjugates are being prepared for delivery by nanoparticles. The Pfs25-EPA conjugate was shown to elicit favorable immune response and well-tolerated in 2016 (Duffy & Patrick Gorres, 2020).

While much progress has been made towards malaria vaccine development, the enormous diversity of the parasite antigens, and redundancy in their interaction with the human host pose a great challenge to the current vaccine candidates, particularly in regions with high transmission such as in much of sub-Saharan Africa where the most variable clinical forms of malaria are observed.

2.5. Clinical presentations of malaria

Malaria infection can be either asymptomatic or symptomatic. Asymptomatic malaria is characterized by the presence of asexual parasites I the blood without symptoms of illness. Symptomatic malaria is characterized by high fever, excessive sweating and yellow coloration of the urine, and it can be either uncomplicated malaria (UM) or severe malaria (SM). The characteristic signs and symptoms of symptomatic malaria result from feeding of the parasites on human hemoglobin, detoxifying it to heme. Polymerization of heme into hemozoin initiates some of the pathophysiological

effects of malaria by eliciting proinflammatory molecules such as glycosyl-phosphoinositol (GPI). The secretion of proinflammatory cytokines in addition to cytoadherence and rosetting elicited by host-parasite interaction results in the variable disease severity.

Severe malaria is the life-threatening form caused predominantly by *P. falciparum* (WHO, 2014). Several sub-types exist including: severe malaria anemia (SMA – characterized by hemoglobin level <5 g/dl and parasitemia > 10,000/µl in children <12 years old, and hematocrit < 15%), cerebral malaria (CM – characterized by impaired consciousness, Glasgow coma scale < 11, and Blantyre coma scale < 3, coma persistent for more than 30 minutes after a seizure, and no record of recent severe head trauma, neurological disease or any other cause of coma), respiratory distress (RD), acidosis (characterized by a plasma bicarbonate of <15mM or venous plasma lactate >5mM), hypoglycemia (characterized by blood glucose level < 40mg/dl), renal impairment (characterized by plasma or serum creatinine > 3mg/dl or blood urea >20mM), jaundice (characterized by plasma or serum creatinine > 3mg/dl and parasitemia > 100,000/µl), pulmonary oedema (oxygen saturation <92% on room air with respiratory rate >30/min), hyperparasitemia (*P. falciparum* parasitemia >10%), as well as significant bleeding (WHO, 2014).

Although malaria control measures have helped to curb malaria mortality, almost half a million people still die each year. And these are only the cases that are reported in healthcare settings. In sub-Saharan Africa for instance, the epidemiological survey of severe malaria is hampered by poverty which affects the reporting and documentation of the cases, as well as the ability of patients to seek medical care. A large proportion of cases and deaths usually occur in homes without the knowledge of healthcare professionals (WHO, 2014). Therefore, it is conceivable that the mortality rate of malaria is even higher than current estimates. Yet, epidemiological studies have shown that of the people who get infected with malaria, a small proportion (about 2%) never come down with severe malaria (Kwiatkowski, 2005). Among the reasons advanced to explain this observation, human genetic factors are the subject of my research.

2.6. Host malaria resistance/susceptibility factors

The contribution of human genetic factors to the variable phenotypic expression of malaria is now well established. The first observation was made by the British-Indian population geneticist, J.B.S. Haldane in 1949 (HALDANE, 1949). Based on epidemiological data, Haldane observed increased frequency of the beta-thalassemia in regions with high malaria prevalence and hypothesized that beta-thalassemia affords protection against malaria. He went on to hypothesize based on mathematical calculations of allele frequencies that a slight gain in fitness conferred by the sickle cell trait should protect individuals in malaria-endemic regions from severe malaria. Haldane's hypothesis became known as the 'malaria hypothesis'.

In 1954, A.C. Allison, a South African geneticist observed epidemiological data and carried out experiments in Kenya and Uganda and observed increased frequency in malaria patients who carried the sickle cell trait but did not suffer from severe malaria (Allison, 1954). He concluded, unequivocally, that the sickle cell trait conferred resistance against *Plasmodium falciparum* malaria in particular. Further clinical and epidemiological data have largely confirmed these findings (Mackinnon et al., 2005). Today, the sickle cell mutation is most prevalent in regions where malaria was once endemic or remains endemic: in sub-Saharan Africa, the Indian-subcontinent, the Mediterranean, South East Asia, the Middle East, Oceania (Papua New Guinea), and South America. Due to its ancestral origin (from Africa), the sickle cell gene has become prevalent in countries with historical malaria endemicity. In such populations, the gene is predominantly contributed by individuals of African ancestry (Esoh & Wonkam, 2021).

The beta- and alpha-thalassemia, encoded by the *HBB* and *HBA* gene clusters respectively, also afford protection against severe malaria and are thus highly prevalent in many countries where malaria is also prevalent (Frédéric B Piel & Weatherall, 2014). The Mediterranean and Middle East harbor the largest frequencies of the beta-thalassemia (the so called thalassemia belt), while Oceania (Papua New Guinea) harbors the highest frequencies of alpha-thalassemia (F B Piel et al., 2013; Frédéric B Piel & Weatherall, 2014). Interestingly, the sickle cell mutation is almost absent from Papua New Guinea even though malaria is present throughout the country. A possible explanation to the observation came in 2005 when Williams et al., observed in Kenyan

children that the protection afforded by both the sickle cell trait and alpha-thalassemia against malaria is lost when the two genotypes occur in the same individual (Williams et al., 2005). This negative epistatic interaction was proposed to explain the opposing frequencies of the genotypes in Oceania. In sub-Saharan Africa, the most common alpha-thalassemia mutation is a 3.7 kb deletion that is highly prevalent; 8.5%, 10.0%, 15.0%, 16.7%, 17.1%, 19.1%, 30.7%, 27.4%, 31.5% in Gambia, Mali, Burkina Faso, Ghana, Nigeria, Cameroon, Malawi, Tanzania and Kenya respectively (Malaria Genomic Epidemiology Network, 2019).

While the sickle cell mutation of the *HBB* gene is the most characterized malaria resistance variant, other gene variants have been discovered. Glucose-6-phosphate dehydrogenase (G6PD) is an enzyme in the pentose phosphate pathway encoded on the X chromosome and involved in the control of oxidative damage. Mutations that lead to a reduction in *G6PD* gene expression and cause a deficiency in the enzyme have been reported to protect against severe malaria forms in sub-Saharan Africa (Kariuki & Williams, 2020). However, results have been contradictory; while protection was observed against cerebral malaria in heterozygote females and hemizygote males, increased susceptibility to severe malaria anemia was observed in hemizygote males and homozygote females (Clarke et al., 2017).

Blood group antigens including of the ABO and MN/S group systems are also known to offer protection against severe malaria, notably the Dantu variant (GYPA-B) of the MN/S blood group system prevalent in East Africa (Kariuki & Williams, 2020). Encoded on chromosome 4 in a region of ancient balancing selection, the variant is a rearrangement of the glycophorin genes A (*GYPA*) and B (*GYPB*), which increases tension in the membrane of red blood cells, making it hard for *P. falciparum* parasites to invade the blood cells (Kariuki et al., 2020). These earlier malaria-associated gene variants were uncovered by candidate (single) gene approaches. However, following the observation that these gene variants jointly explain only a small proportion of malaria phenotypic expression, it was suggested that many more variants that act in a polygenic manner contributing small effects may have been unaccounted for.

2.7. Methods used to study host genetic susceptibility factors to malaria

For many years, candidate gene studies were at the forefront of genetic association studies for identifying malaria risk variants (Patnala et al., 2013). In these studies, genes that had been previously linked to malaria [especially genes involved in immune responses such as toll-like receptor genes (Mockenhaupt et al., 2006)] were preselected, and genetic polymorphisms in the genes were determined (genotyped) in hundreds to thousands of individuals using custom-made genotyping assays such as the Sequenome MassArray® (Bradić et al., 2012). Since candidate or targeted gene association studies depend on the prior knowledge of possible disease-associated genes, they are considered to be hypothesis-driven. In candidate gene association studies, two groups of individuals are usually sampled randomly: a ***case*** group comprising of individuals with the disease, and a ***control*** group comprising of individuals from the same population as the case group, but without the disease. Two candidate gene malaria association studies were conducted in Cameroon in the early stages of the MalariaGEN collaboration and multiple markers that were associated with decreased susceptibility to malaria phenotypes were uncovered in *HBB*, *IL10*, *IL17RE*, *NOS2*, and *ADCY9* genes, while markers that were associated with increased susceptibility to malaria phenotypes were uncovered in *G6PD*, *IL17RD*, *EMR1*, and *RTN3* genes in two major ethnic groups in the country, namely the Bantu and Semi-Bantu (Apinjoh et al., 2013, 2014). Another design of candidate gene association study used to discover malaria risk variants involves the sampling of family members—so called pedigree analysis. In 2005, a pedigree-based study in Kenya was the first to estimate the relative contributions of genetic and other factors to the variability in malaria incidence—so called heritability of malaria (Mackinnon et al., 2005). Candidate gene studies are cheap, and involve only a few hundred polymorphisms that make them quick to perform and tractable to computational tools. The major limitation of candidate gene studies is that they require pre-knowledge of genes involved or suspected to be involved in malaria pathology. Therefore, polymorphisms in multiple unknown genes that play significant roles in malaria pathology remain hidden (so called missing/hidden heritability) (Manolio et al., 2009). In addition, candidate gene association studies are low resolution, meaning that only a handful of variants can be studied.

Genome-wide association study (GWAS) is now one of the most powerful tools for assessing a wider range of host malarial risk variants. GWAS requires no pre-knowledge of related genes (hence not hypothesis-driven), but rather screens the entire genome for any risk variants in the case and control groups (hence GWAS is inherently designed as a hypothesis-generating case-control study). In principle, millions of polymorphisms are genotyped in large samples in order to achieve appreciable power as defined below:

$$P = N\beta^2 f(1-f)r^2$$

where P = power; N = sample size (cases + controls); β = variant effective size; f = minor allele frequency (MAF); r^2 = Pearson correlation coefficient (or linkage disequilibrium – LD between markers)

In GWAS, SNPs are genotyped in larger platforms of several hundred thousands to millions of polymorphisms including the Affymetrix 100k, 500k, (consisting of 100,000 and 500,000 SNPs respectively) and Illumina 300k, 650k, 2.5M (consisting up to 2,000,000 SNPs). In addition, SNPs from public databases obtained via whole-genome sequencing of multiple populations are routinely added to study data by imputation (statistical inference using patterns of LD) to increase the power of the studies. The databases commonly used include those from the 1000 Genomes Project (Altshuler et al., 2010), the HapMap (Belmont et al., 2003), Human Reference Consortium, and the National Heart Lung and Blood Institute (NHLBI)'s Trans-omics for Precision Medicine (TOPMed) program (Taliun et al., 2019). Due to the sheer number of SNPs tested for association, the threshold for statistical significance has been estimated to be at least $5x10^{-7}$ or $5x10^{-8}$ (usually estimated per study by dividing the nominal *p*-value of 0.05 by the total number of SNPs used for association testing in the study) (Spencer et al., 2009). The tools that are commonly used to test for association in candidate gene association studies or GWAS implement basic Chi-Square or linear regression for test of association such as PLINK, Haploview, etc. However, because multiple factors can influence the association of variants with a disease (including age, sex, and ethnicity), more complex models that account for population structure and cryptic relatedness are often needed. In such cases, logistic

regression or linear mixed models are often implemented, for instance in PLINK, EMMAX, BOLT-LMM, and GCTA.

2.8. Methodological challenges to large-scale malaria genetic association studies in Africa

Although GWASs offer a powerful avenue for uncovering disease risk loci, the major methodological challenge faced in such studies in Africa is population genetic structure which is characterized by low linkage disequilibrium (LD) and extensive genetic variation among and within African populations (Teo et al., 2010). Due to low LD between markers in African populations, GWAS on the continent usually requires larger sample sizes to achieve as much power as GWAS in European and Asian populations. The extensive genetic heterogeneity in Africa which is usually accompanied by language, cultural, and religious disparities, usually leads to high false discovery rates (FDRs) and deficiency of significant SNPs in GWAS (Teo et al., 2010).

Human genetic diversity in Africa is complex, with roots from ancient and recent migration events giving rise to enormous genetic mixture (admixture) amidst numerous eco-geographic barriers to gene flow (Busby et al., 2016; Uren et al., 2016). The genomes of African populations have also been shaped by evolutionary and selection pressures from infectious diseases and the environment; this is evident from the relatively high frequency of the Sickle cell trait and the G6PD deficiency (Leffler et al., 2017; Medicine et al., 2004). The extent of the heterogeneity in the continent is perhaps captured by the existence of thousands of local ethnicities.

In Cameroon for instance, there are over 250 tribes distributed within 3 broad ethnic groups; Bantu (BA), Semi-Bantu (SB), and Sudanese (which includes the Foulbe). Previous genetic studies revealed that chunks of the genome of individuals of the BA and SB ethnicities are shared with their African counterparts (an expected finding) (Busby et al., 2016). Yet, the ancestral relationship and extent of genetic differentiation between individuals of different ethnicities in Cameroon have not been explored. Cameroon is considered the World's most culturally diverse nation (J. D. Fearon, 2003; Gardinier et al., 2001), meaning that Cameroon's populations may have been subject to extensive genetic admixture and numerous barriers to gene flow,

leading to differences in allele frequency and haplotype structure (such as in the *HBB* gene cluster) between ethnic populations.

Previous genetic association analyses in Africa have relied upon principal component analysis (PCA) to correct for population structure (Patterson et al., 2006). The top (five to ten) principal components (PCs) are us included as covariates in the association analyses (Malaria Genomic Epidemiology Network, 2019; Ojewunmi et al., 2019). More recently, approaches like mixed models (MM) (Kang et al., 2010; Loh et al., 2015; Yang et al., 2011) and Bayesian statistics (Marchini et al., 2007) have proven particularly effective, in accounting for genetic structure among and within populations in GWASs. Yet, scars of genetic structure remain visible in large-scale genetic studies on the continent given the extensive genetic diversity. Therefore, characterizing the genetic structure of specific populations in Africa will be crucial to the design, analysis, and interpretation of genetic association studies in the continent.

2.9. Treatment and control of malaria

Uncomplicated malaria can rapidly progress to severe malaria which is almost always fatal if not treated. Therefore, treatment of malaria is recommended within 24-48 h of the onset of malaria symptoms. However, due to the risk of antimalarial resistance, treatment is often recommended only to patients who truly have malaria. Effective treatment of malaria is based on rational use of antimalarial agents after a parasitological confirmation of a suspected malaria case, or through a rapid diagnostic test (RDT) (WHO, 2015). The artemisinin-based combination therapies (ACTs) are the first line of action against malaria. Treatment of uncomplicated malaria in children and adults is by any of the following combination therapies for 3 days: artemether + lumefantrine, artesunate + amodiaquine, artesunate + mefloquine, dihydroartemisinin + piperaquine, artesunate + sulfadoxine-pyrimethamine (SP). In special risk groups, however, modifications in treatment regimens are required. For instance, in pregnant women during the first trimester, quinine + clindamycin is recommended for 7 days. In HIV/AIDS patients who are on co-trimoxazole treatment, artesunate + SP is not recommended. For the treatment of uncomplicated *P. vivax*, (or *P. ovale, P. malariae,* and *P. knowlesi*) malaria in chloroquine-susceptible areas, chloroquine could be used or the ACTs (WHO, 2015). However, chloroquine must be avoided in chloroquine-

resistant areas. Severe malaria is treated by intravenous or intramuscular injection of artesunate for at least 24 hours and until the patients can tolerate oral medication (WHO, 2015). After parental treatment of severe malaria, treatment is then completed with 3 days of ACT. Control of malaria is by chemoprevention via the use of seasonal malaria chemotherapy (SMC) plus intermittent preventive treatment in pregnancy by sulfadoxine-pyrimethamine (SP), or by vector control using insecticide-treated nets (ITNs) and indoor residual spray (IRS) (WHO, 2015). Primaquine is usually used to prevent the relapse of *P. vivax* or *P. ovale* infections. Interestingly however, primaquine is often toxic to patients with G6PD deficiency as it can induce hemolysis. Therefore, close monitoring of such patients under primaquine treatment is recommended (WHO, 2015).

2.10. Malaria in Cameroon: the case of the Southwest, Littorale, and Centre regions

Cameroon is usually referred to as a west-central African nation, bothered by Nigeria in the West, Central African Republic in the East, Chad in the North, and Equatorial Guinea, Gabon, and Congo in the South, and situated within the Gulf of Guinea (latitude: 2–13°N, longitude: 9–16°E). With a surface area of approximately 475,000 km^2, the country is home to an estimated 24 million indigens who are all at risk of malaria (**Figure 2.2**) (Antonio-Nkondjio et al., 2019). The Atlantic Ocean forms an approximately 400m coastal border South West of the Country. Ten administrative regions make up the country with different ecological domains. Demographic and health survey (DHS) data, along with data from the malaria indicator survey (MIS) have indicated that vegetation and altitude are important predictors of the geographical distribution of malaria in Cameroon (Massoda Tonye et al., 2018).

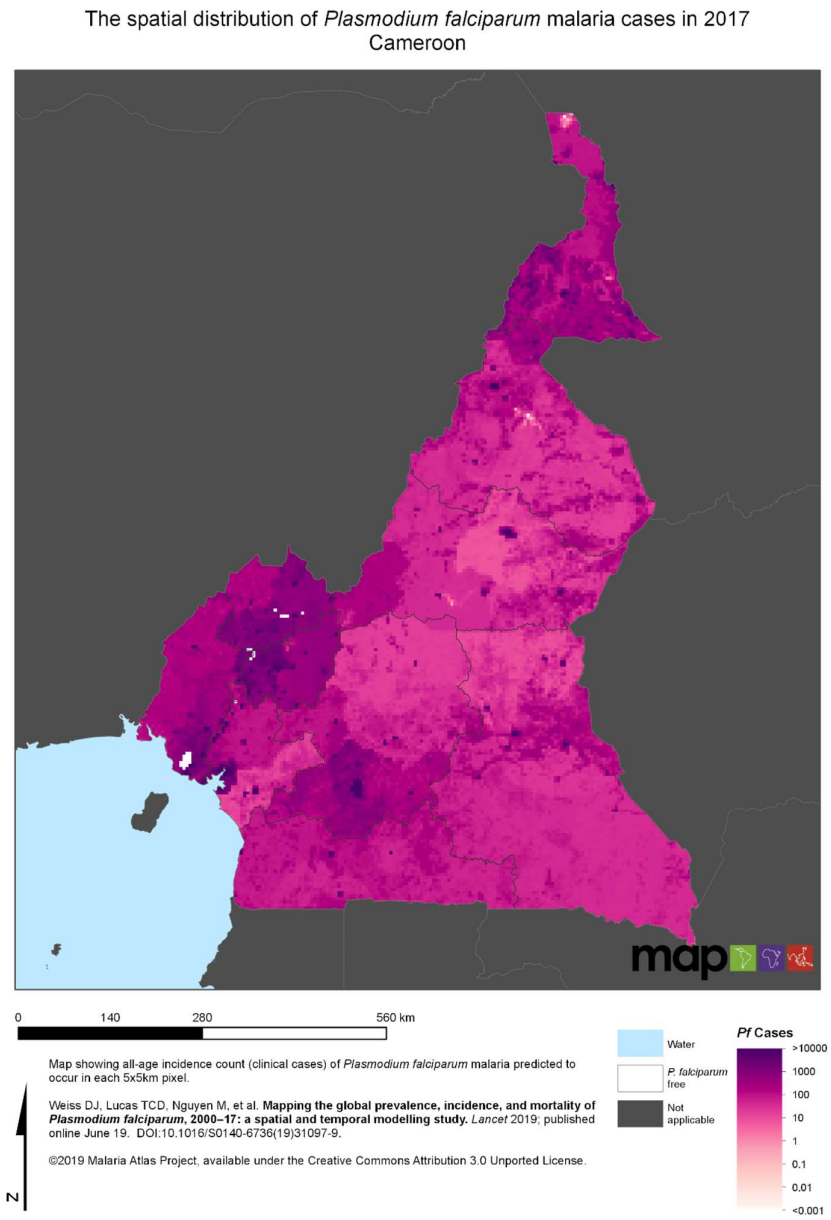


Figure 02: Map of Cameroon showing risk of malaria in the country: This is the most recent catalogue. Retrieved from: Malaria Atlas Project (Welcome to the Malaria Atlas Project - MAP)

In Cameroon, *P. falciparum* makes up about 95% of the parasite populations. *P. vivax*, *P. malariae* and *P. ovale* have also been reported to be circulating in the country (Tabue et al., 2019) with P. vivax only recently reported in the population (Fru-Cho et al., 2014; Russo et al., 2017). Six species are considered as mainly responsible for

malaria transmission in Cameroon: *An. gambiae*, *An. coluzzii*, *An. arabiensis*, *An. funestus*, *An. nili* and *An. Moucheti* (Antonio-Nkondjio et al., 2019). *Anopheles gambiae* is the most effective and wide-spread vector (Mbacham et al., 2019). Other secondary vectors, which are involved in either occasional or temporary malaria transmission have been reported including *An. ovengensis*, *An. paludis*, *An. ziemanni*, *An. coustani*, *An. pharoensis*, *An. marshallii*, *An. rufipes*, *An. carnevalei*, *An. hancocki*, *An. leesoni* and *An. Wellcomei* (Awono-Ambene et al., 2018). The increased use of LLINs and pesticides in agriculture have been associated with expansion of vector resistance to these interventions (Müller et al., 2008). Cameroon is made up of different epidemiologic strata such that transmission varies by stratum. Malaria transmission in Cameroon is intense in the rainy season (peak period) and continues throughout the year with low to moderate transmission in the Central, Littoral (Coastal) and South Western regions (**Figure 2.3**) (Eric A. Achidi et al., 2012).

The central region (Yaounde) has an equatorial climate characterized by constant temperatures between 17-30^{o}C with a mean of 23.1^{o}C, and is located within the rainforest belt of central Africa (Manga et al., 1997). Rainfall is usually abundant (1,500–2,000 mm), humidity usually ranges from 85% to 90%. Two rainy seasons ranging from March to May or June and September to November, and two dry seasons ranging from December to February and June or July to August make up four distinct seasons in the region. Malaria transmission in the region is maximal during and immediately following the two rainy seasons (Quakyi et al., 2000).

The South West and Littoral regions also have equatorial climates characterized by constant temperatures between 18-35^{o}C. Unlike the Centre region, the South West and Littorale regions have two seasons, a short dry season that runs from November to March, and a long rainy season that typically runs from March to November. Rainfall is usually abundant (2,000–10,000 mm) (Manga et al., 1997). The South West region harbours the tallest peak in West Africa, the Mt Cameroon, which endows the region with some peculiar climatic conditions in its high and low altitudes towns. Mean annual rainfall is usually 2625 mm with a relatively constant humidity of 75%–80% (Wanji et al., 2003). Malaria transmission is hyper-endemic in the rainy season, with peak incidence between July and October (E A Achidi et al., 2008).

The Bantu ethnic tribes are the indigenous occupants of all the three regions. However, the Semi-Bantu from the North West and West regions have migrated in huge numbers over the past century into the Littorale and South West for economic reasons (J. Fearon & Laitin, 2005). The Littorale is Cameroon's economic capital with the largest share of industries. The South West host lucrative plantations like tea, palm oil, rubber, cocoa among others which attract workers from other regions. In general, population movements have seen a fair mixture of the ethnic groups in all the regions, with the Centre hosting a larger share of individuals from the North, i.e. the Foulbe.

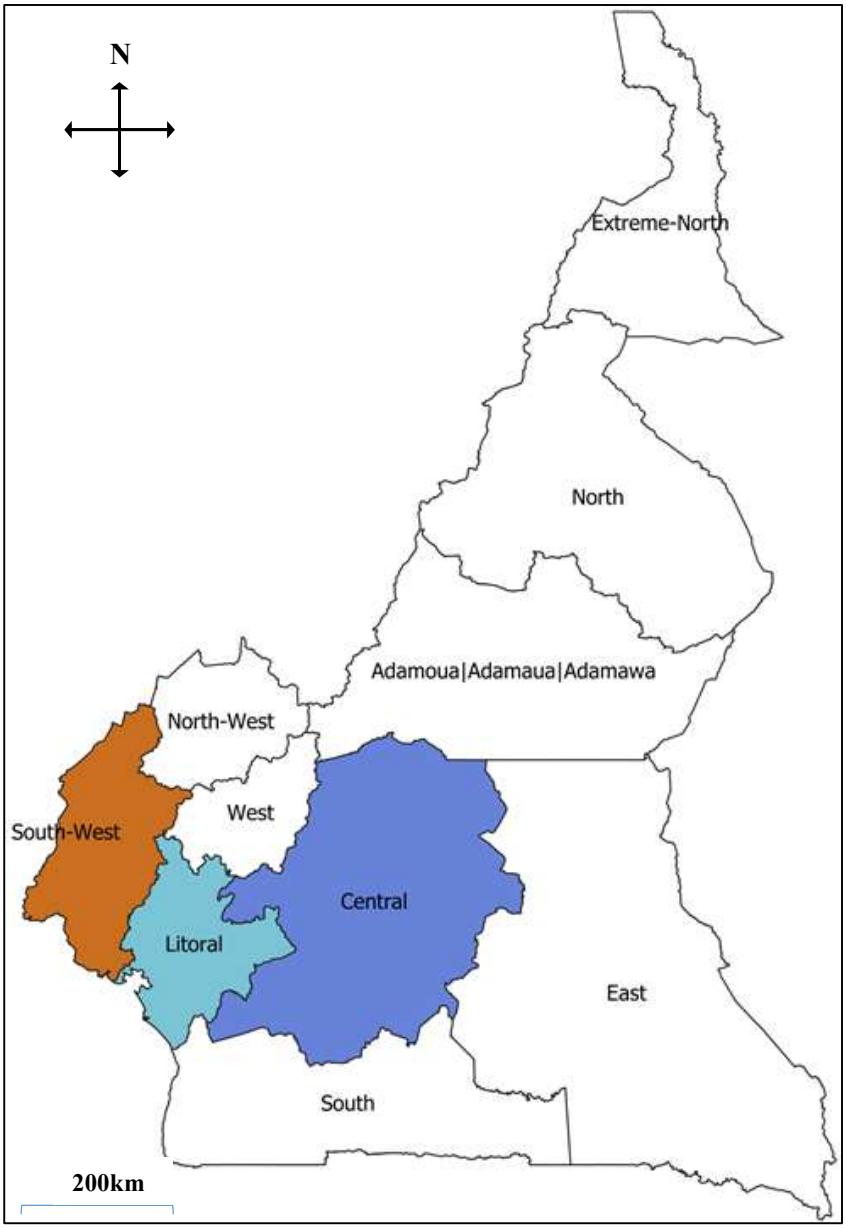


Figure 03: Map of Cameroon showing the three regions of study (Produced using the QGIS version 3.8 software). The colored regions represent the regions from where data for this study was sampled.

2.11. Conclusion

Large scale multi-site genome-wide association studies have proven powerful in revealing human genetic resistant/susceptibility factors to diseases that should be important in future vaccine development strategies. Yet, these methods continue to be

plagued by several challenges in the African setting; of which genetic diversity among African populations is the greatest. Even though more robust methods have been developed to account for genetic diversity among African populations, diversity within the populations still continue to reduce the robustness of these tools in the continent. Therefore, population-specific large-scale genomic analysis that start by first explicitly characterizing the genetic diversity or structure within the population are highly imperative.

CHAPTER THREE

MATERIALS AND METHODS

3.1. Research Design

The current study was an unmatched case-control secondary analysis of human genetic data that was generated between 2003/2005 and 2007/2008 by Achidi *et al.*, as part of a malaria cross-sectional study in three malaria-endemic regions of Cameroon: South West, Littorale and Centre (Eric A. Achidi et al., 2012). In the primary study, DNA samples were extracted in the Malaria Research Unit at the University of Buea in Cameroon and the samples were shipped to the MalariaGEN Oxford Resource Centre in the United Kingdom for genotyping and further processing (Eric A. Achidi et al., 2012). The genotype data contributed to the MalariaGEN Consortial Project 1 (CP1) which was a large-scale multi-site malaria case-control analysis (Malaria Genomic Epidemiology Network, 2005). In addition, two candidate gene association analysis were performed based on pre-selected malaria-associated genes (Apinjoh et al., 2013, 2014).

3.2. Study Area

Samples were collected from four towns across the three regions including Buea and Limbe in the South West Region, Douala in the Littoral Region and Yaounde in the Central Region (Eric A. Achidi et al., 2012). The study sites included hospitals or health centres (where cases were recruited), and primary schools and blood banks (where controls were recruited). Hospitals and health centres included Bota District Hospital and Regional Hospital in Limbe, Laquintinie Hospital (Douala), Mother and Child Hospital (Yaounde), Regional Hospital Annex (Buea), Bokova Health Centre, Mount Mary Health Centre (Buea) and PMI Down Beach (Limbe). All but one (Mount Mary Health Centre) of the health facilities were government institutions which also received patients from neighbouring towns and villages. Primary schools included: Catholic School (CS) Buea Station, CS Great Soppo, CS Muea, Government School (GS) Bolifamba, GS Bonduma, Government Practising School (GPS) Molyko I and II, GPS Muea I and II, HOTPEC Primary School Mile 15 Buea, Oxford Primary School Muea and Government Bilingual Primary School Muea (Eric A. Achidi et al., 2012).

3.3. Ethical Approval

Ethical clearance for the study was obtained from the Institutional Review Board of the Faculty of Health Sciences, University of Buea (proposal number: ID D7.1.A/MPH/SWP/PDPH/PS.CH/2340/811) while administrative authorization was sought from the South West Regional Delegation of Public Health. Authorization to conduct the surveys in primary schools was obtained from the Regional Delegation of Basic Education or the Catholic Education Secretariat. Informed consent was obtained from each participant or their caregiver following a clear explanation of the content of the information sheet for the cases and blood bank donors. Authorization to enroll participants from health facilities or schools was obtained from the Director or Head teacher and only subjects/caregivers who volunteered to participate by signing a written informed consent were enrolled. Access and analysis of the data was done in strict adherence to the MalariaGEN Data Access Policies (Malaria Genomic Epidemiology Network, 2008a, 2009).

3.4. Case definition: inclusion and exclusion criteria

Cases consisted unrelated children with severe malaria (SM) or uncomplicated malaria (UM), aged 1 month to 13 years (See section on Literature Review for definition of UM and SM). Controls were apparently healthy (afebrile) children (aged 1-14 years) and asymptomatic adults (aged 17-52 years) of the Bantu and Semi-Bantu ethnic groups.

3.5. Data retrieval and quality processing

Genotype data was retrieved from MalariaGEN server using secure file transfer protocols (sftp) on approval by site principal investors (PIs) and according to MalariaGEN data access policies (Malaria Genomic Epidemiology Network, 2009). The data consisted per-chromosome VCF files alongside a sample file with case-control information. Sample and SNPs process report files were also retrieved for initial quality information. Briefly, genotyping was performed on the Illumina Omni2.5M array and alignment against the GRCh37 reference genome and genotype calling was performed according to the MalariaGEN three-way genotype calling

algorithm. We carried into our analysis on 2.3 million SNPs from 1471 individuals (693 cases, 778 controls).

3.5.1. Quality control (QC)

The strength of population genetic and association analyses depend on rigorous quality control of the genotype data. Sample QC was performed on the autosomes and the X chromosome using PLINK (Chang et al., 2015). Individuals whose reported nationality was "Non-Cameroonian" or "Missing" and individuals with inconsistent sex information were removed. One individual from each pair of related individuals (2nd, FS, PO, MZ; see appendix 1) was excluded by computing an identity by descent (IBD) report using the KING v2.2.4 software (**Appendix 1**) (Manichaikul et al., 2010). Individuals with outlying heterozygosity (out of the range 0.180 - 0.230) and individuals with >10% missing genotype count were excluded (**Appendix 1**).Using *smartpca* of the EIGENSOFT package (Patterson et al., 2006) samples with outlying ancestries were removed by projecting the dataset against the African populations from the 1000 Genomes phase three version 5 (1KGP3) reference panel (Altshuler et al., 2010). SNP QC involved removing SNPs with minor allele frequency (MAF) < 1%, genotype quality < 95%, SNPs whose genotype quality was significantly ($P < 1x10^{-8}$) different among cases and controls that may indicate a batch effect, and SNPs that failed the Hardy-Weinberg (HWE) equilibrium test at $P < 1x10^{-20}$ were removed. A total of 1,863,254 SNPs of 2,261,351 were left following SNP QC.

3.5.2. Haplotype estimation (phasing) and genotype imputation

Palindromic A/T and C/G SNPs were removed prior to phasing. The remaining SNPs were checked and validated against the 1KGP3 reference panel using the *conform-gt v24May2016* program (Browning, 2016). Phasing and genotype imputation were performed using three strategies; In-house imputation pipeline using EAGLE v2.4 (Loh et al., 2016) for phasing and IMPUTE2 (Marchini & Howie, 2010) for imputation, and web-based strategies using the TOPMed and Michigan Imputation servers (TIS and MIS) using EAGLE v2.4 for phasing and Minimac v4 for imputation (Das et al., 2016; Fuchsberger et al., 2015; Taliun et al., 2019). For downstream analyses, only biallelic SNPs with imputation accuracy (R^2 or IMPUTE info score) ≥ 0.65, MAF > 0.01, and genotype probability ≥ 95% were used.

3.6. Determination of fine scale population structure:

3.6.1. Allele frequency

To generate allele frequency (AF) data, a cluster-stratified AF analysis (*bin size* = 0.05) was performed among randomly selected Cameroonian individuals (BA=50, SB=50, FO=25) using Plink1.9.

3.6.2. Estimation of measures of genetic proximity

Linkage disequilibrium-pruning of SNPs was performed prior to estimation of F_{ST}, PCA, and model-based clustering. In addition, only SNPs with MAF > 0.05 were used. Pairwise F_{ST} estimates among the three ethnic groups within the current data and among the current data and the 1KGP3 reference populations were computed using *smartpca*. Ten (10) axes of genetic variation (principal components - PCs) were computed using *smartpca*. PC plots and F_{ST} heatmaps showing the clustering of the populations into subgroups were generated using R (R Core team, 2016). Model-based clustering was performed using the Admixture algorithm (Alexander et al., 2009). The analysis was done with 5 cross-validation runs (K=1-5) and 300 bootstrap runs. Co-ancestry analysis was performed using ChromoPainter in the linked (LD) mode to summarize the genomic proportions shared among each donor and recipient individual ("Coancestry matrix"). FineStructure was then used to assign individuals into clusters.

3.6.3. Genome scan for signatures of selection

Signatures of selection were investigated to gain insight into the roots of fine-scale population structure by computing the integrated extended haplotype homozygosity (EHH) score (iHS) and cross-population locus-specific integrated (EHH) score (Rsb) using the REHHv3.01 (rehh) package in R (Gautier et al., 2017). Both iHS and Rsb statistics were computed using phased haplotypes with MAF ≥ 0.05. First, iHS was computed on the pooled data set (with all the ethnicities), then Rsb was computed with separate pairs of the different ethnicities (SBvsBA, SBvsFO, and BAvsFO). To assess the significance of selection signatures, rehh computes a two-sided p-value from the Gaussian cumulative distribution function of iHS estimates. The p-values were adjusted by the Benjamin-Hockerberg (BH) and Bonferroni (BF) methods (Benjamini & Hochberg, 1995). The extended Lewontin-Krakauer Fst outlier test (FLK) was also computed (Bonhomme et al., 2010) on a dataset of unlinked loci using hapFLKv1.40

(Fariello et al., 2013). The haplotype variant of FLK test, hapFLK (Fariello et al., 2013) was performed per chromosome using linked loci with MAF ≥ 0.05. HapFLK estimates p-values using a rlm function in R which were adjusted by dividing 0.05 by the total number of SNPs used. The result were visualized using the qqman package in R (D. Turner, 2018).

3.6.4. Investigate of population structure due to malaria pressure: HBB gene cluster haplotypes

Bcftools v1.9 of the SAMTools package (Li et al., 2009) was used to extract haplotype data for HbAA, HbAS, and HbSS chromosomes from the imputed dataset. One hundred and forty six (146) samples with HbS-positive chromosomes were predicted in the phased imputed dataset (Bantu = 79; Semi-Bantu = 67) (Shaikho et al., 2017). Haplotypes were also predicted in the HbS-negative chromosomes and classified them into the classical groups on the basis of four previously described SNPs (*n* = 883; Bantu = 486; Semi-Bantu = 397)—hereafter referred to as 'base' population—in order to gain insight into haplotype conservation in this population.

3.7. Heritability estimation and association analysis:

The relative contributions of genetic and other factors to the variability in malaria incidence in the data (heritability) was performed using EMMAX (Kang et al., 2010), BOLT-LMM v2.3.4 (Loh et al., 2015), and GCTA v1.93.2 (Yang et al., 2011), while association analysis was performed using EMMAX (Kang et al., 2010), BOLT-LMM v2.3.4 (Loh et al., 2015), GCTA v1.93.2 (Yang et al., 2011), and PLINK2 (Chang et al., 2015). Twenty (20) principal components were used for the analyses. For BOLT-LMM, variance component was first estimated using ~500,000 in near linkage equilibrium. For GCTA, a phenotype file with GCTA case-control encoding (1, 0), was first used this to generate a genetic related matrix (*grm*). The *--reml* function was then used to estimate heritability, and the *–mlma* function was used for mixed linear model association analysis. Using EMMAX, a covariates file and a kinship matrix were first generated and used for association analysis. Used the PLINK2 *--glm* function, different models of association were tested: additive [default], genotypic, hethom, dominant, and recessive. All association analyses were adjusted for sex

differences. Association analyses were run on the in-house imputed set and the TOPMed imputed set for the pooled dataset, and the ethnic groups separately.

3.7.1. Annotation of top association signals

The top association signals were annotated using ENSEMBL's variant effect predictor (VEP) (McLaren et al., 2016). Pathway enrichment was done using the STRING server (https://string-db.org).

CHAPTER FOUR

RESULTS

4.1. Characteristics of the study participants

After sample and SNP quality control, the remaining 1073 samples were used for population structure analysis. After population structure analysis, a total of 1029 samples were retained for association analysis. The characteristics of the 1029 participants for whom association analysis was performed are described in **Figure 4.1**. The analysis was based on all malaria cases versus controls for Bantu and Semi-Bantu individuals (Foulbe individuals were excluded due to low sample size). Stratification of cases into sub-phenotypes was deemed to yield insufficient power for GWAS due to low sample sizes after exclusion of low quality samples. Since the most prevalent phenotype was severe malaria (Eric A. Achidi et al., 2012), I did not expect the analysis to be significantly under-powered as a result of pooling the sub-phenotypes together.

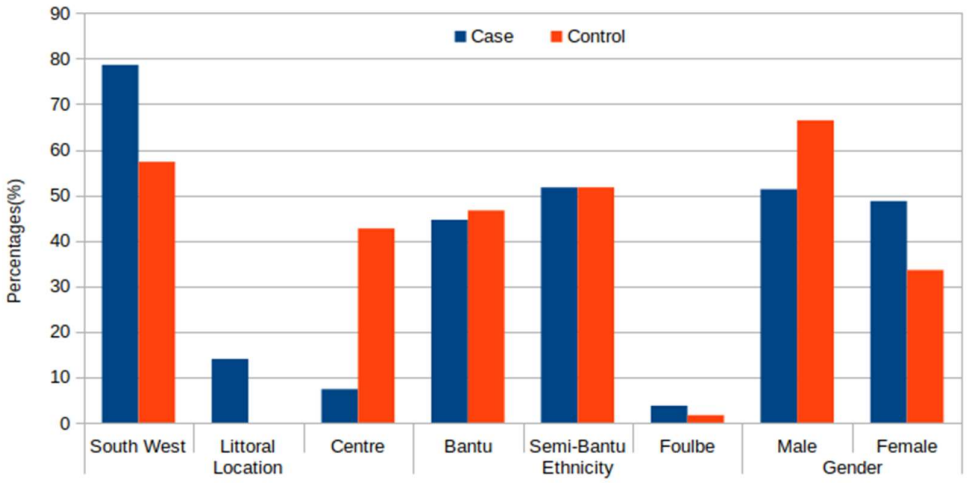


Figure 01. Demographic characteristics of the case-control participants

4.2. Determination of fine scale population structure:

4.2.1. Allele frequency

The FO exhibited a substantial difference in low frequency alleles from the BA and SB populations. The SB and BA had similar AF spectra (**Figure 4.2**). In addition, the BA and SB had a higher proportion of rare alleles than the FO.

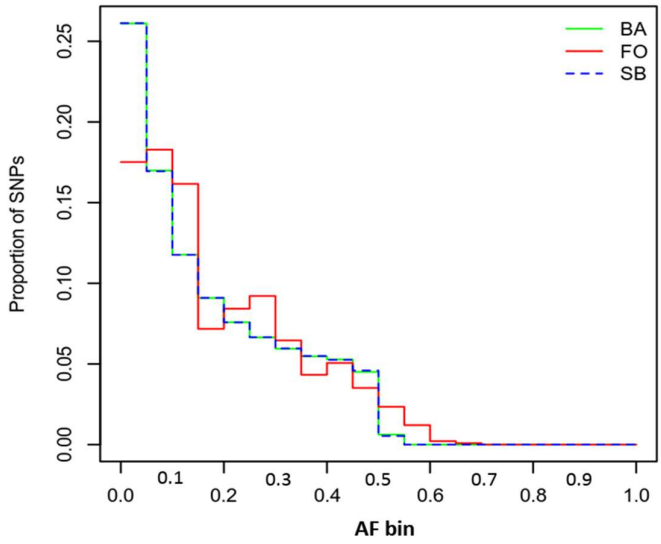


Figure 02. **Allele Frequency Spectrum:** a) Allele frequency spectrum among Cameroonian ethnic groups. The blue line (SB) and green line (BA) are perfectly overlayed such that the blue line is broken to reveal the green**.**

4.2.2. Genetic distance (F*st*)

Estimates for within- and among-continent population comparisons were similar to those previously reported. Here, Cameroonian populations generally clustered with other African populations (**Figure 4.3**). The SB clustered closer to the Yuroba of Nigeria (YRI) (F_{ST} SB vs YRI = 0.002) than did the BA (F_{ST} BA vs YRI = 0.003) contrary to previous estimates (Busby et al., 2016). The FO ethnicity was found to be relatively less genetically related to the YRI (F_{ST} = 0.004) compared to Cameroonian SB and BA populations. Interestingly, the FO, like the LWK population appeared to be more genetically close to populations of European and Asian ancestries when compared to the BA and SB.

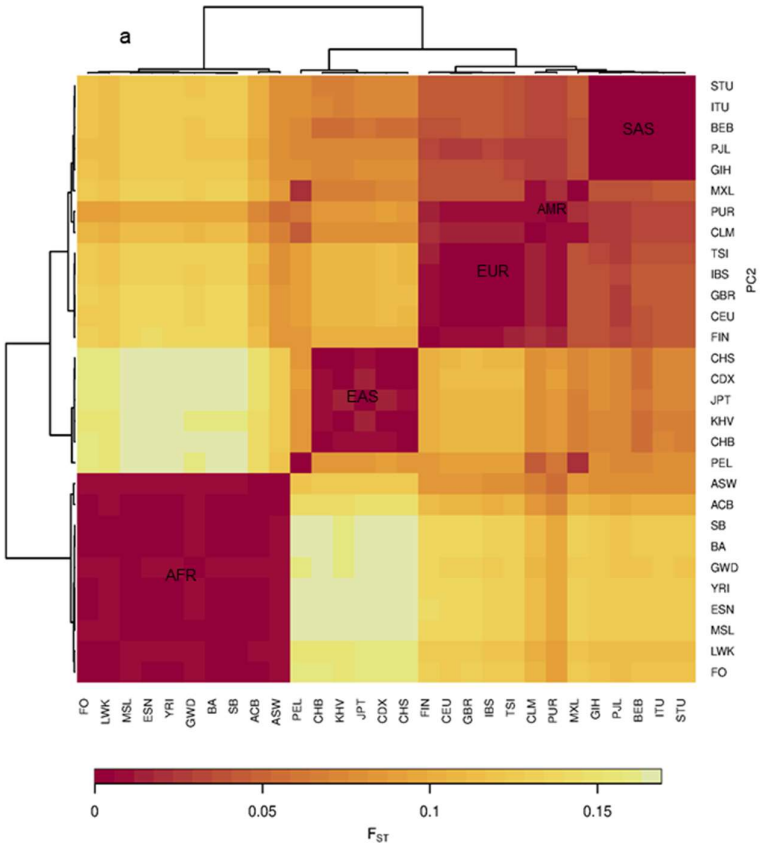


Figure 03. Pairwise Fst and PCA analysis of Cameroonian and world populations: Clustered heatmap showing genetic distance by pairwise population F_{ST} (Hudson) estimation. AFR=African, EAS=East Asian, EUR=European, AMR=American and SAS=South Asian ancestry. The red color denotes closely related population, hence low F_{ST} while the decrease in redness to yellow represents increasing genetic distance (high F_{ST}). Five clusters are apparently corresponding to the five continental proxy ancestry (distinguished broadly by five colors) in the 1,000 Genomes project

4.2.3. Principal component analysis

PCA generally showed positive concordance with F_{ST} results. PCA revealed three clusters in the dataset. Running PCA with "ancestry informative markers" saw an increased resolution of the clusters, clearly separating the FO from the BA and SB

based on PC2 (**Figure 4.4a**). Furthermore, substructures within the FO population and BA population were resolved (**Figure 4.4b**).

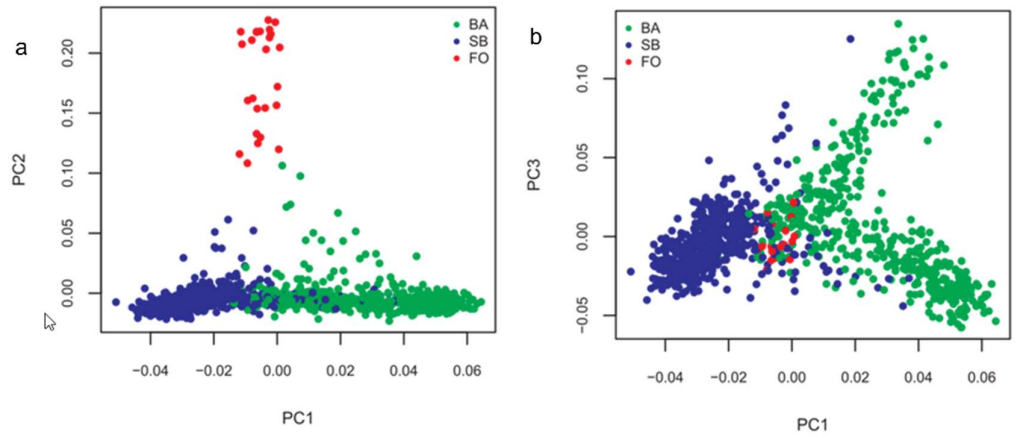


Figure 0.4: PCA of Cameroonian populations: (a) PCA for Cameroonian ethnicities only. PC1 and PC2 separate the three ethnicities, (b) PC1 and PC3 separate the Bantu and Semi-Bantu

4.2.4. Model-based clustering and co-ancestry estimation

At K=2, model-based clustering differentiated the three ethnicities albeit with low resolution. However, at K=3 where the lowest cross-validation error was recorded (**Figure 4.5a**), the three ethnicities were clearly differentiated (**Figure 4.5b**). Ancestral proportions (Q) estimated [green predominant in the BA (~45%), red in the FO (~75%) and blue in the SB (~45%)] (**Figure 4.5c**) show that the different ethnic groups differ by allele frequencies or haplotype structure. Co-ancestry estimation revealed isolation of a subgroup of FO individuals from the BA and SB by PC1 (**Figure 4.5d**). PC2 separated the Bantu and Semi-Bantu, with significant numbers of BA individuals clustering with SB. Generally, all the ethnicities showed a cline into a central cluster that appeared to be a set of admixed individuals, consistent with their longstanding cohabitation while their separation suggests some evidence of ancient genetic isolation and/or gene flow from other populations. The three extreme clusters may represent individuals with the basal ancestry for each respective ethnicity

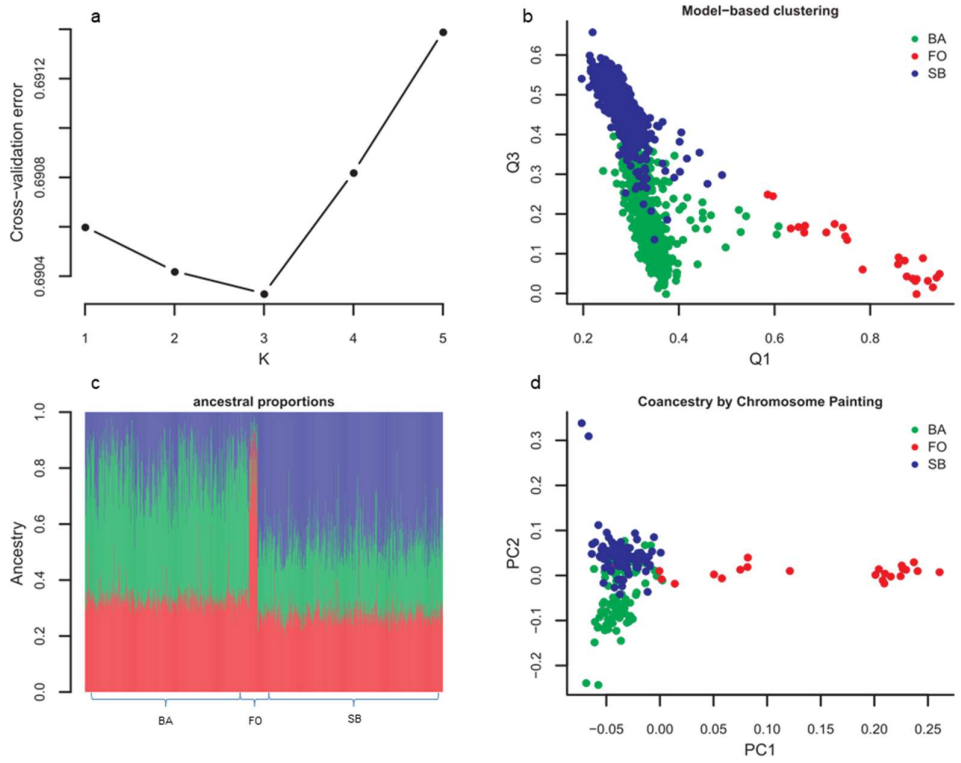


Figure 05: Model-based clustering and Co-ancestry estimation*:* a) Model-based clustering cross-validation (CV) error. Lowest CV error recorded at k=3 indicating three clusters. (b) and (c) show ancestral proportions Q, for each ethnicity colored using the RGB color scheme. (d) Co-ancestry estimation by FineStructure show fine-scale resolution of the clusters according to the ethnic groups. A substructure within the FO ethnic group is apparent.

4.2.5. Genome scan for signatures of selection

Genome-wide scan for signatures of selection by the standardized integrated haplotype score (iHS) which measures the EHH identified strong signatures on multiple chromosomes. This included missense and regulatory region variants in genes overwhelmingly associated with response to infections. The scan identified a total of 133 SNPs within 57 overlapped genes and 173 overlapped transcripts across chromosomes 1 to 12, 14, 16, 17, 19, and 20 with significant signatures of selection at iHS threshold of ±4 (**Figure 4.6**). **Table 4.1** shows the variants with strong signatures of selection that occurred in coding regions. Although the strongest signature occurred on chromosome 1 around the *REG4* gene (iHS = -7.23, p-value = 4.67×10^{-13}), the most

consistent signatures were recorded on chromosome 6 spanning the HLA region which has been reported in several previous studies of selection (Bhatia et al., 2011; dos Santos Francisco et al., 2015; Gineau et al., 2015; Nielsen, 2005). The SNP rs10947368, a missense variant on *HLA-DOA* emerged with the strongest signal within the HLA region (iHS = -6.42, p-value = 1.38×10^{-10}) (**Table 4.1**). In addition, suggestive signatures of selection were recorded in the hemoglobin-beta (*HBB*) gene cluster of chromosome 11, a region with longstanding knowledge of balancing selection under the influence of malaria (Nielsen, 2005). However, the strongest signal on chromosome 11 was a relatively uncommon missense variant (rs7943508) in the *APLNR* gene, implicated in hypertension and some cancers (Lee et al., 2019; Wu et al., 2018). The iHS values obtained were generally normally distributed as expected under neutral evolution with a slight deviation from the expected distribution.

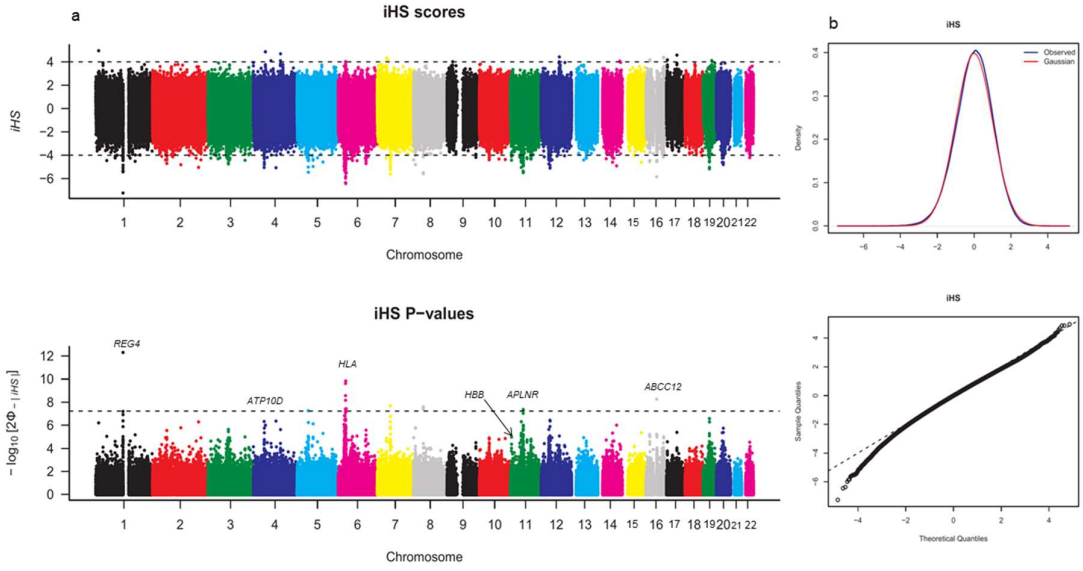


Figure 06: **iHS and corresponding –log10(p-values) Manhattan plots:** a) iHS plot for the autosomes. Negative values signify selection on derived alleles while positive values are associated with selection on ancestral alleles. (b) Distribution of iHS values as observed in the populations (blue) and as expected under neutral evolution (red). Lower plot represents quantile-quantile (Q-Q) plot of iHS p-values. The plot shows that the test statistics are not inflated.

Table 01: Variants with strong signatures of selection in coding genomic regions

rsid	chr:pos	ref	alt	alt.AF	ihs	*p*-value (bh)	*a.a change*	gene
rs10947368	6:32975341	C	T	0.1076	-6.422	5.73e-05	K120N	*HLA-DOA*
rs8192564	6:32191822	G	A	0.06011	-5.494	0.0024	-	*NOTCH4*
rs115261305	6:32793668	C	A	0.1319	-5.385	0.0030	-	*TAP2*
rs1126544	6:33037061	G	C	0.1281	-5.133	0.0085	T121T	*HLA-DPA1*
rs3800326	6:28264717	C	T	0.1039	-5.111	0.0089	P256L	*PGBD1*
rs61737338	6:28227217	C	T	0.08388	-4.922	0.0154	S23F	*NKAPL*
rs7943508	11:57003581	C	T	0.0657	-4.861	0.0176	V300I	*APLNR*
rs6582601	12:38716034	C	T	0.1253	-4.850	0.0183	-	*ALG10B*
rs2233954	6:31105672	G	A	0.08854	-4.802	0.0207	-	*PSORS1C2*
rs34304311	6:28093263	G	A	0.07549	-4.657	0.0298	L14L	*ZSCAN16*
rs6115256	20:25666642	C	T	0.1771	-4.600	0.0348	L48L	*ZNF337*
rs17190762	6:31126992	G	A	0.06291	-4.528	0.0429	-	*TCF19*
rs10896290	11:56128081	A	G	0.2679	-4.524	0.0429	Y120C	*OR8J1*
rs61729683	6:32185818	C	T	0.06337	-4.522	0.0430	A526A	*NOTCH4*
rs78133850	11:57004659	G	A	0.06897	-4.508	0.0442	-	*APLNR*
rs73468666	11:56958933	G	A	0.1761	-4.493	0.0462	-	*LRRC55*
rs75301276	11:55944198	C	T	0.06058	-4.491	0.0462	Y35Y	*OR5J2*
rs58567530	16:48172185	C	A	0.06943	-4.489	0.0464	L311L	*ABCC12*
rs3013106	1:13802437	G	A	0.4455	4.995	0.0116	S254S	*LRRC38*

rsid = Reference SNP ID, chr:pos = Chromosome number and position, ref = Reference allele, alt = Alternate allele, alt.AF = Alternate allele frequency, ihs = Integrated haplotype score, p-value (bh) = Benjamin-Hochberg adjusted p-value, a.a change = Amino acid change

Cross-population selection scan using the Rsb statistic found chromosome 6 to be strongly selected in the BA and SB and only subtly selected in the FO. The BA population showed additional signatures on chromosome 6 involving the missense variant rs9276 on the *HLA-DPB1* and the variant rs1419638 on the *OR5V1* gene, as well as on chromosome 7 not present in the other ethnicities. Likewise, the SB showed specific signatures on chromosomes 16 and 20, while strong signatures specific to the FO population were recorded on chromosomes 1, 7, 9, 10, 16, and 19 (**Appendix 2**). Again, these selection signatures primarily implicated genes involved in disease response.

The extended Lewontin-Krakauer F_{ST} outlier statistic (FLK) for positive selection revealed several genomic regions with subtle allele frequency differences between the ethnicities although none of these regions remained significant after correction for multiple testing by the Benjamin-Hochberg method (Benjamini & Hochberg, 1995; Chen et al., 2017). However, evidence of positive selection remained apparent in the HLA region on chromosome 6. Specifically, positions on chromosomes 2, 6, 8, 10, 17, 18, and 22 were subtly differentiated among the ethnicities. The haplotype variant of the FLK test (hapFLK) revealed strong signature on chromosome 6 as was recorded by iHS, while several other regions on multiple chromosomes showed suggestive signals (**Figure 4.7**). Of note were signatures on chromosomes 10, 16, 17, and 22 occurring in genes associated with food/drug metabolism. We observed signals on chromosome 10 associated with missense variants on the *ACSM6* gene associated with acetyl coenzyme-A production ($p = 1.62 \times 10^{-06}$), and on the *CYP2C8* gene, a cytochrome P450 superfamily enzyme member associated with drug metabolism ($p = 6.09 \times 10^{-06}$). Multiple missense variant signals were also observed in the *ABCC11/12* gene on chromosome 16 ($p = 5.39 \times 10^{-07}$), an ATP binding cassette subfamily member involved in multi-drug resistance. In addition, signals were observed on the *MTTP* gene on chromosome 4 ($p = 3.73 \times 10^{-06}$) involved in triglyceride transfer and lipoprotein assembly, the *TMEM199* gene on chromosome 17 ($p = 5.35 \times 10^{-06}$) whose deficiency is associated with abnormal glycosylation (Jansen et al., 2016), and the *TCN2* gene on chromosome 22 ($p = 6.76 \times 10^{-06}$) involved in the absorption of vitamin B12 (cobalamin). The genome-wide significance threshold was estimated at 6.02×10^{-08}.

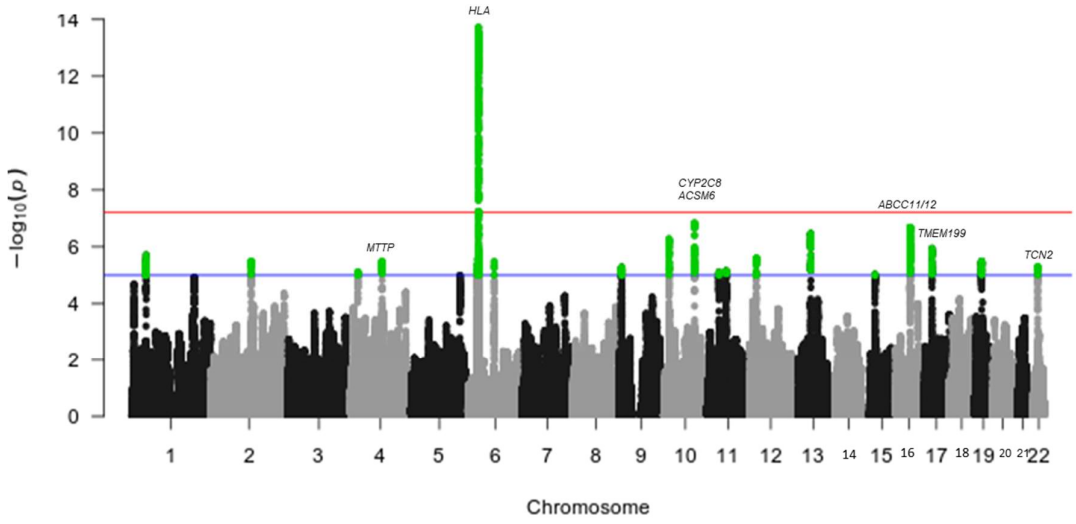


Figure 07: Manahattan plot of hapFLK results: Genome-wide significance threshold (red line), suggestive threshold (blue line).

4.2.6. *HBB* haplotypes among Cameroonian ethnic groups

The high alternate allele frequencies of the SNPs meant that they were imputed with high accuracy (average r^2 = 0.97) (**Table 4.2**). Generally, the 'base' population harbored a higher haplotypic diversity than the HbS chromosome-bearing population (**Figure 4.8a-f**). There was a substantial decrease in haplotype diversity in the HbS chromosome-bearing populations, confirming longstanding knowledge that malaria has been a major force on the human genome. The haplotype frequencies were also largely expected; the BEN is the most prevalent worldwide, the AI was only recently reported in Cameroon (Ngo Bitoungui et al., 2015), and has only been reported again in Egypt (Abou-Elew et al., 2018) and Mauritania (Veten et al., 2012), and predicted in a single chromosome in Kenya (Shriner & Rotimi, 2018). An interesting observation was the absence of OT3 haplotype in the haplotypic background of the HbS chromosome-bearing Semi-Bantu population (**Figure 4.8f**). Two non-classical haplotypes (OT1 and OT2) persisted in the background of all HbS chromosome-bearing populations.

Table 02. *HBB* gene cluster haplotypes in Cameroonians

Haplotype Name	rs3834466 (*HBE1* - HincII)	rs28440105 (*HBG1* - HindIII)	rs10128556 (*HBBP1* - HincII)	rs968857 (*HBBP1* - HincII)	Haplotype
SNP (aaf/R^2)	0.20/0.99	0.87/0.96	0.12/0.94	0.20/0.99	-
AI	GT (1)	C (1)	T (1)	T (0)	1 1 1 0
SEN	G (0)	C (1)	T (1)	T (0)	0 1 1 0
BEN	G (0)	C (1)	C (0)	T (0)	0 1 0 0
CAR	G (0)	C (1)	C (0)	C (1)	0 1 0 1
CAM	G (0)	A (0)	C (0)	T (0)	0 0 0 0
OT1	GT (1)	C (1)	C (0)	T (0)	1 1 0 0
OT2	GT (1)	C (1)	C (0)	C (1)	1 1 0 1
OT3	G (0)	C (1)	T (1)	C (1)	0 1 1 1
OT4	GT (1)	A (0)	C (0)	T (0)	1 0 0 0
OT5	G (0)	A (0)	T (1)	T (0)	0 0 1 0
OT6	G (0)	A (0)	T (1)	C (1)	0 0 1 1
OT7	GT (1)	A (0)	C (0)	C (1)	1 0 0 1
OT8	G (0)	A (0)	C (0)	C (1)	0 0 0 1
OT9	GT (1)	C (1)	T (1)	C (1)	1 1 1 1

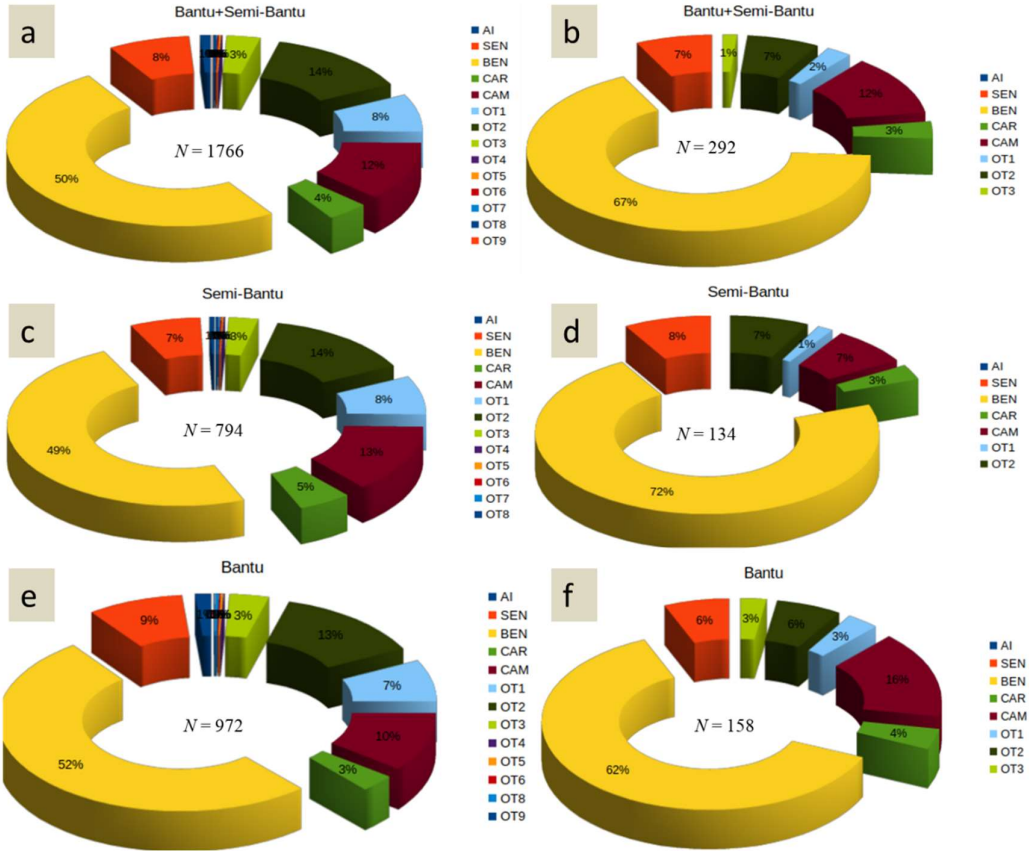


Figure 08: HBB gene cluster haplotype distribution in Cameroonians: The haplotypes are reported for the Bantu and Semi-Bantu only. Left Panel represents haplotype frequencies from individuals carrying no HbS-chromosomes. Right panel represents haplotype frequencies from individuals carrying at least one HbS-bearing chromosome. N = number of chromosomes analyzed for each set.

4.3. Association analysis:

4.3.1. Genotype Imputation performance

Upon alignment of the data to the 1KGP3, ~2% of the SNPs were absent from the 1KGP3 reference panel. A similar observation was made with the Michigan Imputation (MI) server, while the TOPMed Imputation (TI) panel lacked ~4% of the SNPs. The disparate reference allele overlap was reflected by the low squared correlation (r^2) between reference allele frequencies for the TI panel as compared to the 1KGP3 panel. Interestingly, the low reference allele overlap of the TI panel did not appear to hinder its performance as it outperformed the MI and my in-house imputation

strategy, and particularly so at low frequency variants (**Figure 4.9**). TI imputation accuracy was notably better than the recently published MalariaGEN performance. A slightly better imputation performance for the in-house strategy as compared to the MI and MalariaGEN imputation strategies was also observed.

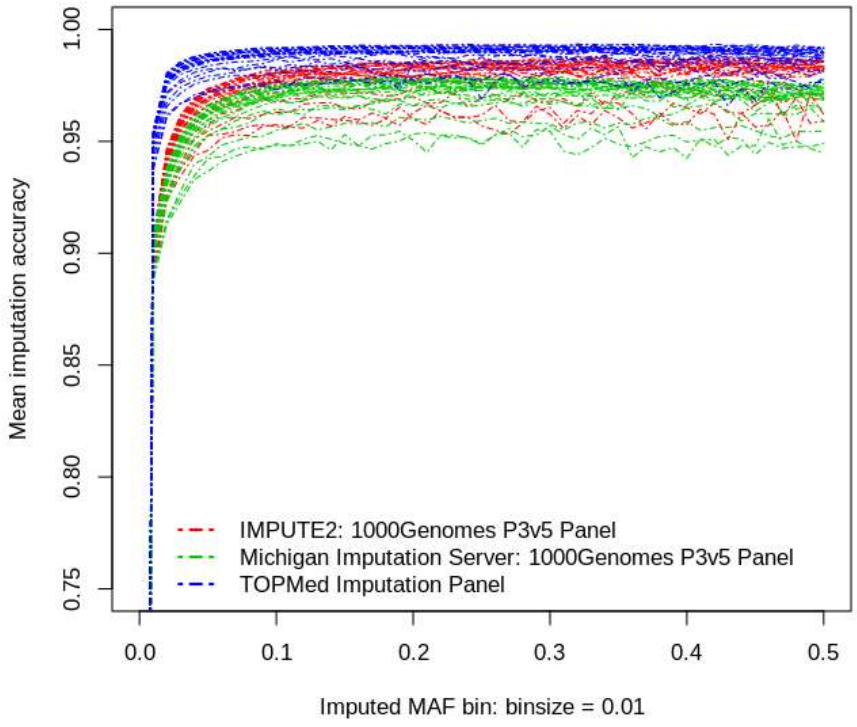


Figure 09: Imputation accuracy*:* Per-chromosome imputation accuracy of our in-house, TOPMed, and Michigan imputation strategies. P3v5 = phase three version 5

4.3.2. Heritability estimation

Narrow sense heritability/pseudo-heritability (heritability contributed by additive genetic variance component) was estimated 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 ~22% using EMMAX in the pre-imputation dataset, consistent with previous estimates (~23%) of severe malaria heritability (Damena & Chimusa, 2020; Malaria Genomic Epidemiology Network, 2019). It thus served as another layer of quality check for the data with respect to case-control ascertainment. The total genetic variants were predicted to explain not more than 9% of the heritability.

4.3.3. Pre-imputation association analysis

Table 4.3 shows candidate loci for which significant and/or marginally significant variants for the pre-imputation set were observed. A single significant variant, rs113508623 in the intergenic region of *CHST15* (distance = 6,400) and *OAX* (distance = 226,266) was observed in the pooled population (*P* = 1.04e-08; OR = 0.42; 95%CI = 0.31 – 0.56; false discovery rate [FDR] = 0.002; genomic control inflation factor [λ_{GC}] = 1.02). This variant was the most significant in the Bantu population (cases = 204, controls = 272, males = 299, females = 177) (**Table 4.3**), while the chromosome 6 variant, rs2842958, on the *SOD2* gene was the most significant in the Semi-Bantu (cases = 262, controls = 291, males = 311, females = 242). No X chromosome associations were observed.

Table 03: Candidate associated loci before imputation

rsid	chr:pos	ref/alt	aaf	p-value (PLINK2) (P_{BH})	OR (95%CI)	Nearest gene	mode	EMMAX	BOLT-LMM	GCTA
					BSB					
rs113508623	10:125859606	T/C	0.18	**2.19e-08 (0.03)**	0.48 (0.37 – 0.62)	*CHST15*	add/het/ dom	**3.79e-08 (0.06)**	**2.2e-08 (0.014)**	**9.82e-08 (0.16)**
rs73547455	11:90154919	A/G	0.04	4.90e-06 (0.98)	0.28 (0.16 – 0.48)	*DISC1FP1*	add/dom	1.35e-06 (0.80)	6.2e-07 (0.52)	9.68e-06 (0.99)
rs7333739	13:71546558	G/A	0.07	2.43e-06 (0.98)	2.26 (1.61 – 3.17)	*LINC00348*	add	2.38e-06 (0.81)	4e-07 (0.36)	4.74e-06 (0.99)
					SB					
rs1172909754	6:160133957	A/G	0.33	1.15e-07 (0.20)	0.36 (0.25 - 0.53)	*SOD2*	dom/het	-	-	-
rs2758352	6:160122921	G/A	0.33	1.22e-07 (0.20)	0.36 (0.25 - 0.53)	*SOD2*	dom/het	-	-	-
rs4902123	14:62746807	A/G	0.41	4.19e-07 (0.79)	2.12 (1.58 - 2.83)	*LOC105370529*	add	4.59e-06 (0.62)	0.01 (0.95)	2.89e-04 (0.99)
					BA					
rs113508623	10:125859606	T/C	0.18	1.64e-06 (0.98)	0.32 (0.20 - 0.51)	*CHST15*	dom	3.44e-06 (0.65)	1.9e-06 (0.56)	3.54e-05 (0.99)
rs73351115	14:96920573	T/G	0.11	2.14e-05 (0.98)	2.69 (1.71 - 4.25)	*AK7*	add	6.13e-07 (0.51)	1.3e-06 (0.56)	4.85e-05 (0.99)
rs13355489	5:156664822	T/C	0.35	8.59e-07 (0.99)	0.32 (0.20 - 0.50)	*ITK*	dom	-	-	-
rs507723	15:46943055	C/T	0.45	**6.20e-08** (0.99)	0.32 (0.21 - 0.48)	*LOC105370803*	dom	-	-	-

4.3.4. Post-imputation association testing

In the imputed set filtered to exclude variants with imputation accuracy $r^2 < 0.65$, MAF < 1%, and genotype probability < 90%, P_{hwe} = 1e-20, the most significant variants (with genome-wide significance) occurred on the *SOD2* gene in Semi-Bantu

individuals and the variants associated with apparently strong malaria protection (**Figure 4.10 & Table 4.4**); rs2842958 being the most significant (*P* = 3.85e-09; OR = 0.31; 95%CI = 0.21 – 0.45; [FDR] = 0.011; λ_{GC} = 0.99). In the Bantu, the chromosome 14 locus *AK7* harbored the most significant signals while several other intergenic and non-genic regions harbored suggestive signals. Multiple loci with significant and/or marginally significant signals were observed in the pooled dataset. The loci observed in the pre-imputation set (notably *CHST17*) were also observed in the imputed set. Of note was the absence of association at the HbS locus in all the analysis sets. Multiple marginally significant variants were identified from the TOPMed imputed set.

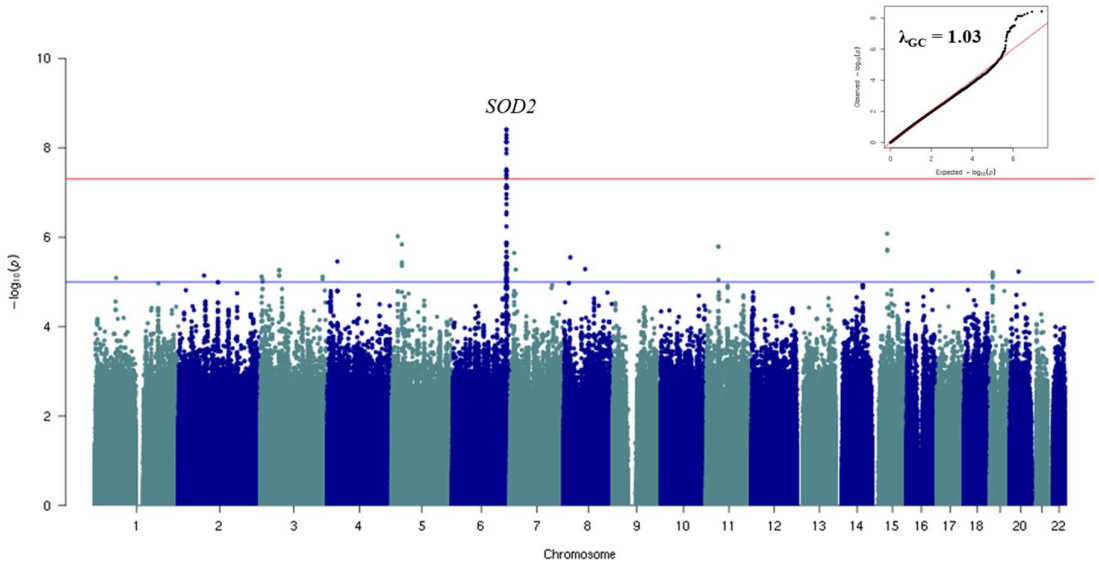


Figure 010: Manhattan plot of association signal in Semi-Bantu individuals: This result was obtained under a dominant model of inheritance at the SOD2 locus. Red line = genome-wide significance (5e-08); blue line = suggestive line (1e-05). Small insert QQ plot shows genomic control inflation factor indicating that test statistics are not inflated.

Table 04: Candidate associated loci after imputation

rsid	chr:pos	ref/ alt	aaf	p-value (PLINK2) (P_{BH})	OR (95%CI)	gene(s)	mode	EMMAX	BOLT-LMM	GCTA
					BSB					
rs514788	11:90065530	T/G	0.04	7.18e-07 (0.98)	0.39 (0.27 – 0.56)	*DISC1FP1*	add/ dom	1.78e-07 (0.60)	1.7e-07 (0.61)	4.01e-06 (0.99)
rs112400941	9:18702156	G/A	0.05	1.20e-06 (0.98)	0.28 (0.16 – 0.46)	*ADAMTSL1*	add	2.85e-07 (0.60)	2.6e-07 (0.61)	6.53e-06 (0.99)
rs113508623	10:125859606	T/C	0.19	1.06e-06 (0.98)	0.54 (0.42 – 0.69)	*CHST15*	add/ het/dom	9.07e-07 (0.60)	9.2e-07 (0.61)	4.18e-06 (0.99)
rs7333739	13:71546558	G/A	0.09	2.43e-06 (0.98)	2.26 (1.61 – 3.17)	*LINC00348*	add/ dom	6.80e-06 (0.96)	7.4e-06 (0.95)	7.02e-06 (0.99)
rs114296724	10:54501974	C/G	0.10	3.52e-07 (0.97)	0.39 (0.27 – 0.56)	*LOC105378305*	add/ dom	2.26e-07 (0.60)	2.7e-07 (0.61)	5.66e-06 (0.99)
rs8083681	18:46218584	A/G	0.77	1.14e-06 (0.98)	1.80 (1.41 – 2.26)	*CTIF*	add	1.28e-06 (0.76)	1.7e-06 (0.95)	5.05e-06 (0.99)
rs529559040	3:34040515	A/G	0.02	2.48e-06 (0.98)	8.05 (3.38 – 19.2)	*LINC01811*	add	3.97e-07 (0.60)	4.2e-07 (0.61)	1.29e-06 (0.99)
rs116089757	8:70562161	T/C	0.04	5.21e-06 (0.98)	3.49 (2.04 - 5.98)	*SULF-1*	add	1.91e-06 (0.89)	2.2e-06 (0.95)	9.43e-07 (0.99)
rs776128	3:77683213	C/T	0.85	4.13e-08 (0.33)	0.39 (0.28 - 0.55)	*ROBO2*	dom/ het	-	-	-
					SB					
rs2842958	6:160108425	A/G	0.68	3.85e-09 (0.01)	0.31 (0.21 - 0.45)	*SOD2*	dom	-	-	-
rs9456440	6:160074463	G/A	0.67	8.045e-08 (0.42)	0.44 (0.32 - 0.59)	*SOD2*	het/add	3.45e-08 (0.25)	6.4e-08 (0.33)	3.69e-07 (0.99)
rs1801253	10:115805056	G/C	0.58	5.72e-07 (0.43)	0.47 (0.35 - 0.63)	*ADRB1*	add	3.52e-07 (0.27)	3.1e-07 (0.33)	9.69e-06 (0.99)
rs80178114	11:36394268	C/T	0.19	6.19e-07 (0.43)	2.55 (1.76 - 3.67)	*PRR5L*	add	2.71e-07 (0.26)	3.3e-07 (0.33)	4.81e-06 (0.99)
rs4902123	14:62746807	C/A	0.40	7.55e-07 (0.43)	2.12 (1.57 - 2.85)	*SYT16*	add	1.31e-06 (0.41)	1.4e-06 (0.52)	1.50e-05 (0.99)
					BA					
rs111614250	14:96930036	G/A	0.10	7.89e-07 (0.95)	0.32 (0.20 - 0.51)	*AK7*	add	-	-	-

4.3.5. Annotation of top signals of association

Annotation revealed that the most significantly enriched pathway was the interleukin 12- (*IL12*) signaling pathway (FDR = 0.0069) involving two genes (*EPB41L4A* and *CCDC144NL*) in the Semi-Bantu only. No hits were observed for the Bantu only or Semi-Bantu+Bantu joint set.

CHAPTER FIVE

DISCUSSION

This current study explored population structure in three ethnic populations within three regions in Cameroon and genetic polymorphisms that may be associated with symptomatic malaria. The study showed that there is significant genetic structure within Cameroon characterized by allele frequency differences among the ethnic groups, distinct clustering patterns, and significant signatures of selection, and by utilizing this information, the study revealed novel malaria resistance loci in the genomes of Cameroonian Semi-Bantu individuals. The meaning of these findings and implication for future genetic studies in Cameroon and Africa are herein discussed.

Population structure analysis revealed extensive genetic structure among the studied ethnicities that has not been previously captured. The mild differences that were observed in pairwise genetic distance (F_{ST}) among Cameroonian ethnicities indicate that the populations have been extensively mixed. Their differences support the existence of distinct ancestral proportions. The FO population belongs to the Sudanese ethnic division with northern African lineage dating back to the ancient Sao civilization that flourished around the shores of Lake Chad around 9^{th}-15^{th} century AD, and a Hausa-Fulani land invasion from Nigeria by the 1800s that led to the establishment of a large Islamic empire involving much of the northern regions of Cameroon related (J. Fearon & Laitin, 2005). The BA, are thought to have been the earliest inhabitants of Cameroon, with traces of their ancient civilization still prominent in the pigmies of the South and East. Some studies have associated the spread of the BA ancestral proportions found in Central, South and East Africa to a Bantu expansion that originated somewhere around South Western Cameroon (Busby et al., 2016; Grollemund et al., 2015; Lipson et al., 2020). The SB individuals mainly inhabit the Western highlands and grass fields of the West and North West of Cameroon. Together with the BA of the South West, the SB of the North West also appear to have endured a four-decade complex cohabitation with Eastern Nigerian populations during colonial era (Gardinier et al., 2001). Therefore, the ancient interactions, and interactions of the recent past of Cameroonian populations with other populations may have paved the way for substantial genetic admixture and drift.

The close similarity in ancestral proportions among the BA and SB was expected. The dissimilarities may be attributable to many factors including varying degrees of contact with different external populations with subsequent genetic drift as could have been possible during their pre-colonial and colonial era as well as following different selective pressures. The considerable dissimilarity in the ancestral proportion of the FO ethnicity from the others is not surprising. However, their splitting into two distinct clusters may provide evidence of genetic heterogeneity within the ethnic group. The separation of the FO ethnicity into two distinct clusters by chromosome painting, one close to the BA and SB and the other quite apart, further indicate that the FO is not homogeneous.

Considering that the genetic differences among the ethnicities would characterize different axes of genetic variation in an association study, one would expect a significant dose of false positive results when all the ethnicities are analyzed together. Therefore, larger sample sizes would be required for association studies in such a highly structured population to be sufficiently powered to identify markers associated with specific phenotypes. Hence, association analysis performed on each ethnicity separately would be more profitable given that ethnic information is accurately captured.

Population genetic approaches that measure genetic distance and quantify shared ancestry are more robust when SNPs are ascertained to be polymorphic in an out-group (Skoglund et al., 2017). However, out-group ascertainment in African populations remains a challenge as the "most recent common ancestor" (MRCA) of African populations remains to be established. Although the roots of anatomically modern humans have recently been traced to Botswana (Chan et al., 2019), the Mende population from Sierra Leone (MSL), shown to harbor the largest proportion of ancestry from a basal West African lineage (Skoglund et al., 2017) fitted well as an out-group in this analysis. A couple of test analyses supported this observation; i) pairwise F_{ST} estimates with the 1000 Genomes populations without SNP ascertainment required either over a million SNPs or > 50,000 SNPs with MAF > 0.35 to observe estimates similar to those previous reported, ii) SNP ascertainment with all African populations except the MSL did not result in F_{ST} estimates as have been previously

reported iii) Finally, SNP ascertainment in the MSL population resulted in similar F_{ST} estimates as have been previously reported using less than a million SNPs with MAF as low as 0.05. Moreover, the MSL population has been previously estimated to have differentiated ~ 300 thousand years ago (ka) – 200ka (Skoglund et al., 2017), about the same time modern humans are thought to have originated from Shum Laka (modern day Cameroon) and Botswana (Southern Africa) (Chan et al., 2019). Hence, in the absence of a publicly available and well-established out-group for African populations, the usage of the MSL population may serve such a purpose.

Signatures of selection observed in this analysis suggest that Cameroonian populations have come under strong disease pressure. The strong signatures targeted primarily immune response and food/drug metabolism genes, suggestive of polygenic adaptation of the population to diseases and changes in diet. Selection, therefore, acts on multiple loci across multiple genes to simultaneously drive phenotypic adaptation (Skoglund et al., 2017), although one would expect, in principle, the core locus affecting a particular trait to be under a selective sweep. Indeed, African populations have had to endure immense pressure from infectious diseases being the oldest populations of anatomically modern humans (formerly hunter-gatherers) (Busby et al., 2016). The effect of the sickle disease on malaria has been well established. Both conditions are thought to have emerged around the same time (4000 - 5000 years ago) coinciding with the adoption of agriculture in Central Africa (Phillipson, 2006). Recent studies however suggest that both malaria emerged earlier ~40,000 – 60,000 years ago, while the sickle cell mutation emerged around 22,000 years ago (Esoh & Wonkam, 2021).

However, such retrospective assessments of the genetic differences among Cameroonian ethnicities and other populations with respect to their demographic histories has limitations; First, the analysis relied on self-reported ethnicity of the father and mother of each participant which may not have been accurate. Nevertheless, these results highlight key differences in the genetic architecture of Cameroonian ethnicities that may have significant bearing on genetic association studies for this population. Informed by population structure analysis, the association strategy herein employed has uncovered potentially novel malaria-associated loci in Cameroonian

individuals. At first glance, the loci, particularly *CHST15* and *SOD2* seem appealing as plausible associations given their biological implication.

CHST15 (carbohydrate sulfotransferase 15; 10q26.13) also known as B-cell Rag-associated gene (*BRAG*) for its co-expression with *RAG1* (recombination-activating gene 1) in B-cells is a type II trans-membrane glycoprotein that plays the following roles; i) induces *RAG1* expressions in B-cell lines (Verkoczy et al., 1998), ii) serves as signaling receptor on the surface of unstimulated mature B-cells (Verkoczy et al., 2000), and iii) possesses sulfotransferase enzymatic activity whereby it catalyzes the transfer of sulfate residues to chondroitin sulfate A (CSA) and dermatan sulfate (DS) (Ohtake et al., 2001). When it transfers sulfates to the C-4 and C-6 hydroxyl groups of CSA, it forms CSE. Recall that CSA is the receptor of choice in the placenta for the *P. falciparum* erythrocyte membrane protein-1 (PfEMP-1) encoded by the *VAR2CSA* gene (Salanti et al., 2004).

Interestingly, *CHST15* is highly expressed in fetal and adult spleen, peripheral blood leukocytes, and lymph node, with modest expression in the heart, ovary, stomach, and brain (Verkoczy et al., 2000). Therefore, given malaria pressure, one can imagine a model in which the gene expression is augmented in one of such tissues as the spleen or ovary with a corresponding increase in sulfotransferase activity, and a concomitant reduction in CSA molecules, thus effectively protecting against pregnancy-associated malaria (PAM). Furthermore, its ability to induce *RAG1* expression may be a means to equip the fetus (and adults perhaps) with the ability to utilize the enormous repertoire of antibody specificities that come with V(D)J recombination (Yu et al., 1999) in order to mount effective defense mechanisms against the malaria parasite. The variant was observed outside of the major *CHST15* gene structure (6492 bp upstream), its derived allele (T) is fixed in most populations outside of Africa, while the ancestral allele (C) is most common in the Esan population of Nigeria (http://www.ensembl.org). There could well be other tag SNPs nearer, or even within the gene that could be uncovered by larger sample sizes. Besides, gene expression under stringent conditions showed that *CHST15* is only expressed in human and baboon DNA, highlighting its high conservation and potential importance

(Verkoczy et al., 2000). Expression studies with further genetic investigations may shed some light.

The role of reactive oxygen and nitrogen species (ROS and RNS respectively) in malaria has been extensively documented and reviewed in (Kavishe et al., 2017). Host defense 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 as tumor necrosis factor alpha (TNF-alpha) 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, a suggestive signal was observed at a TNF variant (rs1800750) in this same population which apparently increased susceptibility to malaria (Apinjoh et al., 2014), suggestive of inflammatory responses and prolonged oxidative stress. Superoxide dismutase 2 (SOD2), a mitochondrial matrix enzyme encoded in nuclear DNA (6q25.3) and highly expressed in many organs including the liver where *Plasmodium* parasites base a significant portion of their cycle, is an effective scavenger of ROS, preventing excess oxidative stress. A great majority of the signals observed in this locus were protective. It would therefore be interesting to investigate further its role in modulating malaria phenotypes in this population.

Although further analysis involving larger sample sizes would generally be required to confirm the contribution of these loci to malaria phenotypic variance in Cameroonians, this analysis serves as a pointer to such studies. Association at the HbS (rs334) locus could not be successfully replicated. This could be due, in part, to pooling the sub-phenotypes together as the variant is known to be most protective against severe malaria. Because the variant was not genotyped in this dataset, and this study depended upon imputation to access it, failure to observe its association may also have been due to the post-imputation filtering schemes wherein it was observed that applying a threshold for genotype probability at 95% excluded the HbS allele.

Subsequent Cameroon-specific GWASs that directly type the variant may be profitable. However, substantial differences in the haplotypic backgrounds of the ethnic groups at the *HBB* gene cluster due to the malaria pressure was observed. The OT3 haplotype was visibly absent from the haplotypic background of HbS-carrying

Semi-Bantu individuals. Because this observation indicates differential evolutionary course of the ethnic groups under malaria pressure, it may serve as a prototype to the genetic architecture at other loci, or perhaps genome-wide. Indeed, population structure analysis showed that the Bantu harbored strong signatures of positive selection at specific loci in the HLA region (*HLA-DPB1*) not present in the Semi-Bantu (Patin et al., 2017). Even the haplotypic backgrounds of the 'base' populations were not entirely identical (OT9 in Bantu and not in Semi-Bantu, while OT7 in Semi-Bantu and not in Bantu). This furthers the knowledge of age old fine-scale genetic structure within Cameroon.

Although the discussion of the specific prevalence and importance of the classical *HBB* gene cluster haplotypes in the various populations in this study falls out of the scope of the current analysis, it must be noted that they can be particular applicable to sickle-cell disease research. For instance, the presence of the AI haplotype in all our 'base' populations and its absence from HbS-carrying populations is particularly interesting given that the haplotype is associated with the most favorable SCD clinical outcomes and would be expected to be prevalent in HbS-carrying chromosomes.

Finally, apart from low sample size, another potential confounding factor in the analysis that could not be controlled with the current data was age. Hence, a larger study that takes into account all these variable would shed more light. The observations herein presented remain useful in informing such future human genetic studies of malaria in Cameroon.

CHAPTER SIX

CONCLUSION AND RECOMMENDATIONS

6.1. Conclusion

This study sought to determine genetic diversity among Cameroonians in three regions and three ethnic groups, and use this information to uncover variants that may be associated with symptomatic malaria susceptibility in Cameroonians, and thus foster the understanding of the disease pathobiology. This in turn is an important first step in identifying novel drug and vaccine targets. The study achieved the underlying objectives by showing that: 1) indeed, fine-scale genetic structure exists within the three regions, and that it affects genetic association signals. Therefore, smarter sampling strategies and analysis designs are needed to effectively detect and significantly minimize/correct population structure effects in Cameroonian, and African populations, 2) potentially novel symptomatic malaria genetic association signals exist in Cameroon that could be uncovered by larger studies, and must therefore be confirmed by future studies. The importance of these findings is that the two significant genes identified open a new door into the investigation of causal SNPs that may explain decreased susceptibility to symptomatic malaria. Subsequent studies should focus on these genes, using candidate gene association techniques and deep sequencing to investigate all variants in these genes, including insertions and deletions (INDELS), structural variants (SVs), and copy number variants (CNVs). This study has thus generated hypotheses for future studies. An important area in which this study has made a significant discovery is in pregnancy associated malaria (PAM) which is most prevalent in central Africa including in Cameroon according to recent WHO reports (WHO, 2020). It is known that the *PfEMP1*-based recombinant VAR2CSA vaccine candidate against *P. falciparum* targets placental malaria (Duffy & Patrick Gorres, 2020). If the *CHST15* enzyme protects against placental malaria by reducing CSA and maintains the normal physiology of the placenta, then it could be an attractive therapeutic target for PAM, for instance by therapeutics that increase its activity.

6.2. Recommendation

Recommendations for future studies in Cameroonian populations

- More samples should be collected and the hypothesis that variants in *SOD2* and *CHST15* confer malaria resistance should be investigated.
- These studies should be candidate gene association and fine-mapping analysis using whole genome sequencing.
- If the samples are collected from different ethnic groups, then genetic associations should be performed on the ethnic groups separately.
- When collecting samples, participant demographic information (ethnicity – mother and father, age, sex, and region) should be accurately captured as much as possible.

Recommendations for similar studies in Africa.

- By comparing different imputation strategies on the new data set informed by population structure, this study confirmed data that showed the superiority of the NHLBI's TOPMed imputation panel over the famed 1000 Genomes reference panel for African populations (Taliun et al., 2019). Also, by observing that our in-house imputation strategy performed slightly better than the MI service and MalariaGEN's large-scale imputation strategy, this study has shown that there are specific benefits to population-specific analysis that are easily lost in multi-site studies. Therefore this study highly recommends that multi-site GWASs be complemented by population-specific analyses.
- The publicly available population of the Mende tribe in Sierra Leone may be used as out-group for human genetic differentiation analysis in Africa while the most recent common ancestor of anatomically modern humans remains to be established
- Heritability estimation could be employed as a method of ascertaining data quality where empirical estimates would suggest accurate case ascertainment while estimates that deviate significantly from empirical values would mean that there is a problem with case definition.

REFERENCES

Abou-Elew, H. H., Youssry, I., Hefny, S., Hashem, R. H., Fouad, N., & Zayed, R. A. (2018). β S globin gene haplotype and the stroke risk among Egyptian children with sickle cell disease. *Hematology*, *23*(6), 362–367. https://doi.org/10.1080/10245332.2017.1403736

Acharya, P., Garg, M., Kumar, P., Munjal, A., & Raja, K. D. (2017). Host-parasite interactions in human malaria: Clinical implications of basic research. *Frontiers in Microbiology*, *8*(MAY), 1–16. https://doi.org/10.3389/fmicb.2017.00889

Achidi, E A, Apinjoh, T. O., Mbunwe, E., Besingi, R., Yafi, C., Awah, N. W., Ajua, A., & Anchang, J. K. (2008). Febrile status, malarial parasitaemia and gastro-intestinal helminthiases in schoolchildren resident at different altitudes, in south-western Cameroon. *ANNALS OF TROPICAL MEDICINE AND PARASITOLOGY*, *102*(2), 103–118. https://doi.org/10.1179/136485908X252287

Achidi, Eric A., Apinjoh, T. O., Anchang-Kimbi, J. K., Mugri, R. N., Ngwai, A. N., & Yafi, C. N. (2012). Severe and uncomplicated falciparum malaria in children from three regions and three ethnic groups in Cameroon: prospective study. *MALARIA JOURNAL*, *11*(1), 215. https://doi.org/10.1186/1475-2875-11-215

Alexander, D. H., Novembre, J., & Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. *Genome Research*, *19*(9), 1655–1664. https://doi.org/10.1101/gr.094052.109

Allison, A. C. (1954). Protection Afforded by Sickle-cell Trait Against Subtertian Malarial Infection. *BMJ*, *1*(4857), 290–294. https://doi.org/10.1136/bmj.1.4857.290

Altshuler, D. L., Durbin, R. M., Abecasis, G. R., Bentley, D. R., Chakravarti, A., Clark, A. G., Collins, F. S., De La Vega, F. M., Donnelly, P., Egholm, M., Flicek, P., Gabriel, S. B., Gibbs, R. A., Knoppers, B. M., Lander, E. S., Lehrach, H., Mardis, E. R., McVean, G. A., Nickerson, D. A., … Peterson, J. L. (2010). A map of human genome variation from population-scale sequencing.

Nature, *467*(7319), 1061–1073. https://doi.org/10.1038/nature09534

Amegashie, E. A., Amenga-Etego, L., Adobor, C., Ogoti, P., Mbogo, K., Amambua-Ngwa, A., & Ghansah, A. (2020). Population genetic analysis of the Plasmodium falciparum circumsporozoite protein in two distinct ecological regions in Ghana. *Malaria Journal*, *19*(1), 1–14. https://doi.org/10.1186/s12936-020-03510-3

Antonio-Nkondjio, C., Ndo, C., Njiokou, F., Bigoga, J. D., Awono-Ambene, P., Etang, J., Ekobo, A. S., & Wondji, C. S. (2019). Review of malaria situation in Cameroon: technical viewpoint on challenges and prospects for disease elimination. *PARASITES & VECTORS*, *12*(1), 501. https://doi.org/10.1186/s13071-019-3753-8

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., Apinjohr, T. O., Anchang-Kimbi, J. K., Njua-Yafi, C., Mugri, R. N., Ngwai, A. N., Rockett, K. A., Mbunwe, E., Besingi, R. N., … Consortium, M. (2013). Association of cytokine and toll-like receptor gene polymorphisms with severe malaria in three regions of cameroon. *PLoS ONE*, *8*(11), e81071. https://doi.org/10.1371/journal.pone.0081071

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. (2014). Association of candidate gene polymorphisms and TGF-beta/IL-10 levels with malaria in three regions of Cameroon: A case-control study. *Malaria Journal*, *13*(1), 236. https://doi.org/10.1186/1475-2875-13-236

Awono-Ambene, P. H., Etang, J., Antonio-Nkondjio, C., Ndo, C., Eyisap, W. E., Piameu, M. C., Mandeng, E. S. E. S. E. S., Mbakop, R. L., Toto, J. C., Patchoke, S., Mnzava, A. P., Knox, T. B., Donnelly, M., Fondjo, E., Bigoga, J. D., & D Bigoga, J. (2018). The bionomics of the malaria vector Anopheles rufipes Gough, 1910 and its susceptibility to deltamethrin insecticide in North Cameroon. *Parasites & Vectors*, *11*(1), 253. https://doi.org/10.1186/s13071-018-2809-5

Band, G., Rockett, K. A., Spencer, C. C. A., Kwiatkowski, D. P., Si Le, Q., Clarke, G. M., Kivinen, K., Leffler, E. M., Cornelius, V., Conway, D. J., Williams, T. N., Taylor, T., Bojang, K. A., Doumbo, O., Thera, M. A., Modiano, D., Sirima, S. B., Wilson, M. D., Koram, K. A., … Johnson, K. J. (2015). A novel locus of resistance to severe malaria in a region of ancient balancing selection. *Nature*, *526*(7572), 253–257. https://doi.org/10.1038/nature15390

Belmont, J. W., Hardenbol, P., Willis, T. D., Yu, F., Yang, H., Ch'Ang, L. Y., Huang, W., Liu, B., Shen, Y., Tam, P. K. H., Tsui, L. C., Waye, M. M. Y., Wong, J. T. F., Zeng, C., Zhang, Q., Chee, M. S., Galver, L. M., Kruglyak, S., Murray, S. S., … Tanaka, T. (2003). The international HapMap project. *Nature*, *426*(6968), 789–796. https://doi.org/10.1038/nature02168

Benjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, *57*(1), 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

Bhatia, G., Patterson, N., Pasaniuc, B., Zaitlen, N., Genovese, G., Pollack, S., Mallick, S., Myers, S., Tandon, A., Spencer, C., Palmer, C. D., Adeyemo, A. A., Akylbekova, E. L., Cupples, L. A., Divers, J., Fornage, M., Kao, W. H. L., Lange, L., Li, M., … Price, A. L. (2011). Genome-wide comparison of African-ancestry populations from CARe and other cohorts reveals signals of natural selection. *American Journal of Human Genetics*, *89*(3), 368–381. https://doi.org/10.1016/j.ajhg.2011.07.025

Bonhomme, M., Chevalet, C., Servin, B., Boitard, S., Abdallah, J., Blott, S., & SanCristobal, M. (2010). Detecting Selection in Population Trees: The Lewontin and Krakauer Test Extended. *Genetics*, *186*(1), 241–262. https://doi.org/10.1534/genetics.104.117275

Bradić, M., Costa, J., & Chelo, I. M. (2012). Genotyping with Sequenom. *Methods in Molecular Biology (Clifton, N.J.)*, *772*, 193–210. https://doi.org/10.1007/978-1-61779-228-1_11

Browning, B. (2016). *conform-gt*. Retrieved from https://faculty.washington.edu/browning/conform-gt.html#example

Buniello, A., Macarthur, J. A. L., Cerezo, M., Harris, L. W., Hayhurst, J., Malangone, C., McMahon, A., Morales, J., Mountjoy, E., Sollis, E., Suveges, D., Vrousgou, O., Whetzel, P. L., Amode, R., Guillen, J. A., Riat, H. S., Trevanion, S. J., Hall, P., Junkins, H., … Parkinson, H. (2019). The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Research*, *47*(1), D1005–D1012. https://doi.org/10.1093/nar/gky1120

Busby, G. B., Band, G., Si Le, Q., Jallow, M., Bougama, E., Mangano, V. D., Amenga-Etego, L. N., Enimil, A., Apinjoh, T., Ndila, C. M., Manjurano, A., Nyirongo, V., Doumba, O., Rockett, K. A., Kwiatkowski, D. P., & Spencer, C. C. (2016). Admixture into and within sub-Saharan Africa. *ELife*, *5*, 1–44. https://doi.org/10.7554/eLife.15266

CDC. (2020). *About Malaria - Biology*. Centers for Disease Control and Prevention. https://www.cdc.gov/malaria/about/biology/index.html

Chan, E. K. F., Timmermann, A., Baldi, B. F., Moore, A. E., Lyons, R. J., Lee, S. S., Kalsbeek, A. M. F., Petersen, D. C., Rautenbach, H., Förtsch, H. E. A., Bornman, M. S. R., & Hayes, V. M. (2019). Human origins in a southern African palaeo-wetland and first migrations. *Nature*, *575*(7781), 185–189. https://doi.org/10.1038/s41586-019-1714-1

Chang, C. C., Chow, C. C., Tellier, L. C., Vattikuti, S., Purcell, S. M., & Lee, J. J. (2015). Second-generation PLINK: Rising to the challenge of larger and richer datasets. *GigaScience*, *4*(1), 7. https://doi.org/10.1186/s13742-015-0047-8

Chen, S.-Y., Feng, Z., & Yi, X. (2017). A general introduction to adjustment for multiple comparisons. *Journal of Thoracic Disease*, *9*(6), 1725–1729. https://doi.org/10.21037/jtd.2017.05.34

Clarke, G. M., Rockett, K., Kivinen, K., Hubbart, C., Jeffreys, A. E., Rowlands, K., Jallow, M., Conway, D. J., Bojang, K. A., Pinder, M., Usen, S., Sisay-Joof, F.,

Sirugo, G., Toure, O., Thera, M. A., Konate, S., Sissoko, S., Niangaly, A., Poudiougou, B., … Kwiatkowski, D. P. (2017). Characterisation of the opposing effects of G6PD deficiency on cerebral malaria and severe malarial anaemia. *ELife*, *6*. https://doi.org/10.7554/eLife.15085

Cowman, A. F., Healer, J., Marapana, D., & Marsh, K. (2016). Malaria: Biology and Disease. In *Cell* (Vol. 167, Issue 3, pp. 610–624). Cell Press. https://doi.org/10.1016/j.cell.2016.07.055

D. Turner, S. (2018). qqman: an R package for visualizing GWAS results using Q-Q and manhattan plots. *Journal of Open Source Software*, *3*(25), 731. https://doi.org/10.21105/joss.00731

Damena, D., & Chimusa, E. R. (2020). Genome-wide heritability analysis of severe malaria resistance reveals evidence of polygenic inheritance. *Human Molecular Genetics*, *29*(1), 168–176. https://doi.org/10.1093/hmg/ddz258

Dara, A., Drábek, E. F., Travassos, M. A., Moser, K. A., Delcher, A. L., Su, Q., Hostelley, T., Coulibaly, D., Daou, M., Dembele, A., Diarra, I., Kone, A. K., Kouriba, B., Laurens, M. B., Niangaly, A., Traore, K., Tolo, Y., Fraser, C. M., Thera, M. A., … Silva, J. C. (2017). New var reconstruction algorithm exposes high var sequence diversity in a single geographic location in Mali. *Genome Medicine*, *9*(1), 1–14. https://doi.org/10.1186/s13073-017-0422-4

Das, S., Forer, L., Schönherr, S., Sidore, C., Locke, A. E., Kwong, A., Vrieze, S. I., Chew, E. Y., Levy, S., McGue, M., Schlessinger, D., Stambolian, D., Loh, P. R., Iacono, W. G., Swaroop, A., Scott, L. J., Cucca, F., Kronenberg, F., Boehnke, M., … Fuchsberger, C. (2016). Next-generation genotype imputation service and methods. *Nature Genetics*, *48*(10), 1284–1287. https://doi.org/10.1038/ng.3656

Deress, T., & Girma, M. (2019). Plasmodium falciparum and Plasmodium vivax Prevalence in Ethiopia: A Systematic Review and Meta-Analysis. In *Malaria Research and Treatment* (2019). Hindawi Limited. https://doi.org/10.1155/2019/7065064

dos Santos Francisco, R., Buhler, S., Nunes, J. M., Bitarello, B. D., França, G. S., Meyer, D., & Sanchez-Mazas, A. (2015). HLA supertype variation across populations: new insights into the role of natural selection in the evolution of HLA-A and HLA-B polymorphisms. *Immunogenetics*, *67*(11–12), 651–663. https://doi.org/10.1007/s00251-015-0875-9

Duffy, P. E., & Patrick Gorres, J. (2020). Malaria vaccines since 2000: progress, priorities, products. In *npj Vaccines 5*(1), 1–9. Nature Research. https://doi.org/10.1038/s41541-020-0196-3

Dundas, K., Shears, M. J., Sun, Y., Hopp, C. S., Crosnier, C., Metcalf, T., Girling, G., Sinnis, P., Billker, O., & Wright, G. J. (2018). Alpha-v–containing integrins are host receptors for the *Plasmodium falciparum* sporozoite surface protein, TRAP. *Proceedings of the National Academy of Sciences*, 201719660. https://doi.org/10.1073/pnas.1719660115

Engle-Stone, R., Williams, T. N., Nankap, M., Ndjebayi, A., Gimou, M.-M. M., Oyono, Y., Tarini, A., Brown, K. H., & Green, R. (2017). Prevalence of Inherited Hemoglobin Disorders and Relationships with Anemia and Micronutrient Status among Children in Yaounde and Douala, Cameroon. *Nutrients*, *9*(7), 693. https://doi.org/10.3390/nu9070693

Esoh, K., & Wonkam, A. (2021). Evolutionary history of sickle cell mutation: implications for global genetic medicine. *Human Molecular Genetics*, *00*. https://doi.org/10.1093/hmg/ddab004

Etang, J., Pennetier, C., Piameu, M., Bouraima, A., Chandre, F., Awono-Ambene, P., Marc, C., & Corbel, V. (2016). When intensity of deltamethrin resistance in Anopheles gambiae s.l. leads to loss of Long Lasting Insecticidal Nets bio-efficacy: a case study in north Cameroon. *PARASITES & VECTORS*, *9*. https://doi.org/10.1186/s13071-016-1420-x

Fariello, M. I., Boitard, S., Naya, H., SanCristobal, M., & Servin, B. (2013). Detecting signatures of selection through haplotype differentiation among hierarchically structured populations. *Genetics*, *193*(3), 929–941.

https://doi.org/10.1534/genetics.112.147231

Fearon, J. D. (2003). Ethnic and cultural diversity by country. *Journal of Economic Growth*, *8*(2), 195–222. https://doi.org/10.1023/A:1024419522867

Fearon, J., & Laitin, D. (2005). *Cameroon*.

Fru-Cho, J., Bumah, V. V., Safeukui, I., Nkuo-Akenji, T., Titanji, V. P. K. K., & Haldar, K. (2014). Molecular typing reveals substantial Plasmodium vivax infection in asymptomatic adults in a rural area of Cameroon. *MALARIA JOURNAL*, *13*(1), 170. https://doi.org/10.1186/1475-2875-13-170

Fuchsberger, C., Abecasis, G. R., & Hinds, D. A. (2015). Minimac2: Faster genotype imputation. *Bioinformatics*, *31*(5), 782–784. https://doi.org/10.1093/bioinformatics/btu704

Gardinier, D. E., DeLancey, M. W., & DeLancey, M. D. (2001). Historical Dictionary of the Republic of Cameroon. *The International Journal of African Historical Studies*, *34*(1), 248. https://doi.org/10.2307/3097360

Gautier, M., Klassmann, A., & Vitalis, R. (2017). rehh 2.0: a reimplementation of the R package rehh to detect positive selection from haplotype structure. *Molecular Ecology Resources*, *17*(1), 78–90. https://doi.org/10.1111/1755-0998.12634

Gineau, L., Luisi, P., Castelli, E. C., Milet, J., Courtin, D., Cagnin, N., Patillon, B., Laayouni, H., Moreau, P., Donadi, E. A., Garcia, A., & Sabbagh, A. (2015). Balancing immunity and tolerance: Genetic footprint of natural selection in the transcriptional regulatory region of HLA-G. *Genes and Immunity*, *16*(1), 57–70. https://doi.org/10.1038/gene.2014.63

Grollemund, R., Branford, S., Bostoen, K., Meade, A., Venditti, C., & Pagel, M. (2015). Bantu expansion shows that habitat alters the route and pace of human dispersals. *Proceedings of the National Academy of Sciences of the United States of America*, *112*(43), 13296–13301. https://doi.org/10.1073/pnas.1503793112

HALDANE, J. B. S. (1949). THE RATE OF MUTATION OF HUMAN GENES.

Hereditas, *35*(1), 267–273. https://doi.org/10.1111/j.1601-5223.1949.tb03339.x

Hien, A. S., Soma, D. D., Hema, O., Bayili, B., Namountougou, M., Gnankine, O., Baldet, T., Diabate, A., & Dabire, K. R. (2017). Evidence that agricultural use of pesticides selects pyrethroid resistance within Anopheles gambiae s.l. populations from cotton growing areas in Burkina Faso, West Africa. *PLOS ONE*, *12*(3). https://doi.org/10.1371/journal.pone.0173098

Howes, R. E., Patil, A. P., Piel, F. B., Nyangiri, O. A., Kabaria, C. W., Gething, P. W., Zimmerman, P. A., Barnadas, C., Beall, C. M., Gebremedhin, A., Ménard, D., Williams, T. N., Weatherall, D. J., & Hay, S. I. (2011). The global distribution of the Duffy blood group. *Nature Communications*, *2*(1), 266. https://doi.org/10.1038/ncomms1265

Jallow, M., Teo, Y. Y., Small, K. S., Rockett, K. A., Deloukas, P., Clark, T. G., Kivinen, K., Bojang, K. A., Conway, D. J., Pinder, M., Sirugo, G., Sisay-Joof, F., Usen, S., Auburn, S., Bumpstead, S. J., Campino, S., Coffey, A., Dunham, A., Fry, A. E., … Kwiatkowski, D. P. (2009). Genome-wide and fine-resolution association analysis of malaria in West Africa. *Nature Genetics*, *41*(6), 657–665. https://doi.org/10.1038/ng.388

Jansen, J. C., Timal, S., van Scherpenzeel, M., Michelakakis, H., Vicogne, D., Ashikov, A., Moraitou, M., Hoischen, A., Huijben, K., Steenbergen, G., van den Boogert, M. A. W., Porta, F., Calvo, P. L., Mavrikou, M., Cenacchi, G., van den Bogaart, G., Salomon, J., Holleboom, A. G., Rodenburg, R. J., … Lefeber, D. J. (2016). TMEM199 Deficiency Is a Disorder of Golgi Homeostasis Characterized by Elevated Aminotransferases, Alkaline Phosphatase, and Cholesterol and Abnormal Glycosylation. *American Journal of Human Genetics*, *98*(2), 322–330. https://doi.org/10.1016/j.ajhg.2015.12.011

Jespersen, J. S., Wang, C. W., Mkumbaye, S. I., Minja, D. T., Petersen, B., Turner, L., Petersen, J. E., Lusingu, J. P., Theander, T. G., & Lavstsen, T. (2016). *Plasmodium falciparum var* genes expressed in children with severe malaria encode CIDRα1 domains. *EMBO Molecular Medicine*, *8*(8), 839–850. https://doi.org/10.15252/emmm.201606188

Kang, H. M., Sul, J. H., Service, S. K., Zaitlen, N. A., Kong, S.-Y., Freimer, N. B., Sabatti, C., & Eskin, E. (2010). Variance component model to account for sample structure in genome-wide association studies. *Nature Genetics*, *42*(4), 348–354. https://doi.org/10.1038/ng.548

Kariuki, S. N., Marin-Menendez, A., Introini, V., Ravenhill, B. J., Lin, Y. C., Macharia, A., Makale, J., Tendwa, M., Nyamu, W., Kotar, J., Carrasquilla, M., Rowe, J. A., Rockett, K., Kwiatkowski, D., Weekes, M. P., Cicuta, P., Williams, T. N., & Rayner, J. C. (2020). Red blood cell tension protects against severe malaria in the Dantu blood group. *Nature*, *585*(7826), 579. https://doi.org/10.1038/s41586-020-2726-6

Kariuki, S. N., & Williams, T. N. (2020). Human genetics and malaria resistance. *Human Genetics*, *0123456789*. https://doi.org/10.1007/s00439-020-02142-6

Kavishe, R. A., Koenderink, J. B., & Alifrangis, M. (2017). Oxidative stress in malaria and artemisinin combination therapy: Pros and Cons. *The FEBS Journal*, *284*(16), 2579–2591. https://doi.org/10.1111/febs.14097

Kwiatkowski, D. P. (2005). How Malaria Has Affected the Human Genome and What Human Genetics Can Teach Us about Malaria. *The American Journal of Human Genetics*, *77*(2), 171–192. https://doi.org/10.1086/432519

Langhorne, J., & Duffy, P. E. (2016). Expanding the antimalarial toolkit: Targeting host–parasite interactions. *The Journal of Experimental Medicine*, *213*(2), 143–153. https://doi.org/10.1084/jem.20151677

Lee, T., Park, C.-K., & Ha, S. Y. (2019). Prognostic Role of Apelin Receptor Expression in Hepatocellular Carcinoma Treated With Curative Surgical Resection. *Anticancer Research*, *39*(6), 3025–3031. https://doi.org/10.21873/anticanres.13435

Leffler, E. M., Band, G., Busby, G. B. J., Kivinen, K., Le, Q. S., Clarke, G. M., Bojang, K. A., Conway, D. J., Jallow, M., Sisay-Joof, F., Bougouma, E. C., Mangano, V. D., Modiano, D., Sirima, S. B., Achidi, E., Apinjoh, T. O., Marsh, K., Ndila, C. M., Peshu, N., … Kwiatkowski, D. P. (2017). Resistance to

malaria through structural variation of red blood cell invasion receptors. *Science*, *356*(6343), eaam6393. https://doi.org/10.1126/science.aam6393

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., & Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, *25*(16), 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Lipson, M., Ribot, I., Mallick, S., Rohland, N., Olalde, I., Adamski, N., Broomandkhoshbacht, N., Lawson, A. M., López, S., Oppenheimer, J., Stewardson, K., Asombang, R. N., Bocherens, H., Bradman, N., Culleton, B. J., Cornelissen, E., Crevecoeur, I., de Maret, P., Fomine, F. L. M., … Reich, D. (2020). Ancient West African foragers in the context of African population history. *Nature*, *577*(7792), 665–670. https://doi.org/10.1038/s41586-020-1929-1

Livingstone, F. B. (1971). Malaria and Human Polymorphisms. *Annual Review of Genetics*, *5*(1), 33–64. https://doi.org/10.1146/annurev.ge.05.120171.000341

Loh, P. R., Palamara, P. F., & Price, A. L. (2016). Fast and accurate long-range phasing in a UK Biobank cohort. *Nature Genetics*, *48*(7), 811–816. https://doi.org/10.1038/ng.3571

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., & Price, A. L. (2015). Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nature Genetics*, *47*(3), 284–290. https://doi.org/10.1038/ng.3190

Lyke, K. E., Ishizuka, A. S., Berry, A. A., Chakravarty, S., DeZure, A., Enama, M. E., James, E. R., Billingsley, P. F., Gunasekera, A., Manoj, A., Li, M., Ruben, A. J., Li, T., Eappen, A. G., Stafford, R. E., Kc, N., Murshedkar, T., Mendoza, F. H., Gordon, I. J., … Seder, R. A. (2017). Attenuated PfSPZ Vaccine induces strain-transcending T cells and durable protection against heterologous controlled human malaria infection. *Proceedings of the National Academy of*

Sciences of the United States of America, *114*(10), 2711–2716. https://doi.org/10.1073/pnas.1615324114

Mackinnon, M. J., Mwangi, T. W., Snow, R. W., Marsh, K., & Williams, T. N. (2005). Heritability of malaria in Africa. *PLoS Medicine*, *2*(12), 1253–1259. https://doi.org/10.1371/journal.pmed.0020340

Malaney, P., Sielman, A., & Sachs, J. (2004). The malaria gap. *American Journal of Tropical Medicine and Hygiene*, *71*(2), 141–146. https://doi.org/https://doi.org/10.4269/ajtmh.2004.71.141

Malaria Genomic Epidemiology Network. (2005). *MalariaGEN Consortial Project 1*. https://www.malariagen.net/projects/consortial-project-1

Malaria Genomic Epidemiology Network. (2008a). *Data Release Policy for Genome-wide Association Data*. *July*. www.malariagen.net/home/downloads/16.pdf

Malaria Genomic Epidemiology Network. (2008b). A global network for investigating the genomic epidemiology of malaria. *Nature*, *456*(7223), 732–737. https://doi.org/10.1038/nature07632

Malaria Genomic Epidemiology Network. (2009). *MalariaGEN Consortium Internal Data Management and Access Policy*. *317*(M), 2007–2010.

Malaria Genomic Epidemiology Network. (2019). Insights into malaria susceptibility using genome-wide data on 17,000 individuals from Africa, Asia and Oceania. *Nature Communications*, *10*(1), 5732. https://doi.org/10.1038/s41467-019-13480-z

Manga, L., Mbingue, S., Etoundi, M. N., & Ngollo, M. (1997). Anopheles namibiensis is anthropophilic and widespread in Cameroon. *Medical and Veterinary Entomology*, *11*(4), 409. https://doi.org/10.1111/j.1365-2915.1997.tb00432.x

Manichaikul, A., Mychaleckyj, J. C., Rich, S. S., Daly, K., Sale, M., & Chen, W. M. (2010). Robust relationship inference in genome-wide association studies.

Bioinformatics, *26*(22), 2867–2873. https://doi.org/10.1093/bioinformatics/btq559

Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., McCarthy, M. I., Ramos, E. M., Cardon, L. R., Chakravarti, A., Cho, J. H., Guttmacher, A. E., Kong, A., Kruglyak, L., Mardis, E., Rotimi, C. N., Slatkin, M., Valle, D., Whittemore, A. S., … Visscher, P. M. (2009). Finding the missing heritability of complex diseases. *Nature*, *461*(7265), 747–753. https://doi.org/10.1038/nature08494

Marchini, J., & Howie, B. (2010). Genotype imputation for genome-wide association studies. *Nature Reviews Genetics*, *11*(7), 499–511. https://doi.org/10.1038/nrg2796

Marchini, J., Howie, B., Myers, S., McVean, G., & Donnelly, P. (2007). A new multipoint method for genome-wide association studies by imputation of genotypes. *Nature Genetics*, *39*(7), 906–913. https://doi.org/10.1038/ng2088

Massoda Tonye, S. G., Kouambeng, C., Wounang, R., & Vounatsou, P. (2018). Challenges of DHS and MIS to capture the entire pattern of malaria parasite risk and intervention effects in countries with different ecological zones: the case of Cameroon. *Malaria Journal*, *17*(1), 156. https://doi.org/10.1186/s12936-018-2284-7

Mbacham, W. F., Ayong, L., Guewo-Fokeng, M., & Makoge, V. (2019). Current Situation of Malaria in Africa. *Methods in Molecular Biology (Clifton, N.J.)*, *2013*, 29–44. https://doi.org/10.1007/978-1-4939-9550-9_2

McLaren, W., Gil, L., Hunt, S. E., Riat, H. S., Ritchie, G. R. S., Thormann, A., Flicek, P., & Cunningham, F. (2016). The Ensembl Variant Effect Predictor. *Genome Biology*, *17*(1), 122. https://doi.org/10.1186/s13059-016-0974-4

Medicine, I. of, Arrow, K. J., Panosian, C., & Gelband, H. (2004). Saving Lives, Buying Time. In *Saving Lives, Buying Time*. National Academies Press. Retrived from https://doi.org/10.17226/11017

Meibalan, E., & Marti, M. (2017). Biology of malaria transmission. *Cold Spring*

Harbor Perspectives in Medicine, *7*(3). https://doi.org/10.1101/cshperspect.a025452

Mockenhaupt, F. P., Cramer, J. P., Hamann, L., Stegemann, M. S., Eckert, J., Oh, N.-R., Otchwemah, R. N., Dietz, E., Ehrhardt, S., Schröder, N. W. J., Bienzle, U., & Schumann, R. R. (2006). Toll-like receptor (TLR) polymorphisms in African children: Common TLR-4 variants predispose to severe malaria. *Proceedings of the National Academy of Sciences*, *103*(1), 177–182. https://doi.org/10.1073/PNAS.0506803102

Mordmüller, B., Surat, G., Lagler, H., Chakravarty, S., Ishizuka, A. S., Lalremruata, A., Gmeiner, M., Campo, J. J., Esen, M., Ruben, A. J., Held, J., Calle, C. L., Mengue, J. B., Gebru, T., Ibáñez, J., Sulyok, M., James, E. R., Billingsley, P. F., Natasha, K., … Kremsner, P. G. (2017). Sterile protection against human malaria by chemoattenuated PfSPZ vaccine. *Nature*, *542*(7642), 445–449. https://doi.org/10.1038/nature21060

Mueller, A. K., Labaled, M., Kappe, S. H. I., & Matuschewski, K. (2005). Genetically modified Plasmodium parasites as a protective experimental malaria vaccine. *Nature*, *433*(7022), 164–167. https://doi.org/10.1038/nature03188

Müller, G. C., Kravchenko, V. D., & Schlein, Y. (2008). Decline of Anopheles sergentii and Aedes caspius populations following presentation of attractive toxic (spinosad) sugar bait stations in an oasis. *Journal of the American Mosquito Control Association*, *24*(1), 147–149. https://doi.org/10.2987/8756-971X(2008)24[147:DOASAA]2.0.CO;2

Ndila, C. M., Uyoga, S., Macharia, A. W., Nyutu, G., Peshu, N., Ojal, J., Shebe, M., Awuondo, K. O., Mturi, N., Tsofa, B., Sepúlveda, N., Clark, T. G., Band, G., Clarke, G., Rowlands, K., Hubbart, C., Jeffreys, A., Kariuki, S., Marsh, K., … Williams, T. N. (2018). Human candidate gene polymorphisms and risk of severe malaria in children in Kilifi, Kenya: a case-control association study. *The Lancet. Haematology*, *5*(8), e333–e345. https://doi.org/10.1016/S2352-3026(18)30107-8

Ngo Bitoungui, V. J., Pule, G. D., Hanchard, N., Ngogang, J., Wonkam, A., Bitoungui, V. J. N., Pule, G. D., Hanchard, N., Ngogang, J., & 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 : A Journal of Integrative Biology*, *19*(3), 171–179. https://doi.org/10.1089/omi.2014.0134

Nielsen, R. (2005). Molecular Signatures of Natural Selection. *Annual Review of Genetics*, *39*(1), 197–218. https://doi.org/10.1146/annurev.genet.39.073003.112420

Ohtake, S., Ito, Y., Fukuta, M., & Habuchi, O. (2001). Human N-Acetylgalactosamine 4-Sulfate 6-O-Sulfotransferase cDNA Is Related to Human B Cell Recombination Activating Gene-associated Gene. *Journal of Biological Chemistry*, *276*(47), 43894–43900. https://doi.org/10.1074/jbc.M104922200

Ojewunmi, O. O., Adeyemo, T. A., Ayinde, O. C., Iwalokun, B., & Adekile, A. (2019). Current perspectives of sickle cell disease in Nigeria: changing the narratives. *Expert Review of Hematology*, *12*(8), 609–620. https://doi.org/10.1080/17474086.2019.1631155

Otto, T. D., Gilabert, A., Crellen, T., Böhme, U., Arnathau, C., Sanders, M., Oyola, S. O., Okouga, A. P., Boundenga, L., Willaume, E., Ngoubangoye, B., Moukodoum, N. D., Paupy, C., Durand, P., Rougeron, V., Ollomo, B., Renaud, F., Newbold, C., Berriman, M., & Prugnolle, F. (2018). Genomes of all known members of a Plasmodium subgenus reveal paths to virulent human malaria. *Nature Microbiology*, *3*(6), 687–697. https://doi.org/10.1038/s41564-018-0162-2

Patin, E., Lopez, M., Grollemund, R., Verdu, P., Harmant, C., Quach, H., Laval, G., Perry, G. H., Barreiro, L. B., Froment, A., Heyer, E., Massougbodji, A., Fortes-Lima, C., Migot-Nabias, F., Bellis, G., Dugoujon, J. M., Pereira, J. B., Fernandes, V., Pereira, L., … Quintana-Murci, L. (2017). Dispersals and genetic adaptation of Bantu-speaking populations in Africa and North America. *Science*

(New York, N.Y.), *356*(6337), 543–546. https://doi.org/10.1126/science.aal1988

Patnala, R., Clements, J., & Batra, J. (2013). Candidate gene association studies: a comprehensive guide to useful in silico tools. *BMC Genetics*, *14*, 39. https://doi.org/10.1186/1471-2156-14-39

Patterson, N., Price, A. L., & Reich, D. (2006). Population structure and eigenanalysis. *PLoS Genetics*, *2*(12), 2074–2093. https://doi.org/10.1371/journal.pgen.0020190

Payne, R. O., Milne, K. H., Elias, S. C., Edwards, N. J., Douglas, A. D., Brown, R. E., Silk, S. E., Biswas, S., Miura, K., Roberts, R., Rampling, T. W., Venkatraman, N., Hodgson, S. H., Labbé, G. M., Halstead, F. D., Poulton, I. D., Nugent, F. L., De Graaf, H., Sukhtankar, P., … Draper, S. J. (2016). Demonstration of the blood-stage plasmodium falciparum controlled human malaria infection model to assess efficacy of the p. falciparum apical membrane antigen 1 Vaccine, FMP2.1/AS01. *Journal of Infectious Diseases*, *213*(11), 1743–1751. https://doi.org/10.1093/infdis/jiw039

Phillipson, D. W. (2006). *African Archaeology* (3rd ed., Vol. 17).

Piel, F B, Howes, R. E., Nyangiri, O. A., Moyes, C. L., Williams, T. N., Weatherall, D. J., & Hay, S. I. (2013). Online Biomedical Resources for Malaria-Related Red Cell Disorders. *Human Mutation*, *34*(7), 937–944. https://doi.org/10.1002/humu.22330

Piel, Frédéric B, & Weatherall, D. J. (2014). The α-Thalassemias. *New England Journal of Medicine*, *371*(20), 1908–1916. https://doi.org/10.1056/NEJMra1404415

Quakyi, I. A., Leke, R. G. F., Befidi-Mengue, R., Tsafack, M., Bonba-Nkolo, D., Manga, L., Tchinda, V., Njeungue, E., Kouontchou, S., Fogako, J., Nyonglema, P., Harun, L. T., Djokam, R., Sama, G., Eno, A., Megnekou, R., Metenou, S., Ndoutse, L., Same-Ekobo, A., … Taylor, D. W. (2000). The epidemiology of Plasmodium falciparum malaria in two Cameroonian villages: Simbok and Etoa. *AMERICAN JOURNAL OF TROPICAL MEDICINE AND HYGIENE*,

63(5–6), 222–230.

R Core team. (2016). R Core Team. In *R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing , Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.R Core Team (2016). R: A language and environment for statistical computing. R Foundation* . https://doi.org/10.1152/ajpgi.00069.2014

Ravenhall, M., Campino, S., Sepúlveda, N., Manjurano, A., Nadjm, B., Mtove, G., Wangai, H., Maxwell, C., Olomi, R., Reyburn, H., Drakeley, C. J., Riley, E. M., & Clark, T. G. (2018). Novel genetic polymorphisms associated with severe malaria and under selective pressure in North-eastern Tanzania. *PLOS Genetics*, *14*(1), e1007172. https://doi.org/10.1371/journal.pgen.1007172

RTS, S. C. T. P. (2012). A Phase 3 Trial of RTS,S/AS01 Malaria Vaccine in African Infants. *New England Journal of Medicine*, *367*(24), 2284–2295. https://doi.org/10.1056/nejmoa1208394

RTS, S. C. T. P. (2015). Efficacy and safety of RTS,S/AS01 malaria vaccine with or without a booster dose in infants and children in Africa: Final results of a phase 3, individually randomised, controlled trial. *The Lancet*, *386*(9988), 31–45. https://doi.org/10.1016/S0140-6736(15)60721-8

Russo, G., Faggioni, G., Paganotti, G. M., Djeunang Dongho, G. B., Pomponi, A., De Santis, R., Tebano, G., Mbida, M., Sanou Sobze, M., Vullo, V., Rezza, G., & Lista, F. R. (2017). Molecular evidence of Plasmodium vivax infection in Duffy negative symptomatic individuals from Dschang, West Cameroon. *Malaria Journal*, *16*(1), 1–9. https://doi.org/10.1186/s12936-017-1722-2

Saiwaew, S., Sritabal, J., Piaraksa, N., Keayarsa, S., Ruengweerayut, R., Utaisin, C., Sila, P., Niramis, R., Udomsangpetch, R., Charunwatthana, P., Pongponratn, E., Pukrittayakamee, S., Leitgeb, A. M., Wahlgren, M., Lee, S. J., Day, N. P. J., White, N. J., Dondorp, A. M., & Chotivanich, K. (2017). Effects of sevuparin on rosette formation and cytoadherence of Plasmodium falciparum infected erythrocytes. *PloS One*, *12*(3), e0172718.

https://doi.org/10.1371/journal.pone.0172718

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., & Theander, T. G. (2004). Evidence for the involvement of VAR2CSA in pregnancy-associated malaria. *Journal of Experimental Medicine*, *200*(9), 1197–1203. https://doi.org/10.1084/jem.20041579

Seder, R. A., Chang, L. J., Enama, M. E., Zephir, K. L., Sarwar, U. N., Gordon, I. J., Holman, L. S. A., James, E. R., Billingsley, P. F., Gunasekera, A., Richman, A., Chakravarty, S., Manoj, A., Velmurugan, S., Li, M. L., Ruben, A. J., Li, T., Eappen, A. G., Stafford, R. E., … Hoffman, S. L. (2013). Protection against malaria by intravenous immunization with a nonreplicating sporozoite vaccine. *Science*, *341*(6152), 1359–1365. https://doi.org/10.1126/science.1241800

Severe Malaria Observatory. (2018). *Severe malaria | Severe Malaria Observatory*. https://www.severemalaria.org/severe-malaria

Shaikho, E. M., Farrell, J. J., Alsultan, A., Qutub, H., Al-Ali, A. K., Figueiredo, M. S., Chui, D. H. K. K., Farrer, L. A., Murphy, G. J., Mostoslavsky, G., Sebastiani, P., & Steinberg, M. H. (2017). A phased SNP-based classification of sickle cell anemia HBB haplotypes. *BMC Genomics*, *18*(1), 608. https://doi.org/10.1186/s12864-017-4013-y

Shriner, D., & Rotimi, C. N. (2018). Whole-Genome-Sequence-Based Haplotypes Reveal Single Origin of the Sickle Allele during the Holocene Wet Phase. *American Journal of Human Genetics*, *102*(4), 547–556. https://doi.org/10.1016/j.ajhg.2018.02.003

Skoglund, P., Thompson, J. C., Prendergast, M. E., Mittnik, A., Sirak, K., Hajdinjak, M., Salie, T., Rohland, N., Mallick, S., Peltzer, A., Heinze, A., Olalde, I., Ferry, M., Harney, E., Michel, M., Stewardson, K., Cerezo-Román, J. I., Chiumia, C., Crowther, A., … Reich, D. (2017). Reconstructing Prehistoric African Population Structure. *Cell*, *171*(1), 59-71.e21. https://doi.org/10.1016/j.cell.2017.08.049

Smith, J. D., Rowe, J. A., Higgins, M. K., & Lavstsen, T. (2013). Malaria's deadly grip: Cytoadhesion of Plasmodium falciparum-infected erythrocytes. *Cellular Microbiology*, *15*(12), 1976–1983. https://doi.org/10.1111/cmi.12183

Soulard, V., Bosson-Vanga, H., Lorthiois, A., Roucher, C., Franetich, J. F., Zanghi, G., Bordessoulles, M., Tefit, M., Thellier, M., Morosan, S., Le Naour, G., Capron, F., Suemizu, H., Snounou, G., Moreno-Sabater, A., & Mazier, D. (2015). Plasmodium falciparum full life cycle and Plasmodium ovale liver stages in humanized mice. *Nature Communications*, *6*(May). https://doi.org/10.1038/ncomms8690

Spencer, C. C. A., Su, Z., Donnelly, P., & Marchini, J. (2009). Designing genome-wide association studies: Sample size, power, imputation, and the choice of genotyping chip. *PLoS Genetics*, *5*(5). https://doi.org/10.1371/journal.pgen.1000477

Spring, M. D., Cummings, J. F., Ockenhouse, C. F., Dutta, S., Reidler, R., Angov, E., Bergmann-Leitner, E., Stewart, V. A., Bittner, S., Juompan, L., Kortepeter, M. G., Nielsen, R., Krzych, U., Tierney, E., Ware, L. A., Dowler, M., Hermsen, C. C., Sauerwein, R. W., de Vlas, S. J., … Heppner, D. G. (2009). Phase 1/2a study of the malaria vaccine candidate apical membrane antigen-1 (AMA-l) administered in adjuvant system AS01B or AS02A. *PLoS ONE*, *4*(4), e5254. https://doi.org/10.1371/journal.pone.0005254

Tabue, R. N., Njeambosay, B. A., Zeukeng, F., Esemu, L. F., Fodjo, B. A. Y. Y. B. A. Y., Nyonglema, P., Awono-Ambene, P., Etang, J., Fondjo, E., Achu, D., Leke, R. G. F. F., Kouambeng, C. C. C., Knox, T. B., Mnzava, A. P., & Bigoga, J. D. (2019). Case Definitions of Clinical Malaria in Children from Three Health Districts in the North Region of Cameroon. *BioMed Research International*, *2019*, 9709013. https://doi.org/10.1155/2019/9709013

Taliun, D., Harris, D., Kessler, M., Carlson, J., Szpiech, Z., Torres, R., Taliun, S., Corvelo, A., Gogarten, S., Kang, H. M., Pitsillides, A., LeFaive, J., Lee, S., Tian, X., Browning, B., Das, S., Emde, A.-K., Clarke, W., Loesch, D., … Abecasis, G. (2019). Sequencing of 53,831 diverse genomes from the NHLBI

TOPMed Program. *BioRxiv*, *2*, 563866. https://doi.org/10.1101/563866

Teo, Y., Small, K. S., & Kwiatkowski, D. P. (2010). *Methodological challenges of genome-wide association analysis in Africa*. *11*(2), 149–160. https://doi.org/10.1038/nrg2731.Methodological

Timmann, C., Thye, T., Vens, M., Evans, J., May, J., Ehmen, C., Sievertsen, J., Muntau, B., Ruge, G., Loag, W., Ansong, D., Antwi, S., Asafo-Adjei, E., Nguah, S. B., Kwakye, K. O., Akoto, A. O. Y., Sylverken, J., Brendel, M., Schuldt, K., … Horstmann, R. D. (2012). Genome-wide association study indicates two novel resistance loci for severe malaria. *Nature*, *489*(7416), 443–446. https://doi.org/10.1038/nature11334

Uren, C., Kim, M., Martin, A. R., Bobo, D., Gignoux, C. R., Van Helden, P. D., Möller, M., Hoal, E. G., & Henn, B. M. (2016). Fine-scale human population structure in Southern Africa reflects ecogeographic boundaries. *Genetics*, *204*(1), 303–314. https://doi.org/10.1534/genetics.116.187369

Verkoczy, L. K., Guinn, B. A., & 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. *Journal of Biological Chemistry*, *275*(28), 20967–20979. https://doi.org/10.1074/jbc.M001866200

Verkoczy, L. K., Marsden, P. A., & 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). *European Journal of Immunology*, *28*(9), 2839–2853. https://doi.org/10.1002/(SICI)1521-4141(199809)28:09<2839::AID-IMMU2839>3.0.CO;2-6

Veten, F. M., Abdelhamid, I. O., Meiloud, G. M., Ghaber, S. M., Salem, M. L., Abbes, S., & Houmeida, A. O. (2012). Hb S [β6(A3)Glu→Val, GAG>GTG] and β-globin gene cluster haplotype distribution in Mauritania. *Hemoglobin*, *36*(4), 311–315. https://doi.org/10.3109/03630269.2012.688782

Wanji, S., Tanke, T., Atanga, S. N., Ajonina, C., Nicholas, T., & Fontenille, D. (2003). Anopheles species of the mount Cameroon region: biting habits, feeding behaviour and entomological inoculation rates. *TROPICAL MEDICINE & INTERNATIONAL HEALTH*, *8*(7), 643–649. https://doi.org/10.1046/j.1365-3156.2003.01070.x

WHO. (2014). Severe malaria. In *Tropical Medicine & International Health 19*, (10) 967. https://doi.org/10.1111/tmi.12313

WHO. (2015). Who Guidelines for the Treatment of Malaria. *Diagnosis of Malaria*, 27–30. https://doi.org/10.1016/0035-9203(91)90261-V

WHO. (2020). World Malaria Report 2020. World Health Organization. In *World Health*.

Williams, T. N., Mwangi, T. W., Wambua, S., Peto, T. E. A., Weatherall, D. J., Gupta, S., Recker, M., Penman, B. S., Uyoga, S., Macharia, A., Mwacharo, J. K., Snow, R. W., & Marsh, K. (2005). Negative epistasis between the malaria-protective effects of α+-thalassemia and the sickle cell trait. *Nature Genetics*, *37*(11), 1253–1257. https://doi.org/10.1038/ng1660

Wright, G. J., & Rayner, J. C. (2014). Plasmodium falciparum Erythrocyte Invasion: Combining Function with Immune Evasion. *PLoS Pathogens*, *10*(3), 1–7. https://doi.org/10.1371/journal.ppat.1003943

Wu, X.-D., Zhang, N., Liang, M., Liu, W.-L., Lin, B.-B., Xiao, Y.-R., Li, Y.-Z., Zeng, K., & Lin, C.-Z. (2018). Gender-specific association between Apelin/APJ gene polymorphisms and hypertension risk in Southeast China. *Gene*, *669*, 63–68. https://doi.org/10.1016/j.gene.2018.05.079

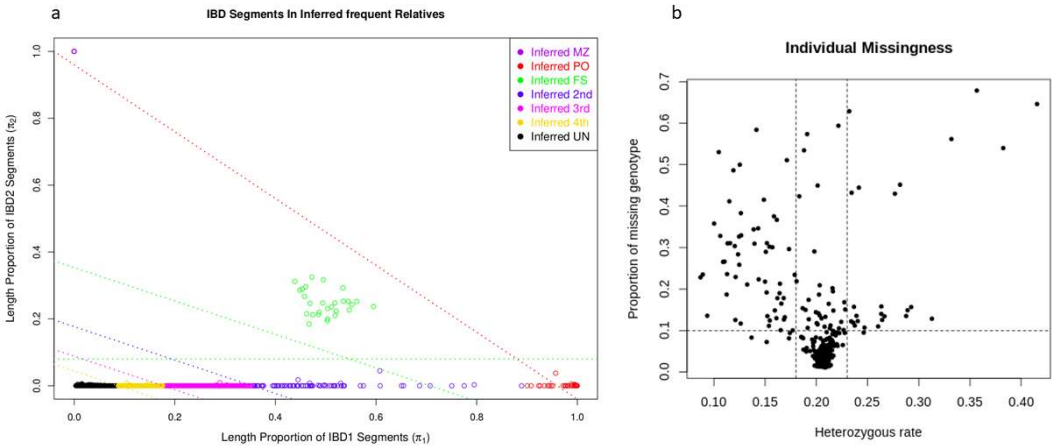
Yang, J., Lee, S. H., Goddard, M. E., & Visscher, P. M. (2011). GCTA: A tool for genome-wide complex trait analysis. *American Journal of Human Genetics*, *88*(1), 76–82. https://doi.org/10.1016/j.ajhg.2010.11.011

Yu, W., Misulovin, Z., Suh, H., Hardy, R. R., Jankovic, M., Yannoutsos, N., & Nussenzweig, M. C. (1999). Coordinate regulation of RAG1 and RAG2 by cell type-specific DNA elements 5' of RAG2. *Science*, *285*(5430), 1080–1084.

https://doi.org/10.1126/science.285.5430.1080

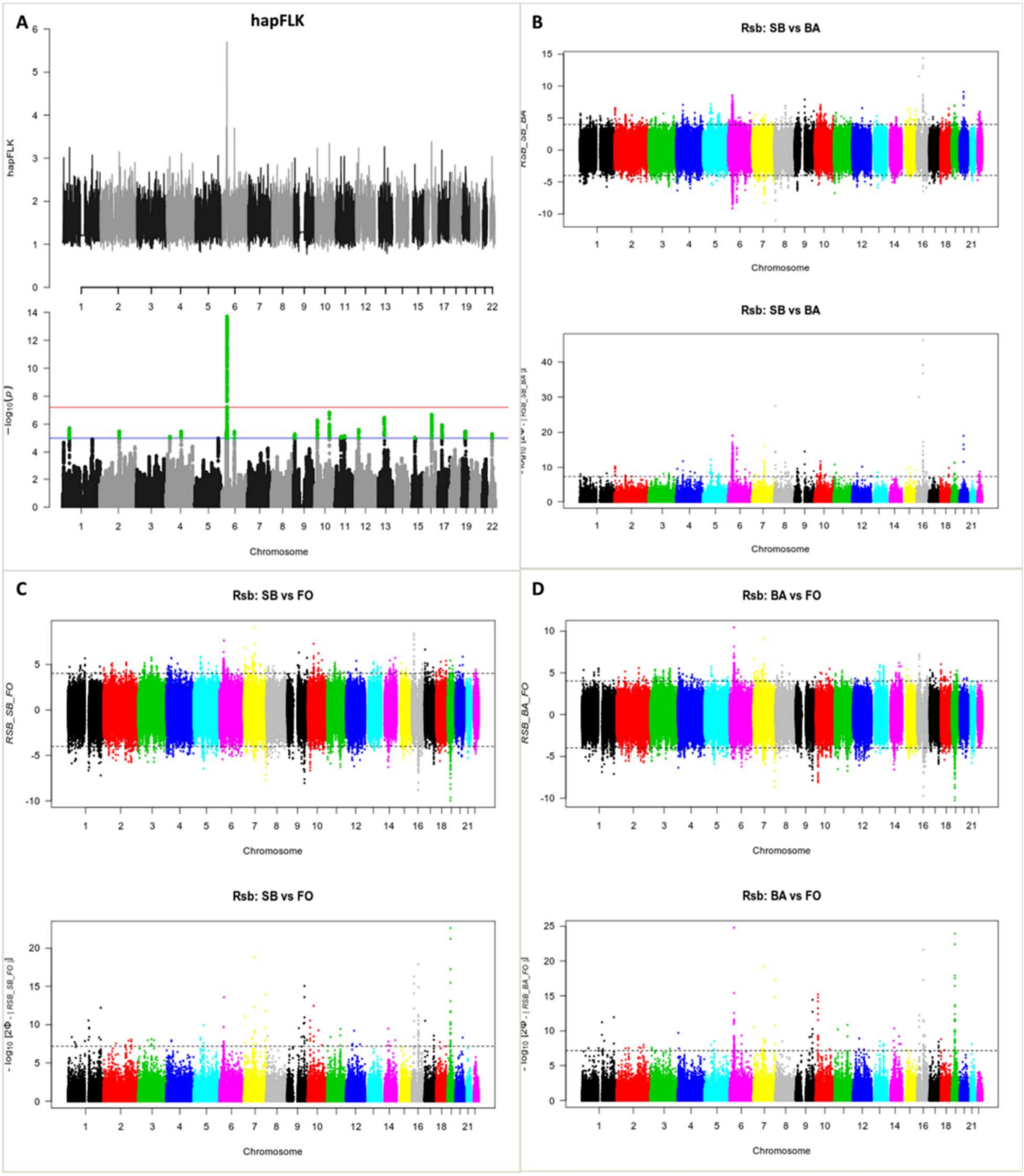
APPENDICES

Appendix I: Quality Control



Sample Quality Control: a) Exclusion of related individuals. MZ = monozygotic twins, PO = parent-offspring, FS = full sibship, 2nd = second degree relation, 3rd = third degree relation, 4th = fourth degree relation, UN = unrelated. One individual from each pair of MZ, PO, FS, and 2nd was excluded. b) Proportion of missing genotype against individual heterozygosity (missingness). Individuals with missing genotype > 10% and heterozygosity out of the range of 0.18 – 0.23 were excluded.

Appendix II: Cross population genome scan for selection and hapFLK results



Appendix III: Imputed Allele frequency at known malaria-associated loci

Table S1 shows the allele frequencies of variants in key malaria-associated loci in our data. The Bantu ethnic group apparently contributed a slightly greater proportion of the alleles in the pooled dataset. Importantly, the general population allele frequencies did not reflect the true allele frequencies per ethnic group. Although not particularly consequential at very common (high frequency) SNPs, this effect can lead to loss of power at less common sites that may appear as rare variants as a result of pooling the ethnic groups together. These rare variants would be filtered out during QC procedures. The Semi-Bantu ethnic group was predicted to have lower HbS and HbC frequencies as compared to the Bantu. This is consistent with a previous finding in various Cameroonian ethnic populations, finding the most predominant Semi-Bantu tribe to have HbS gene frequency of 8.5%, significantly lower than other ethnic groups (Engle-Stone et al., 2017).

Table *S*1: Allele frequency and imputation performance at key malaria-associated loci in Cameroonians

SNP (Gene; Variant)	GRCh37 Coordinate (chr:position)	Strategy	Pooled		Bantu		Semi-Bantu	
			AAF	R^2	AAF	R^2	AAF	R^2
rs334 (*HBB*; HbS)	11:5248232	In-house Imputation	**0.09**	**0.9**	**0.102**	**0.91**	**0.08**	**0.89**
		Michigan	0.086	0.88	-	-	-	-
		BEAGLEv5.1	**0.083**	**0.91**	**0.098**	**0.9**	**0.073**	**0.93**
rs33930165 (*HBB*; HbC)	11:5248233	In-house Imputation	**0.002**	**0.53**	**0.002**	**0.42**	**0.002**	**0.61**
		Michigan	0.0044	0.45	-	-	-	-
		BEAGLEv5.1	0.0029	0.52	0.0052	0.51	0.0015	0.36
rs8176746 (*ABO*)	9:136131322	In-house Imputation	0.17	0.99	-	-	-	-
		TOPMed	0.17	0.99	-	-	-	-
rs10751451 (*ATB2B4*)	1:203650978	In-house Imputation	0.66	1	-	-	-	-
		TOPMed	0.66	0.98	-	-	-	-
rs4951377 (*ATB2B4*)	1:203658471	In-house Imputation	0.66	0.99	-	-	-	-
		TOPMed	0.66	0.98	-	-	-	-
rs62418762 (*EPHA7*)	6:93218698	In-house Imputation	0.049	0.98	-	-	-	-
		TOPMed	0.048	0.99	-	-	-	-
rs184895969 (*FREM3*)	4:144698528	In-house Imputation	0.006	0.87	-	-	-	-
		TOPMed	0.012	0.99	-	-	-	-
rs1050829 (*G6PD*)	X:153763492	Michigan	0.34	0.92	-	-	-	-
rs1050828 (*G6PD*)	X:153764217	Michigan	0.12	0.87	-	-	-	-

Appendix IV: Assessment of power of imputation

Importantly, we observed that BOLT-LMM and EMMAX achieved the greatest power for our data and population. This is consistent with the heritability estimates in which BOLT-LMM and EMMAX had better estimates respectively. We assessed power by the FDR of our association signals (the ability to filter in true signals while minimizing false signals), in which EMMAX and BOLT-LMM consistently had the lowest FDRs. We specifically used the *--lmm* function of BOLT-LMM which fails with low sample size or incorrect case ascertainment, or when it projects no gain in power. Meanwhile, it is expected that EMMAX achieved its power in our analysis on the basis of its structure-correcting ability as in the highly structured 1966 Northern Finland Birth Cohort (NFBC66) (Kang et al., 2010). In general, the association pattern was different for the pooled data set and the ethnic groups analyzed separately. This may be attributable to different haplotypic backgrounds of the ethnic populations as a result of differential evolutionary paths under differential selection pressures.