CHARACTERIZATION OF ERYTHROCYTE RECEPTOR POLYMORPHISMS IN A MALARIA ENDEMIC POPULATION IN KILIFI, KENYA

OMEDO, Irene Akinyi (BSc. Biochemistry, University of Nairobi)

Reg No: I56/64991/2010

Thesis submitted in partial fulfilment for the award of MSc degree in Biotechnology, University of Nairobi.

DECLARATION

I declare that this research is my own work, and has not been submitted for examination in any other university.

Name: Omedo, Irene Akinyi

Registration: I56/64991/2010

Signature: _____Date_____

APPROVAL

This thesis has been submitted with our approval as University Supervisors:

Dr. Isabella L. Oyier

Centre for Biotechnology and Bioinformatics

University of Nairobi

Signature: _____Date_____

Dr. Vincent O. Ochieng'

Department of Biochemistry

University of Nairobi

Signature: _____Date_____

Prof. James O. Ochanda, PhD

Centre for Biotechnology and Bioinformatics

University of Nairobi

Signature: _____Date_____

ACKNOWLEDGEMENT

I would like to acknowledge Dr. Isabella Oyier, my primary supervisor, for the great work she did of helping me to conceptualize this project from an abstract idea and for her guidance throughout the process of developing protocols to data analysis. I would also like to acknowledge Prof. James Ochanda, who first told me the importance of getting a Master's degree and who provided useful and encouraging comments throughout my Master's program. My thanks also go to Dr. Vincent Ochieng' for his supervisory contributions. A big thank you goes to the Center for Biotechnology and Bioinformatics fraternity, for providing an enjoyable learning environment. A special thank you goes to Ann and Rono for their help in the Molecular Biology laboratory.

I also acknowledge the KEMRI-Wellcome Trust Collaborative Research Programme, Center for Geographic Medicine Research-Coast, Kilifi which supported this work through the Malaria Capacity Development Consortium Initiative grant awarded to Dr. Isabella Oyier. Thanks also to the unit for providing the samples and laboratory facilities that made this work possible. Thanks especially to John Okombo, who was an immense help when it came to sequencing.

This work is dedicated to my mother Mary and sister Seline. Sleep with the angels.

Table of Contents

LIST OF FIGURES

LIST OF TABLES

ACRONYMS AND ABBREVIATIONS

WHO	World Health Organization
RBCs	Red Blood Cells
MSP	Merozoite Surface Proteins
GPI	Glycosylphosphatidylinositol
EBL	Erythrocyte Binding-like Ligands
AMA	Apical Membrane Antigen
PTRAMP	Plasmodium Thrombospondin-Related Apical Merozoite Protein
EGF	Epidermal growth factor
EBA	Erythrocyte Binding Antigen
PfRh	*Plasmodium falciparum* Reticulocyte binding-like Protein Homologue
GYP	Glycophorin gene
GP	Glycophorin protein
HbS	Sickle haemoglobin allele
SNP	Single Nucleotide Polymorphism
NCBI	National Centre for Biotechnology Information
PCR	Polymerase Chain Reaction
dNTPs	deoxynucleotide triphosphates
TBE	Tris Borate EDTA

EDTA	Ethylenediaminetetraacetic acid
ExoSAP	Exonuclease Shrimp Alkaline Phosphatase
LD	Linkage Disequilibrium
Hd	Haplotype Diversity
RFLP	Restriction Fragment Length Polymorphism
BLAST	Basic Local Alignment Search Tool
M^KM^K	Glycophorin A and B deficient erythrocytes
DBP	Duffy Binding Proteins
DBL	Duffy Binding Like domain

ABSTRACT

Malaria is a strong selective force in the human genome, selecting genes for resistance to disease in human populations living in malaria endemic areas. Selection by malaria has generated genetic variations, providing evolutionary driving force mediating polymorphisms such as Glucose-6-Phosphate Dehydrogenase deficiency and sickle-cell anaemia. Genes encoding erythrocyte receptors of *Plasmodium falciparum* especially the sialoglycoproteins, glycophorins A, B and C, which are the main parasite receptors, have been shown to be under selection, with Single Nucleotide Polymorphisms (SNPs) and exon deletions being detected. This study analyzed SNPs in glycophorins A, B, and C genes in a population living in the malaria endemic area of Kilifi, Kenya. DNA samples were obtained from malaria positive individuals. Regions encompassing extracellular coding domains of the genes were amplified using gene specific primers. These regions are either putative or confirmed binding domains of the malaria parasite. PCR products were cleaned using EXOSAP-IT enzymatic clean-up procedure and sequenced using BigDye v3.1 terminator chemistry. Genes were analyzed separately for polymorphisms using DNASTAR software. Tajima's D, Fu and Li's D* and F* statistics were calculated to establish if polymorphisms detected were under selection. Elevated haplotype diversity was detected, with glycophorin A having the highest number of haplotypes. Statistically significant ($p<0.05$) SNPs were detected in the second and third exons of glycophorin A, and the fourth exon of glycophorin B. A single polymorphism that was not statistically significant ($p>0.10$) was detected in exon 2 of glycophorin C. Glycophorin A had an excess of intermediate frequency polymorphisms, indicative of balancing selection, while glycophorin B showed significantly higher proportions of low frequency polymorphisms relative to expectation within this population,

indicating a population size expansion or purifying selection. The results indicate that the population is under selection, possibly caused by malaria, as the glycophorins are the main receptors for *P. falciparum*.

CHAPTER ONE

INTRODUCTION AND LITERATURE REVIEW

1.1 General Introduction

Malaria in humans is a deadly and infectious disease caused by several species of the *Plasmodium* protozoan parasite. Disease severity depends on factors such as parasite species and the immune status of the infected individual. Malaria causes more than 1 million deaths annually, mainly in children under 5 years of age and this mortality is attributed mainly to *Plasmodium falciparum*, the most virulent of the parasites (Wipasa *et al*., 2002; WHO, 2008). *Plasmodium* parasites belong to the phylum Apicomplexa which exhibit a complex lifecycle involving a vertebrate host and an arthropod vector, and are characterized by the possession of an apical complex that plays a major role in their invasion of host cells (Cowman and Crabb, 2006). In humans and other mammals, the vector is the female anopheles mosquito (Grassi, 1900; Beier, 1998; Cox, 2010).

During a mosquito blood meal, sporozoites are released into the bloodstream and invade hepatocytes where they undergo a round of asexual replication known as exoerythrocytic schizogony, leading to the formation and release of merozoites into the blood stream. Merozoites invade erythrocytes and develop through the ring, trophozoite and schizont stages. The schizont then ruptures and releases newly formed merozoites into the blood stream and these invade other erythrocytes and undergo another round of multiplication. This cycle continues multiple times, leading to increased parasitemia within the infected host (Wiser, 2009).

Malaria is associated with several clinical symptoms including nausea, headaches and fever (Miller *et al.*, 2002; Wiser, 2009). Fever is caused by the synchronous rupture of red blood cells (RBCs) which leads to the release of antigenic components such as pyrogenic material into the blood stream (Wiser, 2009). These clinical symptoms are caused by the asexual blood stage of the parasite, when merozoites invade, egress and re-invade erythrocytes, thus making this stage in the parasite's lifecycle an important target for interventions to prevent invasion and hence combat malaria (Baum *et al.*, 2005).

1.2 Overview of erythrocyte invasion by *Plasmodium falciparum* merozoites

Though the finer details of erythrocyte invasion by *Plasmodium* merozoites are not completely understood (Iyer *et al*, 2007), the overall mechanism has been relatively well characterized *in P. falciparum* and follows a specific series of steps (Fig 1.1) (Miller, 1977). The first step involves recognition and initial binding of the merozoite to the erythrocyte (Cowman and Crabb, 2006; Cowman *et al*., 2012). How exactly the merozoite recognizes its target cell is still unknown (Cowman and Crabb, 2006), although it has been postulated that the initial interaction between the parasite and target cell may be by random collision and once in contact, proteins located on the surface of the merozoite mediate a reversible interaction between the parasite and erythrocyte (Iyer *et al.*, 2007).

Initial attachment is followed by a step during which the merozoite reorients itself such that its apical region is in contact with the erythrocyte surface, thus allowing a much closer association between the two cells (Cowman and Crabb, 2006). An irreversible tight junction is then formed between the merozoite and erythrocyte and is mediated by specific receptor-ligand interactions (Cowman and Crabb, 2006; Wiser, 2009). The tight junction moves

towards the posterior end of the merozoite as the parasite enters the RBC, and the movement is powered by the parasite's actin-myosin motor (Jones *et al.*, 2006). The merozoite loses its surface coat as it penetrates the erythrocyte (Aikawa *et al.*, 1978) and forms a parasitophorous vacuole which encloses it and in which it undergoes the subsequent rounds of replication to form 16-32 daughter merozoites (Lingelbach and Joiner, 1998; Wickham *et al.*, 2003). The membrane of this vacuole is probably formed by material originating from both the parasite components and the erythrocyte membrane (Dubremetz *et al.*, 1998). Following parasite entry into the erythrocyte, the erythrocyte membrane reseals (Hadley *et al.*, 1986). This step marks the end of merozoite invasion and once inside the erythrocyte, the merozoite develops through the different ring, trophozoite and schizont stages of its life cycle (Wickham *et al.*, 2003; Wiser, 2009).

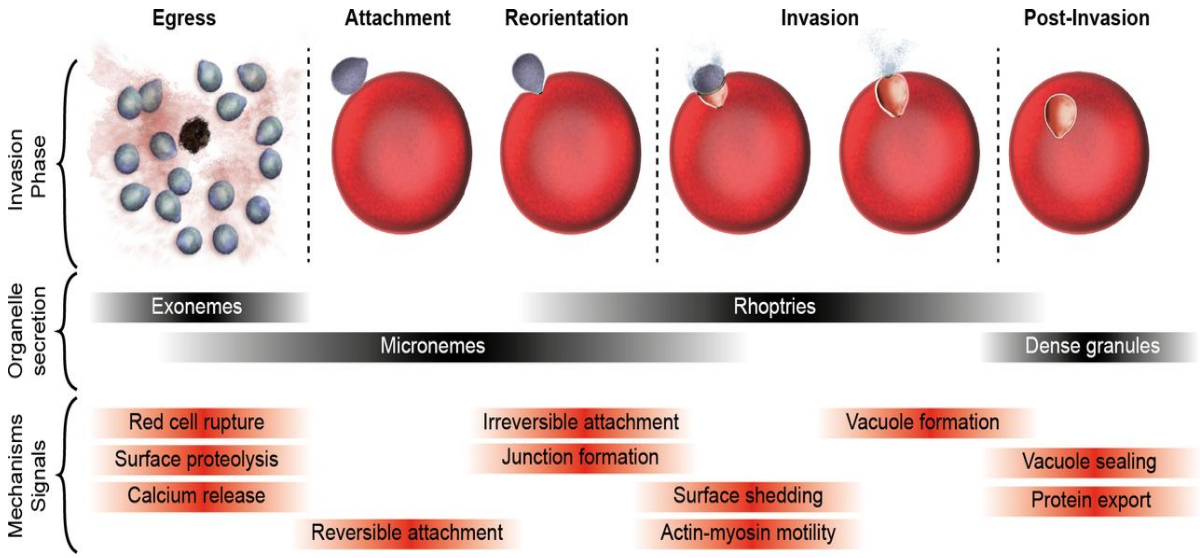


Figure 1.1: A time course of merozoite invasion of the erythrocyte from egress to post-invasion. A cellular overview is given with associated timing of organelle secretion and key mechanistic or signaling steps listed below. Adapted from Cowman *et al.*, 2012.

1.3 Molecular basis of *Plasmodium falciparum* merozoite invasion

The invasion process is a complex series of events requiring the interaction of multiple parasite and erythrocyte proteins (Table 1.1; Fig. 1.2), although little is known about the role of individual proteins in this process (Cowman *et al.*, 2000; Weatherall *et al.*, 2002).

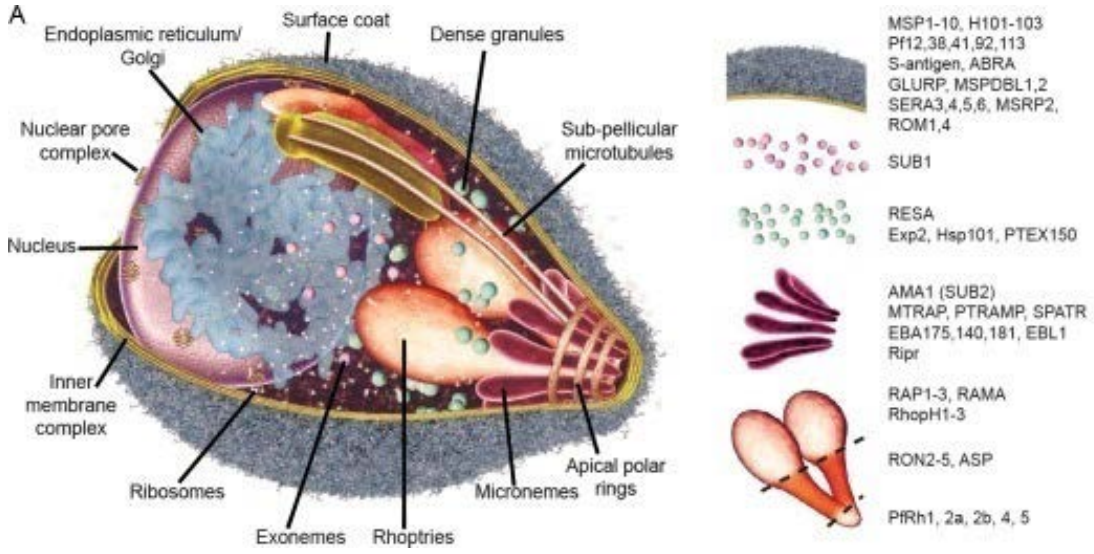


Figure. 1.2 Three-dimensional diagram of *Plasmodium* merozoite showing its core secretory organelles and proteins important for invasion. Adapted from Cowman *et al.*, 2012.

The Merozoite Surface Protein (MSP) family (Fig 1.2) has been implicated in the initial attachment of the merozoite to the erythrocyte surface (Kadekoppala *et al.*, 2008). They are glycosylphosphatidylinositol (GPI)-anchored membrane proteins and peripheral membrane proteins which interact with the membrane via GPI-anchored proteins (Sanders *et al.*, 2005; Kadekoppala*et al.*, 2008). MSP-1 is implicated in mediating the initial contact of parasite and target cell due to its abundance and uniform distribution on the merozoite surface and also because experiments have shown that antibodies against MSP-1 inhibit invasion (Cowman *et al.*, 2000; O-Donnell *et al.*, 2000; Wiser, 2009; Child *et al.*, 2010). MSP-1 interacts with the RBC using Band 3 protein on the erythrocyte surface (Goel *et al.*, 2003; Cowman and Crabb, 2006; Wiser, 2009; Child *et al.*, 2010).

Following reorientation, the invasion process must be activated. Micronemal proteins implicated in this process include Erythrocyte Binding-like Ligands (EBL) (Cowman and Crabb, 2006; Iyer *et al*., 2007; Gaur and Chitnis, 2011). Erythrocyte Binding Antigens (EBAs) belong to this superfamily, which also includes the Duffy Binding Proteins (DBP) of *P. vivax* and *P. knowlesi* (Iyer *et al*., 2007). The EBL gene superfamily contains 2 cysteine rich domains, an amino Duffy Binding-Like (DBL) domain and a cytoplasmic cysteine domain adjacent to a transmembrane domain (Reed *et al*., 2000; Adams *et al*., 2001). Six members of the EBL family have been identified: EBA 175, EBA 181, EBA 140, EBA 165, EBL 1 and MAEBL (Table 1.1) (Cowman and Crabb, 2006; Iyer *et al*., 2007; Gaur and Chitnis, 2011). The DBL domain that characterizes this family is a ligand domain that binds to erythrocyte receptors during invasion (Adams *et al*., 2001; Mayor *et al*., 2005). The multiple numbers of functional EBL genes provide ligand diversity as well as possible use of different receptors on the target cell, thus increasing the parasite's invasion ability (Reed *et al*., 2000; Gilberger *et al*., 2003; Pasvol, 2003). Erythrocyte receptors for some of these ligands have been identified (Table 1.2) (Reed *et al*., 2000; Adams *et al*., 2001; Gilberger *et al*., 2003; Lobo *et al*., 2003; Tolia *et al*., 2005; Iyer *et al*., 2007; Mayer *et al*., 2009; Gaur and Chitnis, 2011). Experiments by Triglia *et al*., (2001) showed that EBA 165 is a transcribed pseudogene.

In addition to the EBL family, the rhoptry neck proteins belonging to the *P. falciparum* Reticulocyte Binding-like Protein homologue (PfRh) family are also involved in invasion (Gaur *et al*., 2003; Triglia *et al*., 2005; Cowman and Crabb, 2006; Iyer *et al*., 2007). Six members of this family have been identified in *P. falciparum*: PfRh1, PfRh2a, PfRh2b, PfRh3, PfRh4, PfRh5 (Table 1.1) (Jennings *et al*., 2007; Hayton *et al*., 2008; Baum *et al*.,

2009). These proteins are type I transmembrane proteins with signal peptide domains, a homology region with varying numbers of cysteine residues and a C-terminal transmembrane domain (Baum *et al*., 2009). /They are very large in size (220-350kDa), except for PfRh5, which has an approximate molecular weight of 63kDa (Gaur *et al*., 2004; Baum *et al*., 2009). PfRh5 also lacks the transmembrane domain present in the others (Baum *et al*., 2009).

Members of the PfRh family have similar gene and protein structure, but their sequence similarity is low (Baum *et al*., 2009), except for PfRh2a and PfRh2b, which have high sequence homology (Jennings *et al*., 2007). This family of proteins binds specifically to reticulocytes and not erythrocytes and is important in mediating the sialic-acid independent invasion pathway, except for PfRh1, which requires sialic acid for invasion (Table 1.2) (Rayner *et al*., 2001; Gaur *et al*., 2004; Triglia *et al*., 2005; Tham *et al*., 2009; Gaur and Chitnis, 2011; Tham *et al*., 2011). Studies by Taylor *et al*., (2001) indicate that PfRh3 is a pseudogene.

Table 1.1: *P. falciparum* merozoite proteins and their properties.

Name	PlasmoDB accession number	Genetic Knockout	Localization in merozoite before/during invasion	Potential function
GPI-anchored MSPs				
MSP-1	PF3D7_1133400	N	Surface/complex shed during invasion with MSP1/19 EGF	Putative Band 3 ligand
MSP-2	PF3D7_0206800	N	Surface	Highly polymorphic; likely structural role as surface coat
Microneme proteins				
AMA 1	PF3D7_1133400	N	Micronemes/surface and binds to RON2	Released on merozoite surface; binds RON complex
EBA-175	PF3D7_0731500	Y	Micronemes/surface and binds to glycophorin A	Binds to glycophorin A, likely signaling role for invasion
EBA-181	PF3D7_0102500	Y	Microneme/Surface and binds to unknown receptor	Binds to unknown receptor on red cell
EBA-140	PF3D7_1301600	Y	Micronemes/surface and binds to glycophorin C	Binds to glycophorin C on red cell
EBL-1	PF3D7_1371600	Y	No data	Binds to glycophorin B
Rhoptry neck proteins				
PfRh1	PF3D7_0402300	Y	Rhoptry neck/surface	Binds to red cells via unknown sialic-dependent receptor Y
PfRh2a	PF3D7_1335400	Y	Rhoptry neck/surface	Binds to red cells via unknown receptor Z
PfRh2b	PF3D7_1335300	Y	Rhoptry neck/surface	Binds to red cells via unknown receptor Z
PfRh4	PF3D7_0424200	Y	Rhoptry neck/surface	Binds to red cells via complement receptor 1
PfRh5	PF3D7_0424100	N	Rhoptry neck/surface forms complex with Ripr.	Binds to red cells via Basigin

N, knockout attempt unsuccessful; Y, knockout generated; ND, knockout not attempted; PV, parasitophorous vacuole; MSP, merozoite surface protein. Adapted from Cowman *et al.*, 2012.

1.4 Alternate pathways of *Plasmodium falciparum* merozoite invasion

Several studies have been conducted that point to the presence of alternate invasion pathways for *P. falciparum* (Table 1.2) (Pasvol, 2003; Gaur *et al*., 2003; Gaur *et al*., 2004; Iyer *et al*., 2007; Tham *et al*., 2009). These include studies on binding of different strains of the parasite to erythrocytes that are deficient in one or more of their surface glycoproteins (e.g. M^kM^k erythrocytes lacking both glycophorins A and B expression), as well as studies on erythrocytes treated with neuraminidase and trypsin (Hadley *et al*., 1987; Gaur *et al*.,

2003; Lobo *et al*., 2003). Targeted gene disruption of the parasite ligands, especially those belonging to the PfRh family, have also been conducted and the results also point to the use of alternate invasion pathways by the parasite (Triglia *et al*., 2005; Iyer *et al*., 2007; Tham *et al*., 2009; Gaur and Chitnis, 2011).

Table 1.2: Alternate invasion pathways of the *P. falciparum* merozoite.

Pathway	**Binding phenotype**	**Parasite ligand**	**Erythrocyte receptor**
Sialic acid-dependent	N_S T_S	EBA 175	Glycophorin A
	N_S T_S	EBA 140	Glycophorin C
	N_S T_R	EBL-1	Glycophorin B
	N_S T_R	EBA 181	Receptor unknown
	N_S T_R	PfRh1	Receptor unknown
Sialic acid-independent	N_R T_R	PfRh2	Receptor unknown
	N_R T_S	PfRh4	Complement Receptor 1
	N_R T_R	PfRh5	Basigin

N=neuraminidase treatment of RBC; T= trypsin treatment of RBC; R=resistant; S=sensitive. Adapted from Gaur and Chitnis, 2011.

1.5 *Plasmodium falciparum* erythrocyte receptors

Studies carried out on human erythrocytes have determined the protein composition of their plasma membranes (Steck, 1974). Several of these proteins have been shown to interact with the *Plasmodium* parasite in a specific manner during invasion by merozoites, though their number is considerably lower than the number of *P. falciparum* ligands identified so far (Iyer *et al*., 2007). These proteins include Glycophorins A, B, and C/D, each of which binds a specific ligand. Glycophorins are transmembrane sialoglycoproteins consisting of a short

cytoplasmic tail domain, a single membrane spanning domain and a variable extracellular domain (Blumenfeld and Huang, 1995; Ko *et al*., 2011). Four glycophorin genes have been identified: GYPA, GYPB, GYPC and GYPE (Colin *et al*., 1989; Blumenfeld and Huang, 1995). The main gene family is made up of GYPA, GYPB and GYPE and numerous variants exist, arising from unequal gene recombinations and splice-site mutations mainly within GYPA and GYPB genes (Fig. 1.3) (Blumenfeld and Huang, 1995; Wang *et al*., 2003; Ko *et al*., 2011).

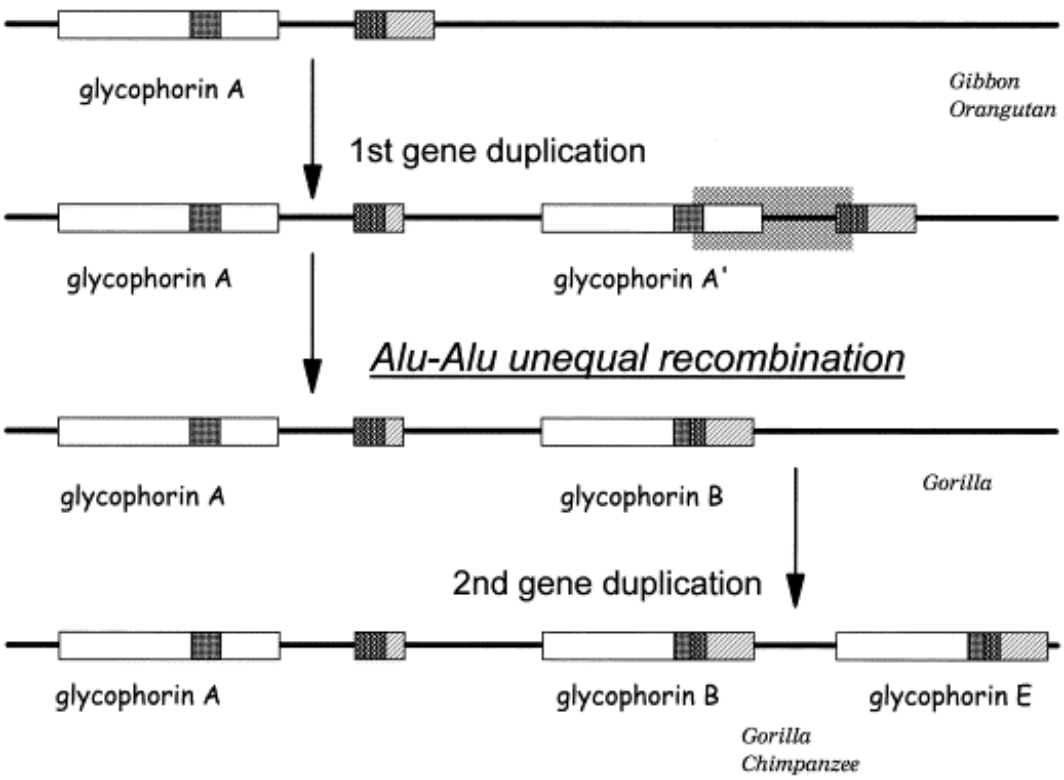


Figure 1.3 Evolution of the primate glycophorin gene family. The primordial glycophorin gene was duplicated and one of the copies gave rise to the glycophorin B gene, through an unequal recombination mediated by Alu sequences. Another duplication led to the glycophorin E gene which is not completely fixed in the Gorilla species. The Alu elements are indicated by shaded boxes, the glycophorin genes by open boxes, and the genomic precursor sequence at the 3′ end of the glycophorin B and E genes is indicated by hatched boxes. (Derived from Makalowski, 2000).

These genes are found on the long arm of chromosome 4 in the order of GYP A, GYP B and GYP E in the 5' to 3' orientation (Onda *et al*., 1993). The three genes are highly homologous, and share over 95% sequence homology spanning over a 1 kilobase region beginning from the 5' end to an *Alu* repeat region in the gene encoding the transmembrane domain (Onda *et al*., 1993; Rearden *et al*., 1993; Ko *et al*., 2011). The glycoproteins encoded by these genes differ in the lengths of their extracellular domains and the structure of their membrane junctions. In addition, both GYP B and GYP E lack the cytoplasmic segment (Blumenfeld and Huang, 1995).

The extracellular domains of glycophorin A molecule (GPA) and glycophorin B molecule (GPB) are made up of about 60% carbohydrates rich in sialic acid (Blumenfeld and Huang, 1995; Hadley *et al*., 1987; Orlandi *et al*., 1992). The presence of sialic acid in these molecules gives erythrocytes a net negative charge, preventing them from adhering to each other and to vascular surfaces (Blumenfeld and Huang, 1995). Sialic acid has also been shown to be the carbohydrate molecule with which *P. falciparum* interacts during invasion of RBCs (Blumenfeld and Huang, 1995; Baum *et al.,* 2002), although invasion also requires interaction of the ligand with the receptor's peptide backbone (Mayer *et al*., 2006).

Glycophorins function as receptors in the sialic acid-dependent pathway, the main invasion pathway for the parasites (Adams *et al*., 2001, Cowman and Crabb, 2006). This has been demonstrated in studies using enzyme treatments and glycophorin deficient erythrocytes which observed a marked reduction in the efficiency of invasion whenever the glycoprotein structure was disrupted or when glycophorin deficient erythrocytes were used (Breuer *et al*., 1983; Vanderberg *et al*., 1985; Orlandi *et al*., 1992; Deans *et al*., 2007; Mayer *et al.,* 2009). Thus far, protein products of glycophorins A, B and C genes have been detected, while that for glycophorin E gene has not yet been detected (Wilder *et al.,* 2009; Ko *et al.,* 2011). Glycophorin E may be an unexpressed pseudogene, as it has several pseudoexons (Ko et al., 2011).

1.5.1 Glycophorin A

GPA is the main glycophorin found on the surface of RBCs, with up to 1 million copies of the molecule per cell (Baum *et al.,* 2002). It carries the MN blood group antigens (Kudo and Fukuda, 1989; Vignal *et al*., 1990; Rearden *et al*., 1993). The M and N antigens are defined by amino acid polymorphisms at positions 1 (Serine to Leucine substitution) and 5 (Glycine

to Glutamate substitution) of the amino acid sequence from 5'-3' location (Blanchard *et al.*, 1987; Blumenfeld and Huang, 1995). Numerous other genetic variations also occur in this molecule, resulting in rare blood group variants (Huang and Blumenfeld, 1995; Baum *et al.*, 2002). The gene coding for this protein has 7 exons (Fig 1.4), with the first and second exons being separated by an intronic region that is about 20kb long (Vignal *et al.*, 1989). The entire first exon and part of the second encode a signal peptide, the remainder of exon 2 as well as exons 3 and 4 encode the extracellular region, exon 5 encodes the transmembrane segment of the protein while exons 6 and 7 encode the cytoplasmic domain(Fig 1.4) (Baum *et al.*, 2002).

Although the full length gene is quite large (over 30kb long), most of it is intronic regions, with the coding exons being relatively small in size (Fig.1.4). The full length protein is 131 amino acids long with the first 26 N-terminal amino acids of glycophorin A (variant exhibiting blood group N specificity) being identical to those of glycophorin B (Blanchard *et al.*, 1987).

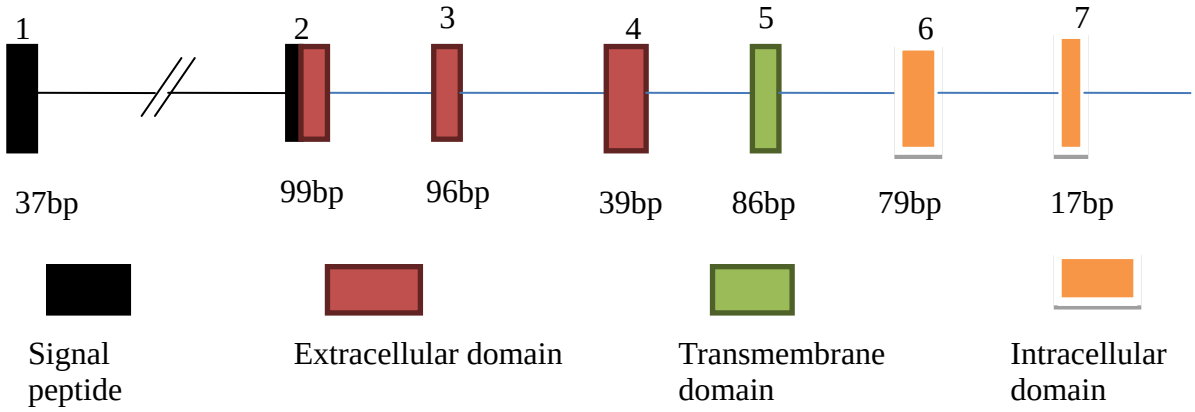


Figure 1.4. Schematic representation of the GYP A gene. Exon sizes in bp are given below the scheme. A signal peptide (encoded by exon 1 and part of exon 2) is cleaved off to leave a 131-amino acid mature protein composed of a glycosylated extracellular domain, transmembrane, and intracellular domain. Forward lines indicate large distance between adjacent exons. Adapted from Baum *et al.*, 2002.

GYP A is the main receptor in the sialic-acid dependent invasion pathway (Cowman and Crabb, 2006). It acts as the receptor for EBA-175 (Wang *et al.,* 2003; Cowman and Crabb, 2006). GYPA, GYPB and GYPE are undergoing rapid molecular evolution (Baum *et al.,* 2002; Wang *et al.,* 2003; Ko *et al.,* 2011). There are two hypotheses that have been put forth to explain the rapid evolution seen in these genes: the 'decoy' hypothesis (Baum *et al.,* 2002) and the 'evasion' hypothesis (Wang *et al.,* 2003). Many pathogens use sugar molecules of glycoproteins on the surface of cells for invasion and the decoy hypothesis postulates that RBC glycoproteins such as glycophorin A function as decoy receptors, attracting pathogens to the RBCs which lack a nucleus and DNA replication mechanisms and away from target tissues (Baum *et al.,* 2002). This hypothesis is supported by evidence that shows that GPA binds to different viruses and bacteria (Paul and Lee, 1987; Brooks *et al.,* 1989). In this case, rapid evolution of this gene is explained by species-specific pathogen pressure and a need by the molecule to target the diverse range of pathogens that it must bind to (Baum *et al.,* 2002; Wilder *et al.,* 2009). In the evasion hypothesis, Wang *et al.,* (2003) postulate that since GPA is the main receptor by which *P. falciparum* invades RBCs, the gene is constantly undergoing evolution to evade the parasite. The hypothesis further postulates that the rapid evolution of the glycophorin genes is also driven by the co-evolution of the parasite ligands and their glycophorin receptors (Wang *et al.,* 2003). Since then, different studies have been conducted that support the evasion hypothesis as opposed to the decoy hypothesis, making the former hypothesis the more likely basis for the rapid evolution seen in these genes.

1.5.2 Glycophorin B

GYP B is a member of the main glycophorin gene family, and is paralogous to GYP A and GYP E, sharing over 95% sequence similarity with these two genes (Onda *et al*., 1993; Rearden *et al*., 1993; Ko *et al.,* 2011). The gene has 5 exons, with the intronic region between exons 1 and 2 being over 18 kb long (Fig.1.5). The gene lacks the cytoplasmic encoding domain that is found in GYP A (Blumenfeld and Huang, 1995).

GYP B encodes the N-antigen at its N-terminus (Tarazona-Santos *et al.,* 2011). It also has two co-dominant alleles; GYP B*S and GYP B*s, coding for the S and s antigens which are defined by amino acid changes at position 29 (Methionine to Threonine substitution) and result in three phenotypes: S+s-, S-s-, S-s+ (Wang *et al.,* 2003; Tarazona-Santos *et al.,* 2011). Recombination and gene conversion between GYP A and GYP B also lead to expression of hybrid gene products that encode other minor glycophorin B antigens such as Dantu (Tarazona-Santos *et al.,* 2011; Faria *et al.,* 2012). Most of these variants are found in African and Asian populations where malaria is endemic, leading to the hypothesis that the variants are selected in these populations due to their relative resistance to *P. falciparum* invasion, as some of them affect the expression of the dominant GYP B*S and GYP B*s alleles, which are used by the parasite as its receptor (Tarazona-Santos *et al.,* 2011).

The U antigen is another member of the MNSs blood group system that is found on GPB and is usually expressed in combination with the Ss antigens, to give the common S+s+U+ phenotype (Tarazona-Santos *et al.,* 2011; Faria *et al.,* 2012). The phenotype S-s-U-, in which there is a deletion in GYP B exons 2 – 5, is characterized by a lack of expression of the GPB protein on erythrocyte surfaces (Tarazona-Santos *et al.,* 2011). These RBCs are relatively resistant to invasion by *Plasmodium* parasites (Tarazona-Santos *et al.,* 2011). In

most cases, the glycophorin B U- RBCs are also S-s- but in some rare cases, the phenotype S-s-U+ ^{var}is observed in individuals (Tarazona-Santos *et al.,* 2011; Faria *et al.,* 2012). This phenotype is characterized by weak expression of the U antigen and absence of the S and s antigens on RBC surfaces and is caused by mutations in exon 5 and intron 5 (Faria *et al.,* 2012). These unique phenotypes are prevalent in populations living in malaria endemic areas (Storry *et al.,* 2003; Tarazona-Santos *et al.,* 2011; Faria *et al.,* 2012).Using S-s-U- (glycophorin B null) cells as well as enzymatic treatment of RBCs, it was determined that glycophorin B is the receptor for *P. falciparum* EBL-1 (Mayer *et al.,* 2009) and recent studies in a Brazilian population show that molecular variations in the GYP B*S and GYP B*s alleles affects susceptibility of individuals to *P. falciparum* infection, with GYPB*S allele being associated with increased susceptibility to infection (Tarazona-Santos *et al.,* 2011).

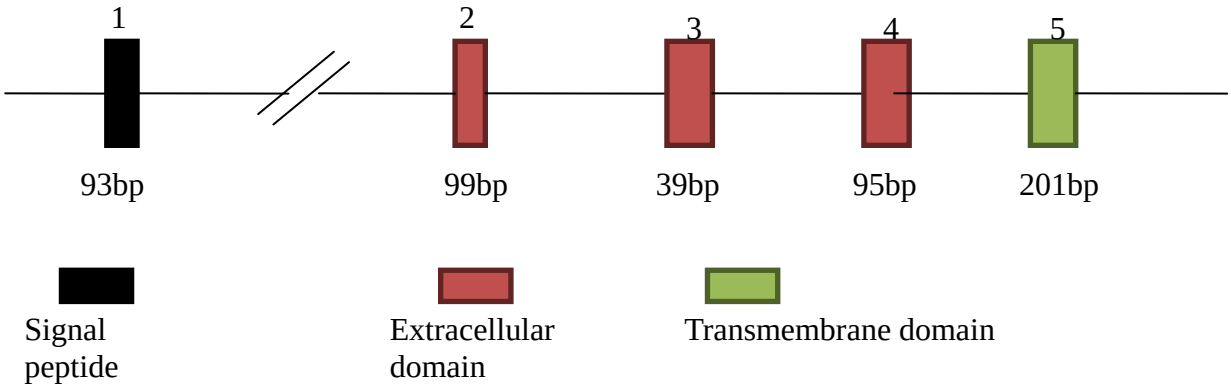


Figure 1.5 Schematic representation of the Glycophorin B gene. Exon sizes are shown below the scheme in bp. The gene lacks the coding segment for the cytoplasmic region. Forward lines indicate large genetic distance between adjacent exons.

1.5.3 Glycophorin C

Glycophorin C molecule (GPC) and glycophorin D molecule (GPD) are encoded by GYPC gene (Fig. 1.6) which is found on the long arm of chromosome 2 (Walker and Reid, 2010). The GYPC gene is made up of 4 coding regions spanning 13.5 kb of DNA and 2 repeated domains that are 3.4 kb in length that may have resulted from a recent duplication event of a single ancestral domain (Colin *et al*., 1989; Reid and Spring, 1994). Exon 1 encodes residues 1-16, exon 2 encodes residues 17-35, exon 3 encodes residues 36-63 and exon 4 residues 64-128 (Reid and Spring, 1994). Exons 2 and 3 share a high nucleotide homology, with less than 5% sequence divergence (Reid and Spring, 1994). The two exons also differ by a 9 amino acid insert at the 3'end of exon 3. Exons 1, 2 and most of exon 3 encode the N-terminal extracellular domain while the remainder of exon 3 and exon 4 encode transmembrane and cytoplasmic domains (Fig. 1.6) (Chasis and Mohandas, 1992; Reid and Spring, 1994).

Despite the similarity in name, GYPC shares no sequence homology with the other glycophorin genes (Wilder *et al.*, 2009). GYPC is also unique in that it encodes two different protein products, GPC and GPD (Wilder *et al.*, 2009), with GPD expressed as a truncated variant of GPC (Colin *et al*., 1989; Wilder *et al.*, 2009). This has been shown to be as a result of leaky translation, a phenomenon that allows the encoding of several protein products from a single gene transcript due to the use of different start codons (Chasis and Mohandas, 1992; Winardi *et al*., 1993; Wilder *et al.*, 2009). In this case, a C to A transversion mutation in the first exon of the gene introduced a novel start codon in the human lineage, allowing for the transcription of a previously untranslated region (Wilder *et*

al., 2009). This trait has however been shown to be uniquely human, with other primates only capable of encoding GPD (Wilder *et al.*, 2009).

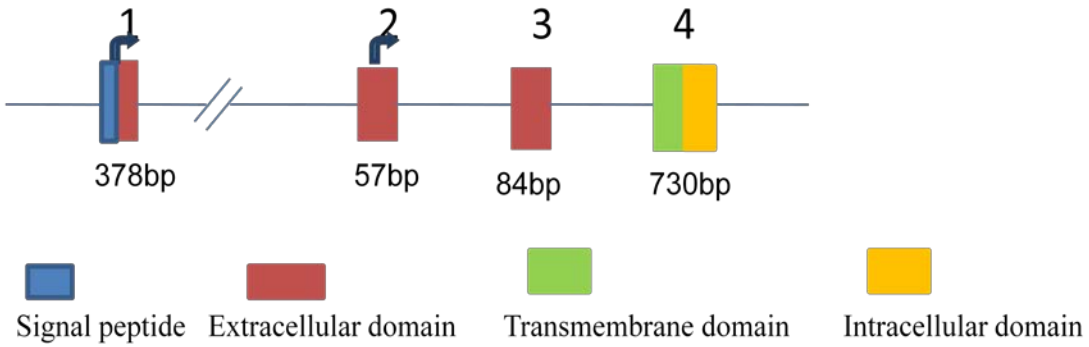


Figure 1.6 Schematic representation of human GYP C gene. The proteins GPC and GPD are encoded via initiation of translation at separate start codons in exons 1and 2, respectively (start sites represented by arrows above the exons). Exons are indicated by boxes and introns by lines. Exons 1 and 2 are separated by an intron that is approximately 34 kb in length (large distance between exons represented by broken lines). Adapted from Wilder *et al*., 2009.

GPC and GPD proteins are expressed on multiple cells, including erythroid, kidney and thymus cells and have several physiological and pathological functions that include maintaining RBC integrity via interaction with Band 4.1 protein (Walker and Reid, 2010). They also contribute to the negative charge of the red blood cell surface (Reid *et al*., 1990; Walker and Reid, 2010). In addition, glycophorin C is involved in *P. falciparum* invasion of RBCs and acts as the receptor for the parasite ligand, EBA-140 (Lobo *et al*., 2003; Maier *et al*., 2003; Pasvol, 2003; Cowman and Crabb, 2006). The binding site on GPC has been putatively identified as lying within exons 2 and 3, although differential binding to variants of GPC lacking either of these exons have also been observed (Lobo *et al*., 2003; Mayer et al., 2004; Wilder et al., 2009).

The two glycoproteins contain the Gerbich (Ge) blood group system with the Gerbich blood group antigens being inherited as autosomal dominant traits (Walker and Reid, 2010). There

are 11 Gerbich antigens, six of which are highly prevalent in malaria endemic regions (Walker and Reid, 2010). Polymorphisms within the glycophorin C gene have been detected, the most common of which are the Gerbich negativity phenotype, where exon 3 is deleted and the Yus phenotype, where exon 2 is deleted (Colin *et al*., 1989; Maier *et al*., 2003; Wilder *et al.*, 2009). Studies conducted by Maier *et al*., (2003) demonstrated that *P. falciparum* cannot use EBA-140 to invade Gerbich-negative erythrocytes because these cells lack a functional GPC, and these cells are therefore fairly resistant to parasite invasion. An earlier study by Mayer *et al.*, (2002), however, showed that polymorphisms in the binding region of EBA-140 lead to variants that bound Gerbich-negative RBCs, thus expanding the invasion capacity of the parasite. The observation that *P. falciparum* cannot use Gerbich-negative glycophorin C molecules to invade RBCs indicates that there are features of the peptide sequence in the deleted region that are vital for parasite ligand binding (Maier *et al.*, 2003; Mayer *et al.*, 2006). In fact, Mayer *et al.*, (2006) suggest that possible interactions between the N-linked and O-linked oligosaccharides found on this exon may be important for the proper exposure of the receptor. The Ge- negative phenotype is prevalent in malaria endemic areas such as Papua New Guinea, and this may represent a selective advantage to populations in these areas (Maier *et al.*, 2003; Pasvol, 2003).

1.6 Malaria as a force of natural selection

Malaria is one of the world's oldest diseases and causes the highest morbidity and mortality (Sabeti, 2008). Due to its debilitating effect, it is hypothesized that humans have had to evolve adaptive traits in order to survive the infection (Kwiatkowski, 2005). As a result, malaria has acted as a strong force of natural selection in the human genome, leading to

selection of genes that confer resistance to the disease in endemic areas (Kwiatkowski, 2005; Ko *et al.,* 2011). The disease has been shown to be the evolutionary driving force behind several genetic diseases such as glucose 6-phosphate dehydrogenase deficiency, sickle-cell disease and alpha-thalassaemia (Kwiatkowski, 2005). The selective pressure of malaria is very strong, as exemplified by the HbS allele (Kwiatkowski, 2005). The allele results from a mutation in the HBB gene which codes for β-globin, causes sickle-cell disease and is lethal in the homozygous state (Kwiatkowski, 2005). In the heterozygous state, HbS confers protection against malaria by as much as 56% in children aged 10 years and below and is widespread in endemic populations (Kwiatkowski, 2005; Williams *et al*., 2005). Different populations have also evolved different independent evolutionary responses to malaria, for example, the lack of the Duffy antigen in certain West African populations prevents RBC infection by *P. vivax* and accounts for the low prevalence of this parasite in these populations, even though other P*lasmodium* species are quite prevalent (Miller *et al,* 1976; Kwiatkowski, 2005).

Natural selection has also been noted on RBC surface proteins (Kwiatkowski, 2005; Ko *et al*., 2011). Sequence analysis of genes coding for molecules functioning as receptors for *P. falciparum* show that these genes are under strong selection, notably the sialoglycoproteins, GYPA and GYPB (Baum *et al.,* 2002; Wang *et al.,* 2003; Kwiatkowski, 2005; Ko *et al,* 2011). GYPC is also under selection and is undergoing evolution, though not as rapidly as the other glycophorin genes (Wilder *et al.,* 2009). Studies by Baum *et al.,* (2002) and Ko *et al.,* (2011) show strong balancing selection in the glycophorin genes, especially glycophorin A. This is evidenced by an excess of intermediate-frequency alleles, which contribute to the maintenance of allelic variation in human populations (Baum *et al*., 2002; Ko *et al*., 2011).

This selection is seen mostly within exon 2 which codes for part of the extracellular domain and which forms part of the binding region for EBA-175 (Baum *et al.,* 2002; Ko *et al.,* 2011). The allelic variation detected, however, was accompanied with little protein divergence, indicating the need to maintain the protein sequence in this region, even in the face of selection (Ko *et al.,* 2011). The protein domain encoded by exon 2 contains binding sites for several O-linked sialoglycans important in parasite ligand binding, and mutations within this region may alter the binding or spatial arrangement of the glycans, thus modifying the binding affinity of *P. falciparum* (Ko *et al.,* 2011). Although a mutation in this case would be advantageous to the host as it could reduce the binding efficiency of the parasite and hence invasion, the protein sequence is maintained and this points to possible importance of the region for a function other than parasite binding.

Several genetic variations have been detected in the glycophorin genes including single nucleotide polymorphisms (SNPs) (in GYPA and GYPB) as well as entire exon deletions (GYPC) (Maier *et al.,* 2003; Wilder *et al.,* 2009; Ko *et al.,* 2011). Association between genetic variations and susceptibility or resistance to disease has been studied in these genes. Exon 3 deletion in GYPC is associated with increased resistance to malaria (Maier *et al.,* 2003; Wilder *et al.,* 2009). The common MNSs polymorphisms in GYPA and GYPB have also been studied, but no association was found between the polymorphisms and susceptibility to severe malaria when a Northern Thailand population was studied (Naka *et al.,* 2007).

1.7 Statistical tests of Neutrality

The current study aims to look at variations in GYPA, GYPB and GYPC in a population living in Kilifi, a malaria endemic region of Kenya. Any variations detected will then be analyzed to determine whether they are under selection using tests of neutrality including Tajima's *D* and Fu and Li's D and Fu and Li's F tests. In population genetics, statistical tests of neutrality are conducted to determine whether the polymorphisms identified are under selection or are evolving neutrally (Akey *et al*., 2004; Duret, 2008). Tests of neutrality measure if data deviate from the expectations under a neutral model, which assumes that many of the variations occurring at the molecular level do not affect fitness, and therefore the evolutionary fate of genetic variation can be explained by stochastic processes such as genetic drift (Simonsen *et al.,* 1995; Akey *et al*., 2004; Duret, 2008). The model assumes a constant population size (Simonsen *et al.,* 1995). This neutral model of evolution is considered a null hypothesis and is rejected in cases where a sequence is under selective pressure (Simonsen *et al.,* 1995; Duret, 2008).

Sequences that do not evolve neutrally are said to be under selection and usually confer some fitness on the organism. Natural selection can be divided into different groups, depending on the effect of the mutation on the organism (Duret, 2008). Positive or directional selection occurs when a mutation that is introduced into a population confers a higher level of fitness and as such the new allele increases in frequency over time, eventually replacing the wild type allele (Duret, 2008). Mutations may also decrease the fitness of an organism, and these will tend to reduce in frequency over time, eventually being wiped out of the population in a process known as negative or purifying selection (Duret, 2008). Mutations that are advantageous in the heterozygous state only, such as the

sickle cell mutation, are maintained at an intermediate frequency in the affected population through a process known as balancing selection (Duret, 2008).

Tests of neutrality such as Tajima's D and Fu and Li's D and F can be used to determine whether a mutation is under selection, and if so, they can also indicate the specific type of selection. The tests essentially measure the same thing, i.e. the hypothesis that all mutations are selectively neutral, but do so using different parameters. With Tajima's D, nucleotide sequence variation is measured using 2 parameters: The number of segregating sites (S) and the average number of pairwise differences (π) which is the sum of pairwise differences divided by the number of pairs. Fu and Li's D* test statistic is based on the difference between η_s, the number of singletons (mutations appearing only once among the sequences), and η, the total number of mutations, while Fu and Li's F* test statistic is based on the differences between η_s, the number of singletons and k, the average number of nucleotide differences between pairs of sequences. In all three cases, negative values signify an excess of low frequency (rare) polymorphisms relative to expectation, indicating population size expansion, such as that occurring after a bottleneck (drastic reduction in population size due to environmental events or human activities) or a selective sweep (reduction of variation in the genetic material due to recent and strong positive selection) (Akey *et al*., 2004). Negative values may also signify purifying selection, which is the removal of deleterious alleles from the population (Akey *et al*., 2004). Positive values signify low levels of both low and high frequency polymorphisms and an excess of intermediate-frequency alleles, indicating a decrease in population size and/or balancing selection (Akey *et al*., 2004). If the value is equal to zero then the sequences are evolving randomly and are not under selection (Akey *et al*., 2004).

Another important test is that of haplotypes diversity and haplotype number. A haplotype refers to a combination of alleles on different loci along a chromosome that are inherited together as a unit (Crawford and Nickerson, 2005). Analysis of haplotypes is important as it gives information about the rate of recombination occurring in a particular gene and this is in turn important in locating disease-causing mutations by linkage methods (Crawford and Nickerson, 2005). Closely tied in with haplotype number is haplotype diversity, Hd, also known as gene/allele diversity or expected heterozygosity, which is a measure of the uniqueness of a particular haplotype in a given population, (Nei and Tajima, 1981; Rozas, 2009). Put in another way, haplotype diversity is the probability that two random sequences are different (Rozas, 2009).

This study is important in identifying variants at these candidate genes which are involved in host-parasite interactions and will shed more light on the impact of malaria on the host genome.

1.8 Project Justification

Malaria acts as a strong selection pressure in human evolution, resulting in numerous genetic variants that confer some level of protection against the disease, especially in areas where it is endemic (López *et al*., 2010; Ko *et al*., 2011). Polymorphisms detected in human RBCs include Southeast Asian Ovalocytosis (SAO) and Gerbich blood group negativity (Wang, 1994; Thathy *et al*., 2005; Khera and Das, 2009). *P. falciparum* also possesses various ligand polymorphisms and there are extensive studies looking into the effects of these polymorphisms on receptor specificity and invasion efficiency (Taylor *et al*., 2002; Mayer *et al*., 2004; Bob *et al*., 2010; Gaur and Chitnis, 2011; Jiang *et al*., 2011). Few studies

aimed at identifying and characterizing RBC polymorphisms and determining their effects on parasite invasion efficiency have been conducted in Africa (Moulds *et al*., 2001; Thathy *et al*., 2005; Migot-Nabias *et al*., 2006; Ko *et al*., 2011) and even fewer in eastern Africa, even though this region is also malaria-endemic. Therefore, studies looking at the RBC receptor polymorphisms are needed in order to get a balanced understanding of the molecular basis of invasion. These studies would also show the impact of malaria disease on the human genome as the long standing struggle for survival between *Plasmodium* parasites and humans are expected to leave footprints in the human genome.

1.9 Objectives

1.9.1 Main objective

To characterize erythrocyte receptor polymorphisms present in a human population living in the malaria endemic region of Kilifi, Kenya.

1.9.2 Specific Objectives

1. To genotype the erythrocyte receptors: Glycophorins A, B and C
2. To identify the polymorphisms present in these receptors and determine their allele frequency distribution
3. To determine whether identified polymorphisms are under natural selection

CHAPTER TWO

MATERIALS AND METHODS

2.1 Study population

The study was conducted in Kilifi County, one of the 47 counties in Kenya. Kilifi County is located in the southern coast of the country, along the Indian Ocean (*Scott et al*., 2012). The area is a malaria endemic region, although transmission has been declining in recent years (*Bejon et al*., 2010). Malaria transmission in this region is seasonal and follows two rainy seasons from April to July and from October to November (*Scott et al*., 2012). The specific study region is part of the Kilifi Health and Demographic Surveillance System (KHDSS) which covers an area of 891km^2 and has a population of 261,919 people (as of March 2011), 18% of whom are under 5 years of age (*Scott et al*., 2012). The surveillance system links community-based surveillance with hospital-based surveillance to provide real-time epidemiological data for different diseases.

2.2 DNA Samples

93 human DNA (2n = 186, representing the diploid number of alleles in all the samples) cross sectional samples extracted in 2001 were used. The DNA was extracted from blood taken from patients admitted to the Kilifi District hospital's High Dependency Unit (HDU) with severe malaria. The individuals had highly variable parasitemia, with a median of 190,000 parasites/µl. The study participants ranged in age from 6 months to 8 years, with an average age of 3.5 years.

2.3 Primer design

Full length sequences for human glycophorin genes A (accession no. NG_007470.3), B (accession no NG_007483.2) and C (accession no. NG_007479.1) were retrieved from the National Center for Biotechnology Information (NCBI) RefSeq gene website (according to NCBI, the glycophorin A sequence was derived from a Taiwanese population, while the population origins of glycophorin B and C sequences are not indicated). The sequences were imported into EditSeq application (DNASTAR, Lasergene Suite) and the regions of interest retrieved. These were: GYPA (exons 2 – 5), GYPB (exons 2 – 5) and GYPC (exons 2 – 3) as well as intronic regions bordering the external exons (Table 2.1). These regions were chosen based on available data identifying them as confirmed or putative binding regions of *P. falciparum* (Baum *et al.*, 2002; Ko *et al.*, 2011) invasion ligands. The regions were used to design gene-specific PCR (F1 and R1 for each gene) and sequencing primers (Table 2.2) using the SeqBuilder application in DNASTAR, Lasergene software suite. The primers were then ordered from Sigma-Aldrich. In total, 8 primers were designed for GYP A, 8 primers for GYP B and 10 primers for GYP C.

Table 2.1: Information on the structure of the various erythrocyte genes. Data derived from NCBI database.

Receptor	Location	Gene size (kb)	accession No.	Exons	Sequenced region (bp)	Exons size (kb)
GPA	4q31.21	31.439	NG_007470	7	2 (99)	4.0
					3 (96)	
					4 (39)	
					5 (86)	
GPB	4q31.21	23.240	NG_007483	5	2 (99)	5.3
					3 (39)	
					4 (95)	
					5 (201)	
GPC	2q14-q21	40.563	NG_007479	4	2 (57)	4.0
					3 (84)	

Table 2.2: PCR and sequencing primers

Gene	Primer name	Sequence
Glycophorin A	GYPA F1	AGGCCAATA ATACAATACTTACCA
	GYPA F2	AGCTGGGTGCTGAGATAAGAGTAA
	GYPA F3	TGG GTCATTAAAAGTAGATACTGA
	GYPA F4	TCATCAACTATTTTGCGACCTTGGA
	GYPA R1	GACAGATTTATATTTAGAGGTTCC
	GYPA R2	ACAACATATGCTCTTCTAAGATAG
	GYPA R3	GGTCACAGTTAATAGTTGTGGGTGC
	GYPA R4	TGCTTATTGTTTATCAGTCACTG
Glycophorin B	GYPB F1	CAATAATACAATACTTACCACGCA
	GYPB F2	TCATCCATGAATACGTGTTGGG
	GYPB F3	ACACATAATAGTATGTTAACTGTAC
	GYPB F4	TGGAGTCTTAGCTCATGGTCAA
	GYPB R1	CATGAGACTTCATGTTATCTTGGA
	GYPB R2	TCTTCTGAGTTTAACTGAACTCAG
	GYPB R3	TTGTCTTTACAATTTCGTGTGA
	GYPB R4	GCTAGAATTCCTCTGTAGTAAGA
Glycophorin C	GYPC F1	GCATACTGCAGAGAACTTAAATG
	GYPC F2	TGAAGGCAGTAGGAAGTTTAG
	GYPC F3	ACCCATTATGACTCCTGACT
	GYPC F4	CAAGCAGAACCTGGTTCCT
	GYPC R1	TACATACATAGATACGTACGATGTA
	GYPC R2	AGTGTCTGATGCTCACACCA
	GYPC R3	AGTGCTATGAGGTCATTCTC
	GYPC R4	GGAAGAAGATAATACAAGGCAC
	GYPC R5	TTACTTTGAAGCTGCAGAGCA
	GYPC R6	TAAGTAGCCTGGTCAGTCTA

2.4 PCR optimizations and gene amplification

Gradient PCR reactions were conducted on the veriti 96 well thermal cycler (Applied Biosystems) to determine the optimal conditions for amplification of the three genes, using a range of annealing temperatures. Optimizations for GYPA and GYPC involved an initial denaturation at 94^{o}C for 2 minutes; 10 cycles of denaturation at 94^{o}C for 15 seconds, annealing temperature range of 45^{o}C - 49^{o}C for 30 seconds, extension at 72^{o}C for 4 minutes; 25 cycles of denaturation at 94^{o}C for 15 seconds, annealing temperature range of 45^{o}C - 49^{o}C for 30 seconds, extension at 72^{o}C for 4 minutes with an increment of 5 seconds per cycle; a final extension at 72^{o}C for 7 minutes. Optimization conditions for GYP B were

similar as for GYP A and GYP C, except for the extension temperature which was carried out at 68^{o}C.

PCR cycles consisted of an initial denaturation at 94^{o}C for 2 minutes; 10 cycles of denaturation at 94^{o}C for 15 seconds, annealing temperature range of 45^{o}C - 49^{o}C for 30 seconds, extension at 68^{o}C for 4 minutes; 25 cycles of denaturation at 94^{o}C for 15 seconds, annealing temperature range of 45^{o}C - 49^{o}C for 30 seconds, extension at 68^{o}C for 4 minutes with an increment of 5 seconds per cycle; a final extension at 68^{o}C for 7 minutes.

Following optimization, a 10µl reaction was set up in two parts as follows: in one microcentrifuge tube, 1X PCR buffer 4 (Roche), 500µM dNTPs (Bioline), 0.3µM of forward and reverse primers (Sigma-Aldrich), 50ng – 150ng of DNA template depending on the original concentration of the template, and DNase free water to give a final volume of 5.0µl. In a second tube, 0.49 units of Expand High Fidelity Taq Polymerase (Roche), 1X PCR buffer 2 (Roche) and DNase free water to make a final volume of 5µl. Contents of the two tubes were then mixed to give a final volume of 10µl per PCR reaction. PCR was carried out in 96 well plates (Applied Biosystems). For the actual PCR reactions, the above optimization reaction conditions were maintained, but with annealing temperatures of 48^{o}C for glycophorins A and C and 49^{o}C for glycophorin B which were determined to be optimal for the genes.

2.5 Gel Electrophoresis

Following PCR, gel electrophoresis was conducted to determine which samples successfully amplified. A 1% agarose gel was prepared using 0.5X Tris/Borate/EDTA (TBE) buffer.

Briefly, 100ml stock of 10X TBE buffer was prepared by mixing 10.8g of tris base and 5.5g of boric acid in 50ml of distilled water. 4ml of 0.5M EDTA (pH 8.0) was added and the mixture made up to 100ml; a 1:20 dilution of the TBE stock solution was made to get a working solution of 0.5X. To make a 1% agarose gel, 1g of agarose powder (Applied Genetic Technologies Corporation) was weighed, added to 100ml of 0.5X TBE and heated to boiling. The solution was then left to cool before adding 2µl of Ethidium Bromide and pouring in a gel tray with combs to set. 1µl of each PCR product, as well as negative and positive control samples were individually mixed with 1µl of 6X Blue Orange loading dye (Promega) and loaded into wells on the gel. DNase-free water was used as the negative control while human DNA sample that was previously shown to amplify these genes was used as a positive control. 1.5µl each of 1kb Hyperladder 1 (Bioline) was loaded into the first and final wells on the gel. The samples were then run, using 0.5X TBE buffer, for 45 minutes at 100 volts. The gels were viewed under ultraviolet light and photos taken on a Molecular Imager Gel Doc (Bio-Rad).

2.6 Purification of PCR products

Purification involved an enzymatic clean-up procedure. Briefly; successfully amplified PCR products were cleaned using ExoSAP-IT reagent (Affymetrix). 3.6µl of ExoSAP-IT reagent was mixed directly with 9.0µl of PCR product. The samples were loaded onto averiti 96 well thermal cycler (Applied Biosystems) and incubated at 37^{o}C for 15 minutes to degrade the remaining primers and nucleotides. The products were then incubated at 80^{o}C for 15 minutes to inactivate the ExoSAP-IT enzymes and this was followed by incubation at 15^{o}C for 5 minutes.

2.7 BigDye Sequencing PCR reaction

For each gene, 1.5ml Eppendorf microcentrifuge tubes corresponding to the total number of PCR and sequencing primers for each gene were set up as follows: each reaction mixture was set up by combining 0.5µl of BigDye terminator ready reaction mix (Applied Biosystems), 2.0µl of 5X sequencing buffer, 1.0µl of 10µM sequencing primer, 5.5µl of DNase free water and 1.0µl of ExoSAP-IT cleaned PCR product to give a final volume of 10µl per reaction. Each sequencing primer used in the reaction was added into a different mastermix tube. The plates were then loaded onto the thermocycler and a sequencing program set up as follows: 25 cycles of denaturation at 96^{o}C for 30 seconds, annealing at 50^{o}C for 15 seconds and extension at 60^{o}C for 4 minutes, with a ramp rate of 1^{o}C per second between the different temperatures.

The BigDye terminator sequencing chemistry is a modification of Sanger sequencing where dideoxynucleotides (ddNTPs) with specific fluorescent dyes attached to each base are added into the reaction. During the sequencing PCR, dNTPs are added to the elongating strand. If a ddNTP is added onto the strand, synthesis is terminated as the ddNTP does not contain a hydroxyl group at its 3' end to allow for continued elongation. This results in PCR products of different lengths, each terminating at different locations along the strand.

2.8 Purification of sequenced PCR products

Purification was carried out using Ethanol/Sodium Acetate precipitation in 96 well plates. This method involved constituting a premix of 3µl of Sodium Acetate, pH 5.2, 62.5µl of 95% ethanol and 24.5µl of distilled water to make a final volume of 90µl per well. 90µl of

the premix was added to each well containing the PCR products. The plates were then sealed with microseals (Bio-Rad) and incubated at -20^{o}C for 30 minutes. Following incubation, the plates were spun at 3000xg for 30 minutes at 4^{o}C on a 5810R bench centrifuge (Eppendorf). The seals were then removed and the plates overlaid with absorbent paper towels and gently inverted. The inverted plates were then spun at 50xg for 1 minute at 4^{o}C. 150µL of ice cold (-20^{o}C) 70% ethanol was then added into each well, the plate sealed and spun at 3000xg for 10 minutes at 4^{o}C. After this the plates were once again inverted over paper towels and excess fluid gently drained. The plates were then overlaid with clean paper towels, inverted and spun at 50xg for 1 minute at 4^{o}C. The plates were then covered with fresh paper towels and left on the bench to air dry.

2.9 Capillary electrophoresis

After air drying the plates, 10µl of thawed HiDi (Formamide) (Applied Biosystems) was added into each well, the plates sealed and heated for 3 minutes at 96^{o}C. Two plates were loaded at a time on an ABI 3130XL capillary sequencer for reading and generation of sequence data trace files. During reading of the sequencing PCR products, different DNA fragments with a dye labeled termination electrophorese through the polymer in the capillary, leading to fragment separation based on molecular weight. The DNA fragments pass a detection window at the end of the capillary where the fluorescent dyes are excited with a laser beam and emit light wavelengths unique to each terminator. The emitted light is then captured, converted to electronic information and sent to a computer which analyzes the signal to generate sequence data.

2.10 Data Analysis

Sequence electropherogram data generated on the genetic analyzer were imported into Seqman application (DNASTAR Lasergene Suite, Version 10) for analysis. For each sample, sequences generated from the different primer extensions were aligned into contigs and each primer trace file assessed for the quality of peaks and base calling. The reference sequences (Glycophorin A, accession no. NG_007470.3; Glycophorin B, accession no NG_007483.2; Glycophorin C, accession no NG_007479.1) were used to scaffold the trace data generated from each primer for each gene. Corrections to base calling were done on the basis of the peaks of the electropherogram and independent of the reference sequence. Clean scaffolds were saved as consensus files.

A multiple alignment of the consensus files for all samples of each gene was carried out in the MegAlign application (DNASTAR, Lasergene Suite) using Clustal W method and the alignments confirmed visually. Mis-aligned sequences were corrected manually. Multi-alignment was done to identify polymorphic positions in the samples for each gene under analysis. The sequencing process did not produce the anticipated exon ranges for most of the samples, thus multi-aligned sequences were saved as fragments encompassing the exons and bordering intronic regions. The alignments were saved in FASTA format and imported into DnaSP V 5.10.01 software for statistical analysis, including Tajima's D, Fu and Li's D and Fu and Li's F using the program's default parameters. Under these three tests, the total number of segregating sites, nucleotide diversity and the total number of nucleotide differences were computed. Using a sliding window, Tajima's D graphs were generated for those regions that showed statistically significant values. Linkage disequilibrium (LD) plots were generated in the DnaSP software, employing two-tailed Fisher's exact test to determine

statistical significance. Allele frequency distribution was calculated for the polymorphic sites in the regions under analysis and Hardy-Weinberg equilibrium calculated to determine whether the population is undergoing evolution or whether it is at equilibrium. Haplotypes circulating in the population were also identified, based on observed SNPs.

CHAPTER THREE

RESULTS

3.1 PCR amplification

Of the 93 DNA samples that were used in the study, successful PCR amplification was achieved for 79 samples for glycophorin A, representing a 4kb amplicon, 79 samples for glycophorin C, representing a 4kb amplicon, and 63 samples for glycophorin B, representing a 5.3kb amplicon (Fig.3.1). Amplification of the rest of the samples was unsuccessful even after three attempts using increasing template volumes. These samples were those that had low template concentrations.

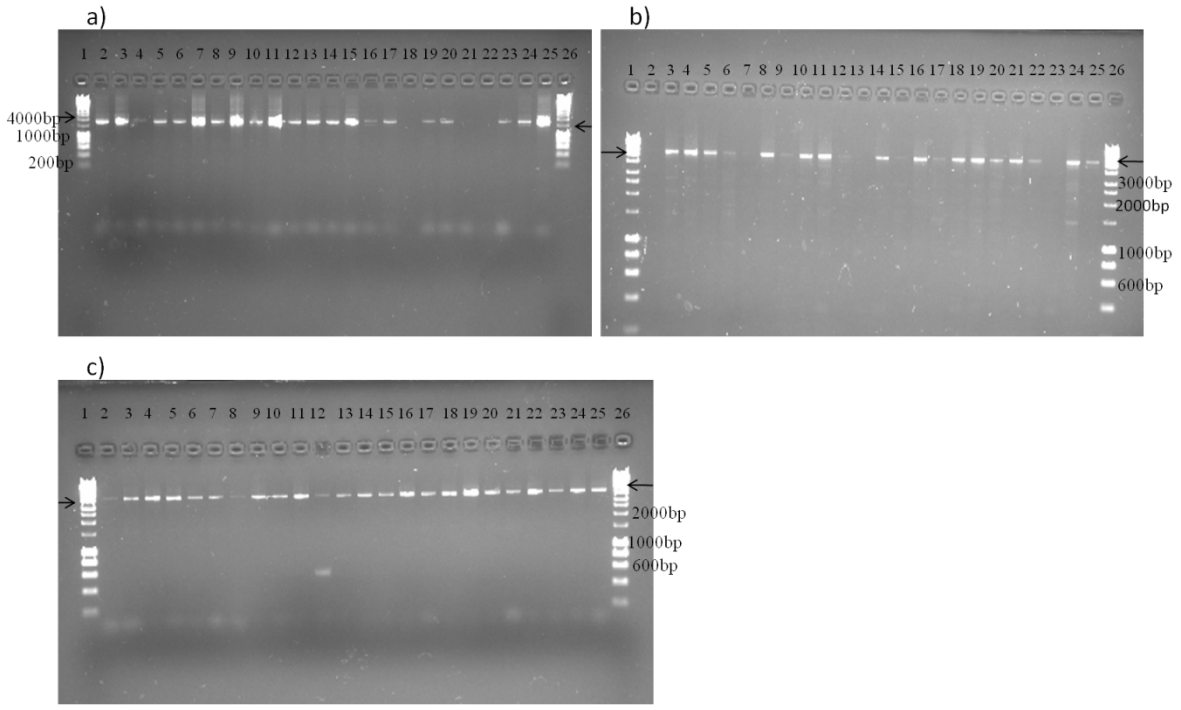


Figure 3.1 Gene amplification products of a) GYP A, b) GYP B and c) GYP C. lanes 1 and 26 in each gel picture represents Bioline's Hyperladder I used to determine amplicon size. Bands between the ladders represent amplified products of different samples. Positions that show no bands represent samples that did not amplify. Arrows mark positions of a) 4000bp, b) 5300bp and c) 4000bp, the sizes of glycophorins A, B and C respectively on the ladder.

In this study, a 1% agarose gel was used to resolve the PCR fragments. The percentage of agarose used in electrophoresis depends on the size of the fragments to be separated. Low percentage agarose gels of between 0.7% - 1.0% are used for resolving large products of between 2kb – 10kb while higher percentage gels of 2% can achieve a good resolution of small fragments of between 0.2 – 1kb (Lewis, 2001).

3.2 Glycophorin Gene Sequencing

Successfully amplified samples were sequenced using BigDye terminator chemistry and the sequences analyzed using the ABI 3130xl genetic analyzer. Following the base calling, primer trace files were imported into Seqman application of DNASTAR Lasergene Suite. Clean primer trace files were generated for 67 samples of glycophorin A, 57 samples for glycophorin B and 67 samples for glycophorin C. For the rest of the sequences, the electropherograms could not be read as some of them contained high background noise while others had multiple peaks at the same positions, making accurate base calling impossible. The samples that failed to yield good sequences were shown to be those that did not have very distinct PCR amplification bands and the problem may have been due to interference from non-specific amplified products during the sequencing reaction, a problem that is common with direct sequencing of PCR products. These primer traces were left out of subsequent analysis and no attempt was made at cloning the samples that did not generate good sequences as the samples that were successfully sequenced were sufficient for downstream analysis. Full length sequence data was not obtained for any of the samples of glycophorins A and B. Consensus files generated from the assembled contigs were thus fragmented. Full length sequences (4.0kb) were obtained for 12 samples of glycophorin C. Due to unavailability of full length sequence data for most of the samples, subsequent analysis was done on fragments of sequences encompassing the exons under analysis together with as much intronic regions bordering the exons as possible (Table 3.1). Attempts to obtain the longest possible sequence data for the different gene regions led to drop-offs in the number of samples used in subsequent analyses from the original clean trace files

generated (Table 3.1). In the case of the 12 full length sequences of glycophorin C, analysis was done for all of them.

Table 3.1: Regions of Glycophorins A, B and C that were sequenced and analyzed in a Kilifi population

Region	Number of samples	Length of region(bp)
Glycophorin A		
Exon 2-3	51	1021
Exon 4	52	240
Exon 5	52	374
Glycophorin B		
Exon 2	45	376
Exon 3	29	865
Exon 4	31	316
Exon 5	40	384
Glycophorin C		
Exon 2	53	550
Intron 2	48	616
Exon 3	39	189
Full length	12	4057

bp, base pairs.

3.3 Analysis of Segregating Sites:

The MegAlign application was used to identify polymorphic or segregating sites within the different regions in the samples. Segregating sites were defined as those differing from the reference sequence at specific positions. Multiple segregating sites were detected in the samples in comparison to the reference glycophorin A (accession no. NG_007470.3), B (accession no NG_007483.2) and C (accession noNG_007479.1) sequences. Glycophorin A showed the highest number of samples with segregating/polymorphic sites among the three genes with a total of 85 segregating sites from the 67 samples. Glycophorin B gave a total of 53 segregating sites from the 57 samples sequenced. Glycophorin C had the least number of segregating sites, with 67 samples segregating at 34 sites. A single base pair insertion was

detected at position 25322 in all the glycophorin A gene samples compared to the reference sequence which was obtained from a Taiwanese population (Fig 3.2). This insertion was not present in the highly homologous glycophorin B gene of the samples analyzed. Glycophorin A also gave 4 different variants at position 25944 in the samples analyzed (Fig 3.2). Of the samples sequenced, 10 had bases TT, 5 had CC, 1 had CT while the remaining 51 had GG at this position.

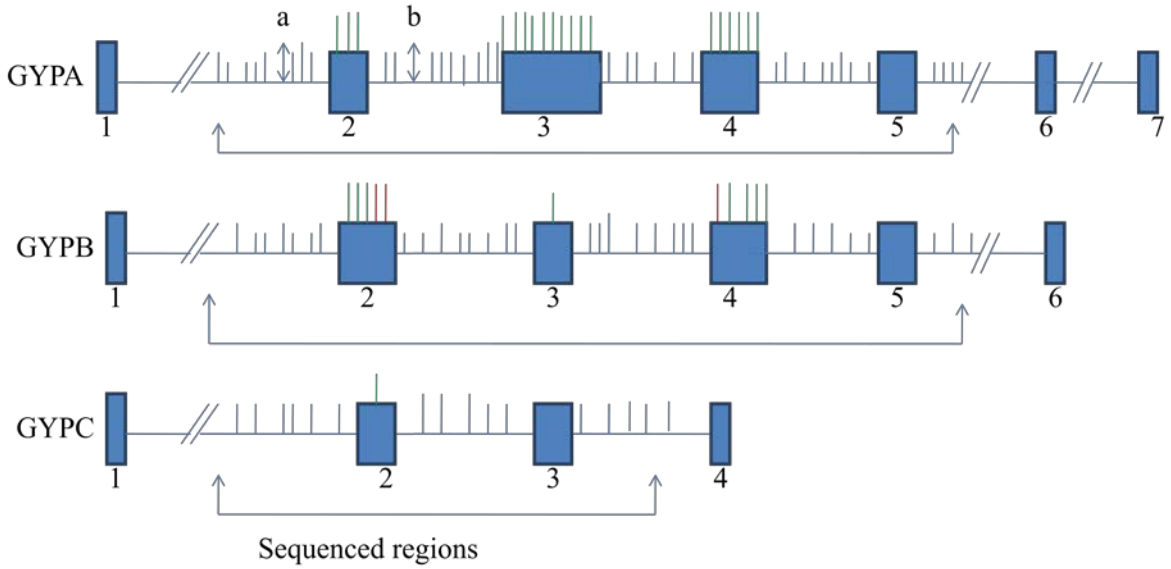


Figure 3.2. Schematic representation of the sequenced regions of glycophorins A, B and C. Numbered blocks represent exons while inter-joining lines represent introns. Red and green lines on top of each exon represent synonymous and non-synonymous SNPs, respectively. Lines on the introns represent SNPs identified in these regions (number of lines do not represent actual numbers of SNPs). [a]position 25322, [b]position 25944. Breaks within introns indicate separation of adjacent exons by large intronic regions.

Most of the polymorphisms detected occurred within the intronic regions of the three genes, with fewer polymorphisms being detected within the exonic regions. 3 non-synonymous SNPs that lead to a change in the encoded amino acids were detected in exon 2 of

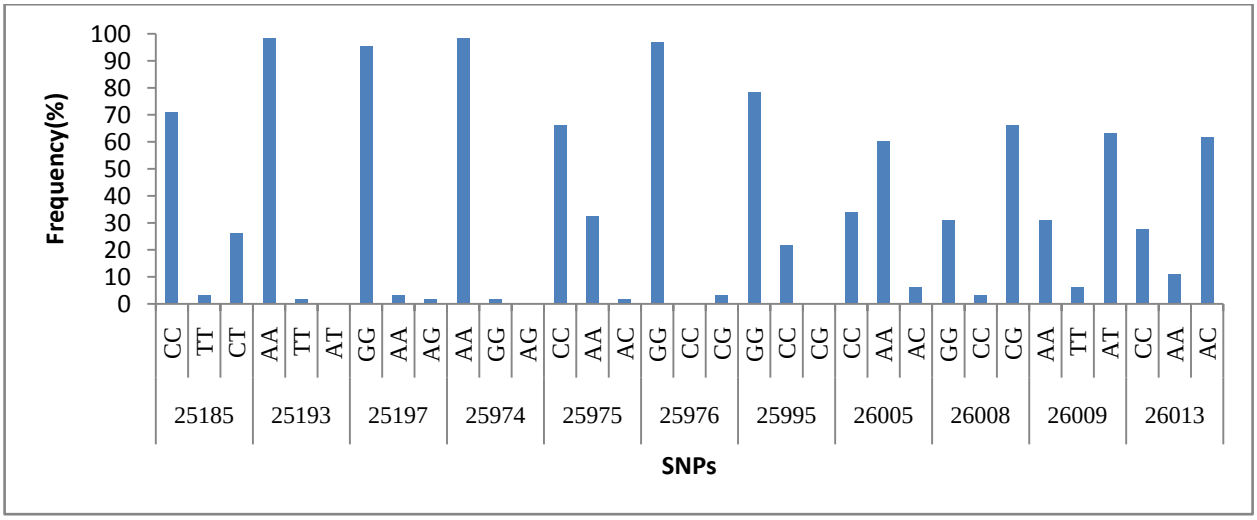
glycophorin A (Table 3.2). 10 non-synonymous SNPs were detected in exon 3 while 6 non-synonymous mutations were detected in exon 4. No amino acid encoding SNP was detected in exon 5 of the analyzed sequences of glycophorin A. Analysis of glycophorin B gene in the samples showed the presence of 3 non-synonymous and 2 synonymous SNPs in exon 2 and 1 non-synonymous SNP in exon 3. Exon 4 of the same gene was shown to have 1 synonymous and 4 non-synonymous SNPs (Table 3.2). Analysis of the less polymorphic glycophorin C showed the presence of a single non-synonymous polymorphism in exon 2 (Table 3.2) while the analysis of exon 3 did not identify any polymorphic sites. In the case of all genes, the non-synonymous mutations were present in high proportions (>50%) within the population.

Table 3.2: Synonymous and non-synonymous variations detected in the glycophorin regions.

Gene region	Ref. codon	Ref. amino acid	Variable codon	Variable amino acid
GYP A				
Exon 2	TCA	Serine	TTA	Leucine
	ACT	Threonine	TCT	Serine
	GGT	Glycine	GAT	Aspartic acid
Exon 3	ACG	Threonine	GAC	Aspartic acid
	GGA	Glycine	CGA	Arginine
	CCT	proline	CAT	Histidine
	AGA	Arginine	ACT	Threonine
	CAT	Histidine	AAT	Asparagine
	AGA	Arginine	ACA	threonine
	TAC	Tyrosine	TCC	Serine
	GAG	Glutamic acid	AAG	Lysine
	GAA	Glutamic acid	AAA	Lysine
	ACC	threonine	AAC	asparagine
Exon 4	AGG	Arginine	ACG	Threonine
	GTA	Valine	GGA	Glycine
	GCC	Alanine	GTC	Valine
	CAT	Histidine	CGT	Arginine
	TCT	Serine	ACT	Threonine
	GAA	Glutamic acid	GTA	valine
GYP B				
Exon 2	TTA	Leucine	TGG	Tryptophan
	ACT	Threonine	TCT	Serine
	GAG	Glutamic acid	GGT	Glycine
	TCT	Serine	TCC	Serine
	TCA	Serine	TCG	Serine
Exon 3	ACG	Threonine	ATG	Methionine
Exon 4	CTC	Leucine	CTG	Leucine
	ATT	Isoleucine	ATG	Methionine
	GGT	glycine	GCT	alanine
	ACG	Threonine	ATG	Methionine
	ACT	threonine	AGT	Serine
GYP C				
Exon 2	CCG	proline	CTG	Leucine

Allele frequencies were calculated for all the SNPs occurring in the intronic and exonic regions of the genes (Fig 3.3). Results indicate that heterozygous SNPs were present at higher frequencies than homozygous SNPs (Fig 3.3).

a)



b)

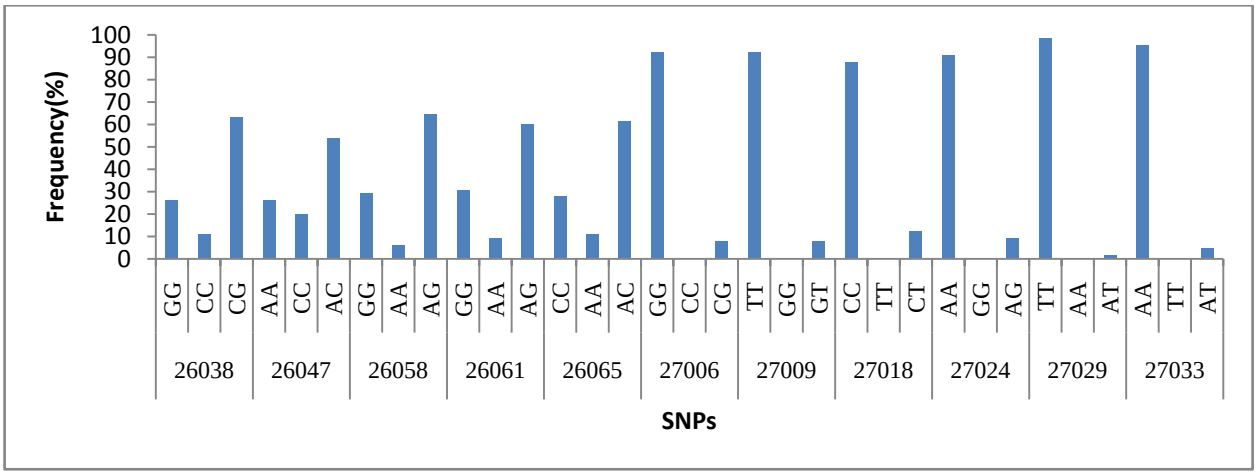
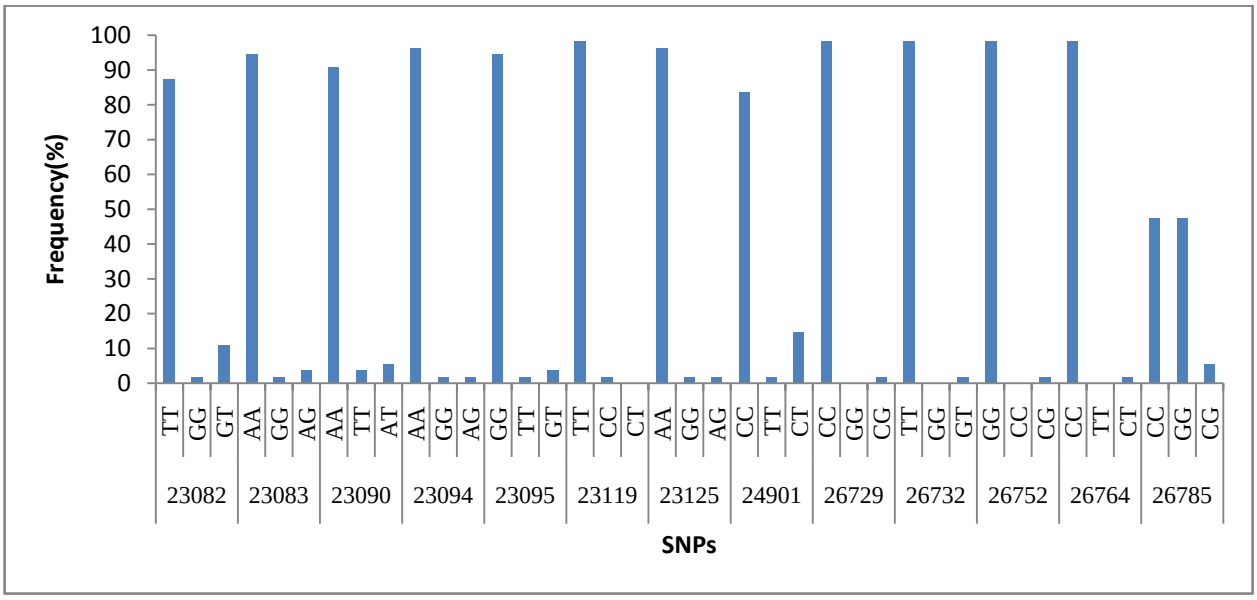


Figure 3.3 Frequencies of SNPs occurring within the coding regions (exons 2 – 5) of glycophorin A. Numbers below the SNPs indicate the SNP position relative to the full length gene. The first base pair under each SNP position represents the homozygous reference alleles, the second base pair represents homozygous non-reference alleles and the third base pair represents heterozygous alleles.

a)



b)

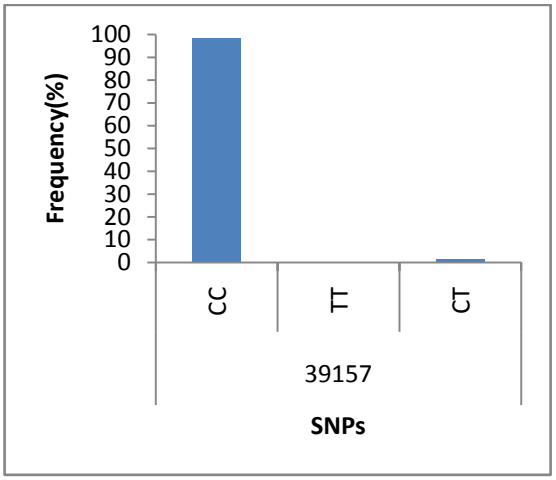


Figure 3.4 Frequencies of SNPs occurring within the coding regions of (a) glycophorin B (exons 2 – 5) and (b) glycophorin c (exon 2). Numbers below the SNPs indicate the SNP position relative to the full length gene. The first base pair under each SNP position represents the homozygous reference alleles, the second base pair represents homozygous non-reference alleles and the third base pair represents heterozygous alleles.

Haplotypes resulting from detected SNPs were also determined by listing the order of SNPs as they appeared in the exons. Glycophorin C had only two haplotypes (Table 3.3), with the

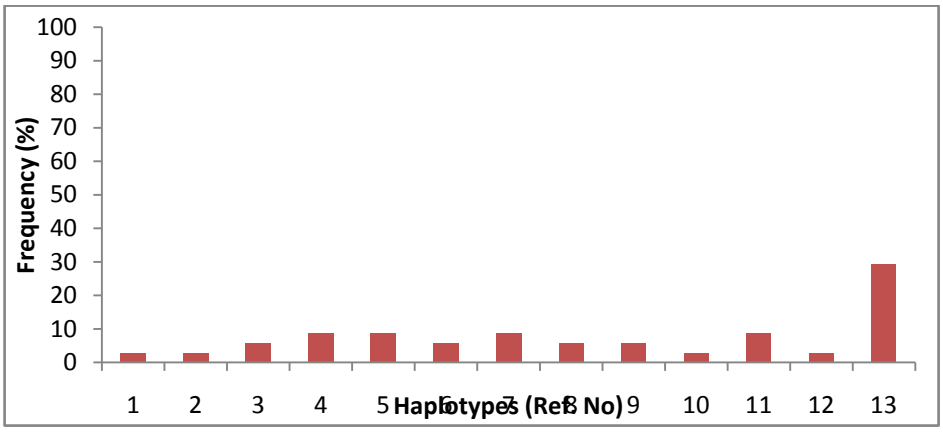
reference haplotype (containing proline) being observed at a higher frequency in the population (98%). Glycophorin B had four haplotypes circulating in the population (Table 3.3), with LTDSSTLIGTS (Ref. No. 3) being the most common haplotype in the population (54%). Glycophorin A had the highest number of haplotypes at 13 (Table 3.3), with STGTGHRHRSKDTRVAHSD (Ref. No.13) being the most prevalent haplotype (29%).

Table 3.3 Circulating haplotypes of glycophorins A, B and C in the Kilifi population.

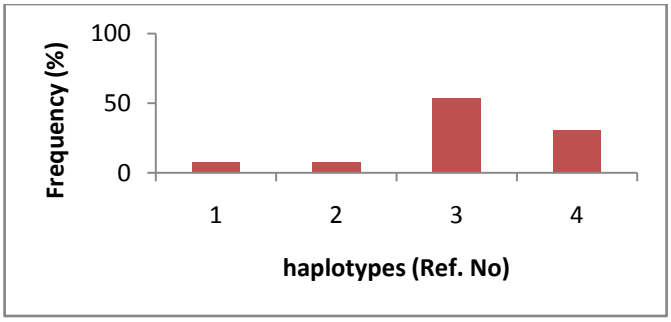
Ref. No	GYP A	Ref. No	GYP B	GYP C
1	LTGTGHRHRSKDTRVAHSD	1	WSGSSTLIGMT	P
2	STDTRHTHTSKDNRVAHSD	2	LSDSSMLIGTS	L
3	STGTGHRHRYDDTRVAHSD	3	LTDSSTLIGTS	
4	STGTGHRHRSKKTRVAHSD	4	LTDSSTLIGTT	
5	STGTGHRHRSKDTRVAHSV			
6	STGTGHRHRSKDTRVVRSV			
7	STGTGHRHRSKDTRVVHSD			
8	STGTRHRHRSKKTRVAHSD			
9	STGTGHTHRSKDTRVAHSV			
10	STGTGPRHRSKKTRVAHSD			
11	STGTGHTNRSKDTRVAHSD			
12	STGTRHTHTSKDNRVAHSD			
13	STGTGHRHRSKDTRVAHSD			

GYP A has the highest number of haplotypes (13) while GYP C has the least (2). Ref. No represents the specific haplotype in the haplotype frequency graphs below (Fig 3.5).

a)



b)



c)

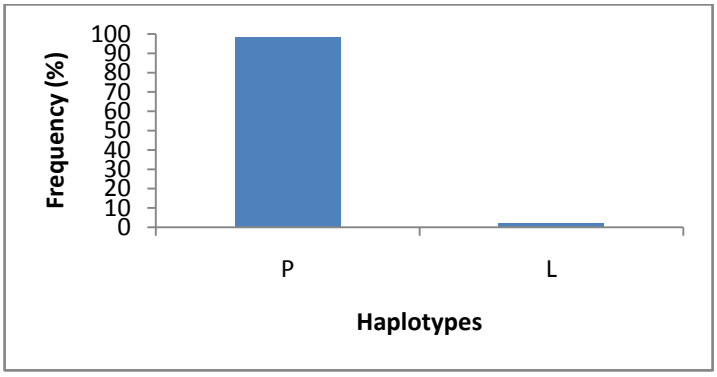


Figure 3.5 Frequencies of a) GYP A, b) GYP B and c) GYP C haplotypes circulating in the Kilifi population. Ref. No refers to individual haplotypes as listed in table 3.3 above. Actual haplotypes were not included in the frequency graphs due to their lengths.

3.4 Statistical Analysis

To determine whether the SNPs detected were under selection, three statistical tests of neutrality were conducted (Table 3.4). These were Tajima's D, Fu and Li's D and Fu and Li's F tests, executed in DnaSP software. Nucleotide diversity, π, and the average number of nucleotide differences, k, were also determined. For all the polymorphisms detected in exons 2-5 of glycophorin A, Tajima's D values calculated were not statistically significant ($p>0.10$) (Table 3.4). However, the segregating sites in exons 2, 3 and 5 and their bordering intronic regions had statistically significant Fu and Li's D values ($p<0.02$) and Fu and Li's F values ($p<0.05$). When the adjacent intronic regions were removed from the analysis, the SNPs in both exons 2 and 3 gave positive statistically significant values of Tajima's D ($p<0.05$), Fu and Li's D ($p<005$) and Fu and Li's F ($p<0.02$). The SNPs in exon 4 of GYP A were not statistically significant ($p>0.10$) under any of the tests. Analysis of glycophorin B exons together with bordering introns produced negative Tajima's D values that were statistically significant ($p<0.05$ and $p<0.02$). Both Fu and Li tests also gave negative, statistically significant values for exon 2 ($p<0.02$). When the exons were analyzed without the introns, only the SNPs in exon 4 were significant ($p<0.05$), giving negative values under all three tests. Glycophorin C analysis showed that the polymorphisms in exons 2 and 3 and their adjacent regions were statistically significant ($p<0.05$ and $p<0.01$ respectively). The single SNP detected in exon 2, however, was not significant when the exon was analyzed on its own (Table 3.4).

Table 3.4: Tajima's D and Fu and Li's values for different regions of the glycophorins A, B and C

	S	Seq. used	π	k	Tajima's D	Fu and Li's D	Fu and Li's F	Hd
Gyp A								
Exons only								
Exon 2	3	102	0.00962	0.95244	2.26*	0.68	1.37	
Exon 3	13	104	0.05503	5.28305	2.38*	1.62*	2.10**	
Exon 4	6	104	0.01317	0.51382	1.24	0.16	0.36	
Exons and bordering introns								
Exon2 and 3	49	102	0.01502	15.33508	1.43	1.80**	1.98*	0.99
Exon 4	15	104	0.00745	1.78734	-1.04	-0.05	-0.49	0.55
Exon 5	20	104	0.01509	5.64451	1.36	1.74**	1.91*	0.98
Gyp B								
Exons only								
Exon 2	9	90	0.00807	0.79850	-1.54	-0.004	-0.64	
Exon 3	2	58	0.00833	0.32486	-0.43	-0.93	-0.91	
Exon 4	7	62	0.00451	0.42834	-1.81*	-2.82*	-2.93*	
Exons and bordering introns								
Exon 2	32	90	0.00582	2.18652	-2.10*	-3.29**	-3.38**	0.7
Exon 3	18	58	0.00186	1.60738	-1.80*	-1.27	-1.73	0.83
Exon 4	15	62	0.00267	0.84241	-2.18**	-2.31	-2.69*	0,51
Exon 5	19	78	0.00424	1.62737	-1.86*	-1.08	-1.63	0.73
Gyp C								
Exons only								
Exon 2	1	106	0.00033	0.01887	-1.02137	-2.0461	-2.02574	
Exons and bordering introns								
Exon 2	14	106	0.00139	0.76586	-1.93*	-3.61**	-3.58**	0.46
Intron 2	17	96	0.00354	2.18092	1.280	-3.40**	-3.56**	
Exon 3	11	78	0.00253	0.47852	-2.35 **	-4.91**	-4.76**	0.3
Full length	31	24	0.00211	8.56884	0.12	0.52	0.46	0.95

The following abbreviations are used: π, Nucleotide diversity; k, Average number of nucleotide differences; S, number of polymorphic (segregating) sites. For Tajima's D, *, $p < 0.05$; **, $p < 0.01$ for Fu and Li's analysis, *, $p < 0.05$; **, $p < 0.02$. Number of sequences used are double the number of samples in each case, reflecting the diploid state of the samples; Hd, Haplotype diversity.

For all regions that returned statistically significant Tajima's D values, Tajima's D was also calculated using a sliding window 100 sites long, with a step size of 25 bases. Subsequent results were used to produce DnaSP graphs.

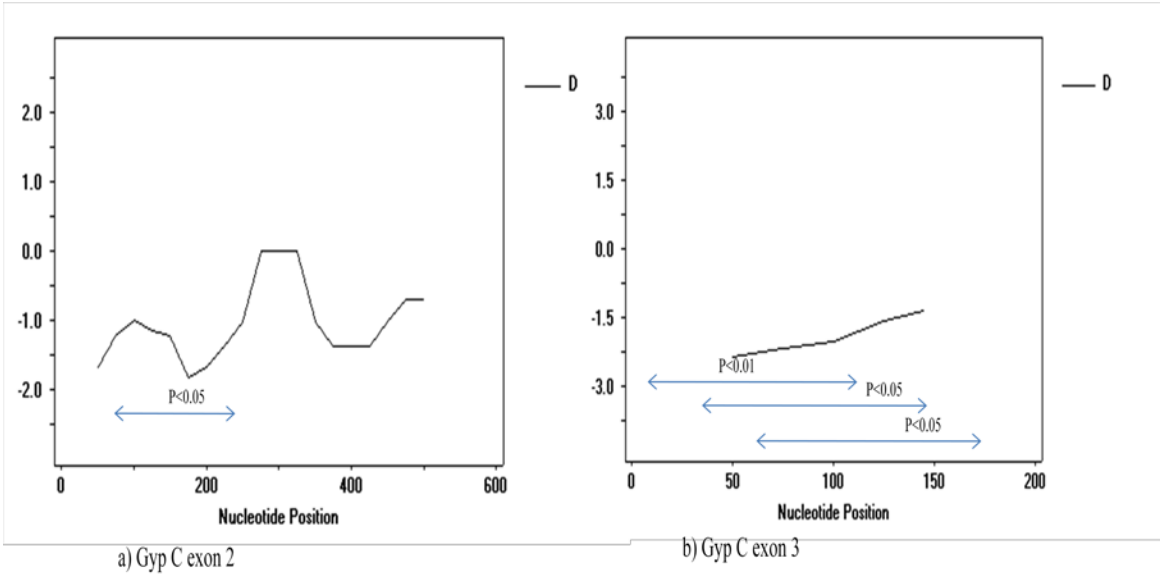


Figure 3.6 Tajima's D graphs for exonic and bordering intronic regions of glycophorin C. The graphs represent those regions that showed statistically significant Tajima's D values. The graphs were constructed using a sliding window 100 sites long and a step size of 25 sites. p refers to the p-value for statistical significance.

Figure 3.7 Tajima's D graphs for exonic and bordering intronic regions of glycophorin B. The graphs represent those regions that showed statistically significant Tajima's D values. The graphs were constructed using a sliding window 100 sites long and a step size of 25 sites. p refers to the p-value for statistical significance.

Degree of Linkage Disequilibrium (LD) was also calculated in the DnaSP software to determine whether the SNPs detected were in linkage disequilibrium. LD was computed for pairs of SNPs, with significance of pairwise comparisons being determined by the two-tailed Fisher's exact test. Results indicate that several of the SNP pairs analyzed showed significant associations, meaning that they were in LD (Fig. 3.8 – 3.10). Glycophorin A had the highest number of significant pairwise associations, with exon 2 and 3 region having 138 significant associations. Exons 4 and 5 of this gene had 32 and 48 significant pairwise

associations respectively (Fig 3.8). Glycophorin B also had a relatively high number of significant associations between the SNPs, with 43, 20, 11 and 13 significant pairwise SNP associations in regions encompassing exons 2, 3, 4 and 5 respectively (Fig. 3.9). Glycophorin C showed the lowest number of pairwise associations, with 17, 4, 12 and 93 significant pairwise associations for regions encompassing exon 2, exon 3, intron 2 and the full length region, respectively (Fig 3.10).

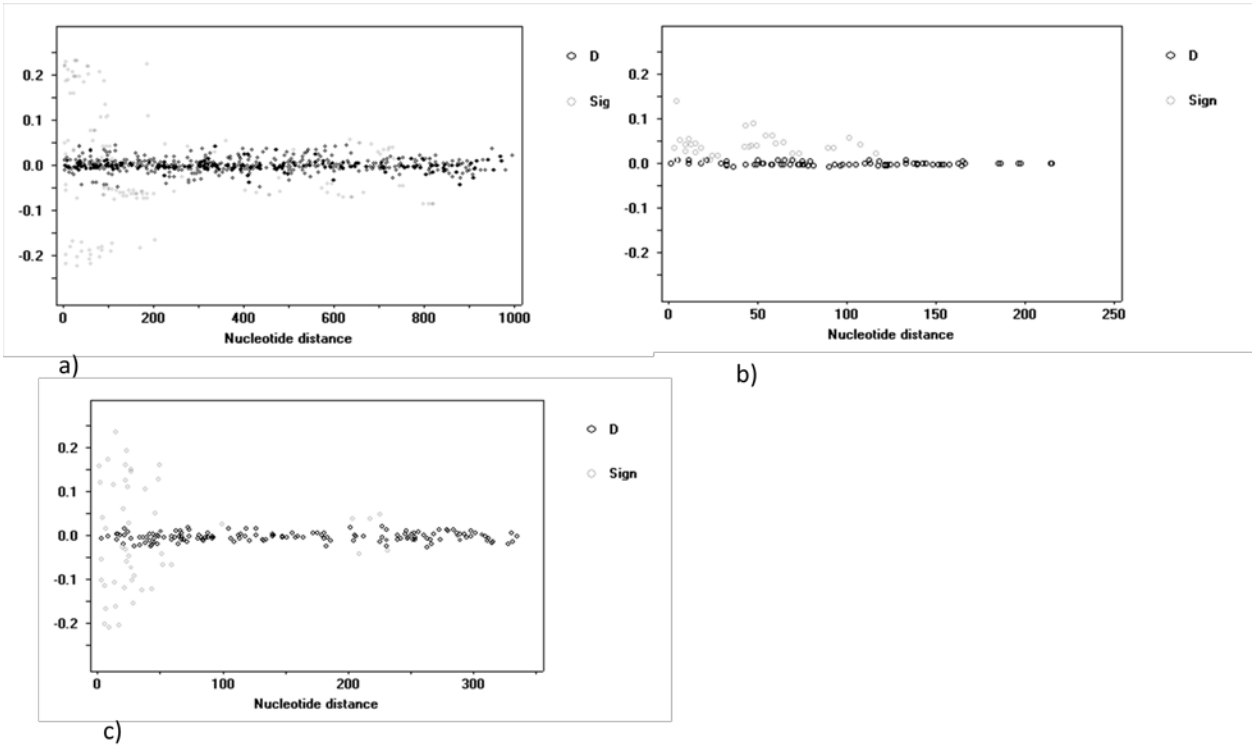


Figure 3.8 LD plots for a) exons 2 and 3, b) exon 4 and c) exon 5 of glycophorin A showing the coefficient of linkage disequilibrium, D, plotted against nucleotide distances. Each circle represents the difference between observed and expected haplotypes frequencies. Sign, represents statistically significant differences as determined by the two-tailed Fisher's test.

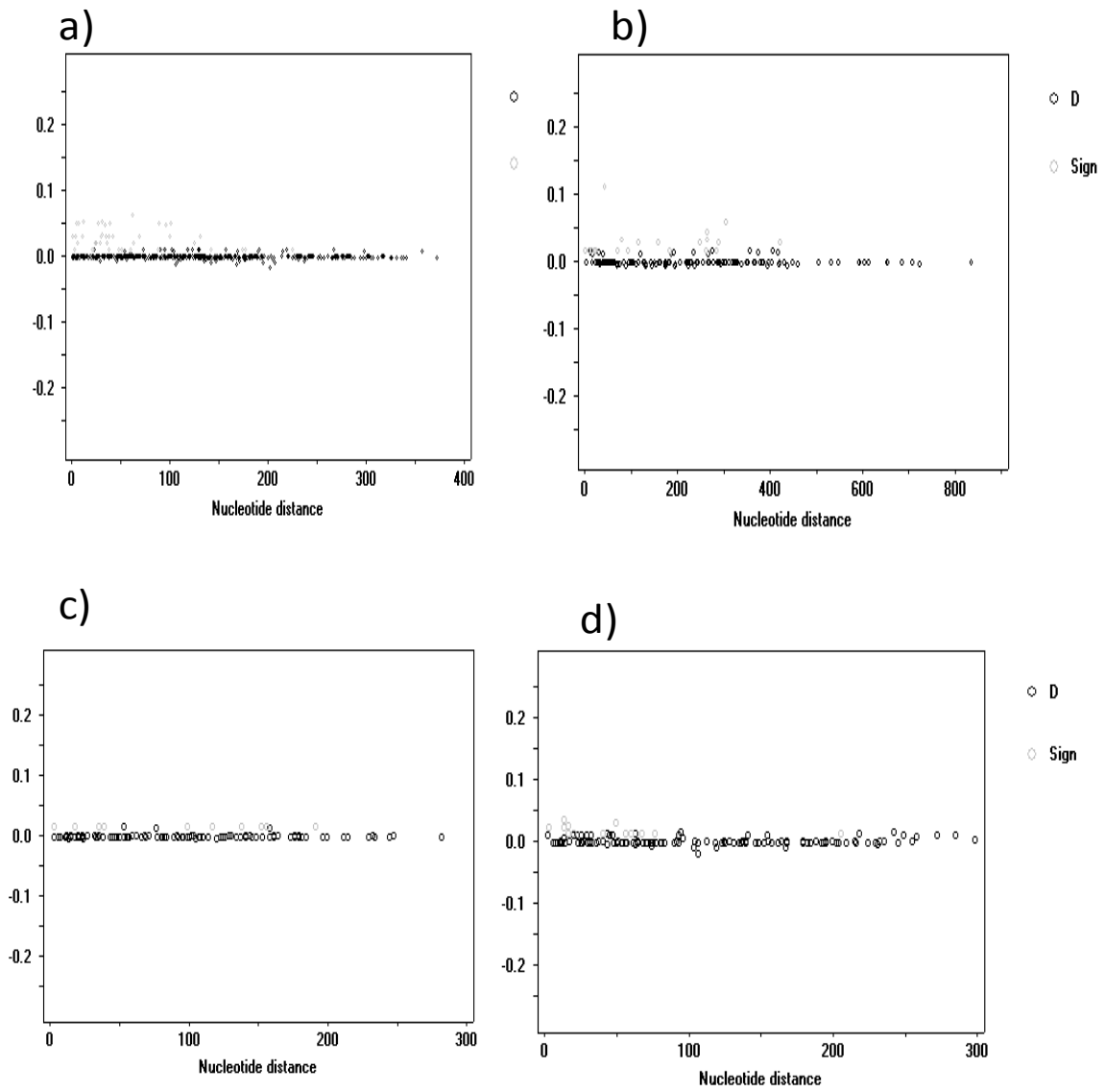


Figure 3.9 LD plots for a) exons 2, b) exon 3, c) exon 4 and d) exon 5of glycophorin B showing the coefficient of linkage disequilibrium D, plotted against nucleotide distances. Each circle represents the difference between observed and expected haplotypes frequencies. Sign, represents statistically significant differences as determined by the two-tailed Fisher's test.

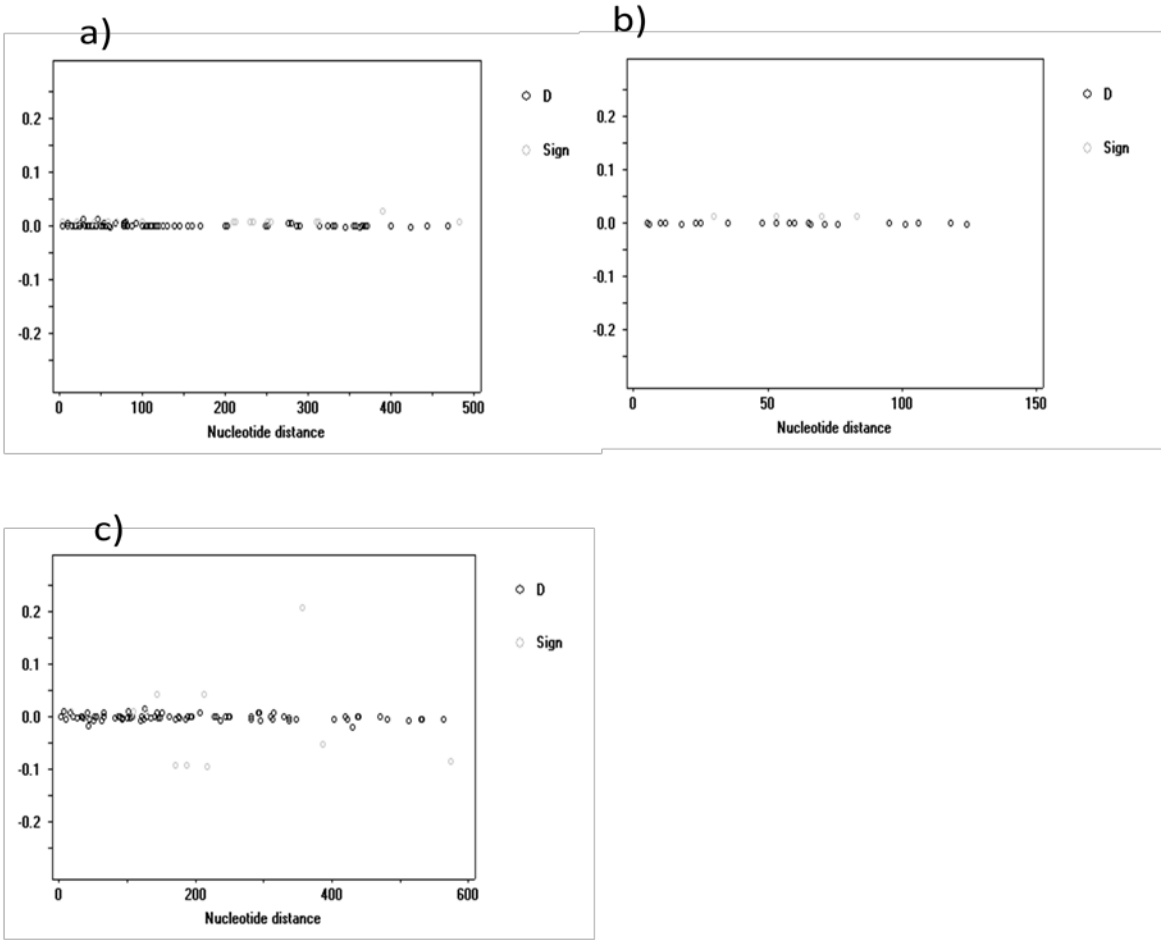


Figure 3.10 LD plots for a) exon 2, b) intron 2 and c) exon 3 of glycophorin C showing the coefficient of linkage disequilibrium D, plotted against nucleotide distances. Each circle represents the difference between observed and expected haplotypes frequencies. Sign, represents statistically significant differences as determined by the two-tailed Fisher's test.

Hardy-Weinberg equilibrium (HWE) was also calculated for all the SNPs that were detected in the three genes (Table 3.5; Table S1). The significance of the HWE values was determined using the Chi-square test with one degree of freedom. SNPs with $p<0.05$ were not consistent with the HWE, and were considered to be evolving. Most of the SNPs did not conform to the HWE ($P<0.05$).

Table 3.5 Hardy-Weinberg equilibrium analysis for exonic regions of glycophorins A, B and C

SNP position	P-Value	SNP position	P-Value	SNP position	P-Value	SNP position	P-Value
GYP A							
25185	P<0.001	25995	P<0.001	26047	0.44	27018	0.59
25193	P<0.001	26005	P<0.001	26058	P<0.001	27024	0.11
25197	P<0.001	26008	P<0.001	26061	0.03	27029	0.95
25974	P<0.001	26009	P<0.001	26065	0.02	27033	0.85
25975	P<0.001	26013	P<0.001	27006	0.75		
25976	0.9	26038	0.01	27009	0.75		
GYP B							
23082	0.03	23095	P<0.001	24901	0.37	26752	0.9
23083	P<0.001	23119	P<0.001	26729	0.9	26764	0.9
23090	0.05	23125	P<0.001	26732	0.9	26785	P<0.001
23094	P<0.001						
GYP C							
39157	0.95						

P-values were calculated using Chi-Square test with one degree of freedom. P<0.05; not consistent with Hardy-Weinberg Equilibrium.

CHAPTER FOUR

DISCUSSION, CONCLUSION AND RECOMMENDATIONS

4.1 DISCUSSION

Glycophorins are important for invasion of erythrocytes by *P. falciparum*, being the receptors for the main parasite ligands, the Erythrocyte Binding-like Antigens (EBAs). Glycophorin molecules are made up of sialic acids and proteins (Orlandi *et al*., 1992). The O- and N-linked sialic acids have been shown to be important for binding to the parasite ligand, but they are not sufficient for optimal binding and a specific amino acid sequence is also required (Sim *et al*., 1994). Because of the need for this specific amino acid sequence, genetic polymorphisms occurring within the parasite's binding regions are expected to affect the specificity and affinity of binding. Such polymorphisms can be detected by various methods including enzymatic approaches such as Restriction Fragment Length Polymorphism (RFLP) and physical methods for discrimination of variants such as differential sequencing with mass spectrometry (Kristensen *et al*., 2001). In this study, direct sequencing of PCR products was used to identify allelic variants in the Kilifi population. The method has a number of advantages including being fast and straight forward. Despite the advantages, direct sequencing of PCR products presents certain challenges, the main one being the difficulty in troubleshooting in cases where sequencing fails as this may be due to different reasons ranging from presence of multiple products to presence of unincorporated primers and nucleotides. To avoid the problem of sequencing unwanted PCR products, only those samples that appeared as clean, single bands on the agarose gels were selected for sequencing. Amplification was relatively easy for glycophorin C, which has unique sequences and has no homologs in the human genome. Amplification of glycophorins A and

B, on the other hand, presented a challenge as the two genes are highly homologous to each other and to another glycophorin gene, glycophorin E (Baum *et al*., 2003; Wang *et al*., 2003; Ko *et al*., 2011). To ensure that the correct genes were amplified and sequenced, gene-specific primers were designed and a BLAST (Basic Local Alignment Search Tool) search against the non-redundant database at NCBI was performed to ensure that the primers picked up the correct genes.

Glycophorin genes are relatively large and range in size from 20 – 40kb. The bulk of the genes are however occupied by non-coding, intronic regions while the coding, exonic regions are relatively small, most of them being less than 100 base pairs long. Glycophorin C, for example, is 40.563kb long and has an intronic region (intron 1) that is 34kb long while the regions encoding the protein's extracellular receptor-binding domains, exons 2 and 3, are 57 and 84 base pairs long, respectively. Due to the relatively large intronic regions separating exons 1 and 2 of the glycophorin genes (intron 1 of glycophorin A is 20kb long while the same region in glycophorin B is 18kb long), it was not possible from the outset to amplify a single, continuous fragment beginning from exon 1. For this reason, and because the first exon acts as a signal peptide (Baum *et al*., 2002) and is cleaved from the final gene product and does not form part of the binding site for *P. falciparum* in any of the genes, it was excluded from the analysis.

For glycophorins A and B, exons 2 – 5 were selected for analysis while exons 2 and 3 were selected in the case of glycophorin C, translating into gene lengths of approximately 4kb for glycophorins A and C and 5.3kb for glycophorin B. These regions were chosen for two main reasons; 1) they code for the extracellular domains of the glycophorins (as well as part of the transmembrane domain in the case of exon 5 in GYP A and GYP B) (Baum *et al*., 2002); 2)

they have been identified as either putative or confirmed binding sites for the P*lasmodium* parasite.

Alignment of the contigs generated from the primer trace files did not generate the expected full length sequences for all of the samples of glycophorin A and B, and most of the samples of glycophorin C. The lack of full length sequences is attributed to insufficient primer coverage of the sequencing regions as well as lack of sufficient overlap as initially predicted. During direct sequencing, it is normally possible to use the PCR primers as sequencing primers and this was done in the case of this study. The PCR primers, however, did not perform well as sequencing primers, and this led to most of the trace files generated from the PCR primers for glycophorins A and B having too much background noise such that the bases could not be confidently called by the KB^{TM} Basecaller software. The high background noise can be attributed to problems that are inherent when PCR primers are used as sequencing primers. These include 1) PCR reactions can be tailored to the primer in terms of temperatures and reagent constituents while this is usually not possible in a sequencing reaction; 2) PCR reactions can use a mismatched priming site while for sequencing reactions which use only one primer, the reaction will always be less efficient if the primer is mismatched (Lyons, 2013). The high background noise seen in the two glycophorin sequences can also be attributed to possible sequencing of non-specific PCR products present in the reaction, a phenomenon which is a disadvantage of direct sequencing and which is possible due to the high homology between the two genes and glycophorin E. To ensure that the sequences used in analyses belonged to glycophorins A and B, random sample sequences generated were blasted against the non-redundant nucleotide database at NCBI and were found to be specific for these two genes, with none of the Blasted sequences

picking up the homologous glycophorin E with significant e-values. Background noise detected in glycophorin C primer trace files was relatively less compared with glycophorins A and B and this is attributable to the fact that the gene is unique, with no homologs in the human genome.

Apart from the PCR primers, several other primers also generated trace files with a lot of noise and low signal intensity, including the F3 primer in glycophorin B, which had background noise peaking after a mononucleotide stretch of T which is caused by Taq polymerase slippage during DNA synthesis and is a recognized limitation of the Sanger sequencing method. This problem can usually be solved by using a poly-mononucleotide primer with a degenerate base at the 3' end, although such a primer was not used in this case. The forward F4 primer used in glycophorin C was also not a good choice as it bound to two different locations within the region of interest, leading to generation of trace files with overlapping peaks in homologous regions. Trace files for this primer were therefore not used in the analysis.

The human genome is diploid and the direct sequencing method used in this study picks up both alleles, meaning that in cases where an individual is heterozygous at a particular position in the glycophorin gene, both alleles would be detected during the sequencing process and would be represented simultaneously on the trace files as double peaks of different colours corresponding to the two bases. Several of these heterozygous sites were detected in the regions analyzed and were called using their corresponding ambiguous codes.

The overall aim of this study was to identify polymorphisms, including single nucleotide polymorphisms (SNPs), which are DNA sequence variations that occur when a single

nucleotide in the genome sequence is altered, as well as insertions and deletions (indels). SNPs are the most common type of genetic variation in humans, making up about 90% of human genetic variation (Crawford and Nickerson, 2005). Multiple alignments of the samples for the individual glycophorin genes identified several polymorphisms in the genes. These polymorphisms were present in both intronic and exonic regions of the genes with most of the variations occurring within the non-coding intronic regions. This is in line with normal expectations as intronic regions do not code for any proteins and therefore are not subjected to strong selective forces to maintain the sequence of the regions. Most of the variations occurring within introns therefore tend to be selectively neutral, meaning that they do not confer any selective fitness or advantage to the organism. Intronic mutations are not always neutral, however, and can affect gene expression (Holland *et al*., 2001). Introns contain splice sites that direct their removal from the gene during gene transcription and mutations that remove or change the sites can alter the gene product and result in a non-functional protein (Holland *et al.*, 2001). Analyses of intronic polymorphisms are therefore important.

Both synonymous and non-synonymous substitutions were present in the exonic regions of the genes examined. Synonymous substitutions are those substitutions in codons which do not lead to a change in the amino acid encoded by the codon, while non-synonymous substitutions lead to a change in the amino acid encoded by the codon. Synonymous substitutions are also known as silent mutations. Non-synonymous mutations have a great implication in the function of proteins as they can lead to a change in the conformation of the protein encoded, making a protein non-functional, or in rare cases, enhancing the functionality of the protein, especially if the amino acid is present in the protein's active site.

Some non-synonymous variations can also lead to a functional amino acid being converted into a stop codon, thus truncating a protein and making it non-functional, a situation which may be deleterious to the organism if the protein encoded is essential to the functioning of the organism.

Non-synonymous substitutions can be either conservative (involve a change to an amino acid with similar physicochemical properties), semi conservative (change from a negatively charged amino acid to a positively charged amino acid and vice versa) or radical (vastly different amino acids). Radical, non-synonymous substitutions have a greater effect on the encoded protein as they are more likely to lead to a change in the conformation of the protein. 3 non-synonymous SNPs were detected in exon 2 of glycophorin A, 2 of which occur at positions determining the MN blood groups. The M blood group has serine and glycine at positions 1 and 5 of the peptide, while the N variant has leucine and glutamic acid at these positions. Most of the haplotypes were the M variant. In the second variant, the polymorphism at position 5 involved a conversion between glycine and arginine, with only three samples having arginine at this position. This may simply be an artifact or it may represent a new mutation at the location in this population. Previous studies have identified the two well known MN variants at positions 1 and 5 in different African populations, including a Nigerian population (Baum *et a*l., 2002; Ko *et al*., 2011). The third polymorphism identified has not been previously characterized and represents a change from glycine to aspartic acid. This polymorphism was relatively rare in this population, being present at a frequency of 3%. The amino acid change represents a conservative substitution of polar amino acid with a charged amino acid. Such substitutions may affect the overall function of the protein as it may lead to a change in the structure of the protein and can be

especially deleterious if it occurs within the binding site of the *Plasmodium* parasite. *Plasmodium* parasites have been shown to bind to the glycophorin molecule through O-linked tetrasaccharides (Orlandi *et al*., 1992), thus SNPs that involve a substitution of threonine or serine with another amino acid could affect parasite binding efficiency. Exon 3 was shown to have 10 non-synonymous substitutions, none of which has been identified in previous studies. Most of the substitutions observed here are conservative. Six non-synonymous mutations, which have also not been previously characterized, were identified in exon 4 (Table 3.2). A single base pair insertion was detected in intron 1 of glycophorin A as well as SNP position with 4 different base pair mutations in intron 3. Insertions and deletions have been detected in introns 2 to 4 of this gene in a Nigerian population (Baum *et al*., 2002). In the Nigerian study, intron 1 was not analyzed.

Three non-synonymous and two synonymous polymorphisms were detected in exon 2 of glycophorin B (Table 3.2). The three non-synonymous polymorphisms have been previously identified in different populations of Cameroon, Nigeria and Kenya (Ko et al., 2011). These polymorphisms are located at positions 1, 4 and 5 of the peptide, and lead to an amino acid change from leucine to tryptophan, threonine to Serine and glutamic acid to glycine at these positions, respectively. The polymorphisms were all in linkage disequilibrium. The non-synonymous polymorphism that determines the Ss blood groups of glycophorin B was also identified in exon 4. The s blood group which is represented by threonine at position 29 was the most common in the Kilifi population, being present at a frequency of 75%. Four other polymorphisms, including a synonymous polymorphism were also detected in this exon and have not been previously identified. The multiple variations present in the two glycophorin genes shows that these genes are rapidly evolving (Wang *et al*., 2003). Glycophorin C had

fewer variations, with only a single non-synonymous variation within exon 2. This polymorphism had been identified in a previous study using a global human sample (Wilder *et al*., 2009). This means that the gene is not as rapidly evolving as the other glycophorin genes analyzed, and this has been previously indicated (Wilder *et al*., 2009).

Statistical analyses show that exons 2 and 3 of glycophorin A (excluding any intronic regions) are under balancing selection, as evidenced by significant positive Tajima's D ($p<0.05$), Fu and Li's D ($P<0.05$) and Fu and Li's F ($P<0.02$) values. Our results are consistent with those of studies previously conducted in endemic populations in Africa (Baum *et al*., 2002; Ko *et al*., 2011). Positive Tajima's D values are indicative of an excess of rare alleles which are selected for and maintained within the population at intermediate frequencies, while positive Fu and Li values are indicative of an absence of singletons among the nucleotides in these exons (Baum *et al*., 2002; Akey *et al*., 2004). When intronic regions were included in the analyses, Tajima's D analysis indicated that the polymorphisms within exons 2 and 3 were not under selection, as the results were not statistically significant ($p>0.10$). This points to a possible dilution of the Tajima's D value when the introns were included in the analysis. However, analyzing the same region using Fu and Li's D and Fu and Li's F tests, showed that the polymorphisms were under selection, giving statistically significant positive results under both Fu and Li's D ($P < 0.02$.) and Fu and Li's F ($P < 0.05$) test statistics. This difference in the significance level outcome may be due to the different parameters employed by the tests (Baum *et al*., 2002; Akey *et al*., 2004).

Glycophorin B showed negative significant Tajima's D values for all the exons (together with bordering intronic regions), and negative Fu and Li's D* and Fu and Li's F* values for exon 2, and exons 2 and 4, respectively, consistent with results from previous studies (Ko *et*

al., 2011). Analyses of the exons on their own gave significant negative values for only exon 4. The negative values indicate that the population may be undergoing expansion or purifying selection. Statistically significant negative values were also obtained for the exons 2, 3, and intron 2 regions of glycophorin C, but analysis of the full length gene did not give statistically significant results for the 12 samples analyzed and neither did analysis of the single polymorphism in exon 2 when the adjacent introns were removed. The variation in results between the full length region and the fragmented regions of the same gene can be attributed to the fact that very few samples were used in the full length analysis (Tajima's D test requires at least 50 sequences for a greater confidence level) as compared to the fragmented regions (Simonsen *et al*., 1995). The use of more sample sequences is thus recommended over the use of longer sequences in these statistical analyses (Simonsen *et al*., 1995).The presence of both statistically significant positive and significant negative values of Tajima's D, Fu and Li's D, and Fu and Li's F may point to selective pressures unique to one population or different demographic histories. Overall, greater significance levels were observed when exons of all three genes were analyzed together with their bordering intronic regions, indicating that the intronic regions may also be subject to selection.

Haplotypes circulating within the Kilifi population were determined (Table 3.3). Due to the high number of SNPs detected here, glycophorin A had the highest number of haplotypes (Table 3.3, Fig 3.5) with the most common haplotype in the population being STGTGHRHRSKDTRVAHSD (Ref. No 13). Glycophorin B had 4 haplotypes (Table 3.3, Fig 3.5), with LTDSSTLIGTS (Ref. No 3) occurring most frequently within the Kilifi population. Glycophorin C had only two haplotypes circulating within the population. Haplotype diversity (Hd) was also calculated for the different genetic regions under analysis

(Table 3.4). Glycophorin A had the greatest haplotype diversity among the three genes, with exons 2, 3 and 5 having the highest diversity among all the regions analyzed. This means that when any two random glycophorin A sequences were compared, it was very likely that the sequences would be different. Exon 3 of glycophorin C had the lowest haplotype diversity. From sequence analysis, this region had the least number of polymorphisms across all samples analyzed and therefore the number of unique alleles/haplotypes would be expected to be quite low. Haplotype diversity within this region would also be quite low as any two samples compared would be likely to have the same sequence due to the low polymorphisms (Rozas, 2009). Overall, the Hd values indicate that most of the haplotypes detected were relatively unique among the sequences analyzed. This, however, may not be representative of the population as the number of samples used in each of the cases was relatively low. Haplotye diversity is greatly affected by linkage disequilibrium, with greater LD leading to lower haplotype diversity for a particular number of SNPs considered. In the case of this study, LD was calculated for all polymorphic sites present and computed for pairwise comparisons, meaning that LD was calculated for only two SNPs at a time. Glycophorin A had the highest number of SNP pairs compared among the three genes. Under the two-tailed Fisher's exact test used to determine significance of the LD values observed, Glycophorin A was shown to have the highest number of significant associations between SNP pairs due to the high number of SNPs detected in this gene and because LD was only considered between two SNPs at a time. Glycophorin C had the least number of significant pairwise associations (Fig.3.6).

Hardy-Weinberg equilibrium calculations showed that some of the SNPs detected in this population are still undergoing evolution, while others have reached equilibrium and remain

constant or are evolving randomly (Table 3.5; Table S1). Several of the SNPs detected within the coding regions of glycophorins A and B were not in Hardy-Weinberg equilibrium, indicating that these SNPs are subject to an evolutionary force (Table 3.5). Since the samples used in this study were all taken from malaria-positive individuals, it may be postulated that the disease is the evolutionary force acting on these genes.

4.2 CONCLUSION

Different polymorphisms have been detected in the human genome that either confer resistance (Maier *et al.,* 2003; Mayer *et al.,* 2006) or predispose individuals to the disease (Tarazona-Santos *et al*., 2011). Several SNPs were detected in glycophorins A, B and C genes. An insertion was also detected in intron 1 of glycophorin A. Several SNPs that have not been previously characterized in these genes were identified. The findings of this study are in agreement with previous studies that have looked at the glycophorin genes in that glycophorin A was shown to have more polymorphisms and to be evolving more rapidly than GYP B and GYP C. The glycophorins are the main receptors of the *P. falciparum* parasite and molecular variations occurring in the gene, especially in the ligand binding domains can affect binding efficiency and hence the severity of disease. Analysis of the genes in the Kenyan population studied showed that some of the SNPs within the exons coding for the extracellular domains were under selection, indicating that malaria, which is endemic in this region, may be acting on the population. Statistical analysis show that the population may be subject to both balancing and purifying selection at these candidate genes.

4.3 RECOMMENDATIONS

The following recommendations are made as part of future work.

1. Association studies should be carried out to determine whether the SNPs detected, and especially those within the ligand binding domains/exons, affect interaction with the parasite and hence influence susceptibility or resistance to malaria. Population genetic studies are affected by different confounding factors such as population structure and other red blood cell polymorphisms, and these will have to be taken into account in any association studies.
2. The sample size used in this study was relatively small and future work can involve the use of a larger sample size to determine if these polymorphisms are common in the population.
3. Similar polymorphisms have been shown to have different effects when different populations were studied, for example, a particular SNP may increase the susceptibility of one population to a disease while the same SNP will have no effect on susceptibility to disease in a different population. It would thus be interesting to compare the Kilifi population with other malaria endemic populations to study the effects of the SNPs detected here on susceptibility to disease.

CHAPTER FIVE

REFERENCES

Adams, J.H., Blair, P.L., Kaneko, O., Peterson, D.S. (2001). An expanding *ebl* family of *Plasmodium falciparum.Trends in Parasitology*, 17: 297-299.

Aikawa, M., Miller, L.H., Johnson, J., Rabbege, J. (1978). Erythrocyte entry by malarial parasites. A moving junction between erythrocyte and parasite. *Journal of Cell Biology,* 77: 72-82.

Akey, J.M., Eberle, M.A., Rieder, M.J., Carlson, C.S., Shriver, M.D., Nickerson, D.A., Kruglyak, L. (2004). Population history and natural selection shape patterns of genetic variation in 132 genes. *Public Library of Science BIOLOGY*, 2 (10): e286. doi:10.1371.journal.pbio.0020286

Baum, J., Chen, L., Healer, J., Lopaticki, S., Boyle, M., Triglia, T., Ehlgen, F., Ralph, S.A., Beeson, J.G., Cowman, A.F. (2009). Reticulocyte-binding protein homologue 5 – An essential adhesin involved in invasion of human erythrocytes by Plasmodium falciparum. *International Journal for Parasitology,* 39: 371 – 380.

Baum, J., Maier, A.G., Good, R.T., Simpson, K.M., Cowman, A.F. (2005).Invasion by *P. falciparum*Merozoites suggests a hierarchy of molecular interactions. *Public Library of Science(PLoS) Pathogen*s, 1:299-309.

Baum, J., Thomas, W.A., Conway, D.J. (2003). Evidence for diversifying selection on Erythrocyte-Binding Antigens of *Plasmoium falciparum* and *P. vivax*. *Genetics,* 163: 1327–1336.

Baum, J., Ward, R.H., Conway, D.J. (2002). Natural selection on the erythrocyte surface. *Molecular Biology and Evolution*, 19:223–229.

Beier, J.C. (1998). Malaria parasite development in mosquitoes. *Annual Review of Entomology*, 43: 519-543.

Bejon P, Williams TN, Liljander A, Noor, A.M., Wambua, J., Ogada, E., Olutu, A., Osier, F.H.A., Hay, S.I., Färnert, A., Marsh, K. (2010). Stable and unstable malaria hotspots in longitudinal cohort studies in Kenya. *PLoS medicine*. 7:e1000304.

Blanchard, D., Dahr, W., Hummel., M., Beyreuther, K., Cartron, J.P. (1987). Glycophorins B and C from human erythrocyte membranes: Purification and sequence analysis. *Journal of Biological Chemistry*.262: 5808 – 5811.

Blumenfeld, O.O., Huang, C. (1995). Molecular genetics of the Glycophorin gene family, the antigens for MNSs blood groups: Multiple gene rearrangements and modulation of splice site usage result in extensive diversification. *Human Mutation*, 6: 199-209.

Bob, N.S., Diop, B.M., Renaud,F., Marrama, L., Durand, P., Tall, A., Ka, B., Ekala, M.T., Bouchier, C., Mercereau-Puijalon, O., Jambou, R. (2010). Parasite polymorphism and severe malaria in Dakar (Senegal): a West African urban area. *Public Library of Science (PLoS)* ONE, 5: e9817. doi:10.1371/journal.pone.0009817.

Breuer, W.V., Kahane, I., Baruch, D., Ginsburg, H., Cabantchik, Z.I. (1983). Role of internal domains of glycophorin in *Plasmodium falciparum* invasion of human erythrocytes. *Infection and Immunity*, 42: 133-140.

Brooks, D.E., Cavanagh, J., Jayroe, D., Janzen, J., Snoek, R., Trust, T.J. (1989). Involvement of the MN blood group antigen in shear-enhanced hemagglutination induced by the *Escherichia coli* F41 adhesin. *Infection and Immunity*, 57: 377 – 383.

Chasis, J.A., Mohandas, N. (1992). Red blood cell polymorphisms. *Blood*, 80:1869 – 1879.

Child, M.A., Epp, C., Bujard, H., Blackman, M.J. (2010). Regulated maturation of malaria merozoite surface protein-1 is essential for parasite growth. *Molecular Microbiology*, 78: 187-202.

Colin, Y., Kim, C.L.V., Tsapis, A., Clerget, M., d'Auriol, L., London, J., Galibert, F., Cartron, J. (1989).Human Erythrocyte Glycophorin C. Gene structure and rearrangement in genetic variants.*Journal of Biological Chemistry*, 264: 3773-3780.

Cowman, A.F., Baldi, D.L., Healer, J., Mills, K.E., O'Donnell, R.A.,, Reed, M.B., Triglia, T., Wickham, M.E., Crabb, B.S. (2000). Functional analysis of proteins involved in *Plasmodium falciparum* merozoite invasion of red blood cells. *Federation of European Biochemical Societies Letters*, 476: 84-88.

Cowman, A.F., Berry, D., Baum, J. (2012).The cellular and molecular basis for malaria parasite invasion of the human red blood cell.*Journal of Cellular Biology*, 17: 961 – 971.

Cowman, A.F., Crabb, B.S. (2006).Invasion of red blood cells by malaria parasites.*Cell*, 124: 755-766.

Cox, F.E.G. (2010). History of the discovery of the malaria parasites and their vectors. *Parasites and Vectors*, 3: 5.

Crawford, D.C., Nickerson, D.A. (2005).Definition and clinical importance of haplotypes. Annual Review of Medicine, 56: 303 – 320.

Deans, A., Nery, S., Conway, D.J., Kai, O., Marsh, K., Rowe, A. (2007).Invasion pathways and malaria severity in Kenyan *Plasmodium falciparum* clinical isolates. *Infection and Immunity,* 75: 3014-3020.

Dubremetz, J.F., Garcia, N.-R., Conseil, V., Fourmaux, M.N. (1998).Apical organelles and host-cell invasion by Apicomplexa.*International Journal for Parasitology*,28: 1007-1013.

Duret, L. (2008) Neutral theory: The null hypothesis of molecular evolution. *Nature Education* 1(1)

Faria, M.A., Martins, M.L., Schmidt, L.C., Malta, M.C.F. (2012). Molecular analysis of the *GYPB* gene to infer S, s, and U phenotypes in an admixed population of Minas Gerais, Brazil. *Revista Brasileira De Hematologia E Hemoterapia,* 34:212-216.

Gaur, D., Chitnis, C.E. (2011).Molecular interactions and signaling mechanisms during erythrocyte invasion by malaria parasites. *Current Opinion in Microbiology*,14: 1-7.

Gaur, D., Mayer, D.C.G., Miller, L.H. (2004).Parasite ligand–host receptor interactions during invasion of erythrocytes by Plasmodium merozoites. *International Journal for Parasitology,* 34: 1413 – 1429.

Gaur, D., Storry, J.R., Reid, M.E., Barnwell, J.W., Miller, L.H. (2003).*Plasmodium falciparum* is able to invade erythrocytes through a trypsin-resistant pathway independent of glycophorin B. *Infection and Immunity,* 71: 6742 – 6746.

Gilberger, T.-W., Thompson, J.K., Triglia, T., Good, R.T., Duraisingh, M.T., Cowman, A.F. (2003).A Novel Erythrocyte Binding Antigen-175 Paralogue from *Plasmodium falciparum* Defines a New Trypsin-resistant Receptor on Human Erythrocytes. J*ournal of Biological Chemistry*, 278: 14480-14486.

Goel, V.K., Li, X., Chen, H., Liu, S., Chishti, A.H., Oh, S.S. (2003)Band 3 is a host receptor binding merozoite surface protein 1 during the Plasmodium falciparum invasion of erythrocytes. *Proceedings of the National Academy of Science. USA*, 100: 5164-5169.

Grassi, B. (1900). Studi di uno Zoologo Sulla Malaria.

Hadley, T.J., Klotz, F.W., Miller, L.H. (1986). Invasion of erythrocytes by malaria parasites: A cellular and molecular overview. *Annual Review of Microbiology*, 40:451-477.

Hadley, T.J., Klotz, F.W., Pasvol, G., Haynes, D.J., McGinniss, M.H., Okubo, Y., Miller, L.H. (1987). Falciparum malaria parasites invade erythrocytes that lack glycophorin A and B (M^KM^K). *The journal of Clinical Investigation*, 80: 1190-1193.

Hayton, K., Gaur, D.Liu, A., Takahashi, J., Henschen, B., Singh, S., Lambert, L., Furuya,T., Boutternot, R., Doll, M., Nawaz, F., Mu, J., Jiang, L., Miller, L.H., Wellems, T.E. (2008). Erythrocyte binding protein PfRH5 polymorphisms determine species-specific pathways of *Plasmodium falciparum* invasion. *Cell Host and Microbe*, 4: 40 – 51.

Holland, J.B., Helland, S.J., Sharapova,N., Rhyne, D.C. (2001). Polymorphism of PCR-based markers targeting introns, promoter regions and SSRs in maize and introns repeat sequences in oat. *Genome*, 44: 1065 – 1076.

Iyer, J., Grüner, A. C., Rénia, L., Snounou, G., Preiser, P. R. (2007). Invasion of host cells by malaria parasites: a tale of two protein families. *Molecular Microbiology*, 65: 231–249.

Jennings, C.V., Ahouidi, A.D., Zilversmit, M., Bei, A.K., Rayner, J., Sarr, O., Ndir, O., Wirth, D.F., Mboup, S., Duraisingh, M.T. (2007). Molecular analysis of erythrocyte invasion in *Plasmodium falciparum* isolates from Senegal. *Infection and Immunity*, 75: 3531–3538.

Jiang, L., Gaur, D., Mu, J., Zhou, H., Long, C.A., Miller, L.H. (2011). Evidence for erythrocyte-binding antigen 175 as a component of a ligand-blocking blood-stage malaria vaccine. *Proceedings of the National Academy of Science. USA*, 108: 7553-7558.

Jones, M.L., Kitson, E.L., Rayner, J.C. (2006). Plasmodium falciparum erythrocyte invasion: a conserved myosin associated complex. *Molecular and Biochemical Parasitology*, 147: 74-84.

Kadekoppala, M., O'Donnell, R.A., Grainger, M., Crabb, B.S., Holder, A.A. (2008). Deletion of the *Plasmodium falciparum* Merozoite Surface Protein 7 gene impairs parasite invasion of erythrocytes. *Eukaryotic Cell*, 7: 2123-2132.

Khera, R., Das, N. (2009). Complement Receptor 1: Disease associations and therapeutic implications. *Molecular Immunology*, 46: 761-772.

Ko, W., Kaercher, K.A., Giombini, E., Marcatili, P., Froment, A., Ibrahim, M., Lema, G., Nyambo, T.B., Omar, S.A., Wambebe, C., Ranciaro, A., Hirbo, J.B., Tishkoff, S.A. (2011). Effects of natural selection and gene conversion on the evolution of human glycophorins coding for MNS blood polymorphisms in malaria-endemic African populations. *The American Journal of Human Genetics*, 88: 741-754.

Kristensen, V.N., Kelefiotis, D., Kristensen, T., Borresen-Dale, A. (2001).High-Throughput methods for detection of genetic variation.*Biochemical Techniques*, 30: 318 – 332.

Kudo, S., Fukuda, M. (1989). Structural organization of glycophorin A and B genes: glycophorin B gene evolved by homologous recombination at Alu repeat sequences. *Proceedings of the National Academy of Science.USA*, 86: 4619 – 4623.

Kwiatkowski, D.P. (2005). How malaria has affected the human genome and what human genetics can teach us about malaria. *American Journal of Human Genetics*, 77: 171 – 192.

Lewis, M. (2001). Agarose gel electrophoresis (basic method). Department of Pathology, University of Liverpool. Accessed on 25/11/2013. www.methodbook.net/dna/agarogel.html.

Lingelbach, K. and Joiner, K.A. (1998). The parasitophorous vacuole membrane surrounding Plasmodium and *Toxoplasma*: an unusual compartment in infected cells. *Journal of Cell Science*, 111: 1467-1475.

Lobo, C., Rodriguez, M., Reid, M., Lustigman, S. (2003).Glycophorin C is the receptor for the *Plasmodium falciparum* erythrocyte binding ligand PfEBA-2 (baebl). *Blood*, 101: 4627-4631.

López, C., Saravia, C., Gomez, A., Hoebeke, J., Patarroyo, M.A. (2010). Mechanisms of genetically-based resistance to malaria. *Gene*, 467: 1-12.

Lyons, R.H. (2013). University of Michigan DNA sequencing core.Last Accessed on 23rd August, 2013. www.seqcore.brcf.med.umich.edu

Maier, A.G., Duraisingh, M.T., Reeder, J.C., Patel, S.S., Kazura, J.W., Zimmerman, P.A., Cowman, A.F. (2003). *Plasmodium falciparum* erythrocyte invasion through glycophorin C and selection for Gerbich negativity in human populations. *Nature Medicine*, 9: 87-92.

Makalowski, W. (2000). Genomic scrap yard: how genomes utilize all that junk. *Gene*, 259: 61 – 67.

Mayer, D.C., Jiang, L., Achur, R.N., Kakizaki, I., Gowda, D.C., Miller, L.H. (2006). The glycophorin C N-linked glycan is a critical component of the ligand for the *Plasmodium*

falciparum erythrocyte receptor BAEBL. *Proceedings of the National Academy of Science, USA*, 103:2358–2362.

Mayer, D.C., Mu, J., Kaneko, O., Duan, J., Su, X., Miller, L.H. (2004). Polymorphism in the *Plasmodium falciparum* erythrocyte-binding ligand JESEBL/EBA-181 alters its receptor specificity. *Proceedings of the National Academy of Science*, USA , 101: 2518–2523.

Mayer, D.C.G., Cofie, J., Jiang, L., Hartl, D.L., Tracy, E., Kabat, J., Mendoza, L.H., Miller, L.H. (2009). Glycophorin B is the erythrocyte receptor of *Plasmodium falciparum* erythrocyte-binding ligand, EBL-1. *Proceedings of the National Academy of Science. USA,* 106: 5248-5352.

Mayer, D.C.G., Mu, J., Kaneko, O., Duan, J., Su, X., Miller, L.H. (2004). Polymorphism in the Plasmodium falciparum erythrocyte-binding ligand JESEBL/EBA-181 alters it receptor specificity. . *Proceedings of the National Academy of Science, USA,* 101*: 2518 – 2523.*

Mayer, D.C.G., Mu, J.B., Feng, X., Su, X, Miller L.H. (2002). Polymorphism in a Plasmodium falciparum erythrocyte-binding ligand changes its receptor specificity. *Journal of Experimental Medicine, 196*: 1523 – 1528.

Mayor, A., Bir, N., Sawhney, R., Singh, S., Pattnaik, P., Singh, S.K., Sharma, A., Chitnis, C.E. (2005).Receptor-binding residues lie in central regions of Duffy-binding-like domains involved in red cell invasion and cytoadherence by malaria parasites. *Blood*, 105: 2557-2563.

Migot-Nabias, F., Pelleau, S., Watier, L., Guitard, J., Toly, C., De Araujo, C., Ngom, M.I., Chevillard, C., Gaye, O., Garcia, A. (2006). Red blood cell polymorphisms in relation to

Plasmodium falciparum asymptomatic parasite densities and morbidity in Senegal. Microbes and Infection, 8: 2352 – 2358.

Miller, L.H. (1977). Hypothesis on the mechanism of erythrocyte invasion by malaria merozoites. *Bulletin of the World Health Organization*, 55: 157-162.

Miller, L.H., Baruch, D.I., Marsh, K., Doumbo O.K. (2002). The pathogenic basis of Malaria. *Nature*, 415: 673-679.

Miller, L.H., Mason, S.J., Clyde, D.F., McGinniss, M.H. (1976). The resistance factor to *Plasmodium Vivax* in blacks. The Duffy-blood group genotype, FyFy. *The New England Journal of Medicine*, 295: 302 – 304.

Moulds, J.N., Zimmerman, P.A., Doumbo, O.K., Kassambara, L., Sagara, I., Diallo, D.A., Atkinson, J.P., Krych-Goldberg, M., Hauhart, R.E., Hourcade, D.E., McNamara, D.T., Birmingham, D.J., Rowe, J.A., Moulds, J.J., Miller, L.H. (2001). *Blood*, 97: 2879- 2885.

Naka, I., Ohashi, J., Patarapotikul, J., Hananantachai, H., Wilairatana, P., Looareesuwan, S., Tokunaga, K. (2007).The genotypes of GYPA and GYPB carrying the MNSs antigensare not associated with cerebral malaria. *Journal of Human Genetics*, 52:476–479.

Nei, M., Tajima, F. (1981).DNA polymorphism detectable by restriction endonucleases. *Genetics*, 97: 145-163.

O'Donnell, R.A., Saul, A., Cowman, A.F., Crabb, B.S. (2000). Functional conservation of the malaria vaccine antigen MSP-1_{19} across distantly related *Plasmodium* species. *Nature Medicine*, 6: 91-95.

Onda, M., Kudo, S., Rearden, A., Mattei, M., Fukuda, M. (1993).Identification of a precursor genomic segment that provided a sequence unique to glycophorin B and E genes. *Proceedings of the National Academy of Science. USA*, 90: 7220-7224.

Orlandi, P.A., Klotz, F.W., Haynes, J.D. (1992). A Malaria Invasion Receptor, the 175-Kilodalton Erythrocyte Binding Antigen of *Plasmodium falciparum* Recognizes the Terminal Neu5Ac(a2-3)Gal- Sequences of Glycophorin A. *Journal of Cell Biology*, 116: 901-909.

Pasvol, G. (2003). How many pathways for invasion of the red blood cell by the malaria parasite? *Trends in Parasitology*, 19: 430-432.

Paul, R.W., Lee, P.W. (1987). Glycophorin is the reovirus receptor on human erythrocytes. *Virology*, 159:94–101

Rayner, J.C., Vagas, E.-S., Huber, C.S., Galinski, M.R., Barnwell, J.W. (2001).A *Plasmodium falciparum* Homologue of *Plasmodium vivax* Reticulocyte Binding Protein (PvRBP1) Defines aTrypsin-resistant Erythrocyte Invasion Pathway. *Journal of Experimental Medicine*, 194: 1571-1581.

Rearden, A., Magnet, A., Kudo, S., Fukuda, M. (1993).Glycophorin B and Glycophorin E Genes Arose from the Glycophorin A Ancestral Gene via Two Duplications during Primate Evolution. *Journal of Biological Chemistry*, 268: 2260-2267.

Reed, M.B., Caruana, S.R., Batchelor, A.H., Thompson, J.K., Crabb, B.S., Cowman, A.F. (2000). Targeted disruption of an erythrocyte binding antigen in Plasmodium falciparum is associated with a switch toward a sialic acid-independent pathway of invasion. *Proceedings of the National Academy of Science. USA*, 97: 7509-7514.

Reid, M.E., Spring, F.A. (1994). Molecular basis of glycophorin C variants and their associated blood group antigens. *Transfusion Medicine*, 4: 139 – 146.

Reid, M.E., Takakuwa, Y., Conboy, J., Tchernia, G., Mohandas, N. (1990).Glycophorin C content of human erythrocyte membrane is regulated by protein 4.1. *Blood*, 75: 2229-2234.

Rozas, J. (2009). DNA Sequence Polymorphism Analysis using DnaSP. Pp. 337-350. *In* Posada, D. (ed.) *Bioinformatics for DNA Sequence Analysis; Methods in Molecular Biology Series* Vol. 537. Humana Press, NJ, USA.

Sabeti, P. (2008). Natural selection: Uncovering mechanisms of evolutionary adaptation to infectious disease. *Nature Education*, 1(1).

Sanders, P.R., Gilson, P.R., Cantin, G.T., Greenbaum, D.C., Nebl, T., Carucci, D.J., McConville, M.J., Schofield, L.,,Hodder, A.N., Yates III, J.R., Crabb, B.S. (2005). Distinct protein classes including novel Merozoite Surface Antigens in Raft-like membranes of *Plasmodium falciparum*. *Journal of Biological Chemistry*, 280: 40169–40176.

Scott, J.A., Bauni, E., Moisi, J.C., Ojal, J., Gatakaa, H., Nyundo, C., Molyneux, C.S., Kombe, F., Tsofa, B., Marsh, K., Peshu, N., Williams, T.N. (2012). Profile: The Kilifi Health and Demographic Surveillance System (KHDSS). International journal of epidemiology 41:650-7.

Sim, K.L., Chitnis, C.E., Wasniowska, K., Hadley, T.J., Miller, L.H. (1994). Receptor and Ligand domains for invasion of erythrocytes by Plasmodium falciparum. *Science*, 264: 1941 – 1944.

Simonsen, K.L., Churchill, G.A., Aquadro, C.F. (1995). Properties of statistical tests of neutrality for DNA polymorphism data. *Genetics*, 141: 413-429.

Steck, T.L. (1974). The organization of proteins in the human red blood cell membrane. *Journal of Cell Biology*, 62: 1-19.

Storry, J.R., Reid, M.E., Fetics, S., Huang, C.H. (2003). Mutations in GYPB exon 5 drive the S−s−U+(var) phenotype in persons of African descent: implications for transfusion. *Transfusion*, 43 :1738–1747.

Tarazona-Santos, E., Castilho, L., Amaral, D.R.T., Costa, D.C., Furlani, N.G., Zuccherato, L.W., Machado, M., Reid, M.E., Rossit, A.R., Santos, S.E.B., Machado, R.L., Lustigman, S. (2011). Population Genetics of GYPB and Association Study between GYPB*S/s Polymorphism and Susceptibility to *P. falciparum* Infection in the Brazilian Amazon.*Public Library of Science ONE*, 6: 1 – 10.

Taylor, H.M., Grainger, M., Holder, A.A. (2002). Variation in the expression of a *Plasmodium falciparum* protein family implicated in erythrocyte invasion. Infection and Immunity, 70: 5779 – 5789.

Taylor, H.M., Triglia, T., Thompson, J., Sajid, M., Fowler, R., Wickham, M.E., Cowman, A.F., Holder, A.A. (2001).*Plasmodium falciparum* homologue of the genes for*Plasmodiumvivax*and *Plasmodium yoelii*adhesive proteins, which is transcribed but not translated. *Infection and Immunity*, 69: 3635–3645.

Tham, W., Schmidt C.Q., Hauhart, R.E., Guariento, M., Tetteh-Quarcoo, P.B., Lopaticki, S., Atkinson,J.P., Barlow, P.N., Cowman, A.F. (2011). *Plasmodium falciparum* uses a key functional site in complement receptor type-1 for invasion of human erythrocytes. *Blood*, 118: 1923- 1933.

Tham, W., Wilson, D.W., Reiling, L., Chen, L., Beeson, J.G., Cowman, A.F. (2009). Antibodies to Reticulocyte Binding Protein-Like homologue 4 inhibit invasion of *Plasmodium falciparum* into human erythrocytes. *Infection and Immunity*, 77: 2427–2435.

Thathy, V., Moulds, J.M., Guyah, B., Otieno, W., Stoute, J.A. (2005). Complement receptor 1 polymorphisms associated with resistance to severe malaria in Kenya. *Malaria Journal*, 4: 54

Tolia, N.H., Enemark, E.J., Sim, B.K.L., Joshua-Tor, L. (2005).Structural Basis for the EBA-175 Erythrocyte Invasion Pathway of the Malaria Parasite *Plasmodium falciparum.* *Cell*, 122: 183-193.

Triglia, T., Duraisingh, M.T., Good, R.T., Cowman, A.F. (2005). Reticulocyte-binding protein homologue 1 is required for sialic acid-dependent invasion into human erythrocytes by *Plasmodium falciparum*. *Molecular Microbiology*, 55: 162-174.

Triglia, T., Thompson, J.K., Cowman, A.F. (2001).An EBA175 homologue which is transcribed but not translated in erythrocytic stages of *Plasmodium falciparum*. *Molecular and Biochemical Parasitology*, 116: 55–63.

Vanderberg, J.P., Gupta, S.K., Schulman, S., Oppenheim, J.D., Furthmayr, H. (1985). Role of the Carbohydrate Domains of Glycophorins as ErythrocyteReceptors for Invasion by *Plasmodium falciparum*Merozoites. *Infection and Immunity*, 47: 201-210.

Vignal, A., Rahuel, C., London, J., Zahar, B.C., Schaff, S., Hattab, C., Okubo, Y., Cartron, J. (1990).A novel gene member of the human glycophorin A and B gene family.Molecular cloning and expression. *European Journal of Biochemistry*, 191: 619-625.

Vignal, A., Rahuel., C., EL Maliki, B., London, J., LeVanKim, C., Blanchard, D., Andre, C., d'Auriol, L., Galibert, F., Blajchman, M.A., Cartron, J. (1989). Molecular analysis of glycophorin A and B gene structure and expression in homozygous Miltenberger class V (Mi.V) human erythrocytes. *European Journal of Biochemistry*, 184: 337-344.

Waler, P.S., Reid, M.E., (2010). The Gerbich blood group system: a review. Immunohematology, 26: 60 – 65.

Wang, D.N. (1994). Band 3 protein: Structure, flexibility and function. *Federation of European Biochemical Societies Letters*, 346: 26-31.

Wang, H., Tang, H., Shen, C..J., Wu, C. (2003). Rapidly Evolving Genes in Human. I. The Glycophorins and Their Possible Role in Evading Malaria Parasites. *Molecular Biology and Evolution*, 20: 1795-1804.

Weatherall, D.J., Miller, L.H., Baruch, D.I., Marsh, K., Doumbo, O.K., Casals-pascual, C., Roberts, D.J. (2002).Malaria and the red cell. *Hematology*, 2002: 35-75.

Wickham, M.E., Culvenor, J.G., Cowman, A.F. (2003). Selective Inhibition of a Two-step Egress of Malaria Parasites from the Host Erythrocyte. *Journal of Biological Chemistry*,278: 37658-37663.

Wilder, J.A., Hewett, E.K., Gansner, M.E. (2009). Molecular evolution of GYPC: Evidence for recent structural innovation and positive selection in humans. *Molecular Biology and Evolution*, 26: 2679–2687

Williams, T.N., Mwangi, T.W., Wambua, S., Alexander, N.D., Kortok, M., Snow, R.W., Marsh, K. (2005).Sickle cell trait and the risk of *Plasmodium falciparum* malaria and other childhood diseases. *Journal of Infectious Diseases*, 192: 178 – 186.

Winardi,R., Reid, M., Conboy, J., Mohandas, N. (1993). Molecular analysis of glycophorin C deficiency in human erythrocytes. *Blood*, 81: 2799-2803.

Wipasa, J., Elliott, S., Xu, H., Good, M.F. (2002). Immunity to asexual blood stage malaria and vaccine approaches. *Immunology and Cell Biology*, 80: 401-414.

Wiser, M.F. (2009). Cellular and molecular biology of *Plasmodium*. Tulane University. Accessed on 23rd August, 2011. http://www.tulane.edu/~wiser/malaria/cmb.html.

World Health Organization.(2008). World malaria report. 190.

APPENDIX

Table S1: Hardy-Weinberg Equilibrium analysis for all SNPs detected in sequenced exonic and bordering intronic regions of glycophorins A, B and C

SNP position	P-Value	SNP position	P-Value	SNP position	P-Value	SNP position	P-Value
GYP A							
25097	P<0.001	25694	0.75	26047	0.44	28658	0.29
25098	P<0.001	25729	P<0.001	26058	P<0.001	28673	0.38
25114	0.9	25732	P<0.001	26061	0.03	28674	0.03
25165	P<0.001	25870	P<0.001	26065	0.02	28679	P<0.001
25185	P<0.001	25894	0.16	26068	0.02	28687	P<0.001
25193	P<0.001	25895	P<0.001	26092	0.57	28696	P<0.001
25197	P<0.001	25900	P<0.001	26868	0.95	28699	0.31
25296	0.9	25907	P<0.001	26886	0.9	28701	P<0.001
25303	P<0.001	25915	P<0.001	26897	0.95	28722	0.23
25304	P<0.001	25939	P<0.001	26918	0.31	28725	0.14
25322	0.14	25940	P<0.001	26929	0.55	28740	0.85
25411	0.04	25942	P<0.001	26961	0.45	28746	P<0.001
25417	P<0.001	25944	P<0.001	26965	0.02	28805	0.85
25418	P<0.001	25974	P<0.001	27006	0.75	28813	0.9
25429	P<0.001	25975	P<0.001	27009	0.75	28904	0.55
25505	P<0.001	25976	0.9	27018	0.59	28926	P<0.001
25508	P<0.001	25995	P<0.001	27024	0.11	28952	0.9
25579	P<0.001	26005	P<0.001	27029	0.95	28985	P<0.001
25597	P<0.001	26008	P<0.001	27033	0.85	28988	P<0.001

25612	P<0.001	26009	P<0.001	27082	0.9		
25672	0.33	26013	P<0.001	27083	0.9		
25673	0.69	26038	0.01	28654	0.64		
GYP B							
22888	0.2	23173	0.9	24889	0.9	28316	0.78
22919	0.6	23176	0.9	24901	0.37	28319	0.9
22948	0.7	23185	0.9	25017	0.9	28326	P<0.001
22994	0.01	23186	0.9	26729	0.9	28328	0.9
22995	0.9	23260	0.9	26732	0.9	28348	0.9
22999	0.9	24485	0.9	26752	0.9	28373	0.72
23056	0.01	24596	P<0.001	26764	0.9	28396	0.83
23082	0.03	24612	P<0.001	26785	P<0.001	28409	0.72
23083	P<0.001	24635	P<0.001	26809	0.9	28419	0.9
23090	0.05	24665	0.9	26905	0.9	28422	P<0.001
23094	P<0.001	24667	0.9	26943	0.9	28423	0.72
23095	P<0.001	24727	0.9	28180	0.9		
23119	P<0.001	24728	0.9	28255	0.9		
23125	P<0.001	24859	0.2	28303	0.83		
GYP C							
39047	0.9	40069	0.9	40849	P<0.001	41770	P<0.001
39080	0.01	40119	0.48	40954	0.9	41872	P<0.001
39157	0.95	40162	0.74	40983	P<0.001	42057	P<0.001
39470	P<0.001	40296	0.95	41022	0.01	42097	0.9
39662	0.9	40306	P<0.001	41074	0.95	42491	0.95
39665	0.9	40434	P<0.001	41267	0.95	42848	P<0.001

39732	0.01	40466	P<0.001	41436	P<0.001	42955	0.9
39825	0.95	40764	0.02	41583	0.9		
39949	P<0.001	40781	0.95	41585	0.9		