MOLECULAR CHARACTERIZATION OF ERYTHROCYTE RECEPTOR GENES, CR1, DARC AND BAND 3 REQUIRED FOR MALARIA PARASITE INVASION

OGOLA CHRISTABEL AWUOR

Reg No. I56/69328/2011

Thesis submitted to the Centre for Biotechnology and Bioinformatics in partial fulfillment for the award of Master of Science degree in Biotechnology, University of Nairobi.

Nov 2014

Declaration

I, Ogola Christabel Awuor, hereby declare that this thesis is my original work and has not been presented for a degree in any other university.

Signature..... Date……..

Ogola Christabel Awuor

Candidate.

Recommendation

We confirm that this thesis has been submitted with our approval as university supervisors:

Signature…… Date……..

Dr. Isabella Oyier

Centre for Biotechnology and Bioinformatics,

University of Nairobi.

Signature…… Date……

Prof James Ochanda,

Center for Biotechnology and Bioinformatics,

University of Nairobi,

Acknowledgements

I would like to express my sincere gratitude to my principal supervisor Dr. Isabella Oyier for the constant guidance and encouragement all through the course of this thesis. I thank her for the role she has played in mentoring me in the field of science. Her meticulous editing greatly contributed to the production of this thesis. I am indebted to Prof. James Ochanda who was my second supervisor for his selfless support and encouragement. The great effort you put in my training right from class work up to the time I was carrying out my research made this thesis possible. My sincere appreciation goes to Irene Omedo. The work she did previously and her advice was helpful during my research work. I would like to thank Anne Owiti and Edwin Rono who were kind enough to offer their advice and help as I worked in the molecular biology laboratory at CEBIB.

I acknowledge the University of Nairobi through the Centre of Biotechnology and Bioinformatics for granting me the scholarship to pursue my Master of Science degree. 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 given to Dr. Isabella Oyier. In July 2013, I went to KWTRP for several weeks to sequence the genes that I had amplified. My time there was very fruitful. John Okombo greatly assisted me in getting the sequence data. I wish to thank my dear friend Phelgona Wasonga for the support she gave me during my stay in Kilifi.

To my parents Alex and Debby, and my husband Chris, I am forever grateful for the endless support.

Table of Contents

List of Figures

List of Tables

Abbreviations

DARC	Duffy Antigen Receptor for Chemokines
CR1	Complement Receptor 1
AE1	Anion-Exchange protein 1
WHO	World Health Organization
RBCs	Red Blood Cells
MSP	Merozoite Surface Proteins
GPI	Glycosylphosphatidylinositol
EBL	Erythrocyte Binding-like Ligands
AMA	Apical Membrane Antigen
PfRh	*Plasmodium falciparum* Reticulocyte binding-like Protein Homologue
HbS	Sickle cell trait allele
SNP	Single Nucleotide Polymorphism
NCBI	National Centre for Biotechnology Information
PCR	Polymerase Chain Reaction
dNTPs	deoxynucleotide triphosphates
ddNTPs	dideoxynucleotides
TBE	Tris Borate EDTA
EDTA	Ethylenediaminetetraacetic acid
ExoSAP	Exonuclease Shrimp Alkaline Phosphatase
BLAST	Basic Local Alignment Search Tool
DBP	Duffy Binding Protein

Abstract

Malaria is still among the most severe infectious diseases at the dawn of the twenty-first century and remains a major global health problem. Five species of the *Plasmodium* parasite cause malaria. The parasite life cycle takes place in two hosts, the mosquito and the mammalian host; however, asexual parasite infection within the blood stream is responsible for the symptoms of the disease. This research focused on understanding the erythrocyte polymorphisms of some of the genes involved in invasion. Mutations that lessen the competence of the merozoite in invading erythrocytes would confer a selective advantage to the host and might be expected to increase in frequency over time through natural selection in malaria endemic regions. This study identified the polymorphisms in DARC, CR1 and Band 3 genes in a malaria endemic population. 93 samples from patients with severe malaria from Kilifi District Hospital were PCR amplified (at the 3 genes) and capillary sequenced and SNPs identified.

Multiple sequence alignments using MEGA genetics software revealed a number of polymorphisms (SNPs) in the three receptor genes. Most of the polymorphisms occurred in the non-coding regions (promoters and introns) and a few were in the coding regions. The few SNPs in the coding region may be due to the need to prevent change in the protein structure and preserve function since these receptors have other biological functions other than acting as receptors for the malaria parasite. DARC sequence analysis showed that all the individuals sampled were duffy negative. Tajima's D statistic, Fu and Li's D and Fu and Li's F test statistics showed that the SNPs in all the three genes were not under selection and therefore the mutations occurring in these genes were evolving randomly and are potentially driven by genetic drift. Since the samples were biased to severe malaria, the results do not fully depict the Kilifi population genetic diversity.

CHAPTER 1

LITERATURE REVIEW

1.1 General introduction

The causative agents of malaria are the *Plasmodium* parasites that belong to the parasitic phylum *Apicomplexa*. These parasites display an amazing level of adaptation; about 200 species of *Plasmodium* infect specific lineages of rodents, birds, reptiles and primates. These species display specificity for particular hosts. However, there have been *Plasmodium* species that have jumped between hosts. This zoonosis is believed to be the explanation behind the human origin of *P.vivax,* an ancestral parasite has been shown to have made a leap from macaque monkeys to humans (Escalante et al., 2005). The same applies to *P.knowlesi* , a simian species that afterwards became a significant cause of malaria in human beings in South-East Asia (Singh et al., 2004; Figtree et al., 2010).

Currently, five species of *Plasmodium* (*P. falciparum*, *P. vivax*, *P. ovale* and *Plasmodium malariae and P. knowlesi)* have been shown to cause malaria in human beings. Singh & Chitnis (2012) noted that out of the five species, *P. falciparum* is the most pathogenic and common cause of malaria-related deaths. Vivax malaria also contributes significantly to the malaria disease burden.

According to the Centres for Disease Control and Prevention (2014), Africa is worst hit due to a number of factors: *Anopheles gambiae* complex is a very efficient vector and is responsible for high transmission. The major species of parasite found in this region is *Plasmodium falciparum* which is the most pathogenic of the *Plasmodium* species. Insufficient resources and the unstable

socio-economic status in Africa also have had a negative impact on activities aimed at malaria control.

The threat posed by malaria to global health is made worse by extensive drug and insecticide resistance both in the parasite and the mosquito vector, respectively and the lack of an effective vaccine (Min-Oo & Gros, 2005). The most advanced vaccine candidate against *P. falciparum* is RTS, S/AS01. This vaccine has shown 51% efficacy in reducing all episodes of clinical malaria in infants aged 5-17 months over 15 months in a phase 2 trial in Kenya (Bejon et al., 2011;WHO, 2012). The most recent results from a clinical trial testing showed that the vaccine lowers the risk of clinical episodes of malaria by only 31% in babies aged between 6 and 12 weeks (Vogel, 2012). The life cycle of the malaria parasite is complex and makes the development of a malaria vaccine challenging.

1.2 Life cycle of the *Plasmodium* parasite

Cowman & Crabb (2006) noted that the entire Apicomplexa phylum adopts a common manner of host-cell access. Nevertheless, each species has unique features and exploit a specific set of ligand-receptor interactions. These adhesins are attached to a parasite actin-based motor, which drives entry of the blood stage of the parasite into the erythrocyte. They further note that while some *Apicomplexa* can invade diverse host cells, the disease-associated blood-stage form of the malaria parasite is confined to erythrocytes.

Additionally, Wiser (2011) noted that the life cycle of *Plasmodium* inside the human body can generally be categorized into 2 phases: pre/exo-erythrocytic and erythrocytic phases (Figure 1.1). It takes place in two hosts, the definitive and intermediate hosts. The sexual stages occur in the mosquito (the definitive host) while the asexual stages of the life cycle occur in the intermediate

host. The life cycle in the intermediate host begins when the parasite is injected with saliva during the time the infected mosquito is feeding (blood meal). It first goes through a round of merogony in the hepatocytes followed by several rounds of merogony in the erythrocytes. In the erythrocytes the parasites goes through the ring, trophozoite and schizont stages, a process that takes about 48 hours in *P. falciparum*.

The red blood cell invasive stages of the parasite, merozoites, are released from the infected liver cells and invade erythrocytes (Figure 1.1) (Wiser, 2011). The repeated rupturing and invasion of erythrocytes is the cause of the clinical symptoms of malaria (Miller et al., 2002). For the continued existence of the parasite in the human host, successful invasion of uninfected erythrocytes by merozoites is a prerequisite. This is a dynamic and refined process, and entails multiple steps of contact with receptors on the red cell and parasite ligands (Cowman & Crabb, 2006).

After invasion, a number of merozoites form the sexual-stage gametocytes which are taken up by the mosquito vector while taking a blood meal. The gametocytes mature in the mosquito to form gametes which fuse to form a zygote. The zygote then develops into an ookinete that infects the midgut of the mosquito. In the midgut the ookinete develops into an oocyst which produces sporozoites. The sporozoites invade the vector's salivary glands in preparation for transmission to the intermediate host.

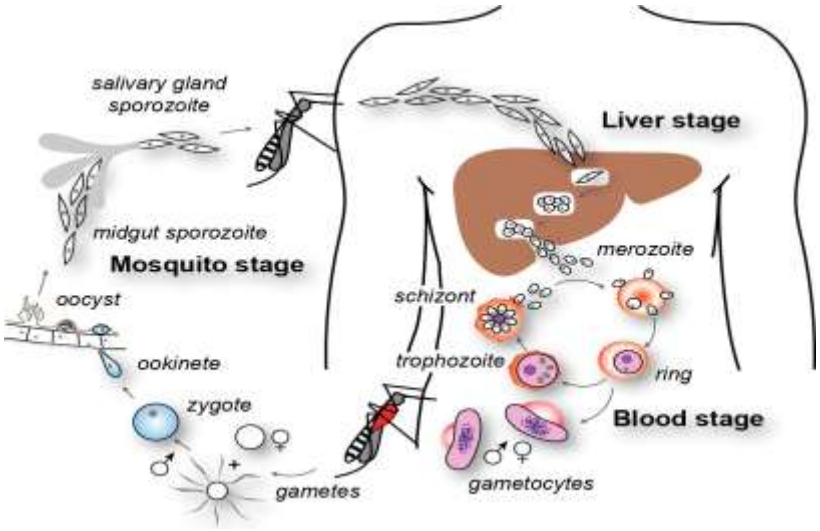

Figure 1.1 Parasite life cycle in the mosquito and intermediate host (Cowman et al., 2012)

1.3 Invasion of erythrocytes

Invasion of red blood cells by the extracellular merozoite includes a highly organized series of molecular interactions and signal transduction events involving the parasite and host erythrocyte (Harvey et al, 2012). In *P. falciparum*, erythrocyte invasion involves the interaction of a number of parasite ligands with receptors on the red cell surface. Key to the success of the *P. falciparum* parasite is its capacity to make use of a number of the red blood cell (RBC) receptors. The use of particular RBC receptor and parasite ligand is therefore not static (Baum et al., 2005).

Proteins on the surface and in specialized secretory organelles at the apical end of the merozoite participate in cell recognition and invasion of the RBCs. Three morphologically distinct apical organelles are detected by electron microscopy: micronemes, rhoptries, and dense granules (Richard et al., 2010). Two key malaria ligand families have been associated with these variable ligand receptor interactions utilized by *P. falciparum* to invade human red cells: the Erythrocyte Binding Ligand (EBL) family, micronemal proteins and Reticulocyte binding Homolog (PfRH) family, rhoptry neck proteins. Ligands from the EBL family mainly direct the sialic acid (SA)

dependent pathways of invasion and the RH family ligands (except for RH1) take part in SA independent invasion (Lobo & Ord, 2012).

Merozoite Surface Protein 1 (MSP1), with its Glycosyl Phosphatidylinositol (GPI) anchor could be engaged in the primary recognition of the erythrocyte in a sialic acid-dependent way (Cowman & Crabb 2006). They further noted that other *P. falciparum*-merozoite surface proteins, such as MSP3, MSP6 and MSP7 exist and are likely to also be involved in the invasion process. When the merozoites come across an erythrocyte, the low-affinity interaction between the merozoite surface and erythrocyte (likely mediated via surface MSPs) is turned into a step of irreversible attachment, perhaps simultaneous with merozoite reorientation to its apex. This step is vital and it triggers all downstream events (Riglar, et al 2011).

Systematic entry into the erythrocyte is orchestrated around the tight junction that has been created and includes rhoptry secretion, generation of a nascent parasitophorous vacuole, and parasitophorous vacuole membrane, activation of actomyosin motor and shedding of the merozoite surface coat (Singh & Chitnis, 2012). The formation of a moving junction which links the membranes of the invading parasite and the host cell is a common characteristic of the Apicomplexan parasites. The two key components of this junction are the surface protein Apical Membrane Antigen 1 (AMA1) and the Rhoptry Neck Protein (RON) complex, which is targeted to the host cell membrane during invasion. Erythrocyte glycophorin receptors and malaria parasite EBL ligands interact in a sialic acid-dependent manner during parasite entry into human red blood cells. EBL-1 interacts with erythrocyte receptor Glycophorin B (GPB) (Mayer et al., 2009). Human erythrocyte Glycophorin A (GPA) which is the most abundant receptor on the RBC surface and Glycophorin C (GPC) were earlier on identified as receptors for *P. falciparum* ligands EBA-175 (Orlandi et al., 1992) and EBA-140 (BAEBL) (Lobo et al., 2003), respectively (Figure

1.2). Other known erythrocyte receptors involved in invasion include Basigin, Complement Receptor 1, Band 3 and Duffy Antigen Receptor for Chemokines (DARC) (Crosnier et al., 2011, Tham et al., 2010, Goel & Li, 2003). They noted that the corresponding ligands for the first three are PfRh5, PfRh4 and MSP1, respectively (Table 1.1). DARC is utilized by *P. vivax* and *P. knowlesi* malaria parasites. These parasites bind to DARC using the Duffy binding Potein (DBP) (Wertheimer & Barnwell, 1989).

Table 1.1 Summary of known malaria parasite ligands and their receptors

PARASITE LIGAND	RBC RECEPTOR	REFERENCE
EBA-175	GPA	Orlandi et al., 1992
EBA-140	GPC	Lobo et al., 2003
EBL-1	GPB	Mayer et al., 2009
PfRh5	Basigin	Crosnier et al.,2012
MSP1	Band 3	Goel & Li, 2003
PfRh4/CD35	CR1	Tham et al., 2010
DBP	DARC	Wertheimer & Barnwell, 1989

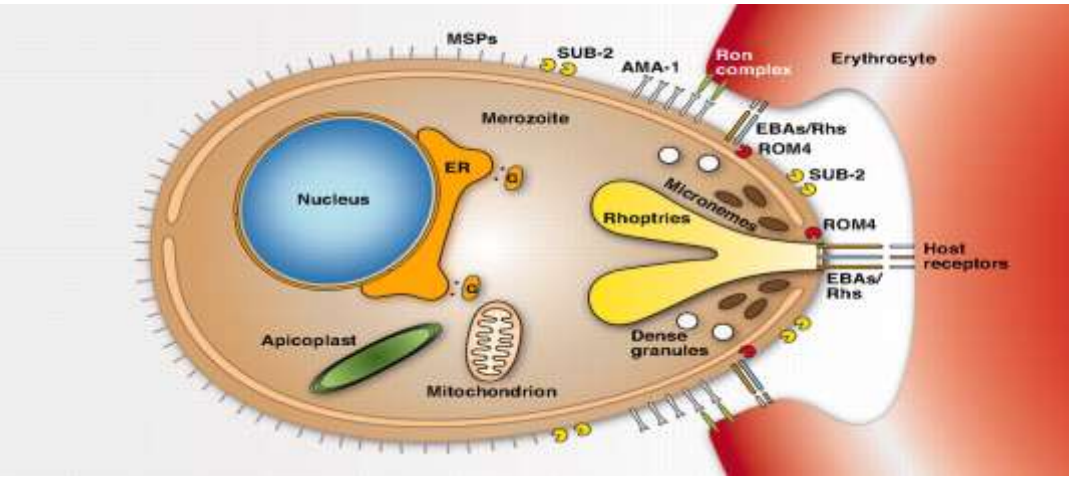

Figure 1.2 Diagram of the merozoite stage showing some of the proteins involved in invasion of the erythrocyte through ligand-receptor interactions (Kappe et al., 2010).

1.4 Erythrocyte genetic polymorphisms

Human erythrocyte polymorphisms have long been thought to take part in resistance to malaria (Williams, 2006). According to Ko et al., (2011), malaria is also the strongest known force for evolutionary selection in the recent history of the human genome. Some of the best described malaria protective polymorphisms are found in genes that code for erythrocyte-specific structural proteins or metabolic enzymes (Table 1.2) which act as a drawback to the blood-stage of the disease (Min-Oo & Gros, 2005). High frequencies of host erythrocyte polymorphisms such as the thalassaemias, sickle cell disease, complement receptor-1 (CR1) deficiency, glucose-6-phosphate dehydrogenase (G6PD) deficiency, south-east Asian ovalocytosis (SAO) and Duffy negative blood groups are found in regions where malaria is prevalent (Table 1.2) (Fowkes et al., 2008).

Table 1.2 Examples of erythrocyte genetic polymorphisms and their mechanism of protection against severe malaria.

Name of host polymorphism	Gene affected	Parasite ligand	Polymorphism	Mechanism of protection	Reference
Receptor Protein on red blood cells					
Duffy negative	FY (DARC)	Duffy binding protein	GATA-1 motif	Duffy null erythrocytes are resistant to invasion by *P.knowlesi* and *P. vivax*	(Miller, Mason, Clyde, & McGinniss, 1976)
Deficiency of Glycophorin C	GYP C	EBA-140	Deletion of Exon3	Shields against invasion by *P. falciparum* that is mediated by EBA-140	(Maier et al., 2003)
Band 3	SLC4A1	MSP- 1	Deletion of 27bps in SLCA4	RBCs with polymorphism are able to resist invasion. It also causes an increase in adhesion of *P. falciparum* infected ovalocytes to CD36 and this causes a reduction in binding of infected RBCs in the brain	(Cortes, Benet, Cooke, Barnwell, & Reeder, 2004)
Complement Receptor proteins	CR1	PfRH4	Sl2 or McCb	Reduces the ability of infected erythrocytes to rosette	(Cockburn et al., 2004)
Enzymes in the red blood cell					
G6PD deficiency	G6PD	Not applicable	A376G/ G202A	Causes premature phagocytosis of infected RBCs	(Cappadoro et al., 1998)
Hemoglobinopathies					
HbS	HBB	Not applicable	Glutamate to valine substitution at position 6 of the HBB gene	Reduced cytoadherence of *Plasmodium falciparum*-infected RBCs carrying sickle hemoglobin	(Cholera et al.,2008)
a-thalasssemia	HBA1/HBA2	Not applicable	3.7-kb deletion	α-thalassemia lessens the pro-inflammatory consequences of cytoadherence.	(Krause et al., 2012)

1.4.1 Anion-Exchange Protein 1 (Band 3)

The product of this gene is an integral red blood cell membrane protein that plays an important role in maintaining the erythrocyte cytoskeleton and shape, and functions as the principal exchanger of bicarbonate for chloride in the process of removing carbon dioxide from tissues (Patel et al., 2004). The erythrocyte polymorphism, referred to as Southeast Asian ovalocytosis (SAO), is found in malaria endemic regions is strongly associated with heterozygosity for a 27-

base pair deletion in exon 11 corresponding to codons 400 to 408 of the band 3 gene (Vasuvattakul et al., 1999). The combined 5ABC and 6A regions of ectodomains 5 and 6 are the regions of the gene involved in interacting with the parasite ligand during the invasion of *P. falciparum* (Figure 1.3). 5ABC interacts with the 42- kDa processing product of merozoite surface protein 1 ($MSP1_{42}$) through its 19-kDa C-terminal domain (Li et al., 2004). Exons 17 and 18 of the gene code for the binding regions 5ABC and 6A (Figure 1.4)

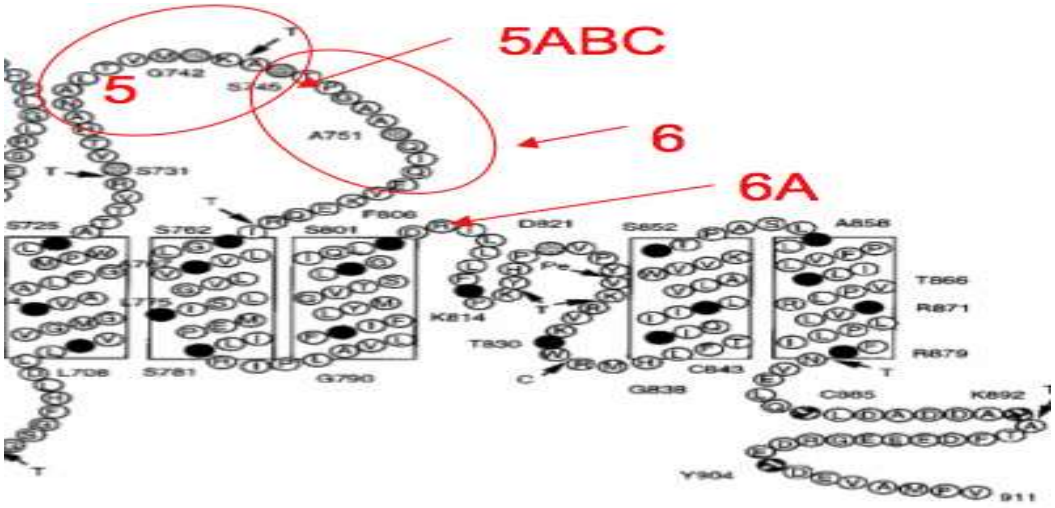

Figure 1.3 The structure of band 3 protein showing regions 5ABC and 6 A that are involved in binding MSP 1 (Fujinaga, 1999).

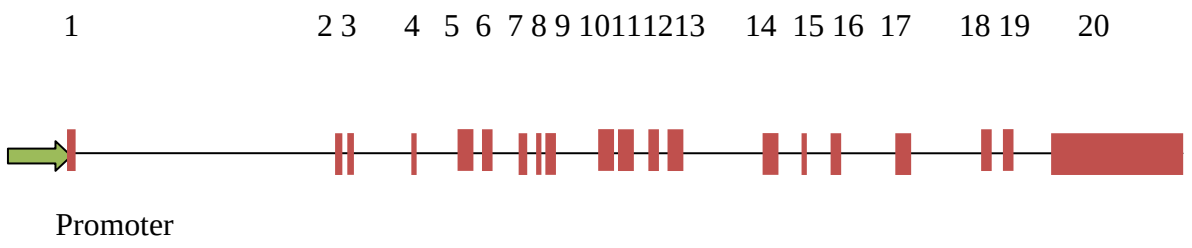

Figure 1.4 Structure of the Anion-Exchange protein 1 (Band 3) gene. The numbers above indicate the exons of the gene (Kalcreuth et al., 2006).

1.4.2 Duffy Antigen Receptor for Chemokines (DARC)

The Duffy blood group antigen is not only a receptor for *Plasmodium vivax* and *P. knowlesi* malaria parasites but it also serves as a blood group antigen and receptor for a family of proinflammatory cytokines termed chemokines (Demogines, Truong, & Sawyer, 2012).

DARC has two main polymorphic sites which are responsible for the FY*O allele and the FY*A/FY*B alleles (Figure 1.7). These are nucleotides T-33C found in the promoter region and governs the FY*O allele and G125A in the second exon that is responsible for the FY*A/ FY*B alleles (Figure 1.6) (Demogines, Truong, & Sawyer, 2011).

The Duffy-null allele (FY B^{ES} / $FY^{*}O$) is almost fixed in sub-Saharan Africa and non-existent elsewhere (Figure 1.5). This enables individuals who originate from this region to be resistant to invasion by *P. vivax* due to the replacement of a nucleotide in the binding site of the transcription factor GATA-1 (Figure 1.6) (Miller et al., 1976). This mutation inhibits DARC gene expression on red blood cells. Individuals with two copies of this version of the SNP in the promoter region completely lack the protein on their red blood cells, but the gene is still transcribed in non-erythroid cells. However, evidence has emerged that *P. vivax* is being transmitted among duffy null individuals (who have the SNP in the GATA-1 binding site) of western Kenya (Ryan et al.,

2006). Similar findings have been observed in Madagascar (Menerd et al., 2010) and Equatorial Guinea (Mendes et al., 2011). This provides evidence that there is a mechanism for erythrocyte invasion that may not depend on DARC alone (Ryan et al., 2006; Menerd et al., 2010).

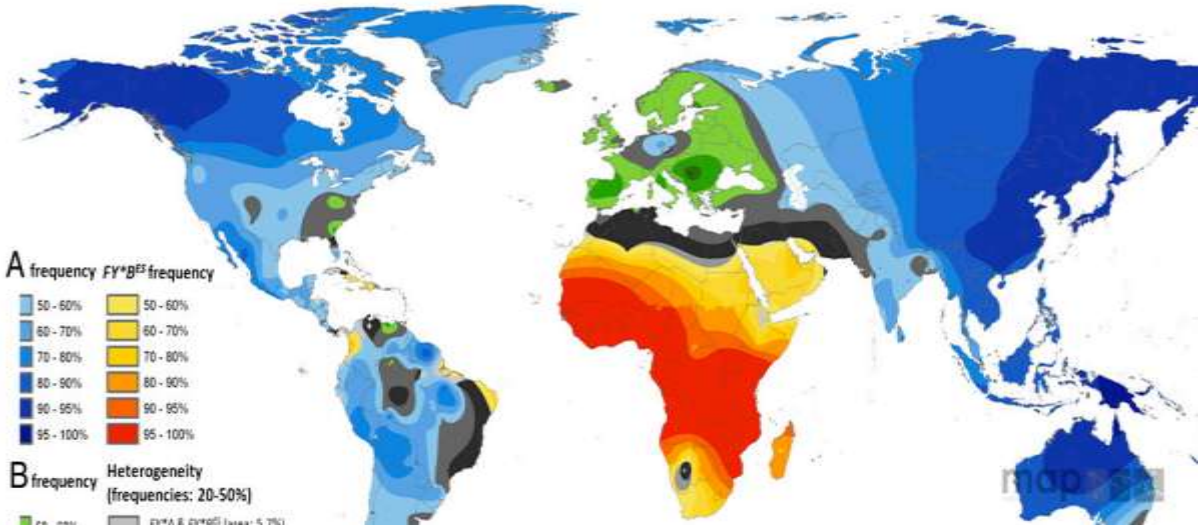

Figure 1.5 Shows the global distribution of DARC alleles. Regions where a single allele is dominant (frequency ≥ 50%) are shown using a colour gradient (red/yellow, FY*B^{ES} green, FY*B; blue, FY*A). Regions that have allelic heterogeneity where no single allele dominates but have two or more alleles each with frequencies ≥ 20% are indicated in the grey-scale. The palest parts in the grey-scale represent areas of heterogeneity between the silent FY*B^{ES} allele and either FY*A or FY*B (when these are inherited together, they do not produce new phenotypes), and darkest part of the grey-scale shows the occurrence of all three alleles together. ((Zimmerman et al., 2013)

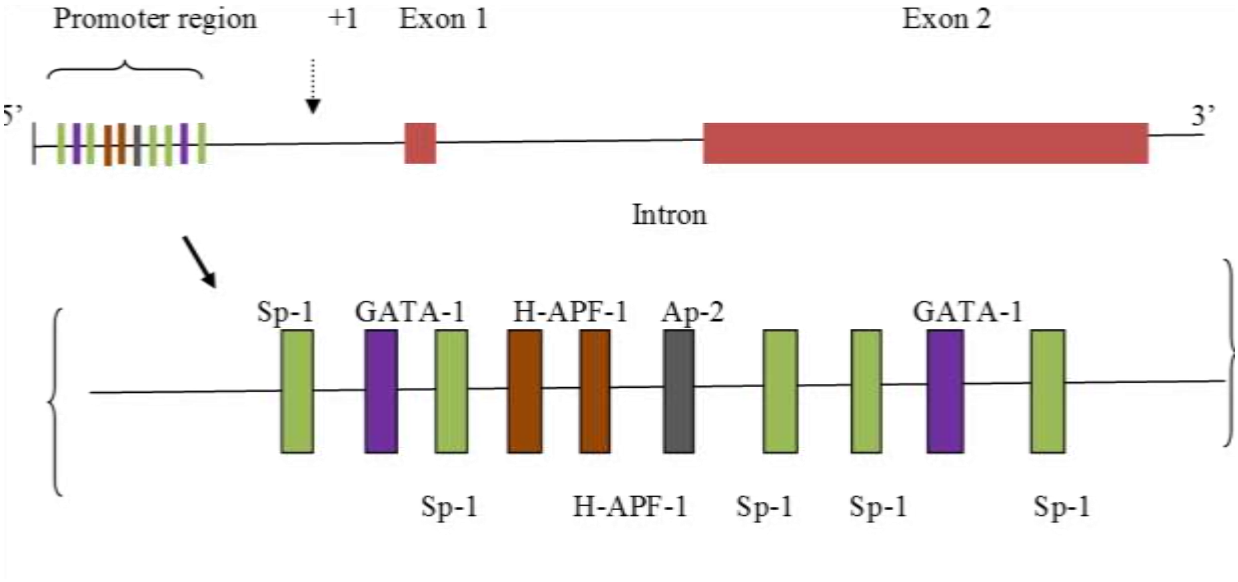

Figure 1.6 Organization of the DARC gene in human beings (A). This figure shows the position of the erythroid GATA site in relation to several other well-known regulatory binding sites in the promoter (Oliveira et al., 2012)

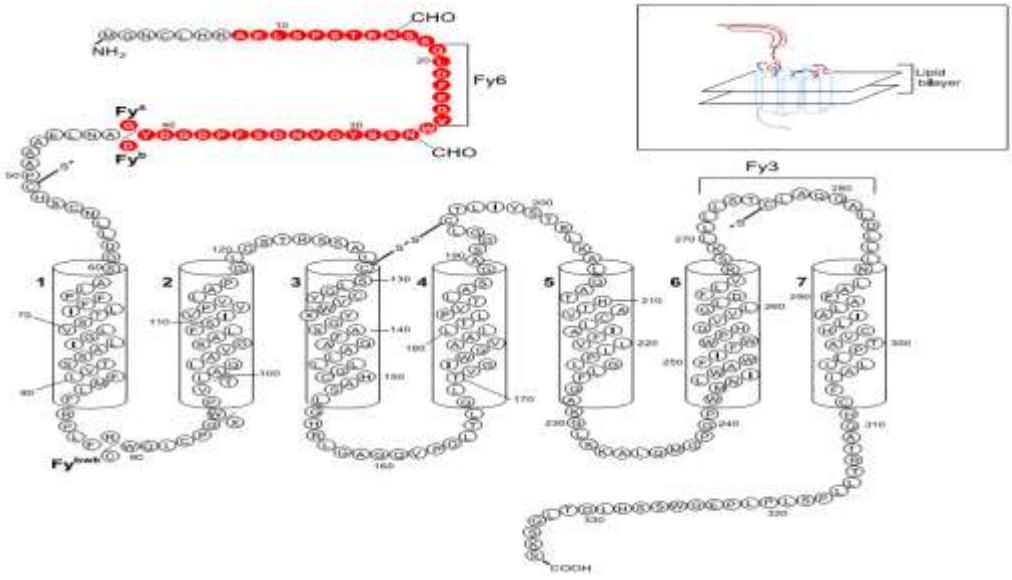

Figure 1.7 The structure of the DARC protein. The region that is involved in binding *P. vivax* is marked in red. Some of the already known polymorphisms in the protein are also shown (Zimmerman et al.,2013).

1.4.3 Complement Receptor 1

According to Liu & Niu (2009), the complement receptor 1 (CR1, CD35) is best characterized as a receptor for the activated form of the complement protein C3, C3b. They note that CR1 was identified as a major receptor for rosetting, a *P. falciparum* virulence phenotype, characterized by the ability of uninfected RBCs to bind to iRBCs to form clumps in vitro, that has been associated with severe malaria in a number of studies in Africa (Teeranaipong & Ohashi, 2008). CR1 mediates rosetting through its interaction with PfEMP-1, a parasite-derived variant erythrocyte membrane protein (Chen et al., 1998).

The ectodomain portion of the most common size allotype of CR1 consists of 30 complement-control-protein repeats (CCPs) or short consensus repeats (Figure 1.9). The 28 N-terminal CCPs are arranged based on the level of homology into four long homologous repeats (LHRs) A–D. The LHRs possibly arose from duplication events of gene sections that encode the constituent Short Consensus Repeats (SCRs) of an LHR. The variation in the size of CR1 mirrors allotypic polymorphism that involves having dissimilar numbers of LHRs. Each LHR consists of seven CCPs. There is 70–95% identity between each LHR. Each CCP has a length of 59–72 amino acids (Liu & Niu, 2009). The last two SCR do not form part of any LHR (Figure 1.9) (Vik & Wong, 1993).

CR1 protein binds PfRh4 using CCP 1-3 (Tham, & Wilson 2010; Tham et al., 2011) and the exons that code for CCP 1-3 are 2,3,4 and 5 (Vik & Wong, 1993) (Figure 1.8).

A number of polymorphisms generated by single nucleotide polymorphisms (SNPs) have been detected in the CR1 protein. Some of these SNPs generate a number of blood group antigen

variants such as the Knops blood group antigens (Liu & Niu, 2009). Studies done by Thathy et al.,(2005) demonstrated that Swain-Langley (*S)l2* and McCoy (*McC)b* of the Sl and McC blood group antigens of CR1 confer a survival advantage in a *P. falciparum* malaria endemic region of western Kenya. The *Sl2* allele is the result of the substitution of glycine, a neutral amino acid, for the basic amino acid arginine at position 1601 (R1601G), whereas the *McCb* allele is the result of the substitution of glutamic acid for lysine at position 1590 (K1590E) in exon 29. A study done by Teeranaipong & Ohashi, 2008 showed that a SNP in intron 27 of CR1 partly affected the quantitative expression of CR1 on erythrocytes. This SNP is composed of A and T, which correspond to high-expression (H) and low-expression (L) alleles, respectively and individuals with the (L) allele were protected from severe malaria.

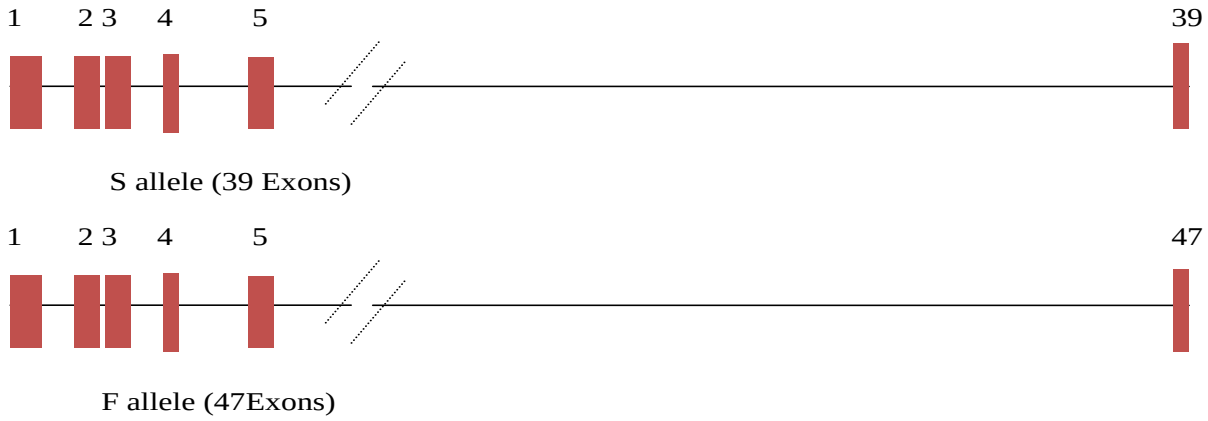

Figure 1.8 Structure of CR1 gene showing the two common alleles, S and F (Atlas of Genetics and Cytogenetics in Oncology and Haematology, 2014). The pink boxes indicate the exons and the numbers above show the exon numbers. Breaks (//) after exon 5 indicate exons 6-38 (or 6-46) and the introns in between.

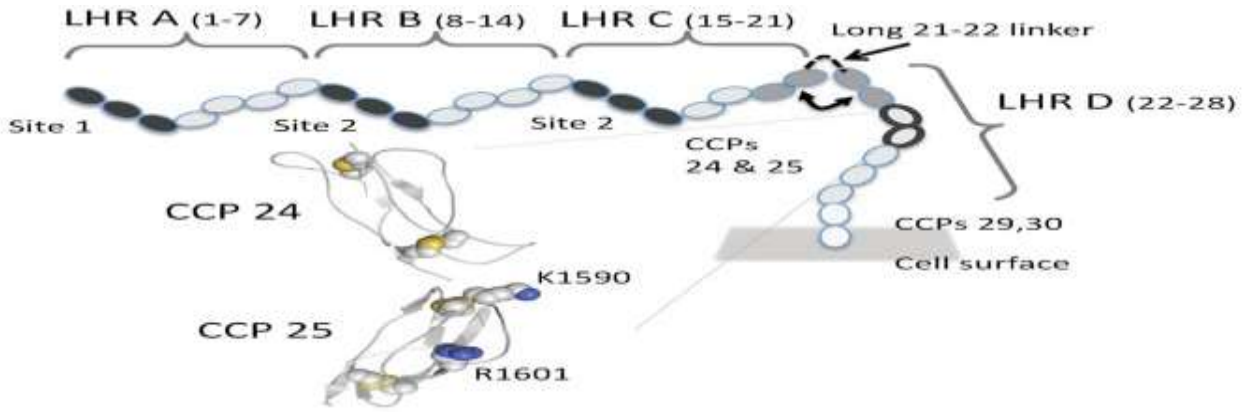

Figure 1.9 Structure of CR1 protein ectodomain portion of the most common size allotype of CR1 (Derived from Tetteh-Quarco et al., 2012)

1.5 Evidence of natural selection during host-parasite interactions

Hosts and pathogens have evolutionary outcomes on each other that lead to co-evolution. Positive natural selection is common in genes that take part in the interaction between host and pathogen (Oliveira et al., 2012).

According to Woolhouse et al., (2002), host resistance may possibly have a cost on fitness in the absence of disease. This cost may become evident in a number of ways such as reduced chances of survival, reduced ability to compete and reduced fertility. A good example is heterozygous advantage that is well illustrated by the association between human sickle cell anaemia and vulnerability to malaria. The mutation that is responsible for sickle cell anaemia is life threatening if it occurs in the two β-globin gene**s** of an individual. However, heterozygous individuals have been shown to be less susceptible to severe malaria (Cholera et al., 2008).

The detection of positive selection in the genes that are involved in interactions between the host and pathogen is one of the ways of determining whether molecular co-evolution is taking place. The underlying principle is that co-evolution is known to accelerate the accumulation of genetic variation. As a consequence, positive selection should be more probable where there is co-

evolution, however this thought remains controversial. One of the ways of identifying positive selection entails the comparison of the proportion of non-synonymous (brings about a change in the amino acid) to synonymous or silent substitution of nucleotides (does not involve a change in the amino acid) (Woolhouse et al., 2002).

Some of the receptors on the red blood cells such as CR1, DARC and Band 3 have other biological functions other than acting as receptors for the malaria parasite. They may therefore be under twofold (internal and external) selective pressures. Purifying selection is a form of internal pressure that tries to preserve a protein's biological role. An external selective pressure (positive selection) may also act due to infection by the malaria parasite. This positive selection drives adaptation that enables the host to be resistant to infections (Oliveira et al., 2012).

Mutations that diminish the efficiency of the invasion process would confer a selective advantage to the host and might be expected to increase in frequency over time through natural selection. The project focuses on understanding the polymorphisms of some of the human genes involved in parasite invasion. Since the invasion process is complex and incompletely understood, studying these polymorphisms may give more insight into the host-parasite relationship.

1.6 Research Question

What polymorphisms exist in CR1, Band 3 and DARC in a malaria endemic population in Kilifi, Kenya?

1.7 Hypothesis

Polymorphisms exist in CR1, Band 3 and DARC genes in the malaria endemic population in Kilifi.

1.8 Objectives

- To determine the genetic variants in three erythrocyte receptor genes: Band 3, Complement Receptor 1 (CR1) and Duffy antigen Receptor for chemokines (DARC).
- To identify the polymorphisms in CR1, band 3 and DARC genes in a malaria endemic population and determine their frequencies.
- To assess whether the identified polymorphisms are under selection.

1.9 Justification

Malaria is a disease that has continued to be a health burden to humans. It is a major cause of death in developing countries with children and expectant women being the most affected (Centers for Disease Control and Prevention, 2014) . Approximately, 3.4 billion people reside in regions that are considered to be at risk of malaria transmission. In 2012 alone, it is estimated that malaria caused 207 million clinical episodes and a further 627,000 fatalities, about 91% of these deaths occurred in Africa (WHO, 2014).

A number of factors such as the intricate biology of the *Plasmodium* parasites, their high polymorphism and their resistance to antimalarial drugs have hindered the successful control of malaria. Since merozoite invasion is one of the stages in the parasite life cycle that is susceptible to immunity, it is an area that has been given attention by many researchers (Leykauf & Treeck, 2010). Researchers have been trying to develop vaccines against the blood stage antigens of the parasite. A vaccine against the malaria parasite has great potential towards conquering the disease. Nevertheless, the quest for an effective vaccine has been a difficult task partly due to the widespread genetic diversity in parasite antigens considered as vaccine candidates (Muramatsu, 2012).

Host immune selection has led to the emergence of antigenic diversity due to allelic polymorphisms in the malaria parasite. A lot of work has been done to characterize the polymorphisms in the genes coding for ligands of the malaria parasite that are required for invasion. These ligands have been considered as vaccine candidates. The identification of polymorphisms in the respective erythrocyte receptors therefore has the potential to give more insights into the host–parasite relationship that could potentially stir new advances in the prevention and treatment of the disease. The association involving the parasite blood-stage infection and the RBC polymorphism will also be essential in the estimation of the populations which are at risk and this will guide the efforts towards malaria elimination.

CHAPTER 2

2.0 MATERIALS AND METHODS

2.1 Study population

This study was carried out in the county of Kilifi (Figure 2.1) in the coastal region of Kenya. It is regarded as a malaria endemic region although studies have shown that transmission has been on the decline recently. The particular study site is the Kilifi District Hospital which is within the Kilifi Health and Demographic Surveillance System (KHDSS) that covers an area of 891km^2 and caters for population of 261,919 people. 18% of this population are below the age of 5 (Scott et al., 2012). The Kilifi District Hospital serves Kilifi County of 1,109,735 inhabitants according to the 2009 census (County Government of Kilifi, 2013). The largest ethnic group in Kilifi County are the Mijikenda. In the urban areas and the areas bordering the Indian Ocean there is a sizeable population of the Swahili and other people who originate from other areas of the country.

The KHDSS maintains extensive epidemiological research based at the KEMRI/ Wellcome Trust Programme. One of their core research areas is on host genetic vulnerability to infectious diseases.

Figure 2.1 The map of Kenya showing the different counties in the country. The county where the study was conducted is Kilifi and it is marked with green on the coast of Kenya bordering the Indian Ocean.

Ninety three human DNA samples were used, the diploid number of alleles being 186. These samples were extracted in 2001 from whole blood taken from patients admitted to the Kilifi District hospital's High Dependency Unit (HDU) suffering from severe malaria. Malaria transmission in this region is seasonal and occurs mainly during two rainy seasons from April to July and from October to November (Scott et al., 2012). The samples were collected during both the dry and wet seasons (Appendix).

Severe malaria in this case was characterized with severe malaria anaemia , cerebral malaria and respiratory distress. Severe malarial anaemia was described as a febrile illness with hemoglobin level< 5g/dl^2. Cerebral malaria was defined as a Blantyre coma score ≤ 2. Fifty of these patients were males while fourty were females. The sex of two patients was not defined. The individuals

had variable parasitemia, with a median of 190,000 parasites /µl. The age of the study participants ranged from 6 months to 8 years, with the mean age being 3.5 years.

2.2 DNA extraction

Genomic DNA from the blood samples of each individual was previously extracted using the DNA blood minikit from QIAGEN, Inc. This kit allows for the generation of DNA that is ready for use following a series of uncomplicated steps. To begin with, the cellular components of blood were lysed using a lysis buffer. DNA was then absorbed onto the QIAamp silica-gel membrane in a brief centrifugation step. PCR inhibitors, for instance divalent cations and proteins were removed completely in two wash steps due to the salt and pH conditions in the buffers that are used, leaving behind pure nucleic acid to be eluted in elution buffer that comes with the kit. The eluted DNA was stored at -20˚C after making a 1:10 dilution.

2.3 Primer Design

Apart from Band 3 promoter primers that were obtained from literature (Kalcreuth et al., 2006), the rest were designed using Editseq and Seqman applications from DNASTAR version 11 (Madison, WI, USA). Reference sequences for human genes were obtained from the National Center for Biotechnology Information (NCBI) website. The Genbank accession numbers are shown in Table 2.1 below.

Table 2.1 Genes that were sequenced and their Genbank accession numbers

Gene	Chromosome number	Chromosomal location	Genbank accession number
DARC	1	1q21-q22	NC_000001.11
CR 1	1	1q32	NC_000001.10
Band 3 gene and Band 3 promoter	17	17q21.31	NC_000017 L35930.1

The sequences were imported into the Editseq application (DNASTAR Lasergene software suite version 11) and CR1 exons 2-5, Band 3 exons 17 and 18 and DARC promoter, and its two exons were marked in uppercase. The rationale for choosing the regions in CR1 and Band 3 was based on available literature that showed that they play a role in binding *P. falciparum* ligands (Tham, & Wilson, 2010;Kalcreuth et al., 2006).

The PCR primers flanked the regions of interest and were designed to be gene specific (Table 2.1) using the Seqman application. The important considerations while designing the primers included the primer length, GC content, specificity and melting temperature (Tm). The primer lengths were between 18-24 bps. Such lengths are long enough for the primers to be specific and short enough for them to bind with ease to the template at the annealing temperature (PREMIER Biosoft, 2014). Primer melting temperature (Tm) is described as the temperature at which half of the double stranded DNA will separate to form single strands. The best results are produced by primers whose Tm are in the range of 52-58^{o}C (PREMIER Biosoft, 2014). All the primers that were designed had melting temperatures of 60^{o}C and below. The GC content in the primers was between 40-60%. The primers were designed in such a way that in the codons at either ends of the primers at least one base was either a G or a C. According to PREMIER Biosoft, 2014, this helps to promote end stability and specific binding due to the stronger bonding of G and C bases.

Regions that were avoided while designing the primers included those that had repeats, a continuous run of a single base and regions that were homologous to other genes. Repeats were avoided because they can cause mispriming. To avoid cross homology that would cause the primers to amplify other genes in the mixture, a BLAST was performed against the sequences in NCBI to test the specificity of the PCR primers. The reverse primers were reverse complemented before ordering. The primers were ordered from Sigma, UK (Table 2.2).

Table 2.2 PCR and sequencing primer sequences.

Gene	Region	Primer name	Primer sequence
Complement Receptor 1	Exon 2	F2	**aagtatggtaatttctccat**
		R3	**aggaactcaaagcagtaacaag**
	Exon 3	F3	**gttgagaccttatgtactaaaa**
		R4	**gttaaagagcagatggtaatag**
	Exon 4	F4	**gtgatagatagtcctttgat**
		R5	**tgaaggacagattgcacagaa**
	Exon 5	F5	**gtttagtgactcatgagatttc**
		R1	**caaatactaatctcctgatccaac**
DARC		F1	**tcggtaaaatctctacttgct**
		F2	tcatttcccgtgctgttt
		F3	tgtaactctgatggcctc
		F4	tgctggatgactctgcac
		F5	tactgacactgcctgtca
		R2	catcagagttacaccgga
		R3	gaaaaggaagagatataaaga
		R4	aagacgggcaccacaatg
		R1	**ttcacaaaggcagtgtagact**
Band 3	Promoter	Forward	**cagctctttagaaccagccagggtc**
		Reverse	**cagggtcccttggtgaagtcctgc**
	Exon 17 and 18	F1	**aactgagctacaaggacacc**
		R2	tgctgggtccgaacaga
		F2	catatggtgcctgtgttt
		R1	**gatgcccgtgaataagtgcatg**

The primers marked in bold were used as PCR and sequencing primers.

2.4 Amplification of DARC, CR1 and Band 3 genes

Expand High Fidelity PCR system from Roche was used to amplify the selected regions in the genes using polymerase chain reaction (PCR) in a final volume of 10µl. The 10µl reaction was in two parts. This helped to avoid the 3'-5' exonuclease activity of the high fidelity Taq polymerase. It has the ability to proofread and this can cause some degradation of templates and primers when the reaction is being set up. All the reagents were vortexed and centrifuged prior to starting to

ensure that they were homogeneous. The two mixes were prepared in sterile eppendorf tubes. The mixes were prepared for multiple reactions and the volumes prepared included one extra tube to take care of the volumes lost while pippeting.

The first reaction mix (for one reaction) contained PCR water, 1.0µl of 25mM $MgCl_2$ stock solution, 0.2µl of dNTP mix (10mM of dGTP, dATP, dTTP and dCTP), 0.3µl of both forward and reverse primers (10uM). The second reaction mix (for one reaction) had 3.86µl of PCR water, 1µl of Expand High Fidelity Buffer (10X) with 15mM $MgCl_2$ and 0.14µl of Expand High Fidelity Enzyme mix. The initial volume of template used was 0.5µl. For samples that failed to amplify, the volumes were increased to 0.75 or 1µl.The volume of PCR clean water (Sigma, UK) in the first reaction mix was adjusted accordingly to give a final volume of 10µl.

Gradient polymerase chain reactions were performed to determine the optimum conditions for the amplification of the three genes. A series of annealing temperatures were used for the optimization to achieve a single band and good yield of PCR products. The thermal profile was as follows: An initial denaturation at 94^{o}C for 2 minutes, 25 cycles of denaturation at 94^{o}C for 15 seconds, annealing at temperatures of between 46- 64^{o}C for 30 seconds (Table 2.3), extension at 72^{o}C . Elongation time was based on how long the fragments were considering that synthesis speed for the enzyme is 1 kb/min, but not less than 45 seconds. A final 7 minute extension was done at 72^{o}C. The elongation temperature used for all the reactions was 72^{0}C. 68^{0}C was used for products that are larger than 3kb. In the second round of PCR, there was a gradual increase of extension time by 5 seconds for each successive cycle. PCR was done in PCR tubes and 96 well plates from Applied Biosystems.

Table 2.3 PCR product sizes and the annealing temperatures used during PCR.

Gene	Region	Length (base pairs)	Annealing temperature (^{0}C)
Complement Receptor 1	Exon 2	401	46
	Exon 3	290	47
	Exon 4	324	54
	Exon 5	547	58
DARC	promoter and full length gene	1869	57
Band 3	Promoter	639	64
	Exon 17 and 18	2170	56

The extension temperature used for all the reactions was 72^0C.

2.5 Gel electrophoresis

The gels were prepared using 0.5X Tris-borate- EDTA buffer pH8.0 (Appendix). 1% agarose gel was made by dissolving 1g of agarose powder (from Applied Genetic Technologies Corporation) in 100ml of 0.5X TBE buffer. The solution was boiled and cooled using running water from the tap. Before the gel set, 2µl of ethidium bromide was added and the conical flask swirled to ensure even mixing. The gel was allowed to set on a gel tray which had a comb at one end.

The combs formed wells where the samples that had been mixed with the loading dye (6X Blue Orange from Promega) were loaded. 1µl of amplified PCR product of each gene fragment was mixed with 1µl of loading dye and run on the gel to check the quality of amplification. The PCR products were run alongside 1.5µl of 1kb DNA ladder (HyperLadder 1, Bioline UK) which was used as the standard DNA marker. 0.5X TBE buffer was used as the running buffer. Electrophoresis was done for a period of 45 minutes at 100 volts. Visualization of the gels was by digital photography under UV light using the Molecular Imager Gel Doc (Bio-Rad., UK). The samples that were considered for sequencing only contained a single band.

2.6 Purification of the PCR products

The amplicons were purified using EXOSAP-IT from Affymetrix. EXOSAP-IT is used as an efficient clean-up process preceding applications such as DNA sequencing that needs PCR products that are devoid of excess primers and nucleotides. It makes use of two hydrolytic enzymes (Exonuclease-I and Shrimp Alkaline Phosphatase) that remove the excess dNTPs and primers that remained in solution after amplification, respectively. For every 5µl of PCR products 2µl of EXOSAP-IT was added. The final volume of the reaction was 12.6µl (accounting for 1µl that was used in the agarose gel electrophoresis). The EXOSAP-IT reagent was added to the PCR product and this was followed by incubation at 37^{o}C for 15 minutes, followed by incubation at 80^{o}C to inactivate the enzymes. An extra step of cooling was done for 15 minutes at 15^{o}C.

2.7 Sequencing with Big Dye Terminators

Fluorescent DNA sequencing method used was based on the Sanger method. The sequencing reaction is comparable to a PCR given that template DNA is reproduced to generate new strands starting at the site of the annealed primer. The difference is that unlike the conventional PCR where two primers are used, in dye terminator reactions only one primer is used and apart from the typical dNTPs, there are other four dye-labelled dideoxynucleotides (ddNTPs). Once a ddNTP is integrated into the growing strand of DNA, synthesis ceases due to the lack of a hydroxyl group. The ultimate products of this reaction consist of a set of different lengths of DNA fragments that are fluorescently labeled at the 3' end.

Sequencing PCR was carried out using the Big Dye Terminator (BDT) mix v3.1 in 96 well plates from Applied Biosystems. The master mix for one sequencing reaction was prepared by mixing the following reagents in a sterile eppendorf tube: 0.5µl of Big Dye Terminator (BDT) ready reaction mix v3.1, 1.75µl of 5X sequencing buffer, 0.3µl of primer (10µM) and PCR clean water

(Sigma, UK). The volume of water and PCR product were adjusted accordingly to give a final volume of 10µl per reaction. For samples that had strong bands, 3µl of purified PCR product was used. The volume was increased to 4µl for samples that had faint bands. The preparation of the mix was done under minimal light because the fluorescently labeled ddNTPs are light sensitive.

The plates were then loaded onto a thermocycler. The cycling profile for the sequencing reaction was 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.

2.8 Big Dye PCR purification using ethanol/sodium acetate mixture

The products of the sequencing reaction were purified using ethanol/sodium acetate precipitation in 96 well plates. This step was necessary to remove excess primers and ddNTPs. A mix consisting of the following was constituted in falcon tubes: 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 for each well. A multichannel pipette was used; the mix was poured into a clean weighing boat and 90µl dispensed into each well. The plates were sealed using adhesive seals from Bio-Rad and placed in a -20^{o}C freezer for 30 minutes for the extension products to precipitate. Centrifugation of the precipitated products was carried out at 3000xg for 30 minutes at 4^{o}C on a 5810R bench centrifuge (Eppendorf). The plates were unsealed and inverted gently on absorbent paper towels that were folded to fit the size of the plate and absorb of the supernatant, during the centrifugation process of the inverted plate at 50xg for 1 minute. Ice-cold 70 % ethanol was poured into a weighing boat and 150µl dispensed into each well using a multichannel pipette. The plates were sealed again and centrifuged at 3000xg for 10 minutes. The plates were once more inverted over clean paper towels

and spun at 50xg for 1 minute at 4^{o}C. The plates were then covered loosely with clean paper towels and left on the bench to air dry for at least 30 minutes.

2.9 Capillary electrophoresis

Once the plates were completely dry, 10µl of Hi-Di formamide was added into each well. Capillary electrophoresis was performed in an automated 3130xl sequencer from Applied Biosystems, UK. The sequencer was able to separate DNA fragments that differ by just one base pair. Each of the four ddNTPs had a special fluorescent dye of a different colour attached to it. These dyes gave light at a different wavelength when excited using a laser beam. The resulting fluorescence was picked out by a charge-coupled device (CCD) camera and converted into a chromatogram (Figure 3.2). As the fluorescently labeled extension products from the sequencing reaction migrated through the polymer passed the laser detector, each base was detected as a colour signal

2.10 Sequence editing and alignments

The sequencing files of the three genes generated by the 3130xl sequencer from Applied Biosystems was analysed using the Seqman program from DNASTAR Lasergene software suite version 11. The program was used to align contigs and identify polymorphic sites (Figure 3.2). Contigs were visually inspected for identification of heterozygous sites and edited manually. The sequences were then saved as consensus files. DNA sequences were aligned using the Clustal W multiple alignment function in MEGA (http://www.megasoftware.net/) to identify SNPs for each individual. The files were saved in a FASTA format. The FASTA files were then opened using the Jalview program to trim the ends since DnaSP software does not accept sequences of unequal lengths. The DnaSP software was used to read unphase (or genotype) the data files (diploid individuals) in FASTA format which included the IUPAC nucleotide ambiguity codes

representing heterozygous sites (for example, a peak of A and T superimposed on each other is W).

2.11 Statistical analysis

In this study, π was computed to determine nucleotide diversity. Nucleotide diversity per site, π, refers to the variance in the average number of nucleotide differences per site between two sequences and is used to estimate genetic diversity in the population. The software allows the user to estimate the nucleotide diversity in the entire sequenced data or in specific regions by choosing the region to analyze. For sequences that had high nucleotide diversity the sliding Window option was used to target specific regions of the sequences. Sliding window allows one to select the size of the sequence block (window length) and the number of times the calculation is repeated (step size).

Tajima's D statistic, Fu and Li's D and Fu and Li's F test statistics were calculated to determine DNA sequences evolving in a random manner (neutral mutations) and those evolving in a process that is non-random. A negative Tajima's D signifies an excess of low frequency polymorphisms relative to expectation and this could be evidence of purifying selection or a recent population expansion. A positive Tajima's D signifies low levels of both low and high frequency polymorphisms and can be due to balancing selection acting on the population and recent shrinking of the population.

CHAPTER 3

3.0 RESULTS

3.1 PCR amplification

Positive and negative controls were included in the different PCR reactions to check whether there was any contamination. No amplification is expected in the negative control, its lanes had no bands. This shows that the reagents were not contaminated.

For DARC, the promoter region through to exon 1 and to the end of the second exon (whole gene) was amplified in 93 samples yielding an amplicon of about 1.8 kb (Figure 3.1). out of the 93 samples that were used in the study,89 were successfully amplified.

Exons 2, 3, 4 and 5 of CR1 were amplified separately. The fragments generated included some parts of the introns that flanked the exons of interest. Initially exons 2 and 3 were amplified in a single PCR because the intron separating them is not very long (630 bp). In this PCR, 85 samples were successfully amplified. The fragment size was approximately 1kb (Figure 3.1). After amplifying the exons separately, the samples that were succeesfully amplified were 87. For the fifth exon**,** 89 samples were amplified successfully.

The band 3 promoter was amplified separately yielding a 600bp product (Figure 3.1) and exons 17 and 18 were amplified in a single PCR yielding a 2170 bp product (Figure 3.1). For the promoter region, 86 samples were successfully amplified. 85 samples were successfully amplified for the exons 17 and 18 region.

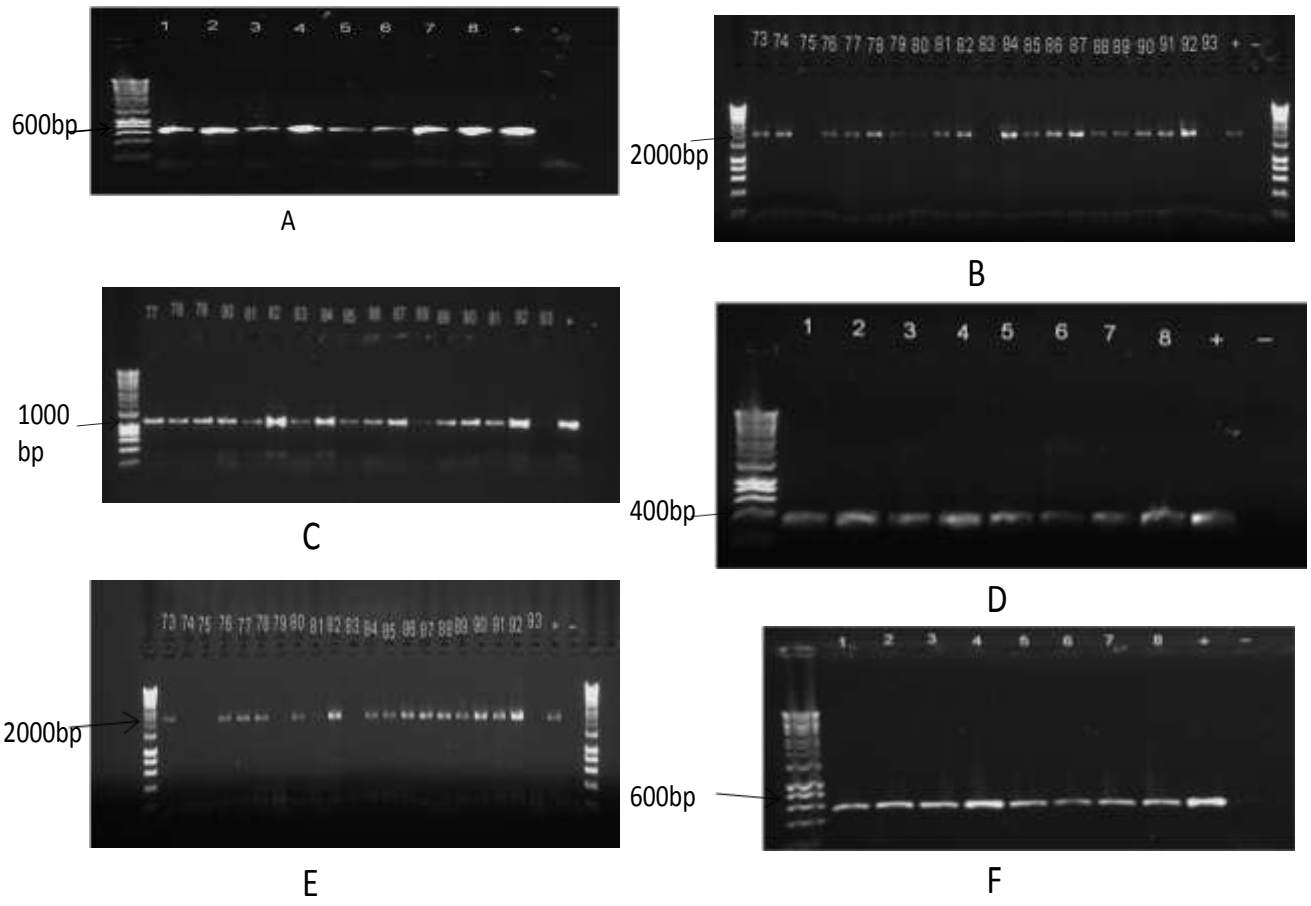

Figure 3**.1** Gel pictures of A) Band 3 promoter (600bp) B) DARC gene (approximately 2kb) C) CR1 exons 2 and 3 when they were amplified together (approximately 1kb) D) CR1 exon 4 (400bp) E) Band 3 exons 17 and 18 (approximately 2kb) F) CR1 exon 5 (600bp). The first lane (and the last one in B and E) contain the DNA hyperladder that was used to show the sizes of the amplicons. The sizes of the fragments are shown by the arrows on the right. The lanes after the ladder were for the 93 different samples and are labelled using different numbers.The same numbers on different gels do not represent the same samples. Lanes with no bands represent samples whose amplification was unsuccessful**.** The lanes marked with + and – signs were for the positive and negative controls respectively.

3.2 Sequencing results

For DARC all the samples that amplified successfully gave good sequence data. The chromatogram had clean and tall peaks.

In CR1, the amplification products of exons 2 and 3 (when they were amplified together) did not give good sequencing data because of the repeats found in the intron. The chromatogram had peaks that overlapped due to a stretch of mononucleotide sequence. Even after amplifying the exons separately, the overlapping peaks in exons 2 and 3 sequence data made it difficult to analyze. The sequences were therefore excluded from further analysis. For exon 4, 75 samples gave good sequence data. For the fifth exon 85 samples gave good sequence reads.

The band 3 promoter 83 samples yielded good sequence reads. The forward primer for exon 17 failed to give good sequence reads. The peaks in the chromatogram were low and had a lot of noise. The sequence was therefore determined using the reverse primer which yielded clean and tall peaks in the chromatogram, thus providing confidence of the 74 samples that yielded good sequence data. For exon 18, 72 of the samples resulted in good sequence reads.

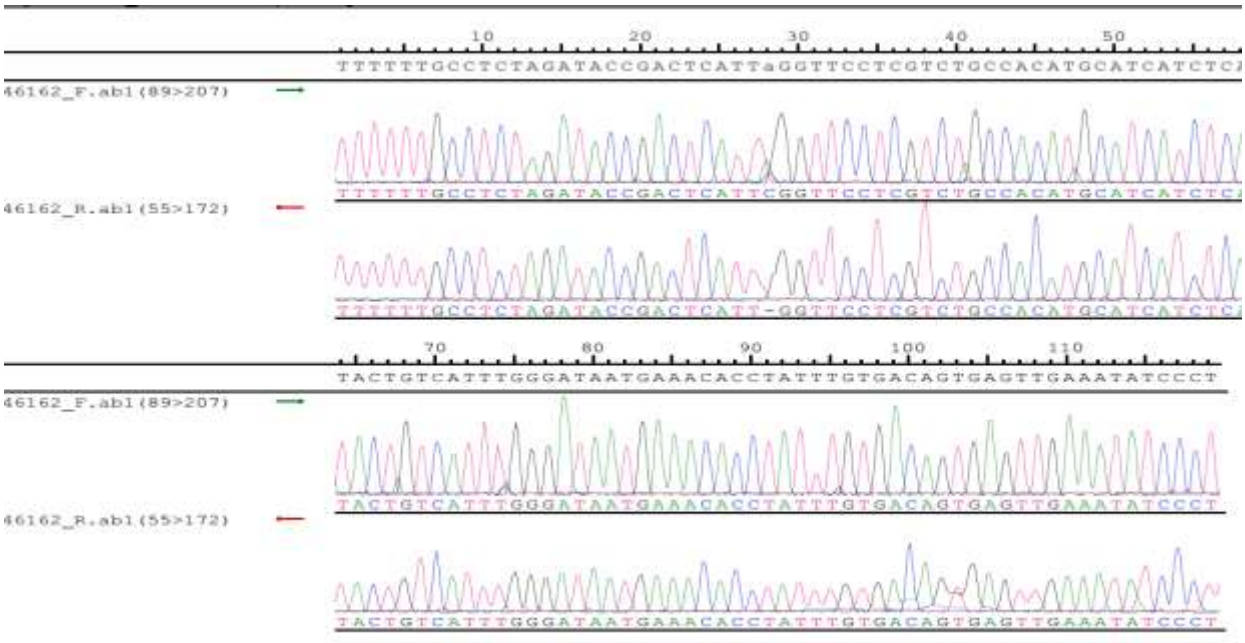

Figure 3.2 Chromatogram image showing an alignment of sequence data for exon 5 of CR1 for visualization of SNPs from 2 primer reads (46162_F and 46162_R).

3.3 Statistical test results

Table 3.1 The nucleotide diversity and neutrality test results of the band 3, DARC and CR1 exons 4 and 5.

Gene	Region	n (bp)	S	Nucleotide diversity (π)	NEUTRALITY TESTS Fu & Li's D	Fu & Li's F	Tajima's D
Complement Receptor 1	Exon 4 and Intron 4	184	1	0.00132	0.46895	0.51576	0.37574
	Exon 5	387	4	0.00039	-0.37596	-0.85753	-1.43282
Band 3	Promoter	449	3	0.00959	0.46364	1.02611	1.80087
	Exon 17	197	1	0.00027	0.46966	0.11257	-0.75057
	Exon 18	253	3	0.00159	-0.6657	-0.69513	-0.43021
DARC		12	1	0.01868	0.45954	0.48215	0.30131

n= total number of sites analysed, S= number of segregating sites

All the Tajima's D, Fu and Li's D and Fu and Li's F values were not significant, $p > 0.10$ (Table 3.1). The Tajima's D analysis was also depicted in 100bp sliding windows across CR1 exon 5 and Band 3 promoter to see which mutations were influencing the high values. However, the values were not significant, $p > 0.10$ (Figure 3.3).

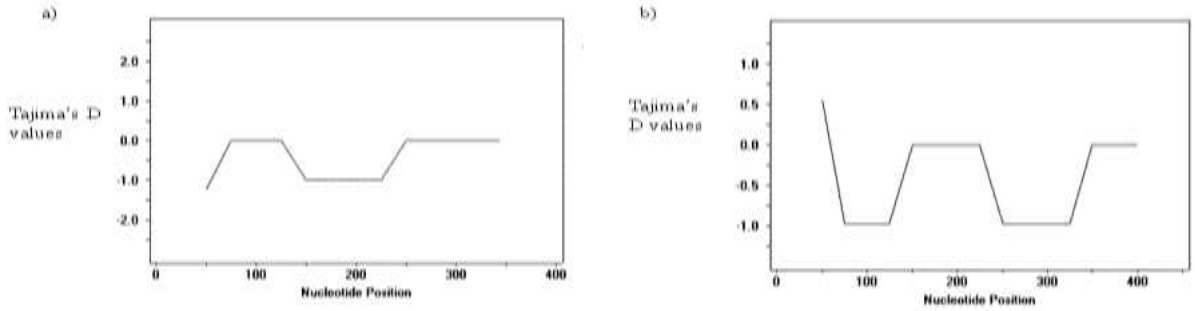

Figure 3.3 Tajima's D sliding window graphs for a) CR1 exon 5 and b) band 3 promoter. Tajima's D values have been plotted against nucleotide positions.

3.4 SNPs found in DARC, CR1 and Band 3

The DARC fragment that was amplified covered the polymorphic nucleotide positions −33T or C in the promoter region and 125G or A in the second exon. Other mutations that were detected occurred in the intron, 1457T or C and 1577delT (Figure 3.4**).** Only one DARC allele was observed in all the 89 samples genotyped, the FY*B^{ES} allele (ES = erythroid silent). Individuals carrying this allele do not express the DARC protein since all the samples had the C nucleotide at position -33 in the promoter region of the FY gene.

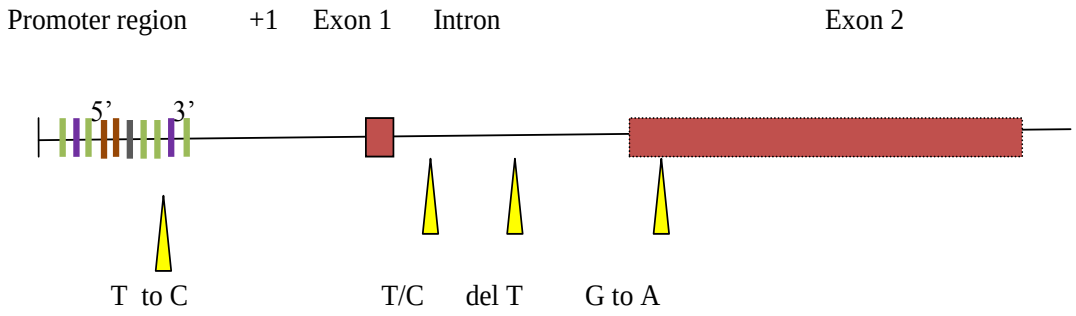

Figure 3.4 Structure of DARC gene depicting the variations (yellow triangles) that were detected in the samples as compared to the reference sequence (NC_000001.11).

In the band 3 gene, the most prominent SNP occurred in the promoter which showed a T/C at position -5699 upstream of the gene. The frequencies of the C and T alleles were 60.84% and 39.15%, respectively (Table 3.2, Figure 3.5 and Figure 3.6). Among the 83 individuals sampled, 34 (41%) had the $AE1^{-5699CC}$ genotype and individuals who had the $AE1^{-5699CT}$and $AE1^{-5699TT}$ genotypes were 33 (40%) and 16(19%), respectively. There were no SNPs found in exon 17 (Figure 3.6). In exon 18, three SNPs were found two of them occurring in the bordering introns on either side of the exon (Table 3.2). In the 17th intron (at position 26534 according to the reference sequence), there was a T/G substitution. The T and G alleles had a frequency of 77.7%

and 22.2%, respectively (Figure 3.6). The SNP occurring in the exon 18 at position 26610 was a singleton and resulted in amino acid change from serine to phenylalanine. In the 18th intron, (between exons 18 and 19) (position 26780), there was an A/C substitution with the C allele being the main allele at 97.9%.

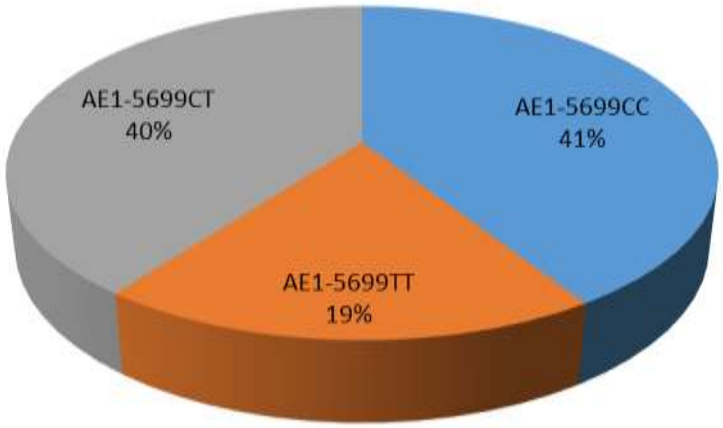

Figure 3.5 Shows the proportions of Band 3 (AE1) genotypes that were found in the promoter region from the analysis of 83 DNA samples.

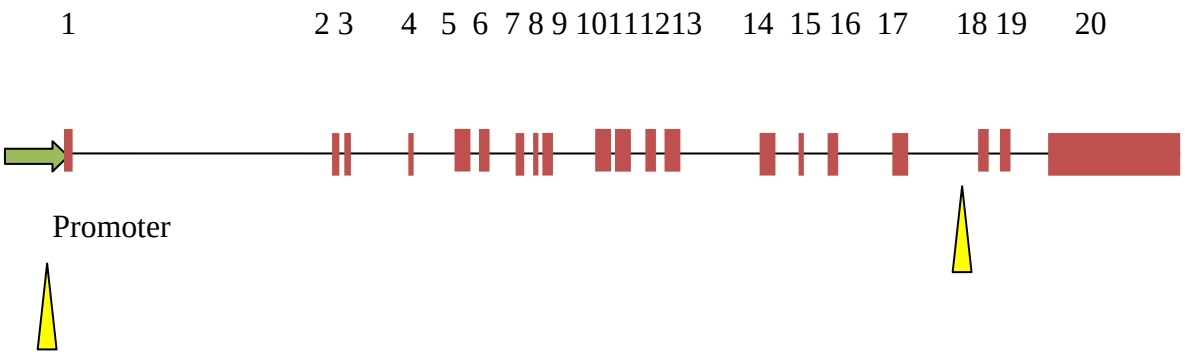

Figure 3.6 Structure of the band 3 gene showing the high frequency (>10%) SNPs that were detected (positions marked using yellow triangles)

No SNPs were detected in exon 4 of CR1. Four mutations were found in the 5th exon and one mutation in intron 4 (Table 3.2). These were 49360 A to G, 49527 C to G, 49530 C to T and 49353 C to A. The fourth mutation was a singleton. The mutation in intron 4 involved a C to T substitution with an allele frequency of 14% (Figure 3.7).

Figure 3.7 Structure of the CR1 gene showing the major SNP (freq of > 10%) that was detected (position marked using a yellow triangle) and regions excluded from analysis are exons 2 and 3 (marked with xx). Breaks after exon 5 indicate exons 6-38 and the introns in between.

Table 3.2 Summary of the SNPs found in DARC, CR1 and Band 3

Gene	SNP position according to the ref seq	Amino acid change	Genotypes	n (%)
DARC				
Intron	1457	_	TT	70 (78.7)
			TC	17 (19.1)
			CC	2 (2.2)
Total number of samples				89
Band 3				
Promoter	96	_	CC	34 (41)
			CT	33 (40)
			TT	16 (19)
Total number of samples				83
17th intron	26535	_	TT	46 (63.9)
			TG	23 (31.9)
			GG	3 (4.2)
Exon 18	26610	G (Ser) to T(Phe)	GG	71 (98.6)
			GT	1 (1.4)
			TT	0 (0)
18th intron	26780	_	CC	69 (95.8)
		_	AC	3 (4.2)
			AA	0 (0)
Total number of samples				72
Complement Receptor 1				
4th intron	20610	_	CC	57 (76)
			TC	15 (20)
			TT	3 (4)
Total number of samples				75
Exon 5	32506	C (Pro) to A (Pro)	CC	84 (98.8)
			CA	1 (1.2)
			AA	0 (0)
	32513	A (Thr) to G (Ala)	AA	82 (96.5)
			AG	3 (3.5)
			GG	0 (0)
	32680	C (Pro) to G (Pro)	CC	78 (91.8)
			GC	7 (8.2)
			GG	0 (0)
	32683	C (Pro) to T (Ser)	CC	83 (97.6)
			TC	2 (2.4)
			TT	0 (0)
Total number of samples				85

n= number of samples. The dashes in the amino acid change column show that the SNPs did not cause any change in the amino acids.

CHAPTER 4

DISCUSSION, CONCLUSION AND RECOMMENDATIONS

4.1 Discussion

The main aim of this study was to identify the polymorphisms in DARC, CR1 and Band 3 genes in the County of Kilifi, a malaria endemic region that lies at the Kenyan coast. Multiple sequence alignments of the samples revealed a number of polymorphisms in the three receptor genes. Most of the polymorphisms occurred in the non-coding regions (promoters and introns) and a few were in the coding regions.

Most of the SNPs were in the non-coding regions and these do not cause amino acid change in the protein. Although an intronic SNP may not alter the protein structure it may affect the manner in which a gene is regulated. SNPs that occur in non-coding regions may affect how a gene is spliced and thus alter its expression (American College of Neuropsychopharmacology, 2008).

DARC is a receptor for *Plasmodium vivax* and *Plasmodium knowlesi*. Contact with DARC is mediated by *P. vivax* Duffy-binding protein (PvDBP) leading to the formation of a junction, a crucial step in the invasion of human erythrocytes (Chitnis & Sharma, 2008). The DARC promoter region contained the C nucleotide at position -33 demonstrating that the erythrocyte silent variant is fixed in the severe malaria population sampled. For the 125^{th} nucleotide position in exon 2, only the A nucleotide was observed implying that all the Duffy negative individuals sampled harboured the FYB allele. The presence of the 'ES' promoter variant and the Fyb coding variant together form the $FY*B^{ES}$ allele. Our results are in agreement with a previous report which showed the -33 C allele that occurs in the GATA 1 transcription-factor binding motif is near fixation in Africa for resistance to infection by the malaria parasite *P. vivax*. In a study

carried out by Miller et al., (1976), all American volunteers of African descent who were duffy negative and resistant to experimental *P. vivax* blood stage infections. The other volunteers who were duffy positive developed blood stage infections caused by *P. vivax*. The fixation of the duffy negative allele in many ethnic groups in Africa has made *P. vivax* transmission rare (Mendis & Sina, 2001; Welch et al., 1977). The high frequency of the duffy null allele in the study region may explain the absence of vivax malaria in Kilifi (Kubasu, 2012).

Similar findings were also obtained in a study carried out in Gambia in West Africa. Red blood cells from 1,168 donors, consisting of nearly the entire populations of two Gambian villages located in the rural area were genotyped for Duffy blood group antigens. This was done using antisera to both the Fya and Fyb variants. All tests were negative for both variants meaning that none of the RBCs had duffy antigens on their surface. A blood film examination was carried out on the same samples and the results confirmed the non-existence of *Plasmodium vivax* parasitaemia. Infections with *P. falciparum*, *P. malariae* and *P. ovale* were however detected (Welch et al., 1977).

Although the duffy null phenotype has for a long time been considered as offering absolute protection against *P. vivax* infection, new reports are however emerging that duffy negative individuals are now showing by *P. vivax* infections (Menerd et al., 2010). This means that the parasite may have acquired another mechanism for erythrocyte invasion that does not depend on DARC alone.

The aspartate-to-glycine amino acid substitution in the 42^{nd} codon in the N-terminal region of DARC forms the basis for Duffy blood groups antigens Fy^b (Asp, A allele) and Fy^a (Gly, G allele). A study carried out by King et al., (2011) indicated that there is a disparity in the level of

vulnerability to vivax malaria between individuals expressing Fy^a and Fy^b. This difference is due to the parasite's capability to bind with Fy^b in a more efficient way as compared to Fy^a. The Fy^b has been shown to be the ancestral allele (King et al., 2011) and the Fy^a allele is mainly found in Asia (Zimmerman et al., 2013). The difference in the distribution of DARC alleles can possibly be attributed to the fact that since *P. vivax* originated in Asia it selected a resistant allele in this region (the Fy^a allele) because the interaction between parasites and humans is likely to have been there for a much longer time and continues to persist (King et al., 2011). It seems that in Africa, the mutation that was selected is the one in the gene's promoter region, resulting in the Duffy negative phenotype to stop the efficient binding of *P. vivax*. It is likely that since these individuals do not express the Fyb protein on the red blood cell, not much selection would be expected in DARC's open reading frame. Instead of mutating the coding region, a mutation occurred in the promoter region to switch off the gene. In the Kilifi population sampled, 100% had the Fy^b (Asp, A allele).

When compared to the reference sequence, the 1577delT occurred in all the samples (the reference sequence had a T at this position). To rule out the possibility of an error in the reference sequence, another sequence (NG_011626.1) was used to confirm that this deletion existed. The same results were obtained; the deletion was in all the samples and the sequence NG_011626.1 had a T at the same position. The population from which this sequence was obtained is not mentioned. The Tajima's D and Fu and Li's F and D were used to evaluate whether the sequencing data showed any indication of departure from neutrality resulting in values 0.30131, 0.45954, and 0.48215 for Tajima's D and Fu and Li's F and D, respectively (Table 3.1). None of the values was significant ($p > 0.10$) suggesting that DARC polymorphisms had evolved randomly. In a duffy negative population, such results are expected since there is no selective pressure from *P. vivax*.

CR1 is the host receptor for PfRh4 (Tham, & Wilson 2010). The regions targeted in CR1 were exons 2, 3, 4 and 5. This study is the first to focus on the polymorphisms in the CR1 gene that code for the region that interacts with parasite ligand, PfRh4. This region is known as site 1 and is encoded by exons 2, 3, 4 and 5. Site 1 consists of complement-control-protein repeats (CCP) 1-3 (Vik & Wong, 1993).

The sequencing data obtained when exon 2 and 3 were amplified together was of poor quality due to stretches of the same nucleotide (T) in the intron which could have caused polymerase slippage during synthesis of the DNA. Long continuous stretches of a single base in a sequence is one of the documented drawbacks of the Sanger method of sequencing that was used in this study. When exons 2 and 3 were amplified separately, the chromatograms had peaks that overlapped one another. A possible explanation for this is that there were regions that were homologous to the two exons in another part of the template where the sequencing primer begins extension. These homologous regions could be potentially due to the fact that CR1 protein is organized into four long homologous repeats (LHRs) A–D (Figure 1.9). The LHRs are believed to have arisen from the duplication of sections of the gene that code for the constituent Short Consensus Repeats (SCRs) of an LHR. The identity between each LHR is estimated to be between 70–95% .The sequences for these two exons were therefore excluded from analysis.

The most notable SNP in CR1 occurred in the fourth intron and involved a C to T substitution at position 20610 with the T allele having a frequency of 14%. The C/T SNP has not been identified before. The other four polymorphisms in CR1 were rare (a frequency of < 10%). Two of them were synonymous mutations (did not cause a change in amino acid) and the other two were nonsynonymous. All the neutrality tests that were done were not significant suggesting that the mutations were evolving randomly (Table 3.1). Since CR1 has other biological functions, it seems

that it is not under selection in order to preserve function and perhaps since the Swain-Langley (*S)l2* and McCoy (*McC)b* polymorphisms exist in the C terminus region, the mutation in the N terminal is limited so as not to allow too many polymorphisms in the gene.

Band 3 protein is also known as anion exchange protein 1. It belongs to the solute carrier family of proteins. The product of the band 3 gene acts as a receptor for the malaria parasite, *P. falciparum*. The 5ABC and 6A regions of the protein interact with the 42-kDa processing product of merozoite surface protein 1 ($MSP1_{42}$) through its 19-kDa C-terminal domain (Li et al., 2004).

A study done by Kalcreuth et al., (2006) in Ghana showed that the T to C exchange was more frequent in Africans than Europeans in the promoter region of the band 3 gene. Although this study looked at the SNPs in African children, the results are comparable. The C allele occurred more frequently (60.8%) than the T allele in the Kilifi population (Figure 4.7). In the study done by Kalcreuth et al., (2006), this mutation was associated with a higher risk of malaria (40%), higher fatality among children with severe anaemia and distinct cases of metabolic acidosis in patients who had cerebral malaria. Contrary to the popular belief that a genetic variant that occurs more in Africans in malaria endemic regions than Europeans would be protective against malaria, this happened not to be the case with the band 3^{-5699C} allele. The possible explanation for the existence of this mutation in the African population in regions that are malaria endemic according to Kalcreuth et al., (2006) is that African conditions other than malaria could have made the band 3^{-5699C} allele more advantageous for human fitness in malaria endemic regions in Africa. As long as this pressure continues to exist, the allele is likely to be maintained at this high frequency. Another reason that could possibly explain the high frequency of the band 3^{-5699C} allele is that the population of study was biased to severe malaria and this SNP may be causing individuals who

have it to be more susceptible to malaria. This SNP in the promoter could possibly be affecting the regulation of this gene (Kalcreuth et al., 2006).

The other SNPs in the band 3 gene were rare with a frequency of <10 %. The sequences obtained from the band 3 promoter, exons 17 and 18 seem to be evolving in a random manner according to the values obtained from Tajima's D and Fu and Li's statistics. All the values were not significant, $p > 0.10$. These coding regions of the gene may be evolving randomly despite being receptors for the parasite ligand MSP1 probably to preserve its other biological functions. Just like in DARC where the SNP in the promoter has a bigger impact to that in the coding region, it is likely that the impact of the SNP in the band 3 promoter is greater hence there are fewer mutations in the coding region.

4.2 Conclusion

The detection of SNPs in genes that are involved in invasion by the malaria parasite needs to be interpreted with caution. These genes have other biological functions and the malaria parasite may not be the only selective pressure. Moreover, other diseases are also found in regions that are malaria endemic and they may also be playing a role in shaping the human genome.

The results of the DARC sequence analysis show the dominance of the $FY*B^{ES}$ allele. The existence of the Fyb coding variant and the 'ES' variant in the promoter together form the $FY*B^{ES}$ allele. This reinforces further the theory that the FY*B is the ancestral allele since the Kilifi population does not express the Fyb protein on the red blood cell, not much selection would be expected in DARC's open reading frame. Recent observation that *P. vivax* has acquired a new capacity to infect duffy null individuals is an indication of the ability of species to evolve when faced with barriers that tend to hinder them from reproducing. More work needs to be done to

discover the machinery that *P.vivax* uses to invade the duffy deficient RBCs. Results from such studies would be valuable in the development of a vaccine against *P. vivax*. The only challenge that needs to be conquered is the reluctance of *P. vivax* to grow in laboratory cultures.

Analysis of the sequences for the regions of Band 3 and CR1 gene that code for regions involved in binding the parasite ligands in the Kilifi population showed that the SNPs were not under selection and mutations occurring in these genes are neutral (have no effect on an individual's fitness). This implies that the mutations are not affected by natural selection and are potentially driven by genetic drift (Duret, 2008). The SNPs in the exons were few and most of the SNPs detected are in the non-coding region. This may be due to the need to prevent change and preserve function since these receptors have other biological functions other than acting as receptors for the malaria parasite.

4.3 Recommendations

The recommendations for future work are as follows:

1. This study mainly focused on exonic regions. Exons only account for a small fraction of the human genome. It may be interesting to carry out research on other parts of the candidate genes such as the intronic SNPs, SNPs on splice site and SNPs in the promoter region (for CR1). Alteration in nucleotide sequence in these regions has also been shown to affect gene regulation.
2. Future studies on the SNPs identified in this study should involve association analysis. The identification of beneficial or detrimental alleles can be done best in association studies which involves looking at variations in genes in two groups (those with disease and healthy controls) to ascertain if there is an association with a particular phenotype.

If statistically significant variation in the allele frequency is found between a population of those with the disease and the controls, then there may be an association linking the allele and disease. In the study done by Kalcreuth et al., (2006), the mutation found in the band 3 promoter was associated with a higher risk of malaria (40%), higher fatality among children with severe anaemia and distinct cases of metabolic acidosis in patients who had cerebral malaria. Since this study involved the use of samples from children with severe malaria, it would be interesting to compare the association of the mutation with the different groups (children with cerebral malaria, severe anaemia and respiratory distress).

3. Further studies should also include a larger number of individuals who are not seriously ill with malaria to add to the knowledge of the Kilifi population genetic diversity.

References

American College of Neuropsychopharmacology. (2008). Intronic Polymorphisms Affecting Alternative Splicing of Human Dopamine D2 Receptor Are Associated with Cocaine Abuse. *American College of Neuropsychopharmacology*. Retrieved July 4, 2014, from http://www.acnp.org/resources/articlediscussionDetail.aspx?cid=8d972238-26c1-4d6d-a0b0-a352b4a27b1d

Atlas of Genetics and Cytogenetics in Oncology and Haematology. (2014). CR1 (complement component (3b/4b) receptor 1 (Knops blood group)). Retrieved April 8, 2014, from http://atlasgeneticsoncology.org/Genes/GC_CR1.html

Baum, J., Maier, A., Good, R., Simpson, K., & Cowman, A. (2005). Invasion by *P. falciparum* Merozoites Suggests a Hierarchy of Molecular Interactions. *PLoS Pathogens, 1*(4).

Bejon, P., Cook, J., Bergmann-Leitner, E., Olotu, A., Lusingu, J., Mwacharo, J., Vekemans, J., et al. (2011). Effect of the Pre-erythrocytic Candidate Malaria Vaccine RTS,S/AS01E on Blood Stage Immunity in Young Children. *Journal of Infectious Diseases, 204*, 9–18.

Cappadoro, M., Giribaldi, G., O'Brien, E., Turrini, F., Mannu, F., & Ulliers, D. (1998). Early Phagocytosis of Glucose-6-Phosphate Dehydrogenase (G6PD)-Deficient Erythrocytes Parasitized by Plasmodium falciparum May Explain Malaria Protection in G6PD Deficiency. *Blood, 92*(7).

Centers for Disease Control and Prevention. (2014, March 26). Malaria. Retrieved April 8, 2014, from http://www.cdc.gov/malaria/malaria_worldwide/impact.html

Chen, Q., Barragan, A., Fernandez, V., Sundstrom, A., Schlichtherle, M., Sahlen, A., … Wahlgren, M. (1998). Identification of *Plasmodium falciparum* Erythrocyte Membrane

Protein 1 (PfEMP1) as the Rosetting Ligand of the Malaria Parasite *P. falciparum*. *J Exp Med*, *187*(1), 15–23.

Chitnis, C., & Sharma, A. (2008). Targeting the *Plasmodium vivax* Duffy-binding protein. *Trends in Parasitology*, *24*(1), 29–34.

Cholera, R., Brittain, N., & Gillrie, M. (2008). Impaired cytoadherence of *Plasmodium falciparum*-infected erythrocytes containing sickle hemoglobin. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(3), 991–996.

Cockburn, I., Mackinnon, M., Donnell, A., Allen, S., Moulds, J., Baisor, M., Rowe, A. (2004). A human complement receptor 1 polymorphism that reduces *Plasmodium falciparum* rosetting confers protection against severe malaria. *PNAS*, *101*(1), 272–277.

Cortes, A., Benet, A., Cooke, B., Barnwell, J., & Reeder, J. (2004). Ability of *Plasmodium falciparum* to invade Southeast Asian ovalocytes varies between parasite lines. *Blood*, *104*(9).

County Government of Kilifi. (2013). County Government of Kilifi. Retrieved June 14, 2014, from http://www.kilifi.go.ke/index.php/about/background1

Cowman, A., & Crabb, B. (2006). Invasion of Red Blood Cells by Malaria Parasites. *Cell, February 24, 2006*, *124*(1), 755–766.

Cowman, A., Berry, D., & Baum, J. (2012). The cellular and molecular basis for malaria parasite invasion of the human red blood cell. *Journal of Cell Biology*, *198*(6), 961–971.

Crosnier, C., Bustamante, L., Bartholdson, J., & Bei, A. (2011). Basigin is a receptor essential for erythrocyte invasion by *Plasmodium falciparum*. *Nature*, *480*, 534–538.

Demogines, A., Truong, K., & Sawyer, S. (2012). Species-Specific Features of DARC, the Primate Receptor for *Plasmodium vivax* and *Plasmodium knowlesi*. *Molecular Biology and Evolution*, *29*(2), 445–449.

Duret, L. (2008). Neutral Theory: The Null Hypothesis of Molecular Evolution. *Nature Education*, *1*(1), 218.

Escalante, A., Cornejo, O., Freeland, D., Poe, A., & Durrego, E. (2005). A monkey's tale: The origin of *Plasmodium vivax* as a human malaria parasite. *PNAS*, *102*(6), 1980–1985.

Figtree, M., Lee, R., Bain, L., Kennedy, T., Mackertich, S., & Urban, M. (2010). *Plasmodium knowlesi* in Human, Indonesian Borneo. *Emerging Infectious Diseases*, *16*(4), 672–674.

Fowkes, F., Michon, P., Pilling, L., & Ripley, R. (2008). Host erythrocyte polymorphisms and exposure to *Plasmodium falciparum* in Papua New Guinea. *Malaria Journal*, *7*(1).

Fujinaga, J. (1999). Topology of the Membrane Domain of Human Erythrocyte Anion Exchange Protein, AE1. *The Journal of Biological Chemistry*, *274*(10), 6626–6633.

Goel, V., & Li, X. (2003). Band 3 is a host receptor binding merozoite surface protein 1 during the *Plasmodium falciparum* invasion of erythrocytes. *PNAS*, *vol. 100*(9), 5161–5164.

Harvey, K., Gilson, P., & Crabb, B. (2012). A model for the progression of receptor–ligand interactions during erythrocyte invasion by *Plasmodium falciparum*. *International Journal for Parasitology*, *42*(1), 567–573.

Kalcreuth, V., Evans, J., Timmann, C., Kuhn, D., Agbenyega, T., Horstmann, R., & May, J. (2006). Promoter Polymorphism of the Anion-Exchange Protein 1 Associated with Severe Malarial Anemia and Fatality. *Journal of Infectious Diseases, 194*, 949–957.

Kappe, S., Vaughan, A., Boddey, J., & Cowman, A. (2010). That Was Then But This Is Now: Malaria Research in the Time of an Eradication Agenda. *Science*, *328*(5980), 862–866.

King, C., Adams, J., Xianli, J., Grimberg, B., McHenry, A., & Greenberg, L. (2011). Fya/ Fyb antigen polymorphism in human erythrocyte Duffy antigen affects susceptibility to *Plasmodium vivax* malaria. *PNAS*, *108*(50), 20113–20118.

Ko, W., Kaercher, K., & Giombini, E. (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*(1), 741–754.

Krause, M., Diakite, S., Lopera-Mesa, T., Amaratunga, C., & Arie, T. (2012). α-Thalassemia impairs the cytoadherence of *Plasmodium falciparum*-infected erythrocytes. *PLoS ONE*, *7*(5).

Kubasu, S. (2012). The vectors of malaria and filariasis in Kilifi and Kwale districts of Kenya. *Kenyatta University Institutional Repository*. Retrieved from http://ir-library.ku.ac.ke/handle/123456789/4142

Leykauf, K., & Treeck, M. (2010). Protein Kinase A Dependent Phosphorylation of Apical Membrane Antigen 1 Plays an Important Role in Erythrocyte Invasion by the Malaria Parasite. *PLoS Pathogens*, *6*(6), 1–11.

Li, X., Chen, H., Oo, T., Daly, T., & Bergman, L. (2004). A Co-ligand Complex Anchors *Plasmodium falciparum* Merozoites to the Erythrocyte Invasion Receptor Band 3. *The Journal of Biological Chemistry*, *279*(7), 5765–5771.

Liu, D., & Niu, Z.-X. (2009). The structure, genetic polymorphisms, expression and biological functions of complement receptor type 1 (CR1/CD35). *Immunopharmacology and immunotoxicology*, *31*(4), 524–535.

Lobo, C., & Ord, R. (2012). Targeting Sialic Acid Dependent and Independent Pathways of Invasion in *Plasmodium falciparum*. *PLoS ONE*, *7*(1), 1–9.

Lobo, C., Rodriguez, M., Reid, M., & Lustigman, S. (2003). Glycophorin C is the receptor for the *Plasmodium falciparum* erythrocyte binding ligand PfEBP-2 (baebl). *Blood*, *101*(11), 4628–4631.

Maier, A., Duraisingh, J., & Reeder, J. (2003). *Plasmodium falciparum* erythrocyte invasion through glycophorin C and selection for Gerbich negativity in human populations. *Nature Medicine*, *9*(1), 87–92.

Mayer, G., Cofie, J., Jiang, L., Hartl, D., Tracy, E., Kabat, J., Mendoza, L., et al. (2009). Glycophorin B is the erythrocyte receptor of *Plasmodium falciparum* erythrocyte-binding ligand, EBL-1. *PNAS*, *106*(13), 5348–5352.

Mendes, C., Dias, F., Figuiredo, J., Gonzalez, V., Cano, J., & De Sousa, B. (2011). Duffy Negative Antigen Is No Longer a Barrier to *Plasmodium vivax* – Molecular Evidences from the African West Coast (Angola and Equatorial Guinea). *PLoS Neglected Tropical Diseases*, *5*(6).

Mendis, K., Sina, B., Marchesini, P., & Carter, R. (2001). The neglected burden of *Plasmodium vivax* malaria. *American Journal of Tropical Medicine and Hygiene*, *64*, 97–106.

Menerd, D., Barnadas, C., Christiane, B., Henry-Halldin, C., Gray, L., Ratsimbasoa, A., Thonier, V., et al. (2010). *Plasmodium vivax* clinical malaria is commonly observed in Duffy-negative Malagasy people. *PNAS*.

Miller, L., Baruch, D., Marsh, K., & Doumbo, O. (2002). The pathogenic basis of malaria. *Nature*, *415*, 673–679.

Miller, L., Mason, S., Clyde, D., & McGinniss, M. (1976). The Resistance Factor to *Plasmodium vivax* in Blacks — The Duffy-Blood-Group Genotype, FyFy. *New England Journal of Medicine*, *295*, 302–304.

Min-Oo, G., & Gros, P. (2005). Erythrocyte variants and the nature of their malaria protective effect. *Cellular Microbiology*, *7*(6), 753–763.

Muramatsu, T. (2012). Basigin: a multifunctional membrane protein with an emerging role in infections by malaria parasites. *Expert Opinion Therapeutic Targets (2012) 16(10)*, *16*(10), 999–1011.

Oliveira, T., Harris, E., Meyer, D., & Silva Jr, W. (2012). Molecular evolution of a malaria resistance gene (DARC) in primates. *Immunogenetics*, *64*(7), 497–505.

Olivier, M. (2004). From SNPs to function: the effect of sequence variation on gene expression. Focus on "A survey of genetic and epigenetic variation affecting human gene expression." *Physiological Genomics*, *16*(182-183).

Orlandi, P., Klotz, F., & Haynes, D. (1992). A Malaria Invasion Receptor , the 175-Kilodalton Erythrocyte Binding Antigen of *Plasmodium falciparum* Recognizes the Terminal Neu5Ac(a2-3)Gal -Sequences of GlycophorinA. *The Journal of Cell Biology*, *116*(4), 901–909.

Patel, S., King, C., Mgone, C., Kazura, J., & Zimmerman, P. (2004). Glycophorin C (Gerbich Antigen Blood Group) and Band 3 Polymorphisms in Two Malaria Holoendemic Regions of Papua New Guinea. *American Journal of Hematology*, *75*, 1–5.

PREMIER Biosoft. (2014). PCR Primer Design Guidelines. *PREMIER Biosoft*. Retrieved June 14, 2014, from http://www.premierbiosoft.com/tech_notes/PCR_Primer_Design.html

Richard, D., MacRaild, C., Riglar, D., Chan, J.-A., Foley, M., Baum, J., Ralph, S., et al. (2010). Interaction between *Plasmodium falciparum* Apical Membrane Antigen 1 and the Rhoptry Neck Protein Complex Defines a Key Step in the Erythrocyte Invasion Process of Malaria Parasites. *The Journal of Biological Chemistry*, *285*(19), 14815–14822.

Riglar, D., Richard, D., Wilson, D., Boyle, M., & Dekiwadia, C. (2011). Super-Resolution Dissection of Coordinated Events during Malaria Parasite Invasion of the Human Erythrocyte. *Cell Host & Microbe*, *9*(1), 9–20.

Ryan, J., Stoute, J., Amon, J., Dunton, R., Mtalib, R., Koros, J., Owuor, B., et al. (2006). Evidence for Transmission of *Plasmodium vivax* Among a Duffy Antigen Negative Population in western Kenya. *American Journal of Tropical Medicine and Hygiene*, *75*(4), 575–581.

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.

Singh, B., Sung, L., & Radhakrishna, A. (2004). A large focus of naturally acquired *Plasmodium knowlesi* infections in human beings. *Lancet*, *363*, 1017–1024.

Singh, S., & Chitnis, C. (2012). Signaling mechanisms involved in apical organelle discharge during host cell invasion by apicomplexan parasites. *Microbes and Infection*, *14*(10), 820–824.

Teeranaipong, P., & Ohashi, J. (2008). A Functional Single-Nucleotide Polymorphism in the CR1 Promoter Region Contributes to Protection against Cerebral Malaria. *The Journal of Infectious Diseases*, *198*, 1880 –91.

Teeranaipong, P., & Ohashi, J. (2008). A Functional Single-Nucleotide Polymorphism in the CR1 Promoter Region Contributes to Protection against Cerebral Malaria. *The Journal of Infectious Diseases*, *198*, 1880 –91.

Tetteh-Quarco, P., Schmidt, C., Tham, W.-H., Hauhart, R., Mertens, H., & Rowe, A. (2012). Lack of Evidence from Studies of Soluble Protein Fragments that Knops Blood Group Polymorphisms in Complement Receptor-Type 1 Are Driven by Malaria. *PLoS ONE*, *7*(4).

Tham, W.-H., Schmidt, C., Hauhart, R., Gueriento, M., Tetteh-Quarco, P., Sash, L., Atkinson, J., et al. (2011). *Plasmodium falciparum* uses a key functional site in complement receptor type-1 for invasion of human erythrocytes. *Blood*, *118*(7), 1923–1933.

Tham,, W.H., & Wilson, D. (2010). Complement receptor 1 is the host erythrocyte receptor for *Plasmodium falciparum* PfRh4 invasion ligand. *PNAS*, *vol. 107*(no. 40), 17327–17332.

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

Vasuvattakul, S., Yenchitsomanus, pa-thai, Vanchuanichsanong, P., & Thuwajit, peti. (1999). Autosomal recessive distal renal tubular acidosis associated with Southeast Asian ovalocytosis. *Kidney International*, *56*(1), 1674–1682.

Vik, D., & Wong, W. (1993). Structure of the Gene for the F Allele of Complement Receptor Type 1 and Sequence of the Coding Region Unique to the S Allele. *Journal of Immunology*, *151*(11), 6214–6224.

Vogel, G. (2012). A setback for Malaria Vaccines. *ScienceNow*. Retrieved December 8, 2013 from http://news.sciencemag.org/sciencenow/2012/11/a-setback-for-...

Welch, S., McGregor, I., & Williams, K. (1977). The Duffy blood group and malaria prevalence in Gambian West Africans. *Transactions of the Royal Society of Tropical Medicine and Hygiene*, *71*(4), 295–6.

Wellems, T., Hayton, K., & Fastirhurst, R. (2009). The impact of malaria parasitism: from corpuscles to communities. *Journal of Clinical Investigation*, *119*(9), 2496–2505.

Wertheimer, S., & Barnwell, J. (1989). *Plasmodium vivax* interaction with the human blood group glycoprotein: iden- tification of a parasite receptor-like protein. *Exp. Parasitology*, *69*, 340–350.

WHO. (2012). Immunization, Vaccines and Biologicals. *World Health Organization*. Retrieved from www.who.int/topics/malaria/

Williams, T. (2006). Red blood cell defects and malaria. *Molecular & Biochemical Parasitology*, *149*(1), 121–127.

Wiser, M. (2011). *Plasmodium Species Infecting Humans. Human Plasmodium Species*. Retrieved July 8, 2012from: www.tulane.edu/~wiser/protozoology/

Woolhouse, M., Webster, J., Domingo, E., & Charlesworth, B. (2002). Biological and biomedical implications of the co-evolution of pathogens and their hosts. *Nature Genetics*, *32*(1), 569–577.

World Health Organization. (2014).10 facts on malaria. Retrieved April 8, 2014, from http://www.who.int/features/factfiles/malaria/en/

Zimmerman, P., Ferreira, M., & Mercereau-Puijalon, O. (2013). Red Blood Cell Polymorphism and Susceptibility to *Plasmodium vivax*. *Advances in parasitology*, *81*(1), 27–76.

APPENDIX

TBE Buffer Preparation

0.5M Ethylenediamine tetraacetic acid solution of pH8.0 that was required for the preparation of this buffer was made first. For a stock solution of 500ml, 93.05 g of EDTA disodium salt was weighed and dissolved in 400ml of deionized water. The pH was adjusted using NaOH and measured using a pH meter. The solution was then topped up to a volume of 500ml. The TBE buffer was prepared in a sterile glass bottle with a volume of 1litre.To make 100ml of concentrated solution of the buffer (10X), 108g of Tris base and 55g of boric acid were weighed and dissolved in 900ml of deionized water using a magnetic stirrer and a hot plate. 40ml of the 0.5M EDTA was added and the final volume brought to 1 litre. To make the working solution, the stock was diluted 20X using deionized water to give a concentration of 0.5X.

Table S1 Summary of patient characteristics

id	year	date	parasitemia	sex	age	hb	SMA	bcs_ey	bcs_m	bcs_v	bcstot	BCS
45332	2000	12/11/2000	10260	M		12.7	0	0	0	1	1	1
45536	2001	1/7/2001	12208	M	0.66	8.3	0	0	1	1	2	1
45537	2001	1/7/2001	5280	M	1.67	6.1	0	0	1	1	2	1
45539	2001	1/9/2001	744	M	3.6	11.4	0	0	1	1	2	1
45678	2001	1/1/2001		F		10.3	0	1	2	2	5	0
45749	2001	1/4/2001	6930	F	8.46	7.3	0	0	1	1	2	1
45814	2001	1/8/2001	9030	M	3.16	3.6	1	1	2	2	5	0
45910	2001	1/12/2001	23595	M	2.18	6.7	0	0	1	1	2	1
45924	2001	1/12/2001	4805	M	0.97	4.2	1	0	1	1	2	1
45938	2001	1/13/2001	2948	F	3.81	8.3	0	1	2	1	4	0
45951	2001	1/13/2001	2363	M	12.32	5.9	0	0	0	1	1	1
45973	2001	1/14/2001	118	M	3.46	6.7	0	0	1	1	2	1
45976	2001	1/15/2001	187860	F	0.45	9.4	0	1	2	2	5	0
45989	2001	1/15/2001	1350	M	1.38	9.2	0	0	1	1	2	1
46016	2001	1/14/2001	72118	F	2.41	11.4	0	0	0	1	1	1
46018	2001	1/16/2001	7680	F	3.01	12.3	0	0	0	0	0	1
46019	2001	1/18/2001	13847	F	2.73	5.4	0	0	2	2	4	0
46022	2001	1/20/2001	3080	M	2.16	7.4	0	0	0	0	0	1
46026	2001	1/29/2001	8800	M	2.77	12.3	0	1	2	2	5	0
46027	2001	1/30/2001	3036	F	3.68	6.1	0	0	1	1	2	1
46028						6.9	0					1
46032	2001	2/3/2001	23520	M	0.99	10	0	0	1	2	3	0
46034	2001	2/8/2001	8855	F	2.63	10.6	0	0	0	0	0	1
46041	2001	2/21/2001	16200	M	1.71	6.1	0	1	2	2	5	0
46043	2001	2/25/2001	5300	M	6.75	7.6	0	0	1	1	2	1
46064	2001	1/17/2001	11316	F	1.26	8.1	0	0	0	0	0	1
46075	2001	1/18/2001	7560	F	0.93	9.1	0	1	2	2	5	0
46116	2001	1/20/2001	4280	F	0.42	7	0	1	1	2	4	0
46162	2001	1/23/2001	1157	F	2.4	7.2	0	1	2	2	5	0
46168	2001	1/24/2001	1584	F	3.09	5.2	0	0	2	1	3	0
46187	2001	1/25/2001	6016	F	1.43	8.8	0	1	2	2	5	0
46206	2001	1/26/2001	4224	F	1.06	5.7	0	0	1	1	2	1
46218	2001	1/26/2001	1136	M	2.61	4.4	1	0	1	1	2	1
46227	2001	1/27/2001		M	1.86	8.8	0	1	2	2	5	0
46274	2001	1/29/2001	10788	M	0.2	5.8	0	0	0	1	1	1
46299	2001	1/30/2001	1017	M	3.06	8.2	0	0	1	1	2	1
46325	2001	2/1/2001	26656	M	2.65	10.2	0	1	2	2	5	0
46335	2001	2/2/2001	770	F	1.16	5.9	0	0	1	1	2	1

46366	2001	2/4/2001	13125	F	6.79	7.5	0	1	2	2	5	0
46378	2001	2/5/2001	22134	F	2.34	7.7	0	0	0	0	0	1
46390	2001	2/5/2001	296	M	1.93	8.6	0	0	1	1	2	1
46397	2001	2/6/2001	750	F	1.88	6	0	1	1	2	4	0
46434	2001	2/8/2001	1530	M	3.04	7.2	0	1	2	2	5	0
46459	2001	2/10/2001	2096	F	2.49	7.2	0	1	2	2	5	0
46479	2001	2/11/2001	35700	M	1.31	8.4	0	1	2	2	5	0
46516	2001	2/15/2001	117600	F	4.11	11	0	1	2	2	5	0
46529	2001	2/15/2001	2744	F	3.7	11.1	0	0	0	1	1	1
46547	2001	2/16/2001	318240	M	4.11	6.3	0	0	1	1	2	1
46555	2001	2/17/2001		M	4.18	7.1	0	0	2	1	3	0
46557	2001	2/17/2001	1859	F	6.67	4.9	1	0	1	1	2	1
46569	2001	2/18/2001	2475	M	2.78	7.2	0	0	1	1	2	1
46577	2001	3/3/2001	15500	M	1.25	7.6	0	0	0	1	1	1
46589	2001	3/22/2001	25872	F	0.52	10.8	0	1	1	0	2	1
46590	2001	3/25/2001	6276	F	4.09	1.8	1	0	0	0	0	1
46591	2001	3/26/2001	2880	M	0.42	8	0	0	1	1	2	1
46608	2001	2/19/2001	10350	M	3.26	4.4	1	0	1	1	2	1
46653	2001	2/23/2001	12714	F	5.22	7.9	0	0	1	1	2	1
46659	2001	2/24/2001	16362	M	2.3	6.7	0	0	2	1	3	0
46700						10.5	0					1
46721	2001	3/1/2001	9776	F	2.63	6.2	0	0	2	1	3	0
46732	2001	3/2/2001	15785	F	5.92	7.5	0	1	2	1	4	0
46743	2001	3/2/2001	1922	M	5.12	8.6	0	1	2	2	5	0
46769	2001	3/5/2001	46498	M	1.8	2.5	1	1	2	2	5	0
46807	2001	3/6/2001	22600	M	0.65	6.6	0	1	2	2	5	0
46866	2001	3/10/2001	3950	F	0.22	5.8	0	0	1	1	2	1
46897	2001	3/12/2001	10080	M	6.84	11.7	0	1	2	1	4	0
46898	2001	3/12/2001	15485	M	5.33	6.9	0	0	1	1	2	1
47033	2001	3/24/2001	20096	M	2.29	7.6	0	0	1	1	2	1
47049	2001	3/26/2001	10388	M	0.4	8.5	0	0	0	0	0	1
47074	2001	3/27/2001	2751	F	7.12	14.5	0	0	1	1	2	1
47085	2001	3/27/2001	1595	F	1.08	6	0	0	0	0	0	1
47094	2001	4/3/2001	9180	M	2.13	10.2	0	0	1	1	2	1
47149	2001	4/9/2001	5814	M	11.03	5.7	0	0	1	1	2	1
47180	2001	4/5/2001		M	2.18	9.9	0	0	2	1	3	0
47194	2001	4/13/2001	640	M	0.82	4.5	1	0	1	1	2	1
47239	2001	4/11/2001	1290	F	0.711	11	0	0	1	1	2	1
47280	2001	4/13/2001	11134	M	1.08	3.6	1	0	1	1	2	1
47319	2001	4/17/2001	1071	M	3.9	9.6	0	0	1	2	3	0
47322	2001	4/18/2001	12720	M	2.62	5.2	0	0	0	0	0	1

47324	2001	4/19/2001	465	M	3.87	11.3	0	0	1	0	1	1
47351	2001	4/20/2001	19900	F	2.96	11.1	0	0	1	1	2	1
47352	2001	4/22/2001	102	M	7.31	10.8	0	0	1	1	2	1
47354	2001	4/25/2001	8850	F	3.86	8.9	0	0	2	1	3	0
47355	2001	4/27/2001	247680	F	2.16	7.4	0	0	0	1	1	1
47429	2001	4/27/2001	5428	F	2.82	8.2	0	1	2	2	5	0
47615	2001	5/16/2001	3828	M	1.94	7.9	0	1	1	1	3	0
48024	2001	6/7/2001	53500	M	7.11	5.9	0	0	1	1	2	1
48044	2001	6/8/2001	25864	F	0.32	2.1	1	0	1	1	2	1
48636	2001	7/9/2001	34113	M	1.22	6.1	0	0	1	1	2	1
48777	2001	7/16/2001	5304	F	4.76	8.2	0	0	1	1	2	1
48878	2001	7/23/2001	1171800	M	2.86	8.7	0	1	2	2	5	0
50971	2002	1/12/2002	9471	F	0.51	8.3	0	1	2	2	5	0
51250	2002	2/4/2002	101140	M	2.84	8.6	0	0	2	2	4	0

Table S2 Explanation of patient characteristics

Variable	Type	Len	Label	Legal Values
SEX	Num	8	Sex	1=Male 2=Female
DOB	Date	8	Date of Birth	
BCS_V	Num	8	Blantyre Coma Score verbal	0= no response 1= abnormal cry 2= Localizing
BCS_M	Num	8	Blantyre Coma Score motor	0= no response 1= abnormal cry 2= normal cry or spee
BCS_EY	Num	8	Blantyre Coma Score eye movements	0= not following 1= following
BCSTOT	Num	8	Blantyre Coma Score total	
AGEMTHS	Num	8	Age in months	
COMA	Num	8	Child in coma	
HB	Num	8	Haemoglobin	
RESPDIS	Num	8	Respiratory distress	0=no 1=yes