UNIVERSITY OF LIVERPOOL

Using genomics and population genetics to understand genetic variation in Malawi *Plasmodium falciparum* clinical isolates

Thesis submitted in accordance with the requirements of the University of Liverpool and College of Medicine, University of Malawi for the degree of

Doctor of Philosophy

Harold Martin Ochola

August 2013

Liverpool School of Tropical Medicine

Declaration

This thesis is the result of my own work except where indicated. Studies in this thesis were done in collaboration with other groups and in some instances work was shared. Contribution towards the work is as follows:

Activity	**Responsibility**
Sample collection and processing	Sole
White blood cell depletion	Sole
DNA extraction	Sole
DNA quantification	Shared
Library preparation and sequencing	Collaborators
Raw data analysis and SNP identification	Shared
Population genetic analysis	Sole
Copy number variation (FREEC)	Sole
Copy number variation (PG)	Shared
Thesis preparation	Sole

The material contained in this thesis has not been presented, nor is currently being presented, either wholly or in part for any other degree or qualification elsewhere.

Harold Martin Ochola

Acknowledgements

I would like to thank the Malaria Capacity Development Consortium (MCDC) for offering me a studentship to undertake this PhD. Special thanks go to the MCDC secretariat, College of Medicine, Malawi and the Malawi-Liverpool-Wellcome Trust for their administrative support and allowing me to use their facilities. This work would not be possible without the ACTia study that provided samples – I thank Dr Anja Terlouw, the study staff, children, parents and guardians who participated in this study. I would also like to thank the Wellcome Trust Sanger Institute, particularly Prof Dominic Kwiatkowski and his team for the sequencing support and providing raw sequence data for analysis.

My heartfelt thanks go to my supervisors: Dr Jacqui Montgomery, Prof Alister Craig, Dr Taane Clark, Prof Anja Jensen and Dr Mipando Mwapasa for the amazing analytical and scientific input, and for guidance. I share my sincere gratitude to Dr Taane Clark and his team for the training in genomics and population genetic analysis. I am also very grateful to my PhD advisors, Drs Themba Mzilahowa and Dean Everett for their mentorship.

I would also like to thank my friends and PhD colleagues at the MCDC, Malawi-Liverpool-Wellcome Trust and Kenya for their sincere support. Many thanks go to Dr Kamija Phiri, Prof Moffat Nyirenda and Louise Afran for the encouragement throughout good and hard times and not to forget Aaron, Hezron, Larry and Ken.

Many thanks go to my mother, brothers and sisters. Lastly, I would like to thank my wife Dr Elise Schieck for her love, support and encouragement throughout my PhD. I love you! I dedicate this work to my son Mateo Schieck – he will be happy that daddy is finally coming back home!

Abstract

The natural selection imposed by host immunity and antimalarial drugs has driven extensive adaptive evolution in *Plasmodium falciparum*, leading to an ever-changing landscape of genetic variation. We have carried out whole-genome sequencing of 93 *P. falciparum* clinical isolates from Malawi and used population genetic methods to investigate the genetic diversity and regions under selection. In addition, by computing *XP-EHH*, *PCA* and F_{ST} we have compared the Malawi isolates to five dispersed others (Kenya, Mali, Burkina Faso, Cambodia and Thailand), and identified genes potentially under positive directional selection. Geographic stratification of genetic diversity in the populations followed continental lines and small population differences were observed within Africa. Positive directional selection signals were identified at or near *pfdhps, pfcrt, pfmdr1* and *pfgch1* (known drug targets) and in several merozoite invasion ligands (e.g., *msp3.8, trap* and *ama1*). We discuss the role of drug selection in promoting fixation of alleles between populations with differing adaptation to local drug pressure. Analysis of copy number variation in Malawi provides a detailed catalogue of new and previously identified gene deletions and duplications with critical roles in cytoadherence, gametocytogenesis, invasion and drug response. This work provides the first genome-wide scan of selection and CNV in Malawi to guide future studies in investigating parasite evolution, changing malaria epidemiology, and monitoring and evaluating impact of malaria interventions as they are deployed.

Abbreviations

$	US Dollar
A+T (AT)	Adenine and Thiamine
ACT	Artemisinin-based Combination Therapy
ACTia	ACTs in action
AMA1	Apical membrane antigen 1
ART	Artemisinin
AS+AQ	Artesunate-Amodiaquine
BAMtools	Binary Alignment and Map tools
BCFtools	Binary Call Format tools
bp	Base pair
cc	Cubic centilitres
CF11	Fibrous Cellulose
CGH	Comparative genomic hybridization
cM	Centimorgan
CM	Cerebral malaria
CMX	Cotrimoxazole
CNVs	Copy Number Variations
C/MOI	Complexity/multiplicity of infections
CQ	Chloroquine
CQR	Chloroquine resistance
CQS	Chloroquine sensitive
CSP	Circumsporozoite protein
DBL-MSP	Duffy binding like - Merozoite surface protein
DHA-PPQ	Dehydroartemisinin-Piperaquine
DNA	Deoxyribonucleic acid
dNTPs	Deoxynucleotide Triphosphates
FAM	Fluorescein amidite
F_{ST}	Fixation Index test
GO	Gene Ontology
HEPES	N-(2-Hydroxylethyl)piperacin-N′-(2-ethylsulphonacid)
HF	Halofantrine
iHS	Integrated Haplotype Score
HIV	Human Immunodeficiency Virus
IDT	Integrated DNA technologies
IE	Infected Erythrocytes
INDELs	Insertions and Deletions
IPTp	Intermittent Prevention Therapy in Pregnancy
IRS	Indoor Residual Spraying
KAHRP	Knob-associated histidine rich protein

kb	Kilo base
LA	Artemether-Lumefantrine
LCRs	Low Complexity Regions
LD	Linkage Disequilibrium
LF	Lumefantrine
LRH	Long Range Haplotype
MADIBA	MicroArray Data Interface for Biological Annotation
MAF	Minor allele frequency
MaRCH	Malaria rebound in children
Mb	Mega base
MIP	Malaria in Pregnancy
ml	Millilitres
MPS	Massive Parallel Sequencing
MQ	Mefloquine
MS	Microsatellite
MSP	Merozoite surface protein
MSP3	Merozoite surface protein 3
NAHR	Non-allelic homologous recombination
NGS	Next Generation Sequencing
PCA	Principal Component Analysis
PCR	Polymerase Chain Reaction
PfAPI	*Plasmodium falciparum* annual parasite incidence
PfEMP1	*P. falciparum* erythrocyte membrane protein 1
$PfPR_{2-10}$	*P. falciparum* parasite rate age-standardized to 2–10 year
PYR	Pyrimethamine
QN	Quinine
RBC	Red Blood Cells
RDT	Rapid Diagnostic Test
RFLP	Restriction Fragment Length Polymorphism
RIFIN	Repetitive Interspersed Family
RPMI	Rosewell Park Memorial Institute
RT	Room temperature
rt-PCR	Real time PCR
RTS,S	Also known as Mosquirix
SAMtools	Sequence Alignment/Map tools
SDX	Sulfadoxine
SM	Severe malaria
SNPs	Single Nucleotide Polymorphisms
SP	Sulfadoxine-Pyrimethamine
STEVOR	Subtelomeric Variable Open Reading Frame
SURFIN4.2	Surface associated interspersed protein 4.2

VSA	Variant Surface Antigen
WHO	World Health Organization
WTSI	Wellcome Trust Sanger Institute
XP-EHH	Cross Population Extended Haplotype Homozygosity

Table of Contents

List of Tables

List of Figures

To my family and MCDC

Chapter 1

Introduction to malaria and malaria genetics

1.1 Global malaria burden

Malaria is caused by eukaryotic parasites of the genus *Plasmodium* (*P. falciparum*, *P. vivax*, *P. ovale*, *P. malariae* and *P. knowlesi*). Human *P. falciparum* malaria is the deadliest, and predominates in Africa while *P. vivax* is the most widespread. *P. falciparum* is transmitted to humans by the bite of an infected female *Anopheles* mosquito (Mendis et al. 2001; Genton et al. 2008). An estimated 3.3 billion people are at risk of malaria worldwide, with populations living in sub-Sahara Africa at the highest risk of malaria (Figure 1.1). There were an estimated 219 million cases of malaria (range 154–289 million) and 660 000 deaths (range 610 000–971 000) in 2010. (WHO. World Malaria Report 2012).

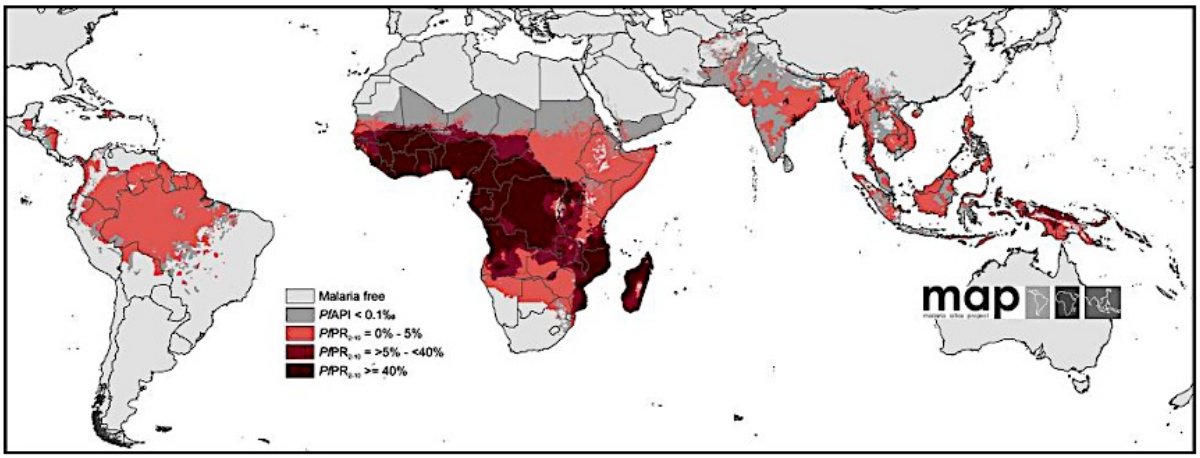


Figure 1.1: Global distribution of *P. falciparum* transmission risk, 2010. Medium grey regions indicate low rates of *P. falciparum* infections (unstable risk, *P. falciparum* annual parasite incidence, PfAPI <0.1 per 1000 people *per annum*) while light grey indicate no risk. Shadings of red indicate levels of infections: low risk, *P. falciparum* parasite rate age-standardized to 2–10 year for endemic mapping ($PfPR_{2-10}$ ≤5%), in light red; intermediate risk, $PfPR_{2-10}$ > 5% < 40%, in medium red; high risk, $PfPR_{2-10}$ ≥40%, in dark red. Adapted from (Gething et al. 2011).

1.2 *P. falciparum* and life cycle

P. falciparum is a eukaryotic pathogen with a complex life cycle, spending part of its lifespan in the *Anopheles* mosquito (definitive host), as mostly diploid, and in the human host as a haploid organism, where it gives rise to numerous clinical manifestations ranging from mild to life- threatening illness (Figure 1.2).

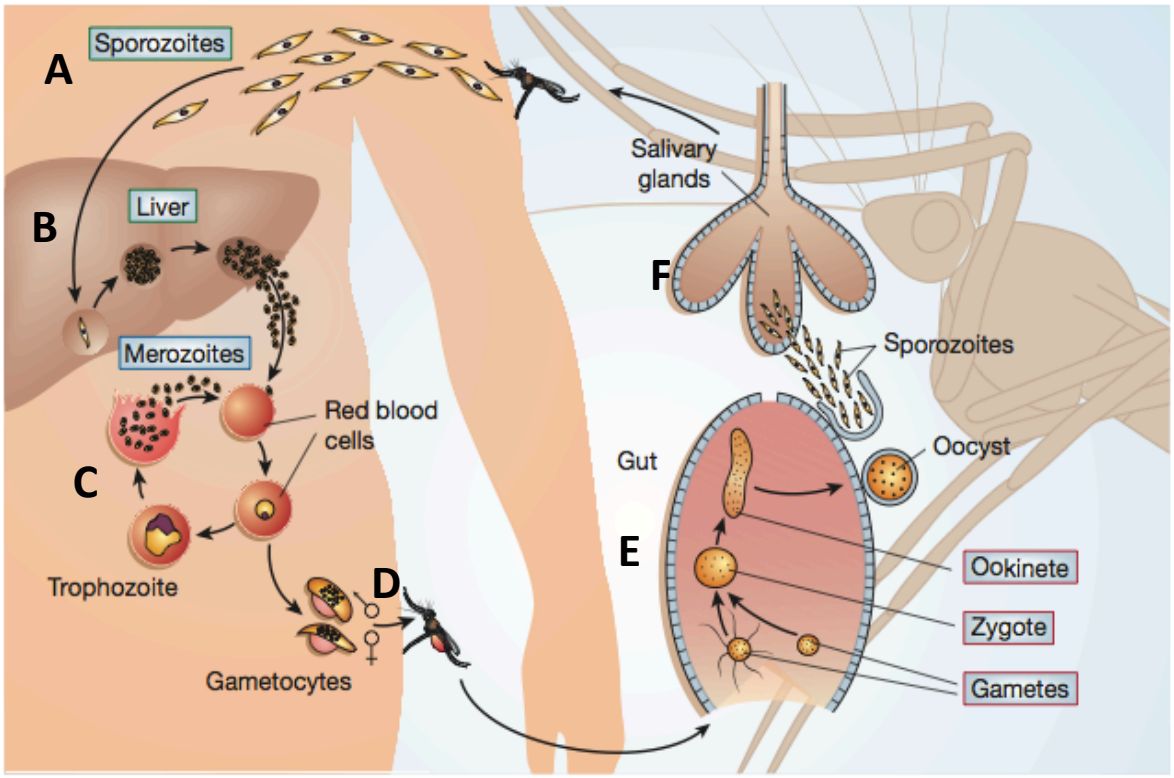


Figure 1.2: *P. falciparum* life cycle adapted from (Ménard 2005).

Within its hosts, the parasite must adapt to varying environmental conditions such as cellular metabolism, temperature, drugs and immune response and as such, it requires specific capabilities to proliferate and transmit.

During a blood meal, female *Anopheles* mosquito injects sporozoites into the human host (A). Within thirty seconds, the sporozoites enter the liver into hepatocytes and are hardly detected in the blood stream after thirty minutes (Essential Malariology. Fourth Edition. David A. Warrell and Herbert M. Gilles). The sporozoites develop into exo-erythrocytic schizonts undergoing asexual reproduction until the hepatocyte bursts (B), to release up to 40,000 merozoites into the human bloodstream, and infect erythrocytes. At this stage, a few differences are observed between different *Plasmodium* species - in the case of *P. vivax* and *P. ovale*, some sporozoites enter hepatocytes and form hypnozoites (dormant stage) instead of directly developing exo-erythrocytic schizonts. The hypnozoites (4-5 µm in diameter) can remain dormant in the liver for years and at some point, through a triggering signal that is still not fully described, the hypnozoites develop into exo-erythrocytic schizonts, producing thousands of merozoites that cause relapse of the disease. Merozoites do not usually remain in the blood stream for long periods (helping to avoid direct contact to the host immune system), infecting erythrocytes as soon as they are released, and undergo an asexual reproduction cycle called the erythrocytic schizogony. At this stage, the nucleus is divided 3 to 5 times, followed by division of the cytoplasm. The ring stage is formed, followed by a more metabolically active trophozoite stage and finally a schizont. Schizonts are fully developed parasite forms, containing a variable number of merozoites, usually 8 to 24, again depending on species and strain. Once the schizonts rupture, the meroziotes are released into the blood stream and directly invade new red

blood cells, thereby starting a new cycle of schizogony (C). The length of the intra-erythrocytic cycle also differs between different *Plasmodium* species and strains, typically 48 hours (in *P. falciparum, P. vivax* and *P. ovale*) or 72 hours (in *P. malariae*), which explains the periodicity of the fever paroxysms experienced by the patient. Some merozoites invading new red blood cells do not develop into schizonts but into sexually differentiated forms called gametocytes (D).

When a female *Anopheles* mosquito feeds on a blood meal from a malaria positive person, it ingests infected erythrocytes (IE), male and female gametocytes and uninfected erythrocytes. In the male gametocytes, the nucleus divides into four to eight nuclei, which ex-flagellate to form microgametes, each being able to fertilize a macrogamete - the mature form of the female gametocyte. Fertilization of the microgamete and macrogamete produces a zygote, which develops into an ookinete that crosses the mosquito gut wall to form an oocyst between the epithelial lining and the basal lamina. As the oocyst grows, the nucleus divides, forming elongated sporozoites, approximately 1000 *per* oocyst (E). The sporozoites actively break through the wall of the oocyst, and reach the salivary glands via the haemolymph. An infected mosquito will inject sporozoites into the human host, to complete the life cycle (F) (Essential Malariology. Fourth Edition. David A. Warrell and Herbert M. Gilles).

The *Plasmodium* lifecycle differs at some stages depending on the species and strains. For example only *P. vivax* and *P. ovale* form hypnozoites in the liver, which cause relapse of the disease. *P. ovale* and *P. vivax* merozoites invade reticulocytes only, whereas *P. falciparum* and *P. malariae* infect erythrocytes of different ages. The length of the

erythrocytic schizogony cycle also varies, which results in different periodicity of the febrile paroxysms. The amount of released merozoites from the hepatocyte schizonts or the erythrocytic schizonts can also determine the severity of the disease. For example, whilst the mortality level for *P. falciparum* is high, accounting for most malaria deaths in the world, *P. vivax* malaria is generally considered only rarely fatal but there is now growing concern of serious and fatal illnesses associated with Vivax malaria (Price et al. 2009; Mendis et al. 2001; Nurleila et al. 2012).

1.3 Malaria control and impact on the *P. falciparum* genome

Malaria transmission can be controlled in many ways. Preventing mosquito bites and killing mosquitoes with insecticide-impregnated bed-nets (ITNs), insect repellents, indoor residual spraying (IRS) or draining standing water where mosquitoes lay their eggs have become common. Also available are antimalarial drugs (e.g., artemisinin-based combination therapies (ACT)), and under development are vaccines whose success would provide a high level of protection for a sustained period (Kilama and Ntoumi 2009; Agnandji et al. 2011; Bejon et al. 2013).

The WHO malaria 2012 report indicated that international funding for malaria control has continued to rise, to a peak of US$ 2 billion in 2011. However, they also noted that this funding still fell short of the projected resources required (US$5 billion per year for the years 2010 - 2015) to achieve malaria control targets. This funding was also projected to remain at these levels or decrease before 2015 unless new sources of funds were identified. This, it's noted, will pose great challenges on the effective and efficient control of malaria. Nevertheless, efforts have been put in place to allow expansion of access to valuable public health tools such as long-lasting ITNs and IRS, as well as early access to rapid diagnosis and

effective antimalarial drugs, to reduce death toll in a number of countries. For example, the percentage of households owning at least one ITN in sub-Saharan Africa stood at 50% in 2010 from 3% in 2000, while the percentage protected by IRS increased to 11% in 2010 from less than 5% in 2005 (WHO. World Malaria Report 2012) Household surveys also indicated that 96% of persons with access to an ITN within the household actually used it. The number of rapid diagnostic tests (RDTs) and ACTs procured has increased, so is the percentage of reported suspected cases receiving a parasitological test, from 67% globally in 2005 to 76% in 2010, with the largest increase in sub-Saharan Africa (WHO. World Malaria Report 2012). Despite this significant progress, more work is needed before the target of universal access is attained. More than 50% reduction in malaria cases was reported between 2000-2010 in 43 of the 99 countries with on-going transmission (WHO. World Malaria Report 2012).

Nonetheless, behind these successes to malaria control lie a worrying reality, that the *P. falciparum* is adapting in response to selection pressures from the host immune system, drug treatment and changes in transmission intensity owing to specific malaria interventions (Anderson et al. 2011; Weedall and Conway 2010). The major consequence of this adaptation has been emergence and spread of drug resistant parasites and difficulties in developing a vaccine. In addition, mosquito resistance to commonly used components in IRS and ITN (pyrimethrin and DDT) is spreading (Kilama and Ntoumi 2009). However, these changes to *P. falciparum* population structure can potentially be used to identify and circumvent survival strategies used by the parasite with an eye towards reducing malaria burden.

It has widely been suggested that a vaccine against *P. falciparum* would probably be the best way to curb malaria and reduce much of the burden associated with it. The RTS,S

vaccine from GlaxoSmithKline Biologicals has so far been the most promising. This recombinant malaria protein (consisting of components of the *P. falciparum* circumsporozoite protein (CSP) joined to hepatitis surface antigen) has been shown to reduce number of severe cases of malaria and delay the time to the first clinical episode (Agnandji et al. 2011; Aponte et al. 2007). A pooled survey of all phase two trials reported that RTS,S efficacy against all episodes of clinical malaria varied by transmission, with high efficacy (60%) in low transmission areas, falling to 4% in high transmission areas (Bejon et al. 2013).

1.4 *P. falciparum* genome

Despite challenges such as high A+T content, the *P. falciparum* genome was one of the first eukaryotic genomes sequenced. The full genome sequence was published in 2002 and was followed by the publication of other *Plasmodium* genome sequences such as *P. vivax* (Carlton et al. 2008). These data have allowed elucidation of basic genome architecture and identification of key structural elements, common metabolic and biosynthesis pathways and unique aspects that are shared among several *Plasmodium* parasites (Gardner et al. 2002; Carlton et al. 2008; Hall et al. 2005). For example, it was observed that a large proportion of *P. falciparum* proteins lack similarity with known proteins from other organisms, suggesting a *Plasmodium*-specific role.

The *P. falciparum* life cycle has a mostly haploid genome containing ~23.3 Mb nucleotides encoding ~ 5500 genes and ~145 pseudogenes (www.genedb.org, Hertz-Fowler et al. 2004) organised into 14 chromosomes, ranging in size from ~640-kb to 3.3-Mb in the nuclear genome, along with two extra chromosomal DNA elements (35-kb circular plasmid and 6-kb mitochondrial genomes) (Gardner et al. 2002, 1993). The nucleotide content

(percentage of the G/A/T/C) is 80.6% AT rich (80.6% in coding regions and ~90% in noncoding regions), greater than the 67.7% in *P. vivax* (Gardner et al. 2002). The A+T rich regions are thought to be more recombinogenic thus helping to produce antigenic variation (Winzeler 2008). Almost one-half of the predicted genes encode conserved hypothetical proteins with unknown functions (Carlton et al. 2008), although recent re-annotation of the genome has assigned putative functions to many additional genes.

There is a high incidence of tandem repeats and low complexity regions (LCRs) in *P. falciparum* antigens. These tandem repeats are thought to be involved in immune evasion mechanisms, for example through antigen diversification (Hughes 2004), they may reduce the host's antibody response to critical epitopes through multiple cross-reactivity amongst them (Anders 1986). LCR expansion also partly contributes to the slightly larger size of *P. falciparum* proteins (Carlton et al. 2008).

1.5 The *P. falciparum* genetic diversity map and function

Genetic variation is key to the survival of *P. falciparum,* as it provides a means for overcoming environmental challenges such as immune and drug pressure. Several factors have contributed to this variation and are described below.

The large numbers of asexual blood stage parasites in a single infection provides a reservoir within which biological selection can act. Chromosomal polymorphisms (e.g., genetic recombination, single nucleotide polymorphism (SNPs), insertions and deletions (INDELs), copy number variation (CNV), inversions, translocations and tandem repeats in surface antigens) occurring primarily during the sexual cycle largely contribute to this variation. In the sexual phase, meiosis (an act of genetic recombination through

independent assortment of chromosomes and crossing over between chromosomes) generates new parasite variants with phenotypic traits such as drug resistance. Outcrossing (characterized by high frequency of multiple and distinct genotype infections) produces an effective recombination rate that breaks down linkage disequilibrium (LD, non random association of alleles), whereas selfing/inbreeding preserves LD, a principle that has been used to map drug resistant markers and virulence in *P. falciparum.* Also occurring at this sexual phase is subtelomeric hypervariability, as a result of unstable AT repeat rich regions producing recombination hotspots (Mu et al. 2005; Volkman et al. 2002; Kidgell et al. 2006; Corcoran et al. 1988; Vernick et al. 1988). Another form of variation is achieved through presence of repetitive sequences in certain genes such as circumsporozoite protein (*csp*), where the size of the repeat sequences vary significantly between clones thus providing antigenic diversity needed to escape immune system. For example, adults have specifically acquired strain specific immunity to the *csp* repeat length variants (Bowman et al. 2013; Zeeshan et al. 2012).

However, to date, point mutations or SNPs are the widely characterized form of variation and the markers of choice. SNPs involve a single base pair variation, are relatively abundant in *P. falciparum* and their frequency vary across the 14 chromosomes and genes. A high frequency of SNPs has been observed in surface molecules and putative transporters, that are likely to be under drug or immune selection, whilst house-keeping genes show lesser SNP frequencies (Jeffares et al. 2006; Volkman et al. 2007a; Mu et al. 2007). Studies of *P. falciparum* genetic diversity have revealed their population structure and involvement with drug response. For example, reduction in SNP diversity accompanied by selective sweeps has been associated with positive selection (e.g., in *pfcrt, pfdhps, pfdhfr*) largely

driven by drug pressure (Nair et al. 2003; Wootton et al. 2002), while diversifying selection produced highly polymorphic regions at candidate vaccine targets such as *ama1*, *msp1* and *2*, *csp* and *eba175* genes (Polley et al. 2003; Baum et al. 2003; Ferreira et al. 2003). A rich SNP density has been observed in *P. falciparum*: Volkman et al. sequenced 14 parasite lines and identified 46,937 SNPs (1/446-bp) and 37,039 indels (1/548-bp) (Volkman et al. 2007); Jeffares et al. compared the genomes of *P. falciparum* and *P. reichenowi* and identified 27,169 SNPs (1/762-bp) and 27,478 indels (1/631-bp) (Jeffares et al. 2006) and Mu et al. identified 3,918 SNPs (1/5.9-kb) and 2,548 microsatellites (MS, 1/1.7-kb) by sequencing 3,539 genomic regions (Mu et al. 2010). This SNP discovery has been revolutionized by the recent advent of next generation sequencing (NGS) which enables the rapid sequencing of whole genomes, making SNP discovery and genotyping less laborious. A recent study detected 86,158 exonic SNPs (1/266-bp) in 227 worldwide isolates using NGS (Manske et al. 2012). Taken together, a SNP density of 1/266-bp within 23 Mb genome of *P. falciparum* (Manske et al. 2012), a population pairwise diversity (π) of 1.29×10^{-3} (Volkman et al. 2007), physical versus genetic distance of 17-kb per cM (Su et al. 1999) and LD range of 1.5 to 16-kb (Volkman et al. 2007) demonstrate a very rich map of diversity in this eukaryotic parasite.

CNV also act as a major source of genome variation, are prevalent in *P. falciparum*, and are increasingly being studied. Some CNV even encompass entire genes, and so can influence gene expression levels as well as phenotypic variation, thereby revealing important functions both in disease and drug response such as *pfmdr1* (on chromosome 5) and *msp3.8* (on chromosome 10) in drug resistance (Carret et al. 2005; Ribacke et al. 2007; Van Tyne et al. 2011). In addition, amplifications on chromosome 12 of GTP cyclohydrolase 1, the first enzyme in the folate biosynthesis pathway is likely due to compensation for the

decreased efficiency of the folate pathway caused by mutations in *pfdhps* and *pfdhfr* (Kidgell et al. 2006; Nair et al. 2008). Subtelomeric deletions on chromosome 2 of the knob-associated histidine rich protein gene (*kahrp*) and chromosome 9 of the cytoadherence linked asexual gene 9 (*clag9*) are associated with the loss of cytoadherence (Biggs et al. 1989; Trenholme et al. 2000). Also detected is a gain in copy number of the reticulocyte-binding protein 1 gene (*pfRh1*), involved in human erythrocyte invasion and linked to fast growth, that may have arisen during culture adaptation of certain parasite lines (Nair et al. 2011). CNV also play a role in altering gene expression throughout the *P. falciparum* genome as demonstrated by Gonzales et al. In this study, a large regulatory locus (269 transcripts) occurred within the *pfmdr1* amplification and 13 other unlinked genes. In addition, drug selection in the Dd2 parental clone not only led to amplicons in *pfmdr1* but also in other putative neighbouring regulatory factors that influence the overall transcriptional network in the *P. falciparum* genome (Gonzales et al. 2008). CNV have also been used to explain gene expression differences between field and lab isolates (Mackinnon et al. 2009).

1.6 Immunity and antigenic variation

In highly endemic areas, acquisition of protective immunity to clinical disease is a slow process involving years of multiple infections by the parasite. A person with protective immunity will usually develop mild malaria even with high parasitemia (Doolan et al. 2009; Day and Marsh 1991), however, young children aged less than 5 years and pregnant mothers are at greater risk of developing severe malarial disease. In regions of unstable/patchy malaria transmission and no significant protective immunity, the whole population is generally susceptible to clinical malaria (Carter and Mendis 2002). Because of immune responses killing the parasite, *P. falciparum* has had to exploit tricks to evade

death, mostly through antigenic variation. Antigenic variation derails acquisition of protective immunity, partly attributed to the parasites highly polymorphic antigens that require a constant production of new antibodies by the host immune system (Rogers et al. 1992; Newbold et al. 1992). Given the fact that in nature no sterile immunity develops, it is challenging to develop a vaccine,

1.6.1 Multi-gene families

The success of *P. falciparum* also depends on its ability to invade host tissues and counter host defence mechanisms. To ensure long-term survival they must also maintain infectivity through transmission. Antigenic variation defined as expression of functionally and antigenic distinct proteins, is a strategy employed by *P. falciparum* to evade host immune responses (Craig and Scherf 2001). As it invades a host cell such as red blood cell (RBC), it expresses antigenic variant surface proteins (VSA) on the surface of infected RBC to help avoid specific recognition from immune system (Craig and Scherf 2001; Dzikowski et al. 2006a, 2006b). Most of these proteins are encoded by various multi-gene families (accounting for approximately 7% of genes in the genome) located mostly in subtelomeric end of the chromosomes, are highly polymorphic, and some expressed in a mutually exclusive way. Widely studied VSA include PfEMP1 (encoded by *var* genes), STEVOR (encoded by *stevor* genes*)* and RIFINS (encoded by *rif* genes*)*. *P. falciparum* is particularly known to use PfEMP1 to adhere to receptors of the host microvasculature leading to sequestration of the parasites in the deep vascular bed, for example in the brain leading to cerebral malaria.

1.7 Antimalarial drug resistance in *P. falciparum*

Emergence and spread of drug resistant malaria parasites have greatly influenced epidemiology of malaria and options for its treatment. For many decades, chloroquine (CQ) was used as first line drug treatment for malaria due to its low cost, safety and efficacy. CQ binds to heme molecules in the parasite food vacuole, interfering with heme detoxification and leading to toxic levels of heme in the parasite (Egan et al. 1999). CQ resistance (CQR) was first reported in two foci in southeast Asia (Thai-Cambodia border in late 1950s) and South America (Columbia in early 1960s) and spread to all malaria endemic regions, mainly in sub-Sahara Africa and parts of south America (Wootton et al. 2002; Hayton and Su 2008; Chen et al. 2003; Vieira 2001). CQR was established through the use of a genetic cross between a CQ sensitive and CQ resistant parasite that led to the discovery of a mutation in a putative transporter, *P. falciparum* chloroquine resistant transporter (*pfcrt*) on chromosome 7, and in particular a substitution of lysine by threonine at codon position 76 (K76T) (Fidock et al. 2000; Sidhu et al. 2002). To some extent its also modulated by mutations and copy number variation (CNV) in *P. falciparum* multidrug resistant (*pfmdr1)* gene on chromosome 5 (Mu et al. 2003). However, the contribution of *pfmdr1* in CQR remains unclear and is thought to be a compensation for fitness costs (Patel et al. 2010). Mutations in *pfcrt* are thought to confer resistance by preventing CQ accumulation in the digestive vacuole to levels required for inhibition of endogenous heme detoxification (Egan et al. 1999).

After emergence and spread of CQR in most countries, the next primary drug to be adopted was the co-formulated synergistic combination of sulfadoxine (SDX) and pyrimethamine (PYR), known as sulfadoxine-pyrimethamine (SP) which targeted and inhibits two enzymes - the dihydropteroate synthase (*dhps*) and dehydrofolate reductase (*dhfr*)

genes respectively of the folate biosynthesis pathway in the parasite (Prajapati et al. 2011). SP resistance (SPR) developed and was also first reported in Thailand-Burma border and spread to fixation in approximately 6 years through mutations in each of the genes encoding *pfdhps* (chromosome 8) and *pfdhfr* (chromosome 4) (Nair et al. 2003; Plowe et al. 1997; Nash et al. 2005; Roper et al. 2004). PYR being a competitive inhibitor of *pfdhfr,* displaces the natural folate substrate and resistance occurs when specific point mutations occur: S108N (serine to asparagine at codon 108) only confer a small amount of resistance, while additional mutations in *pfdhfr* at N51I (asparagine to isoleucine), C59R (cysteine to arginine) and I164R (isoleucine to arginine) act synergistically to increase levels of PYR resistance (Vasconcelos et al. 2000; Plowe et al. 1997). The *pfdhfr* quartet mutant of N51I + C59R + S108N + I164R has the highest levels of PYR resistance. Resistance to SDX is due to mutations on the *pfdhps* gene that decrease the enzyme binding affinity to SDX (Triglia et al. 1998). Similar to the *pfdhfr*, an amino acid change in *pfdhps* at position 437 (A437G) represents the initial mutation for SDX resistance. An additional mutation(s) at positions 436 (S436A, serine to alanine), 540 (K540E, lysine to glutamine), 581 (A581G, alanine to glycine), and/or 613 (A613S, serine to threonine) causes elevated levels of SDX resistance *in vitro* (Triglia et al. 1998).

CNV and SNPs at *pfmdr1* have also been associated with parasite responses to several drugs including mefloquine (MQ), quinine (QN), artemisinin (ART), lumefantrine (LF) and halofantrine (HF) (Sidhu et al. 2006). In addition, mutations in a gene encoding a putative Na+/H+ exchanger (*pfnhe*) and *pfatp6* have been associated with parasite responses to MQ and ART respectively (Jambou et al. 2005; Ferdig et al. 2004), but need further proof.

In the last decade ART have been deployed as first-line treatment of uncomplicated *P. falciparum* malaria in endemic countries in combination with other drugs such as MF, amodiaquine (AMQ), piperaquine (PPQ), PYR/SDX, or LF, collectively known as artemisinin-based combination therapies (ACTs) (WHO. World Malaria Report 2012). ART has a short half-life but acts extremely very quickly in reducing parasite densities and symptoms. Together with other control measures ACTs have resulted in a remarkable decrease in malaria morbidity and mortality (Jambou et al. 2005; Barnes et al. 2005). However, recent confirmed reduced parasite *in vivo* susceptibility to artesunate in western Cambodia has led to concerns that the efficacy of ACT could be declining through emergence of drug resistant genotypes (Noedl et al. 2008, 2009; Das et al. 2009). The resistant phenotype has not been well characterized nor any molecular marker identified (Cheeseman et al. 2012; Miotto et al. 2013) and this impedes surveillance studies to monitor the spread of the resistant phenotype. Thus, there is urgent need to identify molecular markers underlying such phenotypic traits in order to give insight into the mechanism of ART antimalarial action and parasite resistance.

However, mutations in certain genes have been postulated to confer resistance to ACT including: (a) The *pfmdr1,* a 4.2-kb long gene whose decreasing copy number is thought to confer susceptibility to MQ, LF, HFX, QN and ART. *In vivo* selection of Y86N allele has also been observed after artemether-lumefantrine (LA) treatment in Africa. The Y86N was considered as a potential marker for LF resistance *in vivo* (Sidhu et al. 2006; Sisowath et al. 2005). (b) The *pfatp6*, a 4.3-kb gene on chromosome 1 encoding calcium dependent sarcoplasmic/endoplasmic reticulum calcium ATPase, has been shown to be a target of ART drugs in *Xenopus* oocytes through a single amino acid change, L263E (Eckstein-Ludwig et al.

2003; Uhlemann et al. 2005). Another amino acid change S769N is also associated with *in vitro* sensitivity (Jambou et al. 2005). (c) A gene (*cytochrome b*) in the 6-kb mitochondrial genome contain mutations that can potentially change susceptibility to ART (Imwong et al. 2010). (d) The *pfubp1*, a 3.3-kb gene on chromosome 2 contain mutations V739F and V770F that have been shown to confer ART resistance in *P. chabaudi* (Hunt et al. 2007). (e) Potential involvement of CQR mutations acting as a prerequisite for subsequent development of ART resistance as observed in *P. chabaudi* (Hunt et al. 2007). (f) A large region on chromosome 13 is also associated with slow ART clearance rates (Cheeseman et al. 2012).

1.8 Malaria epidemiology in Malawi

Malawi is a landlocked country located in southeast Africa. In Malawi, malaria is a major public health problem, with the entire population (15 million) at risk and an estimated 6 million cases occurring annually. Children less than five years of age and pregnant women are mostly affected. *P. falciparum* is the major *Plasmodium* species, with *Anopheles funestus*, *A. gambiae*, and *A. arabiensis* the primary mosquito vectors (Mzilahowa et al. 2012; Malawi National Malaria Indicator Survey, 2010). The Malawi National Malaria Strategic plan 2005-2010 main goal was to scale up malaria interventions to reduce the burden including prioritizing effective antimalarial usage, ITNs, IRS, and prevention in pregnancy (IPTp). ITNs have been the primary control strategy for malaria both in children and pregnant women, with 58.2% of households having at least one ITN. IPTp is used to prevent MIP and was adopted in 1993. As of 2007, LA has been used as first-line treatment for uncomplicated falciparum malaria, AS+AQ used for treatment failure and QN for

treatment of severe cases. IRS was adopted in 2007 and coverage has reached 40%. (Malawi National Malaria Indicator Survey, 2010; WHO. 2012).

Malawi experiences year round transmission with a peak during the rainy seasons between December and May (Ewing et al. 2011). As at 2010, parasite prevalence rate by slide microscopy was 43.3% nationally and severe anaemia prevalence (haemoglobin < 8g/dl) at 12.3% in children under five years of age. Malaria parasitemia among children was higher in rural (46.9%) than urban areas (14.7%), and parasite prevalence increased with age while prevalence of severe anaemia showed an opposite trend (Malawi National Malaria Indicator Survey 2010).

There is high resistance to SP drug, with over 90% of infections carrying the 'quintuple mutant' (*pfdhfr* mutations N51I, C59R and S108N, and *pfdhps* mutations A437G and K540E) (Nkhoma et al. 2007). Since withdrawal of CQ in 1993, prevalence of K76T *pfcrt* has fallen to 0%, and the re-emergence of CQ sensitive genotypes raises the possibility that CQ might be used again to treat malaria (Nkhoma et al. 2007). However, no report on LA treatment failure/resistance has been observed in Malawi to date.

1.9 Advances in investigating *P. falciparum* genetic diversity

With the advent of *in vitro* culturing systems (Trager and Jensen 1976), studies of parasite diversity took centre stage and initially focused on enzyme and antigen protein variability due to technological limitations. Observations of genetic diversity in *P. falciparum* using enzyme protein polymorphisms (particularly in the 6-phosphogluconate hydrogenase and glucose-6-phosphate isomerase) were based on electrophoretic variations in these enzymes to reflect genetic differences (Knowles et al. 1981; Walliker 1983; Hempelmann

and Dluzewski 1981). Soluble antigens were also used to observe antigenic diversity and this was established in sera of infected patients from the Gambia (Wilson et al. 1969) where 50 patients showed highly diverse S-antigens with 18 distinct forms. These procedures were later improved to incorporate use of monoclonal antibodies and immunofluorescence to study diversity in worldwide parasite populations (McBride et al. 1982).

Developed in 1983, gene amplification using polymerase chain reaction (PCR), was introduced as a way of assessing *P. falciparum* genetic diversity and population structure, first by looking at antigenic locus size polymorphisms (Scherf et al. 1991; Ntoumi et al. 1995). The commonly used polymorphic genes for PCR genotyping became *msp1*, *msp2* and *glurp*, with *msp2* as the most informative and used to differentiate between recrudescence and reinfection (Cattamanchi et al. 2003; Färnert et al. 2001). In 1987, the first *P. falciparum* genetic cross was established between two parent lab clones, HB3 and 3D7 to distinguish the two parents and identify recombinant progeny (Walliker et al. 1987). In this study, PYR resistance was used as a phenotypic marker and was linked to an amino acid change in the *pfdhfr* gene (S108N mutation) (Peterson et al. 1988). This work demonstrated that recombination could provide a mechanism for producing parasites with novel genotypes bearing important clinical phenotypes such as drug resistance. In 1990, another genetic cross between CQ resistant clone, Dd2 and CQ sensitive, HB3 was established (Wellems et al. 1990) and led to identification of CQR molecule (Fidock et al. 2000). Further work through development of restriction fragment length polymorphisms (RFLP) markers localised the CQR locus to a ~ 400-kb segment on chromosome 7 (Wellems et al. 1991) and later with genomic advances, microsatellite (MS) markers were used to fine map CQR gene to a 36-kb segment (Su et al. 1999; Su et al. 1997). MS marker genotyping become a

valuable and widely used approach, because the markers have high mutation rates and polymorphisms, multiple alleles and are selectively neutral (Su et al. 1999).

In 2002, completion of *P. falciparum* 3D7 genome sequence provided further advancement and enabled identification of SNP markers (Gardner et al. 2002) that have since become the marker of choice for high throughput genotyping (Su et al. 2007). SNPs are abundant in the *P. falciparum* genome and their distribution vary greatly (Mu et al. 2007) between chromosomes and genes, with high frequencies seen in surface proteins and putative transporters thought to be under drug and immune selection. With the development of MS and SNPs, came new methods for exploiting *P. falciparum* genetic diversity. Genotyping arrays were developed which provided a fast, relatively inexpensive way to examine genome-wide SNPs and CNV and depending on the technology, genomic variation could now be captured at SNP, MS, INDEL and CNV level and have been successfully used in *P. falciparum* (Kidgell et al. 2006; Jiang et al. 2008a,2008b; Ribacke et al. 2007; Hayton and Su 2008). These arrays explored hybridization techniques where labelled parasite genomic DNA is hybridized to an array containing millions of probes sequences. For example, a difference (substitution) in a probe sequence between 3D7 and test sample would reduce signal from the test sample compared to that of 3D7 and inferred mutation. Some commonly used arrays include PFSANGER tiling array (Sanger Institute, UK) used to identify single feature polymorphisms and CNV (Mu et al. 2007; Jiang et al. 2008b; Kidgell et al. 2006; Mourier et al. 2008) and molecular invasion probe that was used to identify loci under positive selection, variation in population recombination events and population structure (Mu et al. 2010) (Table 1.1).

Table 1.1: List of arrays developed and used in detecting polymorphisms in *P. falciparum*

Type of array	Polymorphism	Reference
Affymetrix array	CNV	Carret et al. 2005
	CNV and SNPs	Dharia et al. 2010
	Single feature polymorphism	Kidgell et al. 2006
UltraGAPS printed glass slides	CNV	Ribacke et al. 2007
PSANGER affymetrix GeneChip/CGH array	SNPs and CNV	Jiang et al. 2008a, 2008b
PSANGER affymetrix GeneChip	CNV	Cheeseman et al. 2005
Affymetrix 3,000 SNP chip	SNPs	Neafsey et al. 2008
4.8 M probe affymetrix CGH array	CNV and SNPs	Dharia et al. 2009
Gene expression array	Transcription and CNV	Mackinnon et al. 2009
Affymetrix microarray/molecular invasion probe	SNPs	Mu et al. 2010
Custom NimbleGen 385K array	CNV	Samarakoon et al. 2011a
Affymetrix array containing 74,656 SNPs	SNPs	Amambua et al. 2012

However, tiling arrays have a disadvantage that they can only type known SNPs and usually require species-specific customization. The high AT rich *P. falciparum* genome with an abundance of repeat elements, makes it difficult to design suitable probes for the majority of noncoding regions (Hayton and Su 2008). Tiling arrays also require large amounts of *P. falciparum* DNA which is often obtained by culture adaptation of sampled parasites which in turn may lead to copy number fluctuations, altering naturally occurring CNV (Anderson et al. 2009; Mackinnon et al. 2009). Contamination of human DNA is another problem as it can inhibit hybridization and since human DNA accounts for approximately 100-fold more than parasite DNA, it is important to deplete samples of human white blood cells prior to DNA extraction. Multiclonal infections of *P. falciparum* offer another challenge (Hayton and Su 2008).

Next generation sequencing techniques (NGS), also called massive parallel sequencing (MPS), are increasingly becoming popular. With the ability to produce dense variation, allowing identification of almost whole genome variation, they are highly informative. The inexpensive production of large volumes of sequence data also adds to its advantage over other methods (such as PCR and arrays). Applications of NGS have ranged

from studies of whole-genome sequencing (studies of entire genomes), exome sequencing (investigations of smaller functional portions of the genome), RNA-seq (analysis of the transcribed genome) and ChIP-seq (protein-DNA binding sites). The widely used whole genome sequencing platforms include Illumina/Solexa (GA1, GA2, HiSeq 2000, HiSeq2500, MiSeq platforms), Ion torrent (PGM) and Roche (454-platforms). These are also referred to as second generation sequencers, and are capable of generating millions of DNA fragments using a highly parallel sequencing-by-synthesis process (Metzker 2010; Sboner et al. 2011). The overall cost of sequencing has immensely dropped, for example, the first human genome was sequenced at an approximate cost of $3 billion and it took 13 years to complete. In 2009, the cost had dropped to $100,000 and achieving the $1,000 genome looks feasible (Sboner et al. 2011). It currently takes approximately 11 days to sequence the *P. falciparum* genome on an Illumina HiSeq 2000. Several factors have contributed to this reduction in cost including improvement in experimental design and sample collection, sample sequencing, data reduction and management (Figure 1.3). As sequencing techniques improve, costs and stringency of DNA sample quality falls, as such new effective and cheaper methods of sample processing are developed while at the same time meeting sequencing threshold. CF11 columns have become common in processing *P. falciparum* samples (either cultured *in vitro* or processed directly from a patient without culturing) to deplete human DNA from parasitized blood and has been implemented in large genome-wide sequencing studies, especially in field sites with minimal facilities (Auburn et al. 2011). In case of low blood volumes or parasite densities resulting into less starting material, whole genome amplification (WGA) after DNA isolation has been employed to boost the starting material (Oyola et al. 2012; Kozarewa et al. 2009).

Table 1.2: Specifications of different Illumina sequencing methods and Ion torrent. Adapted from (Quail et al. 2012).

Platform	Illumina MiSeq	Illumina GAIIx	Illumina HiSeq 2000	Ion Torrent PGM
Instrument Cost	$128 K	$256 K	$654 K	$80 K
Sequence yield per run	1.5-2Gb	30Gb	600Gb	20-50 Mb on 314 chip, 100-200 Mb on 316 chip, 1Gb on 318 chip
Sequencing cost per Gb	$502	$148	$41	$1000 (318 chip)
Run Time	27 hours	10 days	11 days	2 hours
Reported Accuracy	Mostly > Q30	Mostly > Q30	Mostly > Q30	Mostly Q20
Observed Raw Error Rate	0.80 %	0.76 %	0.26 %	1.71 %
Read length	up to 150 bases	up to 150 bases	up to 150 bases	~200 bases
Paired reads	Yes	Yes	Yes	Yes
Insert size	up to 700 bases	up to 700 bases	up to 700 bases	up to 250 bases
Typical DNA requirements	50-1000 ng	50-1000 ng	50-1000 ng	100-1000 ng

As at 2013, Solexa/Illumina Hiseq (Illumina, Inc) technology is the most widely used platform and is capable of generating billions of bases of high quality DNA sequence per run in short stretches of DNA (paired-end reads), which are mapped on to a reference genome (usually 3D7 in *P. falciparum,* (Manske et al. 2012)), thus allowing accurate, reproducible and cost effective whole genome analysis (Table 1.2). Although it covers almost the entire genomic sequence of the parasite, it provides less and inaccurate measurement of mapped sequence in difficult stretches of DNA such as repeats and subtelomeric regions containing highly polymorphic genes. Thus, the repetitive nature of the *P. falciparum* genome and its biased base composition (high AT content) is a great challenge in sequencing, read alignment or assembly producing uneven read coverage across AT and GC rich regions which can lead to problems in variation analyses (Oyola et al. 2012). However, continual improvement in read length and optimization of sequencing library construction has greatly improved data output. The high amount of sequence data produced also require high computing power and capabilities which has been enhanced by the continual advancement in bioinformatics and new computational methods (some adapted from analysis of other

organisms) to help reduce the cost of data processing, storage and summaries of raw data into SNPs, CNV and INDELs.

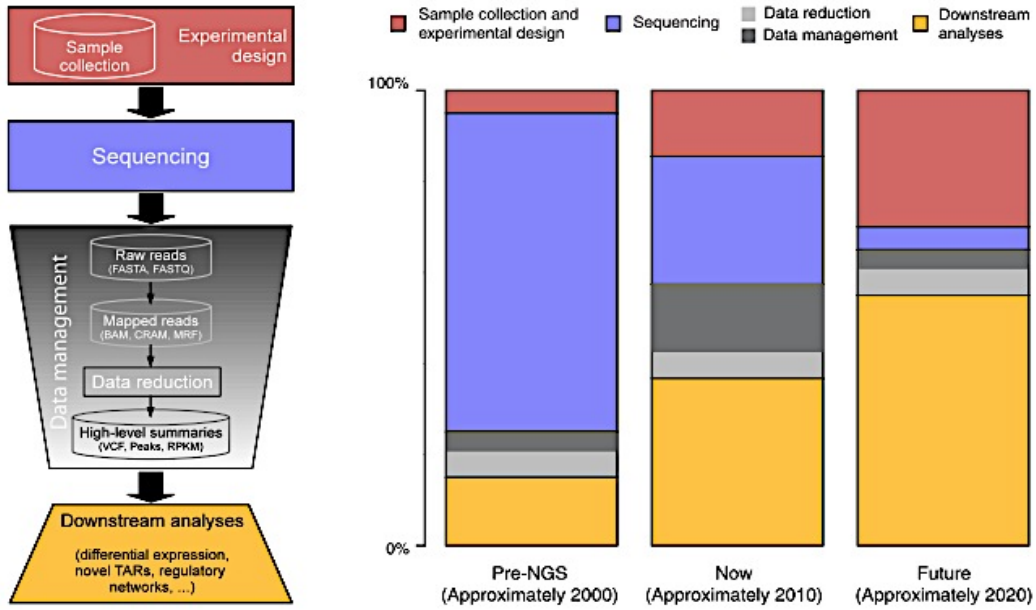


Figure 1.3: Left: the four major components of sequencing: experimental design and sample collection, sequencing, data management and downstream analysis. Right: relative impact of the four major components on sequencing over time. Adapted from (Sboner et al. 2011):

1.10 Illumina sequencing

The Illumina sequencing-by-synthesis technology uses cyclic reversible termination chemistry described by (Bentley et al., 2008) and involves three main steps namely library preparation, cluster generation and sequencing. The length of short sequences (reads) generated by the platform is determined by the number of cycles.

In library preparation, a sample DNA is fragmented using nebulization or sonication followed by an end-repairing step to generate blunt-ended fragments. A single nucleotide

Adenine (A) base is then added to the 3’ end of both DNA strands to produce A-tailed fragments that are then ligated with sequencing adaptors. The adaptors have a 3’ Thymine (T) overhang that complements the A-tails of template fragments. Only fragments containing the sequence adaptors are size-selected and quantified. The sequencing library is then transferred to flow cells that contain oligonucleotides that are complementary to the sequencing adaptors ligated at the end of the templates such that template fragments bind to the surface where both cluster generation and sequencing take place. By sequencing the template fragments from both ends, paired-end reads separated by a known fragment size are generated. There are eight lanes in Illumina’s flow cells that are capable of taking independent samples (libraries) or up to 96 multiplexed libraries.

In cluster generation step, fragments are first denatured into single stranded templates followed by cycles of a bridge-amplification step, where each template is clonally amplified into thousands of templates per cluster and millions of clusters per lane. Each cluster corresponds to a single template molecule.

The cyclic sequencing process involves addition of four fluorescently labelled nucleotides (dNTPs), to the growing chain of sequenced/synthesized bases for each template of a cluster. This hybridization is followed by measuring the intensity of the fluorescent dye via imaging techniques to determine the exact base for each cluster. Once done, the terminators containing the fluorophore are removed to allow for another cycle of hybridization. The signal from each template in a cluster is analysed using a base calling software called Bustard in order to establish the correct base call after each hybridization cycle.

1.11 *P. falciparum* evolution, population structure and LD

Studies of genetic variation in malaria parasites have helped to understand the parasite's adaptive mechanisms to natural and artificial selective pressure, which in turn has had practical significance in advancing vaccine development and surveillance of drug targets. Understanding genetic variation in *P. falciparum* requires knowledge of the parasite population structure, including variation in allele frequencies between populations and how these alleles remain correlated with neighbouring variants by LD. Principal component analysis (PCA), from which one can infer sample relatedness, has been used to determine population structure.

The extensive genetic variation in *P. falciparum* has also been used to deduce its evolution history (Table 1.3), with some studies inferring that *P. falciparum* separated from the chimpanzee parasite, *P. reichenowi*, 5 to 7 million years ago when the humans and chimpanzees diverged (Su et al. 2003). However, a more recent study concluded that it is of gorilla and not of chimpanzee, bonobo or ancient human origin (Liu et al. 2010) and that all known human strains might have resulted from a single cross-species transmission event which could have also resulted in a population bottleneck. What still remains unknown is the exact time when the gorilla *P. falciparum* entered the human population and whether the present ape population provided a source of recurring human infection (Liu et al. 2010).

Table 1.3: Tracing *P. falciparum* population history through genetic variation. Adapted from (Hartl 2004).

Type of genetic variation	Main advantages	Main disadvantages
Antigenic variation	Abundant	Strong selection for diversity to evade the host immune system
Microsatelllites	Abundant	High mutation rate; unknown pattern of mutation
Synonymous nucleotide sites	Low mutation rate; weak selective constraints	Possible selection against variation due to biased codon usage
Intron sites	Low mutation rate; weak sequence constraints	High A/T content; many microsatellites; possible selection against variation due to unrecognized selective forces
Upstream and downstream noncoding regions	Low mutation rate	High A/T content; many microsatellites; possible selection for diversity due to effects on gene expression
Mitochondrial DNA	Low mutation rate; no recombination	Possible selection on some mutations, and selective sweeps of the entire mitochondrial DNA molecule

By looking at genomic variation and structure, *P. falciparum* parasites have been genetically clustered according to their continental origins with major branches observed in Africa, Southeast Asia, Papua New Guinea and South America (Mu et al. 2005, 2010; Volkman et al. 2007; Manske et al. 2012). Within Africa, there are little population differences between countries compared to Southeast Asia and the western hemisphere where it is more pronounced even within a single country (Conway et al. 2001, 1999; Pumpaibool et al. 2009; Mu et al. 2005; Anderson et al. 2000; Manske et al. 2012). Extensive genetic diversity is observed in parasites from Africa and lowest in America and is reflected in observed patterns of LD that is also inversely proportional to the population recombination rate. Observed LD extends for shorter physical distances in Africa (approximately 1.5-kb), higher in Asian parasites and South America (approximately 16-kb) (Mu et al. 2005; Volkman et al. 2007). The difference in this LD also reflects the demographic history of these populations. For example, in high malaria transmission areas there is frequent occurrence of 'super-infection' with multiple genotypes produced through multiple bites from distinct *P. falciparum*-infected mosquitoes resulting in a high multiplicity of infection (MOI). This allows for mixing of genetically distinct gametocytes (through outcrossing) during the sexual phase and thereby producing short block of LD. By contrast,

in low malaria transmission areas, identical gametocytes and the lack of recombination will preserve LD (Figure 1.4) (Conway et al. 1999; Volkman et al. 2012). High LD outside Africa is thought to be as a result of population bottlenecks that eliminate many allele combinations leaving strong correlations (and high LD), a smaller effective population size can only sustain fewer allele combinations.

MOI also has some significance in malaria pathogenesis (D'Alessandro 1997). Human malaria parasites in low endemic areas have evolved a lower level of virulence (due to less co-infection), than those in high endemic areas (Conway et al. 2007). Contrary to this, another study of controlled comparisons within populations in high endemic areas also showed that different numbers of genotypes in multiple infections are not generally associated with different clinical symptoms or severity of malaria (Robert et al. 1996; Conway and McBride 1991; Kun et al. 1998). In fact, they showed that the numbers of genotypes per infection were similar between clinical groups, with severe cases often having only a single clone (Robert et al. 1996). Another study of post-mortem analysis of parasites in multiple organs of Malawian children observed that cerebral malaria cases are less genetically mixed (Montgomery et al. 2006).

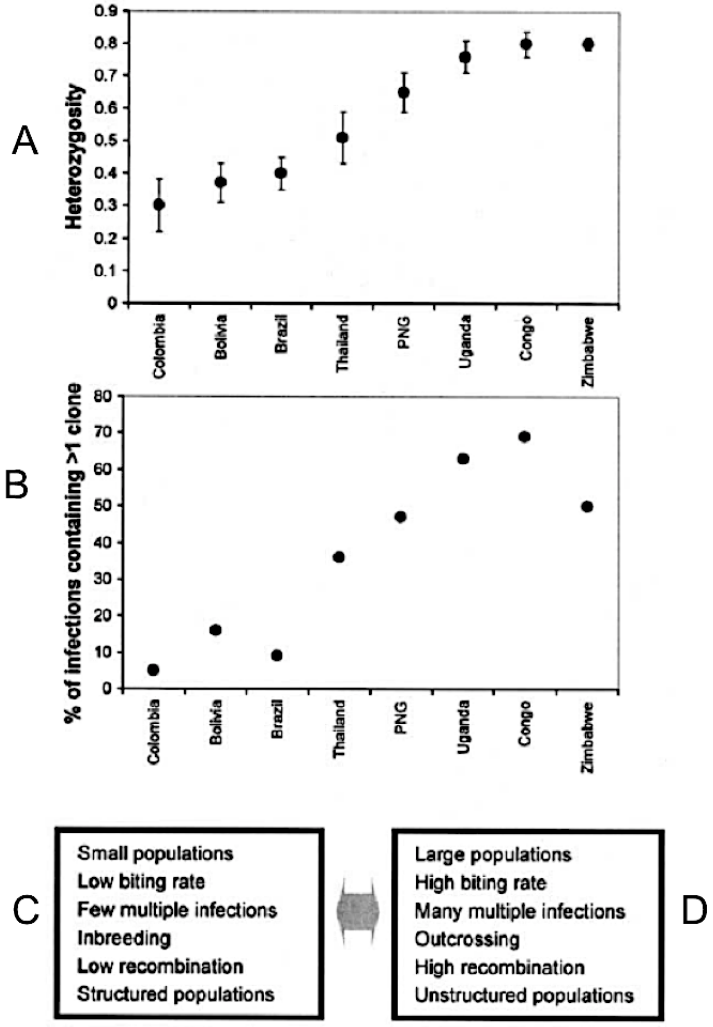


Figure 1.4: *P. falciparum* population genetics and multiplicity of infections. (A): Levels of heterozygosity (the probability that two randomly drawn alleles are different at a locus) rare in South America, intermediate in SE Asia and highest in some African populations. Heterozygosity measures levels of genetic diversity, which correlates to MOI – and is highest in Africa (B) Multiplicity of infections. Percentage of infections containing multiple clones higher in Africa (reaching 70% in Congo) than in South America (<20% in Columbia, Bolivia and Brazil) and South East Asia populations (approximately 40 to 50% in Thailand and Papua New Guinea). (C-D) Dynamics of population genetics between low malaria transmission, C and high transmission settings, D. Low transmission areas are characterised by small effective population size, low biting rates, few multiple infections, inbreeding, low recombination and detectable structured populations the opposite is true for high transmission area. Adapted from (Anderson et al. 2004).

1.12 Signatures of selection in *P. falciparum* genome

P. falciparum has two distinctive hosts, human and mosquito. The major sources of selection acting on the parasite in its hosts are immunity, drugs, vector availability, host death and co-infection (Mackinnon and Marsh 2010). These selective forces have an impact on the parasite's transmissibility and so it must adapt amidst this highly heterogeneous environment in order to survive. In the human host, the parasite is subject to the strongest selection from immune responses and drugs and it undergoes severe population bottlenecks. The two most widely studied classes of selection in *P. falciparum* are balancing and directional selection which are discussed below.

1.12.1 Balancing selection and identification of vaccine targets in *P. falciparum* genome

A major factor in the evolution of malaria parasites is to evade host immune responses. Parasite antigens recognised by the host immune system will usually evolve under frequency dependent immune selection (where rare alleles are favoured simply because they are of low frequency), exhibiting extensive genetic polymorphism coupled with a bias for non-synonymous mutations that is maintained by balancing/diversifying selection. In an infection, host variant-specific antibodies raised against an antigen will afford a new variant a competitive advantage which in turn will rise in frequency enough to induce production of specific antibodies against itself and thus causing diversifying selection. This strategy therefore offers the parasite the potential to encode immunological identities that are kept at intermediate frequencies. The distinctive patterns of genetic polymorphism arising from immune selection should depart from those predicted under

neutral evolution models and can be tested using population genetic methods that test for signatures of balancing selection.

1.12.2 Testing for balancing selection using Tajima's *D* test

Developed by and named after a Japanese researcher Fumio Tajima, Tajima's *D* is a widely used test of neutrality (Tajima 1989). It tests the hypothesis that all mutations are selectively neutral and of no importance. The purpose of this test is to therefore look for departures from this neutral model to distinguish between a DNA sequence evolving randomly ("neutrally") and one evolving under non-random process including balancing and directional selection, population expansion and sub-division and genetic hitch-hiking. Tajima's *D* test is based on the differences between the number of segregating sites (*S*) and the average number of pairwise nucleotide differences (π). Under neutral/null hypothesis model, expectations of S and π are,

$$\begin{aligned} E[\pi] &= \theta \\ E[S] &= a_1\theta \end{aligned}$$

Where θ = 2*N*μ, 2N is the haploid population size and μ is the mutation rate per generation and a_1 defined as,

$$a_1 = \sum_{i=1}^{n-1} 1/i$$

Under the neutral model of DNA sequence evolution the number of segregating sites (*S*) and pairwise differences (π) equals to ϑ ($S/a_1 = \vartheta$ and $\pi = \vartheta$) so they are statistically indistinguishable from one another. If they are not then they deviate from the neutral

model and their difference is the basis of Tajima's *D* test. However, S/a_1 and π vary from sample to sample and this variance is accounted for in the equation below:

$$V = \mathbf{Var}[\pi - S/a_1]$$

Where *V* denotes the sampling variance of the difference between the two estimates. Tajima's *D* is then calculated as:

$$D = \frac{\pi - S/a_1}{\sqrt{V}}$$

Tajima's *D* is therefore based on the difference between S/a_1 and π. High sequence diversity and excess polymorphism/rare alleles selected for and maintained at intermediate frequency increases π above the neutral expectation, more than *S*, so *D* is positive ($\pi - S/a_1$) > 0 (Figure 1.5A) and this suggests balancing selection. However, population admixture can also lead to positive *D*. Low sequence diversity and excess/high frequency of rare alleles for example after a selective sweep or selection against genotypes carrying deleterious variants reduces π, so $D < 0$ (Figure 1.5C). Otherwise, under neutrality, $D = 0$ (Figure 1.5B). Thus, under positive diversifying selection, an excess of intermediate frequency polymorphisms and lower number of singletons make the value of Tajima's *D* positive, with values greater than +2 or less than -2 likely to be significant. However, this is never a critical value of significance and as such *D* values that greatly deviate from the bulk of empirical distribution are thought to be interesting.

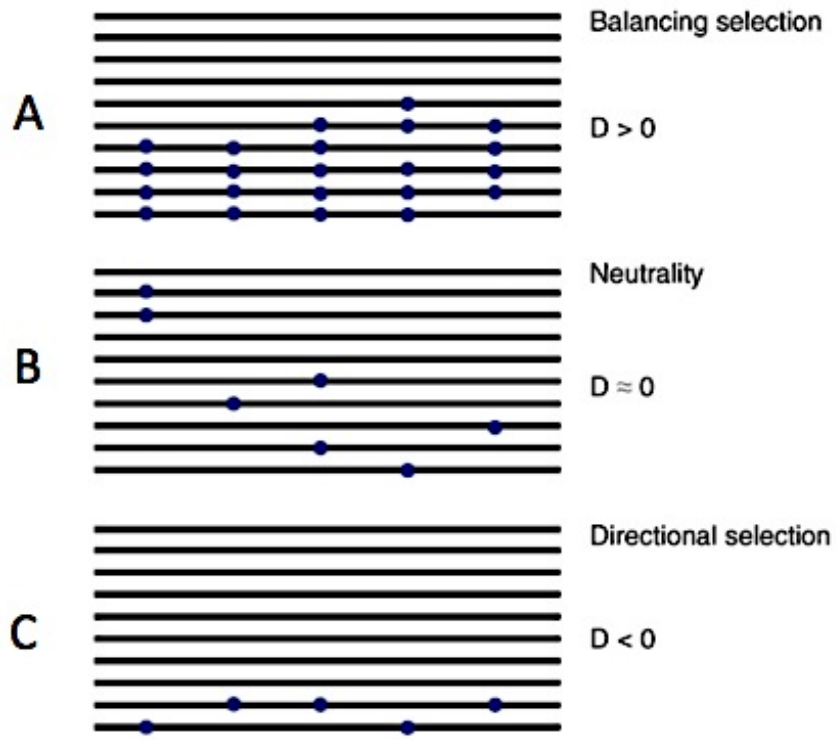


Figure 1.5: Principle of balancing selection and Tajima's *D*. Horizontal lines represent segregating sites (sequences) and dots represent nucleotides variants. (A) High allelic diversity, π >S: *D* >0. (B) Neutral genetic drift, *D*=0. (C) Low allelic diversity, new polymorphisms rare, and π <S: *D* <0. Adapted from (Weedall et al. 2010).

1.12.3 Testing for balancing selection and population differentiation using F_{ST}

Wright's F-statistics, especially F_{ST}, is a measure of population differentiation due to the structure of genetic variation within and between populations. Estimates of F_{ST} can identify genomic regions that are target of selection, and comparing F_{ST} from different genomic regions can infer demographic history of those populations. It is directly related to the variance in allele frequencies among populations and conversely to the resemblance among individuals within populations. F_{ST} is calculated as:

$$F_{ST} = 1-(H_S/H_T)$$

Where, H_T = total gene diversity or expected heterozygosity in the total population as estimated from the pooled allele frequencies and H_S = average expected heterozygosity estimated from each sub-population

If F_{ST} is small (0 – 0.05), it means allele frequencies within each population are very similar/maintained against the tendency for neutral drift to decay, little genetic differentiation; if it is very large (> 0.25), the allele frequencies are very different; F_{ST} is moderate (0.05 – 0.15); F_{ST} is large (0.15 – 0.25). Very large F_{ST} indicate positive directional selection rapidly fixing an advantageous allele in a population thus showing low diversity and high divergence (Figure 1.6B and 1.7). Very low F_{ST} potentially indicate loci evolving under balancing selection showing elevated diversity with low divergence (Figure 1.6C and 1.7). The genes under balancing selection are unlikely to diverge as rapidly as others because the selection prevents differences from differentially fixing between populations.

1.12.4 Other tests of balancing selection

Other tests of balancing selection include Fu and Li's *F* and *D* (Fu and Li 1993; Fu 1997), which is quite similar to Tajima's *D*, McDonald and Kreitman's (*MK*) test (McDonald and Kreitman 1991), Hudson, Kreitman and Aguade's (*HKA*) test (Hudson et al. 1987), *dN/dS* and $\pi N/\pi S$ (Nei and Gojobori 1986). *dN/dS* and $\pi N/\pi S$ compare non-synonymous (NS) nucleotide differences per NS site (dN, πN) to their synonymous equivalents (dS, πS) between species (d) or within species (π) (Weedall and Conway 2010). Because most NS mutations are deleterious $dN/dS < 1$. However, if NS mutation is advantageous (e.g., in the case of an anti-drug allele) it rapidly becomes fixed in the population than neutral ones, then $dN/dS > 1$. Within a species, high recombination rates will make sites evolve independently of one another. In the case of immune selection, balancing selection will

maintain NS mutations for longer than synonymous (S) leading to excess of πN compared to πS. *MK* test compares ratios of NS to S changes within and between species over long periods (Weedall and Conway 2010). *HKA* tests compares ratio of polymorphism to divergence for different loci to detect those with abnormal ratios.

From analyses of Tajima's *D* and $F_{ST,}$ antigens and vaccine candidates have been identified including AMA1 (Polley and Conway 2001; Mu et al. 2007), MSP3.8 and TRAP (Amambua-Ngwa et al. 2012a).

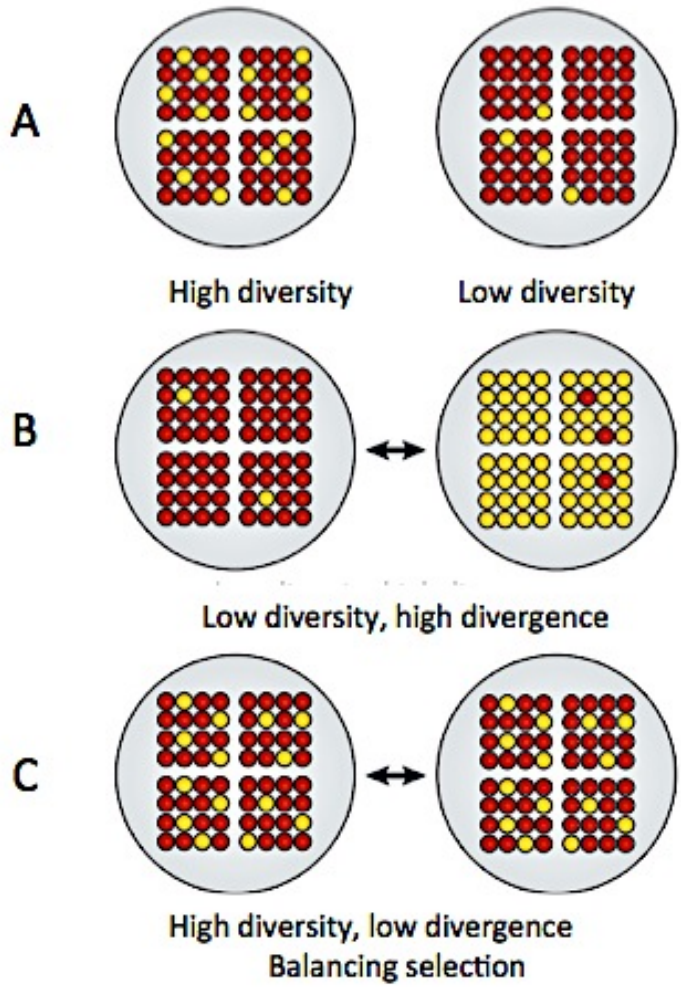


Figure 1.6: Signatures of selection. Large grey circles represent a given population, square matrix of circles represents individuals and red or yellow circles, alleles. (A) Diversity refers to amount of allelic variation among individuals in a given population and divergence refers to amount of variation between populations. (B) Positive directional selection: low allelic diversity and high divergence. (C) Balancing selection: high allelic diversity and low divergence. Adapted from (Volkman et al. 2012).

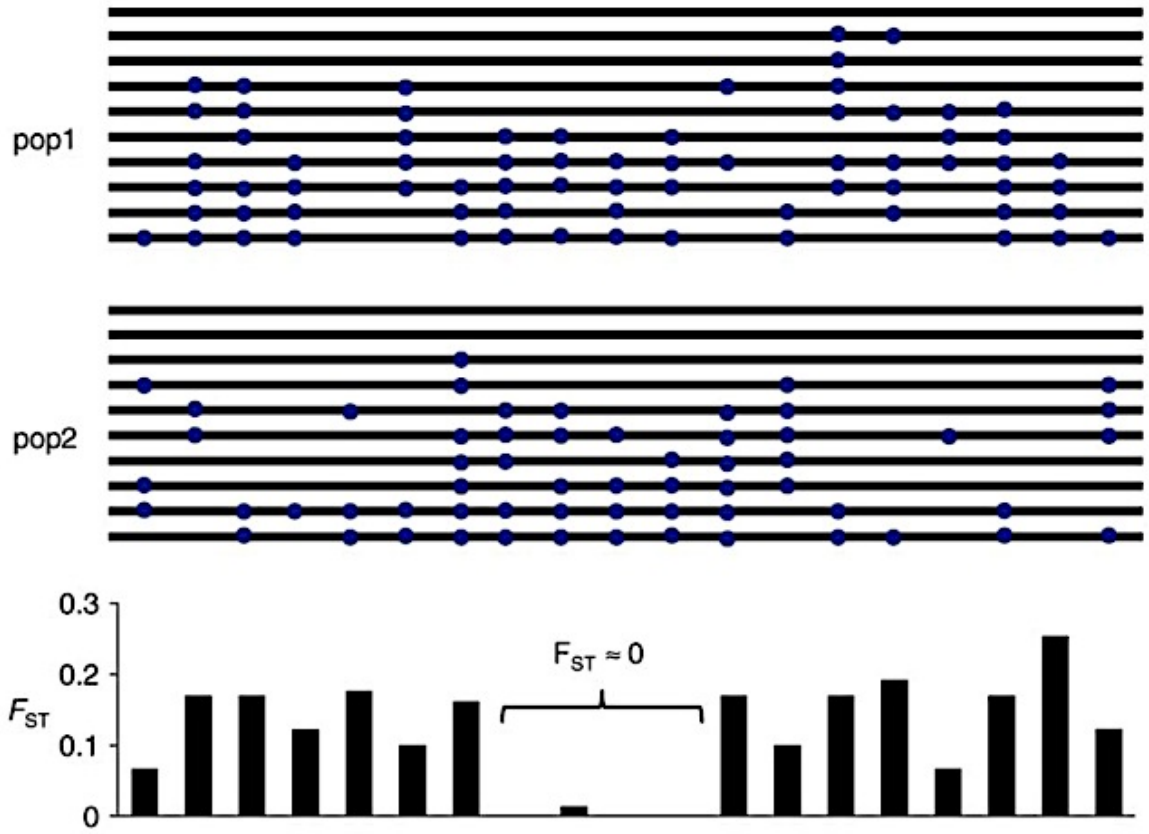


Figure 1.7: Schematic representation of $\mathbf{F_{ST}}$ in two populations. Dots represent nucleotide variants. Allele frequencies in the central region are maintained in both populations by balancing selection, resulting in very low $\mathbf{F_{ST}}$. Detectable between population allele frequencies variation increase $\mathbf{F_{ST}}$. Adapted from (Weedall and Conway 2010).

1.13 Positive directional selection and drug resistance in *P. falciparum*

In *P. falciparum* populations, introduction of a successful drug usually results in a strong positive directional selection on the parasite population. To overcome the drug pressure, the parasite undergoes evolution resulting into emergence of drug resistant alleles. An advantageous allele will spread rapidly (increase in frequency) in the population and become fixed. As this happens neighbouring alleles also hitchhike to high frequencies and spread through the parasite population and remain in LD with the resistance allele, eventually leading to valleys of reduced allelic variation (selective sweeps) with increased LD in the regions surrounding the resistance loci (Figure 1.8). After drug selection, recombination and mutation gradually restore variation and break down LD.

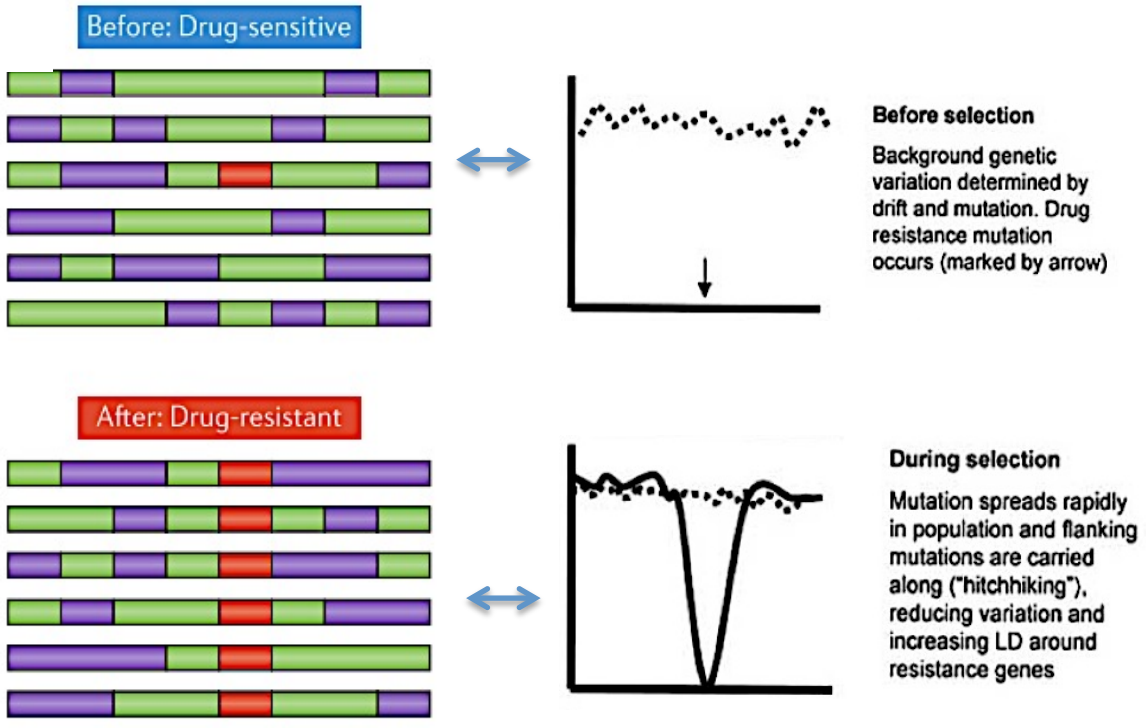


Figure 1.8: Drug selection effect. Results in directional selection leaving a distinctive selective sweep, consisting of reduced polymorphism and enhanced LD. Adapted from (Volkman et al. 2012).

Patterns of genetic variation around drug resistant loci have been revealed by assessing LD, and tests of positive selection such as F_{ST} and 'haplotype-based' tests, e.g., integrated haplotype test (*iHS*), long-range haplotype test (*LRH*) and cross population extended haplotype homozygosity test (*XP-EHH*) (Sabeti et al. 2002, 2007; Voight et al. 2006). There is generally short LD in *P. falciparum* due to high recombination rates (Jiang et al. 2011). But in response to drug pressure long haplotypes with extended LD arise around the advantageous variant (Figure 1.8 and 1.9).

Genome scans for selective sweeps have been performed by several studies and have identified both known and unknown drug targets; some of the most notable and referenced studies are discussed below. Wotton *et al* analysed 342 polymorphic microsatellite markers from 87 parasites from South America, South East Asia, Africa and Papua New Guinea and observed that mutations in *pfcrt* evolved on four independent occasions (2 origins in S. America, single origin in Asia that spread to Africa and one in Papua New Guinea) and that these mutations spread across the globe through selective sweeps (Wootton et al. 2002).

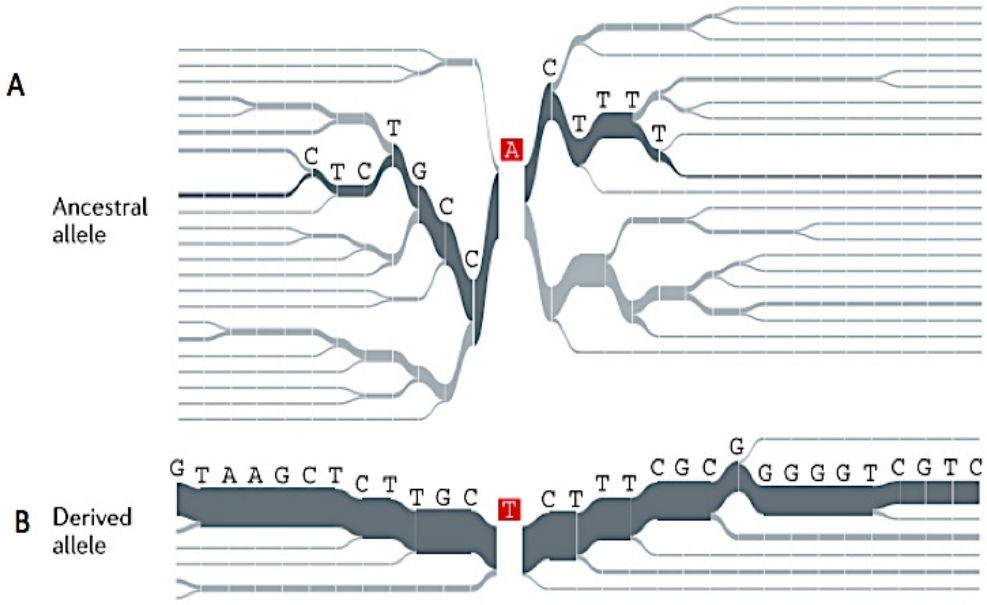


Figure 1.9: Effect of positive selection on haplotype structure visualized using a haplotype bifurcation diagram. During a drug selection (causing A to T SNP shift) long-range haplotypes would arise similar to B and stand out from the normal haplotype structure of the genomic background, A. Adapted from (Volkman et al. 2007).

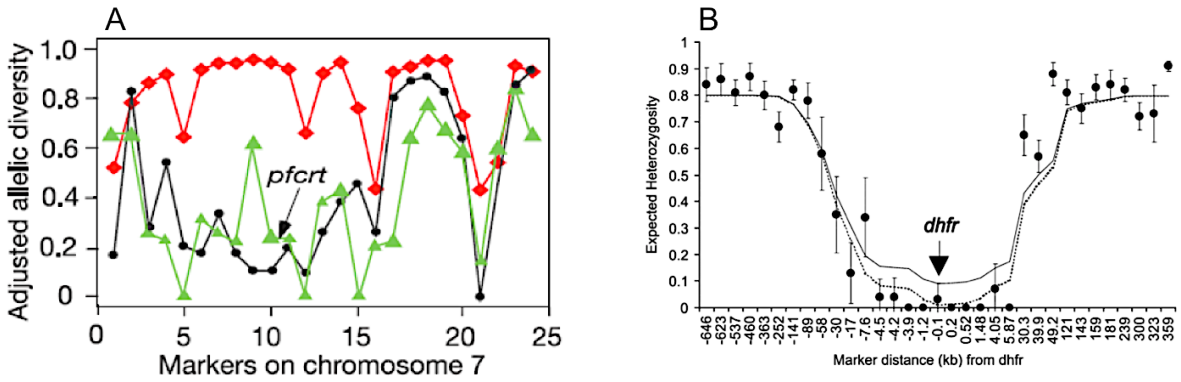


Figure 1.10: Drug selection effect. Reduced allelic/microsatellite diversity around; (A) *pfcrt* in CQR isolates (Black line, Asia and Africa; green line, South America) compared to CQS isolates from Asia and Africa (red line) and (B) 100-kb around the *pfdhfr* locus on chromosome 4 of *P. falciparum* isolates from Thailand-Myanmar border. Adapted from (Wootton et al. 2002; Nair et al. 2003).

By comparing genetic variation in CQR and CQ sensitive (CQS) parasites from these locations, he observed a dramatic reduction in genetic diversity and extensive *LD* around

pfcrt in CQR parasites (Figure 1.10A) (Wootton et al. 2002). The following year, Nair *et al*, also used the same principle to examine MS variation across chromosome 4 while trying to study the impact of PYR resistance around the *pfdhfr* gene. They discovered reduced MS length variation in a 12-kb region flanking the *pfdhfr* and diminished variation for approximately 100-kb around the same gene (Figure 1.10B) (Nair et al. 2003). More recently, genome-wide scans of selective sweeps across *P. falciparum* genomes have been performed to describe regions under positive directional selection. These studies have been successful in validating previously identified anti-drug loci and in detecting novel loci. Mu *et al* used the molecular invasion probe array and tests for selective sweeps (*REHH*, *iHS* and *XP-EHH*) to generate genome-wide maps of selection from parasite populations from Asia, Africa and America and identified multiple loci under significant positive selection including *pfcrt, pfama1, pfmdr1, surfin* (Mu et al. 2010). Additional studies using similar tests have concurring results (Van Tyne et al. 2011; Amambua-Ngwa et al. 2012b; Park et al. 2012).

1.13.1 Identifying positive selection using *EHH, iHS* and *XP-EHH*

Metrics that are used to identify selective sweeps probe for regions with reduced haplotype diversity. The Extended haplotype homozygosity test (*EHH*) measures reduction in haplotype diversity by computing the probability as a function of distance from the core region that two randomly chosen chromosomes that share the core allele are identical (Sabeti et al. 2002). It heavily relies on the relationship between an allele's frequency and the extent of LD surrounding it, such that under neutral evolution a new variant will require a longer time to reach high frequency and recombination will substantially break down LD around it. Under positive selection, the new variant will rapidly reach high frequencies over a short time that recombination does not substantially break down long-range association

or LD with other SNPs. Because of this, *EHH* needs to control for local recombination rates. This is usually achieved by genotyping dense SNPs and collecting samples from same locality. *EHH* will therefore involve finding a core haplotype with a combination of high frequency and high *EHH* score compared to the core haplotypes at the same locus (Sabeti et al. 2002).

For bi-allelic SNP with alleles a (ancestral) and A (derived), the *EHH* is computed as follows:

$$EHH(x) = \frac{\sum_{i=1}^{h_x} \binom{n_i}{2}}{\binom{n_a}{2} + \binom{n_A}{2}}$$

Where n_a and n_A are number of haplotypes with alleles a and A respectively, n_i is the count of the i^{th} haplotype in a population and h_x is the number of distinct haplotypes in a genomic region up to a distance x from the locus.

Integrated haplotype score (*iHS*) compares log of the ratio of the integrated *EHH* score (profiles) for haplotypes containing the ancestral allele to the integrated *EHH* for haplotypes containing the derived allele. Under neutral decay, rates of *EHH* for ancestral allele and derived allele, $EHH_a/EHH_A \approx 1$ and $iHS \approx 0$. High values of *iHS* indicate a reduced and extended haplotype diversity with a slower fall-off *EHH* around selected allele. Large negative *iHS* values indicate unusually long haplotypes containing the derived allele while large positive values indicate long haplotypes containing ancestral allele (for example when selection favours ancestral allele/s or if they hitchhike within the 'sweep' region). Identifying regions that contain clusters of high scoring SNPs gives a better indicator of a selective

sweep because selective sweeps produce clusters of extreme scores across the swept region (Voight et al. 2006).

$$\text{Unstandardized } iHS = \ln(EHH_a/EHH_A)$$

The *iHS* is a local measure of selection and therefore both the ancestral and derived allele will share same genomic environment, thus making it insensitive to variations in population demography such as recombination rates, demographic history and bottlenecks. It's also very reliable when one cannot find a good reference, but has low power to detect alleles that are fixed or close to fixation. However, *XP-EHH* (Sabeti et al. 2007), which compares *EHH* profiles between two populations at the same SNP, is insensitive to allele frequencies, more reliable when a reference population with similar demographics is available and has high power to detect alleles that are fixed in one of the populations in comparison. Computing *XP-EHH* requires computation of *EHH* in each population and is defined as follows:

$$unstandardized\ XP-EHH = \log\left(\frac{\int_D EHH_{pop1}(x)dx}{\int_D EHH_{pop2}(x)dx}\right)$$

Where *pop1* and *pop2* represent the two populations.

1.14 Rationale and objectives

Malaria is a major cause of morbidity and mortality in Malawi, mostly in children under five of age who are particularly at risk of developing severe disease. Genetic variation in *P. falciparum,* is key to its pathogenesis, as it allows the parasite to evolve rapidly and overcome host immune systems, drugs and vaccines, all of which generally aim to curtail infections and reduce the parasite's transmission potential (Volkman et al. 2007; Kidgell et al. 2006; Mackinnon and Marsh 2010). Therefore, understanding and characterizing genetic variation will help explore the complex nature of host-parasite interaction, co-evolution and the parasite response to malaria interventions. Since the first full genome sequence in 2002, rapid developments in genomics and genetics have accelerated studies of *P. falciparum* genetic diversity and several genotyping platforms have demonstrated a rich diversity from SNP discovery, identifying loci potentially under immune and drug selection (Gardner et al. 2002; Kidgell et al. 2006; Mu et al. 2007, 2010; Volkman et al. 2007). It has been observed that positive natural selection at drug-resistant loci (*pfdhps, pfcrt*) causes reduction in allelic diversity while diversifying selection produces highly polymorphic regions at candidate vaccine targets such as *ama1*, *msp1*, *msp2*, *csp* and *eba175* (Nair et al. 2003; Wootton et al. 2002; Polley et al. 2003; Baum et al. 2003; Ferreira et al. 2003). Immune evasion and emergence of drug resistant genotypes are largely driven by acquisition of genetic changes in the parasite genome.

This thesis describes a study designed to observe *P. falciparum* genetic population changes and identification of CNV, with the following features:

1 Children aged between 6-48 months were recruited at one site, Chikwawa (a high transmission area), at a time when maternal immunity to malaria is waning or finished, and at 0-2 months in another site, Zomba (a relatively low malaria transmission but with neighbouring high transmission areas) when the children were still under maternal protection. Therefore, the age range would cover a fairly large variation in immune response with some children near immunological naivety to malaria and could potentially be infected with malaria parasites of any genetic background. However, for this thesis, no Zomba samples were included, as they had not been sequenced at the time of analysis.

2 The longitudinal nature of the study allowed us to investigate multiple malaria episodes in these children. As they develop resistance to particular genetic types, we hypothesise that each malaria episode will consist of parasites of different genetic backgrounds.

3 Recent rollout of LA, bed-net and IRS schemes will put pressure on the parasite population, potentially causing population bottlenecks. Parasites of particular genetic backgrounds will only survive these pressures through advantageous phenotypic traits or through favourable circumstances.

In this study, we aimed to use MPS, leveraging Illumina short read technology:

1. To sequence uncultured *P. falciparum* paediatric isolates obtained from a 3-year longitudinal study of a Malawian population after a selective sweep with SP and continuous selective pressure from intensive use of LA, ITNs and IRS;
2. To identify genetic variants including SNPs, and large structural variants (e.g., CNV, INDELs) and provide a map of genomic variation;

3. To use the SNP variation, to identify regions under selection pressure specific to this *P. falciparum* population;
4. To compare Malawi genetic variation to other populations of disperse origins (Kenya, Burkina Faso, Mali, Thailand and Cambodia);

The sequencing was performed at the Wellcome Trust Sanger Institute (WSTI).

Chapter 2

Whole-genome scans for selection and changing antimalarial drug pressure in Malawi *Plasmodium falciparum* clinical isolates

2.1 Introduction

An estimated 3.3 billion people are at risk of malaria with sub-Sahara Africa having the majority of cases (81%) and deaths (91%), mostly in children less than five years of age and pregnant women (WHO. World Malaria Report 2012). Malawi grapples with a heavy burden of endemic falciparum malaria with year round transmission that peaks during the long rainy season from early December to May (Ewing et al. 2011). Since 2005, use of ITNs in Malawian households has reached ~60% saturation, and the government's campaign for IRS has targeted transmission hotspots. ACT were adopted as first line treatment for uncomplicated malaria in 2007, replacing SP (WHO. World Malaria Report 2012; Roca-feltrer et al. 2012). Despite the scaling up of interventions, malaria still accounts for up to 30 - 40% of all outpatient visits, and childhood cases between 2001 and 2010 (Roca-Feltrer 2012) have not declined. This chapter assesses evolutionary driven genetic changes in the local *P. falciparum* population that may both reflect and affect malaria control interventions, and provide a baseline genetic characterisation of variation and selection to inform further surveys of anti-malarial resistance and population structure.

Genetic variation in *P. falciparum* is central to its survival and threatens to undermine malaria control interventions. This evolutionary process enables parasite populations to rapidly overcome host immune responses and antimalarial drugs to establish persistent

infections and increase transmission (Volkman et al. 2007; Kidgell et al. 2006; Mackinnon and Marsh 2010). Surveying evolutionarily driven genetic changes in *P. falciparum* and investigating parasite responses to antimalarial interventions are therefore crucial to efforts to reduce the malaria burden. Since the first *P. falciparum* genome sequence was completed in 2002, rapid advances in genomics and genetics have accelerated studies of *P. falciparum* genetic diversity. Several genotyping platforms have discovered a rich diversity of SNPs, which have helped to identify loci potentially under drug or immune selection (Gardner et al. 2002; Kidgell et al. 2006; Mu et al. 2010, 2007; Volkman et al. 2007). While positive natural selection at drug-resistance loci (e.g., *pfdhfr* and *pfcrt*) reduces allelic diversity, diversifying selection produces highly-polymorphic regions in genes encoding targets of naturally-acquired immunity and thus of candidate vaccine antigens (e.g., *ama1*, *msp1*, *msp2*, *csp* and *eba175*) (Nair et al. 2003; Wootton et al. 2002; Polley et al. 2003; Baum et al. 2003; Ferreira et al. 2003).

Here I have used massive parallel sequencing (MPS, Illumina short read) technology to identify >100,000 SNPs in an initial set of 93 sequenced clinical *P. falciparum* isolates from Malawi. These parasites were obtained directly from children in a 3-year longitudinal study of a Malawian *P. falciparum* rural population in the Chikwawa district, one of twelve sentinel sites in the country chosen for intensive antimalarial intervention. Samples were collected between December 2010 and July 2011 and patients were young (average age 11 months), with moderate parasitemia (1,000 - 10,000 parasites/µL) and only slightly anaemic with an average haemoglobin concentration of 9 g/dl. All samples except two were uncomplicated malaria cases. In analysing SNP variation, this sought to identify signatures of selection for potential targets of antimalarial drugs, host immune responses and vaccines.

Specifically, we used the allele frequency-based Tajima's *D* approach (Tajima 1989) to detect balancing selection, and the integrated Haplotype Score |*iHS*| to detect selective sweeps (Voight et al. 2006).

2.2 Materials and Methods

2.2.1 Ethics statement

This project was approved by the College of Medicine Research Ethics Committee, University of Malawi (Ref. no. P.10/09/831) and the Liverpool School of Tropical Medicine. All samples were collected with written informed consent from a parent or guardian of the patients. Participants in this study were part of an ACTia clinical trial (Reference number P.10/08/707) in Chikwawa and MaRCH study (Reference number P.05/10/954) in Zomba district.

2.2.2 Study area

This study was carried out in Chikwawa and Zomba districts of Malawi (Figure 2.1) (which generally receives 763 to 1,143 mm rainfall *per annum*). Chikwawa falls within the lowland zones of Malawi, in the lower Shire valley where the altitude is a few metres (approximately 50 metres) above sea level. During the rainy season (December to May), rising of water leads to formation of ponds and marshes, also present are permanent wetlands with stagnant water and together with temperatures that are generally above 21 degrees Celsius (with mean annual temperatures at 25 degrees Celsius), Chikwawa provides good breeding sites for mosquitoes. This leads to high malaria transmission with peaks between December and May. The study area covered 24 villages within a radius of 10 kilometres in Chikwawa. Recruitment of patients (which was part of a larger study in this

area, ACTia study) was continuously done from 2010 – 2011. Samples were collected from 2010 – 2013 and covered three malaria seasons.

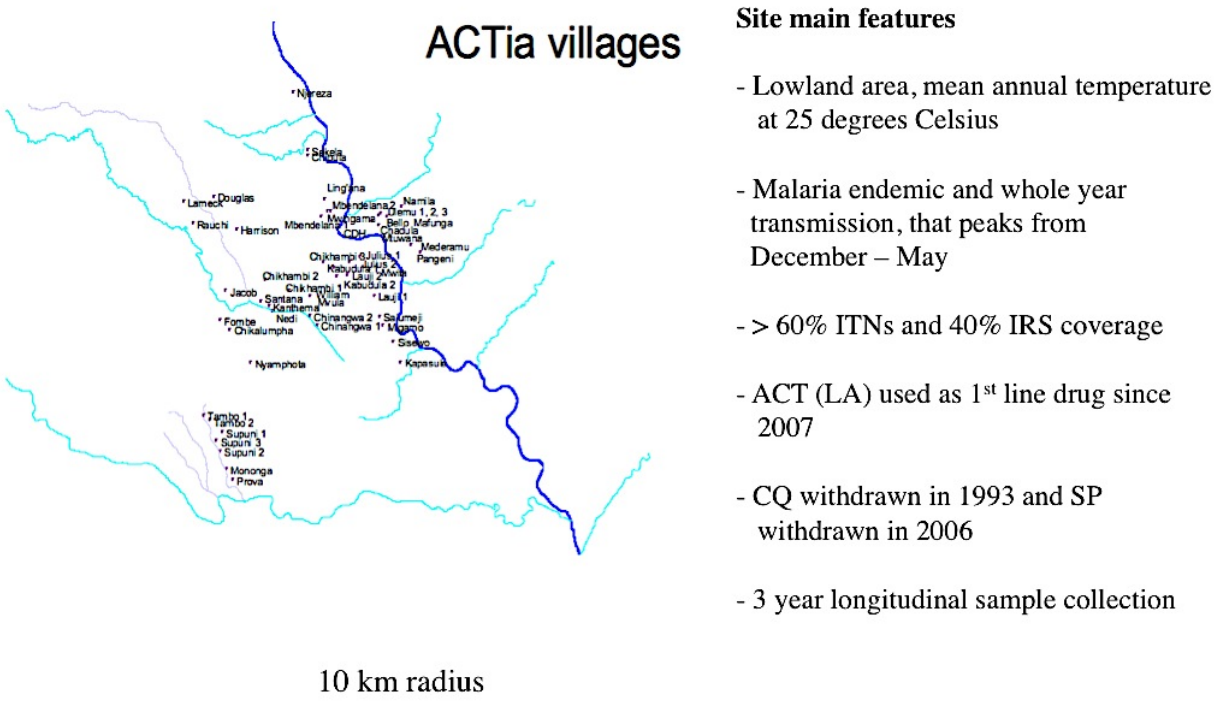


Figure 2.1: Study area showing location of the Chikwawa district. Key features of this study site are described.

Zomba district falls partly within the highlands and lowland zones of Malawi (Figure 2.2). Higher rainfalls are experienced in Zomba where malaria transmission is perennial but increases during the rainy season when pools of stagnant water are created. Between 2010 and 2013, 40 samples were collected from this location but none included in this analysis, as they had not been sequenced.

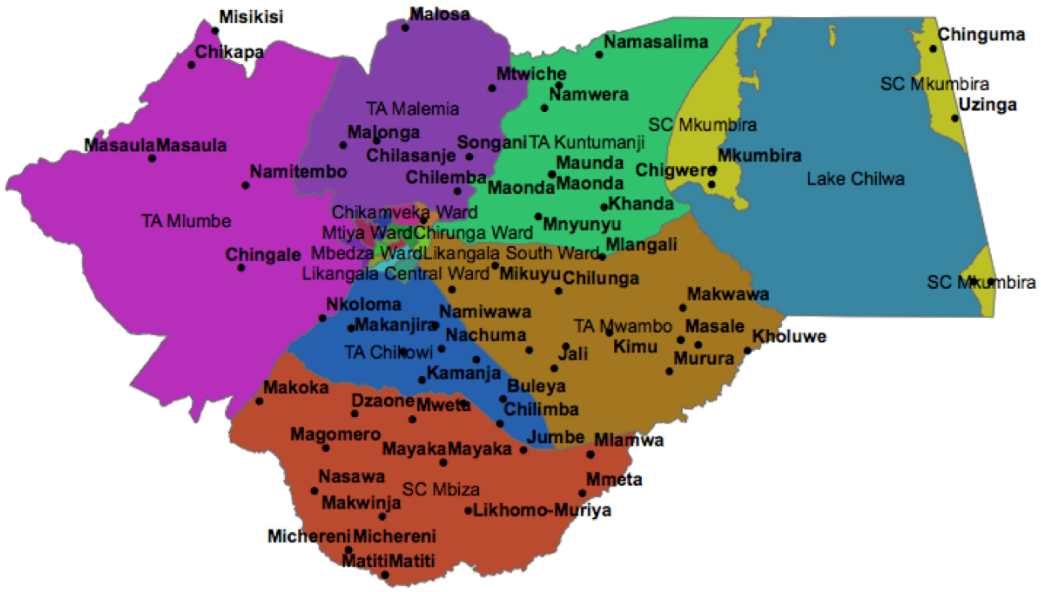


Figure 2.2: Map of Zomba district showing study site. Most samples were collected from TA Mwambo, TA Kuntumanji, TA Chikowi and TA Mkumbira villages.

2.2.3 Type of study

This is an ongoing prospective cohort study nested within a randomised control trial (ACTia) and a cohort observational study (MaRCH). Within ACTia, participants were recruited between 6-48 months and randomized treated with LA or dehydroartemisinin piperaquine (DHA-PPQ) for any confirmed clinical malaria episode. When a participant presented with a febrile illness, a malaria smear was made from a finger-prick blood sample and, if positive, treated according to the study arm. In MaRCH, two cohorts of participants (aged 0-2 months) - an HIV exposed group on cotrimoxazole (CMX) prophylaxis for 12 months and a non-HIV-exposed group - were followed up until 2 years. 500 HIV exposed and 500 non-HIV exposed were recruited simultaneously from the area of residence of each HIV exposed child. Following blood smear reading, any participant with a minimum parasite

density of 100 – 1000 parasites/µL was included in this study. Clinical information on all participants was kept within rules and regulations of both ACTia and MaRCH studies.

Experiments performed at Malawi-Liverpool-Wellcome Trust Labs by myself:

2.2.4 Materials, equipment, reagents and chemicals

- Whatman CF11 cellulose powder (Whatman catalogue no. 4021- 050)
- Whatman Grade 105 lens cleaning tissue (100 x 150 mm, Whatman catalogue no. 2105-841)
- 10 cc plastic syringes, centred (Becton- Dickinson catalogue no. 309604)
- RPMI-1640, with HEPES and L-Glutamine (Invitrogen catalogue no. 21875-091)
- 15 ml centrifuge tubes (Fisher catalogue no. FB55950)
- 5 ml pipettes (Fisher catalogue no. FB51889)
- 10 ml pipettes (Fisher catalogue no. FB55484)
- Cryotubes (Fisher catalogue no. 12-565-298)
- Needles, 23G (NHS Supplies catalogue no. FTR052)
- Vacuette, EDTA, 4 ml (NHS Supplies catalogue no. 454009)
- Blood collection set, 23g with push button adaptor (NHS Supplies catalogue no. KFK323)
- Pipetboy Acu Pipette controller (Fisher catalogue no. PMR-100-070H)
- 10X PBS (Life technologies catalogue no AM9625)
- Centrifuge
- Water bath
- Gloves

- QIAGEN blood midi-kit (Qiagen catalogue no. 51185)
- 1-200 µL tips (Fisher catalogue no. FB34551)
- 1000 µL tips (Fisher catalogue no. FB34611)
- DNA LoBind tubes, 1.5 ml (Fisher catalogue no. TUL-145-020D)

2.2.5 Blood sample collection

During any confirmed clinical malaria episode, 3 mls of whole blood at a minimum parasite density of 100-1000 parasites/µL was collected intravenously in EDTA vacutainer prior to patient treatment with either LA or DHA-PPQ (ACTia), or CMX (MaRCH) under the two study guidelines. Vacutainers were labelled with study ID and patient unique identifier and inverted gently to mix well. These samples were stored at 4 degrees Celsius until time for white blood cell depletion, within 24 hours to avoid haemolysis and release of human DNA. A total of 400 malaria positive paediatric samples were collected by June 2013.

2.2.6 WBC depletion of whole blood using CF11 column (Sriprawat et al. 2009)

2.2.6.1 Purpose and scope

The purpose of this procedure was to reduce the amount of human DNA in *P. falciparum* infected whole blood samples for genomic studies by removing human leukocytes.

2.2.6.2 Preparation of CF11 cellulose columns

This procedure was performed in the hood to avoid inhalation of CF11 powder. Using a razor blade, 1cm^2 square of Whatman lens cleaning tissue (approximately the size of the bottom of the syringe barrel) was cut. The plunger was removed from a syringe and using forceps, two 1cm^2 squares of lens paper were laid flat at the bottom of the syringe

barrel to cover the tip opening. 10 ml of loosely packed CF11 powder was added to the syringe barrel. The plunger was put back and CF11 powder packed down to approximately the 5 ml mark. Pre-made columns were stored until use.

2.2.6.3 Procedure for leukocyte depletion

The plunger was removed from the CF11 column and if the cellulose powder was displaced by air, it was pushed back to restore the cellulose powder within the 5 ml mark. The columns were suspended over an uncapped 15 ml tube using a clamp. 6 mls of RPMI was added to the top of the column and allowed to flow-through. Whole blood (sample) was added to a 15 ml centrifuge tube and an equal volume (same volume as original blood sample) of RPMI added to the blood. The sample was gently pipetted up and down to create a homogenous mix of whole blood and RPMI. Sample was added to the top of the column and allowed to pass through by gravity into a clean 15 ml collection centrifuge tube. Once drops no longer passed into the collection tube, additional 3 mls of RPMI was added to the top of the column and allowed to pass through. The filtered sample was spun at 1000 times *g* at room temperature for 10 minutes, to create a clear separation of pellet from supernatant. The supernatant was discarded without disrupting the red blood cell pellet by leaving a small amount of supernatant. Blood pellet was stored in a cryotube at –70 degrees Celsius until DNA extraction.

2.2.7 DNA extraction from Whole Blood Using the QIAamp Blood Midi Kit (adapted from QIAGEN handbook with minor changes)

2.2.7.1 Steps performed before DNA extraction

- All centrifuge steps were carried out at room temperature (RT) 15 – 25 degrees Celsius.
- DNA samples were equilibrated at room temperature to thaw.
- Water bath at 70 degrees Celsius was prepared.
- Buffer AW1, AW2 and QIAGEN protease were prepared according to manufacturer's instructions.

2.2.7.2 Procedure for DNA extraction

One hundred and twenty µL of QIAGEN protease was pipetted into the bottom of a 15 ml centrifuge tube. Samples with less than 1 ml blood pellet, were topped up with 1X PBS to bring the volume to 1 ml. Sample (PBS and blood pellet) was added to the protease and mixed properly by pipetting up and down. A total of 1.2 mls of Buffer AL was added to the sample and mixed thoroughly to produce a homogenous solution by inverting the tube 15 times, followed by additional vigorous shaking for at least 1 minute. Multiple tubes were inverted simultaneously by clamping them into a rack using another empty rack, grasping both racks. The sample was then incubated at 70 degrees Celsius for 10 minutes, to allow lysis of cells for maximum DNA yield. After incubation, 1 ml of 99% ethanol was added to the sample, and mixed by inverting the tube 10 times, followed by additional vigorous shaking to produce a homogenous solution. The solution was transferred onto a QIAamp Midi column placed in a 15 ml centrifuge tube, taking care not to moisten the rim. The cap

was closed and centrifuged at 1850 times *g* for 3 minutes. If the solution had not completely passed through the membrane, it was centrifuged again at a slightly higher speed (2000 times *g*) for 3 minutes. To avoid any spillage of liquid through the ventilation slots on the rims of the columns we held the QIAamp Midi columns in an upright position. The QIAamp Midi column was removed, the filtrate discarded and placed back into the 15 ml centrifuge tube. Without moistening the rim, 2 mls of Buffer AW1 was added, the cap closed and centrifuged at 4000 times *g* for 2 minutes. Without moistening the rim, 2 mls of Buffer AW2 was added to the QIAamp Midi column, the cap closed and centrifuged at 4000 times *g* for 30 min. This increased centrifugation enabled removal of all traces of Buffer AW2 before elution. The sample was incubated for 10 minutes at 70°C to evaporate residual ethanol. Ethanol in the eluate may cause inhibition in the PCR leading to false positives results. The QIAamp Midi column was placed in a clean 15 ml centrifuge tube and the collection tube containing the filtrate discarded. 300 µl of Buffer AE equilibrated to room temperature, was added directly onto the membrane of the QIAamp Midi column, cap closed and incubated at room temperature for 5 minutes, and centrifuged at 4000 times *g* for 4 minutes. For maximum yield, 300 µl fresh Buffer AE equilibrated to room temperature, was pipetted onto the membrane of the QIAamp Midi column, cap closed and centrifuged at 4000 times *g* for 4 min. The resulting filtrate (DNA solution) was pipetted into a 1.5 ml Lo-Bind eppendorf tube and kept at 4 degrees Celsius ready for shipping to WTSI for sequencing.

Summary of experiments performed by collaborators at WTSI (sequencing):

2.2.8 Sample preparation and sequencing

- Ninety-three purified DNA samples were sent to WTSI malaria laboratories for MPS sequencing. Total DNA in each sample was quantified on the Invitrogen Qubit as per

the manufacturer's standard protocol (Quant-iT^{TM}). Quantitative real-time PCR analysis was undertaken on all DNA samples to determine the quantities of human to Plasmodium DNA, with primers specific to human (PLAT1), *P. falciparum* (AMA1) and *P. vivax* (VIV3) using Lightcycler 480 qPCR.

2.2.9 Quantification of human to *P. falciparum* DNA concentrations

2.2.9.1 Reagents/consumables, primers and probes

- DNA free water
- Q/RT-PCR probe mix (Roche or Kappa)
- Sample DNA (0.1-10ng)
- Control DNA mix (0.1-10ng)
- 96 well qPCR plate (Starlab I1402-9909)
- Optically clear plate seal (Starlab E2796-9795)
- Lightcycler 480 qPCR

Forward and reverse primer sequences for VIV3, AMA1 and PLAT1:

Name	Type	Sequence
VIV3	Primer- forward	AAA GAT TCG TAG CTG TCG GTG GGT
	Primer-reverse	TTC CAT TAA GTG CGC GTA CCG AGA
	Probe	ACA GCG ACG ACT CCA GAT CCG ATT TA
AMA1	Primer-forward	TGC CAT ATA TTC CGT CCA TGG
	Primer-reverse	ACG CAT ATC CAA TAG ACC ACG
	Probe	CGA ACC CGC ACC ACA AGA ACA AAA
PLAT1	Primer-forward	CTT ACC ACA TCC GCT CCA TC
	Primer-reverse	TTC ACA CTC TCC GTC ACA TTG
	Probe	CAC ATC CCC AGT GCC GAG TTA GA

2.2.9.2 Procedure

Master mix:

Component	Volume (µl)
Probe Mix	10
AMA1 Primer/probe mix	1
PLAT1 Primer/probe mix	1
VIV3 Primer/probe mix	1
Water	6

The probes used dyes from integrated DNA technologies (IDT), FAM for VIV3, Lightcycler 640 for AMA1 and HEX for PLAT1. In addition to this, the performance of certain dye types were improved with the addition of an internal quencher, ZEN with both HEX and FAM dyes. Master mix was prepared as above. 19µl of master mix was pipetted into (3x sample) + 6 wells (3 wells each for control mix and water), i.e., for 10 samples, 36 wells were used. 1µl of each sample was pipetted into 3 wells and 1µl of control mix into a separate 3 wells. After adding 1µl DNA/RNA-free water into the rest of the remaining 3 wells, the PCR plate was sealed and loaded into the qPCR machine that was run using the following cycling conditions:

Cycling conditions:

Stage	Temperature (^{o}C)	Time (Min:Sec)
Activation	95	05:00
Amplification (45 cycles)	57	00:20
	72	00:01
	95	00:10
Storage	40	Infinite

Standard curves are generated using the Lightcycler 480 software from a 5 times dilution series (5 replicates of each). A known concentration standard is then included in each run and mapped onto the previously generated standard curves, the unknown samples

concentrations are then subsequently calculated based on this standard. Human percentage is calculated using excel, triplicates are used for each unknown and averages of these for human, *P. falciparum* and (if appropriate) *P. vivax* were analyzed to determine percentage composition.

2.2.10 Library preparation and sequencing

Standard sequencing libraries for all DNA samples were prepared following the manufacturer's recommended protocol by the WTSI core library preparation and sequencing team (Bentley et al. 2008). All samples with less than 60% human DNA contamination and sufficient DNA (> 1µg/1000ng) underwent whole genome sequencing on the Illumina HiSeq2000 machines without any PCR amplification (Kozarewa et al. 2009) with 76 to 100-bp paired-end sequencing reads. The entire short read sequence data have been deposited in European Nucleotide Archive.

2.2.11 Data processing – Alignment, SNP discovery and quality filtering

Short reads for all 93 samples were mapped to 3D7 reference genome (version 3.0) using *Smalt* (http://www.sanger.ac.uk/resources/software/smalt). Twenty-four samples had low coverage (average genome-wide coverage less than 10-fold), and were excluded. For the remaining samples (n=69, average coverage genome-wide >35-fold), SNPs were called using *samtools* and bcf/*vcftools* with default settings (http://samtools.sourceforge.net). This process identified 115,965 SNPs across the 69 samples. We retained 88,655 high quality SNPs in the nuclear genome that met the following criteria: (i) biallelic, (ii) quality scores in excess of 30 (error rate less than 1 per 1000-bp), (iii) average coverage across all samples between 10- and 2000-fold (Li et al. 2009), (iv) not located in sub-telomeric regions, the hypervariable *var*, *rifin* and *stevor* gene families, or regions of low uniqueness, and (v) had

no more than two missing or mixed called genotypes across the samples. Uniqueness was calculated by a sliding 50-bp window of contiguous sequence across the 3D7 reference genome and detecting the presence of this motif elsewhere in the genome. SNPs were only retained if they were in a unique position. Genotypes at SNP positions were called using ratios of coverage and heterozygous calls converted to the majority genotype on 80:20 coverage ratio or greater (Robinson et al. 2011; Manske et al. 2012). Population data from other populations (Kenya, n=37; Burkina Faso, n=40; Mali, n=40; Cambodia, n=80, Thailand, n=80; 294,187 SNPs; Preston et al. 2013, SRA study ERP000190) were processed in the same way. The four mitochondrial SNPs (*mt772, mt1692, mt4179* and *mt4952*) were extracted from the alignments and genotypes called as described above. Public accession numbers for sequence data are contained in SRA studies at www.ebi.ac.uk/ena/data/view/ERP000190 and www.ebi.ac.uk/ena/data/view/ERP000199, as well as accessible from www.malariagen.net. We obtained permission to use additional metadata information from relevant investigators who are included on this thesis.

2.2.12 Population genetics analysis and selection metrics

For the analysis of SNPs (n=88,655) in the Malawian populations, we computed the allele frequency-based Tajima's *D* approach (Tajima 1989) at the gene level to detect balancing selection. The integrated Haplotype Score *iHS* was applied to detect selective sweeps (Voight et al. 2006). For this, a recombination map was generated using LDhat 2.1 as previously described (Chang et al. 2012), and *P*-values computed from standardised values based on a two-tailed conversion from a Gaussian distribution (Chang et al. 2012). A threshold of *iHS* ($P<0.0006$) was determined to account for multiple testing, and calculated using the actual nominal 5% cut-off from re-running the algorithm on 10,000 bootstrap

samples. To assess the effects of low frequency alleles potentially due to population expansion, the *iHS* was also applied to SNP data with a minor allele frequency of 5%. There was no notable difference in results, especially the hits identified. To assess if there was any population structure within the Malawian population, principal component analysis (PCA) was conducted using SMARTPCA (EIGENSOFT 3.0, Patterson, Price, and Reich 2006). We applied a local LD correction (nsnpldregress = 2) and calculated the top 25 eigenvectors, with corresponding *P*-values. For comparisons between populations (Malawi vs. other, n=294,187 SNPs), we applied the across population long range LD method *XP-EHH* (Sabeti et al. 2007) and the population differentiation metric F_{ST} (Holsinger and Weir 2009). *P*-values for the *XP-EHH* estimates were calculated using a Gaussian approximation, and a Bonferroni threshold of $P<0.00006$ was established for reporting. A threshold for F_{ST} (>0.2) was established as a benchmark for reporting by looking at the 99.9-ile across all comparisons. A PCA based on a matrix of pairwise identity by state values was used to cluster all samples. We used the ranked F_{ST} statistics to identify the informative polymorphism driving the clustering.

2.3 Results A

2.3.1 Summary characteristics of sampled Malawi isolates

Between December 2010 and June 2013, duration covering three malaria seasons, we collected 400 malaria positive samples. Out of these, 93 samples collected between December 2010 and June 2012 were sequenced at the WTSI and data included in this thesis. The remaining samples were either still in the sequencing pipeline or their data released late and were not included in this thesis. Samples patients were young (average age 11

months), with high parasitemia (1,000 - 10,000 parasites/µL) and average haemoglobin (9 g/dl). Most samples (> 97%) were from uncomplicated malaria cases.

2.3.2 Quantification of human DNA content using quantitative real time PCR

To date, we have quantified human DNA content in 309 samples (including the 93 samples used in this thesis) out of the 400 available for sequencing. Total DNA concentration ranged from 0 – 238 ng/µL (in 100-500 µL sample volumes) and ratios of human DNA concentrations ranged from 0 – 100%. Samples with less than 60% human DNA contaminations and with at least 1000 ng total DNA were sequenced (Table 2.1).

Table 2.1: Summary of proportion of DNA concentrations in the samples.

	Total DNA concentration (ng/µl)	% Human DNA concentration
Minimum	0	0
Median	3.97	0.78
Mean	16.92	9.8
Maximum	238	100

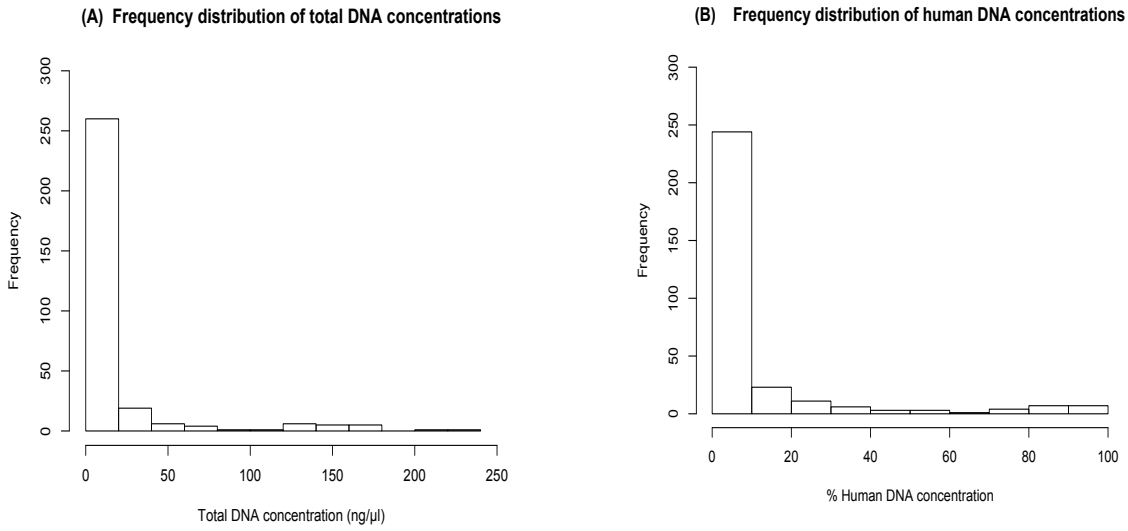


Figure 2.3: Frequency distribution of (A) total DNA concentrations and (B) total human DNA concentrations in the total sample as estimated by quantitative real-time PCR. 90% of samples had <30% human DNA contamination and were sequenced.

2.3.3 Summary of sequence results and SNP quality filtering steps

A total of 115,965 SNPs across 69 samples were identified and quality control was performed as described in the methods section. The median number of individual SNP differences from the reference genome and genomic coverage across samples before and after QC are provided (Table 2.2, 2.3 and Figure 2.4). In addition, distribution of mixed calls and "missingness" in isolates and SNPs are given (Figure 2.5-2.7). Both the non-reference and minor allele frequency spectrum are dominated by rare derived alleles. Only 1,446 (1.1%) SNPs/alleles were > 95% different from the reference (Figure 2.8 and 2.9).

Table 2.2: Summary of sequence results across 93 samples for the nuclear genome. Poor sequencing of some samples resulted in very low coverage and poor SNP calling.

	Coverage	Read depth	Mapped Reads	% Mapped	SNPs	SNPs (Q30)
Mean	34	612.4	13,535,495	97.5	23,411	19,429
Median (range)	34 (0-158)	514 (78-2401)	11,848,299 (99,473-52,221,597)	98.2 (78.5-99.2)	26,834 (23-41,245)	21,633 (9-34,747)

Table 2.3: Summary of sequence results across 69 samples for the nuclear genome.

	Coverage	Read depth	Mapped reads	% Mapped	SNPs	SNPs (Q30)
Mean	44.76	671.6	16,544,779	98.3	28,349	23,509
Median (range)	38 (11-158)	608 (120-2,401)	13,926,835 (4,447,550-52,221,597)	99.2 (97.7-99.2)	28,641 (13,206-41,245)	23,422 (9,973-34,727)

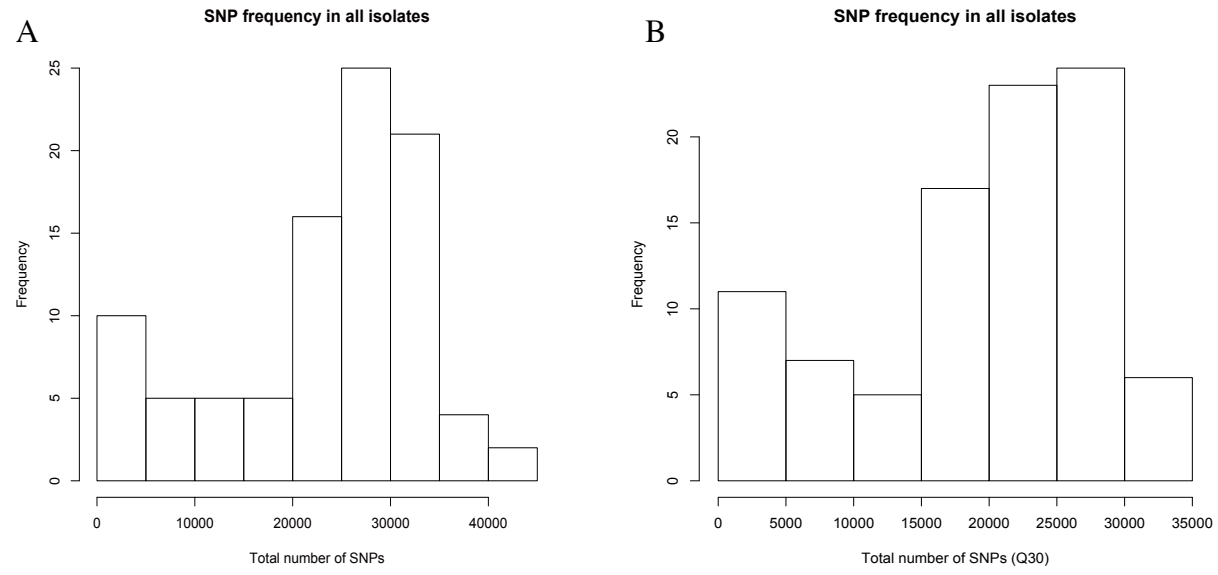


Figure 2.4: Proportion of candidate SNPs in all isolates. (A) Candidate SNPs identified before applying mapping score quality. (B) Candidates SNPs identified with good mapping quality of Q30 (1 error per 1000-bp).

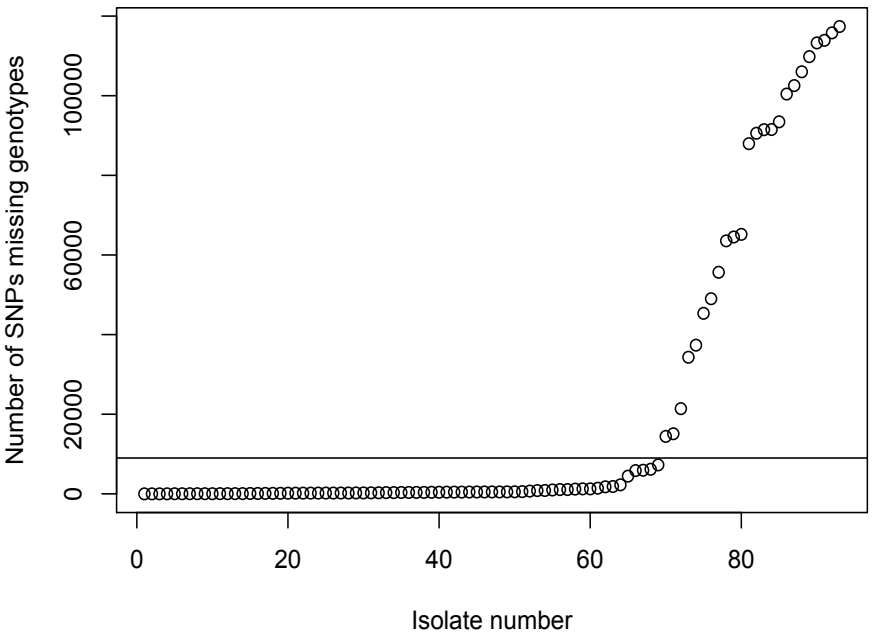


Figure 2.5: Isolates missing genotypes. 69 samples with less than 10,000 missing calls were retained. Samples (n=24) with high missing calls were discarded.

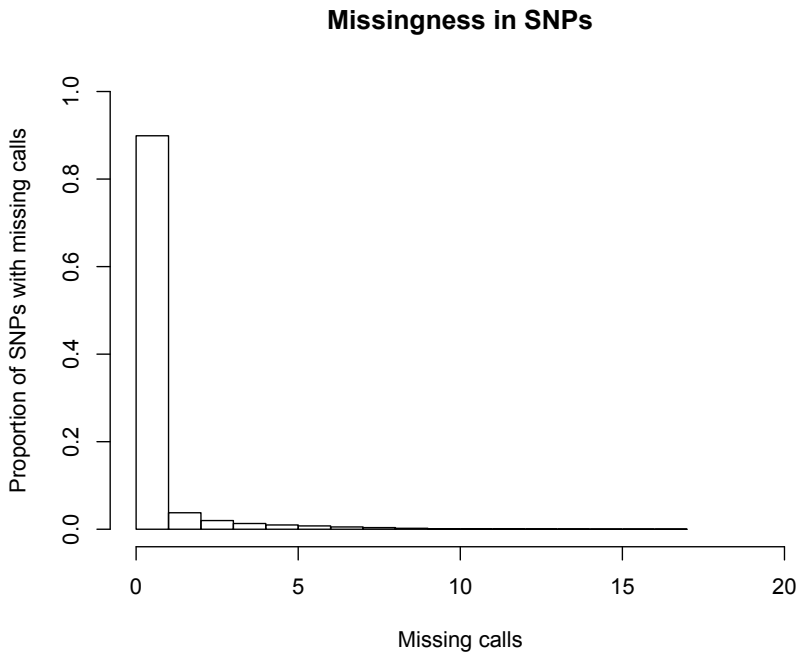


Figure 2.6: Missing SNP calls. 106,515 (92%) with no more than two missing calls were retained.

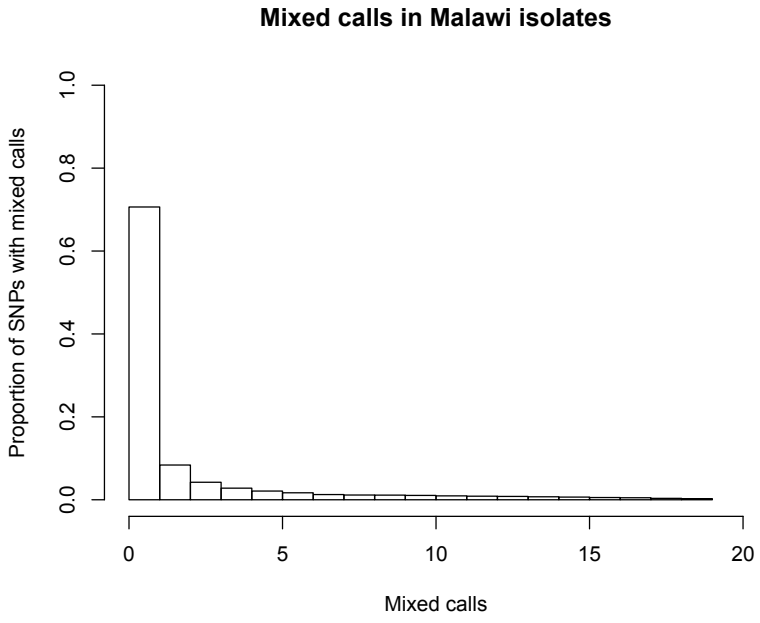


Figure 2.7: Proportion of mixed calls in Malawi isolates. 88,655 (83.2%) biallelic SNPs with no more than two mixed calls were retained.

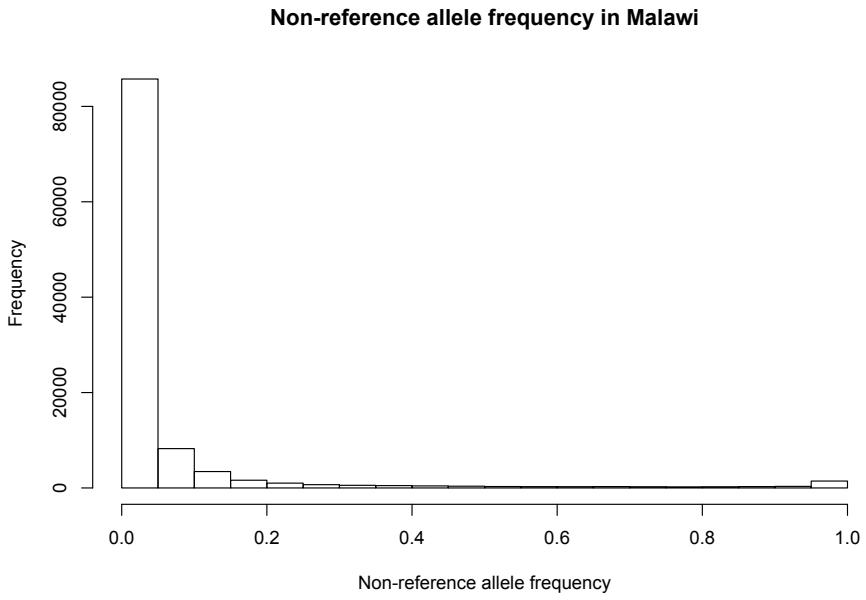


Figure 2.8: Non-reference allele frequency in Malawi isolates. Rare alleles dominate in the population. 1,446 (1.1%) SNPs/alleles were > 95% different from the reference.

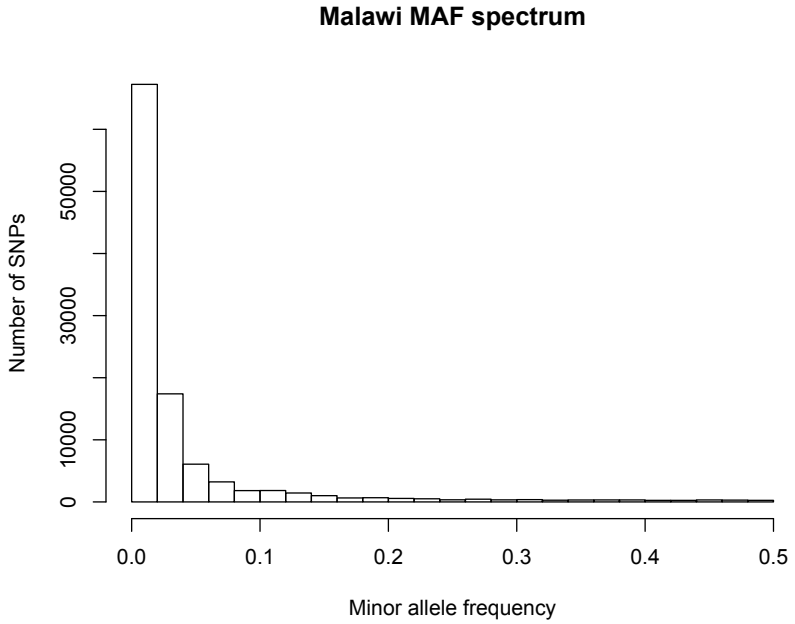


Figure 2.9: Minor allele frequency in Malawi isolates. The allele frequency spectrum is dominated by rare derived SNPs, the majority (82%) of which have a minor allele frequency (MAF) <5%.

2.3.4 Population structure in Malawi P. *falciparum* population

An examination of the structure or stratification within and between populations could reveal insights into their evolutionary history. Previous studies (Manske et al. 2012; Miotto et al. 2013; Preston et al. 2013) show that population structure of *P. falciparum* is geographically dispersed, as a result of adaptation to different environments and differing selection pressures. Before analyzing the sequencing data from the nuclear genome for population structure, mitochondrial (*mt*) genome (~6-kb) SNPs were used to confirm that the Malawian samples were of African origin. Haplotypes were formed using four established continental specific polymorphisms (*mt772, mt1692, mt4179* and *mt4952*) (Conway et al. 2000). Two haplotypes were present CGCC (identical to 3D7, 90.8%) and CACC (9.2%) – both of African origin. High read depth coverage in the *mt* (median/mean ~1560/1245-fold, ~19/23-fold greater than the nuclear genome) is also consistent with the known multiple copies of the organelle in a *P. falciparum* cell (Vaidya and Mather 2009). We investigated population structure within the Malawian population, and detected no strong evidence of structure, SMARTPCA principal components $P>0.1$ (Patterson et al. 2006) (Figure 2.10). This may be expected because all samples were recruited within the same season and geographical region.

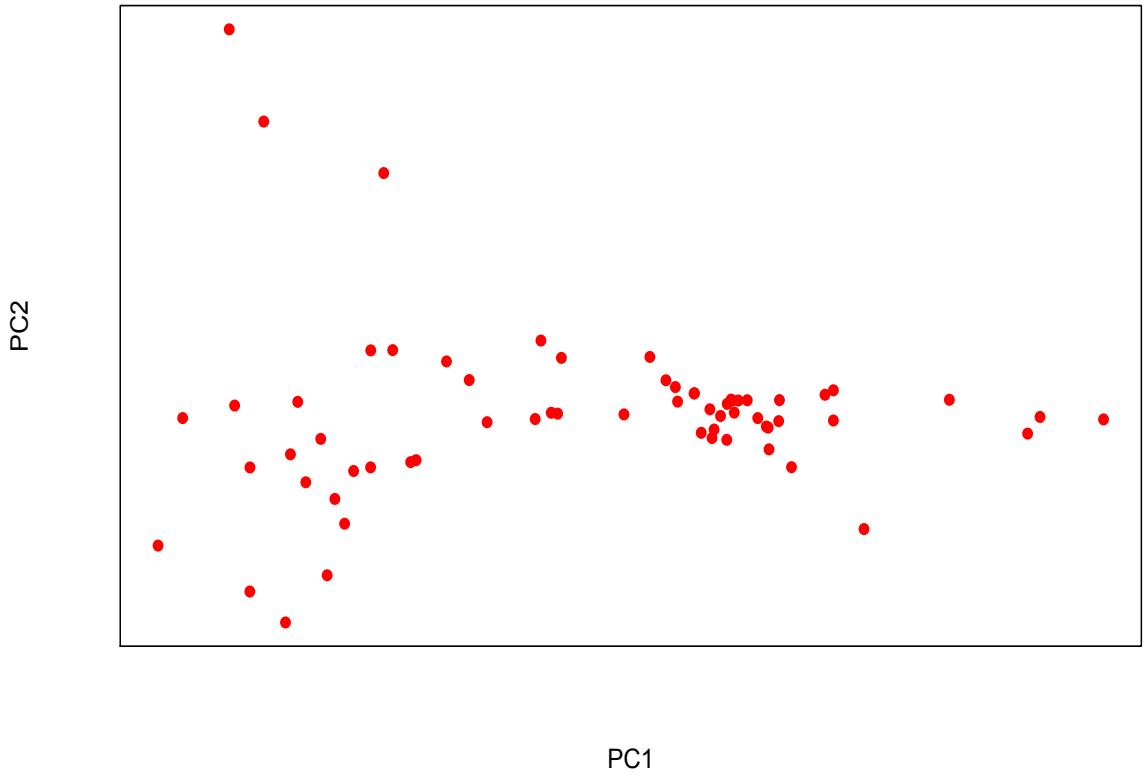


Figure 2.10: Population structure of Malawi parasites assessed by PCA on SNPs. SNPs show no structure.

2.4 Results B

2.4.1 Inferring balancing selection in a Malawi *P. falciparum* population

Whole-genome application of the Tajima's *D* metric reveals a high proportion (83%) of alleles with "negative" Tajima's *D* values, indicative of an excess of low-frequency and singleton polymorphisms, which may have arisen from historical population expansion or purifying selection (Weedall and Conway 2010; Amambua-Ngwa et al. 2012b). This analysis also identified 19 genes (Table 2.5) with positive Tajima's *D* values (i.e., ≥1.0) as loci potentially under balancing selection. These loci encode proteins with roles in pathogenesis

and potential as targets for immunity and drug intervention such as *msp3.8*, *msp3*, *dbl-msp, eba175, ama1* and *surfin4.2* (Table 2.5, Figure 2.11). Many of these genes have had positive indices of balancing selection in previous studies of African populations (Ochola et al. 2010; Amambua-Ngwa et al. 2012b). Immuno-epidemiological studies of MSP and AMA1 proteins have particularly identified allele-specific antibody responses associated with protection from malaria (Polley et al. 2007; Osier et al. 2008). The *surfin4.2* gene encodes SURFIN protein, expressed on infected erythrocytes or merozoites and thus exposed to host immunity (Winter et al. 2005).

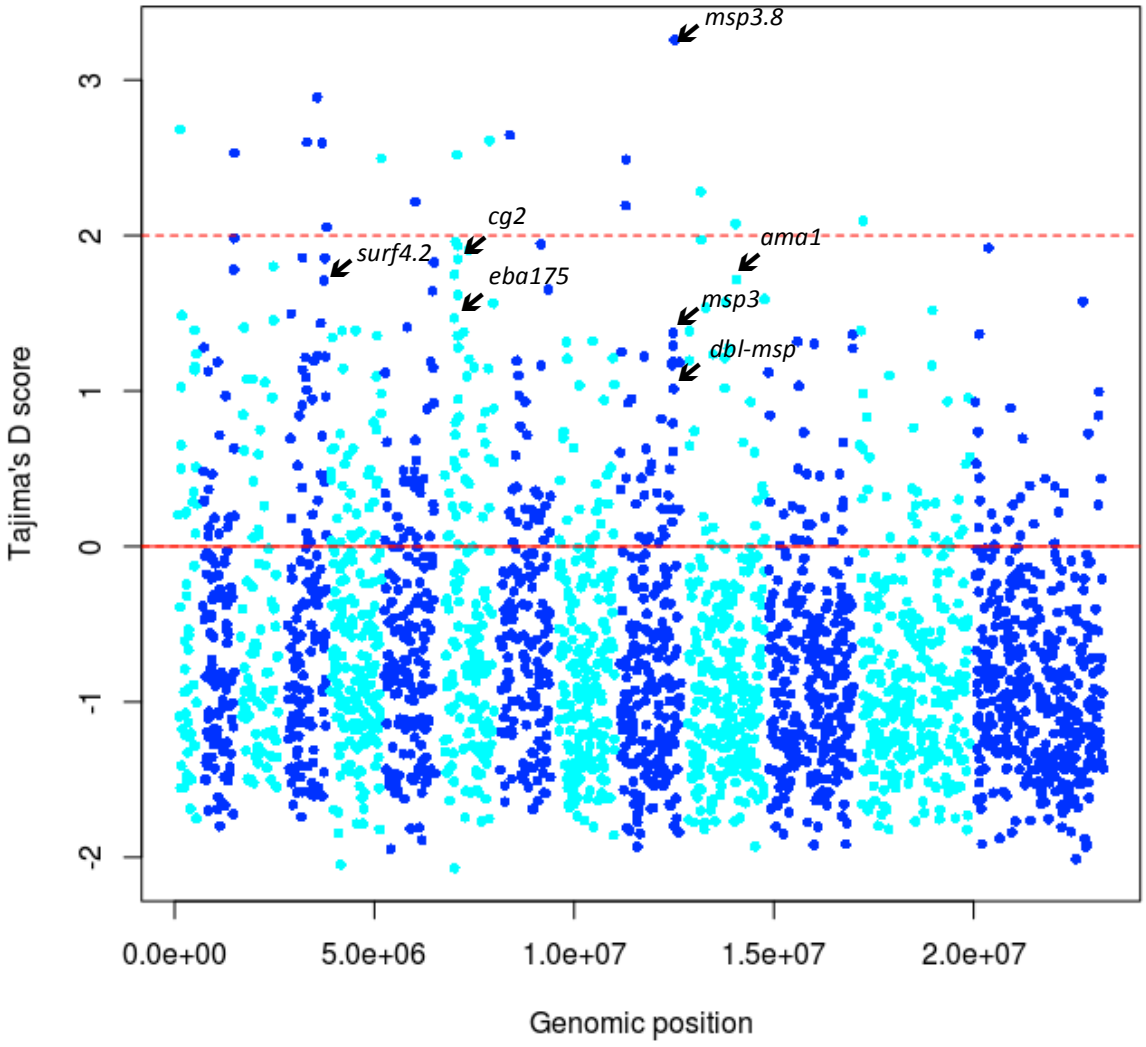


Figure 2.11: Genome-wide distributions of Tajima's *D* across *P. falciparum* genome, showing genes subject to balancing selection in Malawi parasite isolates. Coloured dots represent individual genes. The solid horizontal line indicates genomic threshold for positive Tajimas' *D* ≥0. Dashed line indicates crude significance, *D* ≥+2 values.

Table 2.4: Genetic loci under balancing selection (Tajima's *D* ≥1.0).

Gene	Tajima's *D*	Start	Stop	Gene description
PF3D7_1036300	3.26	1432702	1434553	Merozoite surface protein 3.8
PF3D7_0710000	2.52	447902	457801	Conserved Plasmodium protein, UF
PF3D7_0425400	2.06	1144011	1144822	Plasmodium exported protein (PHISTa), UF
PF3D7_0221000	1.98	848124	849107	Plasmodium exported protein, UF
PF3D7_0709300	1.96	414302	421420	Cg2 protein
PF3D7_0710200	1.94	463705	471598	Conserved Plasmodium protein, UF
PF3D7_0424400	1.85	1100085	1102381	Surface-associated interspersed protein 4.2 (SURFIN 4.2)
PF3D7_0630800	1.83	1288574	1290718	Conserved Plasmodium protein, UF
PF3D7_1133400	1.72	1293957	1295622	Apical membrane antigen 1
PF3D7_1149600	1.59	2001079	2003312	DnaJ protein, P
PF3D7_1126100	1.57	1018557	1021025	Autophagy-related protein 7, P
PF3D7_0731500	1.56	1358502	1362925	Erythrocyte binding antigen-175
PF3D7_0104100	1.49	178094	180554	Conserved Plasmodium membrane protein, UF
PF3D7_0516300	1.39	679096	680745	tRNA pseudouridine synthase, P
PF3D7_1035400	1.37	1404453	1405160	Merozoite surface protein 3
PF3D7_0113800	1.24	527113	536327	DBL containing protein, UF
PF3D7_0630300	1.15	1260750	1269383	DNA polymerase epsilon, catalytic subunit a, P
PF3D7_0103600	1.03	161480	165521	ATP-dependent RNA helicase, P
PF3D7_1035700	1.01	1413250	1415182	Duffy binding-like merozoite surface protein

UF = Unknown function and P = putative

2.4.2 Inferring positive selection in a Malawi *P. falciparum* population

To examine evidence for signatures of positive directional selection integrated haplotype score (*iHS*) was computed. Seven contiguous loci (Table 2.6, Figure 2.12, *P*<0.0006) of strong selection were identified including areas surrounding SP drug resistance loci *pfdhps* and *pfgch1*. Additional regions include areas on chromosomes 1, 2, 4, 10 (*msp6* and *msp3.8*), 11 (*pfama1*), 13 (*trap*) and 14 (*msp7*). Evidence of positive directional selection in antigenic loci, thought to be primarily under balancing selection, is consistent with previous reports (Mu et al. 2010; Amambua-Ngwa et al. 2012a, Borrmann et al. 2013), where long haplotype signals in *ama1* and *trap* have also been detected in Asia, east and west African parasites. It is possible there are non-immune pressures acting on these loci or that positive directional selection may be a result of a variant unfamiliar to the host immune system rising from very low frequencies in the population.

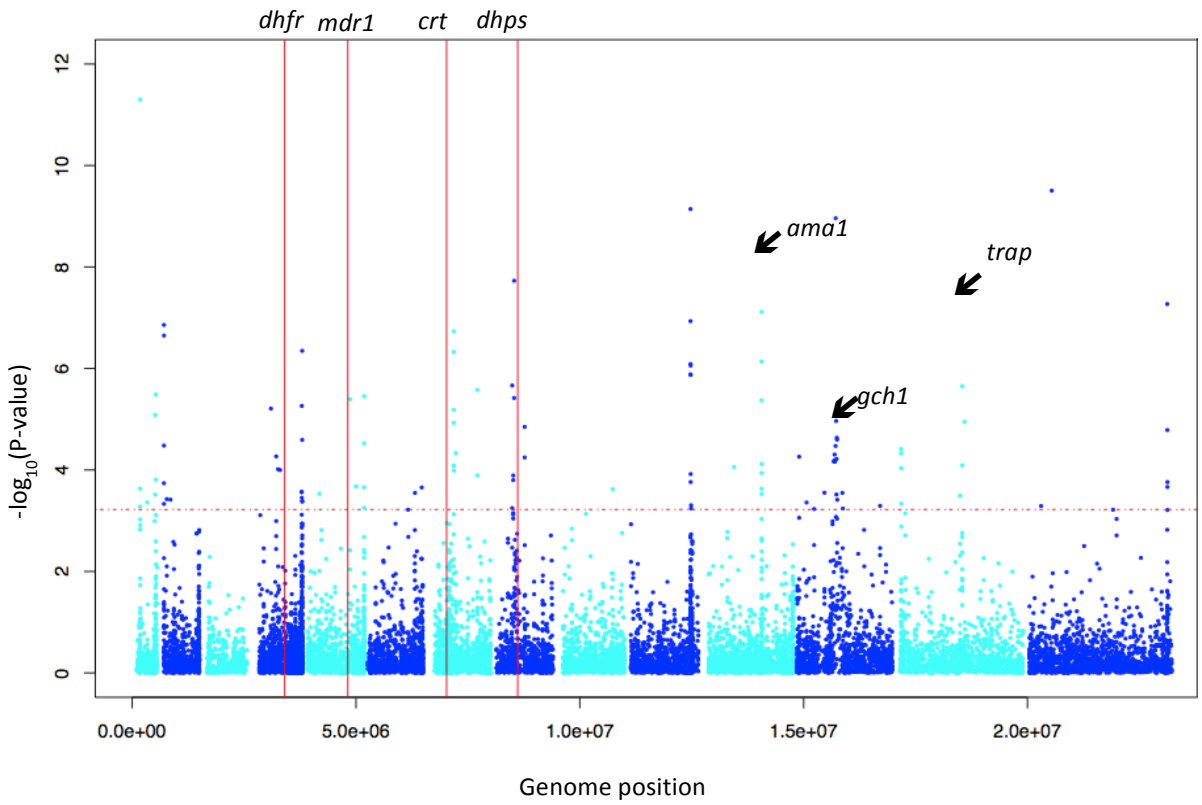


Figure 2.12: Recent positive directional selection in a Malawi *P. falciparum* population. Dashed line indicates genome-wide significant *iHS* values (P<0.0006). Vertical lines indicate (from left) locations of *pfdhfr*, *pfmdr1*, *pfcrt* and *pfdhps* respectively.

Table 2.5: Regions under recent positive directional selection in Malawi, *iHS* (*P*<0.0006).

Chromosome	Start	Stop	Size	Comment
1	178726	180317	1591	
	512851	558256	45405	Contains DBL containing protein, UF
2	842699	855734	13035	
4	1065176	1144415	79239	Contains deveral Plasmodium exported proteins
5	966314	1181373	215059	Approximately 4-kb from pfmdr1
	1320756	1325021	4265	
7	409122	470642	61520	Approximately 3-kb from pfcrt
	507357	665385	158028	
	1358889	1380385	21496	
8	449188	585854	136666	Surrounding pfdhps
10	1389354	1434268	44914	Contains msp6 and msp3.8
11	1294082	1295369	1287	Contains ama1
12	800894	1059078	258184	Surrounds pfgch1; several transcription factors
13	102848	106661	3813	Contains trap
	1419420	1466484	47064	
14	2982003	3149504	167501	Contains msp7

UF = unknown function

2.5 Results C

2.5.1 Placing Malawi parasites in the global population structure of *P. falciparum*

Using the same raw sequence-processing pipeline applied to the Malawi data, SNPs were called across all six populations. A total of 294,187 SNPs were identified across the six populations: Chikwawa, Malawi (n=71, 46% polymorphic); Kilifi, Kenya (n=37, 30%); Bobo-Dioulasso, Burkina Faso (n=40, 33%); Bamako, Mali (n=40, 37%); Pailin, west Cambodia (n=80, 22%) and Mae Sot, Thailand (n=80, 23%), with 8% being specific to Malawi only. A principal component analysis (PCA) of the SNPs revealed modest population differences between Africa and Asia, within Southeast Asia, and small differences within Africa (Figure 2.13). These PCA results are also consistent with those previously published (Manske et al. 2012; Preston et al. 2013). We further applied the SNP-wise F_{ST} metric to measure genomic divergence across the six populations. At a stringent genome-wide cut-off value (F_{ST} ≥0.2), pairwise comparisons identified the most divergent alleles between Malawi and the five

populations (Table 2.7). Overall, there is high level of divergence in *clag3.1* and *clag3.2* (genes implicated in the process of parasite invasion). Alleles encoding known drug targets (e.g., *pfcrt*-K76T, *pfdhps*-K540E, *pfdhps*-A436G and *pfmdr1-N86Y*) showed large divergence between populations (Table 2.8). These inter-population differences may reflect local historical parasite adaptation to anti-malarial drug pressure, leading to fixation or near fixation of the implicated drug resistant alleles.

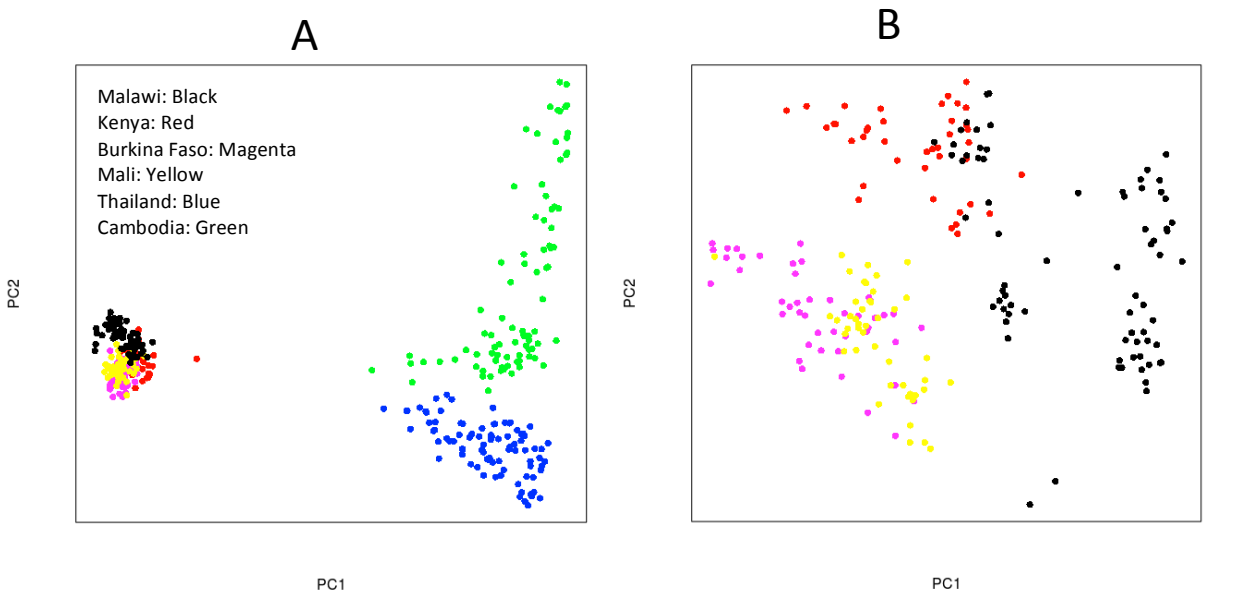


Figure 2.13: Principal components analysis using SNPs differentiates *P. falciparum* isolates by continent and within SEA (A), and between East and West Africa (B).

Table 2.6: Genes with multiple alleles with F_{ST} ≥0.2 (*P*<0.001), stratified by parasite population. Integers refer to the number of alleles *per* gene with F_{ST} ≥0.2. Blank entries refer to lack of differentiating alleles in a given gene between populations compared.

Gene ID	Gene product	KENYA	BFASO	MALI	CAMBODIA	THAI
PF3D7_0113000	Glutamic acid-rich protein		4			
PF3D7_0202000	Knob-associated histidine-rich protein	2				
PF3D7_0209000	6-cysteine protein		3	4		
PF3D7_0302200	Cytoadherence linked asexual protein 3.2					39
PF3D7_0302500	Cytoadherence linked asexual protein 3.1				36	10
PF3D7_0320400	Oocyst capsule protein				11	
PF3D7_0419900	Phosphatidylinositol 4-kinase, P				17	
PF3D7_0501200	Parasite-infected erythrocyte surface protein		3			
PF3D7_0523000	Multidrug resistance protein	2				
PF3D7_0525100	Acyl-CoA synthetase		4	4		
PF3D7_0527200	Ubiquitin carboxyl-terminal hydrolase, P	2				
PF3D7_0628100	HECT-domain (ubiquitin-transferase), P			3	12	12
PF3D7_0629500	Amino acid transporter, P	2				
PF3D7_0629700	SET domain protein, P	2				
PF3D7_0704600	Ubiquitin transferase, P				10	10
PF3D7_0709000	Chloroquine resistance transporter	2		3	1	7
PF3D7_0709100	Cg1 protein			3	12	11
PF3D7_0709300	Cg2 protein				19	16
PF3D7_0710000	Conserved Plasmodium protein, UF				19	14
PF3D7_0810800	Dihydropteroate synthetase		3	3	3	2
PF3D7_0824400	Nucleoside transporter 2		3	3		
PF3D7_0826100	E3 ubiquitin-protein ligase, P				23	27
PF3D7_0911300	Cysteine repeat modular protein 1		3	3		
PF3D7_0927200	Zinc finger protein, P				10	
PF3D7_0929400	High molecular weight rhoptry protein 2		5	5		
PF3D7_1004800	ADP/ATP carrier protein, P					16
PF3D7_1110200	Pre-mRNA-processing factor 6, P					10
PF3D7_1125700	Kelch protein, P		4	4		
PF3D7_1149000	Antigen 332, DBL-like protein					12
PF3D7_1221000	Histone-lysine N-methyltransferase				10	12
PF3D7_1222600	Transcription factor with AP2 domain(s)	2		5		
PF3D7_1223100	cAMP-dependent protein kinase regulatory subunit	2				
PF3D7_1223300	DNA gyrase subunit A	2				
PF3D7_1223400	Phospholipid-transporting ATPase, P	2	4	4		
PF3D7_1223500	Conserved Plasmodium protein, UF		3	3		
PF3D7_1231800	Asparagine-rich protein, P				10	
PF3D7_1431200	Conserved Plasmodium protein, UF				15	14
PF3D7_1452000	Rhoptry neck protein 2		6	5		

KENYA (F_{ST} between Malawi and Kenya), B FASO (F_{ST} between Malawi and Burkina Faso (B FASO)), MALI (F_{ST} between Malawi and Mali), CAMBODIA (F_{ST} between Malawi and Cambodia) and THAI (F_{ST} between Malawi and Thailand (Thai)). Integers refer to number of alleles/gene with F_{ST} ≥0.2 (*P*<0.001). UF = Unknown function, P = Putative.

Table 2.7: F_{ST} of known antimalarial drug-resistance loci. Blanks infer very low F_{ST}.

Locus	Mutation	KENYA	B FASO	MALI	CAMBODIA	THAI
CRT	K76T	0.22	0.25	0.49	1	1
	Q271E			0.46	0.98	1
	N326S				0.82	1
	I356T				0.83	1
DHPS	S436A		0.35	0.47		
	A437G		0.28	0.49		
	K540E		0.91	0.91	0.34	
	A581G				0.37	0.61
MDR1	N86Y	0.37				
	N1226Y					0.37
	D1246Y	0.32				

KENYA (F_{ST} between Malawi and Kenya), B FASO (F_{ST} between Malawi and Burkina Faso (B FASO)), MALI (F_{ST} between Malawi and Mali), CAMBODIA (F_{ST} between Malawi and Cambodia) and THAI (F_{ST} between Malawi and Thailand (Thai)).

Table 2.8: Allele frequencies of common drug resistant SNPs across all six populations.

Locus	Mutation	Malawi	Kenya	Mali	Burkina Faso	Thailand	Cambodia
CRT	K76T	0	0.306	0.615	0.364	1	1
	Q271E	0	0.226	0.59	0.231	1	0.993
	N326S	0	0.048	0.013	0.048	1	0.913
	I356T	0	0	0.183	0.087	1	0.917
DHPS	S436A	0.007	0.025	0.611	0.489	0.178	0.329
	A437G	1	0.825	0.386	0.609	1	0.977
	K540E	0.965	0.808	0	0	0.868	0.406
	A581G	0	0	0	0	0.782	0.565
MDR	N86Y	0.028	0.55	0.256	0.272	0.012	0.0176
	N1226Y	0	0.014	0	0	0.575	0.006
	D1246Y	0	0.425	0.022	0.065	0	0

2.5.2 Inferring positive selection in Malawi *P. falciparum* using *XPEHH*

The across population long-range haplotype method, *XP-EHH,* was applied to compare Malawi to each of the other populations, to identify regions potentially under recent directional selection. Several genes detected with extended haplotypes in Malawi include *pfdhps*, *ron2* and *trap*, and near *pfdhfr*, *pfmdr1* and *pfgch1*, mostly consistent with *iHS* results (Table 2.10 and $P<0.00006$). In the other populations, signals were detected at *pfdhps* and *pfcrt*. Selective sweeps at *pfdhps* and *pfcrt* respectively are probable effects of high SP drug pressure (in Malawi) and chloroquine (in Mali, Burkina Faso, Cambodia and

Thailand). Interestingly, *pfcrt* and *pfdhps* extended haplotypes were observed between Malawi-Kenya and Malawi-Thailand respectively; this is probably indicative of relatively high prevalence of *pfcrt* mutations in Kenya as previously observed (Mwai et al. 2009) and an increase in prevalence of *pfdhps* mutations in Thailand. The lack of evidence for a selective sweep in *pfcrt* in Malawi reflects the withdrawal of CQ and subsequent increase in the ancestral CQ-sensitive allele frequency from the re-expansion of the minority of drug susceptible parasites that may have survived in the population. This observation suggests that parasites carrying the ancestral *pfcrt* alleles have greater fitness in the absence of CQ pressure (Laufer et al. 2006). The observed selective sweep surrounding the GTP cyclohydrolase gene (*pfgch1*, PF3D7_1224000) on chromosome 12 is only unique to Malawi in this study. The *pfgch1* gene is the first gene in the folate biosynthesis pathway, and adaptive selection could result from SP drug pressure (Nair et al. 2008). However, previous studies of Thai parasites have also detected reduced microsatellite diversity and increased LD flanking the *pfgch1* locus (Nair et al. 2008). Across the Malawian population there did not appear to be excess genomic coverage in the *pfgch1* gene (that could explain the positive selection at this locus), but additional investigations should determine whether the locus is amplified because it may indicate rapid recent spread of chromosomes carrying multiple copies of *pfgch1* (Nair et al. 2008). Positive directional selection in the *trap* gene is thought to reflect genetic adaptation to divergent host ligands (Amambua-Ngwa et al. 2012a), whose product have roles in the motility of sporozoites, invasion of hepatocytes and mosquito salivary glands (Ejigiri et al. 2012).

Table 2.9: Regions under directional selection in all six populations identified using *XP-EHH* (*P*<0.00006)

Population	Chromosome	Start	End	Size (bp)	Comment
MALAWI	1	180314	193846	13532	
		487895	489460	1565	
		512383	513033	650	
	4	755433	881703	126270	Approximately 6-kb from *pfdhfr*
		990908	991327	419	
	5	1042259	1109432	67173	Approximately 8-kb from *pfmdr1*
	8	532499	585854	53355	Contains *pfdhps*
	10	1326109	1327397	1288	
	12	461137	473836	12699	
		946416	954490	8074	Approximately 30-kb from *pfgch1*
		983440	1016281	32841	
		1004000	1022661	18661	
	13	1465713	1465965	252	Contains *trap*
	14	1688102	1688881	779	
		2135779	2137007	1228	Contains rhoptry neck protein 2
KENYA	6	1116365	1222963	106598	Acetyl-CoA synthetase, putative
	7	376423	417661	41238	Contains *pfcrt*
	8	467328	468623	1295	Asparagine-rich antigen Pfa55-14
MALI	6	1205649	1290486	84837	
	7	376423	470941	94518	Contains *pfcrt*
		505661	614698	109037	
		1100440	1326844	226404	
	8	468447	469357	910	
	11	1006055	1383842	377787	
B FASO	1	487895	489267	1372	
	7	432780	507284	74504	
		908940	918733	9793	
	8	416971	422505	5534	Plasmepsin X
THAILAND	4	709512	771505	61993	
	6	1114625	1289592	174967	Acetyl-CoA synthetase, putative
	7	339092	451640	112548	Contains *pfcrt*
	8	468447	586054	117607	Contains *pfdhps*
		703454	712742	9288	
		1108501	1117018	8517	
CAMBODIA	6	1109423	1135810	26387	Acetyl-CoA synthetase, putative
		1287468	1289915	2447	
	7	332719	453986	121267	Contains *pfcrt*
		875300	931176	55876	
		989974	993421	3447	
	8	468669	479732	11063	Asparagine-rich antigen Pfa55-14

KENYA (*XP-EHH* between Malawi and Kenya), B FASO (*XP-EHH* between Malawi and Burkina Faso (B FASO)), MALI (*XP-EHH* between Malawi and Mali), CAMBODIA (*XP-EHH* between Malawi and Cambodia) and THAILAND (*XP-EHH* between Malawi and Thailand).

2.6 Discussion

This work has applied current sequencing technologies and genomics to investigate genetic diversity, and the effect of immunity and antimalarial drugs selection on Malawi *P. falciparum* genomes. Studies of this kind have successfully identified drug-resistance mechanisms and targets of naturally-acquired immunity as candidate vaccine targets (Kidgell et al. 2006; Mackinnon and Marsh 2010), but have not yet been conducted in Malawi *P. falciparum* populations. Here examination of genetic variation is provided in parasites from Chikwawa district, Malawi, where the parasite population is exposed to naturally-acquired immunity and was recently exposed to intense pressure from IRS, ITNs, and ACTs. Previous studies have shown that human immune pressure produces genomic regions with high levels of nucleotide diversity, while antimalarial drug pressure results in regions with low diversity and extended haplotypes (Mu et al. 2007, 2010; Sabeti et al. 2006; Dharia et al. 2010; Volkman et al. 2007). This work has used population genetics metrics exploiting these two principles to detect loci under balancing and positive selection in a Malawi *P. falciparum* population. This population was also compared to five geographically dispersed others using F_{ST} and *XP-EHH* to detect regions of genetic divergence and signatures of recent selective sweeps, respectively. In particular, searching for high-scoring SNP clusters gave strong indicators of positive selection.

Analysis using Tajima's *D* identified potential genomic regions under balancing selection, including six genes encoding merozoite invasion ligands: *msp3.8, msp3, dbl-msp, eba175, ama1* and *surfin4.2*. These antigens are exposed to the immune system on the surface of merozoites or during erythrocyte invasion, and are highly polymorphic. Thus, balancing selection at these genes maybe mediated by host immune system as previously reported and have also been listed as possible candidates for vaccines in previous studies

(Alexandre et al. 2011; Baum et al. 2003; Polley and Conway 2001; Ochola et al. 2010; Tetteh et al. 2009; Mu et al. 2010; Amambua-Ngwa et al. 2012b).

Positive directional selection was detected in genomic regions near or surrounding drug targets (*pfmdr1*, *pfcrt, pfdhps* and *gch1*) and in surface antigens such as *trap, ron2, msp3.8*, *ama1* and *msp7* genes with important roles in invasion of host cells (Vulliez-Le Normand et al. 2012; Tufet-Bayona et al. 2009; Ghosh et al. 2009). Interestingly, several F_{ST} test results reflect parasite adaptation to local drug selection. First, low F_{ST} values in *pfcrt*-K76T between Malawi and Kenya may reflect the withdrawal of CQ in these regions and subsequent disparity in the reduction in the prevalence of resistance alleles to 2-4% in Malawi and 60% in Kenya (Nkhoma et al. 2007; Mwai et al. 2009). Second, the F_{ST} values in *pfcrt*-K76T between Malawi and Burkina Faso, and Malawi-Mali are heterogeneous and may suggest varying allele frequencies of K76T allele between Mali and Burkina Faso. Third, high F_{ST} values in *pfcrt*-K76T between Malawi and Cambodia, and Malawi-Mali, suggest that this mutation has reached fixation in Cambodia and Thailand; indeed, CQ remains the first-line treatment for *P. vivax* malaria in these two countries and thus may continue to select for the resistant genotype (Setthaudom et al. 2011). Fourth, fixation of *pfdhps*-K540E between Malawi and Mali, and Malawi-Burkina Faso may reflect the use of SP for the treatment of uncomplicated malaria and as intermittent preventive treatment in the two west-African countries, where the *pfdhps*-K540E mutation is rare (Pearce et al. 2009; Somé et al. 2010; Dicko et al. 2010). High prevalence of *pfdhps*-K540E and A437G is consistent with 90% prevalence of quintuple mutants in Malawi (Nkhoma et al. 2007) and while 437G is found all over Africa the 540E is largely absent in west Africa (Table 2.8). Whilst, F_{ST} may reflect differences in allele frequency due to differential selective pressure, they may also reflect

simply random genetic drift. F_{ST} is dependent on absolute diversity, where regions of low diversity in either population (or both) can result in high values, even if those regions have not been selected differently.

Positive directional selection in chromosome 12 containing *pfgch1* and transcription factors is also particularly interesting. In *P. vivax* it is thought to result from drug selection (Dharia et al. 2010). Mutations in these transcription factors are thought to be a source of increased genetic variability that regulate gene expression whose products may include drug-resistance genes (Levine and Tjian 2003). Increased expression levels of *pvcrt* have been observed in CQ-resistant parasites (Fernández-Becerra et al. 2009), and higher expression levels of *pvdhfr* occurred in *P. vivax* isolates relative to *P. falciparum,* resulting in the proposal that evolution in response to drug and immune pressure might be driven by genetic changes in the corresponding transcription factors (Westenberger et al. 2010).

In conclusion, this chapter describes the sequencing of 93 *P. falciparum* clinical isolates sourced from uncomplicated malaria cases in Malawi and identification of loci under selection. In addition, positive selection signals are identified by comparing Malawi to five other dispersed *P. falciparum* populations. Further work could evaluate the role of these loci in malaria intervention strategies. For example, the genetic variation may enable monitoring of *P. falciparum* transmission dynamics as the epidemiology of malaria changes over time in response to interventions (Volkman et al. 2012). In particular, by using *XP-EHH* and F_{ST} it is shown that selection differences between geographically dispersed populations reflect the history of antimalarial drug use and selection at any given time, whereas during intense drug selection, wild-type alleles are increasingly replaced by mutant alleles. The ability to use this strategy to monitor local adaptation to drug pressure, monitoring

transmission, and inform the type and timing of interventions is appealing. This knowledge will now be used in Malawi to monitor the impact of ACTs, ITNs and IRS on the local parasite population of Chikwawa district over three malaria seasons.

Chapter 3

Genome-wide identification of copy number variations in Malawi *P. falciparum* clinical isolates

3.1 Introduction

Structural variation generally refers to polymorphisms that affect the genomic structure, resulting into abnormal number of copies of one or more sections of the DNA. In humans, it was initially thought that genome variation was mostly due to SNPs, but it later become apparent, that structural variation of large scale duplication and insertions, occur at a much greater extent (12%) than SNPs (Iafrate et al. 2004; Redon et al. 2006; Tuzun et al. 2005; Feuk et al. 2006b). Their roles in determining natural phenotypic variations with human health, disease and evolution have also been highlighted (Zhang et al. 2009; Stankiewicz and Lupski 2010). Structural variation was originally defined as insertions, deletions and inversions greater than 1kb in size (Feuk et al. 2006a) but has since been widened to include smaller structural variants (for example, those >50-bp in length), particularly in the human genome. Currently, >50-bp is used as an operational demarcation between INDELs and copy number variants (CNV) (Feuk et al. 2006a). In *P.* falciparum, structural variation, particularly, CNV are increasingly being studied using techniques and bioinformatics software mostly adopted from human genome projects. Application of these methods has shown that CNV are also prevalent in this eukaryotic cell and act as a major source of genomic variation determining key phenotypes such as drug resistance and disease pathogenesis.

3.2 Causes of structural variation

P. falciparum is known to display vast genetic diversity throughout the genome and especially in the subtelomeric compartments of the 14 chromosomes, accounting for up to 76% of polymorphisms (Gardner et al. 2002; Volkman et al. 2002). In *P. falciparum* these subtelomeric regions are typified by blocks of repeat regions, large multi-gene families (*var, rifin, stevor*), are host-spots of inter-chromosomal recombination events and segmental duplications and share considerable sequence homology amongst chromosomes (Ribacke et al. 2007; Kidgell et al. 2006). Similar to smaller organisms such as *S. cerevisiae* where large polymorphic variations have been mapped to subtelomeric regions, *P. falciparum* also contain a number of large polymorphisms (>500-bp) that are produced by various mechanisms. During mitotic expansion of parasite populations, chromosome breakage and healing provides a mechanism for loss of subtelomeric sequences and genes (Scherf and Mattei 1992). Expansion of large tandem repeated units ('amplicons') also generates large chromosome size polymorphisms (Triglia et al. 1991). Generation of polymorphic variation/genetic diversity has also been observed through homologous recombination among large blocks of subtelomeric repetitive elements (Corcoran et al. 1988).

High recombination rates observed in *P. falciparum* generate high genetic diversity by allowing selection-targeted sites to evolve freely especially in regions containing the large multi-gene families, implicated in cell surface interactions (Bowman et al. 1999; Gardner et al. 2002), that are under strong immune-mediated diversifying selection (Volkman et al. 2002). Apart from sub-telomeres, copy number polymorphisms occur throughout the *P. falciparum* genome (Kidgell et al. 2006), including in drug resistance and invasion genes, suggesting that duplications and deletions are important determinants for parasite survival

and malaria pathogenesis (Sidhu et al. 2006; Van Tyne et al. 2011; Triglia et al. 2005). Correlations between CNV and repetitive sequences in human genomes have also highlighted the role of non-allelic homologous recombination (NAHR) in generating diversity (Hastings and Lupski 2009; Conrad and Hurles 2007) such as amplified genomic regions (amplicons) breakpoints in the *P. falciparum* MQ resistance locus (*pfmdr1*) on chromosome 5 and the GTP cyclohydrolase (*gch1*) occurring within microsatellite regions or repetitive monomeric A/T tracts (Nair et al. 2007, 2008).

Gene deletions usually have little phenotypic effects, owing to several mechanisms of compensation, and are usually purged from the populations (Gu et al. 2003; Conrad et al. 2006). The first is the existence of duplicated genes that ensures that loss of function in one copy can be compensated by other copy or copies. Second, compensation could also result from alternative metabolic pathways or regulatory networks (Gu et al. 2003).

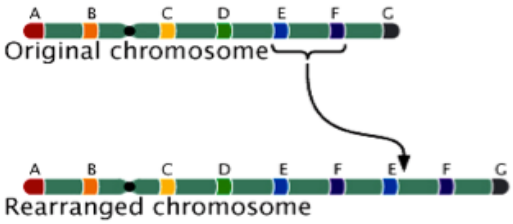
3.3 Types of structural variation

The spectrum of genetic variation in *P. falciparum* ranges from the single base pair (SNP) (Wootton et al. 2002) to large chromosomal events such as INDELs (Anderson et al. 2005), large scale deletions (Biggs et al. 1989), amplifications (Cowman et al. 1994), inversions (Pologe et al. 1990) and translocations (Hinterberg et al. 1994) (Figure 3.1). Discovering the full extent of structural variation, particularly in field isolates and to be able to genotype it routinely in order to understand its effects on malaria pathogenesis and evolution will be important.

The bulk of analysis of diversity in *P. falciparum* has concentrated on detection of SNPs and INDELs using oligonucleotide hybridization microarrays, CGH and DNA re-

sequencing methods (Kidgell et al. 2006; Jiang et al. 2008b; Volkman et al. 2007; Manske et al. 2012). These studies have provided a map of diversity (including of recombination hotspots) in the entire genome that is vast in subtelomeric regions of the 14 chromosomes, higher in genes encoding surface and drug proteins, whereas low in genes encoding mitochondrial, metabolic and cell growth proteins (Mu et al. 2005, 2007; Volkman et al. 2007, 2002; Gardner et al. 2002; Mu et al. 2010; Jiang et al. 2011). SNPs and INDELs, together with large gene deletions and amplifications have been shown to play a significant role in the adaptive biology of the parasite. With the advent of MPS, and a drastic cost reduction of DNA sequencing, there is unprecedented genomic resolution and large sample size applications that have accelerated efforts to map and catalogue structural variation in several genomes (Sepúlveda et al. 2013).

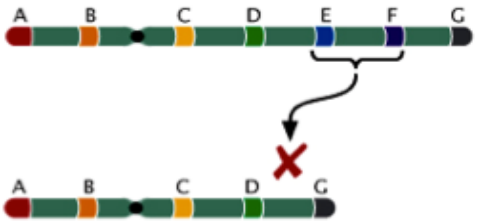
(A) Duplications – increases the number of copies of a chromosomal region



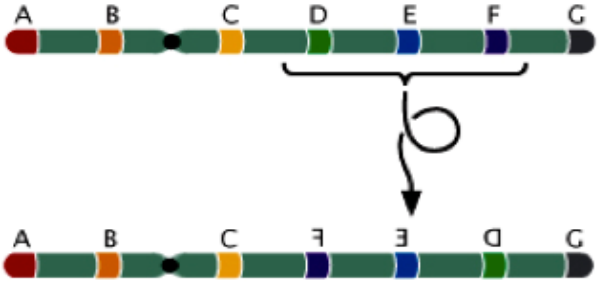
Tandem duplication – duplicated copy reside adjacent to duplicated region

Non-tandem/interspersed duplication –duplicated copy reside further apart from duplicated region

(B) Deletion – removal of a section of DNA

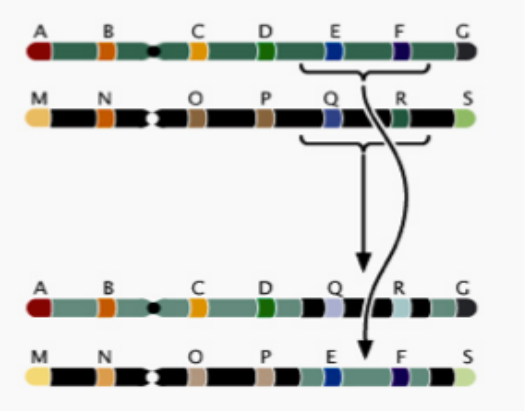


(C) Inversion – half circle rotation of a section of DNA after double-stranded break



Paracentric - does not include the centromere
Pericentric - includes the centromere

(D) Translocation– attachment of part of a chromosome to another



Non-reciprocal – unequal exchanges between non-homologous chromosomes

Reciprocal – two different parts of non-homologous chromosomes switch places

(E) Transposition – movement of part of a chromosome from one position to another

Figure 3.1: Types of structural variation. Adapted from http://www.kean.edu/~ftamari

3.4 Importance of structural variation in *P. falciparum*

3.4.1 Cytoadherence

Unlike other human malaria parasites, mature *P. falciparum*-infected erythrocytes (IE) are known to adhere and block human microvasculature, leading to the development of severe malaria (SM). Ligands expressed on the surface of IE can bind to a number of endothelial cell receptors, including CD36, intercellular adhesion molecule-1, thrombospondin, chondroitin-4-sulfate, vascular cell adhesion molecule-1, E-selectin, and platelet endothelial cell adhesion molecule-1 (Craig et al. 2012; Chakravorty et al. 2008). One molecule identified on the surface of IE, known as *P. falciparum* erythrocyte membrane protein-1 (PfEMP-1) encoded by *var* genes, has been correlated with this cytoadherence (Scherf et al. 2008). An important feature of *P. falciparum* intra-erythrocytic stages is the parasite's ability to export specific proteins to the IE surface, thus modifying the surface of IE to become rigid and inflexible. A reduced flexibility of IE makes their circulation through the microvasculature difficult and favours their adhesion to endothelial cells (Dondorp et al. 2004). Higher levels of IE seen in cerebral vessels of patients dying from CM compared to non-CM cases demonstrated involvement of IE cytoadhesion in the brain to cerebral malaria (MacPherson 1985). The point of contact for this cytoadhesion was identified as a knob-like structure on the surface of IE caused by deposition of knob-associated His-rich protein (KAHRP) under the red cell membrane, that contains several proteins including PfEMP1 (MacPherson 1985; Howard et al. 1990). A major contribution of KAHRP in cytoadherence was illustrated in *in vitro* culture of *P. falciparum* where some parasites lost the ability to produce knobs and subsequently cytoadherence ability. This loss of function was shown to be a consequence of subtelomeric deletions of the region of chromosome 2 containing the

kahrp gene (Biggs et al. 1989). Subtelomeric deletions have also been shown to be involved in a loss of IE cytoadherence through deletion of genes encoding adhesion molecules (Trenholme et al. 2000). In this study they showed that targeted gene disruption of cytoadherence linked asexual gene on chromosome 9 (*clag9*) reduced cytoadherence to C32 human melanoma cells and CD36. Antibodies raised against clones with disrupted *clag9* gene did not react to it.

3.4.2 Anti-malarial drug resistance

Amplifications on chromosome 12 of GTP cyclohydrolase 1, the first enzyme in the folate biosynthesis pathway is likely due to compensation for the decreased efficiency of the folate pathway caused by mutations in *pfdhps* and *pfdhfr* (Kidgell et al. 2006; Nair et al. 2008). The importance of the multidrug resistance (MDR) gene (*pfmdr1*) copy number in determining the *P. falciparum* susceptibility to a variety of antimalarial drugs has been well documented. This gene is a member of ABC transporter family, is located on chromosome 5 and encodes a predicted 12-transmembrane-domain protein (also known as “Pgh-1”) (Foote et al. 1989), which localizes to the parasite digestive vacuole, a site of action of CQ and other quinoline-based antimalarial drugs, including quinine (QN) (Cowman et al. 1991; Sidhu et al. 2006). Development of resistance has rendered several drugs including CQ and SP ineffective, and of concern now, are reports of drug failure of currently used first line drug, ACT.

CQ resistance in *P. falciparum* provided a classic example of how this parasite evolved to overcome drug pressure and has been replicated in latter studies that sort to elucidate resistance to other drugs. A decrease in CQ concentration in resistant parasites more than in sensitive ones was linked to CQ resistance phenotype (Krogstad et al. 1987). A

complex mechanism was observed between *P. falciparum* response to CQ and other drugs, where MQ-selected parasite lines showed increased expression of the *pfmdr1* gene, MQ resistance, decreased CQ resistance and cross-resistance to HF and QN (Cowman et al. 1994). A decreasing *pfmdr1* copy number was reported to heighten susceptibility to ACT regimens including LF and ART drugs (Sidhu et al. 2006)

More recently, a member of MSP3 multigene family, merozoite surface protein encoded by *msp3.8* gene was associated with antimalarial drug resistance. Through functional testing using genome-wide association approach, this study demonstrated that *msp3.8* overexpression in *P. falciparum* decreased sensitivity to HF, MQ, and LF and that increased gene copy number mediated resistance (Van Tyne et al. 2011). Interestingly, positive selection was also detected at this locus in this analysis.

3.5 Detecting structural variation in *P. falciparum*

Originally developed to profile RNA, microarrays have now been used to detect SNPs, gene duplications and loss. Microarray methods are based on hybridization of target DNA to immobilized cDNA oligonucleotide probes in glass slides. Commonly used microarrays for detecting CNV are array-CGH and SNP microarrays with Affymetrix and Illumina being the leading brands. However, sequencing-based methods for the detecting CNV are becoming more cost-effective and popular, and several tools have been developed to detect CNV e.g., Delly (Rausch et al. 2012), FREEC (Boeva et al. 2011), CNVnator (Abyzov et al. 2011).

3.5.1 Hybridization-based SNP micro-array methods

SNP microarrays have been used to determine genomic copy number. In this method

20 matched and mismatched probe pairs that are 25 bases long are designed to each SNP allele. Only the test sample is hybridized, with no competitively hybridizing reference sample (co-hybridization of two DNA) as in array-CGH. To improve the signal-to-noise ratio, the DNA is digested with a restriction enzyme and ligated with fluorescent adapters and then the smaller fragments amplified using universal primers to reduce the complexity of the hybridization. The intensity of the fluorescence upon binding is used as a measure for the matching sequences in the sample. For the detection of CNV, the signal intensities of the match and mismatch probes are clustered and compared with values from another individual (or group of individuals) and relative copy number per locus is determined (Winchester et al. 2009; Carter 2007).

This method has a number of advantages such as its ability to be used in SNP genotyping and targeted copy number analysis. It also requires less sample volume *per* experiment (as compared to CGH) and is cost effective allowing more sample analysis on a limited budget. Disadvantages include being biased towards detecting mostly known CNV (as SNP chip coverage are biased towards known SNPs) and because CNV are most common in regions containing high levels of segmental duplication (such as sub-telomeres) with low SNP coverage, SNP arrays may decrease the number of CNV or polymorphisms typed (Winchester et al. 2009). Cheeseman et al (Cheeseman et al. 2009), successfully used this method, a custom designed Affymetrix PFSANGER GeneChip array to describe CNV throughout the *P. falciparum* genome and reported that they form a major source of genetic variation in *P. falciparum*. He observed that CNV allow adaptive expansion of diverse gene families whilst copy numbers of core “housekeeping” genes are maintained by negative selection.

3.5.2 Array-CGH

Array-CGH is widely used in detecting CNV by competitively hybridizing two fluorescently labelled samples to the target DNA (cohydridization technique). Briefly, the experimental and reference DNA are fragmented in size and differentially fluorescently labelled and hybridized in duplicate together to the array. The resulting fluorescence ratio/signal intensity is measured and quantified. Previously, DNA sequences were used to construct arrays included larger-insert clones (40–200-kb in size), small insert clones (1.5–4.5-kb), cDNA clones (0.5–2-kb), and genomic PCR products (100-bp–1.5-kb) and this resulted to CNV detection at low resolution. However, many developments made in array technology over the past years have led to an increase in resolution at which CNV are detected by increasing number of features ("spots") and using shorter DNA sequences e.g., oligonucleotides (at 25-80bp). However, much of resolution of an array across the whole genome still depends on the number, distribution and lengths of probes. The ability to detect CNV also largely depends on signal-to-noise ratio and probe response characteristics. Normalization is also often needed to correct for experimental biases for GC/AT content in the DNA. The main advantage of the array-CGH to others is the co-hybridization of the experimental and reference DNAs where the reported fluorescent ratios are influenced less by spotted probe concentration (signal intensity) and variations in slide production and processing (Carter 2007; Ylstra et al. 2006).

3.5.3 Sequencing based computational approaches (SBC)

Next generation sequencing techniques (NGS) have revolutionized SNP genotyping and structural variation discovery, and are slowly replacing microarrays. Unlike array-based

methods that can only detect gains and losses in respect to the reference used to design the probes, sequencing based methods can detect previously unknown polymorphisms and variants such as inversions, translocations (Figure 3.2) and thus the higher cost is justified. In SBC, smaller variants and the exact location of a variation breakpoint are also easily detected at high resolution. Although more expensive than array-based methods, SBC can be robustly used in high throughput large-scale studies as shown with human genome projects such as the 1000 Genomes project. However, analysis of sequence-based data requires more computational work, but is superior to array based computational methods which are limited by the size and breakpoint resolution of prediction which is directly correlated to the density of the probes on the array, which in turn is limited by either the density of the array itself (for array-CGH) or by the density of known SNP loci (for SNP arrays) (Carter 2007; Alkan et al. 2011; Medvedev et al. 2009).

There are four general strategies for SV discovery in SBC and largely rely on mapping sequencing reads to the reference genome, Pf3D7 (with very few involving *de novo* assembly), in the case of *P. falciparum* and subsequently identifying genomic differences.

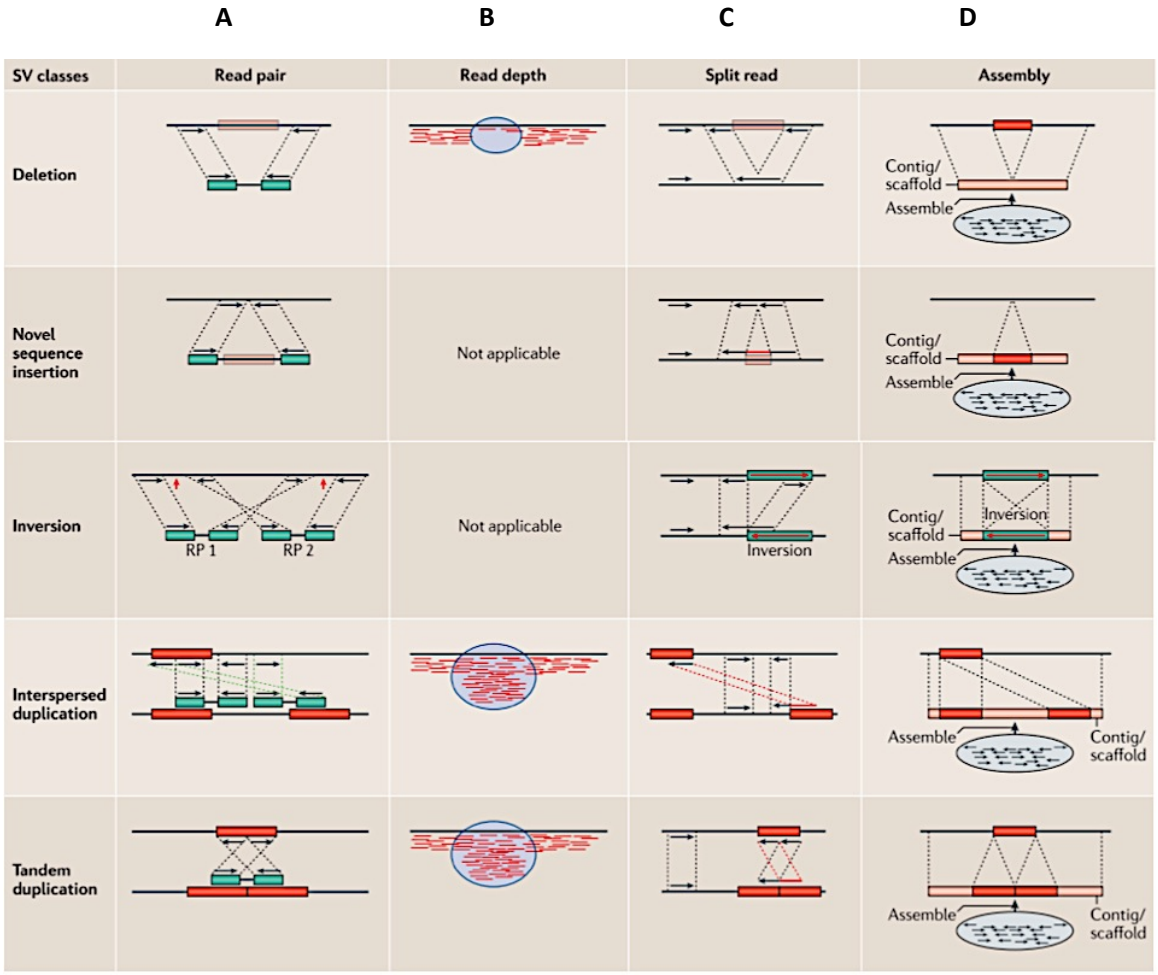


Figure 3.2: Structural variation sequence discovery methods using (A) read pair, (B) read depth, (C) split read and (D) sequence assembly methods. A basic deletion event occurs when a mate pair maps to the corresponding regions of the reference, but the insert size is less than the mapped distance. Conversely, an event is an insertion, if the insert size is greater than the mapped distance. Insertion may involve a novel sequence or a mobile-element. A mobile element insertion can result from translocations or duplications. Duplications can occur as tandem duplications, where the duplicated segment remains adjacent to the source DNA, or interspersed, where the duplicated DNA is incorporated elsewhere in the genome. These events may occur within a single chromosome (intra-chromosomally) or inter-chromosomally (between different chromosomes). A basic inversion signature occurs when the order of the two mates mapping to the reference is preserved but one of them changes orientation. Adapted from (Alkan et al. 2011).

3.5.3.1 Read pair methods

Current sequencing technologies are capable of generating paired-end or mate-paired reads. In read pair sequencing (producing paired-end reads), genomic DNA is fragmented into short segments/libraries (approximately 300-500-bp), followed by sequencing of both ends of the segment. In mate-pair sequencing, a circularized fragment is used. Genomic DNA is fragmented and size-selected inserts (usually, 1.5-20-kb) are circularized and linked by means of an internal adaptor, to create mate-pairs. The circularized fragment is then randomly sheared, and segments containing the adaptor purified and sequenced to generate mate pairs. Mate pair libraries, compared to paired-end, are more DNA intensive owing to the low yield of circularization of large DNA molecules. In this method, variation in insert sizes plays an important role in SV detection, for example the large insert for mate-pair libraries allows for detection of larger structural events than do paired-end libraries. However, paired-end reads provide tighter insert-size distributions and because both ends of library fragments are sequenced, it is valuable for highly repetitive genomes; aligning at least one end read of a pair uniquely onto the reference sequence provides sufficient certainty that the read pair is uniquely mapped to its locus of origin. Equally disadvantageous is aligning short, single end or 'fragment' reads to the genome which results in a higher proportion of non-unique reads that cannot be used for variant discovery (Mardis 2011; Alkan et al. 2011; Medvedev et al. 2009).

Read pairs are mapped back to the reference genome to detect SV. Mapping span/distance and orientation of the reads to the reference genome are assessed and a discordant distance or orientation of the read pair indicates the occurrence of a genomic rearrangement (Figure 3.2A). Read pairs aligning too far apart than the expected mapping

distance suggests a deletion in sample/insertion in reference (where insert size is smaller than mapping distance), whereas read pairs mapping closer together are indicative of an insertion. A novel insertion is produced when only one end of a read pair maps to the reference while the others do not. Orientation inconsistencies predict inversion breakpoints. Interspersed duplications or translocations are detected by more complex patterns where in several pairs one of the reads maps to a different location or chromosome. Finally, tandem duplications can be detected by read pairs that have a correct orientation, but are reversed in their order and have differences in their span (Alkan et al. 2011; Medvedev et al. 2009). Several tools such as Delly (Rausch et al. 2012) are based on read pair technology.

3.5.3.2 Read depth methods

Also known as depth of coverage (DOC), it is probably the best method for predicting copy numbers (Alkan et al. 2009). DOC assumes a random Poisson distribution in mapping depth and any deviation from this distribution is assumed to be due to copy number in the sequenced genome (Bailey et al. 2002). The basic concept is that a duplicated/gain region will show higher read depth while deleted/loss region will show reduced read depth (Figure 3.2B). The main disadvantage of this method is that only CNV can be detected. Because it is largely related to sequence coverage, sequence biases will affect SV detection, for example, in *P. falciparum* with AT rich and repeat regions, are sequenced less reliably with low read depth, while GC-rich will have higher read depth. Mis-mapped reads will also negatively influence SV detection. Larger variants are more reliably detected than smaller ones as statistical power increases with the size (Harismendy et al. 2009). Methods that have used this approach include CNVnator (Abyzov et al. 2011) and FREEC (Boeva et al. 2011).

3.5.3.3 Split-read methods

In split read method un-mappable or partially mappable reads are used to detect SV (Figure 3.2C). A broken read alignment to the reference genome defines SV breakpoint ('split' sequence read signature). Split reads mapping to the reference genome in parts or showing a continuous stretch of alignment gaps indicates a deletion in the sample or an insertion in the reference. Reads spanning tandem duplications will have the split read mapping in reverse order. Interspersed duplications or inter-chromosomal translocations are defined by reads mapping partly to the duplicated region or another chromosome. Clustering of reads is used to increase the reliability of SV detection in read pair method. This method is powerful in detecting SV as small as a single base pair, but the power is greatly reduced by shorter read lengths as the length of a split read generated from shorter reads is rarely uniquely mapped to the reference genome. Therefore, this method is greatly affected by the use of current NGS platforms that produce short reads. However, NGS technologies generating longer reads will make this approach more powerful.

3.5.3.4 Sequence assembly

A complete genome sequencing approach using technologies that produce long sequence reads, would allow an accurate whole genome assembly and thereby identifying all SV (accurately identifying breakpoints) by comparing it to a high quality reference genome (Miller et al. 2010) (Figure 3.2D). This approach is possible using the 'split' read method to detect SV, but its application is difficult due to short reads lengths that are produced by current NGS methods that require a combination of *de novo* and local assembly algorithms to produce contigs that can be compared to a reference genome

(Miller et al. 2010). *De novo* assembly is more accurate when dealing with highly characterisable or unique genomic regions, but its application is more limited when identifying SV in repetitive regions or highly polymorphic regions such as sub-telomeres in *P. falciparum.*

To provide a comprehensive analysis of copy number variation (mainly gains and losses) in a Malawian *P. falciparum* population, two complimentary methods using read depth coverage principle were used, *FREEC* (Boeva et al. 2011) and Poisson hierarchical modelling approach (*PG*, developed by Sepúlveda et al. 2013). In theory these two tests rely on assumptions that deletions show regions with extremely low coverage whereas amplifications are found in regions with exceptionally high coverage. Particularly, the *PG* method was used because of two reasons: that it uses read depth of coverage approach (one of the best in detecting amplification and deletions) and it is flexible and robust in handling different patterns of data, for example those generated by NGS where coverage is never uniform across all samples and individual isolates appear unique. In this scenario, a test that is performed across all samples with default settings is less sensitive as it fails to deal with underlying uniqueness of each sample data. The *PG* method had also been tested in *P. falciparum* lab and clinical isolates and was found to be very effective and sensitive, accurately measuring CNV including previously identified ones, and outperforming other DOC-CNV detecting algorithms (Sepúlveda et al. 2013).

Table 3.1: List of important *P. falciparum* CNV previously detected.

Gene	Name	CNV	Method
PF3D7_0523000	Multidrug resistance protein	Amp	*PG*/PCR
PF3D7_0302200	Cytoadherence linked asexual protein 3.2	Del	*PG*/PCR
PF3D7_0424400	SURFINs	Dup	PCR
PF3D7_0202000	Knob-associated histidine-rich protein	Del	PFGE
PF3D7_0935800	Cytoadherence linked asexual protein 9	Del	PCR/SNP array
PF3D7_0402300	Reticulocyte binding protein homologue 1	Amp	PCR/gene disruption/CGH
PF3D7_1224000	GTP cyclohydrolase I	Amp	*PG*/FREEC/CGH
PF3D7_0930300	Merozoite surface protein 1	Del	*PG*
PF3D7_1148700	Plasmodium exported protein (PHISTc)	Amp	*PG*
PF3D7_1253000	Gametocyte erythrocyte cytosolic protein	Del	*PG*
PF3D7_0522400	Conserved Plasmodium protein	Amp	*PG*
PF3D7_0713400	Serpentine receptor, putative	Del	*PG*
PF3D7_0935900	Ring-exported protein 1	Del	Affymetrix array
PF3D7_0935600	Gametocytogenesis-implicated protein	Del	Affymetrix array
PF3D7_0501300	Skeleton-binding protein 1	Del	Affymetrix array
PF3D7_0501400	Interspersed repeat antigen	Del	Affymetrix array
PF3D7_1401000	Glycophorin binding protein	Del	
PF3D7_0414300	Rab5-interacting protein, putative	Amp	*PG*
PF3D7_1223400	Phospholipid-transporting ATPase, putative	Amp	*PG*
PF3D7_1301200	Glycophorin binding protein	Del	Affymetrix array
PF3D7_1038400	Gametocyte-specific protein	Amp	Affymetrix array
PF3D7_1228300	NIMA related kinase 1	Amp	Affymetrix array
PF3D7_0102200	Ring-infected erythrocyte surface antigen	Amp	Affymetrix array/CGH
PF3D7_1035700	Duffy binding-like merozoite surface protein	Del	Affymetrix array
PF3D7_0935400	Gametocyte development protein 1	Del	Affymetrix array
PF3D7_0522900	Zinc finger protein, putative	Amp	CGH
PF3D7_0722400	GTP binding protein, putative	Amp	CGH

Del=Deletion. Amp=Amplification.

3.6 Methodology

3.6.1 Sequence data

All the Malawi samples (n=93) described above were used in this analysis. BAM files for all the 93 samples generated from the previous section (Chapter 2) were used to detect CNV.

3.6.2 Detecting Copy Number Variation using *FREEC* and *PG*

3.6.2.1 Pre-analysis steps in FREEC

The *FREEC* (version 6.2) algorithm was run on the Malawi samples. Briefly, chromosome length file (chrLen) containing details of the 14 chromosome lengths was generated using a perl script provided with the *FREEC* software. A configuration file containing locations of the chrLen file, fasta files (for individual *P. falciparum* chromosomes) and BAM files (for the 93 samples) were generated, as required by the *FREEC* algorithm. Important parameters included in the configuration file were the minimum and maximum expected GC contents that were set at 0 and 0.6 respectively, with a stepwise window size (500-bp, stepwise at 250-bp) for calculation of raw copy number profile.

3.6.2.2 Estimating coverage profiles and CNV detection using FREEC

Running *FREEC* followed several steps. In brief, for each of the 96 samples, *FREEC* calculates raw copy number (coverage) profile by counting the number of mapped reads in equal-size windows of 500-bp, and a stepwise increment of 250-bp (personal communication from Valentina Boeva – *FREEC* software developer). Within each window, the numbers of mapped reads are counted using their starting mapping positions to provide coverage depth. *P. falciparum* is AT rich, leading to coding regions with higher GC content.

In general, sequence coverage is influenced by GC-content (where a non-linear relationship between read depth and GC-content has been observed, Yoon et al. 2009). *FREEC* assumes a GC-content bias and after calculating this GC-content, it does a polynomial regression to find the dependency function i.e., read count (RC) per window ~ GC-content. It then normalizes the RC by fitting the observed read count by the GC profile (i.e., divides the RC by the corresponding value of function) to produce smooth normalised ratio profiles that are segmented and analysed for genomic gains and losses. During normalization, values close to 1 show there is no copy number alteration (CNA), values significantly lower than 1 (but higher than 0) show loss and gains have values significantly higher than 1. Because of high levels of noise due to mapping differences, *FREEC* does not annotate each window separately, instead it segments normalised ratio profiles, and calculates median values of these normalized RC of each segment. Median value falling closer to 1, show no CNA; closer to 0, show loss and gains are median values closer to 1.5.

3.6.2.3 Detecting copy number variation using PG

To detect copy number variation (CNV) in each *P. falciparum* sample, we used sequence coverage data and a Poisson hierarchical modelling approach for the respective data analysis. This CNV discovery strategy is briefly outlined. First, using *FREEC*, the 3D7 reference genome was divided into non-overlapping windows of 100-bp each and the coverage of each window calculated using *FREEC* by counting the number of reads mapped onto it and using the starting position of the reads. This window size was used as previously reported as it provided a good resolution for estimating copy number profile (Yoon et al. 2009), within which observed read count distribution approached standard normal distribution. Larger window size (e.g., ≥ 1000-bp) resulted in less precision/resolution in

defining CNV breakpoints with smaller CNV (that span one or two windows) being particularly difficult to detect (Sepúlveda et al. 2013; Yoon et al. 2009; Boeva et al. 2011). Second, quality control included removing windows: within 100-kb of subtelomeric regions, containing reads with poor mapping score, those associated with high polymorphic multi-gene families (e.g., *vars, stevors, and rifins*) and those in non-coding regions. *PG* analysis therefore focused only on the *P. falciparum* exome, which on average contained 2-fold coverage. This filtering process resulted into 120,309 100-bp windows accounting for nearly 53% of the 3D7 reference genome. Third, since the underlying GC content can affect coverage, we divided windows according to this genomic variable using the 10%-centiles of its distribution across the whole genome to account for the non-uniform GC-content distribution in the 3D7 reference genome (Gardner et al. 2002), to ensure similar statistical power across the different set of windows. Fourth, each set of windows with similar GC-content is analysed separately by fitting two flexible probability distributions, Poisson-Gamma and Poisson-Lognormal, to the respective coverage data. Fifth, the best model for each individual data set is determined and used to define lower and upper bounds at 99% credible level for the expected coverage under the hypothesis of no copy number variation. Sixth, windows were classified individually into no copy variation, deletions, or amplifications, if the respective coverage values were between those expected bounds, lesser than the lower bound, or greater than upper bound, respectively. At this stage of the analysis we considered false positive deletions if the coverage per nucleotide position of the corresponding windows was greater than zero. The final step is to pool together windows with the same classification in order to identify larger regions containing putative CNV. A more detailed description and discussion of this analysis approach can be found elsewhere (Sepúlveda et al. 2013).

3.7 Results D

3.7.1 Distribution of CNV across chromosomes and *P. falciparum* genomes

Analysis of CNV focused on the *P. falciparum* exome. The sub-telomeres, internal regions associated with highly polymorphic gene families (including *vars*, *stevors*, and *rifins*) were discarded. Non-coding regions and isolates with low genome coverage (<10-fold) were also removed from analysis.

PG model (γ = 99%, where > γ increases the "credibility" of those SV identified or lowers the chances of a false positive) identified 2,374 CNV genes in 49 isolates and at γ = 99.9%, 1,335 genes in 48 isolates were identified (Table 3.2). The median size of CNV detected by *PG* model was 100-bp (range=100-1,700-bp and 100-8,600-bp for amplifications and deletions respectively). Cdc2-related protein kinase 1 and gametocyte specific protein contained the largest amplification (1,700-bp), while the largest deletion (8,600-bp) spans two conserved *Plasmodium* proteins of unknown function on chromosome 14.

FREEC identified a total of 429 CNV genes across 72 isolates (Table 3.2). The median size of CNV is 2,749 (range=499-177,409-bp) and 20,249 (range=499-194,906-bp) for amplifications and deletions respectively. The largest deletion (194,906-bp) on chromosome 10 spans 33 genes while the largest amplification (285,749-bp) spans 52 genes on chromosome 10.

The *PG* method was more sensitive in detecting CNV and detected a larger number of previously identified CNV (compared to *FREEC*). *FREEC*, however, identified larger CNV than the *PG* method (Table 3.2). CNV were detected in all chromosomes and isolates (Table 3.3-3.4 and Figure 3.3). There was an excess of CNV in larger chromosomes. There was a

slight bias of CNV, with amplifications being more frequent than deletions and maybe as a result of positive selection acting on duplications to favour the retention of paralogs (Table 3.2).

CNV size distribution differed between isolates, but in general smaller size CNV occur at higher frequencies than larger ones (Figure 3.3). It has been proposed that CNV evolution (particularly deletions) is largely shaped by purifying selection and would act more effectively in larger populations, such as African *P. falciparum* populations (Dopman and Hartl 2007). Several theories have been proposed as to why smaller CNV could be selected for, such as, high multi-clonal infections creating a greater inter-clonal competition with parasites, with smaller CNV outcompeting those with larger ones and high outbreeding in African populations that could lead to the breakdown of CNV size as well (Dopman and Hartl 2007). Final CNV analysis was focussed to those polymorphisms spanning only a single gene.

Table 3.2: Summary of identified CNV. Hits2 and Genes2 correspond to total hits and CNV genes identified respectively.

					Size (bp)					
		CNV type	Hits	No. of genes	Range	Mean	Median	Hits2	Genes2	No. of isolates
PG	γ =99.9%	Amplifications	11,907	1,771	100-1,800	140	100	15,871	2,374	49
		Deletions	3,964	1,049	100-8,600	148	100			
	γ =99%	Amplifications	4,266	912	100-1,700	140	100	6,889	1,335	48
		Deletions	3,341	611	100-8,600	153	100			
FREEC		Amplifications	2,480	396	499-177,409	8,645	2,749	2,689	429	72
		Deletions	209	51	499-194,906	32,520	20,249			

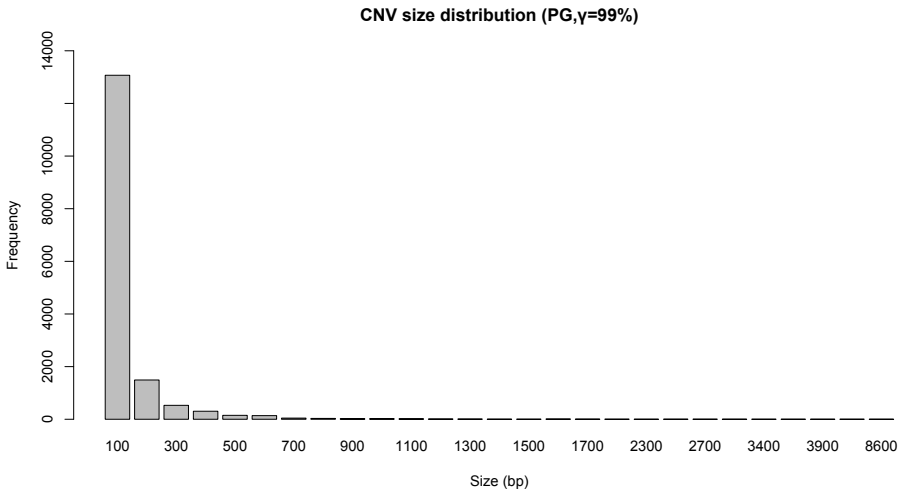
Table 3.3: Chromosomal distribution of CNV detected by *FREEC* and *PG*. CNV were detected across all *P. falciparum* chromosomes.

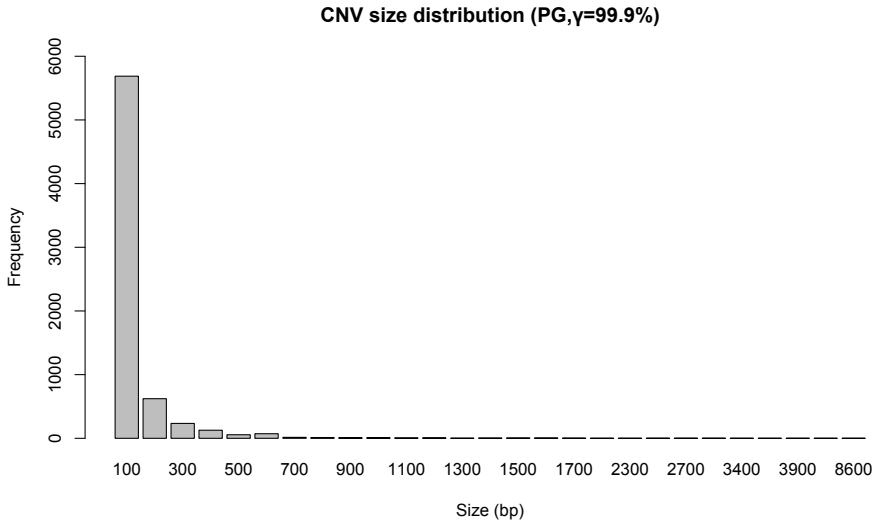
	≥ 500-bp			≥ 100-bp	
Chromosome	*FREEC*	γ=99%	γ=99.9%	γ=99%	γ=99.9%
1	81	37	28	557	296
2	195	6	2	664	275
3	219	48	18	1209	587
4	113	29	12	942	401
5	177	10	0	1017	407
6	68	18	3	726	353
7	258	14	4	1168	504
8	152	9	4	789	243
9	141	46	12	903	349
10	212	63	36	1623	819
11	284	38	25	1481	649
12	173	20	3	1273	549
13	272	72	30	1773	801
14	344	65	42	1746	656

Table 3.4: Distribution of CNV in isolates detected by *FREEC* (≥ 500-bp)and *PG* (≥ 500 and 100-bp) methods.

	≥ 500-bp			≥ 100-bp	
Isolate	*FREEC*	γ=99%	γ=99.9%	γ=99%	γ=99.9%
1	110	0	0	0	0
2	73	35	13	560	189
3	48	15	6	578	165
4	79	23	9	561	247
5	108	54	15	763	255
6	73	34	8	517	183
7	103	0	0	0	0
8	54	0	0	0	0
9	94	42	17	680	281
10	73	29	14	586	244
11	68	1	0	19	1
12	66	26	6	554	183
13	69	39	6	742	216
14	61	10	5	362	109
15	27	9	3	468	241
16	68	0	0	0	0
17	61	8	3	479	206
18	59	10	3	492	209
19	22	8	4	358	168
20	32	5	2	279	89
21	32	0	0	0	0
22	31	9	6	369	173
23	25	7	6	266	154
24	21	0	0	0	0
25	41	2	1	185	100
26	33	1	1	279	112
27	28	5	4	244	179
28	39	0	0	0	0
29	30	1	1	275	105
30	41	0	0	0	0
31	20	0	0	0	0
32	22	0	0	0	0
33	9	0	0	0	0
34	12	34	34	144	106
35	11	5	4	169	119
36	45	0	0	0	0
37	42	2	0	540	191
38	10	0	0	0	0

	≥ 500-bp			≥ 100-bp	
Isolate	*FREEC*	γ=99%	γ=99.9%	γ=99%	γ=99.9%
39	29	1	1	412	141
40	40	0	0	0	0
41	19	2	1	309	145
42	29	0	0	0	0
43	44	0	0	0	0
44	28	4	3	238	125
45	34	0	0	0	0
46	44	0	0	173	92
47	4	4	5	170	99
48	29	7	6	383	213
49	13	0	0	0	0
50	29	3	3	195	123
51	21	4	3	371	223
52	11	6	6	353	158
53	16	0	0	0	0
54	10	1	0	67	50
55	20	1	0	251	135
56	26	3	2	269	127
57	15	0	0	109	42
58	44	0	0	186	123
59	13	0	0	0	0
60	23	4	1	211	106
61	36	0	0	201	93
62	29	0	0	0	0
63	29	0	0	0	0
64	6	2	2	124	62
65	24	0	0	0	0
66	8	5	4	280	67
67	16	2	2	170	54
68	13	5	3	261	131
69	46	7	6	272	180
70	18	0	0	162	60
71	41	0	0	234	115
72	42	0	0	0	0





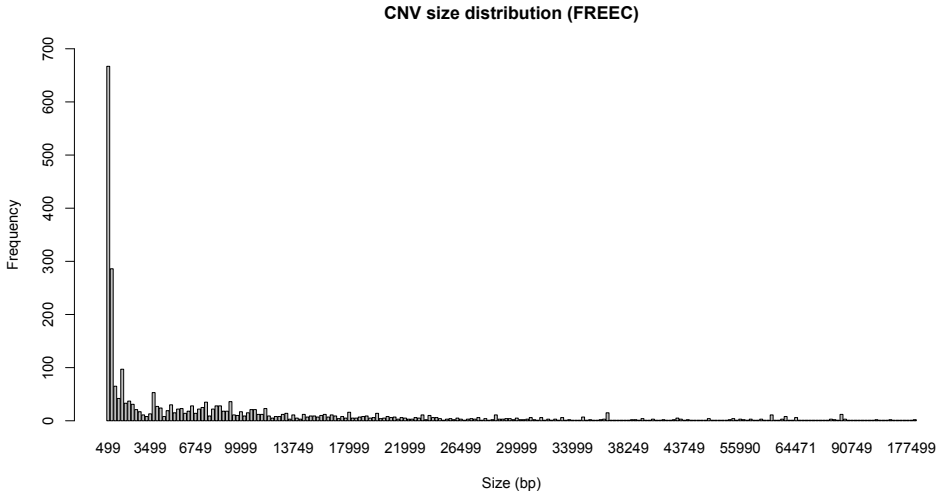


Figure 3.3: CNV size distribution using *FREEC* and *PG*. Smaller size CNV (less than 500-bp) dominate in the population.

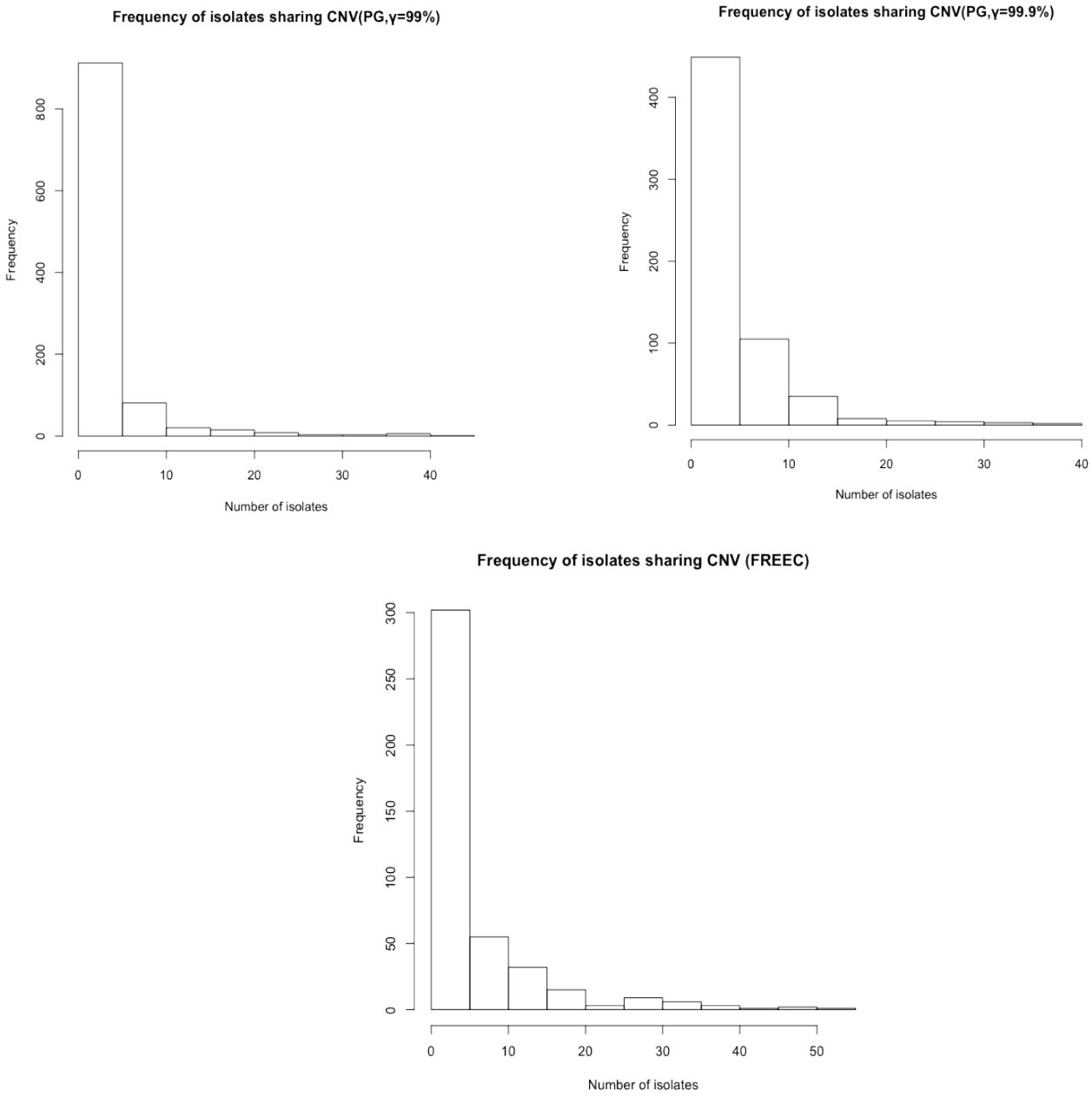


Figure 3.4: Frequency of polymorphic and monomorphic CNV. Polymorphic CNV (occurring in single isolates) are more frequent than monomorphic CNV (occurring in multiple isolates)

3.7.2 Detection of previously identified CNV

These two methods identified important CNV that have previously been identified (Table 3.3). Well-documented deletions with roles in cytoadherence, invasion and pathogenesis were detected. These include cytoadherence linked asexual genes on chromosome 3 (*clag 3.1* and *clag 3.2,* PF3D7_0302500 and PF3D7_0302200 respectively), reticulocyte binding protein 2 homologue b (*pfRh2b,* PF3D7_1335300) on chromosome 13, gametocyte-specific protein (*pf11-1,* PF3D7_1038400) on chromosome 10, and a locus on chromosome 2 containing the knob-associated histidine-rich protein (*kahrp*, PF3D7_0202000) also associated with loss of cytoadherence. Deletions were also detected in two multigene families associated with the merozoite surface, a Duffy binding-like merozoite surface protein (*dbl-msp,* PF3D7_1035700) with a possible role in binding of merozoites with erythrocytes during invasion (Wickramarachchi et al. 2009) and merozoite surface protein (*msp3.8*, PF3D7_1036300). Another deletion was detected in antigen 332 (*Pf332*, PF3D7_1149000). Parasites lacking this gene have been found to be rigid, less adhesive to "Cluster of Differentiation 36" (CD36), and with decreased expression of the major cytoadherence ligand, PfEMP1, on the IE surface and thus could have a role in malaria pathogenesis (Glenister et al. 2009). Amplifications were detected in 8 genes including in *msp3.8*, that had previously been associated with parasite drug resistance, ring-infected erythrocyte surface antigen (RESA, PF3D7_0102200), ring-exported protein 1 (*rex1,* PF3D7_0935900) and *pf11-1*. Residing in Maurer's cleft (a transporter of parasite proteins to IE surface), the *rex1* gene makes a group of four homologous genes (*rex1, rex2, rex3, rex4*) found on chromosome 9 that encode ring-stage proteins exported into the IE (Spielmann and Hawthorne 2006; Hawthorne et al. 2004; Bhattacharjee et al. 2008).

Table 3.5: List of previously identified CNV detected in this study. CNVp represents CNV type in previous studies and CNVn represents CNV type in this study.

Gene	Product description	CNVp	CNVn	Method
PF3D7_0102200	Ring-infected erythrocyte surface antigen	Amp	Amp	*PG*
PF3D7_0202000	Knob-associated histidine-rich protein	Del	Del	*FREEC*
PF3D7_0302200	Cytoadherence linked asexual protein 3.2	Del	Del	*FREEC/PG*
PF3D7_0302500	Cytoadherence linked asexual protein 3.1	Del	Del	*PG*
PF3D7_0321600	ATP-dependent RNA helicase, putative	Amp	Amp	*PG*
PF3D7_0522900	Zinc finger protein, putative	Amp	Amp	*PG*
PF3D7_0722400	GTP binding protein, putative	Amp	Amp	*PG*
PF3D7_0935900	Ring-exported protein 1	Del	Amp	*PG*
PF3D7_1035700	Duffy binding-like merozoite surface protein	Del	Del	*PG*
PF3D7_1036300	Merozoite surface protein	Amp/Del	Del	*PG*
PF3D7_1038400	Gametocyte-specific protein	Amp/Del	Amp/Del	*FREEC/PG*
PF3D7_1149000	Antigen 332, DBL-like protein	Del	Del	*PG*
PF3D7_1228300	NIMA related kinase 1	Amp	Amp	*FREEC*
PF3D7_1335300	Reticulocyte binding protein 2 homologue b	Del	Del	*PG*

Del=Deletion. Amp=Amplification.

3.7.3 Identification of deletions (loss) spanning only a single gene

Apart from deletions listed in Table 3.5 that were previously identified, several others were detected (Table 3.6-3.8), and these include: a deletion on chromosome 10 containing a liver stage antigen 1 (*lsa1,* PF3D7_1036400), with an important role in parasite transition from liver to blood (Mikolajczak et al. 2011). A deletion on chromosome 7 in a well characterized invasion gene, erythrocyte binding antigen (*eba-175,* PF3D7_0731500) (Sakura et al. 2013). Deletion in sporozoite-specific transmembrane protein S6 (*trep,* PF3D7_1442600) that aids in parasites gliding mortility (Combe et al. 2009) and in a circumsporozoite- and TRAP-related protein (*ctrp,* PF3D7_0315200), an essential protein in ookinite infectivity and mosquito transmission of malaria (Dessens et al. 1999). Other deletions were found in two structural RNAs (18S and 28S ribosomal RNA) with possible roles in ribosomal gene expression (Chakrabarti et al. 2007), and in a cysteine repeat modular protein 4 (*pfcrmp4*) on chromosome 14 that mediates host-parasite interaction throughout *P. falciparum* life cycle and specifically transmission from the mosquito to the

host (Douradinha et al. 2011). The largest deletion spanning a single chromosome appeared on *pf11-1* (5,249-bp, *FREEC*) and conserved *Plasmodium* protein, PF3D7_1474400 (3,900-bp, *PG*). Several genes appeared at very high frequencies by *PG* method - a DBL containing protein, PF3D7_0113800 (75%), *clag3.2* (70%), *clag3.1* (50%) and *pf11-1* (57%) and by *FREEC*, *clag3.2* occurred in 24% of all isolates

Table 3.6: List of deletions (≥ 500-bp) identified by *FREEC*.

Chromosome	Gene	Size	Product description
Pf3D7_10_v3	PF3D7_1038400	5249	gametocyte-specific protein
Pf3D7_03_v3	PF3D7_0302200	3749	cytoadherence linked asexual protein 3.2
Pf3D7_07_v3	PF3D7_0726000	1999	28S ribosomal RNA
Pf3D7_02_v3	PF3D7_0202000	1249	knob-associated histidine-rich protein
Pf3D7_14_v3	PF3D7_1442600	999	sporozoite-specific transmembrane protein S6
Pf3D7_10_v3	PF3D7_1036400	999	liver stage antigen 1
Pf3D7_03_v3	PF3D7_0315200	749	circumsporozoite- and TRAP-related protein
Pf3D7_13_v3	PF3D7_1371000	749	18S ribosomal RNA
Pf3D7_14_v3	PF3D7_1473700	749	nucleoporin NUP116/NSP116, putative

Table 3.7: List if deletions ≥ 500-bp identified by *PG* (γ = 99%).

Chromosome	Gene	Size	Product description
Pf3D7_14_v3	PF3D7_1474400	3900	conserved Plasmodium protein, unknown function
Pf3D7_14_v3	PF3D7_1473700	2700	nucleoporin NUP116/NSP116, putative
Pf3D7_03_v3	PF3D7_0302200	1200	cytoadherence linked asexual protein 3.2
Pf3D7_10_v3	PF3D7_1038400	1100	gametocyte-specific protein
Pf3D7_03_v3	PF3D7_0302500	900	cytoadherence linked asexual protein 3.1
Pf3D7_01_v3	PF3D7_0113800	800	DBL containing protein, unknown function
Pf3D7_10_v3	PF3D7_1036300	600	merozoite surface protein
Pf3D7_10_v3	PF3D7_1036400	600	liver stage antigen 1
Pf3D7_13_v3	PF3D7_1335300	600	reticulocyte binding protein 2 homologue b
Pf3D7_14_v3	PF3D7_1442600	600	sporozoite-specific transmembrane protein S6
Pf3D7_07_v3	PF3D7_0731500	500	erythrocyte binding antigen-175
Pf3D7_10_v3	PF3D7_1035700	500	duffy binding-like merozoite surface protein
Pf3D7_11_v3	PF3D7_1149000	500	antigen 332, DBL-like protein
Pf3D7_14_v3	PF3D7_1417900	500	conserved Plasmodium protein, unknown function

Table 3.8: List deletions ≥ 500-bp, identified by *PG* (γ = 99.9%).

Chromosome	Gene	Size	Product description
Pf3D7_14_v3	PF3D7_1474400	3900	conserved Plasmodium protein, unknown function
Pf3D7_14_v3	PF3D7_1475400	3800	cysteine repeat modular protein 4
Pf3D7_14_v3	PF3D7_1473700	2700	nucleoporin NUP116/NSP116, putative
Pf3D7_14_v3	PF3D7_1476300	1300	Plasmodium exported protein (PHISTb), unknown function
Pf3D7_03_v3	PF3D7_0302200	1100	cytoadherence linked asexual protein 3.2
Pf3D7_10_v3	PF3D7_1038400	1100	gametocyte-specific protein
Pf3D7_03_v3	PF3D7_0302500	800	cytoadherence linked asexual protein 3.1
Pf3D7_01_v3	PF3D7_0113800	600	DBL containing protein, unknown function
Pf3D7_10_v3	PF3D7_1036300	600	merozoite surface protein
Pf3D7_10_v3	PF3D7_1036400	600	liver stage antigen 1
Pf3D7_13_v3	PF3D7_1335300	600	reticulocyte binding protein 2 homologue b
Pf3D7_10_v3	PF3D7_1035700	500	duffy binding-like merozoite surface protein
Pf3D7_11_v3	PF3D7_1149000	500	antigen 332, DBL-like protein
Pf3D7_13_v3	PF3D7_1318300	500	conserved Plasmodium protein, unknown function
Pf3D7_14_v3	PF3D7_1417900	500	conserved Plasmodium protein, unknown function
Pf3D7_14_v3	PF3D7_1442600	500	sporozoite-specific transmembrane protein S6

3.7.4 Identification of amplifications (gains) spanning a single gene

All amplifications detected by *FREEC* and *PG* methods are listed in Table 3.9-4.1. There were more amplifications than deletions both by *FREEC* and *PG* method (Tables 3.2). Several genes contained both deletions and amplifications e.g., *pf11-1, ctrp, lsa1,* 28S and 18S ribosomal RNAs. Of interest is a previously unidentified amplification on *kahrp* gene, noting that deletions on this gene have been associated with loss of cytoadherence. Amplification was also detected in a well-characterized malaria vaccine candidate, circumsporozoite protein (*csp,* PF3D7_0304600) (Plassmeyer et al. 2009) a major surface component of *P. falciparum* sporozoites and essential for host cell invasion. Using *FREEC*, the highest copy number gains (223 and 63) were observed in two putative genes; a nucleoporin NUP116/NSP116 (*PfNup116,* PF3D7_1473700) and a eukaryotic translation initiation factor 3 subunit 10 (PF3D7_1212700) respectively. In addition, 23 gains in 6-cysteine protein (*p230,* PF3D7_0209000), 18 gains in *pf11-1*, 16 gains in *kahrp*, and 15 gains in *csp* were observed. In terms of frequency, *lsa1* occurred at 62% by *FREEC* and 68% by *PG*

(γ = 99%), 79% of isolates (*PG*, γ = 99%) contained ABC transporter, (*mrp2,* PF3D7_1229100), *rex1* (72%; *PG*, γ = 99%). On average 64% of isolates contained *pf11-1* gene by *PG* method. Largest amplifications were observed in a *Plasmodium* conserved protein spanning 5,249-bp (PF3D7_0213600; *FREEC*) and *pf11-1* (1,700-bp, *PG*).

Table 3.9: List of amplifications (≥ 500-bp) identified by *FREEC*.

Chromosome	Gene	Gain	Size	Product description
Pf3D7_02_v3	PF3D7_0213600	3	5249	conserved Plasmodium protein, unknown function
Pf3D7_10_v3	PF3D7_1038400	18	4999	gametocyte-specific protein
Pf3D7_07_v3	PF3D7_0726000	2	2999	28S ribosomal RNA
Pf3D7_10_v3	PF3D7_1014600	3	2249	transcriptional coactivator ADA2
Pf3D7_10_v3	PF3D7_1036400	3	2249	liver stage antigen 1
Pf3D7_11_v3	PF3D7_1148600	2	1999	18S ribosomal RNA
Pf3D7_09_v3	PF3D7_0905100	2	1749	nucleoporin NUP100/NSP100, putative
Pf3D7_11_v3	PF3D7_1116000	5	1499	rhoptry neck protein 4
Pf3D7_02_v3	PF3D7_0220000	2	1249	liver stage antigen 3
Pf3D7_09_v3	PF3D7_0909500	2	1249	subpellicular microtubule protein 1, putative
Pf3D7_02_v3	PF3D7_0202000	16	999	knob-associated histidine-rich protein
Pf3D7_05_v3	PF3D7_0527000	2	999	DNA replication licensing factor MCM3, putative
Pf3D7_08_v3	PF3D7_0803400	2	999	DNA repair protein RAD54, putative
Pf3D7_01_v3	PF3D7_0112700	7	749	28S ribosomal RNA
Pf3D7_02_v3	PF3D7_0209000	24	749	6-cysteine protein
Pf3D7_03_v3	PF3D7_0304600	15	749	circumsporozoite (CS) protein
Pf3D7_03_v3	PF3D7_0315200	9	749	circumsporozoite- and TRAP-related protein
Pf3D7_05_v3	PF3D7_0508900	3	749	conserved Plasmodium protein, unknown function
Pf3D7_05_v3	PF3D7_0525200	2	749	conserved Plasmodium protein, unknown function
Pf3D7_08_v3	PF3D7_0807100	3	749	RNA helicase, putative
Pf3D7_11_v3	PF3D7_1110400	2	749	asparagine-rich antigen
Pf3D7_11_v3	PF3D7_1147800	3	749	merozoite adhesive erythrocytic binding protein
Pf3D7_12_v3	PF3D7_1212700	63	749	eukaryotic translation initiation factor 3 subunit 10, putative
Pf3D7_12_v3	PF3D7_1228300	6	749	NIMA related kinase 1
Pf3D7_13_v3	PF3D7_1311900	2	749	vacuolar ATP synthase subunit a
Pf3D7_13_v3	PF3D7_1367800	9	749	secreted ookinete protein, putative
Pf3D7_14_v3	PF3D7_1462800	9	749	glyceraldehyde-3-phosphate dehydrogenase
Pf3D7_14_v3	PF3D7_1473700	223	749	nucleoporin NUP116/NSP116, putative

Table 4.0: List of amplifications ≥ 500-bp, identified using *PG* (γ = 99%).

Chromosome	Gene	Size	Product description
Pf3D7_10_v3	PF3D7_1038400	1700	gametocyte-specific protein
Pf3D7_04_v3	PF3D7_0417800	1700	cdc2-related protein kinase 2
Pf3D7_03_v3	PF3D7_0311300	1700	phosphatidylinositol 3- and 4-kinase, putative
Pf3D7_11_v3	PF3D7_1104600	1700	radial spoke head protein, putative
Pf3D7_03_v3	PF3D7_0310200	1600	phd finger protein, putative
Pf3D7_03_v3	PF3D7_0318200	1600	DNA-directed RNA polymerase II, putative
Pf3D7_10_v3	PF3D7_1033000	1600	conserved Plasmodium protein, unknown function
Pf3D7_07_v3	PF3D7_0716100	1500	large ribosomal subunit assembling factor, putative
Pf3D7_12_v3	PF3D7_1251200	1500	coronin
Pf3D7_13_v3	PF3D7_1309400	1400	HORMA domain protein, putative
Pf3D7_01_v3	PF3D7_0102200	1300	ring-infected erythrocyte surface antigen
Pf3D7_04_v3	PF3D7_0404600	1300	conserved Plasmodium membrane protein, unknown function
Pf3D7_11_v3	PF3D7_1128300	1300	6-phosphofructokinase
Pf3D7_09_v3	PF3D7_0924500	1200	conserved Plasmodium membrane protein, unknown function
Pf3D7_04_v3	PF3D7_0408700	1100	sporozoite micronemal protein essential for cell traversal
Pf3D7_06_v3	PF3D7_0607300	1100	uroporphyrinogen III decarboxylase
Pf3D7_09_v3	PF3D7_0934100	1100	DNA excision-repair helicase, putative
Pf3D7_13_v3	PF3D7_1323800	1100	vacuolar protein sorting-associated protein 52, putative
Pf3D7_14_v3	PF3D7_1423400	1100	conserved Plasmodium membrane protein, unknown function
Pf3D7_03_v3	PF3D7_0321600	1000	ATP-dependent RNA helicase, putative
Pf3D7_07_v3	PF3D7_0716200	1000	conserved Plasmodium protein, unknown function
Pf3D7_08_v3	PF3D7_0828800	1000	GPI-anchored micronemal antigen
Pf3D7_10_v3	PF3D7_1004600	1000	conserved Plasmodium membrane protein, unknown function
Pf3D7_10_v3	PF3D7_1033100	1000	S-adenosylmethionine decarboxylase/ornithine decarboxylase
Pf3D7_11_v3	PF3D7_1104500	1000	conserved Plasmodium protein, unknown function
Pf3D7_11_v3	PF3D7_1125100	1000	vacuolar membrane protein-related, putative
Pf3D7_13_v3	PF3D7_1350400	1000	ubiquitin-activating enzyme E1, putative
Pf3D7_13_v3	PF3D7_1365700	1000	conserved Plasmodium membrane protein, unknown function
Pf3D7_03_v3	PF3D7_0316500	900	conserved Plasmodium protein, unknown function
Pf3D7_03_v3	PF3D7_0318200	900	DNA-directed RNA polymerase II, putative
Pf3D7_04_v3	PF3D7_0404600	900	conserved Plasmodium membrane protein, unknown function
Pf3D7_08_v3	PF3D7_0803400	900	DNA repair protein RAD54, putative
Pf3D7_10_v3	PF3D7_1031300	900	conserved Plasmodium protein, unknown function
Pf3D7_11_v3	PF3D7_1143500	900	conserved Plasmodium protein, unknown function
Pf3D7_12_v3	PF3D7_1249800	900	conserved Plasmodium protein, unknown function
Pf3D7_13_v3	PF3D7_1365700	900	conserved Plasmodium protein, unknown function
Pf3D7_13_v3	PF3D7_1323900	900	conserved Plasmodium protein, unknown function
Pf3D7_13_v3	PF3D7_1350900	900	transcription factor with AP2 domain(s)
Pf3D7_14_v3	PF3D7_1421600	900	conserved Plasmodium protein, unknown function
Pf3D7_14_v3	PF3D7_1432200	900	conserved Plasmodium protein, unknown function
Pf3D7_03_v3	PF3D7_0303500	800	spindle pole body protein, putative
Pf3D7_03_v3	PF3D7_0318300	800	conserved Plasmodium protein, unknown function
Pf3D7_03_v3	PF3D7_0310200	800	phd finger protein, putative
Pf3D7_04_v3	PF3D7_0407300	800	transcription factor, putative
Pf3D7_04_v3	PF3D7_0406700	800	conserved Plasmodium protein, unknown function
Pf3D7_07_v3	PF3D7_0716200	800	conserved Plasmodium protein, unknown function
Pf3D7_08_v3	PF3D7_0818500	800	zinc finger protein, putative
Pf3D7_10_v3	PF3D7_1031200	800	MORN repeat-containing protein 1
Pf3D7_10_v3	PF3D7_1036400	800	liver stage antigen 1
Pf3D7_12_v3	PF3D7_1233200	800	conserved Plasmodium protein, unknown function
Pf3D7_13_v3	PF3D7_1350700	800	N6-adenine-specific methylase, putative
Pf3D7_14_v3	PF3D7_1421800	800	conserved Plasmodium protein, unknown function
Pf3D7_03_v3	PF3D7_0311300	700	phosphatidylinositol 3- and 4-kinase, putative
Pf3D7_05_v3	PF3D7_0504800	700	conserved Plasmodium protein, unknown function

Pf3D7_10_v3	PF3D7_1004600	700	conserved Plasmodium protein, unknown function
Pf3D7_11_v3	PF3D7_1128300	700	6-phosphofructokinase
Pf3D7_13_v3	PF3D7_1348500	700	TBC domain protein, putative
Pf3D7_13_v3	PF3D7_1323800	700	vacuolar protein sorting-associated protein 52, putative
Pf3D7_14_v3	PF3D7_1440500	700	allantoicase, putative
Pf3D7_14_v3	PF3D7_1423400	700	conserved Plasmodium membrane protein, unknown function
Pf3D7_01_v3	PF3D7_0110800	600	transcription initiation factor TFIIB, putative
Pf3D7_02_v3	PF3D7_0212700	600	conserved Plasmodium protein, unknown function
Pf3D7_03_v3	PF3D7_0316600	600	formate-nitrite transporter, putative
Pf3D7_03_v3	PF3D7_0313600	600	conserved Plasmodium protein, unknown function
Pf3D7_04_v3	PF3D7_0407100	600	methyltransferase, putative
Pf3D7_04_v3	PF3D7_0417700	600	conserved Plasmodium protein, unknown function
Pf3D7_04_v3	PF3D7_0417900	600	conserved Plasmodium protein, unknown function
Pf3D7_05_v3	PF3D7_0525500	600	conserved Plasmodium protein, unknown function
Pf3D7_05_v3	PF3D7_0508900	600	conserved Plasmodium protein, unknown function
Pf3D7_06_v3	PF3D7_0615300	600	GPI-anchored wall transfer protein 1, putative
Pf3D7_07_v3	PF3D7_0704100	600	conserved Plasmodium membrane protein, unknown function
Pf3D7_09_v3	PF3D7_0925100	600	conserved Plasmodium protein, unknown function
Pf3D7_11_v3	PF3D7_1128500	600	conserved protein, unknown function
Pf3D7_12_v3	PF3D7_1233700	600	homocysteine S-methyltransferase, putative
Pf3D7_12_v3	PF3D7_1237200	600	conserved Plasmodium protein, unknown function
Pf3D7_12_v3	PF3D7_1229100	600	ABC transporter, (CT family)
Pf3D7_13_v3	PF3D7_1330200	600	conserved Plasmodium protein, unknown function
Pf3D7_13_v3	PF3D7_1322900	600	conserved Plasmodium protein, unknown function
Pf3D7_13_v3	PF3D7_1340300	600	nucleolar complex protein 2, putative
Pf3D7_13_v3	PF3D7_1304600	600	conserved Plasmodium protein, unknown function
Pf3D7_13_v3	PF3D7_1309400	600	HORMA domain protein, putative
Pf3D7_13_v3	PF3D7_1322400	600	conserved Plasmodium protein, unknown function
Pf3D7_01_v3	PF3D7_0111000	500	kinesin-8, putative
Pf3D7_03_v3	PF3D7_0312200	500	TPR domain containing protein
Pf3D7_04_v3	PF3D7_0419900	500	phosphatidylinositol 4-kinase, putative
Pf3D7_04_v3	PF3D7_0418000	500	conserved Plasmodium protein, unknown function
Pf3D7_05_v3	PF3D7_0506500	500	conserved Plasmodium protein, unknown function
Pf3D7_05_v3	PF3D7_0530600	500	XAP-5 DNA binding protein, putative
Pf3D7_07_v3	PF3D7_0722400	500	GTP binding protein, putative
Pf3D7_09_v3	PF3D7_0928200	500	conserved Plasmodium protein, unknown function
Pf3D7_09_v3	PF3D7_0932600	500	apicoplast ribosomal protein S6, putative
Pf3D7_09_v3	PF3D7_0934700	500	UBX domain, putative
Pf3D7_09_v3	PF3D7_0908200	500	conserved Plasmodium protein, unknown function
Pf3D7_09_v3	PF3D7_0924500	500	conserved Plasmodium protein, unknown function
Pf3D7_09_v3	PF3D7_0935900	500	ring-exported protein 1
Pf3D7_10_v3	PF3D7_1018200	500	serine/threonine protein phosphatase, putative
Pf3D7_10_v3	PF3D7_1025000	500	formin 2, putative
Pf3D7_10_v3	PF3D7_1020500	500	conserved Plasmodium protein, unknown function
Pf3D7_11_v3	PF3D7_1117000	500	conserved Plasmodium membrane protein, unknown function
Pf3D7_11_v3	PF3D7_1106800	500	protein kinase, putative
Pf3D7_12_v3	PF3D7_1245600	500	kinesin, putative
Pf3D7_12_v3	PF3D7_1249900	500	apicoplast dimethyladenosine synthase, putative
Pf3D7_13_v3	PF3D7_1307300	500	DEAD box helicase, putative
Pf3D7_13_v3	PF3D7_1350400	500	ubiquitin-activating enzyme E1, putative
Pf3D7_14_v3	PF3D7_1455300	500	conserved Plasmodium protein, unknown function
Pf3D7_14_v3	PF3D7_1422100	500	mitochondrial ribosomal protein L21 precursor, putative
Pf3D7_14_v3	PF3D7_1435700	500	ataxin-2 like protein, putative
Pf3D7_14_v3	PF3D7_1435600	500	conserved Plasmodium protein, unknown function

Table 4.1: List of amplifications ≥ 500-bp, identified using *PG* (γ = 99.9%).

Chromosome	Gene	Size	Product description
Pf3D7_10_v3	PF3D7_1038400	1700	gametocyte-specific protein
Pf3D7_04_v3	PF3D7_0417800	1700	cdc2-related protein kinase 1
Pf3D7_11_v3	PF3D7_1104600	1700	radial spoke head protein, putative
Pf3D7_10_v3	PF3D7_1033000	1600	conserved Plasmodium protein, unknown function
Pf3D7_11_v3	PF3D7_1104500	1200	conserved Plasmodium protein, unknown function
Pf3D7_03_v3	PF3D7_0311300	1100	phosphatidylinositol 3- and 4-kinase, putative
Pf3D7_12_v3	PF3D7_1251200	1100	coronin
Pf3D7_09_v3	PF3D7_0932100	1000	protein MAM3, putative
Pf3D7_03_v3	PF3D7_0318200	900	DNA-directed RNA polymerase II, putative
Pf3D7_06_v3	PF3D7_0607300	900	uroporphyrinogen III decarboxylase
Pf3D7_11_v3	PF3D7_1125100	900	vacuolar membrane protein-related, putative
Pf3D7_13_v3	PF3D7_1350900	900	transcription factor with AP2 domain(s)
Pf3D7_08_v3	PF3D7_0818500	800	zinc finger protein, putative
Pf3D7_08_v3	PF3D7_0828800	700	GPI-anchored micronemal antigen
Pf3D7_10_v3	PF3D7_1031200	700	MORN repeat-containing protein 1
Pf3D7_10_v3	PF3D7_1033100	700	S-adenosylmethionine decarboxylase/ornithine decarboxylase
Pf3D7_11_v3	PF3D7_1108500	700	succinyl-CoA synthetase alpha subunit, putative
Pf3D7_03_v3	PF3D7_0310200	600	phd finger protein, putative
Pf3D7_04_v3	PF3D7_0406700	600	conserved Plasmodium protein, unknown function
Pf3D7_13_v3	PF3D7_1330200	600	conserved Plasmodium protein, unknown function
Pf3D7_13_v3	PF3D7_1330300	600	DnaJ protein, putative
Pf3D7_13_v3	PF3D7_1348500	600	TBC domain protein, putative
Pf3D7_01_v3	PF3D7_0102200	500	ring-infected erythrocyte surface antigen
Pf3D7_01_v3	PF3D7_0110800	500	transcription initiation factor TFIIB, putative
Pf3D7_04_v3	PF3D7_0404600	500	conserved Plasmodium membrane protein, unknown function
Pf3D7_04_v3	PF3D7_0417900	500	conserved Plasmodium protein, unknown function
Pf3D7_06_v3	PF3D7_0615300	500	GPI-anchored wall transfer protein 1, putative
Pf3D7_10_v3	PF3D7_1031300	500	conserved Plasmodium protein, unknown function
Pf3D7_13_v3	PF3D7_1323900	500	conserved Plasmodium protein, unknown function
Pf3D7_14_v3	PF3D7_1422100	500	mitochondrial ribosomal protein L21 precursor, putative
Pf3D7_14_v3	PF3D7_1422200	500	conserved Plasmodium protein, unknown function
Pf3D7_14_v3	PF3D7_1423400	500	conserved Plasmodium membrane protein, unknown function

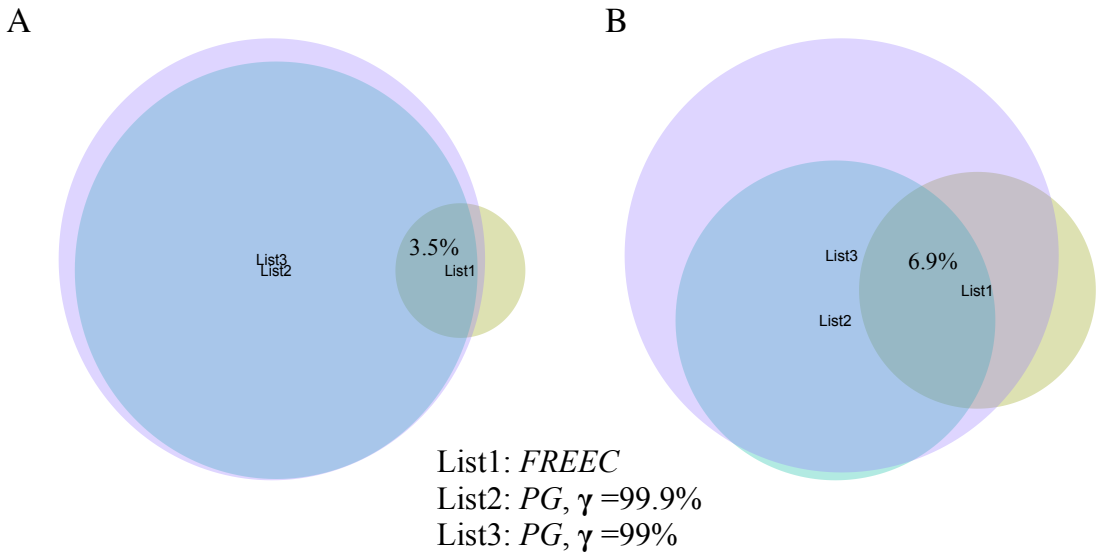


Figure 3.5: Overlap of CNV detected by the three methods. A: 3.5% overlap of deletions. B: 6.9% overlap of duplications. More than 70% overlap of CNV detected by *PG*, γ = 99% and *PG*, γ = 99.9% was observed, as the stringency of the two methods is controlled by γ, where as you increase γ from 99 to 99.9%, the number of false positives is lowered. The low proportion of shared CNV between the three methods is mainly due to the sensitivity of the methods; *PG* method detects hits, most of which occur in loci not targeted by *FREEC*. Less overlap between the three methods could also result from the reduced number of CNV detected by *FREEC* owing to the default settings that could only identify CNV ≥ 500-bp, while *PG* identified CNV ≥ 100-bp.

3.7.5 Using copy number variation to define population structure in Malawian *P. falciparum* isolates

CNV were used to define population structure of Malawian *P. falciparum* parasite populations using a principal component analysis (PCA) (Figure 3.8). As in PCA using SNP variation, there was no structure explained by Malawi CNV.

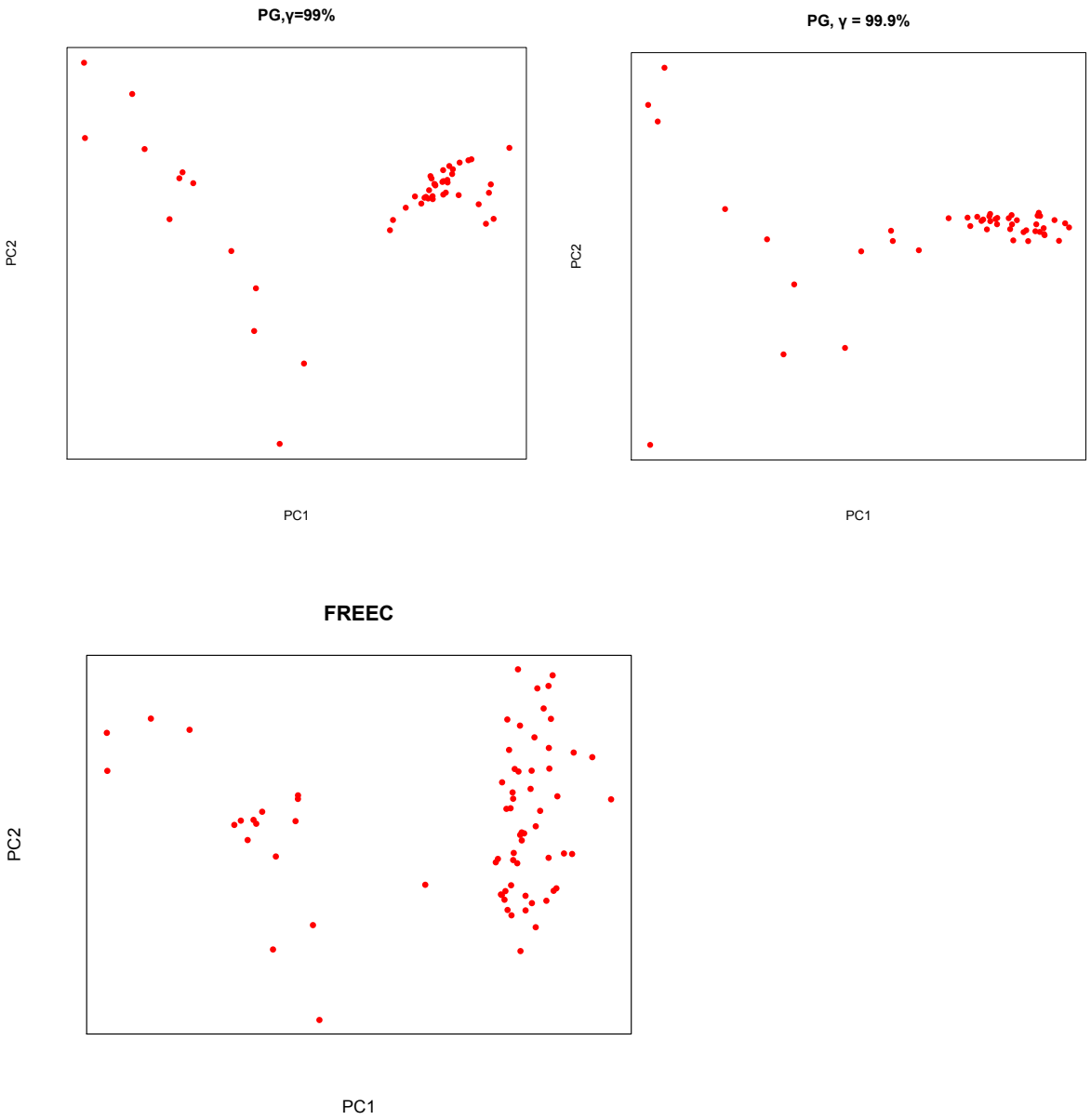
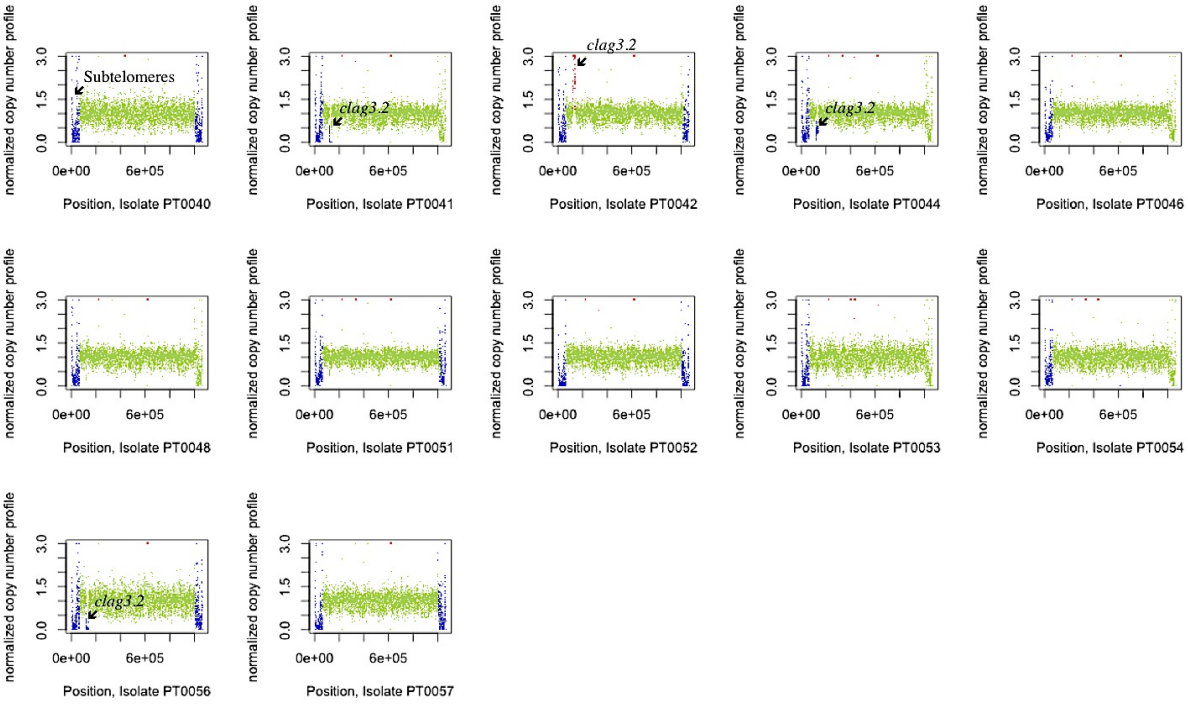


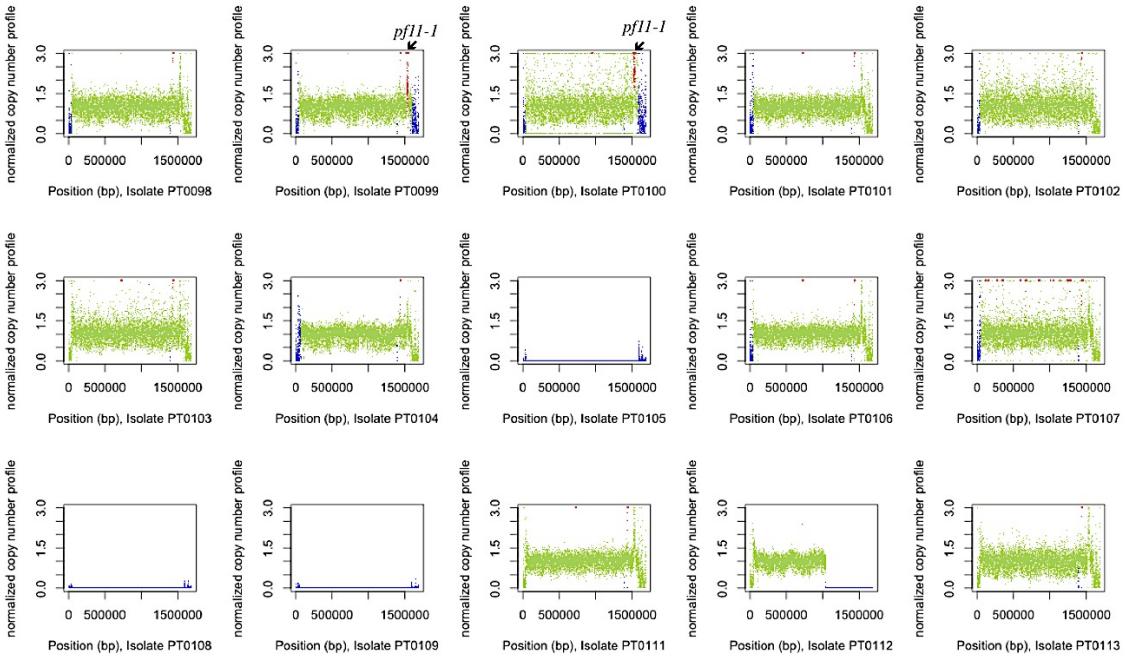
Figure 3.6: Population structure inferred from principal component analysis of CNV identified by *FREEC* (A), *PG*, γ = 99.9% (B) and *PG*, γ = 99% (C). PCA on CNV showed no structure.

3.8 Visual representation of copy number variation

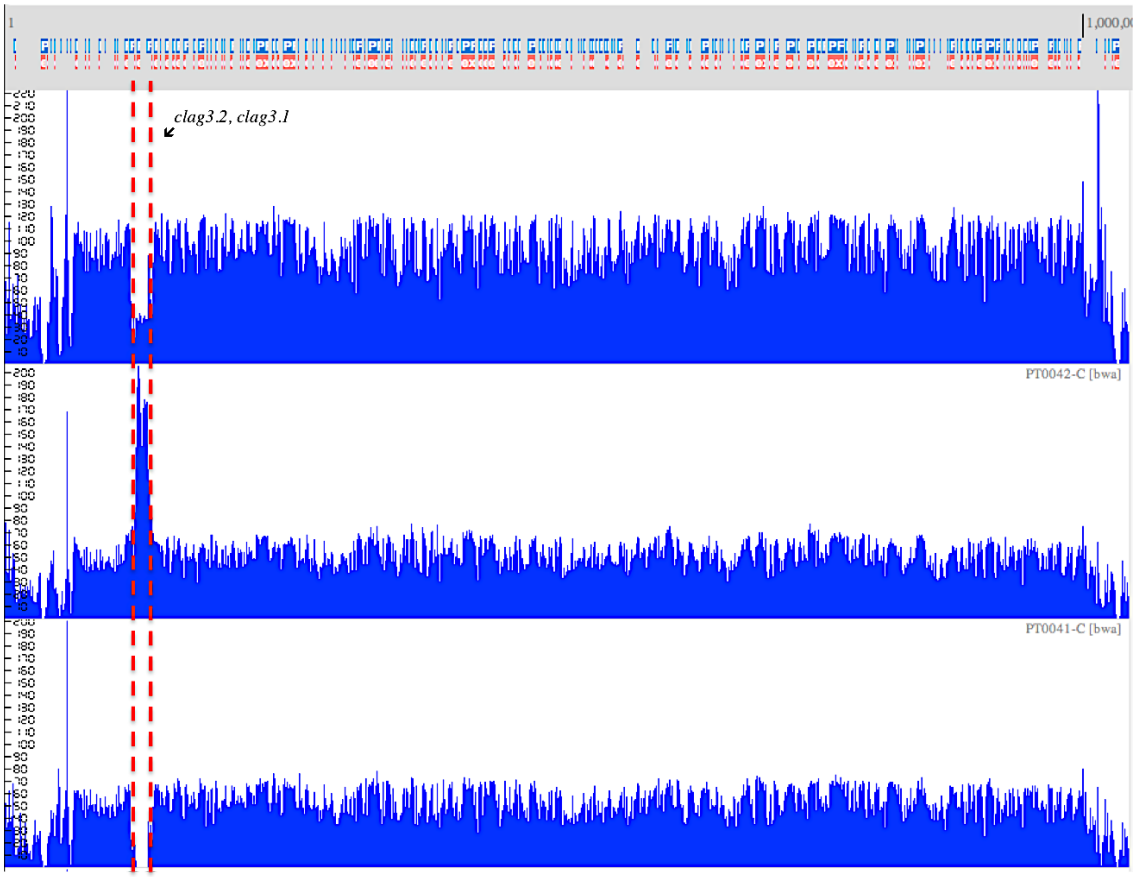
A.



B.



C.


clag3.2, clag3.1
PT0042-C [bwa]
PT0041-C [bwa]

D.

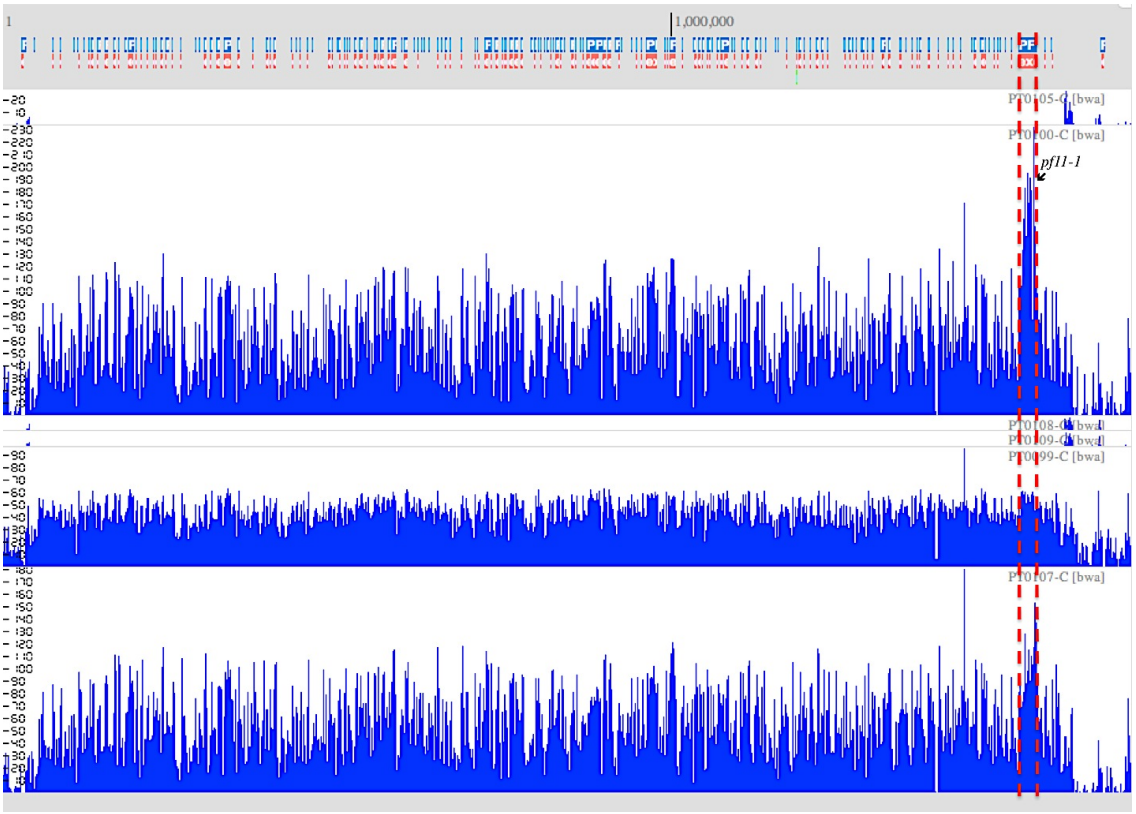

1,000,000
pf11-1
PT0099-C [bwa]
PT0107-C [bwa]

Figure 3.7: Visual representation of copy number variation with *FREEC* software (A-B) and Lookseq browser (C-D). (A-B) GC- content normalized copy number profiles for (A) chromosome 3 and (B) chromosome 10. Automatically predicted copy number gains are shown in red and predicted losses in blue. Because of low 'mappability' in subtelomeric regions, there are predicted losses close to the chromosome ends. (A) Isolates PT0042, PT0044 and PT0056 show loss between 0-1(y-axis, normalized copy number) and amplification between 1.5-3.0 for PT0042. This region contains *clag3.1 and clag3.2* genes. (B) Amplification is detected between 1-3.0 in PT0099 and PT0100 containing *pf11-1* gene. In isolates PT0105, PT0108 and PT0109 very low or no read counts were observed. (C-D) Visual representation in Lookseq browser for deletion and amplification in a region containing *clag3.1* and *clag3.2* (C), and amplification of a region containing *pf11-1* gene (D). Very low or no read counts were observed in some isolates such as PT0105, PT0108 and PT0109 (D).

3.9 Discussion

Several studies (mostly in the human genome) have demonstrated that copy number variants contribute significantly to genetic variation in natural populations of both humans and pathogens, playing a major role in genome evolution, adaptive changes and pathogenesis of disease. In the human genomes, CNV are more prevalent than point mutations or SNPs with an estimated rate of 10,000 to 1,000,000 per generation (compared to mice at 100 – 1,000 per generation (Korbel et al. 2008)), and have been associated with important human clinical phenotypes. In *P. falciparum*, an organism with extensive and rapid rates of genome variation (especially in Africa) and evolution, rates of CNV are also likely to be high. However, studies of CNV in *P. falciparum* have often been ignored (with larger efforts geared towards studying SNPs) but have only lately been well documented. Even with relatively few studies existing in *P. falciparum*, important phenotypic associations to CNV have been confirmed, notably in drug resistance. In addition, repeats in *csp* gene evolve (contract and expand) by internal duplications and deletions and could

influence/modulate host immune responses against this gene (Zeeshan et al. 2012)

To date, perhaps the most important and well-documented CNV is an amplification of the *P. falciparum* multidrug resistance 1 gene (*pfmdr1*) on chromosome 5 that influences parasite susceptibility to a variety of antimalarial drugs including MQ, LF, QN, and ART (Sidhu et al. 2006). Amplification of GTP cyclohydrolase 1 (on chromosome 12) of the folate biosynthesis is also associated with anti-folate drug resistance (Kidgell et al. 2006; Nair et al. 2008), and another amplification, *msp3.8* has recently been documented to be involved in decreasing sensitivity to HF, MF, and LF (Van Tyne et al. 2011). In this present study, none of these amplifications were detected in Malawi population. In fact, a deletion instead of amplification was contained in *msp3.8.* Undetected *pfmdr1* copy changes are not surprising, as this has been documented before, where 100% of isolates analyzed lacked this amplification in Malawi (Nkhoma et al. 2009). This observation may be explained by the low drug pressure (such as QN which has only been used in treating severe malaria) associated with amplifications in this gene in Malawi. ART and LF were recently introduced and thus if there were any selected copies of *pfmdr1* they would be limited/undetectable compared to the effective parasite population size. In contrast, for example, Thailand *P. falciparum* populations has been under intense pressure from almost all *pfmdr1* selective drugs, and thus show *pfmdr1* amplicon frequency at >30% (Nair et al. 2007; Price et al. 1999, 2006). Other well-characterized CNV include deletions on *kahrp* gene (associated with loss of cytoadherence) (Biggs et al. 1989), *clag9* gene (involved in gametocytogenesis and associated with loss of cytoadherence of *in vitro* parasite cultures) (Trenholme et al. 2000). Amplifications of *pfRh1* gene is prevalent in lab lines only, and has been correlated to overexpression that influenced higher growth rates of *P. falciparum* through facilitating

erythrocytes invasion (Triglia et al. 2005; Nair et al. 2011; Jiang et al. 2008a). This study did not detect any amplification or deletion on *pfRh1* and *clag9* genes respectively, but a deletion on reticulocyte binding protein 2-homologue b (*pfRh2b*) was observed. PfRH proteins are involved in a sialic acid-dependent merozoite invasion of erythrocytes, and a 0.58-kb deletion on *pfRh2b* was primarily found in Africa, first observed at high frequencies in parasites from Senegal (Jennings et al. 2007) and later in Tanzania, Malawi, Gambia and Malaysia (Ahouidi et al. 2010). This study also observed that the C-terminal region of *PfRh2b* harbouring the *PfRh2b* deletion does elicit immune responses that may exert directional selection on the polymorphism, but may not necessarily be under balancing selection.

A deletion and amplification was detected in *kahrp* gene. Deletion in this gene is consistent with previous reports (Biggs et al. 1989), while an amplification (16 gains in copy number, making it one of the highest in this population) is previously unreported. Whilst deletion on this gene is associated with loss of cytoadherence, amplification may lead to over-expression of this gene resulting in increased cytoadherence ability of parasites bearing the amplification.

Deletions were also detected on chromosome 3 containing two highly polymorphic and paralogous genes with 96.7% nucleotide similarity, *clag3.1* and *clag3.2* (Iriko et al., 2009). Deletions in these two genes were present at high frequencies in Malawi population and have also been previously reported elsewhere (Iriko et al., 2009; Jiang et al., 2008a; Sepúlveda et al., 2013), particularly in Ghanaian samples (Sepúlveda et al. 2013). Members of this gene family are thought to play an essential role in IE cytoadherence and invasion and have been studied as vaccine targets (Iriko et al., 2009; Kaneko et al., 2005). Polymorphisms in this multigene family, including copy number polymorphisms, are thought

to confer advantage to the parasite e.g., through redundancy of erythrocyte invasion or antigenic variation and could have undergone positive diversifying selection during *P. falciparum* evolution (Iriko et al., 2009).

Also detected was a deletion and amplification on a gametocyte-specific protein (encoded by *pf11-1* gene) on chromosome 10, which plays a role in gametocytogenesis in the rupture of the host erythrocyte and emergence of gametes (Scherf et al. 1992). This gene possessed several characteristics in this study: it contained the largest deletion and a large amplification; both deletion and amplification appeared at high frequencies in Malawi population (57 and 64% respectively) and contained a high number of copy gains (18 copies). Initial reports had suggested that deletion on this gene only occurred in laboratory lines possibly in response to stress of *in vitro* culture and its function in gametocytogenesis (Scherf et al. 1992). These CNV have now been reported in both lab and field isolates (Dharia et al. 2010; Jiang et al. 2008a; Mackinnon et al. 2009; Kidgell et al. 2006).

Other notable CNV observed at high frequencies in this population include amplifications in *lsa1* gene (average frequency by *FREEC* and *PG*, 65%), *rex1* gene (72%, *PG*), ABC transporter, *pfmrp2* (79%, *PG*) and deletion in a DBL containing protein, with unknown function (75%, *PG*). The *pfmrp2* encodes a member of the ABC transport family localized to the plasma membrane in all asexual stages of the parasite (Kavishe et al. 2009). Membrane transport proteins have been shown to mediate translocation of molecules and ions across biological membranes, including the uptake of nutrients into cells, the removal of unwanted metabolic waste products such as drugs. In reference to drug resistance, by pumping out drugs, intracellular drug accumulation is decreased. Two transport proteins multidrug resistance proteins (MDR/MRP) have largely been shown to be responsible for this type of

resistance. It might be reasonable to think that *pfmrp2* amplification and its high frequency should be further studied for possible roles in drug resistance or facilitating growth and replication of the parasite (Kavishe et al. 2009; Martin et al. 2005).

The functional categories of genes involved in CNV were examined for enrichment using gene ontology (GO) analysis in MADIBA with significance of the annotation evaluated using a hypergeometric *P*-value with a false discovery rate (*FDR*) correction (Law et al. 2008). Functional categories that are significantly enriched (*FDR* corrected *P*-values <0.05) are associated with genes involved in biological processes such as cytokinesis, transcription, metabolism, biosynthesis, transport of molecules and ions (including multidrug transport), gamete generation and entry into host cell.

Chapter 4

Final discussions, conclusions and future directions

4.1 Introduction

P. falciparum malaria is an important clinical disease widespread in sub-Saharan Africa and causing the highest morbidity and mortality in children under the age of five years and pregnant mothers (WHO. World Malaria Report 2012). Because of increased deployment of several malaria control interventions in malaria endemic regions there has been a considerable decrease in the number of malaria cases in some areas (Trape et al. 2011). But despite these efforts the overall number of deaths attributable to malaria remains unchanged (WHO. World Malaria Report 2012). There are four main malaria control tools that are used in malaria endemic regions including Malawi, namely: ITNs, IRS, IPTpd and ACTs. In Malawi, even as these interventions are increasingly used, there has been a lack of decline in childhood malaria between 2001-2010 (Roca-Feltrer 2012). To reduce malaria effectively, it is predicted that countries should reach coverage of 70% of the above-mentioned interventions (Rowe and Steketee 2007)

P. falciparum faces strong natural selection its two hosts. The human immune response and mosquito vector provide the strongest natural selective force. Drug treatment and changes in transmission intensity owing to specific intervention drive environmental pressure (Mackinnon and Marsh 2010). Data available suggest that this parasite escapes selective pressure through evolution thereby leaving distinct evolutionary paths that can be scanned by genetic and genomic tools to identify and understand key aspects of its biology

and transmission amidst these pressures. Key to this is to identify targets that are vulnerable to these interventions and whether these targets provide further knowledge of the parasite's survival strategies. Indeed, several approaches have applied population-biology based investigations to either validate or identify genetic loci responding to these pressures including drug targets (e.g., *pfcrt, pfdhps, pfdhfr, pfmdr1*) and vaccine candidates (e.g., *pfama1, msp3.8, eba175*) (Amambua-Ngwa et al. 2012a, 2012b; Mu et al. 2007; Miotto et al. 2013). In addition, most recent, use of genome-wide or whole genome sequencing strategies have taken a center stage owing to the increasingly lower costs of producing large amounts of sequence data (e.g., use of Illumina sequencing technology). This combined with the rising amount of computational tools to analyse genome data has allowed the production of rich/extensive genetic diversity data and population structure of worldwide populations of *P. falciparum* to identify loci under selection or associated with clinical phenotypes (Miotto et al. 2013; Cheeseman et al. 2012). Using similar approaches to identifying tools for monitoring and evaluating malaria interventions will be immensely informative.

In this present study, I have used massive parallel sequencing (MPS, Illumina short read) technology and identified >100,000 SNPs in 69 clinical *P. falciparum* isolates which were obtained directly from children in an on-going 3-year longitudinal study of a Malawian *P. falciparum* rural population in the Chikwawa district. Using the sequence data available for the 93 isolates, and two software packages for detecting copy number variation using read depth coverage, I provide a comprehensive documentation of deletions (losses) and amplifications (gains) in this population. Specifically, the objectives of this thesis were to use second-generation sequencing techniques:

1. To sequence uncultured *P. falciparum* paediatric isolates obtained from a 3-year longitudinal study of a Malawian population after a selective sweep with SP and continuous selective pressure from intensive use of LA, ITNs and IRS;
2. To identify genetic variants including SNPs, and large structural variants (e.g. CNV, INDELs) and provide a map of genomic variation;
3. To use the SNP variation, to identify regions under selection pressure specific to this *P. falciparum* population;
4. To compare Malawi genetic variation to other populations of disperse origins (Kenya, Burkina Faso, Mali, Thailand and Cambodia);

Important findings resulting from this thesis are briefly described below.

4.2 Whole genome sequencing of *P. falciparum* isolates and SNP identification

Five hundred uncultured clinical *P. falciparum* isolates have been sampled, depleted of human white blood cells and genomic DNA extracted. Parasite to human DNA has been quantified using qPCR in 329 samples. 93 samples were successfully sequenced using massive parallel sequencing. The process of depleting white blood cells (using CF11) from whole blood sample was largely successful in the 329 samples with >90% of samples having less than 30% human DNA contamination. Median human DNA concentration was 0.78% (range, 0-100%)

This analysis identified a rich SNP diversity in Malawi *P. falciparum*. 115,965 raw SNPs across the 69 isolates and giving a SNP density of approximately 1/198-bp. Further quality filtering produced 88,655 high quality SNPs (SNP density of 1/259-bp, compared to

1/266-bp in Manke et al. 2012). The principal component analysis using this SNP catalogue indeed showed an extensive genetic variation within the sampling region.

4.3 Local selection in Malawi *P. falciparum* genomes

Using this SNP variation two statistical metrics (haplotype and allele based) were calculated to identify genome regions under strong positive and balancing selection respectively.

4.3.1 Signatures of balancing selection

Using Tajima's *D* metric, genomic regions under positive diversifying selection, consistent with actions of balancing selection that maintains allelic variation in population, were identified. Large negative Tajima's *D* values were observed (>80% alleles), depicting singletons and excess of rare derived alleles at high frequencies, showing a genome mostly under purifying selection (eliminating deleterious mutations) or population expansion possibly after a bottleneck (Aris-Brosou and Excoffier 1996; Swanson 2003; Alexandre et al. 2011; Li 2011). Candidate signatures of balancing selection identified included merozoites invasion ligands contributing to 31.5% of all genes with significant Tajima's *D* (i.e., ≥1.0), showing that considerable number merozoite proteins possess rare alleles that are maintained at intermediate frequencies and are candidate targets of immune responses. Some of these genes are also replicated in other independent studies of different endemic populations, suggesting their roles as potential vaccine candidates, and possible based on multi-allelic antigen formulation (Tetteh et al. 2009; Ochola et al. 2010; Amambua-Ngwa et al. 2012b; Weedall and Conway 2010; Alexandre et al. 2011; Kaewthamasorn 2012; Xangsayarath et al. 2012).

4.3.2 Positive directional selection

Positive directional selection enables species to adapt to their living environments and usually leave footprints in the genomes of living organisms that can be detected. In particular, it allows an allele that is beneficial to either reproduction or survival to increase in frequency (toward fixation) as a result of the individual carrying the allele having an increased fitness. To infer directional selection in this population, haplotype based *iHS* was used to detect strong selection in areas surrounding SP drug resistance loci *pfdhps* and *pfgch1*. Weak selection signals at the *pfcrt* gene is likely due to the replacement of CQ-resistant parasites with CQ-sensitive parasites following CQ withdrawal from Malawi 20 years ago.

4.4 Inferring directional selection in Malawi *P. falciparum* population by comparing to geographically dispersed others

This analysis identified 294,187 SNPs across 6 populations – Malawi, Kenya, Burkina Faso, Mali, Cambodia and Thailand. An analysis of SNP density showed that the African populations were more polymorphic than the Southeast Asian.

4.4.1 Positioning Malawi parasites in the global population structure of *P. falciparum*

A principal component analysis of all populations showed distinct population differences between African and southeast Asian parasites with little genetic differentiation observed in Africa compared to Asia. Within Africa moderate population differences were detected between those parasites from the east (Malawi and Kenya) and west (Burkina Faso and Mali).

4.4.2 Inferring directional selection in Malawi *P. falciparum* population using *XP-EHH*

A search for selective sweeps by comparing Malawi to the six geographically dispersed population detected long-range haplotypes in loci surrounding *pfdhps, pfcrt* and *gch1*. This meant that strong drug selection by SP and CQ is largely responsible for reduction in haplotype diversity at drug targets and thus varying selection signals between different populations.

4.4.3 Inferring positive selection in Malawi *P. falciparum* population using F_{ST}

Using F_{ST} metric, genomic regions that have significantly differentiated between the six populations were identified. Two major observations were made: 1) that this differentiation increased with geographic distance from Malawi, Kenya, Burkina Faso, Mali, Cambodia and Thailand; 2) drug resistant alleles at known drug targets (*pfcrt, pfdhps, pfmdr1*) were highly differentiated and reflect parasite adaptation to historical drug pressure in these regions. There was a high level of differentiation in *clag3.1* and *clag3.2* genes (genes implicated in the process of parasite invasion) between Malawi and the two Asian populations, probably depicting a major underlying survival/adaptive strategy between these populations.

4.5 Copy number variation in Malawi *P. falciparum* genomes

This thesis has examined copy number variation in field isolates of *P. falciparum*. Studies of this kind have involved both *in vitro* adapted and *in vivo* parasites with results showing adaptive differences caused by CNV. Whilst *in vitro* environment is largely buffered in a supplemented medium and allows rapid propagation of asexual stage parasites, evasion

of drugs, host immune defence and transmission to the mosquito vector *via* sexual stages is critical in host environment highlighting the need to examine and characterize the extent of CNV *in vivo*. CNV size, gene content, population frequency is examined, as well as the role of genetic drift and selection in shaping CNV distributions. Polymorphic CNV (unique to individual isolate) are as common as monomorphic ones (occurring in multiple isolates). This analysis identifies both previously known CNV and unknown ones in both lab and field isolates. Polymorphisms associated with important *P. falciparum* biological functions such as cytoadherence, gametocytogenesis and merozoite invasion were observed. For example a deletion on reticulocyte binding protein 2 homologue b (a homolog of a reticulocyte binding protein gene well known to facilitate erythrocyte invasion) could modulate elicitation of immune responses. Amplification in *kahrp* gene may lead to over-expression of this gene leading to increased cytoadherence capability. Deletions were present on *clag3.1* and *clag3.2*, genes with essential roles in cytoadherence and invasion.

4.6 Implications and future directions

The genetic variation detected in Malawi and five other dispersed populations may enable us to monitor *P. falciparum* transmission dynamics as the epidemiology of malaria changes over time in response to interventions. For example, by using *XP-EHH* and F_{ST} selection metrics we have shown that loci containing selective sweeps and allelic divergence reflect the history of antimalarial drug use and selection at any given time, whereas during intense drug selection, wild-type alleles are increasingly replaced by mutant alleles. The ability to use this strategy to monitor local adaptation to drug pressure, monitoring transmission, and informing the type and timing of interventions is appealing. Future work will aim to use this knowledge in Malawi to monitor the impact of ACTs, ITNs and IRS on the

local parasite population of Chikwawa district over the three malaria seasons. We hypothesise a change in population structure e.g., bottlenecks, to occur in this population as successful interventions are used. Further potentially exciting work would involve looking at some of the genes detected under positive directional selection. For example, it would be interesting to look at expression of transcription factors in drug resistant versus sensitive parasites or in different clinical presentations of malaria such as severe malaria/cerebral malaria versus uncomplicated malaria.

The roles of CNV in influencing important clinical phenotypes have been highlighted before. Of importance will be to functionally validate these CNV and correlate them with phenotypes in this population. Since clinical data from the 500 samples are available to us, this will prove to be an exciting prospect. A possibility of complex phenotypes such as virulence influenced by a collection of highly mutable CNV changes as seen in some complex diseases in humans, could be explored (Sebat 2007). In addition, genetics aspects of CNV in *P. falciparum* field isolates maybe investigated, for example the rate and mechanisms of mutation and complexity of CNV evolution (Anderson et al. 2009). To provide a full map of CNV in this population, future work will also involve using other available CNV detection software to detect inversions and translocations.

There are few caveats to the analysis, which could turn into future opportunities. First, there are multiple infections (mean MOI ≈ 2.7, Assefa et al. 2014) in the samples, and in general this can confound genotyping, diversity and population analysis by generating false haplotype calls. The calculation of heterozygosity and detection of signatures of recent positive selection assume clonal samples, and *de novo* assembly of genomes in the presence of MOI can lead to potentially cryptic gene characterisations. To overcome the problem of

multiplicity, we generated a 'majority genotype' file where genotypes at SNP positions were called using ratios of coverage and heterozygous calls converted to the majority genotype on an 80:20 coverage ratio or greater. Over the entire preliminary dataset, 29% mixed calls were observed. Like others (e.g., Manske et al. 2012), we have restricted ourselves to polymorphisms where the frequency is low (83% SNPs with two mixed calls or less), and not biased our population genetic metrics. With longer read lengths it may be possible to reconstruct the individual parasite genomes from the data. Second, some regions of the genome were excluded, namely highly variable regions including *var* genes. The complete reconstruction *de novo* of a *P. falciparum* genome to high quality will be possible in the near future with longer read lengths. This *de novo* assembly approach will enable full genomic variation identification, without the need for a reference genome. Future work could consider the population genetics of these highly variable regions constructed using *de novo* assembly, especially as these regions are important in *P. falciparum* pathogenesis, for example, cytoadherence enables the parasite to evade splenic clearance and has been associated with severe malaria.

Future work will also include examining *P. falciparum* genetic diversity and population structure between the two geographically separated sampling sites i.e., Chikwawa and Zomba districts. Variables that that may affect parasite diversity such as age, seasonality (dry and wet seasons) and drug pressure will also be examined. Because of the nature of the study, it was possible to collect longitudinal samples from the same kid at different time points and thus allowing within patient longitudinal analysis of parasite genetic variation. Because CMX prophylaxis provides effective protection against malaria, it may modulate the development of malaria-specific immunity and potentially increase the

risk of malaria after it is stopped (rebound effect). Future work will investigate whether CMX prophylaxis between HIV exposed and non-exposed with influence infecting parasite types.

4.7 Appendices

Table A1: Genomic regions under balancing selection detected using Tajima's *D* by window approach. Bold refer to genes with *D*≥2.0.

Start	Finish	Tajima's D	Genes		
414001	**418000**	**2.69**	**PF3D7_0709300**		
1430001	**1434000**	**2.64**	**-**	**PF3D7_1036300**	
1432001	**1436000**	**2.55**	**-**	**PF3D7_1036300**	
846001	**850000**	**2.24**	**PF3D7_0220900**	**-**	**PF3D7_0221000**
330001	**334000**	**2.08**	**-**	**PF3D7_0806100**	**PF3D7_0806200**
880001	**884000**	**2.07**	**PF3D7_0720300**	**PF3D7_0720400**	
1286001	**1290000**	**2.03**	**PF3D7_0630700**	**-**	**PF3D7_0630800**
824001	**828000**	**2.02**	**PF3D7_0619500**	**-**	**PF3D7_0619600**
784001	**788000**	**2.02**	**-**	**PF3D7_1219600**	
1324001	**1326231**	**2.01**	**-**	**PF3D7_0532800**	
1144001	1148000	1.96	PF3D7_0425400	-	
408001	412000	1.81	PF3D7_1110200	-	PF3D7_1110300
1100001	1104000	1.79	PF3D7_0424400		
910001	914000	1.79	-	PF3D7_0420200	
514001	518000	1.77	PF3D7_0113600		
1322001	1326000	1.77	-	PF3D7_0532800	
222001	226000	1.75	PF3D7_1004800	-	PF3D7_1004900
1098001	1102000	1.68	-	PF3D7_0424400	
220001	224000	1.63	PF3D7_1004700	-	PF3D7_1004800
1288001	1292000	1.61	-	PF3D7_0630800	PF3D7_0630900
2168001	2172000	1.60	PF3D7_1253000	-	PF3D7_1253100
848001	852000	1.59	-	PF3D7_0221000	PF3D7_0221100
1100001	1104000	1.59	PF3D7_0825500	-	PF3D7_0825700
122001	126000	1.56	PF3D7_0302200		
1986001	1990000	1.52	PF3D7_1149400	-	
908001	912000	1.51	PF3D7_0420100	-	PF3D7_0420200
1036001	1040000	1.51	PF3D7_0422500	-	PF3D7_0422600
1390001	1394000	1.46	-	PF3D7_1035100	
1292001	1296000	1.45	-	PF3D7_1133400	
1294001	1298000	1.45	PF3D7_1133400	PF3D7_1133500	
840001	844000	1.43	PF3D7_0220800		
328001	332000	1.43	PF3D7_0806000	-	PF3D7_0806100
1018001	1022000	1.43	-	PF3D7_1126100	
178001	182000	1.39	PF3D7_0104100	-	
1988001	1992000	1.38	-	PF3D7_1149500	
822001	826000	1.37	PF3D7_0619500	-	PF3D7_0619600
796001	800000	1.36	-	PF3D7_0619100	PF3D7_0619200
844001	848000	1.36	PF3D7_0220800	-	PF3D7_0220900
180001	184000	1.35	PF3D7_0104100	-	PF3D7_0104200
522001	526000	1.34	PF3D7_0113700	-	
528001	532000	1.33	PF3D7_0113800		
678001	682000	1.32	-	PF3D7_0516300	PF3D7_0516400
100001	104000	1.31	-	PF3D7_1301800	
2000001	2004000	1.29	-	PF3D7_1149600	
422001	426000	1.25	PF3D7_0808300	-	PF3D7_0808400
1392001	1396000	1.23	PF3D7_1035100	-	
1198001	1202000	1.23	-	PF3D7_0930300	
1266001	1270000	1.21	PF3D7_0630300	-	
106001	110000	1.20	PF3D7_1301800		
154001	158000	1.19	-	PF3D7_1302900	
104001	108000	1.16	PF3D7_1301800		
1004001	1008000	1.14	PF3D7_0422200		
424001	428000	1.12	PF3D7_0808400	-	PF3D7_0808500
512001	516000	1.12	PF3D7_0113600		
842001	846000	1.11	PF3D7_0220800	-	PF3D7_0220900
1142001	1146000	1.05	PF3D7_0425300	-	PF3D7_0425400
282001	286000	1.03	PF3D7_0905600	PF3D7_0905700	
534001	538000	1.02	PF3D7_0113800	-	PF3D7_0113900
102001	106000	1.02	-	PF3D7_1301800	
1388001	1392000	1.01	PF3D7_1035000	-	PF3D7_1035100

4.8 References

Abyzov A, Urban AE, Snyder M, Gerstein M. 2011. CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. *Genome Res* **21**: 974–84.

Agnandji ST, Lell B, Soulanoudjingar SS, Fernandes JF, Abossolo BP, Conzelmann C, Methogo BGNO, Doucka Y, Flamen A, Mordmüller B, et al. 2011. First results of phase 3 trial of RTS,S/AS01 malaria vaccine in African children. *N Engl J Med* **365**: 1863–75.

Ahouidi AD, Bei AK, Neafsey DE, Sarr O, Volkman S, Milner D, Cox-Singh J, Ferreira MU, Ndir O, Premji Z, et al. 2010. Population genetic analysis of large sequence polymorphisms in *Plasmodium falciparum* blood-stage antigens. *Infect Genet Evol* **10**: 200–6.

Alexandre JS, Kaewthamasorn M, Yahata K, Nakazawa S, Kaneko O. 2011. Positive selection on the *Plasmodium falciparum clag2* gene encoding a component of the erythrocyte-binding rhoptry protein complex. *Trop Med Health* **39**: 77–82.

Alkan C, Coe BP, Eichler EE. 2011. Genome structural variation discovery and genotyping. *Nat Rev Genet* **12**: 363–76.

Alkan C, Kidd JM, Marques-Bonet T, Aksay G, Antonacci F, Hormozdiari F, Kitzman JO, Baker C, Malig M, Mutlu O, et al. 2009. Personalized copy number and segmental duplication maps using next-generation sequencing. *Nat Genet* **41**: 1061–7.

Amambua-Ngwa A, Park DJ, Volkman SK, Barnes KG, Bei AK, Lukens AK, Sene P, Van Tyne D, Ndiaye D, Wirth DF, et al. 2012a. SNP Genotyping Identifies New Signatures of Selection in a Deep Sample of West African *Plasmodium falciparum* Malaria Parasites. *Mol Biol Evol* **29**: 3249–53.

Amambua-Ngwa A, Tetteh KK a., Manske M, Gomez-Escobar N, Stewart LB, Deerhake ME, Cheeseman IH, Newbold CI, Holder A a., Knuepfer E, et al. 2012b. Population Genomic Scan for Candidate Signatures of Balancing Selection to Guide Antigen Characterization in Malaria Parasites. *PLoS Genet* **8**: e1002992.

Anders RF. 1986. Multiple cross-reactivities amongst antigens of *Plasmodium falciparum* impair the development of protective immunity against malaria. *Parasite Immunol* **8**: 529–39.

Anderson T, Nair S, Qin H. 2005. Are Transporter Genes Other than the Chloroquine Resistance Locus (*pfcrt*) and Multidrug Resistance Gene (*pfmdr*) Associated with Antimalarial Drug Resistance. *Antimicrob Agents Chemother* **49**: 2180–2188.

Anderson T, Nkhoma S, Ecker A, Fidock D. 2011. How can we identify parasite genes that underlie antimalarial drug resistance? *Pharmacogenomics* **12**: 59–85.

Anderson T, Patel J, Ferdig M. 2009. Gene copy number and malaria biology. *Trends Parasitol* **25**: 336–343.

Anderson TJ, Haubold B, Williams JT, Estrada-Franco JG, Richardson L, Mollinedo R, Bockarie M, Mokili J, Mharakurwa S, French N, et al. 2000. Microsatellite markers reveal a spectrum of population structures in the malaria parasite *Plasmodium falciparum*. *Mol Biol Evol* **17**: 1467–82.

Anderson TJC. 2004. Mapping drug resistance genes in *Plasmodium falciparum* by genome-wide association. *Curr Drug Targets Infect Disord* **4**: 65–78.

Aponte JJ, Aide P, Renom M, Mandomando I, Bassat Q, Sacarlal J, Manaca MN, Lafuente S, Barbosa A, Leach A, et al. 2007. Safety of the RTS,S/AS02D candidate malaria vaccine in infants living in a highly endemic area of Mozambique: a double blind randomised controlled phase I/IIb trial. *Lancet* **370**: 1543–51.

Aris-Brosou S, Excoffier L. 1996. The Impact of Population Expansion and Mutation Rate Heterogeneity on DNA sequence polymorphism. *Mol Biol Evol* 494–504.

Assefa S, Preston M, Campino S, Ocholla H, Sutherland CJ, Clark TG. 2014. estMOI : Estimating multiplicity of infection using parasite deep sequencing data. *Bioinforma* 1–3. (*Epub ahead of print*).

Auburn S, Campino S, Clark TG, Djimde A a., Zongo I, Pinches R, Manske M, Mangano V, Alcock D, Anastasi E, et al. 2011. An Effective Method to Purify *Plasmodium falciparum* DNA Directly from Clinical Blood Samples for Whole Genome High-Throughput Sequencing ed. G. Dimopoulos. *PLoS One* **6**: e22213.

Bailey JA, Gu Z, Clark RA, Reinert K, Samonte R V, Schwartz S, Adams MD, Myers EW, Li PW, Eichler EE. 2002. Recent segmental duplications in the human genome. *Science* **297**: 1003–7.

Barnes KI, Durrheim DN, Little F, Jackson A, Mehta U, Allen E, Dlamini SS, Tsoka J, Bredenkamp B, Mthembu DJ, et al. 2005. Effect of artemether-lumefantrine policy and improved vector control on malaria burden in KwaZulu-Natal, South Africa. *PLoS Med* **2**: e330.

Baum J, Thomas AW, Conway DJ. 2003. Evidence for diversifying selection on erythrocyte-binding antigens of *Plasmodium falciparum* and *P. vivax*. *Genetics* **163**: 1327–36.

Bejon P, White MT, Olotu A, Bojang K, Lusingu JP a, Salim N, Otsyula NN, Agnandji ST, Asante KP, Owusu-Agyei S, et al. 2013. Efficacy of RTS,S malaria vaccines: individual-participant pooled analysis of phase 2 data. *Lancet Infect Dis* **13**: 319–27.

Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, Hall KP, Evers DJ, Barnes CL, Bignell HR, et al. 2008. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456**: 53–9.

Bhattacharjee S, van Ooij C, Balu B, Adams JH, Haldar K. 2008. Maurer's clefts of Plasmodium falciparum are secretory organelles that concentrate virulence protein reporters for delivery to the host erythrocyte. *Blood* **111**: 2418–26.

Biggs B a, Kemp DJ, Brown G V. 1989. Subtelomeric chromosome deletions in field isolates of *Plasmodium falciparum* and their relationship to loss of cytoadherence in vitro. *Proc Natl Acad Sci U S A* **86**: 2428–32.

Boeva V, Zinovyev A, Bleakley K, Vert J-P, Janoueix-Lerosey I, Delattre O, Barillot E. 2011. Control-free calling of copy number alterations in deep-sequencing data using GC-content normalization. *Bioinformatics* **27**: 268–9.

Bowman NM, Congdon S, Mvalo T, Patel JC, Escamilla V, Emch M, Martinson F, Hoffman I, Meshnick SR, Juliano JJ. 2013. Comparative population structure of *Plasmodium falciparum* circumsporozoite protein NANP repeat lengths in Lilongwe, Malawi. *Sci Rep* **3**: 1990.

Bowman S, Lawson D, Basham D, Brown D, Chillingworth T, Churcher CM, Craig A, Davies RM, Devlin K, Feltwell T, et al. 1999. The complete nucleotide sequence of chromosome 3 of *Plasmodium falciparum*. *Nature* **400**: 532–8.

Carlton JM, Adams JH, Silva JC, Bidwell SL, Lorenzi H, Caler E, Crabtree J, Angiuoli S V, Merino EF, Amedeo P, et al. 2008. Comparative genomics of the neglected human malaria parasite *Plasmodium vivax*. *Nature* **455**: 757–63.

Carret CK, Horrocks P, Konfortov B, Winzeler E, Qureshi M, Newbold C, Ivens A. 2005. Microarray-based comparative genomic analyses of the human malaria parasite *Plasmodium falciparum* using Affymetrix arrays. *Mol Biochem Parasitol* **144**: 177–86.

Carter NP. 2007. Methods and strategies for analyzing copy number variation using DNA microarrays. *Nat Genet* **39**: S16–21.

Carter R, Mendis K. 2002. Evolutionary and Historical Aspects of the Burden of Malaria. *Clin Microbiol Rev* **15**: 564–594.

Cattamanchi A, Kyabayinze D, Hubbard A, Rosenthal PJ, Dorsey G. 2003. Distinguishing recrudescence from reinfection in a longitudinal antimalarial drug efficacy study: comparison of results based on genotyping of *msp-1*, *msp-2*, and *glurp*. *Am J Trop Med Hyg* **68**: 133–9.

Chakrabarti K, Pearson M, Grate L, Sterne-Weiler T, Deans J, Donohue JP, Ares M. 2007. Structural RNAs of known and unknown function identified in malaria parasites by comparative genomics and RNA analysis. *RNA* **13**: 1923–39.

Chakravorty SJ, Hughes KR, Craig AG. 2008. Host response to cytoadherence in *Plasmodium falciparum*. *Biochem Soc Trans* **36**: 221–8.

Chang H-H, Park DJ, Galinsky KJ, Schaffner SF, Ndiaye D, Ndir O, Mboup S, Wiegand RC, Volkman SK, Sabeti PC, et al. 2012. Genomic Sequencing of *Plasmodium falciparum* Malaria Parasites from Senegal Reveals the Demographic History of the Population. *Mol Biol Evol* **29**: 3427–39.

Cheeseman IH, Gomez-Escobar N, Carret CK, Ivens A, Stewart LB, Tetteh KK a, Conway DJ. 2009. Gene copy number variation throughout the *Plasmodium falciparum* genome. *BMC Genomics* **10**: 353.

Cheeseman IH, Miller B a., Nair S, Nkhoma S, Tan a., Tan JC, Al Saai S, Phyo a. P, Moo CL, Lwin KM, et al. 2012. A Major Genome Region Underlying Artemisinin Resistance in Malaria. *Science (80-)* **336**: 79–82.

Chen N, Kyle D, Pasay C. 2003. pfcrt Allelic Types with Two Novel Amino Acid Mutations in Chloroquine-Resistant *Plasmodium falciparum* Isolates from the Philippines. *Antimicrob Agents Chemother* **47**: 3500–3505.

Combe A, Moreira C, Ackerman S, Thiberge S, Templeton TJ, Ménard R. 2009. TREP, a novel protein necessary for gliding motility of the malaria sporozoite. *Int J Parasitol* **39**: 489–96.

Conrad D, Hurles M. 2007. The population genetics of structural variation. *Nat Genet* **39**: 30–36.

Conrad DF, Andrews TD, Carter NP, Hurles ME, Pritchard JK. 2006. A high-resolution survey of deletion polymorphism in the human genome. *Nat Genet* **38**: 75–81.

Conway DJ. 2007. Molecular epidemiology of malaria. *Clin Microbiol Rev* **20**: 188–204.

Conway DJ, Machado RL, Singh B, Dessert P, Mikes ZS, Povoa MM, Oduola a M, Roper C. 2001. Extreme geographical fixation of variation in the *Plasmodium falciparum* gamete surface protein gene *Pfs48/45* compared with microsatellite loci. *Mol Biochem Parasitol* **115**: 145–56.

Conway DJ, McBride JS. 1991. Population genetics of *Plasmodium falciparum* within a malaria hyperendemic area. *Parasitology* **103 Pt 1**: 7–16.

Conway DJ, Roper C, Oduola a M, Arnot DE, Kremsner PG, Grobusch MP, Curtis CF, Greenwood BM. 1999. High recombination rate in natural populations of *Plasmodium falciparum*. *Proc Natl Acad Sci U S A* **96**: 4506–11.

Corcoran LM, Thompson JK, Walliker D, Kemp DJ. 1988. Homologous recombination within subtelomeric repeat sequences generates chromosome size polymorphisms in P. falciparum. *Cell* **53**: 807–813.

Cowman AE, Karcz S, Galatis D, Culvenor JG. 1991. A P-glycoprotein Homologue of *Plasmodium falciparum* Is Localized on the Digestive Vacuole. **113**: 1033–1042.

Cowman AF, Galatis D, Thompson JK. 1994. Selection for mefloquine resistance in *Plasmodium falciparum* is linked to amplification of the *pfmdrl* gene and cross-resistance. **91**: 1143–1147.

Craig a, Scherf a. 2001. Molecules on the surface of the *Plasmodium falciparum* infected erythrocyte and their role in malaria pathogenesis and immune evasion. *Mol Biochem Parasitol* **115**: 129–43.

Craig AG, Khairul MFM, Patil PR. 2012. Cytoadherence and severe malaria. *Malays J Med Sci* **19**: 5–18.

D'Alessandro U. 1997. Severity of malaria and level of *Plasmodium falciparum* transmission. *Lancet* **350**: 362.

Dondorp A, Nosten F, Das D, Phyo AP, Tarning J, Ph D, Lwin KM, Ariey F, Hanpithakpong W, Lee SJ, Ringwald P, Silamut K, et al. 2009. Artemisinin-Resistance Malaria in Asia. 455–467.

Dessens JT, Beetsma a L, Dimopoulos G, Wengelnik K, Crisanti a, Kafatos FC, Sinden RE. 1999. CTRP is essential for mosquito infection by malaria ookinetes. *EMBO J* **18**: 6221–7.

Dharia N V, Bright a T, Westenberger SJ, Barnes SW, Batalov S, Kuhen K, Borboa R, Federe GC, McClean CM, Vinetz JM, et al. 2010. Whole-genome sequencing and microarray analysis of ex vivo *Plasmodium vivax* reveal selective pressure on putative drug resistance genes. *Proc Natl Acad Sci U S A* **107**: 20045–50.

Dicko A, Sagara I, Djimdé A a, Touré SO, Traore M, Dama S, Diallo AI, Barry A, Dicko M, Coulibaly OM, et al. 2010. Molecular markers of resistance to sulphadoxine-pyrimethamine one year after implementation of intermittent preventive treatment of malaria in infants in Mali. *Malar J* **9**: 9.

Dondorp AM, Pongponratn E, White NJ. 2004. Reduced microcirculatory flow in severe falciparum malaria: pathophysiology and electron-microscopic pathology. *Acta Trop* **89**: 309–317.

Doolan DL, Dobaño C, Baird JK. 2009. Acquired immunity to malaria. *Clin Microbiol Rev* **22**: 13–36, Table of Contents.

Dopman EB, Hartl DL. 2007. A portrait of copy-number polymorphism in *Drosophila melanogaster*. *Proc Natl Acad Sci U S A* **104**: 19920–5.

Douradinha B, Augustijn KD, Moore SG, Ramesar J, Mota MM, Waters AP, Janse CJ, Thompson J. 2011. *Plasmodium* Cysteine Repeat Modular Proteins 3 and 4 are essential for malaria parasite transmission from the mosquito to the host. *Malar J* **10**: 71.

Dzikowski R, Frank M, Deitsch K. 2006a. Mutually exclusive expression of virulence genes by malaria parasites is regulated independently of antigen production. *PLoS Pathog* **2**: e22.

Dzikowski R, Templeton TJ, Deitsch K. 2006b. Variant antigen gene expression in malaria. *Cell Microbiol* **8**: 1371–81.

Eckstein-Ludwig U, Webb RJ, Van Goethem ID a, East JM, Lee a G, Kimura M, O'Neill PM, Bray PG, Ward S a, Krishna S. 2003. Artemisinins target the SERCA of *Plasmodium falciparum*. *Nature* **424**: 957–61.

Egan TJ, Marques HM, Town C, Africa S. 1999. The role of haem in the activity of chloroquine and related antimalarial drugs. **192**: 493–517.

Ejigiri I, Ragheb DRT, Pino P, Coppi A, Bennett BL, Soldati-Favre D, Sinnis P. 2012. Shedding of TRAP by a rhomboid protease from the malaria sporozoite surface is essential for gliding motility and sporozoite infectivity. *PLoS Pathog* **8**: e1002725.

Ewing VL, Lalloo DG, Phiri KS, Roca-Feltrer A, Mangham LJ, SanJoaquin M a. 2011. Seasonal and geographic differences in treatment-seeking and household cost of febrile illness among children in Malawi. *Malar J* **10**: 32.

Färnert A, Arez AP, Babiker HA, Beck HP, Benito A, Björkman A, Bruce MC, Conway DJ, Day KP, Henning L, et al. 2001. Genotyping of *Plasmodium falciparum* infections by PCR: a comparative multicentre study. *Trans R Soc Trop Med Hyg* **95**: 225–32.

Ferdig MT, Cooper R a, Mu J, Deng B, Joy D a, Su X, Wellems TE. 2004. Dissecting the loci of low-level quinine resistance in malaria parasites. *Mol Microbiol* **52**: 985–97.

Fernández-Becerra C, Pinazo MJ, González A, Alonso PL, del Portillo H a, Gascón J. 2009. Increased expression levels of the *pvcrt-o* and *pvmdr1* genes in a patient with severe *Plasmodium vivax* malaria. *Malar J* **8**: 55.

Ferreira MU, Ribeiro WL, Tonon AP, Kawamoto F, Rich SM. 2003. Sequence diversity and evolution of the malaria vaccine candidate merozoite surface protein-1 (MSP-1) of *Plasmodium falciparum*. *Gene* **304**: 65–75.

Feuk L, Carson AR, Scherer SW. 2006a. Structural variation in the human genome. *Nat Rev Genet* **7**: 85–97.

Feuk L, Marshall CR, Wintle RF, Scherer SW. 2006b. Structural variants: changing the landscape of chromosomes and design of disease studies. *Hum Mol Genet* **15 Spec No**: R57–66.

Fidock D a, Nomura T, Talley a K, Cooper R a, Dzekunov SM, Ferdig MT, Ursos LM, Sidhu a B, Naudé B, Deitsch KW, et al. 2000. Mutations in the *Plasmodium falciparum* digestive vacuole transmembrane protein PfCRT and evidence for their role in chloroquine resistance. *Mol Cell* **6**: 861–71.

Foote SJ, Thompson JK, Cowman AF, Kemp DJ. 1989. Amplification of the multidrug resistance gene in some chloroquine-resistant isolates of *Plasmodium falciparum*. *Cell* **57**: 921–930.

Fu Y. 1997. Statistical tests of neutrality of mutations against population growth, hitchhiking and background selection. *Genetics* **147**: 915–925

Fu YX, Li WH. 1993. Statistical tests of neutrality of mutations. *Genetics* **133**: 693–709.

Gardner MJ, Feagin JE, Moore DJ, Rangachari K, Williamson DH, Wilson RJ. 1993. Sequence and organization of large subunit rRNA genes from the extrachromosomal 35 kb circular DNA of the malaria parasite *Plasmodium falciparum*. *Nucleic Acids Res* **21**: 1067–71.

Gardner MJ, Hall N, Fung E, White O, Berriman M, Hyman RW, Carlton JM, Pain A, Nelson KE, Bowman S, et al. 2002. Genome sequence of the human malaria parasite *Plasmodium falciparum*. *Nature* **419**: 498–511.

Genton B, D'Acremont V, Rare L, Baea K, Reeder JC, Alpers MP, Müller I. 2008. *Plasmodium vivax* and mixed infections are associated with severe malaria in children: a prospective cohort study from Papua New Guinea. *PLoS Med* **5**: e127.

Gething PW, Patil AP, Smith DL, Guerra C a, Elyazar IRF, Johnston GL, Tatem AJ, Hay SI. 2011. A new world malaria map: *Plasmodium falciparum* endemicity in 2010. *Malar J* **10**: 378.

Ghosh AK, Devenport M, Jethwaney D, Kalume DE, Pandey A, Anderson VE, Sultan A a, Kumar N, Jacobs-Lorena M. 2009. Malaria parasite invasion of the mosquito salivary gland requires interaction between the Plasmodium TRAP and the Anopheles saglin proteins. *PLoS Pathog* **5**: e1000265.

Glenister FK, Fernandez KM, Kats LM, Hanssen E, Mohandas N, Coppel RL, Cooke BM. 2009. Functional alteration of red blood cells by a megadalton protein of *Plasmodium falciparum*. *Blood* **113**: 919–28.

Gonzales JM, Patel JJ, Ponmee N, Jiang L, Tan A, Maher SP, Wuchty S, Rathod PK, Ferdig MT. 2008. Regulatory hotspots in the malaria parasite genome dictate transcriptional variation. *PLoS Biol* **6**: e238.

Gu Z, Steinmetz LM, Gu X, Scharfe C, Davis RW, Li W-H. 2003. Role of duplicate genes in genetic robustness against null mutations. *Nature* **421**: 63–6.

Hall N, Karras M, Raine JD, Carlton JM, Kooij TW a, Berriman M, Florens L, Janssen CS, Pain A, Christophides GK, et al. 2005. A comprehensive survey of the *Plasmodium* life cycle by genomic, transcriptomic, and proteomic analyses. *Science* **307**: 82–6.

Harismendy O, Ng PC, Strausberg RL, Wang X, Stockwell TB, Beeson KY, Schork NJ, Murray SS, Topol EJ, Levy S, et al. 2009. Evaluation of next generation sequencing platforms for population targeted sequencing studies. *Genome Biol* **10**: R32.

Hartl DL. 2004. The origin of malaria: mixed messages from genetic diversity. *Nat Rev Microbiol* **2**: 15–22.

Hastings P, Lupski J. 2009. Mechanisms of change in gene copy number. *Nat Rev Genet* **10**: 551–564.

Hawthorne PL, Trenholme KR, Skinner-Adams TS, Spielmann T, Fischer K, Dixon MWA, Ortega MR, Anderson KL, Kemp DJ, Gardiner DL. 2004. A novel *Plasmodium falciparum* ring stage protein, REX, is located in Maurer's clefts. *Mol Biochem Parasitol* **136**: 181–9.

Hayton K, Su X-Z. 2008. Drug resistance and genetic mapping in *Plasmodium falciparum*. *Curr Genet* **54**: 223–39.

Hempelmann E, Dluzewski AR. 1981. Effect of physostigmine on *Plasmodium falciparum* in culture. *Tropenmed Parasitol* **32**: 48–50.

Hertz-Fowler C, Peacock CS, Wood V, Aslett M, Kerhornou A, Mooney P, Tivey A, Berriman M, Hall N, Rutherford K, et al. 2004. GeneDB: a resource for prokaryotic and eukaryotic organisms. *Nucleic Acids Res* **32**: D339–43.

Hinterberg K, Mattei D, Wellems TE, Scherf a. 1994. Interchromosomal exchange of a large subtelomeric segment in a *Plasmodium falciparum* cross. *EMBO J* **13**: 4174–80.

Holsinger KE, Weir BS. 2009. Genetics in geographically structured populations: defining, estimating and interpreting F(ST). *Nat Rev Genet* **10**: 639–50.

Howard RJ, Handunnetti SM, Hasler T, Gilladoga A, de Aguiar JC, Pasloske BL, Morehead K, Albrecht GR, van Schravendijk MR. 1990. Surface Molecules on *Plasmodium falciparum*-Infected Erythrocytes Involved in Adherence. *Am J Trop Med Hyg* **43**: 15–29.

Hudson RR, Kreitman M, Aguadé M. 1987. A test of neutral molecular evolution based on nucleotide data. *Genetics* **116**: 153–9.

Hughes AL. 2004. The evolution of amino acid repeat arrays in *Plasmodium* and other organisms. *J Mol Evol* **59**: 528–35.

Hunt P, Afonso A, Creasey A, Culleton R, Sidhu ABS, Logan J, Valderramos SG, McNae I, Cheesman S, do Rosario V, et al. 2007. Gene encoding a deubiquitinating enzyme is mutated in artesunate- and chloroquine-resistant rodent malaria parasites. *Mol Microbiol* **65**: 27–40.

Iafrate AJ, Feuk L, Rivera MN, Listewnik ML, Donahoe PK, Qi Y, Scherer SW, Lee C. 2004. Detection of large-scale variation in the human genome. *Nat Genet* **36**: 949–51.

Imwong M, Dondorp AM, Nosten F, Yi P, Mungthin M, Hanchana S, Das D, Phyo AP, Lwin KM, Pukrittayakamee S, et al. 2010. Exploring the contribution of candidate genes to artemisinin resistance in *Plasmodium falciparum*. *Antimicrob Agents Chemother* **54**: 2886–92.

Iriko H. 2009. Diversity and evolution of the *rhoph1/clag* multigene family of *Plasmodium falciparum*. *Gene* **201**: 11–21.

Jambou R, Legrand E, Niang M, Khim N, Lim P, Volney B, Ekala MT, Bouchier C, Esterre P, Fandeur T, et al. 2005. Resistance of *Plasmodium falciparum* field isolates to in-vitro artemether and point mutations of the SERCA-type PfATPase6. *Lancet* **366**: 1960–3.

Jeffares D, Pain A, Berry A, Cox A. 2006. Genome variation and evolution of the malaria parasite *Plasmodium falciparum*. *Nat Genet* **39**: 120–125.

Jennings C V, Ahouidi AD, Zilversmit M, Bei AK, Rayner J, Sarr O, Ndir O, Wirth DF, Mboup S, Duraisingh MT. 2007. Molecular analysis of erythrocyte invasion in *Plasmodium falciparum* isolates from Senegal. *Infect Immun* **75**: 3531–8.

Jiang H, Li N, Gopalan V, Zilversmit MM, Varma S, Nagarajan V, Li J, Mu J, Hayton K, Henschen B, et al. 2011. High recombination rates and hotspots in a *Plasmodium falciparum* genetic cross. *Genome Biol* **12**: R33.

Jiang H, Patel JJ, Yi M, Mu J, Ding J, Stephens R, Cooper R a, Ferdig MT, Su X. 2008a. Genome-wide compensatory changes accompany drug- selected mutations in the *Plasmodium falciparum* crt gene. *PLoS One* **3**: e2484.

Jiang H, Yi M, Mu J, Zhang L, Ivens A, Klimczak LJ, Huyen Y, Stephens RM, Su X-Z. 2008b. Detection of genome-wide polymorphisms in the AT-rich *Plasmodium falciparum* genome using a high-density microarray. *BMC Genomics* **9**: 398.

Kaewthamasorn M. 2012. Stable allele frequency distribution of the polymorphic region of SURFIN4.2 in *Plasmodium falciparum* isolates from Thailand. *Parasitol Int* **61**: 317–323.

Kaneko O, Yim Lim BYS, Iriko H, Ling IT, Otsuki H, Grainger M, Tsuboi T, Adams JH, Mattei D, Holder A a, et al. 2005. Apical expression of three RhopH1/Clag proteins as components of the *Plasmodium falciparum* RhopH complex. *Mol Biochem Parasitol* **143**: 20–8.

Kavishe R a, van den Heuvel JM, van de Vegte-Bolmer M, Luty AJ, Russel FG, Koenderink JB. 2009. Localization of the ATP-binding cassette (ABC) transport proteins PfMRP1, PfMRP2, and PfMDR5 at the *Plasmodium falciparum* plasma membrane. *Malar J* **8**: 205.

Kidgell C, Volkman SK, Daily J, Borevitz JO, Plouffe D, Zhou Y, Johnson JR, Le Roch K, Sarr O, Ndir O, et al. 2006. A systematic map of genetic variation in *Plasmodium falciparum*. *PLoS Pathog* **2**: e57.

Kilama W, Ntoumi F. 2009. Malaria: a research agenda for the eradication era. *Lancet* **374**: 1480–2.

Knowles G, Sanderson A, Walliker D. 1981. *Plasmodium yoelii*: genetic analysis of crosses between two rodent malaria subspecies. *Exp Parasitol* **52**: 243–7.

Korbel J, Kim P, Chen X. 2008. The current excitement about copy-number variation: how it relates to gene duplication and protein families. *Curr Opin Struct Biol* **18**: 366–374.

Kozarewa I, Ning Z, Quail MA, Sanders MJ, Turner DJ. 2009. Amplification-free Illumina sequencing-library preparation facilitates improved mapping and assembly of GC-biased genomes. *Nat Methods* **6**: 291–295.

Krogstad D, Gluzman I, Kyle D, Oduola A, Martin S, Milhous W, Schlesinger P. 1987. Efflux of chloroquine from *Plasmodium falciparum*: mechanism of chloroquine resistance. *Science (80-)* **238**: 1283–1285.

Kun JF, Schmidt-Ott RJ, Lehman LG, Lell B, Luckner D, Greve B, Matousek P, Kremsner PG. 1998. Merozoite surface antigen 1 and 2 genotypes and rosetting of *Plasmodium falciparum* in severe and mild malaria in Lambaréné, Gabon. *Trans R Soc Trop Med Hyg* **92**: 110–4.

Laufer M, Thesing P, Eddington N, Masonga R, Dzinjalamala F, Takala S, Taylor T, Plowe C. 2006. Return of Chloroquine Antimalarial Efficacy in Malawi. *N Engl J Med* **355**: 1959–1966.

Law PJ, Claudel-Renard C, Joubert F, Louw AI, Berger DK. 2008. MADIBA: a web server toolkit for biological interpretation of *Plasmodium* and plant gene clusters. *BMC Genomics* **9**: 105.

Levine M, Tjian R. 2003. Transcription regulation and animal diversity. *Nature* **424**: 147–51.

Li H. 2011. A new test for detecting recent positive selection that is free from the confounding impacts of demography. *Mol Biol Evol* **28**: 365–75.

Liu W, Li Y, Learn G, Rudicell R. 2010. Origin of the human malaria parasite *Plasmodium falciparum* in gorillas. *Nature* **467**: 420–425.

Mackinnon MJ, Li J, Mok S, Kortok MM, Marsh K, Preiser PR, Bozdech Z. 2009. Comparative transcriptional and genomic analysis of *Plasmodium falciparum* field isolates. *PLoS Pathog* **5**: e1000644.

Mackinnon MJ, Marsh K. 2010. The selection landscape of malaria parasites. *Science* **328**: 866–71.

MacPherson GG. 1985. Human Cerebral Malaria. A Quantitative Ultrastructural Analysis of Parasitized Erythrocyte Sequestration. *Am J Pathol* **119**: 385–401.

Malawi National Malaria Indicator Survey. 2010. *Malawi National Malaria Indicator Survey 2010*.

Manske M, Miotto O, Campino S, Auburn S, Almagro-Garcia J, Maslen G, O'Brien J, Djimde A, Doumbo O, Zongo I, et al. 2012. Analysis of *Plasmodium falciparum* diversity in natural infections by deep sequencing. *Nature* **487**: 375–379.

Mardis ER. 2011. A decade's perspective on DNA sequencing technology. *Nature* **470**: 198–203.

Martin RE, Henry RI, Abbey JL, Clements JD, Kirk K. 2005. The “permeome” of the malaria parasite: an overview of the membrane transport proteins of *Plasmodium falciparum*. *Genome Biol* **6**: R26.

McBride JS, Walliker D, Morgan G. 1982. Antigenic diversity in the human malaria parasite *Plasmodium falciparum*. *Science* **217**: 254–7.

McDonald JH, Kreitman M. 1991. Adaptive protein evolution at the Adh locus in Drosophila. *Nature* **351**: 652–4.

Medvedev P, Stanciu M, Brudno M. 2009. Computational methods for discovering structural variation with next-generation sequencing. *Nat Methods* **6**: 13–20.

Ménard R. 2005. Knockout malaria vaccine ? *Nature* **433**: 6–7.

Mendis K, Sina BJ, Marchesini P, Carter R. 2001. The neglected burden of *Plasmodium vivax* malaria. *Am J Trop Med Hyg* **64**: 97–106.

Metzker ML. 2010. Sequencing technologies - the next generation. *Nat Rev Genet* **11**: 31–46.

Mikolajczak SA, Sacci JB, De La Vega P, Camargo N, VanBuskirk K, Krzych U, Cao J, Jacobs-Lorena M, Cowman AF, Kappe SHI. 2011. Disruption of the *Plasmodium falciparum* liver-stage antigen-1 locus causes a differentiation defect in late liver-stage parasites. *Cell Microbiol* **13**: 1250–60.

Miller J, Koren S, Sutton G. 2010. Assembly Algorithms for Next-Generation Sequencing Data. *Genomics* **95**: 315–327.

Miotto O, Almagro-Garcia J, Manske M, MacInnis B, Campino S, Rockett K a, Amaratunga C, Lim P, Suon S, Sreng S, et al. 2013. Multiple populations of artemisinin-resistant *Plasmodium falciparum* in Cambodia. *Nat Genet* **45**: 648–655.

Montgomery J, Milner D a, Tse MT, Njobvu A, Kayira K, Dzamalala CP, Taylor TE, Rogerson SJ, Craig AG, Molyneux ME. 2006. Genetic analysis of circulating and sequestered populations of *Plasmodium falciparum* in fatal pediatric malaria. *J Infect Dis* **194**: 115–22.

Mourier T, Carret C, Kyes S, Christodoulou Z, Gardner PP, Jeffares DC, Pinches R, Barrell B, Berriman M, Griffiths-Jones S, et al. 2008. Genome-wide discovery and verification of novel structured RNAs in *Plasmodium falciparum*. *Genome Res* **18**: 281–92.

Mu J, Awadalla P, Duan J, McGee KM, Joy D a, McVean G a T, Su X. 2005. Recombination hotspots and population structure in *Plasmodium falciparum*. *PLoS Biol* **3**: e335.

Mu J, Awadalla P, Duan J, McGee KM, Keebler J, Seydel K, McVean G a T, Su X. 2007. Genome-wide variation and identification of vaccine targets in the *Plasmodium falciparum* genome. *Nat Genet* **39**: 126–30.

Mu J, Ferdig MT, Feng X, Joy D a., Duan J, Furuya T, Subramanian G, Aravind L, Cooper R a., Wootton JC, et al. 2003. Multiple transporters associated with malaria parasite responses to chloroquine and quinine. *Mol Microbiol* **49**: 977–989.

Mu J, Myers R a, Jiang H, Liu S, Ricklefs S, Waisberg M, Chotivanich K, Wilairatana P, Krudsood S, White NJ, et al. 2010. *Plasmodium falciparum* genome-wide scans for positive selection, recombination hot spots and resistance to antimalarial drugs. *Nat Genet* **42**: 268–71.

Mwai L, Ochong E, Abdirahman A, Kiara SM, Ward S, Kokwaro G, Sasi P, Marsh K, Borrmann S, Mackinnon M, et al. 2009. Chloroquine resistance before and after its withdrawal in Kenya. *Malar J* **8**: 106.

Mzilahowa T, Hastings IM, Molyneux ME, McCall PJ. 2012. Entomological indices of malaria transmission in Chikhwawa district, Southern Malawi. *Malar J* **11**: 380.

Nair S, Miller B, Barends M, Jaidee A, Patel J, Mayxay M, Newton P, Nosten F, Ferdig MT, Anderson TJC. 2008. Adaptive copy number evolution in malaria parasites. *PLoS Genet* **4**: e1000243.

Nair S, Nash D, Sudimack D, Jaidee A, Barends M, Uhlemann A-C, Krishna S, Nosten F, Anderson TJC. 2007. Recurrent gene amplification and soft selective sweeps during evolution of multidrug resistance in malaria parasites. *Mol Biol Evol* **24**: 562–73.

Nair S, Nkhoma S, Nosten F, Mayxay M, French N, Whitworth J, Anderson T. 2011. Genetic changes during laboratory propagation: copy number at the reticulocyte binding protein 1 locus of *Plasmodium falciparum*. *Mol Biochem Parasitol* **172**: 145–148.

Nair S, Williams JT, Brockman A, Paiphun L, Mayxay M, Newton PN, Guthmann J-P, Smithuis FM, Hien TT, White NJ, et al. 2003. A selective sweep driven by pyrimethamine treatment in southeast asian malaria parasites. *Mol Biol Evol* **20**: 1526–36.

Nash D, Nair S, Mayxay M, Newton PN, Guthmann J-P, Nosten F, Anderson TJC. 2005. Selection strength and hitchhiking around two anti-malarial resistance genes. *Proc Biol Sci* **272**: 1153–61.

Nei M, Gojobori T. 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol Biol Evol* **3**: 418–26.

Newbold CI, Pinches R, Roberts DJ, Marsh K. 1992. *Plasmodium falciparum*: the human agglutinating antibody response to the infected red cell surface is predominantly variant specific. *Exp Parasitol* **75**: 281–92.

Nkhoma S, Molyneux M, Ward S. 2007. Molecular surveillance for drug-resistant *Plasmodium falciparum* malaria in Malawi. *Acta Trop* **102**: 138–42.

Nkhoma S, Nair S, Mukaka M, Molyneux ME, Ward S a, Anderson TJC. 2009. Parasites bearing a single copy of the multi-drug resistance gene (*pfmdr-1*) with wild-type SNPs

predominate amongst *Plasmodium falciparum* isolates from Malawi. *Acta Trop* **111**: 78–81.

Noedl H, Se Y, Schaecher K, Smith BL, Socheat D, Fukuda MM. 2008. Evidence of artemisinin-resistant malaria in western Cambodia. *N Engl J Med* **359**: 2619–20.

Noedl H, Socheat D, Satimai W. 2009. Artemisinin-resistant malaria in Asia. *N Engl J Med* **361**: 540–1.

Ntoumi F, Contamin H, Rogier C, Bonnefoy S, Trape JF, Mercereau-Puijalon O. 1995. Age-dependent carriage of multiple *Plasmodium falciparum* merozoite surface antigen-2 alleles in asymptomatic malaria infections. *Am J Trop Med Hyg* **52**: 81–8.

Nurleila S, Syafruddin D, Elyazar IRF, Baird JK. 2012. Serious and fatal illness associated with falciparum and vivax malaria among patients admitted to hospital at West Sumba in eastern Indonesia. *Am J Trop Med Hyg* **87**: 41–9.

Ochola LI, Tetteh KKA, Stewart LB, Riitho V, Marsh K, Conway DJ. 2010. Allele Frequency – Based and Polymorphism-Versus- Divergence Indices of Balancing Selection in a New Filtered Set of Polymorphic Genes in *Plasmodium falciparum*. *Mol Biol Evol* **27**: 2344–2351.

Organization, World Health . 2012. *World Malaria Report*.

Osier FH a, Fegan G, Polley SD, Murungi L, Verra F, Tetteh KK a, Lowe B, Mwangi T, Bull PC, Thomas AW, et al. 2008. Breadth and magnitude of antibody responses to multiple *Plasmodium falciparum* merozoite antigens are associated with protection from clinical malaria. *Infect Immun* **76**: 2240–8.

Oyola SO, Otto TD, Gu Y, Maslen G, Manske M, Campino S, Turner DJ, MacInnis B, Kwiatkowski DP, Swerdlow HP, et al. 2012. Optimizing Illumina Next-Generation Sequencing library preparation for extremely AT-biased genomes. *BMC Genomics* **13**: 1.

Park DJ, Lukens AK, Neafsey DE, Schaffner SF, Chang H-H, Valim C, Ribacke U, Van Tyne D, Galinsky K, Galligan M, et al. 2012. Sequence-based association and selection scans identify drug resistance loci in the *Plasmodium falciparum* malaria parasite. *Proc Natl Acad Sci U S A* **109**: 13052–7.

Patterson N, Price AL, Reich D. 2006. Population structure and eigenanalysis. *PLoS Genet* **2**: e190.

Pearce RJ, Pota H, Evehe M-SB, Bâ E-H, Mombo-Ngoma G, Malisa AL, Ord R, Inojosa W, Matondo A, Diallo D a, et al. 2009. Multiple origins and regional dispersal of resistant dhps in African *Plasmodium falciparum* malaria. *PLoS Med* **6**: e1000055.

Peterson DS, Walliker D, Wellems TE. 1988. Evidence that a point mutation in dihydrofolate reductase-thymidylate synthase confers resistance to pyrimethamine in falciparum malaria. *Proc Natl Acad Sci U S A* **85**: 9114–8.

Plassmeyer ML, Reiter K, Shimp RL, Kotova S, Smith PD, Hurt DE, House B, Zou X, Zhang Y, Hickman M, et al. 2009. Structure of the *Plasmodium falciparum* circumsporozoite protein, a leading malaria vaccine candidate. *J Biol Chem* **284**: 26951–63.

Plowe C V, Cortese JF, Djimde a, Nwanyanwu OC, Watkins WM, Winstanley P a, Estrada-Franco JG, Mollinedo RE, Avila JC, Cespedes JL, et al. 1997. Mutations in *Plasmodium falciparum* dihydrofolate reductase and dihydropteroate synthase and epidemiologic patterns of pyrimethamine-sulfadoxine use and resistance. *J Infect Dis* **176**: 1590–6.

Polley SD, Chokejindachai W, Conway DJ. 2003. Allele frequency-based analyses robustly map sequence sites under balancing selection in a malaria vaccine candidate antigen. *Genetics* **165**: 555–61.

Polley SD, Conway DJ. 2001. Strong diversifying selection on domains of the *Plasmodium falciparum* apical membrane antigen 1 gene. *Genetics* **158**: 1505–12.

Polley SD, Tetteh KK a, Lloyd JM, Akpogheneta OJ, Greenwood BM, Bojang K a, Conway DJ. 2007. *Plasmodium falciparum* merozoite surface protein 3 is a target of allele-specific immunity and alleles are maintained by natural selection. *J Infect Dis* **195**: 279–87.

Pologe LG, de Bruin D, Ravetch J V. 1990. A and T homopolymeric stretches mediate a DNA inversion in *Plasmodium falciparum* which results in loss of gene expression. *Mol Cell Biol* **10**: 3243–6.

Prajapati SK, Joshi H, Dev V, Dua VK. 2011. Molecular epidemiology of *Plasmodium vivax* anti-folate resistance in India. *Malar J* **10**: 102.

Preston MD, Assefa SA, Ocholla H, Sutherland CJ, Nzila A, Michon P, Hien TT, Bousema T, Christopher J. 2013. PlasmoView: A web-based resource to visualise global *Plasmodium falciparum* genomic variation. *J Infect Dis* 1–21.

Price RN, Cassar C, Brockman a, Duraisingh M, van Vugt M, White NJ, Nosten F, Krishna S. 1999. The *pfmdr1* gene is associated with a multidrug-resistant phenotype in *Plasmodium falciparum* from the western border of Thailand. *Antimicrob Agents Chemother* **43**: 2943–9.

Price RN, Douglas NM, Anstey NM. 2009. New developments in *Plasmodium vivax* malaria: severe disease and the rise of chloroquine resistance. *Curr Opin Infect Dis* **22**: 430–5.

Price RN, Uhlemann A-C, van Vugt M, Brockman A, Hutagalung R, Nair S, Nash D, Singhasivanon P, Anderson TJC, Krishna S, et al. 2006. Molecular and pharmacological determinants of the therapeutic response to artemether-lumefantrine in multidrug-resistant *Plasmodium falciparum* malaria. *Clin Infect Dis* **42**: 1570–7.

Pumpaibool T, Arnathau C, Durand P, Kanchanakhan N, Siripoon N, Suegorn A, Sitthi-Amorn C, Renaud F, Harnyuttanakorn P. 2009. Genetic diversity and population structure of *Plasmodium falciparum* in Thailand, a low transmission country. *Malar J* **8**: 155.

Quail M a, Smith M, Coupland P, Otto TD, Harris SR, Connor TR, Bertoni A, Swerdlow HP, Gu Y. 2012. A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC Genomics* **13**: 341.

Rausch T, Zichner T, Schlattl A, Stütz AM, Benes V, Korbel JO. 2012. DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics* **28**: i333–i339.

Redon R, Ishikawa S, Fitch KR, Feuk L, Perry GH, Andrews TD, Fiegler H, Shapero MH, Carson AR, Chen W, et al. 2006. Global variation in copy number in the human genome. *Nature* **444**: 444–54.

Ribacke U, Mok BW, Wirta V, Normark J, Lundeberg J, Kironde F, Egwang TG, Nilsson P, Wahlgren M. 2007. Genome wide gene amplifications and deletions in *Plasmodium falciparum*. *Mol Biochem Parasitol* **155**: 33–44.

Robert F, Ntoumi F, Angel G, Candito D, Rogier C, Fandeur T, Sarthou JL, Mercereau-Puijalon O. 1996. Extensive genetic diversity of *Plasmodium falciparum* isolates collected from patients with severe malaria in Dakar, Senegal. *Trans R Soc Trop Med Hyg* **90**: 704–11.

Roca-Feltrer A. 2012. Lack of Decline in Childhood Malaria, Malawi, 2001-2010. *Emerg Infect Dis* **18**: 272–278.

Rogers WO, Malik a, Mellouk S, Nakamura K, Rogers MD, Szarfman a, Gordon DM, Nussler a K, Aikawa M, Hoffman SL. 1992. Characterization of *Plasmodium falciparum* sporozoite surface protein 2. *Proc Natl Acad Sci U S A* **89**: 9176–80.

Roper C, Pearce R, Nair S, Sharp B, Nosten F, Anderson T. 2004. Intercontinental spread of pyrimethamine-resistant malaria. *Science* **305**: 1124.

Rowe AK, Steketee RW. 2007. Predictions of the impact of malaria control efforts on all-cause child mortality in sub-Saharan Africa. *Am J Trop Med Hyg* **77**: 48–55.

Sabeti PC, Reich DE, Higgins JM, Levine HZP, Richter DJ, Schaffner SF, Gabriel SB, Platko J V, Patterson NJ, Mcdonald GJ, et al. 2002. Detecting recent positive selection in the human genome from haplotype structure. *Nature* **419**: 832–7.

Sabeti PC, Schaffner SF, Fry B, Lohmueller J, Varilly P, Shamovsky O, Palma a, Mikkelsen TS, Altshuler D, Lander ES. 2006. Positive natural selection in the human lineage. *Science* **312**: 1614–20.

Sabeti PC, Varilly P, Fry B, Lohmueller J, Hostetter E, Cotsapas C, Xie X, Byrne EH, McCarroll S a, Gaudet R, et al. 2007. Genome-wide detection and characterization of positive selection in human populations. *Nature* **449**: 913–8.

Sakura T, Yahata K, Kaneko O. 2013. The upstream sequence segment of the C-terminal cysteine-rich domain is required for microneme trafficking of *Plasmodium falciparum* erythrocyte binding antigen 175. *Parasitol Int* **62**: 157–64.

Sboner A, Mu XJ, Greenbaum D, Auerbach RK, Gerstein MB. 2011. The real cost of sequencing: higher than you think! *Genome Biol* **12**: 125.

Scherf a, Mattei D. 1992. Cloning and characterization of chromosome breakpoints of *Plasmodium falciparum*: breakage and new telomere formation occurs frequently and randomly in subtelomeric genes. *Nucleic Acids Res* **20**: 1491–6.

Scherf A, Carter R, Petersen C, Alano P, Nelson R, Aikawa M, Mattei D, Pereira L, Leech J. 1992. Gene inactivation of *pf 11-1* of *Plasmodium falciparum* by chromosome breakage and healing : identification of a gametocyte-specific protein with a potential role in gametogenesis. *EMBO J* **11**: 2293–2301.

Scherf A, Lopez-Rubio JJ, Riviere L. 2008. Antigenic variation in *Plasmodium falciparum*. *Annu Rev Microbiol* **62**: 445–70.

Scherf A, Mattei D, Sarthou JL. 1991. Multiple infections and unusual distribution of block 2 of the *msa1* gene of *Plasmodium falciparum* detected in west African clinical isolates by polymerase chain reaction analysis. *Mol Biochem Parasitol* **44**: 297–9.

Sebat J. 2007. Major changes in our DNA lead to major changes in our thinking. *Nat Genet* **39**: S3–5.

Sepúlveda N, Campino SG, Assefa S a, Sutherland CJ, Pain A, Clark TG. 2013. A Poisson hierarchical modelling approach to detecting copy number variation in sequence coverage data. *BMC Genomics* **14**: 128.

Setthaudom C, Tan-ariya P, Sitthichot N, Khositnithikul R, Suwandittakul N, Leelayoova S, Mungthin M. 2011. Role of *Plasmodium falciparum* chloroquine resistance transporter and multidrug resistance 1 genes on in vitro chloroquine resistance in isolates of *P. falciparum* from Thailand. *Am J Trop Med Hyg* **85**: 606–11.

Sidhu ABS, Uhlemann A-C, Valderramos SG, Valderramos J-C, Krishna S, Fidock D a. 2006. Decreasing pfmdr1 copy number in *Plasmodium falciparum* malaria heightens susceptibility to mefloquine, lumefantrine, halofantrine, quinine, and artemisinin. *J Infect Dis* **194**: 528–35.

Sidhu ABS, Verdier-Pinard D, Fidock D a. 2002. Chloroquine resistance in *Plasmodium falciparum* malaria parasites conferred by *pfcrt* mutations. *Science* **298**: 210–3.

Sisowath C, Strömberg J, Mårtensson A, Msellem M, Obondo C, Björkman A, Gil JP. 2005. In vivo selection of *Plasmodium falciparum pfmdr1 86N* coding alleles by artemether-lumefantrine (Coartem). *J Infect Dis* **191**: 1014–7.

Somé AF, Séré YY, Dokomajilar C, Zongo I, Rouamba N, Greenhouse B, Ouédraogo J-B, Rosenthal PJ. 2010. Selection of known *Plasmodium falciparum* resistance-mediating polymorphisms by artemether-lumefantrine and amodiaquine-sulfadoxine-pyrimethamine but not dihydroartemisinin-piperaquine in Burkina Faso. *Antimicrob Agents Chemother* **54**: 1949–54.

Spielmann T, Hawthorne P. 2006. A Cluster of Ring Stage – specific Genes Linked to a Locus Implicated in Cytoadherence in *Plasmodium falciparum* Codes for PEXEL-negative and PEXEL-positive Proteins Exported into the Host Cell. *Mol Biol Cell* **17**: 3613–3624.

Sriprawat K, Kaewpongsri S, Suwanarusk R, Leimanis ML, Lek-Uthai U, Phyo AP, Snounou G, Russell B, Renia L, Nosten F. 2009. Effective and cheap removal of leukocytes and platelets from *Plasmodium vivax* infected blood. *Malar J* **8**: 115.

Stankiewicz P, Lupski JR. 2010. Structural variation in the human genome and its role in disease. *Annu Rev Med* **61**: 437–55.

Su X. 1999. A Genetic Map and Recombination Parameters of the Human Malaria Parasite *Plasmodium falciparum*. *Science (80-)* **286**: 1351–1353.

Su X, Hayton K, Wellems TE. 2007. Genetic linkage and association analyses for trait mapping in *Plasmodium falciparum*. *Nat Rev Genet* **8**: 497–506.

Su X, Kirkman L a, Fujioka H, Wellems TE. 1997. Complex polymorphisms in an approximately 330 kDa protein are linked to chloroquine-resistant *Plasmodium falciparum* in Southeast Asia and Africa. *Cell* **91**: 593–603.

Su X-Z, Mu J, Joy D a. 2003. The “Malaria’s Eve” hypothesis and the debate concerning the origin of the human malaria parasite *Plasmodium falciparum*. *Microbes Infect* **5**: 891–896.

Swanson WJ. 2003. Adaptive evolution of genes and gene families. *Curr Opin Genet Dev* **13**: 617–622.

Tajima F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123**: 585–95.

Tetteh KK a, Stewart LB, Ochola LI, Amambua-Ngwa A, Thomas AW, Marsh K, Weedall GD, Conway DJ. 2009. Prospective identification of malaria parasite genes under balancing selection. *PLoS One* **4**: e5568.

Trager W, Jensen JB. 2005. Human malaria parasites in continuous culture. 1976. *J Parasitol* **91**: 484–6.

Trape J-F, Tall A, Diagne N, Ndiath O, Ly AB, Faye J, Dieye-Ba F, Roucher C, Bouganali C, Badiane A, et al. 2011. Malaria morbidity and pyrethroid resistance after the introduction of insecticide-treated bednets and artemisinin-based combination therapies: a longitudinal study. *Lancet Infect Dis* **3099**: 1–8.

Trenholme KR, Gardiner DL, Holt DC, Thomas E a, Cowman a F, Kemp DJ. 2000. *clag9*: A cytoadherence gene in *Plasmodium falciparum* essential for binding of parasitized erythrocytes to CD36. *Proc Natl Acad Sci U S A* **97**: 4029–33.

Triglia T, Duraisingh MT, Good RT, Cowman AF. 2005. Reticulocyte-binding protein homologue 1 is required for sialic acid-dependent invasion into human erythrocytes by *Plasmodium falciparum*. *Mol Microbiol* **55**: 162–74.

Triglia T, Foote SJ, Kemp DJ, Cowman AF. 1991. Amplification of the multidrug resistance gene *pfmdr1* in *Plasmodium falciparum* has arisen as multiple independent events. *Mol Cell Biol* **11**: 5244–50.

Triglia T, Wang P, Sims PF, Hyde JE, Cowman a F. 1998. Allelic exchange at the endogenous genomic locus in *Plasmodium falciparum* proves the role of dihydropteroate synthase in sulfadoxine-resistant malaria. *EMBO J* **17**: 3807–15.

Tufet-Bayona M, Janse CJ, Khan SM, Waters AP, Sinden RE, Franke-Fayard B. 2009. Localisation and timing of expression of putative *Plasmodium berghei* rhoptry proteins in merozoites and sporozoites. *Mol Biochem Parasitol* **166**: 22–31.

Tuzun E, Sharp AJ, Bailey JA, Kaul R, Morrison VA, Pertz LM, Haugen E, Hayden H, Albertson D, Pinkel D, et al. 2005. Fine-scale structural variation of the human genome. *Nat Genet* **37**: 727–32.

Uhlemann A-C, Cameron A, Eckstein-Ludwig U, Fischbarg J, Iserovich P, Zuniga F a, East M, Lee A, Brady L, Haynes RK, et al. 2005. A single amino acid residue can determine the sensitivity of SERCAs to artemisinins. *Nat Struct Mol Biol* **12**: 628–9.

Vaidya AB, Mather MW. 2009. Mitochondrial evolution and functions in malaria parasites. *Annu Rev Microbiol* **63**: 249–67.

Van Tyne D, Park DJ, Schaffner SF, Neafsey DE, Angelino E, Cortese JF, Barnes KG, Rosen DM, Lukens AK, Daniels RF, et al. 2011. Identification and functional validation of the novel antimalarial resistance locus PF10_0355 in *Plasmodium falciparum*. *PLoS Genet* **7**: e1001383.

Vasconcelos KF, Plowe C V, Fontes CJ, Kyle D, Wirth DF, Pereira da Silva LH, Zalis MG. 2000. Mutations in *Plasmodium falciparum* dihydrofolate reductase and dihydropteroate synthase of isolates from the Amazon region of Brazil. *Mem Inst Oswaldo Cruz* **95**: 721–8.

Vernick KD, Walliker D, McCutchan TF. 1988. Genetic hypervariability of telomere-related sequences is associated with meiosis in *Plasmodium falciparum*. *Nucleic Acids Res* **16**: 6973–85.

Vieira P. 2001. Analysis of the PfCRT K76T Mutation in *Plasmodium falciparum* Isolates from the Amazon Region of Brazil. *J Infect Dis* **183**: 1832–3.

Voight BF, Kudaravalli S, Wen X, Pritchard JK. 2006. A map of recent positive selection in the human genome. *PLoS Biol* **4**: e72.

Volkman SK, Hartl DL, Wirth DF, Nielsen KM, Choi M, Batalov S, Zhou Y, Plouffe D, Le Roch KG, Abagyan R, et al. 2002. Excess polymorphisms in genes for membrane proteins in *Plasmodium falciparum*. *Science* **298**: 216–8.

Volkman SK, Neafsey DE, Schaffner SF, Park DJ, Wirth DF. 2012. Harnessing genomics and genome biology to understand malaria biology. *Nat Rev Genet* **13**: 315–28.

Volkman SK, Sabeti PC, DeCaprio D, Neafsey DE, Schaffner SF, Milner D a, Daily JP, Sarr O, Ndiaye D, Ndir O, et al. 2007b. A genome-wide map of diversity in *Plasmodium falciparum*. *Nat Genet* **39**: 113–9.

Vulliez-Le Normand B, Tonkin ML, Lamarque MH, Langer S, Hoos S, Roques M, Saul F a, Faber BW, Bentley G a, Boulanger MJ, et al. 2012. Structural and functional insights into the malaria parasite moving junction complex. *PLoS Pathog* **8**: e1002755.

Walliker D. 1983. The genetic basis of diversity in malaria parasites. *Adv Parasitol* **22**: 217–59.

Walliker D, Quakyi IA, Wellems TE, McCutchan TF, Szarfman A, London WT, Corcoran LM, Burkot TR, Carter R. 1987. Genetic analysis of the human malaria parasite *Plasmodium falciparum*. *Science* **236**: 1661–6.

Weedall GD, Conway DJ. 2010. Detecting signatures of balancing selection to identify targets of anti-parasite immunity. *Trends Parasitol* **26**: 363–9.

Wellems T, Panton L, Gluzman I. 1990. Chloroquine resistance not linked to *mdr*-like genes in a *Plasmodium falciparum* cross. *Nature* **345**: 253–255.

Wellems TE, Walker-Jonah a, Panton LJ. 1991. Genetic mapping of the chloroquine-resistance locus on *Plasmodium falciparum* chromosome 7. *Proc Natl Acad Sci U S A* **88**: 3382–6.

Westenberger SJ, McClean CM, Chattopadhyay R, Dharia N V, Carlton JM, Barnwell JW, Collins WE, Hoffman SL, Zhou Y, Vinetz JM, et al. 2010. A systems-based analysis of *Plasmodium vivax* lifecycle transcription from human to mosquito. *PLoS Negl Trop Dis* **4**: e653.

Wickramarachchi T, Cabrera AL, Sinha D, Dhawan S, Chandran T, Devi YS, Kono M, Spielmann T, Gilberger TW, Chauhan VS, et al. 2009. A novel *Plasmodium falciparum* erythrocyte binding protein associated with the merozoite surface, PfDBLMSP. *Int J Parasitol* **39**: 763–73.

Wilson RJ, McGregor IA, Hall P, Williams K, Bartholomew R. 1969. Antigens associated with *Plasmodium falciparum* infections in man. *Lancet* **2**: 201–5.

Winchester L, Yau C, Ragoussis J. 2009. Comparing CNV detection methods for SNP arrays. *Brief Funct Genomic Proteomic* **8**: 353–66.

Winter G, Kawai S, Haeggström M, Kaneko O, von Euler A, Kawazu S, Palm D, Fernandez V, Wahlgren M. 2005. SURFIN is a polymorphic antigen expressed on *Plasmodium falciparum* merozoites and infected erythrocytes. *J Exp Med* **201**: 1853–63.

Winzeler EA. 2008. Malaria research in the post-genomic era. *Nature* **455**: 751–6.

Wootton JC, Feng X, Ferdig MT, Cooper R a, Mu J, Baruch DI, Magill AJ, Su X-Z. 2002. Genetic diversity and chloroquine selective sweeps in *Plasmodium falciparum*. *Nature* **418**: 320–3.

Xangsayarath P, Kaewthamasorn M, Yahata K, Nakazawa S, Sattabongkot J, Udomsangpetch R, Kaneko O. 2012. Positive diversifying selection on the *Plasmodium falciparum surf4.1* gene in Thailand. *Trop Med Health* **40**: 79–89.

Ylstra B, van den Ijssel P, Carvalho B, Brakenhoff RH, Meijer G a. 2006. BAC to the future! or oligonucleotides: a perspective for micro-array comparative genomic hybridization (array CGH). *Nucleic Acids Res* **34**: 445–50.

Yoon S, Xuan Z, Makarov V, Ye K, Sebat J. 2009. Sensitive and accurate detection of copy number variants using read depth of coverage. *Genome Res* **19**: 1586–92.

Zeeshan M, Alam MT, Vinayak S, Bora H, Tyagi RK, Alam MS, Choudhary V, Mittra P, Lumb V, Bharti PK, et al. 2012. Genetic variation in the *Plasmodium falciparum* circumsporozoite protein in India and its relevance to RTS,S malaria vaccine. *PLoS One* **7**: e43430.

Zhang F, Gu W, Hurles ME, Lupski JR. 2009. Copy number variation in human health, disease, and evolution. *Annu Rev Genomics Hum Genet* **10**: 451–81.