

***PLASMODIUM FALCIPARUM* POPULATION
STRUCTURE INFERRED BY MSP1 AMPLICON
SEQUENCING OF PARASITES COLLECTED FROM
AREAS OF DIFFERING TRANSMISSION IN KENYA**

BRIAN KABUNGA ANDIKA

**MASTER OF SCIENCE
(Molecular Biology and Bioinformatics)**

**JOMO KENYATTA UNIVERSITY
OF
AGRICULTURE AND TECHNOLOGY**

2024

***Plasmodium falciparum* Population Structure Inferred by *msp1*
Amplicon Sequencing of Parasites Collected from Areas of Differing
Transmission in Kenya**

Brian Kabunga Andika

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

2024

DECLARATION

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

Signature Date

Brian Kabunga Andika

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

Signature Date

Prof. Naomi Maina, PhD

JKUAT, Kenya

Signature Date

Dr. Victor Mobegi, PhD

UoN, Kenya

Signature Date

Dr. John Waitumbi, PhD

U.S. Army Medical Research Directorate – Africa, Kenya

DEDICATION

This work is dedicated to my family. Their prayers and support have been the driving momentum that has enabled me to achieve the targeted goal in my studies.

ACKNOWLEDGMENT

I express my deep appreciation for the professional guidance and mentorship I was given by my supervisors. It was really a great privilege to work under their tutelage. Prof. Naomi Maina for her dedicated and generous support in all aspects of this work including review of this thesis and offered regular positive criticism throughout the research. Her support was without doubt, crucial throughout this study. I also extend my gratitude to Dr. Victor Mobegi of the Department of Biochemistry, University of Nairobi who together supervised my work for the thorough orientation he gave me at the onset of the project and the extensive correspondence had while doing this work.

I am greatly indebted to the U.S. Army Medical Research Directorate Laboratory Director in Kisumu, Dr. John Waitumbi for giving me the opportunity to work in his laboratory and for his continuous support during my master's program. Dr. Waitumbi formulated the study, funded and supervised the laboratory assays. I am grateful for his valuable scientific advice, constructive criticism, encouragement, deep commitment and guidance throughout the study. I have learnt a lot from him and acknowledge his commitment in ensuring that this work was done well. Special thanks to the team at basic science lab who trained me on various laboratory techniques.

I owe my sincere gratitude to my family for the undying love and support. They truly gave me strength, and support that I needed through those tough years. Finally, my deepest gratitude to God for giving me strength, both physically and mentally throughout the study period.

TABLE OF CONTENTS

DECLARATION	ii
DEDICATION	iii
ACKNOWLEDGMENTS	iv
TABLE OF CONTENTS	v
LIST OF TABLES	viii
LIST OF FIGURES	ix
LIST OF APPENDICES	x
ABBREVIATIONS AND APPENDICES	xi
ABSTRACT	xiii
CHAPTER ONE	1
INTRODUCTION	1
1.1 Background Information	1
1.2 Statement of the Problem	3
1.3 Justification	4
1.4 Hypothesis	4
1.5 Objectives	4
1.5.1 General Objective.....	4
1.5.2 Specific Objectives	4
CHAPTER TWO	6
LITERATURE REVIEW	6
2.1 Malaria Situation in Kenya.....	6
2.2 Molecular Epidemiology of Malaria	7
2.3 Polymorphism and Variation.....	8
2.4 Malaria Biology, Lifecycle and Host Responses	9
2.5 Merozoite Surface Protein 1	11
2.6 Genotyping of <i>Plasmodium</i> Parasites.....	12

2.7 Overview of Next Generation Sequencing.....	14
2.7.1 Haplotype Inference and MOI Estimation.....	14
CHAPTER THREE	16
MATERIALS AND METHODS	16
3.1 Study Design	16
3.2 Ethical Approval.....	18
3.3 Sampling Technique.....	18
3.3.1 Sample Collection.....	18
3.3.2 Extraction of Parasite DNA from Field Samples.....	19
3.4 <i>In vitro</i> Maintenance of <i>P. falciparum</i> 3D7 W2 Continuous Culture.....	19
3.4.1 Preparation of Merozoites.....	19
3.4.2 Synchronization of Parasite Cultures.....	19
3.4.3 Extraction of Parasite DNA from Cultured Lines	20
3.5 Primers and PCR Conditions.....	20
3.6 18S ribosomal RNA Real Time PCR for Cultured Parasites	21
3.7 Amplicon Sequence Library Preparation	22
3.7.1 Primary PCR.....	22
3.7.2 Nested PCR.....	23
3.7.3 Amplicon Library Indexing and Cleaning	23
3.7.4 Normalization and Pooling of Amplicon Libraries	24
3.7.5 Denaturing the Pooled Library and Sequencing.....	24
3.8 Bioinformatics Analysis	25
3.8.1 Haplotype Calling and Determination of Multiplicity of Infections	25
3.8.2 Statistical Analysis.....	25
CHAPTER FOUR.....	27
RESULTS	27
4.1 Limit of Detection and Quality Control	27
4.2 Demography and Malaria Parasitaemia in the Study Population.....	28

4.3 <i>Pfmsp1</i> Genetic Diversity.....	30
4.3.1 Genetic Diversity by Fragment Size Polymorphisms.....	30
4.3.2 Genetic Diversity by Nucleotide and Amino Acids Polymorphisms	31
4.4 Multiplicity of Infection	36
CHAPTER FIVE.....	40
DISCUSSION	40
5.1 Performance of Deep Sequencing Assay	40
5.2 Genetic Structure of <i>P. falciparum</i> Populations.....	41
5.3 Multiplicity of Infection and Genetic Diversity	42
CHAPTER SIX	43
CONCLUSION AND RECOMMENDATIONS	43
6.1 Conclusion.....	43
6.2 Recommendations	43
REFERENCES.....	45
APPENDICES	55

LIST OF TABLES

Table 3.1: List of Primers, Probes and their Corresponding Sequences.....	21
Table 4.1: <i>P. falciparum</i> Clonal Diversity by Age, Gender and Malaria Endemicity...	36

LIST OF FIGURES

Figure 2.1: <i>Plasmodium falciparum</i> Life Cycle.....	10
Figure 2.2: Schematic Structure of the Merozoite Surface Protein.....	12
Figure 3.1: Kenyan Map Showing the Location of the Nine Surveillance Sites.....	17
Figure 4.1: Amplification Plots to Determine Limit of Detection.....	27
Figure 4.2: <i>P. falciparum</i> msp1 Block 2 Resolved on 2% Agarose Gel.....	28
Figure 4.3: Scatter Plots Showing qPCR Ct Values in Different Age Groups.....	30
Figure 4.4: Distribution of <i>Pfmsp1</i> Allelic Families Based on Fragment Sizes.....	31
Figure 4.5: Frequencies of Amino Acid Substitutions Across Block 2 of K1.....	33
Figure 4.6: Frequencies of Amino Acid Substitutions Across Block 2 of MAD20.....	34
Figure 4.7: Frequencies of Amino Acid Substitutions Across Block 2 of RO33.....	35
Figure 4.8: Linear Plot Showing Linear Relationship of MOI Against Ct Values.....	38
Figure 4.9: Temporal Variation in Allelic Families in Regions of Different Malaria En- demicity.....	39

LIST OF APPENDICES

Appendix I: Sequence Alignment of K1 Alleles of msp1 Block 2.....	55
Appendix II: Sequence Alignment of MAD20 Alleles of msp1 Block 2.....	56
Appendix III: Sequence Alignment of RO33 Alleles of msp1 Block 2.....	56
Appendix IV: Ethical Clearance	57
Appendix V: Publication.....	60

ABBREVIATIONS AND ACROMYNS

ACT	Artemisinin-based Combination Therapy
ama1	Apical Membrane Antigen 1
AmpliSeq	Amplicon Sequencing
AS	Artesunate
bp	base pair
CE	Capillary Electrophoresis
LoD	Limit of Detection
glurp	glutamine rich protein
He	Expected Heterozygosity
HRP	Histidine-Rich Protein
IRS	Indoor Residual Spraying
ITN	Insecticide-Treated bed Net
MLST	Multi Locus Sequence Typing
MOI	Multiplicity of Infection
MGB	Minor Groove Binder
Pfmsp1	<i>Plasmodium falciparum</i> merozoite surface protein 1
PlasmoDB	Plasmodium Database
qPCR	quantitative PCR
RBC	Red Blood Cell

RDT	Rapid Diagnostic Test
RFLP	Restriction Fragment Length Polymorphism
SNP	Single Nucleotide Polymorphism
SRST2	Short Read Sequence Typing 2
18S rRNA	18S small subunit ribosomal RNA

ABSTRACT

An important measure of *Plasmodium falciparum* parasite diversity is multiplicity of infection (MOI), usually derived from the highly polymorphic genes such as *msp1*, *msp2*, *glurp* and microsatellites. MOI is used to distinguish recrudescence and new infecting clones to inform malaria control interventions. Conventional methods of deriving MOI lack fine resolution needed to discriminate minor clones. The aim of the current study was to infer *P. falciparum* population structure by *msp1* amplicon sequencing of parasites collected from areas of differing transmissions in Kenya. A total of 264 *P. falciparum* positive blood samples collected from patients with acute febrile illnesses were retrieved from frozen sample repository. Samples were collected over a 10-year period from 2010 to 2019 and originated from areas of varying malaria endemicities. *Pfmsp1* gene was amplified from extracted total DNA, amplicon libraries prepared and sequenced on an Illumina MiSeq platform. Children <5 years had higher parasitaemia (mean=23.5±5 SD, $p = 0.03$) than those ≥5-14 (mean=25.3±5 SD), and those ≥15 (mean=25.1±6 SD). MOIs and haplotype dynamics were derived and stratified by spatial and temporal measures. Of the 1014 alleles detected, 553 (54.5%) were K1, 250 (24.7%) were MAD20 and 211 (20.8%) RO33, that clustered into 19 K1 allelic families (108-270 bp), 14 MAD20 (108-216 bp) and one RO33(153 bp). By amplicon sequencing, the mean MOI was 4.8(±0.78, 95% CI) for the malaria endemic Lake Victoria region Alupe, Kombewa and Kisumu, 4.4(±1.03, 95% CI) for the epidemic prone Kisii highland (Kisii and Nyamira) and 3.4(±0.62, 95% CI) for the seasonal malaria semi-arid regions of Marigat, Isiolo, Lodwar, Iftin and Gilgil. High levels of clonal diversity were identified throughout the different transmission settings ($He=0.98$). This thesis describes additional 176 distinct allelic sequences to this database highlighting the added advantages that a highly sensitive tool such as AmpliSeq brings in malaria surveillance. The key findings of this study provide information that informs malaria control interventions such as *msp1*-based vaccine candidates.

CHAPTER ONE

INTRODUCTION

1.1 Background Information

Malaria is a life-threatening infectious disease caused by parasites of the *Plasmodium* genus transmitted through bites of infected female *Anopheles* mosquitoes. From 2000 to 2016, the World Health Organization (WHO) recorded significant progress in combating malaria in endemic areas (WHO, 2023). However, data from the 2023 WHO world malaria report showed that the progress towards reduction of global malaria cases had stalled in recent years. Approximately 600,000 people died from malaria, with 61% of the cases recorded children under the age of 5 (WHO, 2023). To combat the disease burden, intensive intervention efforts have been enhanced including treatment with anti-malarial drugs, use of insecticide-treated bed nets (ITNs) and indoor residual spraying. WHO estimates the global malaria control efforts to have helped reduce malaria deaths by more than 60% (WHO, 2023). In Kenya, malaria is a leading cause of morbidity and mortality with over 96% of malaria infections caused by *P. falciparum* (Kenya Demographic and Health Survey., 2021). The Ministry of Health estimates that 70% of the population in Kenya live in areas where malaria transmission occurs 8–12 months per year (Kenya Demographic and Health Survey., 2021).

Previous molecular studies revealed the occurrence of multiple genetically diverse *P. falciparum* strains that circulate in malaria endemic regions which contributes to the ability of *P. falciparum* to evade the host immune response and develop resistance to anti-malarial drugs (Anderson *et al.*, 2000; Ferreira *et al.*, 2004; Dzikowski *et al.*, 2009). Multiclonal malaria infections can influence clinical outcomes in a manner that is dependent on transmission intensity (Mahdi *et al.*, 2016) and negatively impacting an individual's response to anti-malarial drug treatment (Mavoko *et al.*, 2016).

RTS, S/AS01 vaccine remains the most advanced malaria vaccine, although its mechanism of action is poorly characterized (Nielsen *et al.*, 2018). Therefore, there is urgent need for development of vaccine against the most virulent *Plasmodium* species, *P. falciparum*, in particular. The vaccines under development, such as those targeting pre-erythrocytic stage proteins and asexual blood stage antigens, are primarily intended to prevent clinical disease (WHO, 2021). However, many blood-stage merozoite proteins that elicit protective immunity against malaria use parallel redundant pathways and are extremely polymorphic (Berzins *et al.*, 2002; Gaur *et al.*, 2004). A polymorphic antigen with strong immunogenicity may still be considered as a component of a multistage polyvalent vaccine and protect vulnerable populations in diverse transmission settings (WHO, 2021).

Merozoites are the most abundant surface antigens in the blood stage of *P. falciparum* and therefore plays a crucial role in the initial low affinity attachment of parasite to RBC membrane during erythrocyte invasion (Lin *et al.*, 2015). Merozoite surface protein (*mSP1*) contains 17 blocks (block 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16 and 17) of which block 2 shows extensive allelic polymorphism worldwide (Tanabe *et al.*, 1987; Miller *et al.*, 1993). Block 2 alleles are mainly represented by three clonal families namely K1, MAD20 and RO33 in the field isolates based on their characteristic tri-peptide motifs (Zhong *et al.*, 2018). Previous studies (Alam *et al.*, 2011; Mahdi *et al.*, 2016; Zhong *et al.*, 2018) show that *mSP1* gene pool is shaped by a localized pattern of parasite transmission and its immunogenic repertoire is furnished with a limited number of conserved epitopes. The suitability of the *mSP1* as a potential vaccine target, as revealed in these studies may have significant implications in the global malaria eradication initiatives.

Some studies have proposed the use of Multiplicity of Infections (MOI) for evaluating changes in malaria transmission intensity after the implementation of malaria control programs (Chang *et al.*, 2012; Niang *et al.*, 2017). Conversely, other studies have demonstrated the lack of correlation between malaria transmission intensity and MOI (Alam *et al.*, 2011; Duah *et al.*, 2016; Touray *et al.*, 2020). One potential confounder in these association studies is the use of different genotyping methods. Given this, the present study

aimed to determine the genetic diversity of *msh1* block 2 using amplicon deep sequencing. In parallel, the question, how the observed allelic variation of this segment affects the distribution of clones temporally and spatially was also addressed. SeekDeep (Hathaway *et al.*, 2018) was used to reveal the genetic diversity of plasmodium circulating in Kenya. This is a bioinformatics pipeline designed for analysis of haplotype frequency from amplicon sequencing data and has been used successfully in several studies investigating malaria population genetics globally (Parobek *et al.*, 2014; Lin *et al.*, 2015; Boyce *et al.*, 2018).

1.2 Statement of the Problem

Falciparum malaria is the biggest health burden in sub-Saharan Africa, of the 249 million cases globally, 233 million (around 94%) were African, malaria is a leading cause of morbidity and mortality with over 96% of malaria infections due to *P. falciparum* (Kenya Demographic and Health Survey., 2021). Numerous studies have sought to establish the impact of MOI in asymptomatic and symptomatic infections, the possible connection with risk of morbidity and the relationship that exist between MOI and parasite density. The last decade however has seen rapid development of next generation sequencing (NGS), also called high throughput sequencing or deep sequencing. The NGS is widely applicable to field samples for epidemiology, epigenetic and microbiome studies. However, performing NGS on field samples, is much more challenging than on samples from laboratory-cultivated parasites and requires more robust analysis methods. In case of *Plasmodium* blood samples, the main challenges for the laboratory work are contamination with host DNA, the large biological variation and different developmental stages of the *Plasmodium* parasite. This may lead to overpopulation of minority clones or sequence with high similarity with host DNA. As a result, often no biological replicates are feasible, because each patient harbors a unique parasite strain and a unique mixture of stages. Most NGS analysis methods are not developed for complex field isolates and therefore need adaptations to be applicable on such field samples.

1.3 Justification

A key epidemiological consequence of *Plasmodium* genetic diversity is the existence of natural infections consisting of multiple clones especially in areas of high malaria endemicity (Trape *et al.*, 1994; Contamin *et al.*, 1995). The number of co-infecting parasite genotypes referred to as multiplicity of infection (MOI) is determined by genotyping polymorphic genes or gene segments to describe the minimum number of variant clones in an infection (Tanabe *et al.*, 1987). Genes routinely analyzed for determination of MOI typically include *Pmsp1*, *Pmsp2* and *glurp* for their size polymorphisms by PCR followed by visualization on gel or fragmentation by Capillary Electrophoresis (CE), this thesis describes amplicon sequencing of *Pfmsp1* block 2 to determine the recrudescence and newly infecting clones already persisting in the Kenyan population. While high prevalence of low parasitaemia makes it difficult to obtain a precise estimate of the MOI (Miller *et al.*, 2017), an additional method to determine the limit of detecting minority clones would further complement that using amplicon deep sequencing method to reveal the true diversity in the different transmission settings of Kenya.

1.4 Hypothesis

We hypothesize that Deep sequencing of *m*sp1 is a reliable marker for studying temporal and spatial changes to establish the genetic diversity in *Plasmodium falciparum*.

1.5 Objectives

1.5.1 General Objective

To assess the temporal and spatial changes in *P. falciparum* population structure using amplicon sequencing (AmpliSeq) of clones already persisting in different transmission settings in Kenya.

1.5.2 Specific Objectives

- (i) To determine the limit of detection of low-density parasites and the utility of deep sequencing in determining *m*sp1 block 2 allele families.

- (ii) To determine the temporal and spatial *P. falciparum* genetic population structure of *msp1* gene by targeted.
- (iii) To determine genetic diversity and multiplicity of infection of *P. falciparum* isolates collected from areas of differing malaria transmission in Kenya.

CHAPTER TWO

LITERATURE REVIEW

2.1 Malaria Situation in Kenya

Of the 249 million cases noted in 2022 globally, 233 million (around 94%) were in the WHO African Region, malaria is a leading cause of morbidity and mortality with over 96% of malaria infections due to *P. falciparum* (Kenya Demographic and Health Survey., 2021). The Ministry of Health estimated that 70% of the population in Kenya lives in areas where malaria transmission occurs 8–12 months per year (Kenya Demographic and Health Survey., 2021). Other Plasmodium species *Plasmodium P. malariae*, *P. ovale*, and *P. vivax* were reported in the country (Okara *et al.*, 2010). Malaria transmission and the risk of infection in Kenya are largely influenced by altitude, temperature and rainfall patterns (Kenya Demographic and Health Survey., 2021). Therefore, prevalence of malaria cases varies considerably according to the season and geographic regions (Kenya Demographic and Health Survey., 2021). The variations in altitude and terrain create contrasts in the country's climate, which ranges from tropical along the coast, to temperate at the interior to very dry in the north and northeast (Kenya Demographic and Health Survey., 2021). The country is divided into the following four eco-epidemiological strata of malaria:

- (i) **Endemic areas:** These areas experience stable malaria have altitudes ranging from zero in the coastal region to 1,300 meters around the malaria endemic Lake Victoria region in western Kenya (Kenya Demographic and Health Survey., 2021). Malaria transmission is intense throughout the year, with *P. falciparum* prevalence between 20-40% and high entomological inoculation rates of 29.2% per year (MoH 2016). The coastal region has malaria prevalence ranging from 5–20%. Of the total Kenyan population, 26% lives in a malaria-endemic zone.
- (ii) **Highland epidemic prone areas:** Epidemics are experienced in western highlands of Kenya where malaria transmission is seasonal with considerable year-to-

year variation (Kenya Demographic and Health Survey., 2021). The entire population is vulnerable and case-fatality rates during an epidemic can be greater than in endemic regions (Waitumbi *et al.*, 2009). Approximately 39% of Kenyans live in these areas (Kenya Demographic and Health Survey., 2021). The malaria prevalence in these areas ranges from 1–10% but some foci have prevalence rates ranging between 10% and 20% (Kenya Demographic and Health Survey., 2021).

(iii) **Seasonal malaria transmission areas:** This epidemiological zone comprises arid and semiarid areas of northern and southeastern parts of the country, which experience short periods of intense malaria transmission during the rainy seasons (Kenya Demographic and Health Survey., 2021). Although this is the largest zone in terms of geographic size, only 14% of the population lives in these areas where the malaria prevalence is less than 5% (Kenya Demographic and Health Survey., 2021).

(iv) **Low malaria risk areas:** This zone covers the central highlands of Kenya including Nairobi. Approximately 21% of the population inhabits this area where there is little to no malaria transmission (Kenya Demographic and Health Survey., 2021).

2.2 Molecular Epidemiology of Malaria

Traditionally, the interest on infectious diseases has been focused on the role of the infectious agents at the origin of the disease in higher organisms (Alam *et al.*, 2011). Scientists' interest has been expanded to include the genetic structure, the immune response and the evolutionary consequences of public health interventions (Alam *et al.*, 2011). Molecular epidemiology can be a potential tool for understanding infectious diseases such as *P. falciparum* malaria, and the evaluation of interventions for their treatment (drug trials) and prevention (exposure-reducing measures and vaccines) (Waitumbi *et al.*, 2009). One practical measure of molecular epidemiology is to identify the infective agents responsible for

the disease for instance to distinguish between the different human pathogenic *Plasmodium* species), and secondly, to determine their biological (phylogenetic) relationships, and their routes of transmission. Genes responsible for their virulence (e.g., var genes in *Plasmodium falciparum*), vaccine-relevant antigens (*Pfmsp1/pfmsp2/pfama1*) and drug resistance (Pyrimethamine/Sulfadoxine resistance) have been characterized (Amin *et al.*, 2007).

A detailed understanding of the overall genetic structure of a pathogen population is not only essential for epidemiological tracking, but used for typing characters that change in a range and at a rate that is informative for the particular question. The knowledge of the genetic structure may be important to understand and predict the responses of pathogen populations to selective pressures imposed by host immunity, both natural and vaccine-induced, and could be important for effective management of anti-parasitic interventions. For tracking the transmission of malaria in a localized area, a relatively fast-changing genetic marker will be useful (Lerch *et al.*, 2019). On the other hand, the same marker would not be valuable for looking at trends in the global population of *P. falciparum* over many decades (Doolan *et al.*, 2009). Polymorphic marker genes have been shown to be very useful tools to evaluate intervention strategies and to monitor prevalence of malarial parasites or changes in frequencies of genotypes (Boyce *et al.*, 2020).

2.3 Polymorphism and Variation

DNA Polymorphism is defined as the expression of distinct alleles of a gene at a single gene locus in different clones of the parasite (Reeder and Brown, 1996), whereas variation shows the ability of a clonal population to switch the antigenic phenotype with unchanged genotype (Snounou *et al.*, 2013). As pointed out by (Polley and Conway, 2001), allelic polymorphism is usually a “between-host” survival strategy, providing an individual pathogen with maximum fitness for successful infection of its host. The biological role and impact of antigenic variation is usually considered as a “within-host” mechanism that allows parasite survival in an immunocompetent host (Polley and Conway, 2001). Antigenic variation creates diverse individuals, by switching its antigenic phenotype during long-

term infection of a single infected host. Both strategies are realized as immune evasion mechanisms in malaria.

2.4 Malaria Biology, Lifecycle and Host Responses

Malaria is caused by parasites of the genus *Plasmodium*. Malaria parasites infect a variety of hosts, ranging from reptiles and birds to mammals, primates, and humans (CDC). *Plasmodium* parasites are transmitted by mosquitoes of the genus *Anopheles*, which are ubiquitous throughout the world (WHO, 2023). Five of these *Plasmodium* species can cause diseases in humans: *P. falciparum*, *P. vivax*, *P. ovale*, *P. malariae*, and *P. knowlesi*. *P. cynomolgi* and *P. simium* have also been recently observed to transmit the disease (Brasil *et al.*, 2017; Hartmeyer *et al.*, 2019; Raja *et al.*, 2020). Together, *P. falciparum* and *P. vivax* cause many malaria cases globally, with *P. falciparum* being more lethal (Doolan *et al.*, 2009). All *Plasmodium* parasites share a similar, but complex, life-cycle. The meiotic and sole diploid stage occurs in the *Anopheles* mid-gut. After meiosis, haploid sporozoites are produced, which migrate to the mosquito salivary glands (Doolan *et al.*, 2009). As the mosquito draws a blood meal from a potential host, sporozoites are deposited into the host, which then migrates into the bloodstream and infect hepatocytes (Rodrigues *et al.*, 2008). After several days of intra-hepatic development, infected hepatocytes release schizonts into the bloodstream. Schizonts rupture, each releasing numerous merozoites, which infect circulating red blood cells (RBCs). Infected RBCs undergo multiple cycles of parasite replication, rupture, and re-infection, causing both a recurring fever and a rapid increase in parasitaemia. As a byproduct of blood stage infection, sexual forms, called gametocytes, are produced, which are ingested by mosquitoes thus the life-cycle repeats (Figure 2.1). With repeated *Plasmodium* infections, hosts develop partial immunity that may be strain specific. However, with continued exposure this immunity fades, though it does not completely disappear (Doolan *et al.*, 2009).

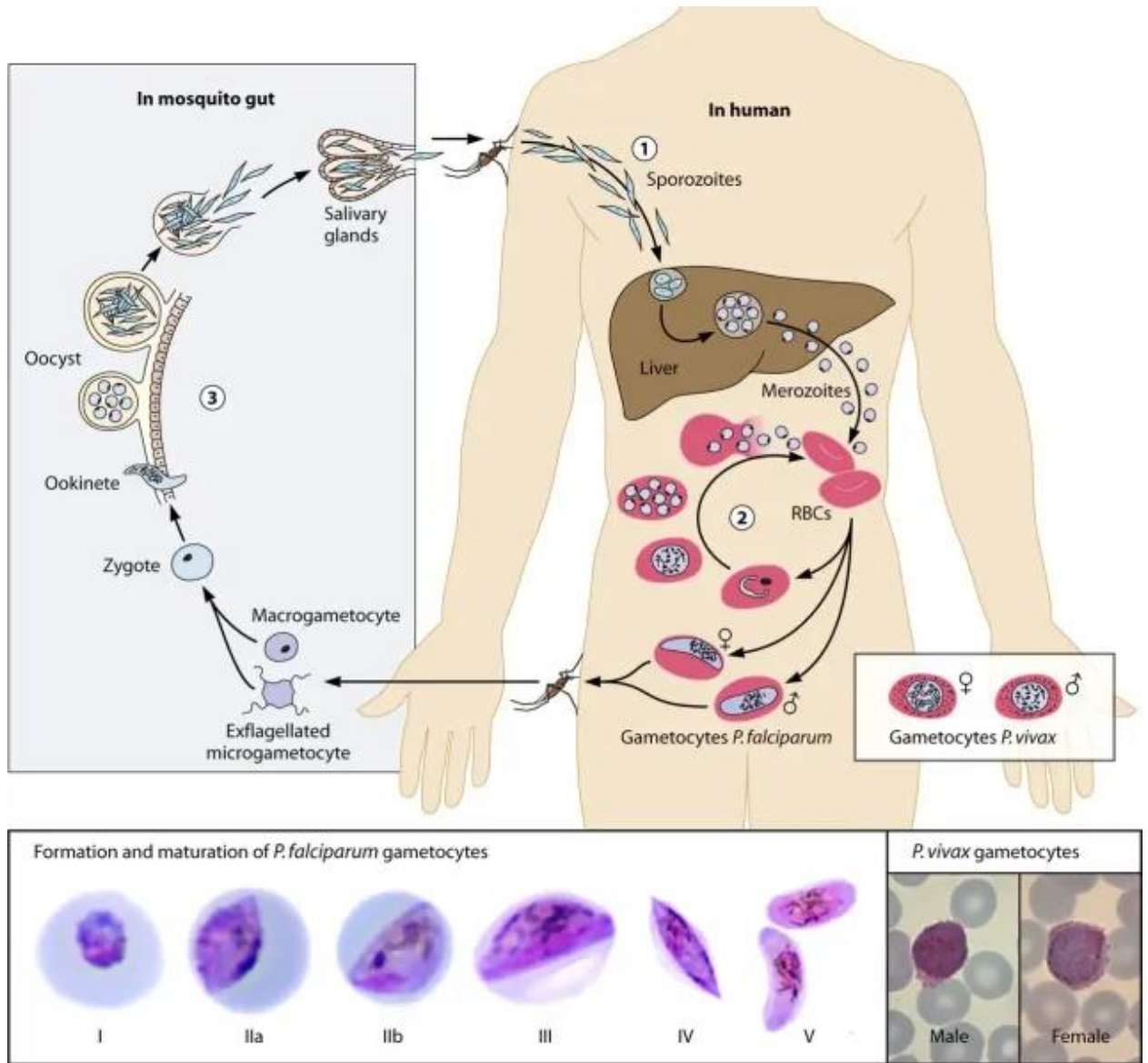


Figure 2.1: *Plasmodium falciparum* Life Cycle in Mosquito and Human Host. The different sexual and asexual stages are highlighted (Source: CDC, 2020)

2.5 Merozoite Surface Protein 1

Merozoite surface protein 1 (*msp1*) is a polymorphic protein that is initially synthesized as a large (~190 kDa) precursor during merozoite development within the infected erythrocyte. This precursor molecule subsequently undergoes post-translational processing that results in a series of fragments of 83, 42, 38 and 28–30kDa, which are then found on the surface of the newly released merozoite as a non-covalent linked protein complex (Holder *et al.*, 1992). Numerous studies have been undertaken to describe sequence diversity of *msp1* in diverse geographical regions, through the sequencing of field isolates from these Sudan, Ethiopia and Uganda, it has been possible to understand the genetic structure of the *msp1* gene, to elucidate the polymorphisms that are characteristic of its constituent segments and define various *msp1* genotypes (Liljander *et al.*, 2009). Inter-species comparisons of nucleotide sequences have led to the description of *msp1* gene in 17 blocks consisting of seven variable blocks, interspersed between five conserved and five semi-conserved blocks (Figure 2.2). Block 1 is a semi conserved region consisting of 55 amino acids, and amino acid variances occur at only three residues at position 44 (G or S), 47(H or Q) and 52 (I or V) (Miller *et al.*, 1993). Block 2 is a major polymorphic region of the *msp1* gene with three alleles K1, MAD20 and RO33. MAD20 and K1 consists of a series of tripeptide repeats, while the RO33 type is a non-repetitive sequence showing little variation between isolates. Within blocks 3, 5 and 7 diversities are generated by intragenic recombination between representative sequences of this region generated by fusion of K1 and MAD20 types, the conserved cysteine rich block 17, also referred to as 19kDa or *msp1*₁₉.

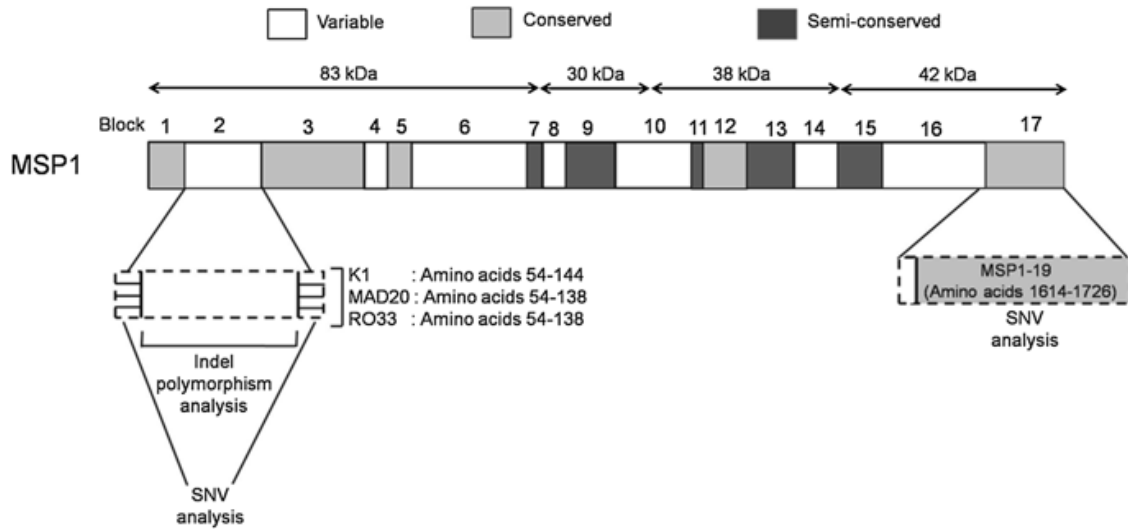


Figure 2.2: Schematic Structure of the Domains of Merozoite Surface Protein. The Regions Subjected to Sequence Analysis were Highlighted using Broken Lines (Source: Ghoshal et al., 2018).

2.6 Genotyping of *Plasmodium* Parasites

Individual parasite clones are identified by genotyping. To determine not only MOI or duration of infection, but also to study population structure or phenotypes like drug resistance (Reeder *et al.*, 1996). Depending on the genotyping application, different marker sets are selected (Lin *et al.*, 2015; Touray *et al.*, 2020). A single marker of high resolution is often sufficient for epidemiological studies where individual clones need to be identified. For studying phenotypes like drug resistance, markers covering all mutations (e.g., several SNPs within a gene, or several genes) associated with resistance must be typed. In population genetic studies, multiple genome-wide markers that are unlinked from each other and not under selection pressure are required. For recrudescence typing in anti-malarial drug efficacy trials, the use of three unlinked markers with high resolution are recommended by the WHO (WHO, 2023). The first methods to genotype *P. falciparum* relied

on the amplification of length-polymorphic merozoite surface protein 2 (*m*sp2) and subsequent sizing either by full length fragment or by restriction fragment length polymorphism (RFLP) (Zakeri *et al.*, 2005). In 2006, PCR-RFLP was modified to capillary electrophoresis (CE). This change increased resolution by using different fluorescent-labels for the FC27 and 3D7 allelic families (Lerch *et al.*, 2019). CE simplified analysis by omitting the interpretation of the RFLP size pattern, which was difficult to analyze especially when RFLP size patterns of multiple concurrent clones were superimposed (Early *et al.*, 2017). Currently, the recommended marker and method for genotyping in drug trials is merozoite surface protein 1 (*m*sp1), *m*sp2 and glutamine rich protein (*glurp*) by CE (WHO, 2021). Another genotyping method is typing of 24-42 SNPs (SNP barcode) that are distributed over the whole genome (Fulakeza *et al.*, 2019). This multi-locus SNP-typing (MLST) can determine genome-wide diversity and is suited for population studies, as selected SNPs are unlinked to each other. Mutations of SNPs are determined by either High Resolution Melting (HRM), Oligonucleotide Ligation or TaqMan (Rieneck *et al.*, 2015). However, SNP-typing requires a lot of DNA template, as each SNP is typed as an independent assay. Another challenge is the haplotype inference in case of multi-clone infections. The haplotypes of sample with mixed infection are difficult to resolve if the genotypes are unlinked to each other (Koepfli *et al.*, 2017). Improvement in next generation sequencing technologies (Illumina, 454/Roche or Ion Torrent) towards longer sequence reads and lower sequencing cost per sample by multiplexing of samples permitted the use of amplicon sequencing in epidemiological studies. Amplicon deep sequencing genotyping has a higher sensitivity, quantifies proportion of different variants and can detect low-abundant clones (minority clones) in samples with multiple concurrent infections. However, the higher sensitivity of amplicon sequencing comes at a cost of calling false alleles caused by sequencing error or PCR artefacts (Goshal *et al.*, 2018). First amplicon sequencing genotyping of *P. falciparum* used the length polymorphic markers *m*sp1 and *m*sp2, as well as the SNP polymorphic region of circumsporozoite protein (*csp*) (Miller *et al.*, 2017; Gruenberg *et al.*, 2019; Castañeda-Mogollón *et al.*, 2023). In the past few

years, whole genome sequencing (WGS) of single clone infections also became an option to determine genotypes (Yavo *et al.*, 2016). However, the cost per sample is high and the sequence library preparation is also quite laborious for large studies. For mixed clone infections, WGS is not feasible as the minority clone can only be detected at high sequencing costs.

2.7 Overview of Next Generation Sequencing

In the last decade, next generation sequencing (NGS), also called high throughput sequencing or deep sequencing, became widely applicable to field collected samples for molecular epidemiology studies. Performing NGS on field collected samples is much more challenging than on samples from laboratory cultivated parasites and requires more robust methods of analysis. In case of *Plasmodium* samples collected from patients the main challenges for the laboratory work are that the amount of input material is limited and contaminated with host DNA or RNA. For data analysis, the large biological variation between field samples is a challenge. Field samples can contain complex mixtures of infecting clones or development stages. As a result, often no biological replicates are feasible, because each patient harbors a unique parasite strain and a unique mixture of stages. Most NGS analysis methods are not developed for complex field isolates and therefore need adaptations to be applicable on such samples.

2.7.1 Haplotype Inference and MOI Estimation

In amplicon sequencing, multiple SNPs are usually linked by a single sequence read. Haplotype inference in such data can be done by clustering of those sequence reads, e.g., SeekDeep (Hathaway *et al.*, 2018), DADA2 (Callahan *et al.*, 2016), and SRST2 (Inouye *et al.*, 2014). The clustering combines similar sequence reads together that differ because of amplification or sequencing errors. However, also sequence reads from closely related clones cluster together, if they differ in only one SNP (Goshal *et al.*, 2018). For data from WGS or SNP barcodes, where SNPs are unlinked or only partly overlapping by sequence reads, the number of co-infecting clones should be estimated before haplotype inference can be performed. Multiplicity of infection (MOI) is defined as the number of co-infecting

parasite clones. Individuals in countries with high transmission of *Plasmodium* are often infected with several clones concurrently (Miller *et al.*, 2017), this superinfection can be caused by multiple infective mosquito bites or by a single mosquito bite injecting multiple genetically distinct parasite clones (Lerch *et al.*, 2019). SeekDeep is currently the most used method to analyze AmpliSeq genotyping data of *Plasmodium*. However, SeekDeep can only be used on a cluster with a large working memory capacity. In this thesis, an in-depth analysis using SeekDeep was used for simple analysis of amplicon sequencing data.

CHAPTER THREE

MATERIALS AND METHODS

3.1 Study Design

This study was part of a cross-sectional surveillance study of *P. falciparum* infection prevalence in populations of across Kenya. To test the hypothesis that *P. falciparum* genetic diversity determined by *msp1* amplicon deep sequencing would reveal the parasites temporal and spatial changes in *P. falciparum* with a high resolution, experimental procedures conducted in this study were categorized into two phases.

Phase one involved developing an amplicon sequencing assay to genotype *P. falciparum*. Here the genotyping principle described by (Lerch *et al.*, 2018) was adopted and which has successfully been applied in typing malaria parasite *P. falciparum*. Selection of *Plasmodium falciparum* merozoite surface protein (*Pfmsp1*) as the marker of choice for this study was done based on existing literature, in which three polymorphic markers and microsatellites described and validated by (Snounou *et al.*, 1999) were considered for typing *P. falciparum*.

Laboratory grown 3D7 *P. falciparum* lines were used to optimize and validate the methodology as well as determine the limit of detection (LOD). To evaluate LOD and the sensitivity of the assay, amplification was done on DNA extracted from red blood cell cultures with varying parasite densities for each *P. falciparum*. Specificity of the developed assay was evaluated by examining its ability to accurately identify the correct alleles in the laboratory lines. All PCR amplifications were re-amplified and re-scored at least twice to assess the reproducibility in detection and size determination of the assay.

Phase two involved genotyping *P. falciparum* field isolates using the validated assay. Field samples identified to be *P. falciparum* positive by a rapid diagnostic test (ABBOTT Bioline Malaria Ag P.f/Pan test kit) and microscopy were selected for genetic analysis.

To estimate genetic complexity of the parasite, each locus in each amplified and sequenced sample were scored using the Seek Deep bioinformatics pipeline.

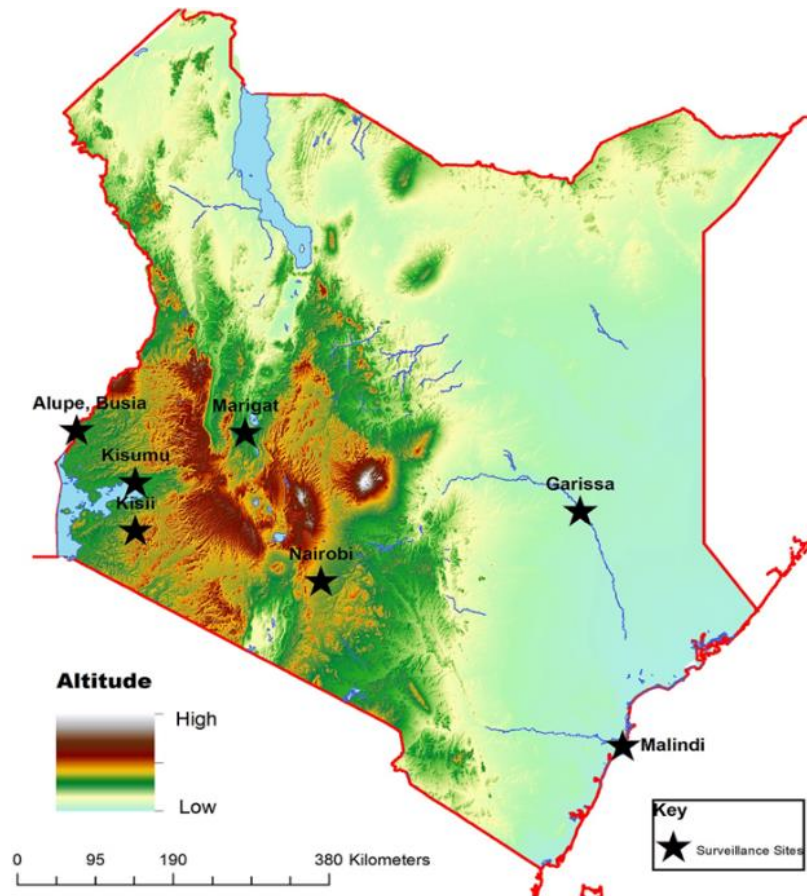


Figure 3.1: Kenyan Map Showing the Location of the Nine Surveillance Sites of the Acute Febrile Illness Study (AFI). Nine Sampling Sites Included: Marigat, Kisii, Lodwar, Gilgil, Alupe, Kombewa, Iftin (Garissa), Kisumu (Kisumu District Hospital) and Isiolo.

3.2 Ethical Approval

Eligible subjects were recruited under a study protocol that was approved by the Ethical Review Committee of the Kenya Medical Research Institute (IRB protocol KEMRI SERU # 1282) (Appendix IV) and the Walter Reed Army Institute of Research Human Subjects Protection Board in the United States of America (WRAIR HSPB # 1402).

3.3 Sampling Technique

This research utilized samples from consenting subjects attending clinics in areas of differing transmissions from the year 2010 to 2019. Inclusion criteria: febrile individuals positive for *P. falciparum* by light microscopy and rapid diagnostic test were considered as potential candidates for the study after giving informed consent. Exclusion criteria: Patients under malaria treatment and those who did not give informed consent.

3.3.1 Sample Collection

This study was part of an ongoing epidemiological survey project where finger prick dried blood spot samples for children under the age of five and whole blood samples for older age groups were obtained from study participants in accordance with the inclusion criteria. In total, 274 archived whole blood and dried blood spot (DBS) samples; 95 were from malaria endemic Lake Victoria region, 52 from the epidemic prone Kisii highland and 117 from semi-arid areas that experience intermittent malaria.

The samples were initially screened in the field for *P. falciparum* by microscopy and rapid diagnostic test. Parasitaemia levels of *P. falciparum* positive samples were determined and recorded for archiving. Whole blood samples were stored in ethylenediaminetetraacetic acid (EDTA) tubes and archived at -80°C until use, and dried blood spots (DBS) from *P. falciparum* positive samples were prepared by absorbing approximately 50 µl of the collected blood onto Whatman filter paper (Whatman, Maidstone, UK) and air-drying. DBS were stored at -80°C in sealed zip-lock bags with desiccant awaiting DNA extraction.

3.3.2 Extraction of Parasite DNA from Field Samples

From whole blood and dried blood spots, genomic DNA was extracted using Qiagen QIAamp DNA Mini Kit (Qiagen, Crawley, UK). In brief, three punches (6mm in diameter) were taken from each blood spot and extracted according to manufacturer's instructions. DNA was eluted in a final volume of 200 μ l, aliquoted and kept under frozen conditions (-80 °C) until use.

3.4 *In vitro* Maintenance of *P. falciparum* 3D7 W2 Continuous Culture

Laboratory cultured *P. falciparum* lines W2 continuous culture lines as described by (Oduola *et al.*, 1988) was used to initiate and maintain a culture as described by (Trager and Jensen, 1976) with minor modifications. Briefly, growth of the 3D7 malaria parasites was initiated in washed group O⁺ human RBC diluted to 5% hematocrit in complete RPMI 1640 media supplemented with 0.2% bicarbonate, 25Mm HEPES, 50 μ g/ml gentamicin and 10% heat inactivated human serum. Culture was maintained in 25cm² corning flasks (Corning incorporated, Corning NY, USA) with daily replacement of growth medium (RPMI + heat inactivated human serum).

3.4.1 Preparation of Merozoites

Merozoites were harvested from the supernatant of centrifuged iRBCs (10 minutes at 800 \times g). The supernatant was transferred to a fresh Falcon tube and centrifuged for ten minutes at 3000 \times g. Pelleted merozoites were re-suspended and used for the preparation of parasite derived proteins or IFA slides (Lambros and Vanderberg, 1979).

3.4.2 Synchronization of Parasite Cultures

To enrich for early ring stages parasites the culture was synchronized with 5% D-sorbitol in distilled water which lyses RBCs containing late ring stages and other mature parasites (Lambros and Vanderberg, 1979). This treatment was repeated every 48 hours until >98% of the parasites were in the ring stage as confirmed by microscopy. Parasites were maintained in culture until 3% ring stage (equivalent to 18,000 parasites/ μ l) was achieved.

3.4.3 Extraction of Parasite DNA from Cultured Lines

Parasite culture genomic DNA was extracted using Qiagen QIAamp DNA Mini Kit (Qiagen, Crawley, UK). Briefly, serial dilutions of the ring stage parasites were made to 0.55 parasites/ μ l 200 μ l of each dilution was used for extraction as recommended by the manufacturer. DNA was eluted to a final volume of 50 μ l, aliquoted and kept under frozen conditions (-80°C) until use.

3.5 Primers and PCR Conditions

A region of *Pfmsp1* covering nucleotides 1201627 to 1202210 (to include whole the expanse of block 2 K1, MAD20 and RO33 alleles) was amplified in a primary PCR. A subsequent nested PCR was performed using a combination of degenerate primers (Table 3.1) to accommodate strain differences in the *Pfmsp1* gene.

Nested PCR primers were laced with Illumina adapter overhang for compatibility with the sequencing platform. Briefly, in the primary PCR, 3 μ l of DNA template, 0.625 μ M of each primer and 1 \times NEB Next HIFI master mix were used in a 25 μ l reaction that included initial denaturation at 95 °C for five minutes, followed by 25 cycles of denaturation at 94 °C for one minute, annealing at 58 °C for two minutes and extension at 72 °C for two minutes, then a single annealing step at 58 °C for two minutes and final extension at 72 °C for five minutes.

In the secondary PCR, 2.5 μ l of DNA template, 0.2 μ M of each primer, 1 \times NEB Next HIFI master mix were used in a 25 μ l reaction. Cycling conditions included initial denaturation at 95 °C for five minutes, followed by 25 cycles of denaturation at 94 °C for 30 seconds, annealing at 55 °C for 30 seconds and extension at 72 °C for 30 seconds, then a final extension at 72 °C for five minutes. Amplicons were visualized on 2% agarose gel stained with gel red (Invitrogen, Carlsbad, CA).

3.6 18S ribosomal RNA Real Time PCR for Cultured Parasites

18S rRNA qPCR was performed as described previously (Murphy *et al.*, 2012; Wampfler *et al.*, 2013) on the serial dilutions of the ring stage parasites to determine the limit of detection (LOD) using primers and probes in a total volume of 20 μ l. The limit of detection (LOD) was defined as the last dilution at which more than 50% of replicates were positive. The amount of target DNA in each sample was calculated from the C_i value using a standard curve as described above in a previous method. Primer and probe sequences, as well as qPCR mixes and cycling conditions were as follows: an initial step of 15 minutes at 95 °C to activate the chemically modified hot-start Taq DNA polymerase, followed by 45 cycles of a 15-second denaturation at 95 °C and then 60 seconds annealing and extension at 60 °C.

Table 3.1: List of Primers, Probes and their Corresponding Sequences

Primer name	Target gene	Sequence (5' to 3' end)
Plasmoprobe	18s rRNA	6FAM-ATGGCCGTTTTTAGTTCGTG-TAMRA
Plasmo-F primer	18s rRNA	GCTCTTTCTTGATTTCTTGGATG
Plasmo-R primer	18s rRNA	AGCAGGTTAAGATCTCGTTCG
Falciprobe	18s rRNA	FAM-TTGCATATGGAAAAGATACCT-MGB
Falci-F	18s rRNA	CCATCAAGAGATTTAGGATCCAGATT
Falci-R	<i>P. falciparum</i> R	GCTACAAGAGGTACCCAAAAATAAAAA

Forward primers mix	pri-	Merozoite surface protein 1 (msp1)	5'TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCTAGAAGCTTTAGAAGATGCAGTATTG-3'
			5'TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGNCTAGAAGCTTTAGAAGATGCAGTATTG-3'
			5'TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGNNCTAGAAGCTTTAGAAGATGCAGTATTG-3'
			5'TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGNNNCTAGAAGCTTTAGAAGATGCAGTATTG
Reverse primers mix	pri-	Merozoite surface protein 1 (msp1)	5'GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGTGATTGGTTAAATCAAAGAGTTCGG-3'
			5'GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGNTGATTGGTTAAATCAAAGAGTTCGG-3'
			5'GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGNNTGATTGGTTAAATCAAAGAGTTCGG-3'
			5'GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGNNNTGATTGGTTAAATCAAAGAGTTCGG-3'

3.7 Amplicon Sequence Library Preparation

3.7.1 Primary PCR

QIAamp DNA Blood Mini Kit (Qiagen) was used to extract DNA from 200 µl of study samples as recommended by the manufacturer. A region of *Pfmsp1* from nucleotides 1201626 to 1202210 (includes the K1, MAD20 and RO33 alleles) was amplified in a primary PCR. Multiplexed primary PCR was performed in a total volume of 25 µl including 3 µl template DNA, 1nM of each primary primer (Invitrogen, USA) and 12.5µl AmpliTaq Ready Mix. Cycling conditions were as follows: initial denaturation 95 °C for three minutes followed by 25 cycles of 30 seconds denaturation at 98 °C, 30 seconds annealing

at 58 °C and 30 seconds elongation at 72 °C plus a final elongation of two minutes at 72 °C.

3.7.2 Nested PCR

A subsequent nested PCR was performed using a combination of degenerate primers (Table 3.1) to accommodate clonal differences in the *Pfmsp1* gene, with Illumina adapter overhang. Briefly, in the primary PCR, 3µl of DNA template, 0.625µM of each primer and 1× NEB Next HIFI master mix were used in a 25 µl reaction that included initial denaturation at 95 °C for five minutes, followed by 25 cycles of denaturation at 94 °C for 1 minute, annealing at 58 °C for two minutes and extension at 72 °C for two minutes, then a single annealing step at 58 °C for two minutes and final extension at 72 °C for five minutes. In the secondary PCR, 2.5 µl of DNA template, 0.2 µM of each primer, 1× NEB Next HIFI master mix were used in a 25 µl reaction. Cycling conditions included initial denaturation at 95 °C for five minutes, followed by 25 cycles of denaturation at 94 °C for 30 seconds, annealing at 55 °C for 30 seconds and extension at 72 °C for 30 seconds, then a final extension at 72 °C for five minutes.

3.7.3 Amplicon Library Indexing and Cleaning

Amplicons were cleaned using 0.6 volumes of AmpureXP beads (Beckman Coulter, USA) followed by a dual indexing PCR to allow multiplexing of samples. For this, a 50µl reaction consisting of 5µl of purified amplicons, 5µl of each NexteraXT i7 and i5 Index Primers (Illumina, USA), 25 µl of Neb Next High-Fidelity 2× PCR Master Mix (New England Bio Labs, MA, US) and 10µl of PCR grade water (ThermoFisher Scientific, CA, USA), with thermocycling at 95 °C for three minutes, followed by 12 cycles of 95 °C for 30 seconds, 55 °C for 30 seconds, and 72 °C for 30 seconds, and a final extension at 72°C for five minutes. The PCR amplification products were visualized by agarose gel electrophoresis; using 5 µl of the amplified reaction product in a 2% agarose gel, containing 3µl gel red and estimated in relation to size standard fragments 1kb DNA Ladder (GeneRuler™). Visual estimation of DNA concentration was also done on Agilent Tape station. In

case the amplification product was not visible in the agarose gel, the samples were not considered for pooling. Samples that were considered were cleaned using AmpureXP beads (Beckman Coulter, USA) using the 16S meta genomics bead-cleaning protocol. Samples were purified with 0.6 volumes of AmpureXP beads (Beckman Coulter, USA) and quantified by Qubit fluorometer (ThermoFisher Scientific).

3.7.4 Normalization and Pooling of Amplicon Libraries

All sequencing library PCR products were pooled in equimolar concentrations. This was done by pooling equal volumes of all products showing similar band intensity complemented by a pool for PCR without visible products on agarose gel. Eventually all pools were combined to a final sequencing library by adjusting the volume used from each pool according to its DNA concentration and number of samples combined in a pool. Final libraries were normalized to a concentration of 4 nM. DNA concentration in nM was calculated based on the size of DNA amplicons using the formula.

$$\frac{(\text{Concentration in ng/}\mu\text{l Post indexing}) \times 10^6}{(660 \text{ g/mol} \times \text{average library size})} = \text{concentration in nM} \quad \dots\dots\dots(1)$$

Volume to pool for each sample:

$$M = n/V \quad \dots\dots\dots(2)$$

Where **M** molarity, **n** number of molecules, **V** volume

3.7.5 Denaturing the Pooled Library and Sequencing

The indexed amplicon libraries were purified with AmpureXP beads according to the manufacturer’s instructions (Beckman Coulter Genomics, USA), and then quantified on Qubit fluorometer using Qubit dsDNA HS (high sensitivity) assay kit according to the manufacturer’s protocol (ThermoFisher Scientific, USA). The pooled amplicon library (PAL) was denatured using 1N NaOH, denatured libraries were diluted to a final concentration of 12 pM and spiked with 5% PhiX (Illumina, USA) as a sequencing control. The diluted amplicon library (DAL) was heat denatured at 95 °C for 2 minutes and sequenced on MiSeq platform (Illumina, USA) using the (301 × 2) bp sequencing chemistry on a MiSeq 600 cycles reagent kit v3 (Illumina, USA).

3.8 Bioinformatics Analysis

3.8.1 Haplotype Calling and Determination of Multiplicity of Infections

Haplotypes of *Pfmsp1* were determined using SeekDeep v2.6.0 (Hathaway *et al.*, 2018). Briefly, raw sequencing reads were filtered and trimmed based on the read length and quality scores using the *extractor* module in SeekDeep with the paired-end feature. After quality filtering, the reads were merged, chimeras removed and the sequences clustered at the sample level by *qluster*, and finally assembled based on the *mSP1* reference gene to generate *mSP1* haplotypes. The assembled haplotypes were analyzed by *processCluster* algorithm which compared sample haplotypes and generated individual and population-level haplotypes and statistics. A final mapping of all sequence reads to selected reference sequences was performed with the CLC Genomics workbench (CLC Inc, Aarhus, Denmark) and queried against the nucleotide database (GenBank) using the Nucleotide Basic Local Alignment Search Tool (BLASTn) (Altschul *et al.*, 1990). A haplotype was defined as a group of sequences within a cluster that represented the same allele of *Pfmsp1*. The MOI was calculated from the number of different alleles (K1, MAD20 and RO33) in the sample. The number of alleles for K1, MAD20 and RO33 was determined for each sample and the largest of these numbers was considered the MOI of that sample. The expected heterozygosity was calculated from the frequencies of the different alleles within the population according to the formula:

$$H_e = 1 - \sum(p_i^2) \dots\dots\dots(3)$$

where **H_e** is the expected heterozygosity, **p_i** is the frequency of the i-th allele in the population.

3.8.2 Statistical Analysis

GraphPad prism and R software were used for visualization and statistical analyses. Box plots comparing the identity between groups were created with GraphPad Prism 9 software, Paired sample t-test was used to compare parasite densities (Ct-values) in age groups. Sequence tables generated by SeekDeep were analyzed using R with the packages

Phyloseq v.1.22 (McMurdie and Holmes, 2013), and vegan v.2.5.2 (Oksanen *et al.*, 2020). A linear regression model was used to assess the relationship between MOI and parasite density. A p-value of <0.05 was considered statistically significant. Figures were generated using the following R packages: ggplot2 v.3.2.1 (Wickham *et al.*, 2016), ggthemes v.4.2.0 (Arnold, 2021), cowplot v.1.0.0 (Wilke, 2016) and viridis v.0.5.1 (Garnier *et al.*, 2018).

CHAPTER FOUR

RESULTS

4.1 Limit of Detection and Quality Control

The 18S rRNA qPCR limit of detection assay consistently detected *P. falciparum* to as low as 0.9 parasite/ μ l, corresponding to 7 parasites in 200 μ l whole blood or Ct = 38 (Figure 4.1).

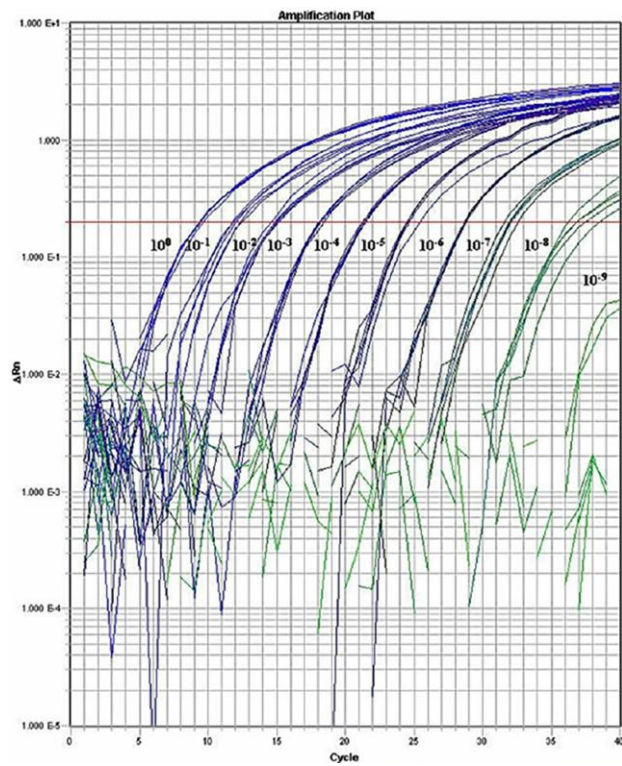


Figure 4.1: Amplification Plots to Determine Limit of Detection. Samples with High Parasitaemia were Captured by the Shown Threshold (red line). The Sigmoid Curves Represent Different Samples which as Visualized in AB 7500 Real-time PCR.

All samples with parasites detected by microscopy were also positive by 18S qPCR, by amplicon sequencing (AmpliSeq) of *Pfmsp1* the lowest parasite density that yielded usable *Pfmsp1* sequence was 9 parasites/ μ l.

The resolution of amplicons on 2% agarose gel electrophoresis gave the expected size of bp falling between 150 bp and 500 bp of 1Kb DNA ladder (Figure 4.2). Since bands of K1, MAD20 and RO33 of the positive controls (synchronized W2 parasites) have clear and uniform sizes corresponding to the expected band size, implies that the primers were specific to the target region of *Pfmsp1*.

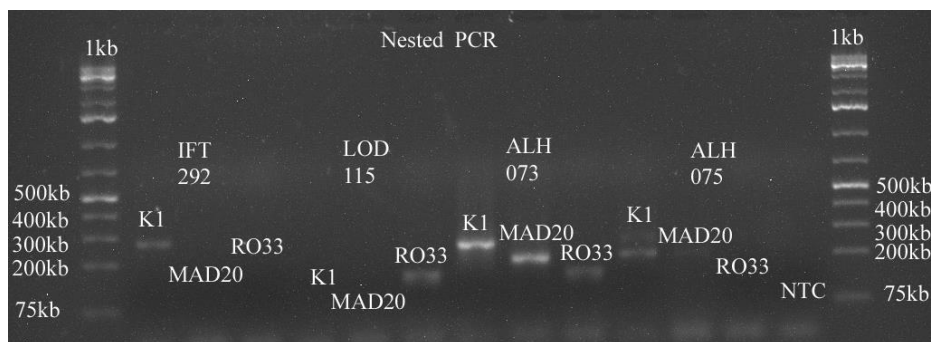


Figure 4.2: *P. falciparum* msp1 Block 2 Resolved on 2% Gel Red Stained Agarose gel. Gel electrophoresis image showing amplified *P. falciparum* controls using nested PCR primers for *Pfmsp1*. The presence of the bands showed successful optimization of the cycling conditions. NTC: Negative template control.

4.2 Demography and Malaria Parasitaemia in the Study Population

A total of 247 samples with *P. falciparum* parasitaemia of ≥ 9 parasites/ μ l were selected for inclusion in data analysis. 122 (49.4%) were from females, 125 (50.6%) were male and the median age was 7 years (interquartile range (IQR): 1-66). 83 (30%) were younger than 5 years, 99 (36%) between 5 and 14 years and 92 (34%) > 15 years. A total of 91

(33%) samples were from the malaria endemic Lake Victoria region, 48 (18%) from the epidemic prone highland's region and 108 (39%) arid regions that have seasonal malaria. qPCR Ct values were used as surrogate for malaria parasite density (Figure 4.3): children <5 years had higher parasitaemia (mean Ct =23.37, SD=±5) compared to 5 and 14 years (mean=25.52, SD=±5) and >15 years (mean=23.72, SD=±6) years). These differences were only significant for under 5 years and 5-to-14-year age groups (p=0.0245). In total, 1014 clones were found of which 212 in epidemic prone highland region of Kisii, 434 from endemic Lake Victoria region and 368 in the semi-arid regions. In the epidemic prone highland region of Kisii, the 212 clones found were classified into 3 different allele types (Table 4.1).

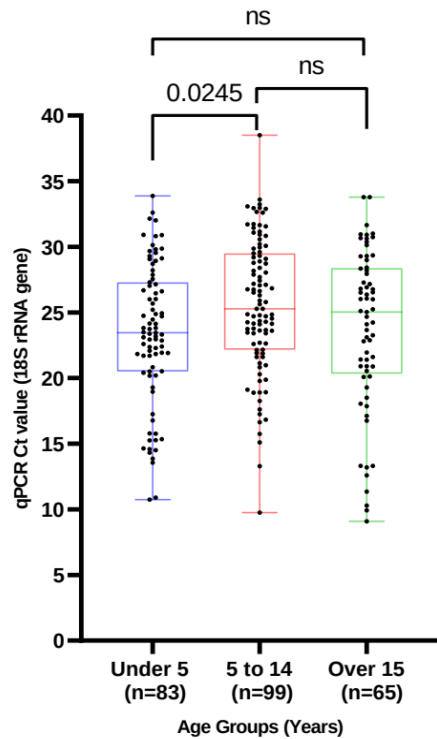


Figure 4.3: Scatter Plots Showing qPCR Ct Values in Different Age Groups. Children under 5 years had higher parasitaemia than older age groups.

4.3 *Pfmsp1* Genetic Diversity

4.3.1 Genetic Diversity by Fragment Size Polymorphisms

After quality filtering *Pfmsp1* sequences from 274 samples, 247 (84%) passed the Q30 scores. The mean number of reads per sample was 30,998 (range 487–49,983). Based on size, 1,014 alleles were obtained of which, 553 (54.5%) were K1, 250 (24.7%) were MAD20 and 211 (20.8%) were RO33 that grouped into 19 K1 allelic families (108-270 bp), 14 MAD20 (108-225 bp) and one RO33 (153 bp) (Figure 4.4).

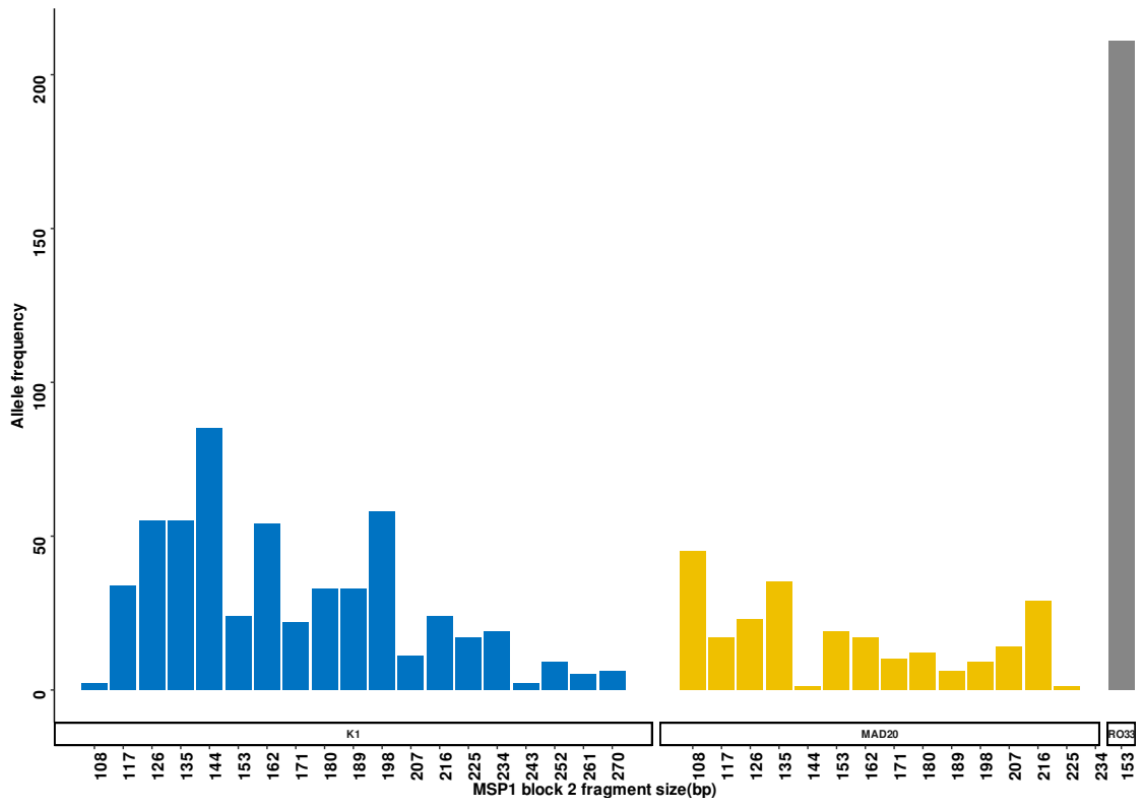


Figure 4.4: Distribution of *Pfmsp1* Allelic Families Based on of Target Alleles. In addition to internal sequence differences (nucleotide and SNPs), AmpliSeq described *Pfmsp1* population structure based on fragment sizes, bar chart plots showing 19 K1 (108-270 bp), 14 MAD20 (108-225 bp) and one RO33 (153 bp) allelic families.

4.3.2 Genetic Diversity by Nucleotide and Amino Acids Polymorphisms

Sequence analysis of the K1, MAD20 and RO33 alleles revealed nucleotide polymorphisms in alleles that had similar sizes. This increased the number of allelic families from 19 to 104 for K1, 14 to 58 for MAD20 and 1 to 14 for RO33. Most K1 diversity was due to duplications and deletions of the repeat amino acid motifs SGT and SGP (Appendix I); no synonymous nucleotide substitutions were observed in K1 repeat motifs. All 104 sequences of K1 were nonsynonymous (Figure 4.5), MAD20 sequences were represented essentially by different combinations of the amino acid motifs SGG, SVA, SVT, and SKG

(Appendix II). Most diversity can be explained by duplications and deletions of the motifs SGG, SVA and SVT. Synonymous nucleotide replacements were observed in the repeat motifs SGG and SVA in 18 out of 58 for MAD20 substitutions (Figure 4.6) and all the 14 substitutions for RO33 were nonsynonymous (Figure 4.7). The 14 RO33 block 2 sequences analyzed exhibited homology. However, six non-synonymous amino acid substitutions were frequently present in codons A63T, A79V, K89N, G90D, G96D and D103N. The distribution and frequency of the substitutions were not random and were highest in the first half of block 2 for K1, middle part for MAD20 and last 1/3 of block 2 for RO33. The K1-like, each allele had different numbers and arrangements of amino acid tripeptide repeat unit SGASAQSGT, SGT or SGPSGT. The MAD20-like were varied by the number of SGG SVA amino acid tripeptide motif. Meanwhile RO33-like was less polymorphic, with limited numbers of amino acid substitutions (Appendix III).

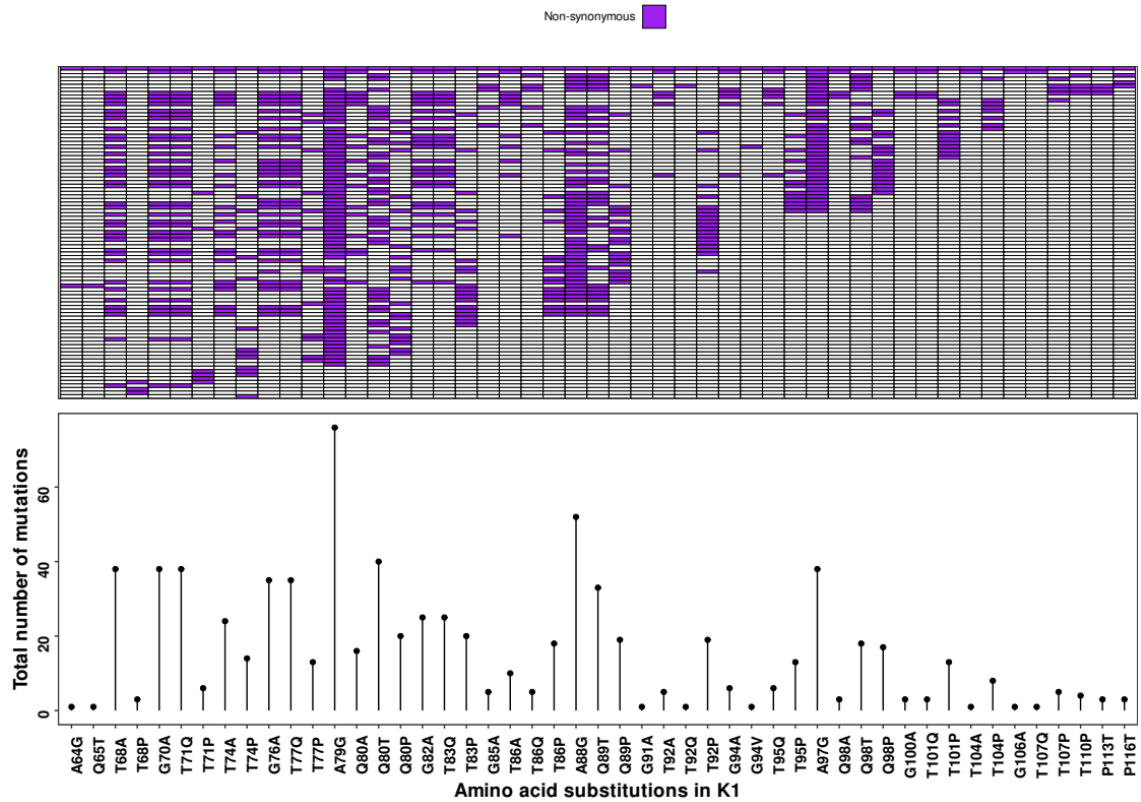


Figure 4.5: Frequencies of Amino Acid Substitutions Across Block 2 of K1

The data shows nonsynonymous amino acid substitutions for K1, the rows represents individual sequences, columns represent the amino acid substitutions. The lollipop plot shows the distribution and frequency of amino acid substitutions.

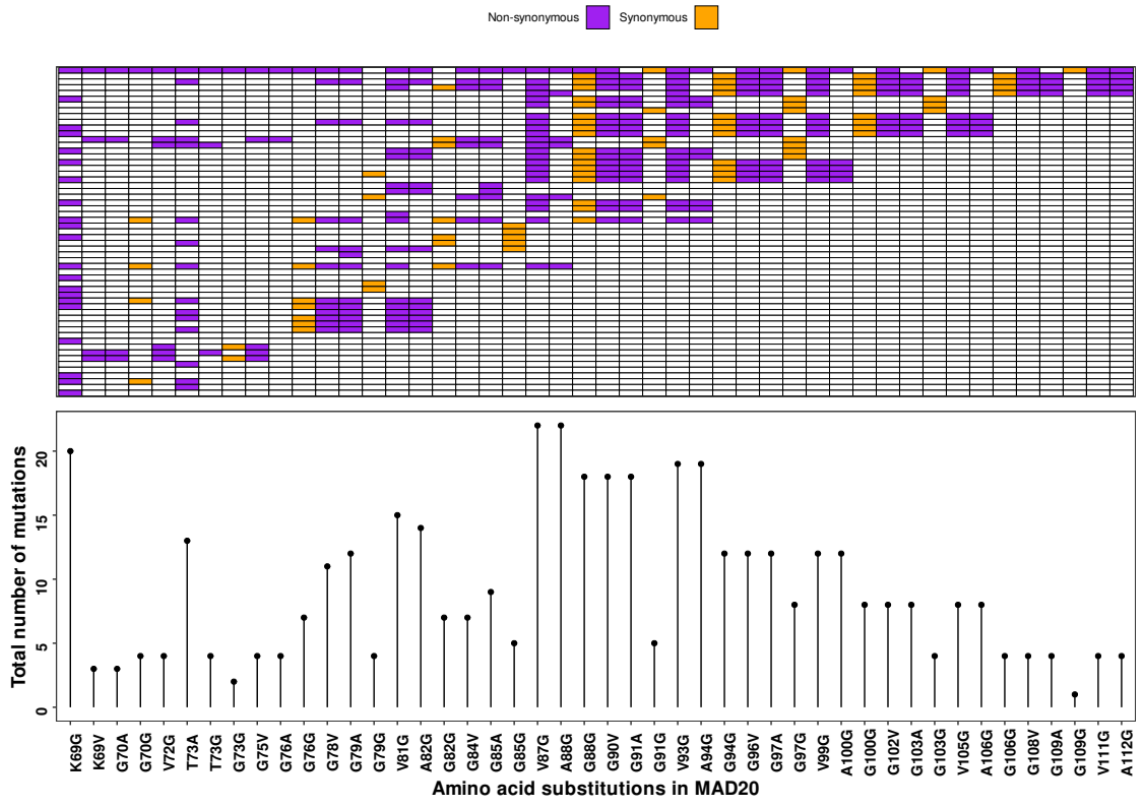


Figure 4.6: Frequencies of Amino Acid Substitutions Across Block 2 of MAD20

The data shows synonymous and nonsynonymous amino acid substitutions for MAD20. The rows represent individual sequences, columns represent the amino acid substitutions. The lollipop plots show the distribution and frequency of amino acid substitutions.

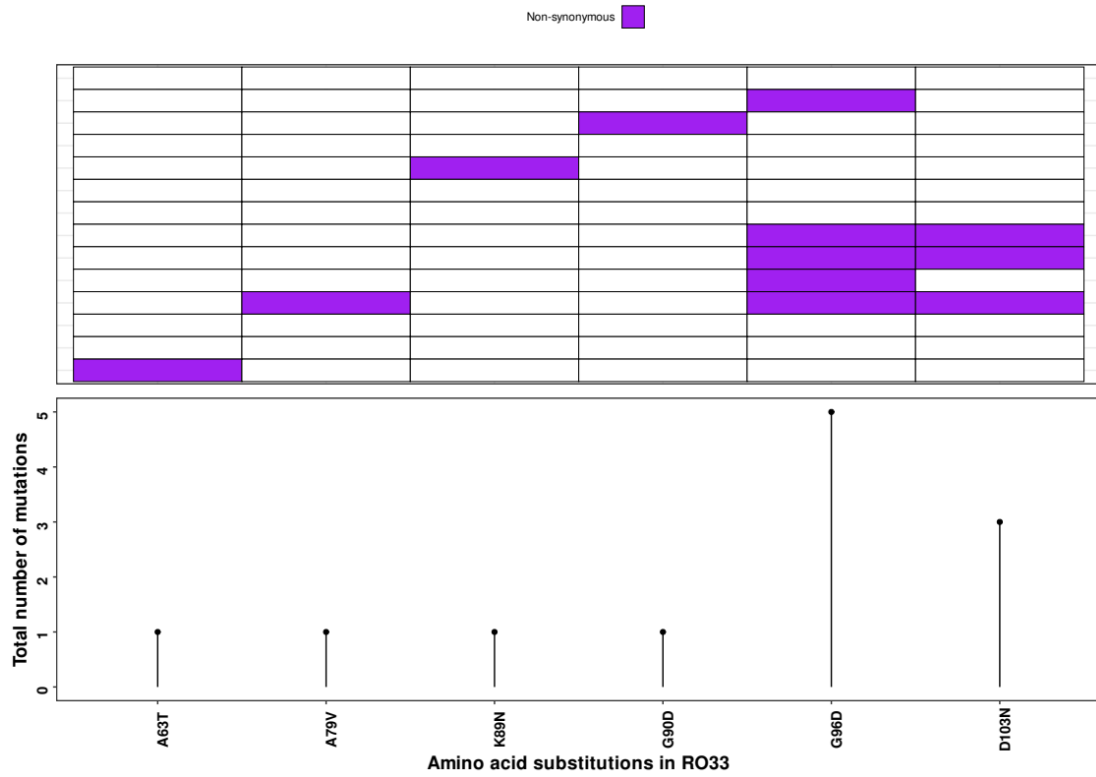


Figure 4.7: Frequencies of Amino Acid Substitutions Across Block 2 of RO33

The data shows nonsynonymous amino acid substitutions of RO33, the rows represent individual sequences, columns represent the amino acid substitutions. The lollipop plots show the distribution and frequency of amino acid substitutions.

4.4 Multiplicity of Infection

Polyclonal infections were detected in 56.3% of samples (139/247), the highest number of alleles detected in a single sample was 14 clones. *P. falciparum* isolates in this study had higher rates of multiple genotypes infection with an overall mean multiplicity of infection of 3.24 (95% CI 2.84–3.63). One thousand and fourteen (1014) sequences were obtained from Seek Deep. The K1 allelic family was the most frequent *msp1* genetic variant circulating in infected patients followed by MAD20 and RO33 representing; K1 (553 clones), MAD20 (250 clones) and RO33 (211 clones). The majority of these alleles especially for the K1 (55%) and MAD20 (24%) occurred at a high frequency, RO33 family (21%) was monomorphic with an amplified fragment size of 153 bp. Based on the genotyping results, 42.1% (108/247) were monoclonal infections, whereas polyclonal infections were most frequented with RO33 only and K1 only. MOI in children <5 was higher in younger age group <5 years (mean=4.5± 0.76, 95% CI) than the 5-14 years (mean=3.9±0.70, 95% CI) and those older >15 years (mean=2.7±0.90, 95% CI).

Table 4.1: *P. falciparum* Clonal Diversity by Age, Gender, and Malaria Endemicity

Variable	Age groups (years)			Gender		Malaria endemicity		
	<5	5-14	≥15	Female	Male	En- demic Lake Victoria region	Epi- demic prone high- lands of Kisii	Sea- sonal malaria arid re- gions
Haplotypes	372	389	252	516	497	434	212	367
Mean MOI	4.5	3.9	2.7	4.2	4.0	4.8	4.4	3.4

There was no distinct allelic pattern for a specific age group. One sample from malaria endemic Lake Victoria region carrying MAD20/RO33 hybrid allele (227bp) was not observed in other regions. The average number of alleles (shown as MOIs) were relatively stable in all the age groups, but slightly higher in the younger age groups (<5 years mean=4.5± 0.76, 95% CI) than the 5-14 years (3.9±0.70, 95% CI) and those older >15 years (2.7±0.90, 95% CI). Females had similar allele frequency to males (mean=4.2±0.66, 95% CI) compared to males (4.0±0.61, 95% CI). The average number of alleles in the malaria endemic Lake region (4.8±0.78, 95% CI) and the epidemic prone highland region (mean=4.4±1.03, 95% CI) were higher than in the seasonal malaria arid regions (mean=3.4 ±0.62, 95% CI). The expected heterozygosity (H_e) is a measure of the probability of infection by two parasites with different alleles at a given locus in all the regions was high (>0.98). MOI varied significantly with a few factors, including parasitaemia, in order to explore the relationship between MOI and parasitaemia, MOI from individual samples were compared to C_t values obtained by real-time qPCR from the samples successfully sequenced.

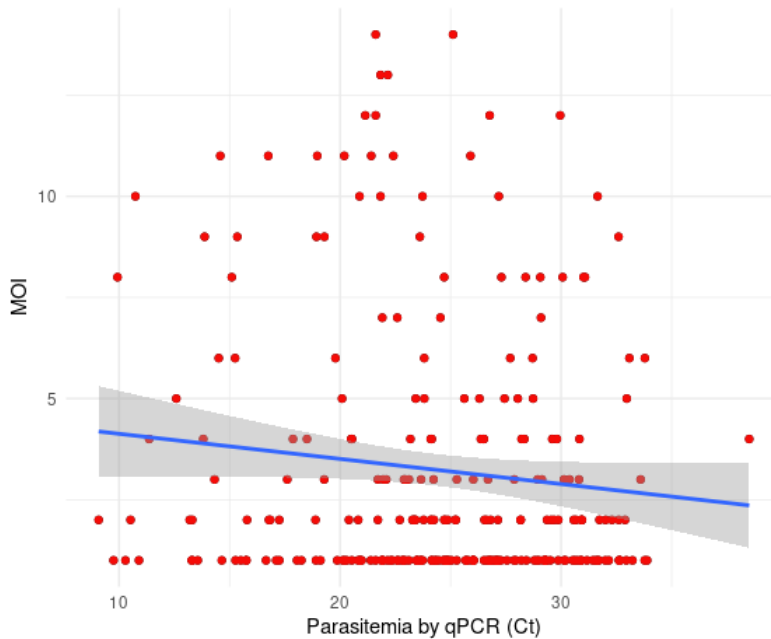


Figure 4.8: Linear Plot Showing Linear Relationship of MOI Against Ct Values. The relationship between the observed *P. falciparum* MOI from individual samples and parasitaemia by qPCR Ct values as surrogate.

There was a non-significant linear trend of increasing MOI with parasitaemia, but the overall variance was high with little accounted for by this model (Pearson coefficient of correlation, $r = -0.1105$, $p = 0.077$ (Figure 4.8)). A non-parametric Spearman rank correlation was also performed, and the results demonstrated no significant relationship between observed MOI and *P. falciparum* parasitaemia as measured by qPCR ($\rho = -0.0867$, $p = 0.165$). In general, the temporal distribution of alleles was least stable in the malaria epidemic prone highland region of Kisii compared to the endemic Lake Victoria region or the seasonal transmission arid region. Overall, alleles increased from the 2014-2015 period and with significant increase in the malaria endemic Lake Victoria region and the seasonal transmission arid region (Figure 4.9).

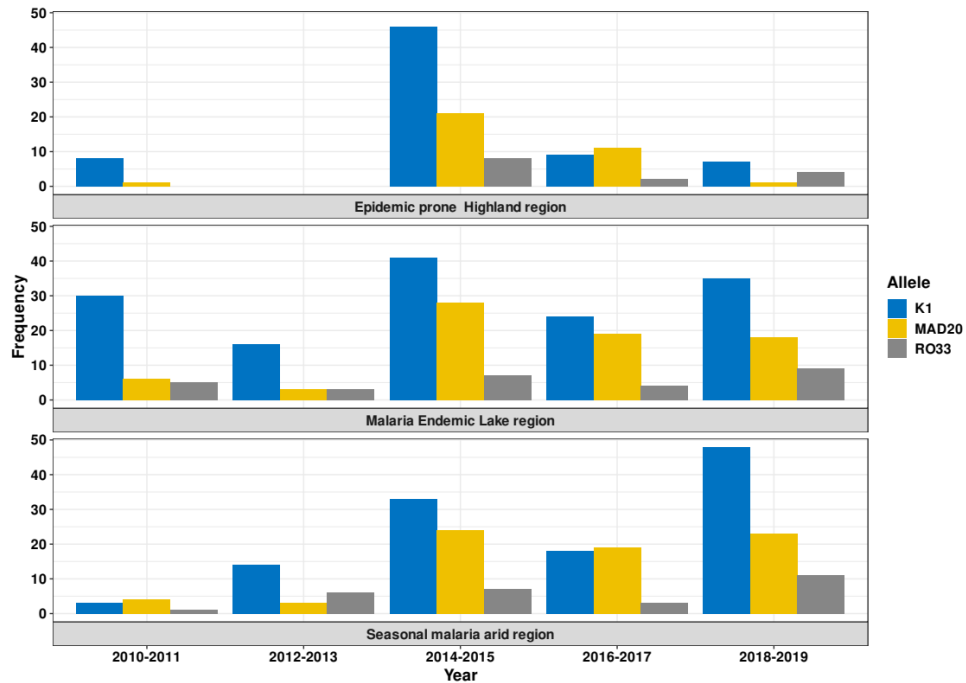


Figure 4.9: Temporal Variation in Allelic Families in Regions of Different Malaria Endemicity. Over time, allele distribution was least stable in the epidemic prone highland region of Kisii, compared to the malaria endemic Lake Victoria region or the seasonal malaria transmission arid region. In general, alleles frequency was low before the 2010-2014 period, and increased thereafter, more so in the Lake Victoria and the seasonal transmission arid regions.

CHAPTER FIVE

DISCUSSION

5.1 Performance of Deep Sequencing Assay

In this study, AmpliSeq of the highly polymorphic *msp1* gene was used to characterize the spatial and temporal allelic structure of *P. falciparum* in three regions of differing malaria endemicities. The choice of *Pfmsp1* was based on several factors. First, it is highly polymorphic, contains several SNPs, likely maintained via balancing selection due to immune pressure in the human host and furthermore previous studies in malaria endemic regions have identified over 60 polymorphic sites within *Pfmsp1* (Lin *et al.*, 2015; Boyce *et al.*, 2018). To demonstrate the applicability of AmpliSeq, the lowest parasitaemia density (LOD) that would give reliable data was determined using an 18S rRNA qPCR assay. 3D7 cultured ring stage parasites placed usable sequences for *msp1* AmpliSeq at a parasitaemia of about 10 parasites/ μ l. Using this parasitaemia cut-off, 264 samples with *P. falciparum* parasitaemia of 10 parasites/ μ l were evaluated. Children under 5 years had statistically significant higher malaria parasitaemia compared to those older than 5 years.

AmpliSeq that combines the size and internal sequence polymorphism improved the power to detect multi-clonal infections. *Pfmsp1* AmpliSeq generated 1,014 size alleles that mapped to K1 (54.5%), MAD20 (24.7%) and RO33 (20.8%) and grouped to 34 allelic families (19 K1, 14 MAD20 and one RO33). By including sequence polymorphisms internal to the sequences, the overall increase in the number of allelic families was by 5 \times (34 to 176), 5.5 \times for K1 (19 to 104), 4.1 \times for MAD20 (from 14 to 58) and 14 \times for RO33 (from 1 to 14). The use of size to deduce clonal multiplicity underestimates the number of clones in an infection. These findings corroborate previous studies that used AmpliSeq for estimating MOI (Juliano *et al.*, 2010; Koepfli and Mueller, 2017; Lerch *et al.*, 2017; Zhong *et al.*, 2018). As has been observed in other studies, K1 was the dominant allelic family convenient with previous findings (Takala *et al.*, 2002; Chenet *et al.*, 2008; Arieu *et al.*,

2014). This is unlike the RO33 that was reported as the dominant allele in parasites collected from Malaysia (Atroosh *et al.*, 2012), Brazil (Kimura *et al.*, 1990), and Gabon (Kun *et al.*, 1998) and unlike MAD20 allele that was the most prevalent in Myanmar (Snounou *et al.*, 1999; Kang *et al.*, 2010), Thailand (Snounou *et al.*, 1999), Iran (Zakeri *et al.*, 2005), Pakistan (Ghanchi *et al.*, 2010), and Colombia (Gómez *et al.*, 2002), Senegal (Niang *et al.*, 2017).

5.2 Genetic Structure of *P. falciparum* Populations

msp1 genotyping for the population genetic structure of *P. falciparum* revealed higher genetic diversity and with a high expected heterozygosity value (0.98). Both synonymous and nonsynonymous amino acids substitutions were identified across the *msp1* block 2. For K1 and RO33, only nonsynonymous substitutions were identified, while for MAD20, both synonymous and nonsynonymous substitutions were identified. The substitutions were not random: For K1, the substitutions were concentrated in the first half of block 2, middle part for MAD20 and last 1/3 for RO33. Previous studies have shown that most alleles fluctuate significantly over the years and can differ across endemic areas (Kiwuwa *et al.*, 2013; Yuan *et al.*, 2013). The present data suggest unequal allelic structure in the three areas of malaria endemicities. In general, the period before 2010 and up to end of 2013 was marked by lowest allele frequencies, and indirectly malaria prevalence. This period coincided with the introduction, adoption and widespread use of artemisinin-based combination therapy (ACT) following withdrawal of sulfadoxine/sulfalene-pyrimethamine (Amin *et al.*, 2007). The epidemic prone Kisii highland region had little or no malaria prior to 2014. Considering that the malaria cases seen in Kisii are largely introduced from the neighboring malaria endemic Lake Victoria region and, therefore, if there are low incidences of malaria in the latter region, there will be even fewer malaria cases in Kisii. Thereafter, there was a big burst in multi clonal infections in 2014-2015 in all the sites. Kisii region bust coincided with an outbreak of what was referred to as highland malaria (Shanks *et al.*, 2005). For unknown reasons, infections in Kisii declined to near zero by

2019. This is unlike in the malaria endemic Lake Victoria region where the alleles' distribution was more stable. In the arid region where malaria has seasonal distribution, the allelic structure has been on upward trajectory, and by 2018 to 2019, the distribution resembled the malaria endemic region. Further studies should be conducted to determine what the factors underlying the stabilization of malaria cases in the arid region and how much of the change is attributable to climate change.

5.3 Multiplicity of Infection and Genetic Diversity

In the current study, the results demonstrated that MOI in children <5 years of age was higher in younger age group <5 years (mean=4.5±0.76, 95% CI) than the 5-14 years (mean=3.9±0.70, 95% CI) and those older ≥15 years (mean=2.7±0.90, 95% CI). MOI was also influenced by malaria intensity, with higher in samples from malaria endemic Lake Victoria basin and the semi-arid region compared to the epidemic prone highlands of Kisii. The average number of alleles in the malaria endemic lake region (mean=4.8±0.78, 95% CI) and the epidemic prone highland region (mean=4.4±1.03, 95% CI) was higher than in the seasonal malaria arid regions (mean=3.4 ±0.62, 95% CI). The expected heterozygosity (He) was very high (0.98) and was independent of transmission pattern. Similar findings have been reported before (Mwingira *et al.*, 2011).

CHAPTER SIX

CONCLUSION AND RECOMMENDATIONS

6.1 Conclusion

This study adds 176 distinct allelic sequences to this database. The *Pfmsp1* antigen has been highly studied as a malaria vaccine candidate and to date, data has demonstrated that *mssl* based vaccine protection against clinical malaria is strain-specific and, therefore, a clear understanding of *mssl* diversity is critical to developing an effective malaria vaccine. Haplotype frequency was influenced by age, gender and transmission settings, highlighting the complexity of determinants of *P. falciparum* population structure. One of the limitations of the study is that the sampling was only possible in patients with parasitaemia cutoff of 10 parasites/ μ l, thus underrepresenting haplotypes from patients with low parasite density, thereby failing to capture the full diversity of haplotypes present in the population. There is probably no way to solve this shortcoming since the technology used is not sensitive enough to sequence very low parasite load. Nevertheless, the analytical depth of AmpliSeq give high confidence that the data obtained is robust and provide a credible overview of the *P. falciparum* population structure in the study populations and regions.

6.2 Recommendations

This study recommends that:

1. Since malaria is for the most part, endemic in developing countries with limited budgets and techniques for use in malaria research, amplicon sequencing would ideally be of great benefit to research groups who perform high-throughput genetic analyses with a high number of markers. It proved to be highly sensitive, specific and reproducible and it would be applicable in determining the genetic diversity, MOI's and population structure of *P. falciparum* populations.

2. On the lead up to malaria elimination goals in Kenya, elimination strategies need to be implemented indiscriminately in the different transmission settings to curtail possible parasite gene flow and subsequent malaria importation through human movement and effects of climate change.

REFERENCES

- Alam, M. T., De Souza, D. K., Vinayak, S., Griffing, S. M., Poe, A. C., Duah, N. O., ... & Koram, K. A. (2011). Selective sweeps and genetic lineages of *Plasmodium falciparum* drug-resistant alleles in Ghana. *Journal of Infectious Diseases*, 203(2), 220-227.
- Atroosh, W.M. Al-Mekhlafi HM, Mahdy MAK, Surin J. *et al.* (2012) 'The detection of pfcrt and pfmdr1 point mutations as molecular markers of chloroquine drug resistance, Pahang, Malaysia', *Malaria Journal*, 11, pp. 1–7.
- Boyce, R.M. Hathaway N, Fulton T, Reyes R, Matte M, Ntaro M, *et al.* (2018) Reuse of malaria rapid diagnostic tests for amplicon deep sequencing to estimate *Plasmodium falciparum* transmission intensity in western Uganda, *Scientific Reports*, 8(1), pp. 1–10.
- Brasil, P. Zalis, M. G., de Pina-Costa, A., Siqueira, A. M., Júnior, C. B., Silva, S., ... & Daniel-Ribeiro, C. T. (2017) Outbreak of human malaria caused by Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of molecular biology*, 215(3), 403-410.
- Amin, A. A., Zurovac, D., Kangwana, B. B., Greenfield, J., Otieno, D. N., Akhwale, W. S., & Snow, R. W. (2007). The challenges of changing national malaria drug policy to artemisinin-based combinations in Kenya. *Malaria journal*, 6, 1-11.
- Anderson, T. J., Haubold, B., Williams, J. T., Estrada-Franco §, J. G., Richardson, L., Mollinedo, R., ... & Day, K. P. (2000). Microsatellite markers reveal a spectrum of population structures in the malaria parasite *Plasmodium falciparum*. *Molecular biology and evolution*, 17(10), 1467-1482.
- Callahan, B. J., McMurdie, P. J., Rosen, M. J., Han, A. W., Johnson, A. J. A., & Holmes, S. P. (2016). DADA2: High-resolution sample inference from Illumina amplicon data. *Nature methods*, 13(7), 581-583.

- Castañeda-Mogollón, D., Toppings, N. B., Kamaliddin, C., Lang, R., Kuhn, S., & Pillai, D. R. (2023). Amplicon Deep Sequencing Reveals Multiple Genetic Events Lead to Treatment Failure with Atovaquone-Proguanil in *Plasmodium falciparum*. *Antimicrobial Agents and Chemotherapy*, e01709-22.
- Chang, H. H., Park, D. J., Galinsky, K. J., Schaffner, S. F., Ndiaye, D., Ndir, O., ... & Hartl, D. L. (2012). Genomic sequencing of *Plasmodium falciparum* malaria parasites from Senegal reveals the demographic history of the population. *Molecular biology and evolution*, 29(11), 3427-3439.
- Chenet, S. M., Branch, O. H., Escalante, A. A., Lucas, C. M., & Bacon, D. J. (2008). Genetic diversity of vaccine candidate antigens in *Plasmodium falciparum* isolates from the Amazon basin of Peru. *Malaria journal*, 7(1), 1-11.
- Contamin, H., Fandeur, T., Bonnefoy, S., Skouri, F., Ntoumi, F., & Mercereau-Puijalon, O. (1995). PCR typing of field isolates of *Plasmodium falciparum*. *Journal of clinical microbiology*, 33(4), 944-951.
- Doolan, D. L., Dobaño, C., & Baird, J. K. (2009). Acquired immunity to malaria. *Clinical microbiology reviews*, 22(1), 13-36.
- Duah, N. O., Matrevi, S. A., Quashie, N. B., Abuaku, B., & Koram, K. A. (2016). Genetic diversity of *Plasmodium falciparum* isolates from uncomplicated malaria cases in Ghana over a decade. *Parasites & vectors*, 9, 1-8.
- Dzikowski, R., & Deitsch, K. W. (2009). Genetics of antigenic variation in *Plasmodium falciparum*. *Current genetics*, 55, 103-110.
- Ferreira, M. U., da Silva Nunes, M., & Wunderlich, G. (2004). Antigenic diversity and immune evasion by malaria parasites. *Clinical and Vaccine Immunology*, 11(6), 987-995.

- Garnier, Simon, Ross, Noam, Rudis, Robert, Camargo, Pedro A, Sciaini, Marco, Scherer, C. (2018) ‘Package “viridis”’, *cran*.
- Gaur, D., Mayer, D. G., & Miller, L. H. (2004). Parasite ligand–host receptor interactions during invasion of erythrocytes by *Plasmodium* merozoites. *International journal for parasitology*, 34(13-14), 1413-1429.
- Ghanchi, N. K., Mårtensson, A., Ursing, J., Jafri, S., Bereczky, S., Hussain, R., & Beg, M. A. (2010). Genetic diversity among *Plasmodium falciparum* field isolates in Pakistan measured with PCR genotyping of the merozoite surface protein 1 and 2. *Malaria Journal*, 9(1), 1-6.
- Ghoshal, S., Gajendra, P., Datta Kanjilal, S., Mitra, M., & Sengupta, S. (2018). Diversity analysis of MSP1 identifies conserved epitope organization in block 2 amidst high sequence variability in Indian *Plasmodium falciparum* isolates. *Malaria Journal*, 17, 1-14.
- Gómez, D., Chaparro, J., Rubiano, C., Rojas, M. O., & Wasserman, M. (2002). Genetic diversity of *Plasmodium falciparum* field samples from an isolated Colombian village. *The American journal of tropical medicine and hygiene*, 67(6), 611-616.
- Gruenberg, M., Lerch, A., Beck, H. P., & Felger, I. (2019). Amplicon deep sequencing improves *Plasmodium falciparum* genotyping in clinical trials of antimalarial drugs. *Scientific Reports*, 9(1), 17790.
- Fulakeza, J., McNitt, S., Vareta, J., Saidi, A., Mvula, G., Taylor, T., ... & Seydel, K. (2019). Comparison of msp genotyping and a 24 SNP molecular assay for differentiating *Plasmodium falciparum* recrudescence from reinfection. *Malaria journal*, 18, 1-8.
- Hathaway, N. J., Parobek, C. M., Juliano, J. J., & Bailey, J. A. (2018). SeekDeep: single-base resolution de novo clustering for amplicon deep sequencing. *Nucleic acids research*, 46(4), e21-e21.

- Holder, A. A., Blackman, M. J., Burghaus, P. A., Chappel, J. A., Ling, I. T., McCallum-Deighton, N., & Shai, S. (1992). A malaria merozoite surface protein (MSP1)-structure, processing and function. *Memorias do Instituto Oswaldo Cruz*, 87, 37-42.
- Inouye, M., Dashnow, H., Raven, L. A., Schultz, M. B., Pope, B. J., Tomita, T., ... & Holt, K. E. (2014). SRST2: Rapid genomic surveillance for public health and hospital microbiology labs. *Genome medicine*, 6(11), 1-16.
- Oksanen, J., Kindt, R., Legendre, P., O'Hara, B., Stevens, M. H. H., Oksanen, M. J., & Suggests, M. A. S. S. (2007). The vegan package. *Community ecology package*, 10 (631–637), 719.
- Juliano, J. J., Porter, K., Mwapasa, V., Sem, R., Rogers, W. O., Ariey, F., ... & Meshnick, S. R. (2010). Exposing malaria in-host diversity and estimating population diversity by capture-recapture using massively parallel pyrosequencing. *Proceedings of the National Academy of Sciences*, 107(46), 20138-20143.
- Kang, J. M., Moon, S. U., Kim, J. Y., Cho, S. H., Lin, K., Sohn, W. M., ... & Na, B. K. (2010). Genetic polymorphism of merozoite surface protein-1 and merozoite surface protein-2 in *Plasmodium falciparum* field isolates from Myanmar. *Malaria journal*, 9(1), 1-8.
- Kenya Demographic and Health Survey. (2021) '2021 Kenya MALARIA Indicator Survey', *Kenya Demographic and Health Survey*.
- Kimura, E., Mattei, D., di Santi, S. M., & Scherf, A. (1990). Genetic diversity in the major merozoite surface antigen of *Plasmodium falciparum*: high prevalence of a third polymorphic form detected in strains derived from malaria patients. *Gene*, 91(1), 57-62.
- Kiwuwa, M. S., Ribacke, U., Moll, K., Byarugaba, J., Lundblom, K., Färnert, A., ... & Wahlgren, M. (2013). Genetic diversity of *Plasmodium falciparum* infections in mild and severe malaria of children from Kampala, Uganda. *Parasitology research*, 112, 1691-1700.

- Koepfli, C., & Mueller, I. (2017). Malaria epidemiology at the clone level. *Trends in parasitology*, 33(12), 974-985.
- Kun, J. F., Schmidt-Ott, R. J., Lehman, L. G., Lell, B., Luckner, D., Greve, B., ... & Kremsner, P. G. (1998). Merozoite surface antigen 1 and 2 genotypes and rosetting of *Plasmodium falciparum* in severe and mild malaria in Lambarene, Gabon. *Transactions of the Royal Society of Tropical Medicine and Hygiene*, 92(1), 110-114.
- Lambros, C., & Vanderberg, J. P. (1979). Synchronization of *Plasmodium falciparum* erythrocytic stages in culture. *The Journal of parasitology*, 418-420.
- Lerch, A., Koepfli, C., Hofmann, N. E., Messerli, C., Wilcox, S., Kattenberg, J. H., ... & Felger, I. (2017). Development of amplicon deep sequencing markers and data analysis pipeline for genotyping multi-clonal malaria infections. *BMC genomics*, 18, 1-13.
- Lerch, A. (2018). Methods for analysis of deep sequencing data from mixtures of *Plasmodium falciparum* clones or stage-specific transcriptomes. Unpublished, PhD Thesis, Basel: University_of_Basel.
- Lerch, A., Koepfli, C., Hofmann, N. E., Kattenberg, J. H., Rosanas-Urgell, A., Betuela, I., ... & Felger, I. (2019). Longitudinal tracking and quantification of individual *Plasmodium falciparum* clones in complex infections. *Scientific reports*, 9(1), 3333.
- Liljander, A., Wiklund, L., Falk, N., Kweku, M., Mårtensson, A., Felger, I., & Färnert, A. (2009). Optimization and validation of multi-coloured capillary electrophoresis for genotyping of *Plasmodium falciparum* merozoite surface proteins (msp1 and 2). *Malaria journal*, 8(1), 1-14.
- Lin, J. T., Hathaway, N. J., Saunders, D. L., Lon, C., Balasubramanian, S., Kharabora, O., ... & Juliano, J. J. (2015). Using amplicon deep sequencing to detect genetic signatures of *Plasmodium vivax* relapse. *The Journal of infectious diseases*, 212(6), 999-1008.

- Mahdi Abdel Hamid, M., Elamin, A. F., Albsheer, M. M. A., Abdalla, A. A., Mahgoub, N. S., Mustafa, S. O., ... & Amin, M. (2016). Multiplicity of infection and genetic diversity of *Plasmodium falciparum* isolates from patients with uncomplicated and severe malaria in Gezira State, Sudan. *Parasites & vectors*, 9, 1-8.
- Muhindo Mavoko, H., Kalabuanga, M., Delgado-Ratto, C., Maketa, V., Mukele, R., Fungula, B., ... & Van Geertruyden, J. P. (2016). Uncomplicated clinical malaria features, the efficacy of artesunate-amodiaquine and their relation with multiplicity of infection in the Democratic Republic of Congo. *PLoS One*, 11(6), e0157074.
- McMurdie, P. J., & Holmes, S. (2013). phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. *PloS one*, 8(4), e61217.
- Miller, L. H., Roberts, T., Shahabuddin, M., & McCutchan, T. F. (1993). Analysis of sequence diversity in the *Plasmodium falciparum* merozoite surface protein-1 (MSP-1). *Molecular and biochemical parasitology*, 59(1), 1-14.
- Miller, R. H., Hathaway, N. J., Kharabora, O., Mwandagaliwa, K., Tshefu, A., Meshnick, S. R., ... & Bailey, J. A. (2017). A deep sequencing approach to estimate *Plasmodium falciparum* complexity of infection (COI) and explore apical membrane antigen 1 diversity. *Malaria Journal*, 16(1), 1-15.
- Murphy, S. C., Prentice, J. L., Williamson, K., Wallis, C. K., Fang, F. C., Fried, M., ... & Cookson, B. T. (2012). Real-time quantitative reverse transcription PCR for monitoring of blood-stage *Plasmodium falciparum* infections in malaria human challenge trials. *The American journal of tropical medicine and hygiene*, 86(3), 383.
- Mwingira, F., Nkwengulila, G., Schoepflin, S., Sumari, D., Beck, H. P., Snounou, G., ... & Muggitu, K. (2011). *Plasmodium falciparum* msp1, msp2 and glurp allele frequency and diversity in sub-Saharan Africa. *Malaria journal*, 10, 1-10.

- Niang, M., Thiam, L. G., Loucoubar, C., Sow, A., Sadio, B. D., Diallo, M., ... & Toure-Balde, A. (2017). Spatio-temporal analysis of the genetic diversity and complexity of *Plasmodium falciparum* infections in Kedougou, southeastern Senegal. *Parasites & vectors*, *10*, 1-9.
- Nielsen, C. M., Vekemans, J., Lievens, M., Kester, K. E., Regules, J. A., & Ockenhouse, C. F. (2018). RTS, S malaria vaccine efficacy and immunogenicity during *Plasmodium falciparum* challenge is associated with HLA genotype. *Vaccine*, *36*(12), 1637-1642.
- Oduola, A. M., Milhous, W. K., Weatherly, N. F., Bowdre, J. H., & Desjardins, R. E. (1988). *Plasmodium falciparum*: induction of resistance to mefloquine in cloned strains by continuous drug exposure in vitro. *Experimental parasitology*, *67*(2), 354-360.
- Okara, R. M., Sinka, M. E., Minakawa, N., Mbogo, C. M., Hay, S. I., & Snow, R. W. (2010). Distribution of the main malaria vectors in Kenya. *Malaria journal*, *9*(1), 1-11.
- Otsyula, N., Angov, E., Bergmann-Leitner, E., Koech, M., Khan, F., Bennett, J., ... & Spring, M. D. (2013). Results from tandem phase 1 studies evaluating the safety, reactogenicity and immunogenicity of the vaccine candidate antigen *Plasmodium falciparum* FVO merozoite surface protein-1 (MSP142) administered intramuscularly with adjuvant system AS01. *Malaria journal*, *12*(1), 1-13.
- Parobek, C. M., Bailey, J. A., Hathaway, N. J., Socheat, D., Rogers, W. O., & Juliano, J. J. (2014). Differing patterns of selection and geospatial genetic diversity within two leading *Plasmodium vivax* candidate vaccine antigens. *PLoS Neglected Tropical Diseases*, *8*(4), e2796.
- Polley, S. D., & Conway, D. J. (2001). Strong diversifying selection on domains of the *Plasmodium falciparum* apical membrane antigen 1 gene. *Genetics*, *158*(4), 1505-1512.
- Raja, T. N., Hu, T. H., Kadir, K. A., Mohamad, D. S. A., Rosli, N., Wong, L. L., ... & Singh, B. (2020). Naturally acquired human *Plasmodium cynomolgi* and *P. knowlesi* infections, Malaysian Borneo. *Emerging infectious diseases*, *26*(8), 1801.

- Reeder, J. C., & Brown, G. V. (1996). Antigenic variation and immune evasion in *Plasmodium falciparum* malaria. *Immunology and Cell Biology*, 74(6), 546-554.
- Rodrigues, C. D., Hannus, M., Prudêncio, M., Martin, C., Gonçalves, L. A., Portugal, S., ... & Mota, M. M. (2008). Host scavenger receptor SR-BI plays a dual role in the establishment of malaria parasite liver infection. *Cell host & microbe*, 4(3), 271-282.
- Shanks, G. D., Biomndo, K., Guyatt, H. L., & Snow, R. W. (2005). Travel as a risk factor for uncomplicated *Plasmodium falciparum* malaria in the highlands of western Kenya. *Transactions of the Royal Society of Tropical Medicine and Hygiene*, 99(1), 71-74.
- Snounou, G., Zhu, X., Siripoon, N., Jarra, W., Thaithong, S., Brown, K. N., & Viriyakosol, S. (1999). Biased distribution of msp1 and msp2 allelic variants in *Plasmodium falciparum* populations in Thailand. *Transactions of the Royal Society of Tropical Medicine and Hygiene*, 93(4), 369-374.
- Takala, S., Branch, O., Escalante, A. A., Kariuki, S., Wootton, J., & Lal, A. A. (2002). Evidence for intragenic recombination in *Plasmodium falciparum*: identification of a novel allele family in block 2 of merozoite surface protein-1: Asembo Bay Area Cohort Project XIV. *Molecular and biochemical parasitology*, 125(1-2), 163-171.
- Tanabe, K., Mackay, M., Goman, M., & Scaife, J. G. (1987). Allelic dimorphism in a surface antigen gene of the malaria parasite *Plasmodium falciparum*. *Journal of molecular biology*, 195(2), 273-287.
- Touray, A. O., Mobegi, V. A., Wamunyokoli, F., & Herren, J. K. (2020). Diversity and Multiplicity of *P. falciparum* infections among asymptomatic school children in Mbita, Western Kenya. *Scientific reports*, 10(1), 5924.
- Trager, W., & Jensen, J. B. (1976). Human malaria parasites in continuous culture. *Science*, 193(4254), 673-675.

- Trape, J. F., Rogier, C., Konate, L., Diagne, N., Bouganali, H., Canque, B., ... & Silva, L. P. D. (1994). The Dielmo project: a longitudinal study of natural malaria infection and the mechanisms of protective immunity in a community living in a holoendemic area of Senegal. *American journal of tropical medicine and hygiene*, 51(2), 123-137.
- Oksanen, J., Kindt, R., Legendre, P., O'Hara, B., Stevens, M. H. H., Oksanen, M. J., & Suggests, M. A. S. S. (2007). The vegan package. *Community ecology package*, 10(631-637), 719.
- Waitumbi, J. N., Anyona, S. B., Hunja, C. W., Kifude, C. M., Polhemus, M. E., Walsh, D. S., ... & Sutherland, C. J. (2009). Impact of RTS, S/AS02A and RTS, S/AS01B on genotypes of *P. falciparum* in adults participating in a malaria vaccine clinical trial. *PLoS One*, 4(11), e7849.
- Wampfler, R., Mwingira, F., Javati, S., Robinson, L., Betuela, I., Siba, P., ... & Felger, I. (2013). Strategies for detection of *Plasmodium* species gametocytes. *PloS one*, 8(9), e76316.
- World Health Organization. World malaria report 2021. Geneva, Switzerland: World Health Organization; 2021.
- World Health Organization. World malaria report 2023. Geneva, Switzerland: World Health Organization; 2023.
- Wickham, H. (2011). ggplot2. *Wiley interdisciplinary reviews: computational statistics*, 3(2), 180-185.
- Wilke, C. O., Wickham, H., & Wilke, M. C. O. (2019). Package 'cowplot'. *Streamlined plot theme and plot annotations for 'ggplot2', 1*.
- Yavo, W., Konaté, A., Mawili-Mboumba, D. P., Kassi, F. K., Tshibola Mbuyi, M. L., Angora, E. K., ... & Bouyou-Akotet, M. K. (2016). Genetic polymorphism of *msp1* and *msp2* in *Plasmodium falciparum* isolates from Côte d'Ivoire versus Gabon. *Journal of parasitology research*, 2016.

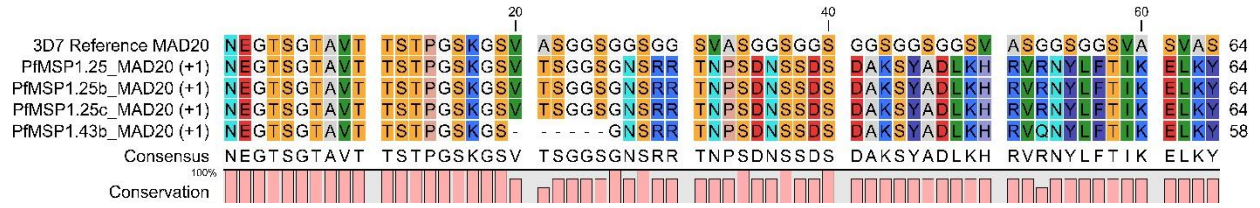
- Yuan, L., Zhao, H., Wu, L., Li, X., Parker, D., Xu, S., ... & Cui, L. (2013). *Plasmodium falciparum* populations from northeastern Myanmar display high levels of genetic diversity at multiple antigenic loci. *Acta tropica*, *125*(1), 53-59.
- Zakeri, S., Bereczky, S., Naimi, P., Pedro Gil, J., Djadid, N. D., Färnert, A., ... & Björkman, A. (2005). Multiple genotypes of the merozoite surface proteins 1 and 2 in *Plasmodium falciparum* infections in a hypoendemic area in Iran. *Tropical Medicine & International Health*, *10*(10), 1060-1064.
- Zhong, D., Lo, E., Wang, X., Yewhalaw, D., Zhou, G., Atieli, H. E., ... & Yan, G. (2018). Multiplicity and molecular epidemiology of *Plasmodium vivax* and *Plasmodium falciparum* infections in East Africa. *Malaria journal*, *17*, 1-14.

APPENDICES

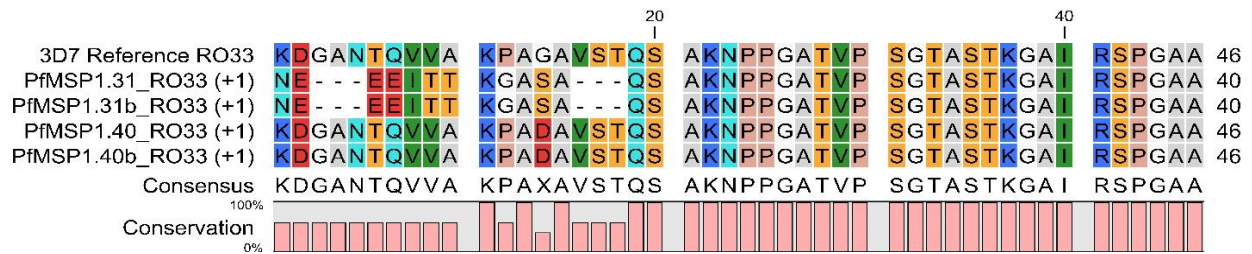
Appendix I: Sequence Alignment of K1 Alleles of *msp1* Block 2.

	20	40	60	80							
3D7 Reference K1	EEITTKGASA	QSGTSGTSGT	SGPS	-----	-----GP	SGPSGTSPSS	RSNTLPRSNT	SSGASPPADA	56		
PiMSP1.00_K1(+1)	EEITTKGASA	QSGT	-----	-----	-----	SGTSPSS	RSNTLPRSNT	SSGASPPADA	41		
PiMSP1.01_K1(+1)	EEITTKGASA	QSGT	-----	-----	-----	SGTSPSS	RSNTLPRSNT	SSGASPPADA	44		
PiMSP1.02_K1(+1)	EEITTKGASA	QSG	-----	-----	-----	TSPSS	RSNTLPRSNT	SSGASPPADA	38		
PiMSP1.03_K1(+1)	EEITTKGASA	QSGT	-----	-----	-----	GT	SGPSGTSPSS	RSNTLPRSNT	47		
PiMSP1.04_K1(+1)	EEITTKGASA	QSGT	-----	-----	-----	GP	SGPSGTSPSS	RSNTLPRSNT	47		
PiMSP1.05_K1(+1)	EEITTKGASA	QSGT	-----	-----	-----	GT	SGTSGTSPSS	RSNTLPRSNT	47		
PiMSP1.06_K1(+1)	EEITTKGASA	QSGTSGTSGT	S	-----	-----	GP	SGPSGTSPSS	RSNTLPRSNT	53		
PiMSP1.07_K1(+1)	EEITTKGASA	QSGTSGTSGP	S	-----	-----	GTSGPS	GTSGPS	RSNTLPRSNT	62		
PiMSP1.08_K1(+1)	EEITTKGASA	QSGASAQSGA	SAQSGASAQ	GT	-----	GP	SGPSGTSPSS	RSNTLPRSNT	65		
PiMSP1.09_K1(+1)	EEITTKGASA	QSGT	-----	-----	-----	GP	SGPSGTSPSS	RSNTLPRSNT	44		
PiMSP1.10_K1(+1)	EEITTKGASA	QSGTSGTSG	-----	-----	-----	P	SGPSGTSPSS	RSNTLPRSNT	50		
PiMSP1.11_K1(+1)	EEITTKGASA	QSG	-----	-----	-----	PSGTSPSS	RSNTLPRSNT	SSGASPPADA	41		
PiMSP1.12_K1(+1)	EEITTKGASA	QSGTSGTSGT	S	-----	-----	GT	SGPSGTSPSS	RSNTLPRSNT	53		
PiMSP1.13_K1(+1)	EEITTKGASA	QSGTSGTSGT	S	-----	-----	GT	SGTSGTSPSS	RSNTLPRSNT	53		
PiMSP1.14_K1(+1)	EEITTKGASA	QSGTSGTSGT	S	-----	-----	PSGTSPSS	RSNTLPRSNT	SSGASPPADA	50		
PiMSP1.15_K1(+1)	EEITTKGASA	QSGASAQSGA	SAQSGT	-----	-----	GP	SGPSGTSPSS	RSNTLPRSNT	59		
PiMSP1.16_K1(+1)	EEITTKGASA	QSG	-----	-----	-----	P	SGPSGTSPSS	RSNTLPRSNT	44		
PiMSP1.17_K1(+1)	EEITTKGASA	QSGT	-----	-----	-----	GP	SGPSGTSPSS	RSNTLPRSNT	47		
PiMSP1.18_K1(+1)	EEITTKGASA	QSGTSGTSGT	S	-----	-----	GTSGTSPSS	RSNTLPRSNT	SSGASPPADA	50		
PiMSP1.19_K1(+1)	EEITTKGASA	QSGTSGTSGP	S	-----	-----	GT	SGPSGTSPSS	RSNTLPRSNT	53		
PiMSP1.20_K1(+1)	EEITTKGASA	QSGASAQSGT	S	-----	-----	GP	SGPSGTSPSS	RSNTLPRSNT	53		
PiMSP1.21_K1(+1)	EEITTKGASA	QSGT	-----	-----	-----	GP	SGPSGTSPSS	RSNTLPRSNT	56		
PiMSP1.22_K1(+1)	EEITTKGASA	QSGT	-----	-----	-----	GT	SGPSGTSPSS	RSNTLPRSNT	56		
PiMSP1.23_K1(+1)	EEITTKGASA	QSGASAQSGT	SAQSGTSAQ	GT	-----	GT	SGPSGTSPSS	RSNTLPRSNT	65		
PiMSP1.24_K1(+1)	EEITTKGASA	QSGASAQSGA	SAQSGASAQ	GASAQSGT	-----	GP	SGPSGTSPSS	RSNTLPRSNT	59		
PiMSP1.26_K1(+1)	EEITTKGASA	QSGASAQSGT	SGTSGT	-----	-----	GP	SGPSGTSPSS	RSNTLPRSNT	71		
PiMSP1.27_K1(+1)	EEITTKGASA	QSGTSGTSGT	S	-----	-----	T	SGPSGTSPSS	RSNTLPRSNT	65		
PiMSP1.28_K1(+1)	EEITTKGASA	QSGASAQ	-----	-----	-----	SGTSPSS	RSNTLPRSNT	SSGASPPADA	44		
PiMSP1.29_K1(+1)	EEITTKGASA	QSGTSGP	-----	-----	-----	GP	SGPSGTSPSS	RSNTLPRSNT	65		
PiMSP1.30_K1(+1)	EEITTKGASA	QSGTSGTSGT	SAQSGTSGT	GTSGTSG	-----	T	SGPSGTSPSS	RSNTLPRSNT	68		
PiMSP1.32_K1(+1)	EEITTKGASA	QSGASAQSGT	SGTSGTSGT	GT	-----	GP	SGPSGTSPSS	RSNTLPRSNT	55		
PiMSP1.33_K1(+1)	EEITTKGASA	QSGTSGTSGT	S	-----	-----	GT	SAQSGTSPSS	RSNTLPRSNT	53		
PiMSP1.34_K1(+1)	EEITTKGASA	QSGT	-----	-----	-----	GP	SGPSGTSPSS	RSNTLPRSNT	56		
PiMSP1.35_K1(+1)	EEITTKGASA	QSG	-----	-----	-----	P	SGPSGTSPSS	RSNTLPRSNT	44		
PiMSP1.36_K1(+1)	EEITTKGASA	QSGASAQSGA	SAQSGASAQ	GT	-----	GP	SGPSGTSPSS	RSNTLPRSNT	56		
PiMSP1.38_K1(+1)	EEITTKGASA	QSGTSGTSGT	SGTSGTSAQ	-----	-----	SGTSPSS	RSNTLPRSNT	SSGASPPADA	56		
PiMSP1.39_K1(+1)	EEITTKGASA	QSGTSGTSGT	S	-----	-----	SGTSPSS	RSNTLPRSNT	SSGASPPADA	56		
PiMSP1.41_K1(+1)	EEITTKGASA	QSGTSGTSGT	S	-----	-----	PSGTSPSS	RSNTLPRSNT	SSGASPPADA	59		
PiMSP1.42_K1(+1)	EEITTKGASA	QSGT	-----	-----	-----	GT	SGPSGTSPSS	RSNTLPRSNT	41		
PiMSP1.44_K1(+1)	EEITTKGASA	QSGASAQSGA	SAQSGASAQ	GASAQSGTSG	TSGTSGP	SGPSGTSPSS	RSNTLPRSNT	SSGASPPADA	80		
PiMSP1.45_K1(+1)	EEITTKGASA	QSGTSGTSGT	S	-----	-----	GTSGTSPSS	RSNTLPRSNT	SSGASPPADA	62		
PiMSP1.46_K1(+1)	EEITTKGASA	QSGASAQSGA	SAQSGTSGT	GT	-----	GP	SGPSGTSPSS	RSNTLPRSNT	65		
PiMSP1.47_K1(+1)	EEITTKGASA	QSGASAQSGA	SAQSGTSGP	GP	-----	GP	SGPSGTSPSS	RSNTLPRSNT	65		
PiMSP1.48_K1(+1)	EEITTKGASA	QSGTSGTSGT	SAQSGTSGT	GTSGTSGT	-----	SGT	SGPSGTSPSS	RSNTLPRSNT	71		
PiMSP1.49_K1(+1)	EEITTKGASA	QSGTSGTSGT	S	-----	-----	T	SGPSGTSPSS	RSNTLPRSNT	65		
PiMSP1.50_K1(+1)	EEITTKGASA	QSGASAQSGA	SAQSGASAQ	GT	-----	GT	SGPSGTSPSS	RSNTLPRSNT	65		
PiMSP1.51_K1(+1)	EEITTKGASA	QSGTSGTSGT	S	-----	-----	SGTSPSS	RSNTLPRSNT	SSGASPPADA	56		
PiMSP1.52_K1(+1)	EEITTKGASA	QSGTSGTSGT	S	-----	-----	GP	SGPSGTSPSS	RSNTLPRSNT	68		
PiMSP1.53_K1(+1)	EEITTKGASA	QSGASAQSGT	SAQSGTSAQ	GTSGTSGT	-----	GP	SGPSGTSPSS	RSNTLPRSNT	71		
PiMSP1.54_K1(+1)	EEITTKGASA	QSGASAQSGA	SAQSGAGT	GT	-----	GP	SGPSGTSPSS	RSNTLPRSNT	65		
PiMSP1.55_K1(+1)	EEITTKGASA	QSGTSGTSGP	S	-----	-----	GP	SGPSGTSPSS	RSNTLPRSNT	68		
PiMSP1.56_K1(+1)	EEITTKGASA	QSGASAQSGT	SAQSGTSAQ	GTSAQSGT	-----	SGT	SGPSGTSPSS	RSNTLPRSNT	71		
PiMSP1.57_K1(+1)	EEITTKGASA	QSGASAQSGT	SGTSGTSGT	GTSGTSGTSG	T	GP	SGPSGTSPSS	RSNTLPRSNT	74		
PiMSP1.58_K1(+1)	EEITTKGASA	QSGT	-----	-----	-----	S	PS	RSNTLPRSNT	38		
PiMSP1.59_K1(+1)	EEITTKGASA	QSGASAQSGT	SAQSGTSAQ	GTSGTSGP	SGP	SGPSGTSPSS	RSNTLPRSNT	SSGASPPADA	77		
PiMSP1.60_K1(+1)	EEITTKGASA	QSGASAQSGT	SAQSGTSAQ	GTSAQSGTSG	T	SGT	SGPSGTSPSS	RSNTLPRSNT	74		
PiMSP1.61_K1(+1)	EEITTKGASA	QSGT	-----	-----	-----	SGTSPSS	RSNTLPRSNT	SSGASPPADA	41		
PiMSP1.62_K1(+1)	EEITTKGASA	QSGASAQSGA	SAQSGP	-----	-----	GP	SGPSGTSPSS	RSNTLPRSNT	59		
Consensus	EEITTKGASA	QSGTSGTSGT	S	---G-S-	S	G	-----	-----GP	SGPSGTSPSS	RSNTLPRSNT	SSGASPPADA

Appendix II: Sequence Alignment of MAD20 Alleles of *msp1* Block 2.



Appendix III: Sequence Alignment of RO33 Alleles of *msp1* Block 2.



Appendix IV: Ethical Approval



KENYA MEDICAL RESEARCH INSTITUTE

P.O. Box 54840-00200, NAIROBI, Kenya
Tel: (254) (020) 2722541, 2713349, 0722-205901, 0733-400003, Fax: (254) (020) 2720030
E-mail: director@kemri.org, info@kemri.org, Website: www.kemri.org

KEMRI/RES/7/3/1

August 01, 2018

TO: **DR. JOHN WAITUMBI,**
PRINCIPAL INVESTIGATOR.

THROUGH: **THE DIRECTOR, CCR,**
NAIROBI.

Dear Sir,

V. S. 02/08/2018

RE: **SSC PROTOCOL No. 1282 (REQUEST FOR ANNUAL RENEWAL): ACUTE FEBRILE ILLNESS SURVEILLANCE IN KENYA.**

Thank you for the continuing review report for the period **July 01, 2017 to June 30, 2018.**

This is to inform you that the Expedited Review Team of the KEMRI Scientific and Ethics Review Unit (SERU) was of the informed opinion that the progress made during the reported period is satisfactory. The study has therefore been granted **approval**.

This approval is valid from **August 18, 2018** through to **August 17, 2019**. Please note that authorization to conduct this study will automatically expire on **August 17, 2019**. If you plan to continue with data collection or analysis beyond this date please submit an application for continuing approval to the **SERU** by **July 20, 2019**.

You are required to submit any amendments to this protocol and other information pertinent to human participation in this study to the SERU for review prior to initiation.

Yours faithfully,

FOR 
THE HEAD,
KEMRI SCIENTIFIC AND ETHICS REVIEW UNIT.

MEMORANDUM FOR Director, Human Subjects Protection Branch (HSPB), Walter Reed Army Institute of Research (WRAIR), 503 Robert Grant Avenue, Silver Spring, Maryland 20910-7500

SUBJECT: Continuing Review Report Acceptance for the Minimal Risk Human Subjects Research Protocol, **WRAIR #1402**, HRPO Log Number A-14327.2


1. The continuing review report, dated 17 November 2017, for the protocol, **WRAIR #1402**, HRPO Log Number A-14327.2, entitled, "Acute Febrile Illness Surveillance in Kenya," (Version 20.7, dated 15 September 2016), submitted by John Waitumbi, DVM, PhD, Laboratory Director, Kondele Research Unit, United States Army Medical Research Unit—Kenya (USAMRU-K), is accepted.
2. The continuing review report covers the reporting period from 3 November 2016 to 31 October 2017. The enrollment for this study is ongoing.
3. As this is a minimal risk protocol, the continuing review report was reviewed by expedited review procedures according to 32 CFR 219.110. In addition, 45 CFR 46.404 continues to apply to this study as research not involving greater than minimal risk to children. Pregnant women may be enrolled and blood collected by venipuncture; therefore, 45 CFR 46.204 also applies. This study continues to meet the requirements for approval under 32 CFR 219.111.
4. The study is sponsored by WRAIR and funded by the Department of Defense Global Emerging Infections Surveillance and Response System (GEIS).
5. The Kenya Medical Research Institute (KEMRI) Scientific and Ethics Review Unit (SERU) reviewed and approved the continuation of this protocol on 14 August 2017 with an expiration date of 18 August 2018.
6. The following documents are approved for continuation:
 - a. Continuing Review Report, dated 17 November 2017;
 - b. Protocol, version 20.7, dated 15 September 2016;
 - c. Informed Consent, version 20.8, dated 18 October 2016;
 - d. Assent Form, version 20.8, dated 18 October 2016; and
 - e. AR Clinical Data Sheet, version 20.7, dated 15 September 2016.
7. Per the current WRAIR Policy #11-49, "Initial and Continuing Human Subjects Protection Education and Training Requirements", an 80% grade on each individual module must be obtained. The Principal investigator is responsible for ensuring each research team member's, to include those listed on the protocol, as well as those who are not explicitly listed but may be providing study/laboratory support, human subjects protection training is current. Additionally, the PI must maintain records of documentation of this training (i.e., a staff log and training files).
8. The expiration date of this study at the WRAIR is **9 January 2019**. A closeout report is due on **9 January 2023**. No changes, amendments, or addenda may be made to the protocol

MCMR-UWZ-C

SUBJECT: Continuing Review Report Acceptance for the Minimal Risk Human Subjects
Research Protocol, **WRAIR #1402**, HRPO Log Number A-14327.2

without prior review and approval by the WRAIR IRB, the KEMRI SERU, and the USAMRMC
Office of Research (ORP) Human Subjects Protection Office (HRPO).

9. The point of contact for this action is Michelle Block, MS, CIP at 301-319-9535 or
michelle.e.block.civ@mail.mil.



LISA M. LEE, PHD, MA, MS
Chair, Institutional Review Board
Walter Reed Army Institute of Research

CF:

Victor Melendez, COL, MS
Douglas Shaffer, PhD
John Waitumbi, DVM, PhD
Stacy Gondi
KEMRI SERU
MCMR-RP

RESEARCH

Open Access

Plasmodium falciparum population structure inferred by *msp1* amplicon sequencing of parasites collected from febrile patients in Kenya



Brian Andika^{1,2}, Victor Mobegi³, Kimita Gathii¹, Josphat Nyataya¹, Naomi Maina², George Awinda¹, Beth Mutai¹ and John Waitumbi^{1*}

Abstract

Background Multiplicity of infection (MOI) is an important measure of *Plasmodium falciparum* diversity, usually derived from the highly polymorphic genes, such as *msp1*, *msp2* and *glurp* as well as microsatellites. Conventional methods of deriving MOI lack fine resolution needed to discriminate minor clones. This study used amplicon sequencing (AmpliSeq) of *P. falciparum msp1* (*Pfmsp1*) to measure spatial and temporal genetic diversity of *P. falciparum*.

Methods 264 *P. falciparum* positive blood samples collected from areas of differing malaria endemicities between 2010 and 2019 were used. *Pfmsp1* gene was amplified and amplicon libraries sequenced on Illumina MiSeq. Sequences were aligned against a reference sequence (NC_004330.2) and clustered to detect fragment length polymorphism and amino acid variations.

Results Children < 5 years had higher parasitaemia (median = 23.5 ± 5 SD, $p = 0.03$) than the > 5–14 (= 25.3 ± 5 SD), and those > 15 (= 25.1 ± 6 SD). Of the alleles detected, 553 (54.5%) were K1, 250 (24.7%) MAD20 and 211 (20.8%) RO33 that grouped into 19 K1 allelic families (108–270 bp), 14 MAD20 (108–216 bp) and one RO33 (153 bp). AmpliSeq revealed nucleotide polymorphisms in alleles that had similar sizes, thus increasing the K1 to 104, 58 for MAD20 and 14 for RO33. By AmpliSeq, the mean MOI was 4.8 (± 0.78, 95% CI) for the malaria endemic Lake Victoria region, 4.4 (± 1.03, 95% CI) for the epidemic prone Kisii Highland and 3.4 (± 0.62, 95% CI) for the seasonal malaria Semi-Arid region. MOI decreased with age: 4.5 (± 0.76, 95% CI) for children < 5 years, compared to 3.9 (± 0.70, 95% CI) for ages 5 to 14 and 2.7 (± 0.90, 95% CI) for those > 15. Females' MOI (4.2 ± 0.66, 95% CI) was not different from males 4.0 (± 0.61, 95% CI). In all regions, the number of alleles were high in the 2014–2015 period, more so in the Lake Victoria and the seasonal transmission arid regions.

Conclusion These findings highlight the added advantages of AmpliSeq in haplotype discrimination and the associated improvement in unravelling complexity of *P. falciparum* population structure.

Keywords Malaria, Multiplicity of infection, *P. falciparum*, *P. falciparum msp1*, Deep sequencing, Genetic diversity

*Correspondence:

John Waitumbi
john.waitumbi@usamru-k.org

Full list of author information is available at the end of the article



© The Author(s) 2023. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.