Determining Plasmodium falciparum Malaria Transmission Networks Through Sequenom And Capillary Electrophoresis Genotyping

Thesis

DETERMINING *PLASMODIUM FALCIPARUM* MALARIA TRANSMISSION NETWORKS THROUGH SEQUENOM AND CAPILLARY ELECTROPHORESIS GENOTYPING.

Irene Akinyi Omedo (BSc, MSc)

A dissertation submitted for the degree of Doctor of Philosophy.

The Open University, UK

Affiliated Research Centre

KEMRI-Wellcome Trust Research Programme, Kilifi, Kenya.

Collaborating Establishment

Wellcome Trust Centre for Human Genetics, University of Oxford, UK.

AUGUST, 2017

Candidate’s contributions

1. Analysis of genome-wide distributed SNP data in *P. falciparum* samples from The Gambia, Kilifi and Rachuonyo South.
2. Sequenom SNP genotyping and analysis of *P. falciparum* parasites sampled from primary school children across Kenya.
3. DNA extraction, PCR amplification and Sanger sequencing of *PfAMA1* and *surf*$_{4.2}$.
4. Analysis of *PfAMA1* and *surf*$_{4.2}$ sequence data.

Publications arising from the work

1. Omedo I, Mogeni P, Bousema T *et al.* Micro-epidemiological structuring of *Plasmodium falciparum* parasite populations in regions with varying transmission intensities in Africa. [version 1; referees: 4 approved] *Wellcome Open Res* 2017, **2**:10 (doi: 10.12688/wellcomeopenres.10784.1).

2. Omedo I, Mogeni P, Rockett K *et al.* Geographic-genetic analysis of *Plasmodium falciparum* parasite populations from surveys of primary school children in Western Kenya [version 1; referees: 2 approved] *Wellcome Open Res* 2017, **2**:29 (doi: 10.12688/wellcomeopenres.11228.1).

ACKNOWLEDGEMENTS

I would like to acknowledge the following people who were instrumental in helping me along my PhD journey.

First and foremost, my deepest gratitude goes to my director of studies, Prof. Philip Bejon. Thank you for taking a chance on me and giving me the opportunity to undertake this project, for your scientific and technical guidance through it all and for your patience as I slowly made my way through “statistics for dummies”. I could not have hoped for a better mentor and teacher.

I would also like to thank my supervisors Dr. Isabella Oyier, for your help with designing and troubleshooting the lab experiments and your guidance on sequence analysis, and Dr Etienne de Villiers, for your valuable bioinformatics input, especially de-bugging my R scripts and running analyses on the server.

To my group mates Alice Kamau, Martha Nginya, David Kangoye, Polycarp Mogeni and Kevin Kamau. Thanks for all the useful discussions around the project, for the moral support and the many conversations on the deeper meaning of life. Big thanks particularly to Alice and Polycarp for your help with the statistical analyses.

Thanks to the immunology laboratory members for making the lab a wonderful place to work in, and for the encouragement in the face of countless failed experiments. Special thanks to Johnstone Makale for helping with sample retrieval and DNA extraction. Special thanks also to Debra Riako for the moral support and for always being on stand-by to babysit. I look forward to returning the favour.

I would also like to acknowledge several members of Prof. Dominic Kwiatkowski’s lab at The Wellcome Trust Centre for Human Genetics, Oxford University. Thank you first to Prof. Kwiatkowski for allowing me to spend time and learn in your lab. I would like to specifically thank Dr. Kirk Rockett, Anne Jeffreys and Christina Hubbart for helping

with the sequenom SNP genotyping and data clean-up. Thanks also to Christa Hendricks for helping organize my trip and coordinating my stay in the UK. Thanks to the whole team for making me feel very welcome.

To my third-party monitor, Prof. Faith Osier, thank you for taking time to talk with me when things seemed to be on a downward spiral and helping me to see, and plan, beyond the PhD.

To my mentor Prof. James Kahindi. Thank you for all the conversations that we had, for always checking up on me to make sure I was still sane, and for showing real concern for my professional and personal wellbeing.

Finally, thank you to Lydia Ogonda and Trizah Koi for all your support and encouragement throughout this PhD journey. Your support made the road much easier. I’m honoured to be considered among your friends.

DEDICATION

To my son Joshua Nathaniel. You came into my life at just the right time. Thank you for letting me see the world through your eyes, and for showing me how to once again find joy in the simple things in life. May your life always be filled with love, happiness and success.

To my Aunt Rose Linda Okudo. Thank you so much for giving me the opportunity, space and support to pursue my education and follow my dreams. I am where I am because of the lessons you taught me. My life is richer and fuller because you existed.

To my mum Mary, dad John and sister Celine. Time flies, and yet for me it stands still. I miss you every day. Sleep with the angels.

ABSTRACT

Declining malaria transmission leads to infection hotspots which need to be targeted to eliminate and eradicate malaria. The degree of parasite mixing in and around transmission foci is likely to impact the effectiveness of targeted interventions and should be considered when developing control programmes. Few studies currently provide empiric evidence on parasite mixing over time and space, making it hard to predict the likely outcomes of targeted interventions. Here, spatio-temporal malaria transmission networks were inferred using genetic data. *P. falciparum* SNP data were analysed at micro-epidemiological scales in two sites in Kenya and one site in The Gambia, and in a subsequent study at macro-epidemiological scales in Western Kenya. Principal component analysis and linear regression were used to analyse population structure and genetic relatedness in time and space, respectively. Study sites were analysed for parasite genotype clusters, barriers to, and directionality in parasite movement. Parasite genetic relatedness was predicted by relatedness in time and space at micro-geographical scales, but no evidence of population structure was seen over larger areas. No barriers to parasite movement were detected at micro or macro-epidemiological scales, although directional movement was observed in two regions of Western Kenya.

PfAMA1 and *surf*$_{4.2}$ capillary sequence data from parasites collected between 1995 – 2014 in Kilifi county were used to validate SNP data results. Sequence data showed high parasite mixing, with no clustering of distinct haplotypes in time or space. Time and distance interacted antagonistically such that distance no longer predicted genetic variation for parasites collected more than 1 year apart.

These findings show parasite populations that are well mixed in time and space, thus targeting hotspots is likely to benefit surrounding communities. However, this high parasite movement is likely to lead to re-introduction of infection from surrounding

regions following “one-off” interventions, although repeated targeted interventions may be effective.

Table of Contents

List of Figures

List of Tables

Chapter 1 Literature Review

1.1 INTRODUCTION

Malaria, an infectious disease with high morbidity and mortality, is endemic in most tropical and sub-tropical regions of the world (WHO, 2016). In humans, it is caused by six species of the *Plasmodium* protozoan parasite (*P. falciparum, P. vivax, P. malariae, P. ovale curtisi, P. ovale wallikeri* and *P. knowlesi*). *P. falciparum,* which is responsible for the highest malaria morbidity and mortality, is found predominantly in Africa (Guerra *et al*., 2008). *P. vivax* is associated with less mortality than *P. falciparum*, but is more geographically widespread, hence increasing the number of people at risk of the disease (Guerra *et al*., 2010). Most *P. vivax* cases occur in Asia (~ 91%) and South America (~ 5.5%), with fewer cases being reported in Africa (3.5%) (Guerra *et al*., 2010). *P. malariae* and the two sub-species of *P. ovale* are less common, but are also found in South America, Asia, Africa and Oceania (Rutledge *et al*., 2017). *P. knowlesi* is a simian parasite transmitted by the long and pig-tailed macaque monkeys in Southeast Asia and can result in severe disease in humans (WHO, 2015). In regions where more than one species occurs, co-infection is common (Rutledge *et al*., 2017, WHO, 2015).

Malaria is transmitted through the bite of an infective female *Anopheles* mosquito. At least 30 – 40 *Anopheles* species that transmit malaria have been identified (Kiszewski *et al*., 2004). One of the most efficient vector species complexes, *Anopheles gambiae*, is found predominantly in Africa (Kiszewski *et al*., 2004). In rare instances, the disease can also be transmitted through transfusion of blood and blood products, organ transplants, needle-sharing among intravenous drug users and congenitally during pregnancy or delivery (Bartoloni and Zammarchi, 2012).

1.2 *P. falciparum* life cycle

Plasmodium parasites belong to the phylum apicomplexa, which exhibit a complex life cycle involving both a vertebrate host and an arthropod vector, and are characterized by the presence of an apical complex that is important in host cell invasion (Cowman and Crabb, 2006). The *P. falciparum* life cycle begins when an infected mosquito takes a human blood meal. While feeding, sporozoites contained in the salivary glands of the mosquito are released into the bloodstream and travel to the liver where they invade hepatocytes and undergo asexual replication (exo-erythrocytic schizogony), leading to the formation of merozoites that are then released into the bloodstream. Merozoites invade red blood cells and develop through three main stages: rings, trophozoites and schizonts. The schizonts rupture and release newly formed merozoites that then invade other red blood cells and undergo another round of maturation and replication. This cycle continues multiple times, leading to increased parasitaemia within the infected host. During the blood stage, some parasites undergo sexual differentiation into male (micro) and female (macro) gametocytes. When these are ingested by mosquitos during a blood meal, the microgametocytes fuse with macrogametocytes to form zygotes in the mosquito midgut. The zygotes develop into motile forms called ookinetes that penetrate the midgut wall and develop into oocysts, which in turn mature and rupture to release sporozoites. Sporozoites then travel to the salivary glands and are released into the bloodstream during the next blood meal, thus perpetuating the life cycle (Wiser, 2017) (figure 1.1).

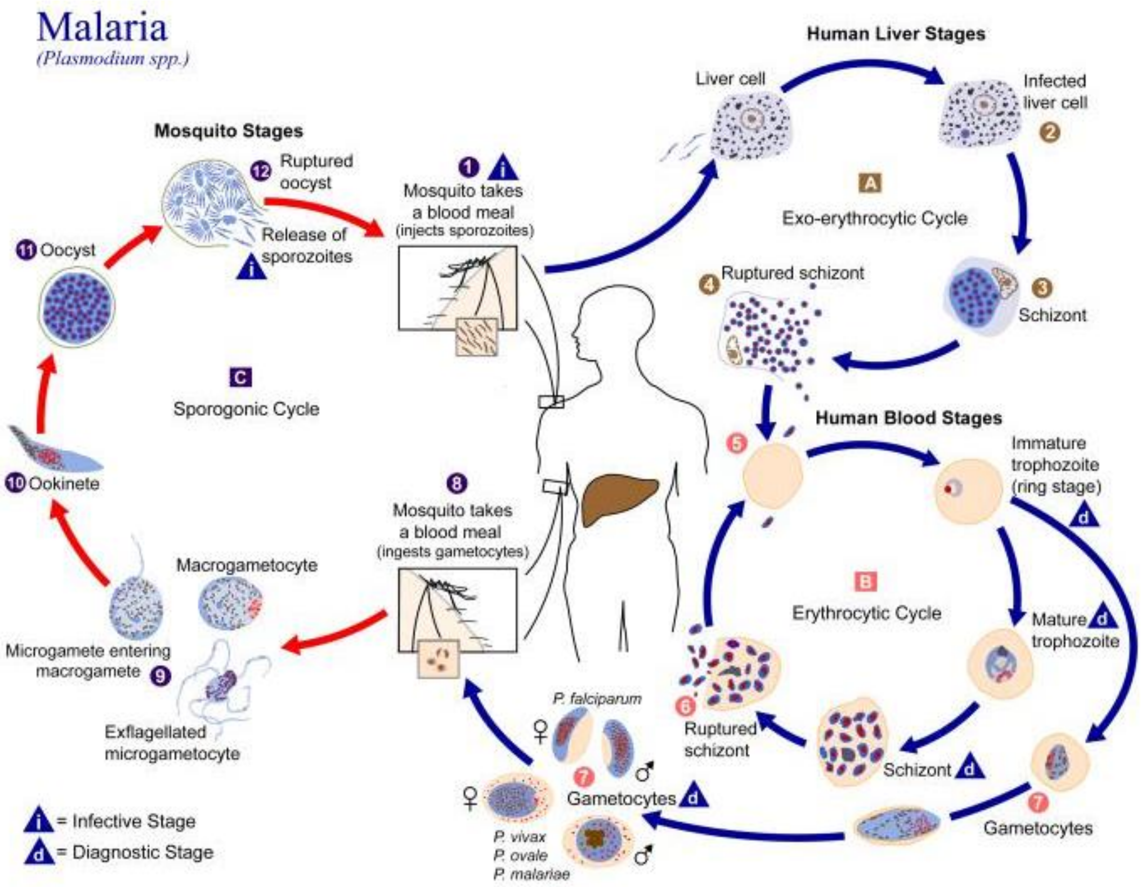


Figure 1.1: ***Plasmodium*** **life cycle**.

The diagram shows the various developmental stages of the parasite in the human host and mosquito vector.

1.3 Clinical manifestations of malaria disease

Following an infective mosquito bite, disease progression in an individual can proceed through successive steps of infection, asymptomatic parasitaemia, uncomplicated disease, severe malaria and death (WHO, 2014). Progression through these stages depends on factors such as the species of the infecting parasite, the host's levels of innate and acquired immunity, host genetic factors as well as the timing and effectiveness of treatment (WHO, 2014). Clinical symptoms associated with malaria are caused by the blood stage of the infection, when merozoites invade, egress and reinvade erythrocytes (Wiser, 2017). *P. falciparum* malaria infection causes an acute febrile illness that was classically said to be characterized by intermittent fever attacks that occur at 48-hour intervals, coincident with the synchronized rupture of infected red blood cells and release of merozoites into the bloodstream (Wiser, 2017). Symptom presentation differs depending on whether the infection leads to uncomplicated mild malaria or proceeds to severe disease. Severe malaria is characterised by variable symptoms including impaired consciousness, acute respiratory distress, multiple convulsions, shock, acute renal failure, prostration and severe anaemia, among others (Marsh *et al.*, 1995, WHO, 2014, Bartoloni and Zammarchi, 2012). The pattern of presentation of severe malaria varies between children and adults, although it is currently unclear whether these variations are attributable only to age differences or whether they result from variations in other aspects such as exposure patterns and health care provision (WHO, 2014). Uncomplicated malaria is characterised by fever and nonspecific, flu-like symptoms including malaise, headache, chills, sweats, vomiting and diarrhoea (Bartoloni and Zammarchi, 2012).

1.4 The epidemiology of malaria in Africa

Globally, malaria was estimated to have caused 212 million clinical cases (range 148 – 304 million) and 429 000 deaths (range 235 000 – 639 000) in 2015 (WHO, 2016). This, however, is likely to be an underestimate of the actual burden of disease, as malaria occurs predominantly in some of the poorest countries in the world, where health systems for identification, documentation and reporting of cases are weakest (WHO, 2014). Additionally, many malaria cases and deaths occur at home, away from health facilities, and even where individuals present to hospital, there may be a misdiagnosis due to the non-specific symptoms of the disease, especially in the absence of confirmatory tests (WHO, 2014). Regionally, most of the morbidity and mortality attributed to malaria occurs in sub-Saharan Africa, with this region accounting for at least 90% of cases and 92% of deaths from the disease (WHO, 2016). The distribution of malaria infections and cases is highly heterogeneous. There are currently 91 countries with ongoing transmission around the world (WHO, 2016). 13 of these, mainly in sub-Saharan Africa, are responsible for 76% of cases and 75% of deaths globally (WHO, 2016). Infants and young children are the most at risk group, with hospital admissions for, and death from, malaria being concentrated in children under the age of 5 years (Carneiro *et al*., 2010). Pregnant women, especially primigravidae (women in their first pregnancy), are also at increased risk of malaria infection, with adverse effects such as low birth weight, preterm birth and foetal and maternal deaths being observed (Takem and D'Alessandro, 2013).

The epidemiology of malaria varies depending on transmission intensity (Snow and Marsh, 2002). At low transmission intensities, all individuals in the population are susceptible due to low or no immunity to the disease. Increasing transmission intensity leads to more frequent exposure to infection, and hence acquisition of immunity, which protects older children and adults in these regions from severe disease, although not infection and uncomplicated disease (Snow and Marsh, 2002). In regions with very high transmission intensities, most members of the population will have acquired anti-disease immunity, and will carry parasites asymptomatically, while severe disease will usually be restricted to infants in their first year of life (Snow and Marsh, 2002).

As transmission declines due to intensified control efforts, the clinical presentations of severe disease also change, with a shift from young children getting predominantly severe malarial anaemia in high transmission regions, to older children getting predominantly cerebral malaria in low transmission regions (Reyburn *et al*., 2005). Several studies provide evidence of this association between age and transmission intensity with disease syndromes. In an analysis of 18 years of surveillance data collected from a paediatric facility in Kilifi, Kenya, the shift from younger children getting severe anaemia to older children getting cerebral malaria was noted as malaria transmission declined in the study area (O'Meara *et al*., 2008). In Tanzania, analysis of admissions to 10 hospitals in areas of different transmission intensities showed the average age of children admitted with severe malaria to be lowest in regions with the highest transmission intensities, and highest in regions with the lowest transmission intensities (Reyburn *et al*., 2005). In this study, severe malarial anaemia predominated in the high transmission areas while cerebral malaria was most common in the low transmission areas (Reyburn *et al*., 2005).

Symptoms associated with pregnancy malaria also vary with transmission intensity. In regions with stable, moderate to high transmission intensities, malaria infections during pregnancy are usually asymptomatic, and symptomatic infections are in the minority (Newman *et al*., 2003, Takem and D'Alessandro, 2013). However, even so-called "asymptomatic" infections are associated with maternal anaemia and low birth weight, a major cause of infant death (Newman *et al*., 2003, Takem and D'Alessandro, 2013). In regions with low, unstable transmission, higher rates of symptomatic infections, including severe disease, are observed in pregnancy, with higher risks of foetal and maternal deaths (Newman *et al*., 2003, Takem and D'Alessandro, 2013).

1.5 Shifting trends in infection and clinical incidence of malaria in Africa

The launch of the Roll Back Malaria initiative in 1998 signalled a recommitment from international funding organizations, governments of malaria endemic countries and the research community to reduce global rates of malaria mortality and morbidity. This was marked by increased funding towards control efforts using available strategies such as long lasting insecticide treated nets (LLINs) and artemisinin combination therapies (ACTs), as well as research into novel intervention strategies such as vaccines (Nabarro and Tayler, 1998). These efforts have led to marked declines in both mortality and morbidity in many endemic regions since 2000. According to the 2016 world malaria report, 17 countries eliminated malaria between 2000 and 2015, although all these countries were situated outside of Africa, which bears the biggest disease burden (WHO, 2016). However, during the same period, the incidence of new malaria cases and deaths reported in Africa fell by 21% and 31% , respectively (WHO, 2016).

Evidence from multiple studies shows a steady decline in malaria incidence in Africa between 2000 and 2015, albeit to variable extents in different geographical areas. At the continental level, studies using both surveillance data and mathematical modelling to quantify the impact of increased control estimated huge gains in reduction in both malaria infection and clinical incidence (Bhatt *et al*., 2015, Gething *et al*., 2016, Noor *et al*., 2014). Within specific countries, declines in malaria incidences have been reported in The Gambia (Ceesay *et al*., 2010), Mozambique (Mayor *et al*., 2015), Senegal (Daniels *et al*., 2015), Eritrea (Graves *et al*., 2008), Zambia (Sutcliffe *et al*., 2011), Zanzibar (Bhattarai *et al*., 2007), Burkina Faso (Geiger *et al*., 2013) and Sao Tome and Principe (Teklehaimanot *et al*., 2009), among others. Greater declines have been reported in East Africa compared to West Africa (Noor *et al*., 2014, O'Meara *et al*., 2008).

In Kenya, declining malaria incidence has been reported along the east African coast (Mogeni *et al*., 2016, Snow *et al*., 2015), as well as in the western part of the country (Kenya national malaria control programme, 2016). In some instances, the decline was coincident with increased control, although in some cases malaria incidence began to decline before widespread interventions were implemented (O'Meara *et al*., 2008, Snow *et al*., 2015, Mogeni *et al*., 2016). Factors that contributed to the decline prior to interventions are currently unknown.

While malaria reduction in Africa is in general a success story, the decline has not occurred uniformly in all countries, or even within individual countries. Several countries recorded little or no changes in malaria incidence over the same period, e.g. in Malawi (Bennett *et al*., 2013) and northern Uganda (Proietti *et al*., 2011), and in some cases even reported an increase in incidence, e.g. in Gabon (Assele *et al*., 2015). In a spatio-temporal analysis of the variations in risks of infection from malaria in Africa between 2000 and 2010, malaria infection was shown to have increased in Malawi and South Sudan, while it remained stable in DR Congo and Chad (Noor *et al*., 2014). Recent data show rebounds in infection in some regions such as coastal Kenya (Mogeni *et al*., 2016, Snow *et al*., 2015), western Kenya (Zhou *et al*., 2016) and Thiés, Senegal (Daniels *et al*., 2015). This is a cause for concern, and indicates the need for novel or improved control interventions to sustain or reduce disease incidence.

1.6 Malaria control

Several strategies have been employed in attempts to control, eliminate and eradicate malaria. The Global Malaria Eradication Programme (GMEP) (1955 – 1969) was the first large scale attempt at malaria eradication and led to the elimination of malaria from some regions, using the long-lasting insecticide, dichloro-diphenyl-trichloroethane (DDT) to kill mosquitoes, and chloroquine to treat clinical malaria (Najera *et al*., 2011). However, the programme failed, facing a combination of the development of resistance to both the insecticide and the drug, political priorities, and a recognition that transmission was too high in much of sub-Saharan Africa for interventions to lead to elimination. The programme was in fact barely implemented in Africa, which bares the biggest burden of the disease. Resurgence of malaria epidemics in regions where it was near elimination saw the abandonment of the GMEP in favour of malaria control in regions where elimination was not feasible, but control remained in a lull until the 1990s. The initiation of the Roll Back Malaria venture led to increased funding for malaria vector control using insecticide treated nets (ITNs), indoor residual spraying (IRS) and larviciding (The malERA Consultative Group on Vector Control, 2011). Controlled ITN trials were carried out in different parts of Africa in the mid to late 1990s and showed the effectiveness of bed nets at reducing morbidity and mortality, especially in children (Alonso *et al*., 1991, D'Alessandro *et al*., 1995, Habluetzel *et al*., 1997, Nevill *et al*., 1996). The results of these and other studies led to increased funding, and consequently to higher bed net coverage across Africa (Bhatt *et al*., 2015, Noor *et al*., 2014, Noor *et al*., 2009).

Apart from bed nets, indoor residual spraying to kill mosquitoes is also widely used as a malaria vector control strategy and has been highly effective in reducing malaria incidence (Curtis and Mnzava, 2000, Pluess *et al*., 2010). These measures, however, are only partially protective and therefore do not eliminate malaria in high transmission areas. Furthermore, challenges such as high costs have prevented the achievement of high levels of coverage, especially among endemic populations living in poor African villages, although coverage has improved markedly in some parts of Africa with consequent public health benefits (Noor *et al*., 2014, Noor *et al*., 2009). Bed nets may also be ineffective against *Anopheles* vectors that are outdoor feeding and resting (The malERA Consultative Group on Vector Control, 2011). Development of resistance to available insecticides has also reduced the efficiency of some of these vector control strategies (The malERA Consultative Group on Vector Control, 2011). Other vector control measures such as poisoning or removing the breeding grounds of the mosquito and aquatic habitats of the larval stages of the vector can also be used, although such environmental modification measures are less often employed due to their associated high costs (Utzinger *et al*., 2001). Cost-effective strategies employing integrated vector management are being encouraged as a better way of controlling malaria, instead of relying on any one control measure alone (The malERA Consultative Group on Vector Control, 2011).

Other than vector control, antimalarial drug therapies are also used as intervention strategies. Quinine was used as far back as the 17th century to treat malarial fevers, and more than 400 years later, it remains effective, although its use is limited due to its adverse side effects (Butler *et al*., 2010). Quinine was the main antimalarial drug used until 1920s, when more effective synthetic drugs were introduced (Achan *et al*., 2011). One such drug was chloroquine, which was widely used beginning in the early 1940s, and was the mainstay drug during the GMEP era (Najera *et al*., 2011). However, resistance to chloroquine developed quickly, and by 1957, resistance had been detected in Thailand and spread through south and southeast Asia, before spreading to East Africa and then on to western and southern Africa (Packard, 2014). Resistance to chloroquine also arose independently in south America in the 1960s (Packard, 2014). Resistance to chloroquine was first reported in Kenya in 1978. An alternative to chloroquine came in the form of sulfadoxine/pyrimethamine (SP) in 1967. Unfortunately, resistance developed rapidly and SP resistance was reported in Thailand in the same year (1967). Resistance to SP remained relatively low in Africa until the 1990s, but spread rapidly once it was established (Packard, 2014).

Artemisinin, isolated from the *Artemisia annua* (sweet wormwood) plant, and its derivatives were later introduced as more effective antimalarial drugs. Artemisinin has a short half-life and rapidly reduces the density of blood stage parasites, although its mode of action is not conclusively known (Cui and Su, 2009). Artemisinin and its derivatives are commonly used in combination with other long-lasting partner drugs such as mefloquine, lumefantrine and piperaquine to reduce the chances of the parasite developing resistance (Cui and Su, 2009). Artemisinin-based Combination Therapies (ACTs) are currently widely employed as first-line treatments against uncomplicated *P. falciparum* malaria and have been highly effective in reducing malaria associated mortality and morbidity (Dondorp *et al*., 2009). However, reports of *P. falciparum* resistance to artemisinin emerged in western Cambodia in 2009, characterised by delayed parasite clearance time (Dondorp *et al*., 2009). Resistance against this drug has now been confirmed in the four Southeast Asian countries of Thailand, Vietnam, Myanmar and Cambodia (Dondorp *et al*., 2009). Although resistance has not been reported in Africa, some of the mutations in the propeller domain of the Kelch 13 gene associated with artemisinin resistance in Southeast Asia have been identified in African parasite populations, albeit at much lower frequencies (MalariaGEN, 2016). Resistance to the ACT partner drug piperaquine has also been detected in Cambodia (Saunders *et al*., 2014). Although ACTs are still effective in Africa, access to such drugs especially by poor, marginalized communities in endemic areas has been relatively low, meaning that treatment cannot be accessed by those who need it most (Noor *et al*., 2009). Intermittent preventive treatment of malaria in infants, young children and pregnant women is also employed as a malaria control strategy in high transmission areas (Nganda *et al*., 2004, White, 2005). Although this may protect the individual, the limited targeting within the population is not sufficient to impact transmission (Nganda *et al*., 2004), and the parasite's resistance to the recommended drugs, sulfadoxine-pyrimethamine (SP), has reduced the effectiveness of this malaria control method (White, 2005). In lower transmission settings, carrying out public awareness and increasing the capacity to detect, respond to and prevent disease during epidemics can also be used as control measures.

Although malaria incidence is declining, it may not be possible to eradicate malaria with the current tools, and novel interventions including vaccines, are needed. Although there are currently no approved malaria vaccines in the market, there are multiple vaccine constructs that are either under development or in clinical trials (WHO, 2016). These vaccines target the different life cycle stages of the *P. falciparum* parasite (Hill, 2011, WHO, 2016). A blood stage vaccine based on the merozoite surface antigen apical membrane antigen 1 (*AMA1*) showed high protection (64.3%) against malaria caused by parasites homologous to the vaccine strain, but much lower protection (17.4%) against heterologous strains in Malian children (Thera *et al*., 2011). The pre-erythrocytic stage vaccine RTS’S which is based on *P. falciparum*’s circumsporozoite protein showed an efficacy of 36.3% in children aged 5 – 17 months old who received four doses of the vaccine (RTS'S Clinical Trials Partnership, 2015), although efficacy waned over time (Olotu *et al*., 2016). The RTS’S vaccine received a positive scientific opinion from the European Medicines Agency and was recently recommended by the world health organization for malaria vaccine pilot programmes in Ghana, Malawi and Kenya. Though promising, the current vaccines give only short term protection and malaria control would be achieved only if the vaccine is given continuously and if all members of the community are vaccinated, a scenario which is currently not feasible. All these control measures are therefore more likely to result in a reduction in transmission intensity, but not in elimination or eradication of malaria (Bhattarai *et al*., 2007, Fegan *et al*., 2007). Newer techniques including gene drives that prevent female mosquitoes from producing eggs or make mosquitoes resistant to *P. falciparum* infection have been touted as technological breakthroughs that could eradicate malaria, but ethical concerns abound about releasing these modified insects into the wild (Hammond *et al*., 2016).

1.7 Spatial and temporal heterogeneity in malaria transmission

Most infectious diseases, including malaria, are heterogeneous in their mode of transmission and follow the 80/20 Pareto principle where 80% of infections occur in only about 20% of the population (Woolhouse *et al*., 1997). Heterogeneity in malaria transmission leads to hotspots of infection which present as a small number of individuals (or clusters of individuals) within a larger population in a defined geographical area who have higher episodes of symptomatic or asymptomatic *P. falciparum* infections than the population outside the hotspot (Bousema *et al*., 2013, Bousema *et al*., 2016). Although heterogeneity in transmission is present at all levels of transmission intensity, the variation is most conspicuous in areas of low and moderate transmission, where a minority of the population may experience multiple episodes of infection while the majority remain infection free (Bousema *et al*., 2010, Bousema *et al*., 2016, Woolhouse *et al*., 1997). In higher transmission settings, heterogeneity may be masked because most of the people in the population will be infected and carry infections asymptomatically (Bousema *et al*., 2012).

Heterogeneity is observed both in time and space (Alemu *et al*., 2013, Mogeni *et al*., 2016). Temporal heterogeneity is most evident in regions with seasonal transmission, where high infection rates are observed during the rainy season when vector densities increase, and infection rates are lower during the dry season when vector densities reduce (Bousema *et al*., 2013), but may also occur from year to year (Mogeni *et al*., 2016). Spatial heterogeneity in malaria transmission has been detected at different geographical scales, including between different regions in a country (Alemu *et al*., 2013), between villages (Bejon *et al*., 2010) and even between homesteads in the same village (Bejon *et al*., 2014). The factors that underlie this variation in transmission are not well understood, although environmental factors such as altitude, cultivation practices, urbanization and proximity to water bodies that act as mosquito breeding sites may play a role (Baidjoe *et al*., 2016). House structural features (Chirebvu *et al*., 2014, Leandro-Reguillo *et al*., 2015, Lwetoijera *et al*., 2013, Wanzirah *et al*., 2015), human behavioural factors such as the amount of time spent outdoors (Chirebvu *et al*., 2014, Liebman *et al*., 2014) as well as genetic factors (Kwiatkowski, 2005, Verhulst *et al*., 2013, Leffler *et al*., 2017) also play roles in the heterogeneity of transmission. Heterogeneity in malaria transmission leads to hotspots of infections which need to be identified and targeted, if malaria elimination is to be achieved (Bousema *et al*., 2016).

1.8 Identification and targeting of hotspots for malaria control

The presence of infection hotspots makes malaria control strategies less effective as they usually persist even after infection has been reduced in surrounding areas (Bousema *et al*., 2012, Ernst *et al*., 2006, Smith *et al*., 2007). Hotspots act as reservoirs of infection and thus a source of disease to the rest of the community, hindering elimination efforts (Bousema *et al*., 2012, Bousema *et al*., 2016). Achieving any meaningful reduction in malaria transmission in areas containing malaria hotspots will require a scale up in the current malaria control activities, including repeated mass drug administration, widespread distribution of LLINs and intensive IRS. These measures are very costly and may not be realistic for universal coverage in most of the resource-poor countries afflicted by the disease. Identification of hotspots for targeted control may therefore be useful to manage scarce resource, as the limited available resources could be targeted to regions with the highest burden, thus ensuring that those most in need get the intervention. Alternatively, and arguably more importantly, targeting control to hotspots is likely to lead to a reduction in incidence in the surrounding community as well if it interrupts the nodes of transmission (Woolhouse *et al*., 1997).

The spatial extents over which hotspots can be identified and targeted vary in size from entire countries, to small geographical areas less than 1km^2 (Bousema *et al*., 2012, Bejon *et al*., 2014). At these scales, malaria hotspots can be identified by measuring asymptomatic parasite prevalence rates, incidence of disease in young children or seroconversion rates in individuals within defined geographical regions (Bousema *et al*., 2010, Bousema *et al*., 2016). Based on the type of metric used, both stable and unstable hotspots can be identified (Bejon *et al*., 2010). Prevalence of asexual parasites and serological markers to malaria-specific antigens can be used to detect stable hotspots, since on the one hand immunity to infection is acquired late in life, or not at all, and on the other, antibodies to malaria are acquired following multiple repeated exposures and are relatively long lived (Bousema *et al*., 2010, Bousema *et al*., 2012). Hotspots defined by incidence of clinical disease are usually temporally unstable, since consistent higher exposure to malaria parasites in any given location would most likely lead to rapid acquisition of immunity against disease. Thus, temporally stable hotspots would be more likely to cause an increased prevalence of infection in the absence of marked increases in disease. Monitoring clinical disease is more linearly related to transmission only in infants or young children, who generally have low immunity regardless of the transmission intensity (Bousema *et al*., 2012, Bousema *et al*., 2016). In the coastal town of Kilifi, Kenya, both stable hotspots of asymptomatic infections and unstable hotspots of febrile disease have been detected within the same study site (Bejon *et al*., 2010).

The choice of marker is also dependent on the transmission intensity, with serological markers and parasite prevalence (measured by PCR) being the most sensitive at low transmission intensities (Kangoye *et al*., 2016). Malaria transmission hotspots have been identified in multiple regions in Africa (Bejon *et al*., 2010, Bousema *et al*., 2010, Bousema *et al*., 2016, Kangoye *et al*., 2016), Asia (Ahmed *et al*., 2013) and South America (Bautista *et al*., 2006).

1.9 Monitoring the effectiveness of control interventions

Different metrics are available to measure changes in malaria transmission and quantify the impact of control interventions. These include both entomological metrics, such as the entomological inoculation rate (EIR), vectorial capacity and sporozoite rate, as well as clinical metrics such as parasite rate, gametocyte rate and the force of infection (FOI) (Tusting *et al*., 2014). Each of these metrics presents its own pros and cons, and the choice of which metric to use is guided by multiple factors including costs, the accuracy and precision desired, the transmission intensity and the availability of expertise needed to measure the metric (Tusting *et al*., 2014). The entomological inoculation rate, which measures the number of infectious bites received by each person per year, has long been considered the gold standard metric of transmission. In line with transmission intensity, EIR varies widely throughout Africa, ranging from less than 1 to higher than 1000 infective bites per person per year (Hay *et al*., 2000). This metric is measured as the product of the human biting rate (the number of bites received by each individual in a year) and the sporozoite rate (the proportion of mosquitoes collected that contain sporozoites) (Tusting *et al*., 2014). Computing human biting rates involves catching and counting mosquitoes attempting to feed on an individual, and this can be done using human landing catches, CDC light traps and pyrethroid spray catches (Tusting *et al*., 2014). Sporozoite rates are computed by examining the caught mosquitoes for sporozoites e.g. using enzyme-linked immunosorbent assays (ELISA) to detect anti-sporozoite antibodies.

Parasite rate (PR) measures the proportion of individuals who are parasite positive in a specific population, at a given timepoint (Tusting *et al*., 2014). It is the most widely collected metric and has been used, both traditionally and currently, to classify geographical regions based on malaria endemicity (Gething *et al*., 2011, Hay *et al*., 2009, Noor *et al*., 2014). It is measured by examining the blood of a sample population under a microscope to detect blood stage parasites (Tusting *et al*., 2014). RDTs and PCR based techniques can also be used to detect parasites. Its usefulness for measuring transmission intensity depends on endemicity, and has been shown to saturate at high transmission intensities due to acquired immunity and multiple infections.

Most of these ‘traditional’ metrics are not standardized between or even within countries, making comparisons difficult. They are also labour intensive, technically challenging to collect and saturate at high transmission levels (Tusting *et al*., 2014). Most importantly, they lack accuracy in low transmission settings where evaluations of effectiveness of control measures are most needed (Yukich *et al*., 2012, Daniels *et al*., 2015).

Modern metrics based on parasite genomics can also be employed to measure transmission intensity, on the premise that parasite genetics reflects the number of different parasite clones infecting individuals in a population, and is representative of the transmission intensity. In high transmission areas, there is higher genetic diversity represented by a higher number of parasite clones, and this number reduces as transmission intensity declines. Metrics such as molecular force of infection (mFOI), which measures the number of new *P. falciparum* clones acquired over time, and multiplicity of infection (MOI), which measures the number of parasite clones in an individual at a given time point, are increasingly being used to measure and track transmission intensity, especially in low transmission areas, due to their higher sensitivity (Mueller *et al*., 2012, Yukich *et al*., 2012). These metrics rely on genetic variations among parasite isolates and can be collected by sequencing or genotyping highly polymorphic loci in the parasite genome.

1.10 Genetic variation and population structure of *P. falciparum.*

The genome of the *P. falciparum* lab strain 3D7 was published in 2002 (Gardner *et al.*, 2002). The A+T rich (~ 81%) nuclear genome is 23MB long and has 14 chromosomes encoding more than 5000 genes (Gardner *et al.*, 2002). The parasite also has a 6kb mitochondrial genome and a 35kb plastid genome (Conway, 2007). Since the publication of the first genome, efforts to improve the genome sequence and gene annotations have continued (Mu *et al.*, 2010b), and the genomes of at least seven other *Plasmodium* species have since been sequenced (MalariaGEN, 2016).

Whole genome studies of *P. falciparum* field isolates from different regions of the world show remarkable genetic diversity within the parasite's genome. This diversity underlies the parasite's ability to evade immune responses and develop resistance to anti-malarial drugs (Jeffares *et al.*, 2007, MalariaGEN, 2016, Miles *et al.*, 2016, Mobegi *et al.*, 2012, Volkman *et al.*, 2007). Studying the extent of this diversity and how it arises is important in understanding how parasites interact and develop resistance against drugs as well as in identifying vaccine targets (Amambua-Ngwa *et al.*, 2012, Ariey *et al.*, 2014, Jeffares *et al.*, 2007). *P. falciparum* genetic variations include SNPs, short insertions and deletions (indels), inversions, non-coding variable number tandem repeats (VNTR), translocations, microsatellites and gene copy number variations (Cheeseman *et al.*, 2009).

Recombination, which refers to allelic rearrangements within chromosomes and has the effect of introducing new allele combinations into the genome, is a major source of genetic variation in the parasite (Jiang *et al.*, 2011, Mu *et al.*, 2010a). The *P. falciparum* genome shows evidence of recombination at relatively high rates which vary between and within populations (Jiang *et al.*, 2011, Mu *et al.*, 2005, Mu *et al.*, 2010a). Within chromosomes, recombination rates also vary, with 'recombination hotspots' located at chromosome ends and centres, although factors that determine the location and activity of such recombination hotspots are still unclear (Jiang *et al*., 2011, Mu *et al*., 2010a). Recombination is important as it allows the parasite to acquire new variations that allow it to evade the host's immune response (Mu *et al*., 2010a). During the diploid sexual stage of the parasite life-cycle in a mosquito, recombination can occur if there are parasites of different genotypes, a situation which arises when mosquitoes feed on an individual infected with multiple genotypes or on different individuals with different genotypes (Jiang *et al*., 2011, Mu *et al*., 2010a). Recombination involves the exchange of genetic material between homologous regions of chromosomes of different parasite isolates, leading to new allele combinations in the progeny (Jiang *et al*., 2011, Mu *et al*., 2010a). Where transmission intensity is low and less within-host genetic diversity is present, "selfing" is more common (Anderson *et al*., 2000, Manske *et al*., 2012).

The genetic variations are present throughout the parasite genome with varying degrees of abundance; for example, microsatellites occur every 2-3kb throughout the genome (Anderson *et al*., 2000). Single nucleotide polymorphisms (SNPs) are single base variations occurring at specific chromosome locations in different members of a species and are common in most organisms (Kwok and Chen, 2003). Although most of the SNPs have no biological functions and generally occur in the non-coding regions of the genome, others, especially those occurring in or near coding regions, may be of biological importance and have in some cases been associated with increased susceptibility to disease (Kwok and Chen, 2003, Tarazona-Santos *et al*., 2011). SNPs are an abundant type of variation in the *P. falciparum* genome, with up to 86,158 SNPs identified in a conservative analysis (Manske *et al*., 2012), and several hundred thousand with less conservative analysis (MalariaGEN, 2016). Most of these SNPs are present at very low frequencies, especially in African parasite populations (MalariaGEN, 2016). Several studies have also identified a high number of indels in *P. falciparum* when comparing either lab strains (Volkman *et al*., 2007) or genetic crosses

(Miles *et al*., 2016). Several SNPs have been associated with resistance against the major antimalarial drugs, including chloroquine (Payne, 1987, Wootton *et al*., 2002), sulfadoxine-pyrimethamine (Ekland and Fidock, 2007, White, 2004), and more recently, artemisinin (Ariey *et al*., 2014, Miotto *et al*., 2015).

P. falciparum population studies of genetically distinct subpopulations within a larger population have been conducted in different regions of the world. Genetic differentiation in subpopulations can occur due to natural selection favouring different genotypes in different environments, random events occurring during transmission of alleles to subsequent generations or initial variation in allele frequencies of the founder subpopulation (Anderson *et al*., 2000, Hartl and Clark, 2007). Population structure leads to reduced heterozygosity within a population as free movement of genes is restricted between different populations of the same organism (Hartl and Clark, 2007). Virtually all organisms have some population structure and in the case of *P. falciparum*, understanding its population genetic structure enables us to know how alleles are distributed within the same parasite population as well as among different parasite populations (Hartl *et al*., 2002). Such an understanding would be important in explaining how parasites acquire drug resistance and how resistance against vaccines could occur, as well as informing on the best epidemiological control strategies (Hartl *et al*., 2002, Manske *et al*., 2012, Mobegi *et al*., 2012, Schultz *et al*., 2010).

Population structure can also be used to measure the effects of control interventions on reducing transmission by studying the level of genetic diversity before, during and after applying control interventions (Daniels *et al*., 2015, Gunawardena and Karunaweera, 2015, Kwiatkowski, 2015). Many *P. falciparum* population genetics studies have been conducted using either microsatellites (Anderson *et al*., 2000, Mobegi *et al*., 2012, Schultz *et al*., 2010) or polymorphic antigenic molecular markers such as merozoite surface proteins (MSP-1 and MSP-2) and glutamate rich protein (GLURP) genes

(Babiker *et al*., 1997, Congpuong *et al*., 2014, Gupta *et al*., 2014, Kyes *et al*., 1997, Mohd Abd Razak *et al*., 2016). Studies looking at MSP-1 and MSP-2 diversity in two villages in Tanzania and Sudan showed greater genetic diversity of parasites in high transmission areas (Babiker *et al*., 1997), while studies of the same genes in a coastal Kenya parasite population found little genetic diversity (Kyes *et al*., 1997). Microsatellites are preferred in population studies because they are presumed to be selectively neutral and thus the resulting observed variations are attributed solely to population history and not natural selection (Anderson *et al*., 2000). Recently SNPs have been used in population genetic studies due to their abundance in the parasite genome and their higher resolution, hence greater ability to distinguish parasite clones (Daniels *et al*., 2008, Manske *et al*., 2012, Roetzer *et al*., 2013, Volkman *et al*., 2007). In one study, Manske and others determined the genetic diversity of *P. falciparum* populations from Africa and Asia/Pacific regions through deep sequencing of parasites from natural infections (Manske *et al*., 2012). In this study, they showed that parasite populations were geographically distinct at the continental level, and parasites from Africa could be easily distinguished from parasites from Papua New Guinea and Southeast Asia, although there was less resolution of parasite populations at the regional level, and parasites from East and West Africa, for example, could not be distinguished from each other (Manske *et al*., 2012). Other studies have also been able to characterize parasites into distinct populations using much fewer SNPs when analysing data over large geographical areas (Campino *et al*., 2011, Daniels *et al*., 2008) as well as over smaller geographical areas (Daniels *et al*., 2015), thus demonstrating the power of SNPs in distinguishing different parasite clones.

P. falciparum population genetics studies show that genetic diversity of this parasite varies greatly worldwide (Anderson *et al*., 2000, Volkman *et al*., 2007). In a detailed analysis, Anderson and others used a set of 12 microsatellites and showed that parasites from Africa have greater within-population genetic variability when compared to parasites from South America and the Asia/Pacific regions (Anderson *et al*., 2000). The study also showed that population structures of *P. falciparum* vary based on the transmission intensities of different geographical areas, with areas of high transmission having more diverse genetic structures compared to areas of low transmission. These results are supported by studies which looked at the same set of microsatellites in a large scale study involving four countries in West Africa (Mobegi *et al*., 2012) and a small scale study involving villages in Papua New Guinea (Schultz *et al*., 2010).

Most population genetic studies have, however, been carried out over relatively large geographical areas, whereas understanding parasite genetic structure at a local level is likely to provide information that best informs targeted control measures for specific regions. An example is the study by Schultz and others, which identified a single village in Papua New Guinea that had high genetic differentiation but little diversity and which could be effectively targeted for malaria control due to its isolation from surrounding villages (Schultz *et al*., 2010).

SNPs are preferred in the analysis of parasite population structure because besides being abundant and widespread across the genome, they are easy to define on algorithmic screening, and can therefore be used as convenient markers. Linkage disequilibrium (LD), which is an important aspect of population structure, has been shown to be strong in low transmission areas due to low recombination rates among parasite clones as most of the zygotes are formed through ‘selfing’ of gametes from the same clone (Anderson *et al*., 2000, Anthony *et al*., 2005, Volkman *et al*., 2007). On the other hand, LD is low in areas of high transmission due to higher effective recombination rates among different parasite clones. This has been validated in studies by both Anderson and others (Anderson *et al*., 2000) and Schultz and others (Schultz *et al*., 2010) which looked at populations with different levels of endemicity using microsatellites, as well as a study by Volkman and others (Volkman *et al*., 2007) which analysed genome-wide SNPs in 16 geographically diverse parasites and showed that LD extends over shorter distances in African parasites as compared to Asian and South American parasites.

The parasite genome is undergoing evolution, and signatures of both balancing and purifying selection are detectable across the genome in parasites from different regions of the world. The selection pressures seem to be specific to different geographical regions and include pressure from anti-malarial drugs and host immunity (Conway *et al*., 2001, Mackinnon and Marsh, 2010). Different techniques are available for detecting these genetic variations.

1.11 Methods of genotyping malaria parasites

Genotyping takes advantage of variations in the genetic make-up of organisms to distinguish parasite clones (Manske *et al*., 2012). Different genotyping methods can be employed based on the specific genetic variation that one wants to detect, sample numbers and available resources (Edenberg and Liu, 2009). Hemi-nested polymerase chain reaction (PCR) has been used as a genotyping method in analysis of microsatellites and highly polymorphic parasite genes such as MSP-1 and MSP-2 (Anderson *et al*., 2000, Mobegi *et al*., 2012, Nyachieo *et al*., 2005). Hemi-nested PCR is a modification of nested PCR where a single primer is used in the second round of amplification as opposed to the usual two primers. The PCR products can then be separated using either capillary electrophoresis or conventional gel electrophoresis where the fragments are separated based on size (Anderson *et al*., 2000, Liljander *et al*., 2009, Nyachieo *et al*., 2005). PCR has been used in combination with restriction fragment length polymorphism (RFLP) to identify different parasite clones (Falk *et al*., 2006). RFLP makes use of differences in homologous DNA sequences which are detected by the presence of DNA fragments of varying lengths after digestion with restriction enzymes (Falk *et al*., 2006). PCR-RFLP followed by agarose gel-electrophoresis provides a relatively inexpensive method of identifying parasite clones and is the recommended method for use in the field during drug efficacy trials to distinguish re-infection from treatment failure (recrudescence) (WHO, 2008). The use of PCR-RFLP in high transmission settings is however difficult as the banding patterns produced on the gel become difficult to analyse due to the multiple complex infections present in these settings (WHO, 2008). In such cases, capillary electrophoresis can be used instead of gel electrophoresis as it provides greater resolution and can better discriminate the different alleles (Liljander *et al*., 2009). The choice of genotyping markers is very important when using these methods as markers that differ only in sequence cannot be used as they would be of the same length and would not distinguish different clones (WHO, 2008). Other genotyping techniques such as southern blotting and hybridization of PCR products using labelled probes can also be employed, but these methods are expensive and time consuming (WHO, 2008). To deal with these issues, different technologies have been adapted to SNP genotyping, with the methods differing in reaction chemistry, sensitivity, throughput and cost (Jenkins and Gibson, 2002, Edenberg and Liu, 2009). Methods such as direct sequencing can be used for SNP detection but are now less commonly used due to their low throughput and high overall costs (Edenberg and Liu, 2009).

Most studies now employ the use of high-throughput techniques such as Sequenom MassARRAY, TaqMan assays and pyrosequencing (Adams, 2008, Gabriel *et al*., 2009, McGuigan and Ralston, 2002, Pourmand *et al*., 2002, Edenberg and Liu, 2009). Sequenom SNP genotyping reaction uses a primer extension method (mini- sequencing) that detects specific alleles based on differences in mass using Matrix-Assisted Laser Desorption/Ionization Time-Of-Flight (MALDI-TOF) mass spectrometry (Gabriel *et al*., 2009). Two platforms are available; the hME (homogeneous MassEXTEND) technology that allows detection of up to 7 SNPs in a pooled assay and the iPLEX GOLD technology that allows detection of up to 40 SNPs (Gabriel *et al*., 2009). Apart from detecting SNPs, Sequenom can also be used to detect small insertions and deletions in the DNA sequence. Sequenom platforms offer several advantages (Gabriel *et al*., 2009). Its multiplex PCR platform allows for the analysis of over 100,000 genotypes per day with up to 40 assays per reaction and scale-up of the number of samples is easy (Gabriel *et al*., 2009). The design of new assays is also relatively easy, depending only on the ability to design primers adjacent to where SNPs of interest are located (Gabriel *et al*., 2009, Edenberg and Liu, 2009). The use of unmodified oligonucleotides further reduces the set-up cost of the assay (Gabriel *et al*., 2009, Edenberg and Liu, 2009). The disadvantages of using the Sequenom platform include a required previous knowledge of the position of SNPs to be studied (Gabriel *et al*., 2009). This technology therefore requires previous sequencing to identify SNPs and cannot be used for SNP discovery studies. The method also returns only genotypic data, thus analyses such as linkage disequilibrium that requires the relation of several SNPs cannot be easily done unless haplotypes are inferred (Gabriel *et al*., 2009).

TaqMan SNP genotyping assays are single tube based PCR systems exploiting the 5' DNA polymerase exonuclease activity (De la Vega *et al*., 2005, McGuigan and Ralston, 2002). The assay consists of locus-specific forward and reverse primers flanking the SNP to be detected and two labelled probes that are allele specific (De la Vega *et al*., 2005). Each probe has a different fluorescent reporter dye at the 5'end and a non-fluorescent quencher dye at the 3'end (De la Vega *et al*., 2005, McGuigan and Ralston, 2002). The proximity of the two dyes on each probe ensures that, when the probe is intact, the quencher dye reduces the fluorescence emitted by the dye at the 5'end, thus reducing the signal (De la Vega *et al*., 2005, McGuigan and Ralston, 2002). During PCR, the allele-specific probe is cleaved at its 5'end by DNA Taq polymerase and with each cycle, cleavage of the probe exponentially increases the fluorescent signal as the fluorophore is separated from the quencher (De la Vega *et al*., 2005). This genotyping platform has several advantages including increased allelic discrimination, flexible assay design and increased signal-to-noise ratio. It also allows one to mask any SNPs that are close to the SNP of interest, thus increasing the success of the assay (McGuigan and Ralston, 2002). Disadvantages include increased costs as it requires the use of labelled probes. The technology can also not be used in SNP discovery as it requires a prior knowledge of the SNPs to be studied (McGuigan and Ralston, 2002).

Pyrosequencing has also been used in genotyping (De la Vega *et al*., 2005, Pourmand *et al*., 2002). Instead of labelled primers or nucleotides, this sequencing technology uses sulfurylase and luciferase enzyme reactions to monitor the release of inorganic pyrophosphate during incorporation of nucleotides (Pourmand *et al*., 2002). Multiplex pyrosequencing has also been done and involves simultaneous detection of multiple target DNA sequences (Pourmand *et al*., 2002).

With advances in genotyping technologies, the use of parasite population genomics is gaining popularity as a way of measuring transmission intensity and tracking the outcomes of control interventions (Daniels *et al*., 2015).

1.12 DNA sequencing

SNP genotyping is easy, convenient and low cost, but it requires prior knowledge of SNPs of interest and is associated with ascertainment bias which can arise based on how SNPs are selected (Lachance and Tishkoff, 2013). As such, sequence data is preferred as it reduces SNP ascertainment bias and enables the typing of many more SNPs, including those present in highly genetically diverse regions of the genome, allowing more accurate analysis of population genetics (Lachance and Tishkoff, 2013).

Sanger sequencing was one of the earliest DNA sequencing methods (Adams, 2008). It uses a chain termination technique where synthesis of a DNA strand is terminated at different points of the synthesis using dideoxynucleotide triphosphates (ddNTPs) that are included in the reaction mix together with deoxynucleotide triphosphates (dNTPs) (Adams, 2008). Earlier methods involved setting up four different reaction tubes, each with a different fluorescently labelled ddNTP, and as the synthesis reaction progressed, DNA polymerase would occasionally incorporate a ddNTP to the strand, thus terminating synthesis (Adams, 2008). This resulted in strands of different lengths being synthesized as the reaction progressed. The different-sized products were then resolved using gel electrophoresis and the sequences determined based on the point of termination of the synthesis (Adams, 2008). This procedure is sensitive enough to distinguish DNA fragments that differ in size by only a single nucleotide (Adams, 2008). Automated sequencers were later introduced, enabling the analysis of multiple samples at a go, and this was followed by the introduction of different fluorescent labels for each ddNTP, allowing the reactions to be carried out in a single tube (Adams, 2008).

Cycle sequencing, which is similar to conventional PCR, was introduced as a modification of the Sanger chain-termination sequencing technique (Murphy *et al*., 2005). In this technique, the target sequence is amplified from a purified DNA sample, the PCR products are then cleaned to remove unincorporated primers and dNTPs, and this is followed by a sequencing PCR reaction incorporating dye-labelled chain terminating ddNTPs (Murphy *et al*., 2005). The sequence products are then subjected to capillary electrophoresis and fluorescence detection (Murphy *et al*., 2005). Cycle sequencing allows the generation of strong sequence signals from small amounts of DNA template due to the multiple rounds of synthesis (Murphy *et al*., 2005).

Although automated Sanger sequencing introduced a method of faster and more efficient detection of DNA sequences, it has low throughput (Adams, 2008). Next generation sequencing (NGS) technologies with high throughput were later introduced and have encouraged studies to move from sequencing of candidate genes to sequencing of whole genomes (Adams, 2008). Examples of NGS platforms that are currently in use include Illumina/Solexa (Illumina), 454 pyrosequencing (Roche Applied Sciences), HeliScope systems (Helicos Biosciences), SoLiD (Applied Biosystems), PacBio (Pacific Biosystems) and Oxford Nanopore (Oxford Nanopore technologies). These techniques differ in their reaction chemistries, length of reads generated and costs (Goodwin *et al*., 2016). The NGS technologies can perform immensely parallel sequencing of PCR products or single DNA molecules. Advances in the development of NGS technologies coupled with the ever-reducing costs of sequencing has made high throughput whole genome sequencing more attractive and raised the possibility of analysing parasite population genomics to allow the assessment of the impact of control interventions in near-real time (Daniels *et al*., 2015).

1.13 Determining disease transmission networks using genetic data

Traditional epidemiological studies to define clusters based on the localization of events in time and space risk reporting associations that are non-existent, or exaggerating existing associations due to bias and confounding (Grimes and Schulz, 2012). Such studies would not accurately determine disease transmission networks, yet identifying these networks is important in helping us understand disease dynamics and informing control strategies. Methods such as social network analysis in outbreaks and transmission of diseases have had limited success, thus necessitating the use of other methods such as genotyping and sequencing. Next generation sequencing techniques have enabled whole genome sequencing and analysis, aiding in the identification of genetic variations such as SNPs in whole organisms. These techniques are fast, therefore disease transmission can theoretically be followed during the course of an outbreak. Due to its high-throughput nature, whole genome sequencing and analyses of hundreds to thousands of samples can be achieved in relatively short periods. Whole genome sequencing and SNP genotyping have been used to map transmission networks of the causative agents of typhoid (*Salmonella typhi* and *Salmonella paratyphi* A), tuberculosis (*Mycobacterium tuberculosis*) and methicillin-resistance staphylococcus aureus (MRSA) (Gardy *et al*., 2011, Harris *et al*., 2013, Harris *et al*., 2010, Janezic *et al*., 2012).

In malaria control, parasite genotypes can be used as a “barcode” for individual isolates, allowing transmission steps to be linked, thus enabling one to track the spread of important parasite characteristics such as drug resistance (Ashley *et al*., 2014, MalariaGEN, 2016). Additionally, parasite genetics can be incorporated into a surveillance system to monitor parasite movement and prevent the re-introduction of malaria following elimination from a specific location (Daniels *et al*., 2015).

1.14 Justification for the study

Malaria transmission continues to decline in many endemic areas, particularly in sub Saharan Africa, in part due to intensified control. As a result, transmission in these regions becomes more heterogenous, leading to hotspots of symptomatic and asymptomatic infection. Effective malaria control and final malaria elimination will require the identification and targeting of these hotspots or reservoirs of infection. However, the level of parasite mixing within and between geographical locations containing hotspots is likely to impact the effectiveness and durability of control interventions and should therefore be taken into consideration when developing control programs. Unfortunately, few studies currently provide empiric evidence on the mixing of parasites over time and space. This study aimed to provide data to fill that gap.

The project involved the use of SNP genotype and sequence data to analyse the genetic diversity and population structure of *P. falciparum* parasites in regions of varying transmission intensities in Kenya and The Gambia. The analyses aimed to describe the spatial and temporal genetic variation of parasite isolates and allow the mapping of parasite movement and interaction within different study sites. The results of the study would show the level of parasite movement and mixing at different geographical scales that may be of relevance to malaria programme officers when designing control interventions. Furthermore, the use of genetic data would enable an inference of the relationship between infections in transmission hotspots and infections in areas surrounding the hotspots, thus informing the design of effective hotspot-targeted interventions. Finally, the findings of this study would enable control programmes to make an inference of the likely outcome of malaria control interventions targeted at different spatial scales.

1.15 Objectives

1.15.1 General objective

To determine *P. falciparum* malaria transmission networks in regions with varying transmission intensities in The Gambia, West Africa and Kenya, East Africa, by analysing the spatial and temporal genetic variations of parasite isolates at different geographical scales.

1.15.2 Specific objectives

1. To analyse the spatial and temporal micro-epidemiological genetic variation in *P. falciparum* parasite populations in three regions with varying malaria transmission intensities: Kilifi, coastal Kenya; Rachuonyo South, western Kenya

and The Kombo coastal districts, The Gambia, using genome-wide distributed SNP genotype data.

2. To analyse spatial and temporal genetic variation of *P. falciparum* parasites at a national and sub-national level using genome-wide distributed SNP genotype data in samples collected from primary school children across Kenya.

3. To analyse spatial and temporal genetic variation in *P. falciparum* parasite isolates collected from children admitted at the Kilifi District Hospital through capillary sequencing of two target genes: Apical Membrane Antigen 1 (*PfAMA1*) and *Surf*$_{4.2.}$

Chapter 2 Temporal and Spatial Micro-Epidemiological Genetic Variation in *P. falciparum* Parasite Populations in Regions with Varying Transmission Intensities in Africa.

2.1 INTRODUCTION

To predict whether targeting hotspots is potentially an effective way of interrupting transmission, there is a need to understand the spatial and temporal scales over which parasite mixing can be observed. Unfortunately, few studies currently provide empiric evidence on the mixing of parasites over space and time. This evidence is important, since targeted malaria control on micro-epidemiological scales is likely to be required to eliminate malaria (Bousema *et al*., 2012). The earliest models of malaria transmission conceived of a completely mixed and homogenous parasite population, and mathematical models based on these show that targeting hotspots may reduce transmission in surrounding areas (Smith *et al*., 2012). These models, however, assume that hotspots are stable and that mosquito mixing in the community is homogenous. However, there has been increasing interest in models allowing for spatial heterogeneity and variably mixed populations of parasites (Perkins *et al*., 2013). These have been guided, in part, by studies showing that certain species of mosquitoes exhibit some level of site fidelity, where they return to the same homesteads to feed (McCall *et al*., 2001). If such behaviour is the norm with very little mixing, then this would greatly reduce the community-wide impact of targeted interventions, and interventions would be beneficial only to individuals within the targeted region. If, however, transmission networks operate freely over large geographical areas, then these interventions would likely have an impact beyond the targeted region. The community-wide impact of targeted control has not been studied extensively. However, early controlled trials showed that bed nets were effective at reducing child morbidity and mortality associated with malaria in villages or communities randomised to the intervention in The Gambia (Alonso *et al*., 1991) and Kilifi (Nevill *et al*., 1996). More recent studies have shown that the use of bed nets in a village randomized to intervention in Asembo, western Kenya, also protected individuals just outside the intervention village who were themselves not using bed nets (Hawley *et al*., 2003). In a recent randomized controlled trial of targeted integrated vector control in Rachuonyo South district in Western Kenya, an initial impact was seen within targeted hotspot areas, but this did not reduce transmission outside the hotspots, and reductions within hotspots were not sustained (Bousema *et al*., 2016). This may have been due to rapid mixing of parasites from areas outside the intervention zones.

Additionally, parasite evolution takes place in a micro-epidemiological context and the spread of drug resistance or new antigenic variants through the population will also be critically dependent on the degree of mixing of parasite populations. Furthermore, declining malaria transmission is associated with increased risk of imported cases of infection and disease from high transmission to low transmission regions, hampering elimination efforts in the low transmission regions (Patel *et al*., 2014) and risking the spread of drug resistant malaria in higher transmission regions (Klein, 2013). Thus, understanding parasite movement and gene flow will provide insights into novel, more targeted approaches to malaria elimination and combating the threats posed by re-introduction.

Under this objective, it was hypothesized that genotyping parasites with fine-scale temporal and spatial data would allow the determination of fine-scale structure to the population and an inference of the degree of parasite mixing in time and space. Genome-wide distributed single nucleotide polymorphisms (SNPs) were genotyped in *P. falciparum* field isolates sampled from three African sites with varying transmission intensities, and analysed to determine the genetic relatedness of parasites within each population. Principal Component Analysis (PCA) was used to detect parasite sub-populations and tests of spatial autocorrelation including Moran's *I* and spatial scan statistics were used to test for autocorrelation among parasite genotypes. The analyses were carried out at different spatial scales ranging from intensive within-village surveillance through to county-wide surveillance.

2.2 MATERIALS AND METHODS

2.2.1 Study population

I was not involved in sample collection or laboratory genotyping of samples analysed in this study. However, the methods used are reviewed here for reference.

P. falciparum infected blood samples were collected from individuals at three sites in two African countries: Kombo coastal districts of The Gambia on the West African coast; Kilifi, Kenya on the East African coast and Rachuonyo South district in the western Kenyan highlands (figure 2.1).

The Gambia has a subtropical climate with a single rainy season between the months of June and October (Ceesay *et al*., 2010) while Kenya has two rainy seasons, experiencing short rains between October and December, and long rains between April and August (Scott *et al*., 2012). In all three sites, *P. falciparum* is the main causative agent of malaria (Bousema *et al*., 2013, Ceesay *et al*., 2010, Scott *et al*., 2012) and transmission occurs almost exclusively during and immediately after the rainy seasons (Ceesay *et al*., 2008, Mwesigwa *et al*., 2015). The common vectors in the Gambia are *Anopheles gambiae s.s.*, *Anopheles arabiensis* and *Anopheles melas* (Caputo *et al*., 2008). The common vectors on the Kenyan coast have historically been *A. gambiae s.s*. and *A. funestus*, although a recent shift to *A. arabiensis* and *A. merus* has been detected

(Mwangangi *et al*., 2013) . In Rachuonyo South district, the main vectors transmitting malaria are *A. gambiae s.s, A. arabiensis and A. funestus* (Stevenson *et al*., 2012). During the study period, temporal trends showed declining malaria transmission in The Gambia and Coastal Kenya (Ceesay *et al*., 2008, Ceesay *et al*., 2010, O'Meara *et al*., 2008, Mogeni *et al*., 2016), although not in western Kenya (Okiro *et al*., 2009). Asymptomatic parasite prevalence is lowest in The Gambia at 8.7% (Sonko *et al*., 2014), intermediate in Kilifi at 14% (Midega *et al*., 2012) and slightly higher in Rachuonyo South at 16% (Stevenson *et al*., 2013). Over the study period, malaria incidence, as measured by malaria slide positivity rate (the proportion of children with fever who are parasite positive based on microscopy), fell from 56% in 1998 to 7% in 2009 in Kilifi (Mogeni *et al*., 2016), and rose slightly in Fajara and Brikama in the Gambia (Ceesay *et al*., 2010). However, transmission intensity is highly heterogeneous both in time and space, with community wide surveys conducted in western Kenya in 2010 showing parasite prevalence rates ranging between 0% - 51.5% (Bousema *et al*., 2013).

2.2.2 Ethics statement

Ethical approval for this study was obtained from Kenya Medical Research Institute (KEMRI)’s Ethical Review Committee (under SSC No. 2239). Written informed consent was obtained from parents/guardians of the study participants. The study methods were carried out in accordance with the approved guidelines.

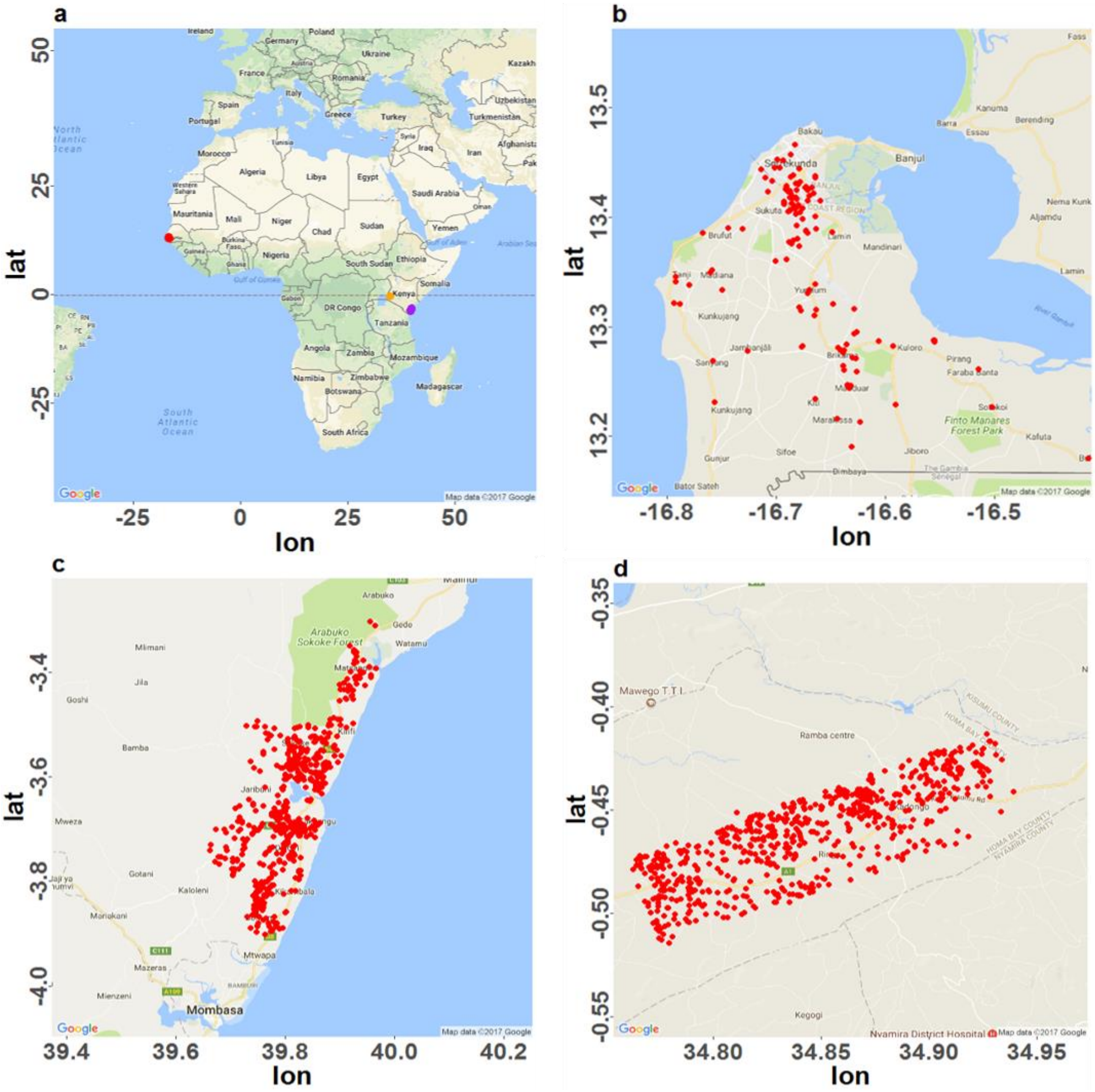


Figure 2.1 Study areas across three sites in Africa.

A) Map of Africa showing the sites of sample collection in The Gambia (red), Kilifi (purple) and Rachuonyo South (orange); and locations of individual sample collection in: **(b)** The Gambia, **(c)** Kilifi and **(d)** Rachuonyo South study sites. Each dot represents an individual sample mapped against the geographical location (homestead) where it was collected.

2.2.3 Sample collection and DNA extraction

A total of 5199 *P. falciparum* infected blood samples were collected in the three sites during hospital admissions and community surveys over a 14-year period from 1998 to

2011 (Table 2.1). In the Gambia, 143 samples were collected from children aged between 8 months to 16 years who presented with either mild or severe malaria at the Medical Research Council (MRC) Fajara clinic or the Brikama government health centre in the Kombo coastal districts in the western side of the country. These children were part of a clinical malaria study in 2007 – 2008 and the geographical coordinates of their residential locations within the Kombo coastal districts had been captured (figure 2.1b) (Ceesay *et al*., 2010). 2312 samples were collected from the Kilifi site, and included children aged 1 to 6 years who had been recruited into a phase 2b randomized trial looking at the efficacy of the Candidate Malaria Vaccines FP9 ME-TRAP (multiple epitope–thrombospondin-related adhesion protein) and MVA ME-TRAP in 2005 (Bejon *et al*., 2006), as well as samples from clinical malaria studies looking at 1) antibody responses to MSP-2 among individuals 3 weeks to 85 years old (Polley *et al*., 2006), 2) the effect of declining transmission on mortality and morbidity in children up to 14 years old (O'Meara *et al*., 2008) and 3) definitions of clinical malaria endpoints (Olotu *et al*., 2010). The 2744 samples from Rachuonyo South district were collected during a community survey conducted in 2011 as part of a trial looking at the impact of targeted control interventions on reducing malaria transmission in the wider community (Bousema *et al*., 2013). Prior to genotyping, DNA was extracted from these samples using either ABI prism 6100 Nucleic Acid prepstation (Applied Biosystems, Waltham, Massachusetts, USA) or Chelex Extraction.

2.2.4 SNP selection and genotyping

276 SNPs in 177 genes were typed in the three parasite populations (Appendix Table 1). The SNPs were selected from a panel of 384 SNPs previously designed for a study of the population structure of *P. falciparum* parasites from Africa, Southeast Asia and Oceania (Campino *et al*., 2011) and were chosen based on three criteria:

a) polymorphic among three of the most studied and well characterised *P. falciparum* strains (3D7, HB3 and IT).

b) uniformly distributed across the parasite genome.

c) ease of typing on the sequenom genotyping platform.

Genes typed included antigen-encoding, housekeeping and hypothetical genes. 52 and 9 SNPs were typed in the antigen-encoding parasite ligands Erythrocyte Binding Antigen 175 (*EBA-175*) and Apical Membrane Antigen 1 (*AMA1*), respectively. The remaining SNPs (herein referred to as “other” SNPs), were distributed more or less evenly across the genome. Between 158 and 226 SNPs were typed in each sample in the Kilifi parasite population, while in The Gambia and Rachuonyo South populations, 131 and 111 SNPs were typed in 143 and 2744 samples, respectively.

Genotyping was done on the Sequenom MassARRAY iPLEX platform that allows multiplexing of up to 40 SNPs in a single reaction well and differentiates alleles based on variations in their masses (Gabriel *et al*., 2009). Briefly, Locus specific PCR and iPLEX extension primers were designed with the sequenom MassARRAY designer software (version 3.1), using 3D7 as the reference genome (PlasmoDB release 9.0). A multiplexed PCR reaction was performed by pooling locus-specific primers, and unincorporated dNTPs and primers were dephosphorylated enzymatically using shrimp alkaline phosphatase (SAP). Extension primers binding immediately adjacent to the SNP site of interest were then extended by a single nucleotide base into the SNP site using mass-modified dideoxynucleotides. The extended products were resin cleaned to remove excess salts and the mass of the different alleles determined using MALDI-TOF mass spectrometry (figure 2.2).

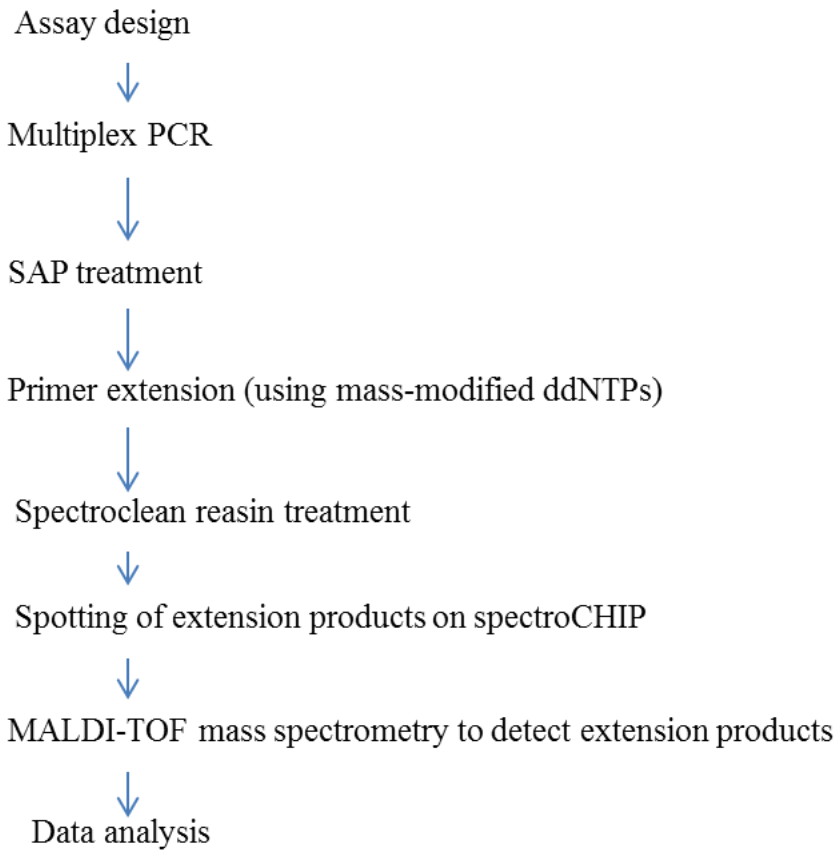

Assay design
Multiplex PCR
SAP treatment
Primer extension (using mass-modified ddNTPs)
Spectroclean reasin treatment
Spotting of extension products on spectroCHIP
MALDI-TOF mass spectrometry to detect extension products
Data analysis

Figure 2.2: Sequenom SNP genotyping process flow chart.

The figure shows the steps involved in SNP genotyping, from designing locus specific PCR primers to conversion of mass-extended products into specific genotype calls.

2.2.5 Sample and SNP cut-off selection criteria

Genotype data were aggregated to determine the distribution of sample and SNP genotyping pass rates, and was carried out separately for each parasite population. Genotyping pass rates were chosen to optimize both the number of samples selected and SNPs typed in each sample. Using both sample and SNP pass rate cut-offs ensured that any SNPs that had high success rates but were typed in only a few samples were excluded from analysis as they would be less useful for comparison purposes. Samples where >40% of SNP typing failed were excluded from analysis, and among the remaining samples, SNP typing for which >30% of samples failed were further excluded from analysis. The criteria for successful SNP typing were based on the SNP intensity values (r) and allelic intensity ratios (theta). Alleles were called as successful if they were above an intensity cut-off value ranging between 0.5 and 1.0, set depending on the performance of the individual SNP assay, and were classified as failed if they were below this cut-off. For those SNPs that were above the cut-off, allelic intensity (theta) ratios ranging between 0 and 1 were used to classify them as homozygous (single parasite genotype infections) or heterozygous (mixed parasite genotype infections). Theta values nearing 0 and 1 indicate different homozygous alleles, while intermediate values indicate heterozygous SNPs, representing mixed parasite populations. Where mixed parasite populations were identified, the majority SNP call at each position was taken to indicate the dominant genotype.

2.2.6 Data analysis

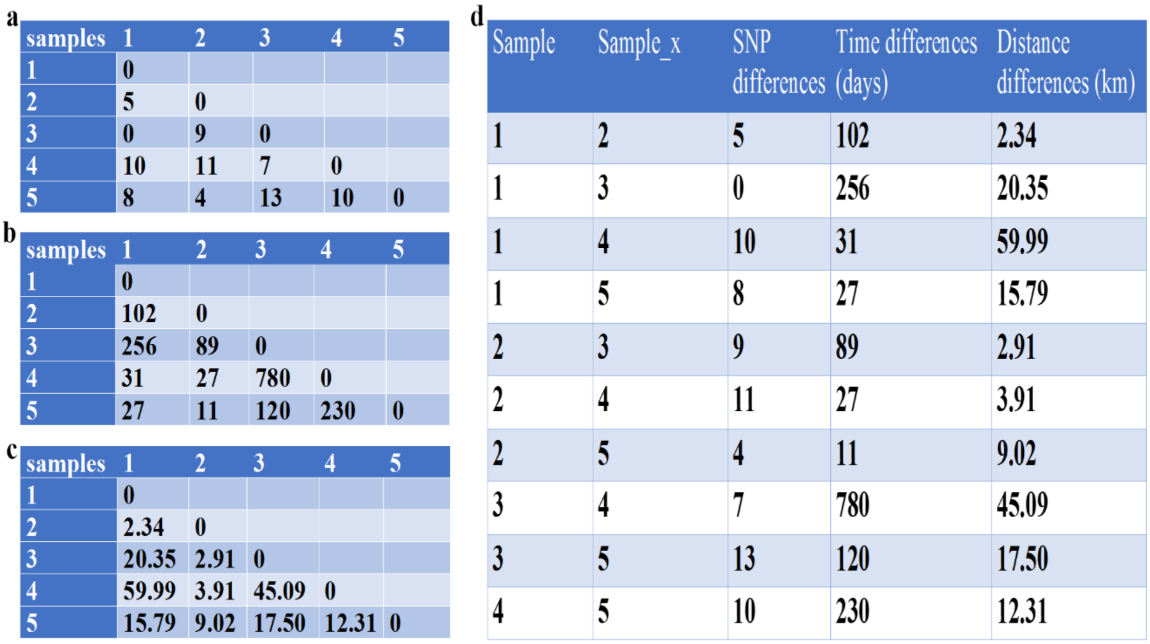
All statistical analyses were conducted in R statistical software (version 3.0.2) (R Core Team, 2013), except for the spatial scan statistics which were computed using SaTScan software (version 9.3) (Kulldorf, 2014). Genotype and geographical data were imported into R software where pre-statistical analyses including removal of negative controls and samples that lacked either genotype and/or spatio-temporal data were undertaken. Analyses conducted included computation of time, distance and SNP differences between parasite pairs, principal component analysis, global and local measurements of spatial autocorrelation, analysis of spatial barriers to parasite movement and variations in genetic differences between parasite pairs over time and space. All analyses were carried out separately for each parasite population (i.e. The Gambia, Kilifi and Rachuonyo South). In each population, the analyses were carried out on the pooled SNP set (all SNPs), as well as three separate SNP subsets (EBA175, *AMA1* and "other"

SNPs). “Other” SNPs represent all SNPs in the dataset, excluding the EBA175 and *AMA1* SNPs.

2.2.6.1 Calculating pairwise time, distance and SNP differences

For each parasite population, time, distance and SNP differences were computed for parasite pairs (figure 2.3), taking half the lower limit of detection of temporal and spatial differences for parasites collected on the same day and/or at the same location. Parasite pairs collected on the same day were assigned a difference of 0.5 days. For older samples in Kilifi (i.e. collected prior to 2004) where location was known to a 5km accuracy, pairs collected at the same location were assigned a difference of 2.5km. Precise geospatial coordinates for recent samples collected in Kilifi (i.e. collected after 2004) as well as all samples from The Gambia and Rachuonyo South were available, so parasite pairs in these three groups collected from the same location (homesteads) were assigned a difference of 0.02km.

SNP differences were computed by comparing genotype data for parasite pairs within each population and counting the number of SNPs between them. Missing SNP data for each parasite was replaced with the major allele in the respective population, after excluding SNP typing where > 30% of assays failed as described above.



a

samples	1	2	3	4	5
1	0				
2	5	0			
3	0	9	0		
4	10	11	7	0	
5	8	4	13	10	0

b

samples	1	2	3	4	5
1	0				
2	102	0			
3	256	89	0		
4	31	27	780	0	
5	27	11	120	230	0

c

samples	1	2	3	4	5
1	0				
2	2.34	0			
3	20.35	2.91	0		
4	59.99	3.91	45.09	0	
5	15.79	9.02	17.50	12.31	0

d

Sample	Sample_x	SNP differences	Time differences (days)	Distance differences (km)
1	2	5	102	2.34
1	3	0	256	20.35
1	4	10	31	59.99
1	5	8	27	15.79
2	3	9	89	2.91
2	4	11	27	3.91
2	5	4	11	9.02
3	4	7	780	45.09
3	5	13	120	17.50
4	5	10	230	12.31

Figure 2.3: Snapshot showing the computation of pairwise SNP, time and distance differences among 5 *P. falciparum* parasites.

The snapshot shows, **(a)** pairwise SNP differences, **(b)** pairwise time differences in days, **(c)** pairwise distance differences in kilometres and **(d)** the final dataset containing SNP, time and distance differences. Samples compared against themselves were excluded from the final analysis.

2.2.6.2 Effect of location and infection status on genetic variation

For the Kilifi parasite population, pairwise SNP differences were also computed between and among parasites collected north and south of the naturally occurring Kilifi creek (the latitude coordinate -3.64 was used to mark the north/south boundary), as well as between and among isolates collected from asymptomatic (community surveys) and symptomatic (hospital cases and short-term laboratory cultured parasites) infections in order to determine whether there were variations in the number of SNP differences based on location of sampling or infection status. In the north/south analysis, each parasite pair was coded into a dummy (categorical) variable based on whether both parasites were collected in the north, both collected in the south, or whether one member of the pair came from the north and the other from the south. Similarly, for the symptomatic/asymptomatic infections, parasite pairs were coded into dummy variables based on whether both members of the pair came from asymptomatic infections, both came from symptomatic infections or whether one member came from an asymptomatic infection and the other from a symptomatic infection. The dummy variables were then included as independent variables in a linear regression analysis with the number of SNP differences between parasite pairs as the outcome variable. The analysis was bootstrapped using 1000 resampling steps to determine the confidence intervals and statistical significance of the observations.

2.2.6.3 Minor Allele Frequency (MAF) distribution

The distribution of minor allele frequencies was computed for all SNP positions in each of the three parasite populations.

2.2.6.4 Population structure and genetic differentiation.

Existence of genetic structure within each parasite population was interrogated using principal components analysis (PCA). PCA is a statistical analysis generally applied to data with high dimensionality, i.e. data with multiple, often correlated variables, to reduce dimensionality of the data while retaining the variables that explain most of the variation in the dataset. The analysis involves transforming the original variables into a new set of variables that are linear combinations of the original variables in the data, are uncorrelated and ordered based on the amount of variation in the original dataset that they explain (Anderson, 2013). The first principal component explains most of the variation in the data, and subsequent components sequentially explain as much of the remaining variation as possible (Ringnér, 2008).

The R function prcomp was used to carry out the PCA. This function takes as its arguments a data matrix which can be either a correlational matrix or a covariance matrix. A covariance matrix is computed if all the variables in the dataset have the same unit of measurement while a correlation matrix is computed when the variables in the data have different units of measurement. In the case of this study, a covariance matrix was used, with individual SNPs representing the variables. PCA can be conducted using spectral decomposition (analyses the covariance and correlation between variables) or singular value decomposition (analyses covariance and correlation among samples) (Anderson, 2013).

Each principal component is a linear combination of the original variables with some associated coefficients (called loadings), which indicate to what extent each variable is correlated with the principal component. The principal components are arranged in order, beginning with the one that explains most of the variation in the data. In this analysis, PCA was computed using singular value decomposition on a covariance matrix of pairwise SNP differences between parasites in each population. Principal component (PC) scores (representing new, uncorrelated parasite genotype values) were computed for the first 3 PCs in each population and the values plotted over geographical maps of the study sites.

Within-population genetic diversity was analysed by computing the average number of pairwise SNP differences in each population. Inter-population genetic differentiation across the three sites was computed based on Weir and Cockrham’s estimate of Wright’s fixation index (F_{ST}), which uses differences in allele frequencies to quantify the level of genetic differentiation between and among populations, with this analysis restricted to 33 SNPs that had been successfully typed in all three populations. The analysis was repeated with an expanded SNP set generated by relaxing the SNP cut-off pass rate to 40%, which increased the number of shared SNPs among the three populations to 40.

Pairwise inter population analyses were also carried out between Kilifi and Rachuonyo South (57 SNPs) and Kilifi and The Gambia (94 SNPs). Analysis between Rachuonyo South and The Gambia was not carried out due to the few number of SNPs (33) that were successfully typed in both populations.

2.2.6.5 Moran's *I* spatial autocorrelation analysis

Moran's *I* is a spatial autocorrelation test that measures the correlation of feature locations and their associated values simultaneously, and is used to determine whether feature attributes are clustered, dispersed or randomly distributed in space. Mathematically, the statistic includes a computation of the mean and variance for each attribute being evaluated (e.g. parasite genotypes represented by PC scores). Deviation from the mean is then computed by subtracting the mean from each parasite PC score. A cross-product is generated by multiplying the deviation values of all features within a specified distance band (e.g. all parasites that are 1km apart), and summing the results. Large deviations from the mean are associated with a larger cross-product which may be positive or negative depending on whether the two features being compared have deviation values that are larger than the mean (positive cross product) or one feature has a deviation value that is less than the mean (negative cross product). Spatial clustering occurs if features with high cross-product values cluster near other features with high cross-product values, and in this case, the Moran's *I* index will be positive. However, if features with high values tend to occur close to features with low values, the Moran's *I* index will be negative, and features will be said to be spatially dispersed. If features with positive cross-product values are balanced with those with negative cross-product values, the index will be zero and the features will be said to be randomly distributed in space.

In computing the Moran's *I* index, latitude and longitude coordinates were used to specify the locations of the features while the scores for the first 3 PCs were used as the feature attribute values. For each principal component, *I* was computed for parasites in 1km, 2km and 5km distance classes, using 100 bootstrap resampling steps to determine the statistical significance of the Moran's *I* correlation coefficients observed.

2.2.6.6 Spatial scan statistics

Spatial scan statistics to detect statistically significant spatial clusters of genetically related parasites were carried out in SaTScan software (version 9.3) (Kulldorf, 2014) and were run separately for each study site. The analysis involved running a purely spatial, retrospective analysis based on a normal probability distribution model using continuous variables (PC scores) and looking for areas with clusters of high PC scores. During the analysis, circular scanning windows centred on each homestead are continuously varied in size, starting from only the homestead (or latitude/longitude point) on which it is centred and gradually increasing to include 50% of the population in the study site. At each window size and location, the ratio of parasites with high PCs inside the window versus outside the window is calculated, and the window with the highest ratio is noted down as a cluster. The statistical significance of this cluster is then determined, taking into account all the multiple tests that are conducted when selecting the optimal window (cluster). This is done by applying random permutations of the PC scores to the spatial coordinates of sample locations, and calculating the log-likelihood statistic of the optimal window detected for each random permutation. A p value is then derived by comparing the log-likelihood statistic for the real data with that of the random permutation, based on 9999 rounds of the random permutation.

2.2.6.7 Relationship between *P. falciparum* population genetics and transmission intensity.

Raster analysis was carried out to test for correlations between population genetics and transmission intensity at fine scale in each study site, with disease incidence and infection prevalence used as a marker of transmission intensity. A raster is a spatial data structure that divides a geographical region into equal sized grids/pixels, and then stores one or more values related to each grid, and allows the representation of 3D information in a 2D format. Raster analysis was carried out for the Kilifi and Rachuonyo South populations only as there was no data measuring transmission intensity (malaria positive fraction or PCR positive fraction) for The Gambian population. Pixels representing different spatial scales (0.5km x 0.5km, 1.0km x 1.0km, 2.0km x 2.0km, 4.0km x 4.0km) were used in the analysis. At each spatial resolution, pixels were assigned the mean of the PC scores and either Malaria Positive Fraction (for Kilifi data) or asymptomatic parasite prevalence by PCR (for Rachuonyo South) of all samples found within that pixel. The correlation between mean PC score and mean MPF or between mean PC score and parasite prevalence was tested by Spearman's rank ordered correlation coefficient.

2.2.6.8 Analysis of spatial barriers to parasite movement

To identify possible spatial barriers to parasite movement and mixing over short distances, each study area was divided into pixels of varying sizes (0.4km x 0.4km, 0.5km x 0.5km, 1.0km x 1.0km) which were then scored with 1 or 0, based on whether or not a straight line linking any two parasites crossed their boundaries. These pixels were used as independent variables in a multivariable linear regression analysis that had the number of SNP differences as the dependent/outcome variable. Significance of the coefficient estimates were determined using non-parametric bootstrap method with 1000 resampling steps. Raster maps showing the mean of the bootstrapped p-values for samples in each pixel were generated.

2.2.6.9 Effects of time and space on *P. falciparum* genetic variations

To analyse the impact of time and distance on the changes in genetic variations between parasite pairs within each population, a multiple fractional polynomial regression analysis was carried out. The product of time (days) and distance (kilometres) was included as a co-variable in the model to test the effect of the interaction of time and distance on changes in genetic variations between parasite isolates. The analyses were run separately for each population. Since the number of days differed for almost all parasite pairs, dummy data were included in the regression analysis to enable the generation of time-distance interaction graphs. For each study site, a distance range of 1 – 10km (with an interval of 0.1km between adjacent distances) was used. Temporal distance with 14 and 10 day intervals were assigned to parasite pairs in Kilifi and the Gambia, respectively, whereas time was not considered for the Rachuonyo South population. Constant SNP differences of 14, 10 and 8 were used for parasite pairs in Kilifi, Rachuonyo South and the Gambia, respectively. Within each population, analyses were carried out separately for all successfully typed SNPs as well as the three SNP subsets (*EBA-175, AMA1*, and “other” SNPs). In the Kilifi population where samples were contributed from different studies, separate analyses were also carried out for samples collected from hospital-based studies, community surveys and short term cultured parasites. The three study groups (hospital cases, laboratory cultured parasites and community surveys) showed similar patterns of variation in SNP differences with time and distance, therefore subsequent analyses were conducted for the individual SNP subsets but on a combined parasite data set containing parasites from all three study groups. For the Rachuonyo South population, SNPs in the *AMA1* gene were not analysed separately due to their few numbers, but were analysed as part of the entire SNP set (all SNPs) for this population. An analysis of the effect of the interaction of time and distance on genetic variation was not carried out for the Rachuonyo South population as the samples were all collected within a few days of each other.

2.3 RESULTS

2.3.1 Summary of study datasets

A total of 5199 *P. falciparum* samples collected from both hospital admissions and community surveys from The Gambia, Kilifi and Rachuonyo South over a 14-year period from 1998 to 2011 were genotyped (Table 1; Table 2). 2769 (53%) of these were selected for further analyses based on SNP and Sample typing pass rates as well as availability of both temporal and spatial data for each sample.

Table 2-1: Summary information for *P. falciparum* infected samples collected from The Gambia, Kilifi and Rachuonyo South.

Region	Study site	Samples genotyped	Samples analysed	Study period	Missing temporal data	Missing spatial data	Missing parasite density	Average parasite density
Kilifi	Community surveys	748	195	Feb - Oct ‘05	34	125	37	4562
Kilifi	KDH	1374	1259	Jul’98-feb’08	2	207	49	352K
Kilifi	Laboratory cultures	190	148	Aug’03 - Apr’10	0	0	190	-
The Gambia	MRC Fajara & Brikama health centre	143	133	Sep’07 – Jan’09	0	0	5	406K
Western Kenya	Rachuonyo South	2744	1034	2010	0	0	2744	-

Table 2-2: Temporal distribution of samples analysed in the *P. falciparum* populations.

Study site	Year of sample collection	No of samples
Kilifi	1998	1
	1999	114
	2000	114
	2001	272
	2002	211
	2003	220
	2004	191
	2005	301
	2006	69
	2007	86
	2008	13
	2009	9
	2010	1
Rachuonyo South	2011	1034
The Gambia	2007	104
	2008	29

Samples from Kilifi and The Gambia exhibited a wide range of parasite densities which invariably affected their genotyping pass rates, with samples having high parasite densities also having high genotyping pass rates (figure 2.4). Genotyping pass rates increased with increase in parasite density, up to a maximum of 10,000 parasites/µl in both Kilifi and The Gambia. Parasite density data were not available for the Rachuonyo South study site.

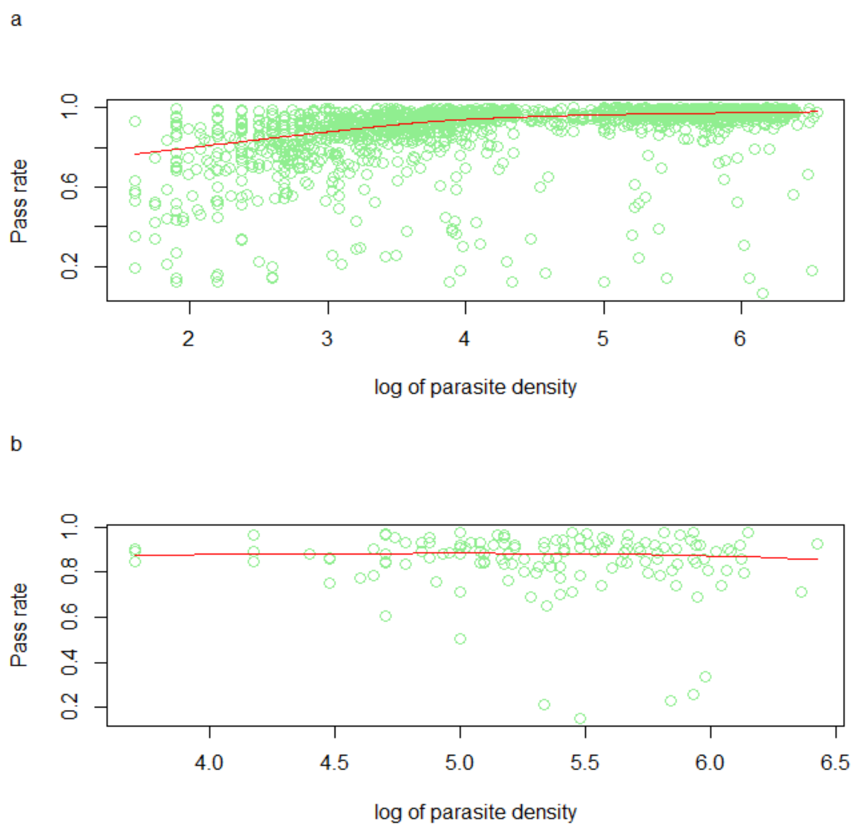


Figure 2.4: Correlation between parasite densities and SNP genotyping pass rates. The relationship was analysed for samples collected in **a)** Kilifi and **b)** The Gambia. Increasing genotyping pass rates were positively correlated with parasite densities, up to a maximum of approximately 10,000 parasites/µl, beyond which parasite density had little impact on genotyping outcome.

Comparisons of the distribution of genotyping pass rates among samples in the three populations showed that hospital cases and short-term cultured isolates had less variation in pass rates compared to community surveys (figure 2.5). Within the Kilifi population, laboratory cultured parasite samples had the highest pass rates and showed the least variation among samples (figure 2.5a). Samples from hospital cases also had high pass rates, with a median value equal to that of the laboratory cultured samples. Those from the community survey had the highest variation in their genotype pass rates. These samples were collected from asymptomatic individuals with a wide range of parasite densities. Samples from The Gambian parasite population (figure 2.5b) had comparable pass rate distribution to that of hospital cases from the Kilifi population, with median pass rate of over 85% and little variation. The Rachuonyo South population (Figure 2.5c) showed variation in distribution of success rates similar to that seen in samples from community surveys in the Kilifi population, but had a lower median success rate of around 40%, indicating lower parasite densities in the Rachuonyo South population compared to the Kilifi population.

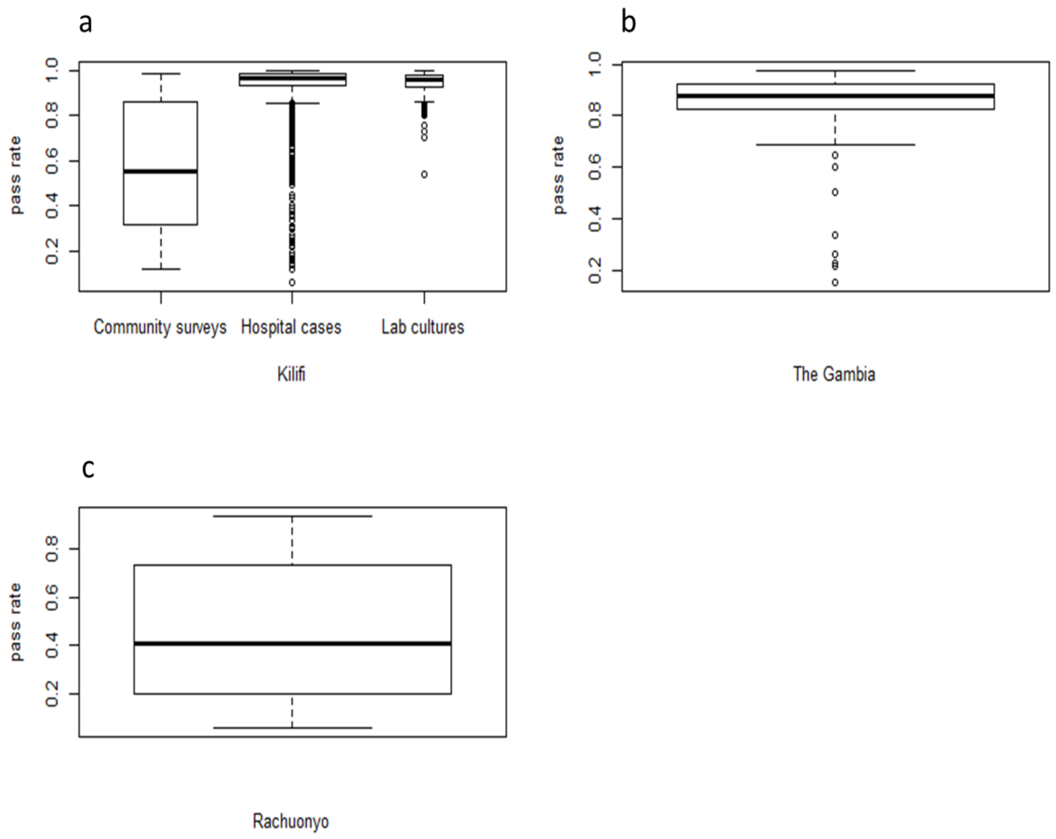


Figure 2.5: Box and Whisker plots showing the distribution of genotyping pass rates for *P. falciparum*.

Plots were produced for parasites from **a)** Kilifi, **b)** The Gambia and **c)** Rachuonyo South populations. Greater variability in genotyping pass rate was seen in the samples collected during community surveys while the hospital cases had a higher number of outliers. Laboratory cultured samples in the Kilifi population had the highest pass rates and showed the least variation.

Genotyping pass rates among samples from the Kilifi population ranged from 5% to 95.7%, while pass rates among SNPs ranged from 0.17% to 96.7%. The SNP that encodes the substitution PfcrtK76T found in the chloroquine resistance transporter gene had the lowest success rate, failing in 99.83% of the samples in which it was typed. This was followed by a SNP found in the EMP1-trafficking protein, which had a failure rate of 99.75%.

Genotyping pass rates among samples from The Gambia population ranged from 15% to 97.7% and pass rates among SNPs ranged from 0 to 100%. Typing of an *EBA-175* SNP and a SNP found in a gene encoding a conserved protein of unknown function, PF11_0353, failed in all the samples while a SNP in the *AMA1* gene was successfully typed in all samples. Genotyping pass rates for both samples and SNPs were generally lower in Rachuonyo South compared to Kilifi and The Gambian populations, possibly due to the lower parasitaemia in Rachuonyo South samples. Sample success rates ranged between 5% and 93.6%, with less than half the samples having 70% or more of their SNPs successfully typed. SNP pass rates ranged from 0% to 98.7%. Typing of three SNPs: two in hypothetical proteins (PF14_0153 and PF11_0347) and one in EBA175 was unsuccessful in all samples in this population. Overall, however, there was no specific bias towards failure of specific samples or SNPs in specific genes, but if there had been, it would have pointed to the possible presence of a previously un-identified SNP within the primer binding site that prevented the primer from binding and prevented detection of the target SNP.

In total, 276 SNPs were typed in parasites from the three populations. Many of these were SNPs distributed throughout the genome and mostly found in genes encoding hypothetical proteins or proteins with unknown functions (Table 2.3; Appendix Table 1). Selection of samples and SNPs for use in further analyses was based on a trade-off between the highest number of samples and SNPs with the least number of missing and failed assays that could be used in the analysis to reach meaningful conclusions.

Table 2-3: single nucleotide polymorphic (SNP) sites typed in Kilifi, Rachuonyo South and Gambian parasites

	EBA	AMA1	“Other” SNPs	total
Kilifi	52	9	175	236
Rachuonyo South	27	3	81	111
The Gambia	45	9	77	131
All	52	9	215	276

Variable numbers of samples and SNPs were selected for further analysis in each site, based on genotyping pass rates (Table 2.4, figure 2.6). In the Gambian population, 131 SNPs were typed in 143 samples and of these, 133 samples and 107 SNPs with at least 70% genotyping pass rates were selected for further analyses. Of the 2312 samples and 236 SNPs typed in the Kilifi population, 1602 samples with a success rate of at least 75% and 177 SNPs with a pass rate of at least 70% were selected for further analyses, while 1034 samples and 82 SNPs with minimum success rates of 60% and 70% respectively were selected from an original set of 2744 samples and 111 SNPs typed in the Rachuonyo South population. Among the samples selected from the Kilifi population, 1259 were from the hospital cases, 195 were from the community surveys and 148 were from the short-term laboratory cultured samples (Table 2.1).

A sufficient number of SNPs were typed in EBA175 and *AMA1* to enable a sub-analysis of these two genes in The Gambian and Kilifi populations (Table 2.3). The sub-analyses were carried out to determine whether patterns of temporal and spatial genetic variation differed when these two genes, which have been shown to be under balancing selection, were analysed, compared to when SNPs in 'neutral' genes were analysed.

Table 2-4: Samples and SNPs analysed in Kilifi, Rachuonyo South and Gambian parasite populations

Study population	Samples analysed (% pass rate cut-off)	SNPs analysed (% pass rate cut-off)
Kilifi	1602 (75)	177 (70)
Rachuonyo South	1034 (60)	82 (70)
The Gambia	133 (70)	107 (70)

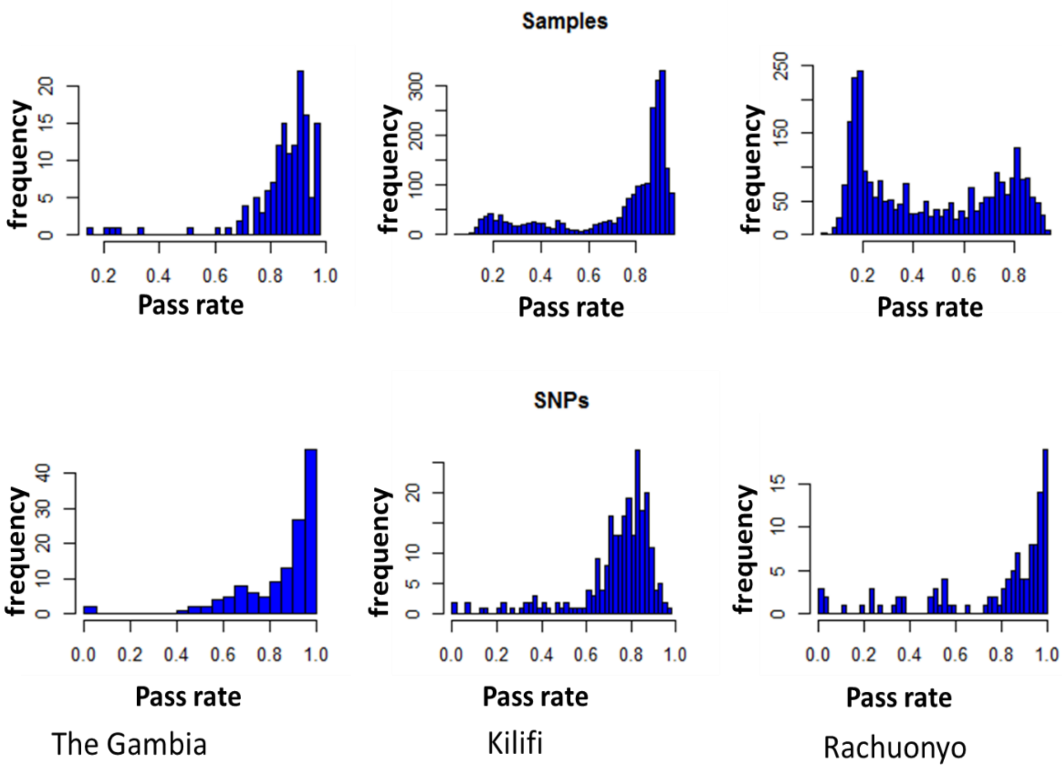


Figure 2.6: Frequency distribution of genotyping pass rates for samples and SNPs in *P. falciparum* parasite populations.

The top panel shows the distribution of genotyping pass rates for samples while the bottom panel shows that of SNPs for The Gambia (left panel), Kilifi (middle panel) and Rachuonyo South (right panel) parasite populations.

2.3.2 Minor Allele Frequency (MAF) distribution

Minor allele frequency (MAF) distribution of SNPs in the different populations ranged from 0% - 47% for parasites in Kilifi and Rachuonyo South populations, and 0% - 49% for parasites in The Gambian population (figure 2.7). Analyses indicate that Kilifi and The Gambia have a similar MAF distribution pattern, where most SNPs are present at low frequencies. For example, over 100 SNPs in the Kilifi population had minor allele frequencies of less than 5%. In comparison, Rachuonyo South parasite population had minor alleles with higher frequencies, indicative of higher genetic diversity.

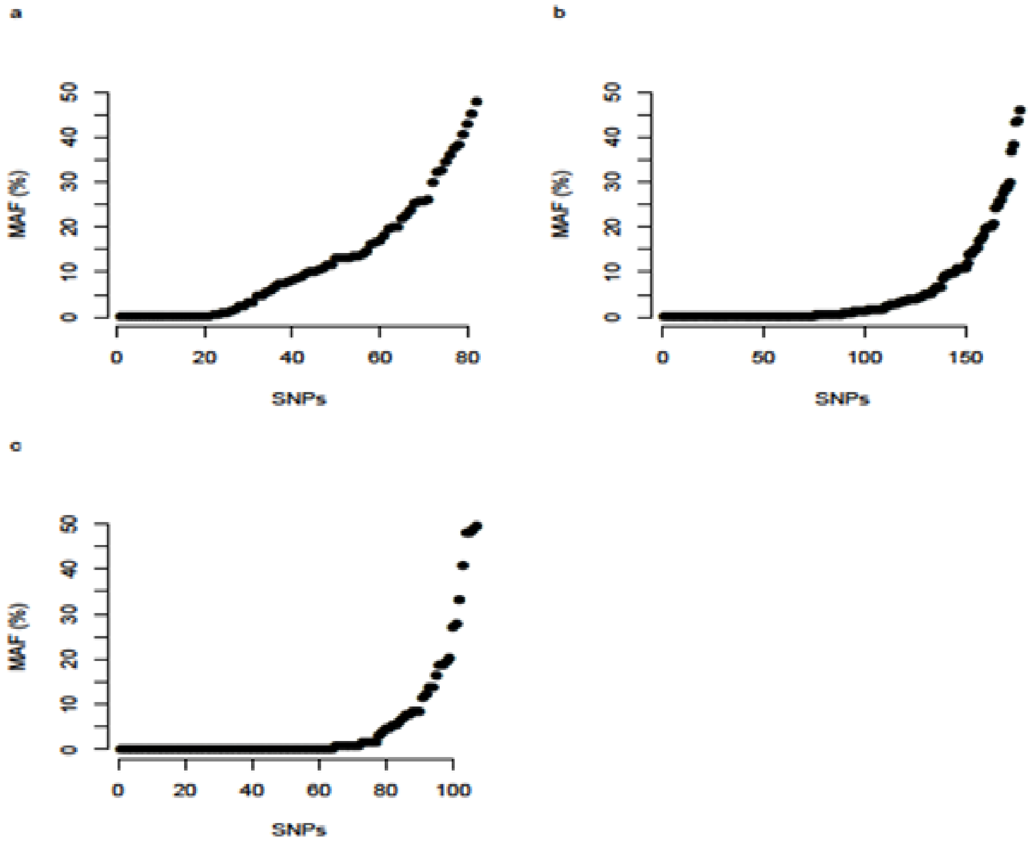


Figure 2.7: Minor allele frequency (MAF) SNP distribution.

MAF was computed for parasites collected in **a)** Rachuonyo South district in western Kenya highlands, **b)** Kilifi in coastal Kenya and **c)** Kombo coastal districts in The Gambia.

Monomorphic positions (100% identical across all samples) were identified in each population. 26% (46 of 177), 20.7% (17 of 82) and 59.8% (64 of 107) of SNP positions in Kilifi, Rachuonyo South and The Gambia, respectively, were identified as monomorphic. Since these positions were identical across all parasite isolates, they did not contribute any information in the analysis of parasite genetic variations. Thus, the actual numbers of polymorphic positions that were detectable within each population, and which contributed to the final analyses were 131 in Kilifi, 65 in Rachuonyo South and 43 in The Gambia (Table 2.5).

Table 2-5: Number of individual SNP positions analysed in *P. falciparum* parasite populations in Kilifi, Rachuonyo South and The Gambia.

	EBA-175	AMA1	'Other' SNPs	Total
Kilifi	36 (17)	8 (8)	133 (106)	177 (131)
Rachuonyo South	20 (10)	3 (3)	59 (52)	82 (65)
The Gambia	39 (11)	9 (8)	59 (24)	107 (43)

Brackets contain the number of polymorphic SNP positions in each SNP subset in each population.

2.3.3 Population differentiation and pairwise SNP differences

Wright's fixation index (F_{ST}) analysis of population differentiation showed the level of differentiation among the three populations to be 0.046 (95% CI: 0.013 – 0.078), comparable with results of other studies of African *P. falciparum* populations using microsatellite typing and whole genome sequencing (Anderson *et al*., 2000, Manske *et al*., 2012). Pairwise population analysis gave F_{ST} values of 0.041 (95% CI: 0.013 – 0.077) between Kilifi and Rachuonyo South, 0.078 (95% CI: 0.016 – 0.149) between

Kilifi and The Gambia and 0.108 (95% CI: 0.001 – 0.195) between The Gambia and Rachuonyo South. Thus, the highest level of differentiation was observed between The Gambian and Rachuonyo South populations and the lowest level of differentiation was observed between The Gambian and Kilifi populations.

The average numbers of SNP differences within and between the three populations were computed using a set of 33 SNPs that were typed in all three populations (Table 2.6). Results of the analysis showed similar levels of diversity for The Gambia and Rachuonyo South (3.407), Kilifi and Rachuonyo South (3.337) and Kilifi and The Gambia (3.264) parasite populations. Within-population genetic diversity showed that parasites in Rachuonyo South had the highest genetic diversity (3.384 SNPs per parasite pair), those in The Gambia had the lowest genetic diversity (2.867 SNPs per parasite pair), while those in Kilifi had intermediate values (3.229 SNPs per parasite pair).

Table 2-6: Average pairwise SNP differences within and between *P. falciparum* parasite populations in Kilifi, Rachuonyo South and the Gambia.

	The Gambia	**Rachuonyo South**	**Kilifi**
The Gambia	2.867		
Rachuonyo South	3.407	3.384	
Kilifi	3.264	3.337	3.229

In the Kilifi parasite population where parasites were stratified by location (north vs south) and infection status (symptomatic vs asymptomatic), analysis of the number of SNP differences between parasites in the north and south showed a higher number of SNP differences between north-north (effect size = 0.357, 95% CI = 0.224 – 0.671, p = <0.001) parasite pairs, and lower number of SNP differences between south-south (effect size = -0.243, 95% CI = -0.578 - -0.119, p=0.002) parasite pairs compared to the north-south parasite pairs. In the symptomatic/asymptomatic infections, there were more SNP differences between symptomatic-symptomatic infections (effect size = 0.807, 95% CI = 0.337 – 1.022, p = < 0.001), and fewer SNP differences between asymptomatic-asymptomatic infections (effect size = -0.781, 95% CI = -1.308 - -0.448, p = < 0.001) compared to the symptomatic-asymptomatic group.

2.3.4 *P. falciparum* population structure

Principal components analysis (PCA) was carried out separately for each population using the 177, 82 and 107 SNPs that were successfully typed in Kilifi, Rachuonyo South and The Gambia parasite populations, respectively. Scree plots showing the amount of variation explained by each principal component were plotted for each parasite population (figure 2.8). In general, each PC accounted for only a small amount of the overall variation in the data. Cumulatively, the first three principal components accounted for 13.2% (PC1=5.1%, PC2=4.4%, PC3=3.7%) of the variability seen in Kilifi, 12.7% (PC1=4.4%, PC2=4.3%, PC3=4%) of the variability seen in Rachuonyo South and 36.1% (PC1=18.4%, PC2=10.4%, PC3=7.3%) of the variability seen in The Gambian parasite populations.

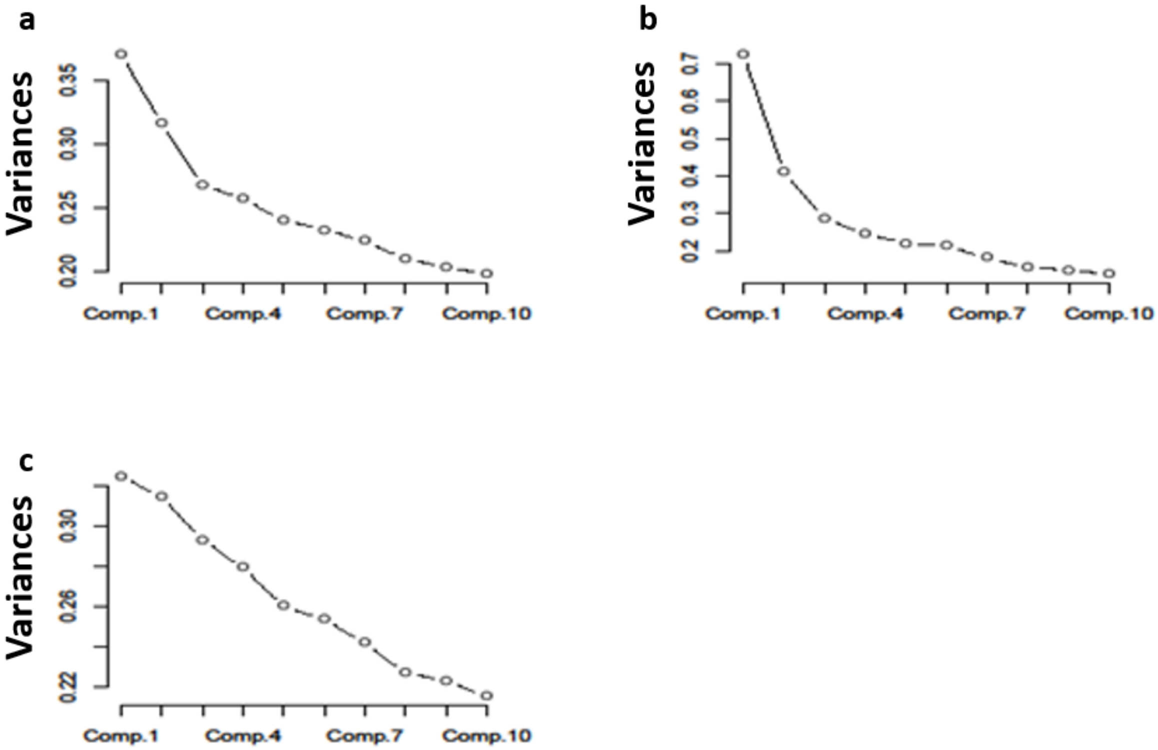


Figure 2.8: Scree plots showing the proportion of total variance in the data accounted for by individual principal components.

In each population, plots were produced for the first 10 principal components in **a)** Kilifi, **b)** The Gambia and **c)** Rachuonyo South *P. falciparum* parasite populations.

Based on general convention, the first three principal components in each population were selected for subsequent analyses. Principal component plots pointed to a high level of within-population mixing among the parasites, and parasite populations could not be resolved into distinct sub-populations using principal component analysis (figure 2.9 - figure 2.12).

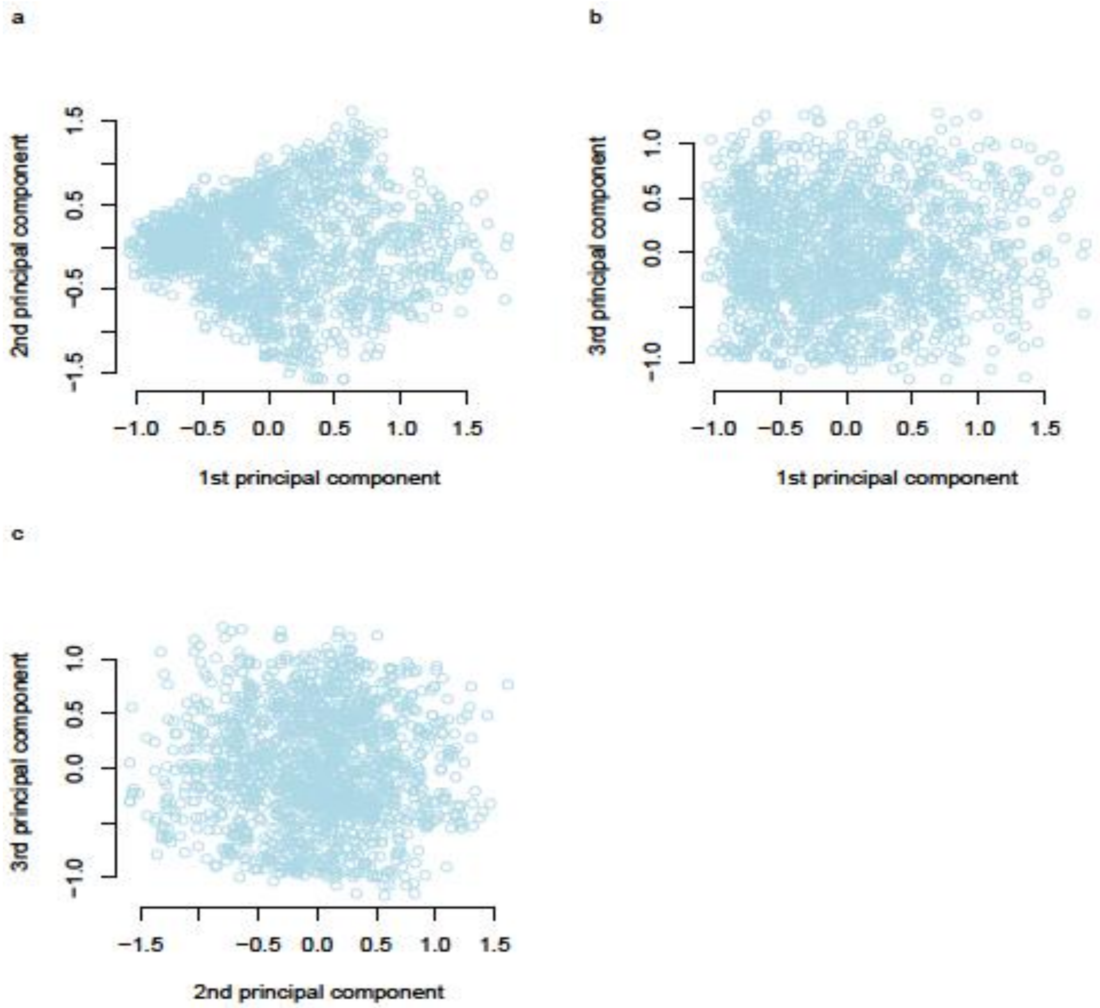


Figure 2.9: Pairwise plots of the first three principal components for Kilifi *P. falciparum* parasite population.

A) 1st PC plotted against 2nd PC, **b)** 1st PC plotted against 3rd PC and **c)** 2nd PC plotted against 3rd PC. Each dot represents an individual sample. No obvious clustering of parasites based on genotypes is evident from these plots.

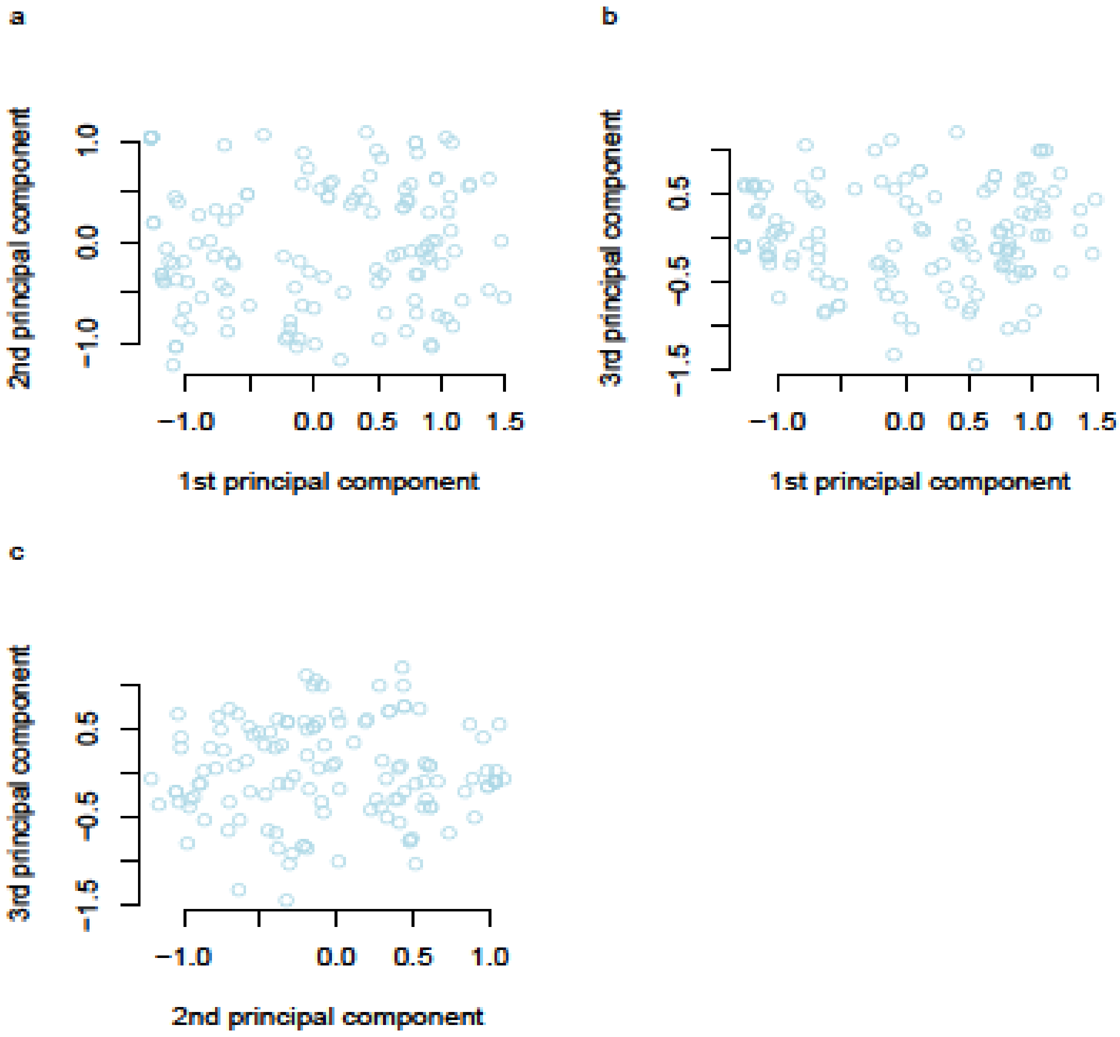


Figure 2.10: Pairwise plots of the first three principal components for The Gambian *P. falciparum* parasite population.

A) 1st PC plotted against 2nd PC, **b)** 1st PC plotted against 3rd PC and **c)** 2nd PC plotted against 3rd PC. Each dot represents an individual sample. No obvious clustering of parasites based on genotypes is evident from these plots.

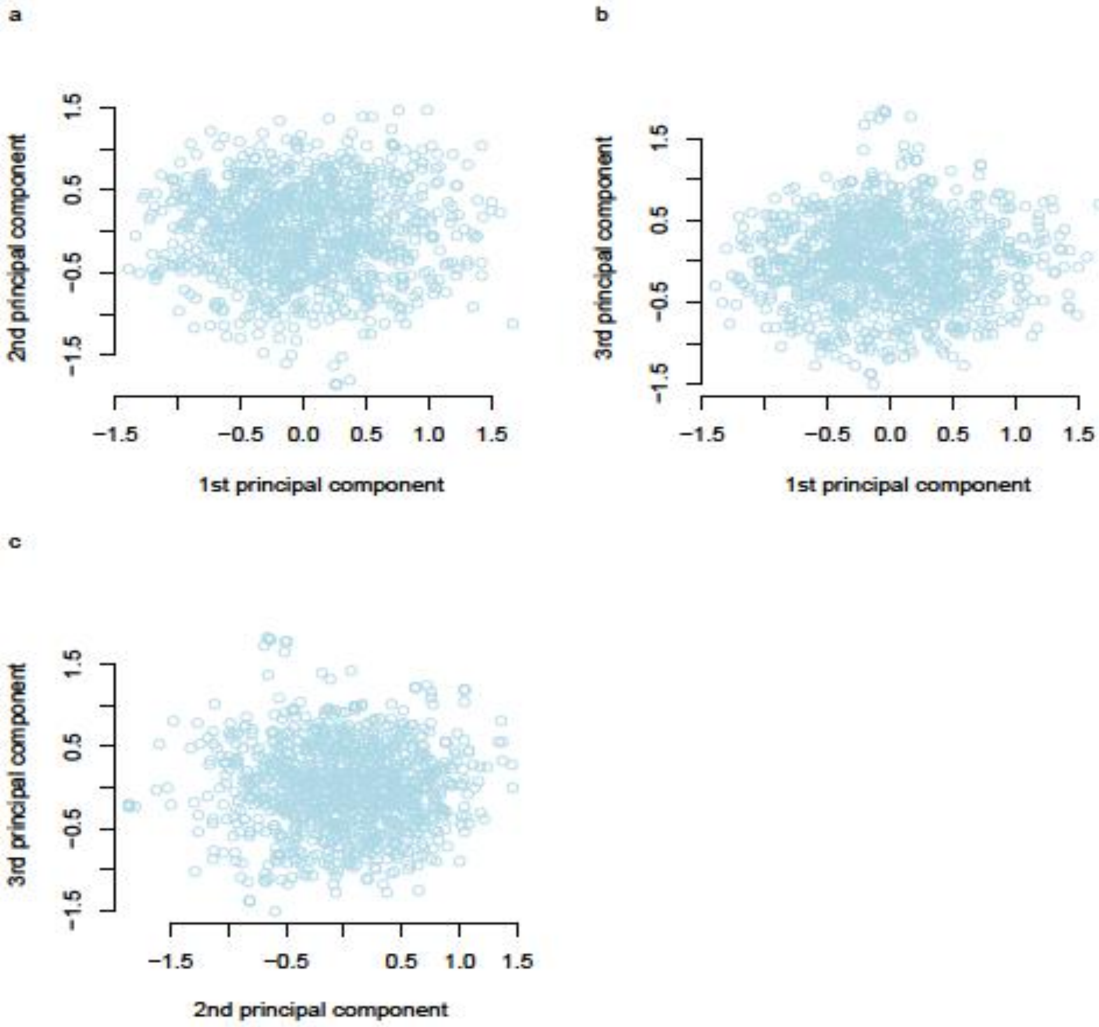


Figure 2.11: Pairwise plots of the first three principal components for Rachuonyo South *P. falciparum* parasite population.

A) 1st PC plotted against 2nd PC, **b)** 1st PC plotted against 3rd PC and **c)** 2nd PC plotted against 3rd PC. Each dot represents an individual sample. No obvious clustering of parasites based on genotypes is evident from these plots.

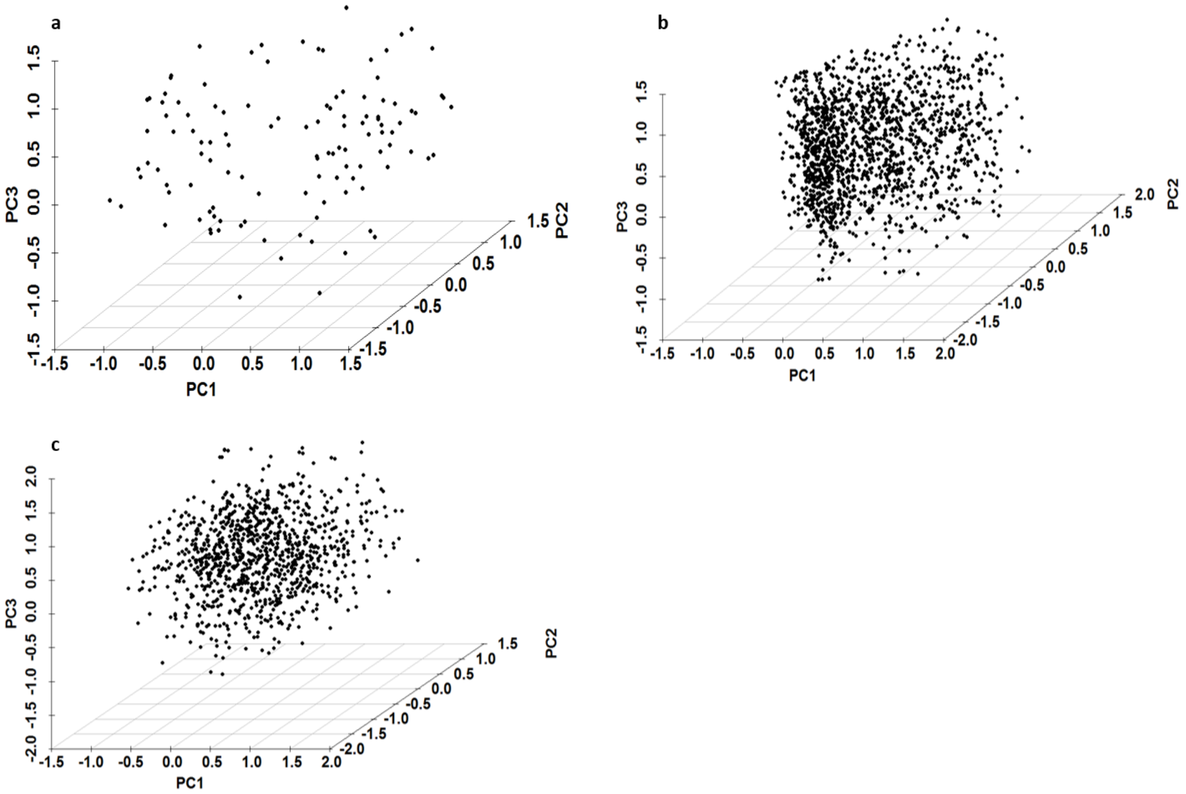


Figure 2.12: 3D Plots of the first three principal components.

Plots were generated for parasite populations in **a)** The Gambia, **b)** Kilifi and **c)** Rachuonyo South. Parasites clustered together and did not separate in space along any of the three principal components.

Scores for the first three PCs were represented on geographical maps of the study locations to show the spatial spread of the parasite isolates (figure 2.13 - figure 2.15). As with the PC plots above, the spatial pattern showed a high level of mixing of parasites, with no obvious clustering of genetically distinct parasites at the geographical scale analysed.

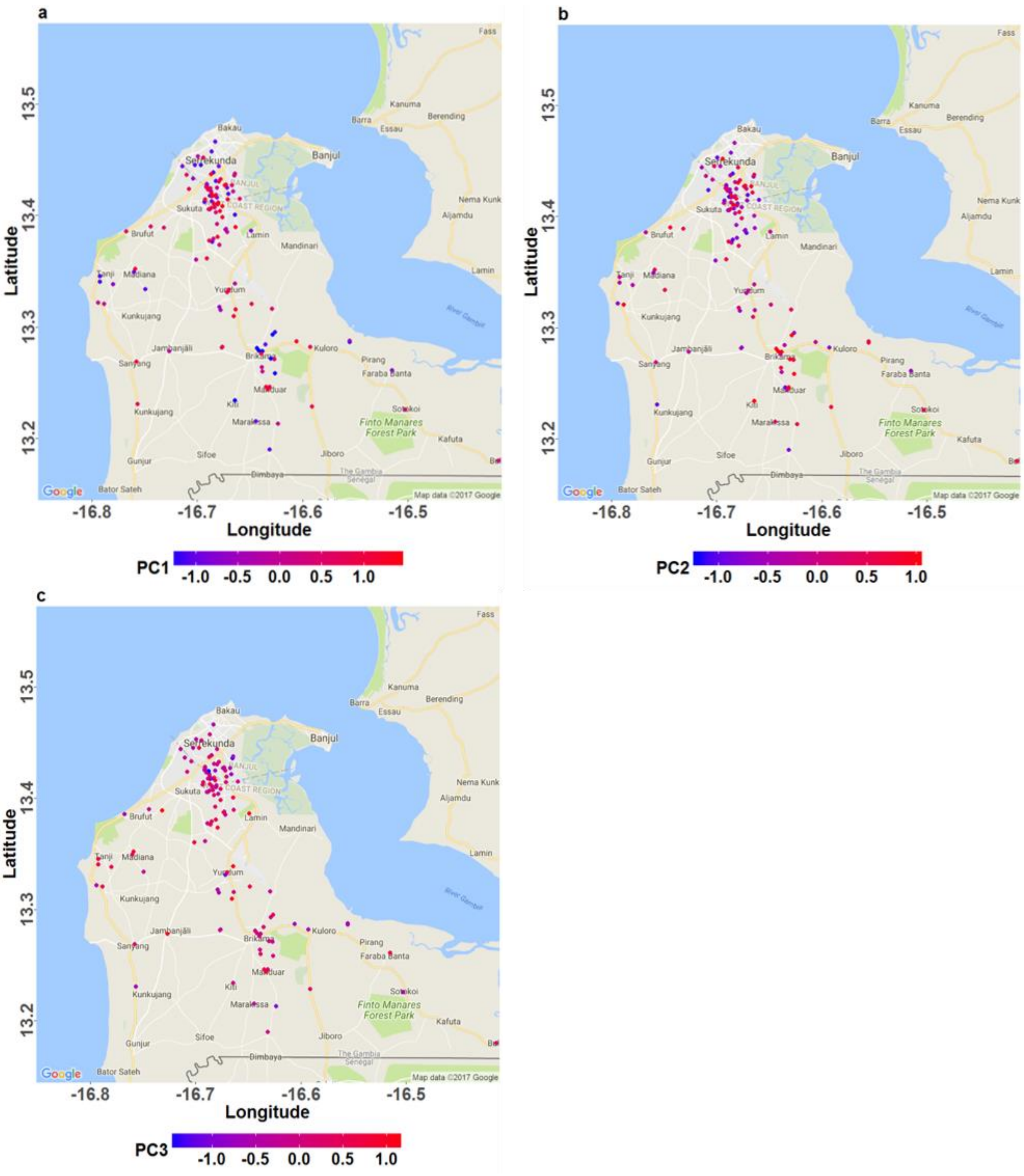


Figure 2.13: Geographic distribution of *P. falciparum* parasite genotypes in The Gambian population.

Each point represents the location of an individual parasite isolate and the colour shading represents distinct genotypes of parasites based on scores for **a)** Principal Component 1, **b)** Principal Component 2 and **c)** Principal Component 3.

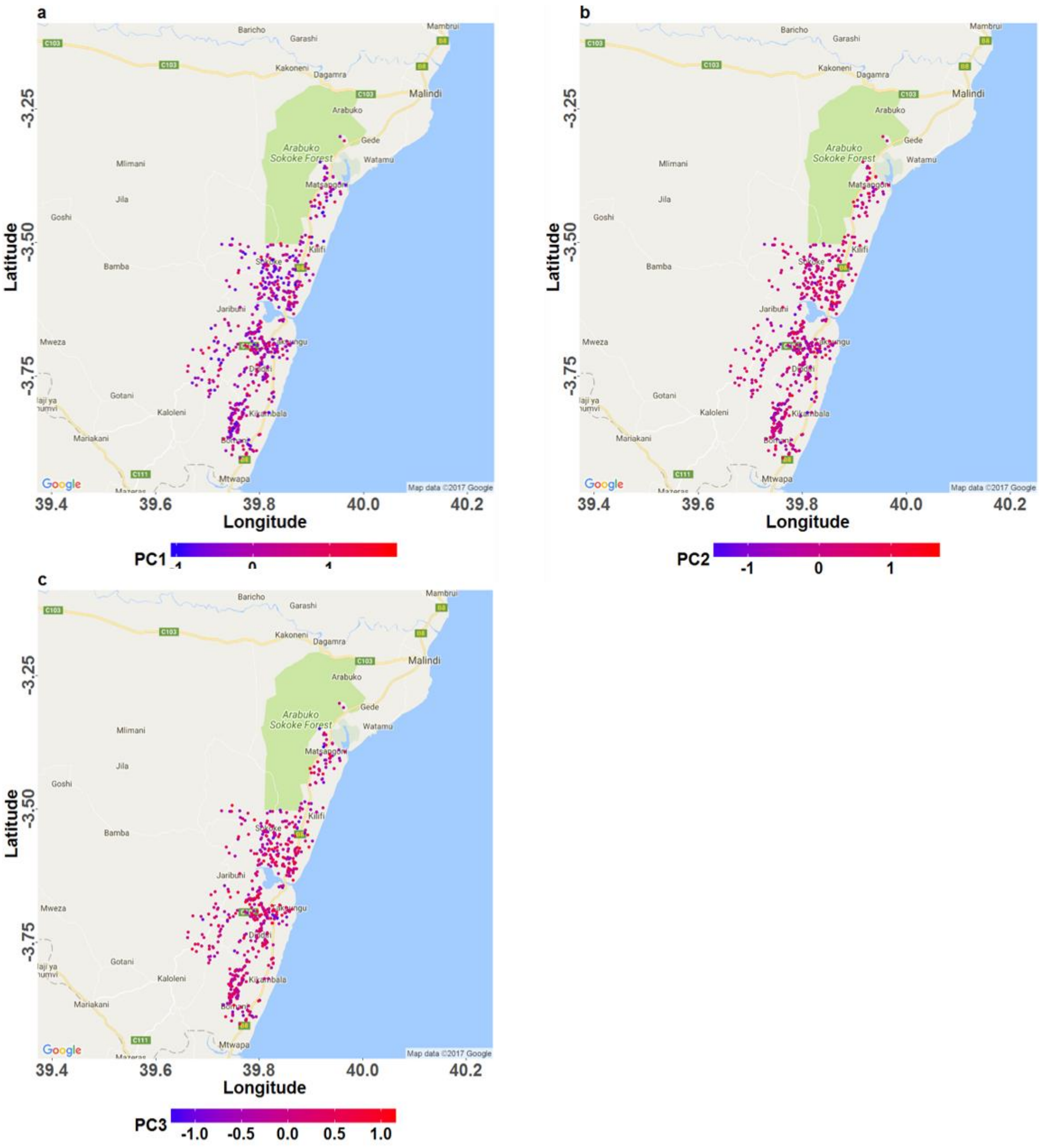


Figure 2.14: Geographic distribution of *P. falciparum* parasite genotypes in the Kilifi population.

Each point represents the location of an individual parasite isolate and the colour shading represents distinct genotypes of parasites based on scores for **a)** Principal Component 1, **b)** Principal Component 2 and **c)** Principal Component 3.

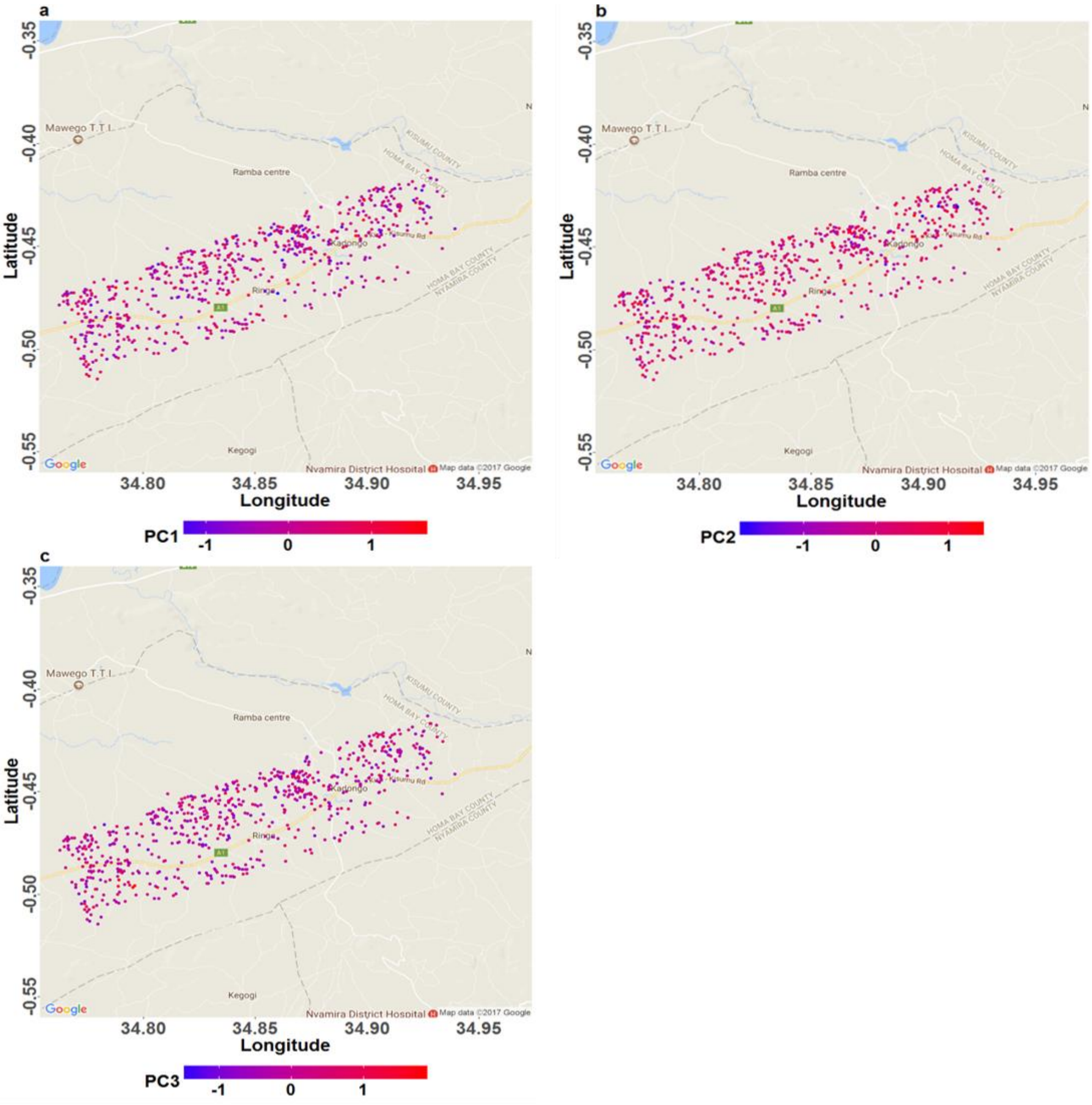


Figure 2.15: Geographic distribution of *P. falciparum* parasite genotypes in the Rachuonyo South population.

Each point represents the location of an individual parasite isolate and the colour shading represents distinct genotypes of parasites based on scores for **a)** Principal Component 1, **b)** Principal Component 2 and **c)** Principal Component 3.

Inter-population comparisons showed a high level of homogeneity, and parasites of East and West African origins could not be resolved based on genotype when 94 SNPs typed in The Gambia and Kilifi were analysed (figure 2.16). Furthermore, Kilifi and Rachuonyo South parasite populations could also not be resolved using the 57 SNPs that were successfully typed in the two populations (figure 2.17). Only 33 SNPs were successfully typed in both The Gambia and Rachuonyo South, and a comparison of these populations based on principal components is not reported due to the few number of SNPs.

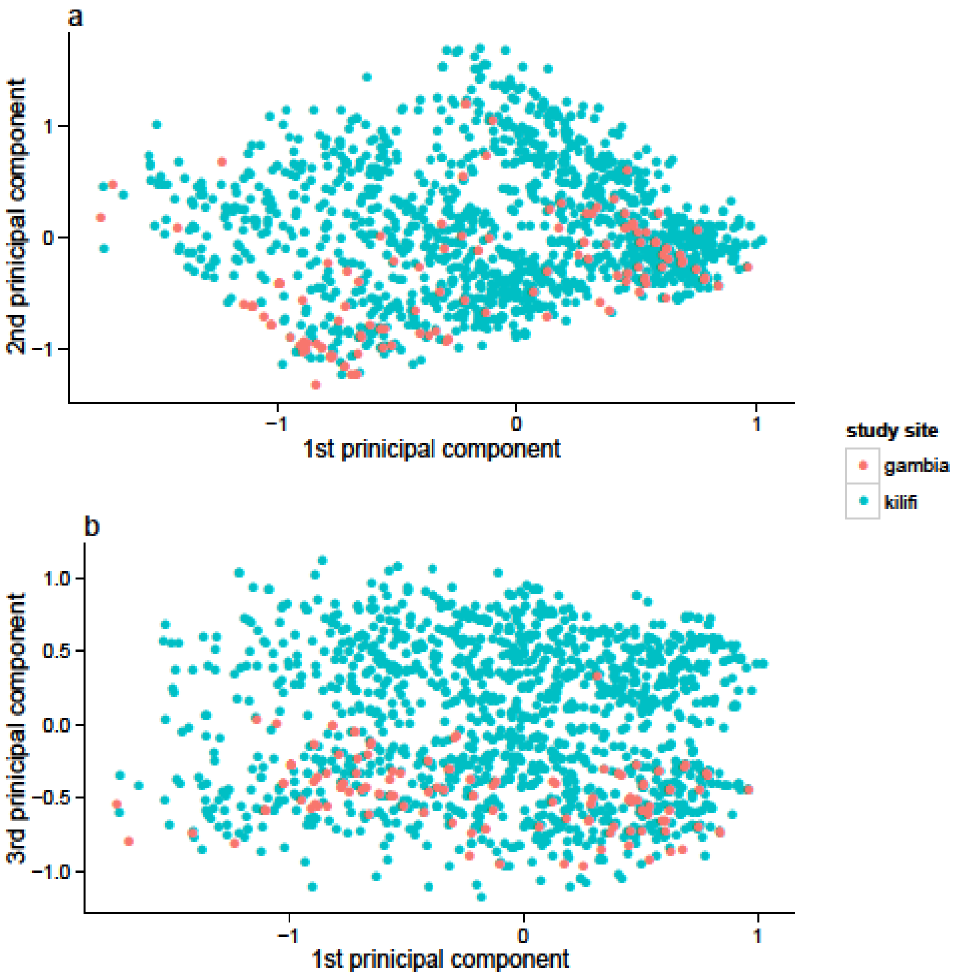


Figure 2.16: Principal Component Analysis plots of 133 Gambian (red) and 1602 Kilifi (green) *P. falciparum* samples based on 94 SNPs typed in both populations. Clear differentiation of samples based on country of origin was not observed when the analysis was carried out along (**a**) the first and second principal components or (**b**) along the first and third principal components.

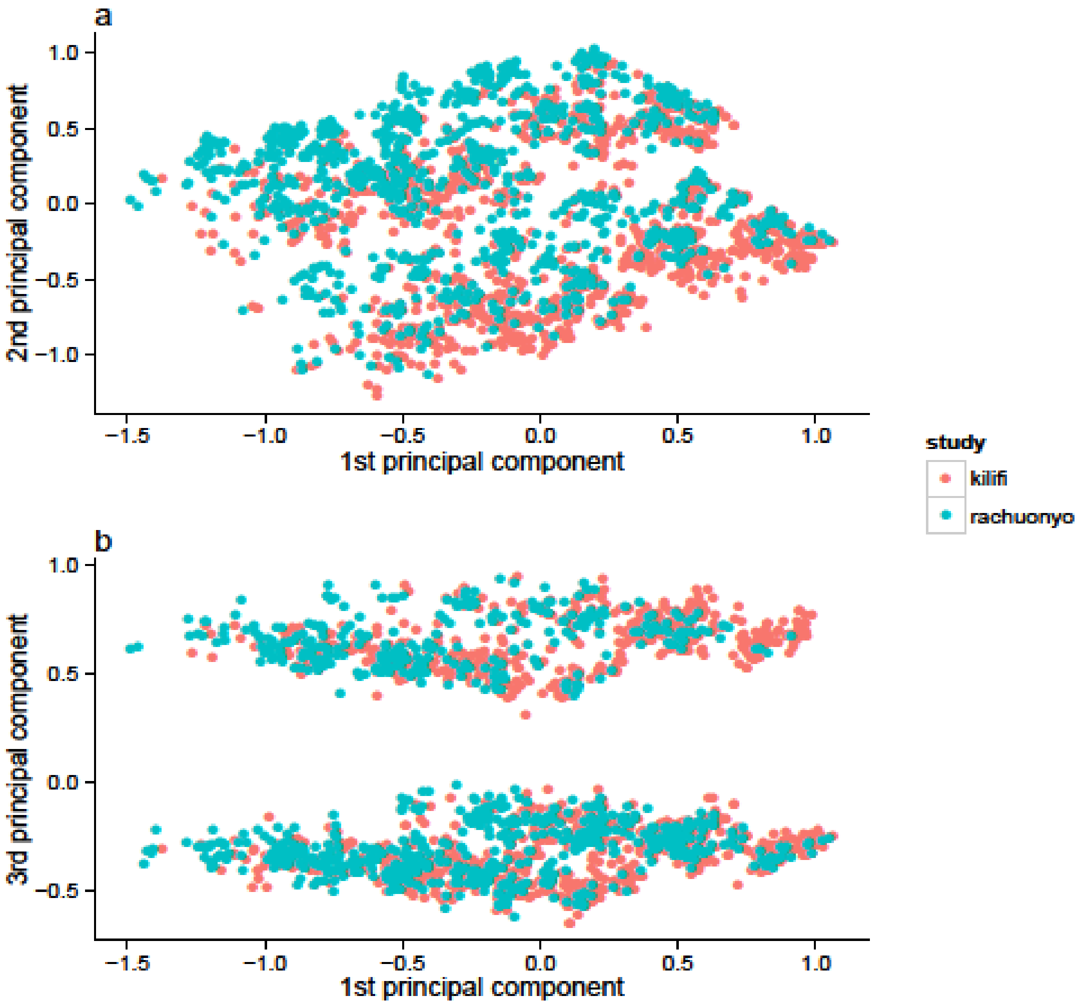


Figure 2.17: Principal Component Analysis plots of 1602 Kilifi (red) and 1034 Rachuonyo South (green) *P. falciparum* samples based on 57 SNPs typed in both populations.

(**a**) Clear differentiation of samples was not observed when the analysis was carried out along the first and second principal components. (**b**) Two clusters containing samples from both populations were observed when the analysis was carried out along the first and third principal components.

2.3.5 Moran's *I* spatial autocorrelation analyses

Having not seen parasite sub-populations by PCA alone, spatial analyses were included to test for spatial structure to the principal component values. Moran's *I* analysis for spatial autocorrelation showed slight positive correlations for parasites spaced over different distance classes of 1km, 2km and 5km. In the Gambian population, statistically significant (p<0.01) positive spatial autocorrelation was detected for parasites that were 6km apart within the 1km distance class, but there was no significant correlation for parasites that were more closely related in space within this distance class (figure 2.18). Statistically significant (p<0.01) spatial auto-correlation was seen for parasites that were 2km apart in the second distance class category while within the 5km distance class, parasites that were approximately 5km and 10km apart were spatially auto-correlated (p<0.01) as well. Only two samples came from the same homestead (0km apart) therefore no analysis was carried at out at this spatial resolution in this parasite population.

Spatially auto-correlated parasites were identified in the Kilifi population at distance classes of 1km, where statistically significant associations were seen at 1km, 3km, 4km, 5km and 6km intervals within this distance class (figure 2.19). Significant associations were also seen for samples that were spaced much further apart (up to 20km apart) showing that parasites move quite freely within this region. In the 2km distance classes, statistically significant spatial auto-correlations were observed for samples that were 2km, 4km and 6km apart, as well as those that were slightly over 20km apart. Within the 5km distance class, most of the associations were very weak and non-significant, although there was one significant association (p<0.01) for parasites that were 5km apart.

A slightly different trend was seen in the Rachuonyo South population (figure 2.20) where parasites that were more closely related in space tended to have a statistically significant (p<0.01) association. This was seen for parasites collected from the same homesteads (0km apart), as well as those that were 1km apart. In some instances, however, statistically significant associations were seen for parasites that were more distantly spaced (over 10km apart).

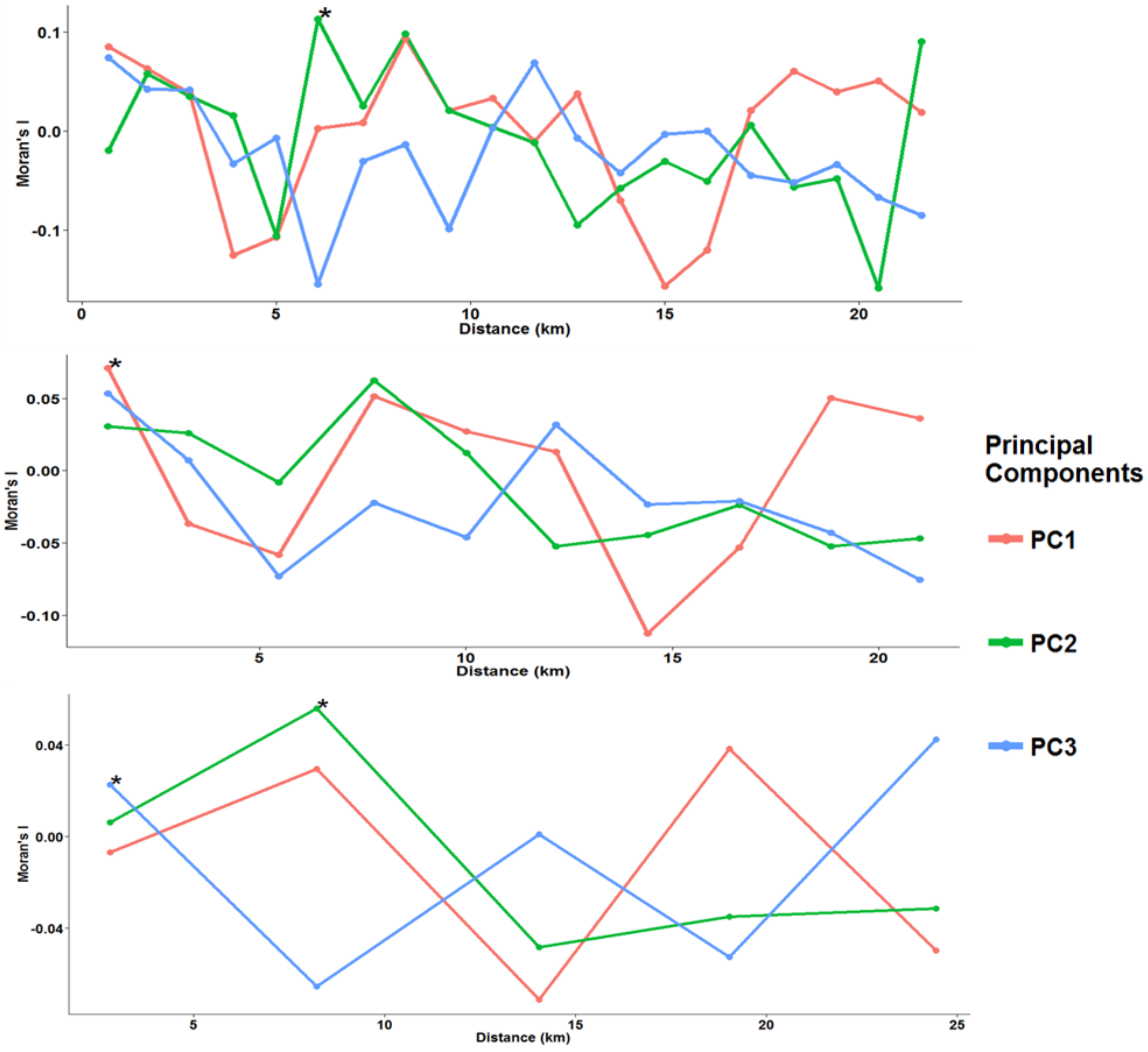


Figure 2.18: Moran’s *I* spatial autocorrelation analysis for the first three principal components in The Gambian *P. falciparum* parasite population.

Correlation coefficients were computed at distance classes of (from top) 1km, 2km and 5km. Asterisks indicate distances at which parasites have significant (p≤0.01) autocorrelations. The lines represent the correlation coefficients obtained when the different principal components were used. Moran’s *I* was not computed for samples collected at the same location (0 km distance difference) due to the few number of samples (2).

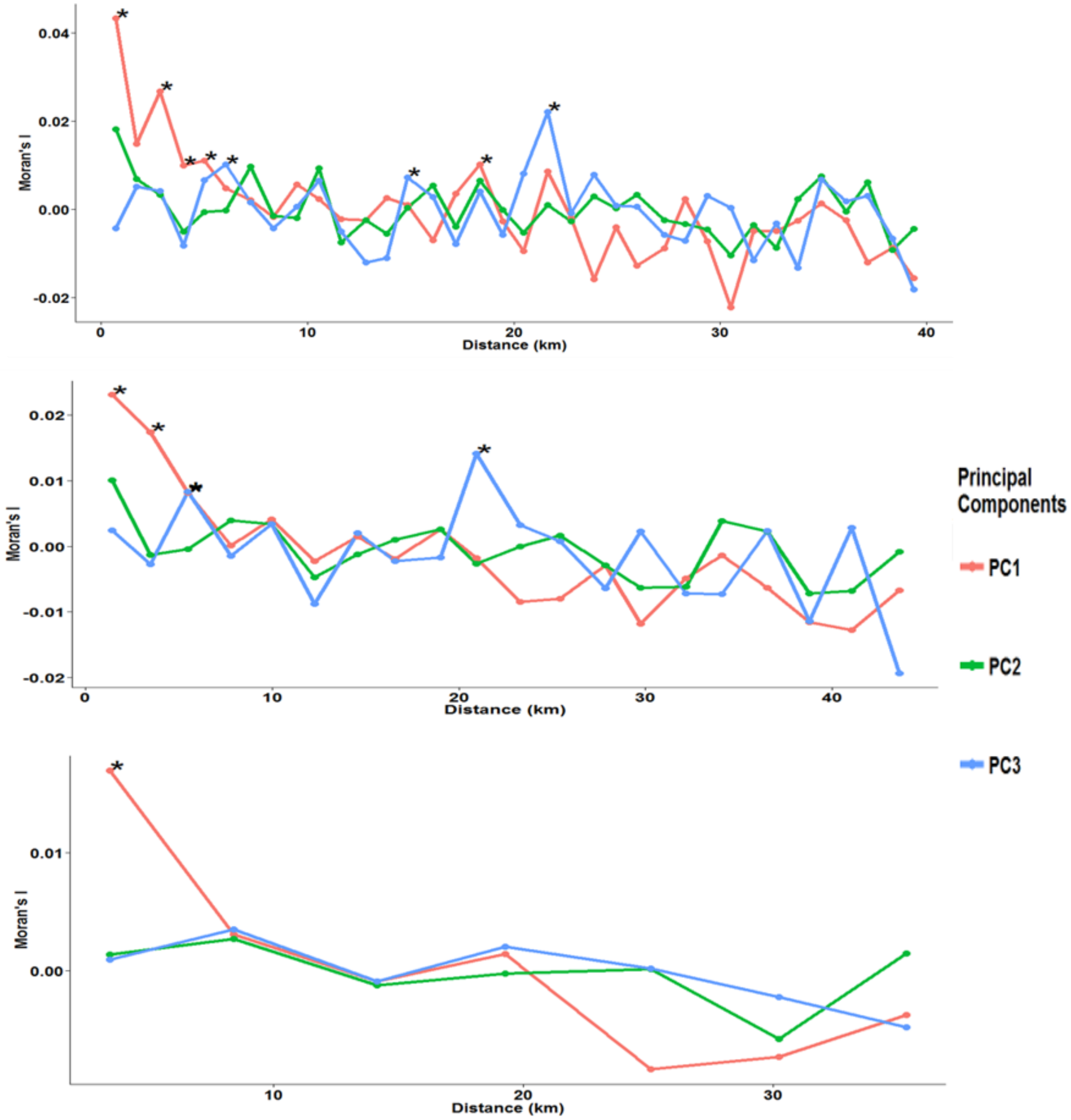


Figure 2.19: Moran's *I* spatial autocorrelation analysis for the first three principal components in Kilifi *P. falciparum* parasite population.

Correlation coefficients were computed at distance classes of (from top) 1km, 2km and 5km. Asterisks indicate distances at which parasites have significant (p≤0.01) autocorrelations. The lines represent the correlation coefficients obtained when the different principal components were used.

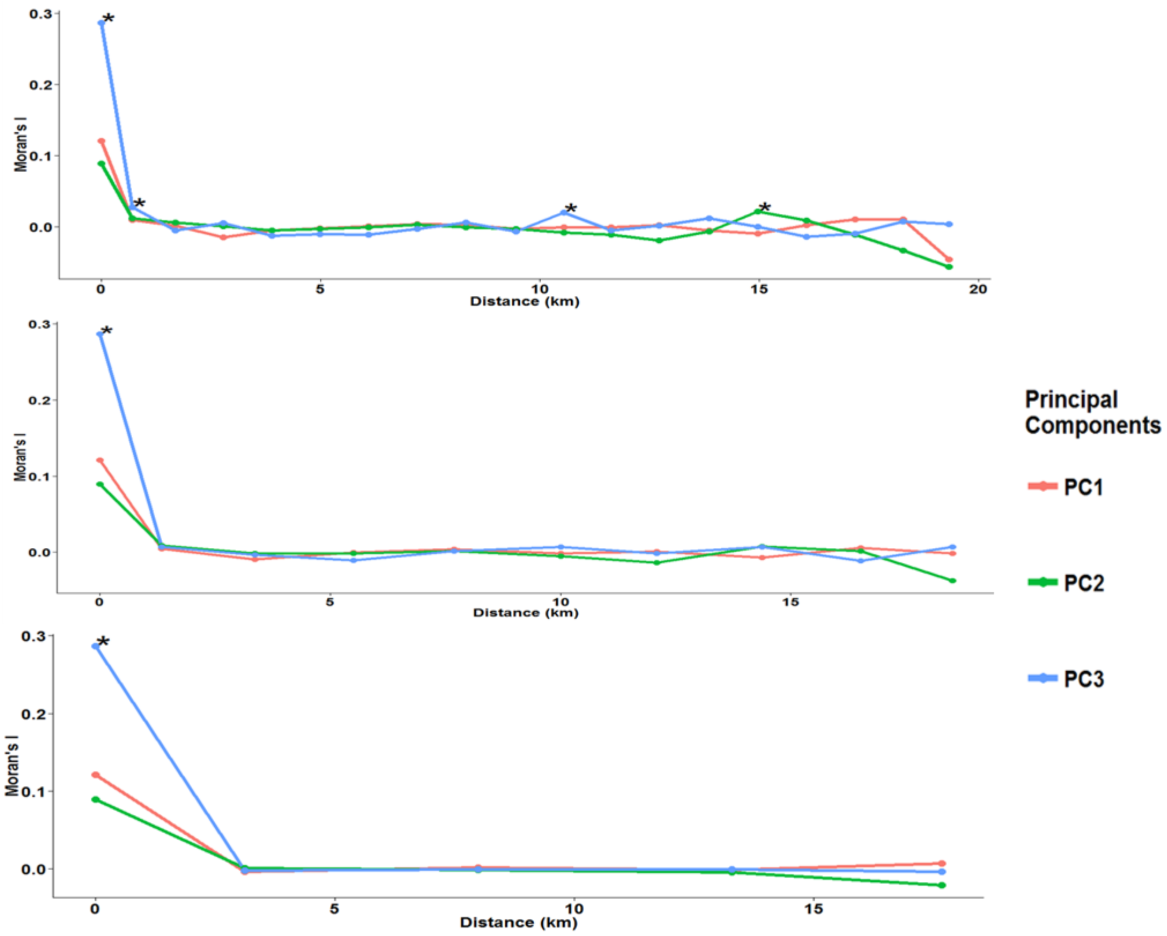


Figure 2.20: Moran's *I* spatial autocorrelation analysis for the first three principal components in Rachuonyo South *P. falciparum* parasite population.

Correlation coefficients were computed at distance classes of (from top) 1km, 2km and 5km. Asterisks indicate distances at which parasites have significant (p≤0.01) autocorrelations. The lines represent the correlation coefficients obtained when the different principal components were used.

2.3.6 SaTScan analysis

Spatial scan statistics based on a normal probability distribution model identified statistically significant (p≤0.01) clusters of genetically distinct *P. falciparum* sub-populations in Kilifi and Rachuonyo South study sites (figure 2.21). In Rachuonyo South, one cluster with a 0.5km radius (p=0.001), containing 14 of the parasite isolates, was detected when analysing the third PC, while in Kilifi, a larger cluster of 15 genetically related parasites was detected with a radius of 1.54km (p=0.011) when analysing the second PC. No statistically significant clusters were detected in The Gambian parasite population.

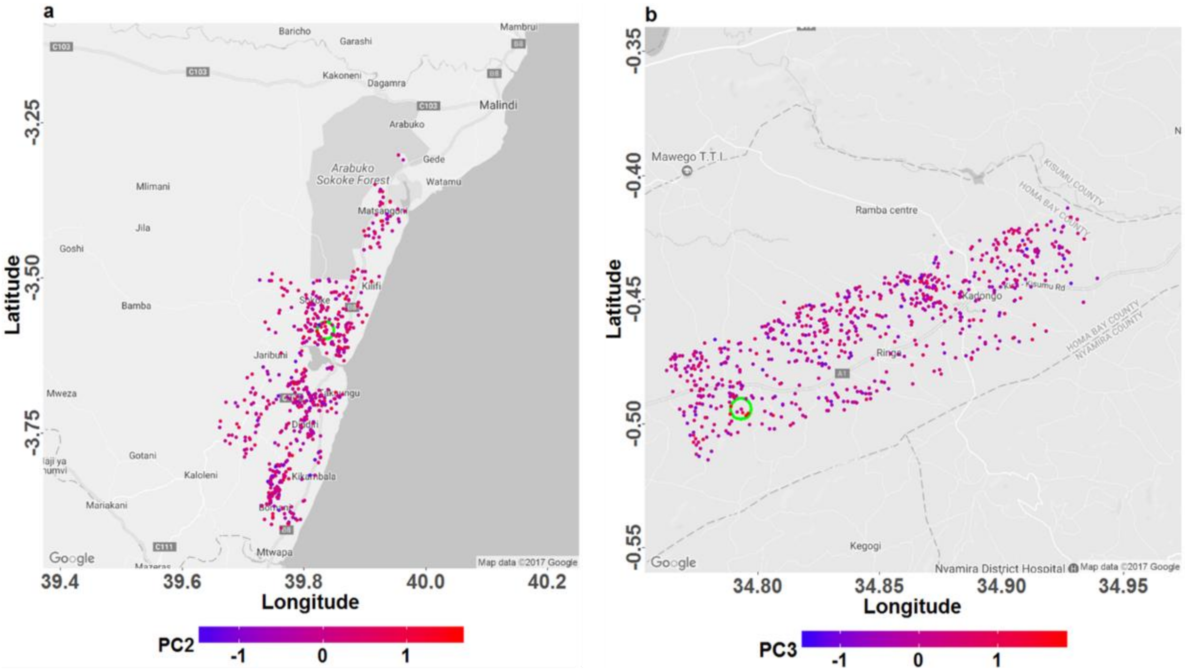


Figure 2.21: Spatial clusters of *P. falciparum* sub-populations in Kilifi and Rachuonyo South populations as identified by spatial scan statistics.

One cluster was identified in Kilifi parasite population (**a**) when the second principal component was analysed and another cluster was identified in the Rachuonyo South parasite population (**b**) when the third principal component was analysed. Each dot represents a sample analysed in the study, mapped against the geographical location where it was sampled. Number of samples in Kilifi cluster = 15; number of samples in Rachuonyo South cluster = 14.

2.3.7 Spatial and temporal changes in *P. falciparum* genetic variation

The effects of distance and time separating parasite pairs on genetic relatedness were examined to determine the spatial extent and rate of parasite mixing (figure 2.22 – figure 2.24). The analyses used regression models where the number of SNP differences between parasite pairs was an outcome predicted by the distance between parasite pairs and the time between parasite pairs. Time was not included for the Rachuonyo South population as the samples were collected in a single cross-sectional survey taken over a few days. Across all three datasets, distance was independently associated with increasing variation in genotype, i.e. the further apart in space any two parasites were, the greater the number of SNP differences between them (Table 2.7). In the Gambian and Kilifi populations, time was also shown to be associated with increasing variation in genotype, with parasite pairs collected further apart in time having a greater number of genetic differences than those collected closer to each other in time. Additionally, in The Gambia and Kilifi populations, time interacted antagonistically with distance to attenuate the effect of distance on genotype relatedness when the time separating samples was greater. This means that the genetic differences between any two parasites increased with distance, but at a decreasing rate when time between these samples increased. This pattern was consistent for all groupings of SNPs (all SNPs, “other” SNPs, *EBA 175* and *AMA1*), with the exception of *AMA1* SNPs in Kilifi, where power was limited due to the low number of SNPs analysed. SNPs were not grouped by gene in Rachuonyo South due to the low number of SNPs typed.

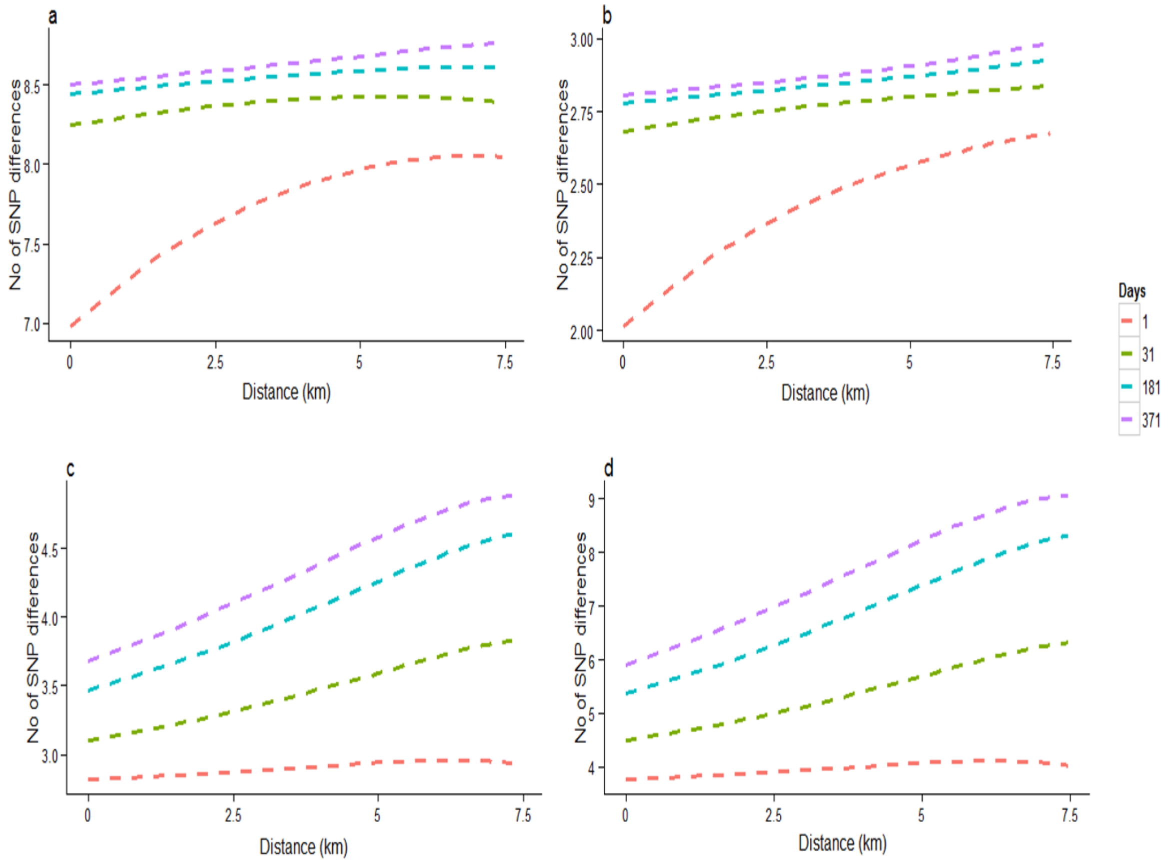


Figure 2.22: Time-distance interaction curves showing the effect of distance on the number of SNP differences between Gambian *P. falciparum* parasite pairs with increasing time.

The analyses were carried out for **a)** all SNPs, **b)** *EBA-175*, **c)** *AMA1* and **d)** "other" SNPs in The Gambian population. Dashed lines represent time intervals separating parasite pairs at 1 day (red), 1 month (green), 6 months (blue) and 1 year (purple).

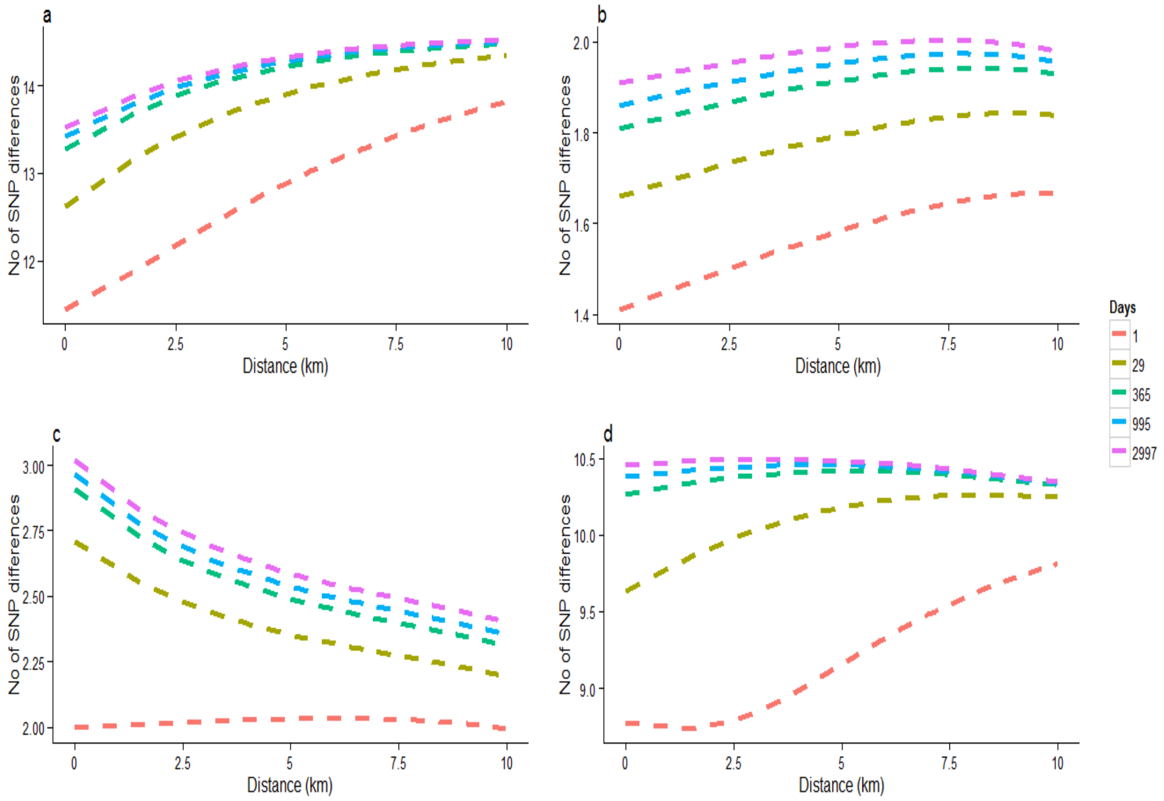


Figure 2.23: Time-distance interaction curves showing the effect of distance on the number of SNP differences between Kilifi *P. falciparum* parasite pairs with increasing time.

The analyses were carried out for **a)** all SNPs, **b)** *EBA-175*, **c)** *AMA1* and **d)** "other" SNPs in the Kilifi population. Dashed lines represent time intervals separating parasite pairs at 1 day (red), 1 month (green), 6 months (blue) and 1 year (purple).

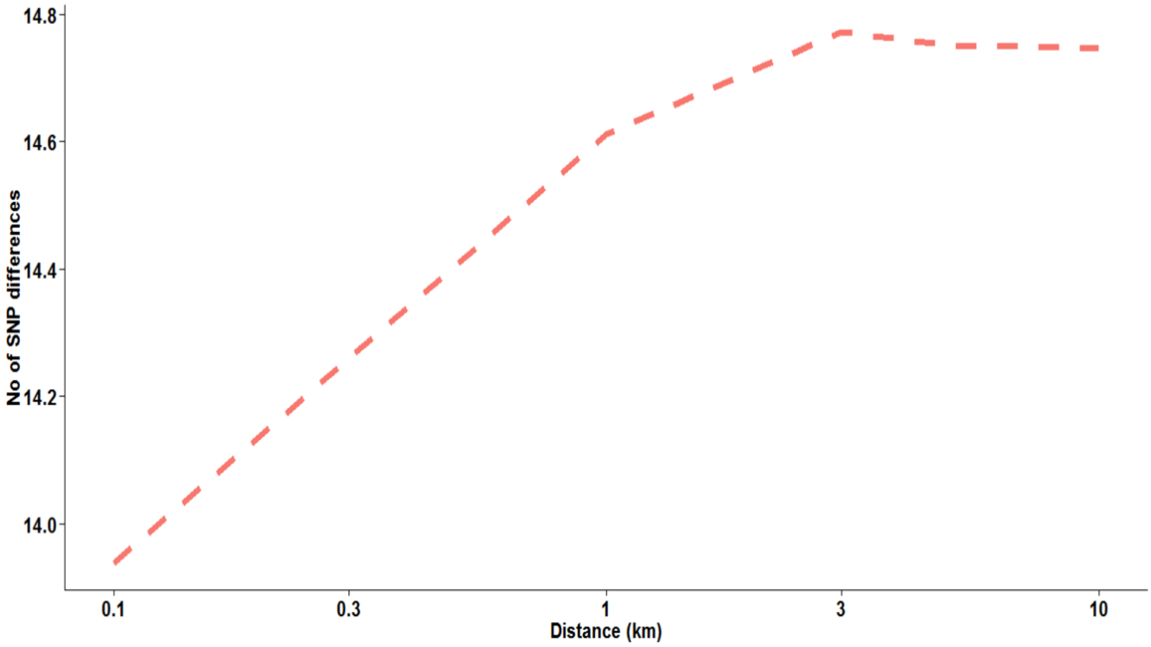


Figure 2.24: Effect of distance on the number of SNP differences between *P. falciparum* parasite pairs in Rachuonyo South.

The analysis was carried out using all the SNPs in the population. The effect of time was not considered because all the samples were collected within a few days of each other.

Bootstrapping the analyses (to account for the linked nature of pairwise observations) gave statistically significant effects of distance, time and the interaction between distance and time on variations in parasite genotypes (Table 2.7).

Table 2-7: 95% bootstrap confidence intervals of the linear effects of time, distance and the interaction of time and distance on changes in SNP differences between *P. falciparum* parasite pairs.

	Time (p value)	**Distance (p value)**	**Distance-Time interaction (p value)**
Kilifi	0.190 – 0.647 (<0.001)	0.297 – 1.363 (0.001)	-0.453 – -0.072 (0.003)
Rachuonyo South	-	0.0104 – 0.275 (0.018)	-
The Gambia	-0.005 - -0.001 (0.004)	0.086 - 0.723 (<0.001)	0.0003 – 0.002 (0.003)

Values represent the change in the number of SNP differences between parasite pairs per day (time), per kilometre (distance) and per day/kilometre (time-distance interaction). Time and distance were log transformed prior to running the regression analyses.

2.3.8 Correlation between *P. falciparum* population genetics and transmission intensity based on different metrics of transmission

Raster analysis by pixels was conducted to examine the spatial relationship between distinct parasite genotypes as represented by principal component scores and either malaria positive fraction (MPF) data (in Kilifi) or PCR positive data (in Rachuonyo South). The range of MPF and parasite prevalence per pixel varied depending on the size of the pixels analysed. In the Kilifi population, MPF ranged from 0 – 100% (0.5km pixels), 0 – 100% (1.0km pixels), 20 – 83% (2.0km pixels) and 33 – 63% (4.0km pixels). In the Rachuonyo South population, PCR positive prevalence varied from 0 – 75% (0.5km pixels), 0 – 47% (1.0km pixels), 3.5 – 35.8% (2km pixels) and 6.2 – 33.4% (4.0km pixels).

Raster analyses showed an overall trend of heterogeneity in both populations, at all spatial resolutions (figure 2.25- figure 2.32). High or low mean PC scores are indicative of parasites that are more closely genetically related within a defined geographical region, while medium mean PC values may be indicative of a mixture of parasites that are less closely related within a geographical region of a similar size. Geographical regions with higher MPF or PCR positive values indicate a higher number of individuals with malaria or asymptomatic parasite infection, respectively. These regions can be classified as hotspots of symptomatic or asymptomatic infection as they have higher infection prevalence compared to surrounding regions. The analysis of principal components did not show any consistent or statistically strong associations with markers of transmission intensity (i.e. malaria positive fraction and prevalence of asymptomatic parasitaemia by PCR).

Spearman’s rank correlation coefficients of association indicated a statistically significant, albeit weak negative association between the mean of principal components and malaria positive fractions at spatial resolutions of 1km x 1km ρ(313)=-0.13, p=0.02

and 2km x 2km, ρ(152)=-0.18, p=0.02 in the Kilifi population (Table 2.8). Within the Rachuonyo South population, clustering was observed at lower resolution of 4km x 4km, with a statistically significant positive association between mean scores of PCR positive fractions and high principal component scores ρ(12)=0.68, p=0.007 (Table 2.8).

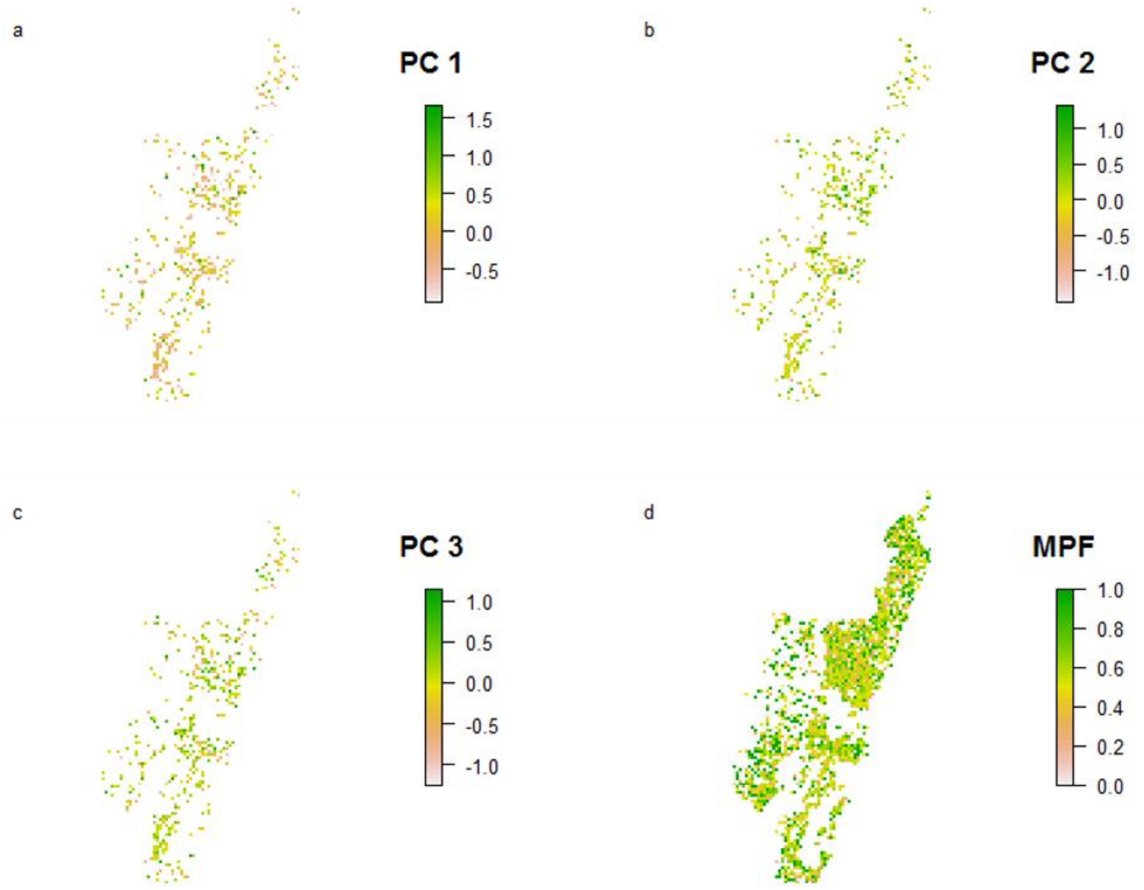


Figure 2.25: Raster analysis by pixels to determine the spatial relationship between *P. falciparum* genotypes as represented by principal component analysis (PCA) and malaria positive fraction over 0.5km^2 in Kilifi.

The study area was divided into 0.5km x 0.5km sized pixels and each pixel assigned the mean PC score or the malaria positive fraction (MPF) of all samples falling within it. (**a**), (**b**) and (**c**) show the distribution of scores for principal components 1, 2 and 3 respectively. (**d**) shows the distribution of MPF across the study site at the specified spatial scale.

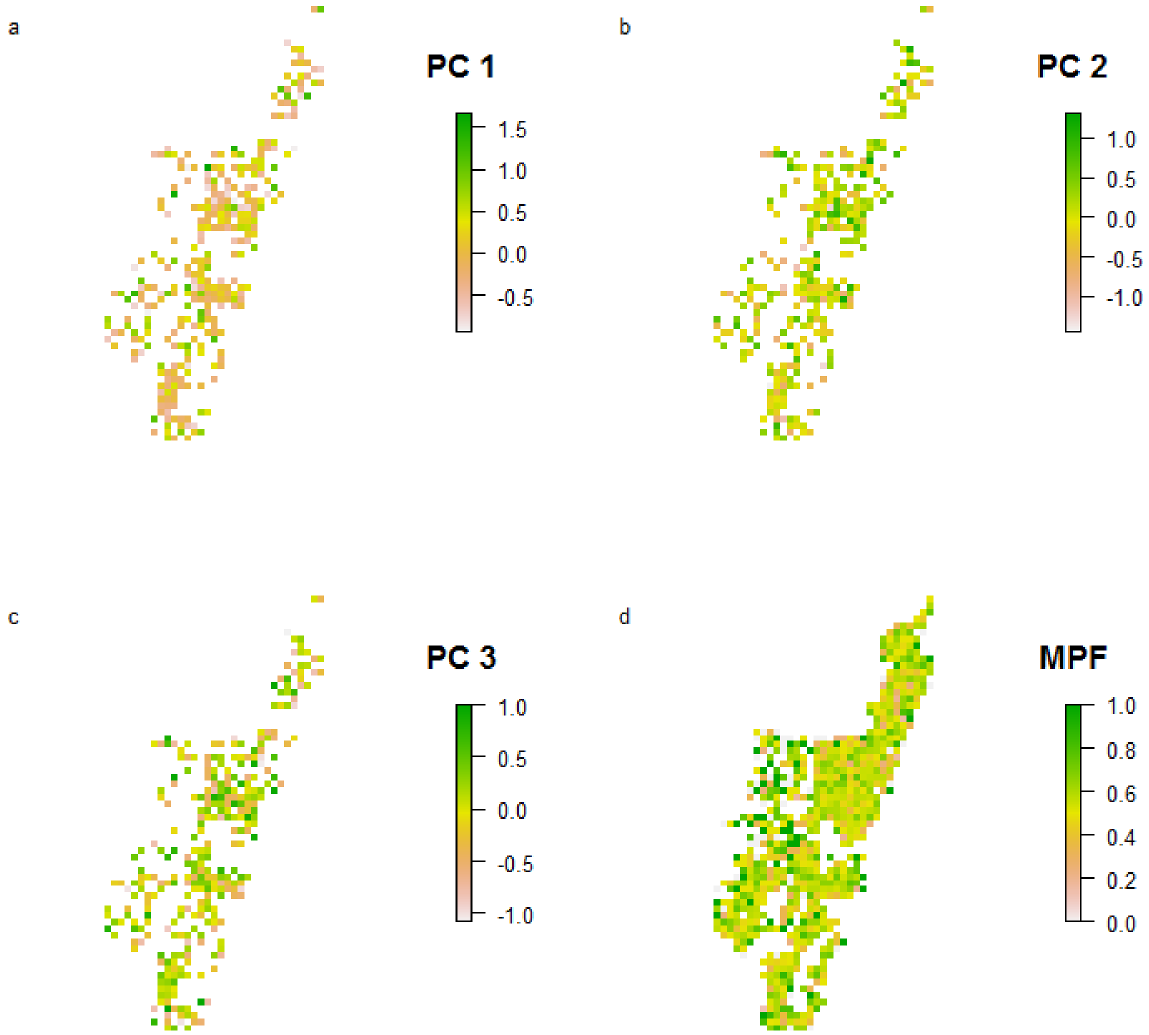


Figure 2.26: Raster analysis by pixels to determine the spatial relationship between *P. falciparum* genotypes as represented by principal component analysis (PCA) and malaria positive fraction over 1.0km^2 in Kilifi.

The study area was divided into 1.0km x 1.0km sized pixels and each pixel assigned the mean PC score or the malaria positive fraction (MPF) of all samples falling within it. (**a**), (**b**) and (**c**) show the distribution of scores for principal components 1, 2 and 3 respectively. (**d**) shows the distribution of MPF across the study site at the specified spatial scale.

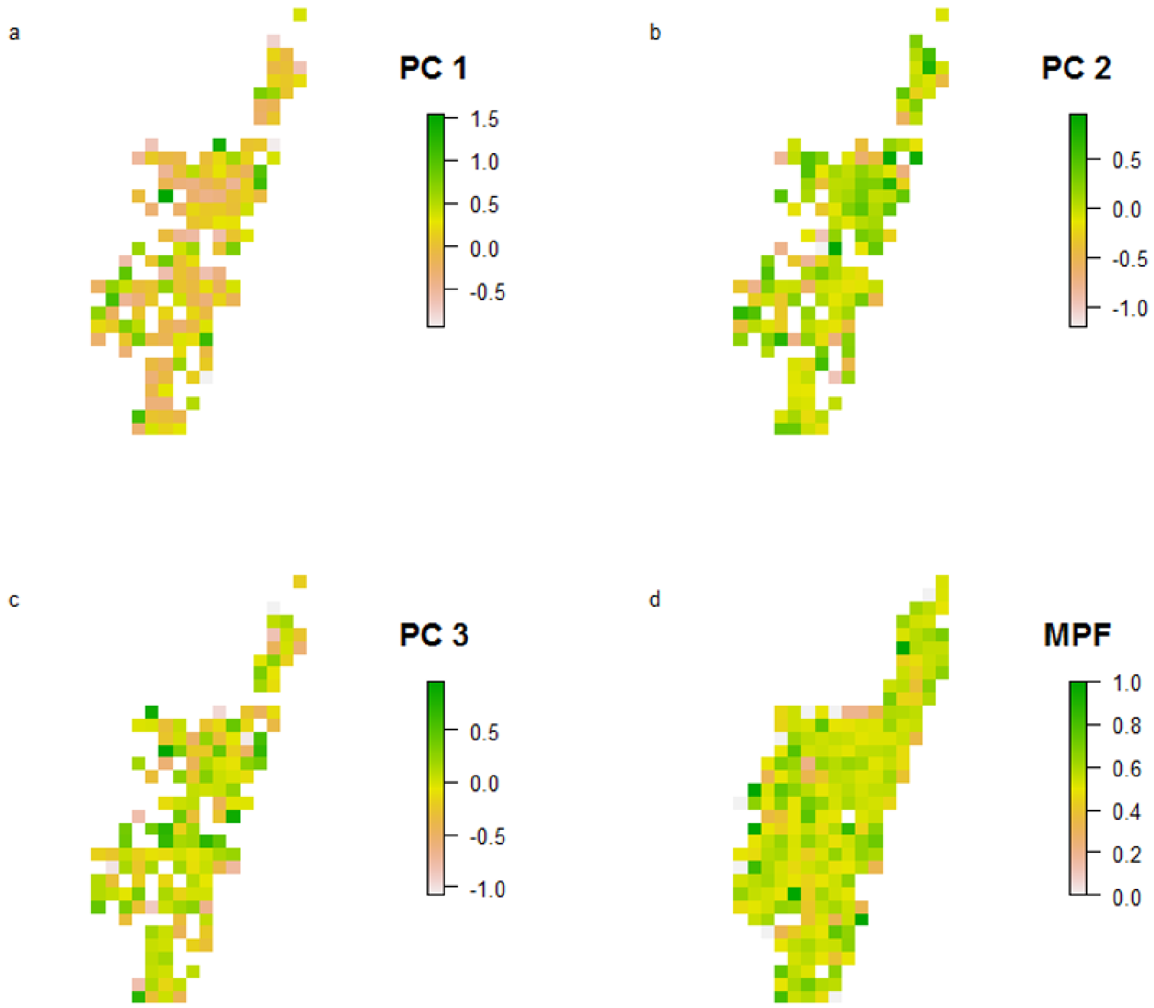


Figure 2.27: Raster analysis by pixels to determine the spatial relationship between *P. falciparum* genotypes as represented by principal component analysis (PCA) and malaria positive fraction over 2.0km^2 in Kilifi.

The study area was divided into 2.0km x 2.0km sized pixels and each pixel assigned the mean PC score or the malaria positive fraction (MPF) of all samples falling within it. (**a**), (**b**) and (**c**) show the distribution of scores for principal components 1, 2 and 3 respectively. (**d**) shows the distribution of MPF across the study site at the specified spatial scale.

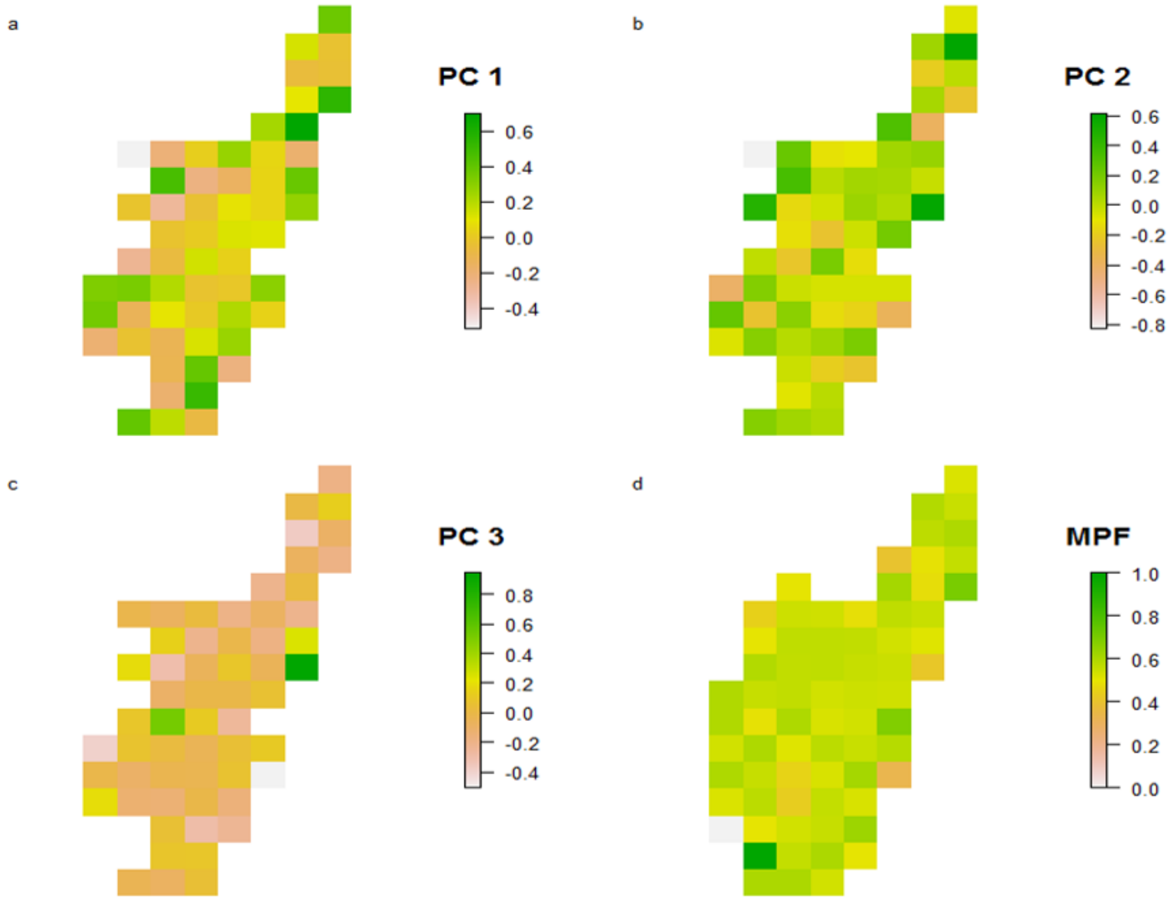


Figure 2.28: Raster analysis by pixels to determine the spatial relationship between *P. falciparum* genotypes as represented by principal component analysis (PCA) and malaria positive fraction over 4.0km^2 in Kilifi.

The study area was divided into 4.0km x 4.0km sized pixels and each pixel assigned the mean PC score or the malaria positive fraction (MPF) of all samples falling within it. (**a**), (**b**) and (**c**) show the distribution of scores for principal components 1, 2 and 3 respectively. (**d**) shows the distribution of MPF across the study site at the specified spatial scale.

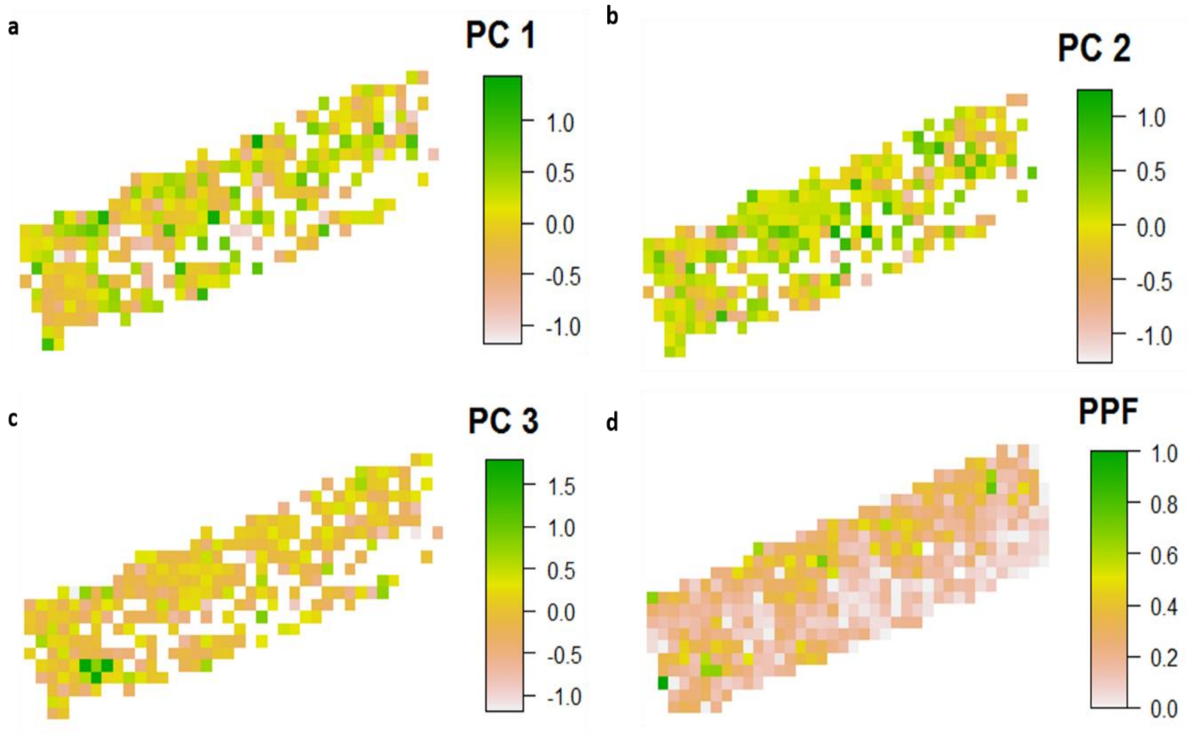


Figure 2.29: Raster analysis by pixels to determine the spatial relationship between *P. falciparum* genotypes as represented by principal component analysis (PCA) and parasite positive fraction over 0.5km^2 in Rachuonyo South.

The study area was divided into 0.5km x 0.5km sized pixels and each pixel assigned the mean PC score or the parasite positive fraction (PPF) of all samples falling within it. (**a**), (**b**) and (**c**) show the distribution of scores for principal components 1, 2 and 3 respectively. (**d**) shows the distribution of PPF across the study site at the specified spatial scale.

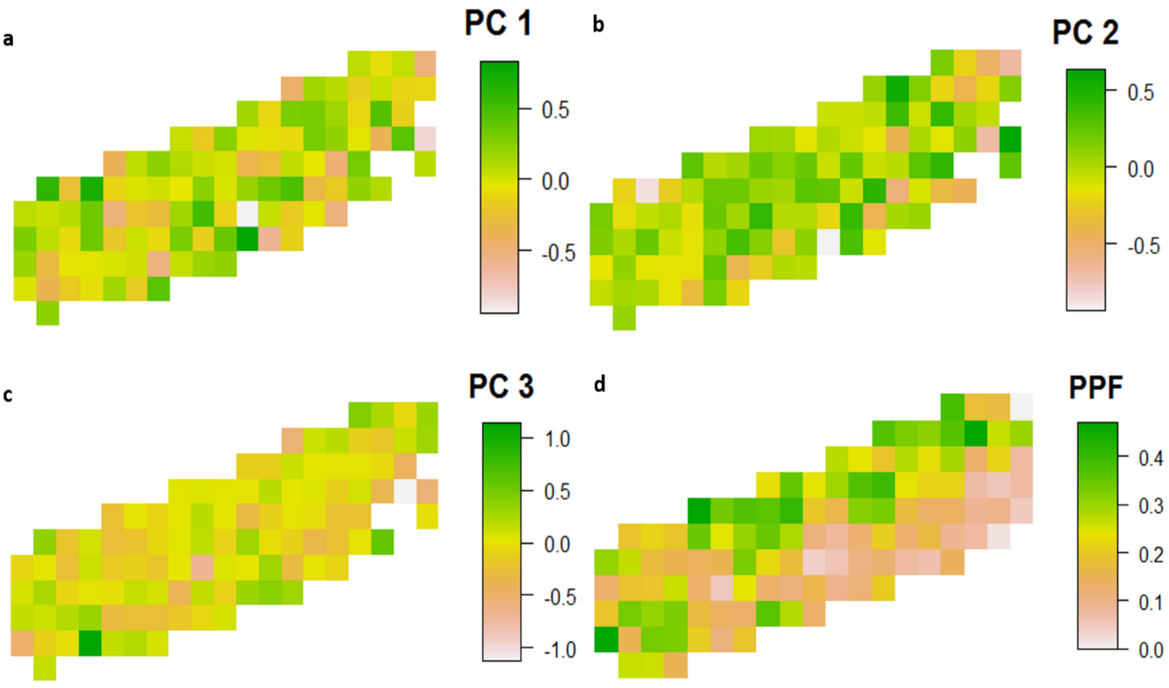


Figure 2.30: Raster analysis by pixels to determine the spatial relationship between *P. falciparum* genotypes as represented by principal component analysis (PCA) and parasite positive fraction over 1.0km^2 in Rachuonyo South.

The study area was divided into 1.0km x 1.0km sized pixels and each pixel assigned the mean PC score or the parasite positive fraction (PPF) of all samples falling within it. (**a**), (**b**) and (**c**) show the distribution of scores for principal components 1, 2 and 3 respectively. (**d**) shows the distribution of PPF across the study site at the specified spatial scale.

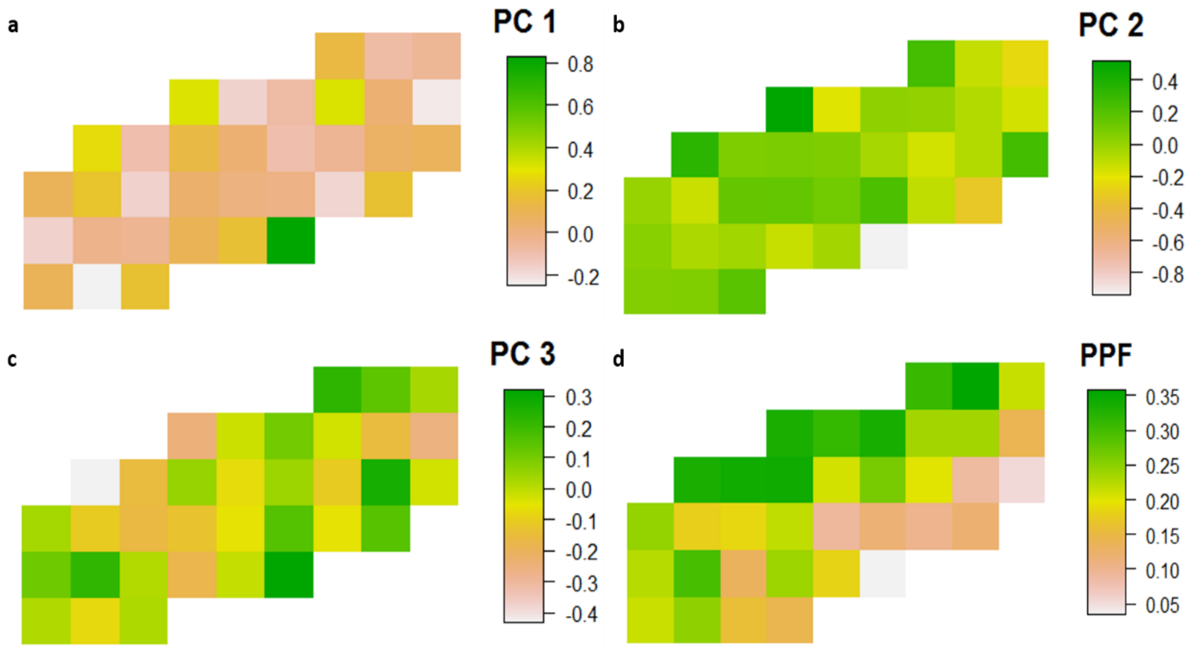


Figure 2.31: Raster analysis by pixels to determine the spatial relationship between *P. falciparum* genotypes as represented by principal component analysis (PCA) and parasite positive fraction over 2.0km^2 in Rachuonyo South.

The study area was divided into 2.0km x 2.0km sized pixels and each pixel assigned the mean PC score or the parasite positive fraction (PPF) of all samples falling within it. (**a**), (**b**) and (**c**) show the distribution of scores for principal components 1, 2 and 3 respectively. (**d**) shows the distribution of PPF across the study site at the specified spatial scale.

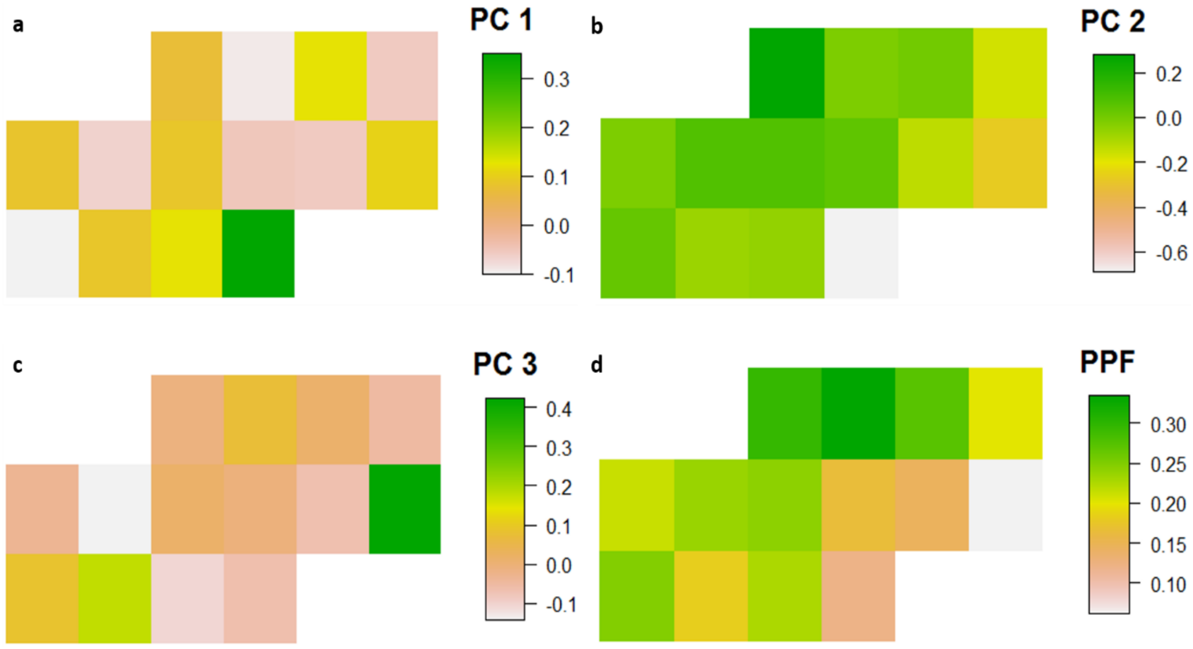


Figure 2.32: Raster analysis by pixels to determine the spatial relationship between *P. falciparum* genotypes as represented by principal component analysis (PCA) and parasite positive fraction over 4.0km^2 in Rachuonyo South.

The study area was divided into 4.0km x 4.0km sized pixels and each pixel assigned the mean PC score or the parasite positive fraction (PPF) of all samples falling within it. (**a**), (**b**) and (**c**) show the distribution of scores for principal components 1, 2 and 3 respectively. (**d**) shows the distribution of PPF across the study site at the specified spatial scale.

Table 2.8: Spearman's rank ordered correlation coefficients between parasite genotypes and infection status in Kilifi and Rachuonyo South.

Kilifi (MPF)	**PC1**	**PC2**	**PC3**
0.5KM (n=432)	-0.07 (0.14)	0.03 (0.53)	0.04 (0.43)
1.0KM (n=313)	**-0.13 (0.02)**	0.03 (0.58)	0.09 (0.11)
2.0KM (n=152)	**-0.18 (0.02)**	0.14 (0.08)	0.01 (0.87)
4.0KM (n=57)	0.04 (0.74)	0.17 (0.21)	-0.01 (0.96)
Rachuonyo South (PPF)	PC1	PC2	PC3
0.5KM (n=272)	0.01 (0.88)	0.02 (0.72)	0.10 (0.11)
1.0KM (n=104)	0.01 (0.93)	0.08 (0.43)	0.03 (0.73)
2.0KM (n=32)	-0.11 (0.61)	0.19 (0.27)	-0.10 (0.57)
4.0KM (n=12)	-0.32 (0.27)	**0.68 (0.007)**	0.10 (0.74)

n= degrees of freedom; p-values in brackets under each PC column. MPF = malaria positive fraction; PPF = Parasite positive fraction.

Scatterplots representing the relationship between the distribution of mean PC scores and malaria positive fractions (Kilifi population) on the one hand, and PCR positive fractions (Rachuonyo South population) on the other, were produced at the 1.0km x 1.0km spatial resolution (figures 2.33 and 2.34).

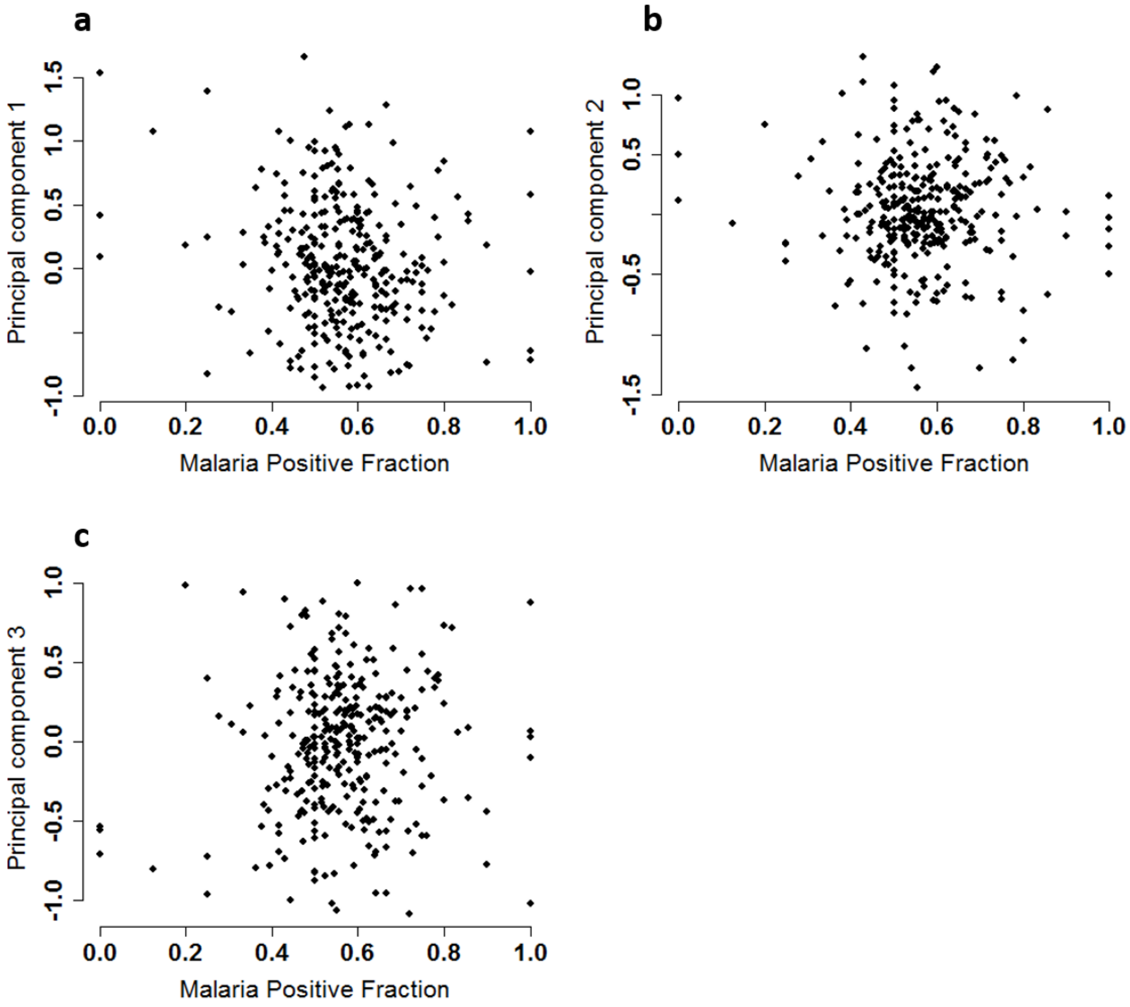


Figure 2.33: Association between mean PC scores and malaria positive fraction in Kilifi *P. falciparum* parasite population.

Scores were computed within 1km x 1km geographical grids to show the relationship between malaria positive fractions and parasite genotypes represented by scores for **a)** Principal component 1, **b)** Principal component 2 and **c)** Principal component 3.

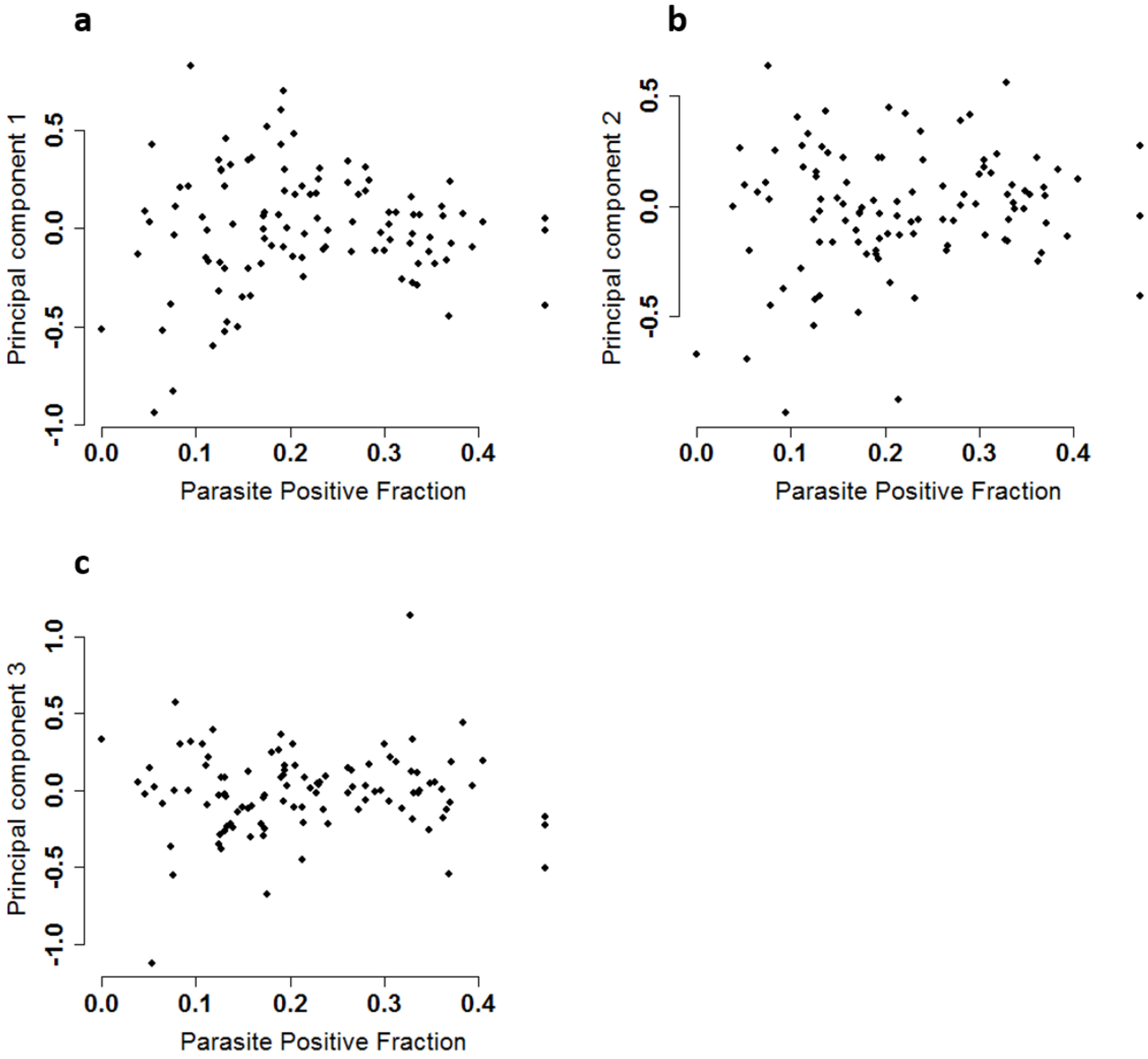


Figure 2.34: Associations between mean PC scores and PCR positive fractions in Rachuonyo South *P. falciparum* parasite population.

Scores were computed within 1km x 1km geographical grids to show the relationship between parasite positive fractions and parasite genotypes represented by scores for **a)** Principal component 1, **b)** Principal component 2 and **c)** Principal component 3.

2.3.9 Analysis of spatial barriers to parasite movement

Raster analysis was also used to examine the study sites for discrete regions that acted as spatial barriers to parasite movement over short distances. This was done by dividing each site into pixels and analysing the number of SNP differences between parasites separated by each pixel. The idea was that pixels that act as barriers to parasite movement and mixing would separate parasites that had a higher number of SNP differences, and conversely, pixels that acted as “gateways” for parasite movement would separate parasites with much fewer SNP differences between them. The analyses were carried out separately in each of the three sites. Additionally, a separate analysis was carried out for samples collected in Junju location in Kilifi county, where all the community survey samples were collected, to determine if there were spatial barriers to parasite movement in this defined region.

Bootstrapping the multivariable linear regression analysis of pairwise comparisons of samples for SNP differences using 189, 703, 340 and 77 pixels for The Gambia, Kilifi, Rachuonyo South and Junju, respectively, showed that the majority of pixels were not significant influences on SNP differences (figure 2.35 – figure 2.38). The few pixels that were significant ($p<0.05$) were not significant after applying Bonferroni correction to account for multiple testing. Furthermore, the distribution of p values was uniform for each dataset (mean p value ~0.5), implying that the null hypothesis of there being no spatial barriers to parasite movement could not be rejected.

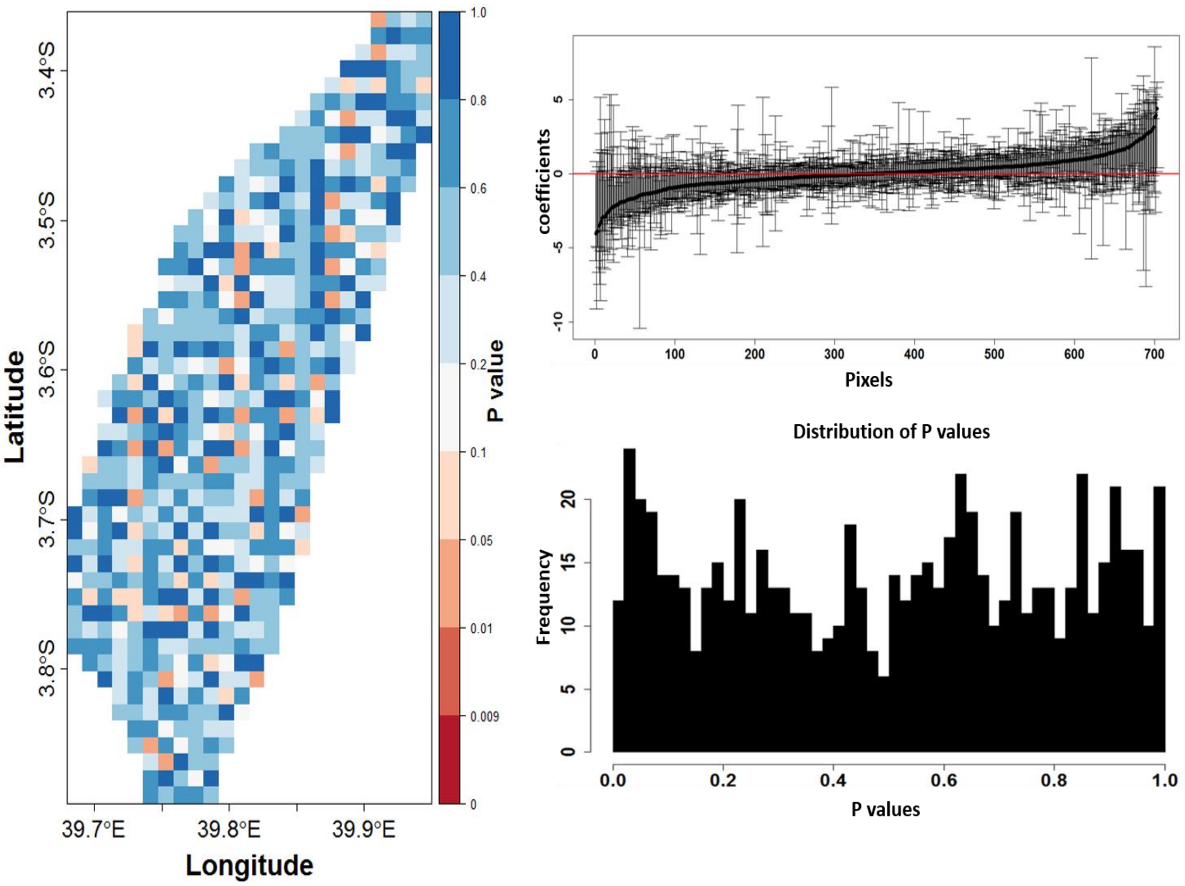


Figure 2.35: Raster analysis by pixels to examine the presence of spatial barriers to parasite movement in Kilifi.

The pixel plot represents p values of bootstrapped linear regression correlation coefficients and show the significance of different geographical locations in acting as barriers to parasite mixing at a spatial scale of 1km x 1km. The colour key indicates the range of p values from > 0 to 1. Significant p values shown on the plot were not significant after applying Bonferroni correction to account for multiple testing. Accompanying the map are a plot showing the 95% confidence interval around the coefficient estimates (with a red line drawn through coefficient estimate 0) and a histogram showing the distribution of bootstrap p values following 1000 resampling steps.

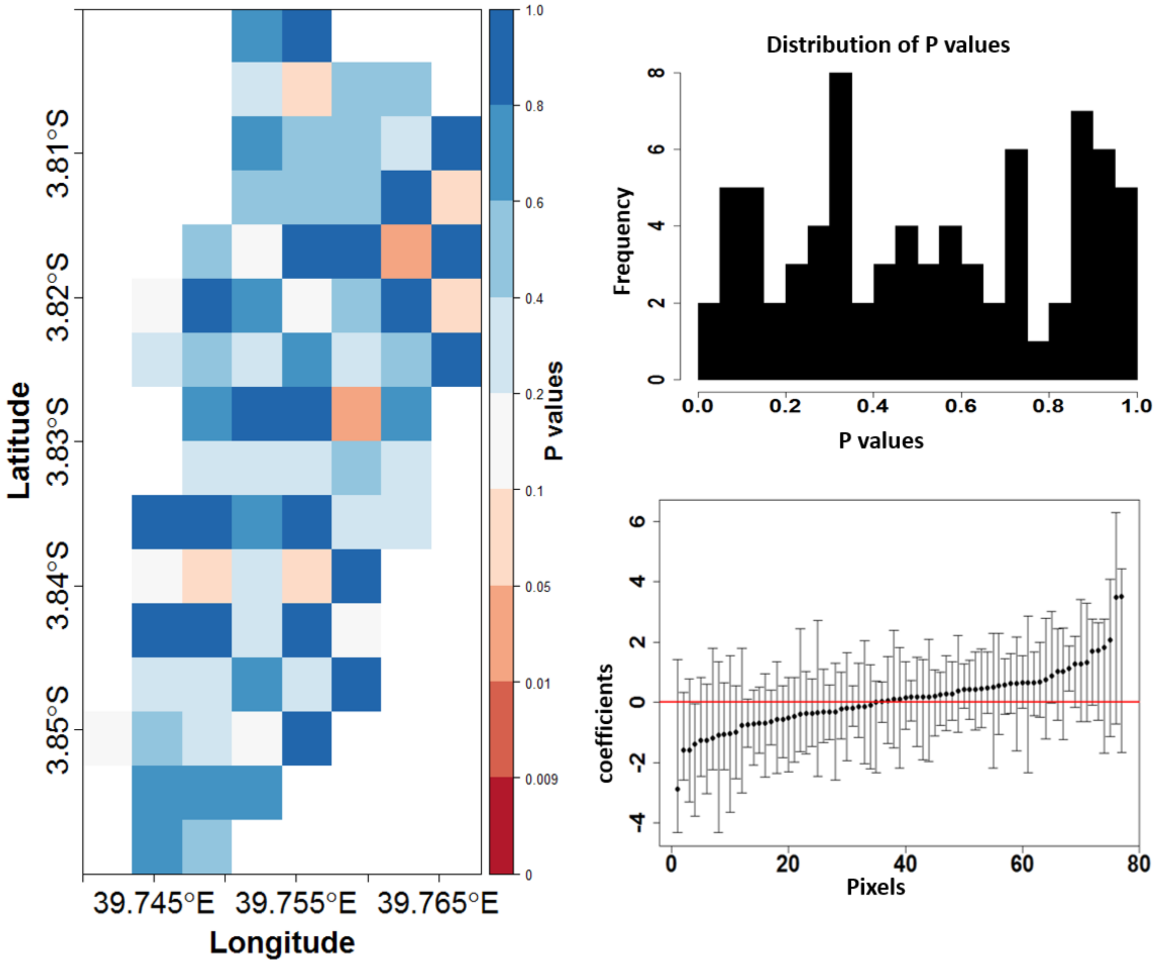


Figure 2.36: Raster analysis by pixels to examine the presence of spatial barriers to parasite movement in Junju location, Kilifi county.

The pixel plot represents p values of bootstrapped linear regression correlation coefficients and show the significance of different geographical locations in acting as barriers to parasite mixing at a spatial scale of 0.4km by 0.4km. The colour key indicates the range of p values from > 0 to 1. Significant p values shown on the plot were not significant after applying Bonferroni correction to account for multiple testing. Accompanying the map are a plot showing the 95% confidence interval around the coefficient estimates (with a red line drawn through coefficient estimate 0) and a histogram showing the distribution of bootstrap p values following 1000 resampling steps.

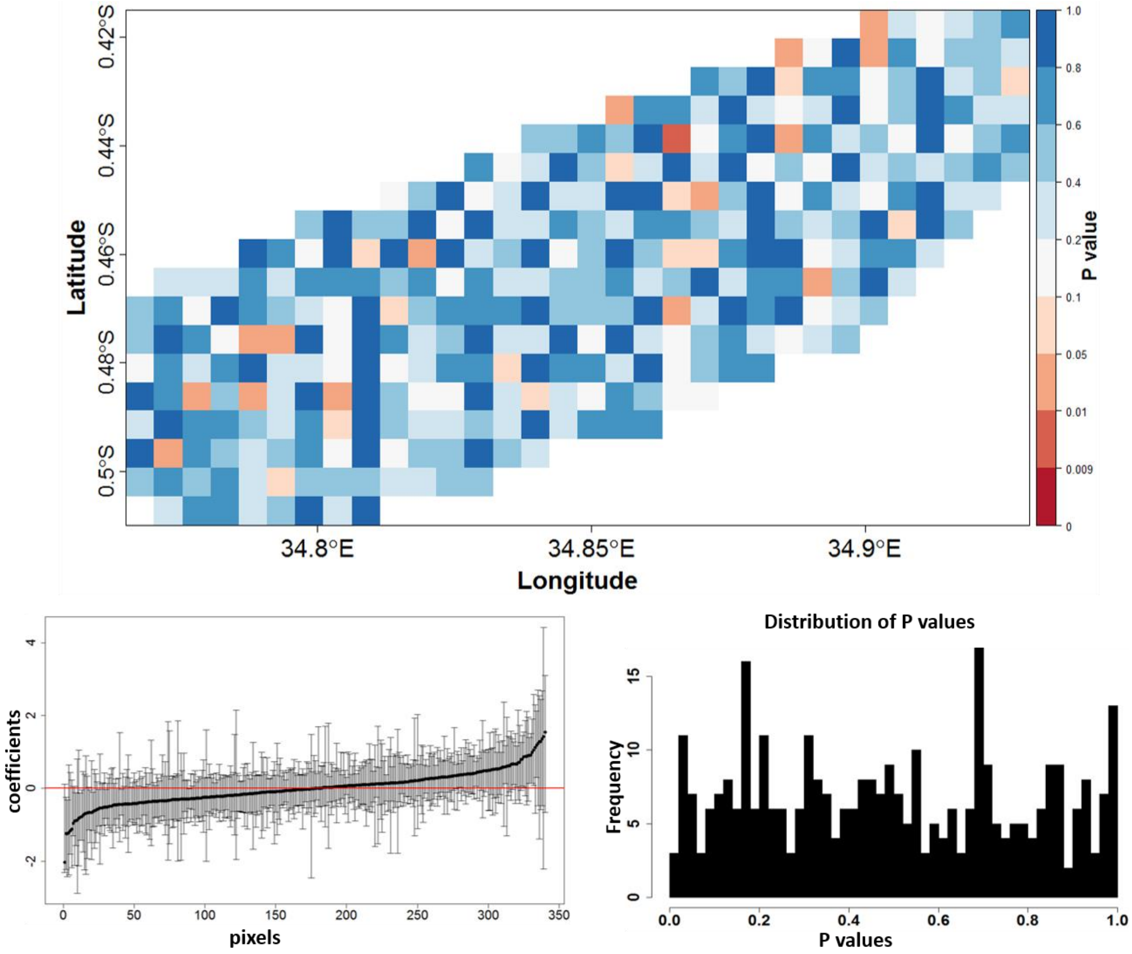


Figure 2.37: Raster analysis by pixels to examine the presence of spatial barriers to parasite movement in Rachuonyo South.

The pixel plot represents p values of bootstrapped linear regression correlation coefficients and show the significance of different geographical locations in acting as barriers to parasite mixing at a spatial scale of 0.5km x 0.5km. The colour key indicates the range of p values from > 0 to 1. Significant p values shown on the plot were not significant after applying Bonferroni correction to account for multiple testing. Accompanying the map are a plot showing the 95% confidence interval around the coefficient estimates (with a red line drawn through coefficient estimate 0) and a histogram showing the distribution of bootstrap p values following 1000 resampling steps.

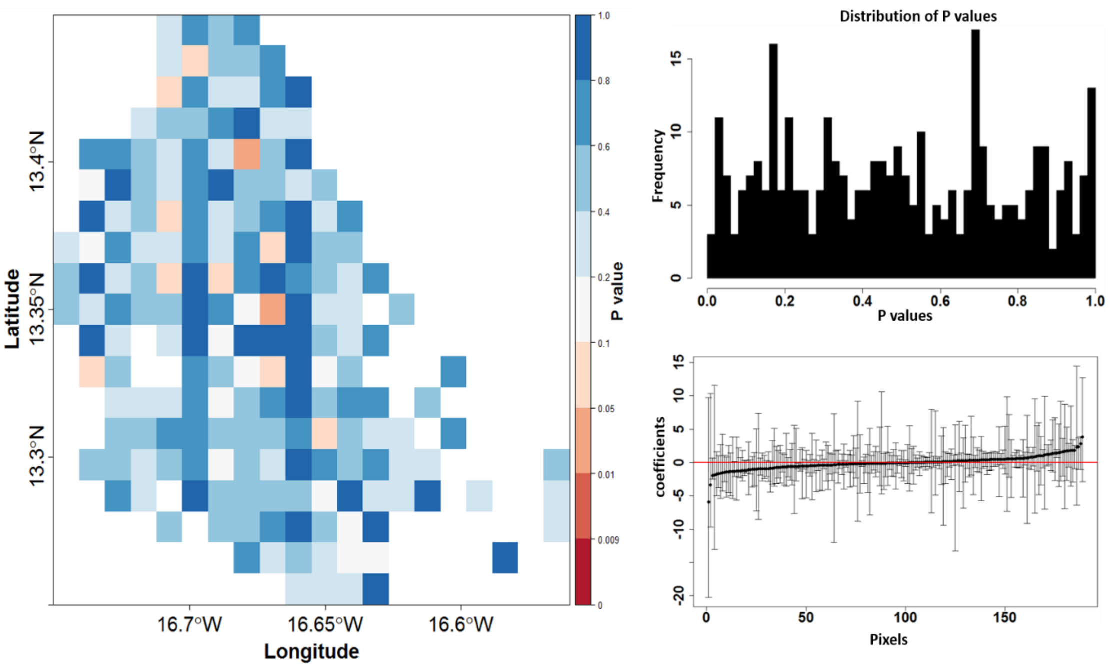


Figure 2.38: Raster analysis by pixels to examine the presence of spatial barriers to parasite movement in The Gambia.

The pixel plot represents p values of bootstrapped linear regression correlation coefficients and show the significance of different geographical locations in acting as barriers to parasite mixing at a spatial scale of 0.5km x 0.5km. The colour key indicates the range of p values from > 0 to 1. Significant p values shown on the plot were not significant after applying Bonferroni correction to account for multiple testing. Accompanying the map are a plot showing the 95% confidence interval around the coefficient estimates and a histogram showing the distribution of bootstrap p values following 1000 resampling steps.

2.4 DISCUSSION

As malaria transmission declines, targeted control at the micro-epidemiological scale is likely to be important in eliminating malaria in any remaining transmission foci. The effectiveness of such targeted measures will depend on the extent of parasite mixing in and around these foci (Bousema *et al*., 2016). In the current study, genome-wide distributed SNPs in *P. falciparum* samples collected from three African sites with varying transmission intensities were typed on the sequenom genotyping platform. Parasites within each site were then analysed for their level of genetic relatedness in time and space, to infer the extent of parasite movement and mixing at micro-epidemiological scales.

The Sequenom genotyping platform used here has high sensitivity and specificity, and samples containing as little as 2.5ng of genomic DNA can be typed successfully, even in the presence of contaminating human DNA (Gabriel *et al*., 2009). However, the genotyping pass rate is DNA concentration dependant, and in this study, was shown to increase with increasing parasite density up to 10,000 parasites/µl. Short-term laboratory cultured parasites and hospital cases from both Kilifi and The Gambia had higher genotyping pass rates compared to samples from community surveys in both Kilifi and Rachuonyo South. Community surveys from Rachuonyo South had the lowest parasite densities, possibly as a result of the higher transmission intensity associated with lower parasitaemia due to higher immunity in these populations (Bodker *et al*., 2006). Other than parasite density, sample and SNP assay failures were random and no bias was observed for specific samples or SNPs, although the existence of such a bias would indicate a possible SNP within the primer binding site which prevented the primer from binding to allow detection of the targeted SNP.

Minor allele frequency (MAF) distributions were comparable in Kilifi and The Gambia where most polymorphisms were rare. Several SNPs had MAF values of zero, meaning that in these populations, the parasites were monomorphic and no SNPs were present at these positions which had been called as polymorphic. This indicates either that the typed positions were wrongly identified as polymorphic or that the polymorphisms were so rare that they could not be identified using the sample size. Given the low frequencies of most minor alleles in African *P. falciparum* populations (MalariaGEN, 2016, Manske *et al*., 2012, Mobegi *et al*., 2014), it is probably the latter.

A simple inspection of the principal components derived from SNP genotyping in The Gambia, Kilifi and Rachuonyo South did not identify any population structure, but instead showed a high level of within-population variation in the three parasite populations. The first three principal components accounted for less than 15% of the variation observed in the parasites in Rachuonyo South and Kilifi, indicating high genetic diversity and low linkage disequilibrium between SNPs, similar to results obtained from analyses of microsatellite data (Anderson *et al*., 2000). However, there was less within-population genetic variation among Gambian parasites, pointing to greater homogeneity within this population. This is supported by the observation in this study that The Gambian parasites were identical in 64 of the 107 SNP positions analysed. High level of homogeneity within a population is a sign of clonal expansion resulting from self-fertilization and a small effective population size (Hartl *et al*., 2002), usually associated with low transmission intensities such as those seen in The Gambia (Anderson *et al*., 2000, Ceesay *et al*., 2010).

The Kilifi-Rachuonyo South and Kilifi-Gambia inter-population genetic differentiation analyses did not resolve the parasites into distinct populations based on geographic location. Instead, parasites from both populations grouped together, indicative of high genetic similarity between these seemingly distantly spaced parasites. The Gambia-Rachuonyo south comparison was not interrogated in detail because of the few SNPs (33) involved. The observation of fewer SNP differences between north-south parasite pairs compared to north-north parasite pairs in Kilifi further supports the inference of high parasite mixing within this parasite population. The fewer number of SNP differences between parasites in the south compared to those in the north indicates that there is a higher level of mixing of parasites in the south compared to the north and is likely a result of the fact that most of the samples from the south were collected from a small area of the study site. This observation of a well-mixed parasite population is in agreement with results of studies using microsatellites, which showed little differentiation of parasites separated by up to 2000km in Africa (Anderson *et al*., 2000, Bakhiet *et al*., 2015, Oyebola *et al*., 2014), immune selected genes (Bodker *et al*., 2006, Duan *et al*., 2008) and SNPs (Mobegi *et al*., 2014). However, other studies have shown population structure when looking at other parasite populations (Anderson *et al*., 2005, Campino *et al*., 2011, Pumpaibool *et al*., 2009), although these analyses were carried out on larger geographical scales than those analysed here. On an international level, for example, some studies have been able to distinguish between Senegalese and Thai parasite isolates using a 24-SNP barcode (Daniels *et al*., 2008), and another study using 4 SNPs out of a set of 384 SNPs was able to resolve East and West African parasites (Campino *et al*., 2011), showing that parasite populations can be resolved on a large geographical scale. A study in Senegal was also able to identify population structure among parasites using a 24 SNP barcode, despite a high level of similarity among the parasites analysed (Daniels *et al*., 2015). It is possible that more detailed genotyping using a larger number of markers, for instance by whole genome sequencing, would start to identify mutations that are private to particular sub-populations at a finer geographical scale, although the degree of mixing observed here suggests that discrete populations are unlikely.

Based on the results of the principal component analysis, it was not concluded that there was a lack of genetic structure to the population, only that this structure could not be identified in the absence of spatial data. The genotype data were thus analysed using spatio-temporal data to detect both global (using Moran's *I* statistic) and local (using spatial scan statistics) clusters of genetically distinct parasite sub-populations.

Global Moran's *I* analysis is used to measure the spatial autocorrelation (spatial dependencies of observations in geographic space) between feature locations (Lat/Lon) and feature attributes (e.g. principal component scores) and is used to determine whether feature attributes are clustered, dispersed or randomly distributed in space. Positive Moran's *I* correlation coefficient values indicate spatial clustering while negative values indicate dispersal of feature attributes in space. The analysis identified weak positive spatial autocorrelation at different spatial scales among parasites in all three populations, some of which were statistically significant. The evidence of spatial autocorrelation of parasite genotypes at different spatial scales points to possible existence of small clusters of genetically related parasites which themselves form part of larger clusters within the study areas. In The Gambian population, clusters of distinct parasite genotypes were detected for parasites separated by up to 6km in each of the distance classes analysed. In the Kilifi and Rachuonyo South populations, most of the statistically significant clusters of parasites were detected over relatively smaller distances of less than 10km. No significant clustering was observed for parasites collected from the same homesteads in the Kilifi population. This could be due to the presence of greater heterogeneity of parasites at this spatial level or the fact that the study consisted of only a few samples that were collected from the same homesteads. This explanation could also suffice for the Gambian population, were only 2 samples were collected from the same homestead. Within the Rachuonyo South population, significant spatial autocorrelation was detected for samples collected from the same homesteads. Making the point for mixing of parasites within the larger population is the fact that some significant clusters were detected in parasites that were more distantly separated in space (i.e. beyond 10km). This points to the possible existence of small clusters of genetically related parasites which themselves form part of larger clusters within the study areas. Previous studies of hotspots of asymptomatic and symptomatic malaria infection have identified hotspots or clusters up to the level of individual homesteads (Bejon *et al*., 2010), and this study indicates that infections in such hotspots may be caused by parasites with different genotypes.

Although Moran's *I* analysis identified significant clustering of parasites based on genotypes in the different populations, the autocorrelation was modest in effect size, signifying weak association of parasites within the identified clusters. Some clustering was observed among parasites spaced over larger distances, but these were not convincing, as shown by the associated low correlation coefficients. Overall, the consistent pattern observed from the Moran's *I* analyses was that of spatial auto-correlation at close-proximity (i.e. at a range of a few km), and little or no auto-correlation at larger distances, with the correlation indices moving up and down around zero (correlation coefficient of zero is indicative of random spatial distribution).

Moran's *I* statistic is a global autocorrelation method of analysis used to determine whether or not parasites separated by specific distances in a defined geographical area are spatially auto-correlated. The statistic does not, however, identify the actual location of these clusters within the study sites. This additional information was derived from the computation of spatial scan statistics, which was used to identify the geographical locations and sizes of statistically significant clusters of genetically related parasite sub-populations (Kulldorff, 2009). Based on this analysis, two parasite clusters were identified, one in Kilifi and the other in Rachuonyo South. The smaller size of the cluster in Rachuonyo South may indicate that parasites mix to high degrees in this population compared to Kilifi, although it may also simply reflect the fact that there was denser sampling in Rachuonyo South. No significant clusters were detected in the Gambian parasite population using spatial scan statistics, indicating that parasites mix freely in this parasite population. This finding is in agreement with those of other studies which failed to detect any genetic differentiation when parasites from this population were compared to parasites from other west African parasite populations (Mobegi *et al*., 2014, Mobegi *et al*., 2012).

Overall, Moran's *I* statistics identified more clusters than the SaTScan analyses. This is because Moran's *I* is a global statistic and looks for patterns on a global scale (over the entire geographical region) and is more likely to pick up global as well as local spatial patterns of clustering, whereas SaTScan looks for local patterns of clustering and is more likely to miss global patterns of clustering. The limited evidence of specific local clusters of parasite populations in the face of evidence of spatial auto-correlation over the whole study site implies that there is a high degree of mixing among parasites within the study sites, leading to limited clustering of parasites into genetically distinct sub-populations at micro-epidemiological scales within the study sites. Previous studies have identified parasite sub-populations based on clustering of serological responses to the important antigen *P. falciparum* Erythrocyte Membrane Protein 1 (*Pf*EMP1) in children in Kilifi (Bejon *et al*., 2011), supporting the observations of parasite sub-populations at this site. In Papua New Guinea, sub-populations of parasites have also been identified at a micro-epidemiological scale using *Pf*EMP1(Tessema *et al*., 2015), indicating that this may be a good marker for population differentiation at the micro-epidemiological level.

The effects of time, distance and time-distance interaction on the variation in SNP differences between parasite pairs within individual study sites were also interrogated. Time and distance were found to be independently associated with increasing variation between parasite genotypes (i.e. the further apart in time or space two parasites were, the greater the genetic differences observed between them). This is because parasites that are further apart in space may not easily interact, thus there are fewer chances for recombination to occur between them. Likewise, due to factors such as mutations and genetic drift, parasites are expected to acquire genetic changes over time, thus parasites with less temporal space between them e.g. 1 month, are likely to have fewer differences than those with greater temporal space between them, e.g. 1 year.

However, in the case of The Gambia and Kilifi populations where longitudinal data were available, time was shown to interact antagonistically with distance, with an increase in time reducing the variations in genetic differences between parasites as distance between the parasites increased. This implies that distance between samples was no longer predictive of genetic variation when there were longer time periods between samples, indicating that, given enough time, even parasites that are separated by large distances would get a chance to interact and recombine, especially if they are not geographically isolated. The number of SNP differences were seen to plateau at approximately 1km in The Gambia, 3km in Kilifi and 10km in Rachuonyo South. This may be attributed to the characteristics of the local parasite population, which in turn may be explained by the distribution of human settlement in the areas sampled, for example in the Gambia, homesteads tend to be clustered together in distinct, autonomous villages whereas in Rachuonyo South there is a denser and more uniform pattern of human settlement over the study area, enabling the interaction of parasites over a much larger distance.

Raster analysis by pixels was carried out to determine the correlation between parasite genotypes and infection prevalence in Kilifi and Rachuonyo South, where data on infection prevalence was available. Statistically significant correlations were observed in both populations, implying the existence of a relationship between clusters of genetically related parasites and clusters of infection prevalence. In a population where disease transmission is heterogeneous, clusters of symptomatic or asymptomatic infections indicate possible "hotspots" of infection in the community as they represent areas where there is a higher than average rate of infection compared to the rest of the population. Studies on hotspots of symptomatic malaria infection have identified hotspots or clusters of infections down to the level of individual homesteads in Kilifi (Bejon *et al*., 2014). However, most of the correlations identified in this study were weak and inconsistent, indicating that infections within higher incidence areas are likely not caused by distinct parasite sub-populations. Instead, such infections are likely caused by parasites that are well mixed within the general population.

Our inability to detect barriers to parasite movement over short distances indicates that parasites move freely within the study areas, and the spatial extent of such parasites may be limited only by the ecology and dispersal range of mosquito vectors. Furthermore, recent examination of the epidemiology of hotspots shows that they occur at the full range of spatial scales, with a pattern of spatial auto-correlation that does not show a discontinuity at any scale (i.e. a smooth semi-variogram) (Bejon *et al*., 2014). This further argues against the existence of discrete "units" of transmission with sub-populations of parasites.

This study has implications for public health interventions that may target transmission hotspots. If hotspots consist of distinct parasite populations that do not mix with parasite populations in the wider parasite community, the impact of hotspot-targeted interventions beyond the hotspot boundaries can be expected to be limited. If parasites mix freely, as suggested by this data, the impact of hotspot-targeted interventions may affect community-wide malaria transmission. This assumes that hotspots can be detected, are stable in time (Bejon *et al*., 2010) and the spread of parasite populations indeed primarily occurs from hotspots to the surrounding community (Bousema *et al*., 2016).

This study had some limitations. First, the number of SNPs typed was relatively small, and this would have limited the power to detect genetic structure among the highly similar parasite populations, especially in The Gambia. Detecting genetic structure in highly similar parasite populations may require either a much larger panel of SNPs or the use of more informative SNPs, as shown in the study by Campino *et al*, (Campino *et al*., 2011). However, despite the small SNP panel used in this study, population structure could still be detected at a micro-epidemiological scale. The analysis suggests that this structure was a uniform spatial and temporal auto-correlation rather than driven by discrete clusters of parasites at specific locations. Despite the limitations of the SNP typing and sample size it can therefore be concluded that any specific clustering is less prominent as a feature than the auto-correlations in space and time that can be detected.

A second limitation is that the study was conducted in only two sites in Kenya and one site in the Gambia. It may be premature to generalize our results more widely and an analysis of more sites will be required to make confident generalizations. On the other hand, the three sites selected demonstrate differing transmission intensities typical of many endemic Sub Saharan African countries, and this was reflected in the level of genetic diversity observed in the populations. Furthermore, the findings are consistent across all three sites. Nevertheless, patterns of parasite mixing may differ between populations based on distinctive features such as geographic isolation and patterns of human movement. Further data are required to make more general conclusions. Furthermore, as transmission continues to decline and malaria programmes gradually shift their focus from control to elimination, the analysis of parasite gene flow between different transmission foci, e.g. Kilifi and Rachuonyo South, will become increasingly important in informing the mitigation measures needed to prevent importation of parasites as a result of human movement and migration. These analyses were not carried out in the current study since the numbers of common SNPs between the two Kenyan sites was low, and we only had parasites from one timepoint in Rachuonyo South, we were therefore unable to conduct an informative analysis of gene flow between sites.

In conclusion, this study has shown that *P. falciparum* parasite populations mix evenly within specific sites in The Gambia, Kilifi and Rachuonyo South and there appear to be no detectable geographical barriers to parasite movement over short distances within these sites. That said, autocorrelations of genotypes were detected at the micro-epidemiological level. It can be concluded that control strategies that efficiently target hotspots will likely benefit the wider community outside the hotspots at the District/County level (I am however unable to comment on larger geographical scales), although this is likely to be affected by factors such as the underlying transmission level, heterogeneity of transmission, and patterns of human movement (Bousema *et al*., 2016). On the other hand, following mass-treatment campaigns it would be predicted that if residual foci of transmission are retained, this will rapidly lead to re-infection of the wider community, and that parasites acquiring mutations conferring drug resistance or immunological escape will be rapidly spread at a micro-epidemiological level. However, these conclusions may not stand if the apparent high level of mixing has emerged over a prolonged period of time, and full genome analyses with greater power to examine low-frequency mutations and recombination events will have greater power to examine the time-scales involved

Chapter 3 Geographic-genetic Analysis of *Plasmodium falciparum* Parasite Populations from Surveys of Primary School Children in Kenya.

3.1 INTRODUCTION

In the previous chapter, I analysed the spatio-temporal genetic variation of *P. falciparum* parasites at micro-epidemiological scales in order to show the degree of parasite mixing over short distances in regions with varying transmission intensities. In that analysis, I identified spatial structure at fine micro epidemiological scales within geographically defined regions in Kenya and The Gambia.

In the current chapter, I carried out a similar analysis at a macro-epidemiological scale to determine whether similar patterns of parasite movement and mixing are observed over larger geographical scales. Limited genetic differentiation between malaria parasite populations has been observed at a regional scale in sub-Saharan Africa (Campino *et al*., 2011, Manske *et al*., 2012, Mobegi *et al*., 2012), as well as within individual countries (Nabet *et al*., 2016).

This study used parasitological data collected during surveys of primary school children across Kenya. Surveys form an important part of monitoring and evaluation of interventions against different infections and diseases. Currently, various malaria indicators are collected through Demographic and Health Surveys (DHS), multiple indicator cluster surveys (MICS) and malaria indicator surveys (MIS) (Brooker *et al*., 2009, Gitonga *et al*., 2010). These surveys mostly target pregnant women and young children under the age of five and collect data on intervention coverage (ITN ownership and use, indoor residual spraying), malaria case management (diagnosis of causes of fever, use of anti-malarial drugs) and prevalence of malaria and anaemia (Kenya national malaria control programme, 2016). Unfortunately, such surveys are not ideal for routine monitoring as they are carried out only every 3-5 years due to the expense, time and labour involved (Brooker *et al*., 2009, Gitonga *et al*., 2010). Additionally, generalizing estimates of parasite prevalence collected in young children and pregnant women to the rest of the population may not be ideal because of parasite sequestration in pregnant women and protective effect of maternal antibodies in young children (Brooker *et al*., 2009, Gitonga *et al*., 2010). Moreover, these surveys are usually powered to provide data at the national or sub-national level, thus making it difficult to use this information for planning targeted control at local levels (Brooker *et al*., 2009, Kenya national malaria control programme, 2016). School surveys provide a cheaper and more rapid alternative to household surveys in the collection of malaria indicators (Brooker *et al*., 2009, Gitonga *et al*., 2010). In addition, school surveys can be used to estimate the community-wide coverage of various malaria interventions such as bed net use and prevalence of anaemia.

This study aimed to examine the geographic-genetic patterns of malaria parasite populations sampled across Kenya, with a view of determining the extent of parasite mixing and genetic adaptation by the parasite population to its local environment. The study also aimed to determine the patterns of flow of parasites around the country, and thereby the main geographical sources and sinks of transmission. 111 single nucleotide polymorphic (SNP) positions were genotyped in 2715 *P. falciparum* isolates collected from children in 146 primary schools in five Kenyan provinces (Western, Eastern, Coast, North Eastern and Nyanza) in order to analyse the genetic relatedness among the parasites. Parasite population structure was examined based on principal component analysis (PCA), and measures of local and global spatial autocorrelation used to test for geographical relatedness among parasite genotypes. Furthermore, the spatial distribution of allele frequencies was analysed to identify any evidence of genetic adaptation of parasite populations to their local environment, and the region examined for spatial barriers to parasite movement, as well as for patterns in the direction of movement in either north/south or east/west directions.

3.2 MATERIALS AND METHODS

3.2.1 Study population

P. falciparum positive samples were collected from children in 146 primary schools in 28 districts spread across five of the eight Kenyan provinces (Western, Nyanza, Coast, Eastern and North Eastern). Kenya has recorded a decline in malaria transmission in the past decade (Bhatt *et al*., 2015, Mogeni *et al*., 2016, Noor *et al*., 2014) and current country-wide malaria prevalence is estimated at 8% (Kenya national malaria control programme, 2016). However, transmission is highly heterogeneous and these five provinces represent varied malaria endemicity zones ranging from semi-arid, seasonal malaria transmission in north eastern province through to endemic transmission in coast province and epidemic transmission in the Western Kenya highlands (Kenya national malaria control programme, 2016). Declines in transmission have been more obvious for the coastal region than for the western region (Okiro *et al*., 2009), two areas of the country that have historically experienced the highest transmission, although recent trends show an increase in transmission in the coast (Mogeni *et al*., 2016, Snow *et al*., 2015).

P. falciparum is the main causative agent of malaria, and is transmitted by different *Anopheles* mosquito species in different parts of the country (Mala *et al*., 2011, Mwangangi *et al*., 2013, Stevenson *et al*., 2012). The highest malaria transmission intensity is currently experienced in western Kenya and is characterized by stable, endemic transmission along the lowlands, and unstable, epidemic transmission within the highlands (Kenya national malaria control programme, 2016, Noor *et al*., 2014, Okiro *et al*., 2009), and despite scale up of interventions, malaria transmission has remained high, or even increased in certain parts of this region (Bayoh *et al*., 2014, Zhou *et al*., 2011).

3.2.2 Ethics statement

During initial sample collection, consent for participation in the surveys was based on passive, opt-out consent by parents rather than written, opt-in consent, due to the routine, low-risk nature of the surveys that were carried out under the mandate of the Ministry of Public Health and Sanitation to conduct disease surveillance. Individual assent from the students was obtained before sample collection (Gitonga *et al*., 2010). Ethical approval for the current genotyping study was provided by Kenya Medical Research Institute (KEMRI) Ethical Review Committee (under SSC No. 2747) and study methods were carried out in accordance with approved guidelines.

3.2.3 Sample collection and DNA extraction

This study used finger prick blood samples collected during a previous nationwide parasitological survey of school children conducted in collaboration with the Division of Vector Borne and Neglected Tropical Diseases, Ministry of Health. The surveys were conducted between September 2008 and March 2010 (Gitonga *et al*., 2010), and were carried out to assess the health of school children and to provide measures of malaria endemicity against which the efficacy of control programmes could be measured. An in-depth description of sample collection was published previously (Gitonga *et al*., 2012, Gitonga *et al*., 2010). Briefly, 480 government primary schools were surveyed and

49,975 samples collected, with a maximum of 110 children randomly sampled in each school. The sampling frame for the selection of schools was based on the national school census carried out in 2008 by the ministry of education, and schools were selected to provide optimal spatial distribution of sampling across the country.

Samples were collected by spotting 3 separate drops of 200µl finger prick blood onto Whatman filter papers. These samples were then air dried and stored, with desiccant, at 4^{o}C. Additionally, each child was tested for *P. falciparum* parasite infection using rapid diagnostic tests. During sample collection, geospatial coordinates for each school were recorded. 2715 children from 146 schools in the five provinces were found to be parasite positive (figure 3.1, table 3.1). Genomic DNA was extracted from these parasite-positive samples.

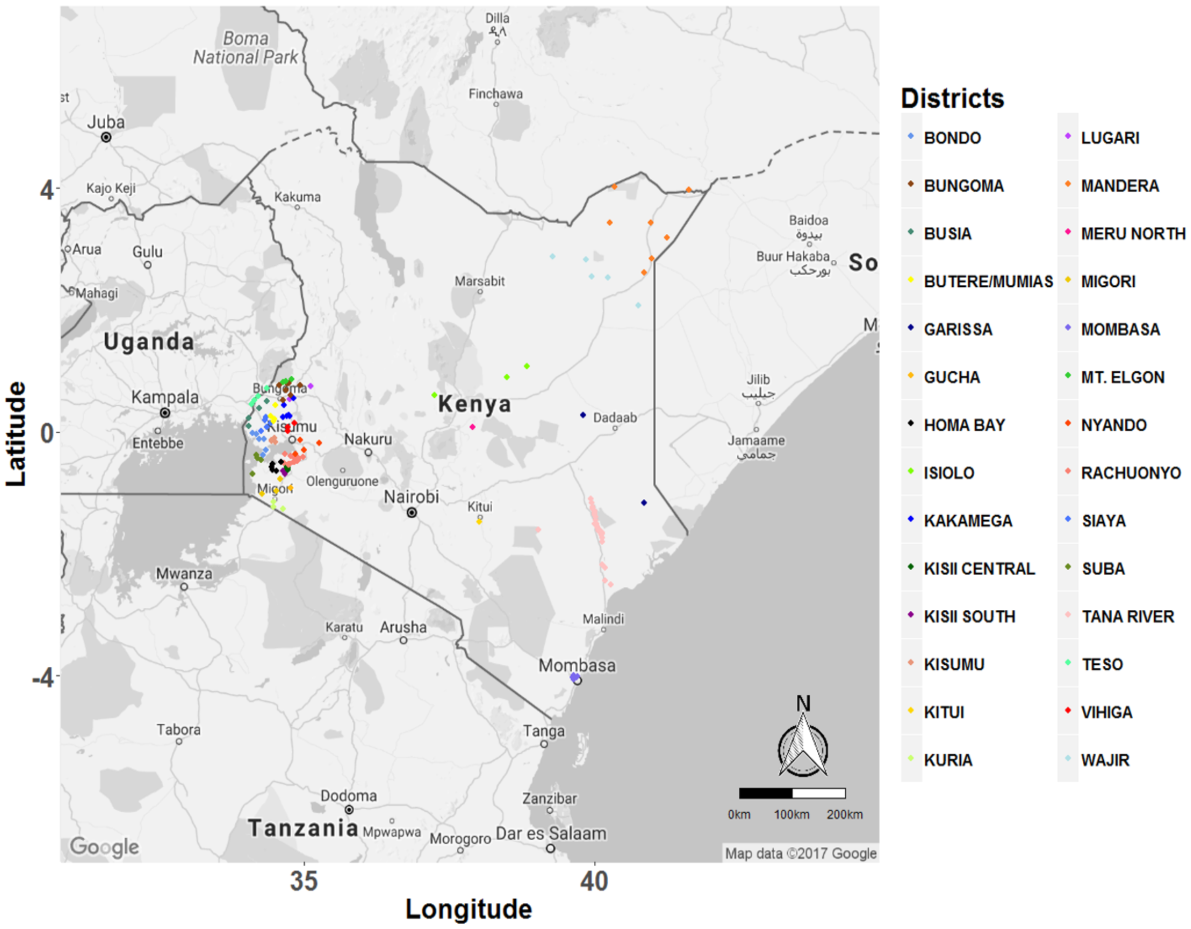


Figure 3.1: Spatial distribution of primary schools surveyed across Kenya.

The map shows the distribution of 146 schools across 28 districts in 5 provinces where 2715 *P. falciparum* positive samples were collected. Each dot represents an individual school, colour-coded by the administrative district in which it is located.

Table 3-1: Number of *P. falciparum* positive samples collected from five Kenyan provinces

Province	Number of samples
Coast	186
Eastern	8
North Eastern	35
Nyanza	1405
Western	1081

During DNA extraction, one of the blood spots was excised from each filter paper, cut into small pieces and placed into 1.5ml flip-top micro-centrifuge tubes (Eppendorf, Stevanage, UK). DNA was extracted using the QIAmp DNA investigator kit (Qiagen, UK), as per the manufacturer's instructions on the Qiagen BioRobot. Picogreen (Fisher Scientific-UK Ltd, Loghborough, UK) was used to determine DNA concentration in each sample.

3.2.4 SNP selection and genotyping

111 exonic SNP positions were typed in 67 *P. falciparum* genes (Appendix Table 3.1) in each of the parasite positive samples. These SNPs were selected from the same 384-SNP set mentioned in chapter two and were chosen based on the same criteria as those in chapter two, i.e.

1) genome-wide distributed.

2) polymorphic between at least two of three *P. falciparum* strains (3D7, HB3 and IT).

3) type-able on the sequenom genotyping platform.

Sequenom SNP genotyping was done at the Wellcome Trust Centre for Human Genetics in Oxford University in 2015, and was carried out following the same procedure described in chapter two. Based on primer compatibility, assays were pooled in groups of 38 or 40 SNPs per plate-well.

3.2.5 Sample and SNP cut-off selection

To determine Sequenom genotyping success rates, a pass/fail criterion was applied to genotyping. In each sample, SNPs that passed genotyping were assigned a 1, and those that failed were assigned a 0. The selection criterion for successful typing was based on individually defined SNP intensity values (R) ranging from 0 to 1. SNPs with intensity values <0.1 were considered low quality and were categorized as failed and excluded from further analyses. In addition, allelic intensity ratios (θ) nearing 0 or 1 were used to classify SNP positions as homozygous, and intensity ratios of intermediate values were used to classify SNPs as heterozygous, representing mixed parasite populations in a single sample. Where mixed parasite populations were identified, the dominant genotype as represented by the majority SNP call was taken forward for further analysis.

Genotype data were then aggregated based on the pass/fail criterion. The criteria were to include samples where at least 60% of SNP typing was successful and, among these, to include SNPs that were successfully typed in at least 60% of samples. Applying these inclusion criteria, the analyses were restricted to 83 SNPs and 1809 samples collected from 88 schools in two provinces in Western Kenya (Nyanza and Western). Samples from the other 3 provinces were excluded either due to complete genotyping failure, i.e. none of the samples meeting the cut-off criteria (Northern Kenya province) or very few numbers of samples and SNPs meeting the cut-off criteria (Eastern and Coast provinces). This likely reflects the lower parasite prevalence in Eastern and Coast provinces (Kenya national malaria control programme, 2016), and lower mean parasite densities among those that were positive (Bousema *et al*., 2014), leading to frequent genotyping failure.

3.2.6 Statistical analyses

Several statistical analyses were carried out either in R statistical software (version 3.3.1) or SaTScan software (version 9.4). Pre-statistical analyses included removing water and G6PD samples which were used as negative and positive controls, respectively. Statistical analyses carried out included analysis of population structure, global and local spatial autocorrelation, distribution of allele frequencies, parasite adaptation to their local environments, spatial barriers to parasite movement and directionality of parasite movement. Each of these analyses (Table 3.2) will be discussed in the next sections.

Table 3-2: Statistical tests conducted on *P. falciparum* isolates from primary school children in Western Kenya.

Statistical analysis	Function
Pairwise SNP and distance differences calculation	To determine the number of SNP differences and distance differences between parasite pairs in the dataset.
Minor allele frequency distribution	Analysis of the numbers and distribution of the minor alleles in the population.
Principal component analysis	Detecting *P. falciparum* population structure at the sub-national level.
Moran's *I* spatial autocorrelation analysis	Analysis of global spatial autocorrelation among parasite pairs at different spatial scales.
Spatial scan statistics	identification of local, spatial clusters or hotspots of parasite sub-populations.
Logistic regression	Analysis of the spatial pattern of distribution of allele frequencies among individual schools.
Bernoulli regression	Identification of spatial clusters of schools with similar allele frequencies of specific SNP loci.
Raster analyses by pixels	Detecting geographical regions that act as spatial barriers to parasite movement; Moran's I analysis was then used to determine if pixels acting as barriers to parasite movement were spatially autocorrelated.
Bearing regression analysis	Analysing directionality in parasite movement within the region.

3.2.6.1 *P. falciparum* pairwise SNP and distance differences calculations

Pairwise SNP and distance differences were computed, comparing each parasite to every other parasite in the study population. For each parasite pair, the number of SNP differences at the 83 polymorphic positions analysed and the distance between sample pairs, based on the geographical coordinates of schools where the samples were collected, were computed. Samples collected from the same school were assigned a distance difference of 2.5km, assuming that schools were at least 5km apart, and taking 2.5km as the lower limit of detection of any two schools (i.e. 2.5km was taken to represent the shortest distance within which any two schools could be separated). Time differences were not computed because the samples were collected over a few months, and samples in the same school were all collected on the same day.

3.2.6.2 P. falciparum population genetics

Heterozygosity in the parasite population was computed as the average number of SNP differences per SNP site between two parasites in all parasite pairwise comparisons. SNP differences between parasite pairs were aggregated at the school level to determine the mean number of SNP differences among parasites per school and the effect of distance on genetic variations analysed by plotting changes in SNP differences against the difference in distance between schools.

Minor allele frequency distributions were computed for all 83 SNPs that had been successfully typed to determine the distribution of common and rare genetic variants in the population. Population structure was interrogated using principal component analysis (PCA), based on singular value decomposition on a covariance matrix of pairwise SNP differences for all samples. SNPs that were included in the analysis but were unsuccessfully typed in individual samples were replaced (in that sample) with the major allele in the population so as not to artificially inflate the allelic diversity in the study population. Where there were mixed genotype infections in an individual sample, the major allele call was taken to represent the dominant genotype at that position in that sample. PCA scores (values of the transformed variables that correspond to a specific data point) representing individual parasite genotypes were computed for the first 3 principal components. Since geospatial positioning information was collected at the school level and samples from the same school were assigned the same geographical coordinates, principal component (PC) scores were later aggregated at the school level, and the mean PC score for parasites in each school plotted on a map of the study region.

3.2.6.3 Tests of spatial autocorrelation

To test the hypothesis that *P. falciparum* genetic structure has a spatial distribution component, both global (Moran's *I*) (Epperson and Li, 1996) and local (scan statistics) (Kulldorf, 2014) spatial autocorrelation analyses among parasite genotypes were carried out. Moran's *I* simultaneously measures the correlation between feature attributes (parasite genotypes, represented here by the scores of the first 3 PCs) and locations (geospatial positioning of schools) to determine whether feature attributes are clustered, dispersed or randomly distributed in space. For each PC, Moran's *I* correlation coefficients were computed for parasite pairs falling within 3 increasing distance bands of 1km, 2km and 5km. The analyses were bootstrapped using 100 resampling steps to determine the statistical significance of the Moran's *I* correlations observed for parasite pairs within each distance class. The analyses were computed for samples separated by up to 60km within each distance class because the data got noisier beyond this point due to fewer pairwise comparisons.

Spatial scan statistics to detect statistically significant local clustering of genetically related parasites were carried out in SaTScan software (version 9.4). The analyses were carried out separately for genotypes represented by each of the first three PCs. For each PC, scores representing individual parasite genotypes were imported into SaTScan, together with the geospatial coordinates of sample collection. A purely spatial, retrospective analysis was then undertaken, based on a normal probability distribution model, to locate geographical clusters of parasites with high PC scores, as described in chapter two. In the normal probability distribution model, each observation or sample is associated with a single negative or positive continuous attribute (the PC score) and the model uses a likelihood function based on the normal distribution. The spatial scan statistics employed here use a circular scanning window that is flexible both in location and size, with a radius that varies continuously from zero (including only a single sample location/school) to an upper limit set by the user (in this case, the largest window was allowed to include up to 50% of the sample locations/schools). For each window location and size, a ratio of observed to expected PC scores is computed for samples found inside and outside the window. Clusters with high ratios are noted and their statistical significance determined after accounting for multiple comparisons using random permutations.

SaTScan outputs the size and geographical midpoints of likely clusters, the number of samples within each cluster, and the p value associated with each cluster. The first cluster, which has the most significant p value, is called the primary cluster. Other clusters that are identified in the population that do not overlap with the primary cluster are reported as secondary clusters. After the analysis, shapefiles containing the spatial coordinates of statistically significant clusters were imported into R to allow for plotting of significant clusters on a map of western Kenya to identify locations of schools with genetically distinct parasite sub-populations.

3.2.6.4 Spatial distribution of allele frequencies

A binomial logistic regression model was used to examine geographic variations in allele frequencies. Each school was included in the model as a categorical independent variable, with the binary (1/0) outcome set as the presence or absence of a specific SNP in parasites within that school. This model was then compared to a null model in a likelihood ratio test to test the goodness of fit of the model containing the SNPs and to identify those SNPs that showed statistically significant ($p < 0.05$) variations in frequency between schools. To keep from inflating allele frequencies, SNPs that failed genotyping in individual samples were excluded from analysis. For those SNPs that showed a significant variation in frequency between schools based on the logistic regression model, the actual frequency of each SNP per school was computed and plotted on a map of the region to visualize the distribution pattern of SNP frequencies in schools within the region.

The Bernoulli model, implemented in SaTScan, was used to examine the study region for clusters of schools with similar allele frequencies for individual SNPs. Only SNPs that showed statistically significant variations in allele frequencies based on the logistic regression model described above were included in the analysis. The Bernoulli model requires that cases and controls be specified prior to the analysis, thus at each SNP position, samples were coded with a 1 if they contained the major allele, and 0 if they contained the minor allele. For each SNP, a case file was generated containing the frequency of cases (total number of 1s) in each school and a control file was generated containing the frequency of controls (total number of 0s) in each school. These two files, together with a coordinate file specifying the geographical location of each school were then imported into SaTScan for the analysis.

The Bernoulli model analyses the distribution of cases (major allele) and controls (minor allele) and tests the null hypothesis that the distribution of cases and controls is random within the geographical area. The spatial scan statistics based on this model involves scanning the geographical space for regions with higher than expected number of cases. To do this, the software computes the fraction of cases/controls inside a specific window versus the cases/controls outside the window. The windows with the highest probability of a case inside versus a case outside are noted down as clusters. The statistical significance of the identified clusters is computed as in the normal distribution model described in chapter two, using random permutations of the cases/control over the spatial coordinates of schools and calculating the log-likelihood statistic of the model fit for the optimal window, then comparing this with the log-likelihood value from the real data to derive a p value after 9999 replication steps. The output of the analysis includes the mid-point coordinates and sizes of primary and secondary clusters, the associated p-value for each cluster, the number of cases as well as the total population (cases + controls) in each cluster. Shapefiles (data file containing location, shape and attributes of clusters) for individual clusters were exported from SaTScan and imported into R for representation on a map of the study area.

3.2.6.5 Analysis of spatial barriers to parasite movement

Raster analysis by pixels was used to test for the presence of spatial barriers to parasite movement and mixing within the study region, which was divided into 192 10km-by-10km grids/pixels. Each pixel was then scored with either 1 or 0 depending on whether or not a straight line linking a school pair crossed the boundaries of that pixel. This was done for all 192 pixels and all school pairs in the study region. These scores were then included as independent variables in a multivariable linear regression analysis with the number of SNP differences between schools as the outcome variable, to test how the presence of a specific pixel affected the nucleotide diversity of parasites in schools separated by that pixel.

To determine the statistical significance of observed differences in nucleotide diversity, the analysis was bootstrapped based on 10,000 resampling steps. The coefficient estimates derived from the pixel analysis were then included in a Moran's *I* analysis to determine whether pixels acting as barriers to or gateways for parasite movement were spatially auto-correlated. Moran's *I* analysis was computed for schools falling within a 10km distance class, representing the same spatial extent as that used to generate the pixels (10km was chosen as this was the spatial scale at which significant results had been observed). The analysis was bootstrapped with 100 resampling steps to determine statistical significance of the Moran's *I* analysis.

Regression analyses based on bearing (direction of motion) were also carried out to examine the parasite population for patterns of directional movement, either in the north/south or east/west directions. For each 10km-by-10km grid, school pairs that crossed its boundaries (scored as 1 in the previous pixel analysis) were selected. Each pair of schools was then individually coded as 1 if the absolute difference in latitude between them was greater than the absolute difference in longitude, indicating a north/south direction of movement, and coded as 0 if the reverse was true, indicating a west/east direction of movement. These new variables were then included in a multivariable linear regression analysis to test the effect of north/south or east/west directional movement on parasite genetic diversity, with the number of SNP differences as the outcome variable. Statistical significance of each pixel in acting as a force in directional movement was tested using bootstrap method with 10,000 resampling steps. The coefficient estimates were plotted on raster grids, highlighting the pixels that were significant after Bonferroni correction. A histogram showing the distribution of bootstrap p values was also generated. The analysis was repeated at 20km-by-20km grid sizes to determine whether similar patterns would be observed at the larger spatial scale. Additional analyses were also carried out for those pixels that had at least 150 school-school pairs crossing them, to rule out spurious significant results based on a small number of points crossing the pixels. The analysis was carried out at 10km-by-10km and 20km-by-20km grid sizes for 63 pixels that had at least 150 points passing through them. Coefficient estimates derived from the bearings' regression analysis were then included as feature attributes in a Moran's *I* analysis to examine the region for spatial autocorrelation in the direction of parasite movement. Moran's *I* analysis was once again carried out for schools falling within 10km distance classes, and with 100 resampling steps to determine statistical significance.

3.3 RESULTS

3.3.1 Sequenom assay performance

111 genome-wide distributed exonic SNPs were genotyped in 2715 *P. falciparum* positive samples collected from 146 primary schools in 28 districts across Kenya (figures 3.2a). Most of these samples (2486) were collected from 95 schools in Western Kenya (figure 3.2b and figure 3. 3). 83 of the 111 SNPs were successfully typed in 1835 samples including 1097, 712, 1 and 25 samples from Nyanza, Western, Eastern and Coast provinces, respectively.

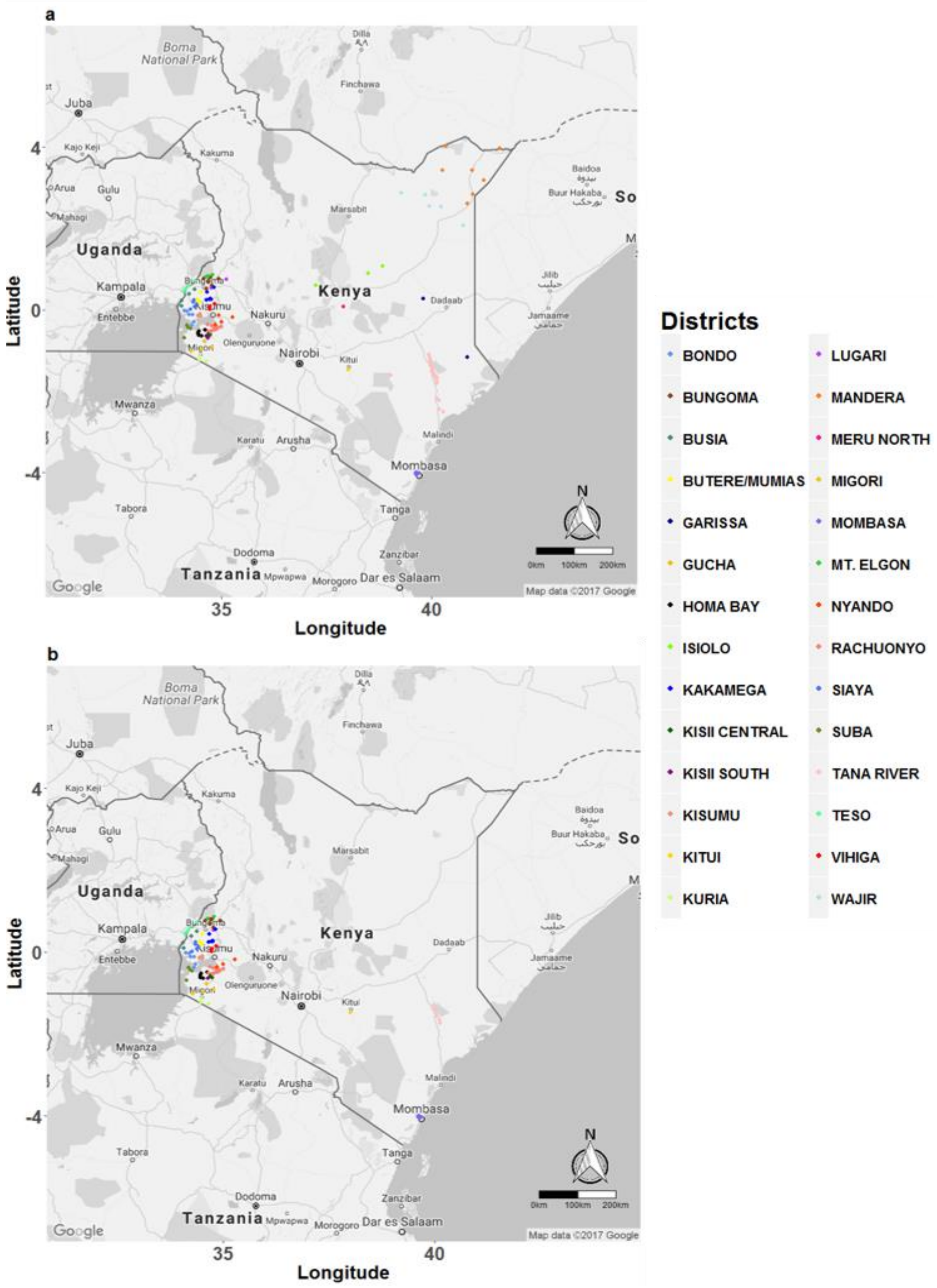


Figure 3.2: Country-wide distribution of primary schools having *P. falciparum* positive infections.

Each dot represents an individual school, colour-coded by the administrative district in which it is located. **(a)** shows the distribution of 146 schools in 28 districts across 5 provinces from where 2715 *P. falciparum* positive samples were collected. **(b)** shows the distribution of 101 schools in 23 districts across 4 provinces containing 1835 samples with at least 60% genotyping pass rate.

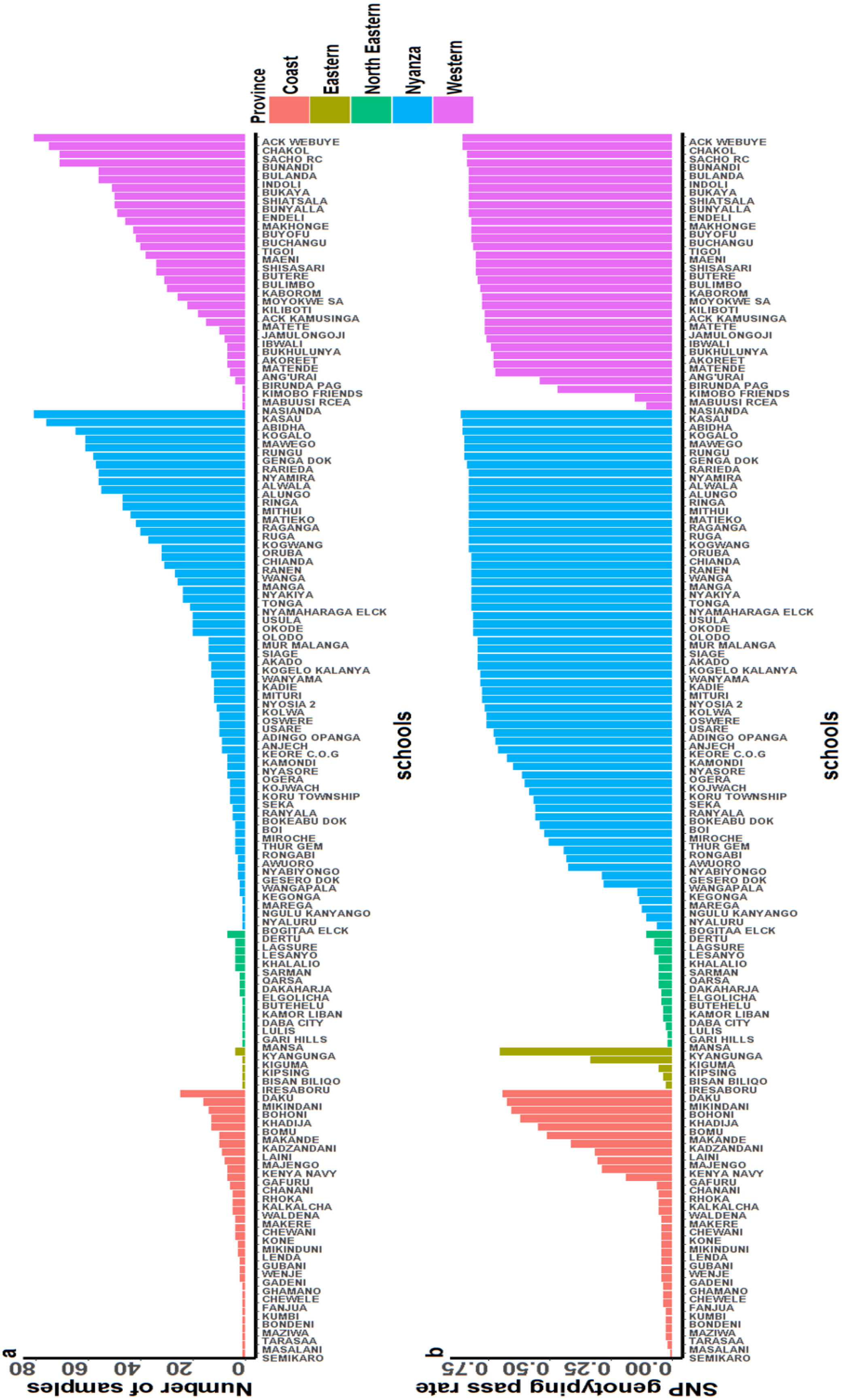


Figure 3.3: Distribution of *P. falciparum* positive samples per school and their associated genotyping success rates.

(a) 2715 samples were collected from 146 schools in 5 provinces across Kenya. The total number of samples varied from 1 to 81 per school. **(b)** 111 single nucleotide polymorphic (SNP) positions were genotyped in all parasite positive samples. Mean genotyping success rates per school ranged from 6-86%.

Genotyping was done on the sequenom platform which allows the identification of genotypes based on variation in the masses of the alleles (figure 3.4)

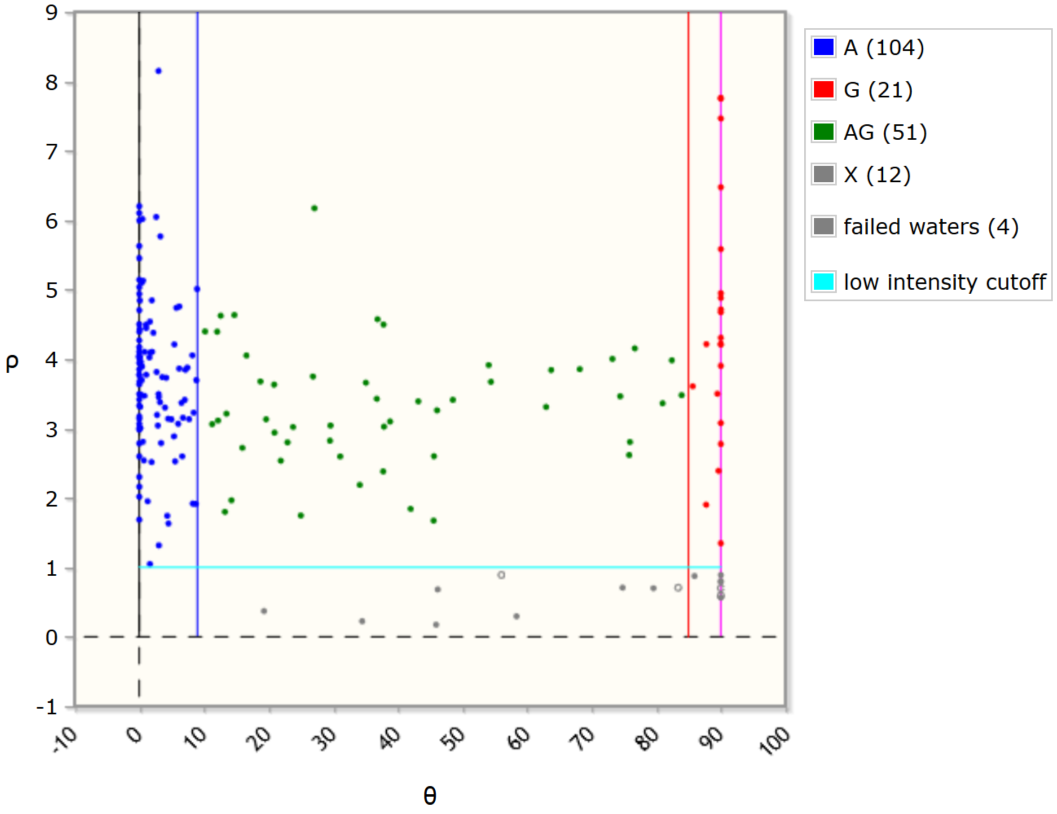


Figure 3.4: Sequenom output plot showing the genotyping results of one SNP position in EBA 175 (MAL7P1_176_MAL7_1413900(AG)).

Single infection/homozygous genotypes are coded as blue (A) or red (G) whereas mixed/heterozygous genotypes (AG) are coded as green. The grey dots represent samples that failed to genotype as well as negative water controls. The low intensity cut-off mark is represented by the light blue line.

Assay genotyping success rates ranged from 0 – 97.6% across the 111 SNPs (figure 3.5). However, the genotyping of 8 SNP positions was unsuccessful in all samples (100% failure rate), including the two SNPs typed in the sulfadoxine/pyrimethamine resistance-associated gene dihydrofolate reductase (dhfr), two SNPs in *EBA-175*, one SNP in *Surf*$_{4.1}$ and the remaining in hypothetical and conserved proteins. Genotyping success rates among samples across the schools ranged from 1.8% - 86.5% (figure 3.5). The 111 SNPs typed in the study were in 67 genes distributed across the parasite genome, and between 1 and 27 SNPs were typed in each of these genes (Appendix Table 2).

SNP genotyping performance was positively correlated with transmission intensity, with parasites from regions with very low transmission intensities (North Eastern) having the lowest genotyping success rates, and those from regions with high transmission intensities (Western Kenya) having the highest genotyping success rates (figure 3.3).

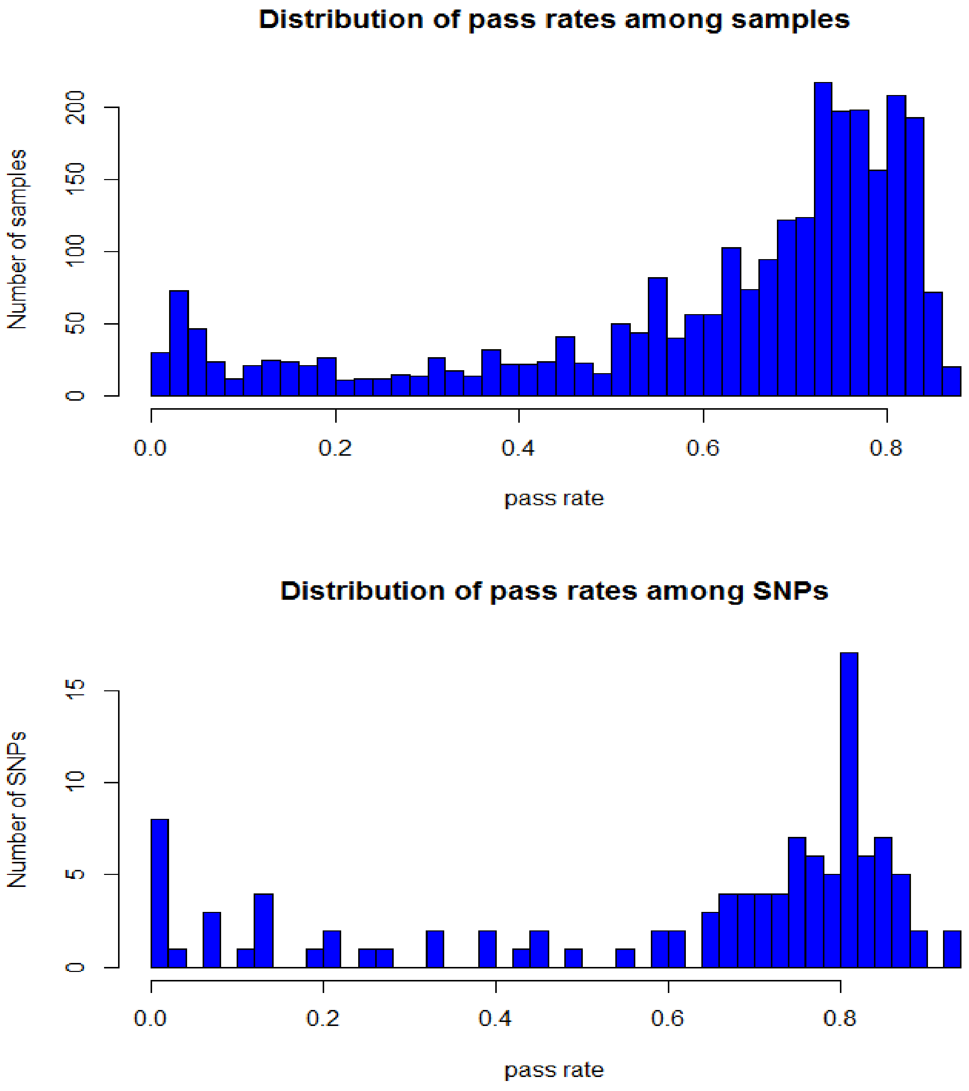


Figure 3.5: Distribution of *P. falciparum* sequenom genotyping success rates among 2715 samples and 111 SNPs.

Success rate ranges were 1.8% - 86.5% among the samples (**top panel)** and 0 – 97.6% among the SNPs (**bottom panel**).

None of the samples from North Eastern province met the 60% sample and SNP inclusion criteria that had been set and were therefore excluded from further analyses. Samples from Coast and Eastern provinces were also excluded from further analyses due to the low number of successfully typed samples (25 samples from Coast province and 1 sample from Eastern province), and subsequent analyses were carried out on 1809 samples from 88 schools in the two Western Kenya provinces (Western and Nyanza).

Variation in parasite prevalence was observed in the region, with areas north and west of Lake Victoria having higher infection prevalence than areas south and east of the lake (figure 3.6).

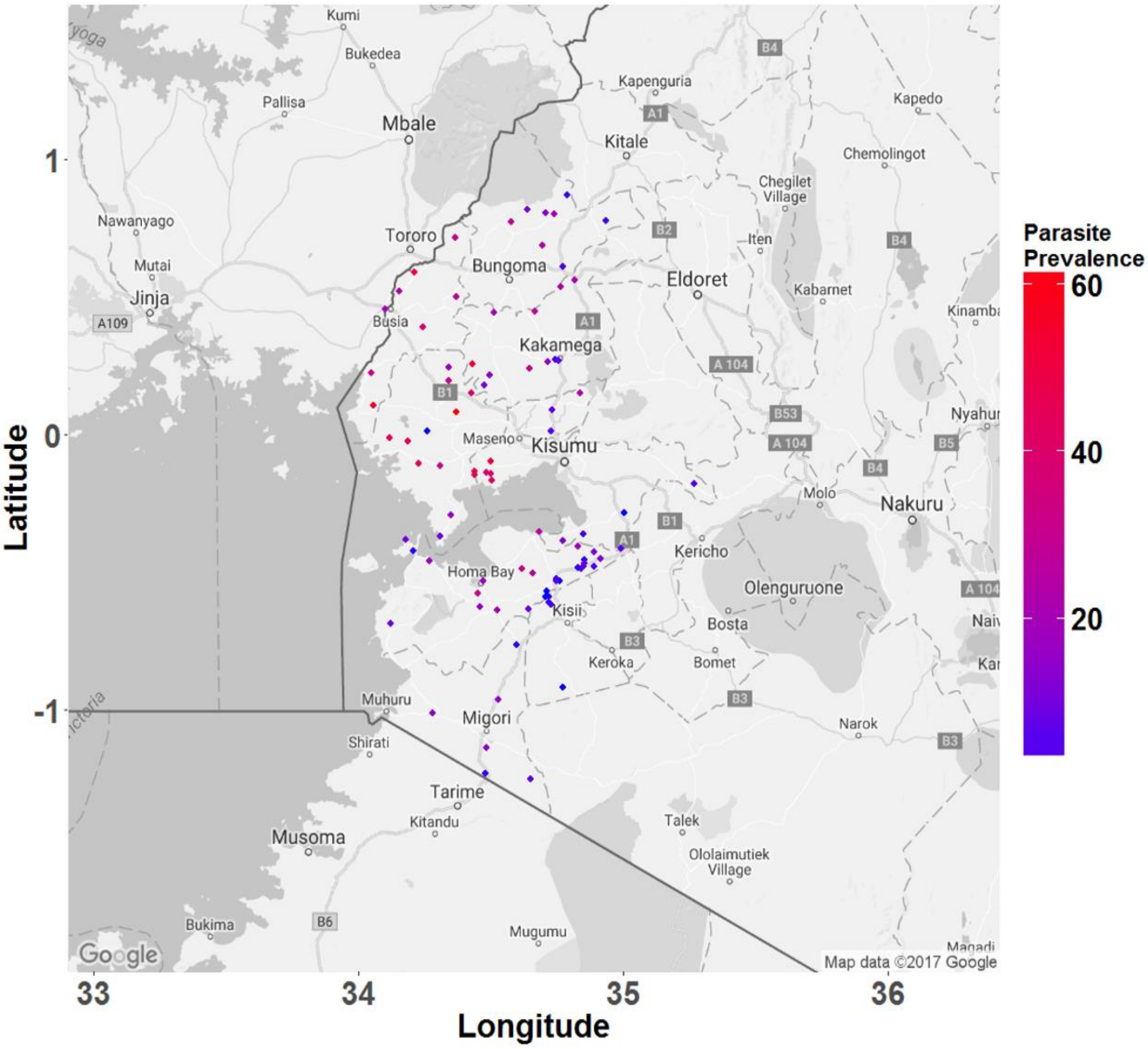


Figure 3.6: ***P. falciparum*** **parasite prevalence in primary schools across Western Kenya.**

Each dot represents an individual school, colour-coded based on parasite prevalence (%). Parasite prevalence ranged from 0.9% - 62%.

3.3.2 Minor allele frequency distribution

Analysis of the minor allele frequency (MAF) distribution showed that most of the genotyped SNPs were present at medium to high frequencies in the parasite population, with 51 of the 83 successfully typed SNPs having minor allele frequencies of 5% or higher (figure 3.7). Although there was a high level of genetic diversity among the isolates, 14 positions that had been identified as polymorphic/ variable based on the 3D7 reference genome were monomorphic (i.e. the same nucleotide was present at that position in all the samples) in this population.

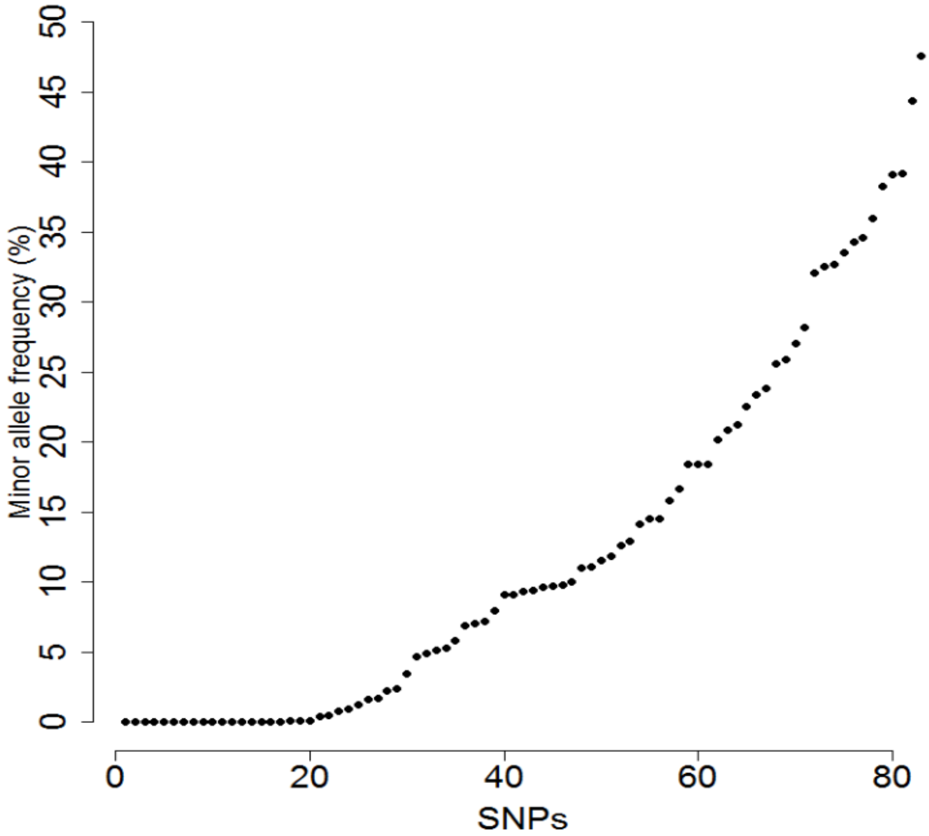


Figure 3.7: Minor allele frequency distribution spectrum of 111 genome-wide distributed SNPs in a western Kenya *P. falciparum* population.

SNP frequencies were calculated based on 1809 samples from 88 schools in Western and Nyanza provinces.

3.3.3 *P. falciparum* population genetics

A high level of within-population genetic diversity was observed in the parasite population, with a heterozygosity score of 0.184 per SNP site. Only 2 parasite pairs were identical in the population. The average number of SNP differences was 15.3 (range 0 - 33) per parasite pair.

Principal Component Analysis (PCA) was carried out to determine the extent of genetic structure within the parasite population. A scree plot showing the amount of variation explained by each of the first 10 principal components was plotted (figure 3.8). Each PC accounted for only a small amount of the genetic variation seen in the parasite population. The first three PCs cumulatively accounted for only 10.78% of the variation (PC1=3.74%, PC2=3.54%, PC3=3.5%), indicating high diversity levels among the parasites. Based on genotype data alone, the parasite population could not be resolved into distinct sub-populations at a sub-national level using PCA (figure 3.9).

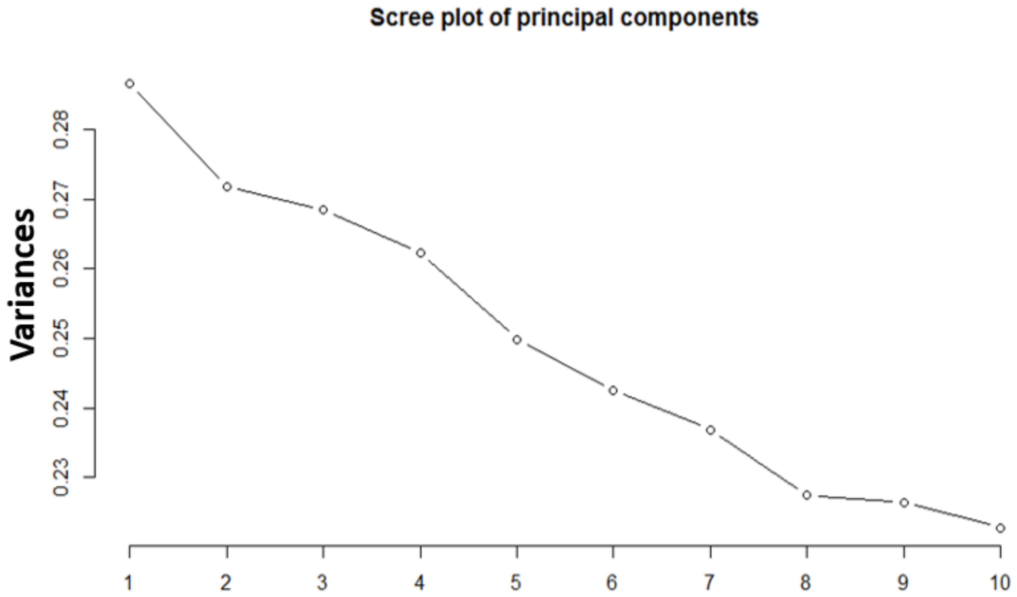


Figure 3.8: Proportion of total variance accounted for by individual principal components in a *P. falciparum* population in western Kenya.

The plot was produced for the first 10 principal components.

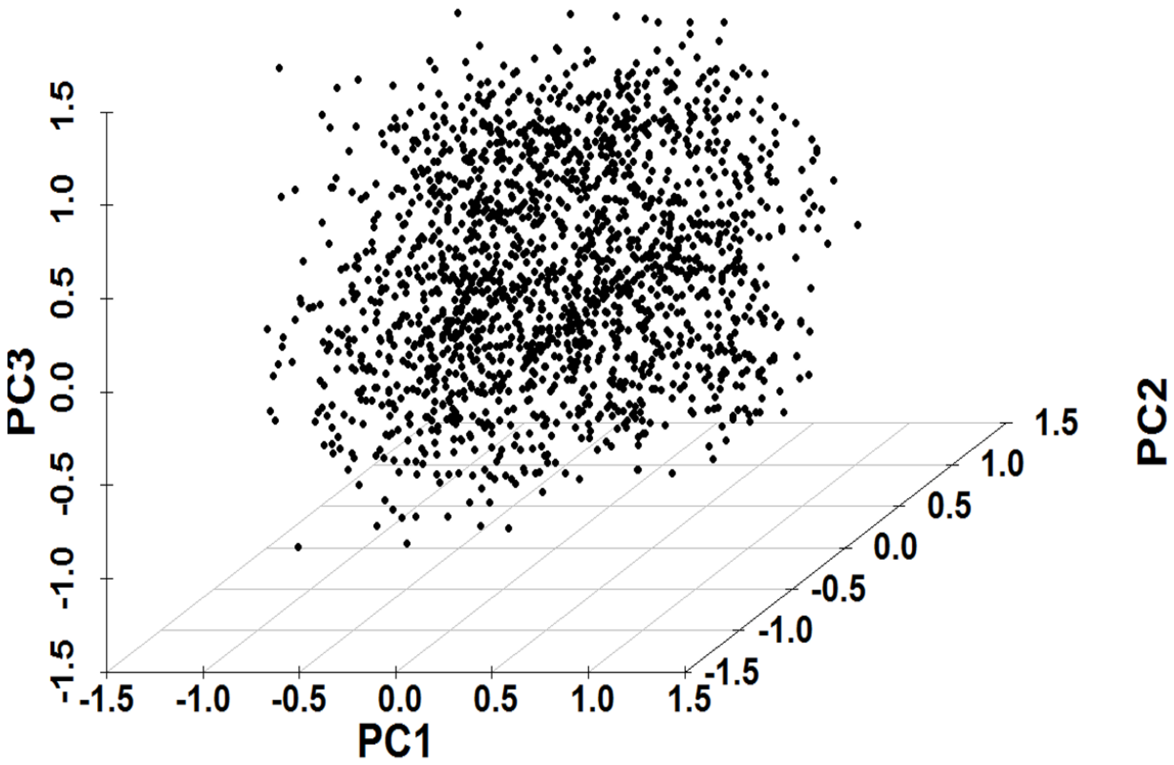


Figure 3.9: 3D Plots of principal component analysis based on scores of the first three principal components.

Each dot represents one of 1809 *P. falciparum* samples collected from 88 schools in western Kenya. Parasites clustered together and did not separate in space along any of the three principal components.

Adding spatial information to the PCA also showed little population structure among parasite isolates at the regional level, and there was no difference in population structure in lower versus higher transmission intensity areas (figure 3.10). These results are indicative of a parasite population that is well mixed within the study area.

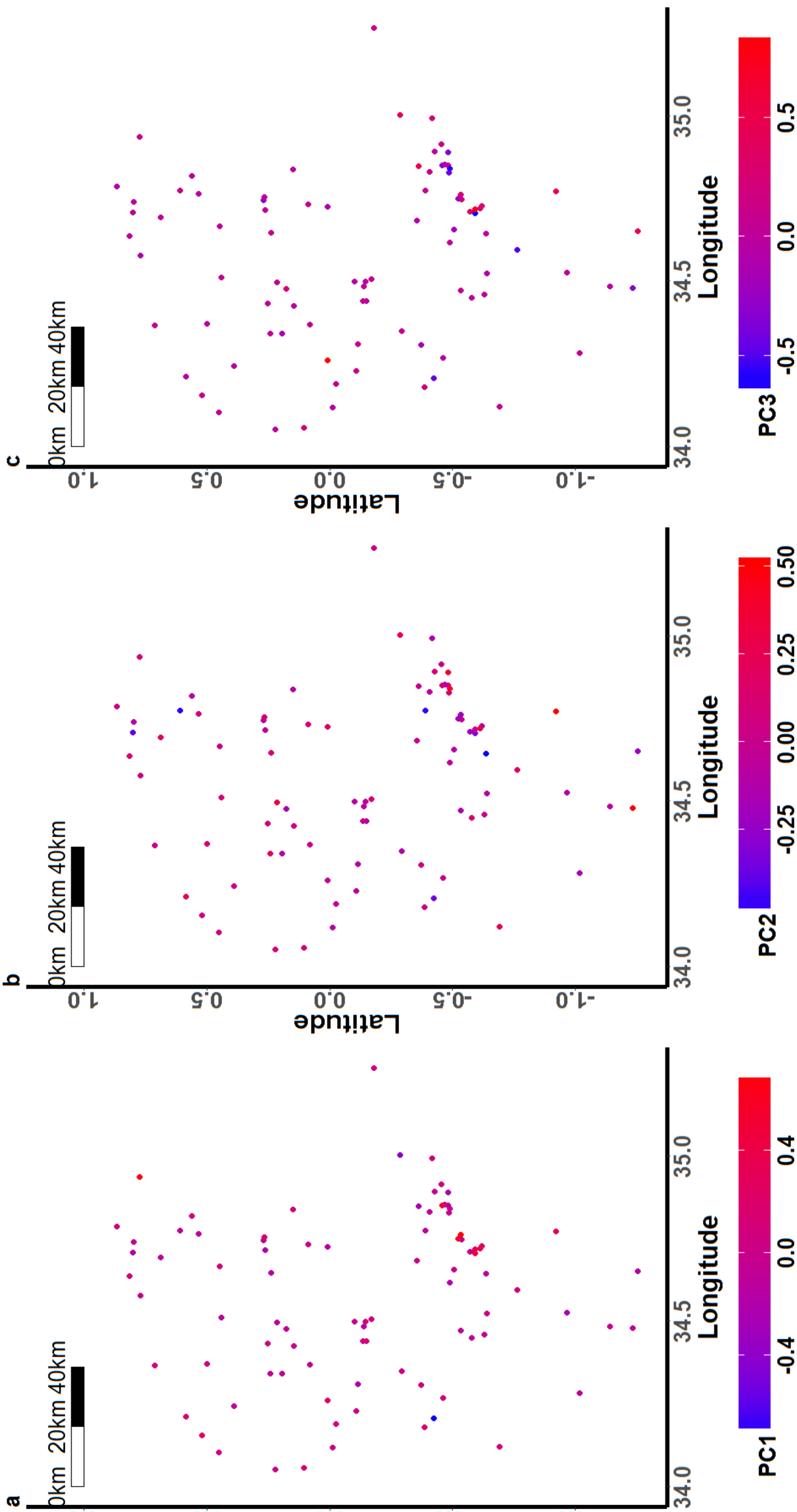

a
b
c
0km 20km 40km
Latitude
Longitude
1.0
0.5
0.0
-0.5
-1.0
34.0
34.5
35.0
PC1
-0.4
0.0
0.4
PC2
-0.25
0.00
0.25
0.50
PC3
-0.5
0.0
0.5

Figure 3.10: Spatial distribution of scores for the first 3 principal components (PCs) representing parasite genotypes.

Geospatial positioning information was collected at the school level, thus PC scores (values of the transformed variables corresponding to a specific data point) were aggregated for all samples in an individual school. Here each dot represents a school, and has been colour-coded based on the mean genotype score of all parasite isolates collected in that school. Cumulatively, the first three PCs accounted for only 10.78% of the variation observed in the genotype data (PC1=3.74%, PC2=3.54%, PC3=3.5%).

3.3.4 Spatial autocorrelation analysis of *P. falciparum* parasites in western Kenya.

To examine structure to the PC scores at a fine scale, both local (spatial scan statistics) and global (Moran's *I*) spatial autocorrelation among parasite isolates were analysed. Moran's *I* analysis showed no statistically significant trends of spatial autocorrelation among parasite pairs that were close to each other in any of the three distance classes (1km, 2km and 5km) analysed. Significant autocorrelations (p<0.01) were observed among parasites that were on average at least 20km apart in space (figure 3.11), but these autocorrelations were associated with very low correlation coefficients (< 0.03) and were not consistently seen in adjacent distance categories. Thus, the overall pattern seen from this analysis was that of little or no spatial autocorrelation in genotypes, even among parasite pairs that were very close to each other in space.

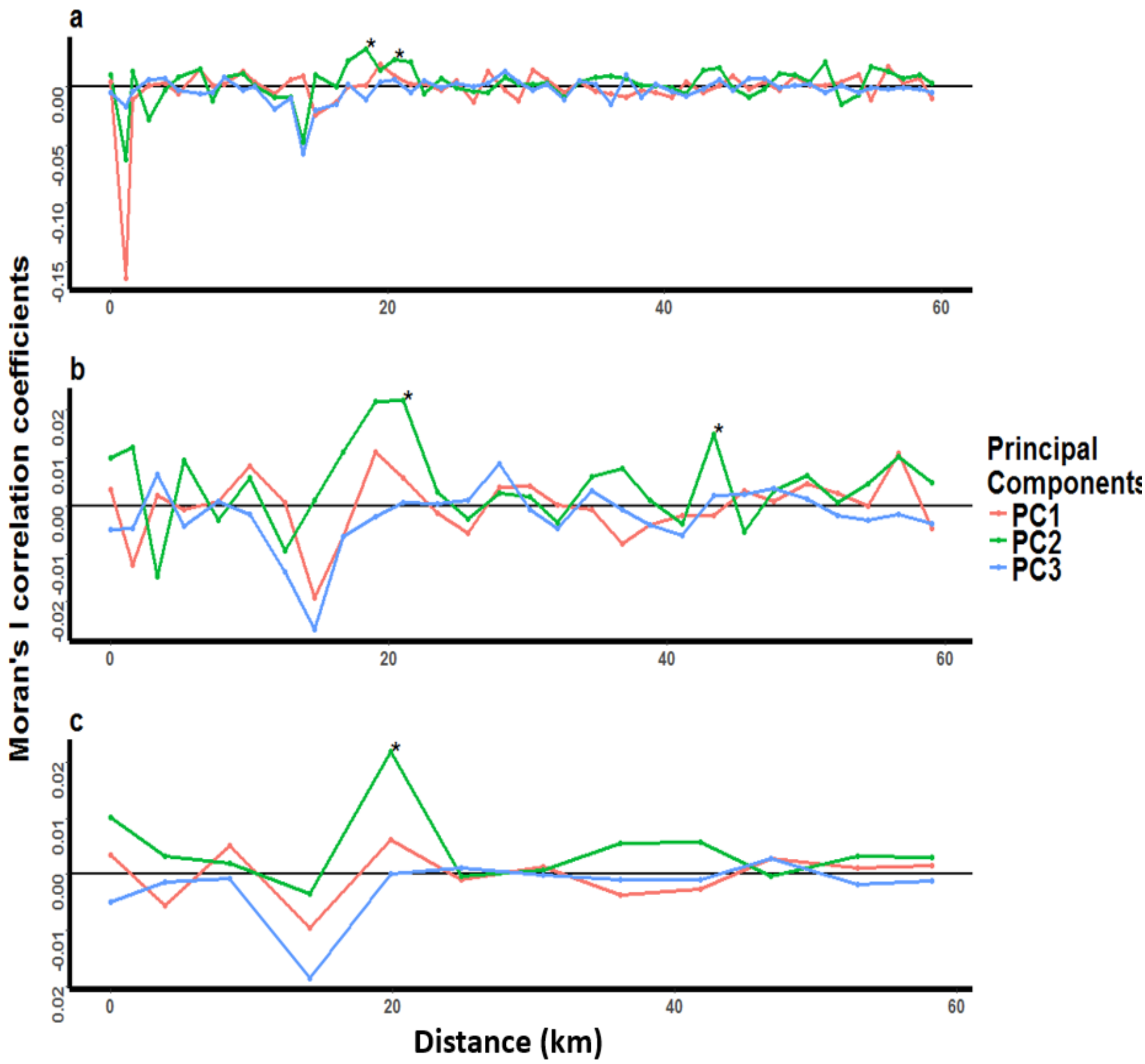


Figure 3.11: Moran's *I* correlation coefficients describing the spatial autocorrelation of genotypes of *P. falciparum* parasite pairs in western Kenya. Spatial autocorrelation was tested separately for parasites grouped into three distance classes of **a)** 1km, **b)** 2km and **c)** 5km. Within each distance class, correlations were computed for each of the first 3 principal components (PCs). The asterisks represent those distances at which statistically significant ($p<0.01$) correlation coefficients were found for parasite pairs within each distance class, indicative of possible clustering of specific parasite genotypes.

However, one statistically significant (p=0.001) cluster was identified based on the second principal component when the data were analysed for local geographical clustering of distinct parasite genotypes using spatial scan statistics (figure 3.12). This cluster was relatively large, with a radius of 67.84km, and included 852 of the 1809 samples. No significant clusters were identified when the first and third PCs were analysed.

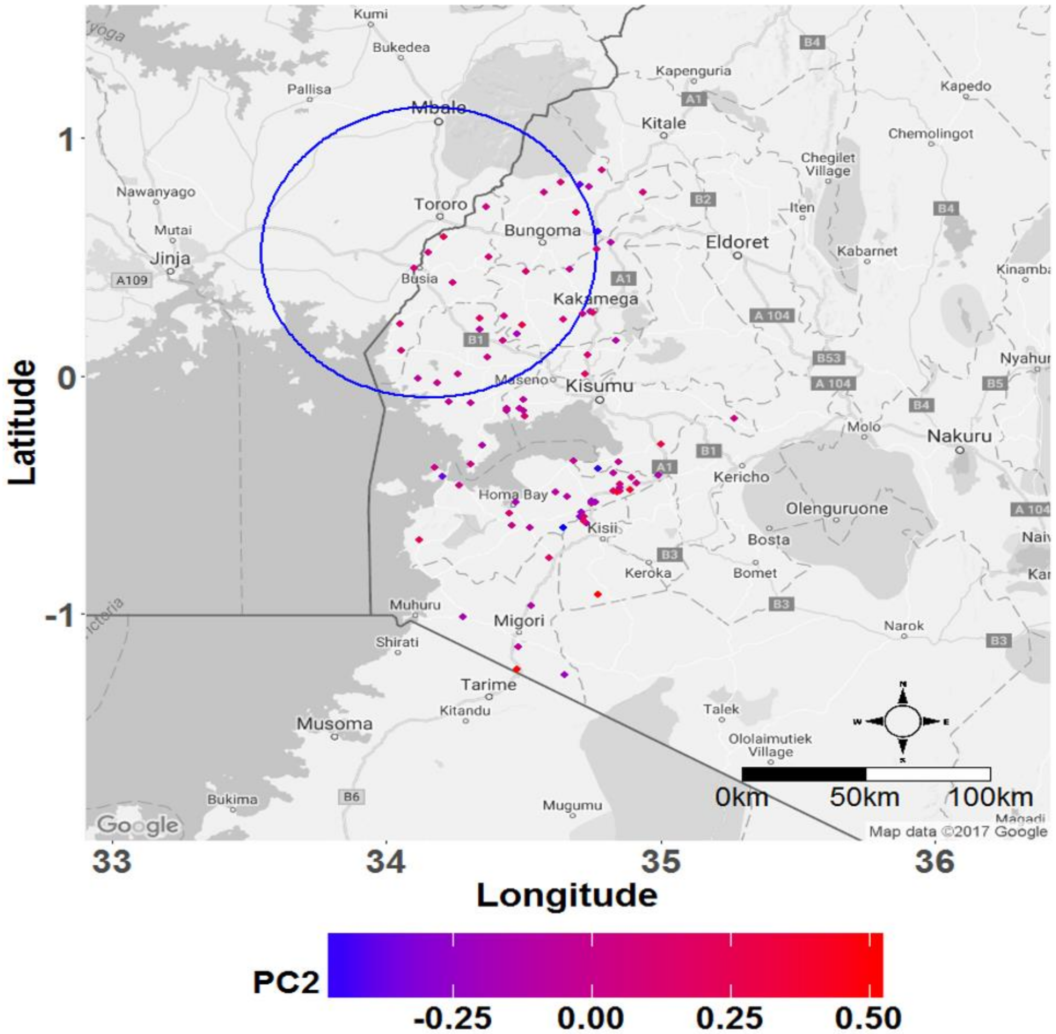


Figure 3.12: Spatial scan statistics to identify local spatially autocorrelated clusters of genetically distinct *P. falciparum* parasite sub-populations in western Kenya. Spatial scan statistics employing the use of multiple circular windows of varying sizes (ranging from covering only 1 sample up to 50% of the sample population) around samples in geographically defined regions was used to compute the ratio between expected and observed number of distinct genotypes within each window. Each window

with higher than expected number of similar genotypes was noted down as a cluster, and its statistical significance determined after accounting for the multiple comparisons. Genotypes for individual parasites were assigned based on scores of the first 3 principal components. Here, each school is colour-coded based on the mean principal component score for all parasite genotypes found within it. One cluster of highly related parasite genotypes (blue circle) was identified when analysing the second principal component.

3.3.5 Allele frequency distribution

A logistic regression model was used to examine the distribution of allele frequencies at each SNP position, and a likelihood ratio test was computed to test for the effect of school on the frequency of each SNP (i.e. determining how the frequency of each SNP varies from school to school). 18 out of 83 SNPs were shown to have statistically significant ($p < 0.05$) variations in frequencies among schools, although none of the SNPs were significant after Bonferroni adjustment for multiple testing.

These 18 SNPs were then included in a spatial scan statistics analysis, using the Bernoulli probability regression model to run a purely spatial analysis to determine the geographic pattern of allele frequency distribution within the parasite population. For each SNP position, cases were represented by the major allele in the population while controls were represented by the minor allele. 5 of the 18 SNPs produced statistically significant geographical clusters containing schools with a higher than expected number of samples with the major allele (figure 3.13; table 3.3). In each of the analyses, only the primary cluster was significant and so was represented on the maps to show how the frequency of specific alleles in schools within the clusters differed from schools outside the clusters.

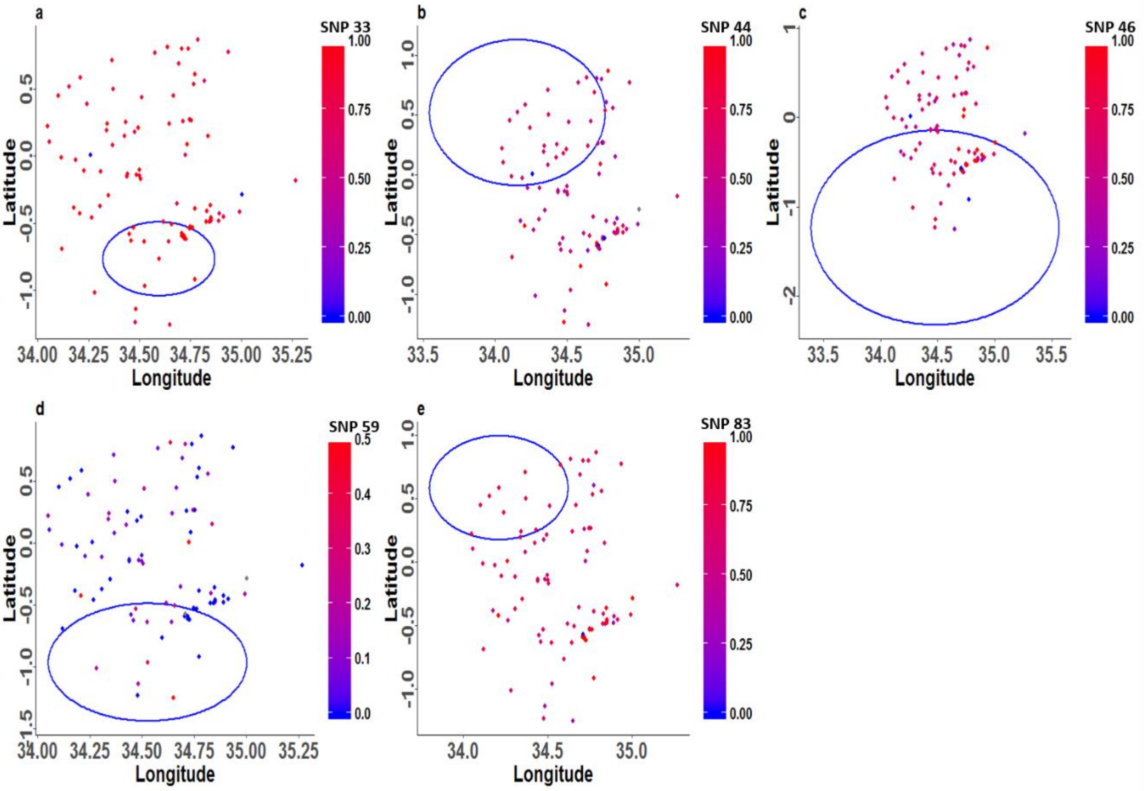


Figure 3.13: Spatial clusters of primary schools with similar allele frequency distributions of specific *P. falciparum* SNPs.

Each dot represents an individual school, colour-coded based on the frequency of a specific allele in the school. Red and blue represent different alleles at each SNP position, and the range between the two colours represent varying frequencies of each allele. Blue circles represent locally distinct clusters of schools identified to have similar frequencies of a particular allele.

Table 3-3: Spatial clusters of primary schools with similar allele frequency distributions of specific *P. falciparum* SNPs.

SNP	Size of cluster (radius in km)	Population size*	No. of cases (samples with major allele)	P value
1	30.55	166	162	0.05
2	67.84	782	495	0.001
3	120.77	454	311	0.053
4	52.87	154	25	0.033
5	45.54	375	284	0.032

*Population size includes both cases (samples with the major allele) and controls (samples with the minor allele). Clusters were generated in SaTScan based on a Bernoulli probability model (Kulldorf, 2014).

3.3.6 Spatial variations in genetic differences between *P. falciparum* parasites

Using a linear regression model to examine the effect of distance on parasite genetic relatedness showed that the number of SNP differences between parasite pairs was positively correlated with distance between the parasites (effect size = 1.85 x 10^$^{-3}$) (figure 3.14). However, bootstrapping the analysis (to take into account the linked nature of pairwise observations) gave no statistically significant effects of distance on genetic relatedness (p=0.347; 95% CI = -0.012 – 0.017). These results therefore provide no evidence for genetic isolation by distance in this parasite population.

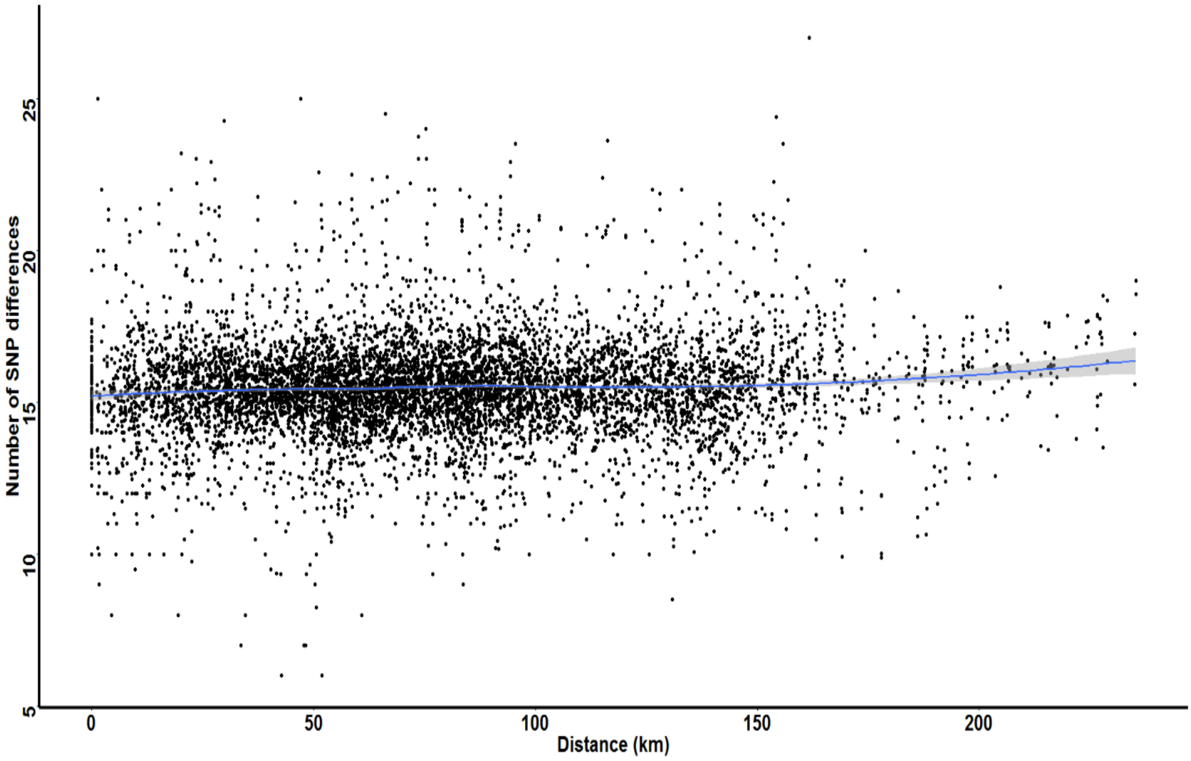


Figure 3.14: Variation in genetic diversity of *P. falciparum* parasites over distance across western Kenya.

Genetic diversity was defined as the average number of single nucleotide polymorphic (SNP) differences between parasites in each pairwise school comparison, and was plotted against the distance between the corresponding school pair. The blue line represents loess-fitted smoothing with 95% confidence intervals (grey area).

3.3.7 Spatial barriers to parasite movement and mixing

Raster analysis using 10km x 10km sized pixels was undertaken to examine the study area for spatial barriers to parasite movement. Most of the pixels were found to have a non-significant influence on the number of SNP differences among parasites, and none were significant after correcting for multiple testing using the Bonferroni correction method (figure 3.15a). Furthermore, a histogram of p values showed a null (uniform) distribution (figure 3.15b), and an analysis of spatial relationships among pixels based on coefficient estimates derived from the pixel regression showed no evidence of autocorrelation among pixels acting as either barriers to or gateways for parasite movement (figure 3.15c). Similar results were observed when the analyses were carried out using 20km-by-20km sized pixels.

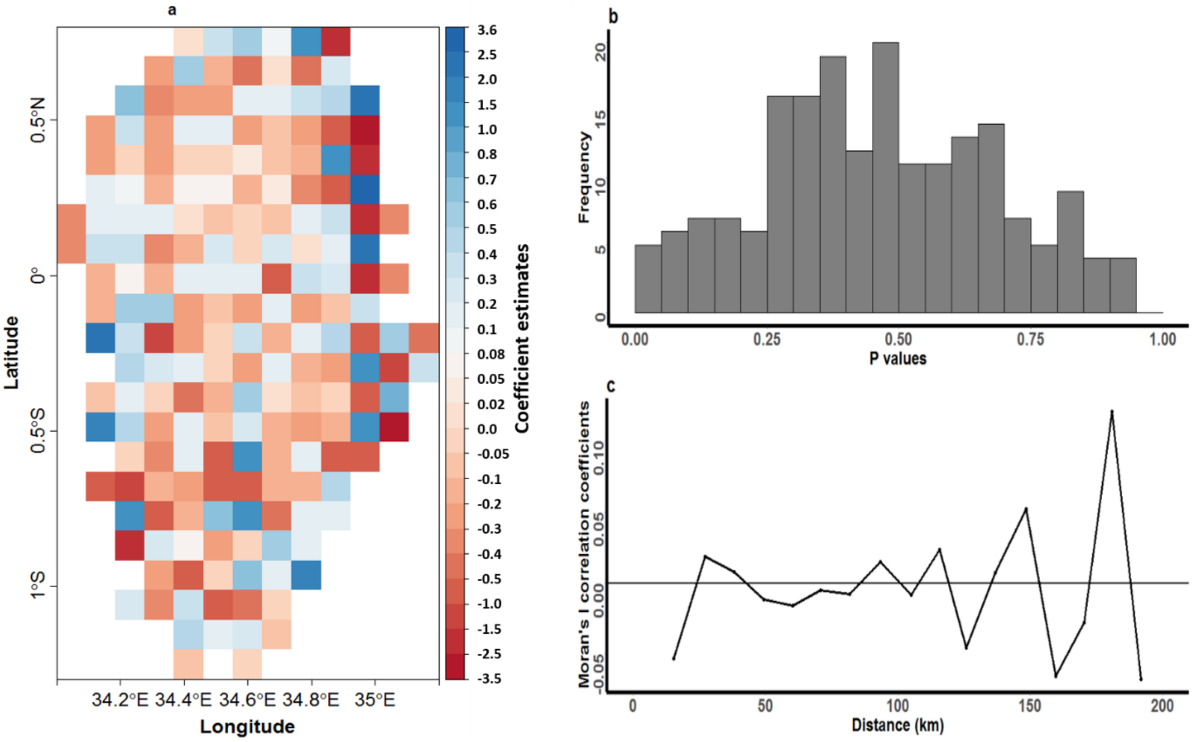


Figure 3.15: Raster analysis by pixels to examine the presence of spatial barriers to *P. falciparum* movement in a geographically defined region of western Kenya.

(a) Each pixel represents a 10km-by-10km area of the region, and is colour-coded based on the coefficient estimates derived from a linear regression analysis that was used to test the impact of each pixel in acting as either a barrier (blue pixels) or gateway (red pixels) to parasite movement within the region. No pixels were significant barriers or gateways to parasite movement after Bonferroni correction to account for multiple comparisons. **(b)** Distribution of p values observed after bootstrapping the regression analysis (with 10,000 resampling steps) to determine the level of significance of pixels in acting as barriers to parasite movement. **(c)** Moran's *I* analysis describing the spatial autocorrelation between geographical locations of pixels and their associated coefficient estimates. Autocorrelation was calculated for parasites grouped in 10km distance bands, and the analysis was bootstrapped 100 times to determine significance.

Raster analysis by pixels was further extended to examine the bearing (direction of movement) of parasites in either the east/west or north/south directions. Individually, most of the 192 pixels were not significant factors in determining directional movement of parasites over the region (figure 3.16a). However, some of the pixels were statistically significant ($p<0.0003$) in representing regions with greater east/west movement, even after accounting for multiple testing. When the regression coefficient estimates derived from analysis of bearing were included in a Moran's *I* analysis to examine the parasite population for spatially auto-correlated direction of movement, there was evidence of statistically significant ($p<0.01$) autocorrelation for school pairs that were separated by up to 40km (Figure 3.16c).

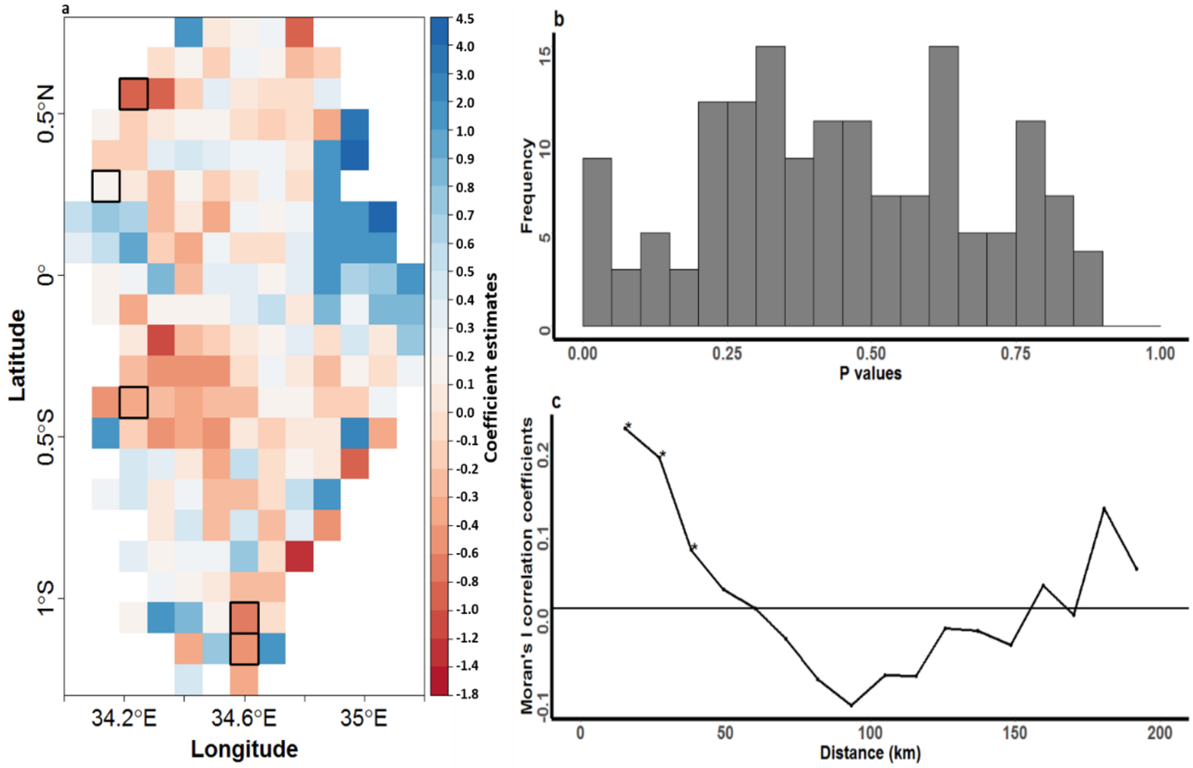


Figure 3.16: Raster analysis by pixels to examine patterns of north/south versus east/west directional movement of *P. falciparum* parasites in western Kenya.

(a) Each pixel represents a 10km-by-10km area of the region, and is colour-coded based on coefficient estimates describing the effect size of each pixel in influencing directional movement. Pixels that were statistically significant after correcting for multiple testing are highlighted with black borders. Grids were colour-coded to represent east/west (red) or north/south (blue) movement. **(b)** Distribution of p values observed after bootstrapping the regression analysis (with 10,000 resampling steps) to determine the level of significance of pixels in influencing parasite directional movement. **(c)** Moran's *I* analysis to describe the spatial autocorrelation of movement within the region. The analysis was computed using geographical coordinates of individual pixels to represent feature locations and coefficient estimates derived from the bearing regression analysis to represent the associated feature values. Autocorrelation was computed for parasites grouped in10km distance bands. Significant positive correlation coefficients (p<0.01; marked by asterisks) were observed for schools separated by up to 40 km within the 10km distance bands.

No statistically significant pixels were observed when the bearing regression analysis was repeated at 10km^2 and 20km^2 spatial scales using only those pixels with at least 150 school-school pairs crossing their boundaries. However, two separate clusters of pixels were identified within the region that showed patterns of specific directional movement, one in the north east (indicative of greater north/south movement) and another in the west (indicative of greater east/west movement) (Figure 3.17).

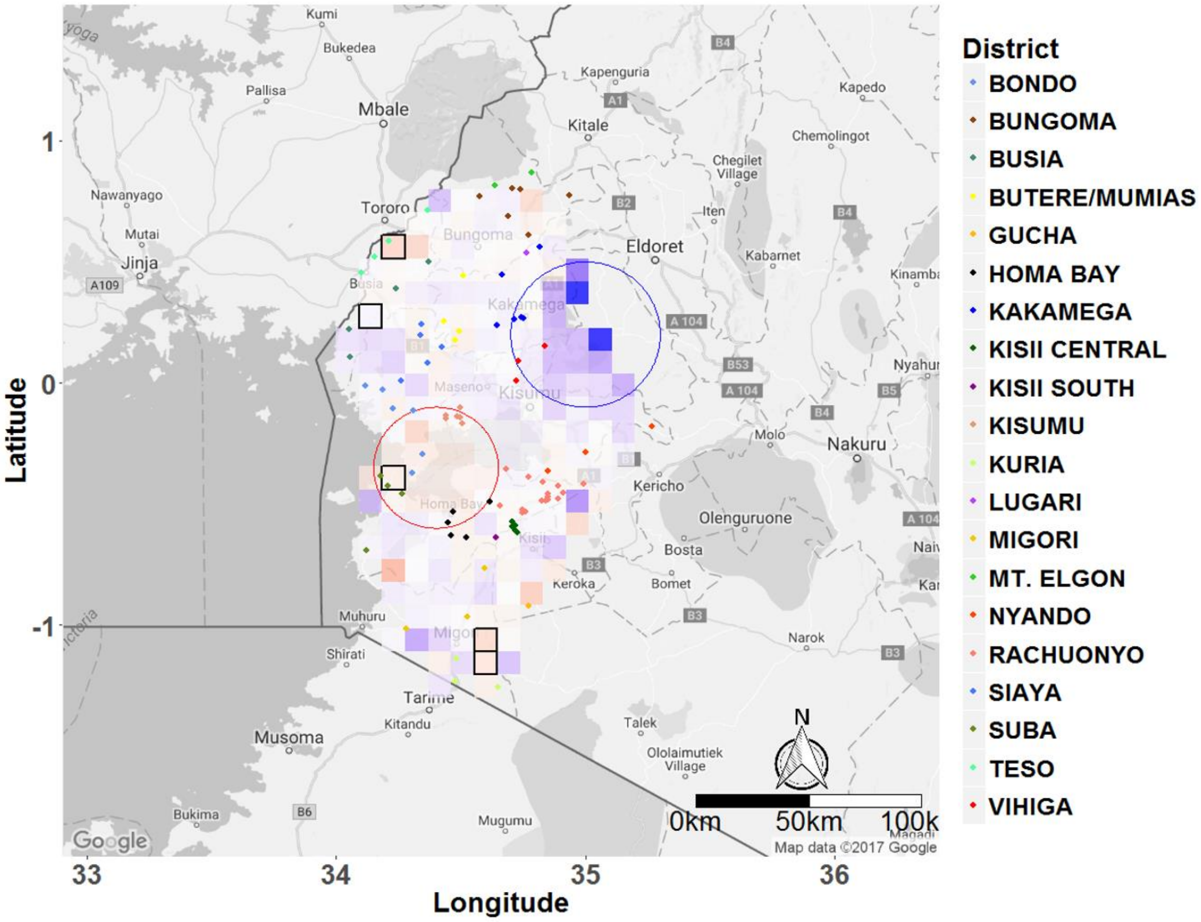


Figure 3.17: Map of the western Kenya study area with raster grids representing bearing analyses superimposed on top of it.

Multivariable linear regression analysis was carried out to determine bearing (directionality of movement) of *P. falciparum* parasites among schools in the region. Grids are colour coded based on the coefficient estimates describing the effect size of that grid in influencing directional movement. Red represents east/west movement, while blue represents north/south movement. The grids with black borders represent those areas that were significant in east/west movement, even after Bonferroni-correction for multiple testing. The blue circle shows the region of the study site that had predominantly north/south movement, while the red circle represents that region that had predominantly east/west movement. Each dot represents a school, colour-coded based on the district in which the school is located.

3.4 DISCUSSION

In chapter two of this thesis, I used SNP genotype data to examine the level of genetic relatedness among *P. falciparum* parasites on a micro-epidemiological scale within three regions with varying transmission intensities in Kenya and the Gambia, and found evidence of spatial sub-structure over short distances (i.e. <10km), despite a high level of parasite mixing (Omedo *et al*., 2017a). In the present analysis, I examined the level of parasite mixing at a sub-national scale in Western Kenya, using parasitological data from primary school surveys to describe the patterns of parasite mixing at a larger geographical scale.

Primary school children were selected as the study population because they are easy to sample. Furthermore, infection diversity peaks at 3 -14 years and then declines in older age groups in high transmission settings (Konate *et al*., 1999, Owusu-Agyei *et al*., 2002, Smith *et al*., 1999), hence our study sample is likely to contain a diverse genetic pool representative of parasites circulating in the region. Sampling only asymptomatic infections in schools may not give the whole range of genetic diversity within the region, as one study identified specific polymorphisms in *AMA1* that could have been more frequent in symptomatic infections compared with asymptomatic infections (Cortes *et al*., 2003). Young children with symptomatic infections would be absent from school and away from the sampling frame. However, the sampling strategy used here was consistent across the different schools and furthermore, the evidence of genomic variation in parasites according to clinical outcome is limited.

This study showed evidence of high genetic diversity in the Western Kenya parasite population, consistent with the high malaria transmission intensity experienced in this region (Ingasia *et al*., 2016, Kenya national malaria control programme, 2016, Noor *et al*., 2014, Okiro *et al*., 2009). Of the five malaria transmission zones in Kenya, Western Kenya currently experiences the highest transmission intensity (Kenya national malaria control programme, 2016) despite efforts to scale up various control interventions, such as long lasting insecticide nets, indoor residual spraying and artemisinin combination therapy, in this region (Gatei *et al*., 2015, Ototo *et al*., 2015, Zhou *et al*., 2016).

Using PCA, I did not identify any genetic structure through inspection of the PC plots derived from SNP genotype data. This indicates an absence of discrete sub-populations within this *P. falciparum* parasite population, and is in agreement with the previous analysis of parasites from the same region (Omedo *et al*., 2017a), and with whole genome data from different African countries (Manske et al., 2012, Mobegi et al., 2012). In South-East Asia, distinct sub-populations associated with antimalarial drug resistance have been detected (Miotto *et al*., 2015). Previous analyses of *P. falciparum* population structure in western Kenya also showed high genetic diversity and little population differentiation in this parasite population (Baliraine *et al*., 2010, Bonizzoni *et al*., 2009, Zhong *et al*., 2007).

In contrast with the previous study of *P. falciparum* parasite populations from The Gambia, Kilifi and Rachuonyo South discussed in the previous chapter (Omedo *et al*., 2017a), analysis of trends in spatial relationships among parasite genotypes identified no significant autocorrelations using Moran's *I* spatial autocorrelation analysis in this study. Overall, the consistent pattern observed across all distance classes was that of no autocorrelation among parasites in schools at all distances, with occasional inconsistent associations that were considered likely to be spurious.

Using the spatial scan statistics, only a single cluster of genetically related parasites was identified based on the second principal component. This limited genetic clustering at both local and global scales, and weak evidence of genetic isolation by distance, are indicative of a parasite population that is well mixed at the sub-national geographical scale. This finding is in contrast to the micro-epidemiological study discussed in the previous chapter, which showed spatial structure to genetic relatedness over short distances (Omedo *et al*., 2017a). In that previous study, however, it was noted that the gradient between spatial separation and genetic relatedness was non-linear, and became less steep with distance such that past 10km there was little genetic differentiation. This observation was hypothesized to be as a result of parasite movement and mixing observed within the study sites, with no geographical areas acting as spatial barriers to parasite movement. It is therefore consistent that no relationship was identified in this study of schools where most pairs of schools were more than 10km apart.

The geographical distribution of allele frequencies for all 83 SNPs in our study population was also examined. Studies of allele frequency distribution have been used to determine parasite population structure and identify patterns of local adaptation of *P. falciparum* isolates (Anderson *et al*., 2005, Gunther and Coop, 2013, Schlotterer, 2002). Such local adaptation may be due to various selection pressures, including environmental pressure and immune selection, and may occur at individual, population or regional scales (Kaltz and Shykoff, 1998, Ochola *et al*., 2010). Of the 83 SNPs examined, 18 SNPs were shown to have statistically significant variations in allele frequencies among schools based on a logistic regression analysis, although none of these 18 SNPs were significant after accounting for multiple testing. These findings suggest that although we were not sufficiently powered to distinguish any individual

SNPs as likely to be significant beyond a Bonferroni correction, on the other hand the fact that 18 SNPs showed a value of $p<0.05$ when only 4 would be expected by chance suggests that there may be some genuine differences in frequency within the group of SNPs.

Reasoning that genuine geographical variation would be likely to show spatial clustering as well as variation by school, the scan statistic was subsequently measured for those SNPs showing significant variation among schools. 5 of the 18 SNPs that were identified showed local clustering among schools. However, these SNPs were not entirely private to a sub-population and occurred in schools inside and outside the clusters. This finding provides weak support for the existence of variable local genetic selection pressures in this parasite population. The identification of SNPs with significant geographical variation in allele frequencies could indicate adaptation of *P. falciparum* parasite populations to their local environment, or more likely may indicate a temporary expansion of a parasite sub-population with a particular SNP simply due to random genetic drift.

An extensive analysis of the study area for spatial barriers to parasite movement using 10km-by-10km and 20km-by-20km sized pixels provided little evidence for the existence of geographical regions that act as barriers to parasite movement at the sub-national scale, and is in agreement with similar analysis in the previous chapter which did not identify any barriers to parasite movement at a micro-epidemiological scale (Omedo *et al*., 2017a). This observation of free movement over the western Kenya region is also supported by a previous analysis of mobile phone data that was used to analyse patterns of human movement within the country (Wesolowski *et al*., 2012) , and which showed substantial movement of people within this region, and further supports the observation here of little or no barriers to parasite movement within the region.

However, a cluster of pixels representing predominantly north/south movement in the north east and another cluster representing predominantly east/west movement in the west of the study area were identified, and when the site was analysed for spatial autocorrelation in the directionality of pixels, statistically significant autocorrelations were observed for school pairs separated by up to 40km. This means that pixels with greater east/west movement were more frequently found next to other pixels with greater east/west movement, and similarly, pixels with greater north/south movement were more frequently found next to pixels with greater north/south movement. Furthermore, some individual pixels showed statistically significant directionality that met Bonferroni-adjusted significance criteria. Although spatial autocorrelation of directionality might simply be because the same data (the same school pairs) cross pixels that are physically close to each other, the observation of two large clusters of pixels with distinct directional patterns of parasite movement is unlikely to have been an artefact of the same school pairs being analysed when all pixels in the clusters were considered, suggesting that the analysis could detect specific migration pathways of parasites.

The findings in this study have several implications for the outcomes of malaria control programmes. Since they show that parasite populations mix to high degrees within the region, with little evidence of geographical clustering, one might conclude that interventions targeting smaller geographical areas within the region are likely to reduce the flow of parasites to regions beyond those targeted. However, the high degree of parasite mixing also means that parasites move relatively freely within the region, and there is therefore a high likelihood of importation of infection from untargeted to targeted regions. This is strongly corroborated by evidence from a cluster-randomized controlled trial in a highland region of western Kenya that showed no impact in reducing transmission inside hotspots 16 weeks after applying interventions, possibly due to the importation of parasites from untargeted surrounding regions (Bousema *et al*., 2016).

This study had some limitations. First, we cannot be definite about the time-scale over which gene flow has occurred. If the gene flow is rapid, this supports this study's conclusions regarding malaria control. On the other hand, it is possible that the well-mixed population emerged over a longer period and that gene flow, while resulting in complete mixing, could be less rapid, in which case targeted interventions would probably not have far reaching effects in the surrounding community. The results of the previous analyses showing spatial and temporal structure at a fine micro-epidemiological scale among parasites in The Gambia, Kilifi and Rachuonyo South suggests rapid gene flow (Omedo *et al*., 2017a). In that study, parasite pairs taken from nearby homesteads had fewer SNP differences between them than parasite pairs that were further apart. However, over the period of a month this distance gradient was attenuated, and was gone by one year such that there were barely any SNP differences between parasite pairs collected one year apart and those collected three years apart, irrespective of the distance between them. However, more definitive work will require an in-depth analysis of whole genome data to identify haplotypes and rare variants in the population, and infer variation over time.

Second, geospatial coordinate data were collected for schools as opposed to individual homesteads, and hence genotype data were aggregated at this level. Structure at micro-epidemiological scales would likely be missed during this analysis. Third, only a small number of SNPs were analysed. This made it impossible to detect relatively rare private SNPs. It is likely that a larger set of genetic markers will be required to identify private SNPs and evidence of local parasite adaptation. SNPs in genes previously shown to be under selection in the parasite genome may additionally be analysed to determine whether population structure is observed based on local variations in selection pressure.

Our previous study showed no population structure when SNPs in EBA175 and *AMA1* were analysed, we therefore did not type and analyse separately SNPs in antigenic genes for the present study.

Fourth, genotype data collected from only one part of the country were analysed, making it impossible to describe patterns of parasite flow across the country, or to generalize these findings to other geographical areas. Additional analyses of samples from other regions of the country that experience malaria transmission such as coast, eastern and north-eastern provinces are recommended. Furthermore, over longer distances human movement becomes more important than mosquito movement in distributing parasites and therefore will need to be taken into consideration when analysing parasite genetic relatedness across large spatial scales. Information on travel distance can be obtained from travel history, or more objectively from mobile phone data, and can be used to track human movement between sources and sinks of parasite transmission. Concordance between spatial parasite genetic relatedness and human movement will further support our hypothesis of high parasite movement and mixing.

In conclusion, this study has shown that parasites mix to high levels within the western Kenya region, with no evidence of parasite sub-populations or geographical barriers to parasite movement, and weak evidence of spatial autocorrelation of parasite genotypes at the local and global scales. It has also shown that directionality of parasite migration can be inferred based on genetic relatedness, and gene flow models, e.g. as implemented in migrate-n software, can be used to determine the migration rates within the region, although such models are likely to prove more useful if distinct parasite populations exist and can be identified within the region.

Chapter 4 Analysis of Spatio-Temporal Genetic Variation among *Plasmodium falciparum* Parasites in Kilifi County Between 1995 and 2014 using Sequence Data of Target Genes.

4.1 INTRODUCTION

In chapter two of this thesis, I used genome-wide distributed single nucleotide polymorphic (SNP) data to study the rate and extent of parasite mixing, by analysing the spatio-temporal changes in genetic variations of *P. falciparum* parasites at micro-epidemiological geographical scales in three sites with differing transmission intensities in Africa (Omedo *et al.*, 2017a). In chapter three, I carried out a similar analysis but at a larger, sub-national scale to determine whether the patterns of parasite movement and mixing that are observed at micro-epidemiological scales are still present at a macro-epidemiological scale (Omedo *et al.*, 2017b). The results of the analyses in the two chapters showed evidence of clustering of distinct parasite sub-populations over short distances (< 10km), and little or no clustering of parasite sub-populations over larger distances, indicating that there is a high degree of parasite mixing within the studied sites.

Both studies used SNP data generated on the sequenom genotyping platform to analyse *P. falciparum* population genetics and arrive at the conclusions drawn about parasite movement and mixing. However, genotype data can show different patterns of population genetics and demographic history based on the SNPs that are analysed, due to SNP ascertainment bias which can arise as a result of the process by which SNPs are chosen (Lachance and Tishkoff, 2013). Therefore, to further validate these results, and to confirm that the observations made were not merely due to the SNP subset selected or the genotyping technique used to generate the data in the two studies, additional analyses of temporal and spatial genetic variations were carried out using Sanger-derived sequences of genes encoding two important *P. falciparum* proteins, SURFIN$_{4.2}$ and Apical Membrane Antigen 1 (*AMA1*). Other than their functional roles, these genes were also selected because they contain a high number of SNPs, allowing different parasite isolates to be distinguished, as well as the absence of structural polymorphisms such as variable number of tandem repeats, which would have made the analysis more complicated. Analysing sequence data also provides the added advantage of allowing the identification of the different haplotypes that were circulating in the population over the study period. Although haplotypes can be inferred using data from isolated SNPs, the method becomes less sensitive when using SNPs that are widely dispersed throughout the genome, due to the high level of recombination that occurs particularly in African *P. falciparum* populations.

4.1.1 *P. falciparum Surf$_{4.2}$* genetic diversity.

SURFINs are a family of 200 – 300 kDa type 1 transmembrane proteins expressed on the surface of merozoites and infected red blood cells (iRBCs) (Mphande *et al*., 2008, Winter *et al*., 2005). They contain domains that share sequence and structural similarity with iRBC-expressed proteins of other *plasmodium* parasites such as *P. falciparum* erythrocyte membrane protein 1 (PfEMP1), VIR family proteins of *P. vivax* and *P. knowlesi*'s variant surface antigen (SICAvar) (Mphande *et al*., 2008, Winter *et al*., 2005). The SURFINs are a polymorphic group of proteins encoded by a 10-member multi-gene family called surface-associated interspersed genes (*surf* genes) in *P. falciparum* 3D7 and include 3 predicted pseudogenes (Winter *et al*., 2005). These genes are located within the subtelomeric regions of 5 chromosomes (1, 4, 8, 13, and 14) (Mphande *et al*., 2008), and are classified as group A or group B depending on whether they contain 2 or 3 exons, respectively (Winter *et al*., 2005). Not much is known about these proteins, and only two members, SURFIN$_{4.1}$ and SURFIN$_{4.2}$, have been described in detail (Gitaka *et al*., 2017, Mphande *et al*., 2008, Winter *et al*., 2005, Xangsayarath *et al*., 2012).

The *P. falciparum* SURFIN$_{4.2}$ encoding gene, *surf*$_{4.2}$ (PlasmoDB ID: PF3D7_0424400), which is the focus of this study, is located in the subtelometric region of chromosome 4. The 7301 nucleotides long gene is composed of 2 exons (exon 1 is 2292 nucleotides long and exon 2 is 4851 nucleotides long) that are separated by a short, 158-base pair intronic region. This gene encodes a 2380 amino acids long protein that is divided into an N-terminal domain (amino acids 1-50), a cysteine rich domain, CRD (amino acids 51 - 195), a variable region (amino acids 196 - 733), a single transmembrane domain (amino acids 734 - 764) and an intracellular domain (amino acids 765 - 2380) (figure 4.1) (Kagaya *et al*., 2015). In terms of genetic diversity, the N-terminal region and cysteine rich domains are relatively conserved, both within and between species. This region is followed by a highly variable (var) region, and a transmembrane domain which is also conserved (Kagaya *et al*., 2015, Winter *et al*., 2005). The C-terminal region of the protein contains 3 to 4 segments of highly conserved tryptophan rich domains (WRDs) composed of about 145 amino acids that are separated by highly variable sequences (Winter *et al*., 2005).

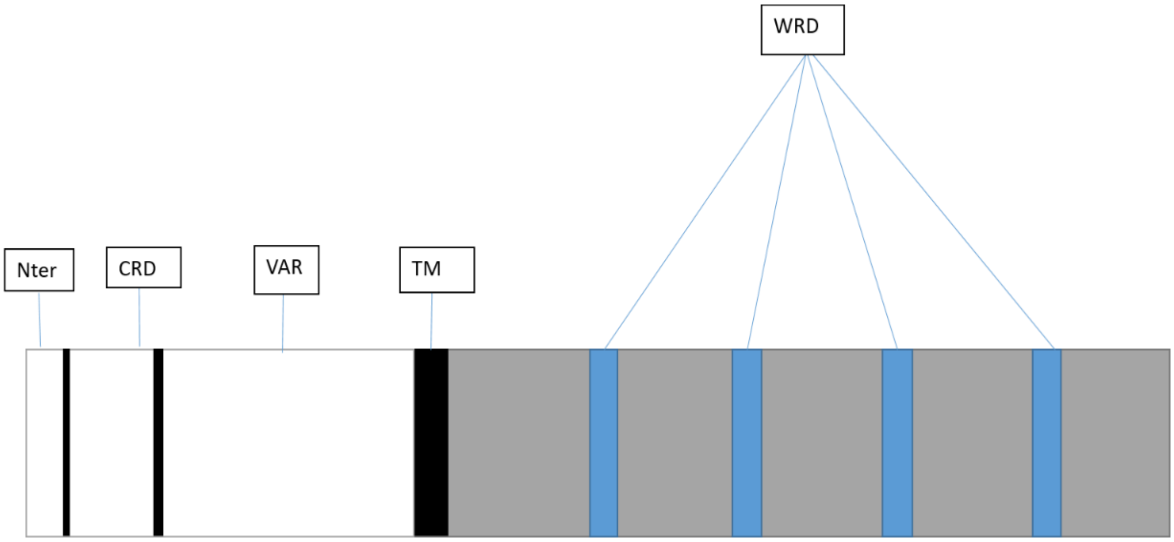


Figure 4.1: Schematic showing the *P. falciparum surf$_{4.2}$* domains.

Nter = N-terminal region; CRD = cysteine rich domain; VAR = variable region, TM = transmembrane domain; WRD = tryptophan rich domains. WRD are found in the intracellular domain and are highly conserved within and between *Plasmodium* species.

The exact role that SURFIN$_{4.2}$ plays has not been determined conclusively. The protein is shown to be co-transported to the surface of the infected red blood cells together with *P. falciparum* erythrocyte membrane protein 1 (*Pf*EMP1) and RIFINs, while in released merozoites, it is present at the apical complex (Winter *et al.*, 2005). Other studies show that SURFIN$_{4.2}$ may be involved in modifying and maintaining the structure of the infected red blood cell, as disrupting the gene lowers the rigidity of iRBCs (Kagaya *et al.*, 2015). iRBC rigidity is important for parasite retention and blockage within micro-capillaries, and would point to an important role for the protein in causing severe disease (Kagaya *et al.*, 2015). Due to its high level of polymorphism and co-localization with *Pf*EMP1 and RIFINs to iRBC surface, the antigen may also play a role in immune evasion (Chan et al., 2014, Ochola et al., 2010).

Population genetics analyses indicate that the gene is under balancing selection in a Kenyan population (Ochola *et al.*, 2010) and Thai population (Kaewthamasorn *et al.*, 2012). These, together with evidence that antibodies against the antigen inhibited erythrocyte invasion (Winter *et al*., 2005) , indicates that this protein may be a potential vaccine candidate.

4.1.2 *P. falciparum* Apical Membrane Antigen 1 (*Pf*AMA1)

P. falciparum Apical membrane antigen 1 (*PfAMA1*) is an important merozoite protein that is one of the leading vaccine candidates against blood stage malaria infection (Remarque *et al*., 2008, Thera *et al*., 2016). It is an 83kDa type 1 integral membrane protein that is structurally conserved and varies in size from 556 to 563 amino acids in most *plasmodium* species (Chesne-Seck *et al*., 2005). In *P. falciparum*, it is 622 amino acids long (Chesne-Seck *et al*., 2005, Remarque *et al*., 2008) and consists of an N-terminal signal peptide (25 amino acids), an extracellular domain (527 amino acids), a single transmembrane domain (20 amino acids) and a short cytoplasmic domain (50 amino acids) (Chesne-Seck *et al*., 2005). The extracellular region can be divided into 3 sub-domains (I, II and III) based on pairwise disulphide bonding of 16 conserved cysteine residues (Hodder *et al*., 1996, Polley and Conway, 2001, Remarque *et al*., 2008).

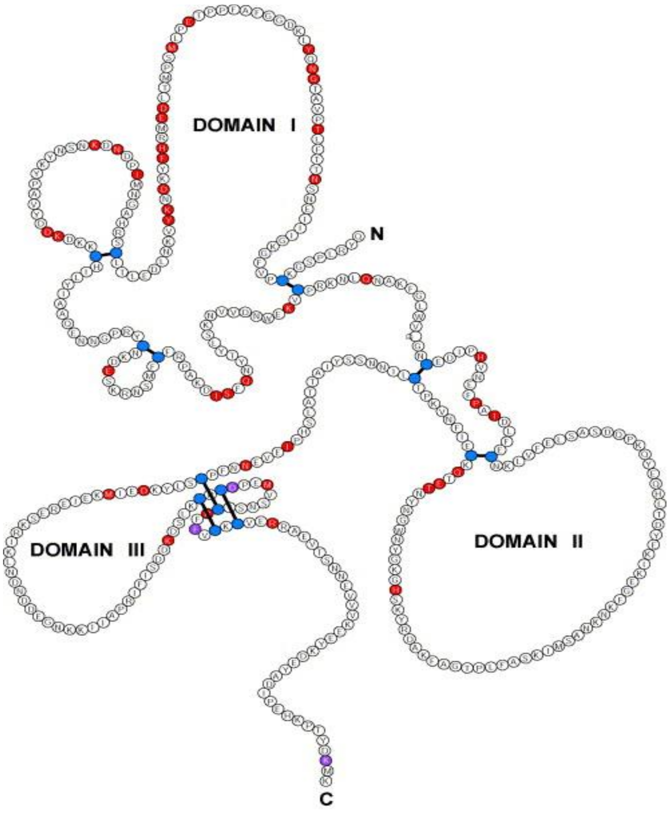


Figure 4.2: Schematic showing the ectodomain structure of *P. falciparum* AMA1 antigen.

The diagram shows the extracellular domain (amino acid residues 141 - 538). The three sub-domains (I, II, III) defined by disulphide bonds (blue) are indicated. Red and Purple shaded positions represent polymorphic amino acids. Figure borrowed from (Nair et al., 2002).

PfAMA1 is synthesised during the parasite's erythrocytic development stage and is translocated within micronemes (Bannister *et al*., 2003). Prior to merozoite invasion of red blood cells, the 83KDa precursor protein is cleaved into a 66kDa mature protein which is then translocated onto the merozoite surface (Narum and Thomas, 1994). Functionally, *PfAMA1* interacts with *Pf*RON2 protein to form the irreversible tight junction which commits the parasite to invasion (Srinivasan *et al*., 2011).

Genetic diversity in *P. falciparum AMA1* presents as single nucleotide polymorphisms (SNPs) (Takala *et al*., 2009). *AMA1* shows a high level of genetic diversity which poses a challenge for the development of a vaccine based on this antigen (Drew *et al*., 2012, Thera *et al*., 2011), with approximately 10% of the amino acid positions being polymorphic (Chesne-Seck *et al*., 2005, Remarque *et al*., 2008, Takala *et al*., 2009). Genetic diversity in the gene is thought to be maintained by balancing selection, driven primarily by host immunity (Polley and Conway, 2001, Thera *et al*., 2008). Evidence of balancing selection has been detected in both domains I and III when analysing parasite populations from Kenya (Osier *et al*., 2010), Nigeria (Polley and Conway, 2001) and Papua New Guinea (Arnott *et al*., 2014, Cortes *et al*., 2003).

The greatest genetic diversity within the gene is found in domain I near a hydrophobic pocket that is thought to be the binding site for *PfAMA1* and proteins forming the erythrocyte invasion machinery (Cortes *et al*., 2003, Takala *et al*., 2009). Domain II has been shown to have the lowest genetic diversity among the three domains (Zhu *et al*., 2016) , and contains an epitope that forms the binding site for a monoclonal antibody that inhibits invasion (Chesne-Seck *et al*., 2005). At a population level, diversity in *PfAMA1* gene varies with *P. falciparum* transmission intensity (Zhu *et al*., 2016). Greater genetic diversity occurs in African parasites, with over 200 haplotypes identified in one village in Mali (Takala *et al*., 2009), and lower haplotype diversity in a Papua New Guinea population (Cortes *et al*., 2003). However, most of these genetic variations occur at low frequency. For example, in the population genetic analysis of 506 *P. falciparum* isolates collected in Mali, nearly half of 214 haplotypes identified were seen only once, i.e. were unique to an individual sample (Takala *et al*., 2009).

At the continental level, genetic structure within *PfAMA1* was detected when comparing sequences of *P. falciparum* parasites from Africa, Southeast Asia, Oceania (Papua New Guinea) and South America (Zhu *et al*., 2016) and a separate study of the global distribution of *PfAMA1* alleles found variations in allele frequencies when analysing different regions of the gene (Takala *et al*., 2009). Additional analysis of genetic diversity among sequences from different regions of the world shows that most of the diversity exists within samples collected in the same geographic region (96.9%), with little variation between samples collected in different regions (3.1%) (Duan *et al*., 2008).

Evidence of the clinical impact of *PfAMA1* polymorphisms is limited and contradictory. Certain polymorphisms have been associated with symptomatic disease in Wosera, Papua New Guinea, indicating that certain strains of *PfAMA1* are a determinant of morbidity associated with the disease (Cortes *et al*., 2003). In a rural population in coastal Kenya, differences in nucleotide frequencies were observed at 16 polymorphic positions when comparing symptomatic (both severe and mild malaria cases) to asymptomatic cases (Osier *et al*., 2010). However, an analysis of *PfAMA1* using restriction fragment length polymorphism (RFLP) did not find an association between specific genotypes and either symptomatic or asymptomatic infections in a rural area of Burkina Faso (Soulama *et al*., 2015).

This current study examined the spatio-temporal genetic variation of *P. falciparum* parasites using sequence data of genes encoding the two parasite proteins *AMA1* and SURFIN$_{4.2}$, described above, based on two different but complementary metrics:

1) Number of SNP differences.
2) Length of DNA segment that is identical by sequence (IBS) between parasite pairs.

As the name suggests, identical by sequence (IBS) is a term used to describe an identical sequence of DNA at a specific locus in two or more parasites. The number of SNP differences was computed as a measure of genetic diversity among parasites, whereas IBS was computed as a measure of genetic similarity among parasites, with an aim of determining whether the two metrics gave similar results regarding changes in parasite genetic variation over time and space. *PfAMA1* and *surf*$_{4.2}$ genes were sequenced and analysed in *P. falciparum* isolates collected from children admitted with either mild or severe malaria at the Kilifi County Hospital between 1995 and 2014.

For each gene, sequences were analysed to identify the number of polymorphic sites, and to determine the level of sequence diversity. The patterns of genetic variation in time and space were visualised using heatmaps, and phylogenetic trees were generated to depict the evolutionary relationship among the sequences. Neutrality tests were performed to determine whether the sequences were evolving randomly or under selective pressure. Additional analyses of linkage disequilibrium and recombination were also carried out. Finally, the effect of time and space on *P. falciparum* parasite genetic variation was interrogated using both pairwise IBS and SNP differences data.

4.2 MATERIALS AND METHODS

4.2.1 Study population

P. falciparum positive blood samples were collected from 1164 children aged 3 months to 13 years old who were admitted to Kilifi County Hospital (KCH) with mild or severe malaria between 1995 and 2014. Most of these children were resident within the Kilifi Health and Demographic Surveillance System (KHDSS) catchment area and their residential locations had been geocoded as part of an on-going surveillance. The KHDSS, which was set up in 2000 to link hospital morbidity and mortality surveillance data to community surveillance data, has been described in detail previously (Scott *et al*., 2012). Changing trends in malaria transmission was recorded during the study period, with high transmission intensity witnessed before 1998, a decline in transmission observed between 1998 and 2009, and a steady increase in transmission after 2009 (Mogeni *et al*., 2016).

4.2.2 Sample selection and DNA extraction

Between 53 and 63 *P. falciparum* parasite positive samples were selected in each year from 1995 to 2014, giving a total of 1164 samples (Table 4.1). Sample selection was based on availability of temporal and/or spatial data and parasitaemia levels of ≥ 5,000 parasites/µl. Total DNA was extracted from the samples using the QIAamp 96 DNA QIAcube HT kit (Qiagen, Manchester, UK), as per the manufacturer's protocol.

Table 4-1: *P. falciparum* positive samples collected from children admitted at Kilifi County Hospital between 1995 and 2014.

		AMA1		SURFIN$_{4.2}$	
Year	**Number sampled**	**Number sequenced**	**Number analysed***	**Number sequenced**	**Number analysed***
1995	60	59	59	44	42
1996	58	55	54	48	47
1997	60	56	52	46	45
1998	58	55	52	53	49
1999	56	52	49	48	45
2000	56	49	44	49	47
2001	63	53	36	52	52
2002	58	53	43	50	50
2003	60	56	52	53	53
2004	63	54	52	55	53
2005	63	55	52	44	43
2006	59	48	47	45	42
2007	53	44	41	42	41
2008	58	51	48	44	39
2009	54	47	47	43	34
2010	56	51	48	49	46
2011	57	54	54	49	49
2012	57	43	41	38	36
2013	56	50	48	51	46
2014	59	59	59	58	38
Total	**1164**	**1044**	**978**	**961**	**897**

*For both genes, the final number of samples analysed varies from the original number of samples collected because some samples failed the PCR amplification (*AMA1*=120; *Surf*$_{4.2}$ = 203), while others produced sequence data with high background noise (*AMA1* = 66; *Surf*$_{4.2}$ = 64) and were excluded from the final analyses**.**

4.2.3 Primer Design

Gene specific polymerase chain reaction (PCR) and sequencing primers for *surf*$_{4.2}$ were designed using 3D7 *surf*$_{4.2}$ gene (PF3D7_0424400) as the reference sequence. A 2660 base pair region encompassing the 2292 base pair long exon 1 was downloaded from PlasmoDB (http://plasmodb.org/plasmo/) and imported into Editseq application (DNASTAR Lasergene version 7) for primer design. PCR primers were designed in the intronic regions around exon 1 of this gene (Table 4.2; figure 4.3). *AMA1* primers used in this study were derived from published literature (Polley and Conway, 2001, Ochola-Oyier et al., 2016), and covered an 1824 base pair region of the 1869 base pair long, single exon gene (Table 4.2; figure 4.4).

Table 4-2: Primers used in PCR amplification and sequencing of *surf*$_{4.2}$ and *AMA1* genes in *P. falciparum* isolates from Kilifi, Kenya.

Gene ID	Product name	Primer name	Primer sequence
PF3D7_0424400	**SURFIN$_{4.2}$**	**F1**	**TAGTAGCTGTAAATGTGGTTAGTC**
		F3B	ATCTGATCAAGGTTCTCAGA
		R1	**CCCTTGAACAATTGTTGTACCAA**
		R2	TTC AAC AAA ATG TGG CAC ATT C
		R5	ACTGGTAGAGACGACATCA
PF3D7_1133400	**AMA1**	**FI**	**GAG CGC CTT TGA GTT TAC**
		F143	GAC TTC CAT CAG GGA AAT GTC C
		F344	TTG AGT GCT TCG GAT CAA CCT AA
		R1	**TCC ATC AGA ACT GGT GTT G**
		R2	CTT TTG ATC ATA CTA GCG TTC T

FI and R1 primers were used in the initial PCR amplification of each of the genes. All primers were used in the sequencing PCR reactions for individual genes.

```
ggtaaataaatacctatatacatatatatattttttattaattcatatagtagctgtaaatgtggttagtcttcatttttatattttcccatttttgataatATGCTTTT
TGTTGTTGAGCTCGACAGCAGATTGGAAAAATCTGCAGATAAAAGAATAAGTGTTGAAAGA
TTTAGGAAAATATTTGAAATTTATGTTGAAGATAAACTTGAAGAATTAAAAAGGTCAGGAT
CTGAAAAGTATGATAAGGATTGTCGAGATTTCAATTATTTTATCGATGATGTAAAAGATGTA
TTTATAAATAATGATTTGGTAAAAATCCCTGTTAAAGTTCGTAAGAGTATTTGGGAAACGCA
TGTTGACAAAAACTTACCGAAACTTATGAAAAACACTACTAGTTGTAAATGTATTAGAAAA
GAACATAATTATAATAAAGAATATAGAGATATGACTCGAACGTTAGAAGATTTTTGTGAAG
AAAAAACACGCAAGCTAGAAATTATATATCAAAAGGATTACGACGAAAGTTTATATGTGAA
TTTTAATGAATGGATAAATAAAAAAAAAGAAGACATTTTAAAGGAATATGAAAAATTAAGT
AATAAAGATAAATATAAACACTTATTAAAAATTAGTGAAACCTGTGATCTTAATCATGTGG
ATAAATTATTCCTTAATATATCAAGTGAGGATATGAAAAAAATGAAGGAAGATGTTAAAAA
ACAACATCTTGAAGTTAAAGTAAGACCACCAGTAAATGAACCGATTGACGATAGAGGTAAA
GAGGATGCTCTAAGTAGAGGGGGTAAAAGCATTATTTTAAATAAGGAGGAAAATTCACCTA
CTGAAGAAATCACGACTGAATATAATCCTGTAAGTGAAATGGGCGTTGGAACTACTATAGC
TCATAGTGAACCTGGGCCCAAAACGGTTAATACTGAAGTTAGAAATGTTCTAAGATCTGAT
GGAAAAATATCTGATCAAGGTTCTCAGAAAAGTCCTCCTAAGGAACTTTCTAATAAACAAA
TGACTCCTGCTCAACGTAAGAATGTGCCACATTTTGTTGAAAGAAGAGGCTATGGAAATAG
TCATGTTAGGGGTAACGCACTTAAAAAAATTAGTAATGGTGATGATAATTATAAAAGTCCTT
CTTCTAATTATATTGAAGTTGATTGTGCTGAAGATAAATATTTTCTATTAGAAGATGGAACT
AATCAATCAGAAAATAGTTGTAAAACAAAATATAACTATTTTGTATCTAATGATTATGATGG
TACTGGGTCTGCTATTTATTCGACAGATCAGGTACCTAGTAGGGAAGAAATTAAAAGCCCTG
ATTCTTTATCTACATTAGATGCCAGAGGAAGTACACATAATTTAAATGTTTCTAATGAGGGA
AATCCTCTTGAGGGTGGGGAGGAAAAAAATAATGTTAAGATAAGTGAACAAAATGGTCGTA
ATCTGGAAAGTTCTGTAGGAACAGATAAGGGTTCAGATAAAAATGAAGAAGAAGTAGCTGC
TACTTGTGATCCAAACGATCGAAACTGTTTTGATGGAAGATATATAAATGTTGATTATATAA
GAGGTTTATTAAAGGGTAAACGTTCTGGTTCAGATGGAAGATCTAATATAAAACACATAAT
TTCTAATAACTTTGATGATAGTAATATGATTTTTCCTACTTTTAATTTTGATTTCGATATTTTA
AAAGCTGAAGAAGAAGTTTTACCTGTGAATAATTCTGATATAATATATGGACATGAAGAAG
TGGAAGAAACTACTCAAGAAGGTGCATCAATATTTGAAAAACATAGTCATACTTCTTCTCAA
CAAAATGATTCATTGCTTCAGGAAATAAATATAGAATGTTATCTACAACTATGGAATTACCT
AATCAACAAGAGGTATTTGGACTATATTCACCTGTATCACGAACGCTTGATAGTGCTATGAG
TTTTTTGCGAAGTATTATTAGTTTGAGTAGTGCTCCAGTTTCACGAAGTGAAGGTCAAAGCA
AAGAAAGCAAACGTGTAGAAATATCAACTACGGTTCAGGATCCTATTGGATATAGGACTTC
CCCTTTACAAATGAATGCTCATAGCGTTGGTGCTGGTATTAATGTATCCTCAATATTATCAAT
GTTAGGTTTGTCTAGTGGACAAGTTCGAAGAAGTGGTGGGCAAGGAAGTGAAACATATATA
GTTGGTACGTCTCAAAGTGGTTTCCATAAAAATGAAGTAATTCCCTCCATAAAAGATAAAA
GTGGTAAAACTCAAATCGTAAGTAATGAAAAAGGAGGGATTTTTTCAAAAGGGATAACATC
AATGATGTCGTCTCTACCAGTTGCATTAGTAACATTTGTATTTCTTTTTATGTTTTTGGTATTT
AATAAGgtaataatatatgataatattatattttgattaaataatatatttatattatgtaataatttttttttttttttttattattaatttaaatataataaattatataa
tgacataatatttaatttaattcattttgtttttatttcattattttgtagatgaatcctttttggtacaacaattgttcaagggaagaaaaaaaaaaaagt
```

Figure 4.3: *P. falciparum surf$_{4.2}$* gene sequence showing the binding sites of the PCR and sequencing primers used.

The sequence is 2660 bases long and includes the 2292 base pair long exon 1 (uppercase bases) and surrounding intronic regions (lowercase bases). Each primer sequence has been colour-coded as in Table 4.2.

```
ATGAGAAAATTATACTGCGTATTATTATTGAGCGCCTTTGAGTTTACATATATGATAAACTT
TGGAAGAGGACAGAATTATTGGGAACATCCATATCAAAATAGTGATGTGTATCGTCCAATC
AACGAACATAGGGAACATCCAAAAGAATACGAATATCCATTACACCAGGAACATACATAC
CAACAAGAAGATTCAGGAGAAGACGAAAATACATTACAACACGCATATCCAATAGACCAC
GAAGGTGCCGAACCCGCACCACAAGAACAAAATTTATTTTCAAGCATTGAAATAGTAGAA
AGAAGTAATTATATGGGTAATCCATGGACGGAATATATGGCAAAATATGATATTGAAGAA
GTTCATGGTTCAGGTATAAGAGTAGATTTAGGAGAAGATGCTGAAGTAGCTGGAACTCAA
TATAGACTTCCATCAGGGAAATGTCCAGTATTTGGTAAAGGTATAATTATTGAGAATTCAA
ATACTACTTTTTTAACACCGGTAGCTACGGGAAATCAATATTTAAAAGATGGAGGTTTTGC
TTTTCCTCCAACAGAACCTCTTATGTCACCAATGACATTAGATGAAATGAGACATTTTTATA
AAGATAATAAATATGTAAAAAATTTAGATGAATTGACTTTATGTTCAAGACATGCAGGAA
ATATGATTCCAGATAATGATAAAAATTCAAATTATAAATATCCAGCTGTTTATGATGACAA
AGATAAAAAGTGTCATATATTATATATTGCAGCTCAAGAAAATAATGGTCCTAGATATTGT
AATAAAGACGAAAGTAAAAGAAACAGCATGTTTTGTTTTAGACCAGCAAAAGATATATCA
TTTCAAAACTATACATATTTAAGTAAGAATGTAGTTGATAACTGGGAAAAAGTTTGCCCTA
GAAAGAATTTACAGAATGCAAAATTCGGATTATGGGTCGATGGAAATTGTGAAGATATAC
CACATGTAAATGAATTTCCAGCAATTGATCTTTTTGAATGTAATAAATTAGTTTTTGAATTG
AGTGCTTCGGATCAACCTAAACAATATGAACAACATTTAACAGATTATGAAAAAATTAAA
GAAGGTTTCAAAAATAAGAACGCTAGTATGATCAAAAGTGCTTTTCTTCCCACTGGTGCTT
TTAAAGCAGATAGATATAAAAGTCATGGTAAGGGTTATAATTGGGGAAATTATAACACAG
AAACACAAAAATGTGAAATTTTTAATGTCAAACCAACATGTTTAATTAACAATTCATCATA
CATTGCTACTACTGCTTTGTCCCATCCCATCGAAGTTGAAAACAATTTTCCATGTTCATTAT
ATAAAGATGAAATAATGAAAGAAATCGAAAGAGAATCAAAACGAATTAAATTAAATGAT
AATGATGATGAAGGGAATAAAAAAATTATAGCTCCAAGAATTTTTATTTCAGATGATAAA
GACAGTTTAAAATGCCCATGTGACCCTGAAATGGTAAGTAATAGTACATGTCGTTTCTTTG
TATGTAAATGTGTAGAAAGAAGGGCAGAAGTAACATCAAATAATGAAGTTGTAGTTAAAG
AAGAATATAAAGATGAATATGCAGATATTCCTGAACATAAACCAACTTATGATAAAATGA
AAATTATAATTGCATCATCAGCTGCTGTCGCTGTATTAGCAACTATTTTAATGGTTTATCTT
TATAAAAGAAAAGGAAATGCTGAAAAATATGATAAAATGGATGAACCACAAGATTATGGG
AAATCAAATTCAAGAAATGATGAAATGTTAGATCCTGAGGCATCTTTTTGGGGGGAAGAA
AAAAGAGCATCACATACAACACCAGTTCTGATGGAAAAACCATACTATTAA
```

Figure 4.4: *P. falciparum AMA1* gene sequence showing the binding sites of PCR and sequencing primers.

The sequence is 1869 nucleotides long and represents the entire single exon gene. Each primer sequence has been colour-coded as in Table 4.2. These primers were derived from published literature (Ochola-Oyier et al., 2016, Polley and Conway, 2001).

4.2.4 Polymerase Chain Reaction optimization and gene amplification.

Gradient PCR reactions for *PfAMA1* and *surf*$_{4.2}$ were run on the verity 96-well thermocycler (Applied Biosystems), using a range of annealing temperatures (Tables 4.3 and 4.4), to identify optimal temperature conditions needed for the amplification of the two genes. *PfAMA1* optimization reactions involved an initial denaturation at 94^{o}C for 2 minutes, 10 cycles of denaturation at 94^{o}C for 15 seconds, annealing (using gradient temperatures of 55^{o}C – 57.7^{o}C) for 30 seconds and extension at 72^{o}C for 2 minutes. This was then followed by 25 cycles of denaturation at 94^{o}C for 15 seconds, annealing using the gradient temperatures (55^{o}C – 57.7^{o}C) for 30 seconds and extension at 72^{o}C for 2 minutes with a 5 second increment on the extension time per successive cycle. A final extension reaction was then carried out at 72^{o}C for 7 minutes (Table 4.3).

Table 4-3: PCR amplification conditions for *P. falciparum AMA1* gene

Step	Temperature (^{0}C)	Time (min: sec)	Cycles
Initial denaturation	94	2:00	1
Denaturation	94	0:15	10
Annealing	55 – 57.7*	0:30	
Extension	72	2:00	
Denaturation	94	0:15	25
Annealing	55 – 57.7*	0:30	
Extension	72	2:00 + 0:05 cycle extension for each successive cycle.	
Final extension	72	7:00	1

*6 different temperatures (55^{0}C, 55.7^{0}C, 56^{0}C, 56.7^{0}C, 57^{0}C, 57.7^{0}C) within the indicated range were tested to identify the optimal annealing temperature for *PfAMA1*.

Optimization of amplification conditions for *surf$_{4.2}$* involved initial denaturation at 94^{o}C for 2 minutes, followed by 10 cycles of denaturation at 94^{o}C for 15 seconds, annealing (using a temperature range of 53^{o}C - 58^{o}C) for 30 seconds, and extension at 68^{o}C for 3 minutes. This was followed by 25 cycles of denaturation at 94^{o}C for 15 seconds, annealing at 53^{o}C - 58^{o}C for 30 seconds, extension at 68^{o}C for 3 minutes, with a 5 second increment on the extension time per successive cycle. A final extension was carried out at 68^{o}C for 7 minutes (Table 4.4.).

Table 4-4: PCR amplification conditions for *P. falciparum surf$_{4.2}$* gene.

Step	Temperature (^{0}C)	Time (min: sec)	Cycles
Initial denaturation	94	2:00	1
Denaturation	94	0:15	10
Annealing	53 – 58*	0:30	
Extension	68	3:00	
Denaturation	94	0:15	25
Annealing	53 – 58*	0:30	
Extension	68	3:00 + 0:05 cycle extension for each successive cycle.	
Final extension	68	7:00	1

*6 different temperatures (53^{0}C, 54^{0}C, 55^{0}C, 56^{0}C, 57^{0}C, 58^{0}C) within the indicated range were tested to identify the optimal annealing temperature for *surf$_{4.2}$*.

Following optimization reactions, a 10µl reaction was set up in two micro-centrifuge tubes as follows: 1µl of 1X PCR buffer 4 (Roche, Basel Switzerland), 0.2µl of 250µM dNTPs (Bioline, Ohio, USA), 0.3µl each of 1µM forward and reverse primers, 0.5µl – 1.5µl of 50ng -500ng of DNA template and DNase free water (Invitrogen, California, USA) were added to the first tube, giving a final volume of 5µl. In the second tube, 0.15µl (0.525 units) of Expand High Fidelity Taq Polymerase (Roche), 1µl of 1X PCR buffer 2 (Roche) and 3.85µl of DNase free water were added to give a final volume of 5µl. The contents of the two micro-centrifuge tubes were then combined to make the 10µl reaction mix. For the actual PCR reactions, the above optimization conditions were maintained, but the annealing temperatures were set at 57^{o}C for *PfAMA1* and 55^{o}C for *surf$_{4.2}$*, as these were found to result in optimal gene amplification.

4.2.5 Gel electrophoretic analysis of PCR products

PCR products were run on electrophoretic gels to identify samples that were successfully amplified. A 1% w/v agarose gel was prepared by weighing and dissolving 1g of agarose powder (AGTC Bioproducts, UK) in 100ml of 1X Tris Borate EDTA (TBE) buffer (Life Technologies, California, USA). The mixture was heated to boiling point in a microwave oven to dissolve the agarose powder in the buffer. The solution was left to cool before adding 5µl of Red Safe gel stain (iNtRON Biotechnology, Korea) and mixing. The mixture was poured into gel trays with combs and left at room temperature for 45 minutes to allow the gels to set.

Once set, the gels were transferred into electrophoresis tanks containing 1X TBE buffer. 1.5µl of each PCR product was separately mixed with an equal volume of 6X Blue Orange loading dye (Promega Corporation, Wisconsin, USA) then loaded into wells in the gel. 2µl of hyperladder I DNA standards (Bioline) was loaded into an extra well on the gel. Positive (3D7) and negative (water) controls were also mixed with the loading dye and loaded into separate wells on the gel. The electrophoresis reaction was then run at 100 volts for 40 minutes (for *PfAMA1*) and 60 minutes (for *surf*$_{4.2}$) to separate out the DNA fragments. After electrophoresis, gels were viewed and photos taken on a molecular imager Gel Doc (Bio-Rad, California, USA).

4.2.6 PCR product purification

Successfully amplified PCR products were purified using EXOSAP-IT (Affymetrix) reagent, based on an enzymatic clean up protocol. The EXOSAP-IT reagent contains 2 enzymes, Exonuclease 1 and Shrimp Alkaline Phosphatase (SAP), which remove unincorporated dNTPs and primers that are likely to interfere with downstream processes such as sequencing. The Exonuclease breaks down the remaining single stranded primers into dNTPs and SAP removes the phosphate group from the dNTPs. Briefly, 3.4µl of the EXOSAP-IT reagent was mixed directly with 8.5µl of successfully amplified PCR product and the mixture incubated in a thermocycler for 15 minutes at 37^{o}C to degrade any remaining primers and nucleotides and then at 80^{0}C for 15 minutes to denature the enzymes. A final incubation at 15^{0}C for 5 minutes was carried out to allow the plates to cool.

4.2.7 BigDye sequencing reaction

BigDye sequencing reactions were set up separately for each primer as follows: 0.5µl of BigDye terminator ready reaction mix (Applied Biosystems, California, USA), 2.0µl of 5X sequencing buffer, 1.0µl of primer, 5.5µl of DNase free water and 1.0µl of purified PCR product were mixed to give a final volume of 10µl. The plates were then loaded into the thermocycler and the products amplified using the following program: 25 cycles of denaturation at 96 ^{0}C for 30 seconds, annealing at 50 ^{0}C for 15 seconds and extension at 60 ^{0}C for 4 minutes.

4.2.8 Purification of sequenced products

The post-sequencing PCR products were purified to remove unincorporated dNTPs and unused primers using ethanol/sodium acetate precipitation. Briefly, a premix solution containing 3µl of sodium acetate, pH 5.2, 62.5µl of 95% ethanol and 24.5µl of distilled water was constituted to make a final volume of 90µl per reaction well. 90µl of the mix was added to each well containing sequenced products. The plates were then sealed with micro-seals (Bio-Rad) and incubated at -20^{o}C for 60 minutes to allow the DNA to precipitate out of solution. After incubation, the plates were spun at 4000 revolutions per minute (rpm) for 30 minutes in the 5810R benchtop centrifuge (Eppendorf) to separate the precipitated DNA from the rest of the solution. The seal covers were removed from the plates, the plates overlaid with clean paper towels and gently inverted to drain the solution mix. The inverted plates were then spun at 200rpm for 1 minute to further remove any remaining solution from the plates without discarding the precipitate. 150µl of ice cold (-20^{0}C) 70% ethanol was then added into each well, the plates sealed and spun at 4000rpm for 10 minutes for a second round of cleaning to remove any remaining salts which could interfere with downstream processes. After spinning, the plates were once again inverted on top of paper towels, and excess ethanol solution drained. The plates were then overlaid with a clean set of paper towels, inverted and gently spun at 200rpm for 1 minute. All centrifugation steps were carried out at 4^{0}C. Finally, the plates were covered with fresh paper towels and left on the bench to air dry, to allow any remaining traces of ethanol to evaporate as these can interfere with subsequent capillary electrophoresis steps, resulting in high background noise.

4.2.9 Capillary electrophoresis of sequenced PCR products

After air drying, 10µl of Formamide HiDi (Applied Biosystems) was added into each plate well, the plates sealed and heated for 3 minutes at 96^{o}C in the thermocycler. The plates were sent to International Livestock Research Institute (ILRI)'s sequencing lab for capillary reading of the sequences and generation of sequence chromatograms.

4.2.10 Data analysis

4.2.10.1 Sequence assembly and confirmation of base calls

Sequence chromatograms (trace files) were imported into seqman application (DNASTAR, Lasergene Version 7) for confirmation of base calling. For each gene and each sample, sequences generated from the 5 different primer extensions were aligned into contigs. 3D7 reference sequences (PF3D7_1133400 and PF3D7_0424400) were used to scaffold the trace data generated from each primer for each gene. Trace files were assessed for the quality of peaks and accuracy of base calling. Corrections to base calling were done on the basis of the peaks of the electropherogram and independently of the reference sequence. Where there was a double peak (peak within a peak) at the same position along the sequence, the major peak was taken to represent the major allele at that position. Sequence chromatograms with high background noise, or in which the two overlapping sequences had a mismatch at a specific location, were excluded. Clean sequences were saved as consensus files after removing the reference sequence.

4.2.10.2 Multiple sequence alignments and identification of segregating sites

To identify polymorphic positions in the gene sequences, a multiple alignment of the consensus files for all samples in each gene was carried out using muscle algorithm implemented in MUSCLE software (Edgar, 2004). The respective 3D7 gene sequences were used as references. The resultant multiple alignment file for each gene was saved in msf file format and imported into Jalview software (Waterhouse *et al*., 2009) where the alignment was confirmed by eye and misaligned sequences manually corrected. Each multiple alignment was then trimmed to get sequences of equal length across all samples in the alignment, with a view of retaining the highest number of samples that provided the longest length of sequence across each gene. Sequences which were too short were removed from the alignment at this point. The final alignments, which consisted of 978 *AMA1* sequences and 897 *surf*$_{4.2}$ sequences were saved in fasta format, and imported into MEGA7 (Kumar *et al*., 2016) and DnaSP v.5.10.01 (Rozas, 2009) software for subsequent analyses.

4.2.10.3 Heatmaps of genetic diversity among *PfAMA1* and *surf$_{4.2}$* sequences.

Temporal and spatial patterns of genetic diversity in *PfAMA1* and *surf$_{4.2}$* genes were visualised by generating heatmaps using the Heatmap function in R's ComplexHeatmap package. The analysis was carried out on the informative (polymorphic) sites only. Briefly, for each gene, a multiple alignment file containing nucleotides at the polymorphic positions was imported into R. For each sample in the alignment, each position was coded as 0 if the base was identical to the 3D7 reference, and 1 if it differed from the reference allele. The resulting 1/0 data matrix was then used to generate heatmaps showing the clustering of sequences based on genetic similarity, with dendrograms illustrating the hierarchical clustering patterns among the sequences. The sequence clustering algorithm works by first converting the 1/0 data matrix into a distance matrix computed using a binary distance measure to calculate the dissimilarity distances between the sequences in the data matrix. The distance matrix is subsequently used in a hierarchical cluster analysis which works by initially assigning each object (sequence) to its own cluster and then sequentially joining the two most similar clusters until there is only one cluster present. A dendrogram representing the sequence clustering pattern was generated. The hierarchical cluster analysis was bootstrapped, based on 1000 resampling steps, using the pvclust() function in the pvclust package to test the stability of the observed clustering pattern represented by the dendrogram.

To discern the patterns of temporal distribution of genetic diversity in the genes, the sequences were clustered based on the year of sample collection. To discern the spatial pattern of genetic diversity, the sequences were clustered based on two categories (north versus south), depending on whether the samples were collected from children living north or south of the naturally occurring Kilifi creek. 978 and 795 sequences were used to generate heatmaps showing the temporal and spatial genetic clustering of *PfAMA1*, respectively. For *surf*$_{4.2}$, 897 and 754 sequences were used to determine the temporal and spatial clustering of sequences, respectively. The disparity in the number of samples used in the temporal and spatial analyses was because some of the samples lacked records on their locations of origin, either because this information was not captured during admission, or because the child lived outside of the KHDSS catchment area. Additional heatmaps were generated using the variable amino acids, to determine whether clustering of genetic diversity showed different patterns when analysing amino acid instead of nucleotide sequence data.

4.2.10.4 Statistical tests of neutrality

Statistical tests of neutrality are used in population genetics to determine whether the observed variations are evolving randomly or under selective pressure. Under neutral evolution, most genetic variations do not affect fitness, thus the fate of such mutations is based on a stochastic process and can either increase to fixation or be lost over time in the population (Duret, 2008).

The Tajima's D statistic is used to measure whether mutations are under selection. This test uses different parameters to measure allele frequency distribution and determine whether observed mutations are selectively neutral. Tajima's D considers the difference between the number of segregating (or polymorphic) sites (*S*) and the average number of pairwise nucleotide differences between sequences (K) (Tajima, 1989). Negative values derived from the test statistic signifies an excess of low frequency or rare polymorphisms, indicating either a population expansion, selective sweep and/or purifying selection (Akey *et al*., 2004). Positive values, on the other hand, signify an excess of intermediate-frequency alleles (alleles present at 5% - 20% frequency), indicating either a decrease in population size (population bottlenecks) or balancing selection, while 0 values signify a population that is evolving randomly (Akey *et al*., 2004).

To carry out this analysis, multiple sequence alignment files were imported into DnaSP5.10.01 software. For *PfAMA1*, the tests were performed separately on:

1) the full region sequenced (nucleotides 70 - 1800),

2) the entire ectodomain (nucleotides 421 - 1641) and

3) each of the individual domains: I (nucleotides 445 - 906), II (nucleotides 958 - 1254) and III (nucleotides 1327 - 1527).

For *surf$_{4.2}$*, the tests were performed on:

1) the entire sequenced region (nucleotides 1 – 2199),

2) the cysteine-rich domain (nucleotides 153 – 585) and

3) the variable region (nucleotides 588 -2199).

A stepwise analysis of Tajima's D was also carried out on the 1731bp and 2199bp regions of *PfAMA1* and *surf$_{4.2}$* genes using a sliding window approach, with a window size of 100 nucleotides and a step size of 25 nucleotides. Additionally, the tests were computed separately for samples collected in individual years, using the 1731bp and 2199bp fragments of *PfAMA1* and *surf$_{4.2}$*, respectively.

4.2.10.5 Linkage disequilibrium and recombination analysis

Linkage disequilibrium (LD), which refers to the non-random association between SNPs at different loci, was measured using the D' (normalised coefficient of linkage disequilibrium) and R^2 (square of the correlation coefficient of allelic states at each loci pair) indices (Mueller, 2004), to test the strength of association between nucleotides at different segregating sites throughout the *PfAMA1* and *surf$_{4.2}$* sequences in the parasite population. For *surf*$_{4.2}$, LD was computed in a pairwise analysis at 323 diallelic sites in 897 samples, after excluding sites with gaps and those that were segregating for three or more nucleotides. For *PfAMA1*, LD was computed in a pairwise analysis for 88 diallelic sites in 978 sequences. For both genes, the significance of each pairwise association was tested using Fisher's exact test and the relationship between LD and nucleotide distance plotted.

The level of recombination was analysed across 897 *surf*$_{4.2}$ sequences and 978 *PfAMA1* sequences to determine the minimum number of recombination events that have occurred in the entire sequence (Hudson and Kaplan, 1985). The recombination parameter, *C*, was also computed between adjacent sites as well as in the entire sequence. *C* is represented by 4Nc, where N represents the effective population size and c represents the probability of recombination between adjacent nucleotides per generation (Hudson, 1987).

4.2.10.6 Identification of circulating ***Pf*****AMA1 and *surf*$_{4.2}$ haplotypes**

PfAMA1 and *surf*$_{4.2}$ haplotypes (combination of SNPs across the gene that are inherited together) circulating in the population during the study period were identified using DnaSP software. Additionally, haplotype diversity, which signifies the uniqueness of a specific haplotype in the population, was also computed.

4.2.10.7 Temporal and spatial genetic variation in ***P. falciparum AMA1*** and ***surf*$_{4.2}$** sequences

To analyse the effects of time and space on *P. falciparum* genetic diversity, two different but complementary measures of genetic diversity were computed: 1) number of SNP differences and 2) identity by state (IBS), which represents identical DNA segments or sequences at a particular locus in two or more parasites. Whereas the number of SNP differences is a proxy for the level of genetic diversity, IBS is a proxy for the level of genetic similarity between sequence pairs. For each gene, each parasite was compared to every other parasite in the dataset, noting the time (day), distance (km), SNP differences and the longest contiguous sequence length shared by the two parasites (IBS).

The number of SNP differences were computed as previously indicated in chapters 2 and 3, i.e. for each gene, pairs of sequences were compared and the number of SNPs at observed polymorphic sites between them counted. Half the lower limit of detection of temporal and spatial differences were taken for parasites collected on the same day and/or at the same location. Parasites collected on the same day were assigned a difference of 0.5 days and those collected from the same location were assigned a difference of 2.5km since location was known to a 5km accuracy.

The relationship between genetic variation and time, distance and the interaction of time and distance was analysed in a multiple fractional polynomial regression model, with the number of SNP differences between parasite pairs as the outcome variable and either time, distance or time-distance interaction between parasite pairs as the independent variables. For *PfAMA1*, the variation in the number of SNP differences was computed over distance for 795 samples that had associated geospatial positioning data, over time for 966 samples that had associated temporal information (i.e. a record of the specific date of admission was available) and over time and space for 795 samples that had both temporal and spatial data. For *surf*$_{4.2}$, variation in the number of SNP differences was computed over distance for 754 samples that had geospatial positioning data, over time for 890 samples that had temporal data and over time and space for 754 samples that had both temporal and spatial data. All the analyses were bootstrapped in linear regression analyses using 1000 resampling steps to determine statistical significance and confidence intervals of the observed results.

Analysis of identity by state involved identifying, for each parasite pair, the longest stretch of contiguous sequence that was shared between the pair. To ensure that the longest stretch of contiguous sequence shared between parasites that was identified was not simply a conserved region across the gene, only polymorphic sites were considered. To do this, each parasite pair was first compared across the polymorphic sites and each nucleotide position coded as 1 if the two parasites had the same base, and 0 if the bases differed. The number of nucleotides in the longest stretch of sequence with identical bases between the two samples was noted and taken to represent the IBS for that parasite pair. The IBS values were then included as outcome variables in a multiple fractional polynomial regression analysis to examine the spatio-temporal relationship between parasite genotypes, using logarithmic transformations of time, distance and time-distance interaction as independent variables. The analyses were bootstrapped in linear regression analyses using 1000 resampling steps to determine statistical significance and confidence intervals of the observed results. For both genes, the IBS analyses were carried out using the same number of samples as those used in the analyses of SNP differences over time, distance and the interaction of time and distance.

For both IBS and SNP differences data, the analyses were carried out on the entire dataset. To determine the rate of change in parasite genotype in more detail, similar analyses were also carried out on subsets of data collected within specific time frames:

1) Samples collected within three years of each other.
2) Samples collected between 4 to 10 years of each other.
3) Samples collected more than 10 years apart.

In each of the above datasets, the analysis was bootstrapped in a linear regression analysis to determine the confidence intervals and statistical significance of the observed results. Finally, a scatterplot of IBS versus SNP differences between sequences was produced to visually represent the relationship between the two parameters.

4.3 RESULTS

4.3.1 *P. falciparum AMA1* and *surf*$_{4.2}$ sequence diversity

A total of 1164 *P. falciparum* samples were collected from children who were admitted to Kilifi County Hospital with mild or severe malaria between 1995 and 2014. *PfAMA1* and *surf*$_{4.2}$ genes were successfully amplified (figures 4.5 and 4.6) and sequenced in 1044 and 961 of these samples, respectively. Sequences containing high background noise and those where one or more primers were unsuccessfully sequenced were excluded from further analyses, and the remaining sequences were aligned to identify polymorphic sites.

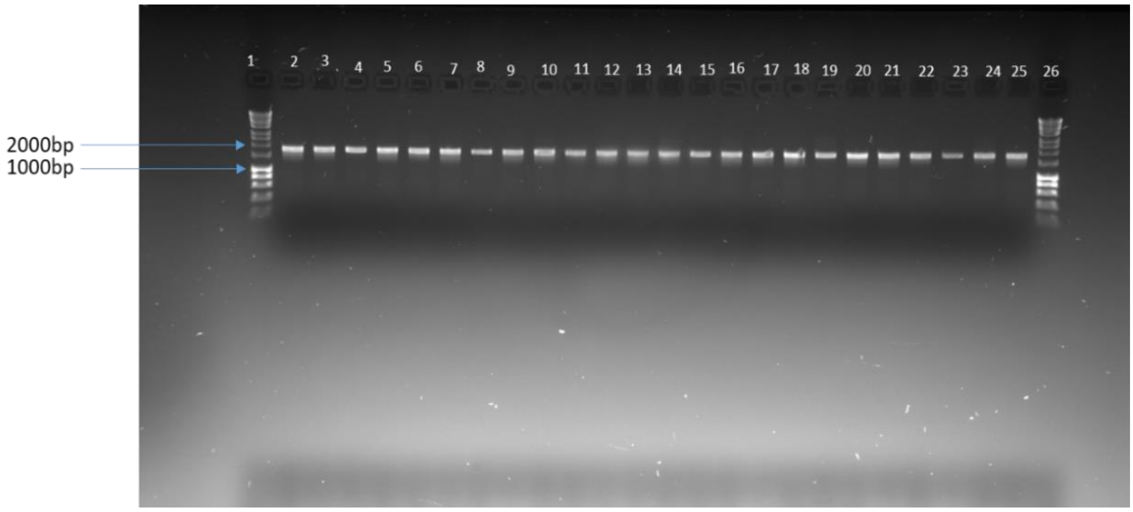


Figure 4.5: Electrophoresis gel photo showing PCR amplification results of *PfAMA1* in 24 samples.

Wells marked 1 and 26 contain Hyperladder I DNA standards. Bands representing DNA fragments of size 1000bp and 2000bp are indicated with arrows.

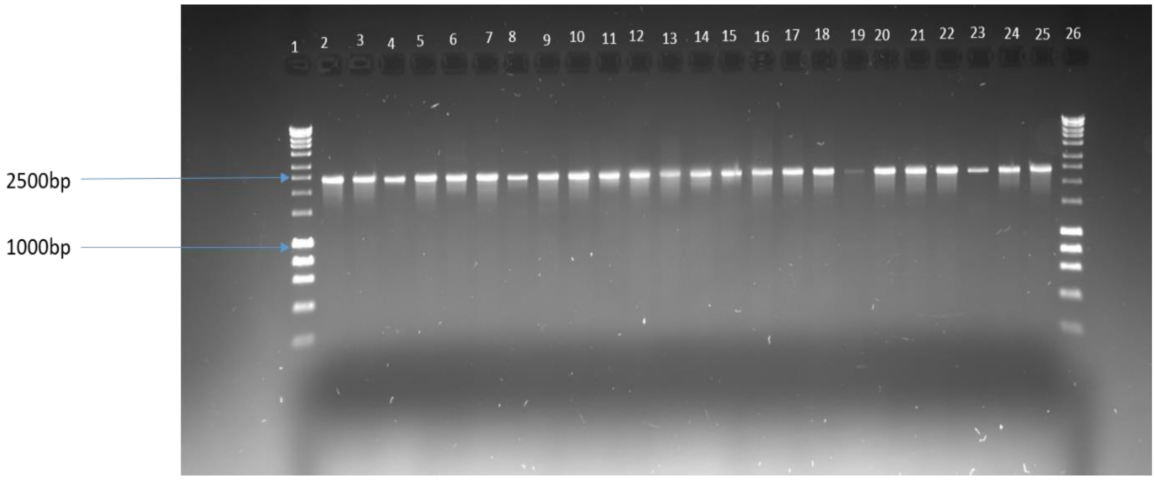


Figure 4.6: Electrophoresis gel photo showing PCR amplification results of *surf*$_{4.2}$ in 24 samples.

Wells marked 1 and 26 contain Hyperladder I DNA standards. Bands representing DNA fragments of size 1000bp and 2500bp are indicated with arrows.

A multiple alignment of each gene was trimmed to get equal sequence lengths across all samples in the alignment, giving a final list of 978 *PfAMA1* sequences that were 1731 nucleotides long (bases 70 – 1800 with reference to the full length *PfAMA1* gene), corresponding to codons 24 – 600, and 897 *surf*$_{4.2}$ sequences that were 2199 nucleotides long (bases 1 – 2199 of exon 1), corresponding to codons 1 – 733.

A total of 156 mutations were identified across 137 polymorphic sites in the 1731 nucleotide region sequenced in *PfAMA1*. 92 of these polymorphic sites were found in the ectodomain region (nucleotide range 421 - 1641). The polymorphic sites included variable numbers of singleton (mutation present in only 1 sample), di-allelic (two alleles) and tri-allelic (three alleles) mutations (Table 4.5). One site was identified which contained a mutation that was fixed in the Kilifi parasite population. In *surf*$_{4.2}$, a total of 487 mutations were identified across 442 polymorphic sites in the 2199 nucleotide region analysed, and included singletons, di-allelic, tri-allelic and tetra-allelic (four alleles) mutations (Table 4.5). Additionally, a 3-base pair deletion was identified at positions 1051 – 1053 in 6 of the samples, and another 3-base pair deletion was identified at positions 1408 – 1410 in 27 of the samples.

Table 4-5: Characteristics of the polymorphic sites identified in *P. falciparum AMA1* and *surf*$_{4.2}$ sequences in a coastal Kenya population.

Mutation type	***Pf*AMA1**	***Surf*$_{4.2}$**
Singletons*	30	76
Diallelic	88	323
Tri-allelic	19	41
Tetra-allelic	0	2
Invariable sites	1594	1751
INDELS	0	2**

*Singletons are polymorphisms that are found in only 1 sample, and may be considered as a sub-category of diallelic SNPs.

**each gap consisted of a 3 base-pair deletion.

Average pairwise nucleotide diversity per site (π) was 0.01487 for *PfAMA1* and 0.04513 for *surf*$_{4.2}$. Among the *PfAMA1* sequences, most of the polymorphic sites were present in domain I (54 SNPs), while domains II and III had 14 and 13 SNPs, respectively. Nucleotide diversity within individual domains showed that domain I had the highest diversity at 0.0276, while domains II and III had comparable diversities at 0.0095 and 0.0166 nucleotide differences per site, respectively. In *surf*$_{4.2}$, the highest diversity was recorded within the variable region, with 375 of the 442 polymorphic sites found here, and an average of 0.0578 nucleotide differences per site (Table 4.6) across this region. To show how genetic diversity varied across different regions of the sequences, nucleotide diversity was computed across the entire sequences of both genes based on a sliding window approach, using a window size of 100 nucleotides and a 25-base pair step size. The sliding window plots showed that most of the diversity was in domain I of *PfAMA1* (figure 4.7) and the variable (VAR) region of *surf*$_{4.2}$ (figure 4.8).

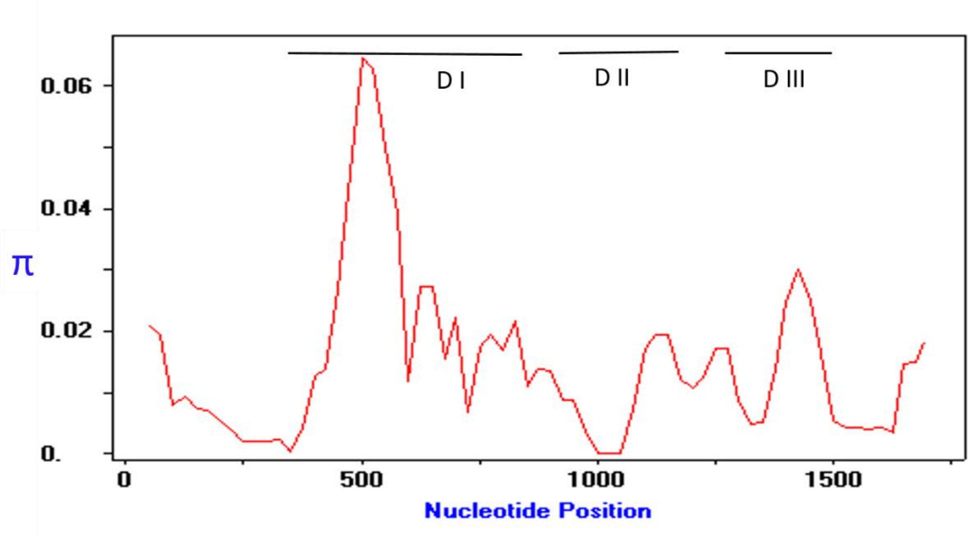


Figure 4.7: Sliding window plot showing the average pairwise nucleotide diversity (π) across a 1731 nucleotide region of *P. falciparum AMA1* in a coastal Kenya parasite population.

A window size of 100 nucleotides with a step size of 25 nucleotides was used to generate the plot. The nucleotide regions corresponding to the three sub-domains (D I, D II, D III) of the gene as determined by 8 disulphide bonds are indicated within the plot. The plot was generated based on 978 sequences collected between 1995 and 2014.

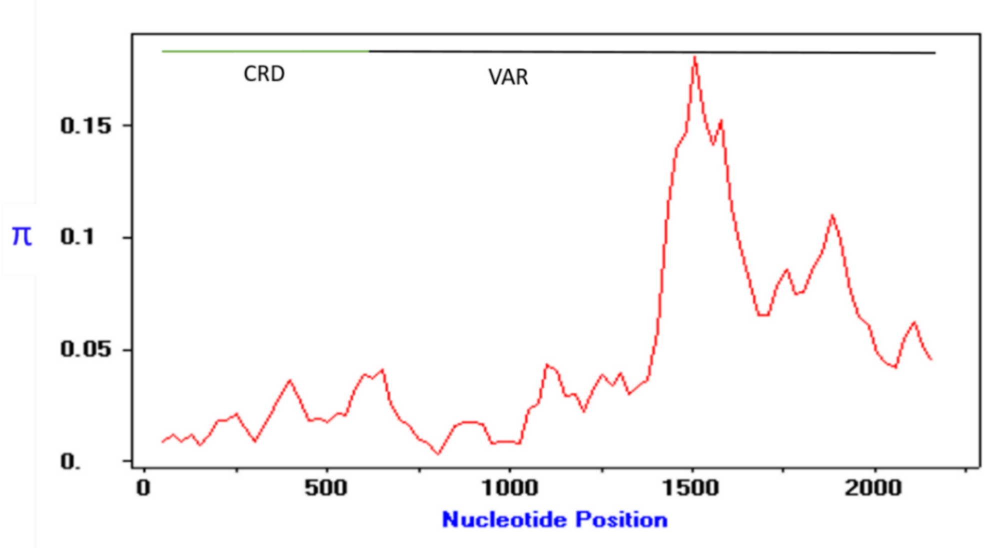


Figure 4.8: Sliding window plot showing the average pairwise nucleotide diversity (π) across a 2199 nucleotide region of *P. falciparum surf*$_{4.2}$ exon 1 encoding the extracellular domain of SURFIN$_{4.2}$ antigen.

A window size of 100 nucleotides with a step size of 25 nucleotides was used to generate the plot. Nucleotide regions corresponding to the cysteine rich domain (CRD) and the variable region (VAR) are indicated within the plot. The plot was generated based on 879 sequences collected between 1995 and 2014.

The high level of diversity was also reflected at the protein level, with 298 and 98 amino acid positions being polymorphic among the *surf$_{4.2}$* and *PfAMA1* sequences, respectively. *PfAMA1* and *surf$_{4.2}$* haplotypes circulating in the Kilifi population during the study period were identified and haplotype diversity (Hd) computed. 641 *PfAMA1* haplotypes were identified out of the 978 sequences, and the corresponding haplotype diversity was very high, at 0.996. 535 of the 641 *PfAMA1* haplotypes identified were unique to individual samples. When the three domains were analysed separately, domain I had 318 unique haplotypes, while domains II and III had 57 and 52 unique haplotypes, respectively. Among the 897 *surf$_{4.2}$* sequences, 685 distinct haplotypes were identified, with a haplotype diversity index of 0.999. 583 of these haplotypes were unique to individual samples. None of the haplotypes could be considered as dominant in either gene, with haplotype frequencies ranging between 1 and 12 (0.15% - 1.7%) among the *surf$_{4.2}$* sequences and 1 and 27 (0.15% – 4.2%) among the *PfAMA1* sequences. For both genes, most of the shared haplotypes were found in samples collected in different years, and identical haplotypes did not cluster within individual years.

4.3.2 Statistical tests of Neutrality

Tajima’s D statistic was calculated in DnaSP5.10.01 to test for departure from neutrality among 978 *PfAMA1* and 897 *surf$_{4.2}$* sequences. The statistic was computed across the entire sequenced region (1731bp for *PfAMA1* and 2199bp for *surf$_{4.2}$*), using a window size of 100 nucleotides with a 25-nucleotide step size. In the case of *AMA1*, additional tests were carried out separately for the three sub-domains as well as the ectodomain region. For *surf*$_{4.2}$, additional tests were carried out on sequences in the variable region and cysteine rich domain. In each gene, the analyses were carried out on the total number of mutations as opposed to the segregating sites to account for multiple mutations at the same site.

Table 4-6: Tajima's D statistic for *Pf*AMA1 and *surf*$_{4.2}$ sequences from children admitted to Kilifi County Hospital, Kenya, between 1995 and 2014.

	L	*S*	η	η_s	π	K	Tajima's D	H (Hd)
***surf*$_{4.2}$**								
Whole sequence	2199	442	487	96	0.045	98.97	1.463	685 (0.999)
VAR	1612	375	418	79	0.055	89.15	1.676	658 (0.998)
CRD	433	52	54	11	0.02	8.541	0.447	288 (0.986)
***Pf*AMA1**								
Whole sequence	1731	137	156	32	0.015	25.74	0.658	641 (0.996)
Ectodomain	1221	92	106	14	0.017	21.25	1.388	450
Domain I	462	54	65	8	0.027	12.75	1.258	318 (0.985)
Domain II	297	14	14	2	0.009	2.83	1.097	57 (0.92)
Domain III	201	13	13	2	0.017	3.34	1.952	52 (0.917)

The analysis was carried out on 978 *PfAMA1* sequences and 897 *surf*$_{4.2}$ sequences. Analysis on *PfAMA1* was carried out on region 70 -1800 of the 1869 base pair long gene. *Surf*$_{4.2}$ analysis was carried out on region 1 – 2199 of the 2299 base pair long extracellular exon. Tajima's D statistic was calculated based on the total number of mutations rather than the number of segregating sites to account for multiple mutation events at a single site. L= sequence length (in base pairs); N = Number of samples; *S* = segregating sites; η = total number of mutations; η_s = number of singleton mutations; π = average nucleotide diversity; K = average number of nucleotide differences; H= number of haplotypes; Hd = haplotype diversity; VAR = variable region; CRD = cysteine-rich domain.

PfAMA1 was associated with positive Tajima's D values when the entire 1731bp region was analysed, as well as when the ectodomain and domains I, II and III were analysed separately (Table 4.6). However, none of the observed Tajima's values were statistically significant. In the *surf*$_{4.2}$ analyses, the entire sequence region, the cysteine rich domain and the variable regions were all associated with positive Tajima's D values.

Sliding window plots were generated to show Tajima's D statistics calculated over the 1731bp *PfAMA1* and 2199bp *surf*$_{4.2}$ regions (figure 4.9). Tajima's D values associated with *PfAMA1* were not significant along the entire sequence, although in *surf*$_{4.2}$, statistically significant values were observed at different points along the sequence (figure 4.9).

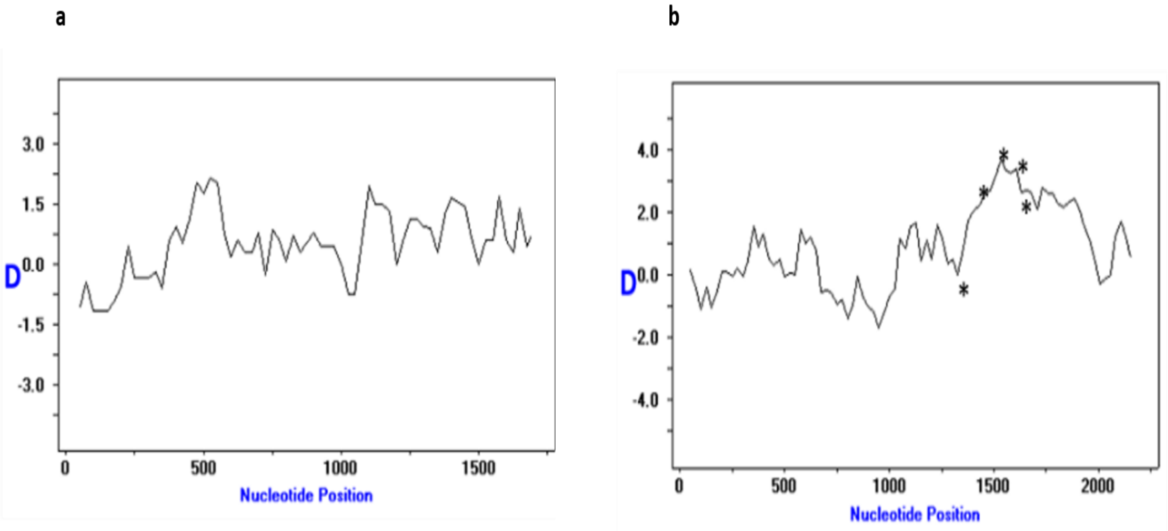


Figure 4.9: Sliding window plots of Tajima's D for *Pf*AMA1 and *surf*$_{4.2}$ sequences. Plots were generated for 978 *PfAMA1* sequences (**a**) and 879 *surf*$_{4.2}$ sequences (**b**). The test statistic was calculated over a sliding window size of 100 nucleotides, with a 25-nucleotide step size. Asterisks indicate regions that were statistically significant ($p < 0.05$).

4.3.3 Analysis of linkage disequilibrium and recombination

Linkage disequilibrium was computed in a pairwise analysis of 323 polymorphic sites in 897 *surf*$_{4.2}$ sequences and 88 polymorphic sites in 978 *PfAMA1* sequences in the Kilifi *P. falciparum* parasite population. In the *surf*$_{4.2}$ analysis, 23509 SNP pairs were found to be statistically significant based on Fisher's exact test ($p<0.05$), with 11590 pairs remaining significant after accounting for multiple testing using Bonferroni correction. In the *PfAMA1* analysis, there were 1376 SNP pairs that were shown to be in LD based on Fisher's exact test ($p<0.05$). 720 of these comparisons were significant after accounting for multiple comparisons using Bonferroni correction method. Both D' and R^2 indices of LD showed a rapid decline of LD with distance among the sequences, with most of the significant non-random associations being present between SNPs that were less than 600 base pairs apart (figures 4.10 and 4.11).

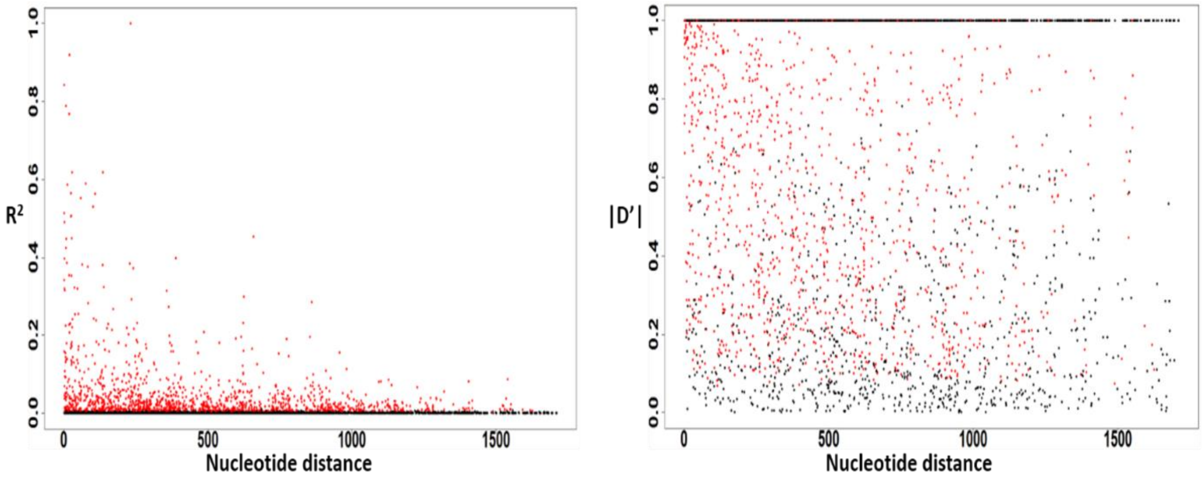


Figure 4.10: Linkage disequilibrium (LD) calculated using R^2 and D' across a 1731 base pair region of *PfAMA1* gene in a parasite population from Kilifi, Kenya. Plots show R^2 and the absolute values of D' (|D'|) plotted against nucleotide distance. Red dots represent those pairs of nucleotide sites that showed significant linkage disequilibrium based on Fisher's exact test, while all other pairs are represented by black dots.

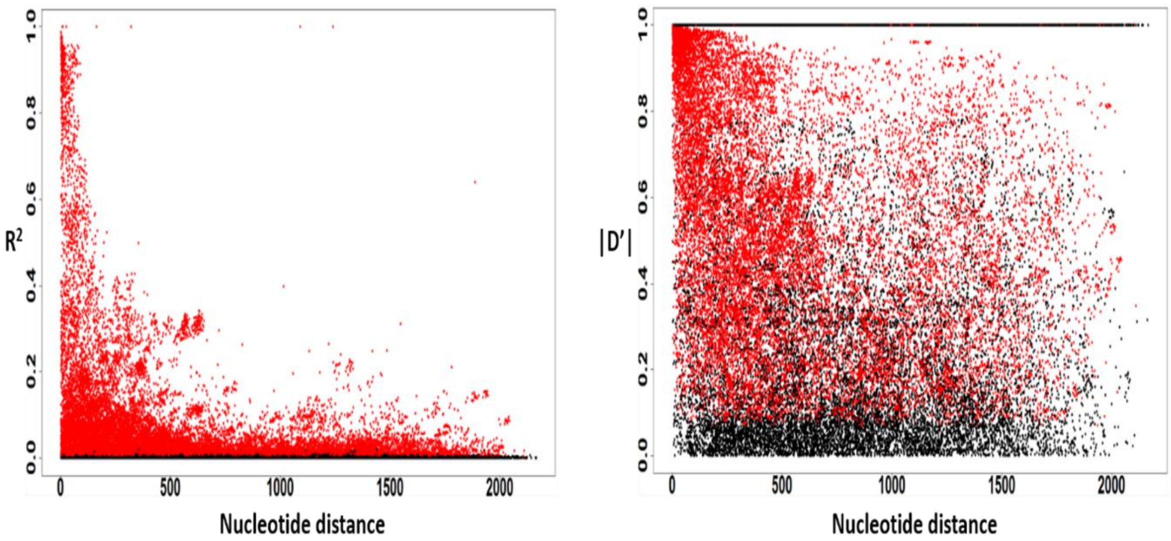


Figure 4.11: Linkage disequilibrium (LD) calculated using R^2 and D' across a 2193 base pair region of *surf*$_{4.2}$ gene in a parasite population from Kilifi, Kenya.
Plots show R^2 and the absolute values of D' (|D'|) plotted against nucleotide distance. Red dots represent those pairs of nucleotide sites that showed significant linkage disequilibrium based on Fisher's exact test, while all other pairs are represented by black dots.

The minimum number of recombination events (Rm) estimated to have occurred in the *PfAMA1* sequences in this parasite population was 47, while the Rm estimated to have occurred among the *surf*$_{4.2}$ sequences was 177. The estimated recombination parameter, C, was 0.1168 between adjacent sites and 202 in the whole *PfAMA1* sequence. When the different gene regions were analysed separately, C across the analysed region and between adjacent sites was determined to be, respectively, 152 and 0.1246 for the entire ectodomain, 102 and 0.2213 for domain I, 101 and 0.3412 for domain II and 28.5 and 0.1425 for domain III. Among the *surf*$_{4.2}$ sequences, the recombination parameter between adjacent sites was estimated at 0.03 between adjacent sites and 65.9 in the whole sequence. When the cysteine rich domain (CRD) and variable regions were considered separately, C across the whole region and between adjacent sites was estimated at 143 and 0.3310, respectively, for the CRD and 53 and 0.0329, respectively, for the variable region.

4.3.4 Temporal and spatial patterns of genetic diversity among *AMA1* and *surf$_{4.2}$* sequences in *P. falciparum* isolates from Kilifi, Kenya.

To provide a visual summary of the genetic diversity present among *P. falciparum* *AMA1* and *surf$_{4.2}$* sequences collected in coastal Kenya between 1995 and 2014, heatmaps were generated based on the segregating sites present in the two genes. The heatmaps showed a high level of genetic diversity spread throughout the sequenced regions of the genes (figure 4.12 – figure 4.23). No discernible pattern of clustering was observed among the *PfAMA1* sequences, either in time when sequences were analysed based on year of collection (figure 4.12), nor in space when they were analysed based on location of origin of the samples (figure 4.15). The addition of a hierarchical clustering algorithm to group the sequences based on genetic similarity did not show patterns of clustering of sequences either in time (figure 4.13) or over space (figure 4.16), and bootstrapping the cluster analysis gave little statistical support for the identified clusters (figures 4.14 and 4.17). Instead, sequences collected from different years (1995 - 2014) and from different geographical regions (north and south) clustered together. This pattern is indicative of a well-mixed parasite population, both in time and space, with the same SNPs being shared by samples collected at different time points, and from different geographical locations.

Similar patterns were observed when the heatmaps were generated using nucleotide or amino acid sequence data, showing that the high level of nucleotide diversity translates into a high level of antigenic diversity in *PfAMA1* in this parasite population.

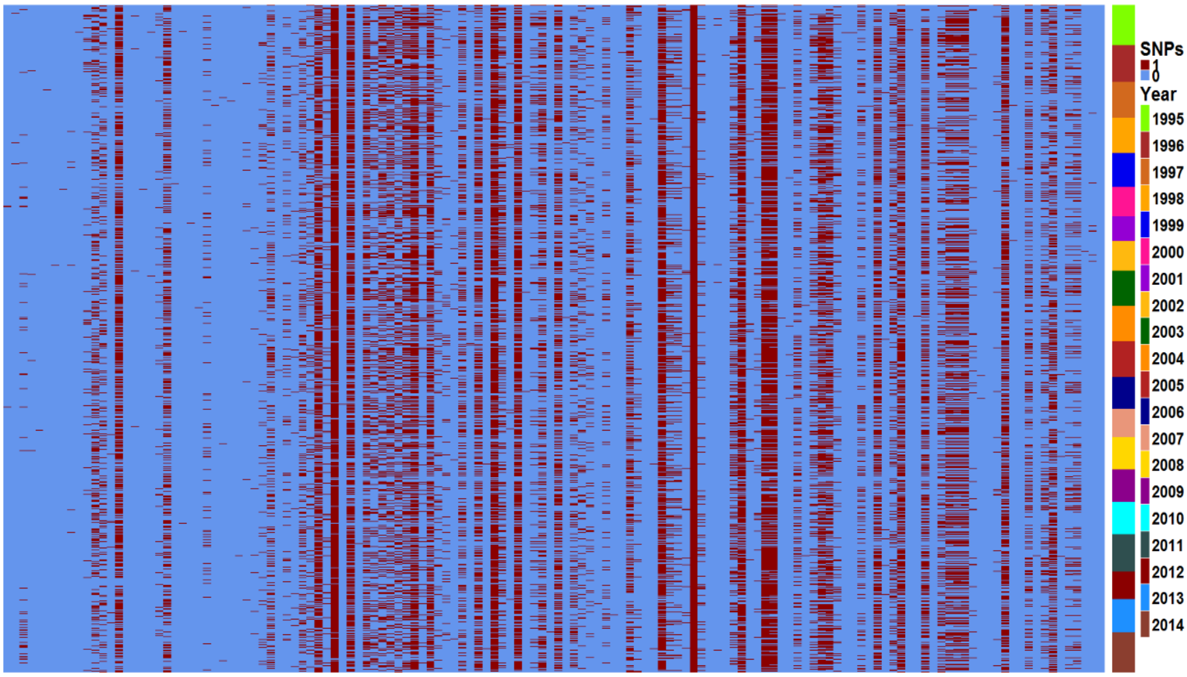


Figure 4.12: Clustering of *AMA1* nucleotide sequences in a *P. falciparum* population from coastal Kenya based on the year of sample collection.

The alignment was based on the segregating sites identified among the sequences. Each row represents one of 978 sequences and each column represents one of 138 polymorphic positions. At each position, nucleotide bases that are identical to *P. falciparum* 3D7 *AMA1* reference sequence are coloured blue, and those that differ from the reference are coloured red. The sequences are grouped in order, based on year of collection from 1995 to 2014.

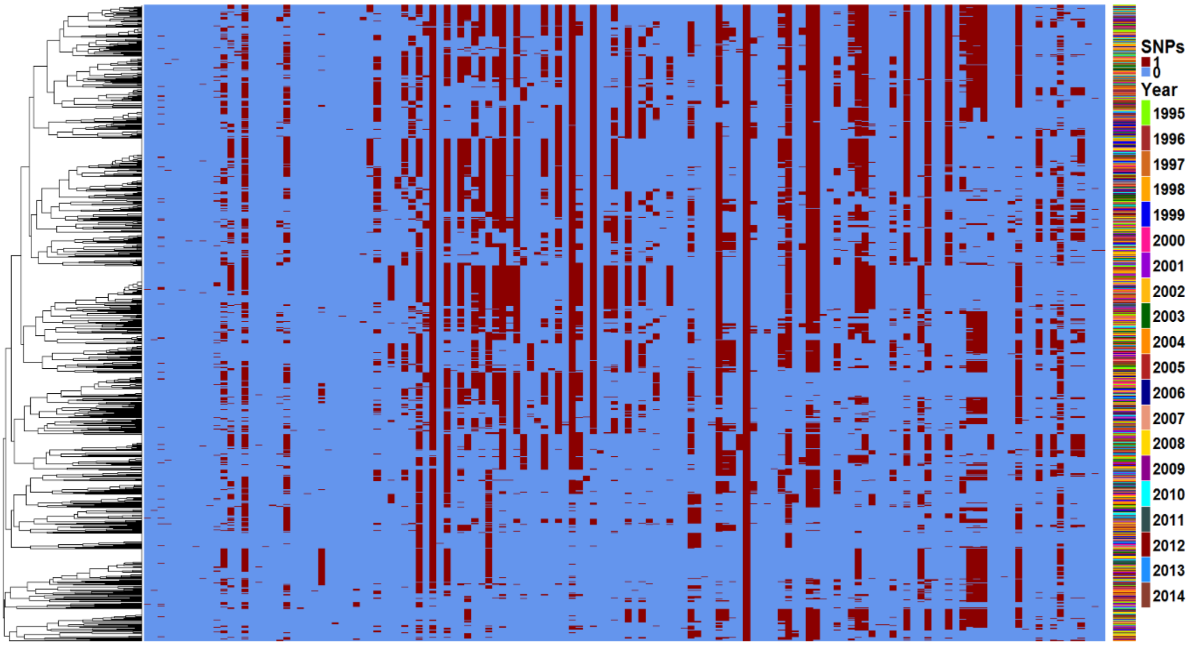


Figure 4.13: Temporal clustering of *AMA1* nucleotide sequences in a *P. falciparum* population from coastal Kenya based on genetic similarity.

The alignment was based on the segregating sites identified among the sequences. Each row represents one of 978 sequences and each column represents one of 138 polymorphic positions. At each position, nucleotide bases that are identical to *P. falciparum* 3D7 *AMA1* reference sequence are coloured blue, and those that differ from the reference are coloured red. Hierarchical clustering was carried out on the sequences (rows) to group samples that are more similar to each other. A dendrogram is added to the heatmap to show this clustering relationship, with sequences that are more similar to each other appearing in the same clade. Year labels are coloured as in figure 4.12 above and include samples collected between 1995 – 2014.

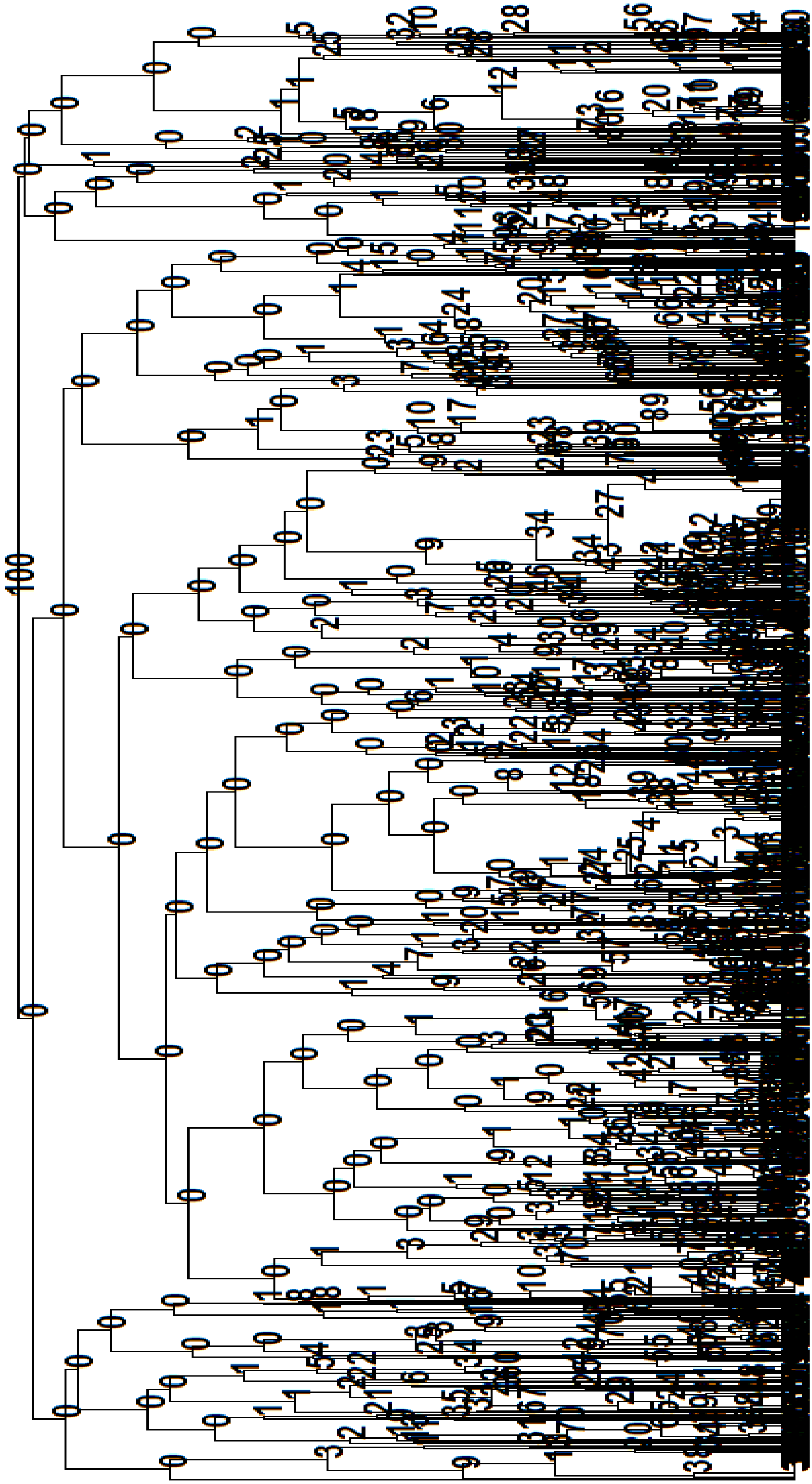


Figure 4.14: Bootstrap support for the *PfAMA1* temporal cluster.

Dendrogram representation of the observed sequence clusters following a bootstrap analysis to test the statistical significance of the hierarchical clustering. Values on each node represent the number of times the specific cluster was observed in 1000 resampling steps. The order of the alignment in the heatmap was maintained in the dendrogram.

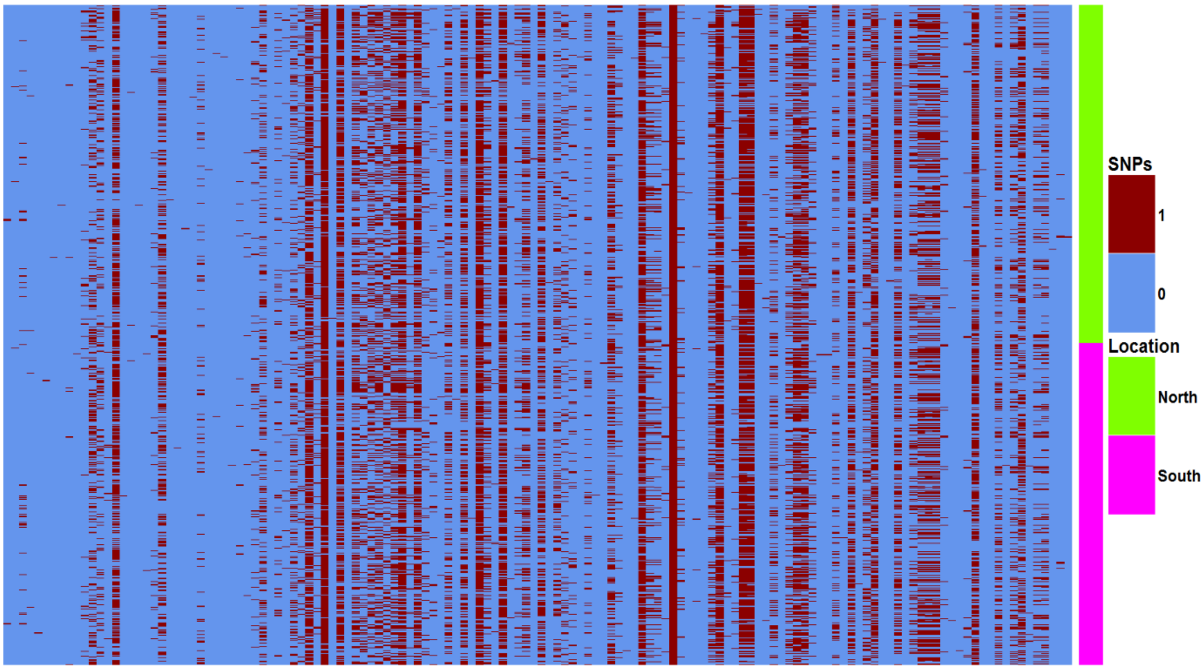


Figure 4.15: Clustering of *AMA1* nucleotide sequences in a *P. falciparum* population from coastal Kenya based on the location of sample collection.

The alignment was based on 138 segregating sites identified among the sequences. Each row represents one of 795 sequences and each column represents one of 138 polymorphic positions. At each position, nucleotide bases that are identical to *P. falciparum* 3D7 *AMA1* reference sequence are coloured blue, and those that differ from the reference are coloured red. The sequences are clustered based on whether they were collected north (green) or south (pink) of the Kilifi Creek (the latitude position -3.64

was used to mark the north/south boundary). 407 of the samples were collected in the north and 388 were collected in the south.

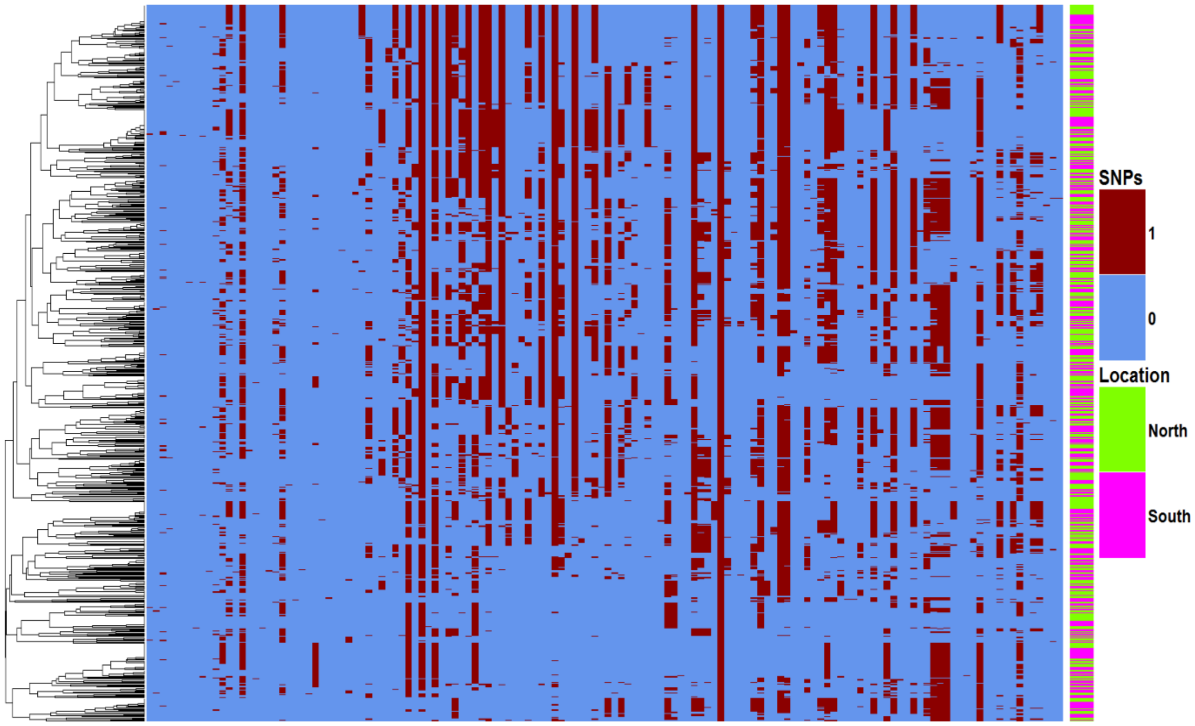


Figure 4.16: Spatial clustering of *AMA1* nucleotide sequences in a *P. falciparum* population from coastal Kenya based on genetic similarity.

The alignment was based on the segregating sites identified among the sequences. Each row represents one of 795 sequences and each column represents one of 138 polymorphic positions. At each position, nucleotide bases that are identical to *P. falciparum* 3D7 *AMA1* reference sequence are coloured blue, and those that differ from the reference are coloured red. Hierarchical clustering was carried out on the sequences (rows) to group samples that are more similar to each other. A dendrogram is added to the heatmap to show this clustering relationship, with sequences that are more similar to each other appearing in the same clade. The sequences are labelled based on whether they were collected north (green) or south (pink) of the Kilifi Creek (the latitude position -3.64 was used to mark the north/south boundary). 407 of the samples were collected in the north and 388 were collected in the south.

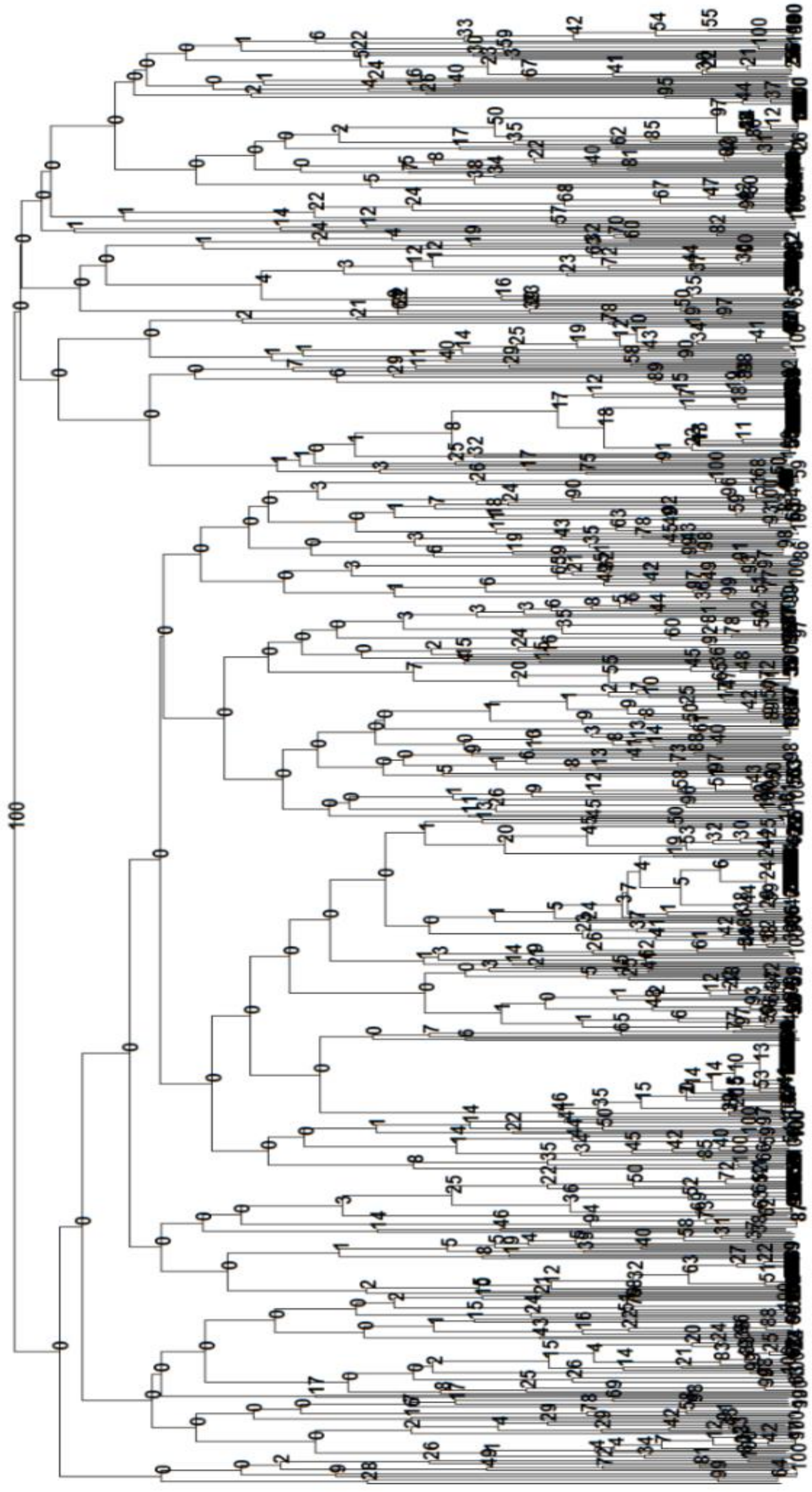

100

Figure 4.17: Bootstrap support for the *PfAMA1* spatial cluster.

Dendrogram representation of the observed sequence clusters following a bootstrap analysis to test the statistical significance of the hierarchical clustering. Values on each node represent the number of times the specific cluster was observed in 1000 resampling steps. The order of the alignment in the heatmap was maintained in the dendrogram.

Heatmaps showing patterns of diversity were also generated for the *surf*$_{4.2}$ gene, with the nucleotide and amino acid sequences being clustered by year and location of collection as well as by sequence similarity in time and space. Patterns of distribution of genetic diversity similar to those seen in *PfAMA1* were shown to exist among the *surf*$_{4.2}$ sequences, with no specific groupings of SNPs among samples collected in the same year (figure 4.18) or in the same location (figure 4.21). However, a different pattern was observed among the sequences when they were clustered based on sequence similarity in time and space. The sequences could be divided into three separate clades, based on the presence or absence of two distinct blocks of SNPs. One of these blocks covered nucleotide region 1427 – 1896, and the second covered nucleotide region 1427 – 2114, with both blocks falling within the variable domain region of the exon. The SNP blocks were found in samples collected at different time points (figures 4.19 and 4.20), and from different geographical locations (figures 4.22 and 4.23), indicating a high level of mixing of parasites across time and space.

Similar patterns were observed when the corresponding polymorphic amino acid residues were used to generate the heatmaps instead of nucleotide sequences. However, bootstrapping the analysis showed little statistical support for the observed clustering pattern, with most of the clusters having bootstrap values of 0, meaning that the specific node clusters were not obtained in any of the resampling steps.

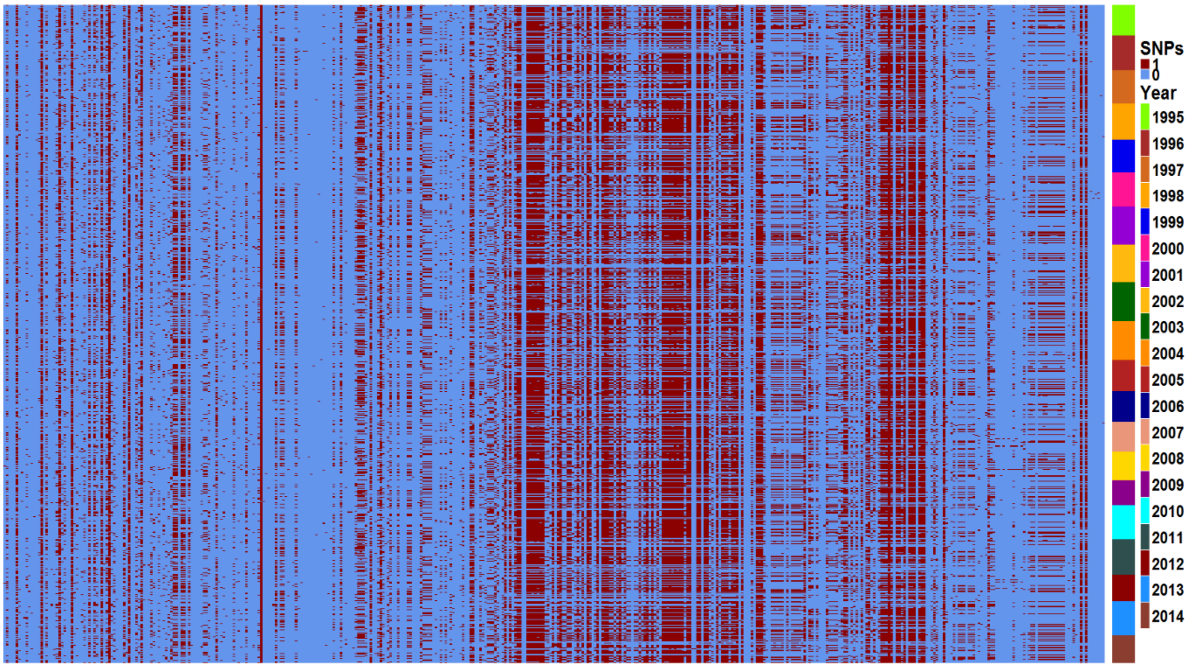


Figure 4.18: Clustering of *surf$_{4.2}$* nucleotide sequences in a *P. falciparum* population from coastal Kenya based on the year of sample collection.

The alignment was based on the segregating sites identified among the sequences. Each row represents one of 879 sequences and each column represents one of 442 polymorphic positions. At each position, nucleotide bases that are identical to *P. falciparum* 3D7 *surf$_{4.2}$* reference sequence are coloured blue, and those that differ from the reference are coloured red. The sequences are grouped according to the year of sample collection from 1995 to 2014.

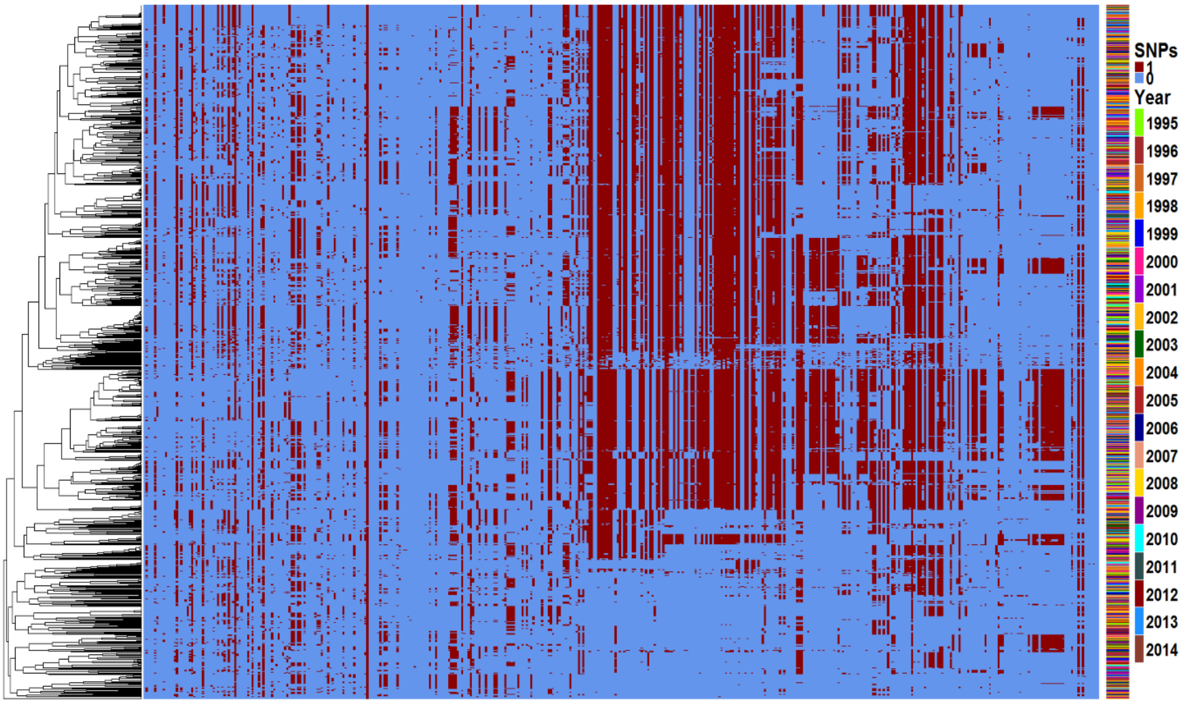


Figure 4.19: Temporal clustering of *surf$_{4.2}$* nucleotide sequences in a *P. falciparum* population from coastal Kenya based on genetic similarity.

The alignment was based on the segregating sites identified among the sequences. Each row represents one of 879 sequences and each column represents one of 442 polymorphic positions. At each position, nucleotide bases that are identical to *P. falciparum* 3D7 *surf$_{4.2}$* reference sequence are coloured blue, and those that differ from the reference are coloured red. Hierarchical clustering was carried out on the sequences (rows) to group samples that are more similar to each other. A dendrogram is added to the heatmap to show this clustering relationship, with sequences that are more similar to each other appearing in the same clade. Year labels are coloured as in figure 4.15 above and include samples collected between 1995 – 2014.

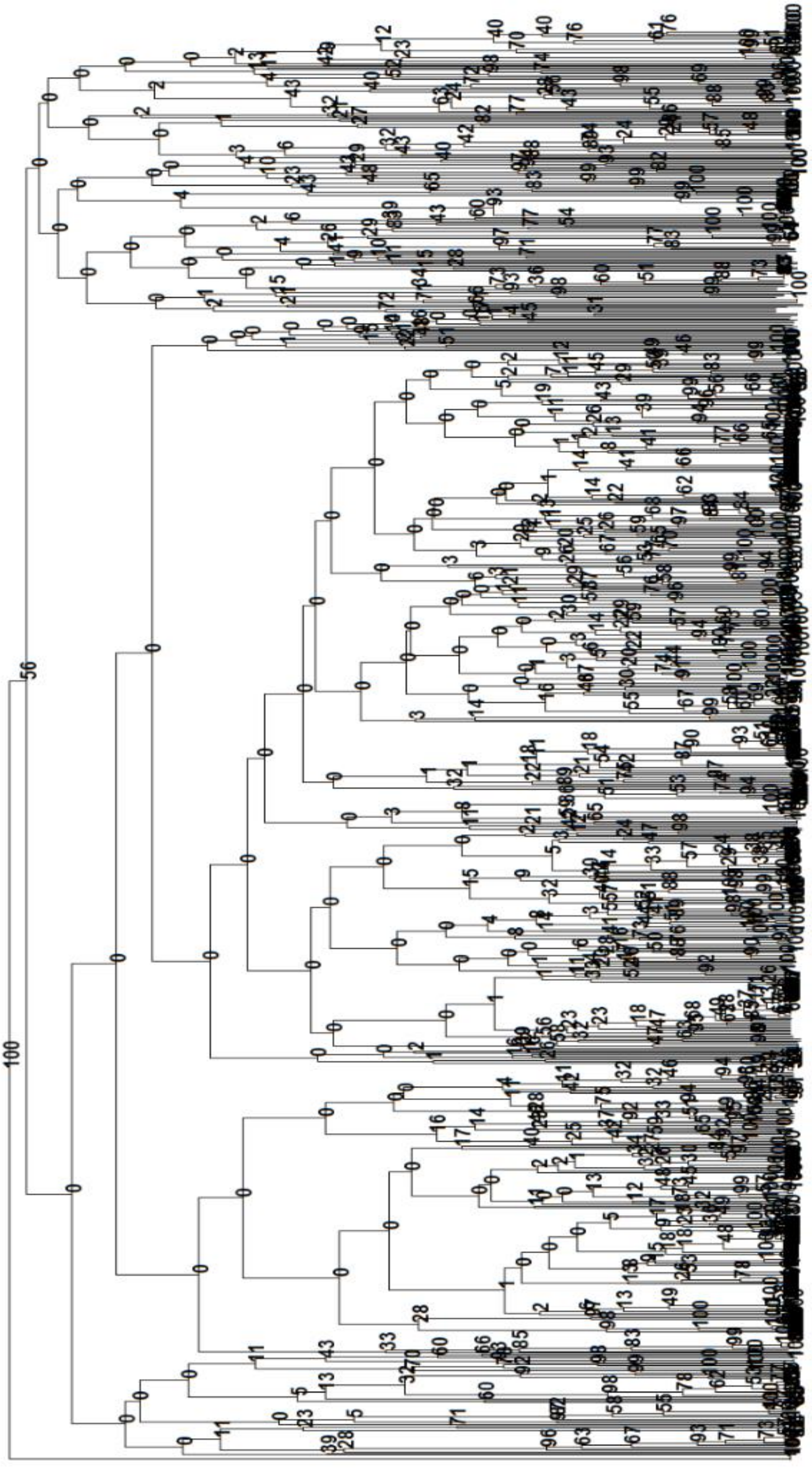

100
56

Figure 4.20: Bootstrap support for the *surf$_{4.2}$* temporal cluster.

Dendrogram representation of the observed sequence clusters following a bootstrap analysis to test the statistical significance of the hierarchical clustering. Values on each node represent the number of times the specific cluster was observed in 1000 resampling steps. The order of the alignment in the heatmap was maintained in the dendrogram.

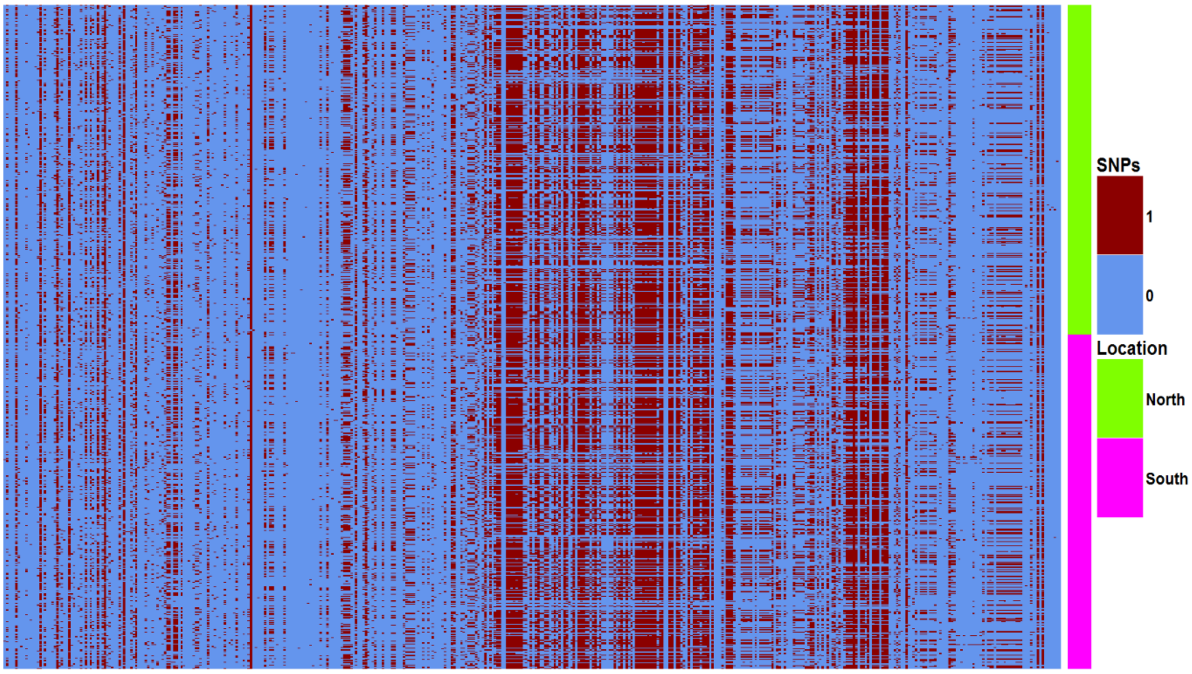


Figure 4.21: Clustering of *surf$_{4.2}$* nucleotide sequences in a *P. falciparum* population from coastal Kenya based on the location of sample collection.

The alignment was based on the segregating sites identified among the sequences. Each row represents one of 754 sequences and each column represents one of 442 polymorphic positions. At each position, nucleotide bases that are identical to *P. falciparum* 3D7 *surf$_{4.2}$* reference sequence are coloured blue, and those that differ from the reference are coloured red. The sequences are clustered based on whether they were collected north (green) or south (pink) of the Kilifi Creek (the latitude position -3.64 was used to mark the north/south boundary). 374 of the samples were collected in the north and 380 were collected in the south.

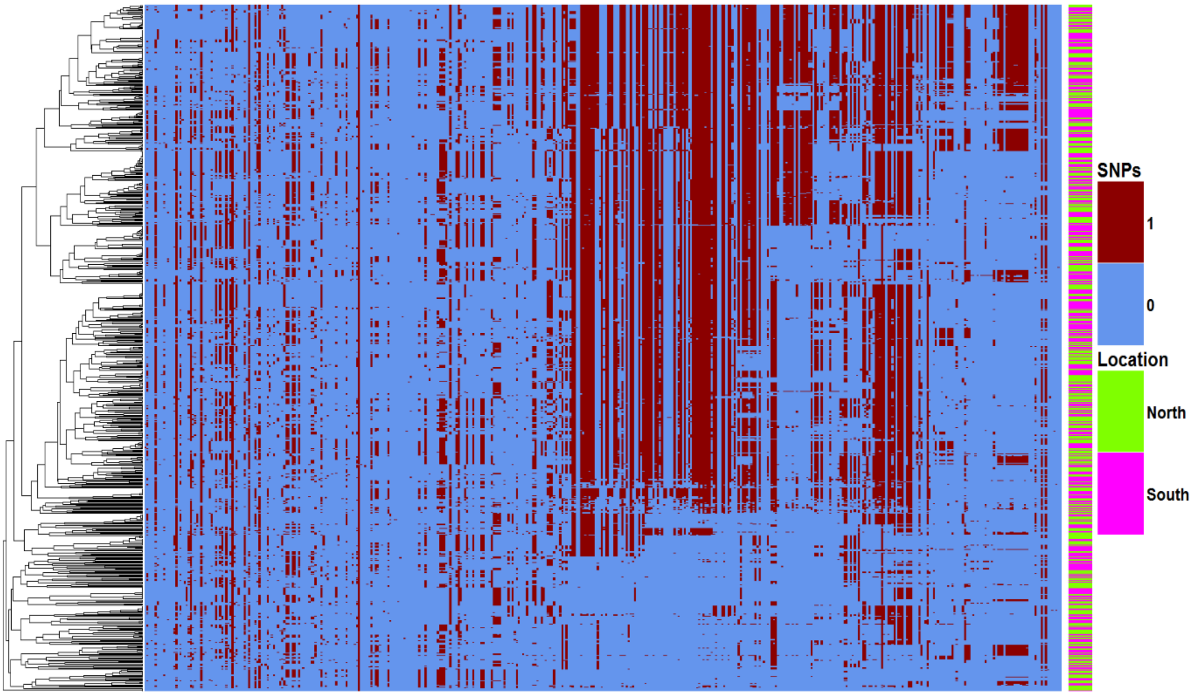


Figure 4.22: Spatial clustering of *surf$_{4.2}$* nucleotide sequences in a *P. falciparum* population from coastal Kenya based on genetic similarity.

The alignment was based on the segregating sites identified among the sequences. Each row represents one of 754 sequences and each column represents one of 442 polymorphic positions. At each position, nucleotide bases that are identical to *P. falciparum* 3D7 *surf$_{4.2}$* reference sequence are coloured blue, and those that differ from the reference are coloured red. Hierarchical clustering was carried out on the sequences (rows) to group samples that are more similar to each other. A dendrogram is added to the heatmap to show this clustering relationship, with sequences that are more similar to each other appearing in the same clade. The sequences are labelled based on whether they were collected north (green) or south (pink) of the Kilifi Creek (the latitude position -3.64 was used to mark the north/south boundary). 374 of the samples were collected in the north and 380 were collected in the south.

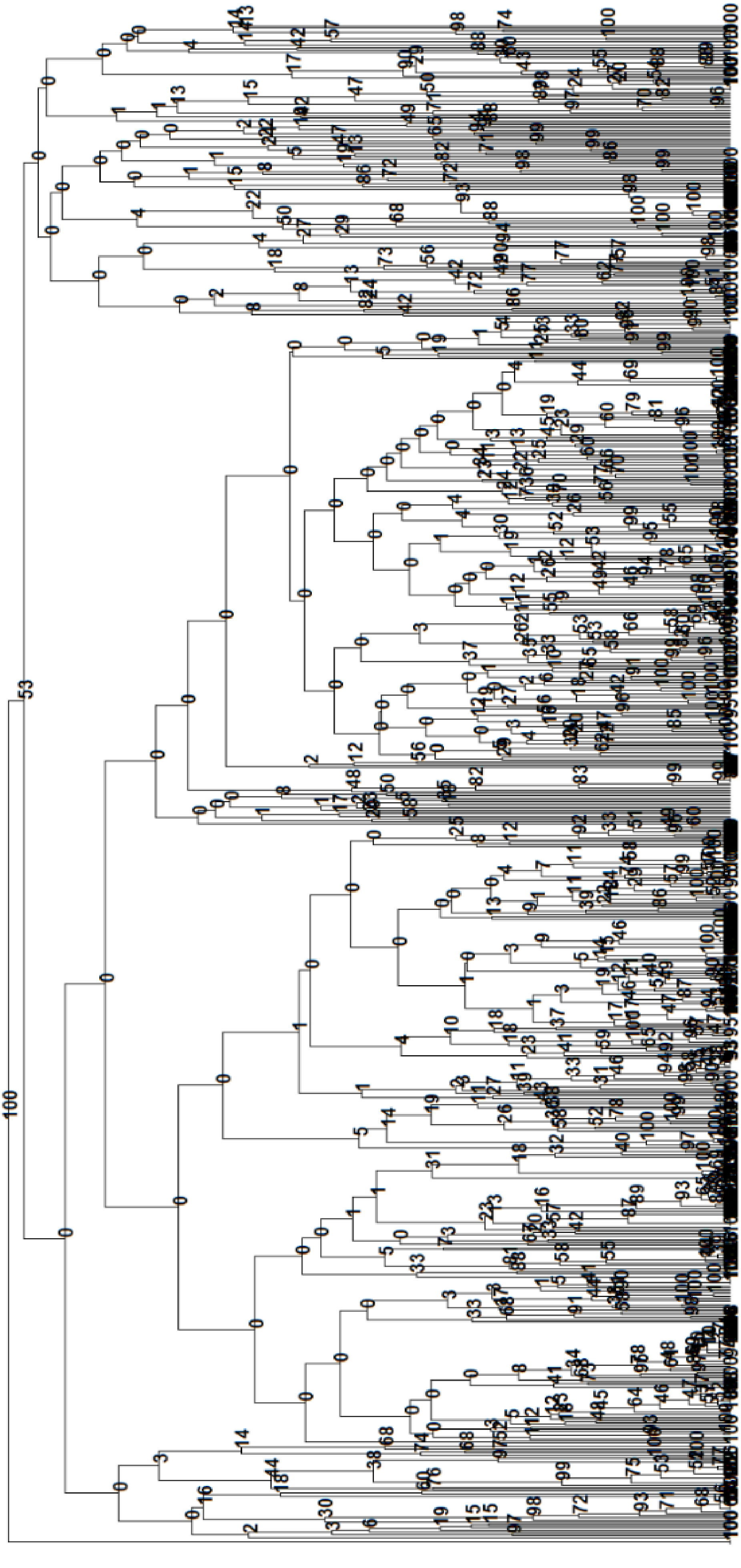


Figure 4.23: Bootstrap support for the *surf$_{4.2}$* spatial cluster.
Dendrogram representation of the observed sequence cluster following a bootstrap resampling analysis to test the statistical significance of the hierarchical clustering. Values on each node represent the number of times the specific cluster was observed in 1000 resampling steps. The order of the alignment in the heatmap was maintained in the dendrogram.

In summary, spatio-temporal genetic analyses of *PfAMA1* and *surf$_{4.2}$* sequences using heatmaps and dendrograms displayed a high degree of mixing within the population, with no obvious clustering of distinct genotypes either in time or space. Instead, parasites with identical or similar genotypes were sampled from different time periods and different geographical locations. Overall, there was little statistical support for most of the dendrogram clusters observed, indicating that the clustering patterns observed where mainly due to chance.

4.3.5 SNP differences versus Identity by State (IBS) for analysis of spatial and temporal genetic variation in *PfAMA1* and *surf$_{4.2}$* genes

Two different but correlated metrics; SNP differences and identity by state (IBS), were used to measure the trends of genetic variation between *P. falciparum* parasite pairs in the Kilifi population. *AMA1* and *surf$_{4.2}$* had an average number of 25.901 and 99.37 SNP differences per parasite pair, respectively and mean IBS of 24.23 and 51.64 nucleotides for *AMA1* and *surf$_{4.2}$*, respectively (figure 4.24). There were several parasite pairs which were identical along their entire sequence lengths (i.e. had no genetic variations).

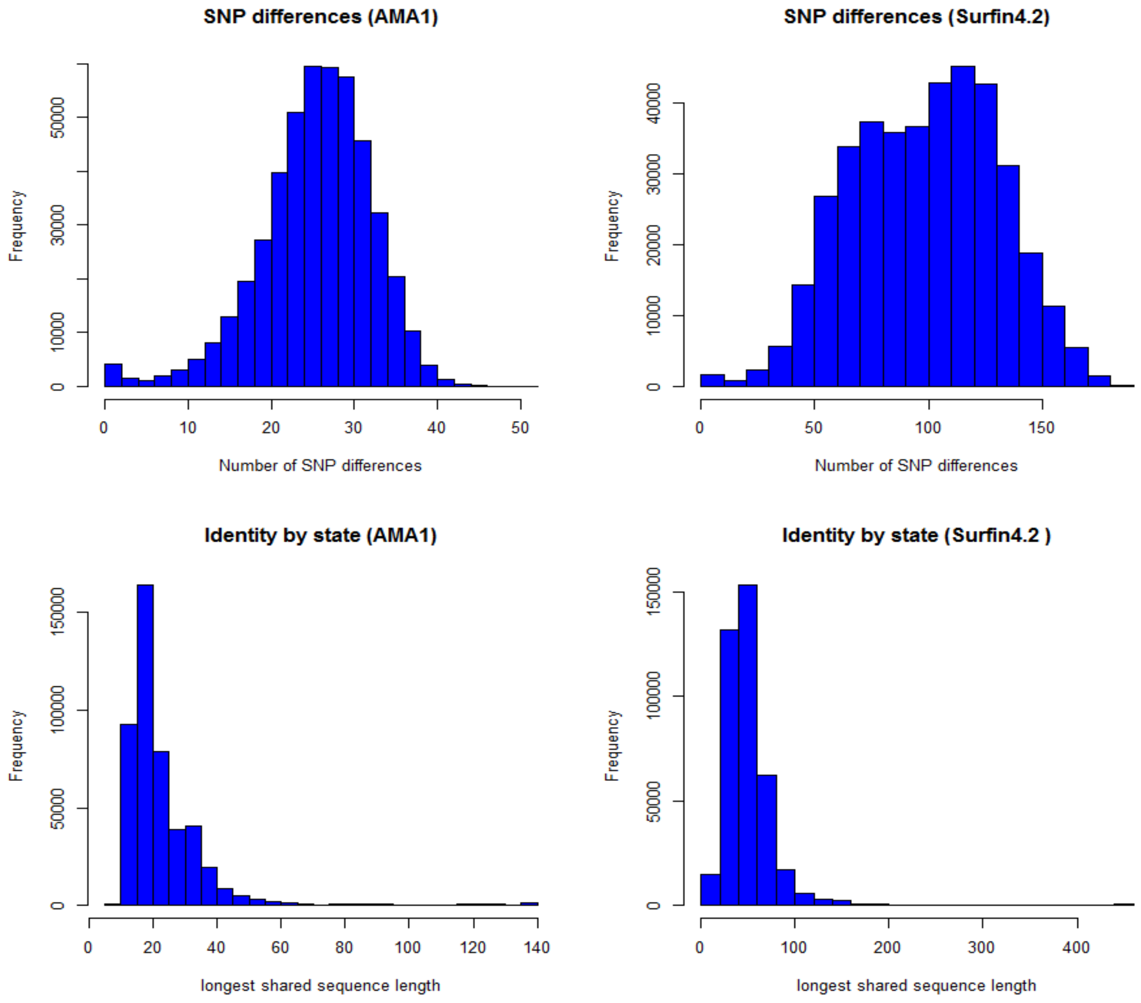


Figure 4.244: Distribution of the number of SNP differences and the longest shared contiguous sequence length (measure of IBS) in AMA1 and surf$_{4.2}$ sequences among P. falciparum parasite pairs in Kilifi, Kenya.

The top panels represent the distribution of the number of SNP differences between *PfAMA1* (**left**) and *surf$_{4.2}$* (**right**) and the bottom panels represent the distribution of the longest shared contiguous sequence length (measure of identity by state) for *PfAMA1* and *surf$_{4.2}$*.

Variations in the number of pairwise SNP differences was inversely correlated with the length of the longest shared sequence between parasite pairs when both *PfAMA1* (figure 4.25) and *surf$_{4.2}$* (figure 4.26) were analysed. However, although a correlation was observed, there was substantial scatter, indicating that there is a lot of variation in the data.

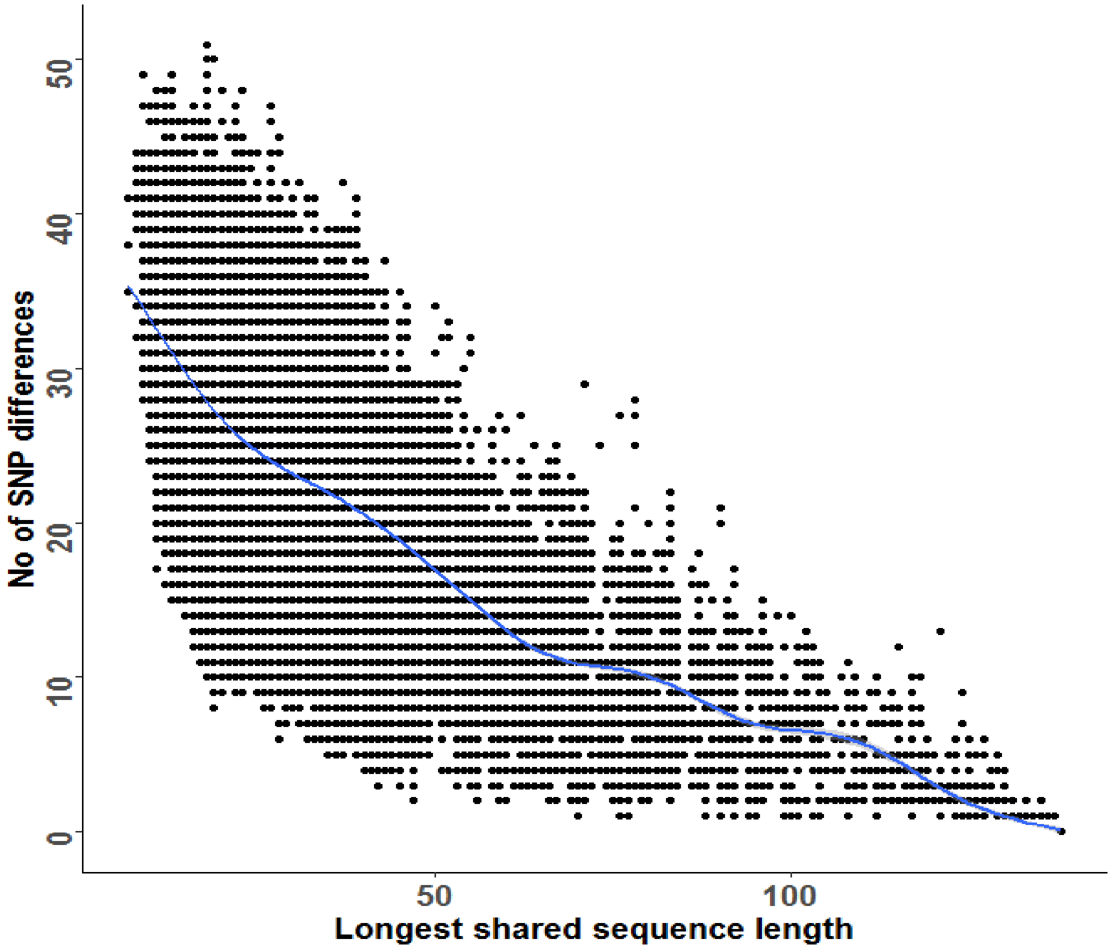


Figure 4.255: Relationship between the number of SNP differences and the longest shared sequence length (IBS) among AMA1 sequences from a P. falciparum parasite population in Kilifi, Kenya.

978 parasites were compared in a pairwise analysis when computing both SNP differences and IBS. The blue line represents a loess smoothed line showing the relationship between the two metrics.

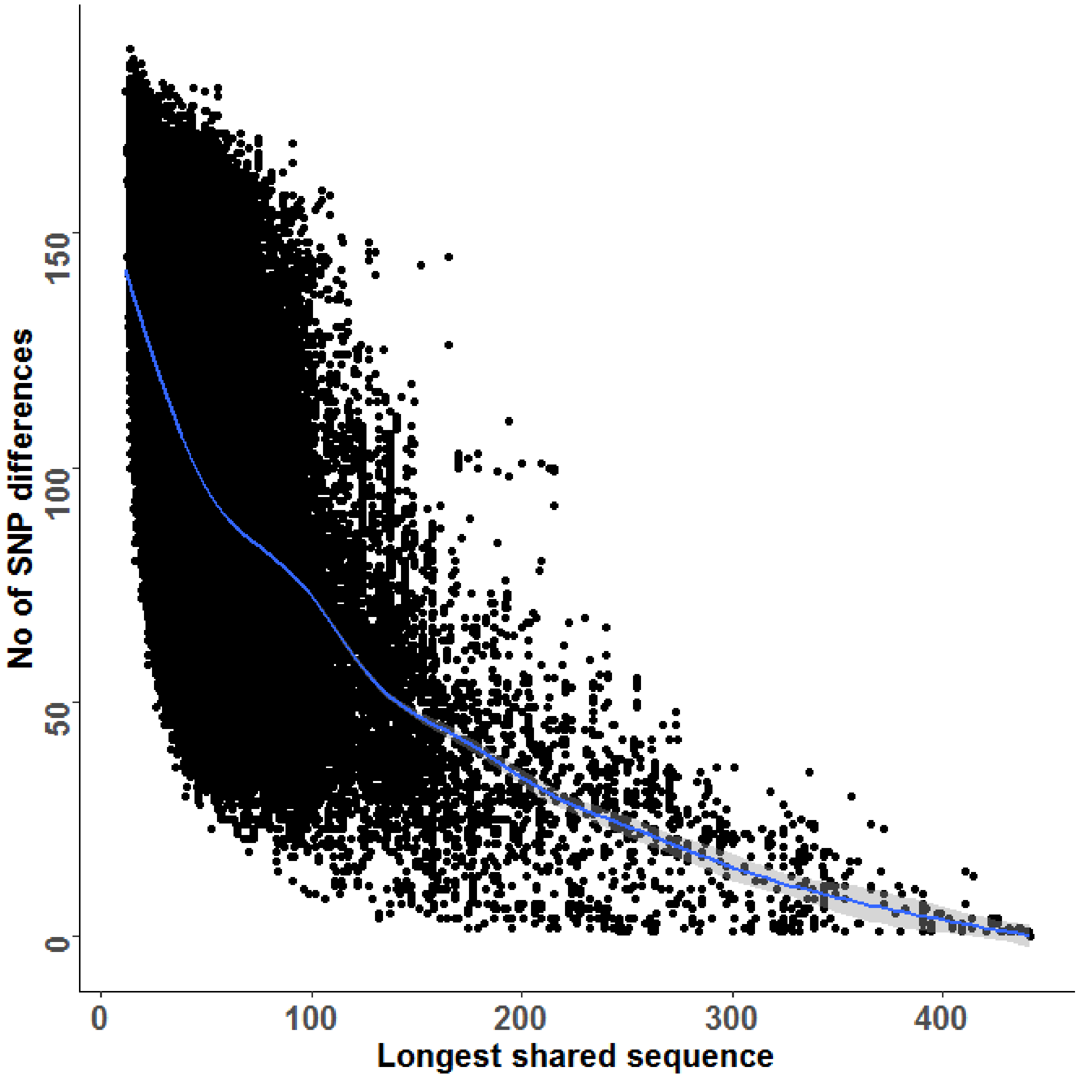


Figure 4.266: Relationship between the number of SNP differences and the longest shared sequence length (IBS) among *surf$_{4.2}$* sequences from a *P. falciparum* parasite population in Kilifi, Kenya.

897 parasites were compared in a pairwise analysis when computing both SNP differences and IBS. The blue line represents a loess smoothed line with 95% confidence intervals (grey shaded area) around it.

Multiple fractional polynomial regression analysis was used to interrogate the independent effects of time and distance on genetic variation in the parasite population, using both SNP and IBS data. Time between sampling (in days) was associated with increasing number of SNP differences between parasite pairs in both *PfAMA1*

(6.338x10^{-5} SNP differences per day) and *surf*$_{4.2}$ (8.544x10^{-5} SNP differences per day) datasets (figures 4.27 and 4.28), and this association was significant in *PfAMA1* (p = 0.002; 95% CI: 2.3x10^{-5} – 1.11x10^{-4}), although not in *surf*$_{4.2}$ (P = 0.3; 95% CI: -2.09x10^{-4} – 4.05x10^{-4}).

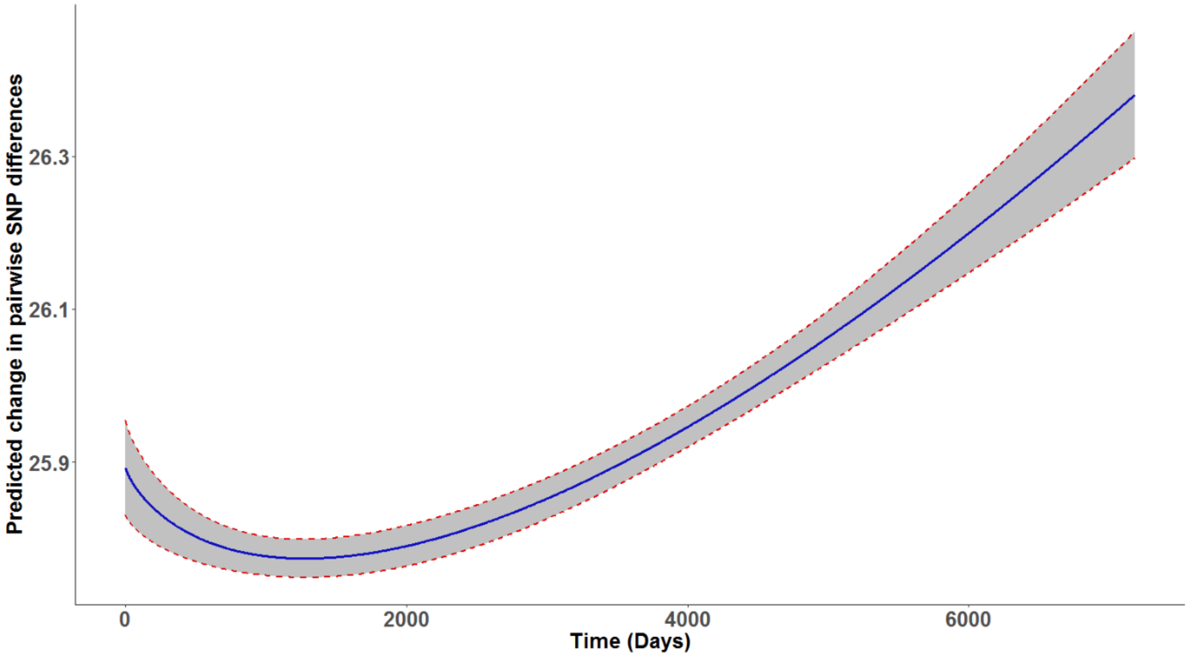


Figure 4.27: Effect of time on variations in the number of SNP differences in *AMA1* sequences between *P. falciparum* parasite pairs.

The analysis was carried out on 966 samples collected from children admitted to the Kilifi County Hospital between 1995 and 2014. The relationship between the number of SNP differences and time between sample collection was modelled in a multiple fractional polynomial regression. The relationship was statistically significant (p=0.002; 95% CI: 2.3x10^{-5} – 1.11x10^{-4}) based on a bootstrapped linear regression analysis with 1000 resampling steps. The blue line represents the estimated change in the number of pairwise SNP differences over time and the grey shaded regions represent the 95% confidence intervals, with the lower and upper bounds defined by the red dotted lines.

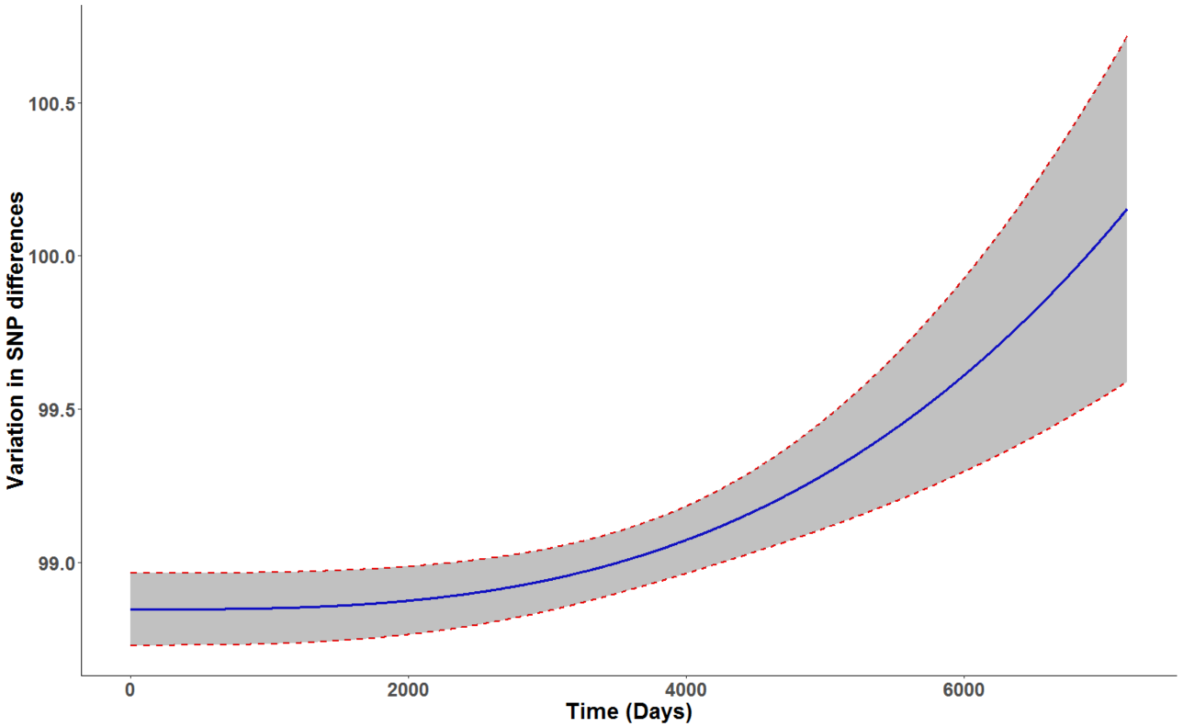


Figure 4.28**: Effect of time on variations in the number of SNP differences in** ***surf$_{4.2}$*** **sequences between** ***P. falciparum*** **parasite pairs.**

The analysis was carried out on 890 samples collected from children admitted to the Kilifi County Hospital between 1995 and 2014. The relationship between the number of SNP differences and time between sample collection was modelled in a multiple fractional polynomial regression. The relationship was not statistically significant (P= 0.3; 95% CI: -2.09x10^{-4} – 4.05x10^{-4}) based on a bootstrapped linear regression analysis with 1000 resampling steps. The blue line represents the estimated change in the number of pairwise SNP differences over time and the grey shaded regions represent the 95% confidence intervals, with the lower and upper bounds defined by the red dotted lines.

Geographical distance (in kilometres) between individuals in the study location was also associated with increasing number of SNP differences between parasite pairs when *PfAMA1* (effect size = 0.006 SNP differences per kilometre) and *surf*$_{4.2}$ (effect size = 0.013 SNP differences per kilometre) sequences were analysed (figures 4.29 and 4.30). However, bootstrapping the analyses to determine significance of the observed results showed non-statistically significant associations for both *PfAMA1* (P=0.082; 95% CI: -0.003 – 0.017) and *surf*$_{4.2}$ (P = 0.38; 95% CI: -0.052 – 0.076).

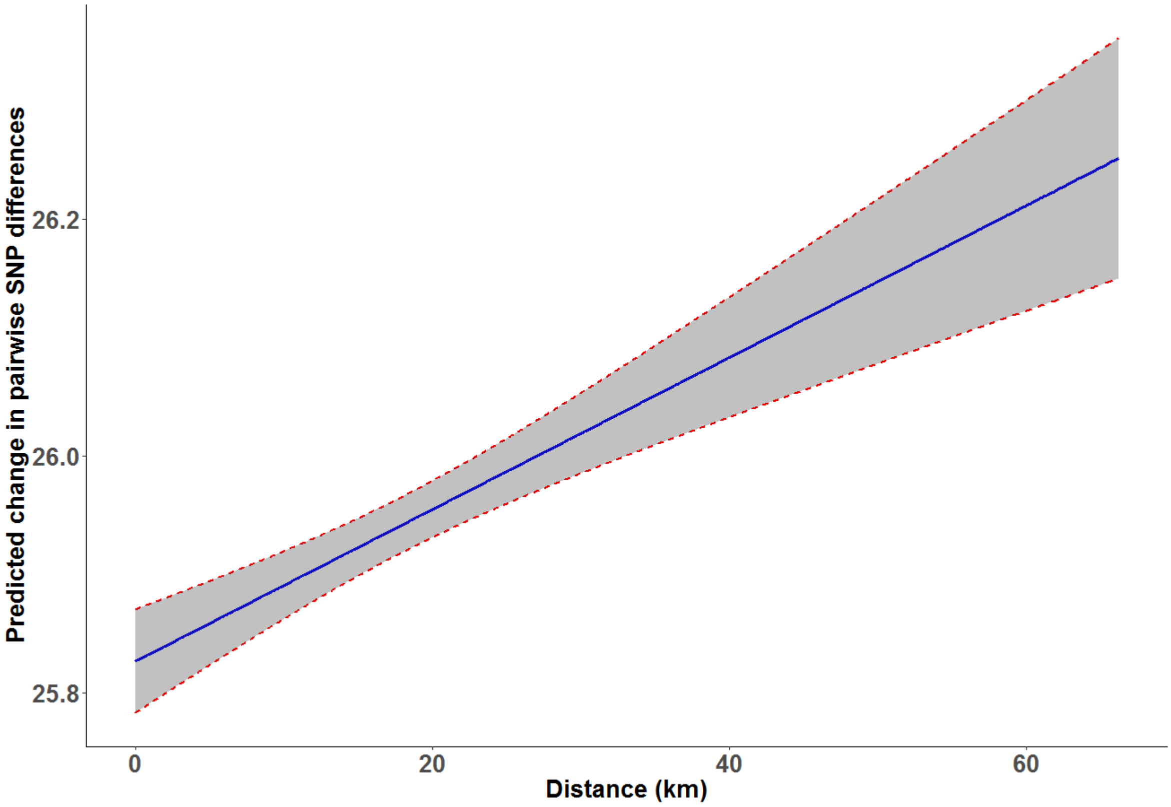


Figure 4.2928: Effect of distance on variations in the number of SNP differences in *AMA1* sequences between *P. falciparum* parasite pairs.

The analysis was carried out on 795 samples collected from children admitted to the Kilifi County Hospital between 1995 and 2014. The relationship between the number of SNP differences and distance between sample collection points was modelled in a multiple fractional polynomial regression. The relationship was not statistically significant (p=0.082; 95% CI: -0.003 – 0.017) based on a bootstrapped linear regression analysis with 1000 resampling steps. The blue line represents the estimated change in the number of pairwise SNP differences over distance and the grey shaded regions represent the 95% confidence intervals, with the lower and upper bounds defined by the red dotted lines.

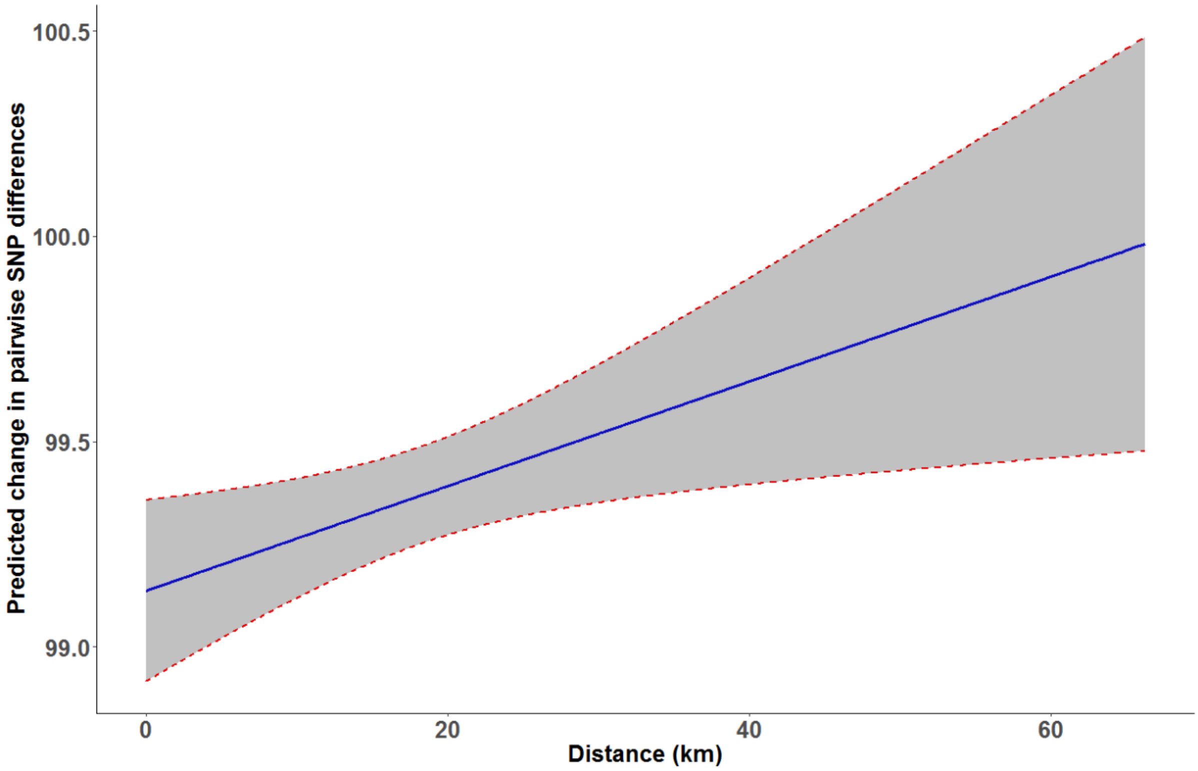


Figure 4.30: Effect of distance on variations in the number of SNP differences in *surf*$_{4.2}$ gene between *P. falciparum* parasite pairs.

The analysis was carried out on 754 samples collected from children admitted to the Kilifi County Hospital between 1995 and 2014. The relationship between the number of SNP differences and spatial distance between samples was modelled in a multiple fractional polynomial regression. The relationship was not statistically significant (P= 0.38; 95% CI: -0.052 – 0.076) based on a bootstrapped linear regression analysis with 1000 resampling steps. The blue line represents the estimated change in the number of pairwise SNP differences over distance and the grey shaded regions represent the 95% confidence intervals, with the lower and upper bounds defined by the red dotted lines.

Using the IBS data instead of the SNP differences data, increasing time between sampling was shown to correlate with decreasing IBS in the *PfAMA1* dataset (effect size = -0.17), i.e., the longest shared sequence length between any parasite pair decreased as the time between sampling of the parasites increased (figure 4.31). However, this effect was seen only in the first 3 years, after which IBS between parasite pairs began to increase as sampling time between the parasites increased. Bootstrapping the analysis gave statistically significant effect of time on IBS (P = 0.03; 95%CI: -0.345 - 0.006). This observation of decreasing IBS over time was inversely correlated with increasing number of SNP differences between the same parasite pair.

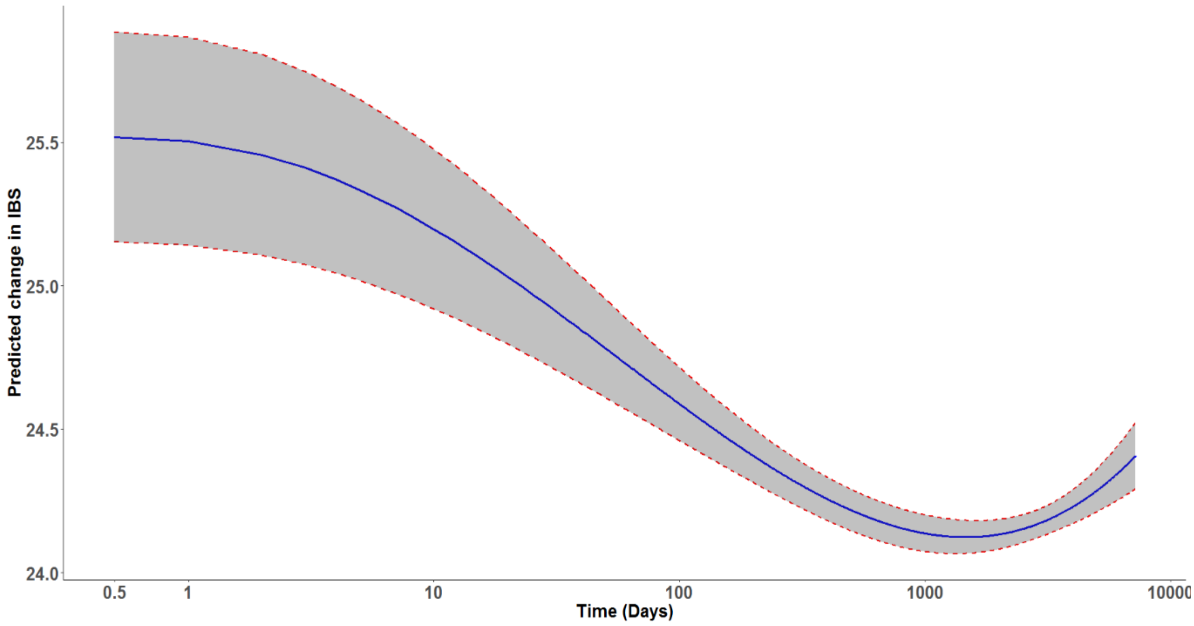


Figure 4.291: Effect of time on IBS in *AMA1* sequences between *P. falciparum* parasite pairs.

The analysis was carried out on 966 samples collected from children admitted to the Kilifi County Hospital between 1995 and 2014. The relationship between the number of SNP differences and time between sampling was modelled in a multiple fractional polynomial regression analysis. The relationship was statistically significant (P=0.03; 95% CI: -0.345 - 0.006) based on a bootstrapped linear regression analysis with 1000 resampling steps. The blue line represents the estimated change in IBS over time and the grey shaded regions represent the 95% confidence intervals, with the lower and upper bounds defined by the red dotted lines.

Interrogating the effect of distance on variation in IBS among the *PfAMA1* sequences gave a negative association (effect size = -0.102) that was not statistically significant (p = 0.3; 95% CI: -0.537 – 0.341) (figure 4.32).

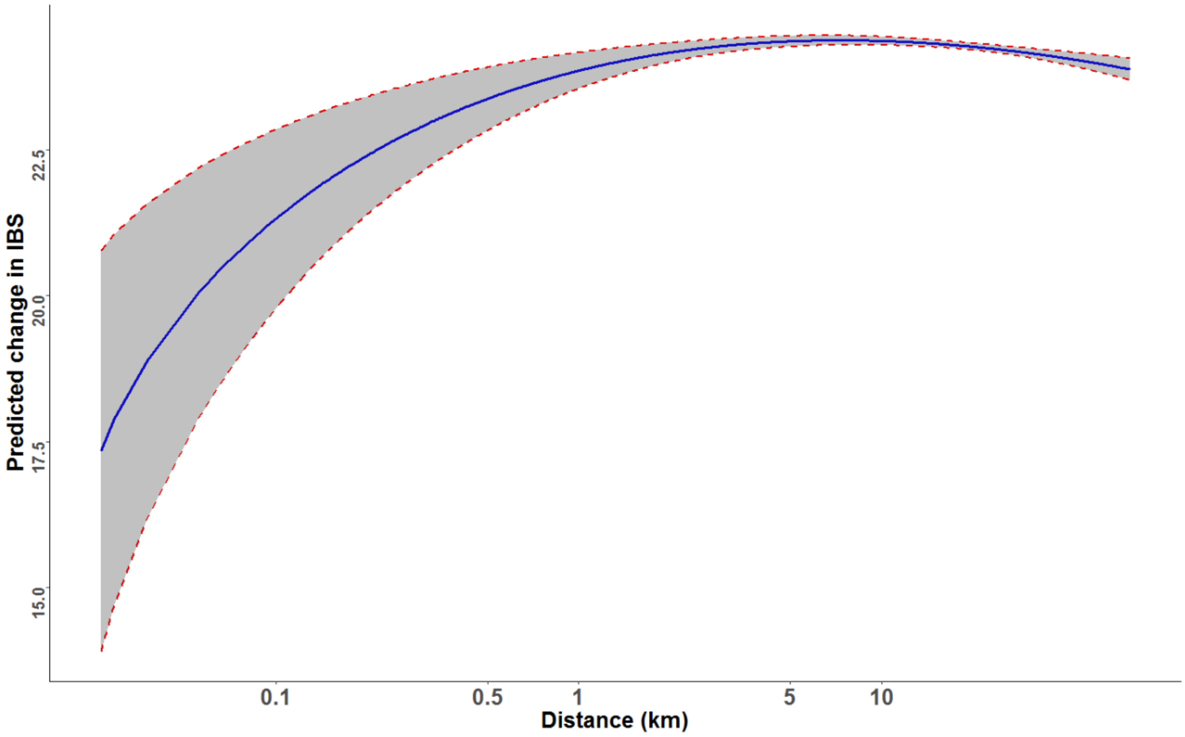


Figure 4.302: Effect of distance on IBS in *AMA1* sequences between *P. falciparum* parasite pairs.

The analysis was carried out on 795 samples collected from children admitted to the Kilifi County Hospital between 1995 and 2014. The relationship between the number of SNP differences and distance between sample collection was modelled in a multiple fractional polynomial regression analysis. The relationship was not statistically significant (p=0.3; 95% CI: -0.537 – 0.341) based on a bootstrapped linear regression analysis with 1000 resampling steps. The blue line represents the estimated change in

IBS over distance and the grey shaded regions represent the 95% confidence intervals, with the lower and upper bounds defined by the red dotted lines.

In the *surf*$_{4.2}$ dataset, IBS was positively correlated with time (effect size = 0.1812), although this observation was not statistically significant (P = 0.18; 95% CI: -0.276 – 0.665) (figure 4.33). Increasing distance between sampling was associated with decreasing IBS among the *surf*$_{4.2}$ sequences (effect size = -0.030), although this effect was also not statistically significant (P = 0.52; 95% CI: -1.156 – 0.982) (figure 4.34).

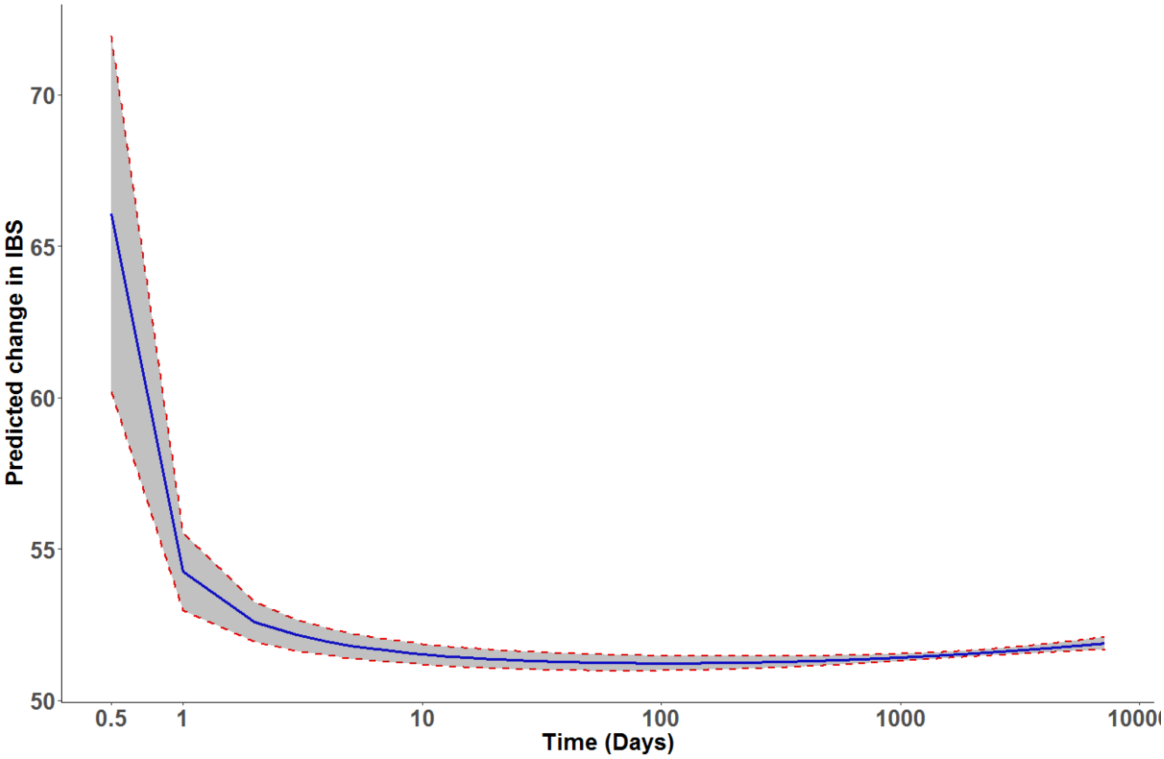


Figure 4.31: Effect of time on IBS in *surf*$_{4.2}$ sequences between *P. falciparum* parasite pairs.

The analysis was carried out on 890 samples collected from children admitted to the Kilifi County Hospital between 1995 and 2014. The relationship between the number of SNP differences and distance between sample collection points was modelled in a multiple fractional polynomial regression. The relationship was not statistically significant (P = 0.18; 95% CI: -0.276 – 0.665) based on a bootstrapped linear regression analysis with 1000 resampling steps. The blue line represents the estimated change in IBS over time and the grey shaded regions represent the 95% confidence intervals, with the lower and upper bounds defined by the red dotted lines.

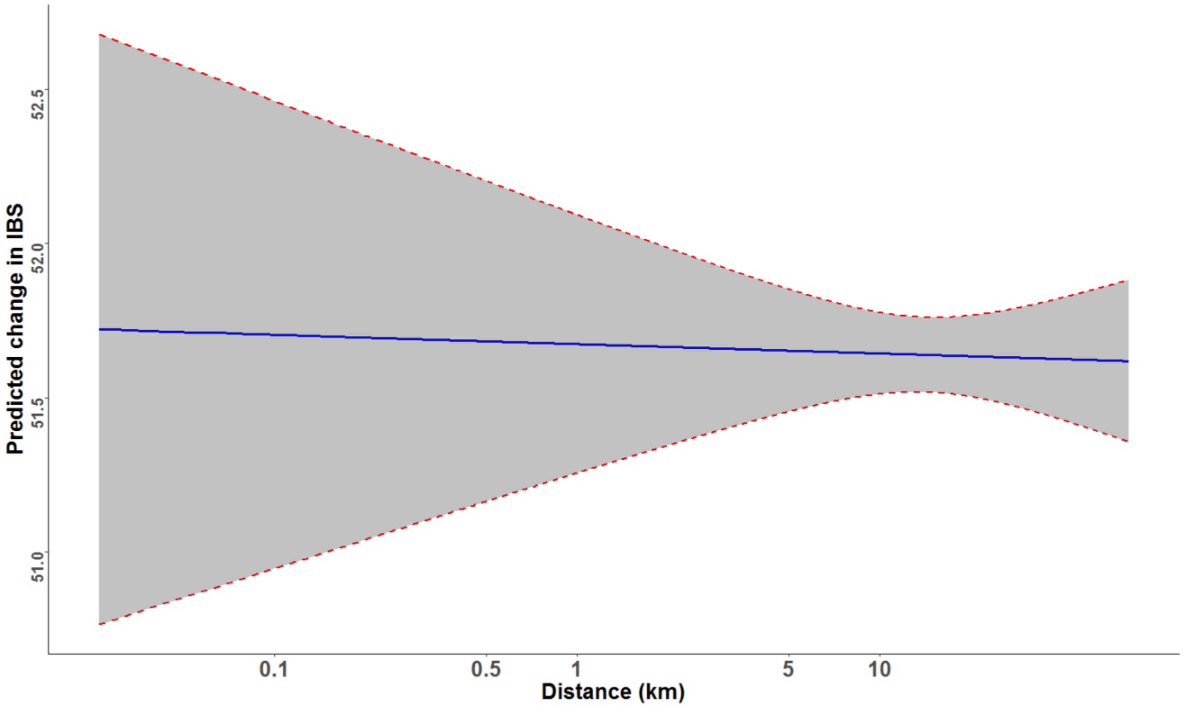


Figure 4.34: Effect of distance on IBS in *surf*$_{4.2}$ sequences between *P. falciparum* parasite pairs.

The analysis was carried out on 890 samples collected from children admitted to the Kilifi County Hospital between 1995 and 2014. The relationship between the number of SNP differences and distance between sample collection points was modelled in a multiple fractional polynomial regression. The relationship was not statistically significant (p=0.52; 95% CI: -1.156 – 0.982) based on a bootstrapped linear regression analysis with 1000 resampling steps. The blue line represents the estimated change in IBS over distance and the grey shaded regions represent the 95% confidence intervals, with the lower and upper bounds defined by the red dotted lines.

Table 4-7: Linear effects of time and distance on changes in the number of SNP differences and IBS between *P. falciparum* parasites based on *AMA1* and *surf*$_{4.2}$ sequences.

Gene	Data type	Time effect size (95% CI); **p value**	Distance effect size (95% CI); **P value**
***Pf*AMA1**	SNPs	6.338x10^{-5} (2.3x10^{-5} – 1.11x10^{-4}); **0.002**	0.006 (-0.003 – 0.017); **0.082**
	IBS	-0.17 (-0.345 – 0.006); **0.03**	-0.102 (-0.537 – 0.341); **0.3**
***surf*$_{4.2}$**	SNPs	8.544x10^{-5} (-2.09x10^{-4} – 4.05x10^{-4}); **0.3**	0.013 (-0.0519 – 0.076); **0.38**
	IBS	0.1812 (-0.276 – 0.665); **0.18**	-0.030 (-1.156 – 0.982); **0.52**

The effects of time's interaction with distance on genetic variation in *PfAMA1* and *surf*$_{4.2}$ were also investigated in this study. In addition to analysing the rate of change in parasite genotypes within the study site during the entire study period using the whole dataset, separate analyses were also carried out for samples collected within 3 years of each other, samples collected 4 – 10 years apart, and samples collected more than 10 years apart. The latter analyses were conducted to determine whether the rate of change in genetic variation was similar throughout the study period or if it varied over time.

Based on the SNP differences data of both genes, time and distance interacted to increase the number of SNP differences between parasite pairs (figure 4.35 and figure 4.36). When the whole dataset was analysed, time was shown to interact antagonistically with distance to attenuate the effect of distance on genotype relatedness, and the effect was significant for *PfAMA1* (effect size = 0.164, $p< 0.001$), but not for *surf*$_{4.2}$ (effect size = 0.079, $p = 0.456$). The results also show that the *PfAMA1* gene accumulates changes quite slowly in this parasite population, with an estimated average of 0.704 SNP differences for parasite pairs collected over the 20-year study period. Analysis of the different subsets of the data showed that most of the SNP differences seen among the sequences were present in parasite pairs that were closely spaced in time, specifically those collected within a month of each other (figure 4.35b and figure 4.36b). However, these effects were not statistically significant for either *PfAMA1* (effect size = 0.0676, p=0.203) or *surf*$_{4.2}$ (effect size = 0.1745, p=0.353). The other data subsets of parasite pairs separated by 4 – 10 years or more than 10 years had very few SNP differences between them, with an estimated average change of 0.24 SNP differences in *PfAMA1* sequences from parasites separated by 4 – 10 years, and 1.022 SNP differences for parasites pairs collected more than 10 years apart (figures 4.35c – d). Among the *surf*$_{4.2}$ genes, parasite pairs separated by between 4 – 10 years had an estimated average change of 0.898 SNP differences while those collected more than 10 years apart had an estimated average change of 1.468 SNP differences (figure 4.36c – d).

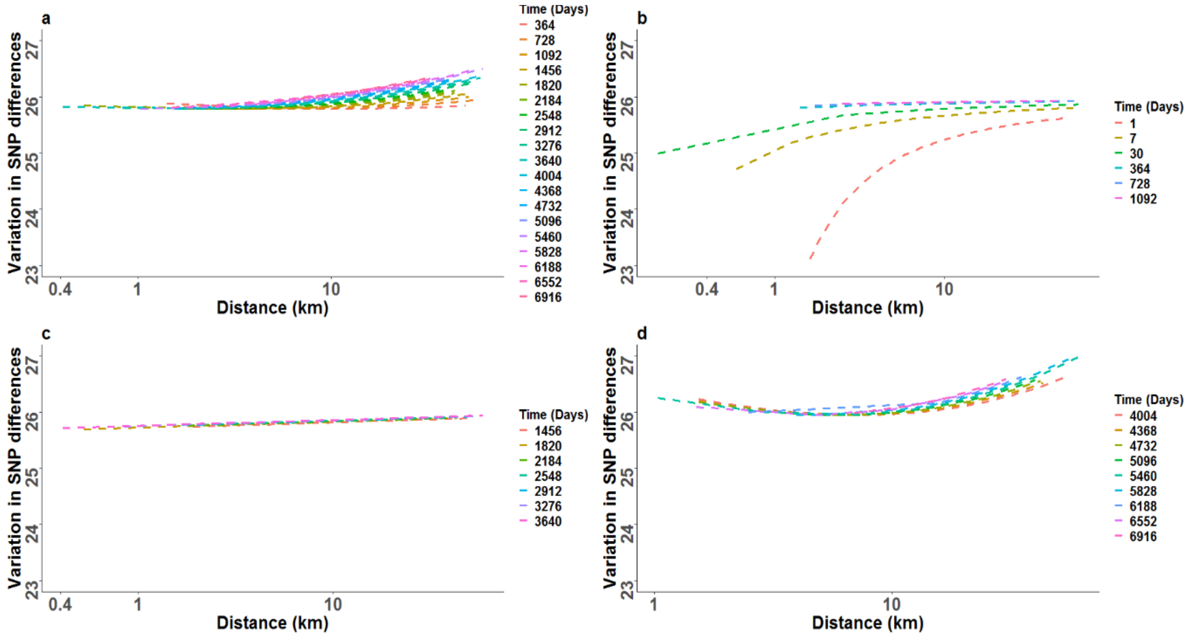


Figure 4.35: Effect of time-distance interaction on the number of SNP differences in *AMA1* sequences between *P. falciparum* parasite pairs.

The analysis was carried out on 795 samples collected from children admitted to the Kilifi County Hospital between 1995 and 2014. Dashed lines represent time intervals separating parasite pairs. The analysis was carried out for (**a**) parasite pairs in the whole dataset, (**b**) parasites collected up to three years apart, (**c**) parasite pairs collected between 4 and 10 years apart, and (**d**) parasite pairs collected more than 10 years apart.

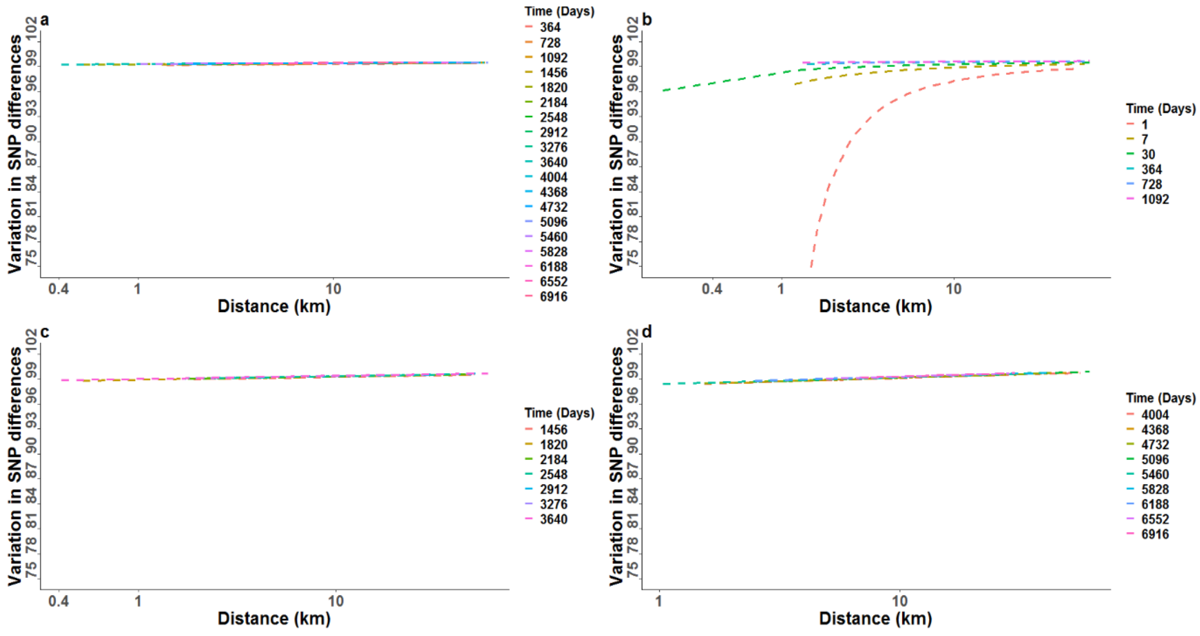


Figure 4.36: Effect of time-distance interaction on the number of SNP differences in *surf*$_{4.2}$ sequences between *P. falciparum* parasite pairs.

The analysis was carried out on 754 samples collected from children admitted to the Kilifi County Hospital between 1995 and 2014. Dashed lines represent time intervals separating parasite pairs. The analysis was carried out for (**a**) parasite pairs in the whole dataset, (**b**) parasites collected up to three years apart, (**c**) parasite pairs collected between 4 and 10 years apart, and (**d**) parasite pairs collected over more than10 years apart.

Analysis of the IBS data for both genes showed that time’s interaction with distance had the effect of reducing the longest contiguous stretch of sequence that was shared between parasites when *PfAMA1* was considered, although this interaction was not statistically significant (effect size = -0.1175, p=0.106) (Table 4.8). Among *surf*$_{4.2}$ sequences, time interacted with distance to increase IBS, meaning that parasites became more similar over time, although this observation was also not statistically significant (effect size = 0.0579, p=0.37). Like the observations made when analysing the SNP differences data, most of the changes in IBS were found among parasites collected within a month of each other (figures 4.37 and figure 4.38), with fewer changes between parasites separated by longer time intervals. Analysis of the different parasite subsets showed little differences among parasites separated by 4 – 10 years (1.17) and those separated by more than 10 years (0.177) when *PfAMA1* sequences were analysed. Similar results of changes in IBS were also observed when *surf*$_{4.2}$ sequences were analysed in parasites separated by 4 – 10 years (0.042) and more than 10 years (1.254).

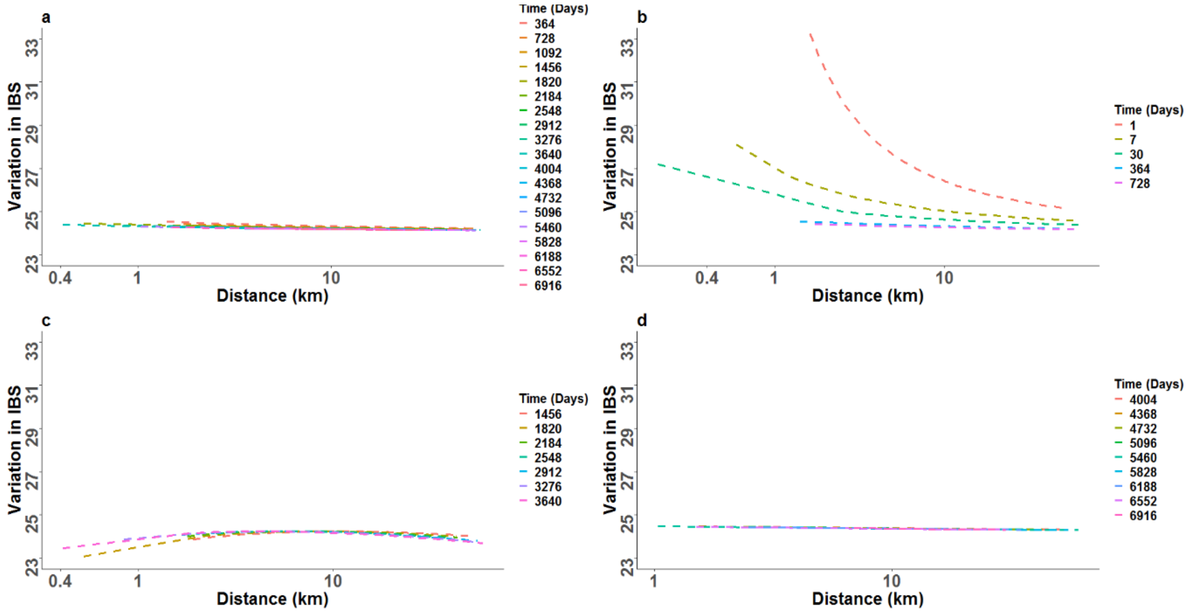


Figure 4.32: Effect of time-distance interaction on IBS in *AMA1* sequences between *P. falciparum* parasite pairs.

The analysis was carried out on 795 samples collected from children admitted to the Kilifi County Hospital between 1995 and 2014. Dashed lines represent time intervals separating parasite pairs. The analysis was carried out for (**a**) parasite pairs in the whole dataset, (**b**) parasites collected up to three years apart, (**c**) parasite pairs collected between 4 and 10 years apart, and (**d**) parasite pairs collected over 10 years apart.

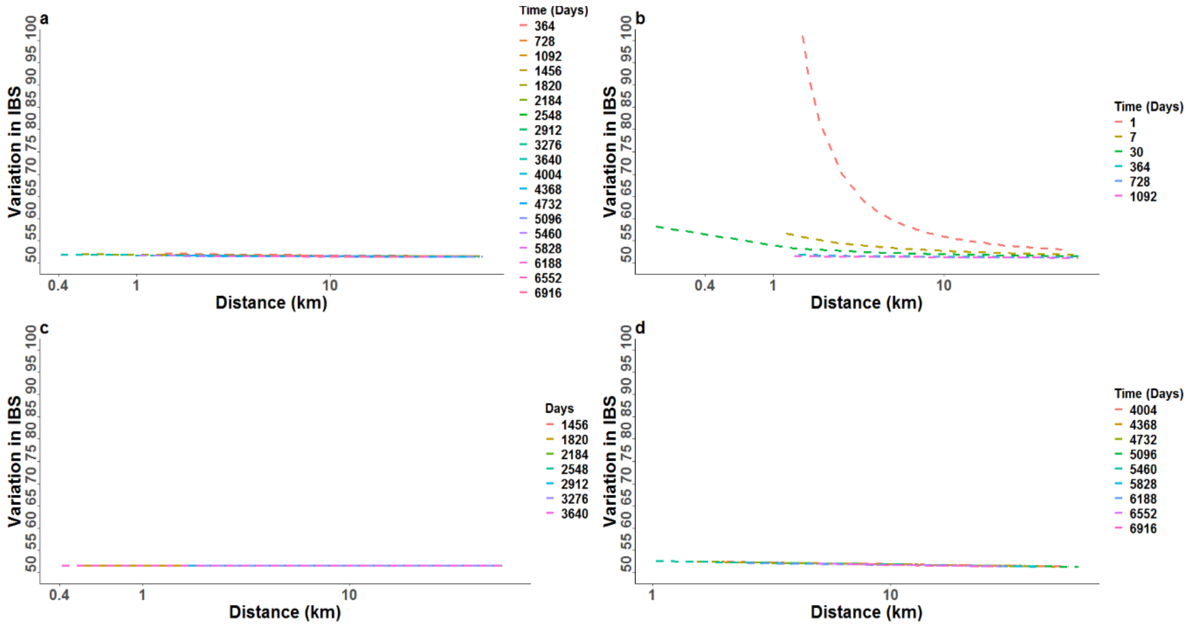


Figure 4.33: Effect of time-distance interaction on IBS in *surf$_{4.2}$* sequences between *P. falciparum* parasite pairs.

The analysis was carried out on 754 samples collected from children admitted to the Kilifi County Hospital between 1995 and 2014. Dashed lines represent time intervals separating parasite pairs. The analysis was carried out for (**a**) parasite pairs in the whole dataset, (**b**) parasites collected up to three years apart, (**c**) parasite pairs collected between 4 and 10 years apart, and (**d**) parasite pairs collected more than 10 years apart.

Table 4-8: Linear effects of the interaction of time and distance on *P. falciparum* genetic variation based on SNP differences and IBS.

GENE	Data type	Dataset	Effect size (95% CI)	P value
PfAMA1	**SNPs**	whole	0.164 (0.063 – 0.288)	<0.001
		0 – 3 years	0.068 (-0.885 – 0.228)	0.203
		4 – 10 years	0.104 (-0.113 – 0.470)	0.113
		> 10 years	0.441 (-0.098 – 0.920)	0.054
	IBS	whole	-0.118 (-0.328 – 0.073)	0.106
		0 – 3 years	-0.282 (-0.675 – 0.045)	0.056
		4 – 10 years	-0.084 (-0.646 – 0.345)	0.287
		> 10 years	-0.100 (-1.100 – 0.753)	0.388
Surf$_{4.2}$	**SNPs**	whole	0.079 (-0.563 – 0.789)	0.456
		0 – 3 years	0.175 (-0.699 – 0.997)	0.353
		4 – 10 years	0.379 (-1.713 – 1.757)	0.527
		> 10 years	0.843 (-1.481 – 5.767)	0.104
	IBS	whole	0.058 (-0.434 – 0.553)	0.37
		0 – 3 years	0.055 (-0.777 – 0.774)	0.437
		4 – 10 years	0.018 (-1.025 – 1.477)	0.36
		> 10 years	-0.720 (-3.630 – 1.089)	0.185

IBS = Identity by state, representing the longest shared contiguous sequence length between parasite pairs; SNPs = single nucleotide polymorphisms.

4.4 DISCUSSION

This chapter aimed to determine the spatio-temporal genetic variation in *P. falciparum* field isolates collected over a 20-year period in Kilifi County, using sequences of genes encoding two antigens, apical membrane antigen 1 (*PfAMA1*) and SURFIN$_{4.2}$, by measuring two complementary metrics of genetic variation: pairwise number of SNP differences as a proxy for genetic diversity, and identity by state as a proxy for genetic similarity. The study represents the single largest analysis of sequences from these two genes in *P. falciparum* parasites from a single geographical location.

Sequence analysis showed high levels of genetic diversity in both *PfAMA1* and *surf$_{4.2}$* genes, with most of this diversity localised to the ectodomain region of *PfAMA1* and the variable region of *surf$_{4.2.}$* Similarly high genetic diversity among *PfAMA1* sequences have been reported in previous studies of global *P. falciparum* isolates, including those from Mali (Takala *et al*., 2009), Nigeria (Polley and Conway, 2001), Kenya (Osier *et al*., 2010), Papua New Guinea (Arnott *et al*., 2014), Thailand (Polley *et al*., 2003), the China-Myanmar border (Zhu *et al*., 2016) and Venezuela (Ord *et al*., 2008). Although *surf*$_{4.2}$ has not been studied in detail, high genetic diversity has been noted in previous analyses of Kenyan (Ochola *et al*., 2010) and Thai (Kaewthamasorn *et al*., 2012) *P. falciparum* populations. For *surf*$_{4.2}$, the average pairwise nucleotide diversity between parasite pairs in the current study (π=0.045) was similar to that from a previous analysis of 69 samples from the Kilifi population (π=0.043) (Ochola *et al*., 2010) and 74 samples from the Thai population (π=0.044) (Kaewthamasorn *et al*., 2012).

The degree of genetic diversity identified among *PfAMA1* sequences in this study was similar to that identified in other African sites but was higher than that seen in southeast Asian or South American parasite populations (Zhu *et al*., 2016), and is in agreement with the higher *P. falciparum* transmission intensity experienced in Africa compared to southeast Asia and South America.

The high level of genetic diversity was also reflected in the number of haplotypes circulating in the region during the study period, with more than 600 distinct haplotypes identified when each gene was analysed separately. Previous studies have identified a large number of *P. falciparum* haplotypes based on *AMA1*, including 229 haplotypes when a global set of 956 sequences were analysed (Zhu *et al*., 2016), and 214 haplotypes when a set of 506 sequences from one Malian village were analysed (Takala *et al*., 2009). Our study currently represents the largest number of *PfAMA1* and *surf*$_{4.2}$ haplotypes identified to be circulating in a single study site, and points to the existence of greater genetic diversity than previously thought in this parasite population. There were no dominant haplotypes identified in this parasite population, and in fact most of the haplotypes were unique to individual samples. This is similar to a study of *P. falciparum* populations in a Malian village (Takala *et al*., 2009), and indicates that most genetic variations are rare, particularly in African parasite populations. An excess of rare genetic variants in African parasites has been shown in previous studies using whole genome sequence data (MalariaGEN, 2016, Manske *et al*., 2012).

The genetic diversity observed within the *surf*$_{4.2}$ gene was similar to that seen in the Thai parasite population (Kaewthamasorn *et al*., 2012), despite differing transmission intensities in the two sites. The highest level of genetic diversity was observed in the variable region, in agreement with studies of other parasite populations (Kaewthamasorn *et al*., 2012, Ochola *et al*., 2010). Two positions, each with a 3 base-pair deletion, were identified in a subset of the *surf*$_{4.2}$ sequences analysed here. These deletions have been identified in a previous study of isolates from the same region (Ochola *et al*., 2010), but were absent when the same gene was analysed in a Thai parasite population (Kaewthamasorn *et al*., 2012), indicating that the gene may be evolving under different selective pressures in the two populations.

Tajima's D was computed to test for departure from neutrality among sequences of both genes. The entire sequenced region, the ectodomain, as well as the individual domains of *Pf*AMA1 were associated with positive, albeit not significant Tajima's D values, meaning that the null hypothesis of random evolution of the genes could not be rejected. This is in contrast to multiple other studies that have shown the ectodomain, and particularly domains I and/or III of *PfAMA1* to be under balancing selection in different parasite populations (Osier *et al*., 2010, Polley *et al*., 2003, Polley and Conway, 2001, Cortes *et al*., 2003, Ord *et al*., 2008), although not in a parasite population along the China-Myanmar border (Zhu *et al*., 2016). Similarly, unlike previous studies showing evidence of selection in the *surf*$_{4.2}$ gene in Kilifi (Ochola *et al*., 2010) and Thai (Kaewthamasorn *et al*., 2012) parasite populations, none of the tests were associated with evidence of selection in the current study despite the large sample sizes used. Previous studies of *P. falciparum* populations using whole genome sequence data show that a majority of genes in the parasite genome are associated with negative Tajima's D values. For example, in a study of over 4000 genes in 152 *P. falciparum* isolates from The Gambia and Republic of Guinea, only 2.5% had positive Tajima's D values associated with balancing selection, while most of the remaining genes were associated with negative values (Mobegi et al., 2014). In a separate study, similar results were obtained in an analysis of over 2800 genes in 65 *P. falciparum* isolates from The Gambia (Amambua-Ngwa et al., 2012). These results are reflective of the historical population expansion that occurred in this parasite, especially in African populations. Thus, even in the absence of statistical significance, the highly positive Tajima's D values observed in this study provides strong evidence of balancing selection, in line with previous studies that have analysed the same genes. Apart from balancing selection, the positive Tajima's D values observed may also be interpreted in terms of population demographics, signifying a decrease in the population size (population bottleneck) resulting from the declining transmission intensity that has been observed in this parasite population during much of the study period (Mogeni *et al*., 2016).

Linkage disequilibrium (LD) was computed in a pairwise comparison for all diallelic sites in each gene. LD was detected across the *PfAMA1* sequence, although it was shown to decline rapidly as the distance between nucleotide pairs increased, and most of the significant associations were observed for nucleotide pairs that were less than 600 bases apart. These results are consistent with those of previous studies which identified strong LD between closely spaced nucleotide sites in the *PfAMA1* gene in other parasite populations in Africa and Southeast Asia (Polley *et al*., 2003, Polley and Conway, 2001, Zhu *et al*., 2016, Osier *et al*., 2010). LD was also observed among *Surf$_{4.2}$* SNPs, with high LD values for nucleotide sites that were less than 1000 bases apart, although LD was also observed for some sites that were more distantly spaced. These results are similar to those from an earlier study of a smaller sample set from the same Kilifi population (Ochola *et al*., 2010). Previous analysis of LD in *surf*$_{4.2}$ from a Thai parasite population showed evidence of LD for SNPs that were more than 1.5kb apart (Kaewthamasorn *et al*., 2012). The extent of LD in *P. falciparum* has been shown to vary based on malaria endemicity, with little or no LD in high transmission areas, and stronger LD in low transmission areas (Anderson *et al*., 2000, Conway, 2007). Using microsatellites data, Anderson and his colleagues showed that there was strong LD in regions with low prevalence (< 1%) and little or no LD in regions with higher transmission intensities (Anderson *et al*., 2000). In high transmission areas, strong LD has been shown to exist generally between nucleotide sites that are less that 1kb apart, with rapid decay to very low levels for nucleotide sites beyond this distance (Polley *et al*., 2003, Polley and Conway, 2001). This pattern of LD is attributed to the multiple mixed infections found in high transmission areas, which increase chances of recombination between gametes of different genotypes during sexual replication in mosquitoes. On the other hand, regions with low transmission intensity have fewer parasite genotypes and thus fewer mixed infections, meaning that selfing between parasites with the same genotype is more common (Anderson *et al*., 2000, Conway, 2007).

Recombination was shown to occur often in this parasite population, based on the rapid decline in LD with increasing distance between nucleotide sites and the high estimates of the recombination parameter, C, for both genes. In the Thai population, a minimum of 35 recombination events were estimated to have occurred in the extracellular region of *surf*$_{4.2}$ (Kaewthamasorn *et al*., 2012), while this value was much higher in the current study, at a minimum of 177 recombination events among the sequences. When the cysteine rich domain and variable regions of the gene were considered separately, the variable region was shown to have a lower number of recombination events, an observation also seen among the Thai isolates. This lower number of recombination events in the variable region is likely to be due to the multiple polymorphisms observed in this region which prevents effective recombination between sequences that are too diverse (Kaewthamasorn *et al*., 2012). The *Pf*AMA1 sequence analysed in this study was estimated to have undergone an estimated 47 recombination events, which was higher than that estimated for a different parasite population from the same area (Osier *et al*., 2010), as well as a parasite population in Nigeria (Polley and Conway, 2001). Recombination rate is closely linked with transmission intensity, and is highest in Africa, followed by Southeast Asia and finally South America (Zhu *et al*., 2016, Ord *et al*., 2008, Anderson *et al*., 2000). The high recombination and rapid decline in LD observed in this study are indicative of high meiotic recombination within this parasite population. These results also indicate a high degree of parasite mixing within the study site and concurs with the observations of high mixing among parasites from this population reported in chapter 2 (Omedo *et al*., 2017a).

Heatmaps displaying visual representations of genetic diversity showed a lack of obvious clustering of parasite genotypes in time and space using nucleotide and amino acid sequences of both *PfAMA1* and *surf*$_{4.2}$. This points to a high level of parasite mixing in the study area, and is further supported by the little to no statistical support observed by bootstrapping the hierarchical clustering of both genes.

This high level of mixing and an inability to resolve parasite genotypes into allele clusters has previously been shown using *PfAMA1* sequences. For example, in an analysis of *P. falciparum* isolates from Kilifi, no clustering was observed among parasites collected from children with asymptomatic, mild or severe malaria (Osier *et al*., 2010) and in western Kenya, parasites did not cluster based on *PfAMA1* genotypes either (Escalante *et al*., 2001). In previous analyses of samples from the same western Kenya region, clustering was not observed among parasite genotypes at a subnational scale when analysing SNP genotype data (Omedo *et al.,* 2017b), although clusters of parasite sub-populations were identified at micro-epidemiological scales of 10km and less (Omedo *et al*., 2017a). However, these clusters were weak and consisted of parasites with different genotypes. Weak clusters of parasite genotypes have also been identified in the Kilifi parasite population based on analysis of the S-antigen (Kyes *et al*., 1997). On a continental level, however, parasite populations have been shown to cluster into distinct groups. An analysis of a global set of *P. falciparum* populations clustered parasites based on *AMA1* haplotypes according to their geographical regions of origin, although some haplotypes were shown to cluster with parasites from a different area (Duan *et al*., 2008). Using whole genome sequence data, *P. falciparum* parasite populations have been shown to cluster into distinct groups along continental lines (Manske *et al*., 2012). However, there has generally been lower clustering of parasites at the regional level when African parasites were analysed. For example, low clustering among samples in individual countries was observed using a set of 12 microsatellite markers, which showed little or no differentiation among parasites that were separated by up to 2000km across Africa (Anderson *et al*., 2000). Mobegi and others showed similar results using microsatellite data to analyse parasite populations across four West African countries (Mobegi et al., 2012). Similar results were observed when SNP data were used across Africa, as well as within West Africa (Campino et al., 2011, Mobegi et al., 2014).

Neither *PfAMA1* nor *surf*$_{4.2}$ sequences clustered based on relatedness over time, and most of the haplotypes were present at only one timepoint. This is similar to a study of parasite populations from Mali, where sequences of parasites collected at one time point did not overlap with those of parasites collected at two other time points (Duan *et al*., 2008), as well as to temporal analyses carried out on *PfAMA1* parasites from the China-Myanmar border (Zhu *et al*., 2016). The temporal mixing of parasites observed among the *surf*$_{4.2}$ sequences in our study are in contrast to the analysis of the same gene in Thai parasites, where the same haplotypes were seen to be circulating up to 14 years apart (Kaewthamasorn *et al*., 2012). This difference in results may be explained by the lower transmission intensity experienced in Thailand, where selfing of the same parasite clones is common compared to the higher transmission setting in Kilifi where recombination among different parasite clones is more common (Anderson *et al*., 2000, Conway, 2007, Hartl *et al*., 2002).

Among the *surf*$_{4.2}$ sequences, polymorphic SNPs were concentrated in the variable region, and could be clustered based on the three distinct patterns in this region. Similar clusters were observed in a Thai parasite population when this gene was analysed (Kaewthamasorn *et al*., 2012). The functional significance of these clusters is currently unknown. In our analysis, specific SNP patterns were not restricted to parasites from a specific year or location, but were instead found in sequences from different time points and different geographical locations.

Although SNP genotype data are often used to study genetic diversity in parasite populations, these can be affected by ascertainment bias since most SNPs chosen for genotyping are usually present at intermediate to high frequency, and it therefore biases against rare SNPs. For this reason, the conclusions reached regarding demographic history and natural selection of the parasite population may be very different depending on whether one uses genotype or sequencing data (Lachance and Tishkoff, 2013). To confirm that the observations made in this study about parasite mixing and movement over the study sites were not simply due to the SNP subsets used but were a true reflection of the underlying population genetic dynamics, the genes encoding two important *P. falciparum* antigens: *AMA1* and SURFIN$_{4.2}$ were sequenced on the sanger sequencing platform. Two different but complementary metrics of parasite diversity: SNP differences and identity by state (IBS), were then measured. IBS is used in genetics to describe two identical alleles or sequences of DNA. It is different from the more commonly measured metric, identity by descent (IBD), in that IBD refers to similarity between a pair of DNA sequences due to common ancestry, whereas IBS does not consider a common ancestry between the parasites. IBS was used in this study because of the observed high level of recombination and high parasite mixing which makes it difficult to accurately determine IBD.

Temporal and spatial genetic variation in the parasite population was measured using the number of pairwise SNP differences and IBS to determine whether the two metrics were in concordance. The metrics showed similar results of variations in parasite genotypes when *PfAMA1* and *surf*$_{4.2}$ were analysed over time, with the number of pairwise SNP differences increasing as IBS declined over time. Similarly, SNP differences increased and IBS decreased as the distance separating parasite pairs increased. This inverse relationship is expected because as the number of SNP differences increases between a parasite pair, IBS, represented by the longest contiguous stretch of sequence shared between the parasite pair, is expected to decline. This is because mutations introduce variations in nucleotide sequences, increasing SNP differences, and recombination leads to the break-up and exchange of segments of DNA, reducing the length of the shared stretch of sequence between parasite pairs. The decrease in IBS over time further shows that recombination and mutation act across the entire sequence in both genes, and is not confined to specific locations in the sequences. However, our analysis of the recombination parameter across both genes showed a lower rate of recombination in the highly variable regions, similar to observations made in the *surf*$_{4.2}$ gene in parasites from Thailand, indicating that recombination occurs much less efficiently in regions of high variability.

Among the *PfAMA1* sequences, the increase in genetic variation over time was shown to be statistically significant. This was similar to our previous analysis of parasites from this site, as well as parasites from two other locations in Kenya and The Gambia

(Omedo *et al.*, 2017a). This increase in genetic diversity over time indicates that, to a large extent, genetic drift is driving the evolution of the gene in this population, with the number of SNPs increasing over time, and supports our observation of non-statistical significance in the tests of neutrality. The results of the changes in genetic variation was corroborated by IBS data, which showed statistically significant reduction in genetic variation over time. However, the pattern of increasing genetic diversity over distance observed in the *PfAMA1* sequences was not statistically significant, in contrast to our earlier analysis which showed statistically significant increase in variation when a set of genome-wide distributed SNPs were used (Omedo *et al.*, 2017a). Concordant trends of decreasing, albeit non-significant genetic variation over distance were noted when IBS data were used instead of SNP data.

Among the *surf*$_{4.2}$ sequences, parasites showed increased genetic variation over both time and distance when SNP differences data were used. However, neither of these observations were statistically significant. When IBS data were used, genetic variation between parasite pairs was shown to decrease over both time and space, although these decreases were also not statistically significant. Furthermore, time was shown to interact antagonistically with distance to affect changes in genetic variation between parasite pairs. Here, genetic variations between parasite pairs increased with distance between the parasites, but this increase was rapidly attenuated as the time separating parasites also increased, and was gone within one year. Analysing subsets of parasites collected within different temporal ranges showed that most of the genetic variation was found among parasites collected within a few days of each other (1 – 30 days), and there were barely any differences between samples that were collected more than one year apart from each other. In fact, beyond one year, few SNP differences were observed between sequences, indicating that there were few genetic variations acquired over time during the 20-year study period. These same patterns were observed when both *PfAMA1* and *surf*$_{4.2}$ genes were analysed.

These results are similar to observations made in the study of parasite pairs from Kilifi, Rachuonyo South and The Gambia that was reported in chapter two, which used SNP genotype data to show a rapid decline in changes in genetic variation with increasing distance when time was taken into account (Omedo *et al*., 2017a). As in the current analysis, most of the genetic variations between parasites observed in that study were observed between parasite pairs that were collected within a month of each other, and most of the genetic variation was gone after one year.

When IBS data were used in place of SNP differences data, a pattern complementary to that observed when using SNP differences emerged. Time interacted antagonistically with distance, leading to a reduction in IBS as the distance separating samples increased. However, this reduction was attenuated over time, and was gone within one year. These observations using both SNP and IBS data show that there is substantial gene flow within the study site such that distance no longer predicts genetic variation for *P. falciparum* parasites collected more than 1 year apart.

The findings in this study have several implications regarding the outcomes of malaria control programmes. As shown in chapters two and three, *P. falciparum* parasites mix to high degrees within Kilifi county, thus targeting control interventions to transmission hotspots within the region is likely to lead to a reduction in transmission in un-targeted, surrounding areas. Unfortunately, as noted in those other studies as well, the high mixing of parasites within the study site means that there is a high chance of importation of infection from untargeted to targeted regions. The analysis of temporal variations in genetic diversity over a 20-year period shows that gene flow between and among parasites occurs all the time in this parasite population, and is not due to historical demographic events that occurred decades or centuries earlier. This further supports our conclusion of a well-mixed parasite population and is supported by our previous analysis of parasite pairs from the same site, which showed that parasites collected from nearby homesteads had fewer SNP differences between them than parasite pairs that were further apart, although this distance gradient was attenuated over a month and was gone within a year.

This study had some limitations. First, some samples lacked exact geospatial (homestead) positioning data and their geographical locations had to be aggregated at a location or sub-location level. This may have led to fine-scale patterns of genetic variation over distance being missed. However, we used imputed location data in the analysis reported in chapter two and we were able to identify genetic variation and population structure at fine-scales (Omedo *et al*., 2017a). A similar data imputation technique was used in this study, thus increasing our confidence in being able to pick up the fine scale spatial patterns of genetic variation. Second, only a small number of samples (between 34 – 59) were sequenced in each year, thus limiting our ability to detect genetic changes and selection between parasites in a year-by-year analysis. Additional sequences are needed from each of the years analysed in order to provide a complete picture of temporal genetic variation in this parasite population. Third, the analyses were carried out on two genes which are expressed on the surface of the merozoite and are under host immune pressure. This selection pressure may have influenced the spatio-temporal pattern of genetic variation observed. Inclusion of a neutral gene in the analysis to determine whether the same patterns of parasite mixing and changes in genetic variation are observed in the neutral gene would provide an additional measure of validation. However, in our previous analyses (chapter two), we observed similar results when we analysed genome-wide distributed SNPs including a large number of SNPs from neutral genes, and when we separately analysed SNPs typed in *AMA1* and erythrocyte binding antigen 175 (*EBA175*), showing that similar patterns of changes in genetic variation are observed regardless of whether “neutral” or immune system-selected genes are analysed.

In conclusion, I have used sequence data derived from genes encoding two important antigens to show that there was a large repertoire of genetic diversity among parasites circulating in Kilifi county between 1995 and 2014. I have also shown that there is a high level of meiotic recombination in the two genes based on the high number of recombination events and rapidly declining LD observed among the sequences. A high degree of parasite mixing was observed within the study area, with little or no clusters of parasite genotypes in time and space, further supporting the previous observation of no geographical barriers to parasite movement within this study site (Omedo *et al*., 2017a). Additionally, similar patterns of temporal and spatial genetic variation of *P. falciparum* parasites were identified using pairwise SNP differences and IBS data, meaning that either metric can be used to study population genetics of *P. falciparum* parasites and further showing that there was no ascertainment bias in the selection of the subset of genome-wide distributed SNPs that were typed in chapters two and three. Time’s effect of attenuating genetic variation with distance shows that parasites mix rapidly within the study site, with little genetic variation between samples that are separated by more than one year. These results support our conclusions from the studies in chapters two and three that targeted control is likely to have an impact in reducing transmission in the surrounding community.

Chapter 5 Concluding Remarks

This thesis focused on analysing the spatial and temporal genetic variation in *P. falciparum* parasites in several sites in two sub-Saharan African countries: Kenya and The Gambia, and involved the use of genome-wide distributed SNPs as well as capillary sequence data from two antigen-encoding genes: *surf*$_{4.2}$ and *PfAMA1*. The different studies described in the various chapters reported in the thesis all fed into an overall aim of using parasite genetic data to determine parasite relatedness in time and space. This would, in turn, be used to infer the rate and spatial extent of parasite movement and mixing within individual study sites. The study was informed by the trends of declining malaria incidence reported over the last 15 years in most parts of sub-Saharan Africa where the disease is endemic, a phenomenon which has put elimination back on the forefront of the malaria research agenda for some of these countries. However, increasing decline means that malaria transmission becomes more heterogenous, leading to hotspots that often act as reservoirs of infection and fuel transmission to the rest of the community. As malaria control efforts enter pre-elimination and elimination phases, these hotspots need to be identified and targeted for more effective control to get rid of the reservoirs and prevent re-introduction into areas where transmission has been interrupted. The attractiveness of hotspot-targeted interventions to national malaria control programmes depends on their ease of identification, their stability over time and whether they seed transmission to the rest of the community such that eliminating malaria in the hotspot leads to malaria elimination in areas surrounding the hotspot. This last point was tested using several statistical measures in this thesis.

Under the first objective, SNP data were used to analyse parasite genetic relatedness in time and space as a proxy for parasite movement and mixing within three study sites with varying transmission intensities. In all three cases, several lines of investigation led to a conclusion of high degrees of mixing among parasites within each study site. First, spatial autocorrelation analyses showed little clustering of parasites into distinct sub-populations on large scale geographical levels, although weak clusters of parasite sub-populations were identified at micro-epidemiological scales of 5km or less. Second, analysis of the effect of the interaction of time and space in affecting changes in genetic variation between parasite pairs also pointed to rapid mixing of parasites within the study sites, with an increase in genetic differences between parasites being observed most conspicuously for parasites separated by up to 1 month, and little or no changes in genetic variation for parasite pairs separated by longer time intervals, regardless of the distance between them. A further analysis of spatial barriers to parasite movement showed no geographical barriers to parasite movement, and parasites were free to move and mix within individual study sites.

The implications of these observations to malaria control programmes are that targeting interventions to hotspots is likely to lead to a reduction in transmission in the surrounding regions. However, these control efforts will need to be sustained until all reservoirs are cleared, failure to which would increase the likelihood of re-introduction of infection due to the high mixing of parasites, e.g. if another hotspot exists nearby which also fuels transmission to the rest of the community. Having achieved malaria elimination in a specific area, additional effort will need to be put in place to prevent re-introduction of malaria infection due to importation fuelled by human migration.

Under the second objective, genome-wide distributed SNP data were once again used to analyse the spatio-temporal parasite genetic relatedness, this time on a sub-national scale, using *P. falciparum* samples collected from primary school children residing predominantly in western Kenya. This analysis also showed evidence of a high degree of parasite mixing, with little spatial autocorrelation among parasites, no local geographical adaptation of parasites to their environment and no evidence of spatial barriers to parasite movement within the region. Interestingly, additional analyses showed evidence of directionality in parasite movement for some parasites in the east/west and north/south directions. Analysis of human movement in this region, e.g. using mobile phone data, is recommended as part of further studies to determine whether concordant patterns in the directionality of human movement is observed. Such an observation would provide support for the observations made in this study.

The results reported in the two studies relied on analysis of a small number of SNPs which may have limited the power to detect genetic structure among highly similar parasites within individual study sites. Future studies that include the analysis of whole genome sequence data instead of SNP genotype data are recommended for a more detailed understanding of genetic variation, population structure and barriers to and directionality in parasite movement at finer geographical scales which may be targeted as part of malaria control efforts. Additionally, coupling whole genome sequence data with detailed temporal data will be required to determine the rate of parasite movement, evidence which would be useful in estimating how long targeted control would need to be sustained to achieve and maintain malaria elimination.

To ensure that the observations made in the two objectives above were accurate and not subject to SNP ascertainment bias resulting from the SNP subsets chosen, *surf$_{4.2}$* (exon 1) and *PfAMA1*, two highly polymorphic genes, were sequenced using capillary electrophoresis. Analyses of these sequences showed high genetic diversity and high levels of parasite mixing, with little evidence of clustering in time and space, similar to previous results observed using SNP genotype data. Analysis of both SNP and identity by state (IBS) data derived from the sequences showed high rates of parasite mixing over the 20-year study period, and for both genes, most of the variations in the sequences were observed between parasite pairs collected within a few days of each other. The fact that the results were similar to those obtained in the previous two analyses using SNP genotype data gives us more confidence in the conclusions reached in the previous analyses. The analysis shows high levels of parasite mixing in the study area. However, the analyses were carried out on genes that have previously been determined to be under immune selection. Future analysis may require the analysis of “neutral” genes or microsatellites, to determine whether similar patterns of genetic variation over time and distance are observed.

The data consistently point to a high level of parasite mixing, with a pattern showing increasing genetic differences between parasite pairs over distance where the pairs are collected within a few weeks of each other, but where distance is uninformative for genetic difference once the pairs are collected within several months or more of each other.

With parasites simply showing genetic drift one might expect increasing genetic differences over increasing distance without an interaction with time. I therefore propose a different model, where parasites exist in discrete parasite sub-populations, with newly acquired parasites multiplying and becoming established within these sub-populations and then being displaced by incoming parasites. Under this scenario, when initially comparing parasites sampled within a short time span, e.g. one week, the comparisons would be among parasites in individual sub-populations (A, B, C or D), and between parasites in non-mixing sub-populations (A vs B, A vs C, etc) (figure 5.1). Due to this non-mixing, there would be a high level of genetic diversity between parasites in different sub-populations, with diversity increasing with distance between the four sub-populations. However, over time, due to free movement, parasites in different populations would interact and exchange genetic material, thus decreasing genetic differences between parasites in different sub-populations. However, the introduction of new parasite genotypes into an existing population would be expected to be transient, with the new genotypes being either rapidly displaced by superinfecting strains or suppressed by host immunity, thus we see genetic diversity declining rapidly.

In the absence of these local factors acting on the survival of specific parasite genotypes, genetic diversity would be expected to increase with distance, without an interaction over time. This hypothesis can be tested using a mathematical model that incorporates the loss of genotypes from a population e.g. due to superinfections or host immunity.

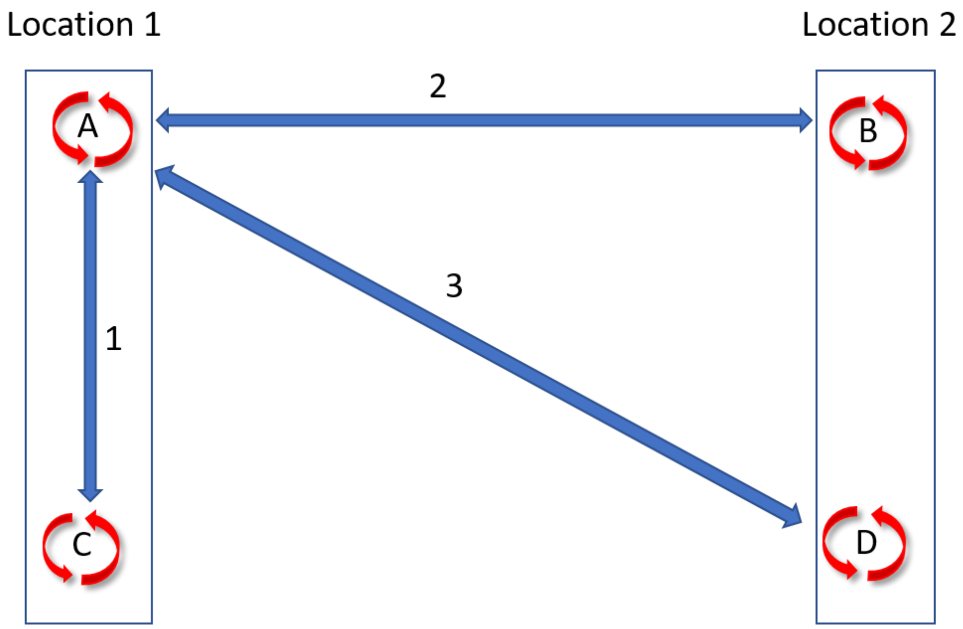


Figure 5.1: Possible representation of parasite movement and mixing explaining the pattern of genetic variations observed over time and space.

A, B, C, and D represent parasite sub-populations within two locations (1 and 2) that are initially non-mixing. Over short time spans, differences are computed within a population (represented by red block arrows). Blue arrows show comparisons between populations. Over time, parasites from different populations move and mix, thus reducing the genetic differences observed between these parasites. Hence when parasite pairs are sampled within a short period of time, the regression model of genetic difference versus distance is comparing parasite pairs at short distances, e.g. populations within location 1 (e.g. within population A) with parasite pairs across locations 1 and 2 (e.g. populations A and B), thus an effect of distance on genetic differences is seen. On the other hand, where parasite pairs are sampled with a long time interval between them, then the regression model of genetic difference versus distance is comparing parasites separated in time within location 1 (e.g. across populations A and C) with parasites across locations and time (e.g. across population A vs D), thus an effect of distance is no longer seen when the pairs collected are separated in time.

5.1 CONCLUSION

As malaria transmission declines, targeting control interventions to high incidence areas (hotspot) become increasingly important in resource-poor regions where the disease predominates. How effective such targeted control measures are depend on the extent of parasite mixing within and around the targeted areas. I have used SNP genotype and Sanger sequence data to analyse the rate and spatial extent of parasite mixing and have shown that there is a high level of parasite mixing in study sites with different transmission intensities, with no detectable geographical barriers to parasite movement over short distances. The various studies reported in this thesis show that parasites mix to high levels within the individual study sites analysed, with no barriers to parasite movement and weak clustering of parasite sub-populations over short spatial scales. This high level of parasite mixing means that targeting hotspots is likely to lead to a reduction in transmission in areas surrounding the hotspot. However, on a cautionary note, residual transmission will likely lead to rapid re-infection of the wider community, hence the need for sustained control until elimination is achieved and thereafter continuous monitoring to prevent parasite re-introduction.

Chapter 6 References

ACHAN, J., TALISUNA, A. O., ERHART, A., YEKA, A., TIBENDERANA, J. K., BALIRAINE, F. N., ROSENTHAL, P. J. & D'ALESSANDRO, U. 2011. Quinine, an old anti-malarial drug in a modern world: role in the treatment of malaria. *Malar J,* 10**,** 144.

ADAMS, J. 2008. DNA sequencing technologies *Nature Education,* 1.

AHMED, S., GALAGAN, S., SCOBIE, H., KHYANG, J., PRUE, C. S., KHAN, W. A., RAM, M., ALAM, M. S., HAQ, M. Z., AKTER, J., GLASS, G., NORRIS, D. E., NYUNT, M. M., SHIELDS, T., SULLIVAN, D. J. & SACK, D. A. 2013. Malaria hotspots drive hypoendemic transmission in the Chittagong Hill Districts of Bangladesh. *PLoS One,* 8**,** e69713.

AKEY, J. M., EBERLE, M. A., RIEDER, M. J., CARLSON, C. S., SHRIVER, M. D., NICKERSON, D. A. & KRUGLYAK, L. 2004. Population history and natural selection shape patterns of genetic variation in 132 genes. *PLoS Biol,* 2**,** e286.

ALEMU, K., WORKU, A. & BERHANE, Y. 2013. Malaria infection has spatial, temporal, and spatiotemporal heterogeneity in unstable malaria transmission areas in northwest Ethiopia. *PLoS One,* 8**,** e79966.

ALONSO, P. L., LINDSAY, S. W., ARMSTRONG, J. R., CONTEH, M., HILL, A. G., DAVID, P. H., FEGAN, G., DE FRANCISCO, A., HALL, A. J., SHENTON, F. C. & ET AL. 1991. The effect of insecticide-treated bed nets on mortality of Gambian children. *Lancet,* 337**,** 1499-502.

AMAMBUA-NGWA, A., TETTEH, K. K., MANSKE, M., GOMEZ-ESCOBAR, N., STEWART, L. B., DEERHAKE, M. E., CHEESEMAN, I. H., NEWBOLD, C. I., HOLDER, A. A., KNUEPFER, E., JANHA, O., JALLOW, M., CAMPINO, S., MACINNIS, B., KWIATKOWSKI, D. P. & CONWAY, D. J. 2012. Population genomic scan for candidate signatures of balancing selection to guide antigen characterization in malaria parasites. *PLoS Genet,* 8**,** e1002992.

ANDERSON, G. B. 2013. *Principal component analysis in R: An examination of the different functions and methods to perform PCA* [Online]. Available: https://www.ime.usp.br/~pavan/pdf/MAE0330-PCA-R-2013 [Accessed November, 30th 2016].

ANDERSON, T. J., HAUBOLD, B., WILLIAMS, J. T., ESTRADA-FRANCO, J. G., RICHARDSON, L., MOLLINEDO, R., BOCKARIE, M., MOKILI, J., MHARAKURWA, S., FRENCH, N., WHITWORTH, J., VELEZ, I. D., BROCKMAN, A. H., NOSTEN, F., FERREIRA, M. U. & DAY, K. P. 2000. Microsatellite markers reveal a spectrum of population structures in the malaria parasite *Plasmodium falciparum*. *Mol Biol Evol,* 17**,** 1467-82.

ANDERSON, T. J., NAIR, S., SUDIMACK, D., WILLIAMS, J. T., MAYXAY, M., NEWTON, P. N., GUTHMANN, J. P., SMITHUIS, F. M., TRAN, T. H., VAN DEN BROEK, I. V., WHITE, N. J. & NOSTEN, F. 2005. Geographical distribution of selected and putatively neutral SNPs in Southeast Asian malaria parasites. *Mol Biol Evol,* 22**,** 2362-74.

ANTHONY, T. G., CONWAY, D. J., COX-SINGH, J., MATUSOP, A., RATNAM, S., SHAMSUL, S. & SINGH, B. 2005. Fragmented population structure of *Plasmodium falciparum* in a region of declining endemicity. *J Infect Dis,* 191**,** 1558-64.

ARIEY, F., WITKOWSKI, B., AMARATUNGA, C., BEGHAIN, J., LANGLOIS, A. C., KHIM, N., KIM, S., DURU, V., BOUCHIER, C., MA, L., LIM, P., LEANG, R., DUONG, S., SRENG, S., SUON, S., CHUOR, C. M., BOUT, D. M., MENARD, S., ROGERS, W. O., GENTON, B., FANDEUR, T., MIOTTO, O., RINGWALD, P., LE BRAS, J., BERRY, A., BARALE, J. C., FAIRHURST, R. M., BENOIT-VICAL, F., MERCEREAU-PUIJALON, O. & MENARD, D. 2014. A molecular marker of artemisinin-resistant *Plasmodium falciparum* malaria. *Nature,* 505**,** 50-5.

ARNOTT, A., WAPLING, J., MUELLER, I., RAMSLAND, P. A., SIBA, P. M., REEDER, J. C. & BARRY, A. E. 2014. Distinct patterns of diversity, population structure and evolution in the *AMA1* genes of sympatric *Plasmodium falciparum* and Plasmodium vivax populations of Papua New Guinea from an area of similarly high transmission. *Malar J,* 13**,** 233.

ASHLEY, E. A., DHORDA, M., FAIRHURST, R. M., AMARATUNGA, C., LIM, P., SUON, S., SRENG, S., ANDERSON, J. M., MAO, S., SAM, B., SOPHA, C., CHUOR, C. M., NGUON, C., SOVANNAROTH, S., PUKRITTAYAKAMEE, S., JITTAMALA, P., CHOTIVANICH, K., CHUTASMIT, K., SUCHATSOONTHORN, C., RUNCHAROEN, R., HIEN, T. T., THUY-NHIEN, N. T., THANH, N. V., PHU, N. H., HTUT, Y., HAN, K. T., AYE, K. H., MOKUOLU, O. A., OLAOSEBIKAN, R. R., FOLARANMI, O. O., MAYXAY, M., KHANTHAVONG, M., HONGVANTHONG, B., NEWTON, P. N., ONYAMBOKO, M. A., FANELLO, C. I., TSHEFU, A. K., MISHRA, N., VALECHA, N., PHYO, A. P., NOSTEN, F., YI, P., TRIPURA, R., BORRMANN, S., BASHRAHEIL, M., PESHU, J., FAIZ, M. A., GHOSE, A., HOSSAIN, M. A., SAMAD, R., RAHMAN, M. R., HASAN, M. M., ISLAM, A., MIOTTO, O., AMATO, R., MACINNIS, B., STALKER, J., KWIATKOWSKI, D. P., BOZDECH, Z., JEEYAPANT, A., CHEAH, P. Y., SAKULTHAEW, T., CHALK, J., INTHARABUT, B., SILAMUT, K., LEE, S. J., VIHOKHERN, B., KUNASOL, C., IMWONG, M., TARNING, J., TAYLOR, W. J., YEUNG, S., WOODROW, C. J., FLEGG, J. A., DAS, D., SMITH, J., VENKATESAN, M., PLOWE, C. V., STEPNIEWSKA, K., GUERIN, P. J., DONDORP, A. M., DAY, N. P. & WHITE, N. J. 2014. Spread of artemisinin resistance in *Plasmodium falciparum* malaria. *N Engl J Med,* 371**,** 411-23.

ASSELE, V., NDOH, G. E., NKOGHE, D. & FANDEUR, T. 2015. No evidence of decline in malaria burden from 2006 to 2013 in a rural Province of Gabon: implications for public health policy. *BMC Public Health,* 15**,** 81.

BABIKER, H. A., LINES, J., HILL, W. G. & WALLIKER, D. 1997. Population structure of *Plasmodium falciparum* in villages with different malaria endemicity in east Africa. *Am J Trop Med Hyg,* 56**,** 141-7.

BAIDJOE, A. Y., STEVENSON, J., KNIGHT, P., STONE, W., STRESMAN, G., OSOTI, V., MAKORI, E., OWAGA, C., ODONGO, W., CHINA, P., SHAGARI, S., KARIUKI, S., DRAKELEY, C., COX, J. & BOUSEMA, T. 2016. Factors associated with high heterogeneity of malaria at fine spatial scale in the Western Kenyan highlands. *Malar J,* 15**,** 307.

BAKHIET, A. M., ABDEL-MUHSIN, A. M., ELZAKI, S. E., AL-HASHAMI, Z., ALBARWANI, H. S., ALQAMASHOUI, B. A., AL-HAMIDHI, S., IDRIS, M. A., ELAGIB, A. A., BEJA-PEREIRA, A. & BABIKER, H. A. 2015. *Plasmodium falciparum* population structure in Sudan post artemisinin-based combination therapy. *Acta Trop,* 148**,** 97-104.

BALIRAINE, F. N., AFRANE, Y. A., AMENYA, D. A., BONIZZONI, M., VARDO-ZALIK, A. M., MENGE, D. M., GITHEKO, A. K. & YAN, G. 2010. A cohort study of *Plasmodium falciparum* infection dynamics in Western Kenya Highlands. *BMC Infect Dis,* 10**,** 283.

BANNISTER, L. H., HOPKINS, J. M., DLUZEWSKI, A. R., MARGOS, G., WILLIAMS, I. T., BLACKMAN, M. J., KOCKEN, C. H., THOMAS, A. W. & MITCHELL, G. H. 2003. *Plasmodium falciparum* apical membrane antigen 1 (PfAMA-1) is translocated within micronemes along subpellicular microtubules during merozoite development. *J Cell Sci,* 116**,** 3825-34.

BARTOLONI, A. & ZAMMARCHI, L. 2012. Clinical aspects of uncomplicated and severe malaria. *Mediterr J Hematol Infect Dis,* 4**,** e2012026.

BAUTISTA, C. T., CHAN, A. S., RYAN, J. R., CALAMPA, C., ROPER, M. H., HIGHTOWER, A. W. & MAGILL, A. J. 2006. Epidemiology and spatial analysis of malaria in the Northern Peruvian Amazon. *Am J Trop Med Hyg,* 75**,** 1216-22.

BAYOH, M. N., WALKER, E. D., KOSGEI, J., OMBOK, M., OLANG, G. B., GITHEKO, A. K., KILLEEN, G. F., OTIENO, P., DESAI, M., LOBO, N. F., VULULE, J. M., HAMEL, M. J., KARIUKI, S. & GIMNIG, J. E. 2014. Persistently high estimates of late night, indoor exposure to malaria vectors despite high coverage of insecticide treated nets. *Parasit Vectors,* 7**,** 380.

BEJON, P., MWACHARO, J., KAI, O., MWANGI, T., MILLIGAN, P., TODRYK, S., KEATING, S., LANG, T., LOWE, B., GIKONYO, C., MOLYNEUX, C., FEGAN, G., GILBERT, S. C., PESHU, N., MARSH, K. & HILL, A. V. 2006. A phase 2b randomised trial of the candidate malaria vaccines FP9 ME-TRAP and MVA ME-TRAP among children in Kenya. *PLoS Clin Trials,* 1**,** e29.

BEJON, P., TURNER, L., LAVSTSEN, T., CHAM, G., OLOTU, A., DRAKELEY, C. J., LIEVENS, M., VEKEMANS, J., SAVARESE, B., LUSINGU, J., VON SEIDLEIN, L., BULL, P. C., MARSH, K. & THEANDER, T. G. 2011. Serological evidence of discrete spatial clusters of *Plasmodium falciparum* parasites. *PLoS One,* 6**,** e21711.

BEJON, P., WILLIAMS, T. N., LILJANDER, A., NOOR, A. M., WAMBUA, J., OGADA, E., OLOTU, A., OSIER, F. H., HAY, S. I., FARNERT, A. & MARSH, K. 2010. Stable and unstable malaria hotspots in longitudinal cohort studies in Kenya. *PLoS Med,* 7**,** e1000304.

BEJON, P., WILLIAMS, T. N., NYUNDO, C., HAY, S. I., BENZ, D., GETHING, P. W., OTIENDE, M., PESHU, J., BASHRAHEIL, M., GREENHOUSE, B., BOUSEMA, T., BAUNI, E., MARSH, K., SMITH, D. L. & BORRMANN, S. 2014. A micro-epidemiological analysis of febrile malaria in Coastal Kenya showing hotspots within hotspots. *Elife,* 3**,** e02130.

BENNETT, A., KAZEMBE, L., MATHANGA, D. P., KINYOKI, D., ALI, D., SNOW, R. W. & NOOR, A. M. 2013. Mapping malaria transmission intensity in Malawi, 2000-2010. *Am J Trop Med Hyg,* 89**,** 840-9.

BHATT, S., WEISS, D. J., CAMERON, E., BISANZIO, D., MAPPIN, B., DALRYMPLE, U., BATTLE, K. E., MOYES, C. L., HENRY, A., ECKHOFF, P. A., WENGER, E. A., BRIET, O., PENNY, M. A., SMITH, T. A., BENNETT, A., YUKICH, J., EISELE, T. P., GRIFFIN, J. T., FERGUS, C. A., LYNCH, M., LINDGREN, F., COHEN, J. M., MURRAY, C. L., SMITH, D. L., HAY, S. I., CIBULSKIS, R. E. & GETHING, P. W. 2015. The effect of malaria control on *Plasmodium falciparum* in Africa between 2000 and 2015. *Nature,* 526**,** 207-11.

BHATTARAI, A., ALI, A. S., KACHUR, S. P., MARTENSSON, A., ABBAS, A. K., KHATIB, R., AL-MAFAZY, A. W., RAMSAN, M., ROTLLANT, G., GERSTENMAIER, J. F., MOLTENI, F., ABDULLA, S., MONTGOMERY, S. M., KANEKO, A. & BJORKMAN, A. 2007. Impact of artemisinin-based combination therapy and insecticide-treated nets on malaria burden in Zanzibar. *PLoS Med,* 4**,** e309.

BODKER, R., MSANGENI, H. A., KISINZA, W. & LINDSAY, S. W. 2006. Relationship between the intensity of exposure to malaria parasites and infection in the Usambara Mountains, Tanzania. *Am J Trop Med Hyg,* 74**,** 716-23.

BONIZZONI, M., AFRANE, Y., BALIRAINE, F. N., AMENYA, D. A., GITHEKO, A. K. & YAN, G. 2009. Genetic structure of *Plasmodium falciparum* populations between lowland and highland sites and antimalarial drug resistance in Western Kenya. *Infect Genet Evol,* 9**,** 806-12.

BOUSEMA, T., DRAKELEY, C., GESASE, S., HASHIM, R., MAGESA, S., MOSHA, F., OTIENO, S., CARNEIRO, I., COX, J., MSUYA, E., KLEINSCHMIDT, I., MAXWELL, C., GREENWOOD, B., RILEY, E., SAUERWEIN, R., CHANDRAMOHAN, D. & GOSLING, R. 2010. Identification of hot spots of malaria transmission for targeted malaria control. *J Infect Dis,* 201**,** 1764-74.

BOUSEMA, T., GRIFFIN, J. T., SAUERWEIN, R. W., SMITH, D. L., CHURCHER, T. S., TAKKEN, W., GHANI, A., DRAKELEY, C. & GOSLING, R. 2012. Hitting hotspots: spatial targeting of malaria for control and elimination. *PLoS Med,* 9**,** e1001165.

BOUSEMA, T., OKELL, L., FELGER, I. & DRAKELEY, C. 2014. Asymptomatic malaria infections: detectability, transmissibility and public health relevance. *Nat Rev Microbiol,* 12**,** 833-40.

BOUSEMA, T., STEVENSON, J., BAIDJOE, A., STRESMAN, G., GRIFFIN, J. T., KLEINSCHMIDT, I., REMARQUE, E. J., VULULE, J., BAYOH, N., LASERSON, K., DESAI, M., SAUERWEIN, R., DRAKELEY, C. & COX, J. 2013. The impact of hotspot-targeted interventions on malaria transmission: study protocol for a cluster-randomized controlled trial. *Trials,* 14**,** 36.

BOUSEMA, T., STRESMAN, G., BAIDJOE, A. Y., BRADLEY, J., KNIGHT, P., STONE, W., OSOTI, V., MAKORI, E., OWAGA, C., ODONGO, W., CHINA, P., SHAGARI, S., DOUMBO, O. K., SAUERWEIN, R. W., KARIUKI, S., DRAKELEY, C., STEVENSON, J. & COX, J. 2016. The Impact of Hotspot-Targeted Interventions on Malaria Transmission in Rachuonyo South District in the Western Kenyan Highlands: A Cluster-Randomized Controlled Trial. *PLoS Med,* 13**,** e1001993.

BROOKER, S., KOLACZINSKI, J. H., GITONGA, C. W., NOOR, A. M. & SNOW, R. W. 2009. The use of schools for malaria surveillance and programme evaluation in Africa. *Malar J,* 8**,** 231.

BUTLER, A. R., KHAN, S. & FERGUSON, E. 2010. A brief history of malaria chemotherapy. *J R Coll Physicians Edinb,* 40**,** 172-7.

CAMPINO, S., AUBURN, S., KIVINEN, K., ZONGO, I., OUEDRAOGO, J. B., MANGANO, V., DJIMDE, A., DOUMBO, O. K., KIARA, S. M., NZILA, A., BORRMANN, S., MARSH, K., MICHON, P., MUELLER, I., SIBA, P., JIANG, H., SU, X. Z., AMARATUNGA, C., SOCHEAT, D., FAIRHURST, R. M., IMWONG, M., ANDERSON, T., NOSTEN, F., WHITE, N. J., GWILLIAM, R., DELOUKAS, P., MACINNIS, B., NEWBOLD, C. I., ROCKETT, K., CLARK, T. G. & KWIATKOWSKI, D. P. 2011. Population genetic analysis of *Plasmodium falciparum* parasites using a customized Illumina GoldenGate genotyping assay. *PLoS One,* 6**,** e20251.

CAPUTO, B., NWAKANMA, D., JAWARA, M., ADIAMOH, M., DIA, I., KONATE, L., PETRARCA, V., CONWAY, D. J. & DELLA TORRE, A. 2008. *Anopheles gambiae* complex along The Gambia river, with particular reference to the molecular forms of *An. gambiae s.s. Malar J,* 7**,** 182.

CARNEIRO, I., ROCA-FELTRER, A., GRIFFIN, J. T., SMITH, L., TANNER, M., SCHELLENBERG, J. A., GREENWOOD, B. & SCHELLENBERG, D. 2010. Age-patterns of malaria vary with severity, transmission intensity and seasonality in sub-Saharan Africa: a systematic review and pooled analysis. *PLoS One,* 5**,** e8988.

CEESAY, S. J., CASALS-PASCUAL, C., ERSKINE, J., ANYA, S. E., DUAH, N. O., FULFORD, A. J., SESAY, S. S., ABUBAKAR, I., DUNYO, S., SEY, O., PALMER, A., FOFANA, M., CORRAH, T., BOJANG, K. A., WHITTLE, H. C., GREENWOOD, B. M. & CONWAY, D. J. 2008. Changes in malaria indices between 1999 and 2007 in The Gambia: a retrospective analysis. *Lancet,* 372**,** 1545-54.

CEESAY, S. J., CASALS-PASCUAL, C., NWAKANMA, D. C., WALTHER, M., GOMEZ-ESCOBAR, N., FULFORD, A. J., TAKEM, E. N., NOGARO, S., BOJANG, K. A., CORRAH, T., JAYE, M. C., TAAL, M. A., SONKO, A. A. & CONWAY, D. J. 2010. Continued decline of malaria in The Gambia with implications for elimination. *PLoS One,* 5**,** e12242.

CHAN, J. A., FOWKES, F. J. & BEESON, J. G. 2014. Surface antigens of *Plasmodium falciparum*-infected erythrocytes as immune targets and malaria vaccine candidates. *Cell Mol Life Sci,* 71**,** 3633-57.

CHEESEMAN, I. H., GOMEZ-ESCOBAR, N., CARRET, C. K., IVENS, A., STEWART, L. B., TETTEH, K. K. & CONWAY, D. J. 2009. Gene copy number variation throughout the *Plasmodium falciparum* genome. *BMC Genomics,* 10**,** 353.

CHESNE-SECK, M. L., PIZARRO, J. C., VULLIEZ-LE NORMAND, B., COLLINS, C. R., BLACKMAN, M. J., FABER, B. W., REMARQUE, E. J., KOCKEN, C. H., THOMAS, A. W. & BENTLEY, G. A. 2005. Structural comparison of apical membrane antigen 1 orthologues and paralogues in apicomplexan parasites. *Mol Biochem Parasitol,* 144**,** 55-67.

CHIREBVU, E., CHIMBARI, M. J. & NGWENYA, B. N. 2014. Assessment of risk factors associated with malaria transmission in tubu village, northern botswana. *Malar Res Treat,* 2014**,** 403069.

CONGPUONG, K., SUKARAM, R., PROMPAN, Y. & DORNAE, A. 2014. Genetic diversity of the *msp-1*, *msp-2*, and *glurp* genes of *Plasmodium falciparum* isolates along the Thai-Myanmar borders. *Asian Pac J Trop Biomed,* 4**,** 598-602.

CONWAY, D. J. 2007. Molecular epidemiology of malaria. *Clin Microbiol Rev,* 20**,** 188-204.

CONWAY, D. J., MACHADO, R. L., SINGH, B., DESSERT, P., MIKES, Z. S., POVOA, M. M., ODUOLA, A. M. & ROPER, C. 2001. Extreme geographical fixation of variation in the *Plasmodium falciparum* gamete surface protein gene *Pfs48/45* compared with microsatellite loci. *Mol Biochem Parasitol,* 115**,** 145-56.

CORTES, A., MELLOMBO, M., MUELLER, I., BENET, A., REEDER, J. C. & ANDERS, R. F. 2003. Geographical structure of diversity and differences between symptomatic and asymptomatic infections for *Plasmodium falciparum* vaccine candidate *AMA1*. *Infect Immun,* 71**,** 1416-26.

COWMAN, A. F. & CRABB, B. S. 2006. Invasion of red blood cells by malaria parasites. *Cell,* 124**,** 755-66.

CUI, L. & SU, X. Z. 2009. Discovery, mechanisms of action and combination therapy of artemisinin. *Expert Rev Anti Infect Ther,* 7**,** 999-1013.

CURTIS, C. F. & MNZAVA, A. E. 2000. Comparison of house spraying and insecticide-treated nets for malaria control. *Bull World Health Organ,* 78**,** 1389-400.

D'ALESSANDRO, U., OLALEYE, B. O., MCGUIRE, W., LANGEROCK, P., BENNETT, S., AIKINS, M. K., THOMSON, M. C., CHAM, M. K., CHAM, B. A. & GREENWOOD, B. M. 1995. Mortality and morbidity from malaria in Gambian children after introduction of an impregnated bednet programme. *Lancet,* 345**,** 479-83.

DANIELS, R. F., SCHAFFNER, S. F., WENGER, E. A., PROCTOR, J. L., CHANG, H. H., WONG, W., BARO, N., NDIAYE, D., FALL, F. B., NDIOP, M., BA, M., MILNER, D. A., JR., TAYLOR, T. E., NEAFSEY, D. E., VOLKMAN, S. K., ECKHOFF, P. A., HARTL, D. L. & WIRTH, D. F. 2015. Modeling malaria genomics reveals transmission decline and rebound in Senegal. *Proc Natl Acad Sci U S A,* 112**,** 7067-72.

DANIELS, R. F., VOLKMAN, S. K., MILNER, D. A., MAHESH, N., NEAFSEY, D. E., PARK, D. J., ROSEN, D., ANGELINO, E., SABETI, P. C., WIRTH, D. F. & WIEGAND, R. C. 2008. A general SNP-based molecular barcode for *Plasmodium falciparum* identification and tracking. *Malar J,* 7**,** 223.

DE LA VEGA, F. M., LAZARUK, K. D., RHODES, M. D. & WENZ, M. H. 2005. Assessment of two flexible and compatible SNP genotyping platforms: TaqMan SNP Genotyping Assays and the SNPlex Genotyping System. *Mutat Res,* 573**,** 111-35.

DONDORP, A. M., NOSTEN, F., YI, P., DAS, D., PHYO, A. P., TARNING, J., LWIN, K. M., ARIEY, F., HANPITHAKPONG, W., LEE, S. J., RINGWALD, P., SILAMUT, K., IMWONG, M., CHOTIVANICH, K., LIM, P., HERDMAN, T., AN, S. S., YEUNG, S., SINGHASIVANON, P., DAY, N. P., LINDEGARDH, N., SOCHEAT, D. & WHITE, N. J. 2009. Artemisinin resistance in *Plasmodium falciparum* malaria. *N Engl J Med,* 361**,** 455-67.

DREW, D. R., HODDER, A. N., WILSON, D. W., FOLEY, M., MUELLER, I., SIBA, P. M., DENT, A. E., COWMAN, A. F. & BEESON, J. G. 2012. Defining the antigenic diversity of *Plasmodium falciparum* apical membrane antigen 1 and the requirements for a multi-allele vaccine against malaria. *PLoS One,* 7**,** e51023.

DUAN, J., MU, J., THERA, M. A., JOY, D., KOSAKOVSKY POND, S. L., DIEMERT, D., LONG, C., ZHOU, H., MIURA, K., OUATTARA, A., DOLO, A., DOUMBO, O., SU, X. Z. & MILLER, L. 2008. Population structure of the genes encoding the polymorphic *Plasmodium falciparum* apical membrane antigen 1: implications for vaccine design. *Proc Natl Acad Sci U S A,* 105**,** 7857-62.

DURET, L. 2008. Neutral theory: The null hypothesis of molecular evolution. *Nature Education,* 1**,** 218.

EDENBERG, H. J. & LIU, Y. 2009. Laboratory methods for high-throughput genotyping. *Cold Spring Harb Protoc,* 2009**,** pdb.top62.

EDGAR, R. C. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res,* 32**,** 1792-7.

EKLAND, E. H. & FIDOCK, D. A. 2007. Advances in understanding the genetic basis of antimalarial drug resistance. *Curr Opin Microbiol,* 10**,** 363-70.

EPPERSON, B. K. & LI, T. 1996. Measurement of genetic structure within populations using Moran's spatial autocorrelation statistics. *Proc Natl Acad Sci U S A,* 93**,** 10528-32.

ERNST, K. C., ADOKA, S. O., KOWUOR, D. O., WILSON, M. L. & JOHN, C. C. 2006. Malaria hotspot areas in a highland Kenya site are consistent in epidemic and non-epidemic years and are associated with ecological factors. *Malar J,* 5**,** 78.

ESCALANTE, A. A., GREBERT, H. M., CHAIYAROJ, S. C., MAGRIS, M., BISWAS, S., NAHLEN, B. L. & LAL, A. A. 2001. Polymorphism in the gene encoding the apical membrane antigen-1 (AMA-1) of *Plasmodium falciparum*. X. Asembo Bay Cohort Project. *Mol Biochem Parasitol,* 113**,** 279-87.

FALK, N., MAIRE, N., SAMA, W., OWUSU-AGYEI, S., SMITH, T., BECK, H. P. & FELGER, I. 2006. Comparison of PCR-RFLP and Genescan-based genotyping for analyzing infection dynamics of *Plasmodium falciparum*. *Am J Trop Med Hyg,* 74**,** 944-50.

FEGAN, G. W., NOOR, A. M., AKHWALE, W. S., COUSENS, S. & SNOW, R. W. 2007. Effect of expanded insecticide-treated bednet coverage on child survival in rural Kenya: a longitudinal study. *Lancet,* 370**,** 1035-9.

GABRIEL, S., ZIAUGRA, L. & TABBAA, D. 2009. SNP genotyping using the Sequenom MassARRAY iPLEX platform. *Curr Protoc Hum Genet,* Chapter 2**,** Unit 2.12.

GARDNER, M. J., HALL, N., FUNG, E., WHITE, O., BERRIMAN, M., HYMAN, R. W., CARLTON, J. M., PAIN, A., NELSON, K. E., BOWMAN, S., PAULSEN, I. T., JAMES, K., EISEN, J. A., RUTHERFORD, K., SALZBERG, S. L., CRAIG, A., KYES, S., CHAN, M. S., NENE, V., SHALLOM, S. J., SUH, B., PETERSON, J., ANGIUOLI, S., PERTEA, M., ALLEN, J., SELENGUT, J., HAFT, D., MATHER, M. W., VAIDYA, A. B., MARTIN, D. M., FAIRLAMB, A. H., FRAUNHOLZ, M. J., ROOS, D. S., RALPH, S. A., MCFADDEN, G. I., CUMMINGS, L. M., SUBRAMANIAN, G. M., MUNGALL, C., VENTER, J. C., CARUCCI, D. J., HOFFMAN, S. L., NEWBOLD, C., DAVIS, R. W., FRASER, C. M. & BARRELL, B. 2002. Genome sequence of the human malaria parasite *Plasmodium falciparum*. *Nature,* 419**,** 498-511.

GARDY, J. L., JOHNSTON, J. C., HO SUI, S. J., COOK, V. J., SHAH, L., BRODKIN, E., REMPEL, S., MOORE, R., ZHAO, Y., HOLT, R., VARHOL, R., BIROL, I., LEM, M., SHARMA, M. K., ELWOOD, K., JONES, S. J., BRINKMAN, F. S., BRUNHAM, R. C. & TANG, P. 2011. Whole-genome sequencing and social-network analysis of a tuberculosis outbreak. *N Engl J Med,* 364**,** 730-9.

GATEI, W., GIMNIG, J. E., HAWLEY, W., TER KUILE, F., ODERO, C., IRIEMENAM, N. C., SHAH, M. P., HOWARD, P. P., OMOSUN, Y. O., TERLOUW, D. J., NAHLEN, B., SLUTSKER, L., HAMEL, M. J., KARIUKI, S., WALKER, E. & SHI, Y. P. 2015. Genetic diversity of *Plasmodium falciparum* parasite by microsatellite markers after scale-up of insecticide-treated bed nets in western Kenya. *Malar J,* 13 Suppl 1**,** 495.

GEIGER, C., AGUSTAR, H. K., COMPAORE, G., COULIBALY, B., SIE, A., BECHER, H., LANZER, M. & JANISCH, T. 2013. Declining malaria parasite prevalence and trends of asymptomatic parasitaemia in a seasonal transmission setting in North-Western Burkina Faso between 2000 and 2009-2012. *Malar J,* 12**,** 27.

GETHING, P. W., CASEY, D. C., WEISS, D. J., BISANZIO, D., BHATT, S., CAMERON, E., BATTLE, K. E., DALRYMPLE, U., ROZIER, J., RAO, P. C., KUTZ, M. J., BARBER, R. M., HUYNH, C., SHACKELFORD, K. A., COATES, M. M., NGUYEN, G., FRASER, M. S., KULIKOFF, R., WANG, H., NAGHAVI, M., SMITH, D. L., MURRAY, C. J., HAY, S. I. & LIM, S. S. 2016. Mapping *Plasmodium falciparum* Mortality in Africa between 1990 and 2015. *N Engl J Med,* 375**,** 2435-2445.

GETHING, P. W., PATIL, A. P., SMITH, D. L., GUERRA, C. A., ELYAZAR, I. R., JOHNSTON, G. L., TATEM, A. J. & HAY, S. I. 2011. A new world malaria map: *Plasmodium falciparum* endemicity in 2010. *Malar J,* 10**,** 378.

GITAKA, J. N., TAKEDA, M., KIMURA, M., IDRIS, Z. M., CHAN, C. W., KONGERE, J., YAHATA, K., MUREGI, F. W., ICHINOSE, Y., KANEKO, A. & KANEKO, O. 2017. Selections, frameshift mutations, and copy number variation detected on the *surf 4.1* gene in the western Kenyan *Plasmodium falciparum* population. *Malar J,* 16**,** 98.

GITONGA, C. W., EDWARDS, T., KARANJA, P. N., NOOR, A. M., SNOW, R. W. & BROOKER, S. J. 2012. *Plasmodium* infection, anaemia and mosquito net use among school children across different settings in Kenya. *Trop Med Int Health,* 17**,** 858-70.

GITONGA, C. W., KARANJA, P. N., KIHARA, J., MWANJE, M., JUMA, E., SNOW, R. W., NOOR, A. M. & BROOKER, S. 2010. Implementing school malaria surveys in Kenya: towards a national surveillance system. *Malar J,* 9**,** 306.

GOODWIN, S., MCPHERSON, J. D. & MCCOMBIE, W. R. 2016. Coming of age: ten years of next-generation sequencing technologies. *Nat Rev Genet,* 17**,** 333-51.

GRAVES, P. M., OSGOOD, D. E., THOMSON, M. C., SEREKE, K., ARAIA, A., ZEROM, M., CECCATO, P., BELL, M., DEL CORRAL, J., GHEBRESELASSIE, S., BRANTLY, E. P. & GHEBREMESKEL, T. 2008. Effectiveness of malaria control during changing climate conditions in Eritrea, 1998-2003. *Trop Med Int Health,* 13**,** 218-28.

GRIMES, D. A. & SCHULZ, K. F. 2012. False alarms and pseudo-epidemics: the limitations of observational epidemiology. *Obstet Gynecol,* 120**,** 920-7.

GUERRA, C. A., GIKANDI, P. W., TATEM, A. J., NOOR, A. M., SMITH, D. L., HAY, S. I. & SNOW, R. W. 2008. The limits and intensity of *Plasmodium falciparum* transmission: implications for malaria control and elimination worldwide. *PLoS Med,* 5**,** e38.

GUERRA, C. A., HOWES, R. E., PATIL, A. P., GETHING, P. W., VAN BOECKEL, T. P., TEMPERLEY, W. H., KABARIA, C. W., TATEM, A. J., MANH, B. H., ELYAZAR, I. R., BAIRD, J. K., SNOW, R. W. & HAY, S. I. 2010. The international limits and population at risk of *Plasmodium vivax* transmission in 2009. *PLoS Negl Trop Dis,* 4**,** e774.

GUNAWARDENA, S. & KARUNAWEERA, N. D. 2015. Advances in genetics and genomics: use and limitations in achieving malaria elimination goals. *Pathog Glob Health,* 109**,** 123-41.

GUNTHER, T. & COOP, G. 2013. Robust identification of local adaptation from allele frequencies. *Genetics,* 195**,** 205-20.

GUPTA, P., SINGH, R., KHAN, H., RAZA, A., YADAVENDU, V., BHATT, R. M. & SINGH, V. 2014. Genetic profiling of the *Plasmodium falciparum* population using antigenic molecular markers. *ScientificWorldJournal,* 2014**,** 140867.

HABLUETZEL, A., DIALLO, D. A., ESPOSITO, F., LAMIZANA, L., PAGNONI, F., LENGELER, C., TRAORE, C. & COUSENS, S. N. 1997. Do insecticide-treated curtains reduce all-cause child mortality in Burkina Faso? *Trop Med Int Health,* 2**,** 855-62.

HAMMOND, A., GALIZI, R., KYROU, K., SIMONI, A., SINISCALCHI, C., KATSANOS, D., GRIBBLE, M., BAKER, D., MAROIS, E., RUSSELL, S., BURT, A., WINDBICHLER, N., CRISANTI, A. & NOLAN, T. 2016. A CRISPR-Cas9 gene drive system targeting female reproduction in the malaria mosquito vector *Anopheles gambiae*. *Nat Biotechnol,* 34**,** 78-83.

HARRIS, S. R., CARTWRIGHT, E. J., TOROK, M. E., HOLDEN, M. T., BROWN, N. M., OGILVY-STUART, A. L., ELLINGTON, M. J., QUAIL, M. A., BENTLEY, S. D., PARKHILL, J. & PEACOCK, S. J. 2013. Whole-genome sequencing for analysis of an outbreak of meticillin-resistant *Staphylococcus aureus*: a descriptive study. *Lancet Infect Dis,* 13**,** 130-6.

HARRIS, S. R., FEIL, E. J., HOLDEN, M. T., QUAIL, M. A., NICKERSON, E. K., CHANTRATITA, N., GARDETE, S., TAVARES, A., DAY, N., LINDSAY, J. A., EDGEWORTH, J. D., DE LENCASTRE, H., PARKHILL, J., PEACOCK, S. J. & BENTLEY, S. D. 2010. Evolution of MRSA during hospital transmission and intercontinental spread. *Science,* 327**,** 469-74.

HARTL, D. L. & CLARK, A. G. 2007. *Principles of population genetics.* , Sinauer Associates, Inc.

HARTL, D. L., VOLKMAN, S. K., NIELSEN, K. M., BARRY, A. E., DAY, K. P., WIRTH, D. F. & WINZELER, E. A. 2002. The paradoxical population genetics of *Plasmodium falciparum*. *Trends Parasitol,* 18**,** 266-72.

HAWLEY, W. A., PHILLIPS-HOWARD, P. A., TER KUILE, F. O., TERLOUW, D. J., VULULE, J. M., OMBOK, M., NAHLEN, B. L., GIMNIG, J. E., KARIUKI, S. K., KOLCZAK, M. S. & HIGHTOWER, A. W. 2003. Community-wide effects of permethrin-treated bed nets on child mortality and malaria morbidity in western Kenya. *Am J Trop Med Hyg,* 68**,** 121-7.

HAY, S. I., GUERRA, C. A., GETHING, P. W., PATIL, A. P., TATEM, A. J., NOOR, A. M., KABARIA, C. W., MANH, B. H., ELYAZAR, I. R., BROOKER, S., SMITH, D. L., MOYEED, R. A. & SNOW, R. W. 2009. A world malaria map: *Plasmodium falciparum* endemicity in 2007. *PLoS Med,* 6**,** e1000048.

HAY, S. I., ROGERS, D. J., TOOMER, J. F. & SNOW, R. W. 2000. Annual *Plasmodium falciparum* entomological inoculation rates (EIR) across Africa: literature survey, Internet access and review. *Trans R Soc Trop Med Hyg,* 94**,** 113-27.

HILL, A. V. 2011. Vaccines against malaria. *Philos Trans R Soc Lond B Biol Sci,* 366**,** 2806-14.

HODDER, A. N., CREWTHER, P. E., MATTHEW, M. L., REID, G. E., MORITZ, R. L., SIMPSON, R. J. & ANDERS, R. F. 1996. The disulfide bond structure of *Plasmodium* apical membrane antigen-1. *J Biol Chem,* 271**,** 29446-52.

HUDSON, R. R. 1987. Estimating the recombination parameter of a finite population model without selection. *Genet Res,* 50**,** 245-50.

HUDSON, R. R. & KAPLAN, N. L. 1985. Statistical properties of the number of recombination events in the history of a sample of DNA sequences. *Genetics,* 111**,** 147-64.

INGASIA, L. A., CHERUIYOT, J., OKOTH, S. A., ANDAGALU, B. & KAMAU, E. 2016. Genetic variability and population structure of *Plasmodium falciparum* parasite populations from different malaria ecological regions of Kenya. *Infect Genet Evol,* 39**,** 372-80.

JANEZIC, S., OCEPEK, M., ZIDARIC, V. & RUPNIK, M. 2012. *Clostridium difficile* genotypes other than ribotype 078 that are prevalent among human, animal and environmental isolates. *BMC Microbiol,* 12**,** 48.

JEFFARES, D. C., PAIN, A., BERRY, A., COX, A. V., STALKER, J., INGLE, C. E., THOMAS, A., QUAIL, M. A., SIEBENTHALL, K., UHLEMANN, A. C., KYES, S., KRISHNA, S., NEWBOLD, C., DERMITZAKIS, E. T. & BERRIMAN, M. 2007. Genome variation and evolution of the malaria parasite *Plasmodium falciparum*. *Nat Genet,* 39**,** 120-5.

JENKINS, S. & GIBSON, N. 2002. High-throughput SNP genotyping. *Comp Funct Genomics,* 3**,** 57-66.

JIANG, H., LI, N., GOPALAN, V., ZILVERSMIT, M. M., VARMA, S., NAGARAJAN, V., LI, J., MU, J., HAYTON, K., HENSCHEN, B., YI, M., STEPHENS, R., MCVEAN, G., AWADALLA, P., WELLEMS, T. E. & SU, X. Z. 2011. High recombination rates and hotspots in a *Plasmodium falciparum* genetic cross. *Genome Biol,* 12**,** R33.

KAEWTHAMASORN, M., YAHATA, K., ALEXANDRE, J. S., XANGSAYARATH, P., NAKAZAWA, S., TORII, M., SATTABONGKOT, J., UDOMSANGPETCH, R. & KANEKO, O. 2012. Stable allele frequency distribution of the polymorphic region of SURFIN(4.2) in *Plasmodium falciparum* isolates from Thailand. *Parasitol Int,* 61**,** 317-23.

KAGAYA, W., MIYAZAKI, S., YAHATA, K., OHTA, N. & KANEKO, O. 2015. The Cytoplasmic Region of *Plasmodium falciparum* SURFIN4.2 Is Required for Transport from Maurer's Clefts to the Red Blood Cell Surface. *Trop Med Health,* 43**,** 265-72.

KALTZ, O. & SHYKOFF, J. A. 1998. Local adaptation in host–parasite systems. *Heredity,* 81**,** 361–370.

KANGOYE, D. T., NOOR, A., MIDEGA, J., MWONGELI, J., MKABILI, D., MOGENI, P., KERUBO, C., AKOO, P., MWANGANGI, J., DRAKELEY, C., MARSH, K., BEJON, P. & NJUGUNA, P. 2016. Malaria hotspots defined by clinical malaria, asymptomatic carriage, PCR and vector numbers in a low transmission area on the Kenyan Coast. *Malar J,* 15**,** 213.

KENYA NATIONAL MALARIA CONTROL PROGRAMME 2016. Kenya Malaria Indicator Survey 2015. .

KISZEWSKI, A., MELLINGER, A., SPIELMAN, A., MALANEY, P., SACHS, S. E. & SACHS, J. 2004. A global index representing the stability of malaria transmission. *Am J Trop Med Hyg,* 70**,** 486-98.

KLEIN, E. Y. 2013. Antimalarial drug resistance: a review of the biology and strategies to delay emergence and spread. *Int J Antimicrob Agents,* 41**,** 311-7.

KONATE, L., ZWETYENGA, J., ROGIER, C., BISCHOFF, E., FONTENILLE, D., TALL, A., SPIEGEL, A., TRAPE, J. F. & MERCEREAU-PUIJALON, O. 1999. Variation of *Plasmodium falciparum msp1* block 2 and *msp2* allele prevalence and of infection complexity in two neighbouring Senegalese villages with different transmission conditions. *Trans R Soc Trop Med Hyg,* 93 Suppl 1**,** 21-8.

KULLDORF, M. 2014. SaTScan v9.3: Software for the spatila and space-time scan

statistics.

KULLDORFF, M. A. I. M. S., INC 2009. SaTScan^{TM} v8.0: Software for the spatial and space-time scan statistics.

KUMAR, S., STECHER, G. & TAMURA, K. 2016. MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. *Mol Biol Evol,* 33**,** 1870-4.

KWIATKOWSKI, D. 2015. Malaria genomics: tracking a diverse and evolving parasite population. *Int Health,* 7**,** 82-4.

KWIATKOWSKI, D. P. 2005. How malaria has affected the human genome and what human genetics can teach us about malaria. *Am J Hum Genet,* 77**,** 171-92.

KWOK, P. Y. & CHEN, X. 2003. Detection of single nucleotide polymorphisms. *Curr Issues Mol Biol,* 5**,** 43-60.

KYES, S., HARDING, R., BLACK, G., CRAIG, A., PESHU, N., NEWBOLD, C. & MARSH, K. 1997. Limited spatial clustering of individual *Plasmodium falciparum* alleles in field isolates from coastal Kenya. *Am J Trop Med Hyg,* 57**,** 205-15.

LACHANCE, J. & TISHKOFF, S. A. 2013. SNP ascertainment bias in population genetic analyses: why it is important, and how to correct it. *Bioessays,* 35**,** 780-6.

LEANDRO-REGUILLO, P., THOMSON-LUQUE, R., MONTEIRO, W. M. & DE LACERDA, M. V. 2015. Urban and architectural risk factors for malaria in indigenous Amazonian settlements in Brazil: a typological analysis. *Malar J,* 14**,** 284.

LEFFLER, E. M., BAND, G., BUSBY, G. B. J., KIVINEN, K., LE, Q. S., CLARKE, G. M., BOJANG, K. A., CONWAY, D. J., JALLOW, M., SISAY-JOOF, F., BOUGOUMA, E. C., MANGANO, V. D., MODIANO, D., SIRIMA, S. B., ACHIDI, E., APINJOH, T. O., MARSH, K., NDILA, C. M., PESHU, N., WILLIAMS, T. N., DRAKELEY, C., MANJURANO, A., REYBURN, H., RILEY, E., KACHALA, D., MOLYNEUX, M., NYIRONGO, V., TAYLOR, T., THORNTON, N., TILLEY, L., GRIMSLEY, S., DRURY, E., STALKER, J., CORNELIUS, V., HUBBART, C., JEFFREYS, A. E., ROWLANDS, K., ROCKETT, K. A., SPENCER, C. C. A. & KWIATKOWSKI, D. P. 2017. Resistance to malaria through structural variation of red blood cell invasion receptors. *Science,* 356.

LIEBMAN, K. A., STODDARD, S. T., REINER, R. C., JR., PERKINS, T. A., ASTETE, H., SIHUINCHA, M., HALSEY, E. S., KOCHEL, T. J., MORRISON, A. C. & SCOTT, T. W. 2014. Determinants of heterogeneous blood feeding patterns by *Aedes aegypti* in Iquitos, Peru. *PLoS Negl Trop Dis,* 8**,** e2702.

LILJANDER, A., WIKLUND, L., FALK, N., KWEKU, M., MARTENSSON, A., FELGER, I. & FARNERT, A. 2009. Optimization and validation of multi-coloured capillary electrophoresis for genotyping of *Plasmodium falciparum* merozoite surface proteins (*msp1* and 2). *Malar J,* 8**,** 78.

LWETOIJERA, D. W., KIWARE, S. S., MAGENI, Z. D., DONGUS, S., HARRIS, C., DEVINE, G. J. & MAJAMBERE, S. 2013. A need for better housing to further reduce indoor malaria transmission in areas with high bed net coverage. *Parasit Vectors,* 6**,** 57.

MACKINNON, M. J. & MARSH, K. 2010. The selection landscape of malaria parasites. *Science,* 328**,** 866-71.

MALA, A. O., IRUNGU, L. W., SHILILU, J. I., MUTURI, E. J., MBOGO, C. M., NJAGI, J. K., MUKABANA, W. R. & GITHURE, J. I. 2011. *Plasmodium falciparum* transmission and aridity: a Kenyan experience from the dry lands of Baringo and its implications for *Anopheles arabiensis* control. *Malar J,* 10**,** 121.

MALARIAGEN 2016. Genomic epidemiology of artemisinin resistant malaria. *Elife,* 5.

MANSKE, M., MIOTTO, O., CAMPINO, S., AUBURN, S., ALMAGRO-GARCIA, J., MASLEN, G., O'BRIEN, J., DJIMDE, A., DOUMBO, O., ZONGO, I., OUEDRAOGO, J. B., MICHON, P., MUELLER, I., SIBA, P., NZILA, A., BORRMANN, S., KIARA, S. M., MARSH, K., JIANG, H., SU, X. Z., AMARATUNGA, C., FAIRHURST, R., SOCHEAT, D., NOSTEN, F., IMWONG, M., WHITE, N. J., SANDERS, M., ANASTASI, E., ALCOCK, D., DRURY, E., OYOLA, S., QUAIL, M. A., TURNER, D. J., RUANO-RUBIO, V., JYOTHI, D., AMENGA-ETEGO, L., HUBBART, C., JEFFREYS, A., ROWLANDS, K., SUTHERLAND, C., ROPER, C., MANGANO, V., MODIANO, D., TAN, J. C., FERDIG, M. T., AMAMBUA-NGWA, A., CONWAY, D. J., TAKALA-HARRISON, S., PLOWE, C. V., RAYNER, J. C., ROCKETT, K. A., CLARK, T. G., NEWBOLD, C. I., BERRIMAN, M., MACINNIS, B. & KWIATKOWSKI, D. P. 2012. Analysis of *Plasmodium falciparum* diversity in natural infections by deep sequencing. *Nature,* 487**,** 375-9.

MARSH, K., FORSTER, D., WARUIRU, C., MWANGI, I., WINSTANLEY, M., MARSH, V., NEWTON, C., WINSTANLEY, P., WARN, P., PESHU, N. & ET AL. 1995. Indicators of life-threatening malaria in African children. *N Engl J Med,* 332**,** 1399-404.

MAYOR, A., BARDAJI, A., MACETE, E., NHAMPOSSA, T., FONSECA, A. M., GONZALEZ, R., MACULUVE, S., CISTERO, P., RUPEREZ, M., CAMPO, J., VALA, A., SIGAUQUE, B., JIMENEZ, A., MACHEVO, S., DE LA FUENTE, L., NHAMA, A., LUIS, L., APONTE, J. J., ACACIO, S., NHACOLO, A., CHITNIS, C., DOBANO, C., SEVENE, E., ALONSO, P. L. & MENENDEZ, C. 2015. Changing Trends in *P. falciparum* Burden, Immunity, and Disease in Pregnancy. *N Engl J Med,* 373**,** 1607-17.

MCCALL, P. J., MOSHA, F. W., NJUNWA, K. J. & SHERLOCK, K. 2001. Evidence for memorized site-fidelity in *Anopheles arabiensis*. *Trans R Soc Trop Med Hyg,* 95**,** 587-90.

MCGUIGAN, F. E. & RALSTON, S. H. 2002. Single nucleotide polymorphism detection: allelic discrimination using TaqMan. *Psychiatr Genet,* 12**,** 133-6.

MIDEGA, J. T., SMITH, D. L., OLOTU, A., MWANGANGI, J. M., NZOVU, J. G., WAMBUA, J., NYANGWESO, G., MBOGO, C. M., CHRISTOPHIDES, G. K., MARSH, K. & BEJON, P. 2012. Wind direction and proximity to larval sites determines malaria risk in Kilifi District in Kenya. *Nat Commun,* 3**,** 674.

MILES, A., IQBAL, Z., VAUTERIN, P., PEARSON, R., CAMPINO, S., THERON, M., GOULD, K., MEAD, D., DRURY, E., O'BRIEN, J., RUANO RUBIO, V., MACINNIS, B., MWANGI, J., SAMARAKOON, U., RANFORD-CARTWRIGHT, L., FERDIG, M., HAYTON, K., SU, X. Z., WELLEMS, T., RAYNER, J., MCVEAN, G. & KWIATKOWSKI, D. 2016. Indels, structural variation, and recombination drive genomic diversity in *Plasmodium falciparum*. *Genome Res,* 26**,** 1288-99.

MIOTTO, O., AMATO, R., ASHLEY, E. A., MACINNIS, B., ALMAGRO-GARCIA, J., AMARATUNGA, C., LIM, P., MEAD, D., OYOLA, S. O., DHORDA, M., IMWONG, M., WOODROW, C., MANSKE, M., STALKER, J., DRURY, E., CAMPINO, S., AMENGA-ETEGO, L., THANH, T. N., TRAN, H. T., RINGWALD, P., BETHELL, D., NOSTEN, F., PHYO, A. P., PUKRITTAYAKAMEE, S., CHOTIVANICH, K., CHUOR, C. M., NGUON, C., SUON, S., SRENG, S., NEWTON, P. N., MAYXAY, M., KHANTHAVONG, M., HONGVANTHONG, B., HTUT, Y., HAN, K. T., KYAW, M. P., FAIZ, M. A., FANELLO, C. I., ONYAMBOKO, M., MOKUOLU, O. A., JACOB, C. G., TAKALA-HARRISON, S., PLOWE, C. V., DAY, N. P., DONDORP, A. M., SPENCER, C. C., MCVEAN, G., FAIRHURST, R. M., WHITE, N. J. & KWIATKOWSKI, D. P. 2015. Genetic architecture of artemisinin-resistant *Plasmodium falciparum*. *Nat Genet,* 47**,** 226-34.

MOBEGI, V. A., DUFFY, C. W., AMAMBUA-NGWA, A., LOUA, K. M., LAMAN, E., NWAKANMA, D. C., MACINNIS, B., ASPELING-JONES, H., MURRAY, L., CLARK, T. G., KWIATKOWSKI, D. P. & CONWAY, D. J. 2014. Genome-wide analysis of selection on the malaria parasite *Plasmodium falciparum* in West African populations of differing infection endemicity. *Mol Biol Evol,* 31**,** 1490-9.

MOBEGI, V. A., LOUA, K. M., AHOUIDI, A. D., SATOGUINA, J., NWAKANMA, D. C., AMAMBUA-NGWA, A. & CONWAY, D. J. 2012. Population genetic structure of *Plasmodium falciparum* across a region of diverse endemicity in West Africa. *Malar J,* 11**,** 223.

MOGENI, P., WILLIAMS, T. N., FEGAN, G., NYUNDO, C., BAUNI, E., MWAI, K., OMEDO, I., NJUGUNA, P., NEWTON, C. R., OSIER, F., BERKLEY, J. A., HAMMITT, L. L., LOWE, B., MWAMBINGU, G., AWUONDO, K., MTURI, N., PESHU, N., SNOW, R. W., NOOR, A., MARSH, K. & BEJON, P. 2016. Age, Spatial, and Temporal Variations in Hospital Admissions with Malaria in Kilifi County, Kenya: A 25-Year Longitudinal Observational Study. *PLoS Med,* 13**,** e1002047.

MOHD ABD RAZAK, M. R., SASTU, U. R., NORAHMAD, N. A., ABDUL-KARIM, A., MUHAMMAD, A., MUNIANDY, P. K., JELIP, J., RUNDI, C., IMWONG, M., MUDIN, R. N. & ABDULLAH, N. R. 2016. Genetic Diversity of *Plasmodium falciparum* Populations in Malaria Declining Areas of Sabah, East Malaysia. *PLoS One,* 11**,** e0152415.

MPHANDE, F. A., RIBACKE, U., KANEKO, O., KIRONDE, F., WINTER, G. & WAHLGREN, M. 2008. SURFIN4.1, a schizont-merozoite associated protein in the SURFIN family of *Plasmodium falciparum*. *Malar J,* 7**,** 116.

MU, J., AWADALLA, P., DUAN, J., MCGEE, K. M., JOY, D. A., MCVEAN, G. A. & SU, X. Z. 2005. Recombination hotspots and population structure in *Plasmodium falciparum*. *PLoS Biol,* 3**,** e335.

MU, J., MYERS, R. A., JIANG, H., LIU, S., RICKLEFS, S., WAISBERG, M., CHOTIVANICH, K., WILAIRATANA, P., KRUDSOOD, S., WHITE, N. J., UDOMSANGPETCH, R., CUI, L., HO, M., OU, F., LI, H., SONG, J., LI, G., WANG, X., SEILA, S., SOKUNTHEA, S., SOCHEAT, D., STURDEVANT, D. E., PORCELLA, S. F., FAIRHURST, R. M., WELLEMS, T. E., AWADALLA, P. & SU, X. Z. 2010a. *Plasmodium falciparum* genome-wide scans for positive selection, recombination hot spots and resistance to antimalarial drugs. *Nat Genet,* 42**,** 268-71.

MU, J., SEYDEL, K. B., BATES, A. & SU, X. Z. 2010b. Recent Progress in Functional Genomic Research in *Plasmodium falciparum*. *Curr Genomics,* 11**,** 279-86.

MUELLER, I., SCHOEPFLIN, S., SMITH, T. A., BENTON, K. L., BRETSCHER, M. T., LIN, E., KINIBORO, B., ZIMMERMAN, P. A., SPEED, T. P., SIBA, P. & FELGER, I. 2012. Force of infection is key to understanding the epidemiology of *Plasmodium falciparum* malaria in Papua New Guinean children. *Proc Natl Acad Sci U S A,* 109**,** 10030-5.

MUELLER, J. C. 2004. Linkage disequilibrium for different scales and applications. *Brief Bioinform,* 5**,** 355-64.

MURPHY, K. M., BERG, K. D. & ESHLEMAN, J. R. 2005. Sequencing of genomic DNA by combined amplification and cycle sequencing reaction. *Clin Chem,* 51**,** 35-9.

MWANGANGI, J. M., MBOGO, C. M., ORINDI, B. O., MUTURI, E. J., MIDEGA, J. T., NZOVU, J., GATAKAA, H., GITHURE, J., BORGEMEISTER, C., KEATING, J. & BEIER, J. C. 2013. Shifts in malaria vector species composition and transmission dynamics along the Kenyan coast over the past 20 years. *Malar J,* 12**,** 13.

MWESIGWA, J., OKEBE, J., AFFARA, M., DI TANNA, G. L., NWAKANMA, D., JANHA, O., OPONDO, K., GRIETENS, K. P., ACHAN, J. & D'ALESSANDRO, U. 2015. On-going malaria transmission in The Gambia despite high coverage of control interventions: a nationwide cross-sectional survey. *Malar J,* 14**,** 314.

NABARRO, D. N. & TAYLER, E. M. 1998. The "roll back malaria" campaign. *Science,* 280**,** 2067-8.

NABET, C., DOUMBO, S., JEDDI, F., KONATE, S., MANCIULLI, T., FOFANA, B., L'OLLIVIER, C., CAMARA, A., MOORE, S., RANQUE, S., THERA, M. A., DOUMBO, O. K. & PIARROUX, R. 2016. Genetic diversity of *Plasmodium falciparum* in human malaria cases in Mali. *Malar J,* 15**,** 353.

NAIR, M., HINDS, M. G., COLEY, A. M., HODDER, A. N., FOLEY, M., ANDERS, R. F. & NORTON, R. S. 2002. Structure of domain III of the blood-stage malaria vaccine candidate, *Plasmodium falciparum* apical membrane antigen 1 (*AMA1*). *J Mol Biol,* 322**,** 741-53.

NAJERA, J. A., GONZALEZ-SILVA, M. & ALONSO, P. L. 2011. Some lessons for the future from the Global Malaria Eradication Programme (1955-1969). *PLoS Med,* 8**,** e1000412.

NARUM, D. L. & THOMAS, A. W. 1994. Differential localization of full-length and processed forms of *PF83/AMA-1* an apical membrane antigen of *Plasmodium falciparum* merozoites. *Mol Biochem Parasitol,* 67**,** 59-68.

NEVILL, C. G., SOME, E. S., MUNG'ALA, V. O., MUTEMI, W., NEW, L., MARSH, K., LENGELER, C. & SNOW, R. W. 1996. Insecticide-treated bednets reduce mortality and severe morbidity from malaria among children on the Kenyan coast. *Trop Med Int Health,* 1**,** 139-46.

NEWMAN, R. D., HAILEMARIAM, A., JIMMA, D., DEGIFIE, A., KEBEDE, D., RIETVELD, A. E., NAHLEN, B. L., BARNWELL, J. W., STEKETEE, R. W. & PARISE, M. E. 2003. Burden of malaria during pregnancy in areas of stable and unstable transmission in Ethiopia during a nonepidemic year. *J Infect Dis,* 187**,** 1765-72.

NGANDA, R. Y., DRAKELEY, C., REYBURN, H. & MARCHANT, T. 2004. Knowledge of malaria influences the use of insecticide treated nets but not intermittent presumptive treatment by pregnant women in Tanzania. *Malar J,* 3**,** 42.

NOOR, A. M., KINYOKI, D. K., MUNDIA, C. W., KABARIA, C. W., MUTUA, J. W., ALEGANA, V. A., FALL, I. S. & SNOW, R. W. 2014. The changing risk of *Plasmodium falciparum* malaria infection in Africa: 2000-10: a spatial and temporal analysis of transmission intensity. *Lancet,* 383**,** 1739-47.

NOOR, A. M., MUTHEU, J. J., TATEM, A. J., HAY, S. I. & SNOW, R. W. 2009. Insecticide-treated net coverage in Africa: mapping progress in 2000-07. *Lancet,* 373**,** 58-67.

NYACHIEO, A., C, V. A. N. O., LAURENT, T., DUJARDIN, J. C. & D'ALESSANDRO, U. 2005. *Plasmodium falciparum* genotyping by microsatellites as a method to distinguish between recrudescent and new infections. *Am J Trop Med Hyg,* 73**,** 210-3.

O'MEARA, W. P., BEJON, P., MWANGI, T. W., OKIRO, E. A., PESHU, N., SNOW, R. W., NEWTON, C. R. & MARSH, K. 2008. Effect of a fall in malaria transmission on morbidity and mortality in Kilifi, Kenya. *Lancet,* 372**,** 1555-62.

OCHOLA-OYIER, L. I., OKOMBO, J., WAGATUA, N., OCHIENG, J., TETTEH, K. K., FEGAN, G., BEJON, P. & MARSH, K. 2016. Comparison of allele frequencies of *Plasmodium falciparum* merozoite antigens in malaria infections sampled in different years in a Kenyan population. *Malar J,* 15**,** 261.

OCHOLA, L. I., TETTEH, K. K., STEWART, L. B., RIITHO, V., MARSH, K. & CONWAY, D. J. 2010. Allele frequency-based and polymorphism-versus-divergence indices of balancing selection in a new filtered set of polymorphic genes in *Plasmodium falciparum*. *Mol Biol Evol,* 27**,** 2344-51.

OKIRO, E. A., ALEGANA, V. A., NOOR, A. M., MUTHEU, J. J., JUMA, E. & SNOW, R. W. 2009. Malaria paediatric hospitalization between 1999 and 2008 across Kenya. *BMC Med,* 7**,** 75.

OLOTU, A., FEGAN, G., WAMBUA, J., NYANGWESO, G., LEACH, A., LIEVENS, M., KASLOW, D. C., NJUGUNA, P., MARSH, K. & BEJON, P. 2016. Seven-Year Efficacy of RTS,S/AS01 Malaria Vaccine among Young African Children. *N Engl J Med,* 374**,** 2519-29.

OLOTU, A., FEGAN, G., WILLIAMS, T. N., SASI, P., OGADA, E., BAUNI, E., WAMBUA, J., MARSH, K., BORRMANN, S. & BEJON, P. 2010. Defining clinical malaria: the specificity and incidence of endpoints from active and passive surveillance of children in rural Kenya. *PLoS One,* 5**,** e15569.

OMEDO, I., MOGENI, P., BOUSEMA, J. T., ROCKETT, K., AMAMBUA-NGWA, A., OYIER, I., STEVENSON, J., BAIDJOE, A., DE VILLIERS, E. P., FEGAN, G., ROSS, A., HUBBART, C., JEFFREYS, A., WILLIAMS, T. N., KWIATKOWSKI, D. & BEJON, P. 2017a. Micro-epidemiological structuring of *Plasmodium falciparum* parasite populations in regions with varying transmission intensities in Africa. [version 1; referees: 3 approved] *Wellcome Open Research,* 2.

OMEDO, I., MOGENI, P., ROCKETT, K., KAMAU, A., HUBBART, C., JEFFREYS, A., OCHOLA, L. I., DE VILLIERS, E. P., GITONGA, C. W., NOOR, A., SNOW, R. W., KWIATKOWSKI, D. & BEJON, P. 2017b. Geographic-genetic analysis of *Plasmodium falciparum* parasite populations from surveys of primary school children in Western Kenya [version 1; referees: awaiting peer review]. *Wellcome Open Research,* 2.

ORD, R. L., TAMI, A. & SUTHERLAND, C. J. 2008. *ama1* genes of sympatric *Plasmodium vivax* and *P. falciparum* from Venezuela differ significantly in genetic diversity and recombination frequency. *PLoS One,* 3**,** e3366.

OSIER, F. H., WEEDALL, G. D., VERRA, F., MURUNGI, L., TETTEH, K. K., BULL, P., FABER, B. W., REMARQUE, E., THOMAS, A., MARSH, K. & CONWAY, D. J. 2010. Allelic diversity and naturally acquired allele-specific antibody responses to *Plasmodium falciparum* apical membrane antigen 1 in Kenya. *Infect Immun,* 78**,** 4625-33.

OTOTO, E. N., MBUGI, J. P., WANJALA, C. L., ZHOU, G., GITHEKO, A. K. & YAN, G. 2015. Surveillance of malaria vector population density and biting behaviour in western Kenya. *Malar J,* 14**,** 244.

OWUSU-AGYEI, S., SMITH, T., BECK, H. P., AMENGA-ETEGO, L. & FELGER, I. 2002. Molecular epidemiology of *Plasmodium falciparum* infections among asymptomatic inhabitants of a holoendemic malarious area in northern Ghana. *Trop Med Int Health,* 7**,** 421-8.

OYEBOLA, M. K., IDOWU, E. T., NYANG, H., OLUKOSI, Y. A., OTUBANJO, O. A., NWAKANMA, D. C., AWOLOLA, S. T. & AMAMBUA-NGWA, A. 2014. Microsatellite markers reveal low levels of population sub-structuring of *Plasmodium falciparum* in southwestern Nigeria. *Malar J,* 13**,** 493.

PACKARD, R. M. 2014. The origins of antimalarial-drug resistance. *N Engl J Med,* 371**,** 397-9.

PATEL, J. C., TAYLOR, S. M., JULIAO, P. C., PAROBEK, C. M., JANKO, M., GONZALEZ, L. D., ORTIZ, L., PADILLA, N., TSHEFU, A. K., EMCH, M., UDHAYAKUMAR, V., LINDBLADE, K. & MESHNICK, S. R. 2014. Genetic Evidence of Importation of Drug-Resistant *Plasmodium* *falciparum* to Guatemala from the Democratic Republic of the Congo. *Emerg Infect Dis,* 20**,** 932-40.

PAYNE, D. 1987. Spread of chloroquine resistance in *Plasmodium falciparum*. *Parasitol Today,* 3**,** 241-6.

PERKINS, T. A., SCOTT, T. W., LE MENACH, A. & SMITH, D. L. 2013. Heterogeneity, mixing, and the spatial scales of mosquito-borne pathogen transmission. *PLoS Comput Biol,* 9**,** e1003327.

PLUESS, B., TANSER, F. C., LENGELER, C. & SHARP, B. L. 2010. Indoor residual spraying for preventing malaria. *Cochrane Database Syst Rev***,** Cd006657.

POLLEY, S. D., CHOKEJINDACHAI, W. & CONWAY, D. J. 2003. Allele frequency-based analyses robustly map sequence sites under balancing selection in a malaria vaccine candidate antigen. *Genetics,* 165**,** 555-61.

POLLEY, S. D. & CONWAY, D. J. 2001. Strong diversifying selection on domains of the *Plasmodium falciparum apical membrane antigen 1* gene. *Genetics,* 158**,** 1505-12.

POLLEY, S. D., CONWAY, D. J., CAVANAGH, D. R., MCBRIDE, J. S., LOWE, B. S., WILLIAMS, T. N., MWANGI, T. W. & MARSH, K. 2006. High levels of serum antibodies to merozoite surface protein 2 of *Plasmodium falciparum* are associated with reduced risk of clinical malaria in coastal Kenya. *Vaccine,* 24**,** 4233-46.

POURMAND, N., ELAHI, E., DAVIS, R. W. & RONAGHI, M. 2002. Multiplex Pyrosequencing. *Nucleic Acids Res,* 30**,** e31.

PROIETTI, C., PETTINATO, D. D., KANOI, B. N., NTEGE, E., CRISANTI, A., RILEY, E. M., EGWANG, T. G., DRAKELEY, C. & BOUSEMA, T. 2011. Continuing intense malaria transmission in northern Uganda. *Am J Trop Med Hyg,* 84**,** 830-7.

PUMPAIBOOL, T., ARNATHAU, C., DURAND, P., KANCHANAKHAN, N., SIRIPOON, N., SUEGORN, A., SITTHI-AMORN, C., RENAUD, F. & HARNYUTTANAKORN, P. 2009. Genetic diversity and population structure of *Plasmodium falciparum* in Thailand, a low transmission country. *Malar J,* 8**,** 155.

R CORE TEAM 2013. R: A language and environment for statistical computing. R Foundation for Statistical Computing. Vienna, Austria.

REMARQUE, E. J., FABER, B. W., KOCKEN, C. H. & THOMAS, A. W. 2008. Apical membrane antigen 1: a malaria vaccine candidate in review. *Trends Parasitol,* 24**,** 74-84.

REYBURN, H., MBATIA, R., DRAKELEY, C., BRUCE, J., CARNEIRO, I., OLOMI, R., COX, J., NKYA, W. M., LEMNGE, M., GREENWOOD, B. M. & RILEY, E. M. 2005. Association of transmission intensity and age with clinical manifestations and case fatality of severe *Plasmodium falciparum* malaria. *Jama,* 293**,** 1461-70.

RINGNÉR, M. 2008. What is principal component analysis? *Nature Biotechnology,* 26**,** 303 - 304.

ROETZER, A., DIEL, R., KOHL, T. A., RUCKERT, C., NUBEL, U., BLOM, J., WIRTH, T., JAENICKE, S., SCHUBACK, S., RUSCH-GERDES, S., SUPPLY, P., KALINOWSKI, J. & NIEMANN, S. 2013. Whole genome sequencing versus traditional genotyping for investigation of a *Mycobacterium tuberculosis* outbreak: a longitudinal molecular epidemiological study. *PLoS Med,* 10**,** e1001387.

ROZAS, J. 2009. DNA sequence polymorphism analysis using DnaSP. *Methods Mol Biol,* 537**,** 337-50.

RTS'S CLINICAL TRIALS PARTNERSHIP 2015. Efficacy and safety of RTS,S/AS01 malaria vaccine with or without a booster dose in infants and children in Africa: final results of a phase 3, individually randomised, controlled trial. *Lancet,* 386**,** 31-45.

RUTLEDGE, G. G., BOHME, U., SANDERS, M., REID, A. J., COTTON, J. A., MAIGA-ASCOFARE, O., DJIMDE, A. A., APINJOH, T. O., AMENGA-ETEGO, L., MANSKE, M., BARNWELL, J. W., RENAUD, F., OLLOMO, B., PRUGNOLLE, F., ANSTEY, N. M., AUBURN, S., PRICE, R. N., MCCARTHY, J. S., KWIATKOWSKI, D. P., NEWBOLD, C. I., BERRIMAN, M. & OTTO, T. D. 2017. *Plasmodium malariae* and *P. ovale* genomes provide insights into malaria parasite evolution. *Nature,* 542**,** 101-104.

SAUNDERS, D. L., VANACHAYANGKUL, P. & LON, C. 2014. Dihydroartemisinin-piperaquine failure in Cambodia. *N Engl J Med,* 371**,** 484-5.

SCHLOTTERER, C. 2002. Towards a molecular characterization of adaptation in local populations. *Curr Opin Genet Dev,* 12**,** 683-7.

SCHULTZ, L., WAPLING, J., MUELLER, I., NTSUKE, P. O., SENN, N., NALE, J., KINIBORO, B., BUCKEE, C. O., TAVUL, L., SIBA, P. M., REEDER, J. C. & BARRY, A. E. 2010. Multilocus haplotypes reveal variable levels of diversity and population structure of *Plasmodium falciparum* in Papua New Guinea, a region of intense perennial transmission. *Malar J,* 9**,** 336.

SCOTT, J. A., BAUNI, E., MOISI, J. C., OJAL, J., GATAKAA, H., NYUNDO, C., MOLYNEUX, C. S., KOMBE, F., TSOFA, B., MARSH, K., PESHU, N. & WILLIAMS, T. N. 2012. Profile: The Kilifi Health and Demographic Surveillance System (KHDSS). *Int J Epidemiol,* 41**,** 650-7.

SMITH, D. L., BATTLE, K. E., HAY, S. I., BARKER, C. M., SCOTT, T. W. & MCKENZIE, F. E. 2012. Ross, macdonald, and a theory for the dynamics and control of mosquito-transmitted pathogens. *PLoS Pathog,* 8**,** e1002588.

SMITH, D. L., MCKENZIE, F. E., SNOW, R. W. & HAY, S. I. 2007. Revisiting the basic reproductive number for malaria and its implications for malaria control. *PLoS Biol,* 5**,** e42.

SMITH, T., BECK, H. P., KITUA, A., MWANKUSYE, S., FELGER, I., FRASER-HURT, N., IRION, A., ALONSO, P., TEUSCHER, T. & TANNER, M. 1999. Age dependence of the multiplicity of *Plasmodium falciparum* infections and of other malariological indices in an area of high endemicity. *Trans R Soc Trop Med Hyg,* 93 Suppl 1**,** 15-20.

SNOW, R. W., KIBUCHI, E., KARURI, S. W., SANG, G., GITONGA, C. W., MWANDAWIRO, C., BEJON, P. & NOOR, A. M. 2015. Changing Malaria Prevalence on the Kenyan Coast since 1974: Climate, Drugs and Vector Control. *PLoS One,* 10**,** e0128792.

SNOW, R. W. & MARSH, K. 2002. The consequences of reducing transmission of *Plasmodium falciparum* in Africa. *Adv Parasitol,* 52**,** 235-64.

SONKO, S. T., JAITEH, M., JAFALI, J., JARJU, L. B., D'ALESSANDRO, U., CAMARA, A., KOMMA-BAH, M. & SAHO, A. 2014. Does socio-economic status explain the differentials in malaria parasite prevalence? Evidence from The Gambia. *Malar J,* 13**,** 449.

SOULAMA, I., SERME, S. S., BOUGOUMA, E. C., DIARRA, A., TIONO, A. B., OUEDRAOGO, A., KONATE, A. T., NEBIE, I. & SIRIMA, S. B. 2015. Clinical Variation of *Plasmodium falciparum eba-175*, *ama-1*, and *msp-3* Genotypes in Young Children Living in a Seasonally High Malaria Transmission Setting in Burkina Faso. *J Parasitol Res,* 2015**,** 985651.

SRINIVASAN, P., BEATTY, W. L., DIOUF, A., HERRERA, R., AMBROGGIO, X., MOCH, J. K., TYLER, J. S., NARUM, D. L., PIERCE, S. K., BOOTHROYD, J. C., HAYNES, J. D. & MILLER, L. H. 2011. Binding of *Plasmodium* merozoite proteins RON2 and *AMA1* triggers commitment to invasion. *Proc Natl Acad Sci U S A,* 108**,** 13275-80.

STEVENSON, J., ST LAURENT, B., LOBO, N. F., COOKE, M. K., KAHINDI, S. C., ORIANGO, R. M., HARBACH, R. E., COX, J. & DRAKELEY, C. 2012. Novel vectors of malaria parasites in the western highlands of Kenya. *Emerg Infect Dis,* 18**,** 1547-9.

STEVENSON, J. C., STRESMAN, G. H., GITONGA, C. W., GILLIG, J., OWAGA, C., MARUBE, E., ODONGO, W., OKOTH, A., CHINA, P., ORIANGO, R., BROOKER, S. J., BOUSEMA, T., DRAKELEY, C. & COX, J. 2013. Reliability of school surveys in estimating geographic variation in malaria transmission in the western Kenyan highlands. *PLoS One,* 8**,** e77641.

SUTCLIFFE, C. G., KOBAYASHI, T., HAMAPUMBU, H., SHIELDS, T., KAMANGA, A., MHARAKURWA, S., THUMA, P. E., GLASS, G. & MOSS, W. J. 2011. Changing individual-level risk factors for malaria with declining transmission in southern Zambia: a cross-sectional study. *Malar J,* 10**,** 324.

TAJIMA, F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics,* 123**,** 585-95.

TAKALA, S. L., COULIBALY, D., THERA, M. A., BATCHELOR, A. H., CUMMINGS, M. P., ESCALANTE, A. A., OUATTARA, A., TRAORE, K., NIANGALY, A., DJIMDE, A. A., DOUMBO, O. K. &

PLOWE, C. V. 2009. Extreme polymorphism in a vaccine antigen and risk of clinical malaria: implications for vaccine development. *Sci Transl Med,* 1**,** 2ra5.

TAKEM, E. N. & D'ALESSANDRO, U. 2013. Malaria in pregnancy. *Mediterr J Hematol Infect Dis,* 5**,** e2013010.

TARAZONA-SANTOS, E., CASTILHO, L., AMARAL, D. R., COSTA, D. C., FURLANI, N. G., ZUCCHERATO, L. W., MACHADO, M., REID, M. E., ZALIS, M. G., ROSSIT, A. R., SANTOS, S. E., MACHADO, R. L. & LUSTIGMAN, S. 2011. Population genetics of *GYPB* and association study between GYPB*S/s polymorphism and susceptibility to *P. falciparum* infection in the Brazilian Amazon. *PLoS One,* 6**,** e16123.

TEKLEHAIMANOT, H. D., TEKLEHAIMANOT, A., KISZEWSKI, A., RAMPAO, H. S. & SACHS, J. D. 2009. Malaria in Sao Tome and principe: on the brink of elimination after three years of effective antimalarial measures. *Am J Trop Med Hyg,* 80**,** 133-40.

TESSEMA, S. K., MONK, S. L., SCHULTZ, M. B., TAVUL, L., REEDER, J. C., SIBA, P. M., MUELLER, I. & BARRY, A. E. 2015. Phylogeography of *var* gene repertoires reveals fine-scale geospatial clustering of *Plasmodium falciparum* populations in a highly endemic area. *Mol Ecol,* 24**,** 484-97.

THE MALERA CONSULTATIVE GROUP ON VECTOR CONTROL 2011. A research agenda for malaria eradication: vector control. *PLoS Med,* 8**,** e1000401.

THERA, M. A., COULIBALY, D., KONE, A. K., GUINDO, A. B., TRAORE, K., SALL, A. H., DIARRA, I., DAOU, M., TRAORE, I. M., TOLO, Y., SISSOKO, M., NIANGALY, A., ARAMA, C., BABY, M., KOURIBA, B., SISSOKO, M. S., SAGARA, I., TOURE, O. B., DOLO, A., DIALLO, D. A., REMARQUE, E., CHILENGI, R., NOOR, R., SESAY, S., THOMAS, A., KOCKEN, C. H., FABER, B. W., IMOUKHUEDE, E. B., LEROY, O. & DOUMBO, O. K. 2016. Phase 1 randomized controlled trial to evaluate the safety and immunogenicity of recombinant *Pichia pastoris*-expressed *Plasmodium falciparum* apical membrane antigen 1 (*PfAMA1-FVO* [25-545]) in healthy Malian adults in Bandiagara. *Malar J,* 15**,** 442.

THERA, M. A., DOUMBO, O. K., COULIBALY, D., DIALLO, D. A., KONE, A. K., GUINDO, A. B., TRAORE, K., DICKO, A., SAGARA, I., SISSOKO, M. S., BABY, M., SISSOKO, M., DIARRA, I., NIANGALY, A., DOLO, A., DAOU, M., DIAWARA, S. I., HEPPNER, D. G., STEWART, V. A., ANGOV, E., BERGMANN-LEITNER, E. S., LANAR, D. E., DUTTA, S., SOISSON, L., DIGGS, C. L., LEACH, A., OWUSU, A., DUBOIS, M. C., COHEN, J., NIXON, J. N., GREGSON, A., TAKALA, S. L., LYKE, K. E. & PLOWE, C. V. 2008. Safety and immunogenicity of an *AMA-1* malaria vaccine in Malian adults: results of a phase 1 randomized controlled trial. *PLoS One,* 3**,** e1465.

THERA, M. A., DOUMBO, O. K., COULIBALY, D., LAURENS, M. B., OUATTARA, A., KONE, A. K., GUINDO, A. B., TRAORE, K., TRAORE, I., KOURIBA, B., DIALLO, D. A., DIARRA, I., DAOU, M., DOLO, A., TOLO, Y., SISSOKO, M. S., NIANGALY, A., SISSOKO, M., TAKALA-HARRISON, S., LYKE, K. E., WU, Y., BLACKWELDER, W. C., GODEAUX, O., VEKEMANS, J., DUBOIS, M. C., BALLOU, W. R., COHEN, J., THOMPSON, D., DUBE, T., SOISSON, L., DIGGS, C. L., HOUSE, B., LANAR, D. E., DUTTA, S., HEPPNER, D. G., JR. & PLOWE, C. V. 2011. A field trial to assess a blood-stage malaria vaccine. *N Engl J Med,* 365**,** 1004-13.

TUSTING, L. S., BOUSEMA, T., SMITH, D. L. & DRAKELEY, C. 2014. Measuring changes in *Plasmodium falciparum* transmission: precision, accuracy and costs of metrics. *Adv Parasitol,* 84**,** 151-208.

UTZINGER, J., TOZAN, Y. & SINGER, B. H. 2001. Efficacy and cost-effectiveness of environmental management for malaria control. *Trop Med Int Health,* 6**,** 677-87.

VERHULST, N. O., BEIJLEVELD, H., QIU, Y. T., MALIEPAARD, C., VERDUYN, W., HAASNOOT, G. W., CLAAS, F. H., MUMM, R., BOUWMEESTER, H. J., TAKKEN, W., VAN LOON, J. J. & SMALLEGANGE, R. C. 2013. Relation between HLA genes, human skin volatiles and attractiveness of humans to malaria mosquitoes. *Infect Genet Evol,* 18**,** 87-93.

VOLKMAN, S. K., SABETI, P. C., DECAPRIO, D., NEAFSEY, D. E., SCHAFFNER, S. F., MILNER, D. A., JR., DAILY, J. P., SARR, O., NDIAYE, D., NDIR, O., MBOUP, S., DURAISINGH, M. T., LUKENS, A., DERR, A., STANGE-THOMANN, N., WAGGONER, S., ONOFRIO, R., ZIAUGRA, L., MAUCELI, E., GNERRE, S., JAFFE, D. B., ZAINOUN, J., WIEGAND, R. C., BIRREN, B. W., HARTL, D. L., GALAGAN, J. E., LANDER, E. S. & WIRTH, D. F. 2007. A genome-wide map of diversity in *Plasmodium falciparum*. *Nat Genet,* 39**,** 113-9.

WANZIRAH, H., TUSTING, L. S., ARINAITWE, E., KATUREEBE, A., MAXWELL, K., REK, J., BOTTOMLEY, C., STAEDKE, S. G., KAMYA, M., DORSEY, G. & LINDSAY, S. W. 2015. Mind the gap: house structure and the risk of malaria in Uganda. *PLoS One,* 10**,** e0117396.

WATERHOUSE, A. M., PROCTER, J. B., MARTIN, D. M., CLAMP, M. & BARTON, G. J. 2009. Jalview Version 2--a multiple sequence alignment editor and analysis workbench. *Bioinformatics,* 25**,** 1189-91.

WESOLOWSKI, A., EAGLE, N., TATEM, A. J., SMITH, D. L., NOOR, A. M., SNOW, R. W. & BUCKEE, C. O. 2012. Quantifying the impact of human mobility on malaria. *Science,* 338**,** 267-70.

WHITE, N. J. 2004. Antimalarial drug resistance. *J Clin Invest,* 113**,** 1084-92.

WHITE, N. J. 2005. Intermittent presumptive treatment for malaria. *PLoS Med,* 2**,** e3.

WHO 2008. Methods and techniques for clinical trials on antimalarial drug efficacy: genotyping to identify parasite populations. Informal consultation organized by the Medicines for Malaria Venture and cosponsored by the World Health Organization, 29–31 May Amsterdam, The Netherlands. Medicines for Malaria Venture and World Health Organization.

WHO 2014. Severe malaria. *Tropical medicine & international health* 19**,** 7-131.

WHO 2015. Annex 6 Treatment of *Plasmoium vivax*, *P. ovale*, *P. malariae* and *P. knowlesi* infections in: Guidelines for Treatment of Malaria. 3rd edition. .

WHO 2016. World Malaria Report.

WINTER, G., KAWAI, S., HAEGGSTROM, M., KANEKO, O., VON EULER, A., KAWAZU, S., PALM, D., FERNANDEZ, V. & WAHLGREN, M. 2005. SURFIN is a polymorphic antigen expressed on *Plasmodium falciparum* merozoites and infected erythrocytes. *J Exp Med,* 201**,** 1853-63.

WISER, M. F. 2017. *Cellular and molecular biology of Plasmodium* [Online]. Available: http://www.tulane.edu/~wiser/malaria/cmb.html [Accessed 26/05/2017].

WOOLHOUSE, M. E., DYE, C., ETARD, J. F., SMITH, T., CHARLWOOD, J. D., GARNETT, G. P., HAGAN, P., HII, J. L., NDHLOVU, P. D., QUINNELL, R. J., WATTS, C. H., CHANDIWANA, S. K. & ANDERSON, R. M. 1997. Heterogeneities in the transmission of infectious agents: implications for the design of control programs. *Proc Natl Acad Sci U S A,* 94**,** 338-42.

WOOTTON, J. C., FENG, X., FERDIG, M. T., COOPER, R. A., MU, J., BARUCH, D. I., MAGILL, A. J. & SU, X. Z. 2002. Genetic diversity and chloroquine selective sweeps in *Plasmodium falciparum*. *Nature,* 418**,** 320-3.

XANGSAYARATH, P., KAEWTHAMASORN, M., YAHATA, K., NAKAZAWA, S., SATTABONGKOT, J., UDOMSANGPETCH, R. & KANEKO, O. 2012. Positive diversifying selection on the *Plasmodium falciparum surf4.1* gene in Thailand. *Trop Med Health,* 40**,** 79-89.

YUKICH, J., BRIET, O., BRETSCHER, M. T., BENNETT, A., LEMMA, S., BERHANE, Y., EISELE, T. P., KEATING, J. & SMITH, T. 2012. Estimating *Plasmodium falciparum* transmission rates in low-endemic settings using a combination of community prevalence and health facility data. *PLoS One,* 7**,** e42861.

ZHONG, D., AFRANE, Y., GITHEKO, A., YANG, Z., CUI, L., MENGE, D. M., TEMU, E. A. & YAN, G. 2007. *Plasmodium falciparum* genetic diversity in western Kenya highlands. *Am J Trop Med Hyg,* 77**,** 1043-50.

ZHOU, G., AFRANE, Y. A., VARDO-ZALIK, A. M., ATIELI, H., ZHONG, D., WAMAE, P., HIMEIDAN, Y. E., MINAKAWA, N., GITHEKO, A. K. & YAN, G. 2011. Changing patterns of malaria epidemiology between 2002 and 2010 in Western Kenya: the fall and rise of malaria. *PLoS One,* 6**,** e20318.

ZHOU, G., LEE, M. C., GITHEKO, A. K., ATIELI, H. E. & YAN, G. 2016. Insecticide-Treated Net Campaign and Malaria Transmission in Western Kenya: 2003-2015. *Front Public Health,* 4**,** 153.

ZHU, X., ZHAO, Z., FENG, Y., LI, P., LIU, F., LIU, J., YANG, Z., YAN, G., FAN, Q., CAO, Y. & CUI, L. 2016. Genetic diversity of the *Plasmodium falciparum* apical membrane antigen I gene in parasite population from the China-Myanmar border area. *Infect Genet Evol,* 39**,** 155-62.

Chapter 7 Appendix

Appendix Table 1: SNPs and genes typed in The Gambia, Kilifi and Rachuonyo South *P. falciparum* parasite population.

gene product	Gene symbol	Chr	No.
Dehydrofolate reductase	DHFR	4	3
dihydropteroate synthetase	DHPS	8	1
Erythrocyte Binding Antigen - 165	EBA165	4	1
Hypothetical protein	MAL13P1_147	13	1
myosin C (MyoC)	MAL13P1_148	13	1
DNA repair protein RAD5, putative (RAD5)	MAL13P1_216	13	1
conserved Plasmodium protein, unknown function	MAL13P1_234	13	1
step II splicing factor, putative	MAL13P1_242	13	1
guanylyl cyclase beta (GCbeta)	MAL13P1_301	13	1
exonuclease, putative	MAL13P1_311	13	1
mitochondrial ribosomal protein L9 precursor, putative	MAL13P1_318	13	1
DNA repair endonuclease, putative	MAL13P1_346	13	1
Hypothetical protein	MAL13P1_380	13	1
conserved Plasmodium protein, unknown function	MAL13P1_39	13	1
conserved Plasmodium protein, unknown function	MAL13P1_88	13	1
Hypothetical protein	MAL7P1_17	7	1
EMP1-trafficking protein	MAL7P1_172	7	3
Plasmodium exported protein, unknown function	MAL7P1_173	7	1
erythrocyte binding antigen -175	MAL7P1_176	7	52
Plasmodium exported protein (hyp9), unknown function	MAL7P1_177	7	1
alpha/beta hydrolase, putative (GEXP08)	MAL7P1_178	7	4
DNA mismatch repair protein MSH2, putative	MAL7P1_206	7	2
Chloroquine resistance transporter (CRT)	MAL7P1_27	7	1
Regulator of chloroquine condensation, putative	MAL7P1_38	7	1
zinc finger protein, putative (ERF2)	MAL7P1_68	7	1
Conserved Plasmodium protein, unkown function	MAL7P1_97	7	1
protein phosphatase, putative	MAL8P1_109	8	1
peptidase family C50, putative	MAL8P1_113	8	1
tyrosine--tRNA ligase (TyrRS)	MAL8P1_125	8	1
Hypothetical protein	MAL8P1_139	8	1
E3 ubiquitin-protein ligase, putative	MAL8P1_23	8	1
Hypothetical protein	MAL8P1_29	8	1
small heat shock protein HSP20, putative (HSP20)	MAL8P1_78	8	1
Multidrug resistance Protein 1	MDR1	5	1
Merozoite surface protein 1	MSP-1	9	1
Plasmodium exported protein, unknown function	PF07_0004	7	1
Conserved Plasmodium protein, unkown function	PF07_0044	7	1
Hypothetical protein	PF07_0053	7	2
RNA binding protein, putative	PF07_0066	7	1
acyl-CoA synthetase (ACS5)	PF07_0129	7	4
Surface-associated interspersed protein 8.2 (SURFIN 8.2)	PF08_0002	8	5
GDP-mannose 4,6-dehydratase, putative	PF08_0077	8	1
Hypothetical protein	PF08_0091	8	1

dihydropteroate synthetase (DHPS)	PF08_0095	8	1
inositol polyphosphate kinase, putative (IPK1)	PF10_0078	10	1
protein phosphatase, putative	PF10_0124	10	1
Hypothetical protein	PF10_0205	10	1
Hypothetical protein	PF10_0212	10	1
4-hydroxy-3-methylbut-2-en-1-yl diphosphate synthase (GcpE)	PF10_0221	10	1
Hypothetical protein	PF10_0292	10	1
merozoite surface protein (H101)	PF10_0347	10	1
syntaxin, Qa-SNARE family (SYN13)	PF11_0052	11	1
vacuolar sorting protein 35, putative	PF11_0112	11	1
mitogen-activated protein kinase 2 (MAP2)	PF11_0147	11	1
cysteine proteinase falcipain 3 (FP3)	PF11_0162	11	1
folate transporter 2 (FT2)	PF11_0172	11	1
DNA mismatch repair protein MLH (MLH)	PF11_0184	11	1
Hypothetical protein	PF11_0185	11	2
Hypothetical protein	PF11_0342	11	4
Apical Membrane Antigen 1	PF11_0344	11	9
conserved Plasmodium protein, unknown function	PF11_0345	11	1
Hypothetical protein	PF11_0347	11	2
RNA (uracil-5-)methyltransferase, putative	PF11_0348	11	2
Hypothetical protein	PF11_0349	11	2
heat shock protein 70 (Hsp70-3)	PF11_0351	11	2
Conserved protein, unknown function	PF11_0353	11	1
Farnesyltransferase beta subunit, putative	PF11_0483	11	1
Conserved protein, unknown function	PF11_0528	11	1
Hypothetical protein	PF13_0018	13	1
sodium/hydrogen exchanger, Na , H antiporter (NHE)	PF13_0019	13	1
membrane integral peptidase, M50 family, putative	PF13_0028	13	1
Surface-associated interspersed protein 13.1 (SURFIN 13.1)	PF13_0075	13	1
U4/U6.U5 tri-snRNP-associated protein 2, putative (USP39)	PF13_0096	13	1
Hypothetical protein	PF13_0237	13	1
transcription factor with AP2 domain(s) (ApiAP2)	PF13_0267	13	1
Hypothetical protein	PF13_0352	13	1
mitochondrial import inner membrane translocase subunit Tim9, putative (TIM9)	PF13_0358	13	1
Hypothetical protein	PF14_0045	14	1
conserved Plasmodium protein, unknown function	PF14_0046	14	1
conserved Plasmodium protein, unknown function	PF14_0047	14	1
GTPase-activating protein, putative	PF14_0048	14	2
DNA mismatch repair protein, putative	PF14_0051	14	2
COBW domain-containing protein 1, putative (CBWD1)	PF14_0052	14	2
ribonucleotide reductase small subunit (RNR)	PF14_0053	14	1
Hypothetical protein	PF14_0054	14	1
conserved Plasmodium protein, unknown function	PF14_0093	14	1
WD repeat-containing protein, putative	PF14_0101	14	1
Hypothetical protein	PF14_0152	14	3
Hypothetical protein	PF14_0153	14	2

Hypothetical protein	PF14_0154	14	1
liver specific protein 1, putative (LISP1)	PF14_0179	14	1
DNA-directed DNA polymerase, putative	PF14_0234	14	1
ataxin-2 like protein, putative	PF14_0338	14	1
apicoplast 1-acyl-sn-glycerol-3-phosphate acyltransferase, putative (AGPAT)	PF14_0421	14	1
histone deacetylase, putative	PF14_0690	14	1
Trailer hitch homolog, putative (CITH)	PF14_0717	14	1
Plasmodium exported protein (PHISTb), unkown function	PF14_0746	14	1
surface-associated interspersed protein 14.1 (SURFIN 14.1)	PF14_0747	14	1
calcium-transporting ATPase (ATP6)	PFA0310c	1	1
conserved Plasmodium protein, unknown function	PFA0410w	1	1
conserved Plasmodium protein, unknown function (GEXP19)	PFA0550w	1	1
multidrug resistance-associated protein 1	PFA0590w	1	1
surface-associated interspersed protein 1.1 (SURFIN 1.1)	PFA0625w	1	1
knob-associated histidine-rich protein (KAHRP)	PFB0100c	2	1
DNA repair endonuclease, putative	PFB0265c	2	1
Merozoite surface protein 5	PFB0305c-a	2	1
ABC transporter B family member 4, putative (ABCB4)	PFC0125w	3	1
DNA-directed RNA polymerase subunit I, putative	PFC0155c	3	1
conserved Plasmodium protein, unknown function	PFC0325c	3	1
Hypothetical protein	PFC0345w	3	1
T-complex protein 1 subunit (CCT7)	PFC0350c	3	1
phosphoglycerate mutase, putative	PFC0430w	3	1
formate-nitrite transporter, putative	PFC0725c	3	1
conserved Plasmodium protein, unknown function	PFC0790w	3	1
surface-associated interspersed protein 4.1 (SURFIN 4.1)	PFD0100c	4	4
Hypothetical protein	PFD0320c	4	1
Hypothetical protein	PFD0340c	4	1
DEAD box ATP-dependent RNA helicase, putative	PFD0565c	4	1
DNA polymerase alpha	PFD0590c	4	2
Phosphoglucomutase-2 (PGM2)	PFD0660w	4	1
apicoplast ribosomal protein L10 precursor, putative	PFD0675w	4	1
Conserved protein, unkown function	PFD0705c	4	1
conserved Plasmodium protein, unknown function	PFD0735c	4	1
Hypothetical protein	PFD0860w	4	1
reticulocyte binding protein homologue 5 (RH5)	PFD1145C	4	3
Surface-associated interspersed protein 4.2 (SURFIN 4.2)	PFD1160w	4	4
Hypothetical protein	PFE0230w	5	1
guanidine nucleotide exchange factor (RCC1)	PFE0420c	5	1
Pre-mRNA-processing ATP-dependent RNA helicase prp5, putative (PRP5)	PFE0430w	5	1
RNA pseudouridylate synthase, putative	PFE0570w	5	1
cation-transporting ATPase 1 (ATPase1)	PFE0805w	5	2
ATP-dependent RNA helicase DDX23, putative (DDX23)	PFE0925c	5	1
Hypothetical protein	PFE1015c	5	1
Hypothetical protein	PFE1045c	5	1

Hypothetical protein	PFE1095w	5	1
ubiquitin carboxyl-terminal hydrolase, putative	PFE1355c	5	1
Conserved Plasmodium protein, unkown function	PFE1520c	5	1
ATP-dependent RNA helicase, putative	PFF0100w	6	1
conserved Plasmodium protein, unknown function	PFF0175c	6	1
phenylalanyl-tRNA synthetase subunit, putative	PFF0180w	6	1
DNA helicase, putative	PFF0225w	6	1
nucleoside diphosphate kinase, putative	PFF0275c	6	1
Conserved Plasmodium protein, unkown function	PFF0325c	6	1
translation initiation factor IF-2, putative	PFF0345w	6	1
Hypothetical protein	PFF0480w	6	1
6-cysteine protein (P12)	PFF0615c	6	1
Hypothetical protein	PFF0660w	6	1
Hypothetical protein	PFF0695w	6	1
Hypothetical protein	PFF0990c	6	1
SNF2 helicase, putative (ISWI)	PFF1185w	6	1
DNA polymerase 1, putative	PFF1225c	6	1
Plasmodium exported protein, unknown function	PFI0086w	9	1
serine/threonine protein kinase, FIKK family (FIKK9.3)	PFI0105c	9	1
DEAD/DEAH box helicase, putative	PFI0165c	9	1
Copper-transporting ATPase (CuTP)	PFI0240c	9	1
mitochondrial carrier protein, putative	PFI0255c	9	1
Conserved Plasmodium protein, unkown function	PFI0275w	9	1
subpellicular microtubule protein 1, putative (SPM1)	PFI0460w	9	1
cysteine repeat modular protein 1 (CRMP1)	PFI0550w	9	1
Hypothetical protein	PFI0690c	9	1
Hypothetical protein	PFI1010w	9	1
para-aminobenzoic acid synthetase (pBAS)	PFI1100w	9	1
Conserved Plasmodium protein, unkown function	PFI1120c	9	1
Acyl-CoA synthetase (ACS7)	PFL0035c	12	1
IMP-specific 5'-nucleotidase, putative,haloacid dehalogenase hydrolase, putative	PFL0305c	12	1
cysteine repeat modular protein 3 (CRMP3)	PFL0410w	12	1
Arginine-tRNA ligase, putative	PFL0900c	12	1
Conserved Plasmodium protein, unkown function	PFL1025c	12	1
isoleucine--tRNA ligase, putative	PFL1210w	12	1
RAP protein, putative	PFL1280w	12	1
Conserved Plasmodium protein, unkown function	PFL1305c	12	1
Cyclin	PFL1330c	12	1
Hypothetical protein	PFL1430c	12	1
Myosin D (MyoD)	PFL1435c	12	1
mitochondrial ribosomal protein L23 precursor, putative	PFL1895w	12	1
3-hydroxyisobutyryl-coenzyme A hydrolase, putative	PFL1940w	12	1
Hbeta58/Vps26 protein homolog, putative	PFL2415w	12	1

The No. column represents the number of SNPs typed in each gene.

Appendix Table 2: genes and SNPs analysed in *P. falciparum* samples collected in primary schools in western Kenya.

gene product	gene symbol	chr	No.
Dehydrofolate reductase	DHFR	4	2
Erythrocyte Binding Antigen - 165	EBA165	4	1
DNA repair protein RAD5, putative (RAD5)	MAL13P1_216	13	1
step II splicing factor, putative	MAL13P1_242	13	1
guanylyl cyclase beta (GCbeta)	MAL13P1_301	13	1
DNA repair endonuclease, putative	MAL13P1_346	13	1
Hypothetical protein	MAL7P1_17	7	1
erythrocyte binding antigen -175	MAL7P1_176	7	27
Plasmodium exported protein (hyp9), unknown function	MAL7P1_177	7	1
alpha/beta hydrolase, putative (GEXP08)	MAL7P1_178	7	1
Regulator of chloroquine condensation, putative	MAL7P1_38	7	1
protein phosphatase, putative	MAL8P1_109	8	1
small heat shock protein HSP20, putative (HSP20)	MAL8P1_78	8	1
Multidrug resistance Protein 1	MDR1	5	1
Conserved Plasmodium protein, unkown function	PF07_0044	7	1
Hypothetical protein	PF07_0053	7	1
RNA binding protein, putative	PF07_0066	7	1
acyl-CoA synthetase (ACS5)	PF07_0129	7	3
Surface-associated interspersed protein 8.2 (SURFIN 8.2)	PF08_0002	8	5
inositol polyphosphate kinase, putative (IPK1)	PF10_0078	10	1
syntaxin, Qa-SNARE family (SYN13)	PF11_0052	11	1
mitogen-activated protein kinase 2 (MAP2)	PF11_0147	11	1
cysteine proteinase falcipain 3 (FP3)	PF11_0162	11	1
Apical Membrane Antigen 1	PF11_0344	11	3
Hypothetical protein	PF11_0347	11	2
Hypothetical protein	PF11_0349	11	1
heat shock protein 70 (Hsp70-3)	PF11_0351	11	1
Farnesyltransferase beta subunit, putative	PF11_0483	11	1
sodium/hydrogen exchanger, Na , H antiporter (NHE)	PF13_0019	13	1
Surface-associated interspersed protein 13.1 (SURFIN 13.1)	PF13_0075	13	1
conserved Plasmodium protein, unknown function	PF14_0093	14	1
WD repeat-containing protein, putative	PF14_0101	14	1
Hypothetical protein	PF14_0153	14	2
liver specific protein 1, putative (LISP1)	PF14_0179	14	1
Plasmodium exported protein (PHISTb), unkown function	PF14_0746	14	1
surface-associated interspersed protein 14.1 (SURFIN 14.1)	PF14_0747	14	1
conserved Plasmodium protein, unknown function	PFA0410w	1	1
surface-associated interspersed protein 1.1 (SURFIN 1.1)	PFA0625w	1	1
DNA repair endonuclease, putative	PFB0265c	2	1
Merozoite surface protein 5	PFB0305c-a	2	1
ABC transporter B family member 4, putative (ABCB4)	PFC0125w	3	1

conserved Plasmodium protein, unknown function	PFC0325c	3	1
T-complex protein 1 subunit (CCT7)	PFC0350c	3	1
surface-associated interspersed protein 4.1 (SURFIN 4.1)	PFD0100c	4	4
DEAD box ATP-dependent RNA helicase, putative	PFD0565c	4	1
DNA polymerase alpha	PFD0590c	4	1
Phosphoglucomutase-2 (PGM2)	PFD0660w	4	1
Conserved protein, unkown function	PFD0705c	4	1
conserved Plasmodium protein, unknown function	PFD0735c	4	1
Hypothetical protein	PFD0860w	4	1
Surface-associated interspersed protein 4.2 (SURFIN 4.2)	PFD1160w	4	4
Pre-mRNA-processing ATP-dependent RNA helicase prp5, putative (PRP5)	PFE0430w	5	1
cation-transporting ATPase 1 (ATPase1)	PFE0805w	5	2
Hypothetical protein	PFE1015c	5	1
ubiquitin carboxyl-terminal hydrolase, putative	PFE1355c	5	1
conserved Plasmodium protein, unknown function	PFF0175c	6	1
Hypothetical protein	PFF0695w	6	1
Hypothetical protein	PFF0990c	6	1
DNA polymerase 1, putative	PFF1225c	6	1
Copper-transporting ATPase (CuTP)	PFI0240c	9	1
mitochondrial carrier protein, putative	PFI0255c	9	1
subpellicular microtubule protein 1, putative (SPM1)	PFI0460w	9	1
Hypothetical protein	PFI1010w	9	1
Acyl-CoA synthetase (ACS7)	PFL0035c	12	1
cysteine repeat modular protein 3 (CRMP3)	PFL0410w	12	1
isoleucine--tRNA ligase, putative	PFL1210w	12	1
Myosin D (MyoD)	PFL1435c	12	1

The No. column represents the number of SNPs typed in each gene.