A multifaceted investigation of the genomics of malaria, from parasite to host, using next-generation sequencing technologies

Ashley A. Osborne

Thesis submitted in accordance with the requirements for the degree of
Doctor of Philosophy

University of London
July 2023

Department of Infection Biology
Faculty of Infectious and Tropical Diseases

LONDON SCHOOL OF HYGIENE & TROPICAL MEDICINE

School of Tropical Medicine and Global Health
NAGASAKI UNIVERSITY

Funded by the Japanese Ministry of Education, Culture, Sports,
Science and Technology WISE Program

Research Group Affiliations:
Professor Taane G. Clark, Professor Susana Campino,
Professor Akira Kaneko & Professor Kiyoshi Kita

I, Ashley Osborne, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Abstract

Despite decades of progress and a drastic reduction in malaria burden worldwide since the implementation of the *Global Malaria Strategy* in 1992 and *Roll Back Malaria* in 1998, malaria incidence rates and deaths in 2020 had their most dramatic increase since the start of the millennium. *Plasmodium* species exhibit remarkably flexible genomes that allow them to adapt and evolve to any range of selective pressures within their environment, from host immune evasion and the acquisition of protective phenotypes to the development of a wide array of drug resistance mechanisms. The aim of this thesis is to demonstrate the potential application of next-generation sequencing technologies within the field of malaria disease surveillance, as well as present novel information surrounding East African *Plasmodium falciparum* populations. Despite accounting for 10% of all malaria cases worldwide annually, and the recent emergence of locally spreading clinically artemisinin resistant *P. falciparum* parasites in Uganda, East African parasite populations have been largely underrepresented in whole genome population studies. To supplement the availability of sequencing data and provide insights into the population structure of *P. falciparum* parasites in East Africa, I generated the first baseline assessment of malaria parasites from the Kenyan region of Lake Victoria, which identified the presence of unique ancestral origins for Lake Victoria isolates compared to other Kenyan parasite populations, as well as a potential genetic subgroup within the wide East African population. To further investigate the genetic structure of East African *P. falciparum* parasites, I generated sequencing data for parasites collected along the Kenya-Uganda border and created a genomic dataset using publicly available data from regions across Kenya, Tanzania, and Uganda. This expanded dataset confirmed the presence of subpopulations within the East African parasite population, including a distinct genetic structure amongst isolates from Western Kenya, the Kenyan region of Lake Victoria, and Central Uganda, confirming our initial findings.

Although the benefits of implementing WGS-based analyses within this region of high transmission cannot be understated, the cost of sequencing can still be cost prohibitive in low-resource settings, such as those presented within this thesis. I supplemented whole genome sequencing (WGS)-based analyses within this thesis with low-cost dual-indexing (e.g., DNA barcoding) amplicon sequencing to achieve high throughput coverage of not only parasite drug resistance markers, but also human genetic variants associated with malaria disease severity or protection. WGS demonstrates limited efficacy when attempting to sequence sub microscopic or low-density infections due to low levels of available template DNA. Ngodhe island, located within the Kenyan region of Lake Victoria, has a unique malaria transmission profile with asymptomatic and sub-microscopic infections accounting for most cases. To overcome this barrier, I utilised the low-cost targeted sequencing methods to generate a drug resistance profile of *P. falciparum* parasites from Ngodhe island for the first time. Malaria is a dynamic parasitic

infection that will require a multifaceted approach towards control, encompassing not only the parasite and vector-based methods, but also increased insight into the human genetics of malaria infection protection and risk. To demonstrate the versatility of this technology, I presented a proof-of-concept method for profiling the human genetic determinants of malarial disease in an at-risk population in Northeast Tanzania. This thesis presents a multifaceted approach to disease surveillance by using both whole genome sequencing and custom targeted sequencing to characterise the genomic and genetic diversity of *P. falciparum* populations in high transmission regions of East Africa, as well as the genetic determinants of malarial disease in at-risk human populations.

Acknowledgements

I would first like to thank my LSHTM supervisors Taane G. Clark and Susana Campino for giving me the opportunity to become a part of their team, as well as providing me with the intellectual and emotional support I needed to successfully pursue a PhD. I would also like to extend my thanks to my Nagasaki University supervisors, Akira Kaneko and Kyoshi Kita, for giving me the chance to be a part of an international PhD programme, linking together many teams and researchers from around the world.

The biggest thanks goes to the member of "The Hub" who acted as the best support network anyone could ask for. As they say, nothing fosters friendships quite like those forged within the face of copious amounts of stress. Through being a part of this group, I've made life-long friendships that I will continue to cherish far beyond the completion of my PhD or my time at LSHTM. From the many nights during the COVID-19 pandemic playing age of empires with Matt, Dan, Gary Jody, and Ernest, almost dying on a back road in Italy with Emma, Holly and Sophie trying to see Leen get married, or having pint-induced life chats at the Pumphandle bar with Emilia, Amy, Anna, Alicia, Gabbie, and the lovely CAT3 girls, you guys have made this PhD such an enjoyable experience.

Many thanks to my colleagues in Kenya (Wataru, James, Laura, Mtakai, and Jesse) who showed me amazing kindness and kept me safe during unplanned run-ins with law-enforcement, as well as all the people on Mfangano island, Ngodhe island, and Mbita who allowed me into your homes and communities.

To one of the best friends I've ever had, thank you for encouraging the pyromaniac within me, the many Friday night fires during lockdown, and for making me falsely believe that I am the funniest person in the world. You mean the world to me Becca! (Remind me to buy celebration firewood).

A shout out to Lord Farquaad and Monster Energy™, one for fully supporting this PhD and the other one for constantly keeping me on my toes.

To my family, for supporting me even as I made the decision to move thousands of miles away to pursue this PhD and being understanding throughout all the missed birthdays and celebrations. I miss you all and I know I cause you great anxiety by travelling to disease-prone areas. I am sorry! I look forward to the times we are all reunited and can't wait to see you soon.

Finally, Alberto. Thank you for being the perfect combination of support and tough love through the final months of this PhD and for being the ball of chaos that you are. Never did I anticipate someone like you would walk into my life and I am grateful for you every day (well… most days). See you in Barcelona Fren.

Thesis Publications and Manuscripts

1. Osborne A, Manko E, Takeda M, et al (2021) Characterizing the genomic variation and population dynamics of Plasmodium falciparum malaria parasites in and around Lake Victoria, Kenya. Sci Rep 11:19809
2. Osborne A, Phelan JE, Kaneko A, et al (2022) Drug resistance profiling of asymptomatic and low-density Plasmodium falciparum malaria infections on Ngodhe island, Kenya, using custom dual-indexing next-generation sequencing. Scientific Reports (Submitted).
3. Osborne A, Manko E, Waweru H, Kaneko A, Kita K, Campino S, Gitaka J, Clark TG (2023) A high-resolution analysis of Plasmodium falciparum population dynamics in East Africa and genomic surveillance along the Kenya-Uganda border. PLOS Genetics (Submitted).
4. Osborne A, Phelan JE, Vanheer LN, Manjurano A, Drakeley CJ, Kaneko A, Kita K, Campino S, Clark TG (2022) High throughput human genotyping for variants associated with malarial disease outcomes using custom targeted amplicon sequencing. Scientific Reports (Submitted).

Additional Publications

1. Moss S, Mańko E, Vasileva H, Texeira de Silva E, Goncalves A, **Osborne A**, Phelan J, Rodrigues A, Djata P, D'Alessandro U, Mabey D, Krishna S, Last A, Clark TG, Campino S (2023). Population dynamics of Plasmodium falciparum on the Bijagós Archipelago, Guinea-Bissau. Scientific Reports. (In Press)
2. Vanheer LN, Mahamar A, Manko E, Phelan J, **Osborne A**, Spadar A, Lanke K, Stone W, Bousema T, Clark TG, Drakeley C, Dicko A, Campino S (2023). Genome-wide genetic variation and molecular surveillance of drug resistance in Plasmodium falciparum isolates from asymptomatic individuals in Ouélessébougou, Mali. Scientific Reports. (Submitted)
3. Kagaya W, Chan CW, Kongere J, Kanoi BN, Ngara M, Omondi P, **Osborne A**, Barbieri L, Kc A, Minakawa N, Gitaka J, Kaneko A (2023). Evaluation of the protective efficacy of Olyset®Plus ceiling net on reducing malaria prevalence in children in Lake Victoria basin, Kenya: study protocol for a cluster-randomized controlled trial. Trials TRLS-D-23-00298. (Submitted)

Table of Contents

Abbreviations

ACT	Artemisinin Combination Therapy
BCS	Blantyre coma score
CQ	Chloroquine
DNA	Deoxyribonucleic acid
F_{st}	Fixation index statistic
F_{ws}	Inbreeding coefficient
Fy-	Duffy negative phenotype
G6PD	Glucose-6-phosphate dehydrogenase
GWAS	Genome-wide association study
HBB	Haemoglobin subunit beta
HbS	Haemoglobin S (sickle-cell haemoglobin)
HbSS	Sickle-cell disease
HbC	Haemoglobin C
HbE	Haemoglobin E
IBD	Identity-by-descent
iHS	Integrated haplotype score
IPTi	Intermittent preventative treatment in infants
IPTp	Intermittent preventative treatment in pregnancy
IRS	Indoor residual spraying
LV	Lake Victoria
MDA	Mass drug administration
Pf	*Plasmodium falciparum*
Pfap2mu	AP-2 complex subunit mu
Pfcrt	Chloroquine resistance transporter
Pfdhfr	Dihydrofolate reductase
Pfdhps	Dihydropteroate synthase
Pfk13	Kelch protein 13
Pfmdr1	Multidrug resistance protein 1
Pfubp1	Ubiquitin carboxyl-terminal hydrolase 1
Pv	*Plasmodium vivax*
RDT	Rapid diagnostic test
SNP	Single nucleotide polymorphism
SP	Sulphadoxine pyrimethamine
SWGA	Selective whole genome amplification
WGS	Whole genome sequencing
WHO	World Health Organization

Chapter 1 – Introduction

Malaria: A global health concern

Malaria is a vector borne disease that continues to be a major global health burden, particularly in many low-income countries and across sub-Saharan Africa. The causative agents of malaria are parasites belonging to the genus *Plasmodium*, of which there are 6 species known to infect humans (*Plasmodium falciparum, P. vivax, P. malariae, P. ovale curtisi* (*Poc*)*, P. ovale wallikeri* (*Pow*)*, and P. knowlesi*), with much of the burden resulting from *P. falciparum* [1, 2]. *P. falciparum* is estimated to be responsible for 99.7% of cases in the World Health Organization (WHO) African region, with the remainder of cases caused by *P. vivax, P. malariae, and P. ovale spp.*, in descending order [2]. Despite decades of progress and a drastic reduction in malaria burden worldwide since the implementation of the *Global Malaria Strategy* in 1992 and *Roll Back Malaria* in 1998, malaria incidence rates and deaths in 2020 had their most dramatic increase since the start of the millennium [2, 3].

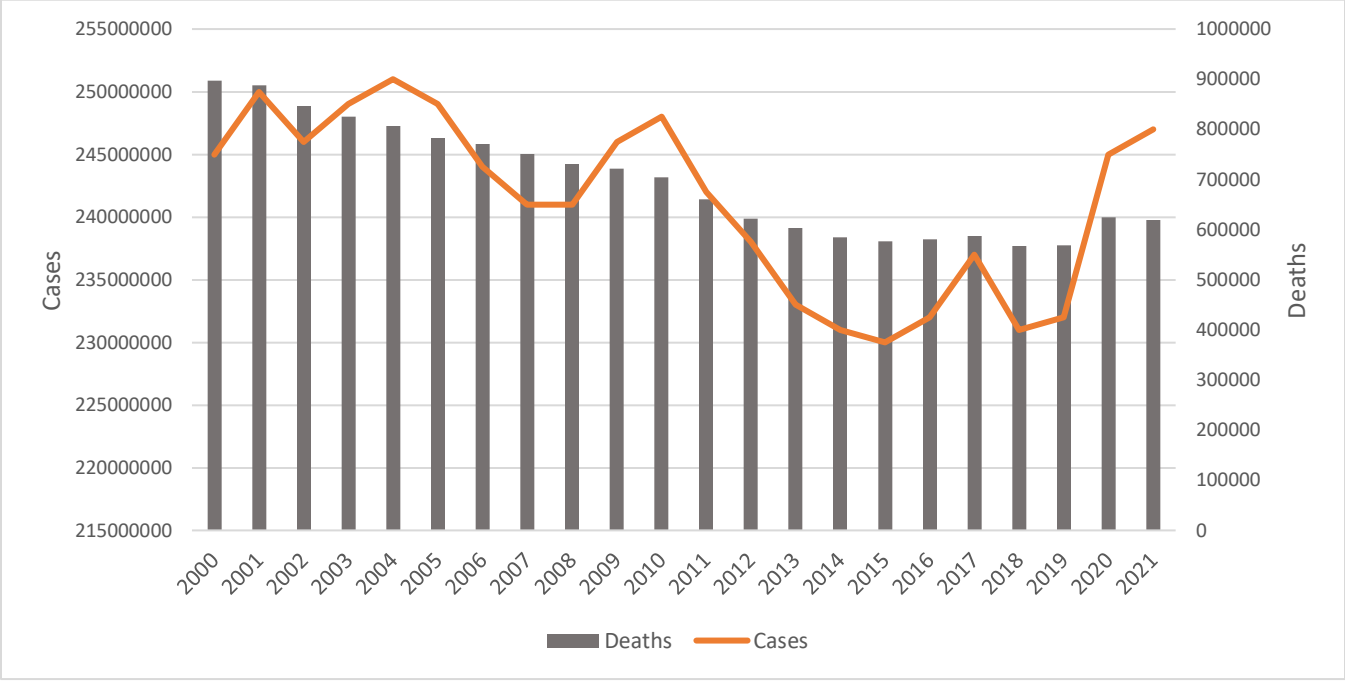


Figure 1: Estimated malaria cases and deaths in the WHO African Region between 2000 and 2021 highlighting the recent increase in malaria incidence.

Increased malaria incidence in 2020 resulted in an estimated 254 million cases globally, up from 232 million in 2019; similarly, there were 625,000 deaths in 2020, up by 57,000 from the previous year (**Figure 1**). These numbers continue to remain elevated in 2021, with 247 million cases and 619,000 deaths estimated to have occurred worldwide. The WHO Africa Region alone accounted for approximately 95% of cases and 96% of deaths, with 80%

of those deaths among children under 5 years of age [2, 3]. This increase in malaria burden is largely believed to be due to disruptions in malaria control programmes and supply chain failures caused by the COVID-19 pandemic, as well as a handful of other biological and ecological threats to malaria control programmes [4, 5]. As efforts to decrease the burden of malaria on public health in low- and middle-income countries (LMICs) are met with stagnating results, and malaria continues to disproportionately impact pregnant women and children under 5, the need for technical advancements in disease control and surveillance has never been more apparent. Malaria is a multifaceted and complex infection, requiring insight into not only the malaria parasite itself, but also the human host and mosquito vector dynamics, as they each pose their own set of unique challenges towards achieving global malaria elimination.

The life cycle of malaria

The life cycle of malaria is a complex process made possible by the ability of *Plasmodium spp.* Parasites to change their cellular and molecular makeup between life cycle stages, utilising upwards of 5,000 characterised genes [6]. This cellular and molecular flexibility allows for continuous host-habitat changes, enabling malaria parasite development across mosquito vectors and a variety of mammalian hosts.

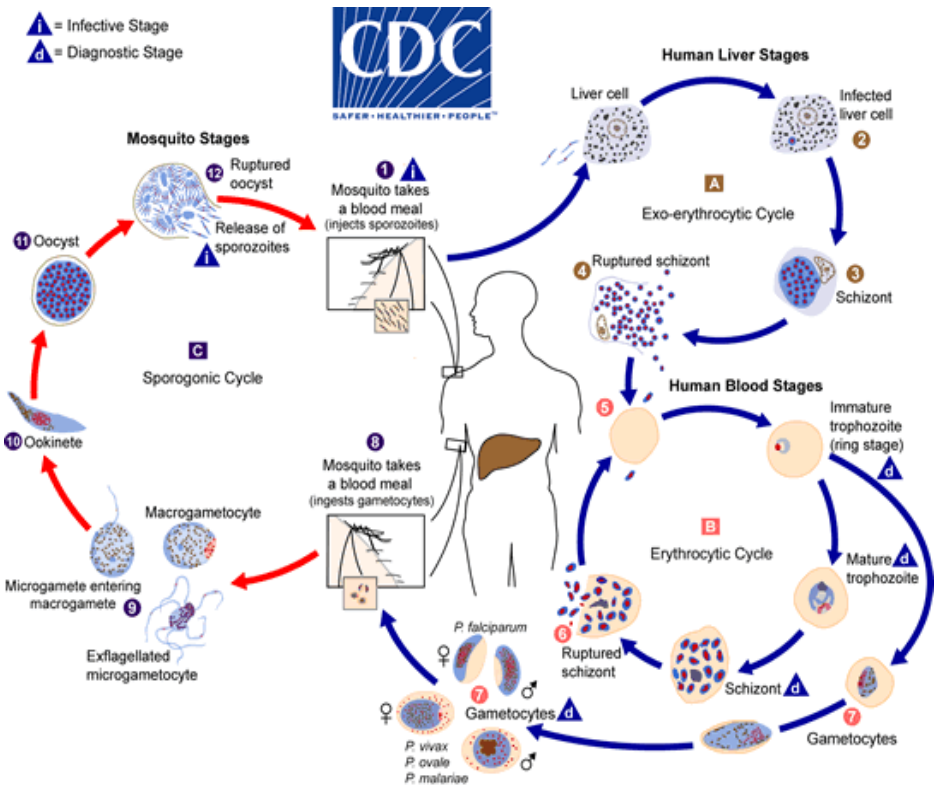


Figure 2: Malaria parasite life cycle between mosquito vector ad human host [7]

Mosquito vectors

Sexual replication of malaria parasites occurs in the mosquito vector and takes place once during a complete life cycle (**Figure 2**, stages 8 to 12). Male and Female *Plasmodium spp.* gametocytes are taken up by mosquitoes following a bloodmeal from an infected mammalian host. The process of ingestion triggers the formation of gametes in the lumen of the mosquito midgut where they develop into macro- and microgametes, female and male gametes, respectively [8]. These gametes undergo meiosis and genetic recombination to produce diploid zygotes, the only time the genome of *Plasmodium spp.* exists in a diploid state, before differentiating into ookinetes [9]. Ookinetes penetrate the mosquito inner midgut epithelium and traverse the epithelial cells through to the opposite side of the epithelium. This process activates the ookinetes transformation into oocysts and their subsequent development into sporozoites which, upon completion, leave thousands of sporozoites waiting to be released [10]. Following oocyst rupture and sporozoite release into the mosquito haemolymph, sporozoites are circulated throughout the entirety of the mosquito's tissues before eventually making their way to the salivary glands [11]. Sporozoites traverse the plasma membrane of the salivary glands using transit vacuoles [12]. The sporozoites that make their way to the salivary gland duct go on to infect humans and other mammalian hosts during the mosquito's next bloodmeal.

Human infection and clinical manifestations

Malaria infections in humans occur when a mosquito carrying *Plasmodium spp. s*porozoites feeds on a human and the sporozoites are passed from the mosquito's salivary gland into the human host's bloodstream (Figure 2, stages 1 to 8). Once in the bloodstream, the sporozoites travel to the liver to infect hepatocytes, the main functional cells within the liver, where they mature and start replicating asexually [6]. During this process, the human host generally remains asymptomatic, however, at this point in the lifecycle, some species of *Plasmodium spp.* (*P. vivax* and *P. ovale spp.*) can establish a dormant hypnozoite stage within the liver [13]. This dormant stage can often remain undetected and, if left untreated, can cause infection relapses. The liver stage of infection, independent of any dormant stage that may occur, generally lasts between 2 and 7 days [14, 15].

Matured parasites are transported out of the liver via vesicles, and back into the bloodstream, where the merozoite stage parasites begin invading the red blood cells (RBCs) to start the asexual blood stage of the infection. Merozoites multiply and mature within the RBCs, ultimately leading to its rupture and the release of gametocytes or merozoites, the sexual and asexual life cycle stages, respectively [13]. Gametocytes, as discussed previously, will be taken up during a mosquito blood meal to continue the mosquito life cycle stages, whereas the trophozoites will create schizonts that continue the process of invading and rupturing RBCs [6]. This rupturing of

the RBCs is what induces the clinical symptoms and manifestations of malaria infections, including anaemia and cyclical fevers. Due to the differing rates at which the parasite species undergo asexual reproduction, and the synchronous rupturing of RBCs, the timing of the cyclical fevers differs depending on the causative species of the infection [13]. Although symptoms of infection vary from mild to severe, as well as differ between species, the classic presentation of malaria infection can include cyclical fevers, fatigue, nausea, and chills [16, 17].

Clinical immunity and protection against symptomatic malaria is well categorised in human populations residing in malaria endemic regions, with the development of protection following repeated bouts of infection and recovery [13, 16]. This repeated exposure reduces the risk of symptomatic or severe malaria infections, conferring protection to older individuals and adults, however it does not prevent the risk of infection entirely. Infections in these protected groups of individuals will still occur but are likely to be mild and asymptomatic [16]. Asymptomatic cases make up most malaria cases worldwide, with the majority going undetected and untreated. Due to this, the true numbers of asymptomatic cases, and global malaria cases, remains unknown, as well as to what degree asymptomatic cases contribute to maintaining malaria transmission chains [2].

Plasmodium falciparum

Responsible for 99.7% of cases within the WHO African region, *P. falciparum* causes the highest burden of mortality and morbidity of the *Plasmodium* species [2, 18]. Despite its stronghold in sub-Saharan Africa, *P. falciparum* is not limited in its distribution and accounts for 50% of malaria cases in the WHO South-East Asia region, 71% in the WHO Eastern Mediterranean region, and 65% of the WHO Western Pacific region [2]. *P. falciparum* is the main cause of severe malarial anaemia and cerebral malaria, the deadliest clinical manifestations of the disease, and remains one of the leading causes of death for children under 5 years of age in sub-Saharan Africa [16, 19]. High transmission remains across much of sub-Saharan Africa due to various environmental and geopolitical factors, including, but not limited to, the warm climate, political instability, and low resource availability.

P. vivax and the neglected malaria parasites

Although limited in its distribution across the African continent, *P. vivax* has an incredibly wide distribution across the remainder of the world, with an estimated one third of the population at risk of infection, and is the second most virulent, human-infective, malaria parasite [14]. The reduced prevalence of *P. vivax* across Africa has been linked to the Duffy-negative blood group, a blood group variant fixed in many sub-Saharan African populations, in which individuals lack an antigen on the surface of their red blood cells believed to be required for the establishment of infection [20]. In addition to having distinct ancestral origins from *P. falciparum*, *P. vivax* also has

the capacity to establish dormant infections in the liver that cannot be treated using first-line treatment methods for *P. vivax* and lead to recrudescent infections if left untreated or undetected [14].

Although generally neglected and understudied due to their mild symptoms, *P. malariae*, *P. ovale curtisi*, and *P. ovale wallikeri* are commonly detected as co-infections alongside *P. falciparum*, as well as with one another. In general, *P. malariae* and *P. ovale spp.* are not detectable using common diagnostic tests so the true number of co-infections are likely to be severely under-reported [15]. *P. malariae* is well documented to cause persistent malaria infections and can latently reside within human hosts for years, remaining in red blood cells for long periods of time at low densities [21, 22].

Malaria control programmes

Malaria control programmes generally consist of three core objectives: prevention, detection, and treatment. Prevention of malaria primarily focuses on vector control, including the distribution and use of insecticide treated bed nets (ITNs), as well as the implementation of indoor residual spraying (IRS) [23]. Breakdowns in the dissemination of prevention measures against the mosquito vectors during the COVID-19 pandemic has played a role in the upsurge of malaria cases currently being observed. Prevention can also include human behaviour modifications, such as keeping outside areas free of standing water and limiting time spent outdoors during peak mosquito feeding times, as well as community education and engagement in malaria control programmes [2, 24]. The detection and treatment objectives are interlinked, with rapid detection required for treatment measures to be taken. Quick and effective treatment of malaria is essential to reduce transmission intensity in high transmission regions, as well as to prevent any resurgence of disease in areas with low transmission, particularly those targeting elimination [25]. Despite gains made over the past three decades by malaria control programmes around the world, long-term efficacy of these programmes is being jeopardised by a variety of biological and environmental threats, such as parasite resistance to treatment methods, vector resistance to insecticides, and climate change.

Emergence and spread of drug resistance in *P. falciparum*

The ability of malaria parasites to undergo sexual reproduction, or genetic recombination, results in a high degree of genetic variability and the ability to rapidly evolve mechanisms to combat clearance by antimalarial drugs. These have a genetic underpinning and can be accessed through whole genome or targeted sequencing of the *P. falciparum* genome.

Chloroquine

Chloroquine (CQ), a derivative of quinine, was first introduced as a treatment for malaria in 1947, but by the 1980s resistance to CQ was widespread in *P. falciparum* populations worldwide. The main mechanism of action for CQ is believed to be its ability to sequester in the digestive vacuole of *Plasmodium spp.* and prevent the degradation of haematin into hemozoin, which results in the build-up of toxic by-products and parasite cell death. Resistance to CQ has been linked to a number of single nucleotide polymorphisms (SNPs) in a digestive vacuole transmembrane protein, now termed the chloroquine resistance transporter gene, or *Pfcrt* [26] (gene size: 3,956(+) bp). A non-synonymous SNP at codon 76, resulting in a substitution of lysine with threonine, is generally used as the main molecular marker for CQ resistance, however the five amino acids at positions 72 to 76 are important and can be utilised to confer the origin of a resistance phenotype [27]. Mutations and copy number variation in the multi-drug resistance gene, *Pfmdr1* (gene size: 7,140(+) bp), and *Pfplasmepsin* 2/3 (gene size: 3,017(+) bp / 2,418(+) bp), have also been documented to confer resistance to CQ, as well as amplify existing resistance [28–30].

Antifolates

Sulphadoxine-pyrimethamine (SP) became the first line treatment for uncomplicated *P. falciparum* malaria following the spread of CQ resistance. Despite being a combination therapy, resistance to SP emerged rapidly, estimated to be as early as 10 years following its introduction, and ultimately resulted in widespread ineffectiveness of the drug. SP inhibits the malaria parasite's folate biosynthesis pathway by targeting two important enzymes, dihydrofolate reductase and dihydropteroate synthetase. SNPs in *Pfdhfr* (gene size: 2,168(+) bp) and *Pfdhps* (gene size: 3,161(+) bp) confer resistance to pyrimethamine and sulphadoxine, respectively, with the addition of subsequent mutations progressively enhancing a parasite's level of resistance [31, 32]. SP resistant parasites commonly present with a triple mutation in *Pfdhfr* (N51I, C59R, and S108N) and a double mutation in *Pfdhps* (A437G and K540E). Although no longer used as a first line treatment, SP is still used as an intermittent preventative treatment in pregnancy (IPTp) and intermittent preventive treatment (IPTi) for at risk infants [33]. Using first line treatments, or ACTs, for chemoprophylaxis would inevitably increase selective pressure for drug resistance therefore SP is still used for preventative treatments. Additionally, ACTs have not yet been approved for women in their first trimester of pregnancy [34, 35]. The K540E mutation on *Pfdhps* is generally used as a proxy measure for the presence of 5 key mutations associated with SP resistance and is used by the WHO to inform decision-making surrounding IPTp guidance, with IPTp implementation only recommended in regions where K540E prevalence is <50%. However, SP as a method of IPTp and IPTi is still documented to be effective even in regions where K540E prevalence is >50% [2, 36].

Artemisinin combination therapies

To preserve the efficacy of the last line of antimalarial drugs, the current first line treatment for uncomplicated *P. falciparum* malaria is artemisinin combination therapies (ACTs). To reduce the risk of resistance emerging, ACTs combine the antimalarial properties of artemisinin with drugs differing in their mechanism of action and half-lifes [24]. Artemisinin is generally able to achieve drastic parasite reduction on its own but has a short half-life in the bloodstream, ranging from 1 to 3 hours, so it is paired with a partner drug that is not metabolised as quickly. This ensures any remaining parasites not cleared by artemisinin can be cleared by the partner drug, ideally before they are able to develop any reduced susceptibility to artemisinin [37, 38]. However, it is important to monitor resistance to partner drugs, as well as artemisinin, to ensure continued efficacy. To date, there have been no new effective antimalarial drugs since the introduction of artemisinin and artemisinin-derivatives, making ACTs the last line of defence against malaria.

Despite the rapid clearance of parasites and short half-life of artemisinin, reduced susceptibility to artemisinin has been observed throughout the Greater Mekong subregion and has recently been reported in a handful of countries within sub-Saharan Africa, including Rwanda and Uganda [39–41]. Resistance is believed to be conferred by mutations in the Kelch-13 propeller domain gene (*Pfk13;* gene size: 3,277(-) bp), with a list of approved, and candidate, molecular markers to monitor available from the WHO [2, 24]. An amino acid substitution in codon 580 (C580Y) has been identified as the primary biomarker for resistance in the Greater Mekong region, while reduced susceptibility in Rwanda (R561H) and Uganda (A675V and C469Y) is believed to have arisen independently [39–41]. Reduced efficacy of ACT partner drugs has been documented across sub-Saharan Africa, however a majority of reports documenting high rates of treatment failures due to compromised partner drug efficacy were determined to have significant deviations from the WHO standard protocol [24, 42]. Despite conflicting evidence in sub-Saharan Africa, resistance to ACT partner drugs piperaquine and mefloquine has been documented in Southeast Asia and suggests the capacity for the introduction of resistance variants to African parasite populations, as observed with CQ and SP, or independent evolution, as witnessed with artemisinin [43, 44]. Reduced efficacy of the partner drug lumefantrine has been linked to variants Y184F and D1246Y

Efficacy of diagnostic tests

Individuals infected with *P. falciparum* malaria are generally diagnosed using rapid diagnostic tests (RDTs) in field settings, however the gold standard of diagnosis remains detection using light microscopy [24]. This is, in part, due to the emergence and spread of deletions in *Pfhrp2/3*, or the histidine rich protein antigen gene, which can result in false negative RDT results and calls into question the long-term efficacy of these tests [45]. First detected

in Peru, *Pfhrp2/3*-deletion parasites have since been detected across South America, Southeast Asia, and sub-Saharan Africa, varying in frequency from 2% up to 80% in symptomatic patients in Eritrea [46, 47].

Malaria and the human genome

Prior to the advent of whole genome sequencing (WGS) and genome-wide comparison studies, it was largely believed that Plasmodium parasites evolved alongside their human, chimpanzee, or gorilla hosts over millions of years and that host species had inherited *P. falciparum*-like infections from their most recent common ancestors [48]. However, genome-wide comparisons of chimpanzee parasites *P. gaboni* and *P. reichenowi* identified a within-species genetic diversity approximately 10-fold higher than is observed within a global sample of *P. falciparum* [49, 50]. This disparity in genetic diversity between chimpanzee parasites and human parasites suggests that *P. falciparum* has undergone a severe genetic bottleneck, consistent with a more recent origin in humans. Following these observations, the current prevailing theory is that *P. falciparum* is relatively new to the human species, potentially due to a gorilla-to-human cross-species transmission event, with mutation and replication rates estimating its origin to be within the past 10,000 years [50].

Despite its relatively recent origin as a human pathogen, *P. falciparum* has exhibited the strongest known selective pressures on the human genome in recent history, largely driven by its high impact on childhood mortality and morbidity. Human populations in regions with intense malarial transmission have been under immense selection pressures over the past 10,000 years, resulting in the evolution of a wide array of polymorphisms associated with protection from parasite invasion and reduced likelihood of severe disease [51, 52]. These variants include a variety of haemoglobinopathies, such as the sickle-cell trait, thalassemia, glucose-6-phosphatase deficiency (G6PD), and countless other erythrocyte defects, that, all together, comprise some of the most common Mendelian diseases observed in humans (**Table 1**) [51, 53, 54].

Table 1: Common human genetic variants observed in malaria endemic regions and their associations with malaria disease infection and severity *[51]*.

Gene	Protein	Function	Genetic associations with malaria
ACKR1	Duffy antigen	Chemokine receptor	Duffy-negative genotype Fy(a-b-) confers protection against *P. vivax* infection; Duffy antigen essential to establishment of *P. vivax* infection
G6PD	glucose-6-phosphate dehydrogenase	Enzyme protects RBCs from reactive oxygen species and oxidative stress	G6PD deficiency observed to confer protection in heterozygous females and hemizygous males (X Chromosome-linked deficiency); parasite maturation inhibition by premature RBC breakdown
GYPA/B	Glycophorin A/B	Sialo glycoprotein	Rearrangement of *GYPA* and *GYPB* genes results in two copies of a hybrid gene that encodes the Dantu blood group antigen; Dantu confers increased RBC tension and limits parasite invasion of RBCs
HBB	haemoglobin subunit beta (β-globin)	Component of haemoglobin (Hb) metalloprotein in RBCs	HbS and HbC alleles protect against severe malaria; HbE allele reduces parasite invasion; Beta thalassaemia confers varying degree of protection in individuals due to impact on RBC development and lack of functional haemoglobin
HBA	haemoglobin subunit alpha (α-globin)		Alpha thalassaemia confers protection against severe malaria but is also observed to enhance mild malaria episodes; results in abnormal or reduced function haemoglobin

Sickle-cell Anaemia

Regarded as the classic paradigm of balancing selection in human populations, the sickle-cell trait has evolved independently in multiple malaria-endemic regions due to its ability to confer up to a 10-fold reduced risk of severe malaria [55, 56]. Sickle haemoglobin (HbS) is an abnormal variant of haemoglobin that polymerises under

low oxygen tension, resulting in rigid erythrocytes and vaso-occlusion, blocked blood flow. The precise mechanism through which the sickle-cell trait protects against malaria remains unknown, although biochemical, immune-mediated, and structural mechanisms of protection have been proposed. It is believed that the protective effect of the sickle haemoglobin is conferred through a combination of these theories, which include parasite tolerance mediated by heme oxygenase-1 (HO-1), reduced parasite growth due to translocation of host micro-RNA to the parasite, and polymerisation-induced growth inhibition following parasite cytoadherence [57–59]. Despite its pathogenic effects in homozygous carriers (HbSS), heterozygous carriers of the sickle-cell trait (HbAS) are granted a relatively harmless protective effect [55, 56]. This has resulted in the HbS allele being maintained across malaria-endemic regions of sub-Saharan Africa at a frequency of 15% or higher, while frequencies in European and East Asian populations remain at 0.006% and 0%, respectively [55].

Duffy-negative blood group

The limited impact and distribution of *P. vivax* across Africa, in particular sub-Saharan Africa, has been linked to the Fy- blood group, or the Duffy-negative blood group, in which individuals lack an antigen on the surface of their red blood cells required for the establishment of infection. This blood group variant is almost fixed in sub-Saharan African populations and, until recently, it was believed that this phenotype conferred complete protection against *P. vivax* infection [20, 60]. Recent incidences of confirmed *P. vivax* infections in individuals harbouring this phenotype have demonstrated limitations in our understanding of how *P. vivax* establishes infection and the role the Duffy antigen plays in that process [61].

Glocose-6-phosphatase deficiency

Glucose-6-phosphatase dehydrogenase (G6PD) is an enzyme encoded by a highly polymorphic gene on the X chromosome and is required by red blood cells to withstand oxidative stress [54, 62]. Variants located within the G6PD gene have been linked to diminished G6PD enzyme activity and result in a deficiency that can have severe clinical manifestations. A majority of individuals carrying this genetic disorder remain asymptomatic, however severely G6PD deficient individuals can develop haemolytic anaemia, and this can be exacerbated by treatment with primaquine, an antimalarial drug used for the treatment of *P. falciparum* gametocytes and relapses of *P. ovale* [2]. It is believed that G6PD genetic variants have arisen due to selection pressure exhibited by malaria on the human genome, particularly due to the geographical distribution of G6PD deficiencies overlapping with malaria endemic regions. This hypothesis has been corroborated by studies that have identified negative associations with severe malaria in hemizygous males and heterozygous females [54, 63].

Thalassaemia and other haemoglobinopathies

Approximately 200 million people around the world are estimated to have an erythrocytic enzyme deficiency, with 700 haemoglobin disorders identified to date. Variations affecting erythrocytes, a crucial component within the malaria parasite lifecycle, have been largely assumed to play a role in determining malaria disease severity in endemic regions [64, 65]. Thalassaemia syndromes are inherited blood disorders resulting from the absence or dysfunction of one or both haemoglobin subunits. The primary thalassaemia syndromes are alpha (α) thalassaemia and beta (β) thalassaemia in which there is variation or deletion of the α-globin or β-globin subunits [66]. Despite the high correlation between the frequency of thalassaemia syndromes and the presence of intense malaria selection, the degree to which thalassemia syndromes may protect against severe disease, nor the exact mechanism of protection, has not been characterised and large-scale studies have yet to identify consistent results [66].

Genomics as a method of disease surveillance

As disease surveillance moves towards genomics, the feasibility of gathering comprehensive insights into all aspects of the lifecycle of malaria to tailor and implement malaria control programmes becomes more accessible.

Plasmodium genomes

Plasmodium species are remarkably adept at adapting and evolving under selective pressures within their environment, from drug resistance to host specifications, largely thanks to their highly flexible genomes. The genomes of Plasmodium species consist of both a nuclear genome and an organellar genome, with GC content and size varying species to species (**Table 2**) [67]. The nuclear genome of Plasmodium species, consisting of up to 14 chromosomes, is constantly undergoing high rates of recombination, allowing for rapid and constant adaptation, including effective host immune evasion. In contrast, the organellar genomes of Plasmodium, including the mitochondria and apicoplast, tend to remain relatively stable and conserved [68]. For all Plasmodium species known to infect humans, the reference genomes have been published and made available for genomic and genetic studies of the parasite species (**Table 2**).

Table 2: Genomic characteristics and life cycle properties of *Plasmodium* species known to infect and cause disease in humans *[69–72]*.

***Plasmodium* species**	**Primary Host**	**Reference Genome**	**Genome Length (Mb)**	**GC Content (%)**	**Protein-encoding genes**	**Distinctive Life Cycle Properties**
P. falciparum	Human	3D7	23.5	19.4	5,300	No dormant stage
P. vivax	Human	PvP01/Sal1	29.0	39.8	6,642	Liver dormant stage; can cause relapse
P. malariae	Human	PmUG01 (Uganda)	31.9	24.74	6,462	Persistent and long-lasting infections
P. ovale curtisi (Poc)	Human	Poc1 (Nigeria)	34.5	28.46	7,950	Relapse possible up to 4 years post-infection
P. ovale wallikeri (Pow)	Human	Pow1 (Gabon)	35.2	28.91	8,582	--
P. knowlesi	*Macaca spp.*	H Strain	23.5	37.5	5,188	Zoonotic infection

Advancements in high throughput sequencing technologies, involving the use of "short-read sequencing", and selective whole genome amplification (SWGA) methodologies has led to more rapid and robust whole genome sequencing (WGS) of Plasmodium species parasites. This rapid expansion of available genomic data has led to the creation of large data repositories of Plasmodium genomic data available in the public domain [69]. These data repositories have made it possible to study the genomic diversity of Plasmodium species and isolates from around the world, including large-scale studies analysing parasite population dynamics and genome-wide association studies (GWAS) targeting vaccine candidates or host-pathogen interactions [18, 73, 74]. Despite the attractiveness of WGS for extensive data yields and large-scale informative studies, the high cost of sequencing and volume of parasite DNA required for successful sequencing results in WGS not being practical for all sequencing needs. As molecular techniques and sequencing technologies advance, lower cost targeted sequencing methods are becoming more widely available and more attractive methods for extracting genetic data from parasite samples, particularly field isolates. Targeted amplicon sequencing, a type of next-generation sequencing that utilises polymerase chain reactions (PCR) to create short sequences of DNA, has recently been employed by researchers as an alternative method to screen for *P. falciparum* loci associated with drug resistance [75]. This can be implemented using long read Oxford Nanopore Technology (ONT) or short read MiSeq Illumina sequencing [76,

77]. MiSeq Illumina sequencing, making use of short read sequencing technology, was used throughout this thesis for the sequencing of both *P. falciparum* parasites and *Homo sapiens* DNA.

Thesis structure

In this thesis, I present a multifaceted investigation into the genomics of *P. falciparum* malaria, from parasite to host, within the East African region to assist future malaria control programme design.

Chapter 2 (published paper)

In Chapter 2, I provide the first baseline assessment of the genomic diversity of *P. falciparum* isolates in the Lake Victoria region of Kenya, which has sparse genetic data. Isolates I collected within the Lake Victoria basin were placed within the context of African-wide populations using Illumina WGS data and population genomic analyses. My analysis revealed that *P. falciparum* isolates from Lake Victoria form a cluster within the East African parasite population and appear to have distinct ancestral origins, containing genome-wide signatures from both Central and East African lineages.

Chapter 3 (manuscript submitted)

In Chapter 3, continuing my work in the Lake Victoria basin, I demonstrate the feasibility of using Illumina next generation sequencing-technology and custom dual-indexing (called amplicon sequencing), to generate a resistance profile of asymptomatic and low-density *P. falciparum* infections from Ngodhe island. Ngodhe island, Kenya, presents a unique malaria profile, with lower *P. falciparum* incidence rates than the surrounding region, and a high proportion of sub-microscopic and low-density infections which are notoriously difficult to sequence. I was able to utilise amplicon sequencing to quantify molecular markers of resistance on the *Pfcrt*, *Pfmdr1*, *Pfdhps*, *Pfdhfr*, and *Pfk13* genes, establishing this method as a viable means of malaria surveillance suitable for regions with typical sub-microscopic infections. Overall, I present a low-cost and expandable approach that can provide timely genetic profiling data to inform clinical and surveillance activities seeking malaria elimination.

Chapter 4 (manuscript submitted)

To establish genomic surveillance along the Kenyan-Ugandan border and produce a high-resolution analysis of *P. falciparum* population dynamics occurring within East Africa, in Chapter 4 I generated sequencing data for isolates collected along the Kenya-Uganda border in Bungoma county. I identified drug resistance biomarkers associated with chloroquine resistance at significantly reduced frequencies compared to wider East African populations and a single isolate was identified to contain a non-synonymous SNP on *PfK13*, resulting in a variant within a WHO candidate marker of reduced susceptibility to artemisinin. My analysis also revealed that *P. falciparum* parasites

from East Africa form subpopulations with distinct genetic structure and diverse ancestral origins, with ancestral admixture analysis suggesting seemingly independent ancestral populations from other major African populations.

Chapter 5 (manuscript submitted)

Malaria has exhibited the strongest known selective pressure on the human genome in recent history and is the evolutionary driving force behind genetic conditions, such as sickle-cell disease, glucose-6-phosphatase deficiency, and countless other erythrocyte defects. Genomic studies (e.g., The 1000 Genomes project) have provided an invaluable baseline for human genetics, but, with an estimated two thousand ethno-linguistic groups thought to exist across the African continent, our understanding of the genetic differences between indigenous populations and their implications on disease is still limited. In Chapter 5, I expand upon the versatility of amplicon sequencing to generate a genetic profile of human polymorphisms associated with malaria pathology in northeast Tanzania. For individuals diagnosed with severe malaria, variants were successfully characterised on the haemoglobin subunit beta (HBB), glucose-6-phosphate dehydrogenase (G6PD), atypical chemokine receptor 1 (ACKR1) genes, as well as the intergenic Dantu genetic blood variant, and then validated using pre-existing genotyping data. High sequencing coverage was observed across all amplicon targets in *HBB*, *G6PD*, *ACKR1*, and the Dantu blood group, with variants identified at frequencies previously observed within this region of Tanzania, and sequencing data exhibited high concordance rates to pre-existing genotyping data (>99.5%).

Thesis Chapter	Manuscript Title	Authors	Status
Chapter 2	Characterizing the genomic variation and population dynamics of *Plasmodium falciparum* malaria parasites in and around Lake Victoria, Kenya	**Ashley Osborne**, Emilia Manko, Mika Takeda, Akira Kaneko, Wataru Kagaya, Chim Chan, Mtakai Ngara, James Kongere, Kiyoshi Kita, Susana Campino, Osamu Kaneko, Jesse Gitaka & Taane G. Clark	Published *Sci Rep*. 2021; Oct 6; 11(1):19809.
Chapter 3	Drug resistance profiling of asymptomatic and low-density *Plasmodium falciparum* malaria infections on Ngodhe island, Kenya, using custom dual-indexing next-generation sequencing	**Ashley Osborne**, Jody E. Phelan, Akira Kaneko, Wataru Kagaya, Chim Chan, Mtakai Ngara, James Kongere, Kiyoshi Kita, Jesse Gitaka, Susana Campino & Taane G. Clark	Submitted
Chapter 4	A high-resolution analysis of *Plasmodium falciparum* population dynamics in East Africa and genomic surveillance along the Kenya-Uganda border	**Ashley Osborne**, Emilia Manko, Harrison Waweru, Akira Kaneko, Kiyoshi Kita, Susana Campino, Jesse Gitaka & Taane G. Clark	Submitted
Chapter 5	High throughput human genotyping for variants associated with malarial disease outcomes using custom targeted amplicon sequencing	**Ashley Osborne**, Jody E. Phelan, Leen N. Vanheer, Alphaxard Manjurano, Christopher J. Drakeley, Akira Kaneko, Kiyoshi Kita, Susana Campino & Taane G. Clark	Submitted

References

1. Antinori S, Galimberti L, Milazzo L, Corbellino M (2012) Biology of Human Malaria Plasmodia Including Plasmodium Knowlesi. Mediterr J Hematol Infect Dis. https://doi.org/10.4084/MJHID.2012.013

2. Geneva: World Health Organization (2022) World malaria report 2022. Licence: CC BY-NC-SA 3.0 IGO

3. World Health Organization (2020) World malaria report 2020: 20 years of global progress and challenges. World Health Organization, Geneva

4. Weiss DJ, Bertozzi-Villa A, Rumisha SF, et al (2021) Indirect effects of the COVID-19 pandemic on malaria intervention coverage, morbidity, and mortality in Africa: a geospatial modelling analysis. The Lancet Infectious Diseases 21:59–69

5. Ryan SJ, Lippi CA, Zermoglio F (2020) Shifting transmission risk for malaria in Africa with climate change: a framework for planning and intervention. Malar J 19:170

6. Aly ASI, Vaughan AM, Kappe SHI (2009) Malaria Parasite Development in the Mosquito and Infection of the Mammalian Host. Annu Rev Microbiol 63:195–221

7. Centers for Disease Control and Prevention (2020) CDC - Malaria - About Malaria - Biology. https://www.cdc.gov/malaria/about/biology/index.html. Accessed 31 Jan 2023

8. Vlachou D, Schlegelmilch T, Runn E, Mendes A, Kafatos FC (2006) The developmental migration of Plasmodium in mosquitoes. Curr Opin Genet Dev 16:384–391

9. Kooij TW, Matuschewski K (2007) Triggers and tricks of Plasmodium sexual development. Curr Opin Microbiol 10:547–553

10. Al-Olayan EM, Beetsma AL, Butcher GA, Sinden RE, Hurd H (2002) Complete development of mosquito phases of the malaria parasite in vitro. Science 295:677–679

11. Hillyer JF, Barreau C, Vernick KD (2007) Efficiency of salivary gland invasion by malaria sporozoites is controlled by rapid sporozoite destruction in the mosquito hemocoel. Int J Parasitol 37:673–681

12. Pimenta PF, Touray M, Miller L (1994) The journey of malaria sporozoites in the mosquito salivary gland. J Eukaryot Microbiol 41:608–624

13. Mawson AR (2013) The pathogenesis of malaria: a new perspective. Pathog Glob Health 107:122–129

14. Chu CS, White NJ (2021) The prevention and treatment of Plasmodium vivax malaria. PLoS Med 18:e1003561

15. Collins WE, Jeffery GM (2005) Plasmodium ovale: Parasite and Disease. Clin Microbiol Rev 18:570–581

16. Perkins DJ, Were T, Davenport GC, Kempaiah P, Hittner JB, Ong'echa JM (2011) Severe Malarial Anemia: Innate Immunity and Pathogenesis. Int J Biol Sci 7:1427–1442

17. Bartoloni A, Zammarchi L (2012) Clinical Aspects of Uncomplicated and Severe Malaria. Mediterr J Hematol Infect Dis. https://doi.org/10.4084/MJHID.2012.026

18. Amambua-Ngwa A, Amenga-Etego L, Kamau E, et al (2019) Major subpopulations of Plasmodium falciparum in sub-Saharan Africa. Science 365:813–816

19. Rénia L, Wu Howland S, Claser C, Charlotte Gruner A, Suwanarusk R, Hui Teo T, Russell B, Ng L (2012) Cerebral malaria. Virulence 3:193–201

20. Howes RE, Patil AP, Piel FB, et al (2011) The global distribution of the Duffy blood group. Nat Commun 2:266

21. Collins WE, Jeffery GM (2007) Plasmodium malariae: Parasite and Disease. Clinical Microbiology Reviews 20:579

22. Badiane AS, Diongue K, Diallo S, et al (2014) Acute kidney injury associated with Plasmodium malariae infection. Malaria Journal 13:226

23. Lobo NF, Achee NL, Greico J, Collins FH (2018) Modern Vector Control. Cold Spring Harb Perspect Med. https://doi.org/10.1101/cshperspect.a025643

24. World Health Organization (2019) Compendium of WHO malaria guidance: prevention, diagnosis, treatment, surveillance and elimination. World Health Organization, Geneva

25. World Health Organization (2017) A framework for malaria elimination: key points and Q&A. World Health Organization, Geneva

26. Fidock DA, Nomura T, Talley AK, et al (2000) Mutations in the P. falciparum digestive vacuole transmembrane protein PfCRT and evidence for their role in chloroquine resistance. Mol Cell 6:861–871

27. Ecker A, Lehane AM, Clain J, Fidock DA (2012) PfCRT and its role in antimalarial drug resistance. Trends Parasitol 28:504–514

28. Djimdé A, Doumbo OK, Cortese JF, et al (2001) A Molecular Marker for Chloroquine-Resistant Falciparum Malaria. New England Journal of Medicine 344:257–263

29. Ansbro MR, Jacob CG, Amato R, et al (2020) Development of copy number assays for detection and surveillance of piperaquine resistance associated plasmepsin 2/3 copy number variation in Plasmodium falciparum. Malaria Journal 19:181

30. Loesbanluechai D, Kotanan N, de Cozar C, et al (2019) Overexpression of plasmepsin II and plasmepsin III does not directly cause reduction in Plasmodium falciparum sensitivity to artesunate, chloroquine and piperaquine. International Journal for Parasitology: Drugs and Drug Resistance 9:16–22

31. Gatton ML, Martin LB, Cheng Q (2004) Evolution of Resistance to Sulfadoxine-Pyrimethamine in Plasmodium falciparum. Antimicrob Agents Chemother 48:2116–2123

32. Sibley CH, Hyde JE, Sims PFG, Plowe CV, Kublin JG, Mberu EK, Cowman AF, Winstanley PA, Watkins WM, Nzila AM (2001) Pyrimethamine–sulfadoxine resistance in Plasmodium falciparum: what next? Trends in Parasitology 17:582–588

33. Mosha D, Chilongola J, Ndeserua R, Mwingira F, Genton B (2014) Effectiveness of intermittent preventive treatment with sulfadoxine–pyrimethamine during pregnancy on placental malaria, maternal anaemia and birthweight in areas with high and low malaria transmission intensity in Tanzania. Tropical Medicine & International Health 19:1048–1056

34. Amimo F, Lambert B, Magit A, Sacarlal J, Hashizume M, Shibuya K (2020) Plasmodium falciparum resistance to sulfadoxine-pyrimethamine in Africa: a systematic analysis of national trends. BMJ Glob Health 5:e003217

35. Divala TH, Cohee LM, Laufer MK (2019) The remarkable tenacity of sulfadoxine-pyrimethamine. The Lancet Infectious Diseases 19:460–461

36. World Health Organization (2014) WHO policy brief for the implementation of intermittent preventive treatment of malaria in pregnancy using sulfadoxine-pyrimethamine (IPTp-SP). 13

37. Lin JT, Juliano JJ, Wongsrichanalai C (2010) Drug-Resistant Malaria: The Era of ACT. Curr Infect Dis Rep 12:165–173

38. Visser BJ, van Vugt M, Grobusch MP (2014) Malaria: an update on current chemotherapy. Expert Opin Pharmacother 15:2219–2254

39. Uwimana A, Legrand E, Stokes BH, et al (2020) Emergence and clonal expansion of in vitro artemisinin-resistant Plasmodium falciparum kelch13 R561H mutant parasites in Rwanda. Nat Med 26:1602–1608

40. Imwong M, Suwannasin K, Kunasol C, et al (2017) The spread of artemisinin-resistant Plasmodium falciparum in the Greater Mekong subregion: a molecular epidemiology observational study. Lancet Infect Dis 17:491–497

41. Tumwebaze PK, Conrad MD, Okitwi M, et al (2022) Decreased susceptibility of Plasmodium falciparum to both dihydroartemisinin and lumefantrine in northern Uganda. Nat Commun 13:6353

42. Ebong C, Sserwanga A, Namuganga JF, et al (2021) Efficacy and safety of artemether-lumefantrine and dihydroartemisinin-piperaquine for the treatment of uncomplicated Plasmodium falciparum malaria and prevalence of molecular markers associated with artemisinin and partner drug resistance in Uganda. Malaria Journal 20:484

43. Boonyalai N, Vesely BA, Thamnurak C, et al (2020) Piperaquine resistant Cambodian Plasmodium falciparum clinical isolates: in vitro genotypic and phenotypic characterization. Malaria Journal 19:269

44. Price RN, Uhlemann A-C, Brockman A, et al (2004) Mefloquine resistance in Plasmodium falciparum and increased pfmdr1 gene copy number. Lancet 364:438–447

45. McCaffery JN, Nace D, Herman C, Singh B, Sompwe EM, Nkoli PM, Ngoyi DM, Kahunu GM, Halsey ES, Rogier E (2021) Plasmodium falciparum pfhrp2 and pfhrp3 gene deletions among patients in the DRC enrolled from 2017 to 2018. Sci Rep 11:22979

46. Gatton ML, Chaudhry A, Glenn J, et al (2020) Impact of Plasmodium falciparum gene deletions on malaria rapid diagnostic test performance. Malaria Journal 19:392

47. Kaaya RD, Kavishe RA, Tenu FF, Matowo JJ, Mosha FW, Drakeley C, Sutherland CJ, Beshir KB (2022) Deletions of the Plasmodium falciparum histidine-rich protein 2/3 genes are common in field isolates from north-eastern Tanzania. Sci Rep 12:5802

48. Escalante AA, Ayala FJ (1994) Phylogeny of the malarial genus Plasmodium, derived from rRNA gene sequences. Proc Natl Acad Sci U S A 91:11373–11377

49. Sundararaman SA, Plenderleith LJ, Liu W, et al (2016) Genomes of cryptic chimpanzee Plasmodium species reveal key evolutionary events leading to human malaria. Nat Commun 7:11078

50. Loy DE, Liu W, Li Y, Learn GH, Plenderleith LJ, Sundararaman SA, Sharp PM, Hahn BH (2017) Out of Africa: origins and evolution of the human malaria parasites Plasmodium falciparum and Plasmodium vivax. Int J Parasitol 47:87–97

51. Kwiatkowski DP (2005) How Malaria Has Affected the Human Genome and What Human Genetics Can Teach Us about Malaria. The American Journal of Human Genetics 77:171–192

52. Mangano VD, Modiano D (2014) An evolutionary perspective of how infection drives human genome diversity: the case of malaria. Curr Opin Immunol 30:39–47

53. Ashley-Koch A, Yang Q, Olney RS (2000) Sickle hemoglobin (HbS) allele and sickle cell disease: a HuGE review. Am J Epidemiol 151:839–845

54. Gampio Gueye NS, Peko SM, Nderu D, Koukouikila-Koussounda F, Vouvoungui C, Kobawila SC, Velavan TP, Ntoumi F (2019) An update on glucose-6-phosphate dehydrogenase deficiency in children from Brazzaville, Republic of Congo. Malaria Journal 18:57

55. Williams TN (2016) Sickle Cell Disease in Sub-Saharan Africa. Hematol Oncol Clin North Am 30:343–358

56. Rees DC, Williams TN, Gladwin MT (2010) Sickle-cell disease. The Lancet 376:2018–2031

57. Ferreira A, Marguti I, Bechmann I, Jeney V, Chora Â, Palha NR, Rebelo S, Henri A, Beuzard Y, Soares MP (2011) Sickle Hemoglobin Confers Tolerance to Plasmodium Infection. Cell 145:398–409

58. LaMonte G, Philip N, Reardon J, et al (2012) Translocation of sickle cell erythrocyte microRNAs into Plasmodium falciparum inhibits parasite translation and contributes to malaria resistance. Cell Host Microbe 12:187–199

59. Archer NM, Petersen N, Clark MA, Buckee CO, Childs LM, Duraisingh MT (2018) Resistance to Plasmodium falciparum in sickle cell trait erythrocytes is driven by oxygen-dependent growth inhibition. Proceedings of the National Academy of Sciences 115:7350–7355

60. Langhi DM, Bordin JO (2006) Duffy blood group and malaria. Hematology 11:389–398

61. Golassa L, Amenga-Etego L, Lo E, Amambua-Ngwa A (2020) The biology of unconventional invasion of Duffy-negative reticulocytes by Plasmodium vivax and its implication in malaria epidemiology and public health. Malaria Journal 19:299

62. Cappellini MD, Fiorelli G (2008) Glucose-6-phosphate dehydrogenase deficiency. Lancet 371:64–74

63. Manjurano A, Sepulveda N, Nadjm B, et al (2015) African Glucose-6-Phosphate Dehydrogenase Alleles Associated with Protection from Severe Malaria in Heterozygous Females in Tanzania. PLoS Genet. https://doi.org/10.1371/journal.pgen.1004960

64. Flint J, Harding RM, Boyce AJ, Clegg JB (1993) The population genetics of the haemoglobinopathies. Baillieres Clin Haematol 6:215–262

65. Koralkova P, van Solinge WW, van Wijk R (2014) Rare hereditary red blood cell enzymopathies associated with hemolytic anemia - pathophysiology, clinical aspects, and laboratory diagnosis. Int J Lab Hematol 36:388–397

66. Bougouma EC, Sirima SB, Bougouma EC, Sirima SB (2020) Inherited Disorders of Hemoglobin and Plasmodium falciparum Malaria. Human Blood Group Systems and Haemoglobinopathies. https://doi.org/10.5772/intechopen.93807

67. Jiang H, Li N, Gopalan V, et al (2011) High recombination rates and hotspots in a Plasmodium falciparum genetic cross. Genome Biology 12:R33

68. Le Roch KG, Chung D-WD, Ponts N (2012) Genomics and Integrated Systems Biology in Plasmodium falciparum: A Path to Malaria Control and Eradication. Parasite Immunol 34:50–60

69. MalariaGEN, Ahouidi A, Ali M, et al (2021) An open dataset of Plasmodium falciparum genome variation in 7,000 worldwide samples. Wellcome Open Res 6:42

70. Benavente ED, Manko E, Phelan J, et al (2021) Distinctive genetic structure and selection patterns in Plasmodium vivax from South Asia and East Africa. Nat Commun 12:3160

71. Rutledge GG, Böhme U, Sanders M, et al (2017) Plasmodium malariae and P. ovale genomes provide insights into malaria parasite evolution. Nature 542:101–104

72. Hocking SE, Divis PCS, Kadir KA, Singh B, Conway DJ (2020) Population Genomic Structure and Recent Evolution of Plasmodium knowlesi, Peninsular Malaysia. Emerg Infect Dis 26:1749–1758

73. Osborne A, Manko E, Takeda M, et al (2021) Characterizing the genomic variation and population dynamics of Plasmodium falciparum malaria parasites in and around Lake Victoria, Kenya. Sci Rep 11:19809

74. Band G, Le QS, Clarke GM, et al (2019) Insights into malaria susceptibility using genome-wide data on 17,000 individuals from Africa, Asia and Oceania. Nat Commun 10:5732

75. Nag S, Dalgaard MD, Kofoed P-E, Ursing J, Crespo M, Andersen LO, Aarestrup FM, Lund O, Alifrangis M (2017) High throughput resistance profiling of Plasmodium falciparum infections based on custom dual indexing and Illumina next generation sequencing-technology. Sci Rep 7:2398

76. Jain M, Olsen HE, Paten B, Akeson M (2016) The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. Genome Biology 17:239

77. Ravi RK, Walton K, Khosroheidari M (2018) MiSeq: A Next Generation Sequencing Platform for Genomic Analysis. In: DiStefano JK (ed) Disease Gene Identification: Methods and Protocols. Springer, New York, NY, pp 223–232

Chapter 2: Characterising the genomic variation and population dynamics of *Plasmodium falciparum* malaria parasites in and around Lake Victoria, Kenya

RESEARCH PAPER COVER SHEET

Please note that a cover sheet must be completed <u>for each</u> research paper included within a thesis.

SECTION A – Student Details

Nagasaki Student No	59719003	**Title**	Miss
LSHTM Student ID No	Lsh1807687		
First Name(s)	Ashley Alexandra		
Surname/Family Name	Osborne		
Thesis Title	A multifaceted investigation of the genomics of malaria, from parasite to host, using next-generation sequencing technologies		
Nagasaki Supervisor(s)	Akira Kaneko, Kiyoshi Kita		
LSHTM Supervisor(s)	Taane Clark, Susana Campino		

If the Research Paper has previously been published please complete Section B, if not please move to Section C.

<u>SECTION B – Paper already published</u>

Where was the work published?	Scientific Reports

When was the work published?	2021-10-06		
If the work was published prior to registration for your research degree, give a brief rationale for its inclusion			
Have you retained the copyright for the work?*	**Yes**	Was the work subject to academic peer review?	**Yes**

*If yes, please attach evidence of retention. If no, or if the work is being included in its published format, please attach evidence of permission from the copyright holder (publisher or other author) to include this work.

SECTION C – Prepared for publication, but not yet published

Where is the work intended to be published?	
Please list the paper's authors in the intended authorship order:	
Stage of publication	Choose an item.

SECTION D – Multi-authored work

For multi-authored work, give full details of your role in the research included in the paper and in the preparation of the paper. (Attach a further sheet if necessary)	I carried out laboratory work preparing samples for WGS, including selective whole genome amplification, DNA clean-up, and shipment of samples. I performed bioinformatic analyses and interpreted the results under the supervision of my supervisors. I wrote and prepared the first draft of the manuscript that was circulated to my supervisors and co-authors.

SECTION E – Names and affiliations of co-author(s)

Please list all the co-authors' names and their affiliations.

Ashley Osborne – Faculty of Infectious and Tropical Diseases, London School of Hygiene and Tropical Medicine & School of Tropical Medicine and Global Health, Nagasaki University Emilia Manko - Faculty of Infectious and Tropical Diseases, London School of Hygiene and Tropical Medicine Mika Takeda – Department of Protozoology, Institute of Tropical Medicine, Nagasaki University Akira Kanko - Department of Parasitology, Osaka City University Wataru Kagaya - Department of Parasitology, Osaka City University Chim Chan – Department of Parasitology, Osaka City University Mtakai Ngara – Department of Microbiology, Karolinska Institutet James Kongere - Department of Parasitology, Osaka City University Kiyoshi Kita - School of Tropical Medicine and Global Health, Nagasaki University Susana Campino - Faculty of Infectious and Tropical Diseases, London School of Hygiene and Tropical Osamu Kaneko - Department of Protozoology, Institute of Tropical Medicine, Nagasaki University Jesse Gitaka – Directorate of Research and Innovation, Mount Kenya University Taane G. Clark – Faculty of Epidemiology and Population Health, London School of Hygiene and Tropical Medicine & Faculty of Infectious and Tropical Diseases, London School of Hygiene and Tropical Medicine

SECTION F

I confirm that all co-authors have agreed that the above paper will be included in my PhD thesis.

Student Signature	[redacted]
Date	06/07/23

LSHTM Supervisor Signature	[redacted]
Date	06/07/23

Nagasaki University Supervisor Signature	
Date	06/07/23

scientific reports

Check for updates

OPEN

Characterizing the genomic variation and population dynamics of *Plasmodium falciparum* malaria parasites in and around Lake Victoria, Kenya

Ashley Osborne[1], Emilia Manko[1], Mika Takeda[2], Akira Kaneko[3,4], Wataru Kagaya[3], Chim Chan[3], Mtakai Ngara[4], James Kongere[3,5], Kiyoshi Kita[6], Susana Campino[1,10], Osamu Kaneko[2,6,10], Jesse Gitaka[7,8,10] & Taane G. Clark[1,9,10]✉

Characterising the genomic variation and population dynamics of *Plasmodium falciparum* parasites in high transmission regions of Sub-Saharan Africa is crucial to the long-term efficacy of regional malaria elimination campaigns and eradication. Whole-genome sequencing (WGS) technologies can contribute towards understanding the epidemiology and structural variation landscape of *P. falciparum* populations, including those within the Lake Victoria basin, a region of intense transmission. Here we provide a baseline assessment of the genomic diversity of *P. falciparum* isolates in the Lake region of Kenya, which has sparse genetic data. Lake region isolates are placed within the context of African-wide populations using Illumina WGS data and population genomic analyses. Our analysis revealed that *P. falciparum* isolates from Lake Victoria form a cluster within the East African parasite population. These isolates also appear to have distinct ancestral origins, containing genome-wide signatures from both Central and East African lineages. Known drug resistance biomarkers were observed at similar frequencies to those of East African parasite populations, including the S160N/T mutation in the *pfap2mu* gene, which has been associated with delayed clearance by artemisinin-based combination therapy. Overall, our work provides a first assessment of *P. falciparum* genetic diversity within the Lake Victoria basin, a region targeting malaria elimination.

Despite decades of research and elimination campaigns, malaria remains a major global health threat and is the sixth leading cause of death in low-income countries, with an estimated 228 million cases and ~ 405,000 deaths worldwide in 2019 alone[1]. The Sub-Saharan African region accounts for more than 93% of all cases and deaths, with children under five years old disproportionately affected[1]. *Plasmodium falciparum* is the most prevalent and deadly malaria parasite in sub-Saharan Africa, as well as the cause of almost all severe disease, including severe malarial anaemia and cerebral malaria[2,3]. Even with progress over the past few decades, reduction in malaria-associated deaths has decelerated since 2016 and progress towards global malaria eradication overall has slowed, in part due to the emergence and spread of drug resistance in malaria parasites and insecticide resistance in their mosquito vectors[1]. Due to their isolated geography and distinctive population dynamics, inhabited islands, such as those within the Lake Victoria basin, offer unique environments to test malaria elimination strategies[4].

[1]Faculty of Infectious and Tropical Diseases, London School of Hygiene and Tropical Medicine, London, UK. [2]Department of Protozoology, Institute of Tropical Medicine, Nagasaki University, Nagasaki, Japan. [3]Department of Parasitology, Graduate School of Medicine, Osaka City University, Osaka, Japan. [4]Department of Microbiology, Tumor and Cell Biology, Karolinska Institutet, Stockholm, Sweden. [5]Centre for Research in Tropical Medicine and Community Development (CRTMCD), Hospital Road Next to Kenyatta National Hospital, Nairobi, Kenya. [6]School of Tropical Medicine and Global Health, Nagasaki University, Nagasaki, Japan. [7]Directorate of Research and Innovation, Mount Kenya University, Thika, Kenya. [8]Centre for Malaria Elimination, Mount Kenya University, Thika, Kenya. [9]Faculty of Epidemiology and Population Health, London School of Hygiene and Tropical Medicine, London, UK. [10]These authors contributed equally: Susana Campino, Osamu Kaneko, Jesse Gitaka and Taane G. Clark. ✉email: taane.clark@lshtm.ac.uk

Studies in low transmission settings have provided valuable insight into malaria elimination strategies that may be viable for islands, as well as an increased understanding of the role human movement plays in malaria transmission in these environments[5]. The implementation of intensive and targeted measures on islands in low-transmission settings, including vector control based strategies (e.g. insecticide-treated nets and indoor residual spraying) and mass drug administration, has led to malaria elimination in places such as the Maldives and Reunion archipelagos[6]. Recently, attention has moved to islands in high-transmission regions where there is an increased rate of gene flow within and between malaria parasite populations, which may aid in propagating novel traits detrimental to malaria treatment outcomes and elimination campaigns[4]. Past failures of malaria elimination campaigns on high-transmission islands, such as Zanzibar and the Comoros archipelago off the coast of East Africa, have highlighted the need for a better understanding of the roles human and vector migration play in settings with intense transmission patterns, and their impact on parasite population genetics and structure[7]. The development of methodologies that allow for the amplification of parasite DNA from infections with low parasite density, alongside the reduction in costs associated with next-generation whole genome sequencing (WGS), has made investigating parasite genomic diversity in malaria endemic settings attainable[8,9].

Malaria transmission and infection risk has decreased in Kenya and cases of the disease are generally not seen in high-elevation regions, such as Nairobi, due to the inability of mosquito vectors to survive at such altitudes. Low-elevation regions, however, such as the Lake Victoria basin and coastal regions along the Indian Ocean, still struggle with high rates of malaria transmission as they provide ideal breeding habitats for *Anopheles* mosquitoes[10,11]. Depending on the season, *P. falciparum* infection rates in the Lake Victoria basin can exceed 40% for those aged 2 to 10 years old[12]. Due to the size of Lake Victoria, and its location along three country borders with differing government policies concerning the implementation and funding of malaria control programmes, applying any targeted vector control in and around Lake Victoria has proven incredibly difficult[4,11]. There are multiple inhabited islands within the Kenyan territory of Lake Victoria, such as Mfangano island (population size 26,000) and Ngodhe island (population size 600–1000), which offer a unique opportunity to study malaria elimination strategies and changes in parasite genetic diversity.

To date, there has been limited research in the Lake Victoria region aimed at characterising the genomic diversity of the malaria parasites present. Previous research has assessed the genetic diversity through targeted genotyping of drug resistance genes or using microsatellite markers to gain insight into population dynamics[4]. However, such approaches underestimate the overall genetic variation of a population as only a small proportion of the genome is represented[13]. Consequently, the genomic diversity and population dynamics of malaria parasites from islands within the region remain largely unexplored. WGS technologies provide a means for generating a comprehensive picture of the epidemiology and structural variation of the *P. falciparum* populations within the Lake Victoria basin[8]. Here, to provide a baseline level of genomic diversity within the Lake region, we generated WGS data for two islands, as well as the mainland sub-county Suba District (population size: North 124,938; South 122,383) and compared the resulting variation to publicly available whole genome sequences, from Kenya and the wider African continent through the Pf3K project (https://www.malariagen.net/parasite/pf3k). Our analysis revealed that *P. falciparum* isolates from the Lake Victoria region of Kenya have distinct ancestral origins, with a high proportion tied to Central and East African lineages, and form a distinct sub-group within East Africa in population structure analyses. Known drug resistance biomarkers were observed in the Lake Victoria isolates at similar frequencies to those of East African parasite populations.

Results

Genome data and multi-clonality. A total of 940,191 high-quality SNPs were identified in the non-hypervariable regions of the *P. falciparum* genome. The final dataset comprised 784 isolates from 9 different countries in Africa (Supplementary Table S1), which included those collected in and around Lake Victoria (Mfangano island, Ngodhe island, and Suba District; N = 48); Kenya (Kilifi, Kisumu, and Kombewa, N = 134); East Africa (Tanzania and Uganda, N = 139); West Africa (The Gambia and Mauritania, N = 159); Central Africa (Cameroon, N = 98); South Central Africa (Democratic Republic of the Congo, N = 97); Southeast Africa (Madagascar and Malawi, N = 119). The samples from Lake Victoria were collected across two inhabited islands and one mainland site (Supplementary Table S1; Supplementary Fig. S1). The average number of pairwise SNP differences, a measure of genetic diversity, were broadly similar within locations around the Lake basin (Mfangano island 7900.8; Suba District 8708.1; Kisumu and Kombewa 9224.7).

Multi-clonality was measured using the F_{WS} metric, which assesses within-host diversity in relation to the local population diversity to characterise the risk of out-crossing/inbreeding, with monoclonal samples exhibiting "high" F_{WS} estimates (i.e. ≥ 0.95)[14,15]. In the samples collected from Mfangano and Ngodhe islands, a mean F_{WS} of 0.944 was observed, with 70.2% of samples exhibiting "high" F_{WS} estimates. When samples from Mfangano island and Ngodhe island were combined with samples from Suba District, a mean F_{WS} of 0.940 was observed (72.9% ≥ 0.95). Most samples from Mfangano (N = 19) and a handful of samples from Suba district (N = 9) were cultured in vitro prior to sequencing, which is consistent with the "high" F_{WS} estimates observed.

Isolates from Kisumu and Kombewa, Kenya had a mean F_{WS} of 0.777 (31.1% ≥ 0.95). The low proportion of samples with high F_{WS} scores is generally associated with a high degree of panmixis between the parasites in the population and a low population sub-structure. Samples from the Lake Victoria Region (i.e., Mfangano island, Ngodhe island, Suba District, Kisumu, Kombewa, and Muleba, Tanzania) and samples from East Africa (i.e., Kenya and Tanzania) had mean F_{WS} scores of 0.840 (48.1% ≥ 0.95) and 0.856 (51.5% ≥ 0.95) respectively.

P. falciparum isolates from Lake Victoria form a distinct sub-group within East Africa. A SNP-based principal components analysis (PCA) and neighbour-joining tree revealed that Lake Victoria isolates (i.e., Mfangano island, Ngodhe island, Suba District, Kisumu, and Kombewa) cluster with the East African subpopu-

lations when compared to larger regional African populations (Fig. 1A,B). To highlight the structure of the East African regional subpopulations, a SNP-based neighbour-joining tree and PCA were compiled using samples from Kenya, Tanzania, Uganda, and the Lake Victoria isolates. As anticipated, the Lake Victoria samples formed a sub-group within the larger East African dataset (Fig. 1C,D). A SNP-based neighbour-joining tree consisting of only samples from Lake Victoria provided a higher resolution of the population dynamics, and demonstrated that the isolates from the islands and Suba district appear to form a distinct genomic sub-group within the larger Lake Victoria population (Fig. 1E,F).

Ancestral admixture analysis provides insight into the ancestral origins of Lake Victoria subpopulation. Spatial modelling of allele sharing using genome-wide SNPs and geographical coordinates was used to evaluate the ancestral origins of Lake Victoria isolates alongside the regional African populations. This ancestral admixture analysis revealed that Lake Victoria isolates contain differing proportions of ancestral genome pieces compared with East African isolates, where the optimum number of ancestral populations (K value) is estimated to be 5 (K1-K5) (Fig. 2). The K5 ancestral population appears to be linked to Central African isolates (Cameroon-like, 88.3%), K4 is linked to East Africa (Kenya-like, 83.0%; Tanzania-like 73.0%), and the K3 population appears to be linked to West Africa (Gambia-like, 78.0%). Lake Victoria isolates appeared to share a high proportion of their ancestral genome with East (K4, 62.3%) and Central (K5, 27.2%) African isolates. Further, a high proportion resemble Ugandan isolates, in addition to those from Kenya or Tanzania (Supplementary Fig. S1). As anticipated, the ancestry of Ethiopia, the Horn of Africa, was quite distinct compared to the other regional African populations, comprising mainly of the K1 ancestral population (92.7%), with limited ancestry in Lake Victoria isolates (2.9%).

Analysis of IBD reveals differences within the Lake Victoria subpopulations. Analysis of identity-by-descent (IBD) was performed to understand the chromosome-level structure within subpopulations in and around Lake Victoria, Kenya. The proportion of pairs identical by descent at each SNP for all isolates was used to determine the extent of genomic relatedness. Isolates from Kisumu and Kombewa exhibited the highest fractions of pairwise IBD across the genome (Kisumu: median = 0.211, range = 0.182–0.238; Kombewa: median = 0.121, range = 0.069–0.200), reflecting high relatedness; while isolates from Mfangano and Suba exhibited lower fractions of IBD (Mfangano: median = 0.032, range = 0.018–0.142; Suba: median = 0.055, range = 0.055–0.132) (Supplementary Table S2). These lower fractions of IBD in the Mfangano and Suba isolates would suggest that relatedness is low, however their laboratory culture may be confounding these results as it is anticipated that samples in these populations would have high fractions of IBD across the genome. Examples of genome-wide IBD along each chromosome for individual sample sites and for regions across Africa are presented (Supplementary Fig. S2).

The top 5% of IBD positions in the island isolates from Lake Victoria were distributed across 31 regions on nine chromosomes [1,3,5,6,7,8,11,12,and13] (Supplementary Table S3). The region on chromosome 8 encompassed the *hydroxymethyldihydropterin pyrophosphokinase-dihydropteroate synthase* gene (*Pfdhps*, PF3D7_0810800), linked to Sulfadoxine–pyrimethamine (SP) antimalarial resistance. The top 5% of IBD positions in the mainland isolates from Lake Victoria were distributed across 13 regions on seven chromosomes [3,6,7,8,12,13,and14]. While the top 5% of IBD positions in East African isolates included 35 regions across eight chromosomes [3,4,5,6,7,8,12,and13]. The IBD region on chromosome 7 encompassed the *chloroquine resistance transporter* gene (*Pfcrt*, PF3D7_0709000), while the region on chromosome 8 included *Pfdhps* linked to SP resistance.

Population differentiation between Lake Victoria isolates and others. Fixation index (F_{ST}) analyses were performed to identify SNPs associated with the distinct subpopulations within the Lake Victoria region, and between other regional populations of the African continent. A genome-wide comparison of the differences in allele frequencies between samples collected from Suba District to those from Kisumu and Kombewa, identified 90 SNPs with an F_{ST} value of 0.5 or higher, suggesting some degree of genetic differentiation between these two subpopulations. In comparison, an analysis comparing Mfangano island with those from Kisumu and Kombewa identified 40 SNPs with an F_{ST} value greater than 0.5. A final comparison between samples collected from Mfangano island with those collected from Suba District resulted in 80 SNPs with an F_{ST} value of 0.5 or higher being identified. Across all island comparisons, most SNPs had an F_{ST} value of 0.5 or lower which suggests there is still frequent and consistent interbreeding between these populations. The lower number of significant SNPs for Mfangano comparisons may be explained by lower sequencing coverage leading to missing SNP genotypes, likely due to the lower parasite DNA concentrations often seen in the asymptomatic infections from which these samples were acquired. To identify genetic markers with the potential to be used to differentiate between isolates within the Lake Victoria region, and with populations across Africa, SNPs with an F_{ST} value of 1 (implying complete population differentiation) were compiled (Table 1). These may be used to determine the origin of a sample and monitor circulating parasites.

Identification of mutations in drug resistance candidate genes. Non-synonymous SNP mutations in resistance-associated genes were identified in historical isolates from coastal (e.g., Suba district) and island parasite populations (e.g., Mfangano island) within Lake Victoria. The frequencies of these mutations in the Lake Victoria isolates were compared to frequencies observed in East African parasite populations (e.g., Tanzania and Kenya) and West African populations (e.g., Mauritania and The Gambia).

Resistance markers for chloroquine and SP were observed in the Lake Victoria isolates at similar frequencies seen in the East African populations, and slightly higher frequencies than in West African populations (Supplementary Table S4). The *Pfcrt* K76T mutation, often used as the main marker of chloroquine resistance, was

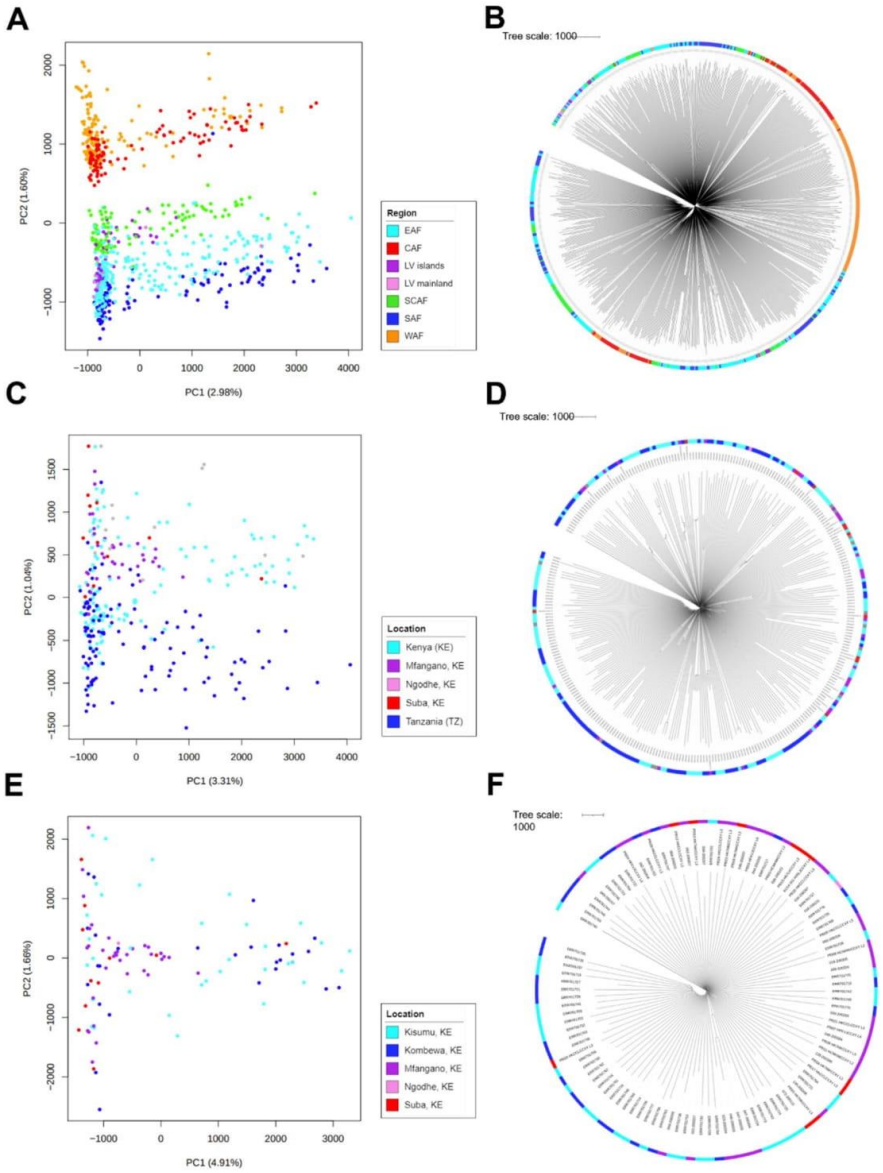


Figure 1. *Plasmodium falciparum* isolates from Lake Victoria (LV) form a distinct sub-group within East Africa. Principal component analysis (PCA) and neighbour-joining (NJ) trees generated from pairwise genetic distance matrices containing 940,191 quality filtered SNPs from 784 *P. falciparum* isolates. (**A**,**B**) A PCA plot and neighbour-joining tree for 784 isolates from East (EAF), Central (CAF), South Central (SCAF), Southeast (SAF), and West (WAF) Africa, LV mainland and islands. (**C**,**D**) A PCA plot and NJ tree for 321 isolates from EAF, LV mainland and islands. (**E**,**F**) A PCA plot and NJ tree for 109 isolates from Kenya, LV mainland and islands.

observed in 17.9% (N = 29) of the Lake Victoria isolates, 16.7% (N = 228) of the East African isolates, and 13.0% (N = 159) of West African isolates. Mutations F938Y and D1246Y in *Pfmdr1* have been associated with resistance to chloroquine, with D1246Y linked to lumefantrine. These mutations were observed in low frequencies in the

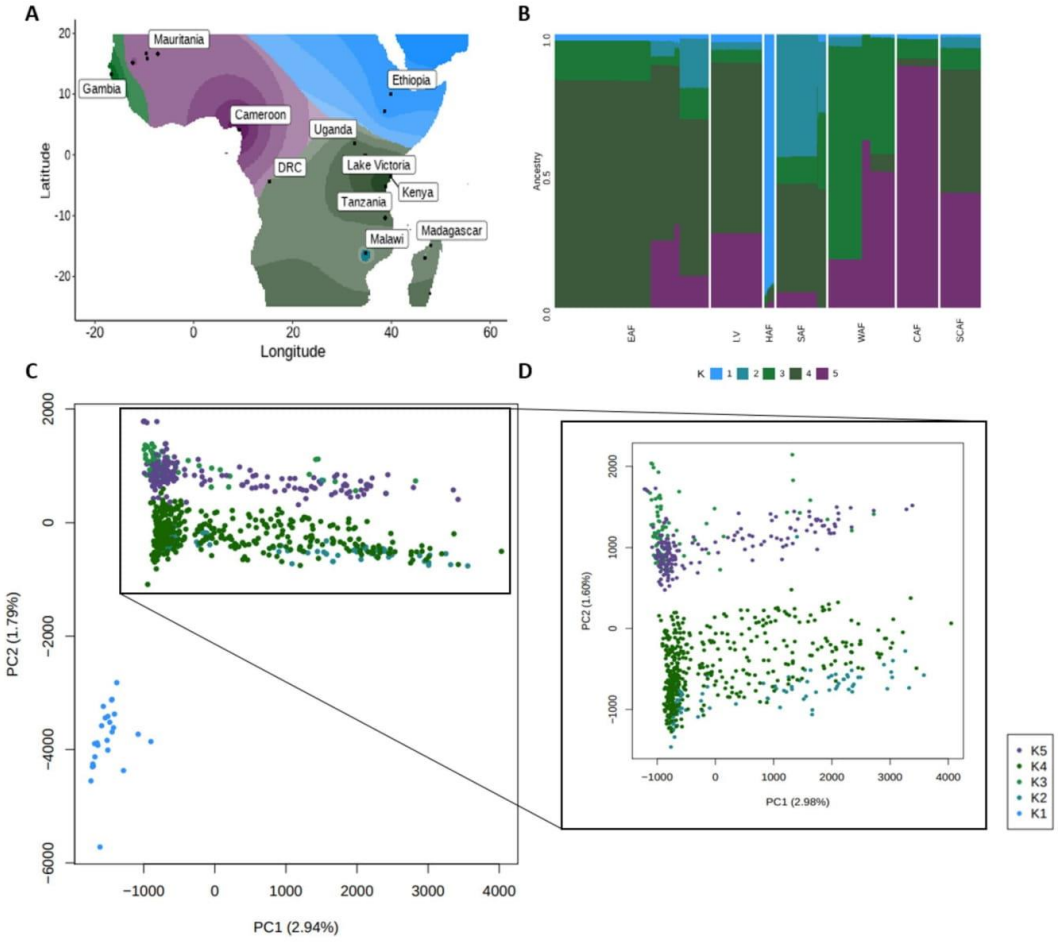


Figure 2. Genome-wide admixture ancestry proportions for regional *P. falciparum* populations across the African continent. (**A**) Geographic map of ancestry coefficients using $K = 5$ ancestral populations across Africa. (**B**) Ancestry per isolate (rows) for each regional population (columns). (**C**) Principal component analysis coloured using the Tess3r ancestry coefficients for the five predicted ancestral populations. (**D**) Cross-section of PCA excluding K1.

Lake Victoria (F938Y 3.6%; D1246Y 7.1%) and East African isolates (7.6%; 11.3%), but at higher frequencies in the West African populations (22.2%; 31.2%).

Pfdhfr mutations associated with SP resistance were observed at very high frequencies in both the Lake Victoria and East African isolates, with 100% of the Lake Victoria isolates containing the N51I and S108N mutations, compared to 93.3% and 99.5%, respectively, in the East African isolates. The C59R (Lake Victoria 92.9%; East Africa 87%) and I164L (Lake Victoria 7.1%; East Africa 2.2%) mutations were also present. The *Pfdhfr* mutations were observed at similar frequencies in the West African parasite populations, however the I164L polymorphism was not observed. The K540E mutation on *Pfdhps* associated with SP resistance was observed in 100% of the Lake Victoria isolates, higher than East (85.8%) and West (30.0%) African isolates.

No known mutations on *Pfk13* associated with artemisinin resistance were observed in the Lake Victoria or East African isolates. The S160N/T polymorphism on *Pfap2mu*, believed to be associated with delayed clearance by artemisinin combination therapies (ACTs), was observed in 17.8% (S160N) of the Lake Victoria isolates, 17.4% (S160N/T) of the East African isolates, and 38.0% (S160N) in West African isolates[16,17].

Regions under selection in Lake Victoria and other subpopulations. Analysis of haplotype structure within the Lake Victoria isolates was performed to determine genomic regions responding to positive directional selection, identified as having high local homozygosity relative to the neutral expectation. SNPs on genes under selective pressure in a distinct subpopulation were identified using the integrated haplotype score

Chromosome	Position	Ref	Population 1 (allele)	Population 2 (allele)
14	2,468,958	A	Mfangano, Ngodhe & Suba (A)	Kisumu & Kombewa (G/AG)
6	67,235	T	Mfangano (A)	Kisumu & Kombewa (T)
14	2,468,958	A	Mfangano (A)	Kisumu & Kombewa (G/AG)
4	1,081,227	T	Suba (T)	Kisumu & Kombewa (G)
9	1,126,737	A	Suba (A)	Kisumu & Kombewa (T)
10	838,875	A	Suba (G)	Kisumu & Kombewa (A)
11	250,576	A	Suba (A)	Kisumu & Kombewa (G)
3	1,004,970	C	Mfangano (C)	Suba (T)
3	1,004,977	T	Mfangano (T)	Suba (C)
5	184,986	G	Mfangano (G)	Suba (A)
7	701,509	G	Mfangano (G)	Suba (A)
11	966,450	C	Mfangano (C)	Suba (T)
13	422,582	T	Mfangano (T)	Suba (G)
14	1,242,566	T	Mfangano (T)	Suba (A)
5	1,116,502	T	Lake Victoria (T)	Kilifi, Kenya (A)
6	59,375	T	Lake Victoria (T)	Kenya, Tanzania and Uganda (G)
6	59,378	C	Lake Victoria (C)	Kenya, Tanzania and Uganda (A)
14	2,468,951	A	Lake Victoria (A)	Kenya, Tanzania and Uganda (T)
5	1,116,502	T	Lake Victoria (T)	Cameroon (A)
6	68,360	C	Lake Victoria (T)	Cameroon (C)
14	2,468,958	A	Lake Victoria (A)	Cameroon (G/AG)
5	1,116,502	T	Lake Victoria (T)	Democratic Republic of Congo (A)
6	68,360	C	Lake Victoria (T)	Democratic Republic of Congo (C)
14	2,468,958	A	Lake Victoria (A)	Democratic Republic of Congo (G)
14	2,468,958	A	Lake Victoria (A)	Malawi and Madagascar (G)
6	68,229	T	Lake Victoria (A)	The Gambia and Mauritania (T)
14	2,468,951	A	Lake Victoria (A)	The Gambia and Mauritania (T)

Table 1. Genetic markers for population differentiation between Lake Victoria isolates and other sub-populations according to the fixation index statistics (F_{ST}). SNPs with an F_{ST} value of 1. Ref = Pf3D7 reference allele; *Lake Victoria = Mfangano, Ngodhe, Suba, Kisumu, and Kombewa; East Africa = Kenya, Tanzania, and Uganda; Central Africa = Cameroon; South Central Africa = Democratic Republic of Congo; Southeast Africa = Malawi and Madagascar; West Africa = The Gambia and Mauritania.

(*iHS*) test statistic, whereas cross-population selective pressures were identified by the *XP-EHH* metric, which contrasts extended haplotype homozygosity (*EHH*) profiles between populations (Supplementary Fig. S3, Supplementary Table S5, Supplementary Table S6).

As anticipated, the most common SNPs under selective pressure were those on genes associated with the host immune response and immune evasion. Within the Mfangano isolates there were 4 genes of interest identified with SNPs that had significant *iHS* values ($(-\log_{10}[1-2|\Phi_{iHS}-0.5|]) > 4.0$), including: *P. falciparum* apical membrane protein 1 (*Pfama1*), one of the leading malaria vaccine candidates, merozoite surface protein 3 (*Pfmsp3*), associated with erythrocyte invasion and host immunity, and the Peptidase family C50 which are associated with the pathogenicity of *P. falciparum* through parasite immune evasion or invasion of host cells (Supplementary Fig. S3, Supplementary Table S5). In isolates from Suba district, only a Plasmodium RNA of unknown function (RUF6) was identified to have significant *iHS* values. Genes of interest within isolates collected from islands in Lake Victoria (e.g., Mfangano and Ngodhe island) included C50, Plasmepsin X (PMX), a mediator of parasite egress and invasion, *Pfmsp3*, and *Pfama1*. *Pfama1* and *Pfmsp3* are vaccine candidates. There was no strong evidence of positive selection (*iHS* values < 4.0) in two other key candidates (reticulocyte-binding protein homologue-5 (*PfRH5*), circumsporozoite protein (*PfCSP*)).

Cross-population analysis of isolates from the Lake Victoria islands with isolates from the Lake Victoria mainland (e.g., Suba, district and Kisumu, Kenya; Muleba, Tanzania) identified 2 genes with significant *XP-EHH* values ((-log10[p-value]) > 5.0), PMX and the Duffy binding-like merozoite surface protein 2 (*Pfdblmsp2*) (Supplementary Fig. S3, Supplementary Table S6). Isolates from the Lake Victoria islands were also compared to isolates collected from East Africa (e.g., Kenya, Tanzania, and Uganda) and 3 genes were identified, including *Pfdblmsp2*, PMX, and a cell traversal protein for ookinetes and sporozoites (CelTOS) which is a conserved antigen believed to have protective potential. Isolates from the Lake Victoria islands and mainland were compared to West and Central Africa to assess regional differences in selection pressures. When compared to West Africa, PMX was found to have significant *XP-EHH* values in the Lake Victoria isolates. When isolates from the Lake Victoria mainland were compared to West Africa, 4 genes were identified to have significant *XP-EHH* values, including several genes associated with drug resistance, including pro-drug activation and resistance esterase

(PARE; pepstatin), *Pfcrt* (chloroquine), and *Pfdhps* (SP), as well as CCR4-associated factor 1 (CAF1), an egress and invasion protein.

Discussion

Advances in next-generation sequencing technologies have provided an increasingly viable means for exploring the genomic variation and population dynamics of malaria parasites in high-transmission regions, where reduction in incidence rates have slowed in recent years. Such platforms can generate comprehensive snapshots of a region's epidemiology and structural variation[1,14,18]. Due to their isolated geography and distinctive population dynamics, islands in high-transmission regions offer a unique environment to study parasite gene flow and test malaria control strategies[4,5]. Here, we have provided a baseline level of genomic diversity within the Kenyan Lake Victoria basin that has revealed the presence of known drug resistance biomarkers. *P. falciparum* isolates from Lake Victoria form a cluster within the East African subpopulation, and island and Suba District isolates appear to form their own sub-group within the Lake Victoria basin. Further, the Lake Victoria parasite populations have central and east African lineage ancestry.

Low levels of genetic differentiation between the island sample sites are typically assumed to be due to the high levels of human traffic between the sampling sites. However, comparisons of isolates obtained from the Lake Victoria basin have highlighted distinct sub-grouping dynamics[4]. Although the movement of people and vectors is common around the islands and adjacent mainland, movement to and from other sections of the lake is limited by transportation infrastructure and geography. A handful of communities on the large island of Mfangano are somewhat isolated from communities with frequent migration to and from the mainland due to their rough terrain and limited access routes, which may account for some degree of differentiation between Mfangano parasites and those from the wider Lake Victoria basin.

Admixture analysis revealed insights into the ancestral origins of the Lake Victoria isolates within the wider African continent context. Lake Victoria isolates contained high proportions of ancestral genome fragments associated with Central African and East African populations. This analysis aligns well with the current prevailing *P. falciparum* origin hypothesis, suggesting that malaria first emerged in Central Africa before spreading out through early human migration, with subsequent more recent migration events throughout the African continent[19]. A closer look at the admixture ancestry of specific country populations demonstrated that isolates collected within the Lake Victoria basin (e.g., Kisumu, Kombewa, Mfangano, Ngodhe, and Suba from Kenya) appeared to share similar cumulative proportions of ancestral genome fragments as isolates from Uganda, but differed from the proportions observed in Kenyan and Tanzanian isolates. Future investigations with a greater density of sampling, including the integration of parasite and human genomics, could provide insights into the co-evolution of host–pathogen and be linked to ethnohistory.

Population differentiation (F_{ST}) analysis identified SNPs specific to the Lake Victoria subpopulations and African regional populations. These SNPs have the potential to be utilised in a molecular surveillance tool to determine the main routes of transmission and parasite migration. Population-specific SNPs were identified within the characterised nuclear genomes and offer a high degree of specificity on a smaller spatial scale, generally to the country of origin. These markers could be combined with population-specific organellar SNPs, which provide regional or continental level resolution, as well as SNPs on merozoite surface protein genes (e.g. *Pfmsp1* and *Pfmsp2*), which can be used to assess transmission intensity[20–22]. Further sample collection and characterization of *P. falciparum* genome data from the Lake Victoria region, alongside a robust African dataset, would be needed to generate a panel of SNPs with the degree of resolution needed to generate an effective molecular surveillance tool.

Analysis of the haplotype structure in the Lake Victoria isolates identified significant SNPs under selective pressure on genes associated with the host immune response and parasite immune evasion, such as *Pfama1*, *Pfmsp3*, and PMX. *Pfama1* is a leading malaria vaccine candidate due to the essential role it plays in parasite invasion of erythrocytes. *Pfama1* and other vaccine candidate genes have been found to be under balancing selection[23]. PMX is another important mediator of parasite invasion and egress[24,25]. *Pfmsp3* is an attractive vaccine candidate due to its role in mediating the host immune response to Plasmodium parasites[26]. Cross-population analysis comparing parasite populations from the Lake Victoria islands with the local mainland and East African populations identified SNPs on genes associated with immune evasion, host immune response, and sporozoite traversal (e.g., *Pfdblmsp2*, PMX, CelTOS). Duffy-binding-like domains are associated with a variety of pathogenic phenotypes in *P. falciparum* and are attractive malaria vaccine candidates given their crucial involvement in erythrocyte invasion. While CelTOS is believed to mediate transmission to mosquito and vertebrate hosts[27,28].

Analysis of identity-by-descent (IBD) revealed that isolates from Kisumu and Kombewa exhibited the highest fractions of pairwise IBD across the genome, reflecting high relatedness. Isolates from Mfangano and Suba exhibited lower fractions of IBD, which would suggest low relatedness between these isolates. However, given their close geographical proximity and the continuous movement of people between these sites, the culturing of some isolates prior to sequencing may be confounding these results[29]. The *Pfdhps* gene, known to be associated with resistance to SP, was found to be in the top 5% of IBD positions in island isolates from Lake Victoria, suggesting this gene is being highly conserved in this population, likely due to the continued use of SP as a method of intermittent preventative treatment in pregnancy (IPTp)[30,31].

Malaria treatment strategies within this region and across the African continent have been tailored to preserve the efficacy of existing antimalarial drugs, while further reducing parasite incidence rates[32]. ACTs are the current first-line treatment for uncomplicated malaria in Kenya, with SP only used as an IPTp, in accordance with WHO guidelines[1,33]. Monotherapy treatments have been largely phased out, with artemisinin only used in combination therapies and chloroquine no longer available over the counter in pharmacies or as a first-line treatment. Despite its discontinued use, resistance markers for chloroquine resistance still persist in these Lake Victoria

isolates and much of East Africa, apart from Malawi which has since seen the return of chloroquine sensitivity in its parasite populations following earlier changes to treatment policies[34,35]. It is anticipated that chloroquine sensitivity will return to parasite populations across Africa when drug selection pressure is completely removed, but the slow reduction in resistance marker frequencies suggests there is a low or negligent fitness cost associated with maintaining these polymorphisms.

A handful of resistance markers for SP were identified in the Lake Victoria isolates, with three established biomarkers (N51I and S108N on *Pfdhfr*; K540E on *Pfdhps*) observed to be almost fixed. The continued use of SP as an IPTp is likely providing enough driving pressure to maintain these polymorphisms within Lake Victoria and in parasite populations across the continent. However, given the high frequencies of these biomarkers, it is also possible that their presence does not come at a large fitness cost to the parasite. The S160N/T mutation in *Pfap2mu*, first documented in Kenyan children following delayed parasite clearance by ACT, was observed to be present at similar frequencies to the rest of East Africa and at a lower frequency than in West Africa[16]. Although further in vivo and in vitro studies are required to confirm the role polymorphisms in *Pfap2mu* play in resistance to ACTs, the presence of this mutation highlights the importance of monitoring parasite populations in high-transmission regions for the emergence of existing and novel resistance methods that could compromise the efficacy of ongoing malaria control and elimination strategies in Africa.

Due to the nature of working with field isolates, there were some limitations with this study. Most samples collected under this survey campaign are from asymptomatic individuals who generally have low parasitaemia levels, which often results in poor quality parasite DNA following extraction from blood spots. This issue was especially apparent in samples collected from Ngodhe island, where many cases of malaria are sub-microscopic and only detectable using PCR[12]. Due to these limitations, asymptomatic infections, in general, tend to be underrepresented in studies aimed at characterizing the genetic and genomic variation of malaria parasites, despite the fact they make up a majority of infections worldwide[36]. To overcome this, we applied selective whole genome amplification methodologies to select and amplify the parasite DNA in these field isolates[8,37,38]. This approach enabled us to increase parasite DNA concentrations in many samples that would normally not meet the requirements for whole genome sequencing, allowing us to provide the first baseline of the genomic variation and population dynamics of the Lake Victoria basin, with the potential for roll-out in other regions in Africa or globally.

Despite the progress made over the last decade, the reduction in malaria incidence rates and deaths has decelerated since 2016 and progress towards global malaria eradication has slowed. These trends, combined with setbacks to malaria control programmes, for example, due to the COVID-19 pandemic, have highlighted the need to generate comprehensive pictures of the epidemiology and structural variation of parasite populations. In high transmission regions, it is important to monitor the efficacy of current malaria control programmes and highlight potential challenges that may undermine their success. Here, we provided a baseline level of the genomic diversity of *P. falciparum* in the Lake Victoria basin of Kenya, a region whose genomic diversity has been previously unexplored. The application of sequencing technologies in malaria endemic regions will assist clinical management and disease control through surveillance activities. Known and putative drug resistance markers in established loci and population-specific markers are detectable using low-cost sequencing-based approaches (e.g., amplicon-based), suitable for a low resource setting. Large-scale approaches using (portable) whole genome sequencing and analysis will assist with understanding the transmission of malaria and provide insights into circulating drug resistance. Such insights will inform the deployment of antimalarial drugs and disease control tools and strategies in a region with currently high malaria burden, but with the potential for disease elimination.

Materials and methods

Study site selection. Parasite DNA sample collection for this study began in 2014 from several inhabited islands in Lake Victoria (Mfangano, N = 36; Ngodhe, N = 1) and the mainland sub-county of Homa Bay County, Suba district (Ungoye, N = 11) (Supplementary Fig. S4). Sample collection in this region was performed within an ongoing bi-annual survey, in accordance with the seasonal malaria prevalence. This seasonal transmission is linked to long (March–May) and short (October–December) rainy seasons. A peak in malaria prevalence is generally observed in June following the long rainy season and remains steady between September and February.

Permission to conduct this study was obtained from the Mount Kenya University Independent Ethics and Research Committee (MKU-IERC) (Approval reference: P609/10/2014) and the Ethics Committee at Osaka City University (Approval number: 3206) and performed in accordance with relevant guidelines and regulations. Workshops and sensitisation meetings were carried out with communities in order to seek community consent to study participation. Written informed consent was obtained from all study participants whose parasite DNA was used in this study.

Study population characteristics. This region of the Lake Victoria basin is occupied by both Luo and Suba ethnic groups, with Ngodhe occupied entirely by the former. The Luo ethnic group are generally migrant fishermen, while the Suba ethnic group rely more heavily on subsistence farming[4]. The main mode of transportation on the lake are dugout boats, but a ferry service is also in place, operating regularly to bring merchants and workers between the islands and the mainland commercial hub, Mbita.

Malaria species identification. Malaria species identification was carried out by microscopy, following WHO guidelines, at the Nagasaki University research station in Kenya by trained microscopists and confirmed at Osaka City University and LSHTM Malaria Reference Laboratory using established nested PCR assays[39,40].

Whole genome sequencing and bioinformatics. Twenty-nine isolates (collected years 2014 to 2015) were sequenced from DNA extracted from short-term (2–3month) cultures using Illumina MiSeq technology

with 300 bp paired end kits organised by Nagasaki University. The DNA for a further twenty-nine isolates (collected year 2020) was extracted from filter papers and amplified using an established selective whole genome amplification (SWGA) primer set and protocols[8,41], before being sequenced on Illumina MiSeq platform with 300 bp paired end kits at the LSHTM. All raw sequence data, following whole genome sequencing (WGS), was mapped to the Pf3D7 (*P. falciparum*) reference (version 3) genome using *bwa-mem* software (default parameters). SNPs and short insertions and deletions (indels) were called using the *samtools* and GATK software suites[28,29]. SNPs occurring in non-unique, highly variable (e.g., *var* genes), low quality or low coverage regions were discarded. Mixed call SNPs were assigned genotypes determined by a ratio of coverage in which nucleotide calls were 80% or higher. Samples with a coverage across the genome averaging less than 5 were not included in any analysis. Of the 58 samples sequenced, 10 were removed, leaving a total of 784 isolates in the final analyses.

Characterising genomic variants and population genetic analyses. Population structure was investigated using neighbour-joining tree and principal component analyses, both using pairwise genetic distance matrices based on SNPs. Nucleotide diversity and population differentiation (F_{ST}) metrics were calculated using the *pegas* R library[42]. The visualization and annotation of neighbour-joining trees was performed using *iTOL*[43]. The variants in drug-resistance genes were extracted from the sequence alignments, and annotated using *snpEff* software, which determined the type of mutation (e.g. non-synonymous, synonymous, or intergenic), as well as the codon and protein shift caused by any non-synonymous mutations and the projected impact of the polymorphism[44]. Grouping of samples into regional populations (e.g., Lake Victoria islands, Lake Victoria mainland, and East Africa) was determined based on patterns or human and vector migration within the region, alongside input from collaborators. WGS data from isolates collected on Mfangano island matched closely with the single isolate obtained from Ngodhe island, and were therefore combined into one population for further analysis.

The R-based *Tess3r* package, which estimates ancestry proportions by modelling on continuous genetic variation across space, was used to calculate admixture based on the spatial modelling of allele sharing using genome-wide SNPs and geographical coordinates from sampling sites (specified in the Pf3k dataset or recorded during sample collection for Lake Victoria isolates)[45–47]. An optimum K value for ancestral admixture coefficients was estimated as 5 from a cross-validation of 1 to 10 dimensions of eigenvalue decay. Bi-allelic sites were included in this analysis, while mixed calls were imputed to alternative and missing calls were left unchanged. The default spatial regularisation parameter ($\sigma = 1$) was used, which attributes equal weight to the loss and penalty functions. *Tess3r* was run 50 times for K5 to K10 and the best Q matrix of ancestry coefficients for every isolate was retained. This analysis used the alternating projected least squares algorithm (APLS; method = "projected.ls"). Plots were visualised using the R-based package *ggplot2* and surfaces were interpolated using the R-based package *Krig*[48–50].

Multi-clonality, or within-host infection complexity, was measured by calculating the F_{WS} metric, using an in-house script, which assesses within-host diversity in relation to the local population diversity in order to characterise the risk of out-crossing/inbreeding by estimating the fixation of alleles, on a scale of 0 to 1, within each infection[14]. Based on previous studies, an $F_{WS} \geq 0.95$ is considered highly indicative of a clonal infection.

Identity-by-descent (IBD) was used to assess connectivity between parasites within the Lake Victoria region and between subpopulations within the region. This was achieved by estimating the pairwise fraction of shared ancestry between genomic segments, the "IBD fraction", that were inferred to have descended from a recent common ancestor without undergoing any intervening recombination. These IBD fractions were calculated using the *hmmIBD* software which accounts for recombination using a hidden Markov model-based approach[51]. The rate of recombination for *P. falciparum* is estimated to be 13.5 Kb per centiMorgan (cM), which should result in chromosomal crossover events occurring at an average rate of approximately 1% per generation.

Regions of the genome under putative positive directional selection were scanned using population-based measures of haplotype diversity within (iHS) or between (XP-EHH) populations using the R-based *rehh* package[52,53].

Data availability

Public accession numbers for raw sequence data analysed are contained in SRA studies ERP000190 and ERP000199, as well as being accessible from the Pf3k project website (https://www.malariagen.net/projects/pf3k). Lake Victoria raw sequences are available from the EBI SRA (Project accession PRJEB46180) and DDBJ (BioProject Accession Number PRJDB12148).

Received: 16 July 2021; Accepted: 22 September 2021
Published online: 06 October 2021

References

1. World Health Organization. *World Malaria Report 2019* (World Health Organization, 2019).
2. Perkins, D. J. *et al.* Severe Malarial Anemia: Innate Immunity and Pathogenesis. *Int. J. Biol. Sci.* **7**(9), 1427–1442 (2011).
3. Rénia, L. *et al.* Cerebral malaria. *Virulence.* **3**(2), 193–201 (2012).
4. Mulenge, F. M. *et al.* Genetic diversity and population structure of plasmodium falciparum in Lake Victoria islands, a region of intense transmission. *Am. J. Trop. Med. Hyg.* **95**(5), 1077–1085 (2016).
5. Lum, J. K. *et al.* Malaria dispersal among islands: Human mediated Plasmodium falciparum gene flow in Vanuatu, Melanesia. *Acta Trop.* **90**(2), 181–185 (2004).
6. World Health Organization, Otten, M. & Williams, R. *World Malaria Report 2009*. World Health Organization; 2009
7. Le Menach, A. *et al.* Travel risk, malaria importation and malaria transmission in Zanzibar. *Sci. Rep.* **1**, 93 (2011).
8. Oyola, S. O. *et al.* Whole genome sequencing of Plasmodium falciparum from dried blood spots using selective whole genome amplification. *Malar. J.* **15**(1), 597 (2016).

9. Gomes, A. R. *et al.* Genetic diversity of next generation antimalarial targets: A baseline for drug resistance surveillance programmes. *Int. J. Parasitol. Drugs Drug Resist.* **7**(2), 174–180 (2017).
10. Okara, R. M. *et al.* Distribution of the main malaria vectors in Kenya. *Malar. J.* **4**(9), 69 (2010).
11. Minakawa, N., Dida, G. O., Sonye, G. O., Futami, K. & Njenga, S. M. Malaria vectors in Lake Victoria and adjacent habitats in Western Kenya. *PLoS ONE* **7**, e32725 (2012).
12. Kagaya, W. *et al.* Malaria resurgence after significant reduction by mass drug administration on Ngodhe Island, Kenya. *Sci. Rep.* **9**(1), 1–11 (2019).
13. Daniels, R. F. *et al.* Modeling malaria genomics reveals transmission decline and rebound in Senegal. *Proc. Natl. Acad. Sci.* **112**(22), 7067–7072 (2015).
14. Manske, M. *et al.* Analysis of Plasmodium falciparum diversity in natural infections by deep sequencing. *Nature* **487**(7407), 375–379 (2012).
15. Auburn, S. *et al.* Characterization of within-host plasmodium falciparum diversity using next-generation sequence data. *PLoS ONE* **7**(2), e32891 (2012).
16. Beshir, K. B. *et al.* Residual Plasmodium falciparum parasitemia in Kenyan children after artemisinin-combination therapy is associated with increased transmission to mosquitoes and parasite recurrence. *J. Infect. Dis.* **208**(12), 2017–2024 (2013).
17. Henriques, G. *et al.* Directional selection at the pfmdr1, pfcrt, pfubp1, and pfap2mu loci of Plasmodium falciparum in Kenyan children treated with ACT. *J. Infect. Dis.* **210**(12), 2001–2008 (2014).
18. Le Roch, K. G., Chung, D.-W.D. & Ponts, N. Genomics and integrated systems biology in plasmodium falciparum: A path to malaria control and eradication. *Parasite Immunol.* **34**(2–3), 50–60 (2012).
19. Patin, E. *et al.* Dispersals and genetic adaptation of Bantu-speaking populations in Africa and North America. *Science* **356**(6337), 543–546 (2017).
20. Daniels, R. *et al.* A general SNP-based molecular barcode for Plasmodium falciparum identification and tracking. *Malar. J.* **7**(1), 223 (2008).
21. Preston, M. D. *et al.* A barcode of organellar genome polymorphisms identifies the geographic origin of Plasmodium falciparum strains. *Nat. Commun.* **5**(1), 4052 (2014).
22. Mwingira, F. *et al.* Plasmodium falciparum msp1, msp2 and glurp allele frequency and diversity in sub-Saharan Africa. *Malar J.* **6**(10), 79 (2011).
23. Samad, H. *et al.* Imputation-based population genetics analysis of plasmodium falciparum malaria parasites. *PLoS Genet.* **11**(4), e1005131 (2015).
24. Remarque, E. J., Faber, B. W., Kocken, C. H. M. & Thomas, A. W. Apical membrane antigen 1: A malaria vaccine candidate in review. *Trends Parasitol.* **24**(2), 74–84 (2008).
25. Nasamu, A. S. *et al.* Plasmepsins IX and X are essential and druggable mediators of malaria parasite egress and invasion. *Science* **358**(6362), 518–522 (2017).
26. Polley, S. D. *et al.* Plasmodium falciparum merozoite surface protein 3 is a target of allele-specific immunity and alleles are maintained by natural selection. *J. Infect. Dis.* **195**(2), 279–287 (2007).
27. Hodder, A. N. *et al.* Insights into duffy binding-like domains through the crystal structure and function of the merozoite surface protein MSPDBL2 from Plasmodium falciparum. *J. Biol. Chem.* **287**(39), 32922–32939 (2012).
28. Kariu, T., Ishino, T., Yano, K., Chinzei, Y. & Yuda, M. CelTOS, a novel malarial protein that mediates transmission to mosquito and vertebrate hosts. *Mol. Microbiol.* **59**(5), 1369–1379 (2006).
29. Brown, A. C. & Guler, J. L. From circulation to cultivation: Plasmodium in vivo versus in vitro. *Trends Parasitol.* **36**(11), 914–926 (2020).
30. Ravenhall, M. *et al.* Characterizing the impact of sustained sulfadoxine/pyrimethamine use upon the Plasmodium falciparum population in Malawi. *Malar. J.* **15**, 1–11 (2016).
31. Turkiewicz, A. *et al.* Genetic diversity of the Plasmodium falciparum GTP-cyclohydrolase 1, dihydrofolate reductase and dihydropteroate synthetase genes reveals new insights into sulfadoxine-pyrimethamine antimalarial drug resistance. *PLoS Genet.* **16**(12), e1009268 (2020).
32. Division of Malaria Control. National Malaria Policy 2010 [Internet]. Republic of Kenya Ministry of Public Health and Sanitation; 2010. Available from: http://www.nmcp.or.ke
33. Ministry of Public Health and Sanitation & Ministry of Medical Services. National guidelines for the diagnosis, treatment and prevention of malaria in Kenya. 2014.
34. Frosch, A. E. P. *et al.* Return of widespread chloroquine-sensitive Plasmodium falciparum to Malawi. *J. Infect. Dis.* **210**(7), 1110–1114 (2014).
35. Kublin, J. G. *et al.* Reemergence of chloroquine-sensitive Plasmodium falciparum malaria after cessation of chloroquine use in Malawi. *J. Infect. Dis.* **187**(12), 1870–1875 (2003).
36. Bousema, T., Okell, L., Felger, I. & Drakeley, C. Asymptomatic malaria infections: Detectability, transmissibility and public health relevance. *Nat. Rev. Microbiol.* **12**(12), 833–840 (2014).
37. Ibrahim, A. *et al.* Selective whole genome amplification of Plasmodium malariae DNA from clinical samples reveals insights into population structure. *Sci. Rep.* **10**(1), 10832 (2020).
38. Benavente, E. D. *et al.* Whole genome sequencing of amplified Plasmodium knowlesi DNA from unprocessed blood reveals genetic exchange events between Malaysian Peninsular and Borneo subpopulations. *Sci. Rep.* **9**(1), 9873 (2019).
39. Isozumi, R. *et al.* Improved detection of malaria cases in island settings of Vanuatu and Kenya by PCR that targets the Plasmodium mitochondrial cytochrome c oxidase III (cox3) gene. *Parasitol. Int.* **64**(3), 304–308 (2015).
40. Mangold, K. A. *et al.* Real-time PCR for detection and identification of Plasmodium spp. *J. Clin. Microbiol.* **43**(5), 2435–2440 (2005).
41. Clarke, E. L. *et al.* swga: A primer design toolkit for selective whole genome amplification. *Bioinforma Oxf. Engl.* **33**(14), 2071–2077 (2017).
42. Paradis, E. pegas: An R package for population genetics with an integrated-modular approach. *Bioinformatics* **26**(3), 419–420 (2010).
43. Letunic, I. & Bork, P. Interactive Tree Of Life (iTOL): An online tool for phylogenetic tree display and annotation. *Bioinformatics* **23**(1), 127–128 (2007).
44. Cingolani, P. *et al.* A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w 1118; iso-2; iso-3. *Fly (Austin).* **6**(2), 80–92 (2012).
45. Caye, K., Deist, T. M., Martins, H., Michel, O. & François, O. TESS3: Fast inference of spatial population structure and genome scans for selection. *Mol. Ecol. Resour.* **16**(2), 540–548 (2016).
46. Caye, K., Jay, F., Michel, O. & François, O. Fast inference of individual admixture coefficients using geographic data. *Ann. Appl. Stat.* **12**(1), 586–608 (2018).
47. Martins, H., Caye, K., Luu, K., Blum, M. G. B. & François, O. Identifying outlier loci in admixed and in continuous populations using ancestral population differentiation statistics. *Mol. Ecol.* **25**(20), 5029–5042 (2016).
48. Wickham, H. *ggplot2: Elegant Graphics for Data Analysis* 2nd edn. (Springer, 2016).
49. Hastie, T. & Tibshirani, R. Generalized additive models. *Stat. Sci.* **1**(3), 297–310 (1986).
50. Cressie, N. A. C. *Statistics for Spatial Data*. Rev. ed. Wiley; 1993. 900 p. (Wiley series in probability and mathematical statistics).

51. hmmIBD: software to infer pairwise identity by descent between haploid genotypes | Malaria Journal | Full Text [Internet]. [cited 2021 Apr 28]. https://doi.org/10.1186/s12936-018-2349-7
52. Gautier, M. & Vitalis, R. rehh: An R package to detect footprints of selection in genome-wide SNP data from haplotype structure. *Bioinforma Oxf. Engl.* **28**(8), 1176–1177 (2012).
53. Gautier, M., Klassmann, A. & Vitalis, R. rehh 2.0: A reimplementation of the R package rehh to detect positive selection from haplotype structure. *Mol. Ecol. Resour.* **17**(1), 78–90 (2017).

Acknowledgements
We wish to thank the study participants and communities. We thank Sachie Takahama for her expertise. Human red blood cells and plasma were obtained from the Nagasaki Red Cross Blood Center. This study was conducted in part at the Joint Usage/Research Center on Tropical Disease, Institute of Tropical Medicine, Nagasaki University, Japan.

Author contributions
A.K., K.K., S.C., O.K., J.G. and T.G.C. conceived and designed the study; MT and JG cultured malaria parasites and contributed parasite DNA for sequencing; M.T., W.K., C.C., M.N., J.K., O.K. and J.G. provided biological materials and data. SC and OK coordinated the sequencing of samples; A.O. and E.M. performed the bioinformatic and statistical analysis, under the supervision of S.C. and T.G.C.; A.O. wrote the first draft of the manuscript, and the final version included edits from all authors. The final manuscript was read and approved by all authors.

Funding
AO is supported by a Nagasaki University—LSHTM PhD studentship funded by the WISE programme of MEXT. SC is funded by the Medical Research Council UK (Grant No. MR/M01360X/1) and BBSRC UK (BB/R013063/1). AK received support from JSPS KAKENHI (Grant Nos. JP18KK0248 and JP19H01080) and JICA/AMED joint research project (SATREPS) (Grant No. 20JM0110020H0002). OK received support from JSPS KAKENHI JP19KK0220, Japan. JG received support from a Tackling Infectious Burden in Africa (TIBA) fellowship, the African Academy of Sciences, and the Japan Society for Promotion of Sciences. TGC is supported by the Medical Research Council UK (Grant Nos. MR/K000551/1, MR/M01360X/1, MR/N010469/1, MR/R020973/1) and BBSRC (BB/R013063/1). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests
The authors declare no competing interests.

Additional information
Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-021-99192-1.

Correspondence and requests for materials should be addressed to T.G.C.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Supplementary Table S1. Summary data for *P. falciparum* isolates included in the SNP-based pairwise genetic distance matrices used to determine trends in population dynamics and identify genetic markers in sub-populations.

Country	Region	Number of samples
Kenya (Lake Victoria)	East Africa	48*
Kenya (other)	East Africa	134
Tanzania	East Africa	125
Uganda	East Africa	14
Madagascar	Southeast Africa	22
Malawi	Southeast Africa	97
Mauritania	West Africa	79
The Gambia	West Africa	80
Cameroon	Central Africa	98
Democratic Rep. of Congo	South Central Africa	97

* Mfangano Island (36), Ngodhe Island (1), Suba District (11)

Supplementary Table S2. Fractions of pairwise identity-by-descent (IBD) across the genome.

Population	Median IBD	Range
Mfangano	0.032	0.018 – 0.142
Suba	0.055	0.055 – 0.132
Kisumu	0.211	0.182 – 0.238
Kombewa	0.121	0.069 – 0.200
Lake Victoria islands*	0.027	0.014 – 0.116
Lake Victoria mainland**	0.046	0.033 – 0.172
Kilifi, Kenya	0.040	0.024 – 0.204
Central Africa	0.042	0.030 – 0.132
East Africa	0.032	0.019 – 0.150
West Africa	0.040	0.025 – 0.210

*Lake Victoria islands (e.g., Mfangano and Ngodhe), ** Lake Victoria mainland (e.g., Suba, Kisumu, and Kombewa, KE), East Africa (e.g., Kilifi, Kenya; Muleba, Tanzania; Nachingwea, Tanzania; Apac, Uganda), Central Africa (e.g., Cameroon), West Africa (e.g., The Gambia and Mauritania)

Supplementary Table S3. Top 5% of identity-by-descent (IBD) regions in Lake Victoria isolates.

Chr	Start	End	Fraction	LV Category*	Location	Gene ID (PF3D7_)	Gene product	Gene name
1	530001	540000	0.015	Islands	537109 – 538025	0113900	CX3CL1-binding protein 1	*CBP1*
3	120001	130000	0.041	Islands	119458 – 124735	0302200	cytoadherence linked asexual protein 3.2	*CLAG3.2*
3	130001	140000	0.033	Islands	135418 – 140660	0302500	cytoadherence linked asexual protein 3.1	*CLAG3.1*
5	880001	890000	0.017	Islands	882376 – 884901	0521700	ATP-dependent RNA helicase DDX1, putative	*DDX1*
5	900001	910000	0.025	Islands	901414 – 902173	0522200	transcription initiation factor TFIID subunit 10, putative	*TAF10*
5	1010001	1020000	0.015	Islands	1013453 – 1014550	0524400	ribosome-interacting GTPase 1, putative	*RBG1*
6	1060001	1070000	0.017	Islands	1061115 – 1062891	0626300	3-oxoacyl-acyl-carrier protein synthase I/II	*FabB/FabF*
6	1070001	1080000	0.015	Islands	1078995 – 1081320	0626800	pyruvate kinase	*PyrK*
6	1080001	1090000	0.015	Islands	1078995 – 1081320	0626800	pyruvate kinase	*PyrK*
6	1110001	1120000	0.056	Islands	1114544 – 1117537	0627800	acetyl-CoA synthetase, putative	*ACS*
6	1200001	1210000	0.025	Islands	1205190 – 1207781	0629300	phospholipase, putative	*PL*
6	1230001	1240000	0.019	Islands	1221941 – 1242922	0629700	SET domain protein	*SET1*
6	1240001	1250000	0.031	Islands	1221941 – 1242922	0629700	SET domain protein	*SET1*
6	1250001	1260000	0.034	Islands	1254907 – 1256940	0630100	alpha/beta hydrolase	*N/A*
7	220001	230000	0.030	Islands	216024 – 229072	0704600	HECT-type E3 ubiq. ligase	*UT*
7	430001	440000	0.049	Islands	435089 – 436195	0709700	Prodrug activation and resistance esterase	*PARE*
7	470001	480000	0.018	Islands	478468 – 479138	0710600	60S ribosomal protein L34	*RPL34*
8	500001	510000	0.114	Islands	508224 – 512428	0809900	JmjC domain-containing protein, putative	*JmjC1*
8	510001	520000	0.116	Islands	508224 – 512428	0809900	JmjC domain-containing protein, putative	*JmjC1*
8	540001	550000	0.018	Islands	548200 – 550616	0810800	hydroxymethyldihydropteri n pyrophosphokinase-dihydropteroate synthase	*PPPK-DHPS*
8	540001	550000	0.018	Islands	541971 – 544796	0810600	ATP-dependent RNA helicase DBP1, putative	*DBP1*
8	550001	560000	0.015	Islands	548200 – 550616	0810800	hydroxymethyldihydropteri n pyrophosphokinase-dihydropteroate synthase	*PPPK-DHPS*
11	530001	540000	0.017	Islands	534973 – 536499	1113900	mitogen-activated protein kinase 2	*MAPK2*
11	550001	560000	0.016	Islands	555514 – 558668	1114700	cyclin-dependent-like kinase CLK3	*CLK3*
12	900001	910000	0.057	Islands	907203 – 914501	1222600	AP2 domain transcription factor AP2-G	*AP2-G*

12	910001	920000	0.057	Islands	907203 – 914501	1222600	AP2 domain transcription factor AP2-G	*AP2-G*
12	920001	930000	0.054	Islands	927825 – 929991	1223100	cAMP-dependent protein kinase regulatory subunit	*PKAr*
12	980001	990000	0.019	Islands	988628 – 991255	1224300	polyadenylate-binding protein 1, putative	*PABP1*
12	990001	1000000	0.029	Islands	998353 – 999275	1224500	histone chaperone ASF1, putative	*ASF1*
12	990001	1000000	0.029	Islands	988628 – 991255	1224300	polyadenylate-binding protein 1, putative	*PABP1*
13	100001	110000	0.024	Islands	99548 – 100521	1301700	CX3CL1-binding protein 2	*CBP2*
3	120001	130000	0.036	Mainland	119458 – 124735	0302200	cytoadherence linked asexual protein 3.2	*CLAG3.2*
6	1110001	1120000	0.152	Mainland	1114544 – 1117537	0627800	acetyl-CoA synthetase	*ACS*
7	470001	480000	0.038	Mainland	478468 – 479138	0710600	60S ribosomal protein L34	*RPL34*
8	410001	420000	0.054	Mainland	416344 – 418065	0808200	plasmepsin X	*PMX*
8	500001	510000	0.167	Mainland	508224 – 512428	0809900	JmjC domain-containing protein, putative	*JmjC1*
8	510001	520000	0.143	Mainland	508224 – 512428	0809900	JmjC domain-containing protein, putative	*JmjC1*
12	910001	920000	0.038	Mainland	907203 – 914501	1222600	AP2 domain transcription factor AP2-G	*AP2-G*
12	920001	930000	0.045	Mainland	927825 – 929991	1223100	cAMP-dependent protein kinase regulatory subunit	*PKAr*
12	970001	980000	0.039	Mainland	974372 – 975541	1224000	GTP cyclohydrolase 1	*GCH1*
12	980001	990000	0.034	Mainland	988628 – 991255	1224300	polyadenylate-binding protein 1, putative	*PABP1*
12	990001	1000000	0.036	Mainland	998353 – 999275	1224500	histone chaperone ASF1	*ASF1*
12	990001	1000000	0.036	Mainland	988628 – 991255	1224300	polyadenylate-binding protein 1, putative	*PABP1*
13	100001	110000	0.040	Mainland	99548 – 100521	1301700	CX3CL1-binding protein 2	*CBP2*
14	2700001	2710000	0.033	Mainland	2709418 – 2713192	1466300	26S proteasome regulatory subunit RPN2, putative	*RPN2*

*Lake Victoria islands (e.g., Mfangano and Ngodhe), Lake Victoria mainland (e.g., Suba, Kisumu, and Kombewa, KE), East Africa (e.g., Kilifi, KE; Muleba, TZ; Nachingwea, TZ; Apac, UG), Central Africa (e.g., Cameroon), West Africa (e.g., The Gambia and Mauritania)

Supplementary Table S4. Non-synonymous single nucleotide polymorphisms (SNPs) in known drug resistance genes. Known resistance-conferring SNPs highlighted in **bold**.

Gene	Position	Ref	Alt	Mutation	LV islands* MAF (n = 29)	East Africa* MAF (n = 228)	West Africa* MAF (n = 167)	Maximum F_{ST}
Pfcrt	403291	G	T	**D24Y**	**0.071**	**0.078**	**0.000**	**0.099**
	403625	A	C	**K76T**	**0.179**	**0.167**	**0.130**	**0.440**
	404407	G	T	**A220S**	**0.179**	**0.155**	**0.127**	**0.373**
	404836	C	G	**Q271E**	**0.179**	**0.164**	**0.128**	**0.413**
	405362	A	G	**N326S**	**0.000**	**0.0075**	**0.048**	**0.014**
	405596	G	A	A355T	0.000	0.0038	0.000	0.017
	405600	T	C	**I356T**	**0.000**	**0.0075**	**0.180**	**0.340**
	405838	G	T	**R371I**	**0.179**	**0.175**	**0.124**	**0.403**
Pfdhfr	748239	A	T	**N51I**	**1.000**	**0.933**	**0.925**	**0.137**
	748262	T	C	**C59R**	**0.929**	**0.870**	**0.930**	**0.047**
	748410	G	A	**S108N**	**1.000**	**0.995**	**0.930**	**0.129**
	748577	A	T	**I164L**	**0.071**	**0.022**	**0.000**	**0.089**
Pfdhps	549256	A	G	N294S	0.000	0.002	0.000	0.005
	549685	G	C	G437A	0.000	0.120	0.213	0.298
	549993	A	G	**K540E**	**1.000**	**0.858**	**0.300**	**0.944**
	550117	C	G	A581G	0.036	0.015	0.000	0.100
Pfmdr1	957908	G	C	E7Q	0.036	0.005	0.000	0.095
	957990	A	G	K34R	0.000	0.005	0.125	0.006
	958145	A	T	**N86Y**	**0.000**	**0.129**	**0.244**	**0.153**
	958440	A	T	**Y184F**	**0.500**	**0.374**	**0.155**	**0.129**
	958484	A	T	T199S	0.036	0.012	0.000	0.167
	959307	A	G	N473S	0.000	0.003	0.000	0.011
	959399	A	T	N504Y	0.000	0.003	0.000	0.011
	959991	C	A	S701Y	0.000	0.003	0.000	0.011
	960258	C	G	T790S	0.000	0.003	0.000	0.011
	960404	A	G	I839V	0.000	0.002	0.000	0.005
	960702	T	A	**F938Y**	**0.036**	**0.076**	**0.222**	**0.026**
	961481	C	A	Q1198K	0.000	0.002	0.000	0.005
	961625	G	T	**D1246Y**	**0.071**	**0.113**	**0.312**	**0.037**
Pfk13	1725266	C	A	A578S	0.036	0.012	0.000	0.005
	1726234	C	T	R255K	0.071	0.027	0.063	0.051
	1726239	C	A	M253I	0.000	0.003	0.000	0.010
	1726349	T	G	N217H	0.000	0.007	0.000	0.008

	1726431	T	A	K189N	0.000	0.012	0.343	0.011
	1726454	A	T	S182T	0.036	0.017	0.000	0.004
	1726592	G	T	H136N	0.000	0.005	0.031	0.006
	1726652	A	T	L116I	0.000	0.003	0.000	0.010
	1726663	C	T	G112E	0.000	0.003	0.125	0.006
	1726676	T	C	K108E	0.000	0.002	0.000	0.005
	1726711	G	T	P96Q	0.036	0.003	0.000	0.005
	1726933	C	T	R22K	0.000	0.002	0.000	0.005
Pfap2mu	718250	G	C	G99A	0.000	0.003	0.000	-0.010
	718333	G	C	V127L	0.000	0.002	0.111	-0.009
	718391	G	A	R146K	0.000	0.033	0.388	0.025
	718433	G	A, C	**S160N/T**	**0.178, 0.000**	**0.170, 0.004**	**0.380, 0.000**	**0.005, 0.001**
	718550	A	C	K199T	0.000	0.072	0.125	0.018
	718969	C	G	R339G	0.000	0.003	0.000	0.010
	719007	A	T	K351N	0.000	0.002	0.000	0.005
	719265	C	A	F437L	0.000	0.019	0.000	0.051
	719288	A	C	N445T	0.000	0.002	0.000	0.005
	719380	T	G	S476A	0.000	0.008	0.000	0.036

MAF = minor allele frequency; Ref = Reference allele; Alt= Alternative allele; *Lake Victoria (LV) islands (e.g., Mfangano and Ngodhe), East Africa (e.g., Kilifi, KE; Muleba, TZ; Nachingwea, TZ; Apac, UG), West Africa (e.g., The Gambia and Mauritania)

Supplementary Table S5. Genes of interest with SNPs showing selection pressure in the population (iHS value ($-\log_{10}[1 - 2 \mid \Phi_{iHS} - 0.5 \mid]) > 4$).

Population	Gene ID	Gene Function
Mfangano	*PF3D7_0809600*	Peptidase family C50; invasion of host cells
	PF3D7_1133400	Apical membrane antigen 1 (PfAMA1); immune evasion against inhibitory antibodies
	PF3D7_1035400	merozoite surface protein 3 (MSP3); generate host antibody response
Lake Victoria Islands*	*PF3D7_0809600*	Peptidase family C50
	PF3D7_0808200	Plasmepsin X (PMX); parasite egress and invasion
	PF3D7_1035400	MSP3
	PF3D7_1133400	PfAMA1
Lake Victoria Mainland*	*PF3D7_0808200*	PMX
	PF3D7_1035400	MSP3
	PF3D7_1301700	CX3CL1-binding protein 2 (CBP2)
	PF3D7_1337800	calcium-dependent protein kinase 5 (CDPK5); regulates parasite egress from erythrocytes
East Africa*	*PF3D7_0104300*	ubiquitin carboxyl-terminal hydrolase 1 (UBP1)
	PF3D7_0103900	parasite-infected erythrocyte surface protein (PIESP15)
	PF3D7_1035400	MSP3
	PF3D7_1301700	CBP2
Central Africa*	*PF3D7_0104300*	UBP1
	PF3D7_0103900	PIESP15
	PF3D7_1035400	MSP3
West Africa*	*PF3D7_0709700*	prodrug activation and resistance esterase (PARE); pepstatin resistance
	PF3D7_0709000	chloroquine resistance transporter (CRT)
	PF3D7_1035400	MSP3
	PF3D7_1301700	CBP2

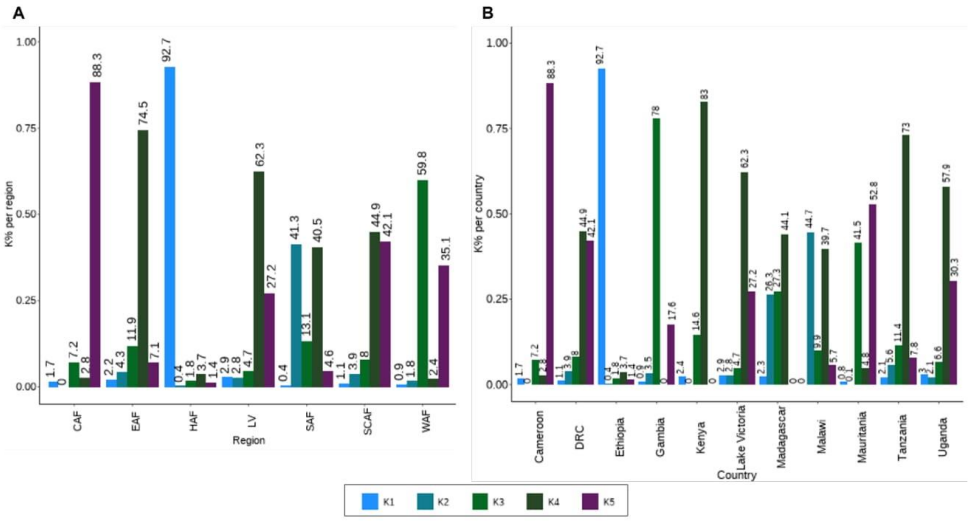
*Lake Victoria islands (e.g., Mfangano and Ngodhe), Lake Victoria mainland (e.g., Suba, Kisumu, and Kombewa, KE), East Africa (e.g., Kilifi, KE; Muleba, TZ; Nachingwea, TZ; Apac, UG), Central Africa (e.g., Cameroon), West Africa (e.g., The Gambia and Mauritania)

Supplementary Table S6. Cross-population analysis of selection pressure on genes of interest within the Lake Victoria (LV) region (XP-EHH (-$\log_{10}$[*p*-value]) > 5).

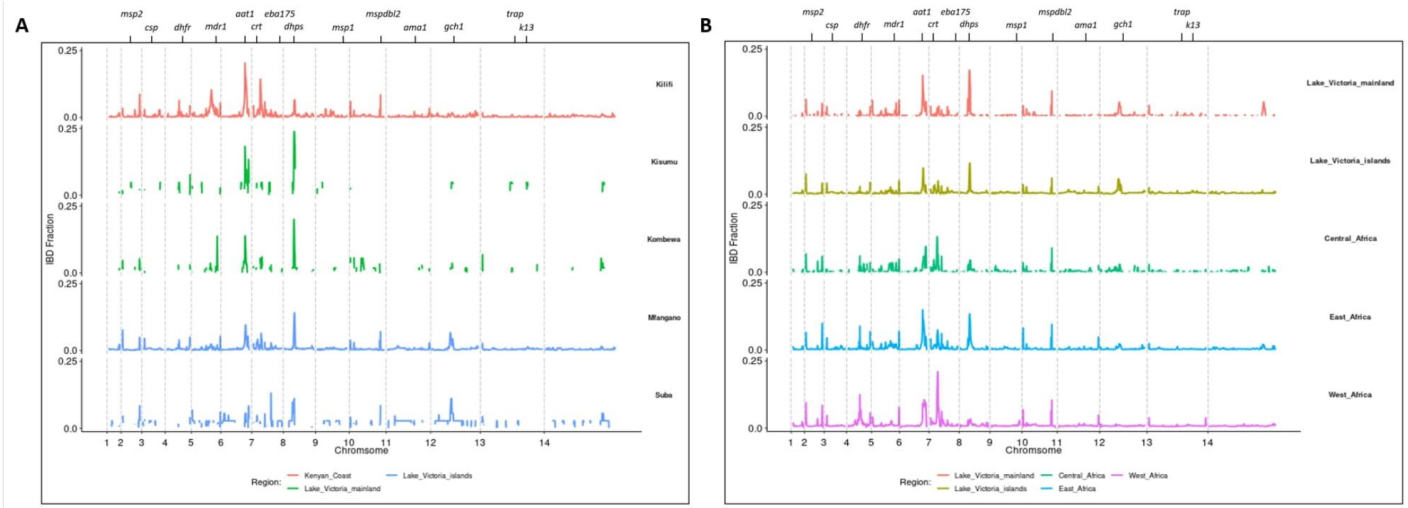
Population*	Gene ID	Gene Function
LV islands vs LV mainland	*PF3D7_0808200*	PMX
	PF3D7_1036300	Duffy binding-like merozoite surface protein 2 (DBLMSP2); A conserved multi-gene family associated with inducing cross-reactive antibodies against *P. falciparum*
LV islands vs East Africa	*PF3D7_1036300*	DBLMSP2
	PF3D7_1216600	Cell traversal protein for ookinetes and sporozoites (CelTOS); a conserved antigen with protective potential
	PF3D7_0808200	PMX
LV islands vs West Africa	*PF3D7_0808200*	PMX
LV mainland vs West Africa	*PF3D7_0709700*	PARE
	PF3D7_0709000	chloroquine resistance transporter (CRT)
	PF3D7_0810800	hydroxymethyldihydropterin pyrophosphokinase-dihydropteroate synthase (DHPS);
	PF3D7_0811300	CCR4-associated factor 1 (CAF1); egress and invasion protein
East Africa vs LV islands	*PF3D7_0808200*	PMX
	PF3D7_1035400	MSP3
East Africa vs LV mainland	*PF3D7_0810800*	DHPS
	PF3D7_0811300	CCR4-associated factor 1 (CAF1)
	PF3D7_1301700	CBP2
Central Africa vs LV islands	*PF3D7_0808200*	PMX
Central Africa vs LV mainland	*PF3D7_0810800*	DHPS
	PF3D7_1035400	MSP3
	PF3D7_1224000	GTP cyclohydrolase 1 (GCH1); antifolate susceptibility

*Lake Victoria islands (e.g., Mfangano and Ngodhe), Lake Victoria mainland (e.g., Suba, Kisumu, and Kombewa, KE), East Africa (e.g., Kilifi, KE; Muleba, TZ; Nachingwea, TZ; Apac, UG), Central Africa (e.g., Cameroon), West Africa (e.g., The Gambia and Mauritania)

Supplementary Figure S1. Cumulative genome-wide admixture ancestry proportions for *P. falciparum* populations across the African continent. (**A**) Cumulative percentages of ancestry per region where K is estimated to be 5. (**B**) Cumulative percentages of ancestry per country where K is estimated to be 5.

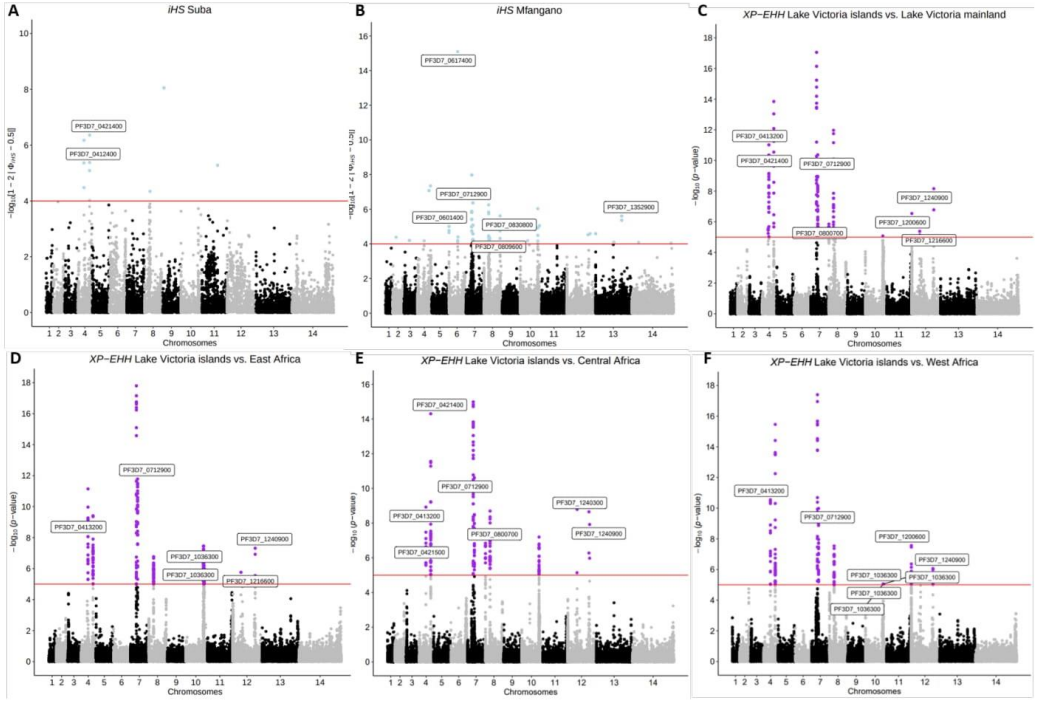


Supplementary Figure S2. Identity by decent (IBD) fractions along each chromosome in *P. falciparum* isolates from (A) Kenya and (B) regional populations across the African continent (e.g. Lake Victoria mainland*, Lake Victoria islands*, Central Africa*, East Africa*, and West Africa*.

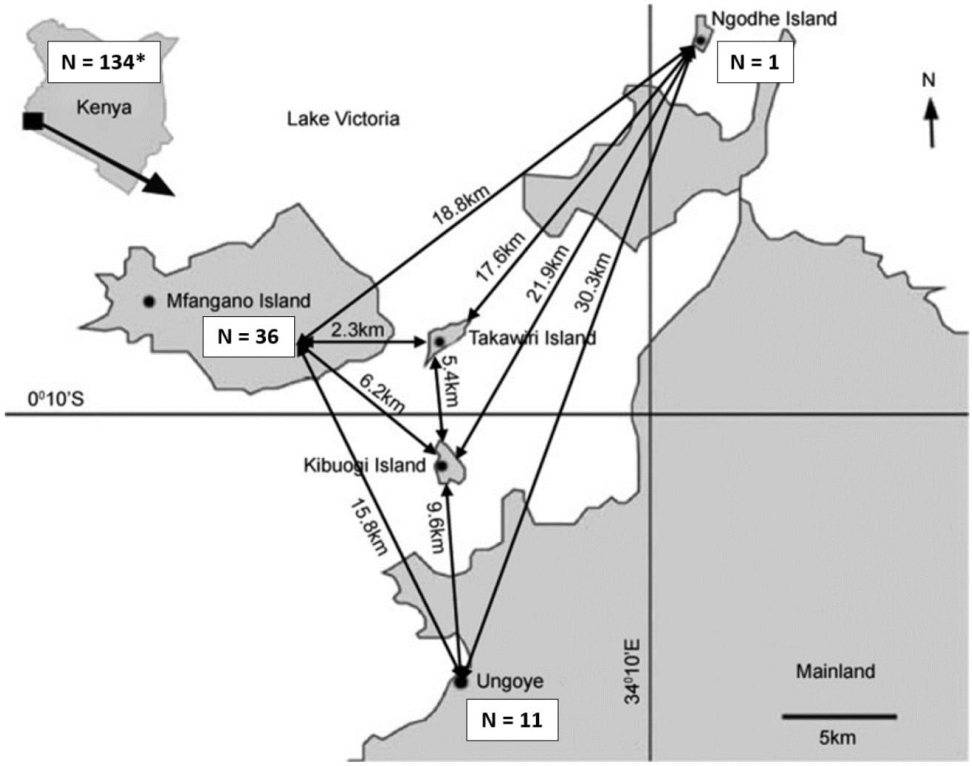


*Lake Victoria islands (e.g., Mfangano and Ngodhe), Lake Victoria mainland (e.g., Suba, Kisumu, and Kombewa, KE), East Africa (e.g., Kilifi, KE; Muleba, TZ; Nachingwea, TZ; Apac, UG), Central Africa (e.g., Cameroon), West Africa (e.g., The Gambia and Mauritania)

Supplementary Figure S3. Signatures of positive selection in Lake Victoria isolates and East African populations. Analysis of haplotype structure to determine genomic regions responding to natural or artificial selection. (**A**) SNPs under selective pressure in Mfangano island isolates with an integrated haplotype score (iHS) >4.0 ($(-\log_{10}[1 - 2 \mid \Phi_{iHS} - 0.5 \mid])$ > 4.0). (**B**) SNPs in Suba District isolates with an iHS value > 4.0. (**C**) SNPs in Lake Victoria isolates with an iHS value > 4.0. (**D**) Cross-population selective pressures identified by comparing SNPs in Mfangano isolates with isolates from Suba District; significant SNPs determined by an *XP-EHH* value > 5.0 ((-log10[p-value]) > 5.0). (**E**) SNPs with an *XP-EHH* value > 5.0 comparing isolates from Lake Victoria with the Lake Victoria mainland. (**F**) SNPs with an *XP-EHH* value > 5.0 comparing Lake Victoria islands with East Africa.



Supplementary Figure S4. Sampling sites and the corresponding number of isolates within the Kenyan region of Lake Victoria. *Total number of Kenyan isolates includes publicly available data from the Pf3K dataset.



Chapter 3: Drug resistance profiling of asymptomatic and low-density Plasmodium falciparum malaria infections on Ngodhe island, Kenya, using custom dual-indexing next-generation sequencing

RESEARCH PAPER COVER SHEET

Please note that a cover sheet must be completed <u>for each</u> research paper included within a thesis.

SECTION A – Student Details

Nagasaki Student No	59719003	**Title**	Miss
LSHTM Student ID No	Lsh1807687		
First Name(s)	Ashley Alexandra		
Surname/Family Name	Osborne		
Thesis Title	A multifaceted investigation of the genomics of malaria, from parasite to host, using next-generation sequencing technologies		
Nagasaki Supervisor(s)	Akira Kaneko, Kiyoshi Kita		
LSHTM Supervisor(s)	Taane Clark, Susana Campino		

If the Research Paper has previously been published please complete Section B, if not please move to Section C.

<u>SECTION B – Paper already published</u>

Where was the work published?			
When was the work published?			
If the work was published prior to registration for your research degree, give a brief rationale for its inclusion			
Have you retained the copyright for the work?*	Choose an item.	Was the work subject to academic peer review?	Choose an item.

*If yes, please attach evidence of retention. If no, or if the work is being included in its published format, please attach evidence of permission from the copyright holder (publisher or other author) to include this work.

SECTION C – Prepared for publication, but not yet published

Where is the work intended to be published?	Scientific Reports
Please list the paper's authors in the intended authorship order:	Ashley Osborne, Jody E. Phelan, Akira Kaneko, Wataru Kagaya, Chim Chan, Mtakai Ngara, James Kongere, Kiyoshi Kita, Jesse Gitaka, Susana Campino & Taane G. Clark
Stage of publication	**Undergoing revision**

SECTION D – Multi-authored work

For multi-authored work, give full details of your role in the research included in the paper and in the preparation of the paper. (Attach a further sheet if necessary)	I carried out laboratory work including primer and multiplex optimisation, as well as prepared samples for WGS, including selective whole genome amplification, DNA clean-up, and shipment of samples. I performed bioinformatic analyses and interpreted the results under the supervision of my supervisors. I wrote and prepared the first draft of the manuscript that was circulated to my supervisors and co-authors

SECTION E – Names and affiliations of co-author(s)

Please list all the co-authors' names and their affiliations.

Ashley Osborne - Faculty of Infectious and Tropical Diseases, London School of Hygiene and Tropical Medicine & School of Tropical Medicine and Global Health, Nagasaki University Jody E. Phelan - Faculty of Infectious and Tropical Diseases, London School of Hygiene and Tropical Medicine Akira Kaneko - Department of Parasitology, Osaka Metropolitan University Wataru Kagaya - Department of Parasitology, Osaka Metropolitan University Chim Chan - Department of Parasitology, Osaka Metropolitan University Mtakai Ngara - Department of Microbiology, Karolinska Institutet James Kongere - Department of Parasitology, Osaka Metropolitan University Kiyoshi Kita - School of Tropical Medicine and Global Health, Nagasaki University Jesse Gitaka - Directorate of Research and Innovation, Mount Kenya University Susana Campino - Faculty of Infectious and Tropical Diseases, London School of Hygiene and Tropical Medicine Taane G. Clark - Faculty of Epidemiology and Population Health, London School of Hygiene and Tropical Medicine & Faculty of Infectious and Tropical Diseases, London School of Hygiene and Tropical Medicine

SECTION F

I confirm that all co-authors have agreed that the above paper will be included in my PhD thesis.

Student Signature	[redacted]
Date	06/07/23

LSHTM Supervisor Signature	[redacted]
Date	06/07/23

Nagasaki University Supervisor Signature	[redacted]
Date	06/07/23

Drug resistance profiling of asymptomatic and low-density *Plasmodium falciparum* malaria infections on Ngodhe island, Kenya, using custom dual-indexing next-generation sequencing

Short title: Drug resistance profiling of low-density *Plasmodium falciparum* infections on Ngodhe island, Kenya

Ashley Osborne [1,2] ashley.osborne@lshtm.ac.uk

Jody E. Phelan [1] jody.phelan@lshtm.ac.uk

Akira Kaneko [3,4] akirakaneko555@gmail.com

Wataru Kagaya [3] wataru.kagaya@gmail.com

Chim Chan [3] aramidus44@gmail.com

Mtakai Ngara [4] ngaravald@gmail.com

James Kongere [2,3,5] jkongere@gmail.com

Kiyoshi Kita [2] kitak@kita-kiyoshi.net

Jesse Gitaka [6,7,*] jgitaka@mku.ac.ke

Susana Campino [1,*] susana.campino@lshtm.ac.uk

Taane G. Clark [1,8,*,**] taane.clark@lshtm.ac.uk

[1] Faculty of Infectious & Tropical Diseases, London School of Hygiene & Tropical Medicine, London, UK

[2] School of Tropical Medicine and Global Health, Nagasaki University, Nagasaki, Japan

[3] Department of Parasitology, Graduate School of Medicine, Osaka Metropolitan Univ., Osaka, Japan

[4] Department of Microbiology, Tumor and Cell Biology, Karolinska Institutet, Stockholm, Sweden

[5] Centre for Research in Tropical Medicine and Community Development (CRTMCD), Hospital Road next to Kenyatta National Hospital, Nairobi, Kenya

[6] Directorate of Research and Innovation, Mount Kenya University, Thika, Kenya

[7] Centre for Malaria Elimination, Mount Kenya University, Thika, Kenya

[8] Faculty of Epidemiology and Population Health, London School of Hygiene & Tropical Medicine, London, UK

* Joint authors

** Corresponding author:

Professor Taane G. Clark

taane.clark@lshtm.ac.uk

ABSTRACT

Malaria control initiatives require rapid and reliable methods for the detection and monitoring of molecular markers associated with antimalarial drug resistance in *Plasmodium falciparum* parasites. Ngodhe island, Kenya, presents a unique malaria profile, with lower *P. falciparum* incidence rates than the surrounding region, and a high proportion of sub-microscopic and low-density infections. Here, using custom dual-indexing and Illumina next generation sequencing, we generate resistance profiles on seventy asymptomatic and low-density *P. falciparum infections* from a mass drug administration program implemented on Ngodhe island between 2015 and 2016. Our assay encompasses established molecular markers on the *Pfcrt, Pfmdr1, Pfdhps, Pfdhfr,* and *Pfk13* genes. Resistance markers for sulfadoxine-pyrimethamine were identified at high frequencies, including a quintuple mutant haplotype (*Pfdhfr/Pfdhps*: N51I,C59R,S108N/A437G,K540E) identified in 62.2% of isolates. The *Pfdhps* K540E biomarker, used to inform decision making for intermittent preventative treatment in pregnancy, was identified in 79.2% of isolates. Several variants on *Pfmdr1*, associated with reduced susceptibility to quinolones and lumefantrine, were also identified (Y184F 47.1%; D1246Y 16.0%; N86 98%). Overall, we have presented a low-cost and extendable approach that can provide timely genetic profiles to inform clinical and surveillance activities, especially in settings with abundant low-density infections, seeking malaria elimination.

Word Count: 196/200

Introduction

Despite rapid progress in disease control and elimination efforts between the years 2000 and 2015, malaria continues to be a major public health burden, particularly in low- and middle-income countries. Caused by parasite species within the *Plasmodium* genus, malaria was responsible for 247 million cases worldwide and contributed to 619,000 deaths in 2021 alone [1]. In Sub-Saharan Africa, *P. falciparum* malaria carries the heaviest burden, accounting for an estimated 95% of all cases and 96% of all deaths, with about 80% of mortality occurring among children less than 5 years of age. Progress towards reduction in disease incidence, and ultimately elimination, has been hampered by the emergence and spread of antimalarial resistance mechanisms in *Plasmodium* parasites, insecticide resistance in mosquito vectors, and ecological threats associated with global climate change. These setbacks have been further exacerbated by malaria control programme and supply chain interruptions caused by the COVID-19 pandemic, as well as shortcomings in global funding for malaria elimination in the wake of the pandemic [2, 3].

Malaria incidence rates and transmission risk have declined across much of Kenya, with national malaria prevalence estimated to be below 10%, although true numbers of cases remain unknown as asymptomatic individuals are far less likely to seek out diagnosis and treatment [4]. Despite an overall reduction in incidence, lower elevation regions in Kenya, including those along the Indian Ocean coast and Lake Victoria, still experience a high malaria burden [5]. Lake Victoria remains a region of intense transmission due to various environmental and geopolitical factors that have made targeted vector control and tailored malaria control programmes difficult to implement region-wide [6–8]. During peak transmission periods, determined by the two rainy seasons in March to May and October to December, *P. falciparum* prevalence in the Lake Victoria basin can reach up to 40% in individuals aged between 2 and 10 years [5, 7, 9]. Overall prevalence, however, can vary drastically amongst the islands and mainland populations [7]. In Homa Bay County, Suba District (population size: North 124,938; South 122,383) and the large island of Mfangano (population size 26,000) generally have the highest

parasite prevalence rates, sometimes even exceeding 40%, while the smaller islands, including Ngodhe island (population size 600-1,000), are generally associated with lower prevalence rates and asymptomatic, sub-microscopic, infections [5, 7, 10].

In 2015, the World Health Organization (WHO) recommended the implementation of mass drug administration (MDA) in low transmission regions, including inhabited islands [11]. Despite its close geographical proximity to regions of intense malaria transmission, Ngodhe island maintains overall low levels of transmission, as well as high proportions of asymptomatic and sub-microscopic infections. This unique transmission profile on Ngodhe island led to the rationalisation of using an MDA strategy in 2015 and 2016 to assess the efficacy of an MDA in reducing malaria on islands with high and heterogeneous transmission [7, 12]. A high compliance MDA with artemisinin-based combination therapies (ACTs) and a single low dose of primaquine, with approximately 90% participation, led to an initial decrease in parasite prevalence from 3% to 0% by microscopy and 10% to 2% by PCR. Despite this initial drastic decrease, prevalence rebounded to baseline levels within six months, and in 2017 was recorded to be 4.6% and 16.0%, respectively [7]. The island's proximity to regions with high malaria transmission and an inability to prevent parasite importation likely prevented sustainable elimination conditions from being achieved [7].

Difficulties in malaria control, and the recent emergence of parasites with reduced susceptibility to artemisinin, a component of ACTs, in Uganda and Rwanda, have highlighted the need for rapid and reliable methods for detecting and monitoring molecular markers associated with drug resistance in *P. falciparum* [13, 14]. ACTs are the first-line treatment for uncomplicated malaria and last-line of defence following the widespread global emergence of chloroquine and sulfadoxine-pyrimethamine (SP) resistance amongst parasite populations [15–18]. Despite widespread resistance, SP is still used as an intermittent preventive treatment during pregnancy (IPTp), in accordance with WHO guidelines [2]. Although chloroquine has been discontinued, molecular markers for chloroquine-resistance

persist in parasite populations across Africa [17]. There is recent evidence of chloroquine-sensitivity returning in countries that adopted more rapid changes in malaria treatment policies, suggesting chloroquine could be reintroduced in the future [19–23]. There are well categorised and studied biomarkers for antimalarial drug resistance, namely single nucleotide polymorphisms (SNPs), on *P. falciparum* genes *Pfcrt, Pfmdr1*, *Pfdhfr*, and *Pfdhps* that confer drug resistance to chloroquine, SP, and ACT partner drugs, as well as SNPs on *PfK13* associated with delayed parasite clearance to ACTs [16, 24–27].

Advancements in low-cost sequencing-based approaches that target molecular markers for drug resistance (e.g., amplicon sequencing) offer a new high throughput method of parasite surveillance, although have not been utilised to look specifically at low-density or asymptomatic infections [28–30]. Due to the high proportion of sub-microscopic and asymptomatic *P. falciparum* infections on Ngodhe island, no genetic information is available for the island's parasite population, including drug resistance frequencies [7]. Here we demonstrate the use of custom dual-indexing and Illumina next generation sequencing-technology, to generate a resistance profile of the *P. falciparum* parasites on Ngodhe island from predominantly asymptomatic and low-density, infections collected during MDA activities between 2015 and 2016. Our analysis was able to quantify molecular markers of resistance on the *Pfcrt*, *Pfmdr1*, *Pfdhps*, *Pfdhfr*, and *Pfk13* genes, establishing this method as a viable means of malaria surveillance suitable for regions with typical sub-microscopic infections.

Results

Illumina amplicon sequencing and coverage

In years 2015 and 2016, a high compliance (>90% participation) MDA was carried out on Ngodhe island. Of the 201 PCR-positive dried blood spot (DBS) samples collected during the 2015/2016 MDA campaign on the island, 102 (53.7%) samples were deemed suitable for sequencing based on low DNA concentration measurements (Qubit HS DNA concentration >0.4 ng/ul). From the 102 *P. falciparum*

samples, 70 (68.7%) were successfully sequenced on an Illumina platform to cover a total of nine 500 base pair amplicons that encompassed five genes (*Pfk13, Pfcrt, Pfmdr1, Pfdhps*, and *Pfdhfr*) (**Table S1; Figure 1**). The 70 sequenced samples were sourced from 49 (70.0%) sub-microscopic infections, 12 (17.1%) low-parasitaemia (<1000 parasites/µl blood) and 9 (12.8%) moderate/high-parasitaemia infections (>1000 parasites/µl blood). Rapid diagnostic test (RDT) data was available for 24 of the sequenced samples, 5 with negative and 19 positive RDT results. Of the 5 samples that were negative by RDT but positive by PCR, 4 were also negative by microscopy. Individuals were classified as asymptomatic with an absence of fever or other acute symptoms [4].

The median depth of read coverage achieved via amplicon sequencing overall was high (range of medians: 69- to 1760-fold; **Table S1**), with *Pfk13* (1760-fold) having noticeably higher average coverage than *Pfcrt, Pfmdr1, Pfdhps*, and *Pfdhfr* (69- to 155-fold) (**Table S1; Figure 2**). In general, regional coverage of 20-fold or higher is considered adequate for genetic studies. Low coverage of codon 581 on *Pfdhps* was linked to a drop in coverage in the middle of the amplicon shared with codons 540 and 613. High coverage of codon 553 on *Pfk13* (median 9,493-fold) was due to primer overlap from adjacent amplicons.

SNP frequencies on drug resistance-associated genes using Illumina sequencing

Across the 70 samples sequenced using the Illumina platform, only SNPs with a minimum read depth >10-fold on *Pfcrt*, *Pfmdr1*, *Pfk13, Pfdhps*, and *Pfdhfr* were included in this analysis to ensure accuracy in SNP reporting. For context and comparison, variant frequencies across these genes were compared to previously categorised frequencies from parasites on Mfangano island, Lake Victoria, as well as East African isolates available via the Pf3k database (**Table 1**) [10].

There were 3 nonsynonymous mutations observed on the *Pfcrt* gene, resulting in amino acid changes M74I, N75E, and K76T. These variants are well documented to be associated with resistance to chloroquine. The *Pfcrt* N74I polymorphism occurred in 3 out of 46 samples while the N75E and K76T polymorphisms were observed in 2 out of 45 samples. The *Pfcrt* K76T biomarker, the primary marker for chloroquine resistance, was observed in 4.4% of the Ngodhe samples, compared to 7.4% of isolates from Mfangano island in 2020 and 17.5% from wider East African populations [10]. The *Pfcrt* C72S and V73V variants were not observed. Analysis of haplotype frequencies on *Pfcrt* for samples with a read depth >10-fold at every position (codons 72, 73, 74, 75, and 76) identified the distributions of mutants and wild type parasites in the Ngodhe parasite population (*n* = 45) (**Table 2**). The wild-type parasites, CVMNK, made up 95.5% of the population while the triple mutant, CV**IET** (mutations underlined), accounted for 4.5% of the population.

On the *Pfmdr1* gene, 7 nonsynonymous SNPs were identified, 3 of which are reported to cause amino acid changes resulting in altered susceptibility to chloroquine and the ACT partner drug lumefantrine (N86Y, Y184F and D1246Y). The *Pfmdr1* N86Y variant, where the wild-type (WT) allele is believed to be selected for by artemether–lumefantrine usage, was identified in 1 out of 50 samples (2.0%), appearing at similar frequencies to Mfangano island isolates (0%) [10]. The D1246Y mutation was observed in 5 out of 31 samples (16.0%), appearing at a higher frequency compared to Mfangano isolates (2.9%). Frequencies in East African isolates were higher for *Pfmdr1* N86Y, 15.7%, while the N1246Y biomarker appeared at a similar frequency to Ngodhe isolates (12.9%). The *Pfmdr1* Y184F biomarker was observed in 47.1% of samples (24 out of 51 isolates). This frequency was slightly higher than frequencies observed in Mfangano island (31.7%), and East African isolates (39.0%) [10]. The *Pfmdr1* S1034C and N1042D variants were not observed in any of the 70 samples with read depth more than 10-fold at this position.

Haplotype analysis of *Pfmdr1* identified the frequencies of single and double mutants within the Ngodhe parasite populations based on codon positions 86, 184, 938, 1034,1042, and 1246 (*n* = 31) (**Table 2**). Wild-type parasites, *Pfmdr1* NYFSND, made up the largest fraction of the population, accounting for 45.2% of isolates, while the single mutant, N**F**FSND, made up 38.7%. The single mutant, *Pfmdr1* NYFSN**Y**, accounted for 9.7% of the population while the double mutant, N**F**FSN**Y**, made up the remaining isolates at 6.4%. No drug resistance-associated polymorphisms were recorded on *Pfk13.* The *Pfk13* variant A578S was recorded but is not thought to confer reduced susceptibility to artemisinin or any other antimalarial drugs.

There were 4 nonsynonymous SNPs identified on *Pfdhfr*, all of which resulting in amino acid changes documented to confer reduced susceptibility to pyrimethamine (N51I, C59R, S108N, and I164L). N51I, C59R, AND S108N changes were seen at similar frequencies amongst the Ngodhe isolates, with prevalence observed at 77.0%, 76.9%, and 80.8%, respectively (**Table 1**)**.** Isolates from Mfangano island and East Africa populations were observed to have all 3 biomarkers at higher frequencies (>87%) than Ngodhe island [10]. The I164L variant was observed in 5.2% of Ngodhe isolates, similar to frequencies to Mfangano isolates (6.9%), and higher than isolates from East Africa (1.5%) and West Africa (0%).

On *Pfdhps*, 3 nonsynonymous SNPs were observed, with 2 known to result in amino acid changes associated with resistance to sulphadoxine, A437G and K540E. The K540E mutant was observed in 79.2% of isolates and is generally used as a proxy measure for the presence of all 5 key mutations for resistance to SP (N51I, C59R, S108N, I164L, A437G, and K540E). The A437G mutation was observed to be fixed at 100% in samples from Ngodhe and Mfangano islands, compared to 84.4% in samples from wider East African populations.

Analysis of the Ngodhe parasite *Pfdhfr/Pfdhps* haplotype frequencies, for samples with a read depth more than 10-fold at every nucleotide position, included *Pfdhfr* codons 51, 59, 108, and 164, as well as *Pfdhps* codons 437, 540, 581 (*n* = 37) (**Table 3**)**.** The quintuple mutant haplotype, **IRN**I**GE**A, was observed in a majority of screened Ngodhe isolates, accounting for 62.2% of the population, as well as a sextuple mutant haplotype, **IRNLGE**A, which was identified in 5.4% of isolates. The single mutant haplotype, NCSI**G**KA, was observed at the second highest frequency, 13.5%, while the wild-type haplotype, NCSIAKA, was not observed in any isolates. Double, triple, and quadruple mutants accounted for the remainder of the population, occurring at frequencies ranging from 2.7% to 5.4%.

Discussion

Advancements in targeted low-cost sequencing approaches provide a viable means for monitoring the emergence and spread of characterised drug resistance-associated polymorphisms in both asymptomatic and low-density *P. falciparum* infections [28, 31]. Asymptomatic infections tend to be under-represented in large-scale genetic analyses, despite accounting for most infections worldwide [32]. This is often due to the difficulties associated with extracting sufficient DNA needed to perform genomic characterisation, such as whole genome sequencing. Additionally, asymptomatic individuals do not tend to seek treatment and go undiagnosed. As countries make progress towards their malaria elimination goals, sub-microscopic and asymptomatic cases will remain the main reservoirs for disease [1]. To ensure accurate and informed disease control policies, molecular surveillance methods need to be sensitive enough to detect regions of interest, such as drug resistance biomarkers, in these types of infections.

To assist in the generation of higher resolution drug resistance profiles of malaria parasites, we demonstrate the use of custom dual-indexing amplicon sequencing technology to identify molecular markers of resistance on *Pfcrt*, *Pfmdr1*, *Pfdhps*, *Pfdhfr*, and *Pfk13* genes from asymptomatic and low-density *P. falciparum* infections on Ngodhe island in Lake Victoria, Kenya. Ngodhe island has a unique

transmission profile compared to surrounding inhabited islands and mainland communities located within western Kenya and Lake Victoria. In a region of intense transmission, asymptomatic and sub-microscopic infections make up most cases detected on Ngodhe island, which also has an overall lower *P. falciparum* incidence [7]. Given these distinctive characteristics, assessing resistance biomarkers within Ngodhe island's *P. falciparum* population for the first time provided an invaluable baseline for drug resistance monitoring, as well as provided data that can be utilised to better inform policy making decisions concerning malaria control programmes within the region.

To preserve the efficacy of antimalarial drugs, monotherapy treatments (e.g., chloroquine) have been phased out for the treatment of uncomplicated malaria and replaced by ACTs in much of world, including Kenya [5, 33]. Despite its discontinued use, chloroquine can still occasionally be found at local pharmacies in parts of Africa as a general treatment for fevers [1]. Monitoring chloroquine resistance mutations in parasite populations can provide insight into the presence of any ongoing drug selection pressure, which could be due to ongoing sales of chloroquine, as well as the return of chloroquine sensitivity, as seen in Malawi, following the complete removal of chloroquine from circulation [23]. The K76T mutation on *Pfcrt*, often used as the main molecular marker for chloroquine resistance, was observed to be at low frequencies in Ngodhe island parasites, with the wild-type K76 allele identified in most of the isolates screened through this study. The same was found to be true for other resistance conferring mutations on *Pfcrt*, with only wild-type alleles observed at codons 72 and 73, and high frequencies of wild-type alleles at codons 73 and 74. These results support the hypothesis that removal of chloroquine drug selection pressure may promote the return of chloroquine sensitive wild-type parasites [19–22]. Additionally, it has been documented that treatment with artemether-lumefantrine, extensively used in East Africa, selects for the chloroquine-susceptible *Pfcrt* K76 allele, suggesting that the K76T mutation may be a drug-specific contributor to enhanced *P. falciparum* susceptibility to lumefantrine [34, 35].

In accordance with WHO guidelines, SP continues to be used as IPTp in pregnant women but is not used as a first-line treatment for uncomplicated malaria [2]. Nonsynonymous polymorphisms on *Pfdhfr* and *Pfdhps* are associated with resistance to pyrimethamine and sulphadoxine, respectively. The degree of resistance to SP increases with each subsequent mutation on these genes, which has led to the emergence of quadruple, quintuple, and, more recently, sextuple mutations [25, 36]. Resistance markers for SP were identified at high frequencies in Ngodhe island isolates, likely due to the continued drug selection pressure within the parasite population. The *Pfdhfr/Pfdhps* quintuple mutant haplotype, encompassing the N51I, C59R, S108N, A437G and K540E mutations, was observed to be the most prominent haplotype, accounting for 62.2% of the screened population. Additionally, a sextuple mutant haplotype, including *Pfdhfr/Pfdhps* variants N51I, C59R, S108N, I164L, A437G and K540E, was identified in two isolates. Given the high frequencies of these SP resistance-associated biomarkers and continued circulation of SP in local parasite populations, there exists the possibility of relatively low fitness costs associated with maintaining these mutations.

The *Pfdhps* K540E variant is used as a proxy measure for the presence of all 5 key mutations associated with SP resistance on *Pfdhfr* and *Pfdhps*. However, following haplotype analysis, the K540E mutation was also observed in double and quadruple mutants at frequencies ranging from 2.7% to 5.4% [37, 38]. In addition to being a proxy measure for SP resistance, the K540E mutation is also used to inform decision making by the WHO surrounding IPTp guidance, with IPTp implementation only recommended in regions where K540E prevalence is <50% [39]. The K540E marker was identified in 79.2% of the Ngodhe island isolates, suggesting there may be reduced efficacy of IPTp-SP therapies within the region. The high frequency of this mutation highlights the need for continued monitoring of malaria infections in pregnant women to prevent adverse malaria-related outcomes, as well as the need for more safe and effective malaria drugs for use in pregnant women [40].

There were no mutations observed on *Pfk13* believed to confer a reduced susceptibility to artemisinin, however there were a handful of variants identified on *Pfmdr1* associated with reduced susceptibility to the ACT partner drug lumefantrine. To prevent artemisinin resistance, ACTs combine artemisinin, a fast acting and highly effective antimalarial drug, with a partner drug that has a longer half-life to clear any remaining parasites [41, 42]. Resistance to ACT partner drugs poses a threat to maintaining the efficacy of these combination therapies and the protection they provide against the emergence of artemisinin resistance. Artemether-lumefantrine is the ACT extensively used across much of Africa, including Kenya where it was introduced in 2006 [27]. The *Pfmdr1* Y184F and D1246Y variants, identified in 47.1% and 16% of isolates respectively, are associated with varying degrees of reduced susceptibility to lumefantrine and quinolines. Recent studies have found evidence that the Y184F mutant allele is not the primary determinant of resistance to lumefantrine but that there may be a genetic correlation between Y184F and the acquisition of a drug-resistance phenotype [26]. The N86Y variant, identified in 2% of isolates, is believed to be associated with changes to susceptibility of chloroquine and amodiaquine. The N86 wild-type codon, conversely, may confer a reduced susceptibility to lumefantrine and was identified in 98% of Ngodhe isolates [43].

There were several limitations with this study, especially related to the use of field isolates. The samples were collected as part of an MDA that occurred throughout 2015 and 2016. Due to the nature of malaria parasites and their ability to adapt quickly to selective pressures, drug resistance polymorphism prevalence may differ from this sample set and current day prevalence. However, there was previously no data available for this region and this study aimed to set a baseline for future studies, as well as demonstrate amplicon sequencing as a viable method for screening low-density malaria infections. Additionally, we would anticipate differences over time in the frequencies of molecular markers for drug resistance regardless of the 2015/2016 MDA as *P. falciparum* populations are always adapting to changes in drug pressures within the environment. Drug resistance linked to copy number variation, particularly relevant for *Pfmdr1*, was not addressed by this technique and

leaves room for future work to address and supplement this methodology. However, amplification of *Pfmdr1* is considered rare in East African parasite populations and is not believed to play a significant role in drug resistance in this region [44]. Finally, only 102 (of the 201) PCR-positive isolates were sequenced. The sequenced isolates had the highest DNA concentrations, and it is possible this could have led to some sampling bias and mutations could have been missed.

Progress towards global malaria eradication over the past few years has been hampered by the emergence and spread of resistance to antimalarial drugs and has highlighted the need for rapid and reliable methods to detect and monitor molecular markers of drug resistance. Here we applied amplicon-based approaches aimed at targeting known and putative drug resistance markers in established loci on *Pfcrt*, *Pfmdr1*, *Pfdhps*, *Pfdhfr*, and *Pfk13* to demonstrate the effectiveness of sequencing-based high throughput surveillance in a low resource setting on asymptomatic, low-density, *P. falciparum* infections. The cost of sequencing has been prohibitive to the implementation of sequencing-based surveillance methods in low-resource settings. However, thanks to the development and application of dual-indexing technology, alongside targeted short-read sequencing, amplicon sequencing presents an affordable alternative to whole genome sequencing, as well as offers the potential for cross-platform capability, including both Illumina and Oxford Nanopore Technology sequencing platforms. In addition to more inclusive pricing (currently < USD 0.5 per amplicon), amplicon sequencing is easily expandable to include a wider range of targets across the Plasmodium genome, as well as targets within other species, suggesting the possibility of an integrated host-pathogen-vector surveillance system.

Methods

Study site selection

In 2015 and 2016, a high compliance (>90% participation) MDA was carried out on Ngodhe island in Lake Victoria (population size 600 – 1,000) [7]. Throughout the duration of the MDA, a total of 3,167 dry blood spot (DBS) samples were collected, including: 184 pre-MDA, 458 on day 0, 391 on day 2, 372 on day 7, 459 on day 35, 387 on day 42, 462 on day 120, and 454 on day 180. Samples were collected from all MDA participants, which encompassed >90% of the island's total population, irrespective of clinical symptoms, gender, or age. Of the 3,167 samples, 201 samples were positive by PCR. PCR positive individuals were treated with Artequick® (artemisinin and piperaquine combination) and primaquine. Rapid diagnostic test data was available for samples collected during the follow-up phase of the MDA, which began on day 120. Of 201 PCR-positive DBS samples collected during the 2015/2016 MDA campaign on Ngodhe island, 102 (53.7%) samples were selected for sequencing based on Qubit HS DNA concentration measurements (DNA concentration >0.4 ng/ul).

Malaria species identification and parasite density assessment were carried out using microscopy, following WHO guidelines, at the Nagasaki University research station in Mbita, Kenya by trained microscopists and confirmed at Osaka Metropolitan University and LSHTM Malaria Reference Laboratory using established nested PCR assays [45, 46]. Ngodhe island is occupied by the Luo ethnic group, migrant fishermen and families. The island is less than 1 km^2 in size and accessible only by small boat; located 3km north of Rusinga island and the mainland town of Mbita. The use of long-lasting insecticide treated bed nets (LLINs) on the island increased between 2012 and 2015. This uptake was due to initiatives introduced by the Kenyan Ministry of Health to provide free LLINs for households and pregnant women. This increase in LLINs usage saw a decrease in malaria prevalence on the island during the 2012-2015 time-period. Indoor residual spraying (IRS) was not implemented on Ngodhe island until 2018.

Primer design

Primers used in this study were designed to investigate polymorphisms in *Pfcrt* (codons 72-76), *Pfmdr1* (codons 86, 184, 1034, 1042, and 1246), *pfdhfr* (codons 16, 51, 59, 108, and 164), *pfdhps* (codons 431, 436, 437, 540, 581, and 613), and the propeller domain of *pfK13* by amplifying fragments, or amplicons, of between 500 and 600 base pairs (bp) (**Table S1**). Primers were designed to contain custom 6-nucleotide indices based on previously described and published methodologies by Nag *et al.* [27]. In this study, the *Pfcrt* primers were further optimised from those published by *Nag et al.* for use in Pf3D7, and field isolates, while the *Pfmdr1* primer set was expanded to include the codon at position 184.

DNA extraction and sample preparation

DNA from isolates collected in years 2015 and 2016 was extracted in 2021 from dried-blood spots (DBS) preserved on filter papers. Filter papers were stored at 4°C from 2015/2016 until blood extraction in 2021. A cold chain was not maintained during sample shipment from Kenya to the London School of Hygiene and Tropical Medicine. For DNA extraction, half of a DBS was used. DNA extraction was performed using the Qiagen QiaSymphony Automated Nucleic Acid Extraction Facility and the Qiagen QIAsymphony DSP DNA kit. Following extraction, DNA was amplified using an established selective whole genome amplification (SWGA) primer set and protocols [31, 47].

PCR reactions and programmes

Simplex PCRs, carried out for *Pfcrt*, were performed using a mastermix containing 5 µl of Q5 Reaction Buffer (New England BioLabs), 0.5 µl of dNTPs (1n nM stocks, New England BioLabs), 0.25 µl Q5 Hot Start High-Fidelity DNA Polymerase (New England BioLabs), and 15.75 µl Milli-Q water (Merck). For each reaction, a total of 1.25 µl of forward and 1.25 µl of reverse primer (10 pmol/µl stocks) were used with 1 µl of DNA for a total reaction volume of 25 µl. Multiplex PCRs (combinations described in **Supplementary Table 2**) were performed using a mastermix containing 5 µl of Q5 Reaction Buffer

(New England BioLabs), 0.5 µl of dNTPs (1n nM stocks, New England BioLabs), 0.25 µl Q5 Hot Start High-Fidelity DNA Polymerase (New England BioLabs), and 15.8 µl Milli-Q water (Merck). For each multiplex reaction, 0.6 µl of both forward primers, for a total of 1.2 µl of forward primer, and 0.6 µl of each reverse primer, for a total of 1.2 µl of reverse primer, were used with 1 µl of DNA for a total reaction volume of 25 µl. The reactions were carried out in a thermocycler consisting of the following steps: Heat activation for 15 minutes at 72°C, 30 cycles of denaturation for 20 seconds at 95°C, annealing for 2 minutes at 55°C, elongation for 2 minutes at 72°C, and a final elongation for 10 minutes at 72°C, followed by a hold at 10°C.

Amplicon purification and pooling

Indices for each amplicon target were conserved according to the sample identifier. Samples not containing shared index combinations were pooled (a maximum of 110 amplicons per pool was achieved) for purification (**Figure 1**). Pooled samples were purified prior to sequencing using bead purification (KAPA), according to the manufacturer's instructions, using a ratio of 1:0.70 of product to bead volume to select for 400 to 500 bp segments of DNA. DNA concentration was measured using Qubit dsDNA HS (Invitrogen) and standardised to a final concentration of 20 ng per 25 µl.

Illumina sequencing and bioinformatics

P. falciparum isolates collected in 2015/2016 (n=102) were sequenced using the Illumina MiSeq platform with 300bp paired end kits at Genewiz (GENEWIZ Germany GmbH). The sequencing reads from pooled samples were demultiplexed, divided into separate files based on their unique indices, using an in-house python script to generate individual FASTQ files necessary for downstream analysis. Following demultiplexing, the raw sequencing data was then mapped to the Pf3D7 (*P. falciparum*) reference (v3) genome using *bwa-mem* software (default parameters for Illumina data) [48]. SNPs and short insertions and deletions (indels) were called using the *samtools*, freebayes, and GATK software suites [49–51]. The minimum base quality for gatk and freebayes was changed to 30 rather than their

default parameters of 10 and 0, respectively. SNPs occurring in low quality or low coverage regions were discarded. Mixed call SNPs were assigned genotypes determined by a ratio of coverage in which nucleotide calls were 80% or higher. SNPs were annotated using *bcftools* consequence calling, which predicts functional variant consequences [49, 52].

Ethical Approval and Consent to Participate

All experimental protocols and research were performed in accordance with relevant guidelines and regulations. Research was approved by the Mount Kenya University Independent Ethics Research Committee (MKU-IERC) (Approval reference: P609/10/2014) and the Ethics Committee at Osaka Metropolitan University (Approval number: 3206). Workshops and sensitisation meetings were carried out with communities to attain community consent to study participation. Written informed consent was obtained from all study participants whose parasite DNA was used in this study.

References

1. World Health Organization (2021) World malaria report 2021. World Health Organization, Geneva

2. World Health Organization (2020) World malaria report 2020: 20 years of global progress and challenges. World Health Organization, Geneva

3. World Health Organization (2019) World Malaria Report 2019. World Health Organization, S.l.

4. Chen I, Clarke SE, Gosling R, Hamainza B, Killeen G, Magill A, O'Meara W, Price RN, Riley EM (2016) "Asymptomatic" Malaria: A Chronic and Debilitating Infection That Should Be Treated. PLOS Med 13:e1001942

5. Githinji S, Noor AM, Malinga J, Macharia PM, Kiptui R, Omar A, Njagi K, Waqo E, Snow RW (2016) A national health facility survey of malaria infection among febrile patients in Kenya, 2014. Malar J 15:591

6. Okara RM, Sinka ME, Minakawa N, Mbogo CM, Hay SI, Snow RW (2010) Distribution of the main malaria vectors in Kenya. Malar J 9:69

7. Kagaya W, Gitaka J, Chan CW, Kongere J, Idris ZM, Deng C, Kaneko A (2019) Malaria resurgence after significant reduction by mass drug administration on Ngodhe Island, Kenya. Sci Rep 9:1–11

8. Minakawa N, Dida GO, Sonye GO, Futami K, Njenga SM (2012) Malaria Vectors in Lake Victoria and Adjacent Habitats in Western Kenya. PLoS ONE. https://doi.org/10.1371/journal.pone.0032725

9. Noor AM, Gething PW, Alegana VA, Patil AP, Hay SI, Muchiri E, Juma E, Snow RW (2009) The risks of malaria infection in Kenya in 2009. BMC Infect Dis 9:180

10. Osborne A, Manko E, Takeda M, et al (2021) Characterizing the genomic variation and population dynamics of Plasmodium falciparum malaria parasites in and around Lake Victoria, Kenya. Sci Rep 11:19809

11. World Health Organization (2015) The role of mass drug administration, mass screening and treatment, and focal screening and treatment for malaria: recommendations.

12. Idris ZM, Chan CW, Kongere J, et al (2016) High and Heterogeneous Prevalence of Asymptomatic and Sub-microscopic Malaria Infections on Islands in Lake Victoria, Kenya. Sci Rep 6:1–13

13. Uwimana A, Legrand E, Stokes BH, et al (2020) Emergence and clonal expansion of in vitro artemisinin-resistant Plasmodium falciparum kelch13 R561H mutant parasites in Rwanda. Nat Med 26:1602–1608

14. Balikagala B, Fukuda N, Ikeda M, et al (2021) Evidence of Artemisinin-Resistant Malaria in Africa. N Engl J Med 385:1163–1171

15. Artemisinin and the antimalarial endoperoxides: from herbal remedy to targeted chemotherapy - PubMed. https://pubmed.ncbi.nlm.nih.gov/8801435/. Accessed 2 Dec 2021

16. Djimdé A, Doumbo OK, Cortese JF, et al (2001) A Molecular Marker for Chloroquine-Resistant Falciparum Malaria. N Engl J Med 344:257–263

17. Olatunde A (1977) Chloroquine-resistant Plasmodium falciparum and malaria in Africa. Trans R Soc Trop Med Hyg 71:80–81

18. Nzila AM, Mberu EK, Sulo J, Dayo H, Winstanley PA, Sibley CH, Watkins WM (2000) Towards an understanding of the mechanism of pyrimethamine-sulfadoxine resistance in Plasmodium falciparum: genotyping of dihydrofolate reductase and dihydropteroate synthase of Kenyan parasites. Antimicrob Agents Chemother 44:991–996

19. Frosch AEP, Laufer MK, Mathanga DP, Takala-Harrison S, Skarbinski J, Claassen CW, Dzinjalamala FK, Plowe CV (2014) Return of widespread chloroquine-sensitive Plasmodium falciparum to Malawi. J Infect Dis 210:1110–1114

20. Kublin JG, Cortese JF, Njunju EM, Mukadam RAG, Wirima JJ, Kazembe PN, Djimdé AA, Kouriba B, Taylor TE, Plowe CV (2003) Reemergence of chloroquine-sensitive Plasmodium falciparum malaria after cessation of chloroquine use in Malawi. J Infect Dis 187:1870–1875

21. Mita T, Kaneko A, Lum JK, Bwijo B, Takechi M, Zungu IL, Tsukahara T, Tanabe K, Kobayakawa T, Björkman A (2003) Recovery of chloroquine sensitivity and low prevalence of the Plasmodium falciparum chloroquine resistance transporter gene mutation K76T following the discontinuance of chloroquine use in Malawi. Am J Trop Med Hyg 68:413–415

22. Mita T, Kaneko A, Lum JK, Zungu IL, Tsukahara T, Eto H, Kobayakawa T, Björkman A, Tanabe K (2004) Expansion of wild type allele rather than back mutation in pfcrt explains the recent recovery of chloroquine sensitivity of Plasmodium falciparum in Malawi. Mol Biochem Parasitol 135:159–163

23. Tumwebaze PK, Katairo T, Okitwi M, et al (2021) Drug susceptibility of Plasmodium falciparum in eastern Uganda: a longitudinal phenotypic and genotypic study. Lancet Microbe 2:e441–e449

24. Chebore W, Zhou Z, Westercamp N, et al (2020) Assessment of molecular markers of anti-malarial drug resistance among children participating in a therapeutic efficacy study in western Kenya. Malar J 19:291

25. Heinberg A, Kirkman L (2015) The molecular basis of antifolate resistance in Plasmodium falciparum: looking beyond point mutations. Ann N Y Acad Sci 1342:10–18

26. Calçada C, Silva M, Baptista V, Thathy V, Silva-Pedrosa R, Granja D, Ferreira PE, Gil JP, Fidock DA, Veiga MI (2020) Expansion of a Specific Plasmodium falciparum PfMDR1 Haplotype in Southeast Asia with Increased Substrate Transport. mBio 11:e02093-20

27. Sisowath C, Ferreira PE, Bustamante LY, Dahlström S, Mårtensson A, Björkman A, Krishna S, Gil JP (2007) The role of pfmdr1 in Plasmodium falciparum tolerance to artemether-lumefantrine in Africa. Trop Med Int Health 12:736–742

28. Nag S, Dalgaard MD, Kofoed P-E, Ursing J, Crespo M, Andersen LO, Aarestrup FM, Lund O, Alifrangis M (2017) High throughput resistance profiling of Plasmodium falciparum infections based on custom dual indexing and Illumina next generation sequencing-technology. Sci Rep 7:2398

29. Talundzic E, Ravishankar S, Kelley J, et al (2018) Next-Generation Sequencing and Bioinformatics Protocol for Malaria Drug Resistance Marker Surveillance. Antimicrob Agents Chemother 62:e02474-17

30. Kunasol C, Dondorp AM, Batty EM, Nakhonsri V, Sinjanakhom P, Day NPJ, Imwong M (2022) Comparative analysis of targeted next-generation sequencing for Plasmodium falciparum drug resistance markers. Sci Rep 12:5563

31. Clarke EL, Sundararaman SA, Seifert SN, Bushman FD, Hahn BH, Brisson D (2017) swga: a primer design toolkit for selective whole genome amplification. Bioinforma Oxf Engl 33:2071–2077

32. Bousema T, Okell L, Felger I, Drakeley C (2014) Asymptomatic malaria infections: detectability, transmissibility and public health relevance. Nat Rev Microbiol 12:833–840

33. Division of Malaria Control (2010) National Malaria Policy 2010.

34. Sisowath C, Petersen I, Veiga MI, Mårtensson A, Premji Z, Björkman A, Fidock DA, Gil JP (2009) In Vivo Selection of *Plasmodium falciparum* Parasites Carrying the Chloroquine-Susceptible *pfcrt* K76 Allele after Treatment with Artemether-Lumefantrine in Africa. J Infect Dis 199:750–757

35. Maiga H, Grivoyannis A, Sagara I, et al (2021) Selection of pfcrt K76 and pfmdr1 N86 Coding Alleles after Uncomplicated Malaria Treatment by Artemether-Lumefantrine in Mali. Int J Mol Sci 22:6057

36. Bacon DJ, Tang D, Salas C, et al (2009) Effects of Point Mutations in Plasmodium falciparum Dihydrofolate Reductase and Dihydropterate Synthase Genes on Clinical Outcomes and In Vitro

Susceptibility to Sulfadoxine and Pyrimethamine. PLoS ONE. https://doi.org/10.1371/journal.pone.0006762

37. van Lenthe M, van der Meulen R, Lassovski M, et al (2019) Markers of sulfadoxine–pyrimethamine resistance in Eastern Democratic Republic of Congo; implications for malaria chemoprevention. Malar J 18:430

38. Turkiewicz A, Manko E, Sutherland CJ, Diez Benavente E, Campino S, Clark TG (2020) Genetic diversity of the Plasmodium falciparum GTP-cyclohydrolase 1, dihydrofolate reductase and dihydropteroate synthetase genes reveals new insights into sulfadoxine-pyrimethamine antimalarial drug resistance. PLoS Genet 16:e1009268

39. World Health Organization (2014) WHO policy brief for the implementation of intermittent preventive treatment of malaria in pregnancy using sulfadoxine-pyrimethamine (IPTp-SP). 13

40. Gikunju SW, Agola EL, Ondondo RO, Kinyua J, Kimani F, LaBeaud AD, Malhotra I, King C, Thiong'o K, Mutuku F (2020) Prevalence of pfdhfr and pfdhps mutations in Plasmodium falciparum associated with drug resistance among pregnant women receiving IPTp-SP at Msambweni County Referral Hospital, Kwale County, Kenya. Malar J 19:190

41. Ashley EA, Dhorda M, Fairhurst RM, et al (2014) Spread of artemisinin resistance in Plasmodium falciparum malaria. N Engl J Med 371:411–423

42. Ouji M, Augereau J-M, Paloque L, Benoit-Vical F (2018) Plasmodium falciparum resistance to artemisinin-based combination therapies: A sword of Damocles in the path toward malaria elimination. Parasite. https://doi.org/10.1051/parasite/2018021

43. Ebong C, Sserwanga A, Namuganga JF, et al (2021) Efficacy and safety of artemether-lumefantrine and dihydroartemisinin-piperaquine for the treatment of uncomplicated Plasmodium falciparum malaria and prevalence of molecular markers associated with artemisinin and partner drug resistance in Uganda. Malar J 20:484

44. Tumwebaze PK, Conrad MD, Okitwi M, et al (2022) Decreased susceptibility of Plasmodium falciparum to both dihydroartemisinin and lumefantrine in northern Uganda. Nat Commun 13:6353

45. Isozumi R, Fukui M, Kaneko A, Chan CW, Kawamoto F, Kimura M (2015) Improved detection of malaria cases in island settings of Vanuatu and Kenya by PCR that targets the Plasmodium mitochondrial cytochrome c oxidase III (cox3) gene. Parasitol Int 64:304–308

46. Mangold KA, Manson RU, Koay ESC, Stephens L, Regner M, Thomson RB, Peterson LR, Kaul KL (2005) Real-Time PCR for Detection and Identification of Plasmodium spp. J Clin Microbiol 43:2435–2440

47. Oyola SO, Ariani CV, Hamilton WL, et al (2016) Whole genome sequencing of Plasmodium falciparum from dried blood spots using selective whole genome amplification. Malar J 15:597

48. Li H (2014) Aligning sequence reads, clone sequences and assembly con*gs with BWA-MEM. 0 Bytes

49. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R (2009) The Sequence Alignment/Map format and SAMtools. Bioinformatics 25:2078–2079

50. Auwera G van der, O'Connor BD (2020) Genomics in the Cloud: Using Docker, GATK, and WDL in Terra. O'Reilly Media, Incorporated

51. Garrison E, Marth G (2012) Haplotype-based variant detection from short-read sequencing. https://doi.org/10.48550/ARXIV.1207.3907

52. Danecek P, McCarthy SA (2017) BCFtools/csq: haplotype-aware variant consequences. Bioinformatics 33:2037–2039

Acknowledgements

We wish to thank all individuals that contributed the MDA and surveillance programs within the Lake Victoria region. AO is supported by a Nagasaki University – LSHTM PhD studentship funded by the WISE programme of MEXT. SC is funded by the Medical Research Council UK (Grant no. MR/M01360X/1). AK received support from JSPS KAKENHI (Grant No. JP18KK0248 & JP19H01080) and JICA/AMED joint research project (SATREPS) (Grant no. 20JM0110020H0002) and Hitachi fund. JG received support from a Tackling Infectious Burden in Africa (TIBA) fellowship, the African Academy of Sciences, and the Japan Society for Promotion of Sciences. TGC is supported by the Medical Research Council UK (Grant no. MR/M01360X/1, MR/N010469/1, MR/R020973/1, MR/X005895/1). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Authors' Contributions

AK, KK, JG, SC, and TGC conceived and designed the study; AK, WK, CC, MN, JK, and JG contributed parasite DNA for sequencing; AK, WK, CC, MN, JK, and JG provided biological materials and data. AO and SC coordinated the sequencing of samples; AO and JP performed the bioinformatic and statistical analysis, under the supervision of SC and TGC; AO wrote the first draft of the manuscript, and the final version included edits from all authors. The final manuscript was read and approved by all authors.

Data Availability

All raw sequence data is available from the ENA (project accession number PRJEB58092).

Competing Interests

There are no conflicts of interest or competing interests.

Figure legends

Figure 1: Primer and fragment design using custom dual indices for amplicon generation prior to Illumina sequencing.

Figure 2: Read depth across targeted mutations within amplicons covering five drug resistance genes; (top) non-*Pfk13*, with a trendline of 50-fold coverage; (bottom) *Pfk13* gene, with a trendline at 1500-fold coverage.

Figure 1: Primer and fragment design using custom dual indices for amplicon generation prior to Illumina sequencing.

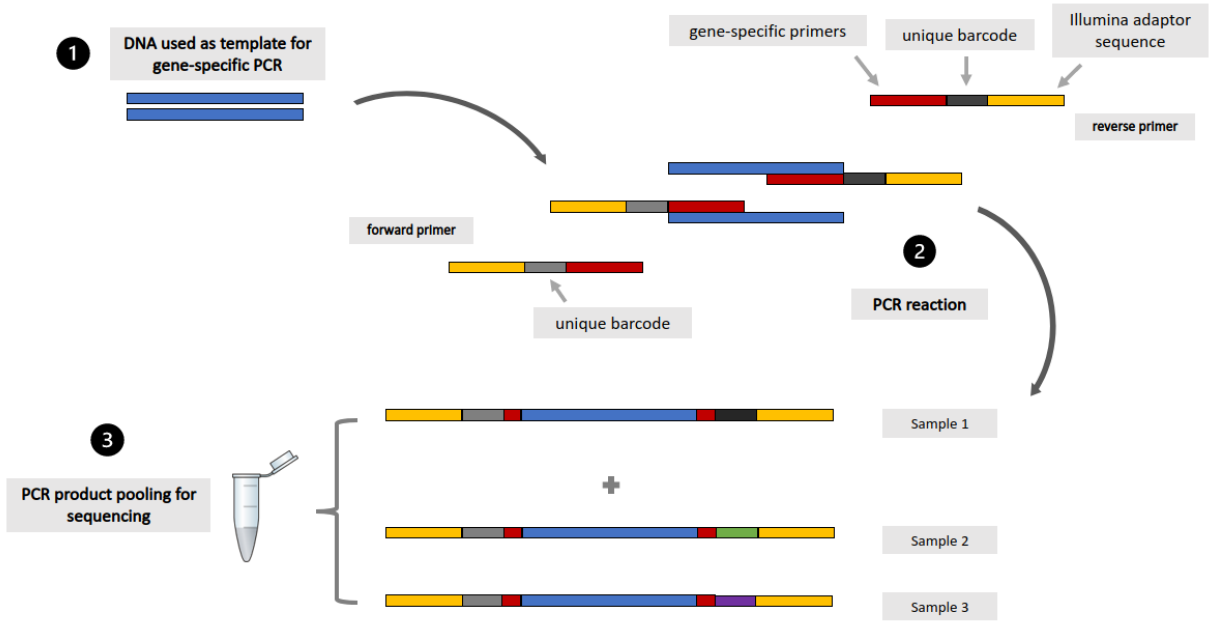
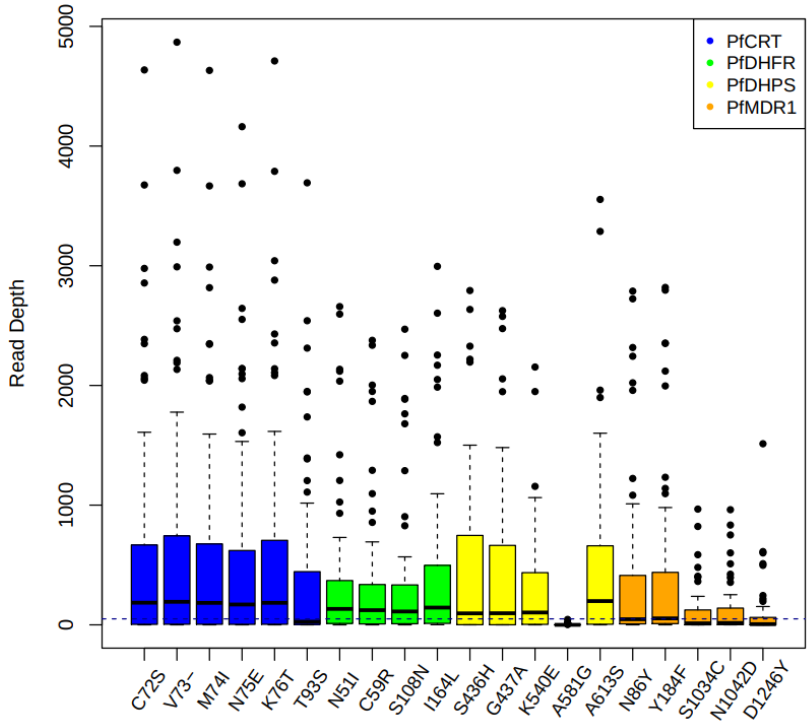


Figure 2: Read depth across targeted mutations within amplicons covering five drug resistance genes; (top) non-*Pfk13*, with a trendline of 50-fold coverage; (bottom) *Pfk13* gene, with a trendline at 1500-fold coverage.



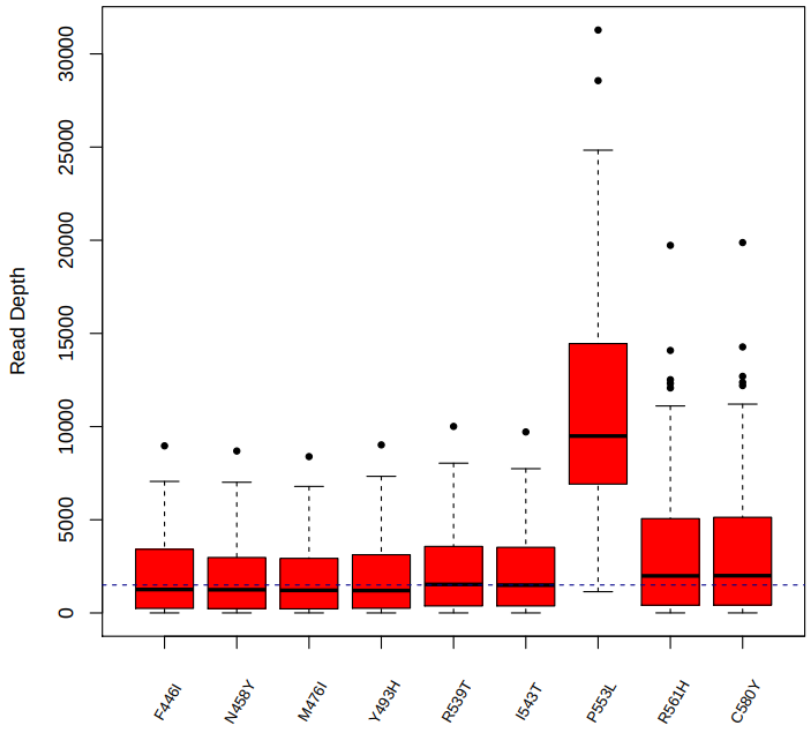


Table 1: Minor allele frequencies (MAF) of variants identified within the *P. falciparum* parasites on Ngodhe island and their respective frequencies across East African and West African parasite populations (minimum read depth >10-fold). [1]East Africa: Kenya, Tanzania, and Uganda; [2]West Africa: The Gambia and Mauritania.

Gene	SNP	MAF	95% CI	*n*	MAF	95% CI	*n*	East Africa[1] MAF (*n* = 228)	West Africa[2] MAF (*n* = 167)
Pfcrt	**M74I**	**6.5**	**2.2 - 18**	**46**	**0**	**0 - 12**	**27**	**0**	**0**
	N75E	**4.4**	**1.5 - 15**	**45**	**0**	**0 - 12**	**27**	**0**	**0**
	K76T	**4.4**	**1.5 - 15**	**45**	**7.4**	**2.1 - 23**	**27**	**17.5**	**13.0**
	H97Y	0	0 – 7.9	45	0	0 - 12	27	0	0
Pfmdr1	F61Y	3.5	0.97 - 12	57	0	0 - 11	30	0	0
	N86Y	**2.0**	**0.35 - 10**	**50**	**0**	**0 - 11**	**30**	**15.7**	**24.4**
	Y184F	**47.1**	**34 - 60**	**51**	**31.7**	**0 - 11**	**30**	**39.0**	**15.5**
	T199S	13.5	6.7 - 25	52	3.1	0.55 - 16	32	0.01	0
	F938Y	3.0	0.83 - 10	66	0	0 - 11	31	8.3	22.2
	F1072Y	2.0	0.37 - 11	48	0	0 - 10	35	0	0
	D1246Y	**16.0**	**7.1 - 33**	**31**	**2.9**	**0.51 - 15**	**35**	**12.9**	**31.2**
Pfk13	A578S	3.1	0.83 - 11	65	0	0 - 10	35	1.1	0
	F451Y	4.5	1.6 - 13	66	0	0 - 10	35	0	0
Pfdhfr	**N51I**	**77.0**	**62 - 85**	**53**	**100**	**90 - 100**	**34**	**91.4**	**92.5**
	C59R	**76.9**	**64 - 86**	**52**	**91.2**	**77 - 97**	**34**	**86.4**	**93.0**
	S108N	**80.8**	**68 - 89**	**52**	**100**	**90 - 100**	**36**	**99.4**	**93.0**
	I164L	**5.8**	**2.0 - 16**	**52**	**6.9**	**2.9 - 22**	**36**	**1.5**	**0**
Pfdhps	S436H	11.9	5.2 - 25	42	22.9	12 - 39	35	0	0
	G437A	**100**	**92 - 100**	**42**	**100.0**	**90 - 100**	**35**	**85.3**	**70.0**
	K540E	**79.2**	**66 - 88**	**48**	**95.7**	**85 - 99**	**35**	**84.4**	**30.0**

Table 2: Drug resistance haplotype frequencies on *Pfcrt and Pfmdr1* identified within the Ngodhe island *P. falciparum* parasite population (minimum read depth >10-fold). Amino acid changes in bold.

No. of samples	*Pfcrt* haplotype					Freq. (%)
	C72S	**V73**	**M741**	**N75E**	**K76T**	
43*	C	V	M	N	K	95.5
2	C	V	**I**	**E**	**T**	4.5
No. of samples	***Pfmdr1* haplotype**					**Freq. (%)**
	N86Y	**Y184F**	**F938Y**	**S1034C**	**N1042D**	**D1246Y**
14*	N	Y	F	S	N	D	45.2
12	N	**F**	F	S	N	D	38.7
3	N	Y	F	S	N	**Y**	9.7
2	N	**F**	F	S	N	**Y**	6.4

*Wild type (WT)

Table 3: Drug resistance haplotype frequencies on *Pfdhfr* / *Pfdhps* identified within the Ngodhe island *P. falciparum* parasite population (minimum read depth >10-fold). Amino acid changes in bold.

No. of samples	*Pfdhfr / Pfdhps*							Freq. (%)
	N51I	C59R	S108N	I164L	A437G	K540E	A581G	
23	**I**	**R**	**N**	I	**G**	**E**	A	62.2
5	N	C	S	I	**G**	K	A	13.5
2	**I**	**R**	**N**	I	**G**	K	A	5.4
2	**I**	**R**	**N**	**L**	**G**	**E**	A	5.4
2	N	C	S	I	**G**	**E**	A	5.4
1	**I**	C	**N**	I	**G**	K	A	2.7
1	**I**	C	**N**	I	**G**	**E**	A	2.7
1	N	**R**	**N**	I	**G**	**E**	A	2.7

Table S1: Read depth of sequencing reads across the 5 sequenced resistance genes (*n* = 70).

Gene	Median	Max
Pfcrt	155	471
Pfmdr1	69	3962
Pfdhps	103	2793
Pfdhfr	126	2604
Pfk13	1760	31280

Table S2: Primers used for the amplification of amplicons on *Pfcrt, Pfmdr1, Pfdhps, Pfdhfr,* and *Pfk13*. (PCR reaction; S = simplex reaction, M = Multiplex reaction).

Amplicon name	Gene ID	Primer Sequence	PCR reaction
Pfcrt	PF3D7_0709000	Forward: TCTTGTCTTGGTAAATGTGCTCA	S1
		Reverse: AGGCCAAAATGACTGAACAGG	
Pfmdr1.1	PF3D7_0523000	Forward: CGTTTAAATGTTTACCTGCACAAC	M1
		Reverse: TGACACCACAAACATAAATTAACGG	
Pfmdr1.7		Forward: TTTGTCCAATTGTTGCAGCTGTATTAACTTT	M4
		Reverse: TGCATTTTCTGAATCTCCTTTTAAGGACATT	
Pfmdr1.8		Forward: GGTAAAGTTGATATTAAAGATGTAAATTTCC	M3
		Reverse: TGGTCCAACATTTGTATCATATTTATTTGG	
Pfdhfr	PF3D7_0417200	Forward: ATGATGGAACAAGTCTGCGACGTTTTCGA	M2
		Reverse: CTAAAAATTCTTGATAAACAACGGAACCTCC	
Pfdhps.3	PF3D7_0810800	Forward: ACCATCAGATGTTTATATAACAAATATGTG	M3
		Reverse: CTGGATTATTTGTACAAGCACTAATATCA	
Pfdhps.4		Forward: AGAATGTGTTGATAATGATTTAGTTGATAT	M4
		Reverse: GATATAAAAGTTGATCCTTGTCTTTCCT	
PfK13.4	PF3D7_1343700	Forward: TAAGTGGAAGACATCATGTAACCAGAGA	M2
		Reverse: CTTCTACATTCGGTATAATAGAAGAGCC	
PfK13.5		Forward: ATGATGGCTCTTCTATTATACCGAATG	M1
		Reverse: GCTATTAAAACGGAGTGACCAAATCTG	

Chapter 4: A high-resolution analysis of Plasmodium falciparum population dynamics in East Africa and genomic surveillance along the Kenya-Uganda border

RESEARCH PAPER COVER SHEET

Please note that a cover sheet must be completed <u>for each</u> research paper included within a thesis.

SECTION A – Student Details

Nagasaki Student No	59719003	**Title**	Miss
LSHTM Student ID No	Lsh1807687		
First Name(s)	Ashley Alexandra		
Surname/Family Name	Osborne		
Thesis Title	A multifaceted investigation of the genomics of malaria, from parasite to host, using next-generation sequencing technologies		
Nagasaki Supervisor(s)	Akira Kaneko, Kiyoshi Kita		
LSHTM Supervisor(s)	Taane Clark, Susana Campino		

If the Research Paper has previously been published please complete Section B, if not please move to Section C.

<u>SECTION B – Paper already published</u>

Where was the work published?	
When was the work published?	

If the work was published prior to registration for your research degree, give a brief rationale for its inclusion			
Have you retained the copyright for the work?*	Choose an item.	Was the work subject to academic peer review?	Choose an item.

*If yes, please attach evidence of retention. If no, or if the work is being included in its published format, please attach evidence of permission from the copyright holder (publisher or other author) to include this work.

SECTION C – Prepared for publication, but not yet published

Where is the work intended to be published?	Scientific Reports
Please list the paper's authors in the intended authorship order:	Ashley Osborne, Emilia Manko, Harrison Waweru, Akira Kaneko, Kiyoshi Kita, Susana Campino, Jesse Gitaka & Taane G. Clark
Stage of publication	**Submitted**

SECTION D – Multi-authored work

For multi-authored work, give full details of your role in the research included in the paper and in the preparation of the paper. (Attach a further sheet if necessary)	I carried out laboratory work preparing samples for WGS, including selective whole genome amplification, DNA clean-up, and shipment of samples. I performed bioinformatic analyses and interpreted the results under the supervision of my supervisors. I wrote and prepared the first draft of the manuscript that was circulated to my supervisors and co-authors.

SECTION E – Names and affiliations of co-author(s)

Please list all the co-authors' names and their affiliations.

Ashley Osborne - Faculty of Infectious and Tropical Diseases, London School of Hygiene and Tropical Medicine & School of Tropical Medicine and Global Health, Nagasaki University Emilia Manko - Faculty of Infectious and Tropical Diseases, London School of Hygiene and Tropical Medicine Harrison Waweru – Centre for Malaria Elimination, Mount Kenya University Akira Kaneko - Department of Parasitology, Osaka Metropolitan University Kiyoshi Kita - School of Tropical Medicine and Global Health, Nagasaki University Susana Campino - Faculty of Infectious and Tropical Diseases, London School of Hygiene and Tropical Medicine Jesse Gitaka - Directorate of Research and Innovation, Mount Kenya University Taane G. Clark - Faculty of Epidemiology and Population Health, London School of Hygiene and Tropical Medicine & Faculty of Infectious and Tropical Diseases, London School of Hygiene and Tropical Medicine

SECTION F

I confirm that all co-authors have agreed that the above paper will be included in my PhD thesis.

Student Signature	[redacted]
Date	06/07/23

LSHTM Supervisor Signature	[redacted]
Date	06/07/23

Nagasaki University Supervisor Signature	[redacted]
Date	06/07/23

Title:

A high-resolution analysis of *Plasmodium falciparum* population dynamics in East Africa and genomic surveillance along the Kenya-Uganda border

AUTHORS

Ashley Osborne[1,2], Emilia Mańko[1], Harrison Waweru[3,4], Akira Kaneko[5,6], Kiyoshi Kita[2], Susana Campino[1,*], Jesse Gitaka[3,4,*], Taane G. Clark[1,7,*]

[1]Faculty of Infectious and Tropical Diseases, London School of Hygiene & Tropical Medicine, London, UK

[2]School of Tropical Medicine and Global Health, Nagasaki University, Nagasaki, Japan

[3]Directorate of Research and Innovation, Mount Kenya University, Thika, Kenya

[4]Centre for Malaria Elimination, Mount Kenya University, Thika, Kenya

[5]Department of Parasitology, Graduate School of Medicine, Osaka Metropolitan University, Osaka, Japan

[6]Department of Microbiology, Tumor and Cell Biology, Karolinska Institutet, Stockholm, Sweden

[7]Faculty of Epidemiology and Population Health, London School of Hygiene & Tropical Medicine, London, UK

* Joint authors

ABSTRACT

Despite decades of progress, malaria incidence rates in 2020 and 2021 had their most dramatic increase since the start of the millennium and these numbers continue to remain elevated. East African countries alone accounted for approximately 10% of all malaria cases worldwide in 2021, with an estimated 24.2 million cases and potentially upwards of 53,000 deaths. Although malaria-related morbidity and mortality remains high within this region, high resolution genome-wide parasite population analyses have been extremely limited in Kenya, Tanzania, and Uganda. The border of Kenya and Uganda has become an area of particular concern in recent years as Uganda struggles with increases in malaria incidence and confirmed the emergence of locally spreading clinically artemisinin resistant *P. falciparum* parasites.

To establish genomic surveillance along the Kenyan-Ugandan border and produce a high-resolution analysis of *P. falciparum* population dynamics occurring within East Africa, we generated WGS sequencing data for 30 parasites from the Western Kenyan border county of Bungoma, as well as generated a comprehensive analysis of East African isolates through a comparison with publicly available sequences (N = 587). Our analysis of *P. falciparum* parasites from East Africa form subpopulations with distinct genetic structure and diverse ancestral origins. Ancestral admixture analysis of these subpopulations alongside isolates from across Africa (N = 363) suggested potential independent ancestral populations from other major African populations. Within isolates from Western Kenya, the prevalence of drug resistance biomarkers associated with chloroquine resistance (e.g. *Pfcrt* K76T) were significantly reduced compared to wider East African populations and a single isolate was identified to contain a non-synonymous SNP on *PfK13*, resulting in a variant within a WHO candidate marker of reduced susceptibility to artemisinin.

INTRODUCTION

Despite decades of progress, and a drastic reduction in malaria burden worldwide since the 1990's, malaria incidence rates in 2020 and 2021 had their most dramatic increase since the start of the millennium [1, 2]. Increased incidence in 2020 led to an estimated 22 million additional malaria cases worldwide, resulting in a total of 254 million cases and 625,000 deaths, exceeding the previous years' estimated deaths by 57,000. Although many malaria control programmes have been resumed following the end to strict COVID-19 pandemic precautions, these numbers have continued to remain elevated in 2021. East African countries bordering Lake Victoria, including Kenya, Tanzania, and Uganda, accounted for approximately 10% of all malaria cases worldwide in 2021, with an estimated 24.2 million cases and potentially upwards of 53,000 deaths [1]. The *Plasmodium falciparum* parasite is estimated to be responsible for almost 100% of cases within the region and is the leading cause of severe disease [3].

Although malaria transmission and infection risk has reduced considerably in Kenya, low-elevation regions, in particular Western Kenya and the Lake Victoria basin, remain areas of intense transmission and malaria continues to be a leading cause of death for children under 5 years of age [4, 5]. This high malaria incidence is driven by a variety of environmental and geopolitical factors, including fluid human migration patterns across country borders for socio-economic reasons, as well as tropical weather and freshwater conditions that favour vector breeding [6, 7]. These factors have complicated the implementation of targeted malaria control within the region. The border of Kenya and Uganda has become an area of particular concern in recent years as Uganda struggles with increases in malaria incidence and mortality, in addition to the confirmed emergence and local spread of clinically artemisinin resistant *P. falciparum* parasites [1, 8]. Artemisinin resistance in Uganda jeopardises malaria control efforts across East Africa and highlights the need to perform regular large-scale surveys of parasite populations to monitor for the emergence or cross-border spread of artemisinin resistant parasites.

East African parasite populations have been well documented to form a genetic cluster when compared to parasite populations from other regions of Africa. In general, East African parasite populations also tend to exhibit strong genetic ancestral linkages with one another and transmission connectivity supported by their geographic proximity [5, 9]. However, high resolution genome-wide parasite population analyses have been extremely limited in Kenya, Tanzania, and Uganda, with most genetic studies focusing on microsatellite markers to assess genetic diversity, and sparse WGS data available from Uganda [6, 10–13]. Although appropriate for establishing baseline genetic diversity metrics, microsatellite markers are not able to provide high resolution insights into the dynamics of a population or across multiple populations [14].

To establish genomic surveillance along the Kenyan-Ugandan border and produce a high-resolution analysis of *P. falciparum* population dynamics occurring within East Africa, we generated WGS sequencing data for parasites from the Western Kenyan border county of Bungoma, as well as generated a comprehensive analysis of East African isolates through a comparison with publicly available sequences [15]. Our analysis revealed that *P. falciparum* parasites from East Africa form subpopulations with distinct genetic structure and diverse ancestral origins, with ancestral admixture analysis suggesting seemingly independent ancestral populations from other major African populations. Within isolates from Western Kenya, the prevalence of drug resistance biomarkers associated with chloroquine resistance were significantly reduced compared to wider East African populations and a single isolate was identified to contain a non-synonymous SNP on *PfK13*, resulting in a variant within a WHO candidate marker of reduced susceptibility to artemisinin.

RESULTS

Genome-wide population data and multi-clonality

To assess the population structure of East African *P. falciparum* populations, SNP variants were extracted from the WGS data of 587 *P. falciparum* isolates with a minimum mapping percentage of 70% and read depth of 5 across 70% of the genome, resulting in a total of 710,552 high-quality SNPs called from non-hypervariable regions of the *P. falciparum* genome. The final dataset to assess East African population structure included isolates from Bungoma county in Western Kenya (N = 30), as well as isolates from multiple points around the Lake Victoria basin (N = 160), Kenya (N = 74), Tanzania (N = 323), and Uganda (N = 12) (**Figure 1A; Supplementary Table S1**). Another dataset was generated to assess the ancestral origins of East African *P. falciparum* populations by placing them within the context of populations from across the African continent, this included 640,596 high-quality SNPs extracted from 363 isolates from East Africa (Kenya, Tanzania, and Uganda; N = 218), West Africa (Guinea and The Gambia; N = 47), Horn of Africa (Ethiopia; N = 25), Central Africa (Cameroon; N = 25); South Central Africa (the Democratic Republic of Congo; N = 25), and Southern Africa (Malawi; N = 25) (**Supplementary Table S2**).

The F_{ws} metric, or mean inbreeding coefficient, was calculated for each East African subpopulation to determine the proportion of complex infections amongst isolates, as well as assess their within-host population diversity or assumed risk of out-crossing/inbreeding. Monoclonal *P. falciparum* isolates exhibit "high" F_{ws} estimates ≥ 0.95. Samples from Western Kenya had a mean F_{ws} coefficient of 0.851, with only 6 out of 29 isolates exhibiting "high" F_{ws} estimates (**Supplementary Figure S1**). Lower mean F_{ws} estimates are generally associated with higher proportions of complex infections and a high degree

of panmixis within the population, common in Kenyan *P. falciparum* isolates from high transmission regions. In general, subpopulations across East Africa had mean F_{ws} coefficients ranging 0.738 to 0.907, with the lowest value (0.738) observed for isolates along the Kenyan Lake Victoria mainland where transmission is high and stable across much of the year.

P. falciparum isolates from highland epidemic outbreaks form distinct population clusters

A pairwise nucleotide matrix generated from 76 samples from Western Kenya (*n* = 30) and Lake Victoria, Kenya (*n* = 46), identified 284,667 high-quality SNPs across non-hypervariable regions of the *P. falciparum* genome. A SNP-based principal component analysis (PCA) and maximum likelihood tree of these populations identified one main population cluster with all 31 isolates from Lake Victoria and 17 isolates from Bungoma county, as well as two distinct clusters of 7 and 4 isolates from Bungoma county (**Supplementary Figure S2**). These observations suggest that the *P. falciparum* isolates in cluster 2 and cluster 3 are nearly genetically identical to, or clones of, one another. Bungoma county is split between both lake-endemic and highland epidemic zones based on malaria endemicity classifications, with highland epidemic zones being prone to malaria outbreaks that can result in distinct *P. falciparum* genetic clusters [16].

For further genome-wide population analyses alongside isolates from other East African populations, as well as populations across Africa, a random sample from cluster 2 and from cluster 3 were chosen to be representative samples of each clade to prevent skewing of population dynamics and genetic structure.

Drug resistance

Non-synonymous mutations on genes associated with the ability to confer resistance to antimalarial drugs were categorised for isolates from Western Kenya with a read depth ≥5 and a maximum F_{st} value for each SNP was calculated by comparing Western Kenyan isolates with East African isolates (N = 460) (**Supplementary Table S3**). Variant frequencies from Western Kenyan isolates were also compared with isolates from the Kenyan region of Lake Victoria (N = 109) and West Africa (N = 50).

A non-synonymous SNP on *PfK13* resulting in the variant V568I was identified in 1 isolate out of 37 within Western Kenya. The V568G variant has been identified as an *in vitro* candidate marker of reduced susceptibility to artemisinin, however the impact of the V568I variant on artemisinin tolerance has not yet been characterised. An in-silico protein model of *Pfk13* was generated with the wild type position V568, WHO candidate variant V568G, and the V568I variant observed within the Western Kenyan isolate (**Supplementary Figure S3**).

Resistance markers associated with resistance to chloroquine were significantly reduced in Western Kenyan isolates compared to other East African isolates. Within isolates from Western Kenya, only 1 isolate out of 36 (2.8%) contained the main *Pfcrt* biomarker for resistance K76T while 14.1% of isolates from East Africa were identified to contain the variant (N = 65/460). On *Pfmdr1* the N86Y variant was not identified within any isolates from Western Kenya (East Africa MAF, 5.9%). It is theorised that the reference N86 is selected for by the use of lumefantrine and may reduce susceptibility to lumefantrine, piperaquine, and mefloquine [17]. Variants Y184F and D1246Y were observed at frequencies of 54.5% (18/33) and 10.8% (4/37), respectively, compared to 37.7% (173/460) and 5% (23/460) in East African isolates. The Y184F variant, although not significantly associated with reduced susceptibility to lumefantrine, is believed to be genetically correlated to the acquisition of a drug-resistance phenotype [17].

Variants associated with resistance to sulphadoxine and pyrimethamine on *Pfdhps* and *Pfdhfr*, respectively, were identified at high frequencies, consistent with East African populations [18]. Variants N51I, C59R, and S108N on *Pfdhfr* were observed in 100% of isolates from Western Kenya, while the I164L variant was only identified in 1 out of 38 isolates (2.2%). Variants on *Pfdhps* were also observed at high frequencies within Western Kenyan isolates, with S436H, G437A, and K540E occurring in 21.6%, 100%, and 87.1% of isolates. Haplotype analysis was done for variants on *Pfdhfr* (N51I, C59R, S108N, and I164L) and *Pfdhps* (S436H, G437A, K540E, and A581G) within parasites containing a read depth >5 at each position (N = 31). The wild-type haplotype, NCSISGKA, was not observed in any of the screened isolates while the quintuple mutant **IRN**I**S**A**E**A accounted for 74.2% of isolates (23/31). The sextuple mutant **IRN**I**HAE**A was observed in 7 isolates while another sextuple mutant, **IRNL**S**AE**A, was observed in a single isolate (1/31).

East Africa has P. falciparum subpopulations with distinct genetic structure

A SNP-based maximum likelihood tree revealed a handful of subpopulations within the larger East African *P. falciparum* population (**Figure 1B**). Lake Victoria isolates from Kenya (i.e., Kisumu, Mfangano island, Ngodhe island, and Suba district) appeared to cluster closely with isolates from Western Kenya and Uganda, separately from other East African subpopulations. Isolates from the Tanzanian region of Lake Victoria and Lake Tanganyika appear to group more closely with one another while isolates from North East Tanzania and Eastern Kenya formed a loose cluster with South East Tanzania interspersed throughout. This population structure identified by the maximum likelihood tree was supported by principal component analysis (PCA) in which separation of Kenyan Lake Victoria isolates alongside Western Kenya and Uganda was further highlighted, as well as the clustering of isolates from North East Tanzania and South East Tanzania (**Figure 1C, D**).

Ancestral admixture analysis reveals diverse ancestral origins of East African subpopulations

To evaluate the ancestral origins of East African *P. falciparum* subpopulations, genome-wide SNPs from isolates collected across the African continent (640,596 SNPs; N = 363 isolates) were used to infer ancestral genotype frequencies and combined with geographical coordinates to produce spatial models of allele sharing. With the optimum number of ancestral populations (K value) estimated to be 6 (K1 – K6), the resulting admixture analysis revealed that isolates from East African subpopulations have distinct ancestral origins from one-another, including three ancestral populations seemingly independent from wider African populations (**Figure 2**). A maximum likelihood tree was generated using the same genome-wide SNPs incorporated into the ancestral admixture analysis and supported the identified population structure (**Figure 2B**).

The K1 ancestral population appears to demonstrate high proportions of shared ancestry between isolates from Western Kenya (proportion K1; 79.6%), Central Uganda (65.4%), and Lake Victoria isolates from Kenya (islands = 79.5%; mainland = 79.7%), whereas the K4 ancestral population appears to differentiate Tanzanian Lake Victoria isolates (proportion K4; 44.3%) from Kenyan Lake Victoria isolates (islands = 0%; mainland = 0.2%), Western Kenya (0%), and Uganda (1.5%) (**Supplementary Figure S4**). The K3 ancestral population appears to link together isolates from Eastern Kenya (proportion K3; 59.6%), North East Tanzania (67.5%), and Lake Tanganyika, TZ (78.3%), whereas South Central Africa appears to share a lower ancestral proportion with K3 (36.1%) and higher ancestral proportions with K2 (proportion K2; 47.1%), alongside Central African (78.5%) and West African isolates (85%). The K6 ancestral population accounted for high proportions of shared ancestry within Southern African (proportion K6; 91.5%) and, to lesser extent, South East Tanzanian isolates (34.6%). South East Tanzanian isolates appeared to have comparatively highly mixed ancestral proportions (K1, 12.1%; K2, 13.8%; K3, 11.3%; K4, 28.2; K6, 34.6%). Alternatively, as has been previously documented, isolates from the Horn of Africa demonstrated distinct ancestral origins from other regional African populations with K5 accounting for their highest proportion of ancestry (81.2%).

IBD analysis of ancestral populations reveals similar regions of homology across East Africa

Identify-by-descent (IBD) was calculated to determine the structure of East African subpopulations at the chromosome level by measuring the proportion of pairs identical by descent at each SNP across all isolates within the population. IBD was measured according to associated ancestral populations K1 – K6 determined via admixture analysis due to the high degree of correlation between geographical proximity and genetic structure. As anticipated, isolates assigned to the K5 ancestral population (i.e. Horn of Africa) had the highest fraction of pairwise IBD across the genome (mean = 0.0849, range =

0.0184 - 0.8945), reflecting high genetic relatedness and conservation across the genome between isolates (**Supplementary Table S4; Supplementary Figure S4**). Alternatively, isolates from K1 (i.e. Western Kenya, Lake Victoria Kenya, and Central Uganda), K4 (i.e. Lake Victoria Tanzania), and K6 (i.e. South East Tanzania and Southern Africa) had the lowest fractions of IBD, suggesting overall lower genetic relatedness or less conservation of genomic regions between isolates (K1: mean = 0.009719, range = 0 – 0.13946; K4: mean = 0.0092520, range = 0 – 0.08748; K6: mean = 0.0093663, range = 0 – 0.87062). A visualisation of genome-wide chromosome-level IBD for ancestral populations K1 – K6 are presented (**Supplementary Figure S5**).

The top 5% of IBD positions within isolates from the K1 ancestral population were distributed across 35 regions on 8 chromosomes while the top IBD positions within the K3 population were distributed across 39 regions on 7 chromosomes (**Supplementary Table 5**). Within the K3 population, the region on chromosome 8 encompassed *hydroxymethyldihydropterin pyrophosphokinase-dihydropteroate synthase* encoding gene (*Pfdhps*, PF3D7_0810800), known to confer partial resistance to the antimalarial sulphadoxine-pyrimethamine (SP). This region on chromosome 8 was also identified within the top 5% of IBD positions within the K4 population, distributed across 22 regions on 7 chromosomes, and K6 population (47 regions on 10 chromosomes).

Evidence of selective sweeps between P. falciparum subpopulations in East Africa

To identify variants under positive directional selection between subpopulations within East Africa, the genome-wide haplotype structure of isolates was analysed to locate regions of high local homozygosity relative to neutral expectation. The integrated haplotype score (*iHS*) test statistic was used to identify regions with high local homozygosity (i.e., positive selection pressure) within a single population while cross-population selection pressure was measured using the *Rsb* metric which compares extended haplotype homozygosity between populations.

As is common, the SNPs most frequently observed to be under within-population positive selection were genes associated with host immune response and parasite immune evasion. Within K1 associated isolates, Plasmepsin X (PMX), characterised as a mediator of parasite invasion and egress, was identified to have a significant *iHS* value ($(-\log10[1-2|\Phi iHS-0.5|]) > 4.0$) (**Figure 3A, B, C; Supplementary Table S6**). Within K3 associated isolates, there were 5 genes of interest identified to have SNPs with significant *iHS* values, including heat shock protein 40 (HSP40), believed to play a role in parasite pathogenicity, and CX3CL1-binding proteins 1 and 2, linked to the cytoadherence of infected erythrocytes, as well as the cAMP-dependent protein kinase regulatory subunit, known to be involved in parasite invasion of erythrocytes [19]. The cAMP-dependent protein kinase regulatory

subunit and CX3CL1-binding protein 3 were also identified to have significant *iHS* values within populations K6.

Cross-population analysis of East African populations revealed two regions under high positive directional selection on genes with known, or theorised, drug resistance associations ($-\log_{10}$(*p*-value) > 5), including *hydroxymethyldihydropterin pyrophosphokinase-dihydropteroate synthase* (*Pfdhps*) and *ubiquitin carboxyl-terminal hydrolase 1* (*Pfubp1*) (**Figure 3D, E, F; Supplementary Table S9**). The K1 population (i.e. Western Kenya, Lake Victoria Kenya, and Central Uganda), when compared to K2 associated isolates (i.e. Central Africa, South Central Africa, and West Africa), had 184 markers on *Pfdhps*, known to confer partial resistance to the antimalarial sulphadoxine-pyrimethamine, with a mean *Rsb* of 6.484. Isolates associated with the ancestral population K3 (i.e. Eastern Kenya, North East Tanzania, and Lake Tanganyika, TZ), when cross-examined against isolates associated with K6 (i.e. South East Tanzania, Southern Africa), was found to have 55 markers on *Pfubp1*, a candidate gene for reduced susceptibility to artemisinin, with a mean *Rsb* of 6.89 [20]. Comparisons of K1 and K2 with K6 also identified positive selection of regions on the merozoite surface protein 3 encoding gene (*Pfmsp3*) known to be an important mediator of antibody responses and associated with naturally acquired immunity, as well as a popular malaria vaccine candidate. Within the K1 population, 320 markers were identified on *Pfubp1* with a mean *Rsb* of 9.844 while 312 markers were identified within the K3 population with a mean *Rsb* of 6.885.

DISCUSSION

Advancements in next-generation sequencing platforms and the curation of large, publicly available datasets, has provided a means to produce high resolution insights into the genetic structure and ancestral origins of parasite populations within high transmission regions. Despite accounting for 10% of all malaria cases worldwide annually, and the recent emergence of locally spreading clinically artemisinin resistant *P. falciparum* parasites in Uganda, East African parasite populations have been largely underrepresented in whole genome population studies [1, 8]. To establish genomic surveillance along the Kenyan-Ugandan border and supplement the limited sequencing data available for Ugandan parasite populations, we generated sequencing data for 30 *P. falciparum* isolates from Bungoma county in Western Kenya. This sequencing data was then combined with newly available *P. falciparum* sequencing data from a public repository to produce a high-resolution analysis of the genetic structure and ancestral origins of parasite subpopulations occurring within East Africa [15]. Sequencing data from Western Kenya identified decreasing frequencies of chloroquine resistance-associated biomarkers compared to other East African subpopulations, as well as a non-synonymous variant on *pfk13* occurring within a codon of concern flagged by the WHO as a candidate for artemisinin resistance [21]. Our analysis also identified subpopulations within East Africa with distinct genetic structure and diverse ancestral populations from other major regional populations across Africa.

Initial analysis of *P. falciparum* isolates from Bungoma county in Western Kenya, alongside isolates from the Kenyan region of Lake Victoria, identified evidence of distinct genetic clusters within a region anticipated to have high genetic homogeneity [5]. A maximum likelihood tree using mapped sequence alignments revealed two small clusters of samples from Bungoma county that were distinct from other isolates from Bungoma county and Lake Victoria. These two clusters also appeared to be nearly genetically identical to, or clones of, one-another. Further investigation revealed that Bungoma county is split between two malaria endemicity classification zones, lake endemic and highland epidemic. Highland epidemic zones are prone to malaria outbreaks that can result in genetic clustering of *P. falciparum* clones within an area while lake endemic is known to sustained transmission throughout the year with peaks in transmission directly following the rainy seasons [4, 16]. Further investigation of *P. falciparum* isolates from Bungoma county and higher resolution location data could provide greater insight into this phenomenon in future work. For our wider population analyses, one isolate from each cluster was included with isolates from the main Bungoma county and Lake Victoria isolates to prevent skewing based on the close genetic relatedness of the isolates.

Following the emergence of widespread resistance, chloroquine (CQ) was discontinued as the first-line treatment for uncomplicated malaria, despite it still being available as an over-the-counter

treatment for fevers in some places [22]. A handful of countries that implemented swift and strict cessation of CQ usage have documented the return of CQ-sensitive parasites, including Malawi and Zambia, however these results have not been observed everywhere [23–25]. The prevailing theory is that illegal over-the-counter sales of CQ may still be driving selection pressure for CQ-resistance parasites in regions where CQ is no longer the primary treatment method. Isolates from Bungoma county in Western Kenya demonstrated a marked reduction in variants associated with CQ resistance, with only one isolate out of 36 containing the main K76T biomarker for resistance, suggesting the possibility that CQ-sensitive parasites may be returning to Kenya. Despite this reduction in CQ resistance biomarkers in Bungoma county, CQ resistance-associated variants still persist in Kenyan Lake Victoria isolates [5]. This may be driven by continued sales of CQ across the lake basin while CQ usage may be reduced in regions along the Kenya-Uganda border.

Although the presence of CQ resistance markers was reduced, variants associated with resistance to SP were observed at high frequencies in Western Kenya, with many variants observed to be fixed within the parasite population. Haplotype analysis for variants on *Pfdhfr* (N51I, C59R, S108N, and I164L) and *Pfdhps* (S436H, G437A, K540E, and A581G) revealed no wild-type parasites (NCSISGKA) within the region, with a quintuple mutant (**IRN**IS**AE**A) identified in approximately 75% of the population. The continued use of SP as an intermittent preventative treatment in pregnancy (IPTp) is likely driving this continued expansion and fixation of SP resistance within the population [1]. In addition to variants on *Pfdhps* and *Pfdhfr*, a non-synonymous variant on *Pfk13* was identified to confer a V to I amino acid substitution at codon 568. The V568G variant has been identified as an *in vitro* candidate marker of reduced susceptibility to artemisinin by the WHO, however the impact of the V568I variant on artemisinin tolerance has not yet been characterised [26]. An in-silico protein model of the wild-type position, WHO candidate, and V568I Western Kenya variant was generated to demonstrate structural variations, although *in vitro* analysis of the V568I variant will be required to determine the impact, if any, it may have on artemisinin susceptibility. Artemisinin resistance in Ugandan *P. falciparum* populations, and the identification of this V568I mutant, highlights the need to perform regular monitoring of parasites within the region to preserve the efficacy of existing malaria treatment therapies.

Previous characterisations of the genomic diversity of East Africa have revealed a largely homogenous population structure linked to transmission connectivity and geographic proximity, as well as high levels of human and vector migration [5, 9]. However, an extended dataset including *P. falciparum* isolates from regions throughout East Africa has revealed multiple parasite subgroups within the region. Isolates from Western Kenya, the Kenyan region of Lake Victoria, and Central Uganda were identified to have a distinct genetic structure when compared with other East African parasites and

did not appear to be closely linked to isolates from the Tanzanian region of Lake Victoria. Isolates from the Tanzanian territory of Lake Victoria appeared to cluster more closely with isolates from Lake Tanganyika, likely due to their relative geographic proximity. The *P. falciparum* incidence rate and proportion of complex infections, or within-host diversity, was higher in isolates from Central Uganda and the Kenyan region of Lake Victoria, when compared to Tanzanian Lake isolates, and may be driving differences in parasite populations despite high human migration throughout the lake basin. Isolates from North East Tanzania and Eastern Kenya seemed to be more closely linked with one another but still appeared to display high levels of homogeneity with other Tanzanian parasite populations.

Admixture analysis identified diverse ancestral origins of East African parasites, including three distinct ancestral populations from wider African parasite populations that offer insight into the observed East African subgroups. Isolates from Western Kenya, Central Uganda, and the Kenya region of Lake Victoria were identified to share high proportions of the same, seemingly independent, ancestral genome with one another, as well as similar cumulative proportions of ancestral genome fragments from other East African and African populations. This independent ancestry would support their distinct genetic structure when compared to other East African parasite populations and may be explained by the migration of the Luo people into Western Kenya and Uganda from South Sudan, whereas the remainder of East Africa and the Lake Victoria basin was largely settled by Bantu peoples originating from Central Africa [27–29]. Interestingly, this ancestral population appeared to donate ancestral genome chunks to not only other Eastern African populations, but also parasite populations from West Africa, Central Africa, South Central Africa, Southern Africa, and the Horn of Africa. Admixture analysis also revealed an independent ancestral population accounting for a majority of the ancestral genome of isolates from Lake Tanganyika that was also shared across East African parasite populations, as well as a large proportion of South Central African isolates (i.e. the Democratic Republic of Congo). This large proportion of shared ancestral genome fragments is likely due to the high rate of trade that occurs on Lake Tanganyika between Tanzania and the DRC.

To date, the prevailing theory concerning the evolution of *P. falciparum* infections in humans suggests a cross-species event occurring between Western Gorillas and humans in Central Africa approximately 10,000 years ago that subsequently spread via human migration routes across Africa [30, 31]. In our analysis, Central African isolates (i.e. Cameroon) were found to share high proportions of their ancestral genome with West African isolates, aligning with westward bantu population migrations from Central Africa, and donated ancestral genome fragments to all East African subpopulations [32]. West African isolates (i.e. the Gambia and Guinea) appeared to share a proportion of their genome with South Central African isolates (i.e. DRC) which may be explained by human migration linked to colonization and slavery during the French occupation of Guinea and the DRC, following its Belgian

occupation [33]. Isolates from the Tanzanian region of Lake Victoria contained a large portion of a unique ancestral genome compared to other populations, although a maximum likelihood tree suggests more samples from this region would be needed to confirm its independence from ancestry associated with Lake Tanganyika.

Analysis of identity-by-descent (IBD) was carried out as another metric of investigating the structure and selection occurring within East African parasite populations. As is common within high transmission populations undergoing intense rates of recombination, fractions of pairwise IBD were low and uneven across genome, with the exception of conserved *Plasmodium* proteins across the genome, a handful of segments including drug resistance associated loci, and genes involved in parasite invasion (e.g. PKAr). A region on chromosome 8 encompassing the drug resistance gene *Pfdhps* was identified in the top 5% of IBD for isolates from Eastern Kenya, North East Tanzania, and Lake Tanganyika (ancestral population K3), as well as the Tanzanian region of Lake Victoria (K4) and South East Tanzanian isolates alongside Southern Africa (K6). The high conservation of *Pfdhps* across East African parasite populations is likely driven by the continued use of SP as a method of IPTp within the region and is supported by the high frequency of drug resistance associated variants nearly fixed within the population. As anticipated, high IBD was observed across the genome for isolates from the Horn of Africa, re-establishing their high levels of genetic relatedness and distinct population structure from other sub-Saharan African parasite populations, providing a practical comparison for East African populations.

To identify regions undergoing recent positive selection within and across East African parasite populations, genome-wide analysis of haplotype structure was assessed according to ancestral populations due to the high degree of correlation between geographical proximity and genetic structure. This analysis identified variants undergoing selection within genes largely associated with host immune response and parasite immune evasion, including PMX in isolates from Western Kenya, Central Uganda, and the Kenyan region of Lake Victoria, as well as HSP40, CX3CL1-binding proteins, and the cAMP-dependent protein kinase regulatory subunit (PKAr) in isolates from Eastern Kenya, North East Tanzania, and Lake Tanganyika. PMX and PKAr have been characterised to play an important role as a mediators of parasite erythrocyte invasion, with PMX also linked to parasite egress, whereas HSP40 is believed to play a role in determining parasite pathogenicity [19, 34]. Cross-population analysis of positive selection revealed two regions under selection pressure with known, or suspected, drug resistance associations. Isolates from Western Kenya, Central Uganda, and the Lake Victoria region of Kenya were identified to have variants under selection within the *Pfdhps* gene, known to contribute to resistance to sulphadoxine, when compared to Central, South Central, and West African populations. Variants were also identified within the same population to be under strong

positive selection within *Pfubp1*, a gene believed to be associated with ACT treatment failures, as well as in isolates from Eastern Kenya, North East Tanzania, and Lake Tanganyika.

Overall, this study has provided an in-depth and high-resolution analysis of the genomic diversity and ancestral origins of *P. falciparum* populations from across East Africa, as well as provided a baseline analysis of parasites from the Kenya-Uganda border. Despite a handful of limitations, including limited data availability from Uganda and genetic clustering of highland epidemic zone isolates from Bungoma county, we were able to quantify changes in drug resistance markers within parasite populations along the Kenya-Uganda border and identify subpopulations with their own distinct genetic structure within East Africa.

MATERIALS AND METHODS

Study site selection

Dried blood spots were collected in 2018 from individuals residing in Bungoma county, Kenya (n = 53). Sample collection within this region was carried out as part of ongoing surveillance within the region which includes regular community *P. falciparum* prevalence surveys.

Bungoma county, situated along the border with Uganda, is defined as having both lake-endemic and highland epidemic zones based on malaria endemicity classifications. Malaria prevalence within the lake-endemic zones is seasonally linked to short (October to December) and long (March to May) rainy seasons. Transmission in the highland epidemic regions results in malaria outbreaks, rather than sustained, seasonal, transmission.

Permission to conduct this study was obtained from the Mount Kenya University Independent Ethics and Research Committee (Approval reference: P609/10/2014) and performed in accordance with relevant guidelines and regulations. Educational workshops and sensitisation meetings were carried out within communities included in this study prior to seeking community consent to study participation. Written informed consent was obtained from all study participants whose parasite DNA was used in this study.

Study population characteristics

Bungoma county is occupied predominantly by Bukusu people, one of the Luhya Bantu tribes, who rely heavily on crop and livestock farming. The main economic crops within the region are sugarcane and maize while paper mills also employ a small proportion of the population. Bungoma county borders Busia County which harbours one of the busiest borders in East Africa, offering access to and from Rwanda, Burundi, The Democratic Republic of Congo, and South Sudan [35].

Species identification

Following WHO diagnostic guidelines, malaria species identification was carried out by microscopy at Mount Kenya University by trained microscopists and confirmed at the LSHTM using established nested PCR assays and in-house species prediction software which uses k-mers in read files specific to Plasmodium species [36].

Whole genome sequencing and bioinformatic processing

P. falciparum DNA from 53 isolates (collected in 2018) was amplified using a selective whole genome amplification method with established protocols and primer sets. Following amplification, DNA was cleaned using KAPA Pure Beads the associated DNA fragment clean-up protocol. Parasite DNA was sequenced on an Illumina Novaseq 6000 platform by Eurofins Genomics in Germany with a minimum of 3.75 million paired reads per sample generated. Raw sequencing data was mapped to version 3 of the PF3D7 *P. falciparum* reference genome using default *bwa-mem* software parameters. Genomic variants were called using *samtools* and GATK software suites, including *HaplotypeCaller* and *ValidateVariants*. Variants in low quality or low coverage regions, as well as highly variable regions, were discarded from analyses. For analysis of drug resistance SNPs, samples with less than 5 read depth were not included. For genome-wide population analyses, samples with less than 70% mapping to the PF3D7 reference genome, or less a minimum coverage of 5 across 70% of the genome, were not included. Of the 53 samples that were sequenced from Bungoma county, 30 were included in population-wide analyses with wider African *P. falciparum* populations, while 33 – 38 isolates were assessed for drug resistance polymorphisms.

Genomic variants and genome-wide population analyses

To assess genome-wide population variation, pairwise genetic distance matrices were produced for each population (East Africa, N = 597; Africa, N = 363) using high-quality SNPs called from isolates within the population. These distance matrices were used to generate maximum likelihood trees with IQ-Tree and principal component analyses (PCA) [37]. Maximum likelihood trees generated with IQ-Tree were visualised and annotated using the *iTOL* web-interface [38]. Variants within genes known to be associated with drug resistance were annotated using *bcftools csq* Haplotype-aware consequence caller to identify functional variants and their inferred coding consequences [39].

The R-based package *malariaAtlas* was used to visualise *P. falciparum* incidence rates across Kenya, Tanzania, and Uganda [40]. Regional populations for cross-population analyses (e.g. Western Kenya, Lake Victoria, West Africa, Horn of Africa, South Central Africa, Central Africa, and Southern Africa) were determined using previously characterised genetic clusters and geographic proximity as well as documented human and vector migration patterns within or between regions [9]. WGS data from

isolates in outbreak clusters 1 and 2 within Bungoma county were classified as nearly identical. To prevent inaccurate population clustering due to close genetic distances, representative samples were taken from each cluster and included in region-wide population analyses.

To estimate ancestry proportions between regional populations across Africa (N = 363), continuous genetic variation across space was estimated using the R-based Tess3r package [41–43]. Admixture calculations were determined with spatial models of allele using genome-wide SNPs and geographical coordinates for the sequencing isolates included in the dataset. GPS coordinates for Bungoma isolates were recorded at the time of collection and was specified in publicly available metadata for isolates from the Pf7 database. Based on a cross-validation of 1 to 10 dimensions of eigenvalue decay, an optimum K value for the ancestral admixture coefficients was estimated to be six. Default Tess3r parameters were used, including spatial regularisation parameters ($\sigma = 1$) which attributes loss and penalty functions to be equal. The Tess3r software was run 50 times for K1 to K10, retaining the best Q matrix of ancestry coefficients for each isolate. The alternating projected least squares algorithm (APLS; method = "projected.ls") was used for this analysis and the final plots were visualised using the R-based package *ggplot2* while surfaces were interpolated using the R package *Krig* [44–46].

Complexity of infection, or the inbreeding coefficient metric (F_{ws}), was assessed using an in-house script that determines the within-host diversity of an infection in relation to the local population diversity. The risk of out-crossing or inbreeding within each infection is calculated by estimating the fixation of alleles on a scale of 0 to 1, with an $F_{ws} \geq 0.95$ considered to be indicative of a clonal infection and lower mean inbreeding coefficients associated with a high degree of panmixis and complex infections [47].

To assess connectivity and conserveness of parasites from Bungoma county with other regional parasite populations, identity-by-descent (IBD) was calculated by estimating the pairwise fraction of shared ancestry between genomic segments. Shared ancestry between genomic segments, or "IBD fractions", are inferred to be descended from recent common ancestors without any intervening recombination. Recombination was accounted for using a hidden Markov model-based approach within the *hmmIBD* software package to calculate IBD fractions [48]. Recombination within the *P. falciparum* species is estimated to be 13.5 Kb per centiMorgan (cM), or chromosomal crossover events occurring at approximately an average rate of 1% per generation.

Putative positive directional selection across the genome was assessed using the R-based *rehh* package that measures haplotype diversity metrics within (*iHS*) or between (*Rsb*) populations. The *iHS* metric, or integrated haplotype score, measures the extended haplotype homozygosity at any given

SNP along an ancestral allele, relative to a derived allele. The *Rsb* metric, or pairwise population genetic distance, is similar to the F_{st} metric but accounts for repeats between microsatellite alleles.

ACKNOWLEDGEMENTS

We wish to thank the study participants and communities. This study was conducted in part at the Joint Usage/Research Center on Tropical Disease, Institute of Tropical Medicine, Nagasaki University, Japan.

AUTHORS CONTRIBUTIONS

AK, KK, SC, JG and TGC conceived and designed the study; HW and JG contributed parasite DNA for sequencing; HW and JG provided biological materials and data. AO and SC coordinated the sequencing of samples; AO and EM performed the bioinformatic and statistical analysis, under the supervision of SC and TGC; AO wrote the first draft of the manuscript, and the final version included edits from all authors. The final manuscript was read and approved by all authors.

FUNDING STATEMENT

AO is supported by a Nagasaki University—LSHTM PhD studentship funded by the WISE programme of MEXT. SC is funded by the Medical Research Council UK (Grant No. MR/M01360X/1) and BBSRC UK (BB/R013063/1). AK received support from JSPS KAKENHI (Grant Nos. JP18KK0248 and JP19H01080) and JICA/AMED joint research project (SATREPS) (Grant No. 20JM0110020H0002. JG received support from a Tackling Infectious Burden in Africa (TIBA) fellowship, the African Academy of Sciences, and the Japan Society for Promotion of Sciences. TGC is supported by the Medical Research Council UK (Grant Nos. MR/K000551/1, MR/M01360X/1, MR/N010469/1, MR/R020973/1) and BBSRC (BB/R013063/1). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

CONFLICTS OF INTEREST

The authors declare no competing interests.

DATA AVAILABILITY

Public accession numbers for raw sequence data analysed are contained in SRA studies ERP000190 and ERP000199, as well as being accessible from the Pf3k project website

(https://www.malariagen.net/apps/pf7/). Bungoma county raw sequences are available the ENA (project accession number not yet released).

REFERENCES

1. Geneva: World Health Organization (2022) World malaria report 2022. Licence: CC BY-NC-SA 3.0 IGO

2. Weiss DJ, Bertozzi-Villa A, Rumisha SF, et al (2021) Indirect effects of the COVID-19 pandemic on malaria intervention coverage, morbidity, and mortality in Africa: a geospatial modelling analysis. The Lancet Infectious Diseases 21:59–69

3. Antinori S, Galimberti L, Milazzo L, Corbellino M (2012) Biology of Human Malaria Plasmodia Including Plasmodium Knowlesi. Mediterr J Hematol Infect Dis. https://doi.org/10.4084/MJHID.2012.013

4. Kagaya W, Gitaka J, Chan CW, Kongere J, Idris ZM, Deng C, Kaneko A (2019) Malaria resurgence after significant reduction by mass drug administration on Ngodhe Island, Kenya. Sci Rep 9:1–11

5. Osborne A, Manko E, Takeda M, et al (2021) Characterizing the genomic variation and population dynamics of Plasmodium falciparum malaria parasites in and around Lake Victoria, Kenya. Sci Rep 11:19809

6. Mulenge FM, Hunja CW, Magiri E, Culleton R, Kaneko A, Aman RA (2016) Genetic Diversity and Population Structure of Plasmodium falciparum in Lake Victoria Islands, A Region of Intense Transmission. Am J Trop Med Hyg 95:1077–1085

7. Minakawa N, Dida GO, Sonye GO, Futami K, Njenga SM (2012) Malaria Vectors in Lake Victoria and Adjacent Habitats in Western Kenya. PLoS One. https://doi.org/10.1371/journal.pone.0032725

8. Balikagala B, Fukuda N, Ikeda M, et al (2021) Evidence of Artemisinin-Resistant Malaria in Africa. New England Journal of Medicine 385:1163–1171

9. Amambua-Ngwa A, Amenga-Etego L, Kamau E, et al (2019) Major subpopulations of Plasmodium falciparum in sub-Saharan Africa. Science 365:813–816

10. Agaba BB, Anderson K, Gresty K, et al (2021) Genetic diversity and genetic relatedness in Plasmodium falciparum parasite population in individuals with uncomplicated malaria based on microsatellite typing in Eastern and Western regions of Uganda, 2019–2020. Malaria Journal 20:242

11. Nderu D, Kimani F, Karanja E, et al (2019) Genetic diversity and population structure of Plasmodium falciparum in Kenyan–Ugandan border areas. Tropical Medicine & International Health 24:647–656

12. Gatei W, Gimnig JE, Hawley W, et al (2015) Genetic diversity of Plasmodium falciparum parasite by microsatellite markers after scale-up of insecticide-treated bed nets in western Kenya. Malaria Journal 14:495

13. Moser KA, Madebe RA, Aydemir O, et al (2021) Describing the current status of Plasmodium falciparum population structure and drug resistance within mainland Tanzania using molecular inversion probes. Mol Ecol 30:100–113

14. Daniels RF, Schaffner SF, Wenger EA, et al (2015) Modeling malaria genomics reveals transmission decline and rebound in Senegal. Proc Natl Acad Sci USA 112:7067–7072

15. MalariaGEN, Abdel Hamid MM, Abdelraheem MH, et al (2023) Pf7: an open dataset of Plasmodium falciparum genome variation in 20,000 worldwide samples. Wellcome Open Res 8:22

16. Zhong D, Afrane Y, Githeko A, Yang Z, Cui L, Menge DM, Temu EA, Yan G (2007) Plasmodium falciparum genetic diversity in western Kenya highlands. Am J Trop Med Hyg 77:1043–1050

17. Calçada C, Silva M, Baptista V, Thathy V, Silva-Pedrosa R, Granja D, Ferreira PE, Gil JP, Fidock DA, Veiga MI (2020) Expansion of a Specific Plasmodium falciparum PfMDR1 Haplotype in Southeast Asia with Increased Substrate Transport. mBio 11:e02093-20

18. Turkiewicz A, Manko E, Sutherland CJ, Benavente ED, Campino S, Clark TG (2020) Genetic diversity of the Plasmodium falciparum GTP-cyclohydrolase 1, dihydrofolate reductase and dihydropteroate synthetase genes reveals new insights into sulfadoxine-pyrimethamine antimalarial drug resistance. PLOS Genetics 16:e1009268

19. Diehl M, Roling L, Rohland L, et al (2021) Co-chaperone involvement in knob biogenesis implicates host-derived chaperones in malaria virulence. PLOS Pathogens 17:e1009969

20. Henriques G, Hallett RL, Beshir KB, et al (2014) Directional selection at the pfmdr1, pfcrt, pfubp1, and pfap2mu loci of Plasmodium falciparum in Kenyan children treated with ACT. J Infect Dis 210:2001–2008

21. World Health Organization (2019) Compendium of WHO malaria guidance: prevention, diagnosis, treatment, surveillance and elimination. World Health Organization, Geneva

22. GOODMAN C, BRIEGER W, UNWIN A, MILLS A, MEEK S, GREER G (2007) MEDICINE SELLERS AND MALARIA TREATMENT IN SUB-SAHARAN AFRICA. Am J Trop Med Hyg 77:203–218

23. Frosch AEP, Laufer MK, Mathanga DP, Takala-Harrison S, Skarbinski J, Claassen CW, Dzinjalamala FK, Plowe CV (2014) Return of widespread chloroquine-sensitive Plasmodium falciparum to Malawi. J Infect Dis 210:1110–1114

24. Mwanza S, Joshi S, Nambozi M, Chileshe J, Malunga P, Kabuya J-BB, Hachizovu S, Manyando C, Mulenga M, Laufer M (2016) The return of chloroquine-susceptible Plasmodium falciparum malaria in Zambia. Malaria Journal 15:584

25. Kublin JG, Cortese JF, Njunju EM, Mukadam RAG, Wirima JJ, Kazembe PN, Djimdé AA, Kouriba B, Taylor TE, Plowe CV (2003) Reemergence of chloroquine-sensitive Plasmodium falciparum malaria after cessation of chloroquine use in Malawi. J Infect Dis 187:1870–1875

26. Geneva: World Health Organization (2020) Report on antimalarial drug efficacy, resistance and response: 10 years of surveillance (2010– 2019). Licence: CC BY-NC-SA 3.0 IGO

27. Tishkoff SA, Reed FA, Friedlaender FR, et al (2009) The Genetic Structure and History of Africans and African Americans. Science 324:1035–1044

28. Kenya National Bureau of Statistics (2019) 2019 Kenya Population and Housing Census Results.

29. Vansina J (1995) New Linguistic Evidence and 'The Bantu Expansion.' The Journal of African History 36:173–195

30. Loy DE, Liu W, Li Y, Learn GH, Plenderleith LJ, Sundararaman SA, Sharp PM, Hahn BH (2017) Out of Africa: origins and evolution of the human malaria parasites Plasmodium falciparum and Plasmodium vivax. Int J Parasitol 47:87–97

31. Liu W, Li Y, Learn GH, et al (2010) Origin of the human malaria parasite Plasmodium falciparum in gorillas. Nature 467:420–425

32. Patin E, Lopez M, Grollemund R, et al (2017) Dispersals and genetic adaptation of Bantu-speaking populations in Africa and North America. Science 356:543–546

33. Ginio R, Sessions J (2016) French Colonial Rule. African Studies. https://doi.org/10.1093/obo/9780199846733-0029

34. Nasamu AS, Glushakova S, Russo I, et al (2017) Plasmepsins IX and X are essential and druggable mediators of malaria parasite egress and invasion. Science 358:518–522

35. Chiuya T, Villinger J, Falzon LC, Alumasa L, Amanya F, Bastos ADS, Fèvre EM, Masiga DK (2022) Molecular screening reveals non-uniform malaria transmission in western Kenya and absence of Rickettsia africae and selected arboviruses in hospital patients. Malaria Journal 21:268

36. Mangold KA, Manson RU, Koay ESC, Stephens L, Regner M, Thomson RB, Peterson LR, Kaul KL (2005) Real-Time PCR for Detection and Identification of Plasmodium spp. J Clin Microbiol 43:2435–2440

37. Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, von Haeseler A, Lanfear R (2020) IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. Molecular Biology and Evolution 37:1530–1534

38. Letunic I, Bork P (2007) Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. Bioinformatics 23:127–128

39. Danecek P, McCarthy SA (2017) BCFtools/csq: haplotype-aware variant consequences. Bioinformatics 33:2037–2039

40. Pfeffer DA, Lucas TCD, May D, et al (2018) malariaAtlas: an R interface to global malariometric data hosted by the Malaria Atlas Project. Malaria Journal 17:352

41. Caye K, Deist TM, Martins H, Michel O, François O (2016) TESS3: fast inference of spatial population structure and genome scans for selection. Molecular Ecology Resources 16:540–548

42. Caye K, Jay F, Michel O, François O (2018) Fast inference of individual admixture coefficients using geographic data. The Annals of Applied Statistics 12:586–608

43. Martins H, Caye K, Luu K, Blum MGB, François O (2016) Identifying outlier loci in admixed and in continuous populations using ancestral population differentiation statistics. Mol Ecol 25:5029–5042

44. Wickham H (2016) ggplot2: Elegant Graphics for Data Analysis, 2nd ed. 2016. https://doi.org/10.1007/978-3-319-24277-4

45. Hastie T, Tibshirani R (1986) Generalized Additive Models. Statistical Science 1:297–310

46. Cressie NAC (1993) Statistics for spatial data, Rev. ed. Wiley, New York

47. Manske M, Miotto O, Campino S, et al (2012) Analysis of Plasmodium falciparum diversity in natural infections by deep sequencing. Nature 487:375–379

48. hmmIBD: software to infer pairwise identity by descent between haploid genotypes | Malaria Journal | Full Text. https://malariajournal.biomedcentral.com/articles/10.1186/s12936-018-2349-7. Accessed 28 Apr 2021

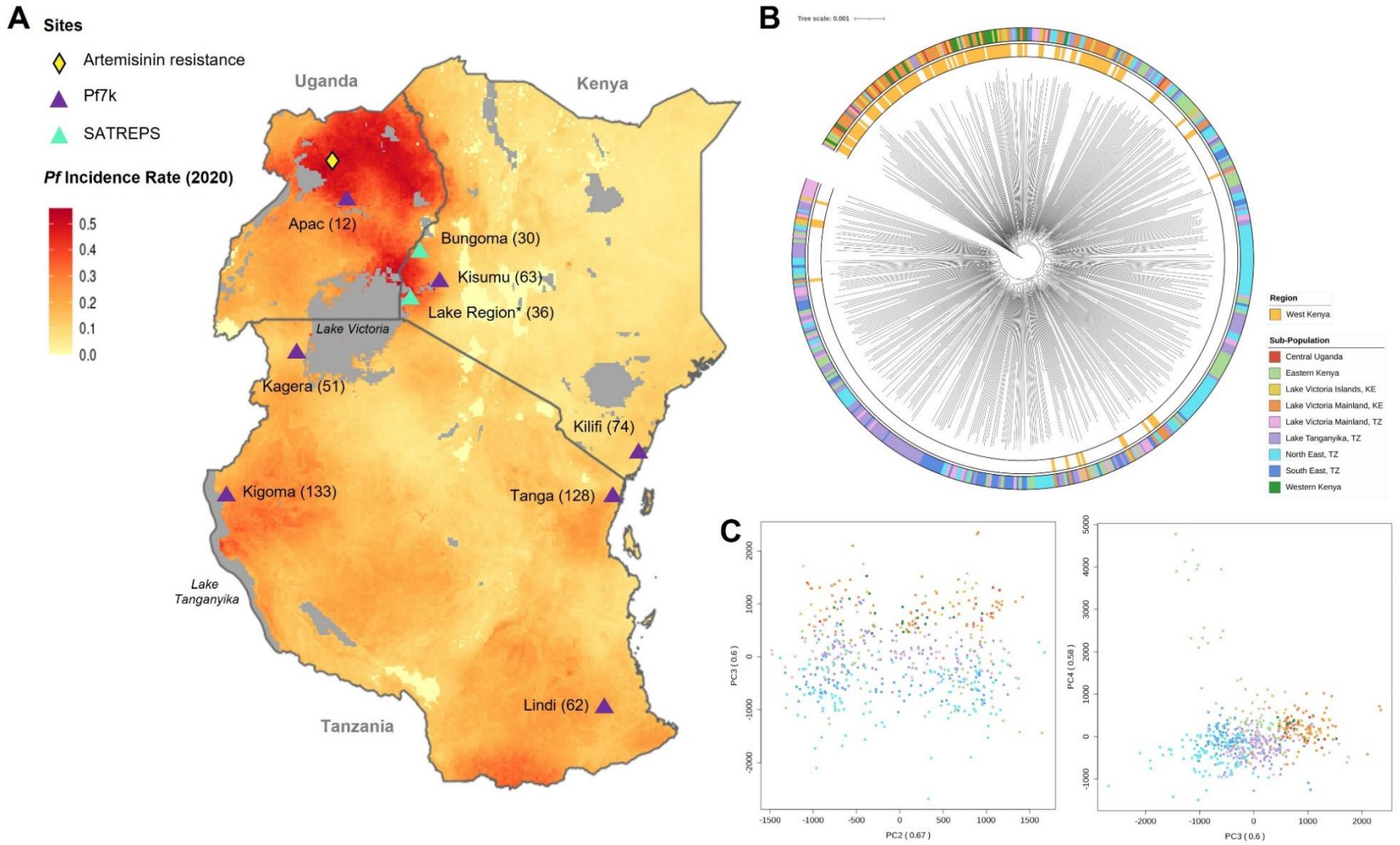
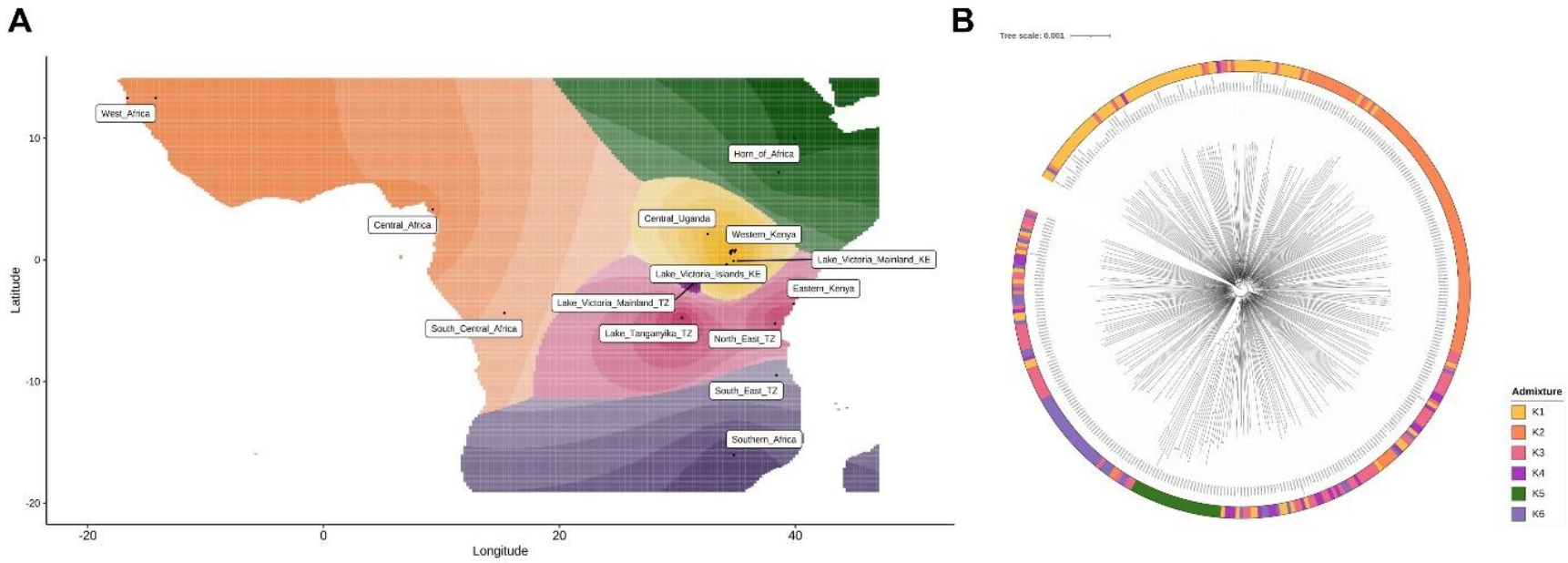


Figure 1: ***P. falciparum*** **isolates from East Africa form subpopulations with distinct genetic structure. Maximum likelihood tree and principal component analysis (PCA) generated from pairwise genetic distance matrix of 710,552 high-quality genome-wide SNPs from 587** ***P. falciparum*** **isolates.** (**A**) Heatmap of *P. falciparum* incidence rates in 2020 across Kenya, Tanzania and Uganda with sampling and artemisinin resistance sites

annotated (*malariaAtlas* R-software). (**B**) a maximum likelihood tree for 587 isolates from Central Uganda, Eastern Kenya, Lake Victoria, Lake Tanganyika, North East Tanzania, South East Tanzania, and Western Kenya. (**C**) PCA of isolates with principal component (PC) 2 and PC3. (**D**) PCA of isolates with PC3 and PC4.



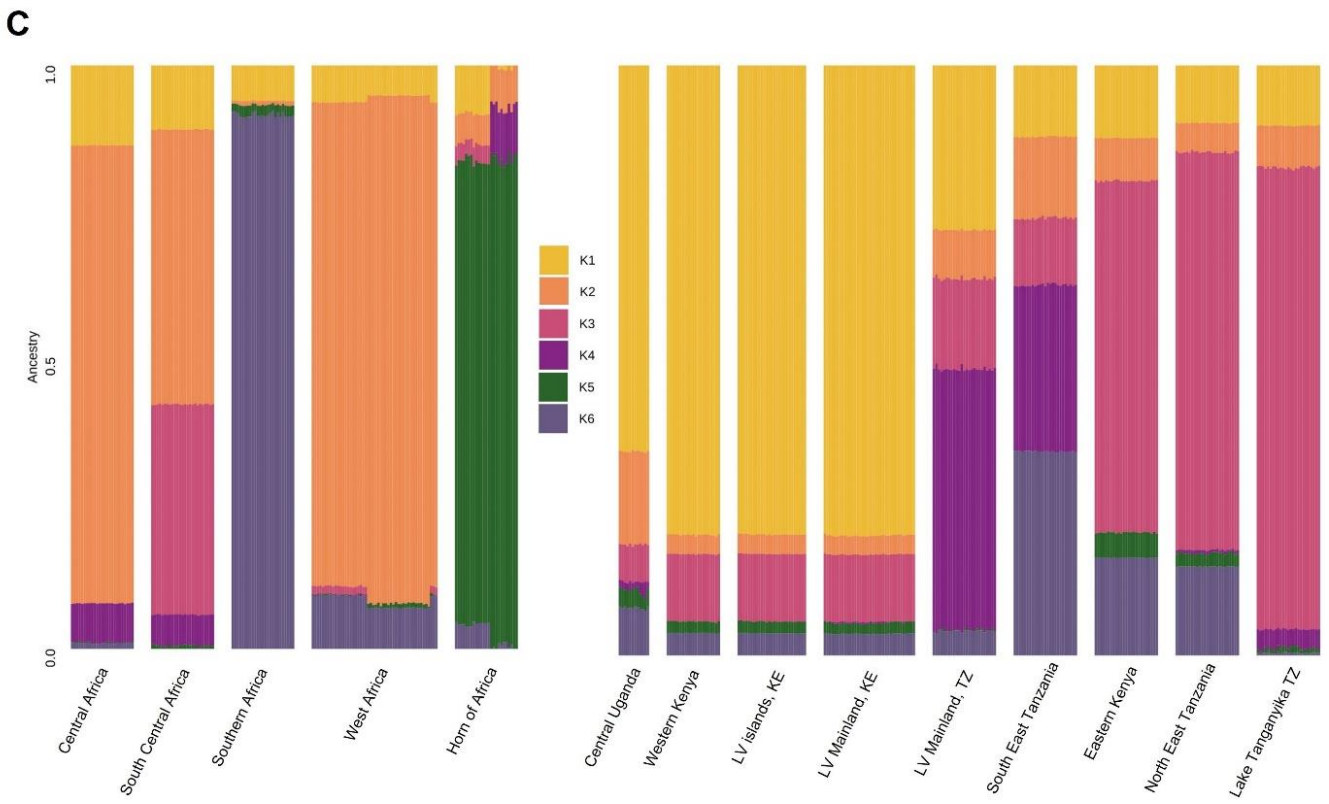


Figure 2: Genome-wide ancestral admixture analysis of East African *P. falciparum* subpopulations alongside regional parasite populations from Across Africa. (**A**) Geographic map of ancestry coefficients where *K* is estimated to be 6 ancestral populations within Africa. (**B**) Maximum likelihood tree of 363 isolates (640,596 genome-wide SNPs) coloured according to their maximum K proportion*. (**C**) Ancestry proportions per isolate (rows) within each subpopulation (columns).

*K1 = Western Kenya, Lake Victoria Kenya, Central Uganda; K2 = Central Africa, South Central Africa, West Africa; K3 = Eastern Kenya, North East Tanzania, Lake Tanganyika, TZ; K4 = Lake Victoria Tanzania; K5 = Horn of Africa; K6 = Southern Africa and South East Tanzania.

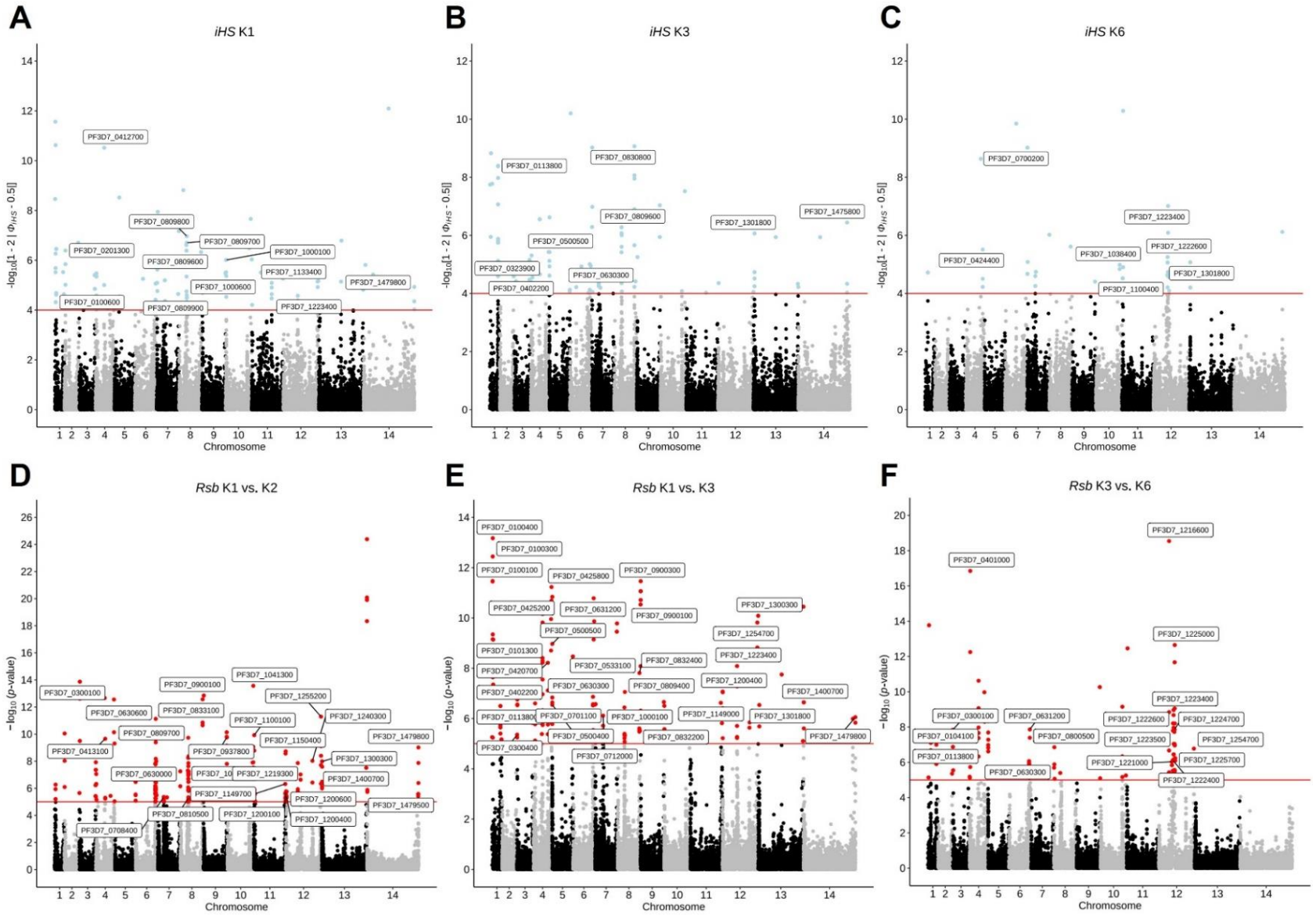
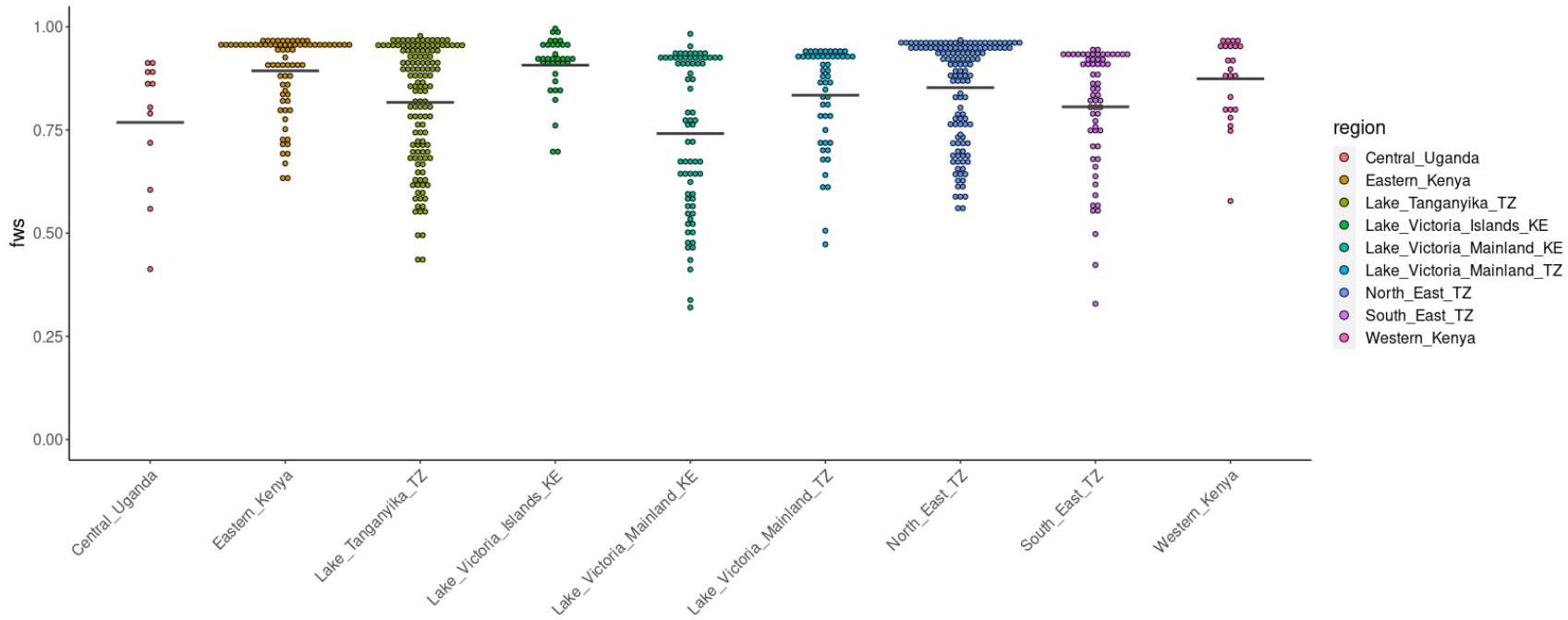

A
iHS K1
B
iHS K3
C
iHS K6
D
Rsb K1 vs. K2
E
Rsb K1 vs. K3
F
Rsb K3 vs. K6
Chromosome
−log10 (p-value)

Figure 3: Signatures of positive selection in East African subpopulations according to their maximum proportion ancestral population* (K-value). The genome-wide haplotype structure of isolates was analysed to locate regions of high local homozygosity relative to neutral expectation. (**A,B,C**) Variants undergoing within-population selection in the inferred K1, K3, and K6 populations with an *iHS* >4.0 ((-log10[1 – 2 | ΦiHS – 0.5 |]) > 4.0). (**D,E,F**) Cross-population analysis of variants under high positive directional selection (-log10(p-value) > 5).

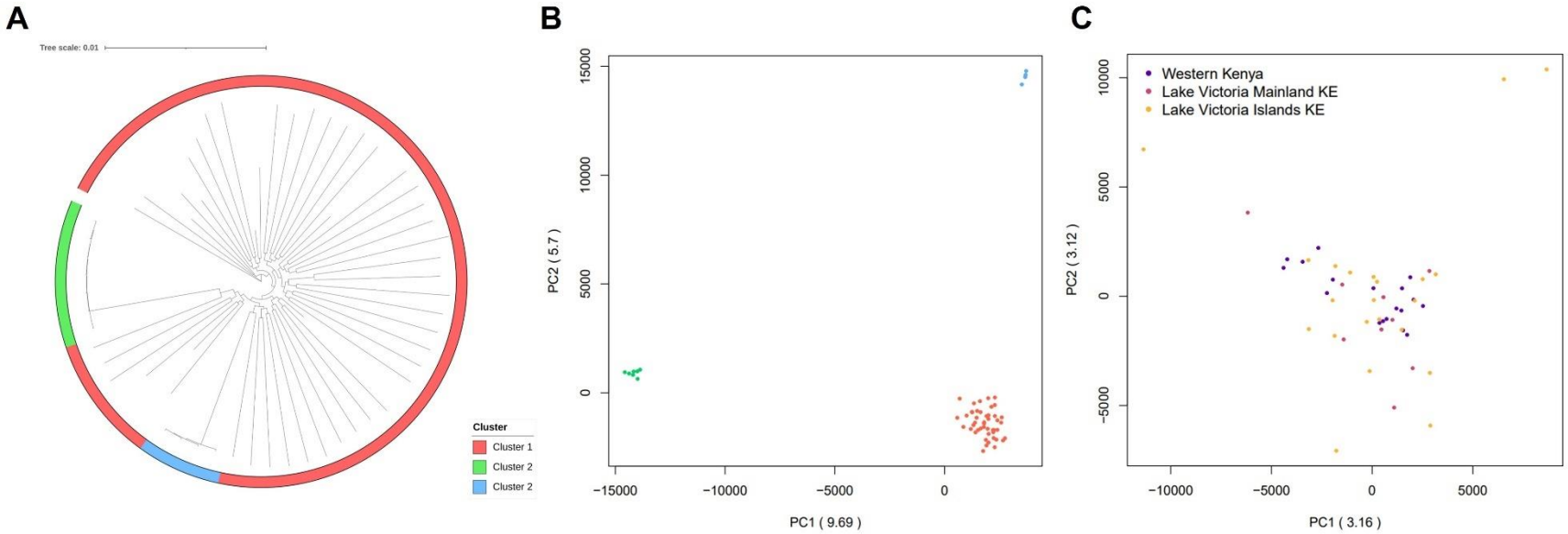
*K1 = Western Kenya, Lake Victoria Kenya, Central Uganda; K2 = Central Africa, South Central Africa, West Africa; K3 = Eastern Kenya, North East Tanzania, Lake Tanganyika, TZ; K6 = Southern Africa and South East Tanzania.

SUPPLEMENTARY FIGURES

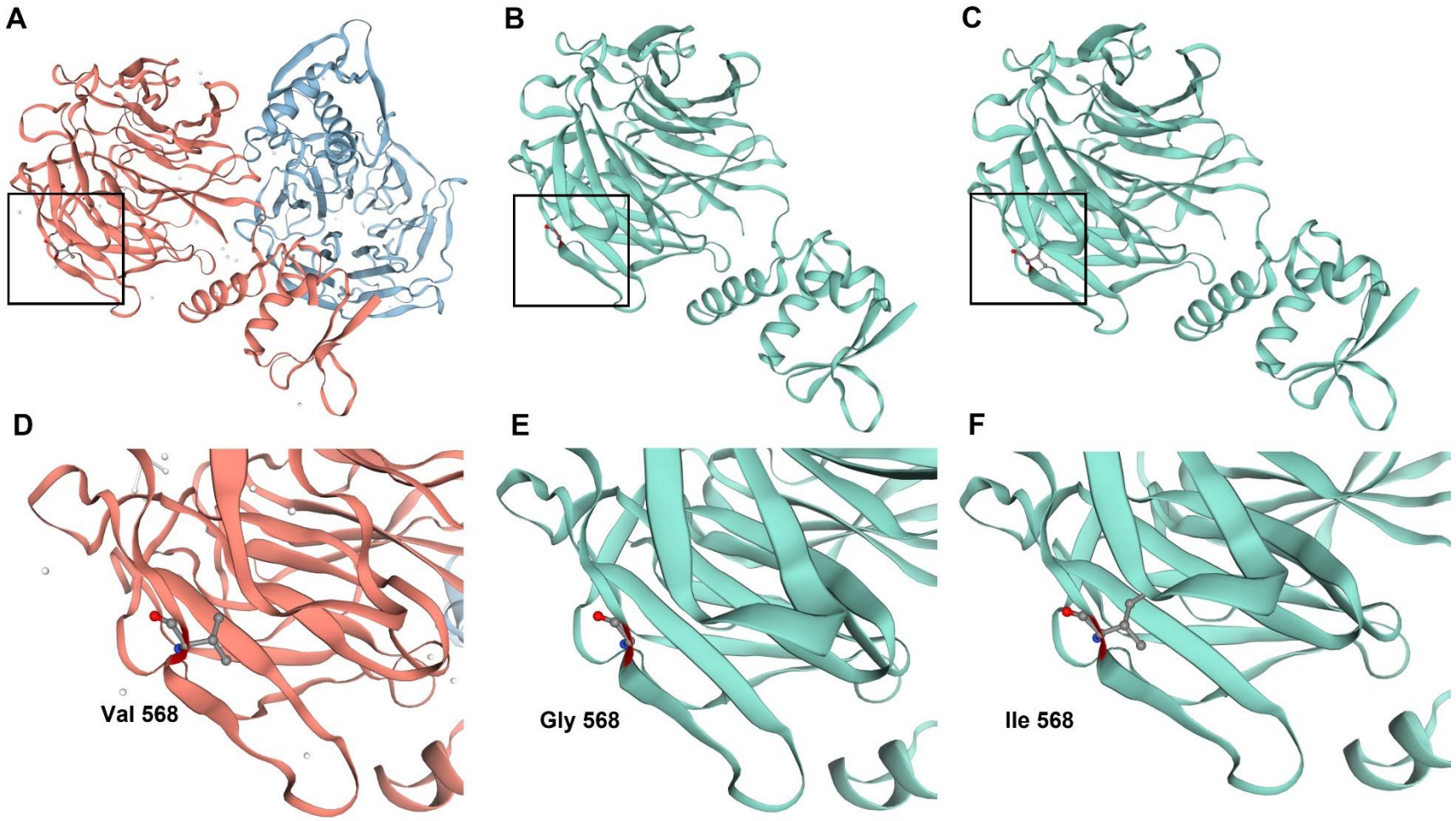
Supplementary Figure S1. Multiplicity of infection across sampling sites included in the East Africa dataset (N = 587) assessed using the inbreeding coefficient (F_{ws} metric).


fws
1.00
0.75
0.50
0.25
0.00
Central_Uganda
Eastern_Kenya
Lake_Tanganyika_TZ
Lake_Victoria_Islands_KE
Lake_Victoria_Mainland_KE
Lake_Victoria_Mainland_TZ
North_East_TZ
South_East_TZ
Western_Kenya
region
Central_Uganda
Eastern_Kenya
Lake_Tanganyika_TZ
Lake_Victoria_Islands_KE
Lake_Victoria_Mainland_KE
Lake_Victoria_Mainland_TZ
North_East_TZ
South_East_TZ
Western_Kenya

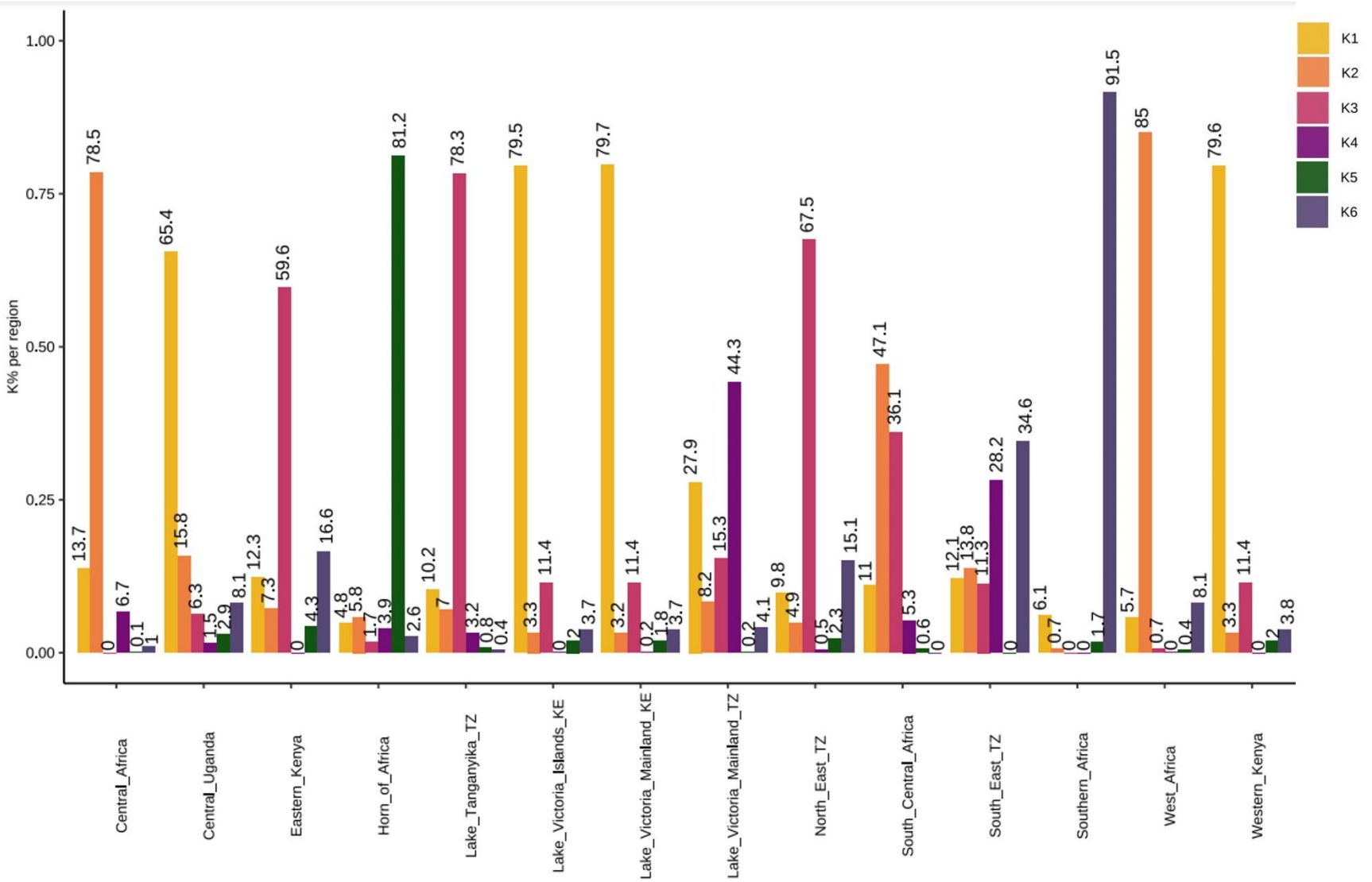
Supplementary Figure S2. Population structure of *P. falciparum* isolates from Western Kenya and Lake Victoria, Kenya. (**A**) Maximum likelihood tree of 74 isolates from Western Kenya (N = 30) and Lake Victoria, Kenya (N = 46) with 284,667 genome-wide SNPs coloured by cluster. (**B**) PCA of genetic pairwise matrix used to generate the maximum likelihood tree coloured by cluster. (**C**) PCA of genetic pairwise matrix retaining all isolates from cluster 1 and a single isolate from cluster 2 and cluster 3 coloured by region (N = 67).



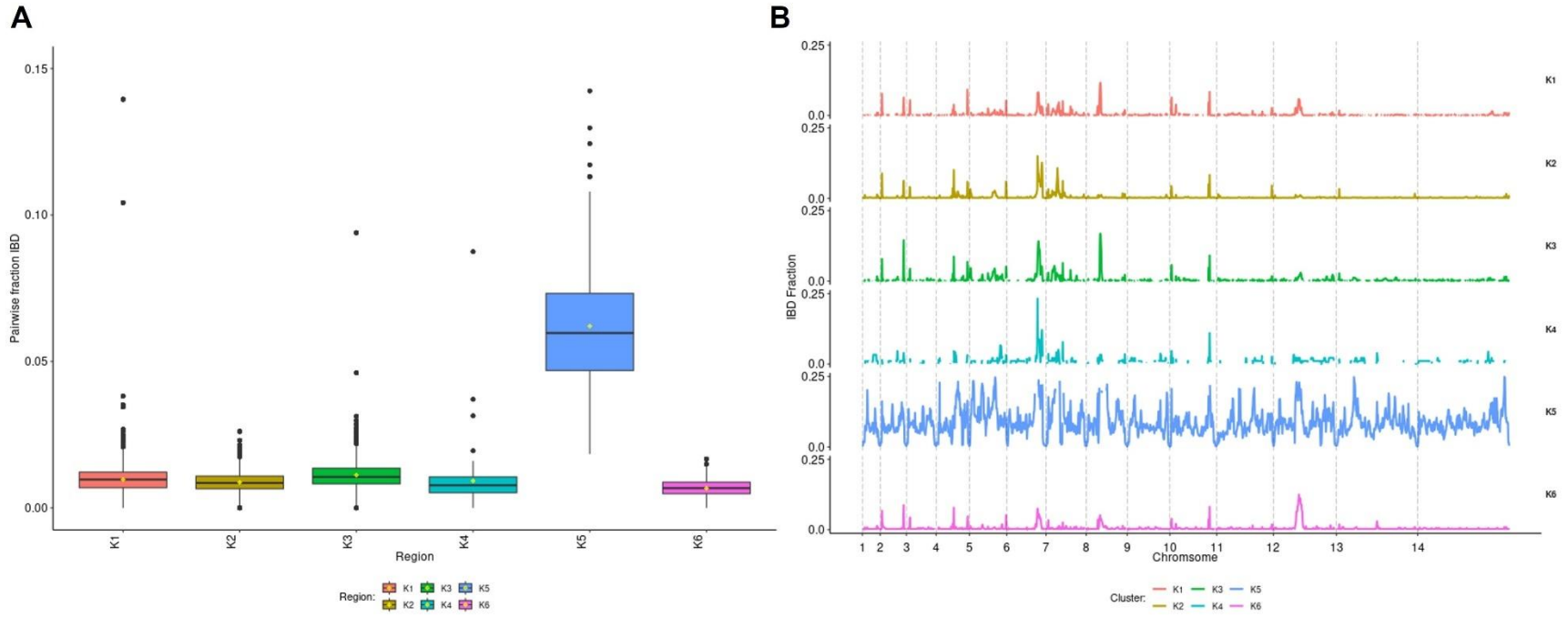
Supplementary Figure S3: An in-silico model of the protein structure of *P. falciparum* Kelch 13 protein highlighting codon 568, a WHO candidate for reduced susceptibility to artemisinin. (**A,D**) Wild-type codon Val568. (**B,E**) Codon 568 Gly [WHO candidate]. (**C,F**) Codon 568 Ile [identified in an isolate from Bungoma county, Kenya].



Supplementary Figure S4. Cumulative genome-wide ancestral admixture proportions for *P. falciparum* populations across Africa and subpopulations within East Africa. Cumulative percentages of ancestry where K is estimated to be 6.



Supplementary Figure S5. Identity-by-descent (IBD) fractions in *P. falciparum* isolates from 6 estimated ancestral populations* within Africa. (**A**) IBD fractions across the genome by ancestral population. (**B**) IBD fractions along each chromosome by ancestral population.



*K1 = Western Kenya, Lake Victoria Kenya, Central Uganda; K2 = Central Africa, South Central Africa, West Africa; K3 = Eastern Kenya, North East Tanzania, Lake Tanganyika, TZ; K4 = Lake Victoria Tanzania; K5 = Horn of Africa; K6 = Southern Africa and South East Tanzania.

SUPPLEMENTARY TABLES

Supplementary Table S1. Summary data for *P. falciparum* isolates included within East African genome-wide population analyses to characterise population dynamics and identify genetic markers of interest within isolates collected from Western Kenya.

Country	Site	Region	Year	Samples (N)
Kenya	Bungoma county	Western Kenya	2018	30
	Suba district	Lake Victoria mainland	2020	10
	Mfangano Island	Lake Victoria islands	2015, 2020	24
	Ngodhe Island	Lake Victoria islands	2016, 2020	12
	Kilifi	Eastern Kenya	2012	74
	Kisumu	Lake Victoria mainland	2014	63
Tanzania	Kagera	Lake Victoria mainland	2013	51
	Kigoma	Lake Tanganyika	2014	133
	Lindi	Southeast Tanzania	2013	62
	Tanga	Northeast Tanzania	2014	128
Uganda	Apac	Central Uganda	2010	12

Supplementary Table S2. Summary data for *P. falciparum* isolates included within African genome-wide population analyses.

Region	Country	Year	Samples (N)
East Africa	Kenya (Western Kenya)	2018	21
	Kenya	2012, 2014, 2015, 2016, 2020	86
	Tanzania	2013, 2014	99
	Uganda	2010	12
West Africa	Guinea	2011	22
	Gambia	2014	25
Horn of Africa	Ethiopia	2013	25
South Central Africa	Democratic Republic of Congo	2014	25
Central Africa	Cameroon	2013	25
Southern Africa	Malawi	2011	25

Supplementary Table S3. Non-synonymous single nucleotide polymorphisms (SNPs) in known drug resistance genes. Known resistance-conferring variants highlighted in **bold.**

Gene	Position	Ref	Alt	Variant	Western Kenya MAF	N (DP >5)	*Maximum F_{st}	East Africa MAF (n = 460)	LV Kenya* MAF (n = 109)	West Africa MAF (n = 50)
Pfcrt	403291	G	T	D24Y	19.4	36	0.00327	7.7	7.3	11.0
	403625	A	C	**K76T**	2.8	36	0.00497	14.1	15.1	35.0
	403715	T	C	F106S	5.5	36	0	0	0.5	0
	404407	G	T	A220S	8.1	34	0.00139	14.5	15.1	35.7
	404836	C	G	**Q271E**	2.9	34	0.00770	14	14.8	35.7
	405362	A	G	N326S	0	34	0	0	0	0
	405600	T	C	I356T	0	34	0	0	0	22.0
	405838	G	T	**R371I**	2.9	34	0.000188	14.8	15.3	35.0
	406295	T	C	V418A	6.1	33	0.000051	0.1	1.0	0
Pfdhfr	748239	A	T	**N51I**	100	33	0.002442	95.1	99.1	96.0
	748262	T	C	**C59R**	100	33	0.007159	86.8	97.2	88.0
	748410	G	A	**S108N**	100	34	0.000207	99.6	99.1	100
	748577	A	T	**I164L**	2.16	38	0.000575	1.2	1.0	2.0
Pfdhps	549681	T	C	**S436H**	21.6	37	0.000059	0	0	0
	549685	G	C	**G437A**	100	37	0.005389	89.8	94.5	92.0
	549993	A	G	**K540E**	97.1	35	0.003367	89.7	95.0	49.0
	550117	C	G	**A581G**	0	37	0.007024	12.9	26.1	5.0
Pfmdr1	958145	A	T	**N86Y**	0	33	0.002974	5.9	3.3	17.0
	958440	A	T	**Y184F**	54.5	33	0.000059	37.7	44.4	48.0
	958484	A	T	T199S	3.1	32	0.010912	0.5	0.5	0
	960702	T	A	F938Y	2.9	34	0.000318	7.1	1.8	1.0
	961625	G	T	**D1246Y**	10.8	37	0.000204	5.0	4.1	5.0
PfK13	1725266	C	A	A578S	5.4	37	0.000365	0.1	0	0
	1725296	C	T	**V568I**	2.7	37	0.010091	0	0	0

	1725752	T	C	I416V	2.8	36	0.000462	0	0	0
	1725848	G	A	H384Y	2.8	36	0.000462	0	0	0
	1726015	T	A	Y328F	2.8	36	0.000462	0	0	0
	1726226	A	T	L258M	5.4	37	0.000361	0	0	0
	1726234	C	T	R255K	2.6	38	0.002195	1.5	1.0	0
	1726432	T	G	K189T	28.6	35	0.012855	16.7	20.2	13.0

Supplementary Table S4. Fraction of pairwise identity-by-descent (IBD) across the genome.

Admixture (K value)	Region	Mean IBD	Range
K1	Western Kenya Lake Victoria islands, KE Lake Victoria mainland, KE Central Uganda	0.009719	0 – 0.13946
K2	Central Africa South Central Africa West Africa	0.0103493	0 – 0.93958
K3	Eastern Kenya North East Tanzania Lake Tanganyika, TZ	0.0112058	0 – 0.09395
K4	Lake Victoria mainland, TZ	0.0092520	0 – 0.08748
K5	Horn of Africa	0.0849156	0 – 0.89465
K6	South East Tanzania Southern Africa	0.0093663	0 – 0.87062

Supplementary Table S5. Top 5% of identity-by-descent (IBD) regions in East African *P. falciparum* populations by ancestral population*.

*K1 = Western Kenya, Lake Victoria Kenya, Central Uganda; K2 = Central Africa, South Central Africa, West Africa; K3 = Eastern Kenya, North East Tanzania, Lake Tanganyika, TZ; K4 = Lake Victoria Tanzania; K5 = Horn of Africa; K6 = Southern Africa and South East Tanzania.

Population	chr	start	end	fraction	gene product	gene name
K1	3	120001	130000	0.041364	cytoadherence linked asexual protein 3.2	CLAG3.2
	3	130001	140000	0.054181	erythrocyte membrane protein 1 (PfEMP1), pseudogene	N/A
	3	120001	130000	0.041364		
	3	130001	140000	0.054181	cytoadherence linked asexual protein 3.1	CLAG3.1
	5	670001	680000	0.024963	40S ribosomal protein S11	RPS11
	5	680001	690000	0.019421	major facilitator superfamily domain-containing protein, putative	MFS1
	5	900001	910000	0.018534	transcription initiation factor TFIID subunit 10, putative	TAF10
	6	1110001	1120000	0.080669	acetyl-CoA synthetase, putative	ACS
	6	1240001	1250000	0.030021	SET domain protein, putative	SET1
	7	390001	400000	0.020566	heat shock protein 110	HSP110c
	7	430001	440000	0.038102	prodrug activation and resistance esterase	PARE
	7	460001	470000	0.04673	conserved Plasmodium protein, unknown function	N/A
	7	470001	480000	0.027067		
	7	470001	480000	0.027067	60S ribosomal protein L34	RPL34
	8	410001	420000	0.028246	plasmepsin X	PMX
	8	510001	520000	0.116119	JmjC domain-containing protein, putative	JmjC1
	8	500001	510000	0.114737		
	10	210001	220000	0.02836	conserved Plasmodium protein, unknown function	N/A
	12	910001	920000	0.058212	AP2 domain transcription factor AP2-G	AP2-G
	12	900001	910000	0.053445		
	12	920001	930000	0.05774	cAMP-dependent protein kinase regulatory subunit	PKAr
	12	970001	980000	0.019743	GTP cyclohydrolase 1	GCH1
	12	990001	1000000	0.027348	polyadenylate-binding protein 1, putative	PABP1
	12	980001	990000	0.018635		

	12	990001	1000000	0.027348	histone chaperone ASF1, putative	ASF1
	13	100001	110000	0.017705	CX3CL1-binding protein 2	CBP2
K2	1	530001	540000	0.013283	CX3CL1-binding protein 1	CBP1
	3	120001	130000	0.040218	cytoadherence linked asexual protein 3.2	CLAG3.2
	3	120001	130000	0.040218	erythrocyte membrane protein 1 (PfEMP1), pseudogene	N/A
	3	130001	140000	0.037887		
	3	130001	140000	0.037887	cytoadherence linked asexual protein 3.1	CLAG3.1
	4	610001	620000	0.013832	26S protease regulatory subunit 6B, putative	RPT3
	4	630001	640000	0.018197	structural maintenance of chromosomes protein 3	SMC3
	4	810001	820000	0.015908	eukaryotic translation initiation factor 3 subunit M, putative	EIF3M
	4	810001	820000	0.015908	conserved Plasmodium protein, unknown function	N/A
	4	820001	830000	0.014191		
	5	60001	70000	0.019001	parasite-infected erythrocyte surface protein	PIESP2
	5	60001	70000	0.019001	skeleton-binding protein 1	SBP1
	5	70001	80000	0.015051		
	5	850001	860000	0.021873	adenosylhomocysteinase	SAHH
	5	880001	890000	0.020505	ATP-dependent RNA helicase DDX1, putative	DDX1
	5	900001	910000	0.023041	transcription initiation factor TFIID subunit 10, putative	TAF10
	5	940001	950000	0.020153	iron-sulfur cluster assembly protein SufA	SufA
	5	950001	960000	0.018212	multidrug resistance protein 1	MDR1
	6	1060001	1070000	0.01395	3-oxoacyl-acyl-carrier protein synthase I/II	FabB/FabF
	6	1070001	1080000	0.014911	conserved Plasmodium protein, unknown function	N/A
	6	1060001	1070000	0.01395		
	6	1080001	1090000	0.018806	pyruvate kinase	PyrK
	6	1070001	1080000	0.014911		
	6	1110001	1120000	0.149757	acetyl-CoA synthetase, putative	ACS
	6	1200001	1210000	0.066279	phospholipase, putative	PL
	6	1210001	1220000	0.043009	polyadenylate-binding protein 3, putative	PABP3
	6	1210001	1220000	0.043009	amino acid transporter AAT1	AAT1
	6	1240001	1250000	0.064681	SET domain protein, putative	SET1

	6	1230001	1240000	0.054759		
	6	1220001	1230000	0.041496		
	7	220001	230000	0.027177	HECT-type E3 ubiquitin ligase UT	UT
	7	210001	220000	0.016629		
	7	300001	310000	0.022052	conserved Plasmodium protein, unknown function	N/A
	7	310001	320000	0.018817		
	7	320001	330000	0.019039		
	7	330001	340000	0.01607		
	7	330001	340000	0.01607	rhoptry-associated membrane antigen	RAMA
	7	370001	380000	0.024128	DNA-directed RNA polymerases I, II, and III subunit RPABC5, putative	RPB10
	7	390001	400000	0.059037	heat shock protein 110	HSP110c
	7	400001	410000	0.076723	chloroquine resistance transporter	CRT
	7	430001	440000	0.065494	prodrug activation and resistance esterase	PARE
	7	460001	470000	0.028383	conserved Plasmodium protein, unknown function	N/A
	7	470001	480000	0.018868		
	7	470001	480000	0.018868	60S ribosomal protein L34	RPL34
	7	480001	490000	0.013788	conserved Plasmodium protein, unknown function	N/A
	9	1410001	1420000	0.012862	cytoadherence linked asexual protein 9	CLAG9
	13	100001	110000	0.033137	CX3CL1-binding protein 2	CBP2
K3	3	120001	130000	0.039811	cytoadherence linked asexual protein 3.2	CLAG3.2
	3	130001	140000	0.044179	erythrocyte membrane protein 1 (PfEMP1), pseudogene	N/A
	3	120001	130000	0.039811		
	3	130001	140000	0.044179	cytoadherence linked asexual protein 3.1	CLAG3.1
	5	60001	70000	0.027917	parasite-infected erythrocyte surface protein	PIESP2
	5	60001	70000	0.027917	skeleton-binding protein 1	SBP1
	5	670001	680000	0.028477	40S ribosomal protein S11	RPS11
	5	850001	860000	0.034266	adenosylhomocysteinase	SAHH
	5	880001	890000	0.036239	ATP-dependent RNA helicase DDX1, putative	DDX1
	5	900001	910000	0.042095	transcription initiation factor TFIID subunit 10, putative	TAF10
	6	1110001	1120000	0.078509	acetyl-CoA synthetase, putative	ACS

	6	1200001	1210000	0.074648	phospholipase, putative	PL
	6	1240001	1250000	0.026327	SET domain protein, putative	SET1
	6	1230001	1240000	0.024727		
	6	1220001	1230000	0.024024		
	7	220001	230000	0.03626	HECT-type E3 ubiquitin ligase UT	UT
	7	300001	310000	0.054023	conserved Plasmodium protein, unknown function	N/A
	7	310001	320000	0.048263		
	7	320001	330000	0.045844		
	7	370001	380000	0.023603	DNA-directed RNA polymerases I, II, and III subunit RPABC5, putative	RPB10
	7	390001	400000	0.029092	heat shock protein 110	HSP110c
	8	510001	520000	0.168469	JmjC domain-containing protein, putative	JmjC1
	8	500001	510000	0.167991		
	8	510001	520000	0.168469	ribosomal protein L33, apicoplast, putative	N/A
	8	540001	550000	0.106039	ATP-dependent RNA helicase DBP1, putative	DBP1
	8	540001	550000	0.106039	hydroxymethyldihydropterin pyrophosphokinase-dihydropteroate synthase	PPPK-DHPS
	8	550001	560000	0.043319		
	8	560001	570000	0.025381	CCR4-associated factor 1	CAF1
	12	980001	990000	0.030177	polyadenylate-binding protein 1, putative	PABP1
	12	990001	1000000	0.027529		
	12	990001	1000000	0.027529	histone chaperone ASF1, putative	ASF1
	13	100001	110000	0.027808	CX3CL1-binding protein 2	CBP2
K4	1	400001	410000	0.032967	bromodomain protein 3, putative	BDP3
	1	410001	420000	0.032967	chromatin assembly factor 1 subunit C, putative	CAF1C
	4	710001	720000	0.032967	histone acetyltransferase, putative	HAT1
	5	1120001	1130000	0.065934	DNA replication licensing factor MCM3, putative	MCM3
	5	1130001	1140000	0.065934	ubiquitin carboxyl-terminal hydrolase 14	USP14
	5	1140001	1150000	0.061424	Hsc70-interacting protein	HIP
	5	1140001	1150000	0.061424	conserved Plasmodium protein, unknown function	N/A
	6	1110001	1120000	0.232108	acetyl-CoA synthetase, putative	ACS
	6	1200001	1210000	0.036396	phospholipase, putative	PL

	6	1240001	1250000	0.048726	SET domain protein, putative	SET1
	6	1230001	1240000	0.037905		
	7	370001	380000	0.039171	DNA-directed RNA polymerases I, II, and III subunit RPABC5, putative	RPB10
	7	460001	470000	0.036551	conserved Plasmodium protein, unknown function	N/A
	8	500001	510000	0.032967	JmjC domain-containing protein, putative	JmjC1
	8	510001	520000	0.032967		
	8	540001	550000	0.032967	ATP-dependent RNA helicase DBP1, putative	DBP1
	8	540001	550000	0.032967	hydroxymethyldihydropterin pyrophosphokinase-dihydropteroate synthase	PPPK-DHPS
	12	740001	750000	0.032967	formin 2	FRM2
K5	5	130001	140000	0.233333	cell division control protein 6, putative	CDC6
	5	130001	140000	0.233333	serine/arginine-rich splicing factor 12	SRSF12
	5	140001	150000	0.215744	actin-depolymerizing factor 1	ADF1
	5	150001	160000	0.233979	60S ribosomal protein L31	RPL31
	5	160001	170000	0.215192	ATP-dependent RNA helicase DDX27, putative	DDX27
	5	440001	450000	0.207598	topoisomerase I	TopoI
	5	880001	890000	0.203333	ATP-dependent RNA helicase DDX1, putative	DDX1
	5	940001	950000	0.236674	iron-sulfur cluster assembly protein SufA	SufA
	5	950001	960000	0.224194	multidrug resistance protein 1	MDR1
	6	1060001	1070000	0.308376	3-oxoacyl-acyl-carrier protein synthase I/II	FabB/FabF
	6	1070001	1080000	0.327739	conserved Plasmodium protein, unknown function	N/A
	6	1060001	1070000	0.308376		
	6	1080001	1090000	0.370637	pyruvate kinase	PyrK
	6	1070001	1080000	0.327739		
	6	1110001	1120000	0.495478	acetyl-CoA synthetase, putative	ACS
	6	1240001	1250000	0.28764	SET domain protein, putative	SET1
	6	1230001	1240000	0.221009		
	7	310001	320000	0.199839	conserved Plasmodium protein, unknown function	N/A
	7	330001	340000	0.206911		
	7	340001	350000	0.251605	rhoptry-associated membrane antigen	RAMA
	7	330001	340000	0.206911		

	7	370001	380000	0.266447	DNA-directed RNA polymerases I, II, and III subunit RPABC5, putative	RPB10
	7	390001	400000	0.251523	heat shock protein 110	HSP110c
	7	400001	410000	0.3	chloroquine resistance transporter	CRT
	7	430001	440000	0.540413	prodrug activation and resistance esterase	PARE
	7	460001	470000	0.611211	conserved Plasmodium protein, unknown function	N/A
	7	470001	480000	0.458015		
	7	470001	480000	0.458015	60S ribosomal protein L34	RPL34
	7	480001	490000	0.316796	conserved Plasmodium protein, unknown function	N/A
	8	500001	510000	0.44766	JmjC domain-containing protein, putative	JmjC1
	8	510001	520000	0.403345		
	8	540001	550000	0.307884	ATP-dependent RNA helicase DBP1, putative	DBP1
	8	540001	550000	0.307884	hydroxymethyldihydropterin pyrophosphokinase-dihydropteroate synthase	PPPK-DHPS
	8	550001	560000	0.200296		
	8	580001	590000	0.289566	histone-arginine methyltransferase CARM1, putative	CARM1
	8	610001	620000	0.266667	GTP-binding protein YihA3	YihA3
	8	620001	630000	0.266667	karyopherin alpha	KARalpha
	8	640001	650000	0.242144	protein KIC7	KIC7
	8	640001	650000	0.242144	conserved Plasmodium protein, unknown function	N/A
	8	670001	680000	0.320068	40S ribosomal protein S16, putative	RPS16
	8	720001	730000	0.277419	eukaryotic translation initiation factor 3 subunit G, putative	EIF3G
	8	730001	740000	0.210779		
	8	730001	740000	0.210779	vacuolar protein sorting-associated protein 9, putative	VPS9
	8	740001	750000	0.200253		
	8	740001	750000	0.200253	ribosome assembly protein RRB1, putative	RRB1
	11	810001	820000	0.250739	exported protein 1	EXP1
	11	1340001	1350000	0.203851	conserved Plasmodium protein, unknown function	N/A
	12	830001	840000	0.23	heterochromatin protein 1	HP1
	12	830001	840000	0.23	histone-lysine N-methyltransferase, H3 lysine-4 specific	SET10
	12	840001	850000	0.23		
	12	910001	920000	0.200475	AP2 domain transcription factor AP2-G	AP2-G

	12	990001	1000000	0.204881	polyadenylate-binding protein 1, putative	PABP1
	12	990001	1000000	0.204881	histone chaperone ASF1, putative	ASF1
K6	1	530001	540000	0.013575	CX3CL1-binding protein 1	CBP1
	3	120001	130000	0.038521	cytoadherence linked asexual protein 3.2	CLAG3.2
	3	130001	140000	0.043575	erythrocyte membrane protein 1 (PfEMP1), pseudogene	N/A
	3	120001	130000	0.038521		
	3	130001	140000	0.043575	cytoadherence linked asexual protein 3.1	CLAG3.1
	4	610001	620000	0.017904	26S protease regulatory subunit 6B, putative	RPT3
	4	630001	640000	0.024259	structural maintenance of chromosomes protein 3	SMC3
	6	1080001	1090000	0.018456	pyruvate kinase	PyrK
	6	1110001	1120000	0.074976	acetyl-CoA synthetase, putative	ACS
	6	1200001	1210000	0.04	phospholipase, putative	PL
	6	1210001	1220000	0.042458	polyadenylate-binding protein 3, putative	PABP3
	6	1210001	1220000	0.042458	amino acid transporter AAT1	AAT1
	6	1220001	1230000	0.034039	SET domain protein, putative	SET1
	6	1230001	1240000	0.033192		
	6	1240001	1250000	0.018921		
	7	460001	470000	0.013448	conserved Plasmodium protein, unknown function	N/A
	8	410001	420000	0.016794	plasmepsin X	PMX
	8	500001	510000	0.051434	JmjC domain-containing protein, putative	JmjC1
	8	510001	520000	0.036514		
	8	540001	550000	0.036449	ATP-dependent RNA helicase DBP1, putative	DBP1
	8	540001	550000	0.036449	hydroxymethyldihydropterin pyrophosphokinase-dihydropteroate synthase	PPPK-DHPS
	8	550001	560000	0.031561		
	8	560001	570000	0.031819	CCR4-associated factor 1	CAF1
	8	570001	580000	0.027692		
	8	570001	580000	0.027692	histone-arginine methyltransferase CARM1, putative	CARM1
	8	580001	590000	0.022653		
	10	210001	220000	0.020698	conserved Plasmodium protein, unknown function	N/A
	12	830001	840000	0.064615	heterochromatin protein 1	HP1

	12	830001	840000	0.064615	histone-lysine N-methyltransferase, H3 lysine-4 specific	SET10
	12	840001	850000	0.064615		
	12	900001	910000	0.104615	AP2 domain transcription factor AP2-G	AP2-G
	12	910001	920000	0.104615		
	12	920001	930000	0.124782	cAMP-dependent protein kinase regulatory subunit	PKAr
	12	970001	980000	0.110712	GTP cyclohydrolase 1	GCH1
	12	980001	990000	0.092608	polyadenylate-binding protein 1, putative	PABP1
	12	990001	1000000	0.091182		
	12	990001	1000000	0.091182	histone chaperone ASF1, putative	ASF1
	13	100001	110000	0.01882	CX3CL1-binding protein 2	CBP2

Supplementary Table S6. SNPs on genes of interest under positive selection pressure within East African *P. falciparum* populations* (*iHS* value (-log10[1 – 2 | ΦiHS – 0.5 |]) > 4).

Population	chr	start	end	Number of Markers	Mean *iHS*	Gene ID	Gene Name	Gene Products
K1	1	20000	110000	80	1.823	PF3D7_0100700	*NA*	Plasmodium exported protein, unknown function, fragment
	3	1000000	1030000	58	1.05	PF3D7_0323800; PF3D7_0324600	*NA*	conserved Plasmodium protein, unknown function; stevor
	6	1330000	1360000	91	1.015	PF3D7_0631900	*NA*	stevor
	8	410000	440000	165	0.679	PF3D7_0808200	PMX	plasmepsin X
	8	470000	530000	121	1.757	PF3D7_0810100; PF3D7_0809900	*NA* JmjC1	ribosomal protein L33, apicoplast, putative; JmjC domain-containing protein, putative
	11	1280000	1310000	74	0.968	PF3D7_1133200	*NA*	conserved Plasmodium protein, unknown function
K2	2	860000	890000	96	0.73	PF3D7_0221800	*NA*	hypothetical protein
	3	990000	1020000	54	1.209	PF3D7_0323800; PF3D7_0323600	*NA*	conserved Plasmodium protein, unknown function; BSD domain-containing protein, putative
	4	120000	150000	285	0.623	PF3D7_0402100	*NA*	Plasmodium exported protein (PHISTb), unknown function
	4	1090000	1120000	210	0.601	PF3D7_0424300	EBA165	erythrocyte binding antigen-165, pseudogene
	6	1240000	1330000	224	1.773	PF3D7_0629700; PF3D7_0631100; PF3D7_0630100	SET1 *NA*	SET domain protein, putative; Plasmodium exported protein (PHISTb), unknown function; alpha/beta hydrolase, putative
	7	370000	390000	40	1.617	PF3D7_0708500; PF3D7_0708100	*NA* RPB10	heat shock protein 86 family protein; DNA-directed RNA polymerases I, II, and III subunit RPABC5, putative
	8	1300000	1330000	261	0.768	PF3D7_0830600	*NA*	Plasmodium exported protein (PHISTc), unknown function
	9	1460000	1490000	106	0.77	PF3D7_0937000	*NA*	Plasmodium exported protein (PHISTb), unknown function
	10	1380000	1410000	413	0.51	PF3D7_1035400; PF3D7_1034900	MSP3 MRScyt	merozoite surface protein 3; methionine--tRNA ligase
	12	640000	670000	47	0.523	PF3D7_1216700	PLP2	perforin-like protein 2
	13	90000	120000	167	1.077	PF3D7_1302000; PF3D7_1301700	PTP6 CBP2	EMP1-trafficking protein; CX3CL1-binding protein 2

K3	1	520000	550000	181	1.294	PF3D7_0113700; PF3D7_0114000; PF3D7_0113900	HSP40 EPF1 CBP1	heat shock protein 40, type II; exported protein family 1; CX3CL1-binding protein 1
	3	990000	1020000	48	1.007	PF3D7_0323800; PF3D7_0323600	*NA*	conserved Plasmodium protein, unknown function; BSD domain-containing protein, putative
	4	120000	150000	232	0.478	PF3D7_0402100	*NA*	Plasmodium exported protein (PHISTb), unknown function
	4	590000	620000	163	0.509	PF3D7_0413600	RPT3	26S protease regulatory subunit 6B, putative
	6	1250000	1280000	74	1.426	PF3D7_0630100	*NA*	alpha/beta hydrolase, putative
	7	450000	470000	143	0.678	PF3D7_0710200	*NA*	conserved Plasmodium protein, unknown function
	8	470000	520000	115	1.35	PF3D7_0810100; PF3D7_0809900	*NA* JmjC1	ribosomal protein L33, apicoplast, putative; JmjC domain-containing protein, putative
	8	1300000	1330000	263	0.762	PF3D7_0830600	*NA*	Plasmodium exported protein (PHISTc), unknown function
	13	90000	120000	166	0.698	PF3D7_1302000 PF3D7_1301700	PTP6 CBP2	EMP1-trafficking protein; CX3CL1-binding protein 2
	14	3110000	3140000	95	0.896	PF3D7_1475600	BDP4	bromodomain protein 4, putative
K6	4	1090000	1120000	170	0.507	PF3D7_0424300	EBA165	erythrocyte binding antigen-165, pseudogene
	12	890000	960000	32	3.348	PF3D7_1223100; PF3D7_1222600	PKAr AP2-G	cAMP-dependent protein kinase regulatory subunit; AP2 domain transcription factor AP2-G
	13	90000	120000	155	0.774	PF3D7_1302000; PF3D7_1301700	PTP6 CBP2	EMP1-trafficking protein; CX3CL1-binding protein 2

*K1 = Western Kenya, Lake Victoria Kenya, Central Uganda; K2 = Central Africa, South Central Africa, West Africa; K3 = Eastern Kenya, North East Tanzania, Lake Tanganyika, TZ; K4 = Lake Victoria Tanzania; K5 = Horn of Africa; K6 = Southern Africa and South East Tanzania.

Supplementary Table S9. Cross-population analysis of selection pressure between ancestral populations* (*Rsb* –log10 (p–value) > 5).

Population	chr	start	end	Number of Markers	Mean *iHS*	Gene ID	Gene Name	products
K1 vs. K2	1	30000	60000	29	2.244	PF3D7_0100700	*NA*	Plasmodium exported protein, unknown function, fragment
	3	1020000	1070000	43	2.164	PF3D7_0324600	*NA*	stevor
	6	1240000	1300000	194	3.248	PF3D7_0629700; PF3D7_0631100; PF3D7_0630100	SET1 *NA*	SET domain protein, putative; Plasmodium exported protein (PHISTb), unknown function; alpha/beta hydrolase, putative
	6	1320000	1350000	105	0.828	PF3D7_0631900	*NA*	stevor
	7	370000	400000	51	2.45	PF3D7_0708500; PF3D7_0708800; PF3D7_0708700; PF3D7_0708100	*NA* HSP110c *NA* RPB10	heat shock protein 86 family protein; heat shock protein 110; Cg8 protein; DNA-directed RNA polymerases I, II, and III subunit RPABC5, putative
	8	460000	560000	185	2.484	PF3D7_0810800; PF3D7_0810100; PF3D7_0810600; PF3D7_0809900	PPPK-DHPS *NA* DBP1 JmjC1	hydroxymethyldihydropterin pyrophosphokinase-dihydropteroate synthase; ribosomal protein L33, apicoplast, putative; ATP-dependent RNA helicase DBP1, putative; JmjC domain-containing protein, putative
	10	1590000	1620000	69	0.907	PF3D7_1040200	*NA*	stevor
	12	750000	780000	73	0.75	PF3D7_1219100; PF3D7_1219000	CHC FRM2	clathrin heavy chain, putative; formin 2
	12	930000	990000	29	3.664	PF3D7_1224000; PF3D7_1224300	GCH1 PABP1	GTP cyclohydrolase 1; polyadenylate-binding protein 1, putative
K1 vs. K3	1	20000	90000	37	3.564	PF3D7_0100700	*NA*	Plasmodium exported protein, unknown function, fragment
	1	520000	550000	201	0.879	PF3D7_0113700; PF3D7_0114000; PF3D7_0113900	HSP40 EPF1 CBP1	heat shock protein 40, type II; exported protein family 1; CX3CL1-binding protein 1
	3	40000	80000	51	1.11	PF3D7_0300900; PF3D7_0300400	*NA*	stevor; stevor
	4	120000	150000	222	0.577	PF3D7_0402100	*NA*	Plasmodium exported protein (PHISTb), unknown function
	4	590000	620000	160	1.355	PF3D7_0413600	RPT3	26S protease regulatory subunit 6B, putative
	4	920000	950000	82	0.897	PF3D7_0420400; PF3D7_0420300	RRF2 ApiAP2	ribosome-recycling factor; AP2 domain transcription factor, putative

	6	1250000	1290000	142	1.741	PF3D7_0630100	*NA*	alpha/beta hydrolase, putative
	11	1940000	1970000	29	1.296	PF3D7_1149000	Pf332	antigen 332, DBL-like protein
	13	90000	120000	164	0.995	PF3D7_1302000; PF3D7_1301700	PTP6 CBP2	EMP1-trafficking protein; CX3CL1-binding protein 2
K1 vs. K6	1	30000	60000	17	2.49	PF3D7_0100700	*NA*	Plasmodium exported protein, unknown function, fragment
	10	1380000	1410000	320	0.683	PF3D7_1035400; PF3D7_1034900	MSP3 MRScyt	merozoite surface protein 3; methionine--tRNA ligase
	12	820000	850000	25	2.221	PF3D7_1221000; PF3D7_1220900	SET10 HP1	histone-lysine N-methyltransferase, H3 lysine-4 specific; heterochromatin protein 1
	12	890000	920000	19	3.242	PF3D7_1222600	AP2-G	AP2 domain transcription factor AP2-G
	12	1000000	1030000	11	4.992	PF3D7_1225200	*NA*	DNA-binding protein, putative
K2 vs. K3	1	520000	550000	209	1.043	PF3D7_0113700; PF3D7_0114000; PF3D7_0113900	HSP40 EPF1 CBP1	heat shock protein 40, type II; exported protein family 1; CX3CL1-binding protein 1
	4	590000	620000	155	0.994	PF3D7_0413600	RPT3	26S protease regulatory subunit 6B, putative
	4	920000	980000	143	0.988	PF3D7_0420400; PF3D7_0420300	RRF2 ApiAP2	ribosome-recycling factor; AP2 domain transcription factor, putative
	6	1270000	1300000	138	2.17	PF3D7_0631100	*NA*	Plasmodium exported protein (PHISTb), unknown function
	6	1330000	1360000	98	0.881	PF3D7_0631900	*NA*	stevor
	7	360000	400000	57	3.387	PF3D7_0708500; PF3D7_0708800; PF3D7_0708700; PF3D7_0707700; PF3D7_0708100	*NA* HSP110c *NA* RPB10	heat shock protein 86 family protein; heat shock protein 110; Cg8 protein; E3 ubiquitin-protein ligase, putative; DNA-directed RNA polymerases I, II, and III subunit RPABC5, putative
	8	480000	560000	100	3.772	PF3D7_0810800; PF3D7_0810100; PF3D7_0810600; PF3D7_0809900	PPPK-DHPS *NA* DBP1 JmjC1	hydroxymethyldihydropterin pyrophosphokinase-dihydropteroate synthase; ribosomal protein L33, apicoplast, putative; ATP-dependent RNA helicase DBP1, putative; JmjC domain-containing protein, putative
	10	1590000	1620000	62	0.849	PF3D7_1040200	*NA*	stevor
	13	90000	120000	172	1.682	PF3D7_1302000; PF3D7_1301700	PTP6 CBP2	EMP1-trafficking protein; CX3CL1-binding protein 2
K2 vs. K6	1	50000	80000	26	1.635	PF3D7_0100700	*NA*	Plasmodium exported protein, unknown function, fragment

	6	1250000	1300000	183	3.559	PF3D7_0631100; PF3D7_0630100	*NA*	Plasmodium exported protein (PHISTb), unknown function; alpha/beta hydrolase, putative
	7	360000	400000	52	3.166	PF3D7_0708500; PF3D7_0708800; PF3D7_0708700; PF3D7_0707700; PF3D7_0708100	*NA* HSP110c *NA* RPB10	heat shock protein 86 family protein; heat shock protein 110; Cg8 protein; E3 ubiquitin-protein ligase, putative; DNA-directed RNA polymerases I, II, and III subunit RPABC5, putative
	12	820000	850000	23	1.911	PF3D7_1221000; PF3D7_1220900	SET10 HP1	histone-lysine N-methyltransferase, H3 lysine-4 specific; heterochromatin protein 1
	12	890000	1030000	53	6.085	PF3D7_1224000; PF3D7_1224500; PF3D7_1225200; PF3D7_1223100; PF3D7_1224300; PF3D7_1224400; PF3D7_1222600	GCH1 ASF1 PKAr PABP1 AP2-G	GTP cyclohydrolase 1; histone chaperone ASF1, putative; DNA-binding protein, putative; cAMP-dependent protein kinase regulatory subunit; polyadenylate-binding protein 1, putative; WD repeat-containing protein, putative; AP2 domain transcription factor AP2-G
K3 vs. K6	1	170000	200000	55	1.703	PF3D7_0104300; PF3D7_0103900	UBP1 PIESP15	ubiquitin carboxyl-terminal hydrolase 1, putative; parasite-infected erythrocyte surface protein
	1	520000	550000	175	0.792	PF3D7_0113700; PF3D7_0114000; PF3D7_0113900	HSP40 EPF1 CBP1	heat shock protein 40, type II; exported protein family 1; CX3CL1-binding protein 1
	4	590000	620000	132	1.208	PF3D7_0413600	RPT3	26S protease regulatory subunit 6B, putative
	6	1250000	1280000	81	2.122	PF3D7_0630100	*NA*	alpha/beta hydrolase, putative
	10	1380000	1410000	312	0.599	PF3D7_1035400; PF3D7_1034900	MSP3 MRScyt	merozoite surface protein 3; methionine--tRNA ligase
	12	640000	670000	32	1.827	PF3D7_1216700	PLP2	perforin-like protein 2
	12	820000	850000	27	1.711	PF3D7_1221000; PF3D7_1220900	SET10 HP1	histone-lysine N-methyltransferase, H3 lysine-4 specific; heterochromatin protein 1
	12	880000	970000	38	4.94	PF3D7_1223100; PF3D7_1222600	PKAr AP2-G	cAMP-dependent protein kinase regulatory subunit; AP2 domain transcription factor AP2-G
	12	980000	1060000	33	4.822	PF3D7_1224500; PF3D7_1225200; PF3D7_1224300; PF3D7_1224400	ASF1 *NA* PABP1 *NA*	histone chaperone ASF1, putative; DNA-binding protein, putative; polyadenylate-binding protein 1, putative; WD repeat-containing protein, putative

*K1 = Western Kenya, Lake Victoria Kenya, Central Uganda; K2 = Central Africa, South Central Africa, West Africa; K3 = Eastern Kenya, North East Tanzania, Lake Tanganyika, TZ; K4 = Lake Victoria Tanzania; K5 = Horn of Africa; K6 = Southern Africa and South East Tanzania.

Chapter 5: High throughput screening of human genetic determinants associated with malarial disease outcomes using dual indexing sequencing technology

RESEARCH PAPER COVER SHEET

Please note that a cover sheet must be completed <u>for each</u> research paper included within a thesis.

SECTION A – Student Details

Nagasaki Student No	59719003	**Title**	Miss
LSHTM Student ID No	Lsh1807687		
First Name(s)	Ashley Alexandra		
Surname/Family Name	Osborne		
Thesis Title	A multifaceted investigation of the genomics of malaria, from parasite to host, using next-generation sequencing technologies		
Nagasaki Supervisor(s)	Akira Kaneko, Kiyoshi Kita		
LSHTM Supervisor(s)	Taane Clark, Susana Campino		

If the Research Paper has previously been published please complete Section B, if not please move to Section C.

<u>SECTION B – Paper already published</u>

Where was the work published?	
When was the work published?	

If the work was published prior to registration for your research degree, give a brief rationale for its inclusion			
Have you retained the copyright for the work?*	Choose an item.	Was the work subject to academic peer review?	Choose an item.

*If yes, please attach evidence of retention. If no, or if the work is being included in its published format, please attach evidence of permission from the copyright holder (publisher or other author) to include this work.

SECTION C – Prepared for publication, but not yet published

Where is the work intended to be published?	Scientific Reports
Please list the paper's authors in the intended authorship order:	Ashley Osborne, Jody E. Phelan, Leen N. Vanheer, Alphaxard Manjurano, Christopher J. Drakeley, Akira Kaneko, Kiyoshi Kita, Susana Campino & Taane G. Clark
Stage of publication	**Undergoing revision**

SECTION D – Multi-authored work

For multi-authored work, give full details of your role in the research included in the paper and in the preparation of the paper. (Attach a further sheet if necessary)	I carried out laboratory work including primer design and optimisation, as well as prepared samples for sequencing, including DNA clean-up and shipment of samples. I performed bioinformatic analyses and interpreted the results under the supervision of my supervisors. I wrote and prepared the first draft of the manuscript that was circulated to my supervisors and co-authors.

SECTION E – Names and affiliations of co-author(s)

Please list all the co-authors' names and their affiliations.

Ashley Osborne - Faculty of Infectious and Tropical Diseases, London School of Hygiene and Tropical Medicine & School of Tropical Medicine and Global Health, Nagasaki University Jody E. Phelan - Faculty of Infectious and Tropical Diseases, London School of Hygiene and Tropical Medicine Leen N. Vanheer - Faculty of Infectious and Tropical Diseases, London School of Hygiene and Tropical Medicine Alphaxard Manjurano – National Institute for Medical Research, Mwanza Research Centre Christopher J. Drakeley - Faculty of Infectious and Tropical Diseases, London School of Hygiene and Tropical Medicine Akira Kaneko Department of Parasitology, Osaka Metropolitan University Kiyoshi Kita - School of Tropical Medicine and Global Health, Nagasaki University Susana Campino - Faculty of Infectious and Tropical Diseases, London School of Hygiene and Tropical Medicine Taane G. Clark - Faculty of Epidemiology and Population Health, London School of Hygiene and Tropical Medicine & Faculty of Infectious and Tropical Diseases, London School of Hygiene and Tropical Medicine

SECTION F

I confirm that all co-authors have agreed that the above paper will be included in my PhD thesis.

Student Signature	[redacted]
Date	06/07/23

LSHTM Supervisor Signature	[redacted]
Date	06/07/23

Nagasaki University Supervisor Signature	[redacted]
Date	06/07/23

High throughput human genotyping for variants associated with malarial disease outcomes using custom targeted amplicon sequencing

Ashley Osborne[1,2], Jody E. Phelan[1], Leen N. Vanheer [1], Alphaxard Manjurano[3], Jesse Gitaka[4,5], Christopher J. Drakeley[1], Akira Kaneko[6,7], Kiyoshi Kita[2], Susana Campino[1,*], Taane G. Clark [1,8,*]

[1] Faculty of Infectious and Tropical Diseases, London School of Hygiene & Tropical Medicine, London, UK

[2] School of Tropical Medicine and Global Health, Nagasaki University, Nagasaki, Japan

[3] National Institute for Medical Research, Mwanza Research Centre, Tanzania & Joint Malaria Program, Kilimanjaro Christian Medical Centre, Moshi, Tanzania

[4] Directorate of Research and Innovation, Mount Kenya University, Thika, Kenya

[5] Centre for Malaria Elimination, Mount Kenya University, Thika, Kenya

[6] Department of Parasitology, Graduate School of Medicine, Osaka Metropolitan University, Osaka, Japan

[7] Department of Microbiology, Tumor and Cell Biology, Karolinska Institutet, Stockholm, Sweden

[8] Faculty of Epidemiology and Population Health, London School of Hygiene & Tropical Medicine, London, UK

* Joint authors

Corresponding authors

Prof. Susana Campino and Prof. Taane G. Clark

Department of Infection Biology

London School of Hygiene and Tropical Medicine

Keppel Street, London WC1E 7HT

susana.campino@lshtm.ac.uk, taane.clark@lshtm.ac.uk

Scientific Reports

ABSTRACT

Malaria has exhibited the strongest known selective pressure on the human genome in recent history and is the evolutionary driving force behind genetic conditions, such as sickle-cell disease, glucose-6-phosphatase deficiency, and some other erythrocyte defects. Genomic studies (e.g., The 1000 Genomes project) have provided an invaluable baseline for human genetics, but, with an estimated two thousand ethno-linguistic groups thought to exist across the African continent, our understanding of the genetic differences between indigenous populations and their implications on disease is still limited. Low-cost sequencing-based approaches have made it possible to target specific molecular markers and genes of interest, leading to potential insights into genetic diversity. Here we demonstrate the versatility of custom dual-indexing technology and Illumina next generation sequencing to generate a genetic profile of human polymorphisms associated with malaria pathology. For 100 individuals diagnosed with severe malaria in Northeast Tanzania, variants were successfully characterised on the *haemoglobin subunit beta (HBB), glucose-6-phosphate dehydrogenase (G6PD), atypical chemokine receptor 1 (ACKR1)* genes, and the intergenic Dantu genetic blood variant, then validated using pre-existing genotyping data. High sequencing coverage was observed across all amplicon targets in *HBB, G6PD, ACKR1,* and the Dantu blood group, with variants identified at frequencies previously observed within this region of Tanzania. Sequencing data exhibited high concordance rates to pre-existing genotyping data (>99.5%). Our work demonstrates the potential utility of amplicon sequencing for applications in human genetics, including to personalise medicine and understand the genetic diversity of loci linked to important host phenotypes, such as malaria susceptibility.

Word count: 248

INTRODUCTION

Despite decades of progress, malaria incidence in 2020 increased for the first time since the start of the millennium and resulted in an estimated 241 million cases and 627,000 deaths, a 12% increase from the number of deaths reported in 2019[1]. This increased incidence of malaria was largely due to disruptions in implementing malaria and vector control programmes, as well as supply chain failures, caused by the COVID-19 pandemic[1,2]. In addition to the COVID-19 pandemic, increased disease incidence has also linked to the continued emergence and spread of drug resistance around the world and the ongoing influence of climate change on vector populations and weather patterns[3,4]. As malaria continues to be a major burden on public health in low- and middle-income countries, where it disproportionately affects pregnant women and children under five years of age, the need for technological advancements in disease control and elimination have never been more apparent[1,5].

Malaria has exhibited the strongest known selective pressure on the human genome observed in recent history and has been proven to be the driving force behind a variety of human polymorphisms associated with malarial disease outcomes and severity, such as the sickle-cell trait, thalassaemia, glucose-6-phosphatase deficiency (G6PD), and other erythrocyte variations[6,7]. Regarded as the classic paradigm of balancing selection in human populations, the sickle-cell allele (HbS) has evolved independently in multiple malaria-endemic regions due to its ability to confer up to a 10-fold reduced risk of severe malaria*,* despite its pathogenic effects in homozygous carriers[8]. While the sickle-cell allele has been associated with protection against severe disease caused by *P. falciparum*, the Duffy negative (Fy-) phenotype has all but eliminated *P. vivax* from much of sub-Saharan Africa and is not associated with any pathogenic phenotypes[9].

Due to the high prevalence of G6PD in sub-Saharan Africa, which results in diminished activity of the G6PD enzyme, it is believed that G6PD genetic variants on the X chromosome arose due to selection pressure exhibited by malaria on the human genome[10,11]. This hypothesis has been corroborated by studies that have identified negative associations with severe malaria in hemizygous males and heterozygous females[7,10]. A majority of individuals carrying this genetic disorder remain

asymptomatic, however clinical manifestations of the deficiency can include haemolytic anaemia which can be exacerbated by treatment with primaquine, an antimalarial drug used for the treatment of *P. falciparum* gametocytes and relapses of *P. ovale*[11,12].

Although progress has been made towards understanding the impacts that variations in the host genome have on malarial disease outcomes, Africa, as a whole, remains underrepresented in genetic studies[13]. Genome-wide studies (e.g., The 1000 Genomes project) have provided invaluable baseline data for human genetics and represented individuals across Indigenous African populations[14]. However, with approximately 2,000 ethno-linguistic groups thought to exist across the African continent, more information is needed to gain a comprehensive understanding of the genetic differences between populations and their implications on disease[13]. Additionally, this underrepresentation of individuals with African ancestry in genetic databases poses challenges towards the successful application of genetically tailored medicine as it becomes more widely available with rapid advances being seen in sequencing technology[13,15].

Genome-wide association studies (GWAS) test for differing genotype or allele frequencies of millions of genetic variants between phenotypes, accounting for the confounding effects of population structure[16]. Variants known to confer phenotypes associated with severe malarial disease, such as HbS and G6PD deficiency, can be measured and their impact on the human genome quantified. This includes investigation into allelic heterogeneity and the implications of differing phenotypes on disease. An example of this being the complexity surrounding G6PD deficiency, with heterozygous females and hemizygous males exhibiting protection from severe malaria, while homozygous individuals range from having no added protection to being more at-risk for severe disease[7]. Perhaps one of most beneficial aspects of GWAS studies is the ability to screen for novel loci with associations to disease outcome, such as in case-control studies, leading to the identification of new “candidate genes” for malaria sensitivity to be investigated further[17,18]

Large-scale, multi-population, GWAS studies have highlighted the high degree of genetic diversity at loci associated with susceptibility to malaria on not only a global scale, but also between African populations within the same country borders[19]. This small-scale diversity further suggests our limited genetic data from across Africa's ethic groups means we have only explored a cross-section of the impact malaria has had on human genetics. Despite the importance of whole genome sequencing (WGS) and GWAS studies, they remain expensive and time consuming, as well as pose a variety of ethical dilemmas when applied to vulnerable populations[20]. Recent advancements in targeted low-cost sequencing-based approaches have made it possible to target specific molecular markers on genes of interest, making use of data and knowledge that has been obtained from extensive whole genome studies[21]. This technology has already been utilised in recent studies as a method of high throughput screening for drug resistance associated loci in malaria parasites, as well as molecular markers for insecticide resistance in mosquito vectors, suggesting the capacity for a cross-species method for surveillance[21,22].

High throughput genotyping of human genetic variants identified to have associations with severe malaria, a biological role in the process of malaria infection, or an impact on the effectiveness of treatment regimens, could be utilised to genetically profile communities with the aim of developing tailored malaria control programmes. Additionally, if combined with advancements in technologies furthering "on-the-go" science, such as Oxford Nanopore Technology's portable MinION, these profiles could be generated in real-time and on a scale never seen before[23,24]. Here we demonstrate the versatility of custom dual-indexing technology and Illumina next generation sequencing by presenting a proof-of-concept method for profiling human genetic determinants of malarial disease in at-risk populations, utilising clinical samples from a well characterised population in Northeast Tanzania[18,25,26].

RESULTS

DNA from a total of 100 patients diagnosed with severe malaria, aged between 2 months and 13 years, at the Tuele Hospital in Muheza, Tanzania were sequenced for 7 amplicon targets across 4 genes associated with malaria sensitivity and disease outcome. The genes are *HBB* (chr. 11; 4 amplicons), *G6PD* (chr. X; 1 amplicon), *atypical chemokine receptor 1* (*ACKR1/DARC*; Duffy Blood group; chr. 1; 1 ampicon), and the Dantu genetic blood variant (chr. 4; 1 amplicon) (**Figure 1**). Samples were multiplexed according to sample-specific indices to promote high throughput and efficient sequencing of large sample sets. Of the 100 DNA samples that were selected for genotyping, 85 were successfully sequenced for all 7 amplicon targets. Of the 15 samples that were not completely profiled, 1 contained inadequate sequencing data for a single amplicon, 12 had inadequate sequencing data for two amplicons, and 1 had inadequate sequencing reads for all targets.

The average coverage across the 7 amplicons targeting gene coding regions ranged from 973- to 2195-fold, while the coverage for the Dantu amplicon, located in an intergenic region, was markedly lower at 184-fold but was still above the recommended minimum coverage of 30-fold for human genetic analyses (**Supplementary Table 1**). The average read depth for specific SNPs of interest across all amplicon targets ranged from 132- to 2003-fold (**Figure 2A**). The coverage for the Dantu blood group variant, rs186873296, was markedly lower than the other SNPs of interest (**Figure 2B**). The lower coverage of the Dantu amplicon is likely explained by a lower primer binding affinity than the other primers used with this method [27].

Variants characterised on HBB, ACKR1, and the Dantu blood variant

A single amplicon was used to characterise variants on the *HBB* gene. The rs334 variant, associated with sickle cell trait (HbS), and the rs33950507 (HbE) and rs33930165 (HbC) variants were used to assess the accuracy of amplicon sequencing and variant calling for *HBB* as pre-existing genotyping data was available. The rs334 variant was recorded in 9.1% (9/99) of participants with 5 individuals identified as heterozygous for the sickle cell allele, carrying both the variant allele and wild-type allele (HbAS), and 4 individuals identified as homozygous positive for the sickle cell allele (HbSS; sickle-cell

disease). As expected, the rs33950507 variant (HbE) was not identified in any of the participants in this study, as this is specific to Southeast Asian populations. Similarly, the rs33930165 variant (HbC) was not identified in any participants, consistent with its high abundance in areas of West Africa. Pre-existing genotyping data was available for rs334, rs33950507 and rs33930165; amplicon sequencing data matched the pre-existing genotyping data for all screened individuals (100% concordance) (**Table 1**)**.** Off-target variants were also classified and presented with their corresponding median read depth (see **Table 1)**.

The *HBB* amplicon encompassed several other polymorphisms not categorised in pre-existing genotyping data in this study population, including rs33972047, rs33915217, and rs713040. The rs713040 variant, associated with benign presentations of beta thalassemia (β+thal) and the fetal haemoglobin quantitative trait locus 1, was identified in 95.8% (92/96) of participants, with 67.7% (65/96) of individuals being homozygous positive for the variant allele. The rs33972047 and rs33915217 variants were not identified in any participants.

The variant rs2814778 on *ACKR1*, which encodes the Duffy blood group antigen, results in a Duffy-negative phenotype, which is predominantly fixed in most populations across Sub-Saharan Africa (**Table 2**). The average coverage for the *ACKR1* amplicon was 1226 and, as expected the rs2814778 variant was identified in 100% (95/95) of study participants and matched pre-existing genotyping data for all individuals. The Dantu blood group variant rs186873296 was identified in 3.1% of study participants (3/98) with three individuals being heterozygous for the variant allele. There was one individual (1/98) that was misidentified as homozygous negative for the variant allele using amplicon sequencing that had been previously identified as heterozygous for the variant allele in the pre-existing genotyping data.

Genotyping for G6PD deficiency

G6PD was covered by four amplicons and overall had high sequencing coverage: average coverage of 1,089-, 973-, 1,212-, and 1,548-fold, respectively. The resulting amplicon sequencing data was used to

determine the G6PD genotype of participants (**Table 1**) and infer the G6PD A- deficiency phenotype of individuals based on alleles present at nucleotide positions 202 (rs1050828) and 376 (rs1050829), as well as others (542, rs5030872; 680, rs137852328; 968, rs76723693) (**Figure 3**). The rs1050828 variant was identified in 18.4% (16/87) of participants, with 8 individuals identified as homozygous positive for the variant allele and 8 individuals identified as heterozygous carriers, while the rs1050829 variant was identified in 44.8% (39/87) of participants, with 26.4% (23/87) being homozygous positive for the variant allele.

Amplicon sequencing genotyping data for rs1050828 matched pre-existing genotyping data for all screened individuals (87/87). There were 2 individuals that were misidentified as heterozygous for the rs1050829 variant allele that had previously been identified as homozygous negative, as well as one individual that was misidentified as heterozygous the rs1050829 allele that had previously been identified to be homozygous positive. The 542, 680, and 968 variants were not identified, however two additional variants, both believed to be benign variants of G6PD deficiency, were identified within the sample population, rs2230036 (12.1%; 12/99) and rs5986875 (1%; 1/99).

Of the 83 individuals with accurate amplicon sequencing genotyping data, available for both the 202 and 376 nucleotide positions, most participants (55.4%; 46/83) were identified to have the wild-type G6PD genotype (Males = B hemizygous; Females = BB homozygous) (**Table 3**). There were 8 individuals (9.6%), 5 males and 3 females, identified to have the severely deficient G6PD phenotype (Males = A-; Females = A-A-).

DISCUSSION

With the stalling progress of malaria control programmes aiming for global elimination, molecular surveillance tools offer a next-generation solution to a major public health burden[1]. Despite the fact that malaria has exhibited the strongest known selective pressure on the human genome in recent history, and its disproportionate public health impact in sub-Saharan Africa, most of the African

continents highly diverse ethnic groups are underrepresented in genetic studies[13]. Not only does this limit our understanding of the evolution and distribution of human genetic variants associated with malaria disease severity across the continent, but it also increases the difficulty of successfully delivering appropriately tailored public health or personalised medicine interventions. One of the main barriers to expanding human genetic profiling to sub-Saharan Africa, behind cost and infrastructure, has been the ethical implications of carrying out WGS and GWAS on vulnerable populations and the difficulties associated with safely storing vast quantities of personal and genomic data for those individuals[15,20].

Advancements in low-cost targeted sequencing technology have made it possible to cheaply target specific genes of interest, eliminating the need to perform WGS or store extraneous human genomic data. Here we presented a proof-of-concept method for the targeted sequencing of human genetic variants associated with malaria disease severity and treatment efficacy using an Illumina sequencing platform and custom dual-indexing technology. The custom 5'- and 3'- indices allowed for individual sample identification across multiple targets following pooling and sequencing in a single reaction. For this study, 10 unique forward indices and 10 unique reverse indices were designed with 8 base pairs each, creating 90 possible dual index combinations. This number is infinitely expandable depending on the requirements of the study. This method was used to sequence and screen for genetic variants on *G6PD, HBB*, and *ACKR1* genes, as well as the Dantu genetic blood variant. In this study, severe malaria cases were chosen with an enrichment for the sickle-cell allele, to assist with validation of amplicon sequencing accuracy, therefore the HbS allele frequency in our dataset (9.1%; 9/99) does not reflect the anticipated population frequency (16.5%; 79/477). For other variants classified within this study, the allele frequencies are broadly similar between the study population and estimates from both Tanzanian and other sub-Saharan African populations.

Due to the size of the *G6PD* gene, located on the X chromosome, four amplicons were used to target five key variants (202, 376, 542, 680, and 968) associated with G6PD deficiency, namely African-type

G6PD A-, which results in diminished activity of the enzyme glucose-6-phosphatase dehydrogenase enzyme[11,28]. All four amplicons were observed to have high sequencing coverage (average >950-fold). The 202 and 376 G6PD variants are common in sub-Saharan Africa, whereas 542 and 968 variants have been observed in West African populations (e.g., The Gambia) and the 680 variant appears rare in the continent [7,29]. This trend was observed in our data where the 202 and 376 variant alleles were observed in 18.4% and 44.8% of study participants, respectively, while the 542, 680, and 968 variants were not identified. Through performing high-resolution sequencing of genetic targets, rather than genotyping for specific loci only, this methodology has the capacity to capture novel, or rare, polymorphisms, such as those with functional consequences that would normally be missed, as well as potentially explain allelic heterogeneity, often confounded in genetic association studies[29]. Two additional variants (rs2230036, rs5986875), both believed to be benign variants of G6PD deficiency, were identified alongside the previously described variants.

Using this method, two individuals were misidentified as being heterozygous for the 376-variant allele, as opposed to being homozygous negative for the allele, and one person was misidentified as heterozygous positive for the allele rather than homozygous positive. These misidentifications could be due to cross-contamination with other samples during the high throughput PCR step of this process. In most instances, small amounts of contamination with low read counts will be removed by bioinformatic filtering steps during downstream analysis, alongside other sequencing artifacts**.** Overall, discordance between amplicon sequencing data and available genotyping data was extremely low. Across our study participants, high quality sequencing data was achieved for 962 genotypes with pre-existing data. Of these 962 genotypes, only 4 were discordant (discordance: 0.42%, 4.2 per 1000 genotypes), including the 3 positions discussed on G6PD. We present this discordance to highlight the need for further testing and optimisation in larger datasets, across multiple populations, before this methodology could be applied within a clinical setting if desired.

For the *HBB* gene, one amplicon was designed to target variants associated with thalassaemias, sickle cell anaemia (HbS), HbC, HbE, as well as a variety of other haemoglobinopathies, and the average coverage for the HBB amplicon was 2195-fold. The HbS allele was identified in 9.1% of study participants with 5 individuals identified to have both the variant allele and wild-type allele (HbAS) and 4 individuals identified to have both variant alleles (HbSS) which results in the clinical manifestation of sickle cell disease rather than conferring a relatively harmless protective effect[30,31]. The HbC and HbE alleles were not observed in this study population which was to be expected as the HbC allele is more commonly found in people of West African descent and the HbE allele is common in Southeast Asian populations[30,32]. The rs713040 variant, associated with benign presentations of beta thalassemia and the hereditary persistence of foetal haemoglobin, was identified in 95.8% of participants, however there is no known impact of this variant associated with malaria infection or disease.

A single amplicon was used to sequence the rs2814778 variant on the *ACKR1* gene, which encodes the Duffy blood group antigen located on the surface of red blood cells[33]. The rs2814778 variant results in a Duffy negative phenotype, or the absence of the Duffy antigen, and is generally fixed in sub-Saharan African populations due to its ability to protect against *P. vivax* infection, which relies on the protein as a receptor for invasion into the red blood cells, and the lack of any known pathogenic effects in humans who harbour this variant[9,33]. As anticipated, 100% of study participants were identified to be homozygous positive for the rs2814778 variant, which results in the Duffy negative phenotype[9].

The amplicon designed to target the Dantu blood group variant was the only amplicon designed to target an intergenic region. The Dantu blood variant is located on chromosome 4, upstream of the *GYPA* and *GYPB* genes, and was recently identified as the causative polymorphism behind a novel protein expressed on the surface of red blood cells that alters the red blood cell surface tension and makes it more difficult for malaria parasites to invade[34]. To date there is no known evidence of health complications in carriers of this variant allele[34]. The Dantu variant was identified in 3.1% of study

participants, with 3 individuals being heterozygous for the variant allele. There was one individual that was misidentified as homozygous negative for the variant allele using amplicon sequencing that had been previously identified as heterozygous for the variant allele in the pre-existing genotyping data. This misclassification was due to low sequencing reads of the non-reference allele, resulting in the alternate allele not meeting the filtering threshold requiring >20% of reads to be non-reference for a heterozygous classification[35].

The average coverage for the Dantu amplicon was 184-fold which was substantially lower than the other amplicons used in this method. This lower coverage is likely due to lower primer binding affinity when compared to other primers used in this method. All primer sets used in this study were designed to have the same annealing temperatures so all reactions could be run concurrently using one PCR programme. This reduces the overall time and costs associated with processing high volume datasets. A lower annealing temperature may increase the overall coverage of the Dantu amplicon, however the coverage achieved using the annealing temperature described in our methodology achieved sufficient coverage to perform confident variant calling[36].

The successful genotyping of targets across *G6PD, HBB, ACKR1,* and the Dantu blood variant, highlights the versatility of amplicon sequencing and suggests a capacity for easy expansion to other genetic markers of disease. Such targets could include the *ABO* gene, variants linked to drug metabolism, or polymorphisms yet to be identified by future GWAS studies. A cost-effective and highly adaptable genotyping assay, such as the one presented here, has the potential to assist with surveillance and personalised medicine, while simultaneously addressing important ethical concerns surrounding the collection of human data and the need for low-cost sequencing methods to be accessible for low- and middle-income countries to promote in-country research capacity.

CONCLUSION

This study presents a methodology that makes use of advancements in high throughput sequencing, and custom-dual indexing technology, to genetically profile human genetic variants. We focused on variants associated with malaria disease severity and treatment implications in genetically diverse communities, simultaneously cutting down on extraneous data collected through genome-wide studies, as well as the associated costs. This methodology leverages off previously presented techniques used in characterising drug resistance biomarkers in *P. falciparum* infections and suggests the possibly of a cross-species method of surveillance which, together, could inform region-specific, highly specialised, malaria intervention programmes.

MATERIALS AND METHODS

Study site description and sample collection

Human and parasite DNA used in this study was obtained from samples collected in a study conducted in Tanga region, Northeast Tanzania between June 2006 and May 2007 [PMID: 25671784; PMID: 29381699]. Samples had been recruited from individuals, aged between 2 months and 13 years, admitted to Tuele Hospital in Muheza, with severe malaria. Severe malaria cases were defined as those with a history of fever within the 48 hours prior to admission, asexual *P. falciparum* parasitaemia, and one or all of the following: more than 2 seizures within the previous 24hrs; Blantyre coma score (BCS) of less than 3 (repeated if BCS < 5 and convulsion within 1 hour of anticonvulsant given within 6 hours); prostration (the inability to sit unsupported or, if aged <8 months, to drink); respiratory distress (deep breathing, low chest wall indrawing, a respiratory rate greater than 70 bpm, or an oxygen saturation of less than 90%); jaundice (identified by inspection of sclera); severe anaemia (haemoglobin less than 5g/dl), blood glucose level less than 2.5mmol/l, or a blood lactate greater than 5mmol/l.

Individuals included within the study underwent genome-wide and targeted genotyping [7,18,37]. Genotyping was carried out to identify variants with previously reported associations to malaria severity, as well as a biological role in malaria infection and disease, was carried out for *G6PD*, *HBB*,

ACKR1, and the Dantu blood variant (**Table 1**). Genotyping data specifically relevant to this study was available, including for *G6PD* (e.g., rs1050828, rs1050829, rs5030872, rs137852328, rs76723693), *HBB* (e.g., rs334, rs33930165, rs33950507), *ACKR1* (e.g., rs2814778), and the Dantu blood group variant (rs186873296) [18,25]. This data was used to assess the sequencing and variant calling accuracy of this methodology. All experimental protocols were approved by the ethical review board of the London School of Hygiene and Tropical Medicine and the Tanzanian National Medical Research Institute (Proposal number: ID 4093). All methods were carried out in accordance with relevant guidelines and regulations and informed consent was obtained from all subjects and/or their legal guardian(s).

Study population characteristics

The population of Muheza is dominated by the Mzigua and Wasambaa ethnic groups which generally rely on subsistence agriculture, livestock keeping, and fishing as their main sources of income. At the time of sample collection, much of the population had access to primary health facilities, however medicine shortages, high costs for users, and inadequate facility infrastructure limited the scope of these facilities. At the time of collection, transmission of *Plasmodium falciparum* was recorded to be high (~50-700 infected bites per person per year) and followed a seasonal pattern with two distinct seasonal peaks per year. In 2002, community prevalence of *P. falciparum* was 88.2% in children aged 2 to 5 years of age. The dominant malaria transmitting vectors within the area were recorded to be *Anopheles gambiae sensu stricto* and *Anopheles funestus*[38].

Primer design

Primers used in this study were designed to investigate polymorphisms on human genes known to be associated with malaria disease severity and outcomes of disease by amplifying fragments, or amplicons, of between 400 and 600 base pairs (bp). Targeted genes included: (i) *HBB* (rs334 (HbS), rs33950507 (HbE), rs33930165 (HbC), rs33941377, rs33944208, rs33972047); (ii) *ACKR1* (rs2814778); (iii) *G6PD* (rs1050828, rs1050829, rs78365220, rs5030872, rs137852328, rs137852314, rs5030868, rs137852330, rs76723693, rs13785232), (iv) Dantu blood group (rs186873296). Forward and reverse primers included unique 8 bp indices, or barcodes, used to demultiplex sequencing data, described

below. The unique primer indices were designed to be 8 bp, to mitigate chances of recombination and contamination possible with shorter indices. Sequencing adaptors were added at Illumina using adaptor ligation technology.

PCR reactions and programmes

PCRs were performed using a master mix containing 5 µl of Q5 Reaction Buffer (New England BioLabs), 0.5 µl of dNTPs (1n nM stocks, New England BioLabs), 0.25 µl Q5 Hot Start High-Fidelity DNA Polymerase (New England BioLabs), and 15.75 µl Milli-Q water (Merck). For each reaction, a total of 1.25 µl of forward and 1.25 µl of reverse primer (10 pmol/µl stocks) were used with 1 µl of DNA for a total reaction volume of 25 µl. The reactions were carried out in a thermocycler consisting of the following steps: Heat activation for 15 minutes at 72°C, 30 cycles of denaturation for 20 seconds at 95°C, annealing for 2 minutes at 60°C, elongation for 2 minutes at 72°C, and a final elongation for 10 minutes at 72°C, followed by a hold at 10°C.

Amplicon purification and pooling

Combinations of indices were conserved by sample identifier, regardless of amplicon target. Samples without overlapping combinations of indices were pooled together for cleaning and sequencing, with a maximum of 200 amplicons allowed per pool to ensure sufficient sequencing coverage. Pooled samples were cleaned prior to sequencing using KAPA bead purification, following the manufacturer's instructions. A ratio of 1:0.70 of product to bead volume was used to size select DNA fragments the size of amplicons and remove excess primers. DNA was measured (Qubit dsDNA HS) and normalised to 20 ng per 25 µl.

Illumina sequencing and bioinformatics

One hundred human DNA samples were chosen from severe malaria cases (Muheza, TZ) and sequenced following PCR amplification of all the amplicon targets. Sequencing was carried out with the Illumina MiSeq platform using adaptor ligation technology at Genewiz (GENEWIZ Germany

GmbH). Following demultiplexing of pooled sequencing reads according to the unique sample IDs, the raw sequencing data was mapped to the GRCh38.p13 *Homo sapiens* reference genome using the default parameters set for Illumina data within the software *bwa-mem*. SNPs and indels were called using the *samtools*, freebayes, and GATK software suites. Samples with a sequencing coverage <30-fold for positions of interest were discarded and not used in further data interpretation[36]. Samples with <20% of reads for a non-reference allele were classed as homozygous-negative and sample with >80% of reads for a non-reference allele were classed as homozygous-positive. SNPs were annotated using the *SnpSift* software, and information available via ClinVar, to identify variants with known rsID numbers (reference SNP cluster ID).

Genotyping data

The sequence data were compared to genotyping data on the samples from previous work, allowing an assessment of error rates. This includes candidate SNPs for *G6PD* (202 rs1050828, 376 rs1050829, 542 rs5030872, 680 rs137852328, 968 rs76723693), *ACKR1* (Duffy blood group antigen, including rs2814778) and *HBB* (HbS rs334, HbC rs33930165, and HbE rs33950507)[7,25,37], as well as imputed genome-wide polymorphism (from Illumina Omni 2.5 million SNP chip) [18].

ACKNOWLEDGEMENTS

We thank all participants and staff in the original Tanzanian study, and the MalariaGEN resource centre for the generation of the published genetic data.

AUTHORS CONTRIBUTIONS

AK, KK, SC and TGC conceived and designed the study; CD and AM provided biological materials and data. AO, LV, and SC coordinated the sequencing of samples; AO and JP performed the bioinformatic and statistical analysis, under the supervision of SC and TGC; AO wrote the first draft of the manuscript, and the final version included edits from all authors. The final manuscript was read and approved by all authors.

FUNDING STATEMENT

AO is supported by a Nagasaki University – LSHTM PhD studentship funded by the WISE programme of MEXT. SC is funded by the Medical Research Council UK (Grant no. MR/M01360X/1). AK received support from JSPS KAKENHI (Grant No. JP18KK0248 & JP19H01080) and JICA/AMED joint research project (SATREPS) (Grant no. 20JM0110020H0002) and Hitachi fund. TGC is supported by the Medical Research Council UK (Grant no. MR/M01360X/1, MR/N010469/1, MR/R020973/1, MR/X005895/1). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

CONFLICTS OF INTEREST

There are no conflicts of interest or competing interests.

DATA AVAILABILITY

All raw sequence data is available from the ENA (project accession number PRJEB58734).

REFERENCES

1. World Health Organization. *World malaria report 2021*. (World Health Organization, 2021).

2. Weiss, D. J. *et al.* Indirect effects of the COVID-19 pandemic on malaria intervention coverage, morbidity, and mortality in Africa: a geospatial modelling analysis. *Lancet Infect. Dis.* **21**, 59–69 (2021).

3. Mordecai, E. A., Ryan, S. J., Caldwell, J. M., Shah, M. M. & LaBeaud, A. D. Climate change could shift disease burden from malaria to arboviruses in Africa. *Lancet Planet. Health* **4**, e416–e423 (2020).

4. World Health Organization. *Report on antimalarial drug efficacy, resistance and response: 10 years of surveillance (2010–2019)*. (World Health Organization, 2020).

5. Perkins, D. J. *et al.* Severe Malarial Anemia: Innate Immunity and Pathogenesis. *Int. J. Biol. Sci.* **7**, 1427–1442 (2011).

6. Kwiatkowski, D. P. How Malaria Has Affected the Human Genome and What Human Genetics Can Teach Us about Malaria. *Am. J. Hum. Genet.* **77**, 171–192 (2005).

7. Manjurano, A. *et al.* African Glucose-6-Phosphate Dehydrogenase Alleles Associated with Protection from Severe Malaria in Heterozygous Females in Tanzania. *PLoS Genet.* **11**, (2015).

8. Ackerman, H. *et al.* A Comparison of Case-Control and Family-Based Association Methods: The Example of Sickle-Cell and Malaria. *Ann. Hum. Genet.* **69**, 559–565 (2005).

9. Howes, R. E. *et al.* The global distribution of the Duffy blood group. *Nat. Commun.* **2**, 266 (2011).

10. Gampio Gueye, N. S. *et al.* An update on glucose-6-phosphate dehydrogenase deficiency in children from Brazzaville, Republic of Congo. *Malar. J.* **18**, 57 (2019).

11. Mason, P. J., Bautista, J. M. & Gilsanz, F. G6PD deficiency: the genotype-phenotype association. *Blood Rev.* **21**, 267–283 (2007).

12. World Health Organization. *Policy brief on single-dose primaquine as a gametocytocide in Plasmodium falciparum malaria*. https://apps.who.int/iris/handle/10665/338498 (2015).

13. Tucci, S. & Akey, J. M. The long walk to African genomics. *Genome Biol.* **20**, 130 (2019).

14. Auton, A. *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).

15. Gurdasani, D. *et al.* The African Genome Variation Project shapes medical genetics in Africa. *Nature* **517**, 327–332 (2015).

16. Uffelmann, E. *et al.* Genome-wide association studies. *Nat. Rev. Methods Primer* **1**, 1–21 (2021).

17. Band, G. *et al.* A novel locus of resistance to severe malaria in a region of ancient balancing selection. *Nature* **526**, 253–257 (2015).

18. Ravenhall, M. *et al.* Novel genetic polymorphisms associated with severe malaria and under selective pressure in North-eastern Tanzania. *PLoS Genet.* **14**, e1007172 (2018).

19. Maiga, B. *et al.* Glucose-6-phosphate dehydrogenase polymorphisms and susceptibility to mild malaria in Dogon and Fulani, Mali. *Malar. J.* **13**, 270 (2014).

20. Vries, J. de & Pepper, M. Genomic sovereignty and the African promise: mining the African genome for the benefit of Africa. *J. Med. Ethics* **38**, 474–478 (2012).

21. Nag, S. *et al.* High throughput resistance profiling of Plasmodium falciparum infections based on custom dual indexing and Illumina next generation sequencing-technology. *Sci. Rep.* **7**, 2398 (2017).

22. Campos, M. *et al.* High-throughput barcoding method for the genetic surveillance of insecticide resistance and species identification in Anopheles gambiae complex malaria vectors. *Sci. Rep.* **12**, 13893 (2022).

23. Jain, M., Olsen, H. E., Paten, B. & Akeson, M. The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. *Genome Biol.* **17**, 239 (2016).

24. Gómez-González, P. J., Campino, S., Phelan, J. E. & Clark, T. G. Portable sequencing of Mycobacterium tuberculosis for clinical and epidemiological applications. *Brief. Bioinform.* **23**, bbac256 (2022).

25. Manjurano, A. *et al.* Candidate Human Genetic Polymorphisms and Severe Malaria in a Tanzanian Population. *PLOS ONE* **7**, e47463 (2012).

26. Shelton, J. M. G. *et al.* Genetic determinants of anti-malarial acquired immunity in a large multi-centre study. *Malar. J.* **14**, 333 (2015).

27. Kayama, K. *et al.* Prediction of PCR amplification from primer and template sequences using recurrent neural network. *Sci. Rep.* **11**, 7493 (2021).

28. Peters, A. L. & Noorden, C. J. F. V. Glucose-6-phosphate Dehydrogenase Deficiency and Malaria: Cytochemical Detection of Heterozygous G6PD Deficiency in Women. *J. Histochem. Cytochem.* **57**, 1003–1011 (2009).

29. Clark, T. G. *et al.* Allelic heterogeneity of G6PD deficiency in West Africa and severe malaria susceptibility. *Eur. J. Hum. Genet.* **17**, 1080–1085 (2009).

30. Thom, C. S., Dickson, C. F., Gell, D. A. & Weiss, M. J. Hemoglobin variants: biochemical properties and clinical correlates. *Cold Spring Harb. Perspect. Med.* **3**, a011858 (2013).

31. Ashley-Koch, A., Yang, Q. & Olney, R. S. Sickle hemoglobin (HbS) allele and sickle cell disease: a HuGE review. *Am. J. Epidemiol.* **151**, 839–845 (2000).

32. Flatz, G., Sanguansermsri, T., Sengchanh, S., Horst, D. & Horst, J. The 'hot-spot' of Hb E [beta26(B8)Glu-->Lys] in Southeast Asia: beta-globin anomalies in the Lao Theung population of southern Laos. *Hemoglobin* **28**, 197–204 (2004).

33. Hamblin, M. T., Thompson, E. E. & Di Rienzo, A. Complex Signatures of Natural Selection at the Duffy Blood Group Locus. *Am. J. Hum. Genet.* **70**, 369–383 (2002).

34. Kariuki, S. N. *et al.* Red blood cell tension protects against severe malaria in the Dantu blood group. *Nature* **585**, 579–583 (2020).

35. Garrison, E. & Marth, G. Haplotype-based variant detection from short-read sequencing. (2012) doi:10.48550/ARXIV.1207.3907.

36. Sims, D., Sudbery, I., Ilott, N. E., Heger, A. & Ponting, C. P. Sequencing depth and coverage: key considerations in genomic analyses. *Nat. Rev. Genet.* **15**, 121–132 (2014).

37. Manjurano, A. *et al.* USP38, FREM3, SDC1, DDC, and LOC727982 Gene Polymorphisms and Differential Susceptibility to Severe Malaria in Tanzania. *J. Infect. Dis.* **212**, 1129–1139 (2015).

38. Maxwell, C. A. *et al.* Variation of malaria transmission and morbidity with altitude in Tanzania and with introduction of alphacypermethrin treated nets. *Malar. J.* **2**, 28 (2003).

39. Pfeffer, D. A. *et al.* malariaAtlas: an R interface to global malariometric data hosted by the Malaria Atlas Project. *Malar. J.* **17**, 352 (2018).

Tables

Table 3: Distribution of alleles associated with malaria disease severity, as well as off-target non-synonymous variants, identified on *HBB, G6PD*, Dantu, and *ACKR1* loci in Northeast Tanzania using amplicon sequencing.

Gene	Chrom	Position	rs ID[a]	Variant Information	Homozygous Negative (-/-)		Heterozygous (-/+)		Homozygous Positive (+/+)		Mean DP	Concordance[b] %
G6PD	X	154532738	rs2230036	C > T	CC	87	CT	6	TT	6	803	*NA*
		154533025	rs76723693	968 A > G	AA	99	AG	0	GG	0	798	100
		154533122	rs137852327	C > T	CC	99	CT	0	TT	0	798	*NA*
		154534125	rs137852328	680 C > A	CC	99	CA	0	AA	0	629	100
		154534177	rs5986875	G > A	GG	98	GA	1	AA	0	629	*NA*
		154534440	rs5030872	542 T > A	TT	99	TA	0	AA	0	622	100
		154535443	*NA*	G > A	GG	86	GA	1	AA	0	622	*NA*
		154535468	*NA*	G > T	GG	86	GT	1	TT	0	622	*NA*
		154534527	*NA*	T > C	TT	86	TC	1	CC	0	622	*NA*
		154535277	rs1050829	376 T > C	TT	48	TC	16	CC	23	956	96.6
		154536002	rs1050828	202 C > T	CC	71	CT	8	TT	8	530	100
HBB	11	5226867	*NA*	C > G	CC	98	CG	1	GG	0	1020	*NA*
		5226925	rs33915217	β+thal	CC	99	CA	0	AA	0	1020	*NA*
		5226932	rs35578002	G > T	GG	98	GT	1	TT	0	1020	*NA*
		5226943	rs33950507	HbE	CC	99	CT	0	TT	0	1020	100
		5226963	rs33972047	β+thal	TT	99	TC	0	CC	0	1020	*NA*
		5226966	rs35382661	A > C	AA	97	AC	2	CC	0	1118	*NA*
		5227002	rs334	HbS	TT	90	TA	4	AA	5	2003	100

		5227003	rs33930165	HbC	CC	99	CT	0	TT	0	2003	100
		5227013	rs713040	β+thal; HPFH[c]	AA	4	AG	27	GG	65	2002	*NA*
		5227072	rs386134236	A > G	AA	98	AG	1	GG	0	984	*NA*
ACKR1	1	159204646	*NA*	A > C	AA	76	AC	19	CC	0	644	*NA*
		159204893	rs2814778	Fy(a-b-) T > C	TT	0	TC	0	CC	95	623	100
Dantu	4	143781321	rs186873296	Intergenic A > G	AA	95	AG	3	GG	0	132	99.0
		143781342	*NA*	G > T	GG	98	GT	1	TT	0	184	*NA*

Table 2: Allele frequencies of variants associated with malaria disease severity on *HBB, G6PD*, Dantu, and *ACKR1* loci in Northeast Tanzanian, African, European, and global human populations.

Gene	rsID	Amplicon Seq[1,2]		Controls[1]	Cases[1, 2]	African		European		Global	
		%	*n*	(*n* = 477)	(*n* = 506)	%	*n*	%	*n*	%	*n*
G6PD	rs1050828	18.4	87	20.0	16.3	13.5	1003	0	766	3.8	3775
	rs1050829	44.8	87	38.5	37.4	33.8	1003	0.4	766	9.5	3775
	rs5030872	0	99	0	0	1.1	3712	0	69444	<0.1	79538
	rs137852328	0	99	0	0	0	2714	<0.01	13108	<0.01	17548
	rs76723693	0	99	0	0	1.0	1003	0	766	3.2	3775
	rs5986875	1.0	99	*NA*	*NA*	3.6	1003	0	766	1.0	3775
	rs137852327	0	99	*NA*	*NA*	0	1003	0	766	0.2	3775
	rs2230036	12.1	99	*NA*	*NA*	12.2	1003	0	766	3.3	3775
HBB	rs334	9.1	99	16.5	2.0	10.0	1322	0	1006	2.7	5008
	rs33930165	0	99	0	0	1.3	1322	0	1006	0.3	5008
	rs33950507	0	99	0	0	0	1322	0	1006	0.3	5008
	rs33972047	0	99	*NA*	*NA*	0	1690	0	13150	0	15924
	rs33915217	0	99	*NA*	*NA*	0	1322	0	1006	0.1	5008
	rs713040	95.8	96	*NA*	*NA*	88.4	1322	83.0	1006	71.4	5008
ACKR1	rs2814778	100	95	*NA*	*NA*	96.4	1322	0.6	1006	26.6	5008
Dantu	rs186873296	3.1	98	*NA*	*NA*	0.4	1322	0.1	1006	0.1	5008

[1] Muheza, Tanzania

[2] Clinically severe malaria (described in Methods)

Table 3: Frequency of G6PD genotypes* among male and female children with severe malaria from Tuele Hospital (Muheza, Tanzania) determined using amplicon sequencing as a method of genotyping.

Gender	Genotype	Frequency N (%)	
Male	B	25	58.1
	A+	13	30.2
	A-	5	11.6
Female	BB	21	52.5
	BA+	6	15.0
	BA-	2	5.0
	A+A+	7	17.5
	A+A-	1	2.5
	A-A-	3	7.5
Overall	A- and A-A-	8	9.6

*G6PD Genotype: male normal = A+ or B; male hemizygous = A-; female normal = BB or BA+ or A+A+; female heterozygous = BA- or A+A-; female homozygous = A-A-. Severe G6PD deficiency genotypes = A- and A-A.

FIGURES

Figure 4: Amplicon targets for (A) atypical chemokine receptor 1 (ACKR1; n=1), (B) the intergenic Dantu genetic blood variant SNP (n=1), (C) haemoglobin subunit beta (HBB; n=1), and (D) glucose-6-phosphate dehydrogenase (G6PD; n=4). Four fragments for G6PD were designed to encompass five SNPs of clinical relevance, while one fragment designed for each of the three remaining regions of interest.

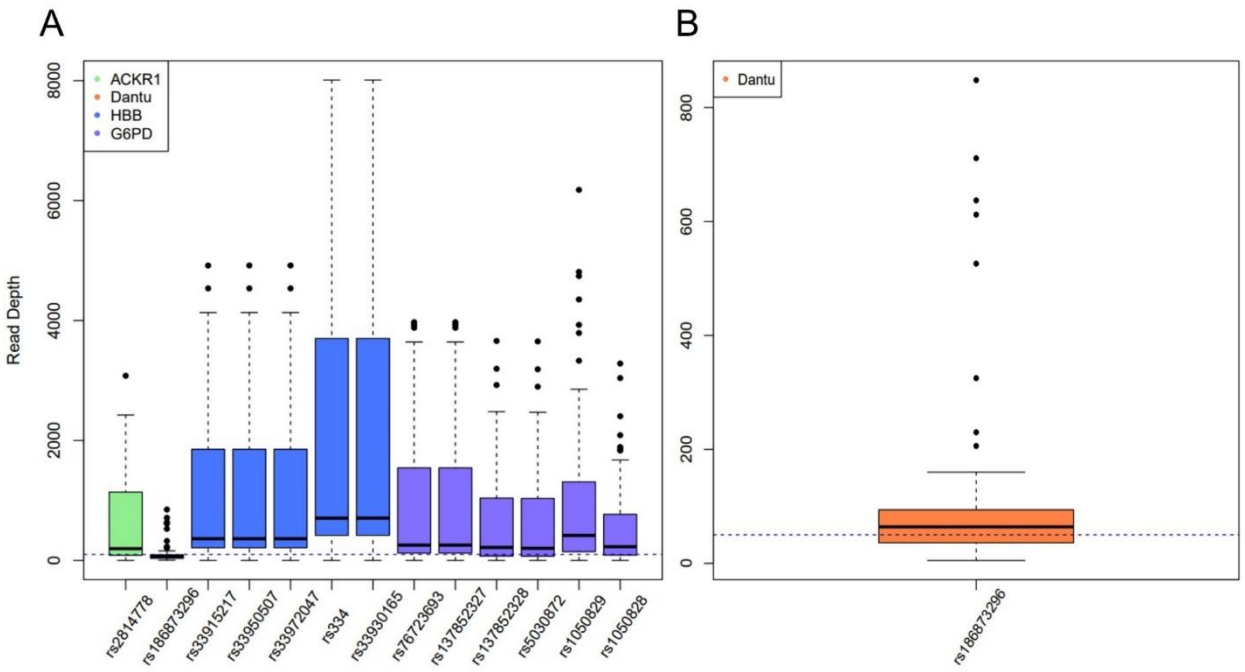


Figure 5: Read depth of SNPs with known associations to malaria disease severity. (A) Read depth of relevant SNPs on ACKR1, Dantu, *HBB*, and *G6PD*; trendline at 100. (B) Read depth of rs186873296, the intergenic Dantu genetic blood variant; trendline at 50.

SUPPLEMENTARY INFORMATION

Table S1: Amplicon coverage across 98 host DNA samples. Average, minimum, and maximum read count per target amplicon.

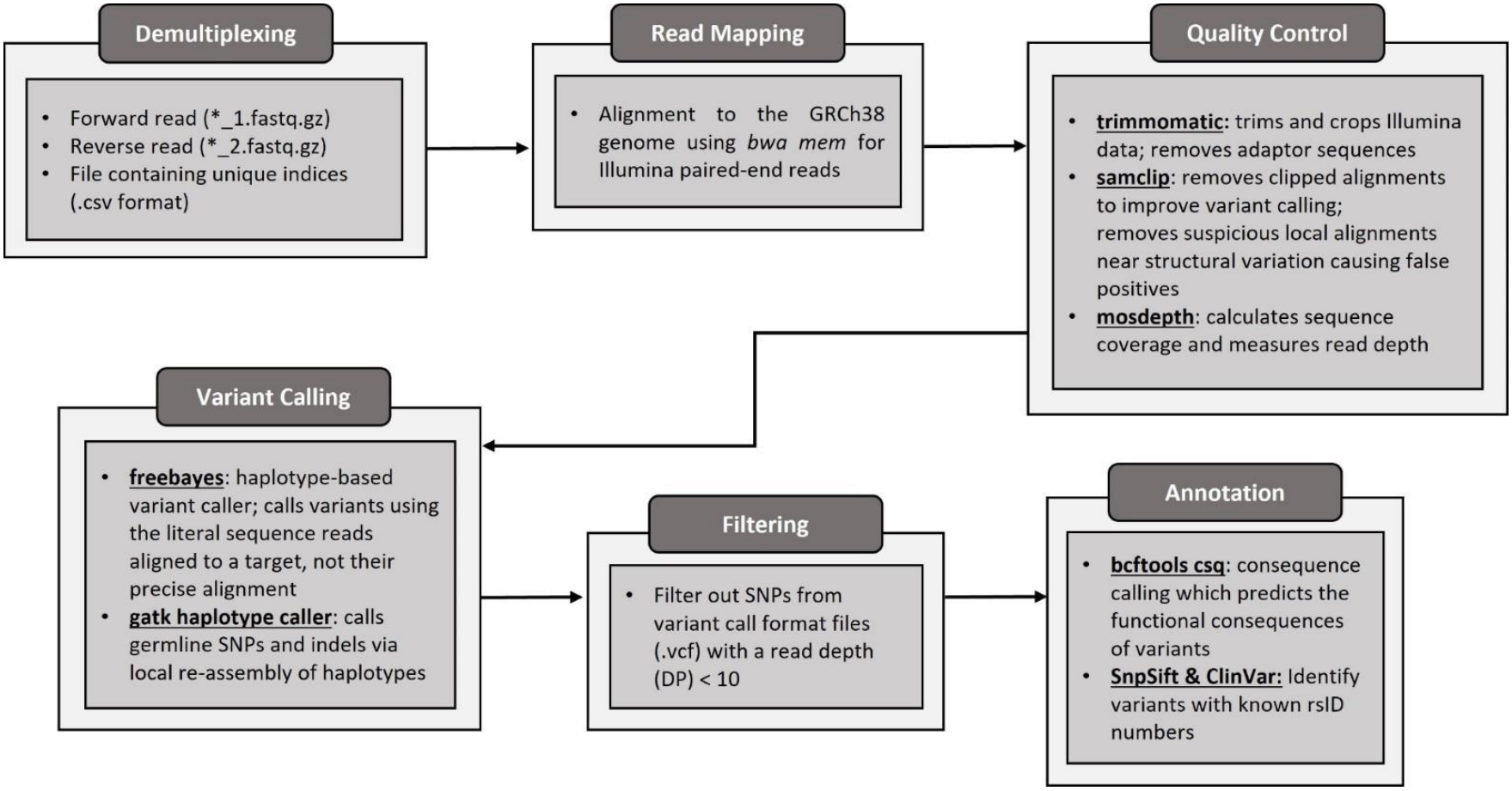
Amplicon	Average (*n* = 99)	Minimum	Maximum
G6PD-1	1089	21	3857
G6PD-2	973	9	3572
G6PD-3	1212	50	3890
G6PD-4	1548	77	4216
HBB	2195	112	5917
ACKR1	1226	97	5292
Dantu	184	5	2729

Table S2: Primers

Gene	RS SNP ID	Primer Sequence
ACKR1	rs2814778	Forward: TGTGCTTGAAGAATCTCTCCTT
		Reverse: CAGGGGAAATGAGGGGCATAG
G6PD-1	rs1050828	Forward: CTGGTAGAGAGGGCAGAACC
		Reverse: GACATGCTTGTGGCCCAGTA
G6PD-2	rs1050829	Forward: CGCTCATAGAGTGGTGGGAG
		Reverse: CACTGACTTCTGAGGGCACC
G6PD-3	rs5030872, rs137852328, rs137852314, rs5030868, rs137852330	Forward: GGATAACGCAGGCGATGTTG
		Reverse: TGATCCTCACTCCCCGAAGA
G6PD-4	rs76723693, rs137852327	Forward: GCCGGCCACATCATGGAAC
		Reverse: CAACTCAACACCCAAGGAGCC
HBB	rs334, rs33950507, rs33972047, rs33930165, rs33941377, rs33944208	Forward: TGGGAAAATAGACCAATAGGCAGA
		Reverse: AAGGACAGGTACGGCTGTCA
*Dantu**	rs186873296	Forward: GCAGATTAGCATTCACCCAG
		Reverse: TGCTCCAGAGTAAGCATCCTTC

*Intergenic region

Figure S1: Flow chart describing the bioinformatic pipeline of analysis from raw sequencing reads to variant calling.



Chapter 6: Discussion and Conclusions

Discussion

As disease surveillance moves towards an era of large-scale genomic studies and genetically tailored medicine due to advancements in next-generation sequencing technologies, implementation of these methodologies in malaria control programmes is becoming increasingly accessible. Plasmodium species exhibit remarkably flexible genomes that allow them to adapt and evolve to any range of selective pressures within their environment, from host immune evasion and the acquisition of protective phenotypes to the development of a wide array of drug resistance mechanisms. Malaria is a dynamic parasitic infection that will require a multifaceted approach towards control, encompassing not only the parasite and vector-based methods, but also increased insight into the human genetics of malaria infection protection and risk. As efforts to reduce the burden of malaria on public health systems in low- and middle-income countries continue to face decreasing returns in effectiveness, malaria control programme design will need to adapt by implementing these next-generation sequencing technologies.

This thesis explores the use of whole genome sequencing (WGS), and genome-wide population analyses, as a method of characterising the genomic diversity of *P. falciparum* parasite populations within East Africa, as well as the implementation of custom dual-indexing technology to screen specific genetic targets of interest across parasites and their human hosts. A large portion of genome studies aimed at characterising the genetic diversity of *P. falciparum* populations around the world have focussed on microsatellite markers or highly variable surface proteins, such as *msp1* and *msp2*, as methods of measuring genetic diversity [1–4]. While invaluable for providing baseline genetic diversity metrics and assessing overall changes in diversity associated with malaria transmission levels, these methods are not able to provide high resolution insights into the dynamics or genetic structure of a population as they ignore large proportions of the genome in their calculations, whereas WGS studies make use of genomic information from the entire genome [5]. Understanding the full genomic diversity of parasite populations, rather than concentrating on specific loci, makes it possible to identify new areas under selection, such as those associated with drug resistance mechanisms or immune evasion, as well as capture diversity in newly sequenced populations that may not align with specific loci identified in past studies. Although the benefits of implementing WGS-based analyses within a region of high transmission cannot be understated, the cost of sequencing can still be cost prohibitive in low-resource settings and suggests the supplemental use of low-cost, targeted, dual-indexing methodologies to achieve high throughput coverage of genes of interest, including parasite drug resistance markers [6].

Despite being a reservoir of intense malaria transmission, *P. falciparum* parasite populations within the Lake Victoria region of Kenya have been largely underrepresented in genetic and genomic diversity studies. Previous research within the Lake Victoria region, using microsatellite markers, identified high levels of genetic diversity and complex infections, consistent with regions of high transmission, as well as low levels of genetic differentiation, however parasite populations were not compared to populations from other regions [2]. In **Chapter 2**, I produced the first assessment of parasite population dynamics and genomic diversity within the Kenyan Lake Victoria basin and how they compare to wider African populations using WGS. Performing WGS makes it possible to characterise and track the flow of genetic material across populations, predominantly through human migration and movement, as well as monitor the spread of adaptive alleles, including both known and novel alleles, through IBD signatures and extended haplotype heterozygosity metrics. I generated WGS data for 48 isolates from Mfangano island, Ngodhe island, and Suba district to compare to publicly available sequences from nine countries across the African continent (N = 736), including isolates from nearby Kisumu and Kombewa. An obstacle for this study was the use of asymptomatic field isolates with low levels of parasite DNA, which resulted in reduced success in generating WGS data for isolates across the samples sites. Although they account for most malaria infections worldwide, asymptomatic infections are not well characterised in existing public datasets due to these low DNA concentrations. Symptomatic infections showcase parasites causing a majority of the high burden of disease, while asymptomatic infections account for most infections around the world. Screening parasites causing asymptomatic infections is essential to our understanding of parasite populations as a whole.

To mitigate the loss of high-quality WGS data, parasite DNA was amplified using SWGA methodologies that select for and amplify *P. falciparum* DNA within the total DNA extracted from a dried blood spot [7–9]. SWGA increases the available *P. falciparum* DNA template to improve WGS success, however the method tends to amplify the most dominant parasite clone within a sample and can result in missing information in populations with high rates of mixed infections, such as the Lake Victoria basin. Comparison of genome-wide SNPs identified within the WGS dataset revealed that *P. falciparum* isolates from Lake Victoria form a genetic cluster within the larger East African population and island isolates, along with isolates from Suba district, appear to form a sub-group within the Lake Victoria subpopulation. My results confirm the initial findings made by Mulenge et al, identifying low levels of genetic differentiation within the region, likely due to high rates of human movement throughout the area and close geographical proximity between sample sites, as well as high levels of genetic diversity and complex infections, commonly observed within regions of intense transmission [2]. To place *P. falciparum* isolates from the Lake Victoria region within a wider context, and build upon previous findings, I carried out cross-population analyses with parasite populations from the across the African

continent. A neighbour-joining tree and PCA highlighted the distinct sub-grouping dynamics of parasites from Lake Victoria and admixture analysis revealed that isolates contained high proportions of ancestral genome fragments from both Central African and East African parasite populations. This observation aligns well with the current prevailing *P. falciparum* origin hypothesis, suggesting that malaria first emerged in Central Africa before spreading out through early human migration, with subsequent more recent migration events throughout the African continent. Further evidence of structural variation within the population was identified using fixation index statistics and regions under positive directional selection were identified on genes associated with host immune response (e.g., *Pfmsp3*) and erythrocyte invasion (e.g., *Pfama1* and PMX). Finally, known drug resistance markers (e.g., *Pfcrt* [K76T]; *Pfdhfr* [N51I, S108N]; *Pfdhps* [K540E]) were identified within the population at similar frequencies to other East African populations.

Although SWGA was applied in **Chapter 2** to increase parasite DNA and high-quality sequencing data, samples from Ngodhe island (N = 18; population size 600 – 1,000) had disproportionately high rates of WGS failure (1 out of 18 samples were successfully sequenced) as a majority of isolates were obtained from low-density infections. Asymptomatic and sub-microscopic infections account for a majority of the cases detected on Ngodhe island, making it difficult to widely implement WGS-based surveillance tools [10]. WGS-based techniques, in addition to demonstrating limited efficacy when processing with low-density infections, can also be cost-prohibitive in low resource settings, such as this one, where timely disease surveillance is crucial in making informed disease control policies. To overcome this limitation, I demonstrated the feasibility of using a low-cost sequencing method, that makes use of custom dual-indexing (i.e. DNA barcoding) amplicon sequencing technology, on low-density and asymptomatic *P. falciparum* infections to generate a drug resistance profile of Ngodhe island for the first time [6] (**Chapter 3)**. The custom barcoding component of this technique allows for isolates to be multiplexed prior to sequencing, reducing both the time and costs associated with high throughput sequencing (less than $0.50 per amplicon). Out of 102 samples with a DNA concentration ≥ 0.4 ng/µl, 70 samples were successfully sequenced using Illumina short-read sequencing for one or more of the nine amplicon targets covering genes with known drug resistance-conferring associations (*PfK13, Pfcrt, Pfmdr1, Pfdhps,* and *Pfdhfr*). Sequencing results identified reduced markers of chloroquine resistance in Ngodhe island parasite populations compared to Mfangano island and other East African isolates, however variants associated with resistance to sulphadoxine-pyrimethamine (SP) were observed to be nearly fixed within the island population alongside variants that may reduce the efficacy of the ACT partner drug lumefantrine. The affordability of this technique makes it a more attainable method for carrying out high throughput drug resistance surveillance in low-resource

settings, including regions with low-density infections and those pursuing elimination where the threat of drug resistance could derail elimination efforts.

East Africa accounts for 10% of the world's malaria cases annually and clinically artemisinin resistant parasite populations have recently been identified to be locally spreading within Uganda, as well as neighbouring Rwanda [11–13]. Although a large public repository of sequencing data is available from sites across the region, apart from Uganda which has limited WGS data availability, an in-depth genome-wide analysis of parasite population dynamics within East Africa has not been conducted. To build upon my work within the Lake Victoria basin and further investigate distinct *P. falciparum* populations within the region, I generated sequencing data for 30 samples from Bungoma county in Western Kenya, located along the Kenya-Uganda border (**Chapter 4)**. This sequencing data was combined with 557 isolates from across East Africa, including the Lake Victoria regions of Kenya and Tanzania, Central Uganda, Eastern Kenya, North East Tanzania, South East Tanzania, and Lake Tanganyika in Tanzania to produce a robust dataset of 710,552 high quality genome-wide SNPs. Isolates from Western Kenya were identified to have a marked reduction in chloroquine resistance associated variants compared to surrounding populations within the Lake Victoria basin as well as a variant on *Pfk13* within a codon of concern flagged by the WHO as a candidate for artemisinin resistance. A maximum likelihood tree and PCA of the expanded dataset confirmed and explored the distinct genetic structure of *P. falciparum* isolates from the Kenyan region of Lake Victoria, alongside Western Kenya and Central Uganda, initially described in **Chapter 2**, as well as identified parasite subpopulations within East Africa, including a distinction between Lake Victoria isolates from Kenya with those from Tanzania.

Admixture analysis carried out in **Chapter 2** highlighted human migration trends consistent with the theory that *P. falciparum* originated within Central Africa and spread during multiple human migration events, which had also been showcased large-scale WGS studies across the African continent [14]. I hypothesised that a greater density of sampling may provide higher resolution insights into the co-evolution of host and pathogen with links to continent-wide, and global, human migration events. To assess the ancestral origins of subgroups appearing to display distinct genetic structure within East Africa, as well as assess population differentiation drivers, a larger dataset was prepared using samples from both East Africa and regional population from across the African continent, including West Africa, Horn of Africa, Central Africa, South Central Africa, and Southern Africa. A dataset of 640,596 SNPs identified diverse ancestral origins for East African subpopulations, including an ancestral population distinct to isolates from Western Kenya, the Lake Victoria region of Kenya, and Central Uganda, as well as an ancestral population distinct to isolates from Lake Tanganyika. Isolates from Western Kenya, Central Uganda, and the Kenya region of Lake Victoria were identified to share high proportions of the

same, seemingly independent, ancestral genome with one another when compared to isolates from other East African and African populations. This independent ancestry may be explained by the migration of the Luo people into Western Kenya and Uganda from South Sudan, whereas the remainder of East Africa and the Lake Victoria basin was largely settled by Bantu peoples originating from Central Africa [2, 14]. This differentiates somewhat from preliminary results I generated, suggesting a higher proportion of ancestry within Lake Victoria isolates donated from Central African lineages when compared to other East African populations. WGS data from populations such as Uganda, Rwanda and Burundi, which is currently not publicly available, may provide greater insight into the role of human migration on *P. falciparum* population dynamics within this region and, perhaps, inform theories pertaining to the origin of *P. falciparum* in human populations. To assess population differentiation and the spread of adaptive alleles between ancestral populations, I quantified Identity by descent (IBD) signatures and extended haplotype heterozygosity metrics for positive selection. Both IBD and positive selection identified strong conservation of *Pfdhps* across East African populations, when compared to wider African populations, and high selection pressures corroborated by high rates of SP resistance within parasite populations. Cross-population analyses identified a gene theorised to be associated with ACT treatment failures within two East African subpopulations [15].

The dynamic nature of *P. falciparum* to adapt and infect its human host has led to a proverbial arms race between parasite and human genetics. Malaria has exhibited the strongest known selective pressure on the human genome in recent history and is the driving force behind an assortment of human genetic variants associated with protection against severe disease, as well as pathogenic mutations with the potential to alter the efficacy of malaria control programmes within a region [16, 17]. Multi-population GWAS studies have identified genetic variants with known associations to malarial disease outcome, as well as potential candidate variants that may impact disease severity [17, 18]. However, high genetic diversity between African populations and ethnic groups, as well as an underrepresentation of African populations in genetic studies, has limited our ability to thoroughly estimate risk within a population when implementing malaria control programmes, particularly when it comes to MDAs [19, 20]. Leveraging off the technology presented in **Chapter 3**, I demonstrate the versatility of high throughput amplicon sequencing by presenting a proof-of-concept method for profiling human genetic determinants of malarial disease in an at-risk population in Northeast Tanzania. This methodology cheaply targeted regions on genes while simultaneously eliminating the need to perform WGS or store sensitive human genomic data from at-risk populations.

To determine the efficacy and accuracy of this method, I screened 100 individuals diagnosed with severe malaria in Northeast Tanzania for genetic variants on *G6PD, HBB,* and *ACKR1*, as well as the

intergenic Dantu blood variant, using primers I designed to target the primary variants of concern within those regions. Genotyping data for all 100 individuals had been previously reported for variants with strong associations for determining malaria disease severity which allowed for the accuracy of amplicon sequencing and variant calling to be assessed. Genotyping accuracy ranged from 96.6% to 100% for variants previously characterised (G6PD 202 and 376, HbS, HbC, Duffy-negative blood group and Dantu blood group). As this technique amplifies the region surrounding the variant of concern on a gene, additional variants that were not provided in previously available GWAS data were also able to be identified, providing greater insight into the genetic profile of the study population. Although the regions targeted in this study account for most of the genetic risk surrounding malaria disease severity and protection, a handful of other targets were not included, such as the ABO blood-grouping gene and Interleukin genes. The ABO was not included in this assay due to difficulties with primer design associated with the high degree of recombination and number of variants that occur within the region, although further primer optimisation may resolve this issue [21]. This technique has been designed to determine population level genetic diversity rather than for individual clinical applications. More in-depth testing of this assay across a larger population from varying geographic locations would be required to determine its suitability for clinical use.

Conclusion

The flexible application of diverse sequencing-based techniques to address region-specific profiles of malaria transmission will be essential in regaining traction with malaria elimination strategies around the world. This thesis presents a multifaceted approach to disease surveillance by using both whole genome sequencing and custom targeted sequencing to characterise the genomic and genetic diversity of *P. falciparum* populations in high transmission regions of East Africa, as well as the genetic determinants of malarial disease in at-risk human populations. I have demonstrated the potential of these technologies to not only generate robust depictions of *P. falciparum* population dynamics within East African subpopulations, but also address the discrepancies in genetic data available for human populations in sub-Saharan Africa while being mindful of the associated ethical concerns.

Future of sequencing-based approaches in malaria elimination

As sequencing-based approaches continue to advance and the associated costs are reduced, the potential application of these techniques in real-time will expand immensely and they will soon become an integral part of malaria control programmes and malaria elimination campaigns around

the world. Between 2020 and 2021, following initial lockdowns surrounding the COVID-19 global pandemic, the most dramatic increase in malaria incidence rates were recorded since the start of the millennium, exceeding the previous years' cases and deaths by 22 million and 57,000, respectively [13]. Although most malaria control programmes have resumed normal operations since the end of strict COVID-19 precautions, the number of cases and deaths have remained elevated and highlight that a reduced rate of progress in malaria incidence reduction leading up to the pandemic has persisted. These trends point to the fact that the fundamental components of malaria disease control, such as vector control, rapid detection and treatment of cases, and the use of preventative treatment seasonally or in pregnant women, are not enough to reach malaria elimination on their own.

Despite the negative implication of recent trends in malaria incidence, they have forced positive movement within the field of malaria research and are driving the investigation and implementation of new surveillance tools, such as those described in this thesis, as well as investment into genomics capacity in low-resource settings. Oxford Nanopore Technology (ONT) portable platforms (e.g., MinION) has introduced sequencing devices that allow long-read sequencing to be generated on site in areas of the world that lack established sequencing facilities. This type of portable sequencing could be implemented in a variety of settings, including clinical facilities in low-resource regions where the rapid identification of an infections' causative species, or the presence of drug resistance markers that may impact treatment outcome, could drastically influence morbidity and mortality rates associated with malaria [22]. Utilisation of portable sequencing technology alongside high throughput amplicon sequencing could revolutionise disease surveillance by providing real-time insight into not only parasite populations or human host genetics of any region of the world, but also the presence of insecticide resistance in vector species that may undermine ongoing control programmes [23]. The molecular mechanisms of insecticide resistance remain poorly categorised due to limited sequence availability, largely associated with the immense size of vector genomes, and diversity of sample isolates. Implementation of long read ONT in large-scale studies across high transmission regions could provide detailed insight into vector genomes and identify targets more easily screened using a high throughput amplicon approach [22, 23].

In **Chapters 2** and **4**, I discuss the distinct genetic structure of *P. falciparum* populations within the Kenyan region of Lake Victoria and identify a handful of subpopulations within East Africa with their on genetic structure. Despite providing insights into the diverse ancestral origins of parasite populations within the region, the specific causes behind the genetic differentiation occurring between Tanzanian Lake Victoria isolates and Kenyan Lake Victoria isolates remains unidentified. Expanded sampling of regions in and around the Lake Victoria basin, including the largely unrepresented populations of Uganda, would provide valuable sequencing data necessary to discern

drivers behind these distinct parasite lineages. Isolates from Western Kenya, the Kenyan region of Lake Victoria, and Central Uganda also appeared to share high proportions of the same ancestral genome unique to the Lake Victoria region. Interestingly, portions of this ancestral genome were donated to parasite populations from across the African continent. Central African isolates have been characterised to share high proportions of their ancestral genome with other regions of Africa and supports the prevailing theory that *P. falciparum* infections in humans occurred following a cross-species event involving Western Gorillas [24, 25]. Further sampling within this region and an expanded dataset of *P. falciparum* isolates from Central Africa may provide insight into the relationship between Central and East African parasite populations, as well as the potential insight into the origin and spread of malaria throughout Africa. Large-scale studies of *P. falciparum* WGS across global populations are difficult to implement and carry out due to the sheer volume of data, however, as technologies improve, understanding global population dynamics using genomic data may provide increased us with a better understanding of the spread of *P. falciparum* out of Africa during human migrations and how parasite populations have changed and adapted to environments around the world. These large-scale analyses may also prove useful for understanding the dynamics of transmission hot spots in environments pursuing elimination.

In **Chapter 5,** I present amplicon sequencing as a method of screening for genetic markers associated with malaria disease outcomes in at-risk populations, as well as describe the disparity in genetic information available for human populations in sub-Saharan Africa [20]. The versatility of this technique to be expanded to include additional markers of disease or identify potential vaccine candidates, specific to or independent from malaria, as well as be applied in low-resource settings, cannot be overstated. One of the biggest barriers to expanding molecular surveillance in disease control programmes within sub-Saharan Africa is the lack of infrastructure and development available for quickly generating and processing sequencing data [20]. Capacity is being strengthened, through the development of new genomic centres and training of bioinformaticians, however, to date, general practice for generating and analysing sequencing data for both malaria parasites and humans has been to export samples to higher income countries with access to the necessary resources. This established workflow has largely fostered the technological divide and stunted the growth of sequencing infrastructure development in low- and middle-income countries, as well as raise ethical concerns regarding the ownership and secure storage of human genomic data. Ideally, mining the African genome for the benefit of African peoples should be carried out and managed in-country [26]. Further optimisation of custom dual-indexing technology to maximise the number of samples that can be sequenced in a single run would continue to reduce costs associated with this technique. Additionally the implementation of portable sequencing technology, such as ONT MinION platform, could be the

push that is needed to bring next-generation sequencing to academic institutions and clinical service providers in sub-Saharan Africa [22].

The work presented in this thesis has the potential to be improved upon by the vast assortment of next-generation sequencing techniques continuing to emerge. These technologies, combined with investment in infrastructure in sub-Saharan Africa, can be used to usher in a new age of malaria control programmes encompassing all aspects of the malaria parasite lifecycle, ultimately leading to global malaria elimination.

References

1. Somé AF, Bazié T, Zongo I, Yerbanga RS, Nikiéma F, Neya C, Taho LK, Ouédraogo J-B (2018) Plasmodium falciparum msp1 and msp2 genetic diversity and allele frequencies in parasites isolated from symptomatic malaria patients in Bobo-Dioulasso, Burkina Faso. Parasites & Vectors 11:323

2. Mulenge FM, Hunja CW, Magiri E, Culleton R, Kaneko A, Aman RA (2016) Genetic Diversity and Population Structure of Plasmodium falciparum in Lake Victoria Islands, A Region of Intense Transmission. Am J Trop Med Hyg 95:1077–1085

3. Agaba BB, Anderson K, Gresty K, et al (2021) Genetic diversity and genetic relatedness in Plasmodium falciparum parasite population in individuals with uncomplicated malaria based on microsatellite typing in Eastern and Western regions of Uganda, 2019–2020. Malaria Journal 20:242

4. Moser KA, Madebe RA, Aydemir O, et al (2021) Describing the current status of Plasmodium falciparum population structure and drug resistance within mainland Tanzania using molecular inversion probes. Mol Ecol 30:100–113

5. Daniels RF, Schaffner SF, Wenger EA, et al (2015) Modeling malaria genomics reveals transmission decline and rebound in Senegal. Proc Natl Acad Sci USA 112:7067–7072

6. Nag S, Dalgaard MD, Kofoed P-E, Ursing J, Crespo M, Andersen LO, Aarestrup FM, Lund O, Alifrangis M (2017) High throughput resistance profiling of Plasmodium falciparum infections based on custom dual indexing and Illumina next generation sequencing-technology. Sci Rep 7:2398

7. Oyola SO, Ariani CV, Hamilton WL, et al (2016) Whole genome sequencing of Plasmodium falciparum from dried blood spots using selective whole genome amplification. Malaria Journal 15:597

8. Ibrahim A, Diez Benavente E, Nolder D, et al (2020) Selective whole genome amplification of Plasmodium malariae DNA from clinical samples reveals insights into population structure. Scientific Reports 10:10832

9. Benavente ED, Gomes AR, De Silva JR, et al (2019) Whole genome sequencing of amplified Plasmodium knowlesi DNA from unprocessed blood reveals genetic exchange events between Malaysian Peninsular and Borneo subpopulations. Scientific reports 9:9873

10. Kagaya W, Gitaka J, Chan CW, Kongere J, Idris ZM, Deng C, Kaneko A (2019) Malaria resurgence after significant reduction by mass drug administration on Ngodhe Island, Kenya. Sci Rep 9:1–11

11. Balikagala B, Fukuda N, Ikeda M, et al (2021) Evidence of Artemisinin-Resistant Malaria in Africa. New England Journal of Medicine 385:1163–1171

12. Uwimana A, Legrand E, Stokes BH, et al (2020) Emergence and clonal expansion of in vitro artemisinin-resistant Plasmodium falciparum kelch13 R561H mutant parasites in Rwanda. Nat Med 26:1602–1608

13. Geneva: World Health Organization (2022) World malaria report 2022. Licence: CC BY-NC-SA 3.0 IGO

14. Amambua-Ngwa A, Amenga-Etego L, Kamau E, et al (2019) Major subpopulations of Plasmodium falciparum in sub-Saharan Africa. Science 365:813–816

15. Henriques G, Hallett RL, Beshir KB, et al (2014) Directional selection at the pfmdr1, pfcrt, pfubp1, and pfap2mu loci of Plasmodium falciparum in Kenyan children treated with ACT. J Infect Dis 210:2001–2008

16. Kwiatkowski DP (2005) How Malaria Has Affected the Human Genome and What Human Genetics Can Teach Us about Malaria. The American Journal of Human Genetics 77:171–192

17. Manjurano A, Sepulveda N, Nadjm B, et al (2015) African Glucose-6-Phosphate Dehydrogenase Alleles Associated with Protection from Severe Malaria in Heterozygous Females in Tanzania. PLoS Genet. https://doi.org/10.1371/journal.pgen.1004960

18. Ravenhall M, Campino S, Sepúlveda N, et al (2018) Novel genetic polymorphisms associated with severe malaria and under selective pressure in North-eastern Tanzania. PLoS Genet 14:e1007172

19. Auton A, Abecasis GR, Altshuler DM, et al (2015) A global reference for human genetic variation. Nature 526:68–74

20. Tucci S, Akey JM (2019) The long walk to African genomics. Genome Biology 20:130

21. Seltsam A, Hallensleben M, Kollmann A, Blasczyk R (2003) The nature of diversity and diversification at the ABO locus. Blood 102:3035–3042

22. Jain M, Olsen HE, Paten B, Akeson M (2016) The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. Genome Biology 17:239

23. Campos M, Phelan J, Spadar A, et al (2022) High-throughput barcoding method for the genetic surveillance of insecticide resistance and species identification in Anopheles gambiae complex malaria vectors. Sci Rep 12:13893

24. Loy DE, Liu W, Li Y, Learn GH, Plenderleith LJ, Sundararaman SA, Sharp PM, Hahn BH (2017) Out of Africa: origins and evolution of the human malaria parasites Plasmodium falciparum and Plasmodium vivax. Int J Parasitol 47:87–97

25. Liu W, Li Y, Learn GH, et al (2010) Origin of the human malaria parasite Plasmodium falciparum in gorillas. Nature 467:420–425

26. Vries J de, Pepper M (2012) Genomic sovereignty and the African promise: mining the African genome for the benefit of Africa. Journal of Medical Ethics 38:474–478