

LONDON  
SCHOOL of  
HYGIENE  
& TROPICAL  
MEDICINE



## LSHTM Research Online

Ravenhall, M; (2019) A bioinformatic analysis of malaria host and pathogen genomics. PhD (research paper style) thesis, London School of Hygiene & Tropical Medicine. DOI: <https://doi.org/10.17037/PUBS.04653632>

Downloaded from: <https://researchonline.lshtm.ac.uk/id/eprint/4653632/>

DOI: <https://doi.org/10.17037/PUBS.04653632>

### Usage Guidelines:

Please refer to usage guidelines at <https://researchonline.lshtm.ac.uk/policies.html> or alternatively contact [researchonline@lshtm.ac.uk](mailto:researchonline@lshtm.ac.uk).

Available under license. To note, 3rd party material is not necessarily covered under this license: <http://creativecommons.org/licenses/by-nc-nd/3.0/>

LONDON  
SCHOOL *of*  
HYGIENE  
& TROPICAL  
MEDICINE



**A bioinformatic analysis of malaria  
host and pathogen genomics**

**Matthew Warwick Ravenhall**

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

**University of London**

**May 2019**

**Department of Pathogen Molecular Biology**

**Faculty of Infectious Tropical Disease**

**LONDON SCHOOL OF HYGIENE & TROPICAL MEDICINE**

Funded by: BBSRC

Research group affiliation(s): Taane G. Clark & Susana Campino

I, Matthew Warwick Ravenhall, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirmed that this has been indicated in this thesis.

Signed \_\_\_\_\_

Date \_\_\_\_\_

## Abstract

Malaria is a significant global disease caused by infection with parasites of the *Plasmodium* genus, which resulted in an estimated 216 million cases and 445,000 deaths in 2016 alone. Co-evolution of *Plasmodium* parasites and their human hosts has shaped both genomes for thousands of years. In this thesis I describe my work identifying and characterising novel genomic variants and selection signals associated with host-pathogen interactions in malaria. For the parasite, analysis of the impact of sustained sulfadoxine/pyrimethamine (SP) use on the Malawian parasite population (n=220) lead to the identification of selection signals associated with SP resistance factors and a novel 436 bp *gch1* promoter region duplication at near-fixation. Next a global approach to copy number variation discovery (n=3,110), based on short read sequencing, identified several novel and geographically specific variants including large 22.9 kbp duplications of *crt* in West Africa. Finally, an inversion discovery pipeline was developed for a long read based approach to inversion detection (n=17). This led to the identification of a novel ‘sandwich inversion’ of *pi4k* in a sample of GB4, similar to inversion-duplication of *gch1* in Dd2. For human genetics within the context of malaria, I conducted a GWAS with a Tanzanian dataset (n=914) and identified novel protective associations, such as for *IL-23R* and *IL-12RB2*. I also identified novel structural variants (SV) with a short-read sequencing based dataset of Tanzanian parent-child trios (n=234) identifying several novel SVs associated with blood antigen systems. Near-fixation deletions in *SEC22B* and *BET1L* were also identified, suggesting an impact on intracellular transportation. Genomic variation associated with host-pathogen interactions is diverse, and SVs represent one overlooked aspect that requires further investigation. Bioinformatic approaches can help identify novel variants but depend upon novel software development, such as those described within this thesis (e.g. SV-Pop).

## **Acknowledgements**

I would like to thank my supervisors Taane Clark and Susana Campino, and all members of ‘The Hub’ including Jody, Ben, Yaa, Neneh, Ernest, Matt H, Dan, Amy, and Pepita, as well as numerous others at LSHTM. My thanks also go to those involved in the collection and sequencing of the datasets analysed within this project, their hard work was the foundation upon which this thesis was built.

Further, I thank the wealth of interesting and enthusiastic students and staff associated with the London Interdisciplinary Doctoral Program (LIDo) and, of course, to the BBSRC for funding the programme. LIDo has been the perfect accompaniment to my research, and I hope that others continue to gain from it.

My thanks also go to colleagues at JPMorgan Chase & Co. whom I worked with during my PIPS including, but certainly not limited to, Sammy Assefa and Ignacio Navarro. Thank you for providing a fresh perspective on my work.

Finally, I would especially like to thank my friends, family, and Becky. You’re all awesome.

## **Additional Publications**

These represent manuscripts to which I made a significant contribution during my PhD.

They are not considered part of my core thesis.

*Ana Rita Gomes, Matt Ravenhall, Ernest Diez Benavente, Arthur Talman, Colin Sutherland,*

*Cally Roper, Taane G. Clark, Susana Campino. Genetic diversity of next generation*

*antimalarial targets: A baseline for drug resistance surveillance programmes. (2017)*

*International Journal for Parasitology: Drugs and Drug Resistance. DOI:*

*10.1016/j.ijpddr.2017.03.001*

*Rebecca Johnson, Matt Ravenhall, Derek Pickard, Gordon Dougan, Alexander Byrne and Gad*

*Frankel. Comparison of *Salmonella enterica* serovars Typhi and Typhimurium reveals*

*typhoidal-specific responses to bile. (2017) Infection and Immunity. DOI: 10.1128/iai.00490-17*

*Matthew Higgins, Matt Ravenhall, Daniel Ward, Jody Phelan, Amy Ibrahim, Matthew S*

*Forrest, Taane G Clark, Susana Campino. PrimedRPA: Primer design for Recombinase*

*polymerase amplification assays. (2018) Bioinformatics (Oxford, England). DOI:*

*10.1093/bioinformatics/bty701*

*Susana Campino; Alejandro Marin-Menendez; Alison Kemp; Nadia Cross; Laura Drought;*

*Thomas D. Otto; Ernest Diez Benavente; Matt Ravenhall; Frank Schwach; Gareth Girling;*

*Magnus Manske; Michel Theron; Kelda Gould; Eleanor Drury; Taane G. Clark; Dominic P.*

*Kwiatkowski; Alena Pance; Julian C Rayner, Ph.D. A forward genetic screen reveals a*

*dominant role for *Plasmodium falciparum* Reticulocyte Binding Protein Homologue 2a and 2b*

*in determining alternative erythrocyte invasion pathways. (2018). PLOS Pathogens. DOI: In*

*Press*

## Table of Contents

Chapter 1: Introduction .....	9
1.1    The global and historical impact of malaria .....	10
1.1.1    Malaria as a significant global disease.....	10
1.1.2    Aetiology of malaria .....	11
1.1.3    A history of human co-evolution with malaria parasites .....	12
1.1.4    Moving towards global eradication.....	13
1.2    The diagnosis, treatment, and control of malaria .....	15
1.2.1    Diagnosis of malaria infection .....	15
1.2.2    Current approaches for treatment and control.....	15
1.2.3    Chloroquine & Quinine-derived Drugs.....	18
1.2.4    Sulfadoxine/Pyrimethamine (SP) & Antifolate Compounds .....	19
1.2.5    Artemether/Lumefantrine & Artemisinin-based Combination Therapies.....	20
1.2.6    Existing and Emerging Resistance.....	22
1.3    The role of <i>Plasmodium</i> parasites .....	23
1.3.1    Species of <i>Plasmodium</i> .....	23
1.3.2 <i>Plasmodium</i> Life Cycle.....	25
1.3.3    The genomics of <i>P. falciparum</i> .....	26
1.3.4    Variation in <i>P. falciparum</i> reference strains .....	28
1.4    Human genetics in the context of malaria .....	29
1.4.1    Determinants of diversity of human response to infection .....	29
1.4.2    Severe malaria subtypes.....	30
1.4.3    Genetics of human susceptibility .....	30
1.5    Methods utilised in exploring genomics.....	33
1.5.1    High throughput sequencing .....	33
1.5.2    Genomic variation in host-pathogen interactions .....	34
1.5.3    Identifying signals of selection .....	35
1.5.4    Identifying structural variation.....	37
1.6    Project Outline.....	39
1.7    References .....	43
Chapter 2: Characterising the impact of sustained sulfadoxine/pyrimethamine use upon the <i>Plasmodium falciparum</i> population in Malawi .....	57
Chapter 3: A global analysis of copy number variation in <i>Plasmodium falciparum</i> identifies a novel duplication of the chloroquine resistance associated gene .....	81
Chapter 4: Analysis of global long read <i>Plasmodium falciparum</i> genomes identifies novel inversions .....	104
Chapter 5: Novel genetic polymorphisms associated with severe malaria and under selective pressure in North-eastern Tanzania.....	130
Chapter 6: Analysis of Tanzanian trios reveals inherited structural variants in genes with roles in ER-Golgi transport and blood antigen systems .....	151

Chapter 7: SV-Pop: Population-based structural variant analysis and visualisation ....	176
Chapter 8: Discussion, Conclusions, and Future Work .....	180
8.1    Discussion .....	181
8.2    Conclusions .....	188
8.3    Future Work .....	189
8.4    References .....	193

## Abbreviations

ACT	Artemisinin Combination Therapy
CNV	Copy Number Variant
CQ	Chloroquine
DEL	Deletion (structural variant)
DUP	Duplication (structural variant)
EHH	Extended Haplotype Homozygosity
FST	Fixation index
GO	Gene Ontology
GWAS	Genome Wide Association Study
iHS	Integrated Haplotype Score
INDEL	Insertion-deletions
INS	Insertion (structural variant)
INV	Inversion (structural variant)
NGS	Next Generation Sequencing
PacBio	Pacific Biosciences
PCA	Principal Component Analysis
PCR	Polymerase Chain Reaction
Pf	<i>Plasmodium falciparum</i>
RBC	Red blood cell
SNP	Single Nucleotide Polymorphism
SP	Sulfadoxine/Pyrimethamine
SV	Structural Variant
WHO	World Health Organisation
XP-EHH	Cross-Population Extended Haplotype Homozygosity

## **Chapter 1:**

### Introduction

## 1.1 The global and historical impact of malaria

### 1.1.1 Malaria as a significant global disease

Today nearly half the global population, approximately 3.7 billion people, is at risk of malaria, with over 90 countries reporting locally acquired transmission. Malaria disease burden is not equal across all countries with 15, 14 of which are in sub-Saharan Africa, accounting for 80% of the total global burden [1]. Malaria therefore remains a significant global health challenge, rivalling the impact of HIV, influenza, and tuberculosis. In 2016 there were an estimated 216 million cases of infection leading to 445,000 deaths [1]. The disease remains one of poverty and infancy with 90% of deaths occurring within Africa, where 99.7% of cases are caused by the *Plasmodium falciparum* parasite [1], and approximately two-thirds of deaths are in children below five years of age [1]. Beyond a significant human cost, the high mortality and morbidity of the disease has been estimated to account for \$12 billion of losses per year in Africa alone, further hampering individual well-being [2].

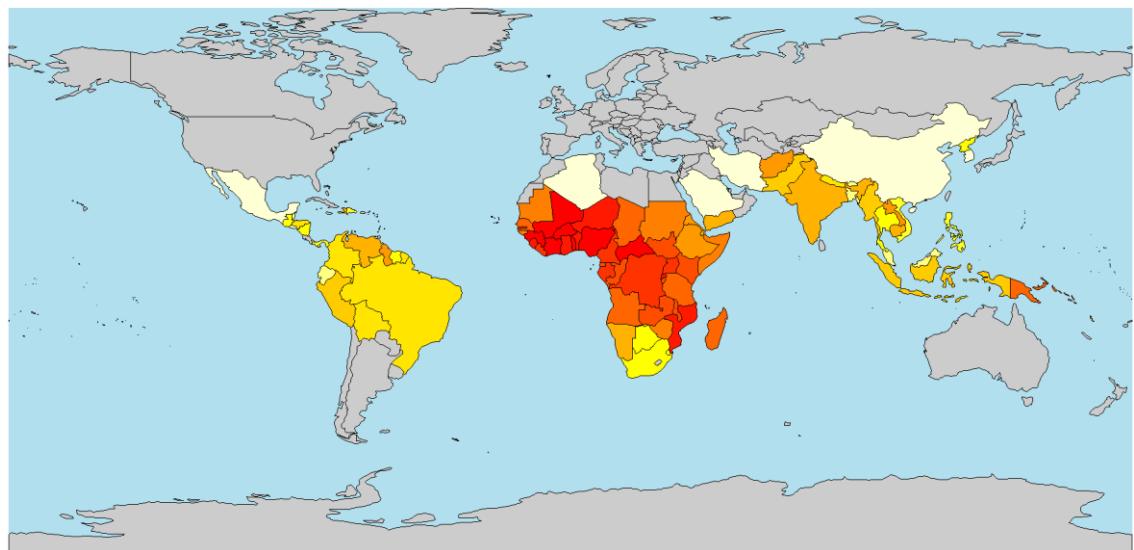


Figure 1: Cases of malaria per capita in 2015. Statistics from the WHO Malaria Report 2016. Yellow indicates lower rates, whilst red represent higher, grey indicates regions with missing data.

### 1.1.2 Aetiology of malaria

Biologically, malaria is caused by infection with protozoan parasites of the *Plasmodium* genus, primarily *Plasmodium falciparum*. Infection begins following transmission through the bite of a female *Anopheles* mosquito during a bloodmeal. At least 70 species of the *Anopheles* genus can carry and transmit the parasites [3], but the disease is most significantly associated with *An. gambiae* and *An. coluzzii*, particularly in Africa [4].

Those infected initially appear asymptomatic until the end of an incubation period approximately 9-30 days after infection with a median of 11 days for *P. falciparum* and longer for *P. malariae* [5,6]. Symptoms of uncomplicated malaria are typically flu-like, including fever, chills, muscle aches, headaches, cough, and diarrhoea [6], and typically occur once parasitaemia has exceeded 100 parasites per µL of blood [7]. Malaria is also characterised by paroxysm, a cycle between coldness and fevers occurring just under every two days for *P. falciparum*, every two day for *P. vivax* and *P. ovale*, and three days for *P. malariae* [8]. The cyclic nature of the parasite blood stage, in which parasites infect and lyse erythrocytes, corresponds to these periodic fevers, with each coinciding with a significant increase in merozoite release into the blood [5].

Typically, an individual is not infectious to subsequent mosquito bites until at least 7-15 days after initial infection, upon which a portion of the parasites will have developed into the mature gametocyte stage [5]. Eventually an individual level of immunity may be acquired, particularly after significant exposure to a variety of parasite strains leading to less severe subsequent infections [9].

Individual responses to infection can be quite diverse, even prior to any infection, with the most extreme forms of response being categorised as severe malaria subtypes. These subtypes include hyperlactatemia, severe malarial anaemia, respiratory distress and cerebral malaria [5]. Within these subsets, individual risk of death or coma is greatly increased [5]. For example, the case fatality rate for treated cerebral malaria is between

10-20%, rising to 50% in pregnant women [7]. In contrast, asymptomatic carriage of *Plasmodium* parasites has been reported in several populations [10,11].

### 1.1.3 A history of human co-evolution with malaria parasites

Malaria has been declared as “the strongest known force for evolutionary selection in the recent history of the human genome” [12], in large part due to the significant burden the disease places upon the global human population today but also due to the sheer length of time humans as a species have lived with the disease. The presence of malaria-causing *Plasmodium* parasites for multiple primate species suggests that the disease predates *Homo sapiens* as a species, and a *Plasmodium* population explosion coinciding with the agricultural revolution approximately 10,000 years ago highlights the co-evolution of humanity and malaria [13].

References to the periodic fevers that characterise malaria have occurred throughout history with these being identified as tertian, quartan, subtertian and quotidian by Hippocrates in 400 BCE [14]. Malaria has also been highlighted as a potential cause of the death of at least four popes (Gregory V 999, Damasus II 1048, Leo X 1521, Sixtus V 1590) [15], of the famous child-Pharaoh Tutankhamen in 1323 BCE [16], and possibly contributing to the fall of the Roman Empire in the 5th century [17]. Notably, the disease was so prevalent in Rome and the Roman Campagna during the Middle Ages that it has historically been referred to as ‘Roman Fever’.

Despite the disease clearly existing for millennia, the term ‘malaria’ did not appear until the 1800s, instead often being referred to as periodic fevers or ague. Discovery of *Plasmodium* parasites did not occur until Alphonse Laveran identified parasites in patient blood smears on the 6th of November 1880, for which he earnt the Nobel Prize in 1907 [18]. Mosquitos were not identified as vectors of *Plasmodium* parasites until Ronald Ross

demonstrated mosquito-borne parasite transmission in birds in 1897 [19], with human transmission via mosquitos being demonstrated two years later in 1899 [20].

#### 1.1.4 Moving towards global eradication

Historically, in addition to the tropical regions affected today, malaria was found across significant hotspots in the Pontine Marshes of Italy, the British Fenlands and Thames Estuary, and Florida, Illinois, and California in the United States of America [21]. Despite being classified as a tropical disease, cases have occurred as far north as Archangelsk, Russia in 1922-23 during a wider epidemic across the Volga basin [22]. Today, these historical regions are considered malaria free due to the removal of potential breeding sites, such as the draining of the British Fens in the 18<sup>th</sup> to 19<sup>th</sup> Centuries [23] and reclamation of the Italian Pontine Marshes in the 1930s [24], except for returning travellers, highlighting the feasibility of eradication. Modern examples of successful eradication include Paraguay [25], Kyrgyzstan [26], and Sri Lanka [27]; all of which were certified malaria free in the last three years.

Globally, rates of infection have reduced by 37% and deaths by 60% between 2000 and 2015 [28]. Yet despite this success, elimination efforts are increasingly at risk of being undermined by various independent emergences of resistance to classic antimalarials such as chloroquine, and more recent artemisinin-based combination therapies (ACTs). Furthermore, it can be argued that those regions where elimination has been possible are so-called ‘low hanging fruits’ with elimination from the remaining, more endemic regions being significantly more difficult.

Looking towards future eradication, in 2016 the WHO identified 21 countries, including Mexico, China, and South Africa, with the potential to eliminate malaria by 2020 [29]. However, those countries within this E-2020 initiative highlights these regions as bearing

relatively low burden with none existing within West, Central or East Africa where most of the malaria burden exists.

The current situation partially reflects previously optimistic malaria control programs such as the Global Malaria Eradication Programme of the 50s and 60s. Whilst initially successful, particularly in Europe and Northern America, the project ultimately failed to effectively eliminate malaria with resurgence of infections occurring in many regions. Whilst success was initially found primarily through the residual spraying of the insecticide DDT, the programme was ultimately undermined by a variety of factors including the rise of anti-malarial resistance in South Asia [30], insecticide resistance to DDT, and the reduction in funding by some countries [31].

As we move again towards the potential elimination of malaria, an emphasis is needed on a diverse range of complementary strategies which consider ecological and chemical control of disease vectors, novel approaches to drug surveillance and policy, the role of human genetics, and a strong emphasis on local health infrastructure empowered by rapid diagnosis tests (and an awareness of emerging ‘resistance’ to these).

Control and eradication efforts must be responsive to adapting parasite populations and tailored to their regional differences. Comprehensive surveillance and characterisation of global *Plasmodium* genomics – a major component of this thesis - represents one key aspect to a modern, empowered approach. Strides can also be made in the detection of asymptomatic human carriers and individuals at greater risk of severe reactions. In general, solutions to these challenges demand an ever-greater understanding of the biological mechanisms relating to malaria in both humans and parasites, particularly those underpinning the host-pathogen relationship.

## **1.2 The diagnosis, treatment, and control of malaria**

### **1.2.1 Diagnosis of malaria infection**

Confirmation of parasitaemia is typically required for a malaria diagnosis and can either be performed via a microscope, rapid diagnosis test, or PCR-based approaches. Examination of stained blood smears under a microscope is the gold standard approach which allow for both confirmation of infection and identification of the specific species. This form of evaluation requires both access to laboratory facilities and a skilled examiner [32] with misdiagnosis being possible due to some shared features between species.

Rapid diagnosis tests (RDTs), in which a blood sample is applied to an antibody treated dipstick, are more suitable for fast diagnosis in the field [32]. Multiple types of RDTs exist though the most common targets include the *P. falciparum* specific gene histidine rich protein II (*Pfhrp2*) and the pan-species lactate dehydrogenase (LDH) [7]. Efficacy of these tests are generally high but can be reduced by the presence of genetic variation of the target proteins [33]. Newer PCR-based approaches are also in development, with these allowing a combination of higher sensitivity with the ability to identify specific variants associated with drug resistance or mixed infections, though they are currently inappropriate for use in the field [32,34].

### **1.2.2 Current approaches for treatment and control**

Approaches towards the control, elimination and treatment of malaria are relatively diverse and include several social, physical and biochemical interventions. Primarily these can be summarised into either prevention of transmission through mosquito control or source reduction, prevention of disease through vaccination, or chemical prophylaxis, and treatment following infection with several antimalarials.

One major component of preventative care is the application of insecticide-based, mosquito-targeted approaches to reduce the number of possible transmission events.

Insecticide-treated mosquito nets (ITNs) and more recently introduced long-lasting insecticide-treated nets (LLINs), primarily act as physical barriers between humans and mosquitoes reducing parasite transmission, with the application of insecticides boosting their efficacy [35]. Similarly, indoor residual spraying (IRS), the internal spraying of households with insecticides such as DDT or pyrethroids to kill mosquitoes is another successful intervention. IRS has been found to be effective at reducing malaria incidence in a 2010 Cochrane review [36], though only if 80% of residences within an area are covered by the intervention [37]. Whilst effective, the insecticides applied to bed nets typically only remain for approximately up to a year for ITNs and at least three years for LLINs, with this being reduced if those nets are actively washed or used inappropriately [38], one example being their occasional misuse as fishing nets [39].

Mosquito control aimed at reducing transmission can also benefit from the active removal of mosquito breeding sites, a technique known as source reduction [40]. *Anopheles mosquitoes* have a diverse range of specific breeding site preferences, but generally require bodies of standing water for their breeding sites [41]. The removal of potential breeding sites can therefore drastically reduce mosquito populations and with it the transmission of malaria, as was historically seen in Europe. Large scale reshaping of landscapes is not necessarily required for successful source reduction interventions. *Anopheles* breeding sites are often man-made, with one study considering two major urban sites in Ghana finding that 74.5% were non-natural [42]. Measures can also be integrated into more broad scale improvements of local infrastructure [43].

RTS,S is the first, and currently only, available malaria vaccine and was approved by European regulators in July 2015, having begun development in the 1980s, [44]. As a recombinant vaccine it contains circumsporozoite protein (CSP) derived epitopes [45], CSP being a protein typically associated with sporozoite adhesion to and invasion of human liver cells [46]. The effectiveness of the vaccine is boosted by co-expression with

hepatitis B surface antigen [47], yet despite this RTS,S has a relatively low efficacy (4.4%) [48]. The development of alternative vaccinations and therefore the identification of novel vaccine candidates therefore remains critical.

Alongside patient treatment, antimalarials are utilised for preventative approaches, particularly for high-risk groups such as pregnant women or travellers from non-endemic regions. Seasonal malaria chemoprevention (SMC), the seasonal application of antimalarials, typically a combination of sulfadoxine-pyrimethamine (SP) and amodiaquine, across the Sahel sub-region was initially recommended by the World Health Organisation in 2012 [49] and has since been shown to be effective at reducing incidence of malaria by 60% [50].

Similarly, intermittent preventative treatment in pregnancy (IPTp), the application of a full course of antimalarials, typically SP, from the second trimester has also been recommended by the WHO across Sub-Saharan Africa from 2012 [51]. Pregnant women represent a key risk group due to significantly reduced level of naturally acquired immunity [52]. It has since been adopted across 36 African countries, despite a lag in expansion in recent years [53].

Application of antimalarials is currently the most effective treatment for malaria, though the emergence of resistance threatens to undermine this efficacy. One classic example is chloroquine, which saw widespread use until the emergence of resistance led to its replacement by alternative treatments such as SP, and more recently artemisinin combination therapies (ACTs).

*Table 1: Summary of antimalarial introduction and resistance. Dates sourced from "Global defence against the infectious disease threat" (WHO 2003)*

Antimalarial	Introduced	Resistance	Associated Genes
Chloroquine	1945	1957	<i>crt</i>
Mefloquine	1977	1982	<i>mdrl</i>
Proguanil	1948	1949	<i>dhfr</i>
Atovaquone	1996	1996	<i>mt-cytb</i>
Sulfadoxine/Pyrimethamine	1967	1967	<i>dhps, dhfr, gch1</i>
Artemether-lumefantrine	2006	Emerging	<i>kelch13</i>

### 1.2.3 Chloroquine & Quinine-derived Drugs

Chloroquine was initially discovered in 1934 as a synthetic alternative to quinine [54], the traditional antimalarial originally isolated from ‘Peruvian bark’, derived from several species of tree from the *Cinchona* genus, in 1820 [55]. The relative safety, efficacy, and affordability of chloroquine quickly led to its widespread use, though the rise of resistance eventually led to its gradual withdrawal. Recently chloroquine-susceptibility has begun to re-emerge in several *Plasmodium* populations though chloroquine use is still generally prohibited [56], except for some Central and South American countries where resistance is especially rare, if not entirely absent, such as Mexico, Panama, Haiti, and Argentina [57].

The mechanism of action for chloroquine against *Plasmodium* parasites is not fully characterised but thought to involve disruption of the parasite’s haem detoxification pathway, located within its digestive vacuole [58]. Digestion of the human host haemoglobin by the infecting parasite produces haem as a by-product. Accumulation of haem is toxic to the parasite, and typically converted to less toxic haemozoin within the parasite digestive vacuole [59]. Chloroquine disrupts the haem detoxification process after passing, non-polarised, into the low pH digestive vacuole through simple membrane permeability, wherein it becomes protonated and resilient to efflux [60]. Once inside the

vacuole, chloroquine concentrations gradually increase, accelerating its binding to haematin and leading to the build-up of toxic haem monomers. Excess levels of haem increase the permeability of the vacuole membrane, resulting in the death of the parasite. Notably mutant parasites without the enzymes required for haemoglobin proteolysis, and therefore without the ability to produce haem, have previously been shown to be resistant to chloroquine [61].

Chloroquine resistance first emerged in Asia and South America in the 1950s and 60s and has emerged independently at least four times including in Southeast Asia (from which it spread to Africa), Papua New Guinea, and twice in South America [62]. Since its initial emergence, chloroquine resistance has spread across the globe, though it has since reduced in several areas, including Africa, following restrictions upon chloroquine use as a front-line anti-malarial [63]. Genetic variants of multi-drug resistance gene 1, *mdr1*, and chloroquine resistance transporter, *crt* are typically associated with chloroquine resistance. For *mdr1*, its decreased expression has been associated with resistance, suggesting that its protein product aids the active transport of chloroquine into the digestive vacuole [64]. In contrast, the K76T variant of *crt* is associated with a reduced chloroquine concentration within the parasite digestive vacuole, suggesting that the wildtype CRT protein actively transports the drug out of the organelle and that the K76T variant reduces the protein's efficiency [65]. The two genes would therefore appear to complement one another, with their combination enhancing resistance.

#### 1.2.4 Sulfadoxine/Pyrimethamine (SP) & Antifolate Compounds

The synthesis of folate cofactors from folic acid through the *Plasmodium* folate pathway is critical to the supply of various molecules required for DNA synthesis [66] and consists of a handful of conserved proteins including DHFR, DHPS, and GCH1 [67]. Antifolate

drugs target this pathway through various targets and mechanisms and include proguanil, dapsone, sulfadoxine, and pyrimethamine [68].

SP is one example of an antifolate combination therapy first introduced in the early 1980s following the emergence of chloroquine resistance, its relative safety makes it one of the most widely used antifolate antimalarials [69]. Whilst sulfadoxine competitively inhibits dihydropteroate synthase (DHPS), encoded by *dhps*, pyrimethamine inhibits dihydrofolate reductase (DHFR), encoded by *dhfr* [70], both being key members of the folate pathway [67].

Resistance to SP first emerged in the late 1980s in Southeast Asia, before spreading to Africa and South America, leading to a shift of its use towards intermittent preventative treatments during pregnancy and infancy [71]. Unsurprisingly, this resistance is associated with variants of the respective target genes with single nucleotide polymorphisms in *dhps*, such as 450E [72], and *dhfr*, including 108D and 164L [73]. Additional resistance has been associated with whole gene duplication of GTP cyclohydrolase 1, *gch1*, with increased expression counter-balancing the fitness cost of *dhps* and *dhfr* point mutations [74]. Notably GCH1 lies upstream of DHPS and DHFR within the folate pathway, suggesting that duplication of *gch1* may act to compensate for fitness costs associated with *dhps* and *dhfr* mutation [75].

### 1.2.5 Artemether/Lumefantrine & Artemisinin-based Combination Therapies (ACTs)

Artemisinin first emerged as a potential anti-malarial during the Vietnam War, when the increasing failure of chloroquine led to requests from North Vietnam to China for an alternative therapy. This resulted in ‘Project 523’, one component of which evaluated folk remedies associated with plants of the *Artemisia* genus, from which the artemisinin compound emerged [76]. An active push away from artemisinin monotherapies began in

2014 following emerging resistance [77]. Today ACTs are the recommended first line anti-malarial for uncomplicated *P. falciparum* malaria.

Artemether-lumefantrine is the most commonly used artemisinin-based combination therapy [78]. Whilst effective, the mechanism of action is unconfirmed, with one leading hypothesis suggesting that artemisinins are activated by interaction with haem and iron(II) oxide within infected red blood cells, resulting in the production of free radicals which kill the intracellular parasite through oxidative stress [79,80].

Initial reports of artemisinin resistance first emerged, as reports of significantly slower parasite clearance, in the 2000s [81]. Since then point mutations in the propeller domain of *kelch13* have acted as a reliable indicator of this reduced clearing in Southeast Asia [82]. Decreased susceptibility has also been associated with variants of *atpase6* in French Guiana [83]. In general, resilience to artemether-lumefantrine has been restricted to Southeast Asia whilst resistance to other forms of ACTs, such as artesunate-sulfadoxine-pyrimethamine (AS+SP), have emerged in India, Somalia, Sudan, Afghanistan, Iran, Pakistan and Yemen, leading these countries to switch to artemether-lumefantrine [84]. Identification of novel drug resistance associated variants has greatly benefited from whole genome-sequencing approaches that utilise computational methods such as association and selection. For example, the role of *kelch13*, and specifically SNPs including the C580 allele, as markers for artemisinin resistance were discovered through a genome-wide association study [85]. Previous approaches, on more limited custom SNP arrays, have allowed for the identification of signals of selection for *crt* resistance with integrated haplotype score (iHS) scans for positive selection on limited arrays [86], future higher resolution genome-wide approaches present the opportunity to widen this net to detect further antimalarial resistance signals at an earlier stage.

## 1.2.6 Existing and Emerging Resistance

Some form of resistance or reduced efficacy exists in the field for almost all known antimalarials, with resistance even emerging for newer treatments such as artemisinin-based combination therapies. Whilst the speed of emergence for resistance varies, historically the time between drug introduction and resistance has been relatively short with resistance to chloroquine appearing just over 20 years after initial use [87] but less than one year for SP [88]. Curiously resistance to chloroquine, SP and ACTs all initially emerged along the Thailand-Cambodian border relatively quickly [81,89]. This likely stems from a combination of early access to artemisinin monotherapies from China during the Vietnam war, infection being limited to the non-continuous exposure of a working age and transient population of forest workers, and a health infrastructure weakened by three decades of war. Further successful WHO surveillance has ensured that resistance was spotted early even if it had earlier emerged independently elsewhere. Mathematical modelling has also suggested that emerging drug resistance is less likely to become established in lower transmission settings, such as Southeast Asia, than high transmission settings, such as sub-Saharan Africa, due to a lower degree of in-host competition between *Plasmodium* strains [90].

Selection for specific parasite variants that convey increased survivability in response to human interventions is not exclusive for antimalarials. The World Health Organisation now recommends that all malaria cases be confirmed by a rapid diagnosis test (RDT) [91] but ‘resistance’ is developing for these. Specifically, the emergence of genetic variants within the target parasite proteins leads to false negatives, thereby causing inappropriate non-application of antimalarials. One key example is deletions in *Pfhrp2/3* that have been shown to cause false positive test results and therefore risk lack of intervention increasing parasite survivability [92].

Genetic variants conveying resistance range of small single base changes (SNPs) such as *crt* K76T to larger structural variants (SVs) such as the duplication of *mdrl* in mefloquine and lumefantrine resistance [93]. Speed of emergence and persistence of resistance also varies (CQ slow to appear, fast to leave; SP fast to appear, slow to leave) which may reflect the underlying complexity of resistance mechanisms. For example, resistance through one specific form (such as K76T for chloroquine) may be less likely to occur and faster to leave (as one specific SNP is required) compared to SP resistance (multiple SNPs convey resistance, fast to acquire any, slow to remove all). The role of compensatory variants is also key as many resistant variants may bear a fitness cost that additional variants may compensate for. One possible example is the duplication of *gch1* as compensatory for *dhfr* and *dhp5* mutations in SP resistance [75].

### 1.3 The role of *Plasmodium* parasites

#### 1.3.1 Species of *Plasmodium*

The *Plasmodium* genus is broad and consists of over 200 species [94], of which the vast majority infect a range of non-human hosts including other primates [95], various mammals [96], several reptile species [97], and over 150 that infect birds [98]. At least five *Plasmodium* species (*P. falciparum*, *P. vivax*, *P. malariae*, *P. knowlesi*, and the two *P. ovale* subspecies) specifically cause malaria in humans with almost all cases being caused by *Plasmodium falciparum*, particularly in Africa where they account for 99.7% of cases [1]. Notably *P. falciparum* is phylogenetically distinct from the other human infecting *Plasmodium* species, clustering closer to the chimpanzee infecting *P. reichenowi* [99]. Their genomes also contain some disease-relevant genes not found in other species, such as the *Rh2a* and *Rh2b* gene pair required for erythrocyte invasion [100].

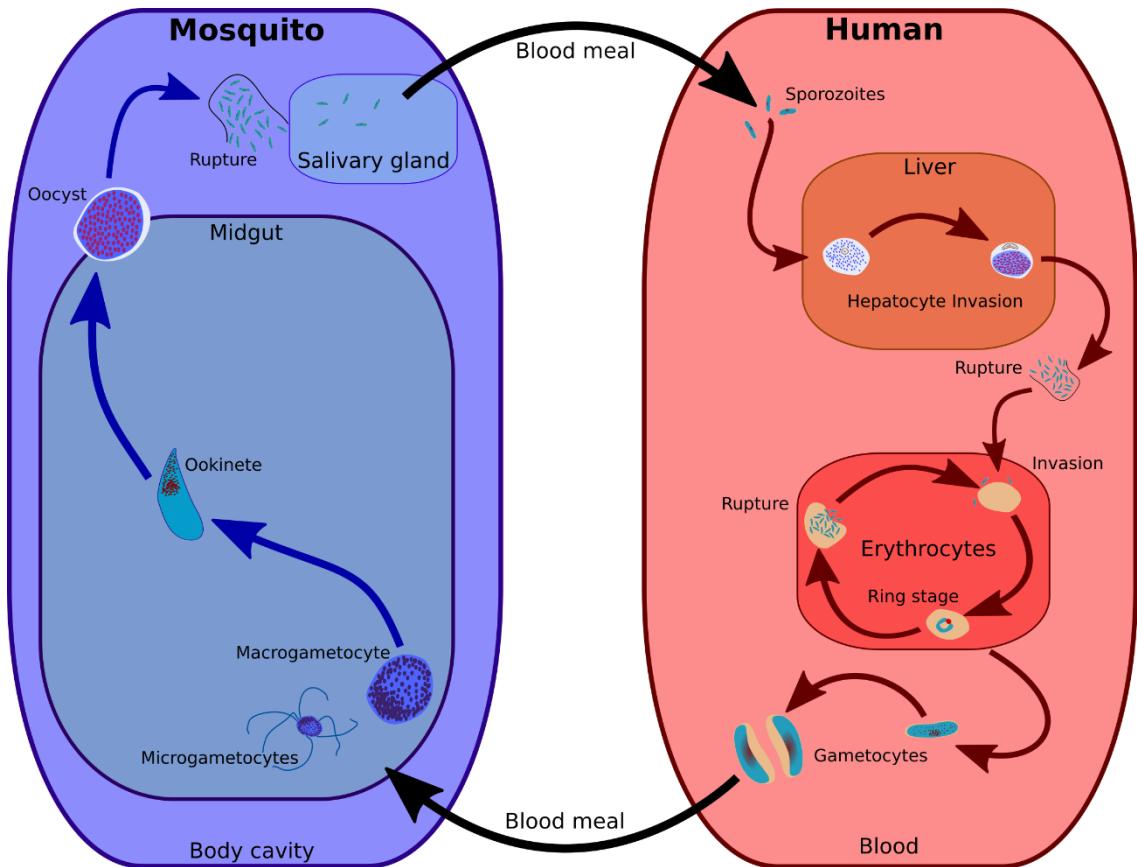
Outside of Africa, *Plasmodium vivax* is more prevalent, accounting for 64% of cases in South America, over 30% in Southeast Asia, and over 40% of cases in the Eastern Mediterranean region [1]. Infection with *P. vivax* or *P. ovale*, is also characterised by an additional dormant liver ‘hypnozoite’ stage, which can allow an infection to ‘hibernate’ for multiple years before symptoms emerge, making this stage a unique target for drug discovery [101].

The more benign *P. malariae*, the zoonotic *P. knowlesi*, and the two *P. ovale* sub-species cause very few cases of malaria. In total, these account for less than 1% of global cases [1]. Though these cases may be underestimated due to common mischaracterisation as *P. falciparum* or *P. vivax* infections due to clinical assumption or visual similarities between *P. ovale* and *P. vivax* when viewed through a microscope, for example due to the presence of shared features such as Schüffner’s dots [102]. Rates of misdiagnosis are often quite significant, with one study based in China reporting false positive rates for *P. ovale* and *P. malariae* that exceeded 70% [103].

A minority of other *Plasmodium* species have also recently been implicated in rarer cases of human infection. These include a 2014 report of a *P. cynomolgi* infection in Peninsula Malaysia [104], and a 2015 report of *P. brasiliense* infection in the Venezuelan Amazon [105]. Typically, such cases were initially misidentified as other forms of *Plasmodium*, suggesting that additional cases of infection with rarer *Plasmodium* spp. may have been overlooked. Other simian *Plasmodium* species, in addition to *P. knowlesi*, have also been demonstrated as capable of infecting humans, including *P. schwetzi* [106], *P. inui* [107], and *P. simium* [108] with cases of the latter being increasingly common in south-eastern Brazil, possibly due to more accurate distinction between cases of *P. vivax* and *P. simium* infection. For example, one recent study found that most cases of malaria in the Rio de Janeiro Atlantic Forest are caused by *P. simium*, leading to the development of specific detection methods to reduce false classification as cases of *P. vivax* infection [109].

### 1.3.2 *Plasmodium* Life Cycle

*Plasmodium* parasites experience a particularly complex life cycle with multiple stages across at least two hosts; a female mosquito and a secondary host, humans in the case of malaria. Initially, the parasite will enter the human host from the mosquito salivary glands, via a bloodmeal, as a sporozoite. These sporozoites then infect hepatocytes within the liver, in which they rapidly replicate forming schizonts that rupture to release the merozoite stage parasite (here *P. vivax* and *P. ovale* may also enter a dormant hypnozoite stage). Those merozoites will then enter the blood and infect erythrocytes, entering a ‘ring-shaped’ trophozoite stage followed by further asexual replication within schizonts until rupturing the host erythrocyte and re-entering the blood. This cyclic process results in the anaemia and periodic fever commonly associated with malaria, and further accounts for the delay in onset of symptoms following infection. Approximately 1.0% of merozoite parasites will progress to the sexual gametocyte stage [110]. Those gametocytes will remain within erythrocytes until taken up into a new mosquito host, in the gut of which they are released to form zygotes and eventually ookinetes that embed into the gut membrane [111]. Once in the gut membrane, ookinetes develop into oocytes that divide multiple times into sporozoites that migrate to the mosquito salivary gland from which they can be injected into a new host during a bloodmeal [112].



*Figure 2: Life cycle of the *Plasmodium falciparum* parasite. Schematic diagram demonstrating the key life cycle stages and their locations within each host for the *P. falciparum* parasite.*

### 1.3.3 The genomics of *P. falciparum*

Efforts to sequence the whole *P. falciparum* genome began in earnest in 1996 [113] with the first draft being completed six years later in 2002 [114]. The core nuclear genome is 23.3 Mbp in size, split across 14 chromosomes and contains approximately 5,300 genes with a mean length of 2.3 kbp [114]. The genome is notable for its significant AT-rich nature, with a GC content less than 20% [114].

*P. falciparum* also contains two forms of non-nuclear DNA in the form of a mitochondrion, and the apicoplast, an organelle unique to the *Apicomplexa* phylum with roles in metabolic pathways such as haem biosynthesis and fatty acid synthesis [115]. Of these, the mitochondrion is 6 kbp in size, and dependent on nuclear genome for tRNA

synthesis [114]. In contrast the apicoplast is far larger, at 34 kbp, and able to encode its own tRNAs [114]. Evolutionarily, the apicoplast is chloroplast-like and thought to be the result of secondary endosymbiosis [116], though it has since lost its photosynthetic capabilities.

Survivability within the host is enhanced through antigenic variation of surface proteins produced by families of highly repetitive genes. These include families of highly variable life stage-specific surface antigens, such as the *var* family [117], the trophozoite-specific adhesins *rifin* family [118], and the schizont stage *stevor* family [119]. Members of these genes recombine to create antigenic diversity on the surface of parasite-infected red blood cells (iRBCs) enhancing immune evasion. Typically, loci for members of these gene families are excluded from standard short read based analysis due to issues resolving their proper sequencing given their highly repetitive nature and their presence within poorly resolved subtelomeric regions.

All three highly variable gene families are also members of the *P. falciparum* exportome, which also includes PHIST-domain containing proteins and ring-infected erythrocyte surface antigen (RESA) [120–122]. As the name suggests, proteins encoded by these genes are exported from the infecting *P. falciparum* into its host. Members of the *P. falciparum* exportome are therefore key to several host-pathogen interactions with their expression being vital for infection of host cells and supporting evasion of the host immune system [123,124].

Many, but not all, exported *P. falciparum* proteins contain the PEXEL (or host-targeting, HT) motif (R/HxL/IxE/D/Q, where x is any amino acid), allowing bioinformatic prediction of exported genes due to the presence of this motif [120,125,126]. Another such motif of bioinformatic interest is the TATA (or Goldberg-Hogness) box, which identifies likely promoter regions in eukaryotes and archaea [127].

### 1.3.4 Variation in *P. falciparum* reference strains

There exists a significant number of established lab-adapted *P. falciparum* reference strains upon which analysis has taken place, the majority of which were isolated in the early 1980s. The primary reference strain for *P. falciparum* is 3D7, whose African origins were only elucidated in 2014 [128], despite having been isolated in the Netherlands in 1981 from an individual whom had never been abroad [129].

Other prominent reference strains include 7G8, a 1981 clone of IMTM22 originally isolated from a 12-year-old male from Manaus, Brazil in 1980 [130], and HB3, a 1983 clone of H1 originally isolated in Honduras in 1980 [131].

One particularly lab-adapted strain is Dd2, a 1996 clone of W2-MEF derived from W2pMCII following six months of selection for mefloquine resistance which was itself derived from 12 months of mefloquine resistance selection of a culture originally isolated from a Lao refugee in 1980 [132,133]. Dd2 is particularly noteworthy for featuring a triplication of *gch1*, which has been associated with SP resistance [70].

Lab and computational-based analyses of these strains has facilitated the identification and characterisation of genomic diversity, virulence, and anti-malarial resistance but have tended to focus on specific lab strains or restricted geographical regions. More recently analyses have shifted from strain-specific contexts to large global datasets. One notable example is the Pf3k project which seeks to take an international approach to the collection, collation, and analysis of *P. falciparum* samples [134]. Release three of this project contained 2,512 samples from 14 countries [135]. This shift towards larger datasets has led to a greater demand for bioinformatic approaches to big data processing and analysis.

## **1.4 Human genetics in the context of malaria**

### **1.4.1 Determinants of diversity of human response to infection**

Individual responses to infection with *Plasmodium* parasites are diverse due to a multitude of environmental, genetic, and immunological factors. Whilst some infected individuals may present as asymptomatic, others can experience combinations of severe subtypes such as hyperlactatemia, respiratory distress, and cerebral malaria [6]. These severe malaria subtypes being almost exclusively caused by *P. falciparum* infection [136]. Some of this diversity reflects expected risk factors for any communicable disease. For example, higher severity is often associated with being especially young or old [137], pregnant [138] immunocompromised, or immunologically unfamiliar to infection. Some diversity also reflects random exposure to particularly virulent strains or regions with higher transmission.

It is also possible to obtain a level of acquired immunity to infection through sustained or repeated infection with Plasmodium parasites [9]. The presence of an acquired level of resistance suggests the circulation of human immunological factors, pathways or mechanisms that, if correctly identified, could underpin efforts towards new antimalarials or new vaccine candidates.

Disease-linked features often represent confounding factors that must be appropriately controlled for in statistical analyses, such as genome-wide association studies, when seeking to identify novel causal genetic variants. These can include gender, where it has been suggested that women (particularly pregnant women) are at higher risk [138], and age, with this possibly reflecting the time-lag associated with developing an acquired immune response [137].

#### 1.4.2 Severe malaria subtypes

A diversity of severe malaria subtypes exists with each being clinically distinct and most often associated with *P. falciparum* infection, though some cases have been associated with *P. vivax* and *P. ovale* [136]. An increased risk of severe response has been associated with specific risk factors including ethnicity both broadly, with Asian and White populations with a higher risk of severe response [139], and specifically, for example associations with the Dantu blood group antigen [140].

Example of severe subtypes include severe malarial anaemia (indicated by a significantly low haemoglobin concentration level, typically below 5 g/dl for under 12s [141]), hyperlactataemia (indicated by blood lactate levels above 5 mmol/L [141]), respiratory distress (which develops in approximately 40% of children with severe *P. falciparum* malaria [142]), and cerebral malaria (typically diagnosed via evaluation with the Blantyre coma scale which considers eye movement, motor responses, and verbal response score < 5 [141]).

#### 1.4.3 Genetics of human susceptibility

Centuries of host-pathogen interactions between human hosts and their malaria parasites have had a significant impact on the human genome, particularly within Africa, leading the claim that the disease has been the largest selective force upon the human genome in recent history [12].

One study estimated that approximately 25% of disease variance for severe malaria is determined by human genetic factors [143]. Multiple variants account for this risk, though the majority are associated with classic examples such as *HbS* (sickle cell trait), alpha-thalassemia, ABO blood type, Duffy negative status, and G6PD deficiencies [144,145].

A wealth of other variants of less significant impact, such as *USP38*, *FREM3*, glycophorins *gypA/B/E*, *DDC*, and *ATP2B4* have also been identified as being associated

with severity of response though are, in sum, less protective than heterozygous carriage of *HbS* [146–150]. Indirect impacts of human genetics also exist, for example twin studies have suggested that a genetic component underpins the likeliness that different individuals are bitten by a mosquito, and therefore the probability of initial infection [151].

Several variants protective against malaria are also associated with a high risk of an alternative disease, predominantly forms of anaemia, highlighting complex patterns of evolutionary selection. The most well-known example of a genetic variant associated with protection against malaria is carrier status for sickle cell anaemia. Heterozygous carriage of the *HbS* sickle variant is associated with an almost 90% reduced risk of severe malaria, due to the reduced ability for *Plasmodium* parasites to infect the malformed sickle-shaped erythrocytes [152]. In contrast, homozygous carriage of the sickle variant of HBB instead leads to sickle cell anaemia, a disease in which half of patients die before they reach fifty [153].

Alpha-thalassemia and G6PD deficiency represents two other anaemia-associated forms of malarial-protective variants. Alpha-thalassemia is an impairment of proper haemoglobin production, where an individual has an irregular ratio of alpha, beta and gamma haemoglobin chains with an excess of inefficient beta tetramer and gamma chain oxygen carriers and is typically caused by deletions of or within *HBA1* or *HBA2* [154]. This disruption results in severe anaemia, alongside resilience to malarial infection, and is therefore selected against in non-malarial regions causing global alpha-thalassemia distribution to be highly correlated with malaria-endemic regions [155]. In contrast, G6PD deficiency is the underproduction of glucose-6-phosphate dehydrogenase (G6PD), an intracellular enzyme that catalyses NADPH production through reduction of NADP+ [156]. Deficiencies in G6PD production result in haemolysis whilst also providing a degree of resistance to uncomplicated *P. falciparum* malaria infection [157]. As with

sickle cell anaemia and alpha-thalassemia, selection for malaria protection over the resulting anaemia results in GP6D deficiency being distributed within malaria-endemic regions [158].

Other variants only convey protection to specific *Plasmodium* species, for example *P. vivax* and *P. knowlesi* utilise host surface antigens such as the Duffy antigen receptor (encoded by *DARC*) to enter human erythrocytes. Deletions of the Duffy antigen receptor therefore provide resistance to merozoite invasion [159], with this being reflected in the global distribution of the Duffy-negative trait to regions with high levels of *P. vivax* infection.

Recent developments in high-throughput sequencing technologies have allowed the expansion from laboratory-based candidate gene studies to genome-wide discovery approaches. The utilisation of genome-wide association studies (GWASs) have led to the identification of several loci associated with the severity of human response to *Plasmodium* infection [145,150] with some being associated with specific subtypes, such as *ADAMTS13* with cerebral malaria [160]. With greater data resolution, our understanding of human genetics in the context of malaria has grown more complex, with a significant number of SNPs only being found in association for specific populations [161]. The search for additional factors is therefore potentially hindered by a lack of African genomes for both primary variant discovery and improving existing reference panels with localised imputation, for example the 1000 Genomes Project only includes seven African populations of which two are based in the USA and Barbados [162]. It is also likely that human genetic factors with more subtle phenotypic impacts differ by geographical region due to inherently weaker selection, for example a variant associated with malaria protection in West Africa may be not be associated in East Africa. This biogeographical nature of some variants highlights the potential benefit of personalising treatment and control approaches to local populations, an incentive that would benefit

from strong, local health infrastructures. Further the use of other whole genome methods, and complementary protein, epigenetics, and other ‘omics approaches is likely to continue identifying novel inherent host factors relevant to malaria through which future treatments and control methods can be better informed.

## 1.5 Methods utilised in exploring genomics

### 1.5.1 High throughput sequencing

‘Next generation’ high throughput sequencing approaches have facilitated the rapid growth of bioinformatic analysis since the turn of the century, having found utility for a wealth of applications across clinical and microbiological settings [163]. Milestone achievements include the sequencing of the human genome in the 2000s [164] and the emergence of novel methods such as large-scale genome-wide association studies [165]. High throughput sequencing technologies can be loosely classified into two forms based on the length of their raw outputs, known simply as ‘short read’ and ‘long read’ sequencing.

Short read based sequencing methods include 454 pyrosequencing [166], Ion Torrent [167], and Illumina platforms [168], amongst others, and typically produce raw reads shorter than 1000 bp. These approaches generally involve fragmentation of the DNA molecule followed by sequencing of those short sections. The result is an abundance of short reads that either require aligning against a reference or *de novo* assembling for interpretation. Recent improvements have centred on cost effectiveness, with Illumina claiming its first ‘\$1000 genome’ in 2015, with that cost being \$200,000 in 2009 [169].

Long read based sequencing methods aim to sequence large sections of DNA, with the unofficial goal of sequencing entire chromosomes without assembly. These approaches tend to be newer and include single-molecule real-time sequencing (SMRT) from Pacific Biosciences, and Nanopore, as provided by Oxford Nanopore. These methods are better

suited towards the resolution of highly repetitive regions, which short read approaches have difficulty resolving [170], and reduce issues related to reference genomes with multiple regions of high similarity [171]. In theory, long read methods should also be able to better resolve complex structural variation events, for example a duplication would be present in the raw reads and not require inference from coverage or breakpoint-based analysis. Recent improvements have focused on enhancing accuracy, with Nanopore specifically having a high stochastic error rate [172].

### 1.5.2 Genomic variation in host-pathogen interactions

Genomic variation underpins many heritable forms of antimalarial resistance for malaria parasites [173], and types of malaria susceptibility for infected humans [174]. These variants range from single base single nucleotide polymorphisms (SNPs) to large genomic regions mutated by structural variations such as deletions, duplications, insertions, and inversions [175]. Once identified, the specific variants present can be used to infer signals of genetic diversity, selection, and association within and between wider populations [176].

Given this range of genetic variants a range of methods for their detection has been developed. For example, SNPs and small insertions/deletions (indels) have classically been detected using lab-based methods such as microarrays [177], or with *in silico* read-alignment based methods such as variant calling [178]. Typically, for an *in silico* approach, short or long reads are aligned to a reference genome with software such as bwa or minimap2 with variants relative to that reference being identified with a tool such as *mpileup* [179–181].

### 1.5.3 Identifying signals of selection

Once detected, SNPs can form the basis of methods to identify signals of selection either as independent features or grouped into inherited sets known as haplotypes. Methods for determining selection therefore range from simple measures of variant frequency, either over time or between populations, to more complex measures of locus conservation. Fixation indexes ( $F_{ST}$ ) are a group of methods that identify differences in population frequencies, they include Nei's, Wright's, and Weir and Cockerham's methods with some optimised for specific uses such as unequal sample sizes [182–184]. All consider relative variant frequencies between two or more populations with larger values (further from 0, closer to 1) indicating that the frequency of a variant in one or more populations is significantly different in the other populations [182–184]. These slight distinctions can make comparison between different forms complex in their minutia, despite them all being referred to as ' $F_{ST}$ '.

Simple haplotype-based methods include Tajima's D, in which variation within pre-defined regions is compared between two or more samples [185]. This method effectively considers the difference between the true average number and the expected number of polymorphisms between the considered sample pairs. More specifically, Tajima's D compares the differences between two estimates of genetic diversity in a population: Tajima's estimator ( $dT$ ), the sum of pairwise-differences between haplotypes over the number of pairwise comparisons (number of samples choose two), and Watterson's estimator ( $dW$ ), the number of segregating sites over the (number of samples - 1)th harmonic number. Tajima's D for a sequence is therefore  $(dT - dW)$  over the standard deviation of  $(dT - dW)$  for all sequences. Positive Tajima's D values indicate an abundance of rare variants and therefore suggesting recent selection, whilst negative values indicate a lack of rare variants suggesting balancing selection [185]. This metric can also be applied on a 'per-gene' basis, rather than for fixed sized windows, meaning

that biological validity may be gained at the cost of more complex comparisons due to different sized SNP sets. Specific limitations of this approach include the requirement for biallelic sequence differences, difficulties in applying the method to complexly structured populations, and biases introduced for populations growing at an exponential rate.

More complex and newer statistical approaches will consider conservation of haplotypes around specific loci. Extended haplotype homozygosity (EHH) based methods, such as integrated haplotype scores (iHS) and cross population extended haplotype homozygosity (XP-EHH), are based on the comparison of conserved haplotype length around specific SNPs for different core alleles (iHS) or different populations (XP-EHH) [186,187]. Similar methods, such as HaploPS, instead identify selection as abnormally long haplotypes for a SNP compared to haplotypes of a similar frequency [188].

EHH is defined as “the probability that two randomly chosen chromosomes carrying the core haplotype of interest are identical by descent (as assayed by homozygosity at all SNPs) for the entire interval from the core region to the point x” [189]. iHS then utilises those base EHH scores and calculates an integrated score under the EHH curve for both ancestral (reference) and derived (alternative) core alleles ( $iHH_A$  and  $iHH_D$  respectively). iHS is then calculated as the ratio of  $iHH_A$  over  $iHH_D$  and values standardised relative to all other iHS scores for a specific genome of interest [186]. XP-EHH takes a similar approach to consider relative differences in haplotype conservative between two population, this is calculated as the natural log of the integrated EHH for a SNP of interest in population A over the equivalent iHH for population B (with the equivalent standardisation) [187].

EHH-based approaches therefore share some specific limitations, for example appropriately phased haplotypes are required and errors introduced during phasing may therefore be inherited. Further these approaches rely upon there being a sufficient density of biallelic SNPs across the genome and will therefore exclude regions with notably high

or low levels of variation, though this is similarly true for other methods that utilise SNP frequencies or associations. This required level of variation also means that EHH methods are generally geared towards the detection of more recent selection events, which may be viewed as a feature rather than a bug. Conversely, genomic regions under consistently high levels of selection, and therefore displaying a low number of SNPs in the population, are generally unsuitable for EHH-based approaches. Strong purifying signals of selection present within the full population may therefore be overlooked, though these may be identified with cross-population approaches, such as XP-EHH.

Notable positives to an EHH approach include the ability to identify the specific allele or population under selection. Further dissection of the EHH curves can also identify skews within that selection, whether up or downstream of the core SNP, allowing the inference of specific sub-regions under strong conservation. EHH-based approaches also incorporate linkage disequilibrium, rather than treating SNPs as purely independent events, and therefore more accurately represent the molecular mechanisms underpinning signatures of selection. Together these factors ensure that EHH-based methods are suitable for the identification of recent selection events relating to relatively frequent variants, such as those associated with emerging drug resistance. Context is also important for selection methods such as these, as selection signals are often normalised over a selection of target loci. For EHH-based methods this can mean that the exclusion of certain regions or SNPs can skew normalisation if it is not appropriately controlled for.

#### 1.5.4 Identifying structural variation

Traditionally structural variants have been identified through lab-based approaches such as PCR and DNA probes, which benefit from being able to directly examine putative variants but are limited to slower analysis of small sample sets [190]. In contrast computational approaches benefit from faster, higher-throughput analysis, allowing

specific regions to be highlighted for follow up investigation in the lab. Most recently developed *in silico* approaches tend to utilise read-based inputs to examine paired read realignment [191,192] or aligned read depth [193,194]. Read depth approaches, such as CNVNator or Control-FREEC, involve mapping reads against a reference genome and identifying duplications and deletions as deviations from an expected read depth [193,194]. Deletions will appear as a genetic region in which nearly zero reads map to the reference, whereas duplications are regions where read depth is double or more of the expected depth [193,194]. This approach requires controlling for regions of poor mapping and accounting for significant GC biases, but tends to be relatively successful at identifying strong signals of significant length. An appropriately sized window is also required, to account for per-base variation, which can limit the size of structural variants that can be identified. Typically, this bottom limit has been around 500 bp leading to a low detection of smaller CNVs. When considering other forms of structural variation, such as inversions and insertions, read depth-based methods can be limited as these forms of structural variation tend not to impact read depth.

Instead, read realignment approaches such as DELLY or LUMPY are considered. These realign each read to the reference genome to determine if specific reads are split across a probable breakpoint, a point where a structural variant begins or ends, or to identify anomalous read pairing. For example, an inversion can be detected when a significant number of reads align in reverse to their paired read, or a duplication might be identified by a read aligning in the same direction but the wrong side of its paired read [191,192]. Many of the limitations of read alignment-based methods, such as the need for enough high-quality reads across the whole genome of interest, are shared with alignment-based genome assembly and read depth SV detection approaches but may present differently. For example, poor quality reads for a specific region may lead to a false positive deletion

through a depth-based method but would require significantly poorly quality reads than for a similar detection in a realignment-based method.

Consideration of the phasing of structural variants may alleviate some of these issues, for example predicted heterozygous genotypes for a haploid organism may suggest contamination, poor quality sequencing or more complex structural variation worthy of further investigation. Typically, methods will utilise specific parameters regarding read quality thresholds, or provide metrics relating to the number of paired or split reads that support the existence of a specific structural variant. These may then be incorporated as a secondary quality filter, post-detection.

Short read based approaches are those most often utilised for structural variant detection given their established presence within bioinformatics, but these can be limited by alignment bias against subtelomeric regions, cryptic alignment due to repetitive regions (such as *var*, *stevor*, or *rifin* genes), and biases involved in GC skew [195]. Long read approaches can resolve some of the issues associated with poor alignment but structural variants detection methods for these are in early development. These newer sequencing technologies also bring new issues such the stochastic errors seen with Oxford Nanopore [172] and low read depth removing the ability to identify significant concordance between reads. Ideally both long and short reads should be considered alongside both read depth and realignment approaches.

## 1.6 Project Outline

This thesis examines the global genomics of host-pathogen interactions in malaria for populations of both the major causal parasite, *P. falciparum*, and humans in malaria-endemic countries, primarily Tanzania. For each species I began by considering the role of SNPs and signals of selection, before extending to large scale explorations of structural variation. This large-scale exploration was facilitated by the development of an analysis

and visualisation tool, *SV-Pop*, and a pipeline for the detection of inversions in long read assemblies.

Beginning with those chapters concerning the *P. falciparum* parasite, **Chapter Two** considers the impact of sustained SP use on the *Plasmodium* population within Malawi relative to its neighbouring countries, and the wider global population. Primarily this meant characterising SNPs known to be associated with anti-malarial resistance, before extending to the identification of novel selection signals and structural variants. Selection signals were identified in several loci including those with known to be associated with SP resistance. Investigation of chloroquine resistance variants revealed their near absence in the Malawian parasite population, demonstrating the ability for this resistance to leave a population when chloroquine use is restricted. In contrast variants associated with SP resistance were present at high frequency, with this being supported by corresponding selection signals. Structural variation was also considered with one striking 436 bp duplication being found at near fixation in the population immediately upstream of *gch1*, a gene for which whole gene duplications have been previously associated with SP resistance [75].

Following discovery of this duplication in Malawi and given the previously unexplored biogeography of structural variation in *P. falciparum*, **Chapter Three** considers my utilisation of over 3,000 samples to explore the global variation of copy number variants in the *P. falciparum* population. This added global resolution to previously identified duplications of *mdr1*, enhanced our understanding of the Central and East African distribution of the 436 bp *gch1* promoter duplication identified in **Chapter Two**, and identified novel duplications of *crt* unique to West Africa which may be associated with the dual carriage of chloroquine resistant and susceptible alleles.

Population specific variants were also identified in genes such as *rh2b*, *rhopH2*, and the ring-infected erythrocyte surface antigen PF3D7\_0102200.

The large-scale approach to structural variant exploration described in **Chapter Three** highlighted intricacies in short read-based methods which undermined successful detection of inversions. To resolve this, **Chapter Four** describes the development of a method for inversion detection from long read assemblies, which better resolve regions that short read approaches find difficult. This method was then applied to a selection of lab cultured and field isolated samples from around the globe. Multiple putative inversions were identified in several highly variable *var*, *rifin*, and *stevor* genes, historical tandem inversions present for *RH2a/b* and elongation factors 1-alpha, and a novel sandwich inversion of *pi4k* identified in GB4, a lab adapted strain of *P. falciparum*.

The human-focused analyses initially considered the severity of malaria response through a case-control genome-wide association study before extending towards an investigation of genomic structural variation within a similar Tanzanian population, with a focus on genes associated with *P. falciparum* interaction. Tanzania is a highly malaria endemic region in which the local population experience a range of symptomatic responses to *P. falciparum* infection including a range of severe malarial subtypes. **Chapter Five** describes work to identify novel genetic variants associated with a higher or lower risk of disease severity. This required application of a mixed model genome-wide case-control association study, with that core analysis being complemented by an examination of signals of selection genome-wide and candidate region exploration of structural variation. Here novel SNPs associated with severe malaria response, including two interleukin receptors (*IL-23R* and *IL-12RBR2*), were identified, alongside signals of selection for loci such as *SYNJ2BP* and *GCLC*.

**Chapter Six** sheds light on the abundance of structural variation within the Tanzanian human genome by up-scaling the structural variation work from **Chapter Five** to a whole genome context. Here I identified just short of 200,000 putative variants in 156 parents, of which 6,932 specific forms were frequent (>5%) and a mean of 16.7% per parent were inherited without mutation to their child (n=78). Gene ontology analysis highlighted enrichment for cell adhesion, drug binding, and intracellular transport functions. Amongst the wide array of variants some, such as a 4,136 bp deletion of *SEC22B* (98.7%) and a 220 bp deletion in *BET1L* (98.7%), were present at near-fixation. Focus was also applied to candidate genes associated with known roles in blood antigen systems and the risk of severe malaria. This led to the identification of novel SVs in *A4GALT*, *ATP2B4*, and *ABCG2*, with the latter potentially being a novel form of the typically highly rare Jr(a-) blood phenotype.

**Chapter Seven** describes *SV-Pop*, a package I developed in Python and R for the high-throughput, multi-population analysis and visualisation of structural variation. Its use was demonstrated in **Chapters Three and Six** in which it was used for variant filtering and analysis, as well as visualisation of the larger datasets. Given the species agnostic nature of this tool, future applications could facilitate exploration of global structural variation in other malaria species, such as *P. vivax*, or mosquito populations.

## 1.7 References

1. World Health Organization. World Malaria Report 2017. WHO Press. 2017.
2. Gallup JL, Sachs JD. The economic burden of malaria. *Am J Trop Med Hyg*. 2001;64:85–96.
3. Molina-Cruz A, Zilversmit MM, Neafsey DE, Hartl DL, Barillas-Mury C. Mosquito Vectors and the Globalization of Plasmodium falciparum Malaria. *Annu Rev Genet*. Annual Reviews ; 2016;50:447–65.
4. Miles A, Harding NJ, Bottà G, Clarkson CS, Antão T, Kozak K, et al. Genetic diversity of the African malaria vector *Anopheles gambiae*. *Nature*. Nature Publishing Group; 2017;552:96.
5. Trampuz A, Jereb M, Muzlovic I, Prabhu RM. Clinical review: Severe malaria. *Crit Care*. BioMed Central; 2003;7:315–23.
6. Bartoloni A, Zammarchi L. Clinical aspects of uncomplicated and severe malaria. *Mediterr J Hematol Infect Dis*. Catholic University in Rome; 2012;4:e2012026.
7. Ashley EA, Pyae Phyo A, Woodrow CJ. Malaria. *Lancet*. Elsevier; 2018;391:1608–21.
8. Ferri F. Protozoal Infections. *Ferri's Color Atlas Text Clin Med*. Elsevier Health Sciences; 2009. p. 1159.
9. Doolan DL, Dobaño C, Baird JK. Acquired immunity to malaria. *Clin Microbiol Rev*. American Society for Microbiology (ASM); 2009;22:13–36, Table of Contents.
10. Babiker HA, Abdel-Muhsin AM, Ranford-Cartwright LC, Satti G, Walliker D. Characteristics of Plasmodium falciparum parasites that survive the lengthy dry season in eastern Sudan where malaria transmission is markedly seasonal. *Am J Trop Med Hyg*. 1998;59:582–90.
11. Zwetyenga J, Rogier C, Spiegel A, Fontenille D, Trape JF, Mercereau-Puijalon O. A cohort study of Plasmodium falciparum diversity during the dry season in Ndiop, a Senegalese village with seasonal, mesoendemic malaria. *Trans R Soc Trop Med Hyg*. 1999;93:375–80.
12. Kwiatkowski DP. How malaria has affected the human genome and what human genetics can teach us about malaria. *Am J Hum Genet*. Elsevier; 2005;77:171–92.
13. Harper K, Armelagos G. The changing disease-scape in the third epidemiological transition. *Int J Environ Res Public Health*. Multidisciplinary Digital Publishing Institute (MDPI); 2010;7:675–97.
14. Hippocrates. On The Sacred Disease.
15. Retief FP, Cilliers L. Diseases and causes of death among the popes. *Acta Theol Suppl*. 2005;233–46.
16. Hawass Z, Gad YZ, Ismail S, Khairat R, Fathalla D, Hasan N, et al. Ancestry and Pathology in King Tutankhamun's Family. *JAMA*. American Medical Association; 2010;303:638.
17. Sallares R, Bouwman A, Anderung C. The spread of malaria to Southern Europe in antiquity: new approaches to old problems. *Med Hist*. Cambridge University Press; 2004;48:311–28.

18. Bruce-Chwatt LJ. Alphonse Laveran's discovery 100 years ago and today's global fight against malaria. *J R Soc Med.* Royal Society of Medicine Press; 1981;74:531–6.
19. Rajakumar K, Weisse M. Centennial year of Ronald Ross' epic discovery of malaria transmission: an essay and tribute. *South Med J.* 1999;92:567–71.
20. Baccetti B. History of the early dipteran systematics in Italy: from Lyncei to Battista Grassi. *Parassitologia.* 2008;50:167–72.
21. Warren M. The Making of a Tropical Disease: A Short History of Malaria. Emerg. Infect. Dis. Centers for Disease Control and Prevention; 2008.
22. Hackett L. Malaria in Europe: an ecological study. Oxford: Oxford University Press; 1937.
23. Kuhn KG, Campbell-Lendrum DH, Armstrong B, Davies CR. Malaria in Britain: past, present, and future. *Proc Natl Acad Sci U S A.* National Academy of Sciences; 2003;100:9997–10001.
24. Majori G. Short history of malaria and its eradication in Italy with short notes on the fight against the infection in the mediterranean basin. *Mediterr J Hematol Infect Dis.* Catholic University in Rome; 2012;4:e2012016.
25. World Health Organization. WHO certifies Paraguay malaria-free. WHO Press. 2018;
26. World Health Organization. Kyrgyzstan receives WHO certification of malaria elimination. WHO Press. World Health Organization; 2016.
27. Bagcchi S. Sri Lanka declared malaria free. *BMJ.* British Medical Journal Publishing Group; 2016;354:i5000.
28. World Health Organization. World Malaria Report 2015. WHO Press. 2015.
29. World Health Organization. Eliminating Malaria. WHO Press. 2015.
30. Nájera JA, González-Silva M, Alonso PL. Some Lessons for the Future from the Global Malaria Eradication Programme (1955–1969). *PLOS Med.* Public Library of Science; 2011;8:e1000412.
31. Nájera JA. Malaria control: achievements, problems and strategies. *Parassitologia.* 2001;43:1–89.
32. Tangpukdee N, Duangdee C, Wilairatana P, Krudsood S. Malaria diagnosis: a brief review. *Korean J Parasitol.* Korean Society for Parasitology; 2009;47:93–102.
33. Murray CK, Gasser RA, Magill AJ, Miller RS, Miller RS. Update on rapid diagnostic testing for malaria. *Clin Microbiol Rev.* American Society for Microbiology (ASM); 2008;21:97–110.
34. Mahende C, Ngasala B, Lusingu J, Yong T-S, Lushino P, Lemnge M, et al. Performance of rapid diagnostic test, blood-film microscopy and PCR for the diagnosis of malaria infection among febrile children from Korogwe District, Tanzania. *Malar J.* BioMed Central; 2016;15:391.
35. Yang G, Kim D, Pham A, Paul C, Yang G, Kim D, et al. A Meta-Regression Analysis of the

- Effectiveness of Mosquito Nets for Malaria Control: The Value of Long-Lasting Insecticide Nets. *Int J Environ Res Public Health*. Multidisciplinary Digital Publishing Institute; 2018;15:546.
36. Pluess B, Tanser FC, Lengeler C, Sharp BL. Indoor residual spraying for preventing malaria. *Cochrane Database Syst Rev*. John Wiley & Sons, Ltd; 2010;
37. World Health Organization. Global Malaria Programme: Indoor residual spraying Use of indoor residual spraying for scaling up global malaria control and elimination. WHO Press. 2006.
38. World Health Organization. WHO Guidance Note for Estimating the Longevity of Long-Lasting Insecticidal Nets in Malaria Control. WHO Press. 2013.
39. Short R, Gurung R, Rowcliffe M, Hill N, Milner-Gulland EJ. The use of mosquito nets in fisheries: A global perspective. Munderloh UG, editor. *PLoS One*. Public Library of Science; 2018;13:e0191519.
40. Sarwar M. Source Reduction Practices for Mosquitoes (Diptera) Management to Prevent Dengue, Malaria and Other Arboborne Diseases. *Am J Clin Neurol Neurosurg*. 2015;1:110–6.
41. Wanji S, Mafo Ac FF, Tendongfor N, Tanga MC, Tchuente F, Bilong Bilong CF, et al. Spatial distribution, environmental and physicochemical characterization of Anopheles breeding sites in the Mount Cameroon region. *J Vector Borne Dis*. 2009;46:75–80.
42. Mattah PAD, Futagbi G, Amekudzi LK, Mattah MM, de Souza DK, Kartey-Attipoe WD, et al. Diversity in breeding sites and distribution of Anopheles mosquitoes in selected urban areas of southern Ghana. *Parasit Vectors*. BioMed Central; 2017;10:25.
43. Kitron U, Spielman A. Suppression of Transmission of Malaria Through Source Reduction: Antianopheline Measures Applied in Israel, the United States, and Italy. *Clin Infect Dis*. Oxford University Press; 1989;11:391–406.
44. RTS SCTP. Efficacy and safety of RTS,S/AS01 malaria vaccine with or without a booster dose in infants and children in Africa: final results of a phase 3, individually randomised, controlled trial. *Lancet*. Elsevier; 2015;386:31–45.
45. Ballou WR. The development of the RTS,S malaria vaccine candidate: challenges and lessons. *Parasite Immunol*. Wiley/Blackwell (10.1111); 2009;31:492–500.
46. Rathore D, Sacci JB, de la Vega P, McCutchan TF. Binding and invasion of liver cells by *Plasmodium falciparum* sporozoites. Essential involvement of the amino terminus of circumsporozoite protein. *J Biol Chem*. American Society for Biochemistry and Molecular Biology; 2002;277:7092–8.
47. Mahmoudi S, Keshavarz H. Efficacy of phase 3 trial of RTS, S/AS01 malaria vaccine: The need for an alternative development plan. *Hum Vaccin Immunother*. 2017;13:2098–101.
48. Olotu A, Fegan G, Wambua J, Nyangweso G, Leach A, Lievens M, et al. Seven-Year Efficacy of

- RTS,S/AS01 Malaria Vaccine among Young African Children. *N Engl J Med.* Massachusetts Medical Society; 2016;374:2519–29.
49. WHO Global Malaria Programme. WHO policy recommendation: Seasonal Malaria Chemoprevention for Plasmodium falciparum control in highly seasonal transmission areas of the Sahel sub-region in Africa. WHO Press. 2012.
50. Cissé B, Ba EH, Sokhna C, NDiaye JL, Gomis JF, Dial Y, et al. Effectiveness of Seasonal Malaria Chemoprevention in Children under Ten Years of Age in Senegal: A Stepped-Wedge Cluster-Randomised Trial. Noor AM, editor. *PLOS Med.* Public Library of Science; 2016;13:e1002175.
51. Tunçalp Ö, Pena-Rosas J, Lawrie T, Bucagu M, Oladapo O, Portela A, et al. WHO recommendations on antenatal care for a positive pregnancy experience-going beyond survival. *BJOG An Int J Obstet Gynaecol.* 2017;124:860–2.
52. Menendez C. Malaria during pregnancy. *Curr Mol Med.* 2006;6:269–73.
53. World Health Organization. WHO | Intermittent preventive treatment in pregnancy (IPTp) [Internet]. WHO Press. World Health Organization; 2018 [cited 2018 Aug 19]. Available from: [http://www.who.int/malaria/areas/preventive\\_therapies/pregnancy/en/](http://www.who.int/malaria/areas/preventive_therapies/pregnancy/en/)
54. Manson P, Cook GC (Gordon C, Zumla A, Manson P. Manson's tropical diseases. Saunders; 2003.
55. Butler AR, Khan S, Ferguson E. A brief history of malaria chemotherapy. *J R Coll Physicians Edinb.* 2010;40:172–7.
56. Mwanza S, Joshi S, Nambozi M, Chileshe J, Malunga P, Kabuya J-BB, et al. The return of chloroquine-susceptible Plasmodium falciparum malaria in Zambia. *Malar J.* BioMed Central; 2016;15:584.
57. Steinhardt LC, Magill AJ, Arguin PM. Review: Malaria chemoprophylaxis for travelers to Latin America. *Am J Trop Med Hyg.* The American Society of Tropical Medicine and Hygiene; 2011;85:1015–24.
58. Kapishnikov S, Berthing T, Hviid L, Dierolf M, Menzel A, Pfeiffer F, et al. Aligned hemozoin crystals in curved clusters in malarial red blood cells revealed by nanoprobe X-ray Fe fluorescence and diffraction. *Proc Natl Acad Sci.* 2012;109:11184–7.
59. Ginsburg H, Famin O, Zhang J, Krugliak M. Inhibition of glutathione-dependent degradation of heme by chloroquine and amodiaquine as a possible basis for their antimalarial mode of action. *Biochem Pharmacol.* 1998;56:1305–13.
60. Martin RE, Marchetti R V., Cowan AI, Howitt SM, Broer S, Kirk K. Chloroquine Transport via the Malaria Parasite's Chloroquine Resistance Transporter. *Science (80- ).* 2009;325:1680–2.
61. Lin J, Spaccapelo R, Schwarzer E, Sajid M, Annoura T, Deroost K, et al. Replication of Plasmodium

- in reticulocytes can occur without hemozoin formation, resulting in chloroquine resistance. *J Exp Med.* 2015;212:893–903.
62. Wootton JC, Feng X, Ferdig MT, Cooper RA, Mu J, Baruch DI, et al. Genetic diversity and chloroquine selective sweeps in *Plasmodium falciparum*. *Nature*. Nature Publishing Group; 2002;418:320–3.
63. Frosch AE, Venkatesan M, Laufer MK. Patterns of chloroquine use and resistance in sub-Saharan Africa: a systematic review of household survey and molecular data. *Malar J*. BioMed Central; 2011;10:116.
64. Barnes DA, Foote SJ, Galatis D, Kemp DJ, Cowman AF. Selection for high-level chloroquine resistance results in deamplification of the *pfdmdr1* gene and increased sensitivity to mefloquine in *Plasmodium falciparum*. *EMBO J*. 1992;11:3067–75.
65. Martin RE, Kirk K. The Malaria Parasite's Chloroquine Resistance Transporter is a Member of the Drug/Metabolite Transporter Superfamily. *Mol Biol Evol*. 2004;21:1938–49.
66. Ferone R. Folate metabolism in malaria. *Bull World Health Organ*. World Health Organization; 1977;55:291–8.
67. Hyde JE. Exploring the folate pathway in *Plasmodium falciparum*. *Acta Trop*. 2005;94:191–206.
68. Nzila A. The past, present and future of antifolates in the treatment of *Plasmodium falciparum* infection. *J Antimicrob Chemother*. Oxford University Press; 2006;57:1043–54.
69. Luzzatto L. The rise and fall of the antimalarial Lapdap: a lesson in pharmacogenetics. *Lancet*. 2010;376:739–41.
70. Heinberg A, Kirkman L. The molecular basis of antifolate resistance in *Plasmodium falciparum*: looking beyond point mutations. *Ann N Y Acad Sci*. NIH Public Access; 2015;1342.
71. Cowman AF, Morry MJ, Biggs BA, Cross GA, Foote SJ. Amino acid changes linked to pyrimethamine resistance in the dihydrofolate reductase-thymidylate synthase gene of *Plasmodium falciparum*. *Proc Natl Acad Sci U S A*. 1988;85:9109–13.
72. Wang P, Read M, Sims PFG, Hyde JE. Sulfadoxine resistance in the human malaria parasite *Plasmodium falciparum* is determined by mutations in dihydropteroate synthetase and an additional factor associated with folate utilization. *Mol Microbiol*. Wiley/Blackwell (10.1111); 1997;23:979–86.
73. Bacon DJ, Tang D, Salas C, Roncal N, Lucas C, Gerena L, et al. Effects of Point Mutations in *Plasmodium falciparum* Dihydrofolate Reductase and Dihydropteroate Synthase Genes on Clinical Outcomes and In Vitro Susceptibility to Sulfadoxine and Pyrimethamine. Sutherland CJ, editor. *PLoS One*. Public Library of Science; 2009;4:e6762.
74. Heinberg A, Siu E, Stern C, Lawrence EA, Ferdig MT, Deitsch KW, et al. Direct evidence for the

- adaptive role of copy number variation on antifolate susceptibility in *Plasmodium falciparum*. *Mol Microbiol.* NIH Public Access; 2013;88:702–12.
75. Nair S, Miller B, Barends M, Jaidee A, Patel J, Mayxay M, et al. Adaptive copy number evolution in malaria parasites. Przeworski M, editor. *PLOS Genet.* Public Library of Science; 2008;4:e1000243.
76. Guo Z. Artemisinin anti-malarial drugs in China. *Acta Pharm Sin B.* Elsevier; 2016;6:115–24.
77. Global Malaria Programme. Emergence and spread of artemisinin resistance calls for intensified efforts to withdraw oral artemisinin-based monotherapy from the market. WHO Press. 2014.
78. World Health Organization. Artemisinin and artemisinin-based combination therapy resistance. WHO Press. 2016.
79. Winzeler EA, Manary MJ. Drug resistance genomics of the antimalarial drug artemisinin. *Genome Biol.* 2014;15:544.
80. Cravo P, Napolitano H, Culleton R. How genomics is contributing to the fight against artemisinin-resistant malaria parasites. *Acta Trop.* 2015;148:1–7.
81. Noedl H, Se Y, Schaecher K, Smith BL, Socheat D, Fukuda MM, et al. Evidence of Artemisinin-Resistant Malaria in Western Cambodia. *N Engl J Med.* 2008;359:2619–20.
82. Ariey F, Witkowski B, Amaratunga C, Beghain J, Langlois A-C, Khim N, et al. A molecular marker of artemisinin-resistant *Plasmodium falciparum* malaria. *Nature.* Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved.; 2014;505:50–5.
83. Jambou R, Legrand E, Niang M, Khim N, Lim P, Volney B, et al. Resistance of *Plasmodium falciparum* field isolates to in-vitro artemether and point mutations of the SERCA-type PfATPase6. *Lancet.* 2005;366:1960–3.
84. World Health Organization. Artemisinin resistance and artemisinin-based combination therapy efficacy (Status report -- August 2018).
85. Miotto O, Amato R, Ashley EA, MacInnis B, Almagro-Garcia J, Amaratunga C, et al. Genetic architecture of artemisinin-resistant *Plasmodium falciparum*. *Nat Genet.* Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved.; 2015;47:226–34.
86. Mu J, Myers RA, Jiang H, Liu S, Ricklefs S, Waisberg M, et al. Plasmodium falciparum genome-wide scans for positive selection, recombination hot spots and resistance to antimalarial drugs. *Nat Genet.* Nature Publishing Group; 2010;42:268–71.
87. Harinasuta T, Suntharasamai P, Viravan C. Chloroquine-resistant falciparum malaria in Thailand. *Lancet.* Elsevier; 1965;286:657–60.
88. Wernsdorfer WH, Payne D. The dynamics of drug resistance in *Plasmodium falciparum*. *Pharmacol*

Ther. 1991;50:95–121.

89. Verdrager J. Epidemiology of the emergence and spread of drug-resistant falciparum malaria in South-East Asia and Australasia. *J Trop Med Hyg.* 1986;89:277–89.
90. Bushman M, Antia R, Udhayakumar V, de Roode JC. Within-host competition can delay evolution of drug resistance in malaria. Riley S, editor. *PLOS Biol.* Public Library of Science; 2018;16:e2005712.
91. World Health Organization. *Treatment of Severe Malaria.* WHO Press. 2015.
92. World Health Organization. WHO | False-negative RDT results and implications of new reports of *P. falciparum* histidine-rich protein 2/3 gene deletions. WHO Press. World Health Organization; 2018.
93. Borges S, Cravo P, Creasey A, Fawcett R, Modrzynska K, Rodrigues L, et al. Genomewide scan reveals amplification of *mdr1* as a common denominator of resistance to mefloquine, lumefantrine, and artemisinin in *Plasmodium chabaudi* malaria parasites. *American Society for Microbiology Journals;* 2011;55.
94. Martinsen ES, Perkins SL. The Diversity of *Plasmodium* and Other Haemosporidians: The Intersection of Taxonomy, Phylogenetics and Genomics. In: Carlton JM, Perkins SL, Deitsch KW, editors. *Malar Parasites Comp Genomics, Evol Mol Biol.* New York: Caister Academic Press; 2013. p. 1–15.
95. Faust C, Dobson AP. Primate malarias: Diversity, distribution and insights for zoonotic *Plasmodium.* One Heal. Elsevier; 2015;1:66–75.
96. Templeton TJ, Martinsen E, Kaewthamasorn M, Kaneko O. The rediscovery of malaria parasites of ungulates. *Parasitology.* Cambridge University Press; 2016;143:1501–8.
97. Zug GR. *Herpetology : an introductory biology of amphibians and reptiles.* Academic Press; 1993.
98. Valkiūnas G (Gediminas). Avian malaria parasites and other haemosporidia. CRC Press; 2005.
99. Martinsen ES, Perkins SL, Schall JJ. A three-genome phylogeny of malaria parasites (*Plasmodium* and closely related genera): Evolution of life-history traits and host switches. *Mol Phylogenet Evol.* Academic Press; 2008;47:261–73.
100. Gunalan K, Gao X, Yap SSL, Huang X, Preiser PR. The role of the reticulocyte-binding-like protein homologues of *Plasmodium* in erythrocyte sensing and invasion. *Cell Microbiol.* Wiley/Blackwell (10.1111); 2013;15:35–44.
101. Campo B, Vandal O, Wesche DL, Burrows JN. Killing the hypnozoite – drug discovery approaches to prevent relapse in *Plasmodium vivax.* *Pathog Glob Health.* Taylor & Francis; 2015;109:107–22.
102. Matsuda S, Matsumoto Y, Yoshida Y. Ultrastructure of Human Erythrocytes Infected with *Plasmodium Ovale.* *Am J Trop Med Hyg.* The American Society of Tropical Medicine and Hygiene; 1986;35:697–703.
103. Cao Y, Wang W, Liu Y, Cotter C, Zhou H, Zhu G, et al. The increasing importance of *Plasmodium*

- ovale and Plasmodium malariae in a malaria elimination setting: an observational study of imported cases in Jiangsu Province, China, 2011–2014. *Malar J. BioMed Central*; 2016;15:459.
104. Ta TH, Hisam S, Lanza M, Jiram AI, Ismail N, Rubio JM. First case of a naturally acquired human infection with Plasmodium cynomolgi. *Malar J. BioMed Central*; 2014;13:68.
105. Lalremruata A, Magris M, Vivas-Martínez S, Koehler M, Esen M, Kempaiah P, et al. Natural infection of Plasmodium brasiliense in humans: Man and monkey share quartan malaria parasites in the Venezuelan Amazon. *EBioMedicine*. Elsevier; 2015;2:1186–92.
106. Rodhain J, Dellaert R. Studies on Plasmodium schwetzi E. Brumpt. III. Plasmodium schwetzi infection in humans. *Ann la Soc belge Med Trop*. 1955;35:757–75.
107. Coatney GR, Chin W, Contacos PG, King HK. Plasmodium inui, a Quartan-Type Malaria Parasite of Old World Monkeys Transmissible to Man. *J Parasitol*. Allen PressThe American Society of Parasitologists; 1966;52:660.
108. Deane LM, Deane MP, Ferreira Neto J. Studies on transmission of simian malaria and on a natural infection of man with Plasmodium simium in Brazil. *Bull World Health Organ*. World Health Organization; 1966;35:805–8.
109. de Alvarenga DAM, Culleton R, de Pina-Costa A, Rodrigues DF, Bianco C, Silva S, et al. An assay for the identification of Plasmodium simium infection for diagnosis of zoonotic malaria in the Brazilian Atlantic Forest. *Sci Rep*. Nature Publishing Group; 2018;8:86.
110. Sinden RE. Sexual development of malarial parasites. *Adv Parasitol*. 1983;22:153–216.
111. Bray RS, Garnham PCC. The life-cycle of primate malaria parasites. *Br Med Bull*. Oxford University Press; 1982;38:117–22.
112. Mueller A-K, Kohlhepp F, Hammerschmidt C, Michel K. Invasion of mosquito salivary glands by malaria parasites: prerequisites and defense strategies. *Int J Parasitol*. NIH Public Access; 2010;40:1229–35.
113. Hoffman SL, Bancroft WH, Gottlieb M, James SL, Burroughs EC, Stephenson JR, et al. Funding for malaria genome sequencing. *Nature*. 1997;387:647.
114. Gardner MJ, Hall N, Fung E, White O, Berriman M, Hyman RW, et al. Genome sequence of the human malaria parasite Plasmodium falciparum. *Nature*. Europe PMC Funders; 2002;419:498–511.
115. Sheiner L, Vaidya AB, McFadden GI. The metabolic roles of the endosymbiotic organelles of Toxoplasma and Plasmodium spp. *Curr Opin Microbiol*. NIH Public Access; 2013;16:452–8.
116. Ralph SA, Foth BJ, Hall N, McFadden GI. Evolutionary Pressures on Apicoplast Transit Peptides. *Mol Biol Evol*. Oxford University Press; 2004;21:2183–94.

117. Scherf A, Lopez-Rubio JJ, Riviere L. Antigenic Variation in *Plasmodium falciparum*. *Annu Rev Microbiol.* 2008;62:445–70.
118. Goel S, Palmkvist M, Moll K, Joannin N, Lara P, R Akhouri R, et al. RIFINs are adhesins implicated in severe *Plasmodium falciparum* malaria. *Nat Med.* 2015;21:314–7.
119. Niang M, Yan Yam X, Preiser PR. The *Plasmodium falciparum* STEVOR Multigene Family Mediates Antigenic Variation of the Infected Erythrocyte. Rogerson SJ, editor. Public Library of Science; 2009;5:e1000307.
120. Hiss JA, Przyborski JM, Schwarte F, Lingelbach K, Schneider G. The *Plasmodium* export element revisited. *PLoS One.* Public Library of Science; 2008;3:e1560.
121. Oberli A, Slater LM, Cutts E, Brand F, Mundwiler-Pachlatko E, Rusch S, et al. A *Plasmodium falciparum* PHIST protein binds the virulence factor PfEMP1 and comigrates to knobs on the host cell surface. *FASEB J.* The Federation of American Societies for Experimental Biology; 2014;28:4420–33.
122. Berzins K, Perlmann H, Wählén B, Carlsson J, Wahlgren M, Udomsangpetch R, et al. Rabbit and human antibodies to a repeated amino acid sequence of a *Plasmodium falciparum* antigen, Pf 155, react with the native protein and inhibit merozoite invasion. *Proc Natl Acad Sci U S A.* National Academy of Sciences; 1986;83:1065–9.
123. Maier AG, Rug M, O'Neill MT, Brown M, Chakravorty S, Szestak T, et al. Exported Proteins Required for Virulence and Rigidity of *Plasmodium falciparum*-Infected Human Erythrocytes. *Cell.* 2008;134:48–61.
124. Pasternak ND, Dzikowski R. PfEMP1: An antigen that plays a key role in the pathogenicity and immune evasion of the malaria parasite *Plasmodium falciparum*. *Int J Biochem Cell Biol.* 2009;41:1463–6.
125. Schulze J, Kwiatkowski M, Borner J, Schlüter H, Bruchhaus I, Burmester T, et al. The *Plasmodium falciparum* exportome contains non-canonical PEXEL/HT proteins. *Mol Microbiol.* 2015;97:301–14.
126. Osborne AR, Speicher KD, Tamez PA, Bhattacharjee S, Speicher DW, Haldar K. The host targeting motif in exported *Plasmodium* proteins is cleaved in the parasite endoplasmic reticulum. *Mol Biochem Parasitol.* NIH Public Access; 2010;171:25–31.
127. Ruvalcaba-Salazar OK, Ramírez-Estudillo M del C, Montiel-Condado D, Recillas-Targa F, Vargas M, Hernández-Rivas R. Recombinant and native *Plasmodium falciparum* TATA-binding-protein binds to a specific TATA box element in promoter regions. *Mol Biochem Parasitol.* 2005;140:183–96.
128. Preston MD, Campino S, Assefa SA, Echeverry DF, Ocholla H, Amambua-Ngwa A, et al. A barcode of organellar genome polymorphisms identifies the geographic origin of *Plasmodium falciparum* strains.

Nat Commun. Nature Publishing Group; 2014;5.

129. Ponnudurai T, Leeuwenberg AD, Meuwissen JH. Chloroquine sensitivity of isolates of Plasmodium falciparum adapted to in vitro culture. *Trop Geogr Med*. 1981;33:50–4.
130. Burkot TR, Williams JL, Schneider I. Infectivity to mosquitoes of Plasmodium falciparum clones grown in vitro from the same isolate. *Trans R Soc Trop Med Hyg*. 1984;78:339–41.
131. Bhasin VK, Trager W. Gametocyte-forming and non-gametocyte-forming clones of Plasmodium falciparum. *Am J Trop Med Hyg*. 1984;33:534–7.
132. Guinet F, Dvorak JA, Fujioka H, Keister DB, Muratova O, Kaslow DC, et al. A developmental defect in Plasmodium falciparum male gametogenesis. *J Cell Biol*. 1996;135:269–78.
133. Campbell CC, Collins WE, Nguyen-Dinh P, Barber A, Broderson JR. Plasmodium falciparum gametocytes from culture in vitro develop to sporozoites that are infectious to primates. *Science*. 1982;217:1048–50.
134. MalariaGEN. Pf3k pilot data release 5 | MalariaGEN [Internet]. MalariaGEN. 2016 [cited 2018 Aug 19]. Available from: <https://www.malariagen.net/data/pf3k-5>
135. MalariaGEN. Pf3k pilot data release 3 | MalariaGEN [Internet]. MalariaGEN. 2015 [cited 2018 Aug 19]. Available from: <https://www.malariagen.net/data/pf3k-pilot-data-release-3>
136. Svenson JE, MacLean JD, Gyorkos TW, Keystone J. Imported malaria. Clinical presentation and examination of symptomatic travelers. *Arch Intern Med*. 1995;155:861–8.
137. Nhabomba AJ, Guinovart C, Jiménez A, Manaca MN, Quintó L, Cisteró P, et al. Impact of age of first exposure to Plasmodium falciparum on antibody responses to malaria in children: a randomized, controlled trial in Mozambique. *Malar J*. BioMed Central; 2014;13:121.
138. Jenkins R, Omollo R, Ongecha M, Sifuna P, Othieno C, Ongerl L, et al. Prevalence of malaria parasites in adults and its determinants in malaria endemic area of Kisumu County, Kenya. *Malar J*. BioMed Central; 2015;14:263.
139. Phillips A, Bassett P, Zeki S, Newman S, Pasvol G. Risk Factors for Severe Disease in Adults with Falciparum Malaria. *Clin Infect Dis*. Oxford University Press; 2009;48:871–8.
140. Ndila CM, Uyoga S, Macharia AW, Nyutu G, Peshe N, Ojal J, et al. Human candidate gene polymorphisms and risk of severe malaria in children in Kilifi, Kenya: a case-control association study. *Lancet Haematol*. Elsevier; 2018;5:e333–45.
141. World Health Organization. WHO | Severe malaria. WHO Press. World Health Organization; 2016.
142. Taylor WRJ, Hanson J, Turner GDH, White NJ, Dondorp AM. Respiratory manifestations of malaria. *Chest*. Elsevier; 2012;142:492–505.

143. Mackinnon MJ, Mwangi TW, Snow RW, Marsh K, Williams TN. Heritability of malaria in Africa. Foote S, editor. PLOS Med. Public Library of Science; 2005;2:e340.
144. Manjurano A, Clark TG, Nadjm B, Mtobe G, Wangai H, Sepulveda N, et al. Candidate human genetic polymorphisms and severe malaria in a Tanzanian population. Lafrenie R, editor. PLoS One. Public Library of Science; 2012;7:e47463.
145. Manjurano A, Sepúlveda N, Nadjm B, Mtobe G, Wangai H, Maxwell C, et al. African glucose-6-phosphate dehydrogenase alleles associated with protection from severe malaria in heterozygous females in Tanzania. Sirugo G, editor. PLOS Genet. Public Library of Science; 2015;11:e1004960.
146. Manjurano A, Sepúlveda N, Nadjm B, Mtobe G, Wangai H, Maxwell C, et al. USP38, FREM3, SDC1, DDC, and LOC727982 Gene Polymorphisms and Differential Susceptibility to Severe Malaria in Tanzania. J Infect Dis. 2015;
147. Network MGE, Malaria Genomic Epidemiology Network. A novel locus of resistance to severe malaria in a region of ancient balancing selection. Nature. Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved.; 2015;advance on:253–7.
148. Leffler EM, Band G, Busby GBJ, Kivinen K, Le QS, Clarke GM, et al. Resistance to malaria through structural variation of red blood cell invasion receptors. Science. American Association for the Advancement of Science; 2017;356:eaam6393.
149. Jallow M, Teo YY, Small KS, Rockett KA, Deloukas P, Clark TG, et al. Genome-wide and fine-resolution association analysis of malaria in West Africa. Nat Genet. Nature Publishing Group; 2009;41:657–65.
150. Timmann C, Thye T, Vens M, Evans J, May J, Ehmen C, et al. Genome-wide association study indicates two novel resistance loci for severe malaria. Nature. Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved.; 2012;489:443–6.
151. Fernández-Grandon GM, Gezan SA, Armour JAL, Pickett JA, Logan JG. Heritability of Attractiveness to Mosquitoes. Hansen IA, editor. PLoS One. Public Library of Science; 2015;10:e0122716.
152. Williams TN, Mwangi TW, Wambua S, Alexander ND, Kortok M, Snow RW, et al. Sickle Cell Trait and the Risk of Plasmodium falciparum Malaria and Other Childhood Diseases. J Infect Dis. 2005;192:178–86.
153. Platt OS, Brambilla DJ, Rosse WF, Milner PF, Castro O, Steinberg MH, et al. Mortality In Sickle Cell Disease -- Life Expectancy and Risk Factors for Early Death. N Engl J Med. 1994;330:1639–44.
154. Galanello R, Cao A. Alpha-thalassemia. Genet Med. Nature Publishing Group; 2011;13:83–8.
155. Flint J, Hill AVS, Bowden DK, Oppenheimer SJ, Sill PR, Serjeantson SW, et al. High frequencies of

- $\alpha$ -thalassaemia are the result of natural selection by malaria. *Nature*. 1986;321:744–50.
156. Efferth T, Schwarzl SM, Smith J, Osieka R. Role of glucose-6-phosphate dehydrogenase for oxidative stress and apoptosis. *Cell Death Differ*. Nature Publishing Group; 2006;13:527–8.
157. Mbanefo EC, Ahmed AM, Titouna A, Elmaraezy A, Trang NTH, Phuoc Long N, et al. Association of glucose-6-phosphate dehydrogenase deficiency and malaria: a systematic review and meta-analysis. *Sci Rep*. Nature Publishing Group; 2017;7:45963.
158. Howes RE, Battle KE, Satyagraha AW, Baird JK, Hay SI. G6PD Deficiency. *Adv Parasitol*. 2013. p. 133–201.
159. Ryan JR, Stoute JA, Amon J, Dunton RF, Matalib R, Koros J, et al. Evidence for transmission of *Plasmodium vivax* among a duffy antigen negative population in Western Kenya. *Am J Trop Med Hyg*. 2006;75:575–81.
160. Kraisin S, Naka I, Patarapotikul J, Nantakomol D, Nuchnoi P, Hananantachai H, et al. Association of ADAMTS13 polymorphism with cerebral malaria. *Malar J*. 2011;10:366.
161. MalariaGEN, Rockett KA, Clarke GM, Fitzpatrick K, Hubbard C, Jeffreys AE, et al. Reappraisal of known malaria resistance loci in a large multicenter study. *Nat Genet*. Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved.; 2014;46:1197–204.
162. Gibbs RA, Boerwinkle E, Doddapaneni H, Han Y, Korchina V, Kovar C, et al. A global reference for human genetic variation. *Nature*. Nature Publishing Group; 2015;526:68–74.
163. Behjati S, Tarpey PS. What is next generation sequencing? *Arch Dis Child Educ Pract Ed*. Royal College of Paediatrics and Child Health; 2013;98:236–8.
164. International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature*. Nature Publishing Group; 2001;409:860–921.
165. Visscher PM, Wray NR, Zhang Q, Sklar P, McCarthy MI, Brown MA, et al. 10 Years of GWAS Discovery: Biology, Function, and Translation. *Am J Hum Genet*. Elsevier; 2017;101:5–22.
166. Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, et al. Genome sequencing in microfabricated high-density picolitre reactors. *Nature*. Nature Publishing Group; 2005;437:376–80.
167. Rusk N. Torrents of sequence. *Nat Methods*. 2011;8:44–44.
168. Meyer M, Kircher M. Illumina sequencing library preparation for highly multiplexed target capture and sequencing. *Cold Spring Harb Protoc*. Cold Spring Harbor Laboratory Press; 2010;2010:pdb.prot5448.
169. Sheridan C. Illumina claims \$1,000 genome win. *Nat Biotechnol*. 2014;32:115–115.
170. Treangen TJ, Salzberg SL. Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nat Rev Genet*. NIH Public Access; 2011;13:36–46.

171. Khost DE, Eickbush DG, Larracuente AM. Single-molecule sequencing resolves the detailed structure of complex satellite DNA loci in *Drosophila melanogaster*. *Genome Res.* Cold Spring Harbor Laboratory Press; 2017;27:709–21.
172. Jain M, Koren S, Miga KH, Quick J, Rand AC, Sasani TA, et al. Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nat Biotechnol.* Nature Publishing Group; 2018;36:338–45.
173. Patel P, Bharti PK, Bansal D, Ali NA, Raman RK, Mohapatra PK, et al. Prevalence of mutations linked to antimalarial resistance in *Plasmodium falciparum* from Chhattisgarh, Central India: A malaria elimination point of view. *Sci Rep.* Nature Publishing Group; 2017;7:16690.
174. Weatherall DJ, Clegg JB. Genetic variability in response to infection: malaria and after. *Genes Immun.* Nature Publishing Group; 2002;3:331–7.
175. Frazer KA, Murray SS, Schork NJ, Topol EJ. Human genetic variation and its contribution to complex traits. *Nat Rev Genet.* Nature Publishing Group; 2009;10:241–51.
176. Cadzow M, Boocock J, Nguyen HT, Wilcox P, Merriman TR, Black MA. A bioinformatics workflow for detecting signatures of selection in genomic data. *Front Genet. Frontiers*; 2014;5:293.
177. Salathia N, Lee HN, Sangster TA, Morneau K, Landry CR, Schellenberg K, et al. Indel arrays: an affordable alternative for genotyping. *Plant J.* Wiley/Blackwell (10.1111); 2007;51:727–37.
178. Nielsen R, Paul JS, Albrechtsen A, Song YS. Genotype and SNP calling from next-generation sequencing data. *Nat Rev Genet.* Nature Publishing Group; 2011;12:443–51.
179. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics.* Oxford University Press; 2009;25:1754–60.
180. Li H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics.* 2018;
181. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics.* Oxford University Press; 2009;25:2078–9.
182. Nei M. Analysis of gene diversity in subdivided populations. *Proc Natl Acad Sci U S A.* National Academy of Sciences; 1973;70:3321–3.
183. Wright S. The genetic structure of populations. *Ann Eugen.* Wiley/Blackwell (10.1111); 1949;15:323–54.
184. Weir BS, Cockerham CC. Estimating F-Statistics for the Analysis of Population Structure. *Evolution (N Y).* Society for the Study of Evolution; 1984;38:1358.
185. Tajima F. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics.* Genetics Society of America; 1989;123:585–95.
186. Voight BF, Kudaravalli S, Wen X, Pritchard JK. A map of recent positive selection in the human

- genome. Hurst L, editor. PLOS Biol. Public Library of Science; 2006;4:e72.
187. Sabeti PC, Varilly P, Fry B, Lohmueller J, Hostetter E, Cotsapas C, et al. Genome-wide detection and characterization of positive selection in human populations. *Nature*. Nature Publishing Group; 2007;449:913–8.
188. Liu X, Kanduri C, Oikkonen J, Karma K, Raijas P, Ukkola-Vuoti L, et al. Detecting signatures of positive selection associated with musical aptitude in the human genome. *Sci Rep*. Nature Publishing Group; 2016;6:21198.
189. Sabeti PC, Reich DE, Higgins JM, Levine HZP, Richter DJ, Schaffner SF, et al. Detecting recent positive selection in the human genome from haplotype structure. *Nature*. Nature Publishing Group; 2002;419:832–7.
190. Lupski JR, de Oca-Luna RM, Slaugenhoupt S, Pentao L, Guzzetta V, Trask BJ, et al. DNA duplication associated with Charcot-Marie-Tooth disease type 1A. *Cell*. Elsevier; 1991;66:219–32.
191. Rausch T, Zichner T, Schlattl A, Stutz AM, Benes V, Korbel JO, et al. DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics*. Oxford University Press; 2012;28:i333–9.
192. Layer RM, Chiang C, Quinlan AR, Hall IM. LUMPY: a probabilistic framework for structural variant discovery. *Genome Biol*. BioMed Central; 2014;15:R84.
193. Boeva V, Popova T, Bleakley K, Chiche P, Cappo J, Schleiermacher G, et al. Control-FREEC: a tool for assessing copy number and allelic content using next-generation sequencing data. *Bioinformatics*. Oxford University Press; 2012;28:423–5.
194. Abyzov A, Urban AE, Snyder M, Gerstein M. CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. *Genome Res*. Cold Spring Harbor Laboratory Press; 2011;21:974–84.
195. Benjamini Y, Speed TP. Summarizing and correcting the GC content bias in high-throughput sequencing. *Nucleic Acids Res*. Oxford University Press; 2012;40:e72.

## **Chapter 2:**

Characterising the impact of sustained  
sulfadoxine/pyrimethamine use upon the *Plasmodium falciparum*  
population in Malawi

## RESEARCH PAPER COVER SHEET

***PLEASE NOTE THAT A COVER SHEET MUST BE COMPLETED FOR EACH RESEARCH PAPER INCLUDE IN A THESIS.***

### **SECTION A – Student Details**

<b>Student</b>	Matt Ravenhall
<b>Principal Supervisor</b>	Taane Clark
<b>Thesis Title</b>	A bioinformatic analysis of malaria host and pathogen genomics

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

### **SECTION B – Paper already published**

Where was the work published?	Malaria Journal		
When was the work published?	29th November 2016		
If the work was published prior to registration for your research degree, give a brief rationale for its inclusion	n/a		
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?	n/a
Please list the paper's authors in the intended authorship order:	n/a
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 conducted all post-alignment data cleaning and analysis, produced all figures, and co-wrote the final manuscript under Taane Clark and Susana Campino's joint supervision.
--	--

**Student Signature:** \_\_\_\_\_

**Date:** \_\_\_\_\_

**Supervisor Signature:** \_\_\_\_\_

**Date:** \_\_\_\_\_

RESEARCH

Open Access



# Characterizing the impact of sustained sulfadoxine/pyrimethamine use upon the *Plasmodium falciparum* population in Malawi

Matt Ravenhall<sup>1</sup> , Ernest Diez Benavente<sup>1</sup>, Mwapatsa Mpando<sup>2</sup>, Anja T. R. Jensen<sup>3</sup>, Colin J. Sutherland<sup>1</sup>, Cally Roper<sup>1</sup>, Nuno Sepúlveda<sup>1,4</sup>, Dominic P. Kwiatkowski<sup>5</sup>, Jacqui Montgomery<sup>6,7</sup>, Kamiya S. Phiri<sup>8</sup>, Anja Terlouw<sup>6,7</sup>, Alister Craig<sup>6</sup>, Susana Campino<sup>1,\*†</sup>, Harold Ocholla<sup>7,8\*†</sup> and Taane G. Clark<sup>1,9\*†</sup>

## Abstract

**Background:** Malawi experienced prolonged use of sulfadoxine/pyrimethamine (SP) as the front-line anti-malarial drug, with early replacement of chloroquine and delayed introduction of artemisinin-based combination therapy. Extended use of SP, and its continued application in pregnancy is impacting the genomic variation of the *Plasmodium falciparum* population.

**Methods:** Whole genome sequence data of *P. falciparum* isolates covering 2 years of transmission within Malawi, alongside global datasets, were used. More than 745,000 SNPs were identified, and differences in allele frequencies between countries assessed, as well as genetic regions under positive selection determined.

**Results:** Positive selection signals were identified within *dhps*, *dhfr* and *gch1*, all components of the parasite folate pathway associated with SP resistance. Sitting predominantly on a *dhfr* triple mutation background, a novel copy number increase of ~twofold was identified in the *gch1* promoter. This copy number was almost fixed (96.8% frequency) in Malawi samples, but found at less than 45% frequency in other African populations, and distinct from a whole gene duplication previously reported in Southeast Asian parasites.

**Conclusions:** SP resistance selection pressures have been retained in the Malawian population, with known resistance *dhfr* mutations at fixation, complemented by a novel *gch1* promoter duplication. The effects of the duplication on the fitness costs of SP variants and resistance need to be elucidated.

## Background

Malawi suffers a heavy burden of endemic falciparum malaria with year-round transmission that peaks during the long rainy season from early December to May [1]. Malaria still accounts for 40% of hospitalizations in children under 5 years of age and 30% of all outpatient visits [2]. The malaria mortality rate is 63 per 100,000

population, and amongst the highest in East Africa [3] despite the roll out of control measures, such as insecticide-treated bed nets (ITNs), intensive indoor residual spraying (IRS), and artemisinin-based combination therapy (ACT) [2]. As one of the first African countries to switch from chloroquine to sulfadoxine/pyrimethamine (SP) in 1993, and the last to switch from SP to ACT in 2007, Malawi stands out from the rest of Africa in having a significantly prolonged exposure to SP [4]. Whilst this meant the reduced frequency of chloroquine resistance alleles in the *Plasmodium* population [5], the same cannot currently be said for SP resistance [6]. Resistance to SP is thought to be a cumulative process whereby mutations are successively acquired in both the *dhfr* (S108N, N51I, C59R, then I164L) and *dhps* (A437G then K540E)

\*Correspondence: susana.campino@lshtm.ac.uk;

harold.ocholla@gmail.com; taane.clark@lshtm.ac.uk

<sup>†</sup>Susana Campino, Harold Ocholla and Taane G. Clark are joint senior authors

<sup>1</sup> Faculty of Infectious and Tropical Diseases, London School of Hygiene and Tropical Medicine, London, UK

<sup>2</sup> Malawi-Liverpool-Wellcome Trust Clinical Research Programme, College of Medicine, University of Malawi, Blantyre, Malawi

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



© The Author(s) 2016. This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated.

genes. These *dhps* and *dhfr* polymorphisms persist at high frequency in the Malawian *Plasmodium* population despite exposure to SP being reduced. Retention of these variants may be due in part to the use of SP in intermittent preventive treatment for pregnant women (IPTp) or a lower than expected fitness cost associated with these variants [6]. The scaling up of the distribution of ACT and IPTp contributed to a 36% drop in the mortality rate between 2004 and 2014 for children under 5 years of age, to an estimated 85 deaths per 1000 live births [2]. However, the control efforts could be derailed by the use of SP in IPTp strategies in sites where parasites resistant to SP persist, as well as any future emergence and spread of ACT-resistant parasites (as seen in Southeast Asia).

Genetic variation in *Plasmodium falciparum* is central to the parasite's survival and can potentially undermine malaria control interventions. The evolutionary process enables parasite populations to select for variants that rapidly overcome host immune responses and anti-malarial drugs to establish persistent infections and increase transmission. Therefore, surveying evolutionarily driven genetic changes in *P. falciparum* and investigating parasite responses to anti-malarial interventions are crucial to efforts to reduce the malaria burden. Where regions are working towards pre-elimination, longitudinal samples are required to monitor parasite transmission dynamics and the efficacy of interventions to control malaria.

Previous work in a rural Malawian *P. falciparum* population in the Chikwawa district in a single 'baseline' season ( $n = 69$ , December 2010 to July 2011) identified several genes encoding merozoite invasion ligands as being retained in the population due to balancing selection [7]. This type of selection actively maintains multiple alleles in the gene pool of a population. Further, by comparing the Malawian *P. falciparum* population to others in Africa and Southeast Asia, signals of recent positive selection were identified at known drug targets (e.g., *dhps*, *crt* and *mdr1*), metabolic enzymes (e.g., *gch1*), and in several invasion ligands (e.g., *msp3.8*, *trap* and *ama1*) [7]. Initial analysis provided evidence of population divergence presumably driven by drug selection on *crt*, *dhps* and *mdr1* genes, and reflects the adaptation of parasite populations to local drug pressure, especially SP. The *dhps* (sulfadoxine target) and *dhfr* (pyrimethamine) genes are on the folate biosynthesis pathway of *P. falciparum*, and in Southeast Asian populations a copy number variant in *gch1* (first gene in pathway) is thought to be associated with SP resistance and its persistence [8].

The initial work in Malawi [7] was followed up by including additional 'baseline' season samples ( $n = 29$ , total  $n = 98$ ) and comparing the genetic diversity in 122 isolates collected in the subsequent 2012 dry and wet

seasons in the Chikwawa and Zomba districts, located 100 km apart. These regions are sentinel sites in Malawi, chosen for intensive anti-malarial intervention involving ACT, ITNs and IRS. The aim was to identify changes in allele frequency within individual, intra- and inter-season (wet and dry seasons) and identify regions under selection pressure. The findings demonstrate limited variation between the Malawian sub-populations over time and the impact of prolonged exposure of parasites to SP. Particularly, fixation of several known SP resistance SNPs and a novel copy number increase of the *gch1* promoter region were identified. Cross-population analysis revealed selective pressure for chloroquine resistance in non-Malawian populations. These populations have experienced prolonged use of chloroquine. Overall, the findings support the use of parasite and population genetic approaches to monitor transmission and the adaptation to drug pressure, and thereby inform the timing and type of interventions to be applied. Existing surveillance could be enhanced with rapid, field-based, genomic tests which genotype the *gch1* promoter region as a proxy for SP resistance in an African setting.

## Methods

### Study sites and sample collection

Whole blood samples were collected from October 2010 to November 2012 from children aged 5–28 months recruited in an ongoing ACTia[abbrev?] study within the high-transmission Chikwawa and Zomba regions in Malawi [7]. All individuals recruited had clinical falciparum malaria and received artemether/lumefantrine (AL) or dihydroartemisinin/piperaquine (DHA) treatment post-collection. Written informed consent was obtained from a parent or guardian of each child with the ethics committees of the University of Malawi's College of Medicine and the Liverpool School of Tropical Medicine both approving the study.

### Whole-genome sequencing and quality control

Human DNA contamination was reduced through leukocyte-depletion of the blood samples using CF11 column filtration [9]. Purified DNA samples ( $n = 220$ ) containing less than 30% human DNA were sequenced at the Sanger Institute using Illumina HiSeq2500 technology, with a minimum of 76-base, paired-end, fragment sizes. All short reads were mapped to the 3D7 reference genome (version 3.0) using *bwa-mem* [10]. SNPs and small indels were called using samtools and bcf/vcftools with default settings [10]. Only those variants with quality scores in excess of 30 (indicating an error rate less than one per 1000 bp) and with minimum coverage of ten were retained [10]. Genotypes at SNP positions were called using ratios of coverage and heterozygous calls were

converted to the majority genotype on a 70:30 coverage ratio or greater [7, 11, 12]. SNPs were excluded from analysis if they had more than 5% mixed or missing genotype calls, or they were positioned within non-unique regions, sub-telomeric regions or within the hypervariable *var*, *rifin* and *stevor* gene families.

Raw sequencing data were also mapped for previously published *P. falciparum* strains (3D7, HB3, DD2, 7G8, GB4) [11] and isolates from East Africa (Kenya, Tanzania, n = 33), West Africa (Burkina Faso, The Gambia, Ghana, Guinea, Mali, Nigeria, n = 430), Central Africa [Democratic Republic of Congo (DRC), n = 56], South America (Colombia, Peru, n = 21), South Asia (Bangladesh, n = 54) and Southeast Asia (Cambodia, Laos, Myanmar, Papua New Guinea, Thailand, Vietnam, n = 1187) [7, 11–14] using the pipeline described above. 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.malariaigen.net/projects/pf3k>). In total, the isolate dataset contains 745,913 high quality SNPs; 245,215 SNPs have no missing genotype calls, 77.1% within genes and 9.1% have a minor allele frequency greater than 1%.

#### Statistical analysis

Population stratification in the isolates was investigated using a principal component analysis (PCA) of the pairwise SNP distances between samples. This approach identified distinct African, Asian and South American clusters, with a further African-only investigation identifying West, Central and East African clusters. Differences in allele frequencies at each SNP were estimated using fixation indexes ( $F_{ST}$ ), with genes ranked by their maximum scores [15]. Tajima's D [16] was implemented to identify genomic regions under balancing selection, with a score greater than two suggesting strong balancing selection. Scores were calculated for each gene containing at least four SNPs. All SNP-based analyses were performed using base R functions. Copy number variation in isolates and strains was analysed with DELLY using default settings [17].

Extended haplotype homozogosity (EHH)-based selection analyses, intra-population iHS and inter-population XP-EHH, were performed using selscan [18], using the default minor allele frequency and EHH truncation values of 0.05. Pair-wise country XP-EHH analyses used Malawi as the reference group, and both iHS and XP-EHH values were normalized genome-wide. P values for iHS and XP-EHH estimates were calculated using a Gaussian approximation. A significance threshold of  $P < 0.00006$  was established for both iHS (>4) and XP-EHH (>6), using a simulation approach.

## Results

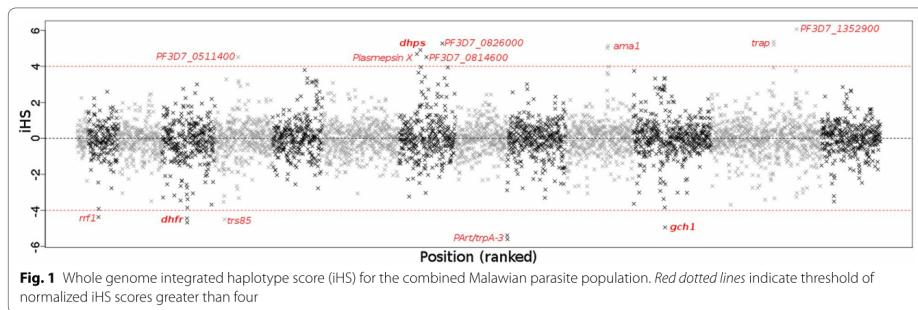
### Malawi sub-population analysis over time and location

Potential stratification within the Malawi dataset, across season, year and location of collection, was explored before consideration of shared signals. Of the 220 parasite isolates collected in this study, 85.9% were from the Chikwawa region whilst only 14.1% were from the Zomba region. Season-wise 43.2% were collected in a wet season, 56.8% in a dry season, across three years (2010 5.0%, 2011 39.5%, 2012 55.5%). Allele frequency differences between the location (median  $F_{ST}$  0.003, max. 0.11) and year (median  $F_{ST}$  0.004, max. 0.07) sub-populations were small. Within and between the seasons and location, the number of SNP differences was similar (~9400 SNPs) (see Additional file 1: Table S1). Twenty-six SNPs have an  $F_{ST}$  value greater than 0.15 for within-Malawi year, location and season-based sub-populations. Seven of these are intergenic and four within 'conserved unknown' genes (see Additional file 1: Table S2). The remaining 15 SNPs are within putative or known genes, including *surfin14.1* (max.  $F_{ST}$  0.172), *heat shock protein 90* (max.  $F_{ST}$  0.166) and the immune evasion antigen *PfEMP1* (max.  $F_{ST}$  0.157) [19]. Notably, all top pair-wise  $F_{ST}$  values above 0.15 were from season-based comparisons. No major differences were detected in allele frequency for known drug resistance mutations (see Additional file 1: Table S3).

### Selection within the combined Malawian parasite sub-populations

Given the absence of strong stratification between the sub-populations, the Malawian datasets were combined to identify selection signals within population. Signatures of potential recent positive selection were identified within 13 genes (iHS > 4). Five of these genes are currently uncharacterized, and three within close proximity to genes of known function and established selection sites (see Fig. 1; Additional file 1: Table S4). For example, the *PF3D7\_1223400* signal (iHS: -4.948) is within 2 kbp of *gch1*, and loci were identified within 60 kbp of *dhp5* and 8.5 kbp of *dhp7*, both members of the *P. falciparum* folate pathway [20]. Selection in these genomic regions within Malawi has been suggested previously [11] and relates to selective pressure due to use of SP.

Additional allele selection signals are present within the erythrocyte invasion critical protein *ama1* [21] (iHS: 5.129, 5.017) and the sporozoite surface protein *trap* (iHS: 5.366, 5.209), both representing hits that have been suggested previously [22, 23]. Non-reference allele selection loci were also found within *Plasmepsin X* (iHS: 4.686), which has a role in ookinete invasion [24]. In contrast, reference allele selection was identified centred on four loci within *TRS85* (iHS: -4.502), *RRF1* (iHS: -4.376) and



*PArt/TrpA-3* (iHS:  $-5.604$ ,  $-5.396$ ). Three genes show potential balancing signals within the Malawian *P. falciparum* population (Tajima's D > 2; see Additional file 1: Table S5). These are *ama1*, required for erythrocyte invasion [25], *surf8.2*, a group A SURFIN protein not found to be expressed post-invasion [26] but associated with susceptibility to pyrimethamine (inhibits folic acid metabolism via *dhfr*) [27], and a conserved unknown protein (*PF3D7\_0710200*) within chromosome 7. The *ama1* locus has both balancing and positive selection, suggesting positive selection for one specific haplotype alongside the retention of multiple others. Such complex selection has previously been identified as differing between *ama1* domains [27, 28].

#### Malawi within a global context

Parasite genetic diversity within Malawi was further contextualized within a global setting. A PCA approach identified distinct African, Asian and South American clusters, as previously reported using large SNP datasets [7, 11, 12, 14, 29]. Within the African-only analysis, distinct West, Central and East African clusters were also distinguished, and Malawi was separated from Tanzania and Kenya. (see Additional file 2: Figure S1).

Pair-wise  $F_{ST}$  values were calculated genome-wide for Malawi against each country with at least 14 samples. The median  $F_{ST}$  values per region per SNP were then estimated to identify those that contained Malawi-unique variants (see Table 1). The top hits included the *dhps* K540E causal mutation, likely reflecting the prevalence of SP resistance in Malawi and other East African populations considered, compared to elsewhere (Table 2; Additional file 1: Table S6). Of the remaining top hits, the study identified the *P47* and *P230* genes that share roles in gametocyte fertility [30], *P48/45* reflecting differences in mosquito vectors [30], and *PF3D7\_0913900* a putative arginine-tRNA ligase. When considering drug-resistant candidate polymorphism, *crt*

mutations (K76T, Q271E, N326S, I356T) were absent in Malawi, reflecting early withdrawal of chloroquine compared to other African countries (see Table 2; Additional file 1: Table S6). K76T and Q271E mutations are near fixed in Asia, and known to have undergone 'hard' selective sweeps [31]. Compared to other African populations, Malawi has a higher frequency of *dhfr* triple mutant haplotypes (N51I/C59R/S108N) and *dhfr/dhps* quintuple mutant genotypes (*dhfr* N51I/C59R/S108N haplotype and *dhps* A437G/K540E haplotype). The *dhps* S436A mutation was at high frequency in West Africa, and almost absent in Malawi. The contributing *dhfr* quadruple mutation (I164L) and *dhfr/dhps* sextuple mutant genotypes (*dhfr/dhps* quintuple mutant genotype and *dhfr* I164L) were only present in Asia. In Malawi, the *dhps* A581G mutation, which has been shown to reduce the effectiveness of SP preventive therapy [32] was present at low frequency, leading to the presence of an alternative *dhfr/dhps* sextuple genotype (*dhfr/dhps* quintuple genotype and *dhfr* I164L). In the Malawian population, no variants of the *kelch13* gene, previously described in Southeast Asia to be associated with artemisinin resistance [33], were identified. All alternative alleles in *kelch13* are present at low frequencies, the maximum at 0.091 for K189T, and reflect previous SNPs frequencies for the rest of Africa [34].

The XP-EHH method was used to identify regions under selection in the Malawi population compared to others. Positive values suggest relative selection in Malawi, whilst negative values suggest selection in non-Malawi (see Additional file 2: Figure S2; Additional file 1: Table S7). 31 genes ( $|XP-EHH| > 6$ ) were identified, of which 18 are uncharacterized and 14 appear against only one other country. The most striking signals were within *PF3D7\_1223400* and *PF3D7\_1223500*, both uncharacterized but within 2 kbp of *gch1*, and indicate relative selection within the Malawian population when compared to Burkina Faso, the DRC, The Gambia, Ghana, Guinea,

**Table 1 Differences in allele frequencies between Malawi and other populations (based on pairwise  $F_{ST}$  scores)**

Gene ID	Position	Gene	Other East Africa	DRC	West Africa	South Asia	Southeast Asia	South America
PF3D7_0810800	549,993	<i>dhps</i>	0.053	0.873	0.974	0.118	0.415	0.628
PF3D7_0209000	375,427	<i>P230</i>	0.023	0.035	0.689	0.722	0.569	0.775
PF3D7_1016500	663,199	<i>PHISTc</i>	0.023	0.077	0.504	0.463	0.174	0.775
PF3D7_0913900	596,674	Arginine-tRNA ligase	0.017	0.014	0.769	0.665	0.719	0.518
PF3D7_1339700	1,595,988	Conserved unknown	0.013	0.017	0.323	0.555	0.731	0.213
PF3D7_1032100	1,293,621	<i>dcp1</i>	0.021	0.132	0.709	0.578	0.659	0.485
PF3D7_1223500	958,593	Conserved unknown (near <i>gch1</i> )	0.093	0.285	0.535	0.668	0.667	0.668
PF3D7_1361800	2,481,275	Conserved unknown	0.02	0.019	0.533	0.630	0.606	0.665
PF3D7_0708500	385,921	<i>hsp86</i>	0.292	0.297	0.428	0.542	0.625	0.016
PF3D7_1346800	1,880,114	<i>P47</i>	0.055	0.26	0.434	0.590	0.590	0.590
PF3D7_0113000	489,337	<i>garp</i>	0.024	0.064	0.562	0.587	0.478	0.104
PF3D7_0716700	730,051	Conserved unknown	0.034	0.079	0.366	0.533	0.567	0.567
PF3D7_0307900	350,293	Conserved unknown	0.016	0.169	0.524	0.267	0.487	0.524
PF3D7_1248700	1,997,660	Conserved unknown	0.131	0.211	0.457	0.426	0.457	0.459
PF3D7_1223400	942,564	Phospholipid-transferring ATPase (near <i>gch1</i> )	0.172	0.237	0.380	0.444	0.444	0.444
PF3D7_1223300	938,341	<i>gyrA</i> (near <i>gch1</i> )	0.146	0.237	0.374	0.444	0.308	0.444
PF3D7_1223400	941,821	Phospholipid-transferring ATPase (near <i>gch1</i> )	0.171	0.216	0.381	0.442	0.442	0.442
PF3D7_1426700	1,036,865	<i>pepc</i>	0.022	0.001	0.429	0.251	0.419	0.430
PF3D7_0215300	629,060	<i>acs8</i>	0.174	0.104	0.402	0.393	0.426	0.132
PF3D7_1346700	1,876,606	<i>P48/45</i>	0.120	0.042	0.324	0.426	0.402	0.220
PF3D7_0615900	665,589	Conserved unknown	0.002	0.261	0.37	0.311	0.343	0.373
PF3D7_1433900	1,362,042	Putative protein kinase	0.051	0.012	0.233	0.368	0.340	0.368
PF3D7_0811600	586,054	Conserved unknown	0.003	0.065	0.340	0.358	0.304	0.358
PF3D7_0724900	1,056,801	Putative kinesin-19	0.010	0.099	0.220	0.358	0.344	0.358
PF3D7_1414200	564,437	Conserved unknown	0.003	0.083	0.278	0.333	0.333	0.333
PF3D7_1021700	883,159	Conserved unknown	0.027	0.201	0.317	0.300	0.331	0.331
PF3D7_0627800	1,115,191	Putative acetyl-CoA synthetase	0.042	0.325	0.325	0.325	0.325	0.325
PF3D7_1218300	718,254	<i>Ap2mu</i>	0.029	0.312	0.309	0.312	0.312	0.312
PF3D7_1223100	928,407	<i>PKAr</i>	0.231	0.306	0.306	0.306	0.306	0.306
PF3D7_1222600	911,963	<i>AP2-G</i>	0.218	0.274	0.304	0.304	0.297	0.304
PF3D7_1223300	935,411	<i>gyrA</i> (near <i>gch1</i> )	0.213	0.306	0.306	0.306	0.306	0.306
PF3D7_0932800	1,306,240	Conserved unknown	0.063	0.274	0.304	0.304	0.297	0.304

Values are the median  $F_{ST}$  values for each regional population. In italics are regional medians above 0.5

Kenya, and Mali (see Additional file 1: Table S7; Additional file 2: Figure S2). Strong shared negative XP-EHH scores were also present within *acs8*, the uncharacterized *PF3D7\_1421100* and the SP resistance gene *dhps*. Strong positive signals, suggestive of selection in the non-Malawian populations, were present within *msp10*, *trap* and three genes directly downstream of *crt* (*cgl*, *glp3*, *cgl2*).

#### Adaptive copy number selection in the *gch1* promoter

Given the hypothesized role for copy number variation (CNV) in *gch1*-mediated SP resistance [35] and signals

of selection observed, potential CNVs at this locus were investigated. Evidence of a promoter duplication was found (genomic region 973,804–974,240 bp; see Fig. 2) in almost all Malawian samples ( $n = 213$ , 96.8%; see Table 2; Additional file 1: Table S6). Similar promoter duplications were found in samples from Ghana ( $n = 59$ , 29.2%), Guinea ( $n = 43$ , 45.3%), DRC ( $n = 25$ , 44.6%), The Gambia ( $n = 5$ , 9.1%), and Asia ( $n = 5$ , < 0.4%). In contrast, the whole *gch1* gene duplication [8] was predicted in samples from Thailand ( $n = 29$ , 13.8%), Cambodia ( $n = 23$ , 4.4%), Vietnam ( $n = 20$ , 10.7%), Ghana ( $n = 10$ , 5%), Myanmar ( $n = 7$ , 7.4%), and Bangladesh ( $n = 6$ , 11.1%) (see Fig. 2;

**Table 2 Drug resistance allele frequencies**

SNP	Malawi	Other East Africa	DRC	West Africa	South Asia	South East Asia	South America
Sample size	220	33	56	430	54	1133	21
<i>dhps</i>							
S436A	0.005	0.061	0.107	0.505	0.509	0.313	0
A437G	0.998	0.894	0.902	0.735	0.843	0.971	0.286
K540E	0.995	0.894	0.063	0.014	0.778	0.477	0.190
A581G	0.027	0.091	0.027	0.002	0.156	0.407	0.238
<i>dhfr</i>							
N51I	0.991	0.909	0.982	0.658	0.471	0.897	0.381
C59R	0.991	0.939	0.821	0.745	0.972	0.996	0
S108N	1	1	1	0.781	1	0.998	0.952
I164L	0	0	0	0	0.167	0.438	0
Double <i>dhfr</i> mutant <sup>a</sup>	1	1	1	0.728	0.982	0.995	0.381
Triple <i>dhfr</i> mutant <sup>b</sup>	0.977	0.848	0.750	0.563	0.392	0.887	0
Quintuple <i>dhfr</i> Mutant <sup>c</sup>	0	0	0	0	0.167	0.438	0
<i>dhfr-dhps</i>							
Quintuple genotype <sup>d</sup>	0.968	0.788	0.036	0.007	0.278	0.421	0
Quintuple + <i>dhfr</i> I164L	0	0	0	0	0.130	0.266	0
Quintuple + <i>dhps</i> A581G	0.027	0.091	0.027	0	0.111	0.198	0
<i>crt</i>							
K76T	0	0.485	0.661	0.416	0.889	0.960	1.000
Q271E	0	0.485	0.643	0.430	0.907	0.935	0
N326S	0	0	0	0.002	0.241	0.630	0
I356T	0	0	0.196	0.140	0.833	0.652	0
<i>kelch13</i>							
K189T	0.091	0.061	0.196	0.502	0.130	0.007	0.714
K189 N	0.005	0	0	0.026	0	0	0
Y493H	0	0	0	0	0	0.044	0
C580Y	0	0	0	0	0	0.224	0
<i>gch1</i>							
Promoter duplication	0.968	0	0.446	0.251	0.037	0.003	0
Whole gene duplication	0	0	0.020	0.023	0.111	0.076	0

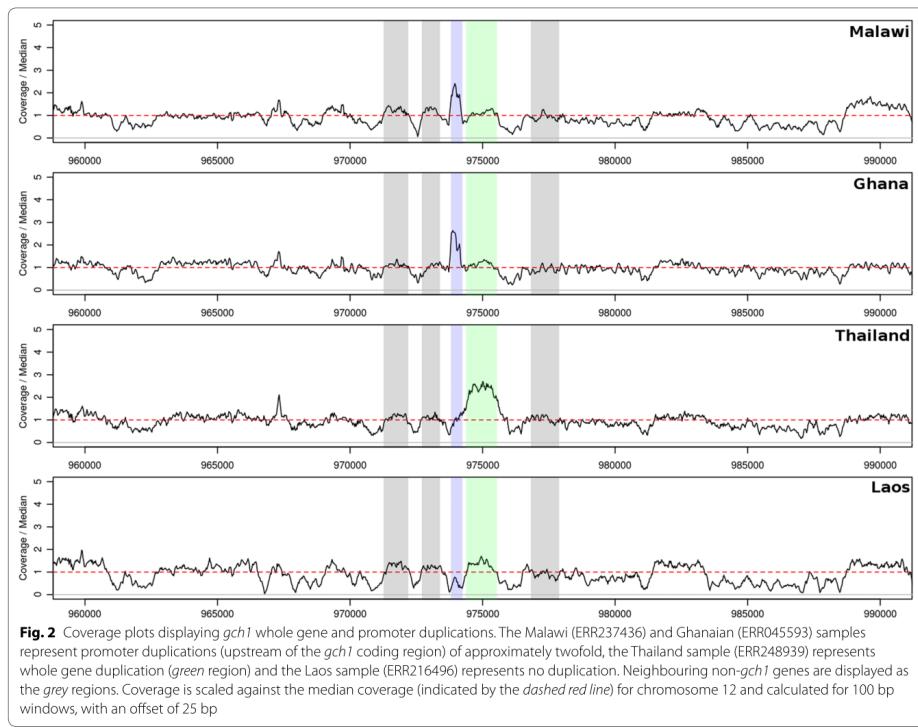
DRC Democratic Republic of Congo

<sup>a</sup> Any two of N51I, C59R or S108N<sup>b</sup> N51I, C59R & S108N<sup>c</sup> Triple *dhfr* mutant haplotype with I164L<sup>d</sup> *dhfr* N51I/C59R/S108N haplotype + *dhps* A437G/K540E haplotype

**Table 2:** Additional file 1: Table S6). The gene duplication appears predominantly on the background of *dhfr* triple mutant haplotype in West Africa (79.2%) and a quadruple mutant haplotype in Southeast Asia (80.3%) and South Asia (50.0%) (see Table 3). The promoter duplication is seemingly not linked to the *dhfr* I164L quadruple mutation and sits predominantly on a triple mutant haplotype background in Malawi (99.1%), DRC (87.9%) and West Africa (85.8%) (see Fig. 3). In these populations, the haplotype background of *dhps* S436A, A437G and K540E appears less important with a number of mutation

combinations present (Table 3). The exception is Malawi where the promoter sits on a predominantly *dhps* A437G/K540E haplotype background, leading to the high frequency of quintuple mutant genotypes.

Coverage analysis of the *P. falciparum* reference strains 3D7, HB3, DD2, 7G8, and GB4 identified no upstream promoter duplication. Previously identified whole gene duplications in 3D7, 7G8, DD2, and GB4, and the absence of duplication in HB3, were confirmed [36] (see Additional file 2: Figure S4). 38 SNPs (eight non-synonymous) were identified across the



*gch1* coding region using the global dataset, all with non-reference allele frequencies (<4%), except a SNP in South America (position 974,633, 28.3%). In the *gch1* promoter region there were 11 SNPs occurring only in African populations, all at low frequency (<4%), except for one polymorphism, which was almost always found on a non-duplication background (position 974,046: East Africa 10.0%, DRC 18.3%, West Africa 18.5%). The overall low levels of nucleotide variation and sequence homogeneity support the argument that *gch1* copy number variants, rather than associated coding SNPs, are targeted by selection. Examination of the EHH revealed differences in linkage disequilibrium (LD) between continents and within Africa (see Fig. 4; Additional file 2: Figure S2). When considering African populations, there is evidence in non-Malawian populations of near symmetrical decays of EHH to a level close to zero within 25 kb. Malawi LD extends much wider and is consistent with a sweep around the promoter duplication (see Fig. 4).

To determine whether the *gch1* promoter duplication was under relative positive selection in the DRC, Ghana and Guinea, a sub-population was applied XP-EHH analysis, where those with the duplication were compared to those without. This approach identified no significant differential sub-population selection in either the DRC, Ghana or Guinea ( $|XP-EHH| < 4$ ), although there are near-significant positive selection signals for the duplication-positive sub-populations in Ghana and Guinea (see Additional file 2: Figure S3). This contrasts with Malawi, where the duplication appears to be under active selection, potentially reflecting differences in SP usage (see Fig. 1; Additional file 2: Figure S2).

## Discussion

Malaria surveillance is crucial to informing and supporting disease prevention, control and elimination strategies. Access to technological advances and their reduced costs mean monitoring approaches involving genomic data are being implemented within malaria programmes

**Table 3 Co-association frequencies of *dhfr* mutations and *gch1* duplications**

<i>dhfr</i> mutants <sup>a</sup>	<i>gch1</i> duplication	<i>dhps</i> mutants <sup>b</sup>	Malawi	Other East Africa	DRC	West Africa	South Asia	Southeast Asia	South America
None	None	None	0	0	0	0.067	0	0.002	0.048
None	None	Single	0	0	0	0.095	0	0	0
Single	None	None	0	0	0	0.002	0.019	0.002	0.524
Single	None	Single	0	0	0	0.040	0	0	0.048
Double	None	None	0	0.030	0.018	0.030	0.019	0.031	0.143
Double	None	Single	0	0	0.071	0.112	0.019	0.028	0
Double	None	Double	0.005	0.121	0	0.002	0.278	0.021	0.143
Double	Promoter	None	0	0	0	0.002	0	0	0
Double	Promoter	Single	0	0	0.054	0.028	0	0	0
Double	Promoter	Double	0.009	0	0	0	0	0	0
Double	Gene	Single	0	0	0	0.005	0	0	0
Double	Gene	Double	0	0	0	0	0.037	0.001	0
Triple	None	None	0	0.061	0.036	0.086	0.019	0.078	0
Triple	None	Single	0	0	0.393	0.328	0.056	0.190	0.048
Triple	None	Double	0.032	0.788	0.036	0.005	0.389	0.185	0.048
Triple	Promoter	None	0	0	0.036	0.012	0	0	0
Triple	Promoter	Single	0	0	0.304	0.160	0	0	0
Triple	Promoter	Double	0.950	0	0.054	0.009	0	0	0
Triple	Gene	Single	0	0	0	0.014	0	0.007	0
Triple	Gene	Double	0	0	0	0	0	0.006	0
Quadruple	None	None	0	0	0	0	0.019	0.004	0
Quadruple	None	Single	0	0	0	0	0.019	0.144	0
Quadruple	None	Double	0	0	0	0	0.093	0.244	0
Quadruple	Gene	Single	0	0	0	0	0	0.014	0
Quadruple	Gene	Double	0	0	0	0	0.037	0.042	0

Frequencies greater than 0.2 are shown in italics

<sup>a</sup> *dhfr* mutant haplotypes: double—any two of *dhfr* N51I, C59R or S108N; triple—*dhfr* N51I, C59R and S108N; quadruple—*dhfr* triple mutant haplotype with I164L; DRC Democratic Republic of Congo

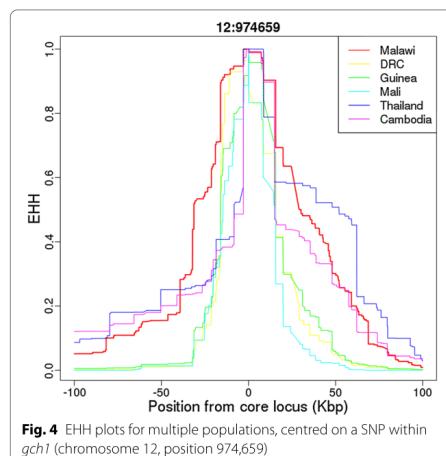
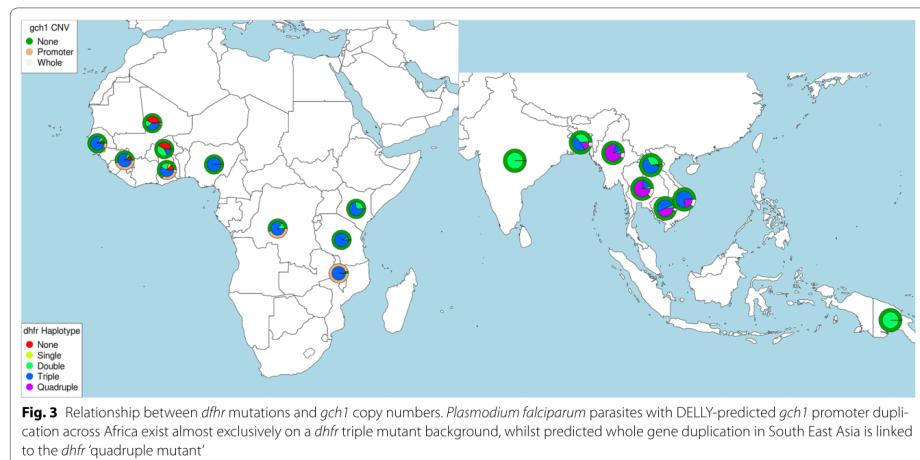
<sup>b</sup> *dhps* double mutant haplotypes consist of *dhps* A437G and K540E

[7, 12, 14]. To inform such programmes, genomic differences were investigated in Malawian *P. falciparum* parasites from two regions across multiple seasons. No major differences were observed between the two regional or temporal sub-populations, potentially due to small sample sizes, a narrow two-year study period, the removal of a minority of SNPs with a high frequency of missing or heterozygous genotypes, and study regions that are only 100 km apart with high human migration between them.

Because of the effects of control measures on *Plasmodium* genomes, signals of selection and differences in allele frequencies were also investigated across Malawi parasites and compared with those from global datasets. Signals in drug resistance-associated genes, and surface-associated proteins that are exposed to the host immune system were detected, including previously described selection in *msp10*, *trap* and *ama1* [7, 23]. Antibodies against the pre-erythrocyte TRAP and erythrocyte stages MSP10 and AMA1 proteins have been detected

in anti-malarial-acquired immunity individuals living in malaria-endemic areas [37, 38]. Vaccines using TRAP, AMA1 and MSP proteins (e.g., MSP1 and MSP3) are being tested with promising results [39]; as such the MSP10 protein may represent a viable vaccine candidate.

Consistent with previous reports, *crt* mutations (K76T, Q271E, N326S, I356T) were absent from Malawi [40, 41]. Neighbouring countries (as well as others in Africa) that did not switch away from chloroquine as early as Malawi continue to report higher levels of *crt* mutations. This observation has led to a setting likened to a Malawian island of chloroquine drug sensitivity in a sea of resistance [41]. Three genes, *cgl1*, *glp3* and *cgl6*, immediately downstream of the *crt* gene, displayed high relative positive selection in multiple non-Malawian populations, consistent with hard selective sweeps [31]. In contrast, SP resistance-associated *dhfr* and *dhps* alleles are retained at high frequencies in Malawi. Compared to other African populations there is a significantly higher level of



both the *dhfr* C59R mutation and the quintuple *dhfr/dhps* mutant genotypes. These frequencies are consistent with previous analyses highlighting the pan-African distribution of the *dhfr* triple mutant haplotypes in contrast to the East–West African differences in distribution of the A437G and K540E *dhps* variants [42, 43]. The SNP-based work is consistent with the results from microsatellite-based studies, considering the dynamics of strong selection for mutations conferring SP resistance, including

support for the observation of the independent origin of sulfadoxine-resistant alleles across a number of regions [42, 44, 45]. Together, these observations reflect the early decline in chloroquine usage and early introduction and prolonged use of SP in Malawi when compared to other African populations.

In almost all Malawian samples (96.8%), a novel duplication was identified in the promoter of the *gch1* gene (973,804–974,240 bp), a member of the folate pathway that includes targets for sulfadoxine and pyrimethamine. This duplication was also detected in other parasites, particularly from West Africa and the DRC, but is near-absent in Asia. Whole gene duplications have previously been detected in Thailand and Cambodia [33] and were detected in this dataset. Positive selection signals are present across the *gch1* region in Malawi, with relative selection in Malawi compared to other African populations where the promoter duplication is also present. Further, samples with the duplication were not found to be under relative selection in the DRC, Ghana or Guinea. Together, the strong evidence of selection for this promoter duplication in Malawi supports its role as being advantageous for *P. falciparum* parasites in regions with high SP use, particularly where there is a higher frequency of resistance-associated *dhfr* and possibly *dhps* variants.

The function of the promoter duplication remains to be established. It is possible that both promoter and whole gene duplications increase *gch1* expression in vivo thereby reducing the fitness cost associated with *dhfr* and *dhps* variants, which convey resistance to SP. *In silico*, functional predictions for the promoter duplication

identified multiple TATA, TATAA and TGTAA PftBp binding motifs [46]. If this duplication acts to reduce the fitness cost of other SP resistance-associated variants, its presence may suggest a more persistent form of resistance which further surveillance will need to confirm. Ongoing fieldwork in Malawi will allow to survey the parasite population during longer periods and to detect genomic changes following the introduction of ACT. Further work should also consider the genomic landscape within other African countries to determine the frequency of the *gch1* duplication and other variants associated with SP resistance.

### Conclusion

This study reports the persistence of genetic variants associated with SP and chloroquine resistance within the *P. falciparum* population in Malawi, despite withdrawal of these anti-malarials from front-line use. Signals of positive selection were also identified, which suggest retention of these resistance-associated variants, as well as various life stage-specific surface antigens. Investigation of *gch1* copy number variation identified the near fixation of a specific 436 bp promoter duplication within Malawi, present in other African countries but absent from Asian populations. It is most likely that this promoter duplication acts in a similar fashion to the whole gene duplication present in Asian populations, although further experimental work is required to elucidate any functional impact.

### Additional files

**Additional file 1.** Additional tables.

**Additional file 2.** Additional figures.

### Authors' contributions

AC, SC, HO, and TGC conceived and designed the study; MM, ATRJ, CJS, CR, NS, JM, KSP, AT, and HO performed laboratory experiments, contributed biological samples, sequencing, epidemiological or phenotypic data; MR, EDB and TGC performed the statistical analysis; DPK led the sequencing effort. MR, SC, HO, and TGC wrote/drafted and finalized the manuscript with contributions from all other authors. All authors read and approved the final manuscript.

### Author details

<sup>1</sup> Faculty of Infectious and Tropical Diseases, London School of Hygiene and Tropical Medicine, London, UK. <sup>2</sup> Department of Physiology, College of Medicine, University of Malawi, Blantyre, Malawi. <sup>3</sup> Centre for Medical Parasitology, University of Copenhagen, Copenhagen, Denmark. <sup>4</sup> Centre for Statistics and Applications of University of Lisbon, Lisbon, Portugal. <sup>5</sup> Wellcome Trust Sanger Institute, Hinxton, UK. <sup>6</sup> Liverpool School of Tropical Medicine, Pembroke Place, Liverpool, UK. <sup>7</sup> Malawi-Liverpool-Wellcome Trust Clinical Research Programme, College of Medicine, University of Malawi, Blantyre, Malawi. <sup>8</sup> School of Public Health and Family Medicine, College of Medicine, University of Malawi, Blantyre, Malawi. <sup>9</sup> Faculty of Epidemiology and Population Health, London School of Hygiene and Tropical Medicine, London, UK.

### Acknowledgements

We thank the parents, guardians and children who participated in this study, and the technical, clinical and nursing staff for assistance. The Medical Research Council UK funded eMedLab computing resource was used for data analysis.

### Competing interests

The authors declare that they have no competing interests.

### Availability of data and materials

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.malariaigen.net/projects/pf3k>).

### Ethics approval and consent to participate

Written informed consent was obtained from a parent or guardian of each child with the ethics committees of the University of Malawi's College of Medicine and the Liverpool School of Tropical Medicine both approving the study.

### Funding

MR is funded by the Biotechnology and Biological Sciences Research Council (Grant Number B/J014567/1). This work was supported by the Malaria Capacity Development Consortium (to HO), which is funded by The Wellcome Trust (Grant Number WT084289MA). JM was supported by a Wellcome Trust fellowship (Grant Number 080964). TGC is supported by the Medical Research Council UK (GRANT No. MR/K000551/1, MR/M01360X/1, MR/N010469/1, MC\_PC\_15103). SC is funded by the Medical Research Council UK (GRANT No. MR/M01360X/1, MC\_PC\_15103). The Wellcome Trust provides core support to the Sanger Institute (Grant Number 077012/Z/05/Z, 098051), the Resource Centre for Genomic Epidemiology of Malaria (Grant Number 090770/Z/09/Z), and The Malawi-Wellcome Trust Programme.

Received: 29 September 2016 Accepted: 23 November 2016

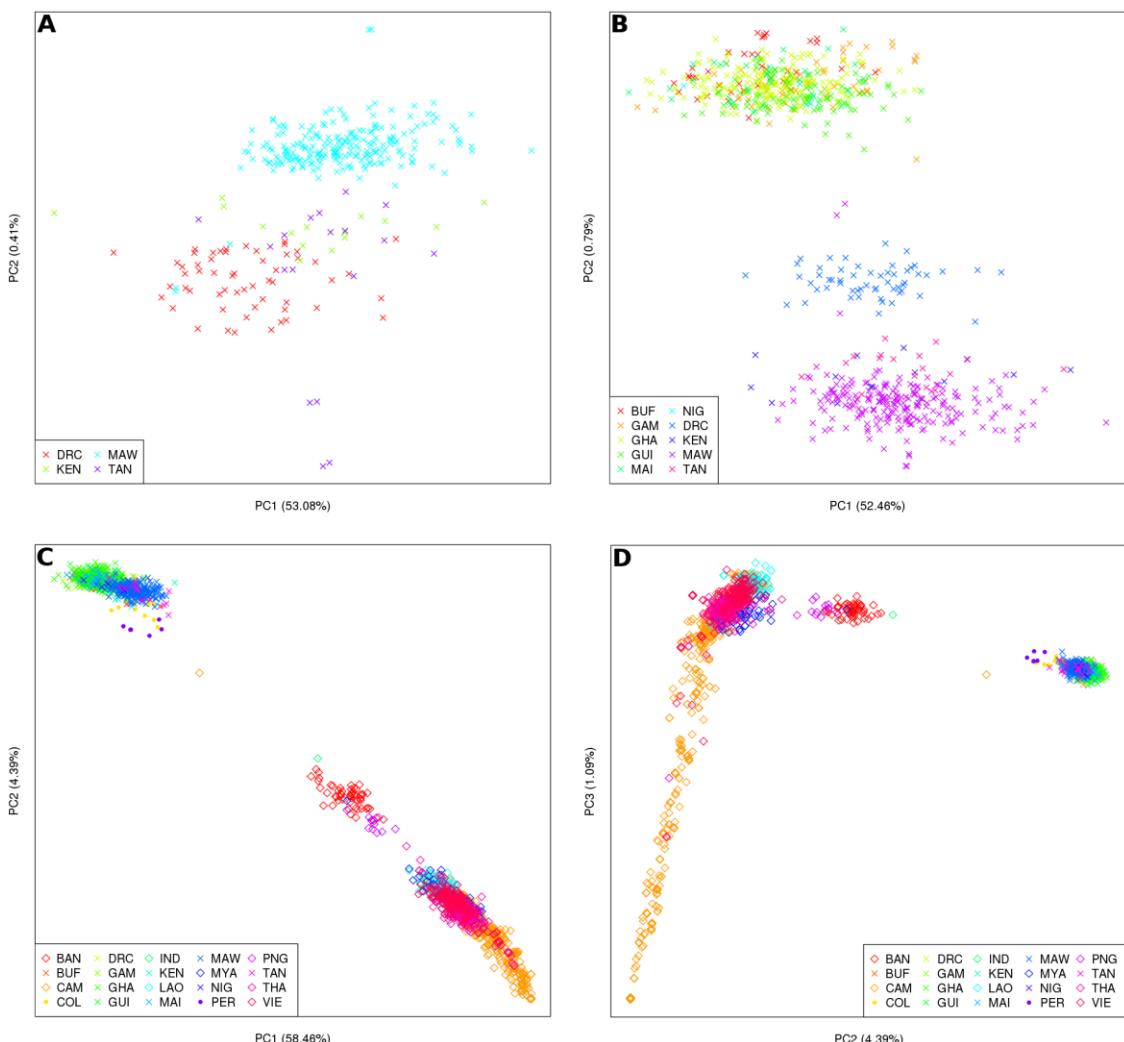
Published online: 29 November 2016

### References

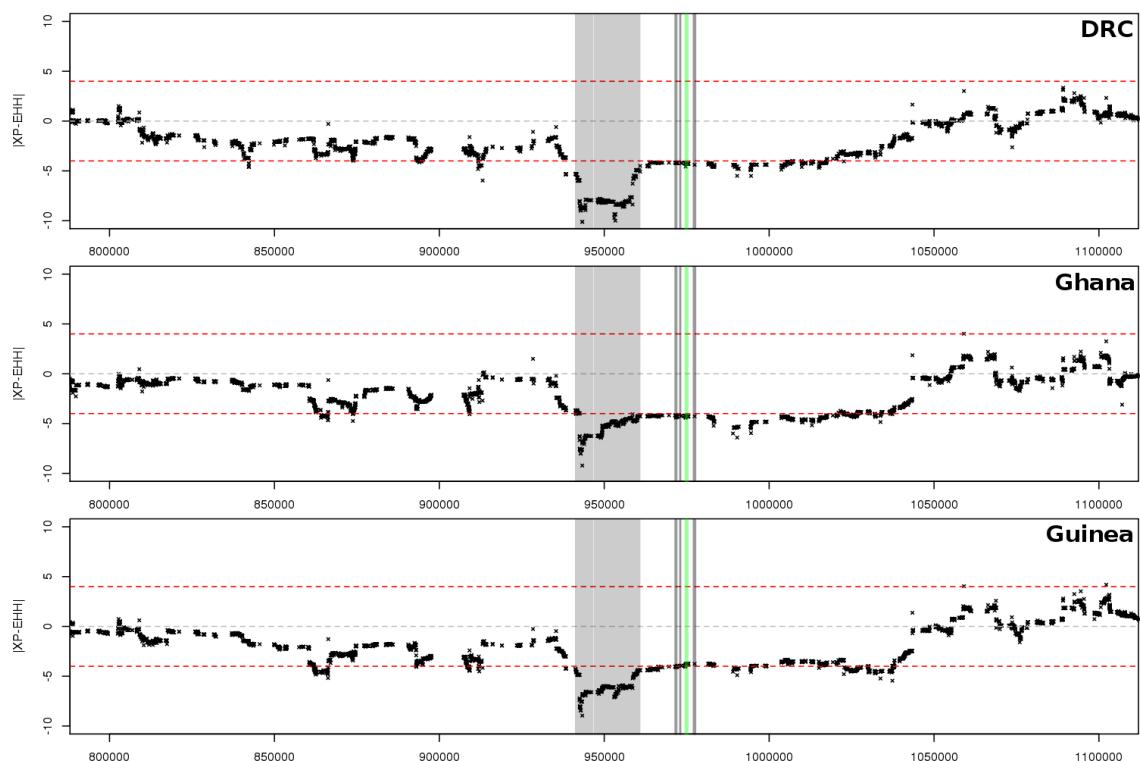
1. The demographic and health surveys program. Malawi malaria indicator survey. 2014. <https://dhsprogram.com/pubs/pdf/MIS18/MIS18.pdf>.
2. President's malaria initiative. The President's malaria initiative A decade of progress. 2016. <https://www.pmi.gov/docs/default-source/default-document-library/pmi-reports/pmi-tenth-annual-report-congress.pdf>.
3. WHO. World Health Statistics 2015. Geneva: World Health Organization; 2015.
4. Flegg JA, Metcalf CJ, Gharbi M, Venkatesan M, Shewchuk T, Hopkins Sibley C, et al. Trends in antimalarial drug use in Africa. Am J Trop Med Hyg. 2013;89:857–65.
5. Kublin JG, Cortese JF, Njunju EM, Mukadam RA, Wirima JJ, Kazembe PN, et al. Reemergence of chloroquine-sensitive *Plasmodium falciparum* malaria after cessation of chloroquine use in Malawi. J Infect Dis. 2003;187:1870–5.
6. Artimovich E, Schneider K, Taylor TE, Kublin JG, Dzinjalama FK, Escalante AA, et al. Persistence of sulfadoxine-pyrimethamine resistance despite reduction of drug pressure in Malawi. J Infect Dis. 2015;212:694–701.
7. Ocholla H, Preston MD, Mpanda M, Jensen AT, Campino S, MacInnis B, et al. Whole-genome scans provide evidence of adaptive evolution in Malawian *Plasmodium falciparum* isolates. J Infect Dis. 2014;20:1991–2000.
8. Nair S, Miller B, Barends M, Jaidee A, Patel J, Mayxay M, et al. Adaptive copy number evolution in malaria parasites. PLoS Genet. 2008;4:e1000243.
9. Auburn S, Marfurt J, Maslen G, Campino S, Ruano Rubio V, Manske M, et al. Effective preparation of *Plasmodium vivax* field isolates for high-throughput whole genome sequencing. PLoS ONE. 2013;8:e53160.
10. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25:1754–60.

11. Samad H, Coll F, Preston MD, Ocholla H, Fairhurst RM, Clark TG. Imputation-based population genetics analysis of *Plasmodium falciparum* malaria parasites. *PLoS Genet.* 2015;11:e1005131.
12. Preston MD, Campino S, Assefa SA, Echeverry DF, Ocholla H, Amambua-Ngwa A, et al. A barcode of organellar genome polymorphisms identifies the geographic origin of *Plasmodium falciparum* strains. *Nat Commun.* 2014;5:4052.
13. Auburn S, Campino S, Miotto O, Djimde AA, Zongo I, Manske M, et al. Characterization of within-host *Plasmodium falciparum* diversity using next-generation sequence data. *PLoS ONE.* 2012;7:e32891.
14. Neher RA, Amato R, Miotto O, Woodrow CJ, Almagro-Garcia J, Sinha I, Campino S, Mead D, Drury E, Kekre M, Sanders M. Genomic epidemiology of artemisinin resistant malaria. *eLife.* 2016;5:e08714.
15. Nei M. Analysis of gene diversity in subdivided populations. *Proc Natl Acad Sci USA.* 1973;70:3321–3.
16. Tajima F. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics.* 1989;113:585–95.
17. Rausch T, Zichner T, Schlattl A, Stütz AM, Benes V, Korbel JO. DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics.* 2012;28:i333–9.
18. Szpiech ZA, Hernandez RD. selscan: an efficient multithreaded program to perform EHH-based scans for positive selection. *Mol Biol Evol.* 2014;31:2824–7.
19. Pasternak ND, Dzikowski R. PfEMP1: an antigen that plays a key role in the pathogenicity and immune evasion of the malaria parasite *Plasmodium falciparum*. *Int J Biochem Cell Biol.* 2009;41:1463–6.
20. Hyde JE. Exploring the folate pathway in *Plasmodium falciparum*. *Acta Trop.* 2005;94:191–206.
21. Triglia T, Healer J, Caruana SR, Hodder AN, Anders RF, Crabb BS, et al. Apical membrane antigen 1 plays a central role in erythrocyte invasion by *Plasmodium* species. *Mol Microbiol.* 2000;38:706–18.
22. Mu J, Myers RA, Jiang H, Liu S, Ricklefs S, Waisberg M, et al. *Plasmodium falciparum* genome-wide scans for positive selection, recombination hot spots and resistance to antimalarial drugs. *Nat Genet.* 2010;42:268–71.
23. Amambua-Ngwa A, Park DJ, Volkman SK, Barnes KG, Bei AK, Lukens AK, et al. SNP genotyping identifies new signatures of selection in a deep sample of West African *Plasmodium falciparum* malaria parasites. *Mol Biol Evol.* 2012;29:3249–53.
24. Li F, Bounkeua V, Pettersen K, Vinetz JM. *Plasmodium falciparum* ookinete expression of plasmeipin VII and plasmeipin X. *Malar J.* 2016;15:11.
25. Srinivasan P, Yasgar A, Luci DK, Beatty WL, Hu X, Andersen J, et al. Disrupting malaria parasite AMA1–RON2 interaction with a small molecule prevents erythrocyte invasion. *Nat Commun.* 2013;4:2261.
26. Mphande FA, Ribacke U, Kaneko O, Kironde F, Winter G, Wahlgren M. SURFIN4.1, a schizont-merozoite associated protein in the SURFIN family of *Plasmodium falciparum*. *Malar J.* 2008;7:16.
27. Borrman S, Strainer J, Mwai L, Abdi A, Rippert A, Okombo J, et al. Genome-wide screen identifies new candidate genes associated with artemisinin susceptibility in *Plasmodium falciparum* in Kenya. *Sci Rep.* 2013;3:3318.
28. Gunasekera AM, Wickramarachchi T, Neafsey DE, Ganguli I, Perera L, Premaratne PH, et al. Genetic diversity and selection at the *Plasmodium vivax* Apical Membrane Antigen-1 (PvAMA-1) locus in a Sri Lankan population. *Mol Biol Evol.* 2007;24:939–47.
29. Manske M, Miotto O, Campino S, Auburn S, Almagro-Garcia J, Maslen G, et al. Analysis of *Plasmodium falciparum* diversity in natural infections by deep sequencing. *Nature.* 2012;487:375–9.
30. van Dijk MR, van Schaijk BCL, Khan SM, van Dooren MW, Ramesar J, Kaczanowski S, et al. Three members of the 6-cys protein family of Plasmodium play a role in gamete fertility. *PLoS Pathog.* 2010;6:e1000853.
31. Wootton JC, Feng X, Ferdig MT, Cooper RA, Mu J, Baruch DI, et al. Genetic diversity and chloroquine selective sweeps in *Plasmodium falciparum*. *Nature.* 2002;418:320–3.
32. Gutman J, Kalliani L, Taylor S, Zhou Z, Wiegand RE, Thwai KL, et al. The A581G mutation in the gene encoding *Plasmodium falciparum* dihydrofumarate synthetase reduces the effectiveness of sulfadoxine-pyrimethamine preventive therapy in Malawian pregnant women. *J Infect Dis.* 2015;211:1997–2005.
33. Ménard D, Khim N, Beghain J, Adegnika AA, Shafiq-Alam M, Amodu O, et al. A worldwide map of *Plasmodium falciparum* K13-propeller polymorphisms. *N Engl J Med.* 2016;374:2453–64.
34. Conrad MD, Bigira V, Kapisi J, Muhindo M, Kamya MR, Havlir DV, et al. Polymorphisms in K13 and falcipain-2 associated with artemisinin resistance are not prevalent in *Plasmodium falciparum* isolated from Ugandan children. *PLoS ONE.* 2014;9:e105690.
35. Heinberg A, Siu E, Stern C, Lawrence EA, Ferdig MT, Deitsch KW, et al. Direct evidence for the adaptive role of copy number variation on antifolate susceptibility in *Plasmodium falciparum*. *Mol Microbiol.* 2013;88:702–12.
36. Sepúlveda N, Campino SG, Assefa SA, Sutherland CJ, Pain A, Clark TG, et al. Poisson hierarchical modelling approach to detecting copy number variation in sequence coverage data. *BMC Genom.* 2013;14:128.
37. Maskus DJ, Bethke S, Seidel M, Kapelski S, Addai-Mensah O, Boes A, et al. Isolation, production and characterization of fully human monoclonal antibodies directed to *Plasmodium falciparum* MSP10. *Malar J.* 2015;14:267.
38. Noland GS, Hendel-Paterson B, Min XM, Moormann AM, Vulule JM, Narum DL, et al. Low prevalence of antibodies to pre-erythrocytic but not blood-stage *Plasmodium falciparum* antigens in an area of unstable malaria transmission compared to prevalence in an area of stable malaria transmission. *Infect Immun.* 2008;76:5721–8.
39. Schwartz L, Brown GV, Genton B, Moorthy VS. A review of malaria vaccine clinical field projects based on the WHO rainbow table. *Malar J.* 2012;11:11.
40. Laufer MK, Takala-Harrison S, Dzinjalama FK, Stine OC, Taylor TE, Plowe CV. Return of chloroquine-susceptible falciparum malaria in Malawi was a re-expansion of diverse susceptible parasites. *J Infect Dis.* 2010;202:801–8.
41. Bridges DJ, Molneyux M, Nkhamo S. Low level genotypic chloroquine resistance near Malawi's northern border with Tanzania. *Trop Med Int Health.* 2009;14:1093–6.
42. Pearce RJ, Pota H, Evehe MS, Ba EH, Mombo-Ngoma G, Malisa AL, et al. Multiple origins and regional dispersal of resistance dhps in African *Plasmodium falciparum* malaria. *PLoS Med.* 2009;6:e1000055.
43. Naidoo I, Roper C. Mapping 'partially resistant', 'fully resistant' and 'super resistant' malaria. *Trends Parasitol.* 2013;29:505–15.
44. McCollum AM, Schneider KA, Griffing SM, Zhou Z, Kariuki S, ter Kuile F, et al. Differences in selective pressure on *dhps* and *dhfr* drug resistant mutations in western Kenya. *Malar J.* 2012;11:77.
45. Vinayak S, Alam MT, Mixson-Hayden T, McCollum AM, Sem R, Shah NK, et al. Origin and evolution of sulfadoxine resistant *Plasmodium falciparum*. *PLoS Pathog.* 2010;6:e1000830.
46. Ruvalcaba-Salazarra OK, del Carmen Ramirez-Estudilloa M, Montiel-Condadoa D, Recillas-Targat F, Vargas M, Hernández-Rivas R. Recombinant and native *Plasmodium falciparum* TATA-binding-protein binds to a specific TATA box element in promoter regions. *Mol Biochem Parasitol.* 2005;140:183–96.

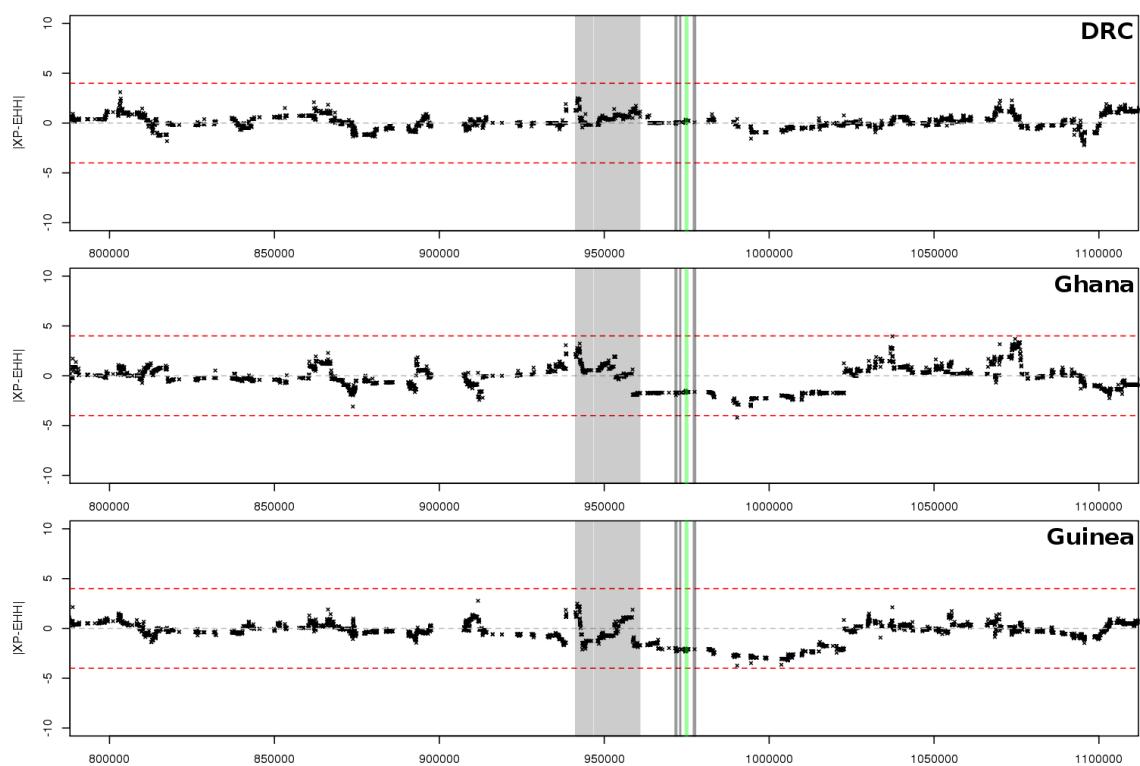
**Additional Figure 1: Clustering through Principal Components Analysis of the combined dataset.** Analysis is based on a pairwise Manhattan distance SNP matrix for all samples; **(A)** East Africa; **(B)** Africa; **(C,D)** Global. Colours indicate countries with point styles indicating continents. West Africa (n=430): Burkina Faso (BUF, 39), Gambia (GAM, 55), Ghana (GHA, 202), Guinea (GUI, 95), Mali (MAI, 35), Nigeria (NIG, 4); Central and East Africa (n=253): Dem. Rep. of Congo (DRC, 56), Kenya (KEN, 15), Malawi (MAW, 220), Tanzania (TAN, 18); South and South-East Asia (n=1187): Bangladesh (BAN, 54), Cambodia (CAM, 526), Laos (LAO, 104), Myanmar (MYA, 95), Papua New Guinea (PNG, 11), Thailand (THA, 210), Vietnam (VIE, 187); South America (n=21): Colombia (COL, 14), Peru (PER, 7).



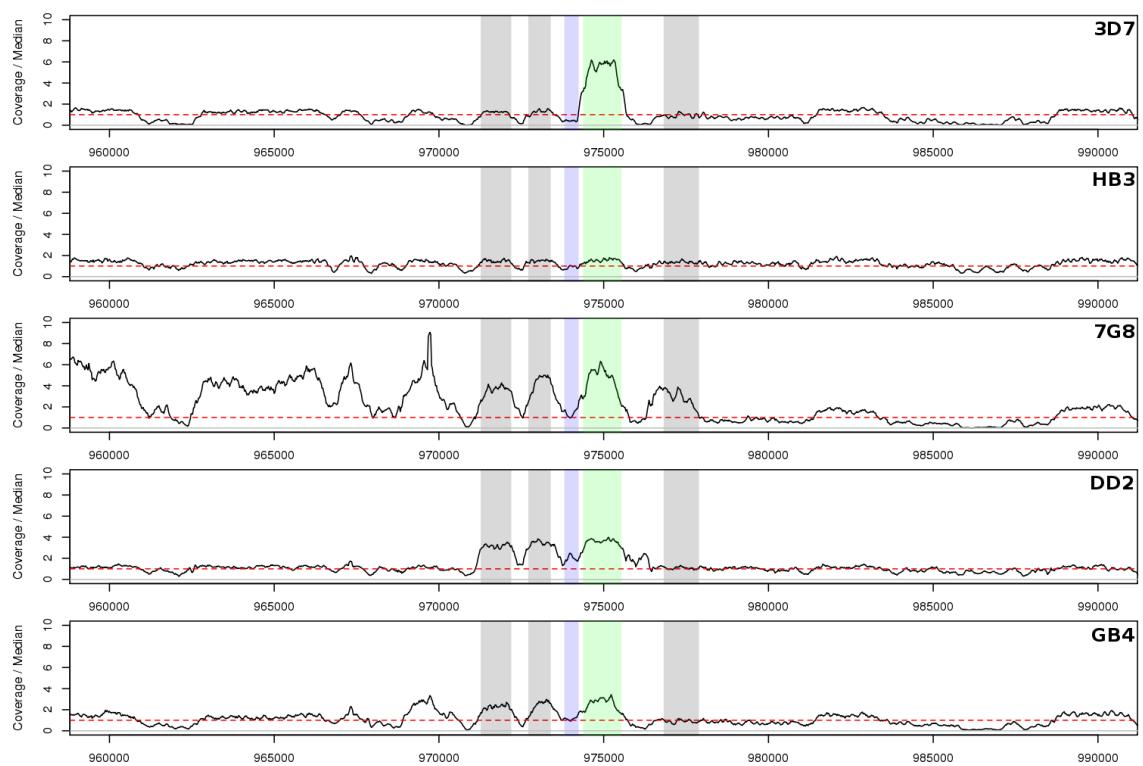
**Additional Figure 2: XP-EHH around *gch1* comparing the Malawi population with those from the DRC, Ghana and Guinea.** Negative values indicate relative fixation in Malawi, whilst positive values indicate relative fixation in the non-Malawian population. Green indicates *gch1* whilst the immediate grey bars indicate the same neighbouring genes as in Figure 2. In addition, the two upstream genes *Pf3D7\_1223400* and *Pf3D7\_1223500* with the most significant XP-EHH are highlighted in grey. Red lines indicate our significance threshold of absolute XP-EHH scores greater than 4.



**Additional Figure 3: XP-EHH selection around *gch1* between duplication-positive and duplication-negative DRC, Ghana and Guinea populations.** Negative values indicate relative fixation in the duplication-positive population, whilst positive values indicate relative fixation in the duplication-negative population. Green indicates *gch1* whilst the immediate grey bars indicate the same neighbouring genes as in Figure 2. In addition, the two upstream genes *Pf3D7\_1223400* and *Pf3D7\_1223500* with the most significant XP-EHH are highlighted in grey. Red lines indicate our significance threshold of absolute XP-EHH scores greater than 4.



**Additional Figure 4: Coverage for the *gch1* region in five reference strains.** Whilst the previously identified *gch1* duplication is present within 3D7 (ERS009999), 7G8 (ERS016318), DD2 (ERS010540) and GB4 (ERS016319) and confirmed absent in HB3 (ERS010539), no reference sample was shown to contain the novel promoter duplication. Green indicates *gch1*, blue indicates the region for the most frequent promoter duplication, and grey indicates neighbouring genes. Coverage is scaled against the median coverage (indicated by the dashed red line) for chromosome 12 and calculated for 100bp windows, with an offset of 25bp.



**Additional Table 1:** Pairwise differences between Malawi seasonal sub-populations.

	<b>Chikwawa 2011 (Wet) (n=22)</b>	<b>Chikwawa 2011 (Dry) (n=64)</b>	<b>Chikwawa 2012 (Wet) (n=34)</b>	<b>Chikwawa 2012 (Dry) (n=57)</b>	<b>Zomba 2012 (Wet) (n=31)</b>
<b>Chikwawa 2011 (Wet) (n=22)</b>	9458 (7607 - 10400)	9434 (6130-11020)	9458 (6229-10660)	9538 (5951-10640)	9522 (6131-10740)
<b>Chikwawa 2011 (Dry) (n=64)</b>	0.01 (0.00-0.19)	9381 (4560 - 11190)	9395 (4484-11140)	9474 (4270-11100)	9442 (4681-11140)
<b>Chikwawa 2012 (Wet) (n=34)</b>	0.01 (0.00-0.20)	0.01 0.00-0.12	9396 (900-10710)	9485 (4440-10810)	9438 (4724-10770)
<b>Chikwawa 2012 (Dry) (n=57)</b>	0.01 (0.00-0.19)	0.01 (0.00-0.12)	0.01 (0.00-0.12)	9557 (4758 - 10700)	9536 (4651-10880)
<b>Zomba 2012 (Wet) (n=31)</b>	0.01 (0.00-0.22)	0.01 (0.00-0.19)	0.01 (0.00-0.15)	0.01 (0.00-0.14)	9448 (522 - 10680)

Upper and diagonal values indicate median (range) number of SNP differences between individuals. Lower values indicate median (range)  $F_{ST}$  values across SNPs for each sub-population pair.

**Additional Table 2:** Top Fst values for pairwise Malawi sub-population comparisons.

Locus	Gene	Max Fst	Other Fst
<i>PF3D7_1038400</i>	Gametocyte specific protein	0.220	0.001 – 0.202
<i>PF3D7_1008500</i>	Conserved unknown	0.199	0.000 – 0.187
<i>PF3D7_0420000</i>	Putative zinc finger	0.198	0.001 – 0.152
8:1309330	Intergenic	0.191	0.001 – 0.097
7:1345508	Intergenic	0.188	0.000 – 0.128
<i>PF3D7_0707300</i>	Rhoptry-associated membrane antigen	0.180	0.002 – 0.118
<i>PF3D7_1326400</i>	Putative translation initiation factor eIF-2B	0.178	0.000 – 0.126
4:1061702	Intergenic	0.178	0.000 – 0.172
<i>PF3D7_0710000</i>	Conserved unknown	0.174	0.000 – 0.050
<i>PF3D7_0525000</i>	Putative zinc finger	0.173	0.009 – 0.118
<i>PF3D7_1446500</i>	Conserved unknown	0.173	0.001 – 0.134
<i>PF3D7_1477600</i>	SURFIN 14.1	0.172	0.000 – 0.117
4:545378	Intergenic	0.170	0.001 – 0.071
<i>PF3D7_0708400</i>	Heat Shock Protein 90	0.166	0.001 – 0.142
<i>PF3D7_0709300</i>	Putative Cg2 protein	0.165	0.005 – 0.084
6:824765	Intergenic	0.164	0.001 – 0.132
8:698268	Intergenic	0.162	0.001 – 0.083
<i>PF3D7_1200700</i>	Acyl-CoA Synthetase	0.158	0.000 – 0.121
<i>PF3D7_1135600</i>	Putative condensin-2 complex subunit D3	0.158	0.000 – 0.053
9:91336	Intergenic	0.157	0.000 – 0.144
<i>PF3D7_0905400</i>	High molecular weight rhoptry protein 3	0.157	0.002 – 0.078
<i>PF3D7_0412400</i>	PfEMP1	0.157	0.000 – 0.076
<i>PF3D7_0310200</i>	Putative phd finger protein	0.156	0.001 – 0.132
<i>PF3D7_1325400</i>	Conserved unknown	0.152	0.001 – 0.105
<i>PF3D7_0629300</i>	Putative phospholipase	0.151	0.001 – 0.128

Threshold of greater than 0.15.

**Additional Table 3:** Allele frequencies for Sulfadoxine-Pyrimethamine resistance mutations within Malawi across all seasons.

Genetic variant	Chikwawa 2011 Wet	Chikwawa 2011 Dry	Chikwawa 2012 Wet	Chikwawa 2012 Dry	Zomba 2012 Wet	Combined population
Sample Size	22	64	38	57	27	220
<i>dhpS</i>						
S436A	0	0	0.026	0	0	0.005
A437G	1	1	0.987	1	1	0.998
K540E	1	0.992	0.987	1	1	0.995
<i>dhfr*</i>						
N51I	1	1	1	0.983	1	0.991
C59R	0.955	0.984	1	0.991	1	0.991
S108N	1	1	1	1	1	1
<i>Triple mutant**</i>	0.955	0.984	1	0.965	1	0.977
<i>gch1***</i>						
Promoter duplication	1	0.938	0.974	0.965	1	0.968

\*I164L (quadruple mutation) is not present; \*\* *dhfr* N51I, C59R & S108N haplotype; \*\*\* Whole gene duplication of *gch1* is absent.

**Additional Table 4:** Top hits for Malawi-only positive selection (iHS) analysis.

Gene ID	Position	iHS	Gene
<i>PF3D7_0208600</i>	355154	-4.376	<i>RRF1</i>
<i>PF3D7_0417400</i>	758290, 758269	-4.678, -4.448	Conserved unknown ( <b>near<sup>1</sup> dhfr</b> )
<i>PF3D7_0505100</i>	226777	-4.502	<i>TRS85</i>
<i>PF3D7_0511400</i>	481921	4.516	Conserved unknown
<i>PF3D7_0808200</i>	417751	4.686	<i>Plasmeprin X</i>
<i>PF3D7_0809600</i>	484954, 490762	4.906, 3.962	Putative petidase family C50 ( <b>near<sup>2</sup> dhps</b> )
<i>PF3D7_0814600</i>	703454	4.519	Conserved unknown
<i>PF3D7_0826000</i>	1111727	5.272	Conserved unknown
<i>PF3D7_1002200</i>	1115470, 1115617	-5.604, -5.396	<i>PArt/TrpA-3</i>
<i>PF3D7_1133400</i>	1294082, 1294982	5.129, 5.017	<i>amaI</i>
<i>PF3D7_1223400</i>	943339	-4.947	Phospholipid-transporting ATPase ( <b>near<sup>3</sup> gchI</b> )
<i>PF3D7_1335900</i>	1466252, 1466264	5.366, 5.209	<i>trap</i>
<i>PF3D7_1352900</i>	2114996	6.066	Exported unknown

Positive scores indicate selection for the alternative core allele, whilst negative scores indicate selection for the reference core allele. All hits are above a threshold of 4 (  $|iHS|>4$  ).

---

<sup>1</sup> Within 10 kbp

<sup>2</sup> Within 50 kbp

<sup>3</sup> Within 5 kbp

**Additional Table 5:** Top Tajima's D values for the combined Malawi population.

Gene ID	Gene	Tajima's D
<i>PF3D7_0710200</i>	Conserved unknown	3.224
<i>PF3D7_0830800</i>	<i>surf8.2</i>	3.196
<i>PF3D7_1133400</i>	<i>ama1</i>	2.916
<i>PF3D7_0424400</i>	<i>surf4.2</i>	1.605
<i>PF3D7_1335900</i>	<i>trap</i>	1.503
<i>PF3D7_0113800</i>	DBL-containing protein	1.475
<i>PF3D7_1475900</i>	Conserved unknown	1.205
<i>PF3D7_1004800</i>	Putative ADP/ATP carrier protein	1.080
<i>PF3D7_1035700</i>	Duffy binding-like merozoite surface protein	0.970

**Additional Table 6:** Drug resistance candidate mutation frequencies

Gene/SNP																				
Sample Size	220	18	15	56	39	55	202	95	35	4	54	527	95	11	104	210	187	14	7	
<i>dhps</i>																				
S436A	0.005	0.056	0.067	—	0.107	0.538	0.100	0.609	0.458	0.657	0.250	0.509	0.342	0.332	0	0.212	0.250	0.366	0	0
A437G	0.998	0.944	0.833	0.902	0.615	0.764	0.705	0.705	0.257	1.000	0.84	0.895	1.000	0.227	0.572	1.000	0.799	0.071	0.714	
K540E	0.995	0.944	0.833	0.062	0	0	0.010	0.042	0	0	0.778	0.341	0.9	0.136	0.178	0.907	0.348	0	0.571	
<i>dhfr</i>																				
N51I	0.991	0.944	0.867	0.982	0.359	0.918	0.592	0.821	0.486	1.000	0.471	0.925	0.901	0	0.644	0.938	0.963	0.214	0.714	
C59R	0.991	1.000	0.867	0.821	0.423	0.845	0.757	0.879	0.486	1.000	0.972	0.996	1.000	0.955	0.976	1.000	1.000	0	0	
S108N	1.000	1.000	1.000	1.000	0.397	0.936	0.817	0.879	0.471	1.000	1.000	1.000	1.000	1.000	0.976	1.000	1.000	0.929	1.000	
I164L	0	0	0.052	0	0	0	0	0	0	0	0.382	0.432	0.861	0	0.008	0.798	0.231	0	0.200	
N51I+C59R	0.977	0.944	0.733	0.750	0.179	0.836	0.490	0.789	0.400	1.000	0.370	0.918	0.905	0	0.606	0.933	0.952	0	0	
N51I+S108N	0.991	0.944	0.867	0.982	0.154	0.909	0.525	0.779	0.371	1.000	0.389	0.916	0.905	0	0.606	0.933	0.947	0.214	0.714	
C59R+S108N	0.986	1.000	0.867	0.768	0.205	0.836	0.713	0.853	0.371	1.000	0.944	0.996	1.000	0.909	0.971	1.000	0.995	0	0	
Triple Mutant*	0.977	0.944	0.733	0.750	0.154	0.836	0.490	0.779	0.371	1.000	0.370	0.916	0.905	0	0.606	0.933	0.947	0	0	
Quadruple Mutant**	0	0	0	0	0	0	0	0	0	0	0.167	0.422	0.811	0	0.010	0.752	0.203	0	0	
<i>crt</i>																				
K76T	0	0.722	0.200	0.661	0.205	0.727	0.223	0.674	0.514	1.000	0.889	0.973	0.989	0.909	0.885	0.981	0.925	1.000	1.000	
Q271E	0	0.722	0.200	0.643	0.205	0.727	0.233	0.663	0.657	1.000	0.907	0.941	0.979	0	0.885	0.986	0.914	0	0	
N326S	0	0	0	0	0	0.018	0	0	0	0	0.241	0.647	0.989	0	0.115	0.952	0.358	0	0	
I356T	0	0	0	0.196	0.026	0.636	0.015	0.126	0.229	0.250	0.833	0.672	0.989	0	0.115	0.990	0.380	0	0	
<i>kelch13</i>																				
K189T	0.091	0.056	0.067	0.196	0.615	0.545	0.530	0.411	0.400	0.500	0.130	0	0.011	0	0	0	0	0.857	0.429	
K189N	0.005	0	0	0	0.026	0.091	0.020	0	0.029	0	0	0	0	0	0	0	0	0	0	
Y493H	0	0	0	0	0	0	0	0	0	0	0	0.087	0	0	0	0	0	0.021	0	0
C580Y	0	0	0	0	0	0	0	0	0	0	0	0.387	0.105	0	0	0	0.138	0.059	0	0
<i>gch1</i>																				
No duplication	0.032	1	1	0.534	1	0.909	0.658	0.547	0.971	1	0.852	0.954	0.926	1	0.990	0.852	0.893	1	1	
Promoter duplication	0.968	0	0	0.446	0	0.091	0.292	0.453	0.011	0	0.037	0.002	0	0	0	0.010	0	0	0	
Whole gene duplication	0	0	0	0.020	0	0	0.050	0	0	0	0.111	0.044	0.074	0	0.010	0.138	0.107	0	0	

\*dhfr N51I, C59R & S108N haplotype. \*\* dhfr N51I, C59R, S108N & I164L haplotype; DRC Democratic Republic of Congo

**Additional Table 7:** Top hits for Malawi pairwise positive selection (XP-EHH) analysis.

Gene ID	Populations	XP-EHH**	Gene
<i>PF3D7_0212500</i>	DRC	6.511	Conserved unknown
<i>PF3D7_0215300</i>	DRC, Ghana, Guinea	-6.087, -6.912, -6.872	<i>acs8</i> ; Acyl-CoA synthetase
<i>PF3D7_0307900</i>	DRC	-6.866	Conserved unknown
<i>PF3D7_0321800</i>	Ghana	6.802	WD repeat-containing protein
<i>PF3D7_0416900</i>	Mali	-6.257	Conserved unknown ( <i>near dhfr</i> )
<i>PF3D7_0417400</i>	Colombia, Ghana	-6.184, -6.323	Conserved unknown ( <i>near dhfr</i> )
<i>PF3D7_0513200</i>	Laos	6.037	Conserved unknown
<i>PF3D7_0525100</i>	Ghana, Guinea	-6.337, -6.514	<i>acs10</i> ; Acyl-CoA synthetase
<i>PF3D7_0526600</i>	Laos, Mali	7.018, -6.181	Conserved unknown
<i>PF3D7_0529000</i>	Bangladesh, Laos, Myanmar	7.162, 6.947, 6.245	Conserved unknown
<i>PF3D7_0620400</i>	DRC, West Africa, Myanmar	5.981, 11.079, 6.125	<i>Msp10</i>
<i>PF3D7_0629700</i>	DRC	7.241	<i>Set1</i>
<i>PF3D7_0709100</i>	Bangladesh, DRC, Gambia, Southeast Asia	7.529, 9.330, 7.759, 8.663	<i>Cg1</i> protein ( <i>near crt</i> )
<i>PF3D7_0709200</i>	Cambodia, Myanmar, Thailand	6.327, 6.259, 6.095	<i>GLP3</i> (Cg6 protein) ( <i>near crt</i> )
<i>PF3D7_0709300</i>	Cambodia, DRC, Thailand, Vietnam	6.759, 8.073, 7.229, 6.470	<i>cg2</i> ( <i>near crt</i> )
<i>PF3D7_0709600</i>	DRC, Gambia	8.498, 6.520	<i>pop1</i>
<i>PF3D7_0710000</i>	Peru	6.019	Conserved unknown
<i>PF3D7_0810200</i>	Gambia	-6.132	<i>ABCK1</i>
<i>PF3D7_0810600</i>	Guinea	-6.477	ATP-dependent RNA helicase DBP1
<i>PF3D7_0810800</i>	Colombia, Guinea, Laos, Vietnam	-6.257, -6.532, -6.257, -6.125	<i>dhp5</i>
<i>PF3D7_0810900</i>	Colombia	-6.257	Conserved unknown ( <i>near dhp5</i> )
<i>PF3D7_0926500</i>	Bangladesh	-6.266	Conserved unknown
<i>PF3D7_1223400</i>	DRC, West Africa, Kenya	-10.114, -9.217, -7.852	<i>near gch1</i>
<i>PF3D7_1223500</i>	DRC, Gambia, Ghana, Guinea, Kenya	-9.998, -6.640, -6.434, -7.130, -6.431	<i>near gch1</i>
<i>PF3D7_1218300</i>	Ghana	6.294	<i>ap2mu</i>
<i>PF3D7_1227500</i>	DRC	-6.231	<i>cyc2</i>
<i>PF3D7_1335800</i>	DRC	6.679	Conserved unknown
<i>PF3D7_1352900</i>	Cambodia, Colombia	6.357, 6.498	Exported unknown, fam-f protein
<i>PF3D7_1324300</i>	Gambia, Ghana, Guinea	6.295, 6.722, 7.758	Conserved unknown membrane
<i>PF3D7_1335900</i>	Colombia, DRC, Mali, Peru, Tanzania, Vietnam	6.682, 6.178, -6.171, 6.662, 6.088, 8.256	<i>trap</i>
<i>PF3D7_1421100</i>	West Africa	-9.650	Conserved unknown

Positive (negative) scores indicate relative selection in the non-Malawi (Malawi) population. Bold indicates genes with known associations with drug resistance. DRC Democratic Republic of Congo.

## **Chapter 3:**

A global analysis of copy number variation in Plasmodium falciparum identifies a novel duplication of the chloroquine resistance associated gene

## **RESEARCH PAPER COVER SHEET**

***PLEASE NOTE THAT A COVER SHEET MUST BE COMPLETED FOR EACH RESEARCH PAPER INCLUDED IN A THESIS.***

### **SECTION A – Student Details**

<b>Student</b>	Matt Ravenhall
<b>Principal Supervisor</b>	Taane Clark
<b>Thesis Title</b>	A bioinformatic analysis of malaria host and pathogen genomics

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

### **SECTION B – Paper already published**

Where was the work published?	n/a		
When was the work published?	n/a		
If the work was published prior to registration for your research degree, give a brief rationale for its inclusion	n/a		
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:	Matt Ravenhall, Ernest Diez Benavente, Colin J. Sutherland, David A. Baker, Susana Campino, 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)	Working with existing BAM files for the primary dataset (and assemblies for PacBio), I conducted the analyses, created all figures, and wrote the manuscript under the supervision of Susana Campino and Taane Clark. I also developed the key analysis tool for this project.
--	--

**Student Signature:** \_\_\_\_\_

**Date:** \_\_\_\_\_

**Supervisor Signature:** \_\_\_\_\_

**Date:** \_\_\_\_\_

**Improving health worldwide**

[www.lshtm.ac.uk](http://www.lshtm.ac.uk)

**A global analysis of copy number variation in *Plasmodium falciparum* identifies a novel duplication of the chloroquine resistance associated gene**

Matt Ravenhall<sup>1</sup>, Ernest Diez Benavente<sup>1</sup>, Colin J. Sutherland<sup>2</sup>, David A. Baker<sup>1</sup>, Susana Campino<sup>1,\*</sup>, Taane G. Clark<sup>1,3,\*</sup>

1. Department of Pathogen Molecular Biology, London School of Hygiene and Tropical Medicine, London WC1E 7HT, UK
2. Department of Immunology and Infection, London School of Hygiene and Tropical Medicine, London WC1E 7HT, UK
3. Department of Infectious Disease Epidemiology, Faculty of Epidemiology and Population Health, London School of Hygiene and Tropical Medicine, London WC1E 7HT, UK

\*Joint last authors

## **Abstract**

The evolution of genetic mechanisms for host immune evasion and anti-malarial resistance has enabled the *Plasmodium falciparum* malaria parasite to inflict high morbidity and mortality on human populations. Most studies of *P. falciparum* genetic diversity have focused on single-nucleotide polymorphisms (SNPs), enabling the identification of drug resistance-associated loci such as the chloroquine related *crt* and sulfadoxine-pyrimethamine related *dhfr*. Whilst larger structural variants are known to impact adaptation, for example, *mdr1* duplications with anti-malarial resistance, no large-scale, genome-wide study on clinical isolates has been undertaken using whole genome sequencing data. By applying a structural variant detection pipeline across whole genome sequence data from 3,110 clinical isolates in 21 malaria-endemic countries, we identified >70,000 specific deletions and >600 duplications. The majority of structural variants are rare (48.5% of deletions and 94.7% of duplications are found in single isolates) with 2.4% of deletions and 0.2% of duplications found in >5% of global samples. A subset of variants was present at high frequency in drug-resistance related genes including *mdr1*, the *gch1* promoter region, and a putative novel duplication of *crt*. Regional-specific variants were identified and a companion visualisation tool has been developed to assist web-based investigation of these polymorphisms by the wider scientific community.

## Introduction

*Plasmodium falciparum* malaria imposes a heavy morbidity and mortality burden, with an estimated 216 million new cases and 446,000 deaths in 2016 alone, with ~90% of the burden in sub-Saharan Africa<sup>1</sup>. An understanding of the genomic diversity of *P. falciparum* parasites could provide insights into novel phenotypes that impact responses to antimalarials and other control measures, as well as host-pathogen interactions. Single nucleotide polymorphism (SNP) based analyses have revealed insights into drug resistance, molecular barcodes for continental origin<sup>2</sup>, transmission dynamics<sup>3</sup>, multiplicity of infection<sup>4</sup> and regions under selective pressure related to immunological and anti-malarial treatment pressure<sup>5</sup>. In comparison, investigations of structural variants (SVs), such as insertions, deletions and duplications, have been sparse. This is despite SVs making an important contribution to genomic diversity and comprising many nucleotides of heterogeneity. In particular, copy number variants (CNVs; large indels and duplications) are thought to be widespread in the *P. falciparum* genome<sup>6</sup> and more abundant than SNPs<sup>7</sup>.

Malaria parasites are exposed to strong selection from the human immune response and treatment with antimalarial drugs. Subsequently, CNVs have often been found in association with specific *P. falciparum* phenotypes, such as drug resistance. Duplications of *mdr1* have been shown to underlie a multi-drug resistance phenotype, with these variants now present at high population frequencies in Southeast Asia<sup>8</sup>, with copy number altering parasite response to multiple anti-malarial drugs<sup>9</sup>. Recently, we identified a novel promoter duplication for *gch1* at near-fixation in a Malawi population<sup>5</sup>, which is distinct from the whole gene duplication observed in Southeast Asia known to contribute to sulfadoxine-pyrimethamine (SP) resistance. In general, such regional genetic variation may arise from differences in drug regimens, mosquito vectors, and host immunity, but is poorly understood.

Given their importance, a genome-wide structural variant map for *P. falciparum* with country and regional resolution should provide insights with which to better understand the impacts of treatment regimes, assess changes in parasite diversity, and ultimately inform the roll-out of anti-malarial drugs and other control initiatives. The advent of microarray technologies, such as

genomic hybridisation arrays (CGH), has improved methods for detecting and confirming known SVs<sup>10,11</sup>. However, studies of this type have typically featured modest sample sizes and focused on the exome of lab-adapted isolates. The largest array-based study in *P. falciparum* clinical samples (n=122) identified 134 high-confidence CNVs across the parasite exome, established they were more common in South American than African or Southeast Asian populations, and identified several loci including *mdr1*, *rh2b*, and histidine-rich proteins II and III to be under positive selection<sup>11</sup>. Recently whole genome sequencing platforms, which produce a greater depth of short or long reads, have been used to detect SVs in *P. falciparum* strains<sup>7,12</sup>, and have potentially finer resolution than array-based approaches. Coupled with bioinformatic advances in detection algorithms, there is now capacity to accurately characterise a broader range of SV types. For example, extremes in coverage can identify duplications and deletions, split sequences and alternative *de novo* assembly-based approaches can detect a number of other types, including inversions and large insertions and deletions<sup>13</sup>.

By analysing whole genome sequencing data from 3,110 clinical isolates and focusing on robust genomic regions (82.6%) of the AT-rich *P. falciparum* genome, we present the first comprehensive genomic map of SVs within the global *P. falciparum* population, with a particular focus on CNVs. We identify a total of 70,257 deletions (mean 403 per sample, median size 26 bp) and 601 duplications (mean 0.39 per sample, median size 1,478 bp), contrasting with an average of 24,495 SNPs and 33,479 small indels (<15 bp) per sample. Several variants were found to be geographically specific and highlight novel structural variants with roles in antigenic variation, drug resistance, and host-pathogen interactions. PacBio sequencing data of *P. falciparum* strains was used to calibrate our bioinformatic pipeline and confirm specific candidates.

## Results

### ***Distribution of variant type, size and location***

The 3,110 samples represented 21 countries across Africa (Central, East, West), Asia (South, Southeast) and South America (Supplementary Table 1), and all displayed little evidence of

multiplicity of infection and non-anomalous coverage (see Methods). Using a bioinformatics pipeline based upon DELLY<sup>13</sup>, we identified more than 1 million putative variants relative to the 3D7 reference genome, which after filtering (see Methods) was reduced to a total of 70,858 high quality variants (70,257 deletions: mean 402.88 per sample, median size 26 bp; 601 duplications: mean 0.39 per sample, median 1,478 bp). Of these, only 1,251 high quality specific large deletions (>500 bp; mean 1.39 per sample, median size 670 bp) and 385 duplications (>500 bp; mean 0.13 per sample, median size 1,478 bp) were detected (Figure 1). The majority of the duplications (94.7%) and half the deletions (34,065 deletions; 48.6%) were unique to single isolates (total: 34,634; 34,065 deletions and 569 duplications). Both deletions and duplications tend to occur within intergenic regions (intergenic/genic ratio: deletions 1.42, duplications 2.15). A (1 kbp) window-based analysis was used to identify regions with overlapping but distinct SVs. For deletions, 24,947 (of 27,388) windows (78.1% genic) were represented, compared to 2,441 windows (80.4% genic) for duplications.

### ***Frequently occurring specific variants***

Previous work has shown that SVs present in a relatively high frequency of the global population are consistent with evidence of phenotypic selection<sup>11</sup>, and therefore we investigated variants identified in at least 5% of the global population (n=3,110). In total we identify 1,676 frequent variants of which only one is a duplication, this being the previously identified 436 bp *gch1* promoter region duplication<sup>5</sup>. Of all frequent variants, 723 (43.1%) are genic and the median length is 27 bp (Supplementary Table 2). Only five variants are greater than 500 bp in length, including four intergenic deletions (size range: 605 to 1,023 bp), and one 553 bp deletion in liver stage antigen 3 (*PF3D7\_0220000*). Interestingly the 553 bp deletion in *PF3D7\_0220000* is present primarily in Southeast Asia, particularly Thailand (14.6%), Laos (11.5%), and Myanmar (11.2%) (Global 5.5%; Africa 0.1%, America 0.0%, Asia 10.2%), and may represent region-specific host-directed selection. Two of the intergenic variants show strong evidence of continental differences (Allele frequency difference: F<sub>ST</sub> score >0.2), including a 1,015 bp deletion in chromosome 9 upstream of *gexp22* (*PF3D7\_0935500*) (F<sub>ST</sub>: 0.227; Africa 13.2%,

America 70.8%, Asia 41.7%) and a 605 bp deletion in chromosome 12 upstream of *ap2mu* (*PF3D7\_1218300*) ( $F_{ST}$  0.249; 0.2% Africa, 0.0% America, 33.6% Asia).

### ***Exploration of structural variation in anti-malarial resistance candidates***

Given that structural variation can have a significant impact on gene expression and anti-malarial resistance, we focused our analysis on the identification of novel structural variants in candidate genes (*dhfr*, *dhps*, *kelch13*, *mdr1*, *gch1*, *crt*). Because it is possible for heterogeneous duplications (i.e. differences in genetic copies) to display as mixed phasing calls, we extended our dataset to include those duplications previously excluded for highly heterozygous phasing, leading to a dataset that included 102,483 putative variants. To minimise the number of false positives, we manually verified the genomic data for all candidate regions. Overall, no high-quality SVs were identified in SP resistance associated *dhfr* (*PF3D7\_0417200*) or *dhps* (*PF3D7\_0810800*) genes, or artemisinin resistance associated *kelch13* (*PF3D7\_1343700*). We identify 115 specific duplication types containing *mdr1* in 189 samples, primarily in Southeast Asia (Global 12.9%; Cambodia 9.5%, Myanmar 11.9%, Papua New Guinea 3.8%, Thailand 29.0%, Vietnam 5.9%), and near absent in Africa (Global 0.10%; Ghana 0.2%) (Figure 2), consistent with previous reports<sup>14</sup>. Similarly, tandem duplications are also present in KE01 (Kenya) and KH01 (Cambodia) within our complementary PacBio-based dataset (n=13).

The whole gene duplication of *gch1* (*PF3D7\_1224000*) and a recently identified 436 bp *gch1* promoter duplication may be linked to SP resistance<sup>5</sup>. We identify 307 samples with 135 distinct forms of whole gene duplication across *gch1* (9.9% of the total dataset) (Figure 3). Similar whole gene tandem duplications were present in PacBio samples for 7G8 and KH02, as a triplication in GB4, and as a triplication with an inverted middle copy in Dd2, this being consistent with the existing literature<sup>7</sup>. In contrast, 491 high quality samples are positive for the previously identified 436bp specific ‘promoter region’ duplication (14.0% of total). We confirm this duplication being present at near-fixation in Malawi (89.5%), frequent in the rest of East Africa (Tanzania 78.5%, Kenya 31.6%), maintained in West Africa (Gambia 6.1%, Ghana 4.3%, Guinea 22.2%) and Central Africa (Democratic Republic of Congo 26.3%), but absent from all Asian and American

samples (Regional  $F_{ST}$  0.554). No such duplication was found in any PacBio sample (n=13), though none of these are from Malawi. These data therefore support the *gchI* promoter duplication being present at notable frequency across Africa, and the need for further functional characterisation of any potential role in SP resistance.

Finally, evidence for a 22.9 kbp duplication of *crt* (*PF3D7\_0709000*) is present and consistent across 32 samples isolated in West Africa (4.3%), specifically sub-populations isolated in Burkina Faso (14 samples, 29.2%), Ghana (15 Samples, 3.4%), Guinea (1 sample, 0.85%) and Mali (1 sample, 1.8%). Those 32 samples correspond to 26 specific variants, the consensus of which suggests that the duplication is most likely around 22,893 bp in length, and therefore includes several genes (*PF3D7\_0708900* (*sco1*), *PF3D7\_0709000* (*crt*), *PF3D7\_0709050* (small nucleolar RNA), *PF3D7\_0709100* (*cgl1*), *PF3D7\_0709200* (*glp3*) and *PF3D7\_0709300* (*cgl2*)) (Figure 4). Three additional samples (2 from Ghana, 1 from Burkina Faso) also display a similar 28.7 kbp duplication, which also includes *PF3D7\_0708800* (heat shock protein 110). No *crt* duplication was present in our secondary PacBio dataset (n=13). To explore the specific variability of *crt* in each sample, we calculated the abundance of resistance-associated haplotypes directly from raw reads, finding that both the CVMNK (chloroquine susceptible) and CVIET (chloroquine resistant) haplotypes were present in all 21 duplication-positive samples in 1:1 ratios, with all other global samples generally featuring only one haplotype. Specific read counts suggested carriage of one chloroquine susceptible and one resistant form. It is unclear, without additional transcriptional analysis, whether these forms are expressed independently though we hypothesise that the presence of both forms may allow individual parasites to benefit from the resistance form whilst reducing associated fitness costs. If so, heterogeneous duplication of this sort may represent a more evolutionarily resilient form of *crt*-associated resistance.

### ***Population-Specific Variants***

Regional differences (across West Africa, Central Africa, East Africa, South Asia, Southeast Asia, South America) in SV frequencies were quantified with  $F_{ST}$  analysis for all high-quality variants (median (range): deletions 0.002 (0 - 0.613); duplications 0.001 (0 - 0.554)). A total of 153 high

quality variants (152 deletions and one duplication) have strong regional differences ( $F_{ST} > 0.2$ ), including: (i) an Asia-specific 59 bp deletion within the hypothetical protein *PF3D7\_0312900* ( $F_{ST}$  0.613, 69.8% South Asia, 70.9% Southeast Asia, 0.0% Rest of the World), (ii) a 40 bp South America-specific deletion in the putative histone deacetylase *PF3D7\_1472200* ( $F_{ST}$  0.497, 54.2% South America, 0.0% Rest of the World), and (iii) the 436 bp *gch1* promoter region duplication ( $F_{ST}$  0.554, 78.0% East Africa, 17.2% Central Africa, 6.8% West Africa, 0.0% Rest of the World) (Supplementary Table 3).

Extending our analysis to include variants within the full dataset, we identify a subset which display significant population specificity whilst also being supported by manual inspection of coverage depth and split read support. Non-drug resistance candidates (described earlier), include a 169 bp deletion within the rhoptry-associated membrane antigen *PF3D7\_0707300* ( $F_{ST}$  0.354; 46.0% Africa, 0.6% Asia, 0.0% America), and a 370 bp deletion in the ring-infected erythrocyte surface antigen *PF3D7\_0102200* ( $F_{ST}$  0.213; Africa 40.3%, Asia 4.2%, America 4.2%) with elevation in West and Central Africa ( $F_{ST}$  0.321; West Africa 66.1%, Central Africa 36.1%, East Africa 4.1%, South Asia 50.9%, Southeast Asia 2.5%, South America 4.2%). We also identify a near Africa-specific 586 bp deletion within the C-terminal of reticulocyte binding protein 2 homologue b (*rh2b*, *PF3D7\_1335300*) ( $F_{ST}$ : 0.334; 58.4% Africa, 12.5% America, 1.3% Asia) and a 29 bp deletion in *rhopH2* (*PF3D7\_0929400*) ( $F_{ST}$  0.288; 73.7% Africa, 100% America, 40.8% Asia), knockdown of which has been shown to inhibit parasite growth within host erythrocytes<sup>15</sup>.

### ***Companion visualisation tool***

To facilitate exploration of the full dataset produced by our analysis pipeline by the wider scientific community, we developed a companion visualisation tool contextualising these SVs within multiple populations (Supplementary Figure 1). This tool and its associated documentation are publicly available at <https://pathogenseq.lshtm.ac.uk/PfGlobalSV.html>.

## **Discussion**

This large and geographically comprehensive study of SVs in *P. falciparum* characterises both known and novel variants, the latter occurring in loci associated with antimalarial resistance, host-pathogen interactions, and disease severity. Deletions represent the bulk of SVs (>99%) identified, primarily due to an abundance of shorter forms (median 26 bp) in comparison to duplications (median 713 bp). We find that 48.6% of high quality deletions and 65.0% of high quality duplications were found in single samples, in line with previous work with smaller sample sizes including the most recent which found that approximately half of structural variants were only present in one of 16 samples<sup>10</sup>. Previous studies have often overlooked the role of smaller structural variants, defining and applying a minimum size of 500 bp. Our results demonstrate that a significant number (97.7%) of high quality variants are present in the 15 to 500 bp size range, indicating that previous studies have under-estimated the full range of genomic variants within the *P. falciparum* genome. This finding that most SVs are under 500bp in size is consistent with previous studies in various species<sup>10,16</sup>.

Population-specific SVs suggest evidence of localised selective pressure<sup>11</sup>. These include the drug resistance associated *mdr1* and *gch1* genes, and a striking novel 22.9 kbp duplication of the chloroquine resistance associated gene *crt*, for which samples are positive for both the CVMNK (chloroquine susceptible) and CVIET (chloroquine resistant) forms of the gene across multiple independent West African sub-populations. It is unclear whether dual-carriage of these variants would allow expression of both or either forms of the *crt* transporter, though it is likely that this could allow individual parasites to benefit from chloroquine resistance with a reduced fitness cost. This finding may also be similar to previously identified alternatively spliced forms of *crt* in eastern Sudan which were hypothesised to facilitate ‘switching’ between chloroquine resistant and susceptible isoforms<sup>17</sup>. Further short ~29 bp and ~430 bp deletions identified here at low frequency in *crt* may reflect the specific deletions identified in that same study. Follow up studies, particularly with culture-adapted clinical isolates in which this duplication is present, are required to properly characterise *in vitro* phenotypes. We also present further characterisation of the promoter region duplication for the SP resistance associated gene *gch1*, previously identified in Malawi<sup>5</sup>. This additional analysis confirms the duplication is at near-fixation in Malawi, and

highlights its presence across Central and East Africa, including notably high frequencies in Tanzania, Kenya, Guinea and the DRC. Further this genetic region has been shown to be under positive selection in Malawi using SNP-based metrics<sup>5</sup>.

Our results demonstrate that application of our pipeline can enhance the speed and capacity of high throughput structural variant discovery. However, this is not without limitations, especially as we rely upon validity of the underlying mapping, for which some regions (such as those which are highly variable or repetitive) are known to be difficult to characterise. In an attempt to resolve this issue, we excluded known highly variable regions from our analysis, such as *var*, *rifin* and *stevor* genes and subtelomeric regions. However, in doing so we prevent discovery of true variants within these regions, including duplications in AT-rich loci<sup>7</sup>. In addition, all identified variants were identified relative to the 3D7 reference strain, consistent with the approach taken in other studies<sup>10,11</sup>. Given that 3D7 is most likely an African strain, SVs within African samples may be artificially under-represented due to those variants also being present within 3D7. Further, the discovery stage of our pipeline inherits the limitations of those tools, such as an inability to infer high quality inversions. This risk was limited by prioritising those variants that were identified by DELLY with support from an alternative discovery software (CNVNator or Control-FREEC).

The approach taken in this study, as with standard SNP discovery, requires single-genotype samples, preventing investigation of more complex isolates. By pre-screening for single clone infections and filtering on rates of predicted phasing, we were able to reduce false positive calls but also removed several highly likely variants that presented with a high prevalence of predicted heterozygous calls and potentially underestimate the total number of duplications. Notable candidate variants excluded from our highest quality dataset but supported by manual inspection of coverage depth or similar variants within the existing scientific literature include 102 putative deletions within the glycophorin A binding, invasion-critical gene *EBA175* (*PF3D7\_0731500*)<sup>18</sup>, the most prominent being a 424 bp deletion in 1,492 samples (30.5% of samples). We also identify a 586 bp deletion elevated in Africa (58.4% Africa, 12.5% America, 1.3% Asia) within *Rh2b*, a gene that plays a key role in erythrocyte invasion<sup>19</sup>, for which similar deletions have previously

been identified (and validated here) in the T996 *P. falciparum* line<sup>20</sup> and samples from Senegal, where it is possibly associated with the utilisation of neuraminidase-sensitive invasion pathways<sup>21</sup>. Another is a 29 bp deletion in *rhopH2* (*PF3D7\_0929400*), with a reduced prevalence in Asian populations, knockdowns of which have been shown to inhibit parasite growth within host erythrocytes<sup>15</sup>.

Our final count of 70,858 high quality specific variants assumes that each SV is distinct by their specific base-pair location. This means that we identify variants which arose from similar evolutionary events but may place insufficient emphasis on variants with a shared phenotypic impact. Previous studies collapsed analysis to a locus level, but risk overlooking complex structural variation within the same gene. This challenge was partially resolved via our secondary windows-based approach, whereby variants are grouped due to their presence within a 1 kbp window. Overall, our work presents a set of high quality structural variants, some population specific, which are likely to have functional consequences for drug resistance and erythrocyte invasion. An extended list of further structural variation requires both technological advances, such as low cost long read platforms with low error rates, as well as computational and algorithmic advances that assemble genomes to high accuracy and require less hands-on filtering. To facilitate further exploration of our full set of global structural variation by the wider scientific community we have developed a visualisation and analysis tool. This resource will assist much-needed genomic investigations into *P. falciparum*, potentially leading to biological insights for the development of disease control measures.

## Methods

### *Sequence data*

Illumina raw sequence data from more than 3,500 samples in the Pf3k project were downloaded from the European Nucleotide Archive (see the project website, <https://www.malariagen.net/projects/pf3k>). The raw sequences were aligned to the *P. falciparum* 3D7 genome using bwa-mem software, (settings: -c 100 -T 50)<sup>22</sup>, with a mean coverage of 70.7x

(Genic: 91.1-fold, Intergenic: 41.8-fold). SNPs and small indels (<15 bp) were called using SAMtools/BCFtools<sup>23</sup> (default settings) and GATK software<sup>24</sup> (settings: "-T UnifiedGenotyper -ploidy 1 -glm BOTH -allowPotentiallyMisencodedQuals 2"). The overlapping set of variants from the two algorithms was retained for further analysis. Samples bearing abnormal coverage less than 20-fold or greater than 300-fold coverage were excluded to reduce false positive rates. Similarly, samples with complex infections were excluded. Multiplicity of infection (MOI) was assessed by estimating mixed SNP call abundance (cut-off >20% genotypes) and using estMOI software<sup>4</sup> (MOI>1). After quality control, our dataset included 3,110 samples representing West Africa (n=738), Central Africa (n=360), East Africa (n=474), South Asia (n=53), Southeast Asia (n=1,461), and South America (n=24). The Illumina data were supplemented by PacBio sequences (ERP009847) from 13 laboratory strains<sup>25</sup>, including 7G8 (Brazil), IT (Brazil), HB3 (Honduras), GA01 (Gabon), GN01 (Guinea), GB4 (Ghana), SN01 (Senegal), CD01 (Congo), KE01 (Kenya), SD01 (Sudan), Dd2 (Indochina), KH01 (Cambodia), and KH02 (Cambodia). These whole genomes were used to validate any putative deletions and duplications detected in discovery pipeline applied to the 3,110 samples. Manual verification of candidate duplications and deletions was facilitated by examination of per-base coverage plots and read pair alignments.

### ***Structural variant discovery***

Structural variants were predicted from short read alignments against the latest 3D7 reference assembly using DELLY (v0.7.3), which has been found to be robust across a range of organisms<sup>13</sup>. Variants longer than 100,000 base pairs were excluded as a conservative filter for erroneous calls, whilst variants identified in a subtelomeric, or highly variable loci such as *var*, *rifin* and *stevor* genes were removed due to established difficulties in accurately mapping these regions. Further, as 7G8 and GB4 have both Illumina and PacBio data, they were used to identify additional highly variable regions for exclusion, as well as assess pipeline parameters. In particular, structural variant discovery and analysis was performed on both the simulated and true Illumina read alignments for each clone, where the paired reads were aligned against their corresponding PacBio references, and structural variant discovery was performed with DELLY. Pipeline parameters were optimised for maximum concordance, and genomic regions (predominantly AT-

rich) with high numbers of false positives excluded. Population-wide filters were also applied to exclude those variants with a median DELLY quality scores below 0.9, missingness >10%, absence of paired read support, or homozygous reference calls frequency >10%. Variants were also removed if they displayed a heterozygous phasing frequency greater than 30%, as these suggest cryptic mixed samples not identified at the SNP calling stage. Regional hotspots were identified using sums of samples with variants for 1 kbp sliding windows with a 500 bp step size. Deletions and duplications were also identified using CNVNator (v0.3.2; bin size of 400 bp)<sup>26</sup> and by Control-FREEC version 11.0 (window size 100 bp, window step 50 bp, ploidy of 1)<sup>27</sup>. Concordance statistics were derived between DELLY and the alternative methods and used to further filter variants that did not have support from at least two methods.

### ***Analysis of Population Statistics***

Multi-population  $F_{ST}$  statistics were calculated between continent (Africa, Asia, South America) and region-based sub-populations (West Africa, Central Africa, East Africa, South Asia, Southeast Asia, South America) for both windows and variants using Nei's method<sup>28</sup>.

### ***Calculation of crt haplotype abundance***

To determine the variability of *crt* in duplication positive samples, we conducted strict match read counts with high quality pre-alignment reads for five specific haplotypes. Haplotype sequences were 25 base pairs long, and included CVMVK (TGTATGTGTAATGAATAAAATTTT), CVIET (TGTATGTGTAATTGAAACAATTTT), CVIDT (TGTATGTGTAATTGATACAATTTT), CVMET (TGTATGTGTAATGGAAACAATTTT), and CVMNT (TGTATGTGTAATGAATACAATTTT).

### **Acknowledgements**

The Medical Research Council UK funded eMedLab computing resource was used for data analysis.

## **Funding**

MR is funded by the Biotechnology and Biological Sciences Research Council (Grant Number BB/J014567/1). TGC received funding from the MRC UK (Grant no. MR/K000551/1, MR/M01360X/1, MR/N010469/1, MR/R020973/1) and BBSRC UK (BB/R013063/1). SC received funding from the Medical Research Council UK grants (MR/R020973/1) and the BBSRC UK (BB/R013063/1).

## **Conflicts of Interest**

The authors declare no conflicting interests.

## **Data Availability**

For the primary short read data set, public accession numbers for the 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>). Raw PacBio sequence data is available from the European Nucleotide Archive (ERP009847).

## References

1. World Health Organization. *World Malaria Report 2016*. WHO Press (WHO Press, 2016).
2. Preston, M. D. *et al.* A barcode of organellar genome polymorphisms identifies the geographic origin of *Plasmodium falciparum* strains. *Nat. Commun.* **5**, (2014).
3. Ankrum, A. & Hall, B. G. Population Dynamics of *Staphylococcus aureus* in Cystic Fibrosis Patients To Determine Transmission Events by Use of Whole-Genome Sequencing. *J. Clin. Microbiol.* **55**, 2143–2152 (2017).
4. Assefa, S. A. *et al.* estMOI: estimating multiplicity of infection using parasite deep sequencing data. *Bioinformatics* **30**, 1292–1294 (2014).
5. Ravenhall, M. *et al.* Characterizing the impact of sustained sulfadoxine/pyrimethamine use upon the *Plasmodium falciparum* population in Malawi. *Malar. J.* **15**, (2016).
6. Ribacke, U. *et al.* Genome wide gene amplifications and deletions in *Plasmodium falciparum*. *Mol. Biochem. Parasitol.* **155**, 33–44 (2007).
7. Miles, A. *et al.* Indels, structural variation, and recombination drive genomic diversity in *Plasmodium falciparum*. *Genome Res.* **26**, 1288–99 (2016).
8. Price, R. N. *et al.* Mefloquine resistance in *Plasmodium falciparum* and increased pfmdr1 gene copy number. *Lancet* **364**, 438–447 (2004).
9. Sidhu, A. B. S., Valderramos, S. G. & Fidock, D. A. pfmdr1 mutations contribute to quinine resistance and enhance mefloquine and artemisinin sensitivity in *Plasmodium falciparum*. *Mol. Microbiol.* **57**, 913–926 (2005).
10. Cheeseman, I. H. *et al.* Gene copy number variation throughout the *Plasmodium falciparum* genome. *BMC Genomics* **10**, 353 (2009).
11. Cheeseman, I. H. *et al.* Population Structure Shapes Copy Number Variation in Malaria Parasites. *Mol. Biol. Evol.* **33**, msv282- (2015).
12. Sepúlveda, N. *et al.* A Poisson hierarchical modelling approach to detecting copy number variation in sequence coverage data. *BMC Genomics* **14**, 128 (2013).
13. Rausch, T. *et al.* DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics* **28**, i333–i339 (2012).
14. Amato, R. *et al.* Genetic markers associated with dihydroartemisinin-piperaquine failure in *Plasmodium falciparum* malaria in Cambodia: a genotype-phenotype association study. *Lancet. Infect. Dis.* **17**, 164–173 (2017).
15. Counihan, N. A. *et al.* *Plasmodium falciparum* parasites deploy RhopH2 into the host erythrocyte to obtain nutrients, grow and replicate. *Elife* **6**, (2017).
16. Conrad, D. F., Andrews, T. D., Carter, N. P., Hurles, M. E. & Pritchard, J. K. A high-resolution survey of deletion polymorphism in the human genome. *Nat. Genet.* **38**, 75–81 (2006).
17. Gadalla, N. B. *et al.* Alternatively spliced transcripts and novel pseudogenes of the

- Plasmodium falciparum resistance-associated locus pfcrt detected in East African malaria patients. *J. Antimicrob. Chemother.* **70**, 116–23 (2015).
- 18. Tolia, N. H., Enemark, E. J., Sim, B. K. L. & Joshua-Tor, L. Structural basis for the EBA-175 erythrocyte invasion pathway of the malaria parasite Plasmodium falciparum. *Cell* **122**, 183–93 (2005).
  - 19. Dvorin, J. D., Bei, A. K., Coleman, B. I. & Duraisingham, M. T. Functional diversification between two related *Plasmodium falciparum* merozoite invasion ligands is determined by changes in the cytoplasmic domain. *Mol. Microbiol.* **75**, 990–1006 (2010).
  - 20. Taylor, H. M., Grainger, M. & Holder, A. A. Variation in the expression of a Plasmodium falciparum protein family implicated in erythrocyte invasion. *Infect. Immun.* **70**, 5779–89 (2002).
  - 21. Jennings, C. V *et al.* Molecular analysis of erythrocyte invasion in Plasmodium falciparum isolates from Senegal. *Infect. Immun.* **75**, 3531–8 (2007).
  - 22. Li, H. Toward better understanding of artifacts in variant calling from high-coverage samples. *Bioinformatics* **30**, 2843–2851 (2014).
  - 23. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
  - 24. McKenna, A. *et al.* The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
  - 25. Otto, T. D. *et al.* Long read assemblies of geographically dispersed Plasmodium falciparum isolates reveal highly structured subtelomeres. *Wellcome Open Res.* **3**, 52 (2018).
  - 26. Abyzov, A., Urban, A. E., Snyder, M. & Gerstein, M. CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. *Genome Res.* **21**, 974–84 (2011).
  - 27. Boeva, V. *et al.* Control-FREEC: a tool for assessing copy number and allelic content using next-generation sequencing data. *Bioinformatics* **28**, 423–425 (2012).
  - 28. Nei, M. Analysis of gene diversity in subdivided populations. *Proc. Natl. Acad. Sci. U. S. A.* **70**, 3321–3 (1973).

## Figure Legends

Figure 1

**High quality variants by position, length and per-chromosome.** A) Distribution of high quality structural variants over each chromosome. B) Distribution of deletions by size categories. C) Distribution of duplications by size categories. D) Distribution of distinct form of deletion across each chromosome. E) Distribution of distinct forms of duplication across each chromosome.

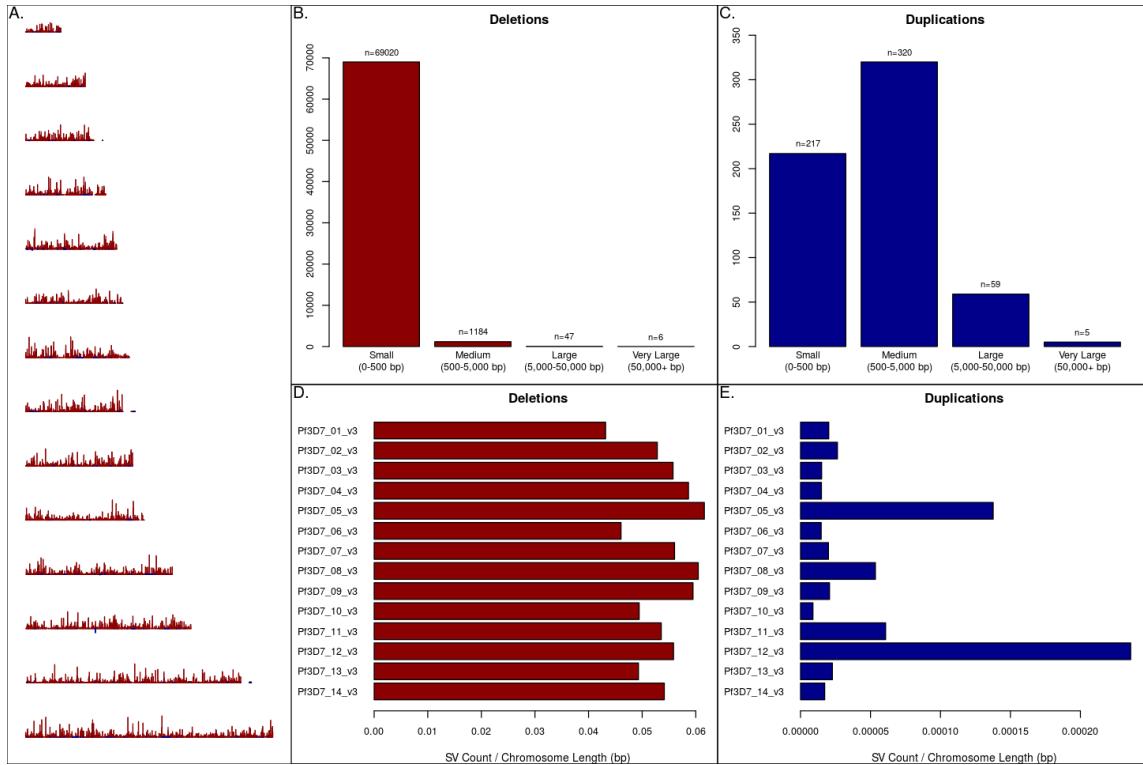


Figure 2

**Coverage plot confirming the whole gene duplication of *mdr1*.** Blue represents the per base coverage, Orange indicates the predicted structural variant, Green indicates the gene of interest, Grey indicates neighbouring genes.

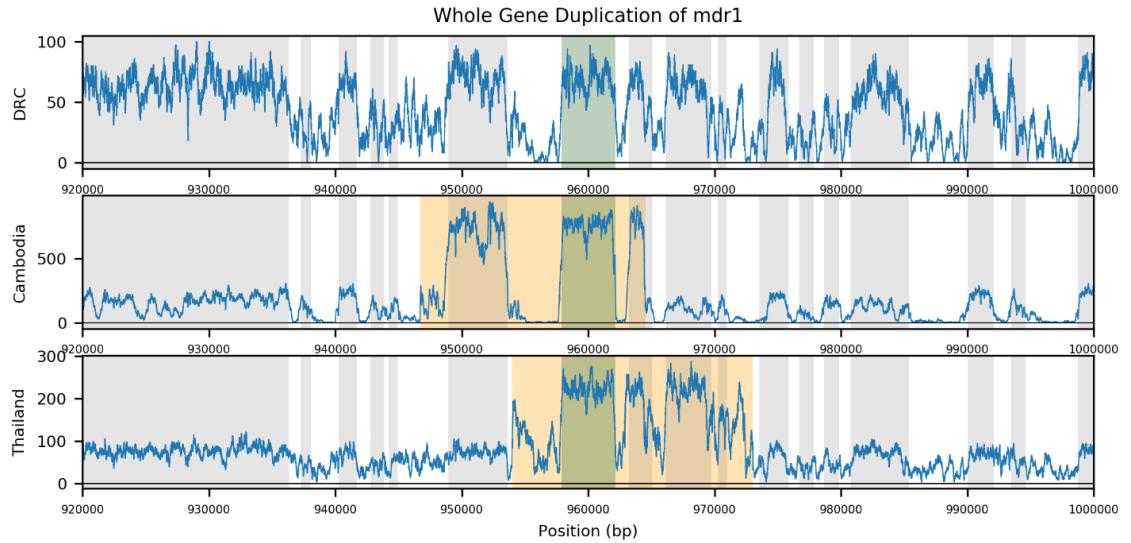


Figure 3

**Coverage plot confirming the *gch1* promoter duplication.** Blue represents the per base coverage, Orange indicates the predicted structural variant, Green indicates the gene of interest, Grey indicates neighbouring genes.

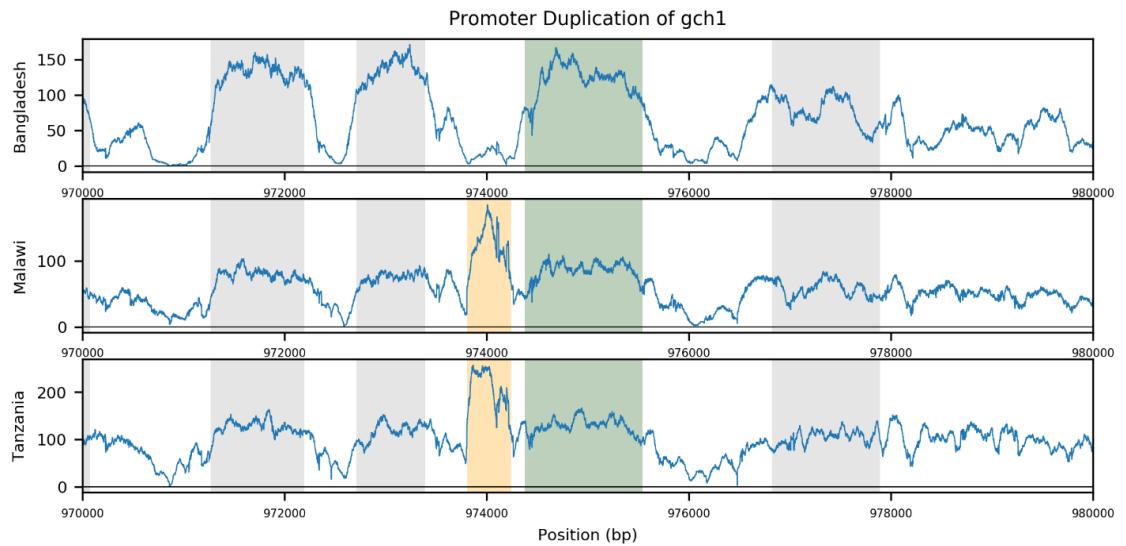
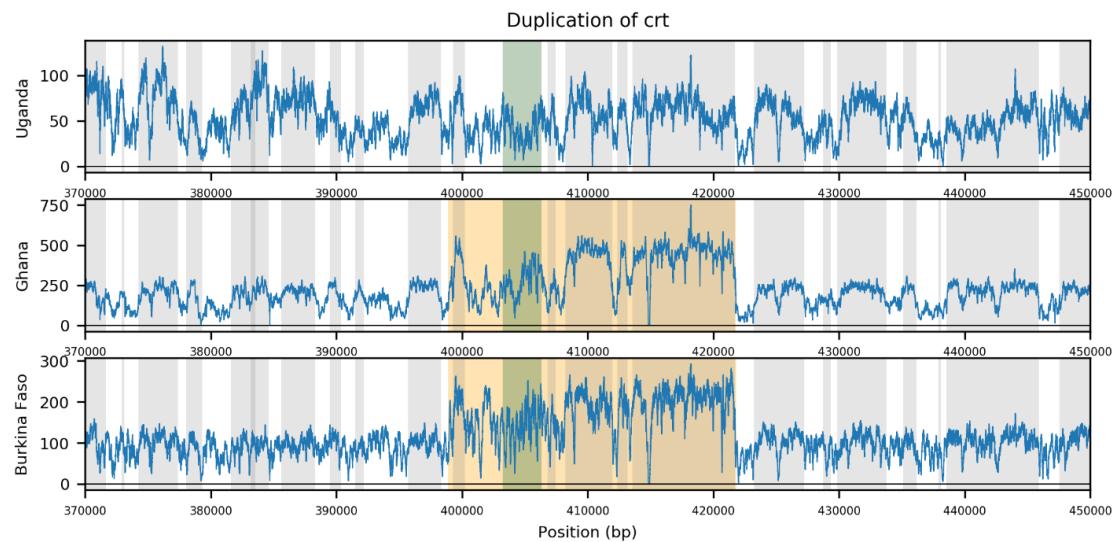


Figure 4

**Coverage plot demonstrating duplication of *crt*.** Blue represents the per base coverage, Orange indicates the predicted structural variant, Green indicates the gene of interest, Grey indicates neighbouring genes.



## Supplementary Information

**Supplementary table 1:** Summary of total sample set (n=3,110).

Continent	Region	Country	Count
Total	-	-	3110
Africa	-	-	1572
Africa	West Africa	-	738
Africa	West Africa	Burkina Faso	48
Africa	West Africa	The Gambia	66
Africa	West Africa	Ghana	446
Africa	West Africa	Guinea	117
Africa	West Africa	Mali	55
Africa	West Africa	Nigeria	6
Africa	Central Africa	-	360
Africa	Central Africa	Cameroon	128
Africa	Central Africa	DRC	232
Africa	East Africa	-	474
Africa	East Africa	Kenya	38
Africa	East Africa	Madagascar	18
Africa	East Africa	Malawi	353
Africa	East Africa	Tanzania	65
Asia	-	-	1514
Asia	South Asia	-	53
Asia	South Asia	Bangladesh	53
Asia	Southeast Asia	-	1461
Asia	Southeast Asia	Cambodia	651
Asia	Southeast Asia	Laos	113
Asia	Southeast Asia	Papua New Guinea	26
Asia	Southeast Asia	Myanmar	134
Asia	Southeast Asia	Thailand	335
Asia	Southeast Asia	Vietnam	202
America	-	-	24
America	South America	-	24
America	South America	Colombia	15
America	South America	Peru	9

**Supplementary Table 2:** Most frequent structural variants by count. [large file]

**Supplementary Table 3:** Most distinct variants by region-based  $F_{ST}$  values. [large file]

**Supplementary Table 4:** Full list of haplotype counts for *crt*. [large file]

**Supplementary Figure 1:** PfGlobalSV visualisation tool screenshot.



## **Chapter 4:**

Analysis of global long read *Plasmodium falciparum* genomes  
identifies novel inversions



## **RESEARCH PAPER COVER SHEET**

***PLEASE NOTE THAT A COVER SHEET MUST BE COMPLETED FOR EACH RESEARCH PAPER INCLUDED IN A THESIS.***

### **SECTION A – Student Details**

<b>Student</b>	Matt Ravenhall
<b>Principal Supervisor</b>	Taane Clark
<b>Thesis Title</b>	A bioinformatic analysis of malaria host and pathogen genomics

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

### **SECTION B – Paper already published**

Where was the work published?	n/a		
When was the work published?	n/a		
If the work was published prior to registration for your research degree, give a brief rationale for its inclusion	n/a		
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?	International Journal of Parasitology		
Please list the paper's authors in the intended authorship order:	Matt Ravenhall, Ernest Diez Benavente, Paola Florez de Sessions, Eloise M. Walker, Martin L. Hibberd, David A. Baker, 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 developed the inversion detection method, conducted all analyses from assembled samples, created all figures, and wrote the manuscript under the supervision of Taane Clark.
--	--

**Student Signature:** \_\_\_\_\_ **Date:** \_\_\_\_\_

**Supervisor Signature:** \_\_\_\_\_ **Date:** \_\_\_\_\_

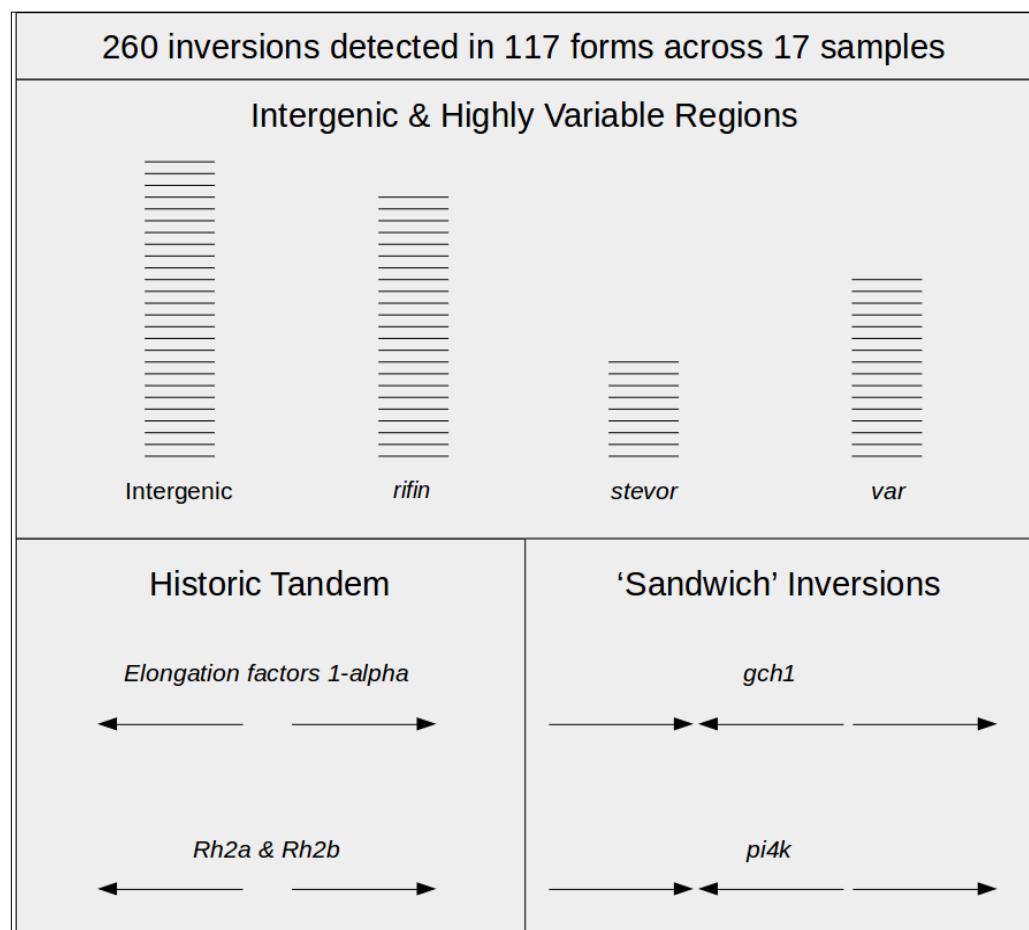
# **Analysis of global long read *Plasmodium falciparum* genomes identifies novel inversions**

Matt Ravenhall<sup>1</sup>, Ernest Diez Benavente<sup>1</sup>, Paola Florez de Sessions<sup>2</sup>, Eloise M. Walker<sup>1</sup>, Martin L. Hibberd<sup>1</sup>, David A. Baker<sup>1</sup>, Susana Campino<sup>1,\*</sup>, Taane G. Clark<sup>1,3,\*</sup>

- 1) Pathogen Molecular Biology Department, London School of Hygiene and Tropical Medicine, London WC1E 7HT, UK
- 2) Genome Institute of Singapore, Singapore
- 3) Department of Infectious Disease Epidemiology, Faculty of Epidemiology and Population Health, London School of Hygiene and Tropical Medicine, London WC1E 7HT, UK

\*Joint last authors

## Graphical Abstract:



## Abstract

Structural rearrangements, including deletions, duplications and inversions, in the *Plasmodium falciparum* malaria genome underpin a range of genetic variation associated with antimarial resistance and host-pathogen interactions. Established examples include the duplication of *mdr1* associated with mefloquine resistance, and duplications of *gch1* associated with sulfadoxine-pyrimethamine resistance. In comparison, knowledge of inversions is incomplete, particularly as they are thought to exist within highly variable or repetitive regions, which are often excluded from genomic data analyses involving short read sequencing data. With the emergence of long read based technologies there is an opportunity to identify novel inversions genome-wide. We developed a pipeline for the robust detection of inversions and, using PacBio assemblies of 17 isolates from 14 countries, identified 260 putative inversions (median: 17; range: 7-20, per sample) in 117 specific forms. These inversions accounted for a median of 506 bp (range: 300-19,068 bp) per sample, compared to a median of 20,055 (range: 841-21,038) SNPs. Whilst, 119 (45.8%) inversions were found in highly variable gene families (*rifin*: 75 (30 genes), *stevor*: 8 (5 genes), *var*: 36 (22 genes)), others involved significant rearrangement of genes associated with anti-malarial resistance such as *gch1* and *pi4k*, and erythrocyte invasion such as *RH2b/RH2a*. Our work provides the first catalogue of polymorphic inversions in *P. falciparum* and will facilitate investigations into their functions.

**Word Count:** 215/300

**Keywords:** Malaria, *Plasmodium*, inversions, structural variation, genomics, PacBio, long-read sequencing, anti-malarial resistance

## 1. INTRODUCTION

The *Plasmodium falciparum* parasite inflicts a high morbidity and mortality on human populations in malaria endemic regions, especially Sub-Saharan Africa. Investigations of *P. falciparum* adaptation for host immune evasion, anti-malarial resistance and other important genetic mechanisms, have focused on single-nucleotide polymorphisms (SNPs), and structural variants such as duplications and deletions (Ravenhall *et al.* 2016, Cheeseman *et al.* 2009). However polymorphic inversions, which involve a change in orientation of a segment of DNA within a chromosome and represent significant genomic rearrangements, have been overlooked in studies of *P. falciparum*. In general, these variants consist of a specific region either being the reverse of an orthologous sequence in a reference genome or, when combined with duplication, of a high identity paralog. Such variants have the potential to impact gene expression to the point of pseudogenisation and underpin phenotype-enhancing events particularly when combined with other chromosomal rearrangements, such as duplications and deletions (Naseeb *et al.* 2016). Inversions are likely to emerge during recombination events or via the actions of transposable elements (Kirkpatrick *et al.* 2010, Mathiopoulos *et al.* 1999), and are thought to underpin the emergence of humans from other primates (Fuek *et al.* 2005). The phenotypic impacts of inversions are not limited to genic disruption as they may, when located within intergenic regions, define novel intrinsic terminator sites or disrupt existing regulatory elements (Gusarov *et al.* 1999).

Only a handful of inversions have been studied in detail in *P. falciparum* and current knowledge about their possible functional effects is still limited. Inverted gene pairs include *RH2a* and *RH2b*, which are erythrocyte invasion ligands and a target of protective immunity (Gunalan *et al.* 2012), and elongation factors 1-alpha involved in binding aminoacyl-tRNAs to the ribosomal acceptor site during translation (Dvorin *et al.* 2010, Riis *et al.* 1990). Inversions are also thought to play a key role in the active rearrangement

of highly variable genes involved in host-pathogen interactions such as those of the *var*, *stevor*, and *rifin* gene families, and have previously been identified in the apicoplast of *Plasmodium chabaudi chabaudi* (Sato *et al.* 2013). Knowledge of the full extent of inversions, their global diversity, and effect upon the evolution of the *P. falciparum* genome, is limited. This is despite the known roles of duplications and deletions in conveying anti-malarial resistance factors such as *mdr1* and *gch1* (Borges *et al.* 2011, Nair *et al.* 2008).

Previously the discovery of putative inversions has been limited to candidate regions and to short read based analytical approaches that have difficulties in resolving particularly repetitive or variable regions, especially for an AT-rich genome such as *P. falciparum*. With the development of newer, long read based sequencing platforms such as PacBio Single Molecule Real Time (SMRT) it is now possible to construct a genome that reduces the impact of poor *de novo* assembly in repetitive regions, but also decreases the reliance on a reference genome where an inversion may not be present. Therefore, we combined a PacBio long read sequencing approach of *P. falciparum* samples with an in-depth inversion discovery pipeline that utilises whole chromosome alignment. Our dataset consists of seventeen strains sourced from eleven countries, four of which were sequenced and assembled in-house. In total we identified 260 putative inversions in 117 specific forms, two of which include genes associated with anti-malarial resistance. Our findings highlight the potential role of putative inversions in known variable genes, such as *var*, *rifin*, and *stevor*, represent a robust method for detecting ancestral paralog inversions such as *RH2a* and *RH2b*, and point towards a novel tandem inversion of *pi4k* that may have an impact on anti-malarial resistance.

## **2.MATERIALS AND METHODS**

### **2.1. Samples and assembly**

DNA was extracted from the cultures of four lab strains (K1, Thailand; D10, Papua New Guinea; NF54, Africa; T996 Thailand), and sequenced on the Pacific Biosciences (PacBio) RSII long read technology at the Genome Institute of Singapore. Sequencing reads were assembled using Hierarchical Genome Assembly Process HGAP2 implemented in the SMRT Portal software suite. Short low confidence contigs (length <1000 or identity < 90%) were removed from subsequent analyses. The overlaps between the start and end of large contigs were found by self-aligning using *Mummer* software (Delcher *et al.* 1999) and removed using in-house scripts. Contigs were aligned, scaffolds inferred, reordered and, if needed, reverse-complemented according to the 3D7 reference using the *mummer* tool and in-house scripts. Following this the reads were realigned to the scaffolds to improve the consensus concordance, leading to complete chromosomes. Chromosome-wide assemblies of PacBio sequence data were also available for a further 13 strains (accession number ERP009847), including five laboratory strains 7G8 (Brazil), Dd2 (Indochina), GB4 (Ghana), HB3 (Honduras), IT (Brazil), and eight field isolates GN01 (Guinea), SN01 (Senegal), CD01 (Congo), GA01 (Gabon), KE01 (Kenya), SD01 (Sudan), and KH01 and KH02 (Cambodia) (see Otto *et al.* 2018).

### **2.2. Inversions discovery and verification**

For each of the 17 strains, the whole chromosome alignments were mapped against the 3D7 reference (v3) using *nucmer* (Delcher *et al.* 1999) to identify genomic regions of high identity (greater than 75%). Additional filtering was based on inclusion of reverse alignments, sequence length greater than 300 bp, and 3D7-strain mapping distances within 50 kbp. The pipeline and associated scripts are available (<https://github.com/mattravenhall/PfINV>). Manual verification was assisted through

alignment of candidate regions with BLAST (Camacho *et al.* 2009), as well as the multiple alignment of orthologous sequences using *ClustalO* (Sievers *et al.* 2011). SNPs were inferred using *snp-sites* software (Page *et al.* 2016), and their impacts were predicted using functions from our ‘bio’ Python module (available at <https://github.com/mattravenhall/bio>). This package was also used for the dissection of palindromic motifs within candidate sequences.

### 3. RESULTS

#### 3.1. Inversions are rare and long variants

We identify 260 putative inversions with greater than 75% sequence identity and 92 with greater than 95% identity (**Figure 1d**). The inversion set (n=260) represents a mean number per sample of 15.3 ( $\pm 4.7$ ) (median: 17; range: 7-20) (**Table 1**) and corresponds to 117 distinct forms with a median length of 506 bp (mean 1,577.3 bp) (**Supplementary Table 1**). In comparison we identified a median of 20,055 SNPs (841 to 21,038) and a median of 29,630 (range: 4,249 to 38,141, mean: 23,281.1) bp of inversions per sample, with a ratio of number of SNPs to total inversion length close to one (mean: 1.38, median: 0.66; range: 0.02-4.85). This suggests an equivalent proportion of the *P. falciparum* genome is impacted by inversions as by SNPs, though SNPs are likely to emerge far more often. There was no significant difference in the number of inversions between the three primary continents (Africa (n=8):  $16.4 \pm 4.4$ , Asia (n=5):  $15.2 \pm 5.9$ , America (n=3):  $13.3 \pm 5.0$ ; Kruskal-Wallis P=0.570), though sample size is limited.

The majority of putative inversions (n=217, 83.5%) are present within intergenic regions or the highly variable members of the *rifin* (75 inversions; 30 genes), *stevor* (8 inversions; 5 genes), and *var* (36 inversions; 22 genes) gene families (**Figure 1b**). These inversions are significantly shorter than the others (median length: intergenic or highly variable

regions 482 bp, vs. other regions: 5,329 bp; Wilcoxon test  $P=2.30 \times 10^{-14}$ ), with generally low identities relative to 3D7 (mean: 87.5%, median: 87.4%) suggesting non-recent or more complex inversion events. We also identify two historic inversion/duplication events present within *RH2a/RH2b* and elongation factor 1-alpha (PF3D7\_1357000 and PF3D7\_1357100). Further, two sample specific inversions are present in Dd2, where we identify the previously characterised *gch1* 'sandwich' inversion (Miles *et al.* 2016), and in GB4 where we identify a novel tandem inversion of anti-malarial drug target *pi4k*. Overall, we identify 119 variants within highly variable genes, 98 within intergenic regions, 42 associated with *RH2a/RH2b*, 11 associated with elongation factor 1-alpha, 1 of *gch1* and 1 of *pi4k*.

### 3.2. Highly variable genes contain numerous putative inversions

Of the putative inversions within 30 different *rifin* genes, 15 are present in single samples. Two genes (PF3D7\_0402500, 12 samples; PF3D7\_0402700, 14 samples) contain inversions in the majority of samples. For PF3D7\_0402500, we identify inversions in 12 samples, six (GB4, KE01, IT, Dd2, NF54, T996) with short forms (INV019; 401 bp) and six (7G8, GA01, GN01, K1, KH01, KH02) with longer forms (INV018; 562 bp), including one in Guinean GN01 with a particularly high identity (98.4%). Similarly, for PF3D7\_0402700, nine samples (IT, KE01, KH02, T996, 7G8, Dd2, GA01, GN01, NF54) contain a short form (INV020; 403 bp), whilst five (CD01, GB4, K1, KH01, SN01) have a longer form (INV021; 570 bp). Given the similar lengths to those in PF3D7\_0402700, these genes may represent a tandem inversion or duplicated domain whereby INV018 pairs with INV021 and INV019 pairs with INV020. Alignment of the two genes supports this, with the pair sharing an approximately 800 bp region (from 350/450 to 1100/1200 bp, 61.6% coverage, 81.6% identity) (**Figure 2**). Of the remaining putative *rifin* inversions, the KH01 strain displays two inverse matches for specific regions in both

PF3D7\_0114700 and PF3D7\_0200500. All four forms (INV005a, INV005b, INV008a, INV008b) have modest identities (PF3D7\_0114700: 84.49% and 84.43%, PF3D7\_0200500: 80.82% for both), suggesting the presence of two KH01-specific imperfect duplications within these genes. Manual inspection of those candidate regions found that INV005a/b aligned to the start and upstream regions of multiple *rifin* genes across several chromosomes, with INV008a/b similarly being found to align to the ends of multiple *rifin* genes. This suggests that these specific regions correspond to shared domains across a subset of *rifin* genes, with INV005a/b perhaps also including shared promoter regions. A similar feature is present within the Papua New Guinea laboratory strain D10 whereby three 329 bp inverted regions (INV073a, INV073b, INV073c) again with acceptable identities (84.8% for all three) are present. Further alignment of the D10 region against 3D7 found that the core sequence matched multiple *var* genes, suggesting that D10 has multiple repeats of a common *var* domain in this copy of PF3D7\_1100300.

Very few isolates share inversions within *var* genes, with thirty-two variants being present once and two being present twice. This diversity, and the generally low identities for these putative inversions (mean: 86.0%, median: 86.1%) reflect the high level of variability within this gene family. Two notable exceptions are PF3D7\_0600600, in which we identify six putative inversions (INV033 to INV038) in five samples, and PF3D7\_0100500, where we identify a 482 bp inversion (INV003; 97.3% identity) in the Gambian GA01 strain. This high identity inversion appears to correspond to a duplicate inversion 18 kb upstream and a number of similar regions around the genome, suggesting that the sequence may represent some form of common motif or transposable element. For PF3D7\_0600600, we observe a heterogeneous group of six putative inversions, two of which (INV033 and INV037) are present in the Kenyan KE01 strain, whilst the other four (INV034, INV035, INV036, and INV038) are in NF54, KH02, K1, and CD01.

Though two variant pairs (INV035 with INV036, and INV037 with INV038) share starting breakpoints in KH02 with K1, KE01 and CD01, there is limited general concordance for either variant positions or length. In general, this highlights the highly diverse variability of PF3D7\_0600600.

Even fewer putative inversions are present within *stevor* gene family members, with five samples containing variants within five genes, though the degree of identity is relatively low (mean: 83.79%, median: 82.91%). Of those, the IT strain has candidate inversions in three genes, a 651 bp region in PF3D7\_0631900 (INV039), a 932 bp region in PF3D7\_1254300 (INV081), and a 1012 bp region in PF3D7\_1300900 (INV093). A similar 959 bp inversion in PF3D7\_1300900 is present in the Guinean GN01 strain. These inversions are all generally low identity with broader alignment showing that these specific sequences are generally shared across multiple *stevor* genes. Finally, three African samples (GB4, GA01, GN01) contain a putative inversion (INV117) within the same region of PF3D7\_1400700, each approximately 1 kbp in length. Poor alignment of these flanking regions and the general repetitiveness of this region suggests that INV117 may represent a detection of an inverted nearby gene with relatively high similarity to PF3D7\_1400700.

### 3.3. Inverted paralogs play a significant role in *P. falciparum* genome

Our approach identified two gene pairs (*RH2a/RH2b* and elongation factors 1-alpha) that have emerged through historical duplication and inversion. *RH2a* and *RH2b* represent the classic example of a tandem inverted duplication in *P. falciparum*, being absent from other *Plasmodium* species (Otto et al. 2014) and bearing a key role in erythrocyte invasion (Duraisingh *et al.* 2003). The pair consists of two highly similar genes that emerged from the inverted duplication of their ancestral gene and therefore represent high identity

inverted paralogs of one another. Putative inversions were identified in 13 samples (7G8, GB4, Dd2, HB3, IT, CD01, KE01, GA01, SD01, GN01, K1, NF54, T996), but absent in the remaining four isolates due to *RH2b* being fully deleted in KH01 and D10, partially deleted in the 3' region in SN01, and having minor deletions in KH02. We also identify putative inversions of the intergenic regions between *RH2a* and *RH2b*, and upstream towards and including a ~200 bp region of *RH6*. These likely represent components of the original inversion duplication between which later variation has occurred.

A lesser-known example of a historic tandem inversion is elongation factors 1-alpha (PF3D7\_1357000, PF3D7\_1357100), which presents with similarly high identities as *RH2a* and *RH2b* (mean: 99.06%, median: 99.51%). This gene pair are, as with *Rh2a* and *Rh2b*, a well-established example of a tandem duplication where one of the pair is inverted relative to the other (Vinkenoog *et al.* 1998). This pair is also intriguing due to the presence of a bi-directional promoter within the intergenic region between the two genes (Fernandez-Becerra *et al.* 2003). Putative inversions were detected in sixteen samples, in which the majority have a 4,433 bp variant. These were absent in T996, and the Congolese CD01 strain has shorter 1,354 and 1,356 bp inversions (INV111 and INV113), which are present for each gene separately. This emphasises the non-inverted nature of the intergenic region between the gene pair, as this region aligns near perfectly with the reference but in only one direction.

3.4. Dd2 features a ‘sandwich’ inversion of YHM2, PF3D7\_1223900, and *gch1*  
Our approach highlighted a more complex rearrangement containing *gch1* in which a tandem triplication event contains an inverted copy (INV080) between two non-inverted copies of the region. We refer to this rearrangement as a ‘sandwich’ inversion for ease of reference. The Dd2 region therefore includes the wildtype *YHM2* (PF3D7\_1223800),

PF3D7\_1223900, and *gch1* (PF3D7\_1224000) region followed by a near-identical inversion of those three genes, completed by a third set of the three genes in wildtype arrangement (**Figure 3a**). This feature represents a 5,306 bp region of Dd2, corresponding to a 5,329 bp region of our 3D7 reference, with 98.6% identity (**Supplementary table 1**). This feature represents the inverted segment of the previously identified Dd2 tandem triplication (Miles *et al.* 2016). This form of inversion appears to be unique to Dd2 and may therefore represent variation acquired within the laboratory after initial collection. Phenotypically the inclusion of PF3D7\_1223800 (*YHM2*) and PF3D7\_1223900 is curious as it could suggest some secondary impact alongside that of *gch1* duplication, or at least a limited negative impact on the preserved genotype. The preservation of all three genes also supports this feature having occurred as a result of a single triplication event, rather than multiple recombination events.

### 3.5. GB4 contains a putative ‘sandwich’ inversion of PF3D7\_0509700, PI4K and PF3D7\_0509900

A ‘sandwich’ inversion is present within a ~7 kbp region of GB4 with the inverted region sharing 97.7% identity (INV027). The central inversion (*PI4K\_B*) includes two genes, one of unknown function with a predicted C-terminal Nse4\_C domain (PF3D7\_0509900) and phosphatidylinositol 4-kinase (*pi4k*, PF3D7\_0509800), the target of imidazopyridines (McNamara *et al.* 2013). The third duplication (*PI4K\_C*) also includes a section of PF3D7\_0509700 (a conserved unknown protein) (**Figure 3b**). The full PF3D7\_0509700 may be duplicated but an assembly gap between the second (*PI4K\_B*, inverted) and third (*PI4K\_C*, longer) duplication makes full resolution of the region difficult. Both *PI4K\_B* and *PI4K\_C* contain SNPs encoding putative premature stop codons, whilst *PI4K\_A* contains some minor deletions relative to 3D7, particularly in repetitive domains. *PI4K\_B* also lacks a 111 bp segment in its 5' region. It seems probable

that *PI4K\_B* and *PI4K\_C* are non-functional alleles of wildtype *PI4K*. If *PI4K\_A* is also shown to be non-functional, this ‘sandwich’ inversion may represent significant disruption of the region rather than an increase in expression, as is classically assumed with duplication events. Evolutionarily, the presence of PF3D7\_0509700 in *PI4K\_C*, but not *PI4K\_B*, suggests that this triple duplication arose through at least two recombination events. Repeated recombination of the region would be consistent with a process of pseudogenisation following the removal of preserving selection.

### 3.6. Intergenic regions contain inversions

Ninety-eight inversions are present in intergenic regions of 3D7, and most bear high identity (mean: 92.3%, median: 94.2%, range: 76.5-99.05%). Within this subset are a number of distinct regions, which reflect variation stemming from low complexity, local inversions as well as constituting parts of larger genic inversions. Twenty-six putative inversions exist within chromosome 13 (128,289 to 128,794 bp) but bear generally lower identity scores (mean: 88.5%, median: 89.6%, range: 76.5-97.7%) with the exception of three particularly high identity variants in Kenyan KE01 (INV095, 300 bp, 97.7%), Congolese CD01 (INV096, 302 bp, 97.7%) and Honduran HB3 (INV096, 302 bp, 96.2%). In general, this core sequence has low complexity and high repetition with manual realignment highlighting similarity between this region and an inverse sequence ~5 kbp upstream. Variants within this region would therefore reflect local similarity with those of high identity seeming to reflect a true local 300 bp inversion in KE01, CD01, and HB3.

A subset of intergenic inversions hint at possible errors in the 3D7 reference, likely due to its assembly from short read sequences. Specifically, whilst manual alignment of INV110, INV114, and INV115 found high identity inverted matches for those specific

sequences (INV110: mean: 98.9%, median: 99.0%, range: 98.5 - 99.0%; INV114: mean: 95.7%, median: 95.9%, range: 91.3 - 97.6%; INV115: mean: 97.3%, median: 97.6%, range: 94.5-99.1%), this did not hold true for their flanking regions. All three candidate inversions are present in multiple sequences (INV110: eleven isolates: GB4, HB3, Dd2, IT, KE01, CD01, SD01, GA01, NF54, K1, T996; INV114 and INV115: twelve isolates: 7G8, GB4, HB3, Dd2, IT, KE01, CD01, SD01, GA01, NF54, K1, T996) with high identity for both the publicly available and in-house PacBio sequences, ruling out the possibility of contamination and the sequence is too specific, extensive and isolated for a possible sequencing error. Realignment and manual inspection of these variants and 1,000 bp flanking regions found high identity alignment (>95%) with *Plasmodium* sp. gorilla clade G1 genome assembly (LT963426), a PacBio assembly for a species closely related to *P. falciparum*. This observation suggests that the corresponding genomic region within 3D7 may be incorrectly assembled, potentially due to repetitive sequences that PacBio is better able to resolve.

#### 4. DISCUSSION

Inversions within the *P. falciparum* genome are generally rare, with an average of ~15 per sample, and can generally be classified into three forms: classic, tandem, and sandwich. Classic inversions refer to those examples where a specific sequence has been flipped *in situ* without duplication. Tandem inversions refer to tandem duplications in which one of the copies is flipped. Whilst, sandwich inversions refer to triplication events or beyond in which the middle copy is reversed relative to the other copies. Classic inversions are particularly rare, with no clear genic examples being identified in this study, though a number of intergenic examples were present (such as INV110); instead inversions seem to occur in combination with duplication events. Tandem inversions include the classic examples of *RH2a/RH2b* and elongation factors 1-alpha, but also an

array of intergenic or highly variable regions. Finally, only two examples of sandwich inversions were observed, the previously identified inversion of *gch1* within a triplication in Dd2, and the novel inversion of *pi4k* in GB4. Both represent variants within laboratory strains with none such examples identified within our field isolates.

As may have been expected given the potentially disruptive consequences of this form of genomic rearrangement, most putative inversions were present within intergenic or highly variable regions. The majority are of lower identity, suggesting the detection of either pseudo-inverted regions or older inversions with depreciated identities. Despite this challenge, we identify several specific, high identity inversions that likely represent true inversion-based recombination of these genes. The use of PacBio SMRT long read sequencing represents a significant opportunity to resolve some genetic regions that have been previously overlooked due to usual exclusion of these regions, and follow-up functional work may present insights into virulence or host immune evasion. The *P. falciparum* genome is host to three families of highly variable genes, the products of which are expressed on the surfaces of infected human erythrocytes conveying key roles in malaria pathology (Niang *et al.* 2009, Gardner *et al.* 1996, Kyes *et al.* 1999). Given their exposure to the human immune system, there is significant selective pressure driving these genes to be hyper-variable with this variability consisting of SNPs, indels, and larger genetic rearrangements. Inversions have been presumed to play a role, but little is known about the extent. A greater understanding of the structural variation for these hyper-variable genes is one key aspect to understanding their role in infection and immune evasion, and therefore in informing future vaccine or anti-malarial strategies.

Inversions were observed across several samples for specific regions of the same genes or intergenic regions, suggesting the detection of specific sub-domains within these

diverse genes. One such example is that of the *rifin* pseudogenes PF3D7\_0402500 and PF3D7\_0402700, for which each contains a pair of mirrored domains approximately 400 and 560 bp in length (INV019 with INV020, and INV018 with INV021). This specificity of similarity is striking, and points towards a possible direct functional relationship between these two genes. Similar selection signals also appear to be present for intergenic inversions such as INV110, a 579 bp inversion downstream of the putative chromosome condensation regulator PF3D7\_1356600. The functional implications of INV110 are unclear, as the specific region is intergenic, but its presence within multiple samples suggests some level of active conservation.

One initially unexpected finding of our study was the detection of novel tandem inverted duplications similar to *RH2a/RH2b* and elongation factors 1-alpha, one example being PF3D7\_0402500 and PF3D7\_0402700. This form of structural variant is uncommon but significant, and arguably represents an under-appreciated component of *P. falciparum* evolution. This is particularly compelling when considering the classic example of *RH2a/RH2b*, a gene pair specific to *P. falciparum*, which have been linked to sialic acid independent invasion, and erythrocyte binding (Gunalan *et al.* 2012, Desimone *et al.* 2009, Sahar *et al.* 2011). Further, two specific regions contain unique sandwich inversions: the inverted duplication of *gch1* in Dd2 (INV080) and of *pi4k* in GB4 (INV027). Curiously both genes have significant roles in anti-malarial resistance; for *gch1*, whole gene and upstream duplications have been associated with late stage resistance to sulfadoxine-pyrimethamine (SP) (Borges *et al.* 2011, Ravenhall *et al.* 2016), whilst *pi4k* is the target of imidazopyrazines (McNamara *et al.* 2013). Both are also lab strains rather than field isolates, suggesting that these inversions may have emerged within the lab after initial isolation. The sandwich inversion of *gch1* in Dd2, in which a triplication of *gch1* features an inverted central copy, has been previously characterised

in some detail (Miles *et al.* 2016), albeit primarily within the context of duplication given the role of similar *gch1* duplications in SP resistance. In contrast, the inverted duplication of *PI4K* appears to be novel, and less clearly resolved as a triplication due to the presence of an assembly gap between *PI4K\_B* and *PI4K\_C*. Manual inspection suggests that this region may be undergoing pseudogenisation in GB4, owing to the presence of putative premature stop codons within *PI4K\_B* and *PI4K\_C*. It is unclear whether the SNPs present within *PI4K\_A* would also lead to a loss of function. The third *PI4K\_C* section of the triplication also includes a copy of PF3D7\_0509700, which has been shown to bind with alternative splicing regulator *PfSRI* (Eschar *et al.* 2015), however the aforementioned assembly gap makes full investigation difficult. It is similarly unclear whether triplication of PF3D7\_0509900 has a phenotypic impact.

Inversions play an understated role in *P. falciparum* genomics, but the phenotypic impact of this variation remains poorly characterised. Our study represents a step towards a broader understanding of *P. falciparum* inversions around the globe for both specific candidates, particularly those associated with anti-malarial resistance, and for identifying patterns within the diverse erythrocyte surface exposed *var*, *stevor*, and *rifin* gene families. We also demonstrate the power for long read sequencing approaches in identifying novel inversions within assemblies through whole chromosome alignment and encourage broader use of similar approaches for discovery in larger sample sets.

*Availability of data and material*

The raw sequence data are available from the European Nucleotide Archive (ERP009847).

*Declaration of Interests*

The authors declare no conflicts of interest.

*Acknowledgements*

The Medical Research Council UK funded eMedLab computing resource was used for data analysis.

*Funding*

MR is funded by the Biotechnology and Biological Sciences Research Council (Grant Number BB/J014567/1). TGC and SC are supported by the Medical Research Council UK (MR/M01360X/1, MR/N010469/1) and BBSRC (BB/R013063/1).

*Ethics Approval*

There are no ethical issues.

## References

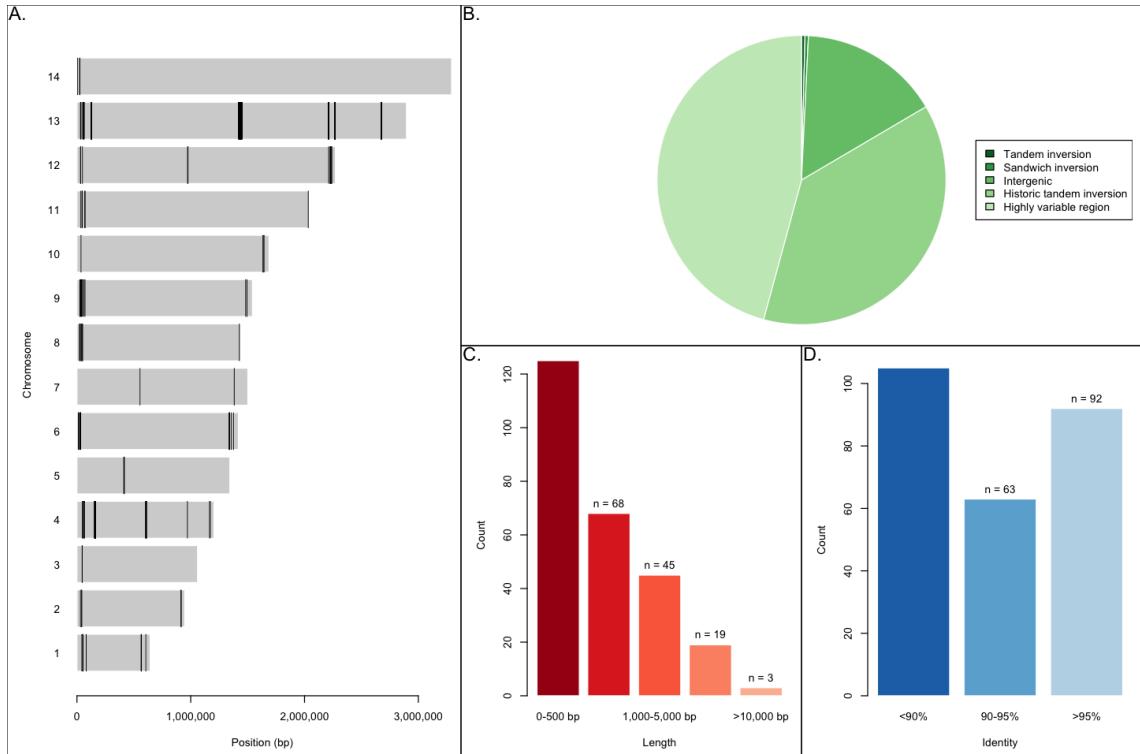
- Borges S, Cravo P, Creasey A, Fawcett R, Modrzynska K, Rodrigues L, Marinelli A, and Hunt P. *Genomewide scan reveals amplification of mdr1 as a common denominator of resistance to mefloquine, lumefantrine, and artemisinin in Plasmodium chabaudi malaria parasites*. Antimicrobial Agents and Chemotherapy. 2011. doi: 10.1128/AAC.01748-10
- Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, and Madden TL. *BLAST+: architecture and applications*. BMC Bioinformatics. 2009. Dec 15;10:421. doi: 10.1186/1471-2105-10-421.
- Cheeseman IH, Gomez-Eschobar N, Carret CK, Ivens A, Stewart LB, Tetteh KK, and Conway DK. *Gene copy number variation throughout the Plasmodium falciparum genome*. BMC Genomics. 2009. 4;10:353. doi: 10.1186/1471-2164-10-353.
- Delcher AL, Kasif S, Fleischmann RD, Peterson J, White O, Salzberg SL. Alignment of whole genomes, Nucleic Acids Res. 1999 Jun 1;27(11):2369-76.
- Desimone, T.M., Jennings, C.V., Bei, A.K., Comeaux, C., Coleman, B.I., Refour, P., et al. *Cooperativity between Plasmodium falciparum adhesive proteins for invasion into erythrocytes*. Molecular Microbiology. 2009. 72: 578–589.
- Duraisingh MT, Triglia T, Ralph SA, Rayner JC, Barnwell JW, McFadden GI, and Cowman AF. *Phenotypic variation of Plasmodium falciparum merozoite proteins directs receptor targeting for invasion of human erythrocytes*. The EMBO Journal. 2003. 22(5):1047-1057. doi: 10.1093/emboj/cdg096
- Dvorin JD, Bei AK, Coleman BI, Duraisingh MT. Functional diversification between two related Plasmodium falciparum merozoite invasion ligands is determined by changes in the cytoplasmic domain. Molecular Microbiology. 2010. 75: 990-1006. doi: 10.1111/j.1365-2958.2009.07040.x
- Eshar S, Altenhofen L, Rabner A, Ross P, Fastman Y, Mandel-Gutfreund Y, Karni R, Llinás M And Dzikowski R. *PfSR1 controls alternative splicing and steady-state RNA levels in Plasmodium falciparum through preferential recognition of specific RNA motifs*. Molecular Microbiology. 96(6). doi: 10.1111/mmi.13007
- Fernandez-Becerra C, de Azevedo MF, Yamamoto MM, and del Portillo HA. *Plasmodium falciparum: new vector with bi-directional promoter activity to stably express transgenes*. Experimental Parasitology. 2003. 103(1–2):88-91. doi: 10.1016/S0014-4894(03)00065-1
- Logan-Klumpler FJ, De Silva N, Boehme U, Rogers MB, Velarde GM, McQuillan JA, Carver T, Aslett M, Olsen C, Subramanian S, Phan I, Farris C, Mitra S, Ramasamy G, Wang H, Tivey A, Jackson A, Houston R, Parkhill J, Holden M, Harb OS, Brunk BP, Myler PJ, Roos D, Carrington M, Smith DF, Hertz-Fowler C, and Berriman M. *GeneDB – An annotation database for pathogens*. Nucleic Acids Research. 2012. 40(D1):D98-D108. doi: 10.1093/nar/gkr1032
- Fuek L, MacDonald JR, Tang T, Carson AR, Li M, Rao G, Khaja R, and Scherer SW. *Discovery of Human Inversion Polymorphisms by Comparative Analysis of Human and Chimpanzee DNA Sequence Assemblies*. PLOS Genetics. 1(4): e56. doi: 10.1371/journal.pgen.0010056
- Gardner JP, Pinches RA, Roberts DJ, AND Newbold CI. Variant antigens and endothelial receptor adhesion in Plasmodium falciparum. PNAS. 1996. 93(8):3503-3508. doi: 10.1073/pnas.93.8.3503
- Gunalan K, Gao X, Lin Yap SS, Huang X, and Preiser PR. *The role of the reticulocyte-binding-like protein homologues of Plasmodium in erythrocyte sensing and invasion*. Cellular Microbiology. 2012. 15(1):35-44. doi: 10.1111/cmi.12038

- Gusarov I and Nudler E. *The mechanism of intrinsic transcription termination*. Molecular Cell. 1999 Apr;3(4):495-504. doi: 10.1016/S1097-2765(00)80477-3
- Kirkpatrick M. How and Why Chromosome Inversions Evolve. PLOS Biology. 8(9): e1000501. doi: 10.1371/journal.pbio.1000501
- Kyes SA, Alexandra Rowe J, Kriek N, and Newbold CI. *Rifins: A second family of clonally variant proteins expressed on the surface of red cells infected with Plasmodium falciparum*. PNAS. 96(16):9333-9338. doi: 10.1073/pnas.96.16.9333
- Martínez-Fundichely A Casillas S Egea R, Ràmia M, Barbadilla A, Pantano L, Puig M, and Cáceres M. InvFEST, a database integrating information of polymorphic inversions in the human genome. Nucleic Acids Res 2014;42:D1027–32.
- Mathiopoulos KD, della Torre A, Santolamazza F, Predazzi V, Petrarca V, and Coluzzi M. *Are chromosomal inversions induced by transposable elements? A paradigm from the malaria mosquito Anopheles gambiae*. Parassitologia. 1999. Sep;41(1-3):119-23.
- McNamara CW, Lee MCS, Lim CS, Lim SH, Roland J, Simon O, Yeung BKS, Chatterjee AK, McCormack SL, Manary MJ, Zeeman A, Dechering KJ, Kumar TRS, Henrich PP, Gagaring K, Ibanez M, Kato N, Kuhen KL, Fischli C, Nagle A, Rottmann M, Plouffe DM, Bursulaya B, Meister S, Rameh L, Trappe J, Haasen D, Timmerman M, Sauerwein RW, Suwanarusk R, Russell B, Renia L, Nosten F, Tully DC, Kocken CHM, Glynne RJ, Bodenreider C, Fidock DA, Diagana TT, and Winzeler EA. *Targeting Plasmodium phosphatidylinositol 4-kinase to eliminate malaria*. Nature. 2013. 12; 504(7479): 248–253. doi: 10.1038/nature12782
- Miles A, Iqbal Z, Vauterin P, Pearson R, Campino S, Theron M, Gould K, Mead D, Drury E, O'Brien J, Rubio VR, MacInnis B, Mwangi J, Samarakoon U, Ranford-Cartwright L, Ferdig M, Hayton K, Su X, Wellemes T, Rayner J, McVean G, and Kwiatkowski D. *Indels, structural variation, and recombination drive genomic diversity in Plasmodium falciparum*. Genome Research. 2016. 26(9): 1288–1299. doi: 10.1101/gr.203711.115
- Naseeb S, Carter Z, Minnis D, Donaldson I, Zeef L, and Delneri D. *Widespread Impact of Chromosomal Inversions on Gene Expression Uncovers Robustness via Phenotypic Buffering*. Molecular Biology and Evolution. 2016. 33(7): 1679–1696. doi:10.1093/molbev/msw045
- Niang M, Yam XY, and Preiser PR. *The Plasmodium falciparum STEVOR multigene family mediates antigenic variation of the infected erythrocyte*. PLOS Pathogens. 2009. 5(2): e1000307. doi: 10.1371/journal.ppat.1000307
- Otto TD, Rayner JC, Boehme U, Pain A, Spottiswoode N, Sanders M, Quail M, Ollomo B, Renaud F, Thomas AW, Prugnolle F, Conway DJ, Newbold C, and Berriman M. *Genome sequencing of chimpanzee malaria parasites reveals possible pathways of adaptation to human hosts*. Nature Communications. 2014. 4754. doi: 10.1038/ncomms5754]
- Otto TD, Boehme U, Sanders M, Reid A, Bruske EI, Duffy CW, Bull PC, Pearson RD, Abdi A, Dimonte S, Stewart LB, Campino S, Kekre M, Hamilton WL, Claessens A, Volkman SK, Ndiaye D, Amambua-Ngwa A, Diakite M, Fairhurst RM, Conway DJ, Franck M, Newbold CI, and Berriman M. *Long read assemblies of geographically dispersed Plasmodium falciparum isolates reveal highly structured subtelomeres*. Wellcome Open Res 2018. 3:52. doi: 10.12688/wellcomeopenres.14571.1
- Page AJ, Taylor B, Delaney AJ, Soares J, Seemann T, Keane JA, and Harris SR. *SNP-sites: rapid efficient extraction of SNPs from multi-FASTA alignments*. 2016. doi: 10.1099/mgen.0.000056

- Ravenhall M, Benavente ED, Mipando M, Jensen ATR, Sutherland CJ, Roper C, Sepúlveda N, Kwiatkowski DP, Montgomery J, Phiri KS, Terlouw A, Craig A, Campino S, Ocholla H, and Clark TG. *Characterizing the impact of sustained sulfadoxine/pyrimethamine use upon the Plasmodium falciparum population in Malawi*. Malaria Journal. 2016. 10.1186/s12936-016-1634-6
- Riis B, Rattan SIS, Clark, BRFC, Merrick WC. *Eukaryotic protein elongation factors*. Trends in Biochemical Sciences. 1990. 15(11):420-424. doi: 10.1016/0968-0004(90)90279-K
- Sahar, T., Reddy, K.S., Bharadwaj, M., Pandey, A.K., Singh, S., Chitnis, C.E., and Gaur, D. *Plasmodium falciparum reticulocyte binding-like homologue protein 2 (PfRH2) is a key adhesive molecule involved in erythrocyte invasion*. PLoS ONE. 2011. 6: e17102.
- Sato S, Sesay AK, and Holder AA. *The unique structure of the apicoplast genome of the rodent malaria parasite Plasmodium chabaudi chabaudi*. PLOS ONE. 8(11). doi:10.1371/annotation/f9f809fc-34b8-42c8-acf3-f8b2616a5f44.
- Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, Lopez R, McWilliam H, Remmert M, Söding J, Thompson JD, and Higgins DG. *Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega*. Molecular Systems Biology. 2011. 7,539. doi: 10.1038/msb.2011.75
- Szpiech ZA, Hernandez RD. *selscan: an efficient multithreaded program to perform EHH-based scans for positive selection*. Mol Biol Evol. 2014;31:2824–7.
- Vinkenoog R, Sperança MA, van Breemen O, Ramesar J, Williamson DH, Ross-MacDonald PB, Thomas AW, Janse CJ, del Portillo HA, and Waters AP. *Malaria parasites contain two identical copies of an elongation factor 1 alpha gene*. 1998. Molecular and Biochemical Parasitology. Jul 1;94(1):1-12. doi: 10.1016/S0166-6851(98)00035-8

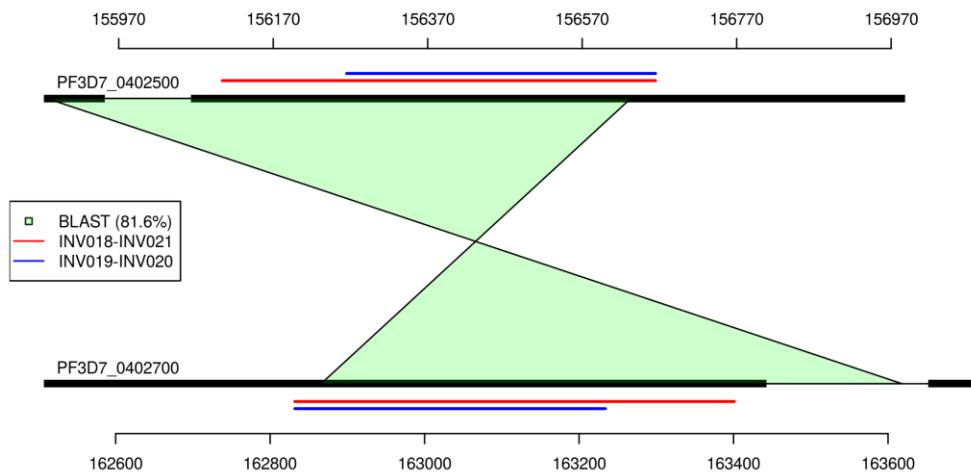
## Figure Legends

**Figure 1: Summary panel of inversions:** (A) Genomic positions of each inversion, (B) Distribution of locus type<sup>4</sup>, (C) Distribution of inversion sizes, (D) Distribution of identities per inversion.

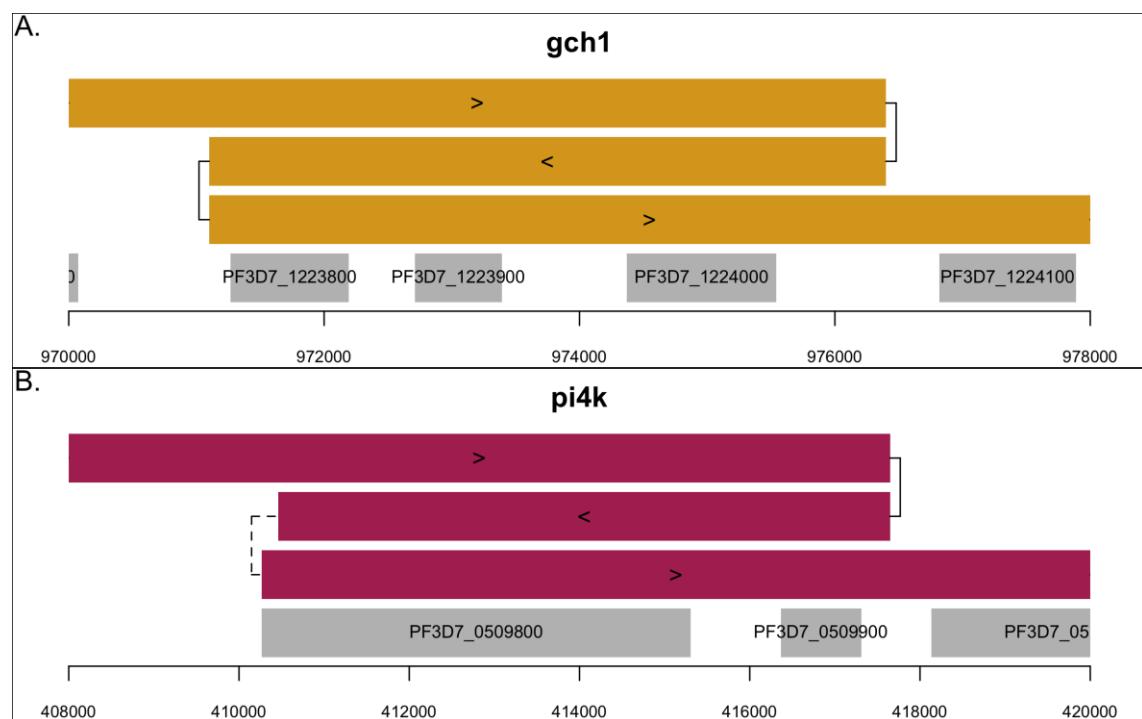


**Figure 2: Schematic representation of orthologous regions of PF3D7\_0402500 and PF3D7\_0402700, including predicted inversion pairs.**

<sup>4</sup> Inversion definitions: Tandem Inversion = Forward copy followed by reverse copy, Sandwich = Reverse copy between two forward copies, Intergenic = Outside a coding region, Historic Tandem Inversion = Previously characterised inversions present in all *P. falciparum* relative to other species, Highly Variable Region = Present within a known highly variable gene.



**Figure 3: Schematic candidate diagrams for inversions of: (A) *gch1* in Dd2, (B) *pik4* in GB4.** Dashed lines indicate a gap between contigs in the strain assembly.



**Table 1: Summary of inversions per sample.** \*long term cultures; PNG = Papua New Guinea.

Isolate	Location	Total Inversions	Intergenic Inversions	Highly Variable Inversions (rifin, var, stevor)	Notable Genic Inversions
7G8*	Brazil	8	2	3 (3, 0, 0)	RH2a/b
IT*	Brazil	18	6	8 (3, 2, 3)	RH2a/b, EF1a
HB3*	Honduras	14	9	2 (1, 1, 0)	RH2a/b, EF1a
NF54*	Africa	20	9	7 (5, 2, 0)	RH2a/b, EF1a
GA01	Gabon	20	10	7 (4, 2, 1)	RH2a/b, EF1a
GN01	Guinea	15	4	10 (5, 3, 2)	RH2a/b
GB4*	Ghana	18	10	4 (2, 1, 1)	RH2a/b, EF1a, PI4K
SN01	Senegal	7	1	6 (4, 2, 0)	-
CD01	Congo	20	9	6 (3, 3, 0)	RH2a/b, EF1a
KE01	Kenya	17	6	7 (4, 3, 0)	RH2a/b, EF1a
SD01	Sudan	14	8	4 (2, 2, 0)	RH2a/b, EF1a
Dd2*	Indochina	19	8	7 (5, 2, 0)	RH2a/b, EF1a, gch1
KH01	Cambodia	11	0	11 (10, 1, 0)	-
KH02	Cambodia	7	2	5 (2, 3, 0)	-
K1*	Thailand	19	5	11 (8, 3, 0)	RH2a/b, EF1a
T996*	Thailand	20	7	10 (8, 1, 1)	RH2a/b
D10*	PNG	13	2	11 (6, 5, 0)	EF1a

**Supplementary table 1: Full list of inversions.** HTI = Historic tandem inversion; HVR = Highly variable region; SI = Sandwich inversion; TI = tandem inversion; Int = intergenic. [large file]

## **Chapter 5:**

Novel genetic polymorphisms associated with severe malaria and  
under selective pressure in North-eastern Tanzania



**Registry**  
 T: +44(0)20 7299 4646  
 F: +44(0)20 7299 4656  
 E: [registry@lshtm.ac.uk](mailto:registry@lshtm.ac.uk)

## **RESEARCH PAPER COVER SHEET**

***PLEASE NOTE THAT A COVER SHEET MUST BE COMPLETED FOR EACH RESEARCH PAPER INCLUDED IN A THESIS.***

### **SECTION A – Student Details**

<b>Student</b>	Matt Ravenhall
<b>Principal Supervisor</b>	Taane Clark
<b>Thesis Title</b>	A bioinformatic analysis of malaria host and pathogen genomics

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

### **SECTION B – Paper already published**

Where was the work published?	PLOS Genetics		
When was the work published?	January 30th 2018		
If the work was published prior to registration for your research degree, give a brief rationale for its inclusion	n/a		
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?	n/a
Please list the paper's authors in the intended authorship order:	n/a
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 conducted the data cleaning and analysis having received the raw genotypes, produced all figures, and co-wrote the final manuscript under Taane Clark and Susana Campino's joint supervision.
--	---

**Student Signature:** \_\_\_\_\_

**Date:** \_\_\_\_\_

**Supervisor Signature:** \_\_\_\_\_

**Date:** \_\_\_\_\_

## RESEARCH ARTICLE

# Novel genetic polymorphisms associated with severe malaria and under selective pressure in North-eastern Tanzania

Matt Ravenhall<sup>1</sup>, Susana Campino<sup>1,2</sup>, Nuno Sepúlveda<sup>2,3</sup>, Alphaxard Manjurano<sup>4,5</sup>, Behzad Nadim<sup>4</sup>, George Mtove<sup>4</sup>, Hannah Wangai<sup>4</sup>, Caroline Maxwell<sup>4</sup>, Raimos Olomi<sup>4</sup>, Hugh Reyburn<sup>2,4</sup>, Christopher J. Drakeley<sup>2,4†</sup>, Eleanor M. Riley<sup>2,4‡</sup>, Taane G. Clark<sup>1,6§\*</sup>, in collaboration with MalariaGEN<sup>1</sup>



**1** Pathogen Molecular Biology Department, London School of Hygiene and Tropical Medicine, London, United Kingdom, **2** Department of Immunology and Infection, London School of Hygiene and Tropical Medicine, London, United Kingdom, **3** Centre for Statistics and Applications, University of Lisbon, Lisbon, Portugal, **4** Joint Malaria Programme, Kilimanjaro Christian Medical College, Moshi, Tanzania, **5** National Institute for Medical Research, Mwanza, Tanzania, **6** Department of Infectious Disease Epidemiology, Faculty of Epidemiology and Population Health, London School of Hygiene and Tropical Medicine, London, United Kingdom

‡ CJD, EMR, and TCC share joint senior authorship of this work.

\* MalariaGEN membership is listed in the Acknowledgments section.

§ Taane.Clark@lshtm.ac.uk

## OPEN ACCESS

**Citation:** Ravenhall M, Campino S, Sepúlveda N, Manjurano A, Nadim B, Mtove G, et al. (2018) Novel genetic polymorphisms associated with severe malaria and under selective pressure in North-eastern Tanzania. PLoS Genet 14(1): e1007172. <https://doi.org/10.1371/journal.pgen.1007172>

**Editor:** Scott M. Williams, Case Western Reserve University School of Medicine, UNITED STATES

**Received:** August 4, 2017

**Accepted:** December 29, 2017

**Published:** January 30, 2018

**Copyright:** © 2018 Ravenhall et al. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

**Data Availability Statement:** Due to patient confidentiality, data are available upon request from MalariaGEN (<https://www.ebi.ac.uk/ega/studies/EGAS00001000638>, <https://www.ebi.ac.uk/ega/studies/EGAS00001000637>, <https://www.ebi.ac.uk/ega/studies/EGAS00001000636>, and <https://www.ebi.ac.uk/ega/studies/EGAS00001001311>). Please see <https://www.malariagen.net/data/terms-use/human-gwas-data> for instructions on how to apply for data access.

## Abstract

Significant selection pressure has been exerted on the genomes of human populations exposed to *Plasmodium falciparum* infection, resulting in the acquisition of mechanisms of resistance against severe malarial disease. Many host genetic factors, including sickle cell trait, have been associated with reduced risk of developing severe malaria, but do not account for all of the observed phenotypic variation. Identification of novel inherited risk factors relies upon high-resolution genome-wide association studies (GWAS). We present findings of a GWAS of severe malaria performed in a Tanzanian population (n = 914, 15.2 million SNPs). Beyond the expected association with the sickle cell HbS variant, we identify protective associations within two interleukin receptors (*IL-23R* and *IL-12RB2*) and the kelch-like protein *KLHL3* (all  $P < 10^{-6}$ ), as well as near significant effects for Major Histocompatibility Complex (MHC) haplotypes. Complementary analyses, based on detecting extended haplotype homozygosity, identified *SYNJ2BP*, *GCLC* and MHC as potential loci under recent positive selection. Through whole genome sequencing of an independent Tanzanian cohort (parent-child trios n = 247), we confirm the allele frequencies of common polymorphisms underlying associations and selection, as well as the presence of multiple structural variants that could be in linkage with these SNPs. Imputation of structural variants in a region encompassing the glycophorin genes on chromosome 4, led to the characterisation of more than 50 rare variants, and individually no strong evidence of associations with severe malaria in our primary dataset ( $P > 0.3$ ). Our approach demonstrates the potential of a joint genotyping-sequencing strategy to identify as-yet unknown susceptibility loci in an African population with well-characterised malaria phenotypes. The regions encompassing

**Funding:** MR is funded by the Biotechnology and Biological Sciences Research Council (grant number BB/J014567/1). The MalariaGEN Project is supported by the Wellcome Trust (WT077383/Z/05/Z) and the Bill and Melinda Gates Foundation through The Foundation for the National Institutes of Health (FNIH, USA) (566) as part of the Grand Challenges in Global Health Initiative. The Resource Centre for Genomic Epidemiology of Malaria is supported by the Wellcome Trust (090770/Z/09/Z). This research was supported by the UK Medical Research Council (G0600718 and G0600230). The Wellcome Trust also provides core awards to the Wellcome Trust Centre for Human Genetics (090532/Z/09/Z) and the Wellcome Trust Sanger Institute (098051/Z/05/Z). TGC is supported by the Medical Research Council UK (Grant no. MR/K000551/1, MR/M01360X/1, MR/N010469/1, MC\_PC\_15103). SC is funded by the Medical Research Council UK (Grant no. MR/M01360X/1, MC\_PC\_15103). NS was funded by the Wellcome Trust grant number 091924 and Fundação para a Ciência e Tecnologia through the project UID/MAT/00006/2013. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

**Competing interests:** The authors have declared that no competing interests exist.

these loci are potential targets for the design of much needed interventions for preventing or treating malaria disease.

### Author summary

Malaria, caused by *Plasmodium falciparum* parasites, is a major cause of mortality and morbidity in endemic countries of sub-Saharan Africa, including Tanzania. Some gene mutations in the human genome, including sickle cell trait, have been associated with reduced risk of developing severe malaria, and have increased in frequency through natural selection over generations. However, new genetic mutations remain to be discovered, and recent advances in human genome research technologies such as genome-wide association studies (GWAS) and fine-scale molecular genotyping tools, are facilitating their identification. Here, we present findings of a GWAS of severe malaria performed in a well characterised Tanzanian population ( $n = 914$ ). We confirm the expected association with the sickle cell trait, but also identify new gene targets in immunological pathways, some under natural selection. Our approach demonstrates the potential of using GWAS to identify as-yet unknown susceptibility genes in endemic populations with well-characterised malaria phenotypes. The genetic mutations are likely to form potential targets for the design of much needed interventions for preventing or treating malaria disease.

### Introduction

Sub-Saharan Africa bears a disproportionately high share of the global *Plasmodium falciparum* malaria burden, with 90% of the estimated 212 million annual cases and 92% of 429,000 annual deaths, mostly in children under five years of age [1]. Whilst the majority of cases of *Plasmodium falciparum* infection are asymptomatic or cause only mild to moderate clinical symptoms, a subset of affected individuals present with severe manifestations such as severe malarial anaemia and cerebral malaria. Risk factors for severe malaria and its various clinical subtypes are poorly understood, although host and parasite genotype, age and immune status have all been established as playing a significant role in individual host susceptibility [2]. *Plasmodium falciparum* has also exerted significant selection pressure upon the human genome, as evidenced by the geographical concurrence of malaria parasite prevalence with sickle cell trait (HbAS) and other haemoglobinopathies, such as the thalassemias and glucose-6-phosphate dehydrogenase (G6PD) deficiency.

Recent studies, set in a region of high malaria transmission in north-eastern Tanzania, estimated that host genetic factors account for approximately 22% of the total variation in severe malaria risk [3], consistent with previous findings in a Kenyan family-based study [2]. Less than half of this variation can be explained by erythrocyte-associated polymorphisms [4], including *HbS* (sickle cell trait), alpha-thalassaemia, ABO blood group [5] and G6PD deficiency [4]. Novel polymorphisms in or around *USP38*, *FREM3* [3], glycophorins *gypA/B/E* [6, 7], *DDC* [8], *MARVELD3* and *ATP2B4* [9] account for additional variation but, in sum, are less protective than heterozygous carriage of *HbS* [3]. Moreover, the effects of some of these loci are subtype-, location-, or population-specific [3, 6, 7, 9], reinforcing the need for targeted genome-wide association studies (GWAS) in different African populations. Utilising such an approach with robust malaria phenotypes in parallel with whole genome sequencing of study

populations is crucial to unravelling host genetic factors that could lead to a greater understanding of protective immunity and development of new tools for disease prevention.

To identify novel loci associated with severe malaria in north-eastern Tanzania, we applied genome-wide association and haplotype-based selection methods to a case-control dataset with extensive phenotypic data for 914 participants and 15.2 million SNPs. In addition to the expected *HbS* (sickle cell) association, our analyses reveal multiple novel loci under association or selection. Association analysis highlighted significant SNP clusters within *IL-23R*, *IL-12RB2*, *LINC00944*, and *KLHL3* whilst lone SNP associations were also present within *TREML4* and *ZNF536*. Further, we reveal loci under recent positive selection including *GCLC* and loci within the Major Histocompatibility Complex (MHC). These analyses were supported by whole genome sequencing of an independent dataset consisting of 247 Tanzanian individuals within parent-child trios, which was used to confirm the allele frequencies of putative associations and determine if there are any linked common structural variants in chromosome regions encoding important polymorphisms.

## Results

### Phenotypic and genotypic data

All severe malaria cases (n = 449) and controls (n = 465) came from the Tanga region of North-Eastern Tanzania. Severe malaria cases presented with varying combinations of hyperlactataemia (57.0%), severe malarial anaemia (49.2%), respiratory distress (27.6%) and cerebral malaria (26.7%) (Table 1). Compared to controls, malaria cases were younger (*t* test P < 2.2 × 10<sup>-16</sup>) and marginally more likely to be male (Chi squared P = 0.012) (Table 1). DNA from all samples (n = 914)

**Table 1. Study participants.**

	Controls (n = 465)		Cases (n = 449)		Difference P-value
Age* (median, range)	2.8	0.9–10.9	1.7	0.2–10.0	<2.2 × 10 <sup>-16</sup>
Female	252	54.2%	205	45.7%	0.012
Ethnicity**					0.52
Mziguia	151	32.5%	146	32.5%	
Wasambara	142	30.5%	135	30.1%	
Wabondei	83	17.8%	86	19.2%	
Mmbena	26	5.6%	23	5.1%	
Mngoni	17	3.7%	18	4.0%	
Pare	11	2.4%	8	1.8%	
Mmakonde	11	2.4%	8	1.8%	
Mgogo	7	1.5%	8	1.8%	
Chagga	9	1.9%	7	1.6%	
Other	8	1.7%	10	2.2%	
Mixed Ethnicity***	150	32.3%	172	38.2%	0.065
Hyperlactatemia/acidosis	-	-	256	57.0%	-
Severe Malarial Anaemia	-	-	221	49.2%	-
Respiratory Distress	-	-	124	27.6%	-
Cerebral Malaria	-	-	120	26.7%	-

\* months

\*\* based on paternal ethnicity

\*\*\* if parental ethnicities were different

<https://doi.org/10.1371/journal.pgen.1007172.t001>

was genotyped on the Illumina Omni 2.5 million SNP chip, and imputed against the 1000 Genomes reference panel (Phase 3) [10] and a Tanzanian parent-child trio panel (see below), using Beagle 4.1 [11], leading to 15.2 high quality SNPs. These markers were complemented by 180 SNPs within malaria candidate genes, including *HBB* (encoding HbS) [3, 4, 5] on the same cases and controls. DNA from a validation cohort of 78 healthy parent and child trios and 13 independent individuals (“Trios dataset”, n = 247) were whole genome sequenced using Illumina HiSeq2500 technology. For the GWAS samples, a principal component analysis (PCA) using all genome-wide SNPs revealed a low degree of stratification by ethnicity and case-control status (S1 Fig) and potential cryptic relatedness due to familial clustering. A similar analysis revealed that GWAS and Trio sample clusters overlap, and there is some separation from the other 1000 Genome African populations, including Yoruba (Nigeria) and Luhya (Kenya) (S1 Fig).

### Association analysis

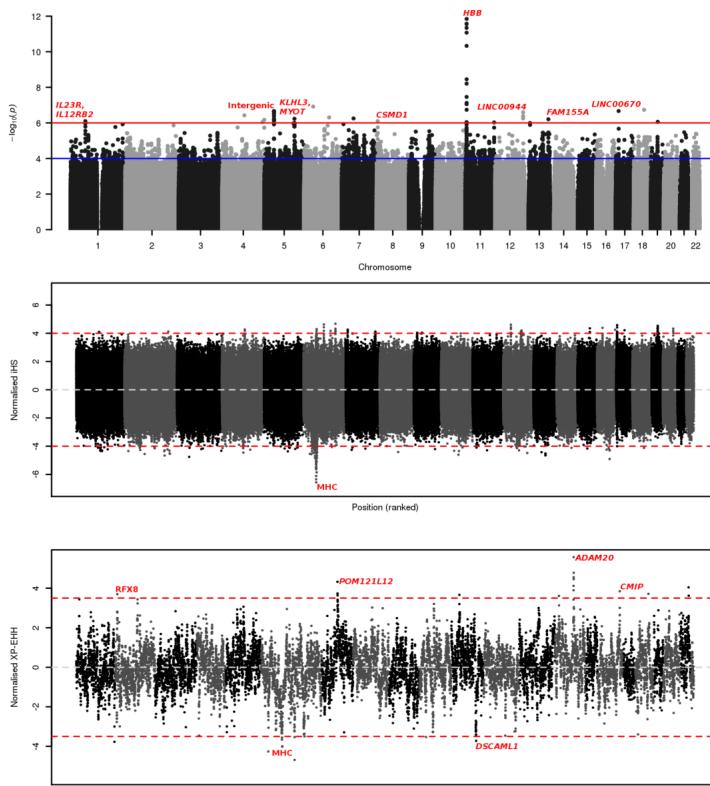
GWAS analysis was undertaken with EMMAX mixed model regression [12], controlling for age as a fixed effect and relatedness (represented in a kinship matrix) as a random effect to account for the cryptic population clustering. Separate models of association were fitted for each SNP (additive, heterozygous, dominant, recessive), with their respective genomic inflation factors all being close to one (see S1 Fig for the heterozygous results), consistent with reliable adjustment for stratification. A total of 53 SNPs (in 16 genomic regions) were identified with a significance level below our threshold ( $P < 1 \times 10^{-6}$ ) (Fig 1, Table 2, S1 Table). Relaxing the stringency would lead to 258 SNPs with a p-value below  $1 \times 10^{-5}$  and 2,322 below a threshold of  $1 \times 10^{-4}$ . As expected, the most significant association was with the sickle cell locus, rs334 ( $P = 8.59 \times 10^{-13}$ , heterozygous odds ratio = 0.07) (Table 2). Controlling for HbS status through a complementary conditional GWAS demonstrated our top associations as robust against linkage with rs334 (Table 2, S1 Table).

Novel associations of note also include SNPs within the *KLHL3-MYOT* region (13 SNPs, Min P =  $5.85 \times 10^{-7}$ , Additive OR = 0.590), the *IL23R-IL12RB2* region (7 SNPs, Min P =  $7.98 \times 10^{-7}$ , Recessive OR = 0.479), *FAM155A* (6 SNPs, Min P =  $6.24 \times 10^{-7}$ , Additive OR = 0.207), and *CSMD1* (5 SNPs, Min P =  $7.98 \times 10^{-7}$ , Additive OR = 4.795). (Fig 2). Three significant SNPs are also found within both *LINC00943/4* and *lncRNA AF146191.4-004*.

Lone SNP associations are present within proximity of *TREML4* ( $P = 1.21 \times 10^{-7}$ , Heterozygous OR = 4.087), zinc finger-containing *ZNF536* ( $P = 8.69 \times 10^{-7}$ , Recessive OR = 0.507), *C4orf17* ( $P = 3.75 \times 10^{-7}$ , Recessive OR = 0.289), and near *LINC00670* ( $P = 2.15 \times 10^{-7}$ , Additive OR = 3.867). And finally, three intergenic regions display clusters of significance, most notably a region within chromosome 5 (43,892,232–43,964,366bp; Min P =  $2.17 \times 10^{-7}$ , Heterozygous OR = 0.354), as well as regions within chromosome 7 and 11.

As expected, allele frequencies of the putative polymorphisms within the Trios dataset are generally equivalent to frequencies in our case and control groups, whilst there were some differences from the 1000 Genomes populations, including within the *HBB* locus (Table 2). Using the Trios dataset, we sought to identify structural variants that could confound the association analysis or be putative hits. We identified no structural variants within *HBB*, *IL-12RB2* or *LINC00943/4*, one deletion (2,904bp) within *IL23R*, and 152 deletions within *KLHL3* (63 distinct variants, all singletons except for one 1,325bp deletion in 91 individuals) (S2 Table). None of the common variants are in linkage disequilibrium with the putative GWAS hits, and eight putative regions had structural variants in the Tanzanian trios, but were absent in the 1000 Genomes populations (S2 Table).

Subtype specific association analyses were undertaken for those SNPs found to be significantly associated with severe malaria in the primary GWAS (Table 2). The majority of



**Fig 1. Genome-wide association and selection plots.** a) Case-Control SNP Associations. Red line indicates genome-wide significance at  $1 \times 10^{-6}$ , blue line indicates genome-wide suggestive significant at  $1 \times 10^{-4}$ ; b) Combined population iHS selection. Red lines indicate significance for absolute iHS scores of 4 or greater; c) Case-Control XP-EHH selection. Red lines indicate significance for absolute XP-EHH scores of 4 or greater.

<https://doi.org/10.1371/journal.pgen.1007172.g001>

significant associations are with the hyperlactataemia subtype, a phenotype that includes 57.0% of cases, with variants within *FAM155A*, and the *HBB* and *KLHL3/MYOT* regions exhibiting associations exceeding our  $1 \times 10^{-6}$  significance threshold. In contrast, variants within *IL-23R*, *IL-12RB2*, *CSMD1*, *ZNF536* and *TREM4* appear to be most significantly associated with severe malarial anaemia, who comprised 49.2% of cases.

### Candidate associations

Candidate SNPs identified in previous studies, with the same individuals, were included to provide appropriate context for novel findings. ABO blood group, *USP38*, *FREM3* and alpha-thalassemia have previously been putatively associated with severe malaria in a Tanzanian

Table 2. More significant SNP associations per locus ( $P < 1 \times 10^{-6}$ ).

SNP	Gene	n SNPs	Location	Minimum P	Conditional P	Subtype P
rs334	HbS (in <i>HBB</i> )	40	11:5248232	HET: 8.59E-13	-	HL: 1.81e-09
rs9296359	<i>TREML4</i>	1	6:41205690	HET: 1.21E-07	HET: 4.42e-07	SMA: 3.29e-07
rs149085856	Intergenic (LINC00670)	1	17:12399526	ADD: 2.15E-07	ADD: 1.06e-06	HL: 2.81e-07
rs113449872	Intergenic	20	5:43909343	HET: 2.17E-07	HET: 2.93e-07	SMA: 2.92e-05
rs11335470	<i>LINC00944</i>	3	12:127237620	HET: 2.52E-07	HET: 1.86e-06	HL: 9.04e-05
rs73832816	<i>C4orf17</i>	1	4:100429757	REC: 3.75E-07	REC: 9.48e-07	CM: 1.02e-06
rs17624383	Intergenic	3	7:53676837	ADD: 5.62E-07	ADD: 3.28e-06	RD: 4.61e-07
rs2967790	<i>KLHL3, MYOT</i>	13	5:137011761	ADD: 5.85E-07	ADD: 2.46e-06	HL: 8.65e-06
rs144312179	<i>FAM155A</i>	6	13:108228013	ADD: 6.24E-07	ADD: 2.92e-06	HL: 1.35e-06
rs114169033	AF1461914-004 (lncRNA)	3	4:190717704	ADD: 6.67E-07	ADD: 1.30e-06	RD: 5.62e-07
rs6682413	<i>IL23R, IL12RB2</i>	7	1:67731614	REC: 7.98E-07	REC: 1.03e-06	SMA: 1.23e-04
rs73505850	<i>CSMD1</i>	5	8:4754838	ADD: 7.98E-07	ADD: 1.42e-06	SMA: 1.20e-05
rs8109875	<i>ZNF536</i>	1	19:31069639	REC: 8.69E-07	REC: 3.57e-06	SMA: 2.80e-05
rs1878468	AC108142.1 (antisense)	1	4:182822332	HET: 8.98E-07	HET: 1.19e-06	HL: 8.10e-07
rs3133394	Intergenic	4	11:130417522	ADD: 9.41E-07	ADD: 1.08e-06	CM: 9.49e-06

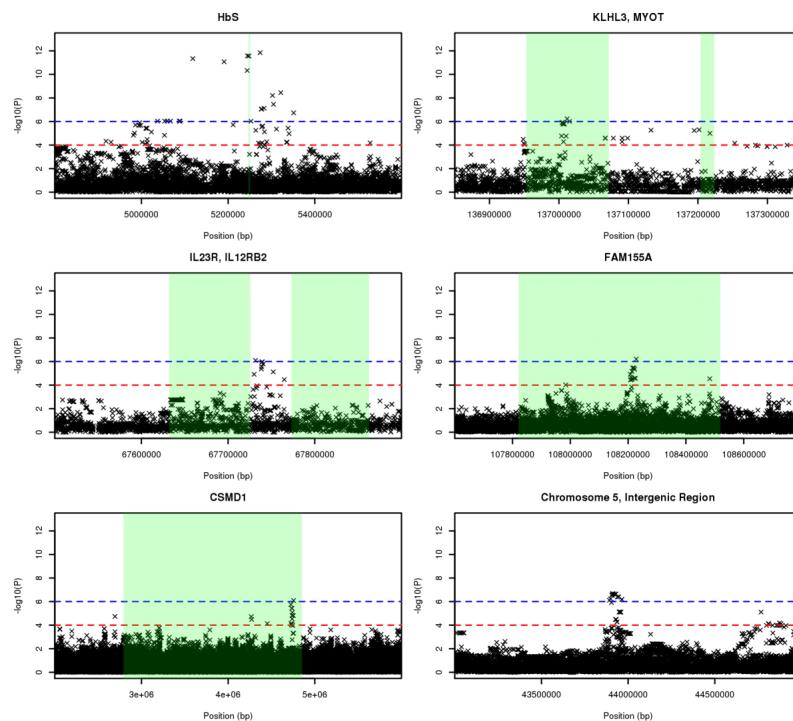
Allele models: ADD Additive, HET Heterozygous, DOM Dominant, REC Recessive. Subtype significances: HL Hyperlactatemia; SMA Severe Malaria Anæmia; RD Respiratory Distress; CM Cerebral Malaria. Locations correspond to the GRCh37 reference genome. Minimum P indicates the most significant P for SNPs in the locus within the case-control GWAS, whilst Conditional and Subtype Ps indicate the most significant P value for those SNPs when controlling for rs334 status, or considering the severe malarial subtypes.

<https://doi.org/10.1371/journal.pgen.1007172.t002>

population [3, 5], but these associations are no longer statistically significant ( $P > 10^{-4}$ ) at a more stringent GWAS significance threshold (S3 Table). We also performed targeted imputation of HLA haplotypes within the MHC, finding the most significant SNP to be rs1264362, which demonstrated a marginal association with hyperlactatemia (additive model  $P = 2.33 \times 10^{-5}$ ).

For the analysis of structural variation within candidate regions in the Trios dataset, we identified 28 distinct deletions within *FREM3*, of which all but one are present in only one individual, and six distinct deletions in *GYPB*, for which copy number variation has previously been identified [6]. Nine distinct variants were identified in *ABO*, including six duplications, one deletion, one insertion and one inversion. All such *ABO* variants are present in single individuals, though 18 individuals have a 23bp insertion. In contrast to a diversity of structural variation present within HLA and the wider MHC region, minor frequency variants were identified in *ATP2B4* (25 deletions across 25 samples), *MARVELD3* (five deletions across five samples), *HBA2* (3 deletions across three samples), and *HBA1* (one sample with one deletion). No structural variants were found in *HBB* or *USP38* (S2 Table).

We imputed structural variants within the wider region of human glycophorin genes (*gypA*, *gypB*, *gypE*) on chromosome four, using 55 distinct large polymorphisms identified in 59 individuals within our Trios dataset (S2 Table). The glycophorin region is structurally highly diverse, and specific individual variants are of low frequency (mean frequency: Case Control dataset = 0.098, Trios dataset = 0.022), consistent with observations in other African populations [7]. Whilst these large variants could be potentially protective against severe malaria, we identified no significant associations looking at each individually ( $P \geq 0.301$ ). Grouping these variants into forms based on genomic location and function may enhance signals within this region, but could also introduce experimenter bias. Further, there exists a multitude of potential variant combinations analysis of which, without specific hypotheses, could risk so-called 'P hacking'. A full and in-depth analysis of this region is required but beyond the scope of this study.



**Fig 2. Region plots for most significant SNP associations.** Green regions highlight genes of interest, as indicated by sub-plot headers. The red line indicates genome-wide significance at  $1 \times 10^{-6}$ , whilst the blue line indicates genome-wide suggestive significant at  $1 \times 10^{-4}$ .

<https://doi.org/10.1371/journal.pgen.1007172.g002>

### Evidence of positive selection

Two approaches were applied to identify regions under recent positive selection within the Tanzanian GWAS population as a whole (Integrated Haplotype Score, iHS) [13], or between the cases and controls (Cross-Population Extended Haplotype Homozygosity, XP-EHH) [14]. A common genome-wide absolute score threshold of 4 (equivalent to  $P = 6.3 \times 10^{-5}$ ) was established for both approaches. At this threshold iHS identified 244 loci, 116 (47.5%) within chromosome 6. Ninety-four of these significant signals are within the MHC, with three loci within *HLA-DOA* having an absolute score greater than 6. Other MHC genes with significant signals include the immunophilin *FKBP5* (35732–37931bp, 9 SNPs, iHS: 4.00–4.84), *SAMD3* (12490–13053bp, 6 SNPs, iHS: 4.00–4.68) and the exocytosis regulator *RIMS1* (72805811–72828559bp, 3 SNPs, iHS: 4.17–4.62) (S4 Table). Most notably, two regions within chromosome 17 (3496105–3689132bp, 6 SNPs, 8 genes including integrin *ITGAE*) and chromosome 19 (38743962–38900106bp, 14 SNPs, 10 genes including two transmembrane channels) represent regions with a high density of selection signals, akin to those within the MHC. Further signals of note include the transcription factor *ZFX3* (*ATBF1*) (chr. 16, 16326–73133bp, 3 SNPs, iHS:

4.27–4.91), *ABHD5* (chr. 3, 43794949bp, 1 SNP, iHS: 4.75), *DUSP19* & *NUP35* (chr. 2, 99180–18528, 3 SNPs, iHS: 4.30–4.66), surface tyrosine-kinase receptor *ERBB4* (chr. 2: iHS: 4.31–4.54), transcription-associated *RORC* (chr. 1, 151792842–151817543bp, 2 SNPs, iHS: 4.31–4.66).

No structural variation was identified in *ABHD5* or *DUSP19*, whilst variants were present but rare for the remaining iHS hits ([S2 Table](#)). In total, two deletions were identified in *RORC*, three in *ZFHX3*, *NUP35*, and *ITGAE*, one of which is an 86bp deletion found in seven individuals, and 14 deletions and a 31bp insertion in *RIMSL*. Particularly variable are *ERBB4* and *FKBP5* for which we identify 49 and 75 distinct variants respectively. *ERBB4* consists of 44 deletions, four insertions, and one inversion, whilst *FKBP5* consists of 70 deletions, three duplications, one insertion, and one inversion.

The between case-control XP-EHH approach identifies 10 significant SNPs across six genetic regions ([S5 Table](#)). Relative selection for the control population lies within three regions, including *POM121L12* (XP-EHH: 4.33), *SYNJ2BP*, *ADAM21*, *ADAM20* (XP-EHH: 4.12 to 5.57) and *ERG* (XP-EHH: 4.04), whilst three regions are under relative selection in the case population, including *MCUR1* (XP-EHH: -4.26), *GCLC* (XP-EHH: -4.69) and the MHC (XP-EHH: -4.02) ([S5 Table](#)). We identify no structural variants within *POM121L12*, *ADAM21*, or *ADAM20*, but a singleton 75bp deletion in *SYNJ2BP*, two deletions within *MCUR1*, three deletions in *GCLC*, and 20 distinct deletions within *ERG*, of which 8 individuals share a 106bp deletion and 6 share a 325bp deletion ([S2 Table](#)).

## Discussion

As expected, the most significant SNP association is the heterozygous protective *rs334* effect ( $P = 2.61 \times 10^{-13}$ ), with thirty-nine further SNPs within *HBB* also being significantly associated with resistance to severe disease. SNPs in other candidate genes, including *FREM3*, *GYPA*, *GYPB* and *USP38* [3], did not exceed a significance threshold of  $1 \times 10^{-6}$ , and their p-values were different (greater) to those previously published because of the use of the more conservative EMMAx mixed model regression [12]. Marginal evidence for a role of HLA association with severe malaria was also identified, and is broadly consistent with previous work in a West African population that demonstrated that carriers of HLA Class I Bw53 and HLA class II DRB1\*1302-DQB1\*0501 were protected against severe malaria [15]. Note that our targeted imputation of HLA utilised a Caucasian reference panel and may therefore overlook further true associations within the HLA locus. Further, we identified signals of positive selection within the MHC region, this being consistent with malaria as a driver of MHC polymorphism in the human population [16, 17].

Of the novel SNP associations identified here, two of the top candidates are located between the interleukin receptors *IL-23R* and *IL-12RB2*, a region that has been identified in GWAS of other inflammatory and immune-linked diseases [18]. *IL-12* and *IL-23* are related pro-inflammatory cytokines that share both the p40 subunit and the *IL-12R $\beta$ 1* receptor subunit. *IL-12* signals through a receptor comprising *IL-12R $\beta$ 1* and *IL-12R $\beta$ 2* and is a potent inducer of *IFN- $\gamma$*  which mediates both clearance of infection and immunopathology in infections with Plasmodium parasites. *IL-23* signalling (through its receptor, comprising *IL-12R $\beta$ 1* and *IL-23R*) promotes transcription of *RORC* which encodes *ROR $\gamma$* , a transcription factor involved in generation of *IL-17*. *RORC* was found to be under recent positive selection in our analysis, further supporting the importance of the pathway. Decreased *IL-12* levels have been associated with progression from uncomplicated malaria to severe disease, specifically an increased risk of severe malarial anaemia in children [19, 20]. Variants in *IL-12B* have been linked to *P. falciparum* parasite density and associated with protection against cerebral malaria in children

whilst, variants in the related *IL-12A* and *IL-12RB1* loci have been associated with protection against severe malarial anaemia among children in western Kenya [19]. Conversely, the *IL-23/IL-17* immune pathway has been implicated in the development of inflammatory reactions in children that develop severe malarial anaemia [21], in multi-organ dysfunction and acute renal failure in adult *P. falciparum* cases from India [22] and with the risk of cerebral malaria in Africa [23]. *IL-23R* haplotypes have also been associated with increased susceptibility to severe malarial anaemia in Kenya [24].

Three significantly associated SNPs are present within *LINC00944*, with one being 80bp from a known CTCF binding site [25]. Structurally, although the LINC00943/4 region is a known deletion site [25], we identified no such deletions within the region in our ‘Trios’ dataset. Broader functionality of this long intergenic non-coding RNA is unclear, given limited experimental characterisation, making it difficult to determine a role for these SNP variants.

A strong association peak was also identified within *KLHL3*, kelch-like protein 3, being a region known to contain an enhancer and various deletions [25]. Correspondingly, we identify 152 such deletions within our Trios reference panel, of which 62 distinct variants are present in only one individual and one 1,325bp deletion is present in 91 individuals. This frequent deletion is located within an open chromatin-containing region between 137,022,562 and 137,023,887bp. Mutations of *KLHL3* have previously been linked with hypertension and metabolic acidosis [26] suggesting that these novel SNP associations and deletions may prime individuals to have a greater risk for severe malarial acidosis (hyperlactataemia).

A number of the most significantly associated SNPs are present as lone, or paired, associations rather than “stacks”. This includes SNPs within or very near to *TREML4* and *ZNF536*. Whilst this may demonstrate false positive outliers, the existence of these SNPs and their minor frequencies are confirmed in our Trios reference panel.

The broad picture of whole population iHS selection is unsurprising, with the MHC region demonstrating the most striking evidence for recent selective sweeps. Our results are also consistent with a number of previously identified iHS signals, such as those for loci containing the alcohol dehydrogenase *ADH7*, cadherin *PCDH15*, synaptotagmin *SYT1*, the nociception receptor *TRPV1*, and the transmembrane protein *SPINT2* [27]. It should also be emphasised that our iHS signals reflect selection within our case-control dataset and therefore oversample, relative to a general Tanzanian population, for those signals associated with susceptibility to severe malaria.

Recent differential selection between the case and control groups, as determined by XP-EHH, identified very few significant signals. There is likely to be limited differential selection between subsets of a closely related population, despite malaria infection being a strong selector. We identified the MHC, *GCLC*, *MCUR1*, *POM121L12* and the *SYNJ2BP-ADAM21-ADAM20* region. The strongest of these signals covers *ADAM20* and *ADAM21*, both members of a larger family of disintegrins and metalloproteinases that are believed to be exclusively expressed in the testis [28]; this association might simply reflect differences in the gender ratio between the cases and controls, for which XP-EHH does not control. Selection for this region is more likely driven by a variant of *SYNJ2BP*, a Synaptojanin-2 binding protein with potential roles in membrane trafficking and signalling [29].

Our previous work has demonstrated that novel associations with potentially significant roles in malaria susceptibility remain to be uncovered [3], and here we show that an integrated approach that identifies signals of association, selection and structural variation can empower such studies. However, with only 914 individuals in this study, sample size is a notable limitation for interpretation. Initial approaches to account for this were pursued through robust contextualisation of novel variants within the secondary ‘Trios’ dataset, and the wider 1000 Genomes project. More generally, it remains vital that further validation, through larger scale

studies, be undertaken to better characterise the SNP and structural variants uncovered. This is particularly true for structural variation such as within *KLHL3*, which may impact gene expression and would therefore benefit from incorporation of transcriptomic data.

Distributions of human genetic variants with putative roles in *P. falciparum* malaria susceptibility are diverse. The HbS sickle cell polymorphism is present across most regions of sub-Saharan Africa but is known to have arisen multiple times leading to a number of distinct haplotypic backgrounds [30]. Similarly, other variants, such as G6PD polymorphism and glycoporphin structural variants vary both in frequency across populations and in their direction of association, leading in some cases to allelic heterogeneity that may be subtype specific. Many protective variants identified within our study, such as *IL-23R* and *KLHL3*, were found at similar frequencies within the 'Trios' dataset but differed from the global 1000 Genomes panel, and may therefore represent examples of Tanzanian- or regional-specific associations. Such variants are informative to our understanding of human-parasite interactions, yet risk being overlooked in inadequately designed studies. Ultimately, human GWAS in parallel with whole genome sequencing of host and parasites in large study populations across Africa will be crucial to unravelling host genetic and parasite interactions that could lead to novel malaria control measures such as vaccines.

## Methods

### Ethics statement

All DNA samples were collected and genotyped following signed and informed written consent from a parent or guardian. Ethics approval for all procedures was obtained from both LSHTM (#2087) and the Tanzanian National Institute of Medical Research (NIMR/HQ/R.8a/Vol.IX/392).

### Study participants and phenotypes

All participants were from the Tanga region of North-Eastern Tanzania, as described previously [3]. Briefly, severe malaria cases ( $n = 449$ ) were recruited in the Teule district hospital and surrounding villages in Muheza district, Tanga region, Tanzania between June 2006 and May 2007. The controls ( $n = 465$ ) were recruited, matched on ward of residence, ethnicity and age (given in months), during August 2008 from individuals without a recorded history of severe malaria [3]. Four severe malaria subtypes were identified within case individuals including hyperlactatemia (Blood lactate  $> 5 \text{ mmol/L}$ ,  $n = 256$ ), severe malarial anaemia (Hemocue Hb  $< 5\text{g/dL}$ ,  $n = 221$ ), respiratory distress ( $n = 124$ ) and cerebral malaria (Blantyre coma score  $< 5$ ,  $n = 120$ ) (Table 1). Parasite infection was initially assessed by rapid diagnostic test (HRP-2-Parascreen Pan/Pf) and confirmed by double read Giemsa-stained thick blood films.

A further 247 anonymously sampled individuals, consisting of 78 healthy parent and child trios (156 parents, 78 children, 13 singletons; 80 (32.4%) Chagga, 77 (31.2%) Pare, 90 (36.4%) Wasambaa), were collected between 2007 and 2008. These individuals are those that had no current illness or no history of malaria. The samples were collected from highland, medium and lowland villages near the Kilimanjaro, Pare and West Usambara mountains in the Tanga region of Tanzania. This is a region that experiences low to medium to high levels of malaria transmission. This dataset was used to confirm allele frequencies and identify candidate region structural variation within the general Tanzanian population, as well as to impute variants onto the case-control set.

### Sample genotyping, sequencing and imputation

DNA was extracted from processed blood samples, as described previously [3, 5]. The DNA was genotyped on the Illumina Omni 2.5 million SNP chip and SNP genotypes called by the MalariaGEN Resource Centre at the Sanger Institute and the Wellcome Trust Centre for Human Genetics, using previously described methods [6,7]. These data were complemented by Iplex genotyping assays that included 180 single nucleotide polymorphisms (SNP) across 50 loci on the same individuals [3]. 107 additional candidate SNPs, including the HbS SNP rs334, were included from previous candidate genotyping of the same case-control individuals; their collection having been described previously [3]. DNA for the individuals in the Trio dataset ( $n = 247$ ) was sequenced using Illumina HiSeq2500 technology at the Sanger Institute, and aligned to the GRCh37 build of the human genome [7]. The minimum genome-wide coverage across the samples was 22-fold. SNPs were called from the alignments using the standard samtools-bcftools pipeline [31]. This process led to 2,788,671 high quality SNPs with quality scores of at least 30 (1 error per 1000bp) and perfect trio-consistent genotype calls. Haplotypes were phased from genotypes using SHAPEIT ([www.shapeit.fr](http://www.shapeit.fr); default settings). Structural variants, including duplications, deletions, insertions and inversions, were identified within the secondary 'Trios' dataset for candidate regions using DELLY version v0.7.3 [32]. This software was applied using default settings, and its use in pipelines has been shown to reliably uncover structural variants from the 1000 Genomes Project, and validation experiments of randomly selected deletion loci show a high specificity [32]. Structural variants greater than 100,000 basepairs in length were removed to conservatively exclude false positives.

To increase genome-wide SNP resolution, our initial case-control dataset was imputed using a combined reference panel of the Phase 3 1000 Genomes project [10] and children within the trio dataset, using Beagle 4.1 [11]. This allowed for the inclusion of 13.5 million additional high quality SNPs, to a total of 15.2 million SNPs. A total of 621,019 SNPs were removed from the pre-imputation dataset due to evidence of: (i) deviations in genotypic frequencies from Hardy-Weinberg equilibrium (HWE) as assessed using a chi-square test ( $>0.0001$ ); (ii) high genotype call missingness ( $>10\%$ ); or (iii) low minor allele frequency ( $<0.01$ ). 51 individuals were removed due to: (i) genotypic missingness ( $>0.1$ ); (ii) abnormal PCA clustering or (iii) missing malaria phenotype data. 849,134 strand flips were identified with snpflip, with these being corrected pre-imputation with Plink v1.07. Raw hybridisation plots were manually verified for all top non-imputed GWAS associations, excluding rs334 for which the data was unavailable. Linkage disequilibrium between SNPs in close genomic distance was calculated using Plink v1.07 [33].

Targeted imputation was performed for HLA haplotypes within the major histocompatibility complex using 9,785 high quality SNPs within the region; for this we utilised SNP2HLA software (version 1.0.3) and the default Caucasian reference panel [34]. Association tests for this targeted analysis were performed through the pipeline described above. Similarly, 1,202 structural variants (698 deletions, 311 duplications, 19 insertions, 174 inversions) within chromosome four were imputed into our primary 'case-control' dataset using IMPUTE2 with default parameters, akin to standard SNP imputation. This approach allowed us to perform association analysis on those structural variants using EMMAX mixed model regression [12]. Trio parental SNP data was also used to provide additional context for our case-control SNPs within the wider Tanzanian population, as seen in [S1 Table](#).

### Association analysis

Case-Control association analysis of SNPs was undertaken with EMMAX mixed model regression [12], controlling for age as a fixed effect and relatedness (represented by a kinship matrix)

as a random effect (to reduce associations relating to familial clustering). Several genotypic models were implemented separately, including additive, heterozygous, dominant and recessive. Minimum P values from each model were utilised for top hit identification. Odds ratios were estimated with Plink v1.07 [33]. Our complementary conditional GWAS shared the pipeline for the main GWAS, but with HbS status added as an additional covariate. To evaluate the statistical potential of our GWAS study, we performed a retrospective power calculation (using <http://zzz.bwh.harvard.edu/gpc/cc2.html>). A study of 460 cases and 460 controls can detect odds ratios of at least 2 for a high risk allele minor allele frequency of 5% with a statistical power of 85% (and type I error of  $10^{-6}$ ). A significance threshold of  $10^{-6}$  was established using a permutation approach [35]. In particular, both the case-control status of the chromosomes were randomly permuted 10,000 times. From each of the 10,000 random experiments, we determined the maximum chi-square statistics (across the four genotypic tests) over all SNPs genotyped. We ordered these statistics and then calculated the 95 percentile. This was the estimate of the 0.05 significance level for the experiment performed, assuming inference is taken with respect to maximum chi-square statistic observed over all genotyped SNPs, and accounts for the linkage disequilibrium between SNPs and correlation between the results from applying the 4 genotypic tests.

### Selection analysis

Whole population Integrated Haplotype Scores (iHS) [13] and case-control Cross-Population Extended Haplotype Homozygosity (XP-EHH) [14] were calculated and normalised over the whole genome using selscan and norm [36]. Core SNPs with a minor allele frequency below 0.01 were excluded from this analysis. In this context, high iHS values indicate a whole population selection signal whilst positive XP-EHH values indicate relative selection within the control population and negative XP-EHH values indicate relative selection within the case population. We looked for structural variants in regions with SNP-based signals of positive selection, as it is possible that selection may actually be driven by structural variants (see [37] for an example).

### Supporting information

**S1 Fig. Population structure.** Visualisation of the first two principal components, by (a) case-control status and (b) father's ethnicity, highlights the existence of cryptic relatedness; (c) Principal component analysis reveals that the 'Trios' and primary 'Case-Control' participants overlap and are within the African cluster of the 1000 Genomes dataset; (d) Quantile-quantile plot for the observed and expected P values of the heterozygous model genome-wide association statistic.  
(PNG)

**S1 Table. Full list of significant SNP associations, including odds ratios and minor allele frequencies.**  
(DOCX)

**S2 Table. Structural variation identified within regions consisting of GWAS associations, known malaria candidates and sites under selection (iHS, XP-EHH).**  
(DOCX)

**S3 Table. Candidate SNP associations.**  
(DOCX)

**S4 Table. Regions under potential whole population positive selection (absolute iHS > 4).**  
(DOCX)

**S5 Table. Regions under potential differential selection between cases and controls (absolute XP-EHH > 4).**  
(DOCX)

### Acknowledgments

We thank the participants and Tanzanian communities who made this study possible, and the healthcare workers who assisted with this work. We would also like to acknowledge members of the MalariaGEN Resource Centre at the Wellcome Trust Sanger Institute and at the Wellcome Centre for Human Genetics at Oxford University who contributed to this study by carrying out the data production and analysis pipelines for genome-wide SNP genotype calling and sequence alignment: in particular, we thank Quang Si Le, Katja Kivinen, Jim Stalker, Anna Jeffreys, Kate Rowlands, Christina Hubbart, Eleanor Drury, Geraldine Clarke, Chris Spencer, Gavin Band, Dominic Kwiatkowski and Kirk Rockett, as well as members of the Sample Management and DNA pipeline teams at the Wellcome Trust Sanger Institute.

### Author Contributions

**Conceptualization:** Christopher J. Drakeley, Eleanor M. Riley, Taane G. Clark.

**Data curation:** Matt Ravenhall.

**Formal analysis:** Matt Ravenhall.

**Funding acquisition:** Taane G. Clark.

**Investigation:** Taane G. Clark.

**Project administration:** Taane G. Clark.

**Resources:** Nuno Sepúlveda, Alphaxard Manjurano, Behzad Nadjm, George Mtove, Hannah Wangai, Caroline Maxwell, Raimos Olomi, Hugh Reyburn, Christopher J. Drakeley, Eleanor M. Riley.

**Software:** Nuno Sepúlveda.

**Supervision:** Susana Campino, Taane G. Clark.

**Writing – original draft:** Matt Ravenhall, Susana Campino, Taane G. Clark.

**Writing – review & editing:** Matt Ravenhall, Nuno Sepúlveda, Alphaxard Manjurano, Behzad Nadjm, George Mtove, Hannah Wangai, Caroline Maxwell, Raimos Olomi, Hugh Reyburn, Christopher J. Drakeley, Eleanor M. Riley, Taane G. Clark.

### References

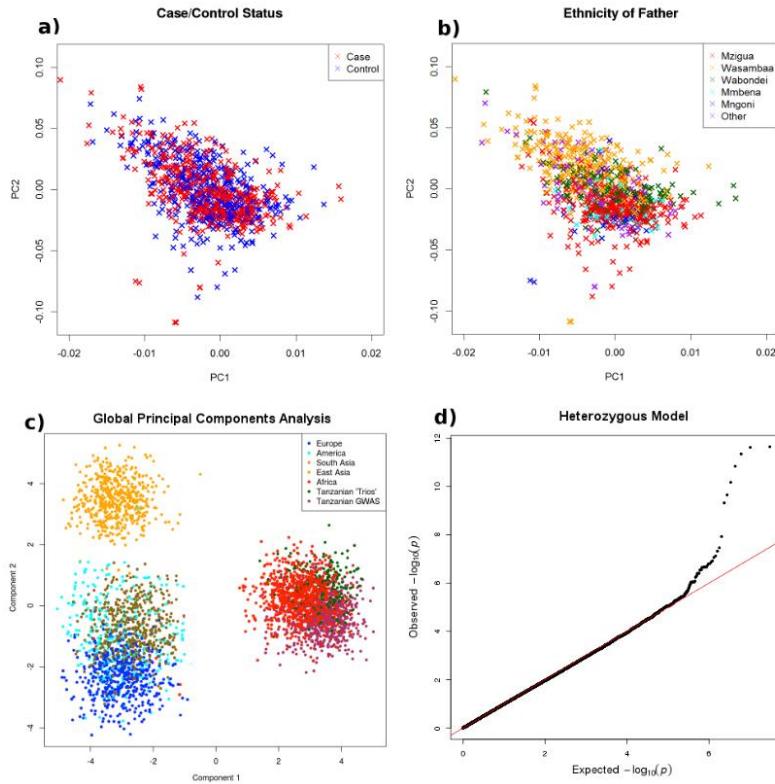
- WHO. (2015). World Malaria Report 2015. WHO Press.
- MacKinnon MJ, Mwangi TW, Snow RW, Marsh K, Williams TN. Heritability of Malaria in Africa. PLoS Medicine 2005; 2(12): e340. <https://doi.org/10.1371/journal.pmed.0020340> PMID: 16259530
- Manjurano A, Sepúlveda N, Nadjm B, Mtove G, Wangai H, Maxwell C, et al. USP38, FREM3, SDC1, DDC, and LOC727982 Gene Polymorphisms and Differential Susceptibility to Severe Malaria in Tanzania. J Infect Dis. 2015; 212:1129–39. <https://doi.org/10.1093/infdis/jiv192> PMID: 25805752
- Manjurano A, Sepulveda N, Nadjm B, Mtove G, Wangai H, Maxwell C, et al. African glucose-6-phosphate dehydrogenase alleles associated with protection from severe malaria in heterozygous females

- in Tanzania. PLoS Genet. 2015; 11(2):e1004960. <https://doi.org/10.1371/journal.pgen.1004960> PMID: 25671784
- 5. Manjurano A, Clark TG, Nadjm B, Mtove G, Wangai H, Sepulveda N, et al. Candidate human genetic polymorphisms and severe malaria in a Tanzanian population. PLoS One 2012; 7(10):e47463. <https://doi.org/10.1371/journal.pone.0047463> PMID: 23144702
  - 6. Malaria Genomic Epidemiology Network. A novel locus of resistance to severe malaria in a region of ancient balancing selection. Nature 2015; 526, 253–257. <https://doi.org/10.1038/nature15390> PMID: 26416757
  - 7. Leffler EM, Band G, Busby GBJ, Kivinen K, Le QS, Clarke GM, et al. Resistance to malaria through structural variation of red blood cell invasion receptors. Science 2017; 10.1126.
  - 8. Jallow M, Teo YY, Small KS, Rockett KA, Deloukas P, Clark TG, et al. Genome-wide and fine-resolution association analysis of malaria in West Africa. Nat Genet 2009; 41(6):657–65. <https://doi.org/10.1038/ng.388> PMID: 19465909
  - 9. Timmann C, Thye T, Vens M, Evans J, May J, Ehmen C, et al. Genome-wide association study indicates two novel resistance loci for severe malaria. Nature 2012; 489, 443–446. <https://doi.org/10.1038/nature11334> PMID: 22895189
  - 10. The 1000 Genomes Project Consortium. A global reference for human genetic variation. Nature 2015; 526, 68–74. <https://doi.org/10.1038/nature15393> PMID: 26432245
  - 11. Browning SR and Browning BL. Genotype imputation with millions of reference samples. Am J Hum Genet 2016; 98:116–126. <https://doi.org/10.1016/j.ajhg.2015.11.020> PMID: 26748515
  - 12. Kang HM, Sui JH, Service SK, Zaitlen NA, Kong S, Freimer NB et al. Variance component model to account for sample structure in genome-wide association studies. Nat Genetics 2010; 42:348–54. <https://doi.org/10.1038/ng.548> PMID: 20208533
  - 13. Voight BF, Kudaravalli S, Wen X, Pritchard JK. A Map of Recent Positive Selection in the Human Genome. PLoS Biology 2006; (3): e72. <https://doi.org/10.1371/journal.pbio.0040072> PMID: 16494531
  - 14. Sabeti PC, Varilly P, Fry B, Lohmueller J, Hostetter E, Cotsapas C, et al. Genome-wide detection and characterization of positive selection in human populations. Nature. 449, 913–918. <https://doi.org/10.1038/nature06250> PMID: 17943131
  - 15. Hill AV, Allsopp CE, Kwiatkowski D, Anstey NM, Twumasi P, Rowe PA et al. Common west African HLA antigens are associated with protection from severe malaria. Nature 1991; 15; 352(6336):595–600. <https://doi.org/10.1038/352595a0> PMID: 1865923
  - 16. Garamszegi LZ. Global distribution of malaria-resistant MHC-HLA alleles: the number and frequencies of alleles and malaria risk. Malar J. 2014; 13:349. <https://doi.org/10.1186/1475-2875-13-349> PMID: 25187124
  - 17. Leffler E, Gao Z, Pfeifer S, Segurel L, Auton A, Venn O et al. Multiple Instances of Ancient Balancing Selection Shared Between Humans and Chimpanzees. Science 2013; 339:1578–1582. <https://doi.org/10.1126/science.1234070> PMID: 23413192
  - 18. Mizuki N, Meguro A, Ota M, Ohno S, Shiota T, Kawagoe T, et al. Genome-wide association studies identify IL23R-IL12RB2 and IL10 as Behcet's disease susceptibility loci. Nat Genet. 2010; 42(8):703–6. <https://doi.org/10.1038/ng.624> PMID: 20622879
  - 19. Zhang L, Prather D, Vandend Eng J, Crawford S, Kariuki S, ter Kuile F, et al. Polymorphisms in genes of interleukin 12 and its receptors and their association with protection against severe malarial anaemia in children in western Kenya. Malar J. 2010; 9:87. <https://doi.org/10.1186/1475-2875-9-87> PMID: 20350312
  - 20. Raballah E, Kempaiah P, Karim Z, Orinda GO, Otieno MF, Perkins DJ, Ong'echa JM, CD4 T-cell expression of IFN- $\gamma$  and IL-17 in pediatric malarial anemia. PLoS One. 2017; 12(4):e0175864. <https://doi.org/10.1371/journal.pone.0175864> PMID: 28426727
  - 21. Oyegue-Liabagui SL, Bouopda-Tuedom AG, Kouma LC, Maghendji-Nzondo S, Nzougue H, Tchitoula-Makaya N, et al. Pro- and anti-inflammatory cytokines in children with malaria in Franceville, Gabon. Am J Clin Exp Immunol. 2017; 6(2):9–20. PMID: 28337387
  - 22. Herbert F, Tchitchek N, Bansal D, Jacques J, Pathak S, Bécavin C, et al. Evidence of IL-17, IP-10, and IL-10 involvement in multiple-organ dysfunction and IL-17 pathway in acute renal failure associated to Plasmodium falciparum malaria. Journal of Translational Medicine 2015; 13:369 <https://doi.org/10.1186/s12967-015-0731-6> PMID: 26602091
  - 23. Marquet S, Conte I, Pouliougou B, Argiro L, Cabantous S, Dessein H, et al. The IL17F and IL17RA Genetic Variants Increase Risk of Cerebral Malaria in Two African Populations. Infect Immun. 2015; 84(2):590–7. <https://doi.org/10.1128/IAI.00671-15> PMID: 26667835
  - 24. Munde EO, Raballah E, Okeyo WA, Ong'echa JM, Perkins DJ, Ouma C. Haplotype of non-synonymous mutations within IL-23R is associated with susceptibility to severe malaria anemia in a P. falciparum

- holoendemic transmission area of Kenya. *BMC Infect Dis.* 2017; 17(1):291. <https://doi.org/10.1186/s12879-017-2404-y> PMID: 28427357
- 25. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* 2012; 6: 489(7414):57–74. <https://doi.org/10.1038/nature1247> PMID: 22955616
  - 26. Boyden LM, Choi M, Choate KA, Nelson-Williams CJ, Farhi A, Toka HR et al. Mutations in kelch-like 3 and culin 3 cause hypertension and electrolyte abnormalities. *Nature* 2012; 482:98–102. <https://doi.org/10.1038/nature10814> PMID: 22266938
  - 27. Grossman SR, Andersen KG, Shlyakhter I, Tabrizi S, Winnicki S, Yen A, et al. Identifying recent adaptations in large-scale genomic data. *Cell.* 2013; 152(4):703–13. <https://doi.org/10.1016/j.cell.2013.01.035> PMID: 23415221
  - 28. Hooft van Huijsduijnen R. ADAM 20 and 21: two novel human testis-specific membrane metalloproteases with similarity to fertilin-alpha. *Gene* 1998; 206(2):273–82. PMID: 9469942
  - 29. Nemoto Y, Arribas M, Haffner C and DeCamilli P. Synaptotagmin 2, a Novel Synaptotagmin Isoform with a Distinct Targeting Domain and Expression Pattern. *J Biol Chem* 1997; 272:30817–30821.
  - 30. Serjeant GR. The Natural History of Sickle Cell Disease. *Cold Spring Harbor Perspectives in Medicine* 2013. <https://doi.org/10.1101/cshperspect.a011783> PMID: 23813607
  - 31. Li H, Handsaker B, Wysoker A, Fennell T, Ruia J, Homer N, et al. The Sequence Alignment/Map Format and SAMtools. *Bioinformatics*. 2009; 25(16):2078–9. <https://doi.org/10.1093/bioinformatics/btp352> PMID: 19505943
  - 32. Rausch T, Zichner T, Schlattl A, Stuetz AM, Benes V, Korbel JO. DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics* 2012; 28: 333–339.
  - 33. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D et al. PLINK: A toolset for whole-genome association and population-based linkage analysis. *Am J Hum Genet* 2007;
  - 34. Jia X, Han B, Onengut-Gumuscu S, Chen W-M, Concannon PJ, Rich SS, et al. Imputing Amino Acid Polymorphisms in Human Leukocyte Antigenes. *PLoS One* 2013; 8(6):e64683. <https://doi.org/10.1371/journal.pone.0064683> PMID: 23762245
  - 35. Sousa I, Clark TG, Holt R, Pagnamenta AT, Mulder EJ, Minderaa RB, et al. Polymorphisms in leucine-rich repeat genes are associated with autism spectrum disorder susceptibility in populations of European ancestry. *Mol Autism*. 2010; 1(1):7. <https://doi.org/10.1186/2040-2392-1-7> PMID: 20678249
  - 36. Szpiech ZA, Hernandez RD. selscan: An Efficient Multithreaded Program to Perform EHH-Based Scans for Positive Selection. *Molec Biol Evol* 2014; 31(10):2824–7. <https://doi.org/10.1093/molbev/msu211> PMID: 25015648
  - 37. Ravenhall M, Benavente ED, Mipando M, Jensen AT, Sutherland CJ, Roper C, et al. Characterizing the impact of sustained sulfadoxine/pyrimethamine use upon the *Plasmodium falciparum* population in Malawi. *Malar J.* 2016 Nov 29; 15(1):575. <https://doi.org/10.1186/s12936-016-1634-6> PMID: 27899115

## Supplementary Information

**Supplementary Figure 1:** Population Structure.



**Supplementary Table 1:** Full list of significant SNP associations, including odds ratios and minor allele frequencies. [large file]

**Supplementary Table 2:** Structural variation identified within regions consisting of GWAS associations, known malaria candidates and sites under selection (iHS, XP-EHH).

SNP	Gene	Location	Total	Del.	Dup.	Ins.	Inv.	Presence in 1000 Genomes populations
<i>IL23R and IL12RB2</i>	1: 67,632,083-67,862,583	GWAS	1 (1)	1 (1)	-	-	-	1 deletion; 1 total case (1 East Asian)
<i>RYR2</i>	1: 237,205,505-237,997,288	GWAS	49 (51)	9 (11)	-	-	40 (41)	8 deletions; 23 total cases (14 Africans, 6 East Asians, 2 Americans, 1 European)
<i>MAP1B</i>	5: 71,403,061-71,505,395	GWAS	2 (9)	2 (9)	-	-	-	-
<i>KLHL3</i>	5: 136,953,189-137,071,779	GWAS	63 (153)	63 (153)	-	-	-	1 deletion; 15 total cases (3 South Asians, 9 Europeans, 3 Americans)
<i>HBB</i>	11: 5,246,694-5,250,625	GWAS/ Candidate	-	-	-	-	-	4 deletions; 14 total cases (3 East Asians, 6 African)
<i>LINC00943/4</i>	12: 127,214,333-127,256,957	GWAS	-	-	-	-	-	2 deletions; 3 total cases (1 South Asian, 1 African, 1 East Asian)
<i>LIPG</i>	18: 47,087,069-47,119,272	GWAS	13 (13)	13 (13)	-	-	-	-
<i>ZNF536</i>	19: 30,719,197-31,204,445	GWAS	2 (3)	2 (3)	-	-	-	4 deletions, 1 ALU deletions; 200 total cases (117 Europeans, 44 Americans, 32 South Asians, 6 Africans, 1 East Asian)
<i>ATP2B4</i>	1: 203,595,689-203,713,209	Candidate	25 (25)	25 (25)	-	-	-	1 deletion; 857 total cases (246 South Asian, 194 East Asian, 177 African, 159 European, 81 American)
<i>USP38</i>	4: 144,106,070-144,144,983	Candidate	-	-	-	-	-	-
<i>FREM3</i>	4: 144,498,455-144,621,828	Candidate	28 (29)	28 (29)	-	-	-	1 deletion; 9 total cases (4 East Asians, 5 Europeans)

<i>GYPE, GYPB, GYPA</i>	4:144,792,020-145,061,904	Candidate	55 (59)	31 (35)	24 (24)	-	-	8 deletions, 2 duplications; 133 total cases (102 Africans, 9 South Asians, 7 Americans, 14 East Asians, 1 European)
<i>Major Histocompatibility Complex (including HLA)</i>	6: 28,477,796-33,448,353	Candidate/iHS/XP-EHH	2333 (3337)	1711 (2552)	1711 (2552)	1711 (2552)	1711 (2552)	101 deletions, 13 duplications, 8 CNVs, 6 ALU deletions; 60,948 total cases (all samples)
<i>ABO</i>	9: 136,125,788-136,150,617	Candidate	9 (21)	1 (1)	6 (6)	1 (13)	1 (1)	1 deletion; 20 total cases (17 Africans, 3 American)
<i>MARVELD3</i>	16: 71,660,064-71,676,017	Candidate	5 (5)	5 (5)	-	-	-	1 deletion; 4 total cases (3 Africans, 1 American)
<i>HBA2</i>	16: 222,846-223,709	Candidate	3 (3)	3 (3)	-	-	-	2 deletions; 274 total cases (192 Africans, 23 East Asians, 23 South Asians, 17 Americans, 4 Europeans)
<i>HBA1</i>	16: 226,679-227,521	Candidate	1 (1)	1 (1)	-	-	-	3 deletions; 297 total cases (192 Africans, 25 East Asians, 23 South Asians, 17 American, 4 Europeans)
<i>RORC, C2CD4D, THEM5</i>	1: 151,792,842-151,817,543	iHS	2 (2)	2 (2)	-	-	-	-
<i>DUSP19, NUP35</i>	2: 183,699,180-185,281,789	iHS	139 (280)	101 (213)	-	-	38 (67)	27 deletions, 5 duplications, 1 CNV, 1 inversion; 1,302 total cases (346 Africans, 209 Europeans, 202 East Asians, 185 Americans, 139 South Asians)
<i>ERBB4</i>	2: 212,380,286-213,576,272	iHS	65 (236)	59 (207)	-	5 (28)	1 (1)	22 deletions, 7 duplications, 2 CNVs, 1 ALU deletion, 1 inversion, 1 insertions; 4,165 total cases (all samples)
<i>MCUR1</i>	6: 13,786,789-13,814,800	XP-EHH	2 (2)	2 (2)	-	-	-	1 deletions; 1 total case (1 African)
<i>PPARD, MKRNP2, FANCE, TEAD3, RPL10A, TULP1, FKBP5, ARMC12</i>	6: 35,337,931-35,732,137	iHS	101 (179)	94 (170)	3 (3)	3 (5)	1 (1)	6 duplications, 4 deletions, 1 CNV; 2,806 total cases (629 Africans, 473 East Asians, 451 Europeans, 443 South Asians, 332 Americans)
<i>GCLC</i>	6: 53,362,139-53,481,768	XP-EHH	3 (4)	3 (4)	-	-	-	2 deletions, 3 total cases (1 African, 2 Europeans)
<i>RIMSI</i>	6: 72,805,811-72,828,559	iHS	1 (1)	1 (1)	-	-	-	1 duplication, 1 CNV; 35 total cases (16 Europeans, 10 Americans, 6 South Asians, 2 Africans, 1 East Asian)
<i>POM121L12</i>	7: 53,103,349-53,104,617	XP-EHH	-	-	-	-	-	-
<i>SYNJ2BP, ADAM21, ADAM20</i>	14: 70,838,148-71,001,732	XP-EHH	21 (21)	21 (21)	-	-	-	1 deletion; 1 total case (1 East Asian)
<i>ZFHX3</i>	16: 72,916,326-73,133,159	iHS	3 (3)	3 (3)	-	-	-	3 deletions; 5 total cases (2 Europeans, 1 South Asian, 1 African, 1 East Asian)
<i>ITGAE</i>	17: 3,632,836-3,689,132	iHS	3 (9)	3 (9)	-	-	-	-
<i>ERG, ETS2</i>	21: 39,751,949-40,196,879	XP-EHH	23 (41)	21 (38)	-	2 (3)	-	7 deletions, 1 ALU deletion, 1 duplication; 2,727 total cases (457 Europeans, 412 Africans, 379 South Asians, 296 Americans, 220 East Asians)

**Supplementary Table 3:** Candidate SNP associations.

SNP ID	Gene	Location	Min. P	Model	Case MAFs	Control MAFs	Trios MAFs	Kenya MAFs	Nigeria MAFs	Global MAFs
<i>rs334</i>	<i>HBB</i>	11:5248232	2.61 $\times 10^{-13}$ *	Heterozygous	0.020	0.080	0.026	0.101	0.139	0.027
<i>rs4951074 rs10900585 rs55868763 rs1541255</i>	<i>ATP2B4</i>	1:203660781 1:203654024 1:203652140 1:203652141	5.19 $\times 10^{-1}$ 2.07 $\times 10^{-1}$ 4.20 $\times 10^{-1}$ 4.20 $\times 10^{-1}$	Dominant Additive Dominant Dominant	0.323 0.338 0.330 0.330	0.340 0.375 0.354 0.354	0.285 0.327 0.298 0.297	0.374 0.414 0.394 0.394	0.421 0.481 0.417 0.417	0.152 0.172 0.150 0.150
<i>rs149914432 rs186790584 rs186873296</i>	<i>FREM3</i>	4:144666678 4:144680140 4:144702474	1.73 $\times 10^{-2}$ 1.74 $\times 10^{-2}$ 2.28 $\times 10^{-2}$	Recessive	0.032 0.031 0.030	0.058 0.054 0.053	0.024 0.024 0.024	0.015 0.005 0.005	0.005 0.005 0.005	0.002 0.001 0.001
<i>rs2466246 rs28459062</i>	<i>USP38</i>	4:143971242 4:144039139	2.21 $\times 10^{-4}$ ; 4.66 $\times 10^{-2}$	Additive Recessive	0.316 0.150	0.229 0.184	0.285 0.126	0.258 0.207	0.204 0.213	0.438 0.092
<i>Deletion</i>	<i>GYPE, B, A</i>	4:144801719-145041744	3.01 $\times 10^{-1}$	Additive	0.001	0.000	0.004	-	-	-
<i>rs1264362 rs2523589 HLA_B_07</i>	<i>HLA</i>	6:30776590 6:31327334 6:31431272	1.90 $\times 10^{-3}$ 9.33 $\times 10^{-3}$ 6.49 $\times 10^{-5}$	Additive Dominant Dominant	0.200 0.394 0.394	0.246 0.483 0.484	- -	0.131 0.288 -	0.032 0.301 -	0.153 0.356 -
<i>rs8176746 rs7853989 rs1053878 rs8176719</i>	<i>ABO</i>	9:136131322 9:136131592 9:136131651 9:136132908	1.89 $\times 10^{-1}$ 2.27 $\times 10^{-1}$ 3.98 $\times 10^{-1}$ 8.69 $\times 10^{-2}$	Dominant Dominant Heterozygous Additive	0.164 0.166 0.261 0.326	0.149 0.155 0.267 0.294	0.156 0.158 0.292 0.311	0.162 0.172 0.202 0.288	0.171 0.171 0.231 0.231	0.153 0.164 0.133 0.344
<i>rs2334880</i>	<i>MARVELD3</i>	16:71653637	7.09 $\times 10^{-1}$	Dominant	0.446	0.449	0.441	0.470	0.407	0.157
<i><math>\alpha</math>-thalassemia deletion</i>	<i>HBA1, HBA2</i>	-	1.80 $\times 10^{-2}$	Recessive	0.264	0.318	-	-	-	-

**Supplementary Table 4:** Regions under potential whole population positive selection

(absolute iHS > 4).

Chromosome	Location	No. of SNPs	Gene
1	101851482	1	RP11-157N3.1 (lincRNA)
1	103755600	1	Intergenic
1	109070167	1	Intergenic
1	114630097-114675076	4	SYT6
1	116717522	1	Intergenic

1	151792842-151817543	2	<i>RORC, C2CD4D, THEM5</i>
1	161968072	1	<i>OLFML2B</i>
1	175850011	1	Intergenic
2	13134854	1	AC064875.2
2	56002607-57936992	3	<i>EFEMP1, CCDC85A, AC007743.1</i>
2	76944275	1	Intergenic
2	137173396-137872490	2	<i>THSD7B</i>
2	183699180-185281789	3	<i>DUSP19, NUP35</i>
2	202847242	1	Intergenic
2	207067503	1	<i>GPR1</i>
2	212380286-213576272	3	<i>ERBB4</i>
3	43794949	1	Intergenic
3	45606651	1	<i>LIMD1</i>
3	105695408	1	Intergenic
3	112913318	1	Intergenic
3	194538730	1	Intergenic
4	4275260	1	<i>LYAR</i>
4	99541944-99548762	2	<i>TSPAN5</i>
4	100334943	1	<i>ADH7</i>
4	107940588-107943491	2	<i>DKK2</i>
4	135796170	1	Intergenic
5	79086960	1	<i>CMYA5</i>
5	99269809	1	Intergenic
5	114127581	1	Intergenic
5	118671874	1	<i>TNFAIP8</i>
5	120965514	1	Intergenic
5	147289856	1	Intergenic
5	156626337	1	<i>ITK</i>
6	21233412	1	<i>CDKAL1</i>
6	25411435	1	<i>LRRC16A</i>
6	27247668-27396321	3	<i>POM121L2, VNIR10P, ZNF204P, ZNF391, MCFD2P1</i>
6	29937493-33853641	94	Major Histocompatibility Complex
6	35337931-35732137	9	<i>PPARD, MKRNP2, FANCE, TEAD3, RPL10A, TULP1, FKBP5, ARMC12</i>
6	72805811-72828559	3	<i>RIMS1</i>
6	106410424	1	Intergenic
6	111924913	1	<i>TRAF3IP2</i>
6	130512490-130537430	6	<i>SAMD3</i>
7	8240341-8243193	3	<i>ICA1</i>
7	20123972	1	AC005062.2
7	22161810	1	<i>RAPGEF5</i>
7	89333692	1	Intergenic
7	141072134-141085654	3	<i>TMEM178B</i>
8	72534277	1	Intergenic
9	8710098	1	<i>PTPRD</i>
9	24423134	1	Intergenic
9	111621283	1	Intergenic
10	56913475	1	<i>PCDH15</i>
10	76833088	1	Intergenic
10	79178467	1	<i>KCNMA1</i>
10	94841988	3	Intergenic
11	15169639-15177816	3	<i>INSC</i>
11	73714650	1	<i>UCP3</i>
12	28214312-28237731	4	Intergenic
12	29659037	1	<i>TMT1</i>
12	58840232	1	Intergenic
12	62396765	1	<i>FAM19A2</i>
12	70951978	1	<i>PTPRB</i>
12	79314798-79741443	3	<i>SYT1</i>
12	83061803-83101314	3	<i>TMT2C</i>
12	96544302	1	Intergenic
12	102331085	1	<i>DRAM1</i>
12	108703455	1	<i>CMKLRI</i>
13	48726060	3	Intergenic
13	69768976-69772154	1	Intergenic
14	81127849	1	<i>CEP128</i>
15	64185344	1	Intergenic
15	77282884-77296134	3	<i>PSTPIP1</i>
16	22943188	1	Intergenic
16	57009165	1	<i>CETP</i>
16	65902516	1	Intergenic
16	72916326-73133159	3	<i>ZFHX3</i>
16	85616985	1	RP11-118F19.1 (lincRNA)
17	3496105	1	<i>SHPK, TRPV1</i>
17	3498411-3527281	2	<i>SHPK, TRPV1</i>
17	3632836-3689132	3	<i>ITGAE</i>
17	45316717	1	Intergenic

18	51448760	1	Intergenic
19	38743962-38900106	14	<i>PPP1R14A, SPINT2, C19orf33, YIF1B, KCNK6, CATSPERG, PSMD8, SPRED3, GGN, FAM98C</i>
20	47403913-47420680	3	<i>PREX1</i>

**Supplementary Table 5:** Regions under potential differential selection between cases and controls (absolute XP-EHH > 4).

Chromosome	Position	No. SNPs	Gene
6	13824087	1	<i>MCUR1</i>
6	31066671	1	Major Histocompatibility Complex
6	53351289	1	<i>GCLC</i>
7	53105223	1	<i>POM121L12</i>
14	70875513-71047754	5	<i>SYNJ2BP, ADAM21, ADAM20</i>
21	40093658	1	<i>ERG, ETS2</i>

## **Chapter 6:**

Analysis of Tanzanian trios reveals inherited structural variants in genes with roles in ER-Golgi transport and blood antigen systems



## RESEARCH PAPER COVER SHEET

***PLEASE NOTE THAT A COVER SHEET MUST BE COMPLETED FOR EACH RESEARCH PAPER INCLUDED IN A THESIS.***

### **SECTION A – Student Details**

<b>Student</b>	Matt Ravenhall
<b>Principal Supervisor</b>	Taane Clark
<b>Thesis Title</b>	A bioinformatic analysis of malaria host and pathogen genomics

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

### **SECTION B – Paper already published**

Where was the work published?	n/a		
When was the work published?	n/a		
If the work was published prior to registration for your research degree, give a brief rationale for its inclusion	n/a		
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?	Genome Biology
Please list the paper's authors in the intended authorship order:	Matt Ravenhall, Nuno Sepúlveda, Alphaxard Manjurano, Behzad Nadim, George Mtove, Hannah Wangai, Caroline Maxwell, Raimos Olomi, Hugh Reyburn, Christopher J. Drakeley, Eleanor M. Riley, Susana Campino, Taane G. Clark
Stage of publication	Not yet 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 conducted data cleaning and analysis from BAM alignment files, created all figures, and wrote the manuscript under the supervision of Taane Clark. I also developed the analysis and visualisation software used in this work, as described in Chapter 7.
--	---

**Student Signature:** \_\_\_\_\_

**Date:** \_\_\_\_\_

**Supervisor Signature:** \_\_\_\_\_

**Date:** \_\_\_\_\_

# **Analysis of Tanzanian trios reveals inherited structural variants in genes with roles in ER-Golgi transport and blood antigen systems**

Matt Ravenhall<sup>1</sup>, Nuno Sepúlveda<sup>2,3</sup>, Alphaxard Manjurano<sup>4,5</sup>, Behzad Nadjm<sup>4</sup>, George Mtove<sup>4</sup>, Hannah Wangai<sup>4</sup>, Caroline Maxwell<sup>4</sup>, Raimos Olomi<sup>4</sup>, Hugh Reyburn<sup>2,4</sup>, Christopher J. Drakeley<sup>2</sup>, Eleanor M. Riley<sup>2,6</sup>, Susana Campino<sup>1,2</sup>, Taane G. Clark<sup>1,7,\*</sup>

- 1 Department of Pathogen Molecular Biology, London School of Hygiene and Tropical Medicine, London WC1E 7HT, UK
- 2 Department of Immunology and Infection, London School of Hygiene and Tropical Medicine, London WC1E 7HT, UK
- 3 Centre for Statistics and Applications of University of Lisbon, Lisbon, Portugal
- 4 Joint Malaria Programme, Kilimanjaro Christian Medical College, Moshi, Tanzania
- 5 National Institute for Medical Research, Tanzania
- 6 The Roslin Institute and Royal Dick School of Veterinary Studies, University of Edinburgh
- 7 Department of Infectious Disease Epidemiology, Faculty of Epidemiology and Population Health, London School of Hygiene and Tropical Medicine, London WC1E 7HT, UK

## Abstract

**Background:** Genomic analyses of human populations tend to consider small variants, such as single nucleotide polymorphisms (SNPs) and insertion-deletions (indels), rather than larger structural variants (SVs) despite their established role in disease. With the recent development of sophisticated bioinformatic techniques for the detection and analysis of SVs, we are now able to conduct high-throughput, in-depth analyses of SVs. Such approaches allow for both the discovery of novel variants generally and, with the use of parent-child paired studies, confirmation of inheritance.

**Results:** We present a comprehensive bioinformatic analysis of nearly 200,000 specific forms of putative structural variation within a human population of trios from Tanzania. 6,932 of which are found in at least 5% of samples, 180,577 being found only once, and approximately 16.7% being inherited without mutation from parent to child. Those most frequent SVs include a heterozygous 4,136 bp deletion of *SEC22B* and its promoter, inversions of *APP*, and 2,494 bp deletions of *ACSL3*. We also highlight novel roles for structural variation in candidate genes associated with malaria risk and blood group antigen systems, including *A4GALT*, *ATP2B4*, and *ABCG2*.

**Conclusions:** Our results demonstrate that SVs are present both broadly within the Tanzanian population and within specific genes with known roles in blood type antigens and risk of severe malaria. Inheritance of a notable portion of these variants suggests phenotypic roles that require investigation in the specific context of susceptibility to severe malaria and other communicable diseases.

**Keywords:** Structural variation, genome diversity, African genomics, Tanzania

## Background

Large studies of global human genomics have led to the identification of variants associated with disease susceptibility and resistance phenotypes. Yet these studies have generally focused upon small variations, such as single nucleotide polymorphisms (SNPs) and insertion-deletion events (indels), whilst overlooking larger variation such as structural variants (SVs). SVs typically exist as deletions, duplications, inversions, and insertions or complex combinations of those, such as sandwich-inversions [1]. Naturally these large forms of variation can be highly disruptive, potentially knocking out a gene through deletion of its promoter or exons. Alternatively, they may enhance expression through whole gene duplication or sequence insertion.

Established examples of structural variation in the human genome include duplications of *AMY1* altering salivary alpha-amylase levels [2] and the 400 kb inversion of factor VIII causing haemophilia A [3]. Deletions of the Duffy antigen-encoding gene *DARC* have also been shown to convey reduced susceptibility to infection by the second deadliest<sup>5</sup> malaria parasite, *Plasmodium vivax* [4]. Given the global distribution of malaria, Duffy-negative status is rare in Europe and the United States of America but present at relatively high frequencies across Africa [5].

African populations are often under-represented in genomic analyses, including large-scale endeavours such as the 1000 Genomes Project [6]. Tanzania represents one such population. Historically, African genomics tend to be less well understood than those of Europeans and North Americans, with the majority of genetic studies identifying disease-relevant variants within Caucasian populations [7]. Several studies have also shed doubt on the validity of disease-association variants when in the context of non-Caucasian populations [8,9]. This includes SNPs that have been identified in association with an

---

<sup>5</sup> As percentage of global deaths attributed to *P. vivax*.

increased risk of breast cancer and atrial fibrillation [10,11]. Together this highlights the need for a broader range of studies into human genomics such that local variations can be identified.

Here we present a case-study analysis of structural variation derived from 156 human genomes collected from Tanzania, enhanced by inheritance analysis to their 78 descendants, and focused towards the role of SVs in blood stage antigens. Whilst the majority of variants were rare, some genes include high frequency SVs with potential roles in intracellular transport, cell adhesion, and increased risk of severe malaria.

## Results

### Overview of identified SVs

Across the 156 parents 197,028 unique variants were identified across four forms of structural variation (deletions, duplications, insertions, and inversions) (Table S1) using a pipeline in which SVs were identified using DELLY and SV-Pop [12,13]. By type, this breaks down to 161,223 deletions, 17,543 duplications, 12,576 inversions and 5,686 insertions. 19,139 distinct genomic features, including genes, enhancers, promoters, promoter flanking regions, and open chromatin regions, contain structural variants (deletions: 14,358; insertions: 3,004; duplications: 3,951; inversions: 2,054). In total, we identified approximately 2,514.7 variants per individual (1911.5 deletions, 128.5 duplications, 365.8 insertions and 108.9 inversions), with those variants having a mean length of 3,537.7 bp (median 934 bp) (Table 1). Broadly, most SVs (59.7%) were inferred as being heterozygous with this being consistent across all models (deletions: 63.3%, duplications: 98.1%, inversions: 83.5%) except for insertions (20.0%). Fully inherited SVs tended to be more homozygous, as is to be expected, with only 29.1% being heterozygous.

6,932 (3.5%) variants were present in at least 5% of parents, consisting of 5,151 deletions, 1,555 insertions, 84 duplications and 142 insertions. 3,033 (43.8%) of these were completely intergenic, whilst the remaining 3,899 SVs (2,887 deletions, 876 insertions, 53 duplications, 83 inversions) at least partially overlapped genomic features. 750 of those features were regulatory elements, including 56 transcription factor binding sites, 176 promoter flanking regions, 80 promoters, 340 open chromatins and 98 enhancers (Table 1).

### Role of inheritance

Detection of structural variants in both parent and child supports their validity and biological relevance. Per-parent a mean of 16.72% (range: 5.60% to 31.58%) of distinct

SVs were inherited without mutation. For the 53,730 instances of parent-child inheritance (of which there are 8,289 distinct SVs), the majority have at least one parent predicted as homozygous for the variant (Total: 69.9%, Homozygous/Homozygous: 15.0%, Homozygous/Heterozygous: 5.8%, Homozygous/Missing: 49.2%), whilst significant rates of heterozygous parents are also present (Total: 30.1%, Heterozygous/Heterozygous: 3.6%, Heterozygous/Missing: 26.5%). Insertions were proportionally more likely to be inherited, though this may be inflated by an inability to directly compare inserted sequences (Table S2). Those most frequently inherited variants generally mirrored the full parent dataset, with deletions of *SEC22B*, inversions of *APP*, and insertions in *DNAH6* all being prominent.

### No significant ethnic stratification

The Chagga, Pare, and Wasambaa ethnic populations display subtle differences in SV frequencies, with Pare (710.6 per sample, 13.7% inherited) having generally more variants (Kruskal-Wallis chi-squared = 5.001, P=0.08) and significantly lower rates of inheritance (Kruskal-Wallis chi-squared = 27.208, P=1.24 $\times 10^{-6}$ ), than Chagga (557.4 per sample, 17.2% inherited) or Wasambaa (620.2 per sample, 18.7% inherited), though these differences were not large enough to suggest distinct ethnic stratification (Table S2). Of those differences, only one variant has an FST score greater than or equal to 0.2: an intergenic 328 bp deletion found at relatively high frequency in the Pare sub-group (FST: 0.2, Tanzania: 17.95%, Chagga: 8.3%, Pare: 43.8%, Wasambaa: 5.0%). Conversely 657 variants were proportionally identified for all sub-groups (FST = 0), including inversions within *APP*, *TBC1D3*, and *LOH12CR1*; deletions within *ZNF665*, *ASIP*, *KCNB2*, *DKK2*, *ACSL3*, *MGAT5*, *SEC22B*, and *PDE4DIP*; and duplications within *AF146191.4*, and *NBPF1*. For SV hotspots, five 1 kbp windows in two loci had an FST greater than or equal to 0.2. These included deletions within a promoter for the potassium channel *KCNJ8* on chromosome 12 (FST: 0.2, Tanzania: 21.15%, Chagga: 0.0%, Pare: 47.9%, Wasambaa:

11.7%), and an intergenic region of chromosome X (Fst: 0.245, Tanzania: 11.54%, Chagga: 0.0%, Pare: 35.4%, Wasambaa: 1.7%).

### **Window-based variant counts identify hotspots of variation**

Genome-wide window-based analysis identified 238,888 1 kbp windows (deletions: 151,922, duplications: 46,976, insertions: 10,778, inversions: 29,212) where at least one individual had a structural variant (Figure 1). 62,779 of those windows contained variants for at least 5% of samples (deletions: 44,406, duplications: 8,663, insertions: 3,110, inversions: 6,600). 255 windows contained a variant for all parents, with 141 of those featuring deletions, and 114 featuring inversions.

Deletion hotspots are found overlapping several genomic features including *ACSL3*, *ASIP*, *DKK2*, *KCNB2*, *ZNF665*, *SEC22B* and its promoter, three open chromatin (ENSR00000176649, ENSR00000181038, ENSR00000225923), a transcription factor binding site (ENSR00000195323), and 74 intergenic windows. In contrast, inversion hotspots include a 55,777 bp inversion on chromosome 17 centered on RP11-1407O15.2 and its neighbouring features: *TBC1D3*, and open chromatin (ENSR00000093575), and a transcription factor binding site (ENSR00000093574) (Figure 2).

### **Gene Ontology enrichment**

Enrichment analysis was applied to the most frequent variants (>5%) within the full parent dataset and identified 399 gene ontology terms (31 under-enriched, 368 over-enriched) with a significant false discovery rate (<0.05) (Table S3). Enrichment fold changes ranged from 0.07 (immunoglobulin production (GO:0002377)) to 5.73 (ionotropic glutamate receptor activity (GO:0004970)).

In general, significant enrichment was observed for neural functions (synapse (GO:0045202): 2.07, FDR=1.69x10-13; neuron part (GO:0097458): 1.76, FDR=1.92x10-13), adhesion (cell adhesion (GO:0007155): 2.08, FDR=6.78x10-13; cell-cell adhesion (GO:0098609): 2.43, FDR=1.14x10-10), and ion and drug binding (ion

binding (GO:0043167): 1.33, FDR=5.99x10-15, drug binding (GO:0008144): 1.57, FDR=1.42x10-8). Fold changes below 1 were present for terms relating to immunity (humoral immune response (GO:0006959): 0.19, FDR=3.27x10-5; adaptive immune response (GO:0002250): 0.41, FDR=1.31x10-3) and olfactory systems (olfactory receptor activity (GO:0004984): 0.36, FDR=1.96x10-3; detection of chemical stimulus involved in sensory perception (GO:0050907): 0.4, FDR=2.81x10-3).

These enrichments were broadly reflected in the full variant dataset suggesting that they correspond to significant roles for structural variation in cell adhesion, neuro-synaptic biology, and drug binding contrasting with a minimal role in olfactory systems. One exception was immune response genes, which were only under-enriched in the frequent variants subset, suggesting that high frequency structural variants in immunological genes are actively selected against - perhaps due to hyper-variability.

### **Structural variants associated with blood groups and susceptibility to malaria**

Given the broad scale of this study, we placed a specific focus on candidate regions previously associated with malaria or erythrocyte surface exposure to allow for a more in-depth, disease-relevant investigation. Malaria was selected due to its significant global impact, recent interest in identifying variants associated with susceptibility to severe subtypes, and our previous work using this dataset for imputation in a genome-wide association study within a severe malaria context [14].

We considered structural variation for genes with established roles in malaria risk and blood antigen groups, of these 15 contained a structural variant in at least one parent (Table 2). Notable examples include *A4GALT* (which encodes the P<sup>k</sup> antigen of the P blood group system), for which 17 individuals have a 369 bp deletion and 17 others have a heterozygous 700 bp deletion, *ATP2B4* (malaria risk factor), for which 24 individuals have non-specific approximately 2,000 bp deletions, and *ABCG2* (Junior (JR) blood antigen), for which 26 individuals have a 336 or 337 bp deletion (Figure 3). No structural

variants were identified in the Duffy gene, *DARC*, and a range of other blood group antigens including *CD44*, *CD55*, and *ART4*. None of these variants were fully inherited.

## **Discussion**

Structural variation underpins a wide range of genetic variation, but few large-scale investigations have been conducted, especially for African genomes. Endeavours such as the 1000 Genomes projects and the African Genome project have sought to better understand human genomics but tend to focus on smaller variants, such as SNPs and indels, and include only a limited range of African populations [6,15]. To our knowledge there has also been no specific consideration of the role of structural variation in blood antigen systems, or the impact of SV inheritance.

We present an in-depth analysis of 156 Tanzanian parents, and further analysis of those variants inherited by their children ( $n=78$ ). The characterisation of these variants is key to placing Tanzanian genomes within a global context and identifying novel variants. Broadly, the Tanzanian population considered here showed few signs of stratification, especially in relation to ethnicity. Structural variants also appear to be abundant, but actively selected against in most cases. This is consistent with previous studies that found an abundance of shorter and rarer variants, with a subset of variants being actively selected for [16–18].

The vast majority of structural variants were infrequent, with 91.65% of variants being present in only one individual. In contrast to the fully inherited variants that were found in a median of 55 individuals. This is consistent with structural variants generally being actively selected against with useful variants quickly being selected for. Broadly, enrichment is observed for gene ontology terms relating to cell adhesion, intracellular transport, and drug binding. Those variants that exist at higher frequencies within the Tanzanian population tend to be deletions, though this is unsurprising given that they are the most common form of variant.

Whilst most SVs are non-specifically conserved in the general Tanzanian population, several genes contain specific variants that appear to have become fixed, or near-fixed, relative to the human reference genome. The most striking examples include a heterozygous 220 bp deletion in *BET1L* (98.7%), a gene associated with ER-Golgi vesicle transport-associated *BET1* [19], and a 4,136 bp deletion in SNARE-complexing *SEC22B* (98.7%) [20]. Notably *BET1* and *SEC22B* both have roles relating to SNARE functionality in the ER to Golgi transportation [21,22], suggesting that deletions in *SEC22B* and *BET1L* may have a linked phenotypic impact. Similar deletions have also been identified in samples from Nigeria, North America, and Kuwait, suggesting widespread pseudogenisation of these genes [23–26]. Several other SVs are also present within genes that encode surface-exposed proteins, such as *GPR156* (probable G-protein receptor, 29 bp insertion in 77.6%), and *TSPAN8* (2,117 bp deletion in 72.4%).

Our focus on known malaria risk factors or blood group antigen genes led to the identification of SV hotspots in P<sup>k</sup> antigen gene *A4GALT*, malaria risk factor *ATP2B4*, and JR blood system antigen *ABCG2*, amongst others. Many of these have been identified in broad studies of structural variation in several populations, though without phenotypic context or specific investigation [23,24,27]. Interestingly, *ABCG2* encodes a transporter protective against xenobiotic molecules as well as the antigen of the JR blood system [28,29]. The wild type Jr phenotype is found at high prevalence in all populations whilst the Jr(a-) phenotype is a highly rare variant found, for example, in 0.05% of Japanese blood donors [30]. Whilst null variants of *ABCG2* can be caused through various short variants [30], it is unclear whether the 336/7 bp deletions detected here reflect a rare JR phenotype.

Several high frequency SVs are also present within genes that have been otherwise linked to potential roles in malaria susceptibility and other diseases. For example, 84.6% of parents have a 40 bp insertion in *RAB7A*, a key regulator of endosomal trafficking [31],

with a similar insertion being present within the Venter genome [27]. *RAB7A* has been found in *P. berghei*-containing host vesicles [32] and is a known target for microbial survival strategies, such as with *Salmonella*, further supporting its role in general host-pathogen interactions [33]. Our identified insertion may therefore convey a protective effect, inhibiting *Plasmodium* invasion or intracellular survival.

Notably, all sampled genomes were aligned to the GRCh37 version of the human reference genome, which was sourced from 13 anonymous individuals in New York [34]. As such, many of the structural variants identified in this study may represent fixed, silent population differences between American and African genomes rather than recently selected variants, although many such selection events likely exist.

## Conclusions

Several novel structural variants were detected within the Tanzanian population, some of which may have roles in intracellular transportation, cell-cell adhesion, and blood antigens. The vast majority of variants were rare, though a significant number were near-fixed or inherited without mutation from parent to child. Our findings are consistent with previous studies that suggest structural variants are frequent to occur, but experience limited positive selection. This work highlights the importance of continued, high-throughput discovery in further populations.

## **Methods**

### **Dataset composition, collection, and sequencing**

Our dataset included 247 anonymously sampled individuals, consisting of 78 healthy parental and child trios (156 parents, 78 children); 80 Chagga, 77 Pare, 90 Wasambaa (Parents: 48 Chagga, 48 Pare, 60 Wasambaa)). 13 singletons were also present but excluded from primary analysis. Samples were collected between 2007 and 2008 villages near the Kilimanjaro, Pare, and West Usambara mountains in the Tanga region of Tanzania, a region with a low to medium level of malaria transmission. Preliminary candidate region structural variation analysis was previously conducted to compliment a genome-wide association study [14]. Samples were sequenced at the Sanger Institute using Illumina HiSeq2500 technology and aligned to the GRCh37 build of the human genome. All co-ordinates refer to the GRCh37.p13 build of the human genome, with annotations being acquired from NCBI.

### **Variant Identification and Filtering**

Structural variants were identified using DELLY version v0.7.3 with default parameters [12]. High-throughput, post-discovery filtering and analysis was undertaken with SV-Pop version 1.0 [13] (available at <http://github.com/mattravenhall/SV-Pop>) and through manual curation. Variants were removed for overlapping known assembly gaps, being abnormally long (>100,000 base pairs), and according to population-level DELLY scores including quality (mean < 0.9), inferred percentage homozygous reference (>10%), and inferred percentage homozygous missingness (>10%). SV discovery was performed for parent and child groups separately, with the inherited subset being those found for both parent and children without inter-generational mutation. Regional hotspots were identified with SV-Pop using default parameters consisting of a sliding window count for a 1000 bp window with a 500 bp step.

### **Gene Ontology**

Gene ontology analysis was performed separately for the parent variants and frequent parent variants (>5%) datasets using the PANTHER classification system [35]. Significance was defined as a false discovery rate less than 0.05.

## **Declarations**

### **Ethics approval and consent to participate**

All DNA samples were collected and genotyped following signed and informed written consent from a parent or guardian. Ethics approval for all procedures was obtained from both LSHTM (#2087) and the Tanzanian National Institute of Medical Research (NIMR/HQ/R.8a/Vol.IX/392).

### **Consent for publication**

Not applicable.

### **Competing interests**

The authors declare no competing interests.

### **Funding**

MR is funded by the Biotechnology and Biological Sciences Research Council (grant number BB/J014567/1). TGC is supported by the Medical Research Council UK (Grant no. MR/K000551/1, MR/M01360X/1, MR/N010469/1, MR/R020973/1) and BBSRC UK (BB/R013063/1). SC is funded by the Medical Research Council UK (Grant no. MR/M01360X/1) and the BBSRC UK (BB/R013063/1).

### **Authors' contributions**

MR conducted the statistical analysis. MR and TC conceived the study. NS, AM, BN, GM, HW, CM, RO, HR, CJD, EMR aided in data collection and resource provision. MR, TC and SC wrote the manuscript and all authors read and approved the final version.

## **Acknowledgements**

We thank the participants and Tanzanian communities who made this study possible, and the healthcare workers who assisted with this work. The Medical Research Council UK funded eMedLab computing resource was used for data analysis.

## References

1. Ravenhall M, Diez Benavente E, de Sessions PF, Walker EM, Hibberd ML, Baker DA, et al. Analysis of global long read *Plasmodium falciparum* genomes identifies novel inversions. *Prep.*
2. Mandel AL, Peyrot des Gachons C, Plank KL, Alarcon S, Breslin PAS. Individual Differences in AMY1 Gene Copy Number, Salivary  $\alpha$ -Amylase Levels, and the Perception of Oral Starch. Kayser M, editor. *PLoS One*. Public Library of Science; 2010;5:e13352.
3. Lakich D, Kazazian HH, Antonarakis SE, Gitschier J. Inversions disrupting the factor VIII gene are a common cause of severe haemophilia A. *Nat Genet*. 1993;5:236–41.
4. Gunalan K, Lo E, Hostetler JB, Yewhalaw D, Mu J, Neafsey DE, et al. Role of *Plasmodium vivax* Duffy-binding protein 1 in invasion of Duffy-null Africans. *Proc Natl Acad Sci*.
5. Howes RE, Patil AP, Piel FB, Nyangiri OA, Kabaria CW, Gething PW, et al. The global distribution of the Duffy blood group. *Nat Commun*. Nature Publishing Group; 2011;2:266.
6. Gibbs RA, Boerwinkle E, Doddapaneni H, Han Y, Korchina V, Kovar C, et al. A global reference for human genetic variation. *Nature*. Nature Publishing Group; 2015;526:68–74.
7. Quansah E, McGregor NW. Towards diversity in genomics: The emergence of neurogenomics in Africa? *Genomics*. Academic Press; 2018;110:1–9.
8. Rosenberg NA, Huang L, Jewett EM, Szpiech ZA, Jankovic I, Boehnke M. Genome-wide association studies in diverse populations. *Nat Rev Genet*. NIH Public Access; 2010;11:356–66.
9. Peprah E, Xu H, Tekola-Ayele F, Royal CD. Genome-wide association studies in Africans and African Americans: expanding the framework of the genomics of human

- traits and disease. *Public Health Genomics*. NIH Public Access; 2015;18:40–51.
10. Ruiz-Narváez EA, Rosenberg L, Cozier YC, Cupples LA, Adams-Campbell LL, Palmer JR. Polymorphisms in the TOX3/LOC643714 locus and risk of breast cancer in African-American women. *Cancer Epidemiol Biomarkers Prev*. American Association for Cancer Research; 2010;19:1320–7.
11. Delaney JT, Jeff JM, Brown NJ, Pretorius M, Okafor HE, Darbar D, et al. Characterization of Genome-Wide Association-Identified Variants for Atrial Fibrillation in African Americans. Toland AE, editor. *PLoS One*. Public Library of Science; 2012;7:e32338.
12. Rausch T, Zichner T, Schlattl A, Stutz AM, Benes V, Korbel JO, et al. DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics*. Oxford University Press; 2012;28:i333–9.
13. Ravenhall M, Campino S, Clark TG. SV-Pop: Population-based structural variant analysis and visualisation. *Prep*.
14. Ravenhall M, Campino S, Sepúlveda N, Manjurano A, Nadjm B, Mtove G, et al. Novel genetic polymorphisms associated with severe malaria and under selective pressure in North-eastern Tanzania. *PLOS Genet*. 2018;14.
15. Gurdasani D, Carstensen T, Tekola-Ayele F, Pagani L, Tachmazidou I, Hatzikotoulas K, et al. The African Genome Variation Project shapes medical genetics in Africa. *Nature*. Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved.; 2014;517:327–32.
16. Sudmant PH, Rausch T, Gardner EJ, Handsaker RE, Abyzov A, Huddleston J, et al. An integrated map of structural variation in 2,504 human genomes. *Nature*. Europe PMC Funders; 2015;526:75–81.
17. Simons YB, Turchin MC, Pritchard JK, Sella G. The deleterious mutation load is insensitive to recent population history. *Nat Genet*. 2014;46:220–4.

18. Cheeseman IH, Gomez-Escobar N, Carret CK, Ivens A, Stewart LB, Tetteh KKA, et al. Gene copy number variation throughout the *Plasmodium falciparum* genome. *BMC Genomics.* BioMed Central; 2009;10:353.
19. Xu Y, Wong SH, Zhang T, Subramaniam VN, Hong W. GS15, a 15-kilodalton Golgi soluble N-ethylmaleimide-sensitive factor attachment protein receptor (SNARE) homologous to rbet1. *J Biol Chem.* American Society for Biochemistry and Molecular Biology; 1997;272:20162–6.
20. Petkovic M, Jemaiel A, Daste F, Specht CG, Izeddin I, Vorkel D, et al. The SNARE Sec22b has a non-fusogenic function in plasma membrane expansion. *Nat Cell Biol.* Nature Publishing Group; 2014;16:434–44.
21. Newman AP, Shim J, Ferro-Novick S. BET1, BOS1, and SEC22 are members of a group of interacting yeast genes required for transport from the endoplasmic reticulum to the Golgi complex. *Mol Cell Biol.* 1990;10:3405–14.
22. Volchuk A, Ravazzola M, Perrelet A, Eng WS, Di Liberto M, Varlamov O, et al. Countercurrent Distribution of Two Distinct SNARE Complexes Mediating Transport within the Golgi Stack. *Mol Biol Cell.* 2004;15:1506–18.
23. 1000 Genomes Project Consortium. A map of human genome variation from population scale sequencing. *Nature.* 2011;467:1061–73.
24. Thareja G, John SE, Hebbar P, Behbehani K, Thanaraj TA, Alsmadi O. Sequence and analysis of a whole genome from Kuwaiti population subgroup of Persian ancestry. *BMC Genomics.* 2015;16:1–14.
25. Arlt MF, Ozdemir AC, Birkeland SR, Lyons RH, Glover TW, Wilson TE. Comparison of Constitutional and Replication Stress-Induced Genome Structural Variation by SNP Array and Mate-Pair Sequencing. *Genet Soc Am.* 2011;187:675–83.
26. Pang AWC, Macdonald JR, Yuen RKC, Hayes VM. Performance of High-Throughput Sequencing for the Discovery of Genetic Variation Across the Complete

- Size Spectrum. *G3 (Bethesda)*. 2014;4:63–5.
27. Levy S, Sutton G, Ng PC, Feuk L, Halpern AL, Walenz BP, et al. The Diploid Genome Sequence of an Individual Human. *PLOS Biol*. 2007;5.
28. Taylor NMI, Manolaridis I, Jackson SM, Kowal J, Stahlberg H, Locher KP. Structure of the human multidrug transporter ABCG2. *Nature*. Nature Publishing Group; 2017;546:504.
29. Castilho L, Reid ME. A review of the JR blood group system. *Immunohematology*. 2013;29:63–8.
30. Tanaka M, Kamada I, Takahashi J, Kimura K, Matsukura H, Tani Y. Defining the Jr(a-) phenotype in the Japanese population. *Transfusion*. Wiley/Blackwell (10.1111); 2013;54:n/a-n/a.
31. Vanlandingham PA, Ceresa BP. Rab7 regulates late endocytic trafficking downstream of multivesicular body biogenesis and cargo sequestration. *J Biol Chem*. American Society for Biochemistry and Molecular Biology; 2009;284:12110–24.
32. Lopes da Silva M, Thieleke-Matos C, Cabrita-Santos L, Ramalho JS, Wavre-Shapton ST, Futter CE, et al. The Host Endocytic Pathway is Essential for Plasmodium berghei Late Liver Stage Development. *Traffic*. 2012;13:1351–63.
33. D’Costa VM, Braun V, Landekic M, Shi R, Proteau A, McDonald L, et al. *Salmonella* Disrupts Host Endocytic Trafficking by SopD2-Mediated Inhibition of Rab7. *Cell Rep*. 2015;12:1508–18.
34. Morgan BQ. *E Pluribus Unum*. Mod Lang J. Nature Publishing Group; 1927;11:485–8.
35. Mi H, Huang X, Muruganujan A, Tang H, Mills C, Kang D, et al. PANTHER version 11: expanded annotation data from Gene Ontology and Reactome pathways, and data analysis tool enhancements. *Nucleic Acids Res*. Oxford University Press; 2017;45:D183–9.

## Figure & Table Legends

**Table 1:** Summary statistics of identified structural variants by model.

Variant Type	Total Count	Mean per Sample	Length (median bp (range))	1 kbp Windows	Frequent (>5%)	Intergenic	Genic
Deletions	161,223	1,911.5	954 (13 to 99,429)	151,922	5,151	64,078	97,145
Duplications	17,543	128.5	958 (138 to 99,894)	46,976	84	6,855	10,688
Insertions	5,686	365.8	21 (15 to 138)	10,778	1,555	2,382	3,304
Inversions	12,576	108.9	1,261 (95 to 99,999)	29,212	142	4,588	7,988

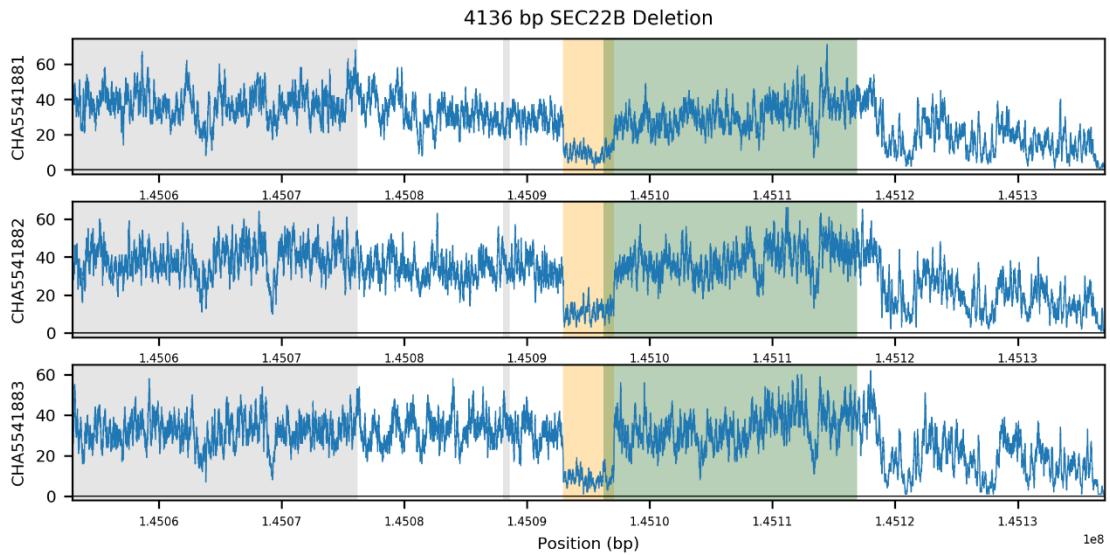
**Table 2:** Specific forms of structural variants associated with malaria risk or blood

antigen genes.

Gene	Qualification	Location	Total	Deletions	Duplications	Insertions	Inversions
<i>A4GALT</i>	P blood system (Pk antigen)	22: 43,088,127-43,117,304	31	31	-	-	-
<i>ATP2B4</i>	Malaria risk factor	1: 203,595,689-203,713,209	24	24	-	-	-
<i>ABCG2</i>	Jr blood group	4: 89,011,416-89,152,474	22	22	-	-	-
<i>FREM3</i>	Malaria risk factor	4: 144,498,455-144,621,828	21	21	-	-	-
<i>AQP1</i>	Colton antigen	7: 30,893,010-30,965,131	11	11	-	-	-
<i>GYPC</i>	Gerbich blood group	2: 127,413,509-127,454,246	4	4	-	-	-
<i>RHD</i>	Rh Protein	1: 25,598,884-25,656,936	3	-	-	-	3
<i>RHCE</i>	Rh Protein	1: 25,688,740-25,756,683	3	-	-	-	3
<i>ABO</i>	ABO blood group	9: 136,125,788-136,150,617	2	1	-	1	-
<i>XG</i>	XG antigen	X: 2,670,091-2,734,539	2	1	-	1	-
<i>XK</i>	Kx antigen	X: 37,545,012-37,591,383	2	2	-	-	-
<i>GYPB</i>	MNS antigen	4: 144,917,257-145,061,844	1	1	-	-	-
<i>SLC14A1</i>	Kidd antigen	18: 43,304,092-43,332,485	1	1	-	-	-
<i>C4B</i>	Chido/Rodgers antigen	6: 31,982,539-32,003,195	1	1	-	-	-
<i>BSG</i>	Ok blood group	19: 571,297-583,493	1	-	-	1	-

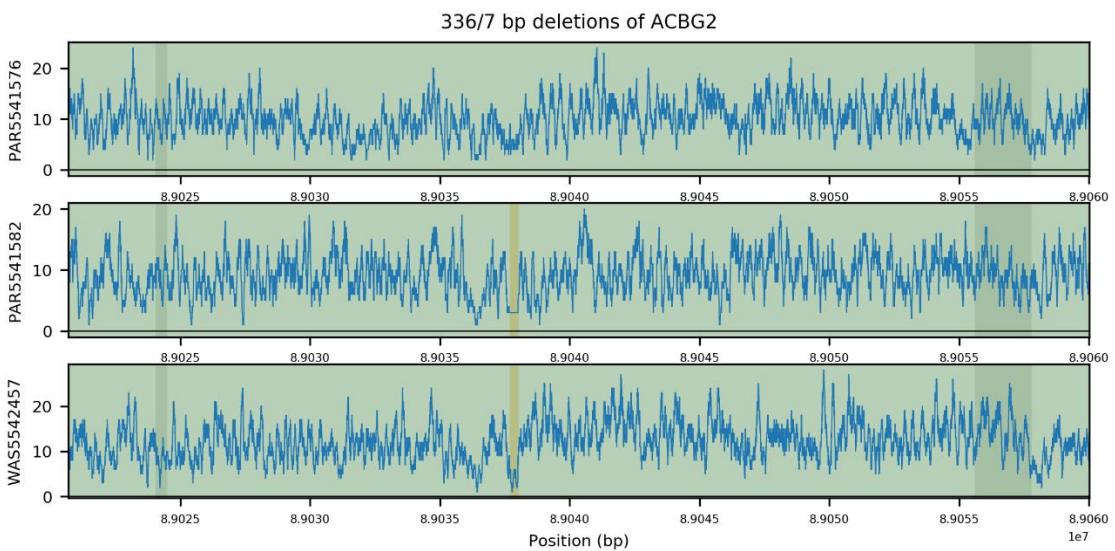
**Figure 1: Coverage plot confirming the 4,136 bp heterozygous deletion in SEC22B.**

Blue represents the per base coverage, Orange indicates the predicted structural variant, Green indicates the gene of interest, Grey indicates genomic features including genes and enhancers.



**Figure 2: Coverage plot confirming the 336/7 bp deletions in ABCG2.** Blue

represents the per base coverage, Orange indicates the predicted structural variant, Green indicates the gene of interest, Grey indicates genomic features including genes and enhancers.



## **Supplementary Figure & Table Legends**

**Supplementary Table 1:** Full list of putative structural variants. [large file]

**Supplementary Table 2:** Per-sample ethnicity, model, and inheritance. [large file]

**Supplementary Table 3:** Significantly enriched gene ontology terms. [large file]

## **Chapter 7:**

SV-Pop: Population-based structural variant analysis and  
visualisation



## **RESEARCH PAPER COVER SHEET**

***PLEASE NOTE THAT A COVER SHEET MUST BE COMPLETED FOR EACH RESEARCH PAPER INCLUDED IN A THESIS.***

### **SECTION A – Student Details**

<b>Student</b>	Matt Ravenhall
<b>Principal Supervisor</b>	Taane Clark
<b>Thesis Title</b>	A bioinformatic analysis of malaria host and pathogen genomics

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

### **SECTION B – Paper already published**

Where was the work published?	n/a		
When was the work published?	n/a		
If the work was published prior to registration for your research degree, give a brief rationale for its inclusion	n/a		
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?	Bioinformatics		
Please list the paper's authors in the intended authorship order:	Matt Ravenhall, Susana Campino, 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 developed SV-Pop to facilitate analysis work for Chapters 3 and 6, created all figures, and wrote the manuscript under the supervision of Taane Clark.
--	--

**Student Signature:** \_\_\_\_\_

**Date:** \_\_\_\_\_

**Supervisor Signature:** \_\_\_\_\_

**Date:** \_\_\_\_\_

---

## Genome analysis

# SV-Pop: Population-based structural variant analysis and visualisation

Matt Ravenhall<sup>1,\*</sup>, Susana Campino<sup>1</sup>, and Taane G. Clark<sup>1,2</sup>

<sup>1</sup>Department of Pathogen Molecular Biology, London School of Hygiene and Tropical Medicine, London, WC1E 7HT, United Kingdom, <sup>2</sup>Department of Infectious Disease Epidemiology, London School of Hygiene and Tropical Medicine, London, WC1E 7HT, United Kingdom.

\*To whom correspondence should be addressed.

Associate Editor: XXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

### Abstract

**Motivation:** Genetic structural variation underpins a multitude of phenotypes, with significant implications for a range of biological outcomes. Despite their crucial role, structural variants (SVs) are often neglected and overshadowed by single nucleotide polymorphisms (SNPs), which are used in large-scale analysis such as genome-wide association and population genetic studies.

**Results:** To facilitate the high-throughput analysis of structural variation we have developed an analytical pipeline and visualisation tool, called SV-Pop. Designed to facilitate downstream analysis and visualisation post-discovery, SV-Pop allows for straight-forward integration of multi-population analysis, method and sample-based concordance metrics, and signals of selection.

**Availability:** SV-Pop is available at <https://github.com/mattravenhall/SV-Pop>.

**Contact:** matt.ravenhall@lshtm.ac.uk

**Supplementary information:** Supplementary data are available at *Bioinformatics* online.

## 1. Introduction

Structural variation (SVs) describes changes to a core genome beyond single nucleotide polymorphisms (SNPs) or very short insertions and deletions (indels). Typically, SVs are large (>500 bp) and consist of four major types: deletions, insertions, duplications, and inversions. All play an important contribution to human and pathogen diversity and disease susceptibility. For example, duplications of the *Plasmodium falciparum* malaria parasite *gch1* have been associated with antimalarial resistance (Heinberg A and Kirkman L 2015), and deletions of the human Duffy antigen convey resistance to malaria infection (Miller L *et al.* 1976). Despite their significant implications, the role of SVs has been overshadowed by SNPs, which are easier and faster to identify. Various SV discovery methods such as DELLY and CNVNator currently exist (Rausch T *et al.* 2012, Abyzov A *et al.* 2011), but there is currently no efficient tool for efficiently identifying concordance between models, up-scaling analysis for multiple populations, or visualising that output.

To assist the identification and investigation of SVs, we have developed a bioinformatics pipeline for high-throughput post-discovery analysis and visualisation.

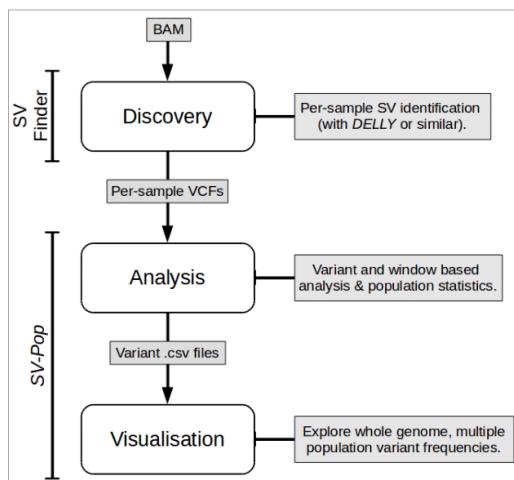
## 2. Implementation

SV-Pop consists of two core modules: (i) population-based *analysis* following individual SV discovery, and (ii) *visualisation* of those variants for dynamic, whole genome exploration. The *analysis module* is a Unix command line tool built in Python (v3.3+) with Pandas (v0.18+), and Numpy (v1.10.4+). The *visualisation module* is built using the R Shiny web framework (Chang W *et al.* 2017), and requires R (v3.3+) alongside the *shiny*, *plotly*, *data.table*, and *dplyr* packages. It can be launched on command line using ‘Rscript easyRun.r’, then explored via your default web browser. Input files should be pre-processed with SV-Pop, using the *PREPROCESS* mode for full compatibility.

### 2.1 Analysis

Input to SV-Pop consists of an array of post-discovery files (vcf format), one per-individual sample. These are typically the output of a run of DELLY or similar (Rausch T *et al.* 2012). Variants across all samples are then processed, identifying and combined those specific variants that are shared across multiple samples and performing appropriate summary statistics. If so desired, variants can be filtered according to their con-

cordance with a secondary discovery method by supplying a csv file of those variants with the *dirConcordance* argument. By default, variants are matched if they overlap at least 80% of the region identified by the primary method.



**Fig. 1.** Overview of the full SV-Pop pipeline.

Once collated, we can consider a rolling window across the sample genome and identify regions with high or low variant overlap. This produces a coverage-like statistic for those underlying SVs. We can then further dissect according to sub-populations, as provided by the user. Specific variant sets can also be annotated, subset, merged, and filtered as required. In addition to this core analysis, we have structured the pipeline to allow seamless integration of various filters and statistics, including method concordance and fixation indexes ( $F_{ST}$ ).

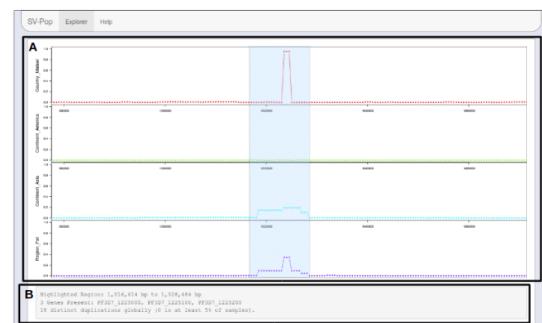
Typically, an *analysis module* run involves calling SVs across multiple models for a population of samples, inputting those individual output vcf files into SV-Pop, and producing per-variant or per-window based statistics (as csv files) for input into the *visualisation module*.

## 2.2 Visualisation

Post-analysis, per-window files can be brought forward to the *visualisation module*, facilitating dynamic investigation of whole genome structural variation. By default, the *visualisation module* will identify variant frequencies and difference metrics (e.g.  $F_{ST}$  values) for all populations within your provided files, allowing the user to easily specify which they are interested in viewing. Similarly, the chromosomes and their sizes are detected allowing the user to specify regions of interest. Users are also able to subset and download specified genomic regions of interest for further analysis.

## 3. Application to *Plasmodium falciparum*

To demonstrate the utility of SV-Pop, *P. falciparum* malaria parasite alignment files from 3,110 samples across 21 countries with published sequence data (Ravenhall M *et al.* 2016) were processed with SV-Pop and loaded into the visualiser. As shown in Figure 2, the previously identified *gch1* promoter duplication is found by both elevated frequencies and a spike in the  $F_{ST}$  metric.



**Fig. 2.** Screenshot of the visualization module displaying duplication frequencies for Malawi, South America, Asia, and region-based  $F_{ST}$  values. a) Variant viewer, displaying per-window frequencies and statistical metrics. The spike in the Malawi track (red) is the previously identified *gch1* promoter region duplication, whilst the ridge in the Asia track (cyan) indicates whole gene duplications. The  $F_{ST}$  track (purple) highlights frequency differences between region groups. b) Region summary, statistics regarding the region highlighted in the viewer. c) Variant and Chromosome selector. d) Population selection. e) Location selection and download.

## 4. Conclusion

SV-Pop dramatically increases the accessibility of large population-based SV studies, allowing for a greater volume of downstream analysis and visualisation. It also establishes a core pipeline upon which to incorporate existing and future metrics such as method concordance and selection statistics. This implementation, which has been tested on a *P. falciparum* dataset, is also species-agnostic ensuring that it can be applied in a wide range of contexts.

## Funding

MR is funded by the Biotechnology and Biological Sciences Research Council (Grant Number BB/J014567/1).

*Conflict of Interest:* none declared.

## References

- Abyzov A, Urban AE, Snyder M, Gerstein M. (2011) CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. *Genome Research*, 21(6):974–84. doi: 10.1101/gr.114876.110
- Chang W, Cheng J, Allaire JJ, Xie Y, McPherson Joe. (2017) shiny: Web Application Framework for R. R package version 1.0.1. <https://CRAN.R-project.org/package=shiny>
- Heinberg A, and Kirkman L. (2015) The molecular basis of antifolate resistance in Plasmodium falciparum: looking beyond point mutations. *Annals of the New York Academy of Sciences*, 1342(1): 10–18. doi: 10.1111/nyas.12662
- Miller LH, Mason SJ, Clyde DF, McGinnis MH. (1976) The resistance factor to Plasmodium vivax in blacks. The Duffy-blood-group genotype, FyFy. *The New England Journal of Medicine*, 295(6):302–304
- Ravenhall M, Benavente ED, Mipando M, Jensen ATR, Sutherland CJ, Roper C, Sepulveda N, Kwiatkowski DP, Montgomery J, Phiri KS, Terlouw A, Craig A, Campino S, Ocholla H, and Clark TG. (2016) Characterizing the impact of sustained sulfadoxine/pyrimethamine use upon the *Plasmodium falciparum* population in Malawi. *Malaria Journal*, 15:575
- Rausch T, Zichner T, Schlatzl A, Stuetz AM, Benes V, Korbel JO. (2012) DELLY: Structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics* 28:i333–i339.

## **Chapter 8:**

### Discussion, Conclusions, and Future Work

## 8.1 Discussion

The work presented here combines two similar approaches to identify and characterise novel genomic variation relevant to host-pathogen interactions in malaria, one focused on the *P. falciparum* parasite and the other on their human hosts. Both components began with single context studies primarily exploring SNP-based variation, with the incorporation of selection methods and candidate structural variation (Chapters Two and Five). Having identified some key novel structural variants when using a candidate gene approach, both human and parasite components progressed to large-scale explorations of structural variation either globally (Chapter Three), in long-read assemblies (Chapter Four), or with consideration of inheritance (Chapter Six). These SV-based studies required the development of novel pipelines and software as detailed in Chapters Four and Seven.

Beginning with the parasite component, Chapter Two covers my work investigating the impact of sustained SP use on the Malawian *P. falciparum* population following fourteen years of primary use following a switch from chloroquine in 1993 and to ACTs in 2007 [1], with those samples being collected between 2010 and 2012. Perhaps unsurprisingly a high frequency of SP resistance associated SNPs remained within the population, with these including classic variants such as *dhps* double mutants (A437G and K540E, 99.5%) and *dhfr* triple mutants (N51I, C59R, and S108N, 97.7%). SNPs associated with chloroquine resistance were entirely absent, suggesting that the *P. falciparum* population was generally chloroquine susceptible, consistent with the existing literature [2].

These SNP-based findings were supported by signals of selection, as demonstrated by extended haplotype homozygosity-based approaches. Within the Malawian population, significant recent positive iHS selection signals were observed around *dhps*, *dhfr*, and

*gch1*, reinforcing their roles in SP resistance [3]. Additional iHS signals, suggestive of conserving selection, were observed around *ama1*, *trap*, and *rrf1*, all of which encode surface antigens crucial for erythrocyte and hepatocyte invasion [4,5]. Relative selection signals between Malawi and other regions were also considered, with signals for Malawi around *gch1*, *dhps*, and *acs8* (acyl-coA synthase), and for non-Malawi around *crt*, *trap*, and *msp10*. Relative selection for *gch1* and *dhps* is consistent with positive selection in Malawi due to SP pressure, which neighbouring countries lack. Similarly, the reverse is true for *crt*, as chloroquine use was halted early in Malawi [1]. Relative selection elsewhere for the merozoite surface protein *msp10* is more curious and may underpin regional differences in host-pathogen interactions. Structural variation was also considered for candidate resistance genes, given the present of *gch1* duplications in Southeast Asia [6]. Those whole gene duplications were found at a low frequency in South (11.1%) and Southeast Asia (7.6%), but absent from Malawi. Instead, novel 436 bp duplications were identified immediately upstream of *gch1* within a probable TATA box, suggesting a potential role in the expression of *gch1*.

For the human component, Chapter Five describes my initial focus on a genome wide association study for a Tanzanian population (n=914) to identify SNPs associated with susceptibility to, or protection against, severe malaria subtypes, such as respiratory distress or cerebral malaria. This dataset included individuals from malaria-endemic regions, divided into those who had experienced a severe response to malaria infection (case), and those who had not (control). This case-control approach confirmed the association of the sickle polymorphism *rs334*, and identified novel associations including a pair of interleukin receptors *IL-23R* and *IL-12RBR2*. Together these suggest a potential role for pro-inflammatory cytokines in severe responses to malaria which may be linked to previous studies that found a protective role for *IL-12* and *IL-23* [7,8]. Crucially those

novel SNPs found in association with severe subtypes were subtly different to variants found in other populations, this being consistent with other GWASs that have identified population-specific associations [9,10]. Many of these variants likely reflect SNPs within local haplotypes that carry a causal variant, or causal variants that are not under strong selection in other populations. As with the Malawian *P. falciparum* approach in Chapter Two, signals of selection were detected using EHH-based iHS and XP-EHH. In this context, iHS considered selection in the combined population, whilst XP-EHH considered selection differences between the case and control group. Due to the high level of relatedness within this dataset, XP-EHH signals were generally low but differential selection signals were identified in *SYNJ2BP*, *GCLC*, and the MHC. Selective differences within the MHC were expected given its key immunological role, however relative selection for *SYNJ2BP*, which encodes a membrane trafficking protein, suggest possible roles in parasite invasion of liver or red blood cells that require laboratory confirmation. Candidate SV discovery was also considered for genes with the most significant association or selection signals, with this work being expanded upon in Chapter Six. Here a range of low frequency SVs were identified and imputed into the primary GWAS dataset, with no SV being significantly associated with risk of severe subtypes.

As my project progressed, the impact of larger structural variation became more apparent, so I sought to conduct large-scale discoveries of SVs in both parasite and human populations. Chapter Three focused on global differences in the distribution of structural variation in 3,110 *P. falciparum* samples from 21 malaria-endemic countries. Over one million putative structural variants were identified, with this being filtered to a high-quality subset of >70,000 specific deletions and >600 duplications. The majority of specific variants were rare (48.5% of deletions and 94.7% of duplications are found in single isolates), only 2.4% of deletions and 0.2% of duplications found in at least 5% of

samples. This study allowed for a greater resolution for the global distribution of previously identified duplications of *mdr1* and the *gch1* promoter region, the latter initially being discovered in the work described in Chapter Two. This confirmed *mdr1* as a primarily Southeast Asian duplication, and the *gch1* promoter duplication as focused within Malawi but also present in several Central and East African populations. A novel 22.9 kbp duplication of *crt* was also identified in West African samples primarily from Burkina Faso and Ghana, but also Guinea and Mali. This large *crt* duplication is particularly intriguing as it displayed high rates of diploid-type phasing, here being used as an approximate proxy for mixed infections in the haploid samples. Follow up haplotype-based analysis of the specific sequencing reads present for the region found high conservation for reads containing both the chloroquine resistance and susceptible types of *crt*. This suggested the presence of two copies of *crt*, one wildtype and one conveying chloroquine resistance. Regional concordance between species, and consistency in the read count ratios (1:1) for chloroquine resistant and susceptible types mean that this *in silico* finding is highly robust. Presumably, carriage for both chloroquine resistant and susceptible forms of *crt* allows the benefit of both phenotypes, perhaps ‘locking in’ chloroquine resistance whilst accounting for associated fitness costs. This reflects similar hypotheses for the role of *gch1* whole gene, and presumably promoter, duplications. By producing more GCH1, resistant but less efficient forms of DHPS and DHFR may have their fitness costs counter-balanced, whilst two forms of *crt* may allow for the expression of a more efficient wildtype CRT alongside or instead of a resistant alternative.

Through investigating the use of short read-based approaches for the inference of structural variation, it became apparent that some types of SV were less successfully identified than others. Specifically, inversions have been poorly detected and

characterised in *P. falciparum* despite several variants of interest, such as the inversion-duplication of *gch1* in Dd2 [11]. To tackle this, I sought to develop a basic method for inversion detection which utilised long sequencing reads. In theory, longer sequencing reads such as PacBio and Nanopore should be able to identify variants in highly similar or repetitive regions more accurately than short read-based methods [12].

Chapter Four details the development of a long-read alignment-based pipeline for the identification of inverted sequences relative to the 3D7 reference genome. Many of the inversions detected in this study were present within highly variable *var*, *rifin*, and *stevor* genes or intergenic regions that are often excluded from short-read based SV discovery due to difficulties resolving their sequences. The abundance of inversions in these highly variable regions is consistent with the existing literature and probably relates to the highly variable nature of these genes [13]. The presence of ancestral inversions was another key finding, with key examples including the *P. falciparum* unique *rh2a/b* gene pair and a pair of elongation factors 1-alpha [14]. Phenotypically these are likely to be well established and highly conserved but highlight inversion-deletions as a key combination of structural variants. Similarly, ‘sandwich inversion’, in which a sequence exists inverted between two other copies are a feature previously identified for *gch1* in the lab-derived Dd2 strain of *P. falciparum* [11]. Alongside robustly discovering that feature in Dd2, a novel ‘sandwich inversion’ of *pi4k* in GB4 was identified. In this variant, the second (inverted) and third copies of *pi4k* appear to contain truncating mutations or partial deletions, likely leading to a complete loss of function. It therefore seems unlikely that this variant produces an increase in relative PI4K expression but perhaps demonstrates a process of pseudogenisation whereby PI4K expression has been lost for all three copies. PI4K itself is of interest as it is the target of imidazopyridines [15], loss of expression may therefore represent an emerging form of resistance.

In contrast, the study described in Chapter Six considered structural variation in 78 trios from malaria endemic Tanzania with a focus applied to genes with known associations with risk of severe malaria and roles in blood antigen systems. I also took this opportunity to explore the role of inheritance in structural variation, finding that approximately 16.7% of each parent's specific structural variants were present without mutation in their children. Broadly, SVs were diverse, but rare, with 91.7% of specific forms being found only once. This low frequency partially reflects uncertainty around the specific position of SVs but is also consistent with active selection against most variants, as has been observed in other studies [16]. Nonetheless a few variants are present at near-fixation suggesting previous selective pressure for these variants. In total only 0.17% (n=328) of variants were found in at least 50% of parents. These include several genes with roles in cell adhesion, intra-cellular transport, and drug binding. Notable examples include *SEC22B* and *BET1*, both of which have significant roles in trafficking from the endoplasmic reticulum to the Golgi complex [17,18]. Additional focus was placed on genes with known roles in blood antigen systems or risk of severe malaria, leading to the identification of SVs in the Junior blood group antigen encoding *ABCG2*, the P<sup>k</sup> antigen forming *A4GALT*, and malaria risk associated *ATP2B4*. Without phenotypic data for the individuals utilised in this study, it was not possible to link these SVs with any specific diseases or phenotypes, though their identification should initiate candidate exploration within the field or laboratory.

In general, the exploration of SVs in the *P. falciparum* and human genomes both identified many rare variants, the majority of which were deletions. A number of these possible constitute false positives, which I sought to filter out via concordance in both detection methods and independent samples. Notably insertions and inversions were excluded from the *P. falciparum* SV project in Chapter Three due to significant levels of

noise within those datasets. In general, the core detection method, DELLY, was more robust for the diploid human genome rather than the haploid *P. falciparum* genome. This distinction is also key to understanding the different uses of predicted phasing scores in both studies. In Chapter Six, phasing is a classic prediction of the dosage for each SV however in Chapter Three it represents a proxy for cryptic mixed infections or poorly resolved loci, as heterozygous status should not be possible for a clonal sample of a haploid species such as the blood stage *P. falciparum*. This hints at the potential for SV-based mixed infection detection methods, similar in approach to existing SNP-based methods such as estMOI [19]. One exception to filtering by predicted phasing was for duplications within *P. falciparum* where the duplication was imperfect, here two copies of a duplication may be assumed as heterozygous when they are highly divergent. This is best demonstrated with *crt*, where divergence between the two copies was supported by haplotype inspection.

Both *Plasmodium* and human focused branches of my work highlighted the poor scaling of existing SV discovery methods, so I sought to develop a method to facilitate multi-population investigation of structural variant in large sample sets. This method is introduced in Chapter Seven in which I highlight the analysis and visualisation components of the project, but partially underpins the work features in Chapters Three and Six. *SV-Pop* consists of two components; an analysis module and a visualisation model. The analysis module takes in structural variants predicted from several discovery methods, integrates population level metadata, and applies both population-based filters and statistical methods on those groups to enhance validity. Those variants can then be passed to the visualisation module, which allows population-population comparison and integration of statistics such as fixation indexes ( $F_{ST}$ ). All metadata is inferred from the inputs provided, reducing the need for manual refinement and allowing the user to jump

straight towards interpretation. The visualisation module can also be used as the foundation for web resources should the user wish to share their dataset with the wider scientific community.

## 8.2 Conclusions

Host-pathogen interactions between *P. falciparum* parasites and their human hosts are underpinned by a range of genomic variation, from single nucleotide polymorphisms to large structural variants. Much focus has been placed upon SNPs, and less on SVs, despite potentially comparable impacts. The work presented here goes some way towards highlighting the need for high-throughput, large-scale investigations of structural variation, as currently occurs for SNPs, and developing methods that facilitate it. It also demonstrates that large genomic changes can be geographically specific, and therefore potentially used in species barcoding alongside SNPs.

For the *P. falciparum* parasite, the work described in Chapter Two, emphasised the ability for chloroquine resistance to be eliminated from a population once that anti-malarial is removed, raising the potential for carefully managed rotations of antimalarials to combat rising resistance. However, the persistence of SP resistance traits several years after the switch to artemisinin-based combination therapies suggests that rotation may not be suitable for all drugs, with some forms of resistance persisting even after the removal of the associated treatment. Novel structural variants were identified that may directly convey antimalarial resistance or compensate for the fitness costs of other resistance variants. Duplications of the *gchl* promoter and the region including *crt* represent significant novel variants that require experimental follow up. Inversions were also explored, with several novel forms potentially conveying resistance. The most significant example of this is the sandwich inversion of *pi4k* in a GB4 sample, which may represent

a disruptive, rather than enhancing, feature given its association with multiple nonsense mutations. Many of these novel findings may help inform future drug recommendations, for example advising against SP use in regions with high levels of the *gch1* promoter duplication if it is found to increase parasite resistance.

For the human studies, SNP associations beyond the well-established examples, such as rs334, were rare and potentially unique to Tanzanians. Key associations were identified between *IL23R* and *IL12RBR2* that suggest a novel, possibly local, immunological protective effect. In general, those variants identified in the GWAS likely reflect both variants of reduced impact, and local responses to specific *Plasmodium* populations. For SVs, many forms exist with only a few being actively selected for. Those at high frequency feature robust signals and exist in roles relating to cell adhesion, intracellular transport, and blood group antigens. Here significant findings were deletions in *SEC22B* and *BET1L*, both of which imply a role in intracellular trafficking from the endoplasmic reticulum to the Golgi complex, and in *ABCG2* which may underpin the rare Jr(a-) blood type. Together these findings may help inform resource allocation by identifying populations at risk of severe response to malaria infection.

### 8.3 Future Work

Being a primarily computational body of work, there is a clear need for laboratory exploration of the variants identified here. Prime examples include the 436 bp *gch1* promoter duplication, sandwich inversions of *pi4k* in GB4, and *SEC22B* deletions in Tanzania. The 436 bp *gch1* duplication is particularly striking as it appears at near fixation in Malawi, whilst also being present at significant frequencies across several other Central (Democratic Republic of Congo 26.3%) and East African populations (Tanzania 78.5%, Kenya 31.6%). PCR confirmation of the duplication, and its specific structure, is required

to confirm this finding *in vitro*. Given the role of *gch1* in the folate pathway, associations of its duplication with SP resistance, and the heavy use of SP in Malawi, it seems likely that this promoter duplication would also be associated with SP resistance. If duplication of the promoter leads to increased expression of the protein, as it likely with *gch1* duplication, this provides a phenotype of compensation for fitness costs associated with *dhps* and *dhfr* mutations which can be experimentally validated. If mutations within *dhps* and *dhfr* genes are associated with reduced efficacy of those proteins, duplications associated with *gch1* may ensure that sufficient levels of folate are produced. Naturally this cannot be properly investigated *in silico*, so classic laboratory-based exploration of the impact of *gch1* promoter duplication on *gch1* expression, for example qPCR comparisons with mutant strains containing zero, one, two or more copies of the promoter region to determine their effect on levels of *gch1* transcription , should be undertaken. If *gch1* expression is found to compensate for *dhps* and *dhfr* mutations, duplications may relate to late-stage ‘locking in’ of SP resistance that would need to be included in active surveillance given the critical role of SP in intermittent treatment of pregnant women and of children (including seasonal malaria chemoprevention).

Another novel duplication identified *in silico* is the putative heterogeneous duplication of the chloroquine resistance transporter, *crt*. Here we find a striking signal in West Africa, and particularly Burkina Faso, where a large 22.9 kbp duplication appears to contain two different forms of the *crt* gene in tandem, one being wildtype and one containing the chloroquine-resistance associated K76T mutation. Given the tendency for K76T to leave *Plasmodium* populations following the removal of chloroquine use, as shown in Malawi but also elsewhere previously, it seems likely that the variant conveys a fitness cost. The presence of two copies of *crt*, one conveying resistance and being beneficial in a chloroquine-rich environment and the other being wildtype and more beneficial in a non-

chloroquine environment, might suggest a compensatory mutation for that fitness cost. Again, follow up classic laboratory work is crucial for characterising this variant, with targeted PCR sequencing being applied to better understand the specific local sequence and its breakpoints, and competition assays to compare the impact of carrying both K76T-positive and K76T-negative copies of *crt* against wildtype and K76T-positive strains in the presence or absence of chloroquine. Tagging of those proteins is also key to determining whether the pair are expressed in parallel, therefore producing heterogeneous expression of CRT, or in response to the presence of chloroquine, suggesting the presence of a potentially novel and complex response to antimalarial detection by the parasite. Once better understood, surveillance of this duplication would need to be undertaken particularly as the presence of both K76T-positive and K76T-negative sequences may cause false positives for studies and diagnosis tests which test for the K76T-negative sequence.

Several novel human variants were also identified in Tanzania, with some being associated with blood antigen and malaria risk genes. These all require follow up confirmation, perhaps through PCR sequencing or additional *in silico* detection, and functional characterisation before further clinical work can be considered. Examples include the 336/7 bp deletions in *ABCG2*, the gene encoding a xenobiotic transporter and the JR blood antigen. Smaller nonsense mutations have been shown to cause the highly rare JR(a-) phenotype [20], and it seems likely that larger deletions could produce a similar phenotype. If so, there may be associations of Jr(a-) status with malaria or similar blood-borne diseases. The GWAS approaches utilised here can also be enhanced through the imputation of further putative SVs, and a move towards a host-pathogen paired-genome analysis. This would require the sequencing of both individuals and the parasites they are infected with. Sequencing both species allows for variation between infecting

parasites, perhaps even for mixed infections, and between human hosts to be adequately controlled for. This would therefore provide additional statistical power for the detection of further novel significantly associated SNPs.

These broad structural variant discovery approaches clearly identify a wealth of novel variants in significant genes. The development of SV-Pop has paved the way for other multi-population studies of structural variation. One clear next step is to perform the approaches utilised in Chapters Three and Six for other *Plasmodium* species, such as *P. vivax* and *P. knowlesi*, and for further human populations. Chapter Seven lays the groundwork for future analysis and exploration of global structural variation of this sort. The same can be said for inversion discovery, as described in Chapter Four, using long-read based assemblies with larger population datasets. Exploration of these datasets, particularly using the SV-Pop visualiser, may help identify further copy number variants with key roles in anti-malarial resistance, host susceptibility, or other disease-relevant phenotypes. Any future studies should seek to integrate advances in long-read based approaches; methods for their utilisation, such as the pipeline described in Chapter Four, therefore need to be actively developed.

## 8.4 References

1. Flegg JA, Metcalf CJE, Gharbi M, Venkatesan M, Shewchuk T, Hopkins Sibley C, et al. Trends in antimalarial drug use in Africa. *Am J Trop Med Hyg*. The American Society of Tropical Medicine and Hygiene; 2013;89:857–65.
2. Laufer MK, Takala-Harrison S, Dzinjalama FK, Stine OC, Taylor TE, Plowe C V. Return of chloroquine-susceptible falciparum malaria in Malawi was a reexpansion of diverse susceptible parasites. *J Infect Dis*. NIH Public Access; 2010;202:801–8.
3. Heinberg A, Siu E, Stern C, Lawrence EA, Ferdig MT, Deitsch KW, et al. Direct evidence for the adaptive role of copy number variation on antifolate susceptibility in *Plasmodium falciparum*. *Mol Microbiol*. NIH Public Access; 2013;88:702–12.
4. Waters AP, Thomas AW, Deans JA, Mitchell GH, Hudson DE, Miller LH, et al. A merozoite receptor protein from *Plasmodium knowlesi* is highly conserved and distributed throughout *Plasmodium*. *J Biol Chem*. 1990;265:17974–9.
5. Akhouri RR, Sharma A, Malhotra P, Sharma A. Role of *Plasmodium falciparum* thrombospondin-related anonymous protein in host-cell interactions. *Malar J*. BioMed Central; 2008;7:63.
6. Nair S, Miller B, Barends M, Jaidee A, Patel J, Mayxay M, et al. Adaptive copy number evolution in malaria parasites. Przeworski M, editor. *PLOS Genet*. Public Library of Science; 2008;4:e1000243.
7. Zhang L, Prather D, Eng J, Crawford S, Kariuki S, ter Kuile F, et al. Polymorphisms in genes of interleukin 12 and its receptors and their association with protection against severe malarial anaemia in children in western Kenya. *Malar J*. BioMed Central; 2010;9:87.
8. Raballah E, Kempaiah P, Karim Z, Orinda GO, Otieno MF, Perkins DJ, et al. CD4 T-cell expression of IFN- $\gamma$  and IL-17 in pediatric malarial anemia. Luty AJF, editor. *PLoS One*. Public Library of Science; 2017;12:e0175864.
9. Timmann C, Thye T, Vens M, Evans J, May J, Ehmen C, et al. Genome-wide association study indicates two novel resistance loci for severe malaria. *Nature*. Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved.; 2012;489:443–6.
10. Griffiths MJ, Shafi MJ, Popper SJ, Hemingway CA, Kortok MM, Wathen A, et al. Genomewide Analysis of the Host Response to Malaria in Kenyan Children. *J Infect Dis*. Oxford University Press; 2005;191:1599–611.
11. Miles A, Iqbal Z, Vauterin P, Pearson R, Campino S, Theron M, et al. Indels, structural variation, and recombination drive genomic diversity in *Plasmodium falciparum*. *Genome Res*. Cold Spring Harbor

Laboratory Press; 2016;26:1288–99.

12. Nakano K, Shiroma A, Shimoji M, Tamotsu H, Ashimine N, Ohki S, et al. Advantages of genome sequencing by long-read sequencer using SMRT technology in medical area. *Hum Cell*. Springer; 2017;30:149–61.
13. Otto TD, Böhme U, Sanders M, Reid A, Bruske EI, Duffy CW, et al. Long read assemblies of geographically dispersed *Plasmodium falciparum* isolates reveal highly structured subtelomeres. *Wellcome Open Res*. 2018;3:52.
14. Otto TD, Rayner JC, Böhme U, Pain A, Spottiswoode N, Sanders M, et al. Genome sequencing of chimpanzee malaria parasites reveals possible pathways of adaptation to human hosts. *Nat Commun*. Nature Publishing Group; 2014;5:4754.
15. McNamara CW, Lee MCS, Lim CS, Lim SH, Roland J, Nagle A, et al. Targeting *Plasmodium* PI(4)K to eliminate malaria. *Nature*. Nature Publishing Group; 2013;504:248–53.
16. Cheeseman IH, Miller B, Tan JC, Tan A, Nair S, Nkhoma SC, et al. Population Structure Shapes Copy Number Variation in Malaria Parasites. *Mol Biol Evol*. Oxford University Press; 2015;33:msv282-.
17. Newman AP, Shim J, Ferro-Novick S. BET1, BOS1, and SEC22 are members of a group of interacting yeast genes required for transport from the endoplasmic reticulum to the Golgi complex. *Mol Cell Biol*. 1990;10:3405–14.
18. Volchuk A, Ravazzola M, Perrelet A, Eng WS, Di Liberto M, Varlamov O, et al. Countercurrent Distribution of Two Distinct SNARE Complexes Mediating Transport within the Golgi Stack. *Mol Biol Cell*. 2004;15:1506–18.
19. Assefa SA, Preston MD, Campino S, Ocholla H, Sutherland CJ, Clark TG. estMOI: estimating multiplicity of infection using parasite deep sequencing data. *Bioinformatics*. 2014;30:1292–4.
20. Tanaka M, Kamada I, Takahashi J, Kimura K, Matsukura H, Tani Y. Defining the Jr(a-) phenotype in the Japanese population. *Transfusion*. Wiley/Blackwell (10.1111); 2013;54:n/a-n/a.