

Inferring the transmission intensity of malaria using genetic data



Jason Alexander Hendry
Green Templeton College
University of Oxford

A thesis submitted for the degree of
Doctor of Philosophy

Michaelmas 2020

Acknowledgements

Heartfelt thanks to:

Gil McVean, Dominic Kwiatkowski, and members of their labs; George Busby, Isaac Ghanai, Kirk Rockett, Christina Hubbart, Davis Mumbengegwi, Dan Bridges, Busiku Hamainza, Jacob Chirwa, Eric Ochomo and Fredros Okumu; the staff and volunteers who contributed to our study in Kaniki; Christophe Fraser and Pete Gething; Grant W. Brown; Tine Goergen; my parents.

Abstract

Despite a recent decline in prevalence, malaria still incurs a substantial burden on global health. Further progress towards elimination will require developing control strategies that are guided by detailed epidemiological surveillance. Yet, many existing surveillance approaches – including estimating parasite prevalence with community-level surveys, or tracking resistance with therapeutic efficacy studies – are time consuming and labour intensive. There is an interest in harnessing the increasing availability of *P. falciparum* genetic data for surveillance, but several challenges remain outstanding. First, the relationships between genetic and epidemiological processes are poorly understood, hindering epidemiological inference from genetic data. Second, the ubiquity of mixed infections complicates analysis of malaria genetic data. Third, how the collection of genetic data can be efficiently integrated into malaria control is still being resolved.

In this thesis, we first elucidate relationships between genetic diversity statistics and epidemiological parameters affecting parasite prevalence, by developing an agent-based simulation that includes a model of parasite DNA. We find many genetic diversity statistics are correlated with parasite prevalence, but that the strength and timelines of these relationships varies. In particular, statistics related to mixed infections are strongly correlated with, and rapidly responsive to, changes in parasite prevalence.

Second, we leverage identity-by-descent patterns present within mixed infections to make inferences about the microepidemiology of malaria. We separate mixed infections comprised of two strains into those generated by a single mosquito bite (co-infection) and those generated by two mosquito bites (super-infection). Applying our approach to the Pf3k data resource, we find evidence for serial co-infection across several hosts, and show that the regional rate of serial co-infection is related to parasite prevalence.

Finally, we develop an approach to genetic data collection that can be deployed in malaria-endemic countries. Using nanopore sequencing and working with local scientists, we infer the drug resistance status of ten samples from Zambia, in-country and in two days. In addition, we present evidence suggesting the approach could be used to identify mixed infections and infer their complexity.

We end by discussing what statistical modelling approaches are most likely to permit parasite prevalence inference in the future, and outline some of the challenges associated with sequencing in the malaria-endemic world.

Contents

List of Abbreviations	xi
1 Introduction	1
1.1 Motivation	1
1.1.1 Thesis aims	3
1.2 Malaria	4
1.2.1 Life cycle of <i>P. falciparum</i>	5
1.2.2 Genome and evolution	8
1.2.3 History and epidemiology	9
1.2.4 Transmission intensity and the environment	10
1.3 Metrics of malaria transmission intensity	13
1.3.1 Basic Reproduction Number, R_0	14
1.3.2 Parasite Rate, PR	14
1.3.3 Entomological Inoculation Rate, EIR	16
1.3.4 Serological Conversion Rate, SCR	17
1.4 Data resources	18
1.4.1 MAP: Transmission intensity data	18
1.4.2 MalariaGEN: Genetic data	22
1.5 Genetic signals of transmission intensity	23
1.5.1 Experimental evidence	23
1.5.2 Theoretical expectations	24
1.6 Conclusion	26
2 Forward-time simulation of <i>P. falciparum</i> transmission and evolution	29
2.1 Introduction	29
2.2 Developing forward-dream	32
2.2.1 Epidemiological layer	34
2.2.2 Infection layer	35
2.2.3 Genome evolution layer	39
2.2.4 Parameterisation and implementation	40
2.3 Genetic diversity and host parasite prevalence at equilibrium	41

2.4	Genetic diversity and host parasite prevalence in times of change . .	46
2.4.1	Malaria control interventions	47
2.4.2	Seasonality	55
2.5	Conclusion	57
3	Epidemiological inference from <i>P. falciparum</i> mixed infections	63
3.1	Introduction	63
3.1.1	Deconvoluting mixed infections	66
3.1.2	Incorporating relatedness between strains	71
3.2	IBD and malaria meiosis	72
3.2.1	Simulating malaria meiosis with pf-meiosis	72
3.2.2	Simulation validation	74
3.2.3	Analysis of oocysts and IBD	76
3.3	IBD and the classes of mixed infections	81
3.3.1	Simulating mixed infection chains with pf-meiosis	81
3.3.2	Classification of mixed infections across Africa and Asia . . .	83
3.4	Mixed infection classification and $PfPR_{2-10}$	86
3.5	Conclusion	88
4	Collecting <i>P. falciparum</i> genetic data from a mobile laboratory in Sub-Saharan Africa	93
4.1	Introduction	94
4.1.1	The evolution of antimalarial drug resistance	94
4.1.2	Monitoring for antimalarial drug resistance in Africa	100
4.1.3	The Land Rover Bursary	102
4.2	Developing a mobile sequencing laboratory	104
4.2.1	Laboratory Protocols	104
4.2.2	Computational Pipeline	113
4.2.3	Mobility, costs, and shipping	115
4.3	Deploying the laboratory in Sub-Saharan Africa	118
4.3.1	Sequencing metrics	120
4.3.2	Error rate analysis	122
4.3.3	Calling known antimalarial drug resistance mutations	125
4.3.4	Downsampling analysis	127
4.4	Discussion	129
4.4.1	Why is artemisinin resistance not already in Africa?	133
4.5	Ethical Approval	135

5	Conclusions	137
5.1	Epidemiology and genetics	138
5.1.1	Model complexity and inference	139
5.2	Mixed infections and transmission	141
5.2.1	Statistical models of mixed infections	142
5.3	Collecting malaria genetic data	144
5.3.1	Data collection and collaboration in Africa	145
	References	147

List of Abbreviations

ABC		Approximate bayesian computation
ACT		Artemisinin-combination therapy
COI		Complexity of infection
DBS		Dried blood spot
DNA		Deoxyribonucleic acid
EIR		Entomological inoculation rate
GWAS		Genome-wide association study
IBD		Identity-by-descent
IBS		Identity-by-state
IRS		Indoor residual spraying
ITNs		Insecticide treated bednets
MCMC		Markov chain Monte Carlo
ONT		Oxford Nanopore Technologies
RBC		Red blood cell
PR		Parasite rate
SCR		Serological conversion rate
SNP		Single-nucleotide polymorphism
SNV		Single-nucleotide variant
<i>spp.</i>		Any species under indicated genus name
WGS		Whole-genome sequencing
WHO		World Health Organization
WSAF		Within-sample allele frequency

The *P. falciparum*, *Anopheles gambiae* and *Homo sapiens* genome sequences have been completed in the last two years, and represent new starting points in the centuries-long search for solutions to the malaria problem.

— Gardner et al. (2002), upon publishing of the *P. falciparum* genome

1

Introduction

Contents

1.1	Motivation	1
1.1.1	Thesis aims	3
1.2	Malaria	4
1.2.1	Life cycle of <i>P. falciparum</i>	5
1.2.2	Genome and evolution	8
1.2.3	History and epidemiology	9
1.2.4	Transmission intensity and the environment	10
1.3	Metrics of malaria transmission intensity	13
1.3.1	Basic Reproduction Number, R_0	14
1.3.2	Parasite Rate, PR	14
1.3.3	Entomological Inoculation Rate, EIR	16
1.3.4	Serological Conversion Rate, SCR	17
1.4	Data resources	18
1.4.1	MAP: Transmission intensity data	18
1.4.2	MalariaGEN: Genetic data	22
1.5	Genetic signals of transmission intensity	23
1.5.1	Experimental evidence	23
1.5.2	Theoretical expectations	24
1.6	Conclusion	26

1.1 Motivation

Malaria is the world's most deadly parasitic disease, exacting a tremendous toll on global health and stymieing development in some of the poorest regions of the world

(WHO 2019). Since 2000, efforts driven by the Roll Back Malaria Initiative and United Nations Millennium Development Goals have produced marked reductions in the global malaria burden, decreasing prevalence by 37% (Bhatt et al. 2015). Yet, despite this impressive effort, in 2018 the World Health Organization (WHO) estimated 228 million cases of malaria and 405 thousand deaths, over 67% of which were children under the age of five (WHO 2019). 93% of these deaths occurred in Africa (WHO 2019), where the prevalence of malaria is still hyper- or holo-endemic ($>50\%$) in several areas (Bhatt et al. 2015). Moreover, regions of Southeast Asia are witnessing a progressive decline in the efficacy of the WHO recommended first-line antimalarial treatment, artemisinin-based combination therapy (ACT), which recent research has attributed to the evolution of several expanding artemisinin- and partner drug resistant sub-populations (Dondorp et al. 2009; Denis et al. 2006; Miotto et al. 2015; Imwong et al. 2017; Amato et al. 2018). Both the persistence of high-transmission areas in Africa and the evolution of resistant sub-populations in Southeast Asia highlight the need for revitalized strategies to combat the disease.

Further progress towards global malaria eradication will require increasingly nuanced implementation of intervention strategies guided by accurate global maps of malaria transmission intensity, migration patterns, and resistance. Such maps will be the product of routine and large-scale malaria surveillance. Indeed, the WHO emphasised the importance of surveillance in their Technical Strategy of Malaria 2016-2030, where it is one of the three central pillars of their overall strategy (WHO 2015). Critically, not only must surveillance approaches be amenable to widespread deployment, but they also need to be robust even in low transmission intensity regimes, where sparse data can undermine existing surveillance approaches (Reiner et al. 2015a; Sturrock et al. 2016).

In light of this need, genomic data stands as a promising yet largely untapped resource. In recent years, global consortial efforts have collected and sequenced thousands of malaria parasites from patient blood samples, providing invaluable insight into the global genetic architecture of malaria (Network 2008; Manske et al. 2012; Amambua-Ngwa et al. 2019). It is clear that this genetic data also

harbours information useful for malaria surveillance: migration patterns, including local/imported case assignments, could be distinguished by analysis of the genetic relatedness between different regions (Tessema et al. 2019a; Tessema et al. 2019b); antimalarial drug resistance can be assessed by querying for mutations in known resistance genes (Ishengoma et al. 2019); and transmission intensity may be related to overall levels of genetic diversity or rates of mixed infection (Anderson et al. 2000; Daniels et al. 2013). In addition, when transmission declines, genetic data allows more information to be derived from a single case, helping counterbalance issues of data sparsity. In recognition of this potential, there has been a call for developing a "genetic epidemiology" of malaria (Volkman et al. 2012; Nkhoma et al. 2013; Wesolowski et al. 2018; Dalmat et al. 2019). But at present, the integration of genetic data with epidemiological inference and mapping is still in its early stages.

1.1.1 Thesis aims

In this thesis, I explore the prospects for *P. falciparum* genetic data to support malaria surveillance. The larger challenge is presented diagrammatically in Figure 1.1 - how can we infer epidemiology from geolocated parasite genetic data? Moreover, how can we collect this genetic data, in a minimally-invasive, efficient and cost-effective manner? Within Figure 1, the specific aims of this thesis are highlighted.

In Chapters 2 and 3, our primary aim is to better understand how transmission intensity could be inferred from genetic data – first by investigating how a broad panel of genetic diversity statistics respond to variation in key epidemiological parameters affecting parasite prevalence (Chapter 2); and then by focusing on mixed infections (Chapter 3). Transmission intensity is a critical surveillance parameter as it is used by many national malaria control programs to stratify interventions to target high burden regions. In Chapter 4, our aim was to develop an approach by which genetic data can be collected from dried blood-spots (DBS) and in-country. Within the resource limits of malaria control we need practicable approaches for genetic data collection, otherwise epidemiological modelling from genetic data will be fruitless.

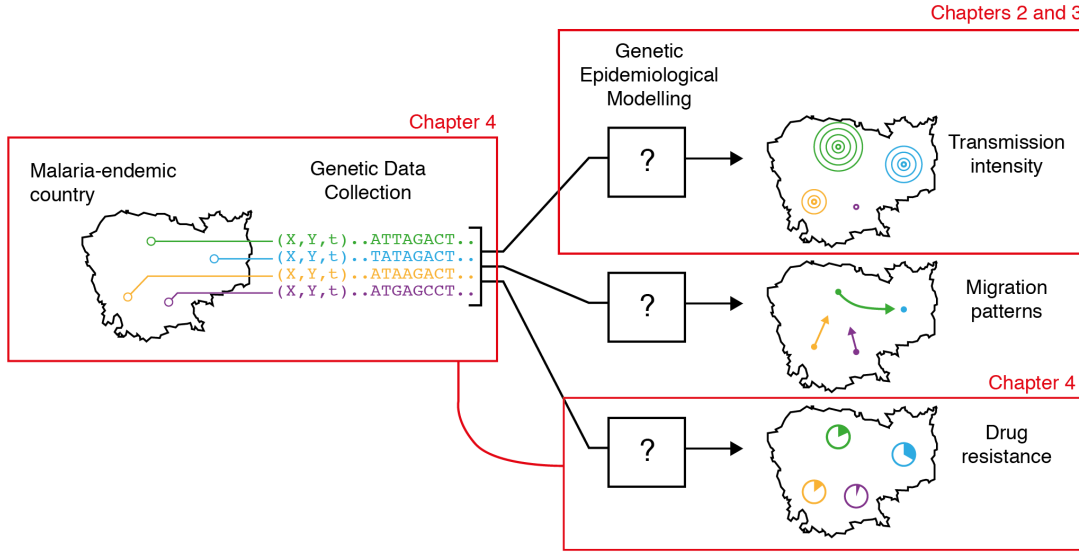


Figure 1.1: The challenge of malaria genetic epidemiology and the focus of this thesis. Left, *P. falciparum* genetic data is collected from patient blood samples from around a malaria endemic country. Metadata includes spatial location X, Y and time of collection t . With genetic data and metadata as input, genetic epidemiological models are used to learn about disease epidemiology across the country, such as transmission intensity, migration patterns, and drug resistance. These models are still under development and represented by "?". Chapters 2 and 3 focus on inferring transmission intensity from genetic data, Chapter 4 focuses on genetic data collection and inferring drug resistance.

The remainder of this chapter provides background material on malaria (Section 1.2), metrics of transmission intensity (Section 1.3), the primary sources of data used in this thesis (Section 1.4), and current published works addressing the inference of transmission intensity from genetic data (Section 1.5).

1.2 Malaria

Malaria is an infection of the blood caused by protozoan¹ parasites of the genus *Plasmodium*, which are transmitted by the bite of female *Anopheles spp.* mosquitoes (White 2014). There are four species of malaria which preferentially infect humans: *P. falciparum*, *P. vivax*, *P. ovale*, and *P. malariae* (White 2014). Of these, both *P. vivax* and *P. falciparum* cause significant morbidity, although nearly all severe disease and death is caused by *P. falciparum* (Miller et al. 1986; White 2014).

¹Protozoa are single-celled eukaryotes.

1.2.1 Life cycle of *P. falciparum*

The life cycle of *P. falciparum* involves both a host stage, which is entirely asexual, and vector stage, during which meiosis occurs. The life cycle is complex and involves *P. falciparum* differentiating through a variety of cell types, as well as infecting a variety of host and vector cells and compartments (Figure 1.2). I provide a very brief outline of this life cycle below. Interested readers can find more information in the following materials: (Prudencio et al. 2006; Cowman and Crabb 2006; Vaughan et al. 2008; J. D. Smith et al. 2013; White 2014; Beier 1998).

Host phase

The host phase of the life cycle begins when a human host is bitten by a female² *Anopheles spp.* mosquito carrying parasites in its salivary glands (Prudencio et al. 2006; Cowman and Crabb 2006; White 2014). During the bite, the mosquito will release saliva into the host, containing anti-coagulants and vasodilators which promote the uptake of blood, as well as *P. falciparum* parasites in the infective, sporozoite form (Cowman and Crabb 2006; Prudencio et al. 2006). Experimental work in rodents has suggested that somewhere between 10 and 1,500 sporozoites will be injected subcutaneously (Medica and Sinnis 2005; Jin et al. 2007), migrate into nearby blood vessels and travel to the liver, where they eventually invade hepatocytes (Prudencio et al. 2006; Vaughan et al. 2008). An approximately six-day incubation period within the liver follows (Bartoloni and Zammarchi 2012; White 2014; Huber et al. 2016), where each sporozoite differentiates into thousands or tens of thousands of merozoites (Prudencio et al. 2006; Cowman and Crabb 2006; Vaughan et al. 2008). After the incubation period, the merozoites burst from the hepatocytes and invade the bloodstream, beginning the blood stage of the infection (Prudencio et al. 2006). The liver stage of the infection is asymptomatic (White 2014).

During the blood stage, merozoites infect erythrocytes (red blood cells, RBCs) and consume the hemoglobin within them (Francis et al. 1997). A single merozoite will infect a single RBC (Cowman et al. 2012). Once the merozoite is inside, it

²Only female vectors seek blood-meals (Bates 1970).

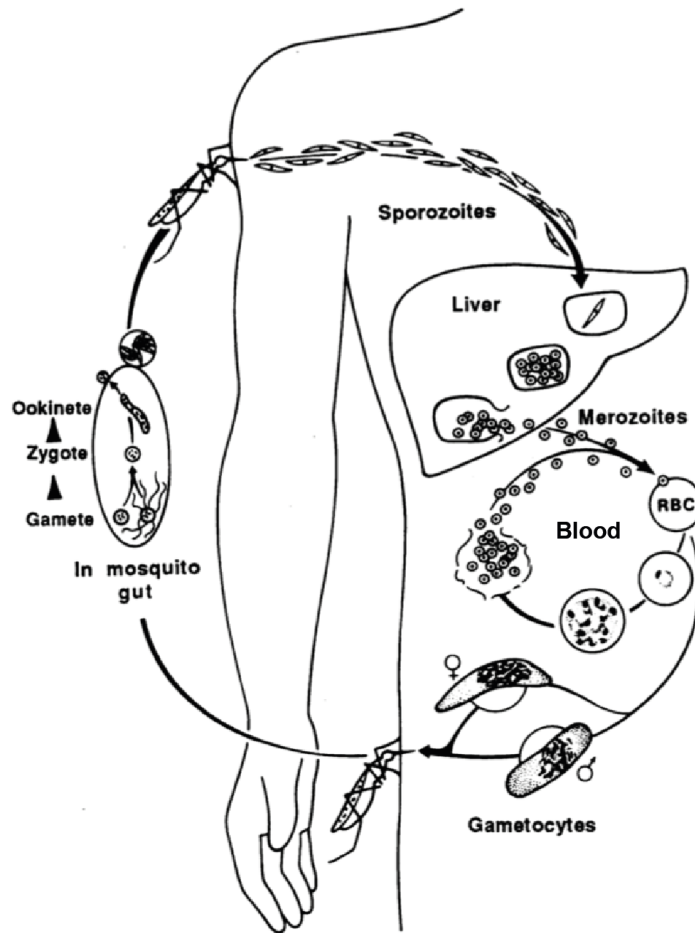


Figure 1.2: *P. falciparum* Life cycle. Source: LH Miller *et al.*. Research toward malaria vaccines. Science (1986), DOI: 10.1126/science.2431481.

catabolizes hemoglobin to promote its replication (Francis et al. 1997). Over the course of the next approximately 48hrs (White 2014), this merozoite will multiply, producing 16 to 32 merozoites (Cowman and Crabb 2006), and passing through a variety of cell stages which are visible microscopically (White 2014). The growth of merozoites causes deformation of the RBC they inhabit, which can lead to their clearance by the spleen (J. D. Smith et al. 2013). To counteract this, *P. falciparum* produces the protein *PfEMP1* that is localized to the RBC surface and which mediates sequestration of RBCs in the microvasculature via cytoadherence (J. D. Smith et al. 2013). Ultimately, the RBC ruptures and for a brief interval (~ 60 s) merozoites enter the plasma before infecting new RBCs (Cowman et al. 2012). This process repeats itself resulting the exponential growth of *P. falciparum*

parasites to parasitemia levels as high as 10^4 - 10^5 per uL (White 2014). Malaria symptoms are associated with the blood-stage, the most characteristic of which are periodic fevers and enlargement of the spleen. Death can occur as a result of coma (known as cerebral malaria) or severe anemia (White 2014).

Within the host blood, merozoite replication is asexual. However, some fraction of the merozoites undergo differentiation into a pre-sexual cell stage, called a gametocyte, which exists in both male and female forms. In *P. falciparum*, mature gametocytes are visible in the blood approximately 7-15 days following merozoite release from the liver (Bousema and Drakeley 2011). There is evidence that their production increases at higher parasite densities or in the presence of antimalarial drugs; factors which have led to the hypothesis that gametocyte formation is induced by conditions unfavourable for asexual replication (reviewed in Bousema and Drakeley (2011)). If an infected host carrying gametocytes receives an additional mosquito bite, there is a possibility that gametocytes of both genders will be imbibed in the mosquito bloodmeal. This initiates the vector phase of the life cycle.

Vector phase

The vector phase is referred to as sporogony (Beier 1998). It begins when a mosquito vector takes a bloodmeal containing gametocytes of both gender (Beier 1998). The bloodmeal is drawn into the midgut (stomach) of the mosquito vector, and here the gametocytes differentiate into gametes and fuse to form a zygote (Ghosh et al. 2000). The zygote then differentiates into a motile cell form known as an ookinete, which penetrates the midgut wall, escaping the digestive processes occurring inside (Beier 1998; Ghosh et al. 2000). The major meiotic steps, including DNA replication, cross-over, and segregation, occur in the zygote and ookinete; this is covered in detail in Section 3.2. By the time the ookinete has reached the outer surface of the midgut wall meiosis will have completed and the ookinete will differentiate into a cyst-like cell form, called an oocyst (Beier 1998; Ghosh et al. 2000). Within the oocyst, sporozoites form through asexual replication over the next 10-12 days (Aly et al. 2009). Afterwards, the oocyst will burst, releasing thousands of sporozoites

which travel to the mosquito salivary glands (Aly et al. 2009). Depending on the number of gametocytes in the bloodmeal, this process can occur multiple times in parallel, and the dissection of infected wild-caught mosquitoes frequently results in the identification of multiple oocysts on the midgut wall (Beier et al. 1991; Collins et al. 1984). A mosquito harbouring sporozoites in its salivary glands is infectious and can continue the life cycle by transmission to a host.

1.2.2 Genome and evolution

The nuclear genome of *P. falciparum* consists of 23.3-Mbp of DNA distributed along 14 chromosomes (Gardner et al. 2002). There is additional DNA material in the mitochondria and the apicoplast³ (Gardner et al. 2002). As described above, *P. falciparum* is haploid for the entire host phase and most of the vector phase of its life cycle.

Plasmodium spp. are amongst the most A+T rich eukaryotic organisms currently known (Videvall 2018) and *Plasmodium falciparum* has a mean A+T content of 80.6% reaching $\sim 90\%$ in intronic regions (Gardner et al. 2002). This results in long and frequent stretches of highly repetitive DNA, which drives an exceptionally high indel mutation rate of 21.1×10^{-10} per ieg⁴ per bp (Hamilton et al. 2017). The single nucleotide variant (SNV) mutation rate ranges between 1.64×10^{-10} and 3.20×10^{-10} per ieg per bp, roughly ten-fold lower than the indel mutation rate (Hamilton et al. 2019; McDew-White et al. 2019).

The meiotic recombination rate has been estimated at 13.5Kbp/cM, which is consistent with a single-cross over event per bivalent⁵, as is necessary for homologous chromosome pairing during meiosis (Miles et al. 2016). The rate of non-crossover recombination events (also known as gene conversion) is roughly half that of cross-over recombination events (Miles et al. 2016).

³An organelle that is homologous to the chloroplast in plants.

⁴Intraerythrocytic generation, or the time from red blood cell infection to lysis, roughly 48hrs.

⁵A bivalent is a four-chromosome structure consisting of the sister chromatids of a pair of homologous chromosomes. It occurs when the cell is at 4n, and before segregation begins.

1.2.3 History and epidemiology

Several hypothesis have been put forth to explain the origin of *Plasmodium falciparum* – including that it derived from ancient humans (Escalante and Ayala 1994), chimpanzees (Rich et al. 2009), or bonobos (Krief et al. 2010). However, a 2010 study of *Plasmodium* DNA, collected from nearly 3,000 wild-primate faecal samples in Africa, found that the entire *Plasmodium falciparum* lineage subtended from within a clade of parasites that infect western gorillas (X. Liu and Fu 2015). This is consistent with *P. falciparum* arising through a single cross-species transmission event from gorillas to humans (W. Liu et al. 2010; X. Liu and Fu 2015). Though difficult to date, the authors have estimated that the observed levels of population genetic diversity in *P. falciparum* could be generated within 10,000 years (Sundararaman et al. 2016).

Traces of malaria can be found throughout ancient history, with a multitude of civilizations having described diseases with its characteristic symptoms. It is described in what is likely the oldest known medical text, the Chinese "Nei Ching" or "Canon of Internal Medicine", from 2700BC (Neghina et al. 2010; Shampo and Kyle 1989). Other civilizations, including the Sumerian, Assyrian, Babylonian, Chinese, Egyptian, Indian, Greek, Roman and Arabic all make references to diseases consistent with malaria over the next several thousand years (Neghina et al. 2010). The name "malaria", which is Italian for "bad air", derives from miasmatic theory, proponents of which believed the disease was caused by noxious vapours emanating from marshes (Neghina et al. 2010). The germ theory of infection was established in the 1870s by Louis Pasteur and Robert Koch and in 1880, Charles Louis Alphonse Laveran, a physician in the French Army working in Algeria, characterised parasites in the blood of patients with malaria (Cox 2010; White 2014). In 1897, Ronald Ross, a military surgeon working in India, established that malaria was transmitted between hosts by mosquito (Cox 2010). He deduced this by experimentally feeding mosquitoes on malaria patients, thereby identifying "spores" on their midgut walls and "rods" within their salivary glands (Cox 2010). Alphonse Laveran and Ronald Ross both received Noble Prizes for their work (White 2014).

It has been suggested that the maximum natural (i.e. pre-intervention) geographical extent of malaria was reached around 1900, when an estimated 53% of the world's land surface was at risk (reviewed in Hay et al. (2004)). This included South America, Africa, India, and Southeast Asia, as well as Central America and parts of North America, and much of Europe and the Middle East (Hay et al. 2004). From 1900, economic development and control of mosquito habitats coincided with malaria decline, resulting in a reduction of the malarious area to an estimated 40% of the world's land surface (Hay et al. 2004). In 1955, the Global Malaria Eradication Program was launched, buoyed by the belief that eradication could be achieved by focusing on vector control with the insecticide dichloro-diphenyl-trichloroethane (DDT), which was first used during World War II (Najera et al. 2011). However, in 1969 the program was abandoned, having eliminated malaria from most of Europe and North America, but not from Africa, Asia or South America (Najera et al. 2011).

From 1980 onward, there was a gradual increase in deaths due to malaria, reaching a peak of approximately 1.8m deaths annually in 2004 (Murray et al. 2012). A significant contributor to this rise was the emergence of resistance to the then first-line antimalarial, chloroquine (Murray et al. 2012). From 2004 onward, malaria prevalence has declined globally (Murray et al. 2012; Bhatt et al. 2015; Weiss et al. 2019), likely driven by the increased funds rallied in 2000 as part of the Roll Back Malaria Initiative and United Nations Millennium Development Goals (Murray et al. 2012; Bhatt et al. 2015). In addition, widespread use of ACT began in 2004 (WHO 2006), and until recently was highly effective. The extensive use of insecticide treated bednets (ITNs) and indoor residual spraying (IRS) has also contributed substantially to the decline, with ITNs averting the most cases overall (Bhatt et al. 2015).

1.2.4 Transmission intensity and the environment

Modern-day malaria is largely confined to the tropics (Figure 1.3) (Weiss et al. 2019). Its geographical extent and transmission intensity is strongly influenced by the climatic and environmental suitability for *Anopheles spp.* vectors (Craig

et al. 1999). There are 476 formally recognised species of the *Anopheles* genus (Harbach 2013) and, of these, 41 transmit malaria at a level considered relevant for public health (Hay et al. 2010). Species can have both differing geographical distributions and behaviours (Sinka et al. 2012). In Africa, the dominant vectors are *An. gambiae*, *An. funestus*, and *An. arabiensis*, whereas in Asia there is a great diversity of *Anopheles spp.* vectors (Sinka et al. 2012). Behavioural differences in species can impact the efficacy of control interventions. For example, in Asia, most vectors are outdoor-biting, limiting the efficacy of indoor residual spraying (IRS) and insecticide treated bednets (ITNs) (St. Laurent et al. 2016).

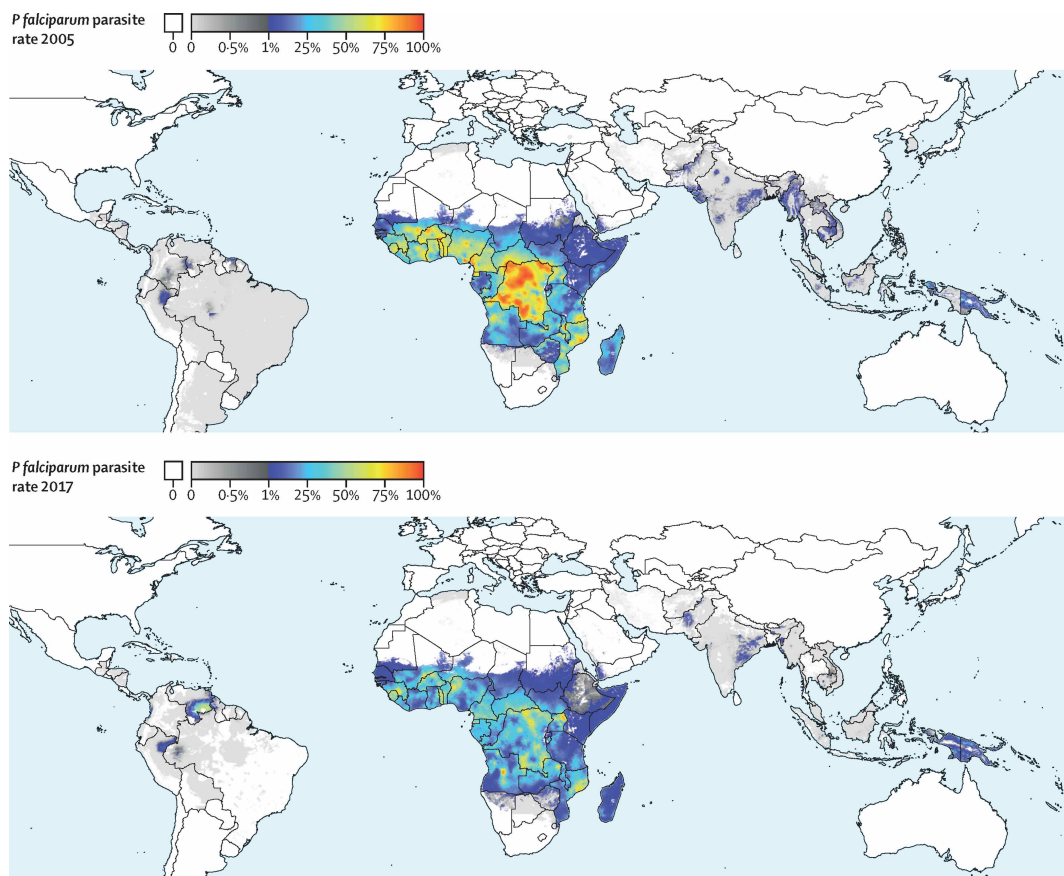


Figure 1.3: Global distribution of *P. falciparum* in 2005 and 2017. Areas without endemic malaria are shown in white. Source: DJ Weiss *et al.* Mapping the global prevalence, incidence and mortality of *Plasmodium falciparum*, 2000-17: a spatial and temporal modelling study. Lancet (2019), DOI: 10.1016/S0140-6736(19)31097-9.

The two factors that most strongly influence a regions suitability for *Anopheles spp.* are its temperature and rainfall (Craig et al. 1999). Endemic *P. falciparum*

requires continuous transmission⁶; critically, infected *Anopheles* must survive long enough for *P. falciparum* to complete its vector-stage development and to deliver a second bite (White 2014). It has been estimated that wild adult *Anopheles gambiae* have a daily survival rate of between 0.91-0.93; a lifespan of around two weeks (Davidson 1954). Sporogony typically takes 8-10 days, but proceeds quicker at higher temperatures and below 16°C does not occur (Craig et al. 1999). Thermal death of *Anopheles spp.* occurs at 40°C, but above 32°C survival rate sharply decrease (Craig et al. 1999). Overall, ideal temperatures are roughly between 20-30°C (White 2014).

Rainfall produces the breeding sites for *Anopheles spp.*, who lay their eggs, and undergo larvae and pupal development on water surfaces (Bates 1970). The relationship between amount of rainfall and vector density is ill-defined for several reasons (Craig et al. 1999). First, different species may prefer different water bodies (reviewed in Rejmnkov et al. (2013)). For example, *An. gambiae* prefers stagnant temporary water, such as puddles, whereas *An. funestus* prefers permanent bodies of water, such as small streams or ponds (Ginnig et al. 2001; Ndenga et al. 2011). Second, though rainfall promotes the formation of breeding sites, excessive rainfall can wash away larvae and pupae (Paaajmans et al. 2007). Third, the small-scale hydrology of an area is of relevance, but is both complex and changeable, influenced by many factors; including human activities like dam-building, deforestation, and irrigation (reviewed in Rejmnkov et al. (2013)). As a result, vector density (and transmission intensity) can be heterogeneous on very small spatial and temporal scales.

The dependence on temperature and rainfall leads to seasonal transmission of malaria in many places of the world, including Sahel and parts of Southern Africa (Reiner et al. 2015b). Typically, malaria transmission peaks shortly after the wet season and falls during the dry season.

⁶Unlike *P. vivax*, which has a dormant phase.

1.3 Metrics of malaria transmission intensity

A central motivation of this thesis is to improve upon existing methods of estimating transmission intensity. Transmission intensity is a general term encompassing a variety of metrics aimed at describing the rate at which malaria parasites are spreading through a population. It plays a central role in many other aspects of malaria epidemiology, influencing the age at first infection, the frequency of different disease syndromes, the prevalence of adaptive immunity, and the incidence of mortality (Snow and Marsh 2002). Moreover, it allows policy-makers to objectively gauge the need for, and efficacy of, malaria control strategies (Tusting et al. 2014).

Despite its overarching importance, measuring transmission intensity in the field is fraught with challenges. Estimating transmission intensity from clinical incidence at health facilities is complicated by variation in treatment seeking behaviour, incomplete medical record keeping and diagnostic errors (Bhatt et al. 2015). In low transmission intensity settings, prohibitively intensive sampling is often needed to make accurate measurements (Sturrock et al. 2016), whereas in high intensity regions many metrics saturate as a result of multiple infections, acquired immunity and heterogeneous biting (Tusting et al. 2014). Metrics derived from sampling mosquitoes suffer as a result of small-scale variability in mosquito density and a lack of standardisation among mosquito capturing methods, while those calculated from the presence of parasites in human blood must grapple with strongly infection-course and age-dependent variations in parasitemia (Tusting et al. 2014).

Genomic data may provide an alternate means to estimating transmission intensity. To this end, an understanding of existing methods for measuring transmission intensity is an important prerequisite to framing how genomic data could benefit its estimation. Below, I review the primary theoretical metric of transmission intensity, the basic reproduction number (R_0), the two most frequently field-collected metrics, parasite prevalence (PR) and the entomological inoculation rate (EIR), and then review an existing molecular metric known as the serological conversion rate (SCR). For a more thorough discussion of all major transmission intensity metrics, readers can consult the review published by Tusting et al. (2014).

1.3.1 Basic Reproduction Number, R_0

Generally, the basic reproduction number is defined as the number of secondary infections that would result from the presence of a single infected individual in a naive population (Dietz 1993). It was first applied to malaria epidemiology in 1952 by Macdonald as the ‘basic reproduction rate’ (Macdonald 1952), but has since been re-termed the ‘basic reproduction number’ as the quantity is unitless and does not explicitly specify a ‘per unit time’ as implied by ‘rate’ (Elandt-Johnson 1975). The term provides a simple threshold condition for areas in which transmission is stable or unstable, namely, $R_0 > 1$ signifies that the number of infections in the population is increasing, and $R_0 < 1$ signifies a decreasing number of infections (Dietz 1993). Additionally, given a specific R_0 value greater than one, the fraction of the population that must be protected from infection in order to promote eventual eradication can be calculated as $1 - 1/R_0$ (Dietz 1993; D. L. Smith et al. 2007a). This clear interpretability has made R_0 a theoretical gold-standard for transmission intensity. In practice, however, R_0 is difficult to estimate accurately: we are largely unable to measure it directly, and classical formula for its calculation are reliant on several poorly characterized parameters (Dietz 1993; D. L. Smith et al. 2007a). As evidence of this, despite the parameters’ long history few estimates were made before 2007, when a benchmark paper by D. L. Smith et al. (2007a) produced 121 estimates across Africa, with values ranging from 1 to over 300,012. These estimates were built by deriving equilibrium relationships between R_0 and other more easily gathered transmission intensity metrics, including parasite rate (PR) and entomological inoculation rate (EIR) as discussed below (D. L. Smith et al. 2007a).

1.3.2 Parasite Rate, PR

Parasite rate, or parasite prevalence⁷, is defined as the proportion of individuals within a population containing parasites in the blood (Tusting et al. 2014). Commonly, blood samples are analysed for the presence of parasites from samples

⁷For the majority of this thesis, parasite prevalence or host parasite prevalence (when there could be ambiguity about whether parasite prevalence refers to hosts or vectors) is used.

collected across a whole community or school (Tusting et al. 2014). Increasingly, national household surveys are including diagnostic tests for malaria and thus can be aggregated to produce PR estimates (Bhatt et al. 2015). The relatively straightforward nature of PR has resulted in it being the traditional metric used to classify regional malaria endemicity and the most ubiquitously collected of all transmission intensity metrics, with over 22,000 estimates made from 1985 to 2010 (Bhatt et al. 2015; Hay et al. 2008).

Despite its ubiquity, PR suffers from several well-documented shortcomings. Firstly, the detection of parasites within blood is limited by sensitivity issues. For example, blood-films used for microscopy-based parasite detection examine only a very small fraction of the total blood volume, and those patients with lower parasitemia are susceptible to being read as false negatives (Tusting et al. 2014). Alternative approaches to parasite detection can be more sensitive, such as PCR-based methods or rapid diagnostic tests, however this variation in sensitivity can be confounding in studies that aggregate PR surveys. Beyond sensitivity issues, blood-based assays for parasite presence must also grapple with infection-course and age-dependent variation in parasitemia (Tusting et al. 2014). Levels of parasitemia during an infection can fluctuate dramatically on daily and even hourly time-scales in response to fever and the interplay between the parasite population and host immune system (O’Meara et al. 2007). As a result, PR measurements are subject to error that tends to underestimate true parasitemia values (O’Meara et al. 2007). Moreover, a strong age-dependency of parasitemia, driven by the presence of maternally-contributed antibodies in young children and adaptive immunity in adults, results in average PR values being influenced by the age-structure of the underlying population (D. L. Smith et al. 2007b).

Modern research has made strides to address many of the described shortcomings with PR , developing both age-standardised measures and adjusting estimates based on season and diagnostic approach (D. L. Smith et al. 2007b; Gething et al. 2011). Still, PR measurements may suffer in regions of very high or very low transmission intensity. In high intensity transmission settings, PR values tend to saturate as a

result of multiple infection, acquired immunity and heterogeneous biting (Tusting et al. 2014). In low intensity settings, specifically in regions where $PR \leq 5\%$, prohibitively extensive sampling is needed to produce confident results (Tusting et al. 2014; Hay et al. 2009).

1.3.3 Entomological Inoculation Rate, EIR

The entomological inoculation rate is defined as the number of infectious bites received by an individual per annum (Tusting et al. 2014). It is a standard measure of transmission intensity and has been collected relatively often, with 223 measurements made between 1980 and 2000 (Hay et al. 2000). Computing EIR relies on collecting mosquitoes at landing catches, of which there are three primary forms: human landing catches, light trap catches, and pyrethrium spray catches (Tusting et al. 2014). The salivary glands of captured mosquitoes are subsequently dissected and probed for sporozoite-stage malaria parasites. Within such a setup, the EIR is calculated as:

$$EIR = \frac{M_{t+}}{M_t} \frac{M_c}{L}$$

where M_{t+} is the total number of sporozoite positive mosquitoes, M_t is the total number of dissected mosquitoes, M_c is the number of collected mosquitoes, and L is the number of landing catches (Tusting et al. 2014). Note that the equation is equal to the product of the sporozoite positive rate (fraction of mosquitoes carrying sporozoites in their salivary glands) by the human biting rate (number of bites per host per unit time).

The EIR is subject to two major sources of variability that considerably affect its precision and accuracy. Firstly, mosquito density is highly variable, both across small spatial distances (e.g. proximity to still water) and in response to environmental factors (such as recent rainfall), complicating the estimation of the human biting rate (Tusting et al. 2014). Secondly, the EIR is lacking in standardised collection guidelines, as exemplified by the three different land catches employed (Kilama et al. 2014). Variability in catch usage (and even indoor- vs.

outdoor- placement) can cause major albeit poorly characterized variation in the number of sampled mosquitoes (Overgaard et al. 2012; Stevenson et al. 2012). Both variability in mosquito density and lack of collection standardisation challenge the comparison of *EIR* values. Additionally, accurate *EIR* estimates are hard to obtain in low transmission intensity settings as a result of difficulties in catching sufficient mosquitoes compounded by generally low sporozoite positive rates ($< 5\%$ even in highly endemic regions) (Tusting et al. 2014; Sturrock et al. 2016).

1.3.4 Serological Conversion Rate, *SCR*

The serological conversion rate describes the prevalence of antimalarial antibodies within a population and thus is an indicator of cumulative exposure to the parasite (Tusting et al. 2014). It is computed by fitting an age-stratified reversible catalytic model of seropositivity (the presence of antibody within the blood) across a population (Corran et al. 2007). Seropositivity is assessed via ELISA against one or a set immunogenic malaria antigens, such as *CSP*, *MSP-119* or *AMA-1* (Corran et al. 2007; Cook et al. 2011).

The longer lasting nature of antimalarial antibodies (from weeks to months) confers a key advantage upon *SCR* as a metric of transmission intensity: the metric integrates population malaria exposure over time making it robust to rapid fluctuations in transmission intensity and allowing better signal detection in low transmission intensity settings. However, the inertial nature of seropositivity also implies that it will be less responsive to meaningful short-term changes in transmission intensity and, in the case of local malaria eradication, may retain signal in accordance with the half-life of the antimalarial antibody being assayed (Tusting et al. 2014). *SCR* has been collected in several countries across Africa and Southeast Asia, though it lacks the widespread usage of *PR* or *EIR* (Drakeley and Cook 2009). Additionally, as *SCR* measures the presence of a host immune response to infection, it is possible that it may be influenced by host or parasite genetics. Suggestive of this, a study of sympatric ethnic groups from Mali and

Burkina Faso found they exhibited differences in the seroprevalence of *CSP* despite experiencing similar levels of malaria exposure (Bolad et al. 2005; Dolo et al. 2005).

Summary

Transmission intensity varies tremendously across malaria endemic regions and has a profound impact on the overall epidemiology of the disease. Though a suite of metrics exist to facilitate its estimation, all have their shortcomings: R_0 , though a theoretical ideal, is difficult to calculate in the field; PR is relatively easy to estimate but is challenged by dynamic changes in blood parasites and accuracy issues in high- and low-intensity settings regions; EIR is confounded by a lack of standardisation and also struggles to produce reliable results in low-intensity regions; in contrast SCR performs well in low-intensity regions, but its time-integrated nature means it is unable to provide real-time responses to changes in endemicity, and may be confounded by the genetics of the underlying host populations.

1.4 Data resources

There are two large-scale projects that have produced public data resources which have been used in this thesis: (1) the Malaria Atlas Project (MAP), which contains data on transmission intensity and intervention coverage, as well as a suite of malaria-relevant co-variates (available at: www.malariagen.net); and (2) the Malaria Genomic Epidemiology Network (MalariaGEN), which hosts genomic data on malaria and its vectors (available at: malariaatlas.org). The two projects are described briefly below.

1.4.1 MAP: Transmission intensity data

The Malaria Atlas Project (MAP) has been leading the effort to integrate data on malaria transmission intensity into global spatiotemporal maps both to guide and assess malaria intervention strategies (Bhatt et al. 2015; Gething et al. 2011; Hay et al. 2009). In the process of producing these maps, the project has made several major contributions to malaria epidemiology including (1) conducting an

exhaustive assembly effort of global *PR* data (Guerra et al. 2007), (2) systematically constructing a set of environmental factors that contribute to spatial variability in *PR* (Weiss et al. 2015), (3) computing global surfaces of *EIR*, R_0 and clinical incidence via their relationship with *PR* (Bhatt et al. 2015; Gething et al. 2011) and (4) implementing modern approaches in model-based geostatistics within a Bayesian statistical framework to compute spatial surfaces and their uncertainty.

Prior to the Malaria Atlas Project, maps of malaria risk were constructed by exploiting environmental dependencies of the vector and parasite (such as mean rainfall and temperature) to define regions of “climatic suitability” for stable malaria transmission (Craig et al. 1999). Though capable of broadly recapitulating historical transmission trends, these maps provided only a course-grained distribution of malaria risk, were built from subjectively selected environmental factors, and failed to incorporate survey data on malaria transmission intensity (Craig et al. 1999). In contrast to this, MAP has been constructing high-resolution (5x5km grid) empirical transmission intensity maps by aggregating *PR* data from three main sources: (1) periodic exhaustive literature searches of geolocated community-level *PR* surveys that conform to strict inclusion criterion, (2) personal communications with field-experts and (3) malaria incidence data derived from demographic household surveys (Bhatt et al. 2015; Gething et al. 2011; Hay et al. 2009; Guerra et al. 2007). From this data, MAP published three global malaria endemicity maps of an age-standardized *Plasmodium falciparum* parasite rate ($PfPR_{2-10}$) in 2007, 2010 and in 2019 (Gething et al. 2011; Hay et al. 2009; Weiss et al. 2019). In addition, a 2015 MAP release specifically covered continental Africa, and was generated from 27,573 geolocated point-observations of $PfPR_{2-10}$ (Bhatt et al. 2015).

In addition to undertaking a massive data assembly effort, the MAP project sought to systematically construct a set of environmental covariates that were most accurately predictive of $PfPR_{2-10}$. Though a large suite of environmental factors – such as temperature, rainfall, land usage, surface moisture, vegetation cover, elevation, etc. – had been previously employed as co-variates in studies of malaria risk, the selection of a particular set of covariates for a given study was often

subjective and without quantitative justification (Weiss et al. 2015). To address this issue, the MAP authors compiled a base set of 32 environmental covariates reflecting all those used in previous published literature (Weiss et al. 2015). These 32 base environmental covariates were then "leveraged" to produce 50 million derived environmental covariates (Weiss et al. 2015). Leveraging involved several steps: all environmental covariates were at a higher resolution than the 5x5km grid size used by MAP and so six different summary statistics (mean, sum, sd., range, max., min., sum) were computed for each pixel, those dynamic environmental covariates (i.e. varying monthly, seasonal, or annually) were temporally summarized and offset with a one- to four-month lag, ten non-linear transformations were applied to each environmental covariate and finally all pairwise interactions between the covariates were created (Weiss et al. 2015). The 50 million covariates that resulted from leveraging were subsequently reduced to a set of twenty which could most accurately predict geolocated $PfPR_{2-10}$ values from a training set of 10,680 surveys (Weiss et al. 2015). Interestingly, the twenty final environmental covariates were all pairwise interaction factors, highlighting the complexity of the relationship between parasite prevalence and the environment.

Another key thrust of the work performed by MAP has been to extend the utility of $PfPR_{2-10}$ maps by using them to calculate other epidemiological metrics globally, such as *Plasmodium falciparum* entomological inoculation rate ($PfEIR$), reproduction number (PfR_0) and clinical incidence. The 2010 global $PfPR_{2-10}$ map was used to calculate $PfEIR$, which was generated by comparing a set 123 co-measured $PfPR_{2-10}$ and $PfEIR$ values and empirically selecting log-linearity amongst a set of candidate relationships; and also to calculate PfR_0 , which was derived from $PfPR_{2-10}$ via a malaria transmission model (Gething et al. 2011). Along the same lines, the 2015 endemicity map of continental Africa included predictions for age-stratified clinical incidence of malaria, in this case estimated from $PfPR_{2-10}$ values using an ensemble of three mathematical models (Bhatt et al. 2015). An important caveat is that although MAP has demonstrated that calculation of $PfEIR$, PfR_0 and clinical incidence from $PfPR_{2-10}$ is possible, the

confidence intervals around other metrics are much wider than those for $PfPR_{2-10}$, as a result of the added uncertainty in the relationships between them.

A final major contribution of MAP to malaria epidemiology comes via the implementation of modern techniques in spatial statistics. Specifically, MAP employed model-based geostatistics (MBG) to go from point observations of malaria prevalence to continuous surfaces of malaria risk calculated over a 5x5km grid (Bhatt et al. 2015). Using MBG within a Bayesian framework allowed for statistically rigorous determination of uncertainty, a critical advancement given the often sparse and imperfectly sampled survey data across regions of interest (Gething et al. 2010). The resulting $PfPR_{2-10}$ estimates are expressed as posterior probability distributions for each pixel in the grid: narrow distributions represent high- confidence estimates of $PfPR_{2-10}$ whilst broader distributions indicate higher uncertainty (Gething et al. 2010). MBG are very computationally expensive and the MAP developed approximations to lessen their computational cost (Gething et al. 2010). In a standard implementation, MBG are calculated for each grid pixel independently, allowing for parallelization (Gething et al. 2010). However, this independent calculation makes estimates of uncertainty over multiple grid pixels impossible, as to do so correctly requires quantitating the correlation between pixel values (Gething et al. 2010). Here, MAP produced an efficient approximation algorithm for joint-estimation of pixel values and thus the quantitation of uncertainty across arbitrary regions within the grid, thereby greatly improving the utility of such maps for policy-makers, who often make decisions based on risk estimated across countries or administrative zones (Gething et al. 2010).

Summary

There are several key take-aways from the work of the MAP project. First and foremost, work from MAP has convincingly exhibited the power of MBG to compute large spatial surfaces from point-estimates of epidemiological parameters. Critically, this means that inference approaches using genetic data can concern themselves, at least initially, with producing simple point estimates of transmission intensity

rather than producing global- or regional- maps; with confidence that methods to integrate such results into spatial surfaces exist. Secondly, MAP has built a set of environmental covariates most predictive of malaria endemicity which ultimately should be incorporated into any estimation method utilizing genetic data. Lastly, MAP exists as an high-quality, global and freely available resource for exploring/benchmarking the accuracy of transmission intensity estimates derived from genetic data, as is done in Chapter 3.

1.4.2 MalariaGEN: Genetic data

The Malaria Genomic Epidemiology Network (MalariaGEN) was established in 2005 to foster genomic research aimed at the elimination of malaria (Network 2008). The network connects research partners around the globe with Second-generation sequencing facilities based at the Sanger Institute in Cambridge, and has as key principles standardisation and open data sharing (Network 2008). Since its establishment, it has facilitated the sequencing of thousands of *Plasmodium falciparum* and *Anopheles gambiae* genomes, as well as hundreds of *Plasmodium vivax* samples and human patients with severe malaria. Its partners span 39 countries from around the world, and the genomic data generated by MalariaGEN has been instrumental in characterizing the global population structure of *P. falciparum* (Manske et al. 2012), *P. vivax* (Pearson et al. 2016) and *An. gambiae* (Miles et al. 2017). It has also been used to conduct genome-wide association studies (GWAS) identifying genetic variants associated with severe malaria in human patients (Band et al. 2019), as well as variants associated with resistance to artemisinin in *P. falciparum* (Miotto et al. 2015).

Chapter 3 of this thesis makes extensive use of MalariaGEN’s Pf3k project, Release 5 of which contains 2,393 samples collected from 14 countries across Africa and Southeast Asia.

1.5 Genetic signals of transmission intensity

1.5.1 Experimental evidence

Can this wealth of genomic data be harnessed to estimate transmission intensity? A series of studies utilizing molecular barcoding – genotyping parasites at a limited set of SNPs or microsatellite markers – have pointed to a relationship between malaria population genetics and transmission levels. Consistently, lower transmission has been associated with the discovery of independent cases harbouring parasites with identical genotypes; suggesting lower genetic diversity (Anderson et al. 2000; Nkhoma et al. 2013; Daniels et al. 2013). Additionally, in low transmission environments, a smaller fraction of infections are found to contain molecular barcodes with heterozygous calls – which, as *P.falciparum* is haploid, reflect cases where a single individual has been infected by multiple distinct parasite strains (a "mixed infection") (Anderson et al. 2000; Nkhoma et al. 2013; Daniels et al. 2013). These trends are visible across continents (Africa has higher transmission, higher genetic diversity, and a greater proportion of mixed infections than Southeast Asia) (Anderson et al. 2000) as well as in longitudinal studies following the genetic consequences of amplified intervention efforts (Nkhoma et al. 2013; Daniels et al. 2013).

Jointly, these studies lend credence to the idea that regions experiencing different levels of malaria transmission will have characteristic genetic signatures. More subtly, several groups proposed that lower transmission intensity likely results in increased clonal propagation of parasite strains – that is transmission between hosts without outcrossing (Anderson et al. 2000; Daniels et al. 2013). In *Plasmodium*, recombination occurs only during the brief meiotic interval within the mosquito, and is only observable if the mosquitoes blood-meal contains two genetically distinct parasites, i.e. it has fed on an individual with a mixed infection (Daniels et al. 2013). In regions with high transmission and high diversity, mixed infections are common, supporting a higher rate of recombination between distinct parasite genomes. Conversely, where transmission is low and diversity is low, mixed infections are less probable, the rate of recombination decreases, and parasite genomes become

more stable (Daniels et al. 2013). Indeed, in low transmission environments cases with identical genotypes can be observed several years apart, suggesting they have been transmitted repeatedly without being fragmented by recombination (Daniels et al. 2013; Echeverry et al. 2013; Daniels et al. 2015).

Building upon the above observations, Daniels et al. (2015) sought to assess whether genetic signals alone were sufficient to infer changes in local transmission intensity. Using a 24-SNP barcode, the authors genotyped 1,007 malaria samples collected from patients reporting to clinic between 2006 and 2013 in Thies, Senegal – during which time Thies witnessed a dramatic reduction and then rebound in parasite prevalence (Daniels et al. 2015). They incorporated a model of SNP barcode dynamics into a stochastic agent-based simulation of malaria transmission which included transmission intensity, in the form of R_0 , as a free parameter. They then employed Approximate Bayesian Computation (ABC) to fit a temporal model of R_0 in Thies by comparing simulated and observed SNP barcodes, minimizing a likelihood metric built from seven summary statistics (including statistics such as the number of barcodes that were observed more than once each year, the number of barcodes that persisted over multiple years, and the number of mixed infections observed). Ultimately, the most likely simulations all mimicked the true transmission history in Thies, supporting the idea that genetic data can be sufficient to infer changes in transmission intensity.

1.5.2 Theoretical expectations

Experimental evidence suggests clear differences in the coalescent process⁸ generating genomes collected from low- or high- transmission intensity regions (Figure 1.4). What are these differences? Firstly, in considering the coalescent process in malaria, note that a lineage will oscillate between vector (mosquito) and host (human) (see Fig. 1, box I). These oscillations define transmission cycles, during which different types of mutations can occur. Cross-over and non-cross over meiotic recombination

⁸Coalescent theory is a formulation of population genetics that describes how a sample of alleles, when looking backwards in time, "coalesce" in pairs until they reach a common ancestor. The model was first described in (Kingman 1982).

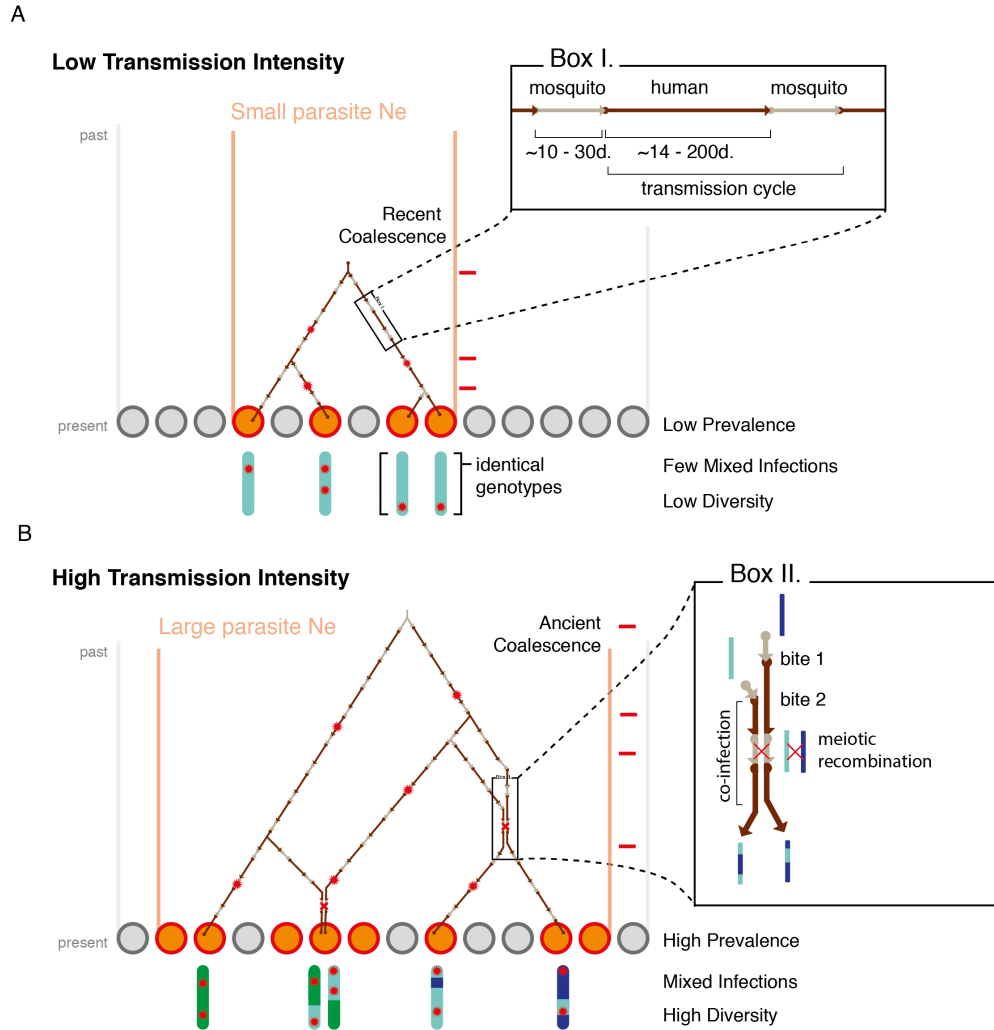


Figure 1.4: The impact of transmission intensity on genetic diversity. (A) Schematic showing hypothesized effects of a low-transmission environment on the coalescent process of a sample of malaria genomes. The population of human hosts is shown as circles; uninfected hosts are in grey and infected hosts are in orange. For some infected hosts, parasite genomes have been sampled and are shown below, with the coalescent process generating them shown above. Box I. Since collected genomes have oscillated between host and vector stages back in time, a coalescent lineage is composed of two phases. (B) Same as in (A), but for a high transmission intensity region. A larger fraction of the host population is infected resulting in a larger parasite N_e and longer time to coalescence. Mixed infections are common, driving recombination, and diversity is higher. Box II. Recombination occurring in a mosquito containing two distinct parasite lineages can reduce linkage disequilibrium and break up haplotypes.

are restricted to the mosquito phase (Miles et al. 2016); the accumulation of SNVs and indels may happen throughout (Hamilton et al. 2017). Next, evidence suggests that the parasite effective population size (N_e) will be a function of regional parasite

prevalence (Nkhoma et al. 2013); higher prevalence regions will have higher effective parasite population sizes, and higher genetic diversity. The time to coalescence between two lineages is proportional to N_e , implying that samples will coalesce in the more distant past in high intensity regions. Additionally, in high intensity regions individuals are more likely to be bit by two infectious mosquitoes, thus mixed infections will be more frequent, facilitating more meiotic recombination (1.4, box II). Consequently, high intensity regions should exhibit lower linkage disequilibrium (Anderson et al. 2000). These are two prominent means by which transmission intensity can be expected to create signals in genetic data: one via the genealogical process (more ancient coalescence causing more genetic diversity) and one via the recombination process (more frequent mixed infections causing more recombination). There are likely others.

In summary, there is experimental support for a link between malaria transmission and genetics. The forces at play can be accommodated by population genetic theory, as we have shown above by framing them within the coalescent. However, at present, these relationships are only understood in a qualitative manner. The only published work (to my knowledge) explicitly estimating malaria transmission intensity from genetic data is that of Daniels et al. (2015), as described above. Their approach of using ABC is very computationally expensive, and moreover their analysis was limited to molecular barcodes, which is of lower-resolution than the genomic data that is now available. Whole-genome sequences have fuelled a plethora of sophisticated demographic inference methods (Gutenkunst et al. 2009; H. Li and Durbin 2011; X. Liu and Fu 2015) and may, similarly, fuel new insights and more accurate estimates of transmission intensity.

1.6 Conclusion

The overarching aim of this thesis is to estimate malaria transmission intensity from *P. falciparum* genetic data. Transmission intensity is a metric of critical importance to malaria control programs, as it allows for targeted deployment and monitoring of control interventions. Yet, all metrics currently used to estimate

transmission intensity have drawbacks, particularly in low transmission settings (Reiner et al. 2015a; Sturrock et al. 2016). Previous studies, as well as theoretical considerations, suggest that genetic diversity and the rate of mixed infection may increase with transmission intensity, which could facilitate inference. To date, estimating transmission intensity from genetic data has only been achieved with ABC (Daniels et al. 2015), a computational expensive method, and without leveraging the abundance of whole genome sequencing data that is now available. In addition to whole-genome sequencing data, a global repository of malaria transmission intensity data, in the form $PfPR_{2-10}$ estimates, is also available (Weiss et al. 2019).

I address the problem of inferring transmission intensity from genetic data most directly in Chapter 2, where I use an agent-based simulation, **forward-dream**, to uncover relationships between genetic diversity statistics and parasite prevalence. One of the findings from this chapter is that mixed infections are especially informative for epidemiological inference. In Chapter 3, I explore mixed infections in greater detail, in particular how the patterns of IBD within them relate to individual transmission events. I make use of both the Pf3k project and data from MAP in Chapter 3. In Chapter 4, I turn to the problem of genetic data collection, developing a methodology to collect genetic data from a dried-blood spot and in-country. I use the method to infer antimalarial resistance status of ten samples in Zambia, but provide preliminary evidence suggesting the method could also be used to characterise mixed infections.

2

Forward-time simulation of *P. falciparum* transmission and evolution

Contents

2.1	Introduction	29
2.2	Developing forward-dream	32
2.2.1	Epidemiological layer	34
2.2.2	Infection layer	35
2.2.3	Genome evolution layer	39
2.2.4	Parameterisation and implementation	40
2.3	Genetic diversity and host parasite prevalence at equilibrium	41
2.4	Genetic diversity and host parasite prevalence in times of change	46
2.4.1	Malaria control interventions	47
2.4.2	Seasonality	55
2.5	Conclusion	57

2.1 Introduction

In this chapter, we explore how the genetic diversity of a *P. falciparum* population may reflect the prevalence of infection across hosts. Our motivation is to establish a set of relationships between genetic diversity statistics and parasite prevalence, PR , that could be leveraged for purposes of statistical inference. More specifically,

we seek to gain insights into how *P. falciparum* genetic data, collected from the blood of infected hosts, could be used to infer parasite prevalence both across different geographical regions and through time.

It is widely accepted that relationships exist between malaria disease epidemiology and the genetics of parasite populations, yet the nature of these relationships is unknown (Volkman et al. 2012). Many features of malaria epidemiology are likely reflected in genetics – including migration and source/sink dynamics, the selection and spread of drug resistance, and changes in transmission intensity (Volkman et al. 2012). In this chapter, we focus specifically on epidemiological parameters that reflect transmission intensity. Inferring the epidemiological parameters of an infectious disease from the genetics of the pathogen is often presented as analogous to making demographic inferences about a population from genetic data, for which a multitude of methods now exist (Pybus et al. 2000; Gutenkunst et al. 2009; H. Li and Durbin 2011; X. Liu and Fu 2015). For example, the Pairwise Sequentially Markovian Coalescent (PSMC) method has been used to infer the effective population size (N_e) of *Homo sapiens* from about ten thousand to two million years ago (H. Li and Durbin 2011). In this line of reasoning, transmission intensity may be, for example, a function of N_e . However, there are at least two important difference that make epidemiological inference from genetic data more challenging than demographic inference using a method like PSMC.

First, it is not immediately clear that the classical models in population genetics¹ which are often employed to make demographic inferences are suitable for *P. falciparum*. Consider that the life cycle of *P. falciparum* involves oscillating between human host and mosquito vector populations, which may be of different sizes, and may also induce different rates of mutation. This structure most resembles a metapopulation model (reviewed in Hastings and Harrison (1994)), with each host and vector containing its own population of parasites. Within both the host and the vector, parasite populations likely experience extreme bottlenecks (for example, as the ookinetes penetrate the midgut wall of the vector), exponential

¹Including the Wright-Fisher, Moran, and all other models that converge to Kingman’s n -Coalescent in the ancestral limit (Kingman 1982).

growth phases (merozoites escaping from hepatocytes and replicating in the blood), and interactions with the host or vector immune system. Indeed, there has been work demonstrating that the malaria life cycle simultaneously intensifies drift and selection; a result contrary to what is expected under a Wright-Fisher Model² (Chang et al. 2013). So, while classical models in population genetics have the advantage of being extensively studied and mathematically tractable, with a known set of relationships between genetic diversity statistics and demographic parameters, they do not readily apply to *P. falciparum*. Finally, even if these models did apply, the relationships between demographic parameters (such as N_e) and epidemiological ones (such as parasite prevalence) would need to be defined.

Alternatively, the epidemiological models that have been developed to reflect malaria biology and transmission do not explicitly incorporate genetic processes. The most well known class of epidemiological models are the so-called "compartment-based" models where individual hosts and vectors transition between compartments which can represent a variety of disease states (such as susceptible, infected, or immune) (reviewed in Mandal et al. (2011) and D. L. Smith et al. (2012)). The overall population is represented by the number of individual hosts and vectors occupying each state. The canonical compartment model is the Ross-Macdonald model, where hosts and vectors can be either susceptible or infected (D. L. Smith et al. 2012). Conveniently, in these models, the equilibrium prevalence of infected hosts and vectors is a function of the parameters specifying the transition rates between compartments. Yet, the absence of genetic processes (and indeed, individual parasites) means they offer no insight into how parasite genetic diversity relates to these transition rates, and by extension, disease epidemiology.

In summary, classical models in population genetics clearly relate demographic parameters to genetic diversity, but do not reflect malaria epidemiology; while canonical epidemiological models appropriately reflect malaria biology and transmission but do not incorporate genetics.

²In the Wright-Fisher, the strength of selection increases as drift decreases.

A second major challenge specific to epidemiological inference from genetic data is that the ultimate aim is often to inform disease control and, as a result, what point in time the inference pertains to is of critical importance. Methods like PSMC are based on inferring the time-to-coalescence between segments of DNA, which is proportional to the harmonic mean of the effective population size until coalescence³. As a result, the estimates of N_e are often historical; in reference to the population as it was on the order of thousands of generations in the past. Such an approach is not suitable for disease control, where policy decisions need to be made on the basis of information about the near-present, or predictions about the future.

In this chapter, we address these issues by first developing a "genetic epidemiological" model: one that incorporates the genetic processes of individual parasites (i.e., mutation and recombination) with population dynamics of the Ross-Macdonald model. Implemented as an agent-based simulation called **forward-dream**, we use the model to explore both equilibrium and dynamic aspects of the relationship between parasite genetic diversity and host parasite prevalence. In particular, our aims are as follows:

1. To develop a genetic epidemiological model of *P. falciparum* in the form of an agent-based simulation (Section 2.2)
2. To elucidate the relationships between a set of genetic diversity statistics and parasite prevalence at equilibrium (Section 2.3).
3. To characterise the response times of different genetic diversity statistics following changes in parasite prevalence (Section 2.4)

2.2 Developing **forward-dream**

We developed an agent-based simulation of malaria incorporating salient features of its transmission and epidemiology, as well as explicitly modelling the genetic material of individual parasites. Our genetic epidemiological simulation is called

³For example, if the sample took on n different sizes until coalescence, then $N_e = n / \sum_{i=1}^n 1/N_e^i$ and $t_{coal} = 2N_e$.

forward-dream (**f**orward-time **d**rift, **r**ecombination, **e**xtinguishment, **a**dmixture and **m**eiosis) and is comprised of three layers: (1) an epidemiological layer, which controls how malaria spreads through a population of hosts and vectors and reaches equilibrium; (2) an infection layer, which controls the dynamics of malaria parasites during individual transmission events and within individual hosts and vectors; and (3) a genome evolution layer, which controls how the genetic material of individual parasites is represented, mutated and recombined. Each layer is described in detail below and all simulation parameters are listed in Table 2.2.

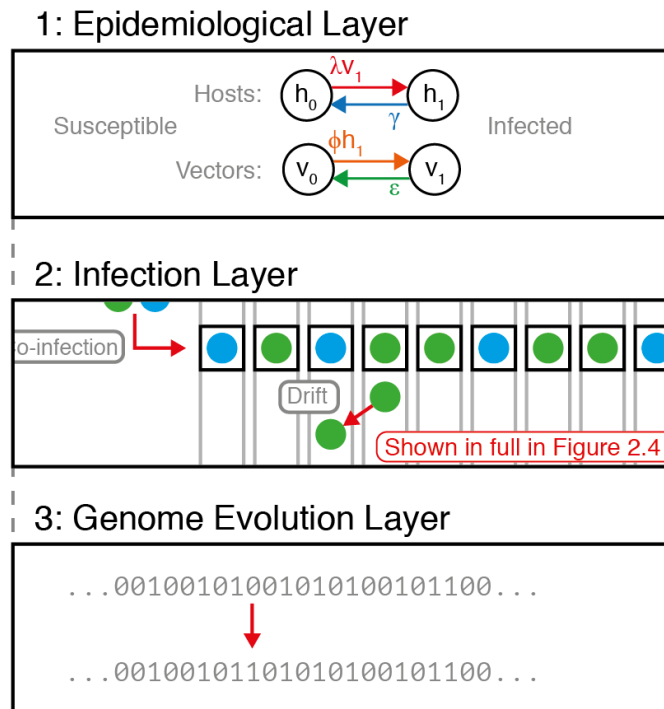


Figure 2.1: Schematic of the three layers of **forward-dream**, an agent-based genetic epidemiological simulation. The epidemiological layer is an implementation of the Ross-Macdonald model, which controls the rate of infection spread and equilibrium across the population. The infection layer controls inter- and intra-host infection processes, including co-infection, super-infection, and genetic drift within hosts and vectors. The genome evolution layer controls how individual parasites genomes are represented, mutated and recombined.

2.2.1 Epidemiological layer

For the epidemiological layer we implemented the Ross-Macdonald Model, in which a fixed number of hosts (N_h) and vectors (N_v) alternate between susceptible and infected based on four fixed rate parameters (Figure 2.2). The model can be described by two coupled differential equations:

$$\begin{aligned}\frac{dh_1}{dt} &= h_0(v_1/N_v)\lambda - h_1\gamma \\ \frac{dv_1}{dt} &= v_0(h_1/N_h)\phi - v_1\epsilon\end{aligned}$$

where h_0 and h_1 are the number of susceptible and infected hosts, respectively, with $N_h = h_0 + h_1$; and v_0 and v_1 are the number of susceptible and infected vectors, respectively, with $N_v = v_0 + v_1$. Note that in this model λ and ϕ are compound parameters representing the rate at which hosts and vectors become infected. In particular,

$$\lambda = b(N_v/N_h)\pi_h$$

where b is the daily vector biting rate, (N_v/N_h) gives the vector density, and π_h is the probability that an infectious bite from a vector produces an infected host. Similarly, we have,

$$\phi = b\pi_v$$

where b is again the daily vector biting rate and π_v is the probability that a vector that bites an infected host becomes infected. We allow for mixed infections, which will be described in Section 2.2.2. However, for the epidemiological layer mixed infections have no consequence: both clonal and mixed infections are assigned to the infected compartments (h_1 or v_1).

The epidemiological layer dictates the equilibrium prevalence of infection in hosts ($X_h = h_1/N_h$) and vectors ($X_v = v_1/N_v$). In particular, the equilibrium prevalence in hosts is given by:

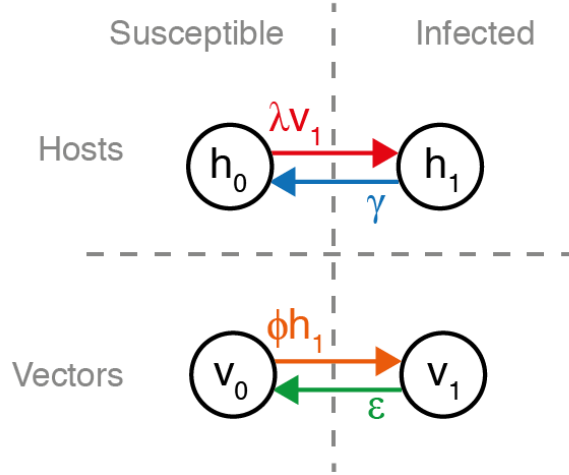


Figure 2.2: Schematic of the epidemiological layer of **forward-dream**, which implements the Ross-Macdonald model. Hosts and vectors transition between susceptible (h_0, v_0) and infected (h_1, v_1) compartments. Rates of clearance (transitioning from infected to susceptible) are defined by the constants γ and ϵ for hosts and vectors, respectively. Rate at which hosts are infected depends on constant λ and the number of infected vectors, v_1 . Similarly, rate at which vectors are infected depends on the constant ϕ and the number of hosts that are infected, h_1 .

$$X_h = \frac{\lambda\phi - \gamma\epsilon}{\lambda\phi + \phi\gamma} \quad (2.1)$$

and in vectors by:

$$X_v = \frac{\lambda\phi - \gamma\epsilon}{\lambda\phi + \lambda\epsilon} \quad (2.2)$$

Note that X_h corresponds to parasite prevalence PR , as described in Section 1.3.2. We have confirmed that our implementation of the Ross-Macdonald Model in **forward-dream** converges to the expected equilibrium values (Figure 2.3). In total, the behaviour of the epidemiological layer is specified by seven parameters.

2.2.2 Infection layer

The infection layer specifies the biology of individual transmission events and infection processes. This includes the nature of the transmission bottleneck, how

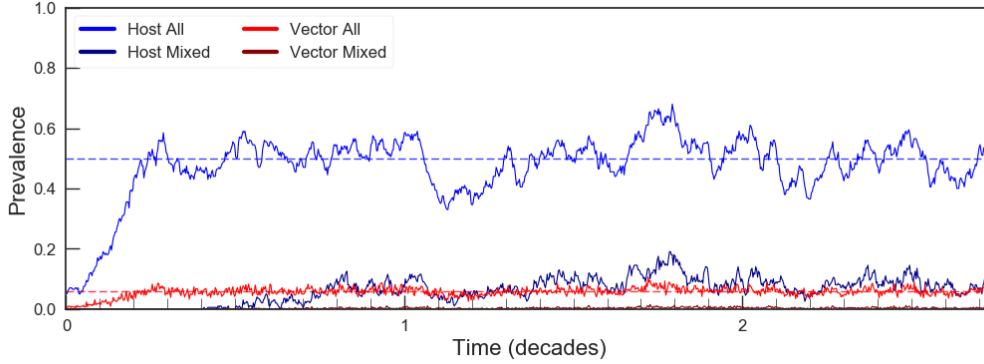


Figure 2.3: Host and vector prevalence in *forward-dream* fluctuates around equilibrium values predict by the Ross-Macdonald model. The first 27 years of a single *forward-dream* simulation are shown. The simulation is seeded with ten infected hosts. Blue and red lines give prevalence of infection across the host and vector populations, respectively. Similarly, dark blue and dark red lines give the prevalence of mixed infections (infections containing > 1 distinct parasite) across the host and vector populations, respectively. Horizontal blue and red dashed lines give equilibrium prevalence values predicted from the Ross-Macdonald model (Equations 2.1 and 2.2).

parasites populate hosts and vectors, how co-infection and super-infection occur, and the rate of drift and mutation within hosts and vectors.

Infection Capacity

The infection layer models an individual host's capacity to be infected with k_{max}^h within-host compartments (Figure 2.4). When an individual host is in the susceptible state (i.e. is uninfected), all k_{max}^h compartments are empty. When an individual host becomes infected, all k_{max}^h compartments simultaneously become populated. Each compartment can harbour a unique parasite genome⁴, although typically multiple compartments will be occupied by identical genomes. For example, if a host is infected by a vector carrying a single parasite genome, all k_{max}^h compartments will be occupied by that genome. Alternatively, if a host is co-infected by a vector carrying two distinct parasite genomes, its compartments will be a mixture of the two genomes.

We note that in actuality, an individual host can harbour up to 10^9 - 10^{11} parasites (White 2004). In all the simulations described in this chapter, k_{max}^h is set to ten (see

⁴Such that the maximum complexity of infection $k = k_{max}^h$.

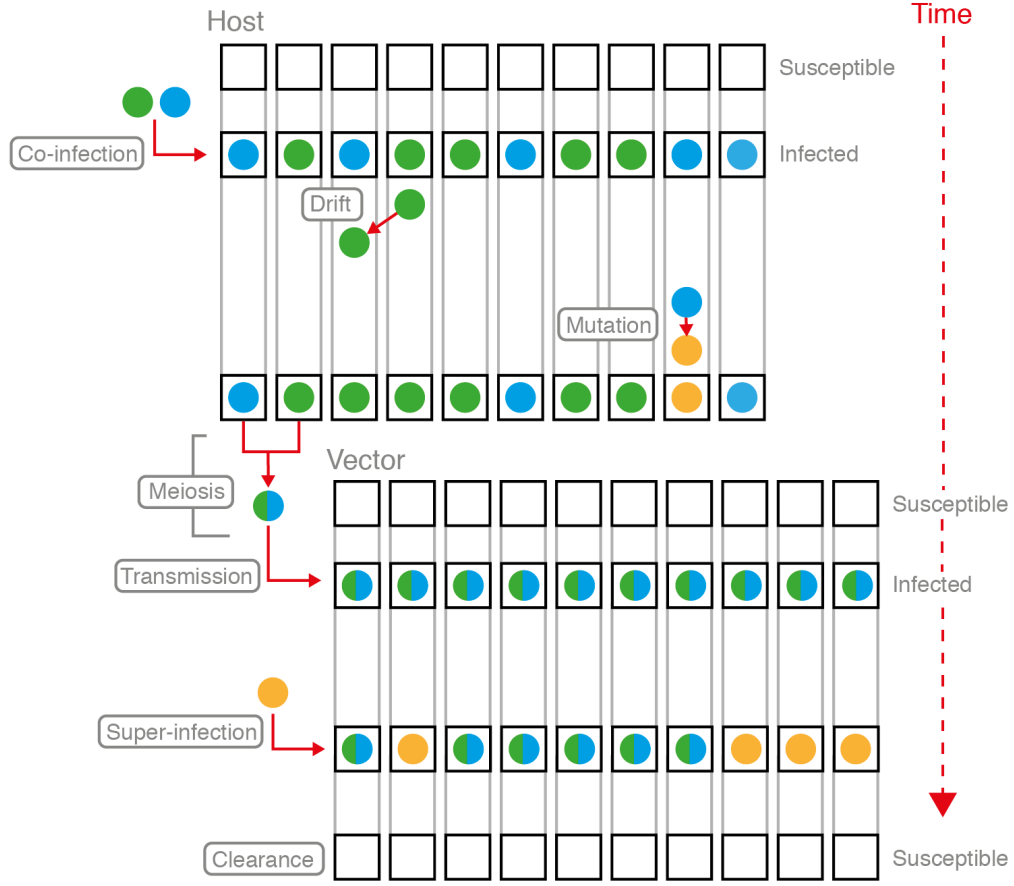


Figure 2.4: Schematic of the infection layer of *forward-dream*. The capacity of individual hosts and vectors to be infected is represented by a fixed number of within-host compartments (represented by black boxes), each which can harbour a unique parasite genome (represented by coloured circles). In a susceptible host/vector, all compartments are empty. Upon infection, all compartments are populated. Drift and mutation can occur between compartments. Meiosis occurs during the transmission of parasites from a host to a vector. Super-infection can occur, resulting in an average of half of all compartments being replaced by newly transmitted parasite strains. Clearance in either hosts or vectors empties all compartments. Note how this layer can be nested within the Ross-Macdonald model.

Table 2.2). This simplification can be justified as follows. First, even if an individual host contains 10^9 - 10^{11} parasites, *P. falciparum* replicates clonally within the host and so the majority of these will have near-identical genomes. A host may, however, be infected by multiple distinct parasite strains, generating a mixed infection. Setting $k_{max}^h = 10$ still accommodates for this possibility. Second, parasites genomes that exist at a frequency much less than 1/10 of the population are unlikely to produce

substantial signal in genetic data. For example, with whole-genome sequencing data, sequencing depths are typically $\simeq 100X$ and so parasites comprising less than $1/10$ of the infection will be represented by only a handful of reads. Finally, limiting the number of parasites per host to k_{max}^h makes simulating the evolution of parasites within a population of hosts computationally tractable.

For vectors, the same strategy is adopted, with each vector's capacity to be infected modelled with k_{max}^v within-vector compartments. For all simulations in this thesis, $k_{max}^v = 10$.

Infection Evolution

The *P. falciparum* infection of a host evolves over time. In particular, the k_{max}^h parasite genomes within the host behave like a continuous-time Moran process for a population with k_{max}^h individuals (Moran 1958). Moving forward in time, drift events occur in continuous time at a rate of d_h per day. When a drift event occurs, two parasite genomes are selected at random from the k_{max}^h available. At this point, either one genome randomly replaces the other, or receives a mutation with probability θ_h . This process continues until the infection is cleared and the host returns to the susceptible state, with all compartments becoming simultaneously empty. For hosts, clearance occurs at a fixed rate γ as specified in the epidemiological layer.

The infection of a vector evolves according to the same process as in the host, but with a set of parameters k_{max}^v , d_v , θ_v , ϵ (Table 2.2).

Infection Transmission

In *forward-dream*, a *P. falciparum* infection is transmitted from a vector to a host through a transmission bottleneck, such that not all of the parasites within the infectious vector establish themselves in the host. In particular, a random subset of all parasites $\bar{v} \subseteq v$ (where $v = \{v_1, v_2, \dots, v_{k_{max}^v}\}$ is the set of all parasites in the vector) is transmitted. The number of transmitted parasites $n = |\bar{v}|$ is drawn from

$$n \sim \max[1, \text{Bin}(k_{max}^v, 1/k_{max}^v)]$$

This results in an average of $1 + (1 - 1/k_{max}^v)^{k_{max}^v} \leq 2$ parasites passing through the transmission bottleneck.

Note that in cases where $n > 1$ and $\bar{v}$ contains unique parasites, co-infection may occur. If the vector is infecting a susceptible (i.e. uninfected) host, k_{max}^h parasites are drawn with replacement from $\bar{v}$ with each having an equal probability ($1/n$) of being drawn. These k_{max}^h parasites then populate the k_{max}^h within-host compartments. Alternatively, super-infection occurs if the host is already infected. Defining the parasites already within the host by the set $h = \{h_1, h_2, \dots, h_{k_{max}^h}\}$, we first create the union $h \cup \bar{v}$. From this set k_{max}^h parasites are drawn, where each parasite has probability $1/2k_{max}^h$ or $1/2n$ of being drawn, if it is from h or $\bar{v}$, respectively. As a consequence, on average super-infection results in half of the within-host compartments being occupied by new parasites.

The transmission from host to vector is the same as above, but with the addition of meiosis. In brief, the n parasite strains selected at random from infected host may undergo meiosis before populating the the k_{max}^v within-vector compartments. The meiosis model is described in greater detail below and in Chapter 3.

In total, the infection layer is specified by six parameters.

2.2.3 Genome evolution layer

Finally, the lowest layer describes the malaria genome model. We represent the genetic material of an individual parasite as a single fixed-length array of zeros and ones, defined by the parameter N_{snps} . In essence, this array represents a single chromosome. Mutation is reversible, and the genome undergoes meiosis during transmission from host to vector, based on a simplified implementation of our previously published meiosis simulator, **pf-meiosis**. The recombination rate is constant and scaled with respect to N_{snps} , such that an average of one cross-over event occurs per bivalent during meiosis. The meiosis model is described in greater detail in Chapter 3. The only parameter specific to the genome evolution layer is N_{snps} .

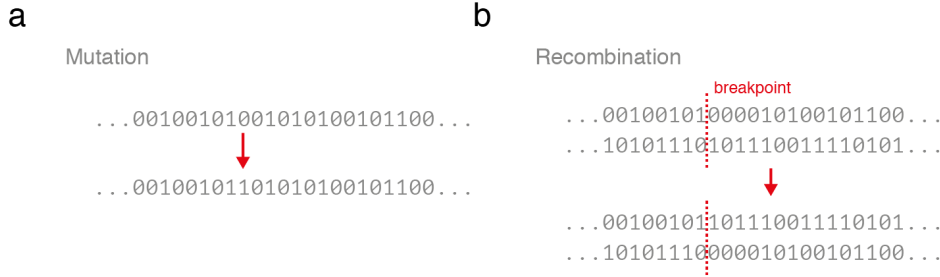


Figure 2.5: Schematic of the genome evolution layer of **forward-dream**, which controls how individual parasite genomes are represented, mutated, and recombined. Individual parasite genomes are represented by an array of 0's and 1's of a fixed length (typically 100-1000). (a) Mutation is reversible, and converts 0 to 1, or 1 to 0. (b) Recombination occurs as part of meiosis. Breakpoints are chosen at random, and material to the right of the breakpoint is swapped between strains.

2.2.4 Parameterisation and implementation

In **forward-dream**, time is continuous and the occurrence of events is dictated by a competing Poisson process. The rate at which any event occurs, r_{total} , is equal to:

$$r_{total} = bN_v + \gamma h_1 + \epsilon v_1$$

which is the sum of the total biting rate across all vectors, the clearance rate across all infected hosts and the clearance rate across all infected vectors. The time until the next event is drawn from an exponential distribution with an expectation of $1/r_{total}$. Conditional on an event occurring, the type of event is sampled proportional to its fractional rate. For example, the probability of a given event being a biting event is:

$$P(bite|event) = \frac{bN_v}{r_{total}}$$

If a biting event occurs, the host and vector involved are sampled at random from the population, and the nature of parasite transmission depends on their infection status. Similarly, for clearance events, the host or vector to be cleared of infection (i.e. returned to the susceptible state) is chosen at random.

The 14 parameter values, along with typical values, are listed in Table 2.2. The values representing biological processes (such as the vector biting rate and host

Parameter	Definition	Value
N_h	Number of hosts	200
N_v	Number of vectors	1000
k_{max}^h	Maximum k for hosts	10
k_{max}^v	Maximum k for vectors	10
b	Biting rate (per vector per day)	0.25
π_h	Probability host is infected after receiving an infectious bite	0.1*
π_v	Probability vector is infected after biting an infectious host	0.1*
γ	Host infection clearance rate	0.005†
ϵ	Vector infection clearance rate	0.2
d_h	Drift rate in hosts (events per day per host)	1
d_v	Drift rate in vectors (events per day per vector)	2
θ_h	Probability of mutation event given drift has occurred in hosts	0.0001
θ_v	Probability of mutation event given drift has occurred in vectors	0.0001
N_{snps}	Number of SNPs modelled per parasite genome	1000

Table 2.1: Complete list of simulation parameters for **forward-dream**. Values given represent those of a typical simulation with a host (X_h) and vector (X_v) prevalence of 0.65 and 0.075, respectively. †Consistent with Sama et al. (2006). *Consistent with Beier et al. (1994).

clearance rate) were selected after a review of the literature. The entire agent-based simulation was implemented in python.

2.3 Genetic diversity and host parasite prevalence at equilibrium

In order to elucidate relationships between host parasite prevalence and genetic diversity, we sought first to vary host parasite prevalence across different **forward-dream** simulations. Within the Ross-Macdonald framework, host parasite prevalence is a

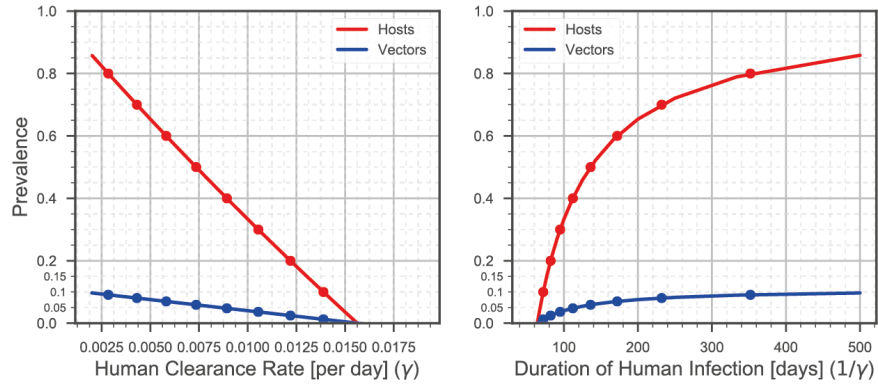
function of the four rate parameters (see Equation 2.1). In nature, what underlies host parasite prevalence differences observed between two geographies or points in time is often unknown, and likely the outcome of a myriad of epidemiological and environmental factors. To achieve different prevalence values in **forward-dream**, we choose three parameters to vary individually: (i) human duration of infection (γ), which may vary between sites if, at one site there is quicker recourse to treatment, or differing proportions of symptomatic and asymptomatic individuals; (ii) vector biting rate (λ), which may vary as a consequence of differences in vector species, environmental conditions, or the presence of bednets; and (iii) vector density (N_V) – which is influenced by climate and weather, geography, and also by insecticide-based interventions (see Onori and Grab (1980)). We tuned each of these parameters to achieve equilibrium prevalence values that varied from 0.3 to 0.8 in a population of two-hundred human hosts⁵ (Figure 2.6). At each prevalence value, **forward-dream** was seeded with ten infected hosts carrying identical parasites and then run until equilibrium⁶. After reaching equilibrium, simulations were continued for an additional 100 years, during which time parasite genomes were collected every ten days from twenty infected hosts selected at random. From these collected parasite genomes we could track the behaviour of a suite of genetic diversity metrics through time and construct their distributions at a given fixed prevalence.

The genetic diversity statistics which we calculated are listed in Table 2.1. The statistics can be divided into three broad categories: (i) those related to mixed infections, which includes the fraction of mixed infections and the mean complexity of infection; (ii) those that are related to the size and shape of the genealogy of the sample, which includes the number of segregating sites, the number of singletons, nucleotide diversity (π), Watterson’s Theta (θ_w), and Tajima’s D ; and (iii) those that summarize the structure of identity-by-descent (IBD) within

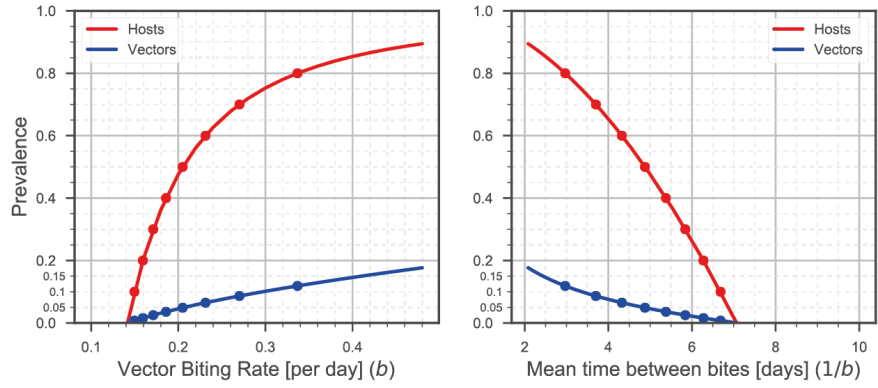
⁵Prevalence values less than 0.3 are of interest. However, they produce frequent stochastic extinction in **forward-dream** (especially over long time windows) and thus were not studied here.

⁶Equilibrium in this context means until the genetic process has reached equilibrium, which has occurred once the number of segregating sites is no longer increasing.

a



b



c

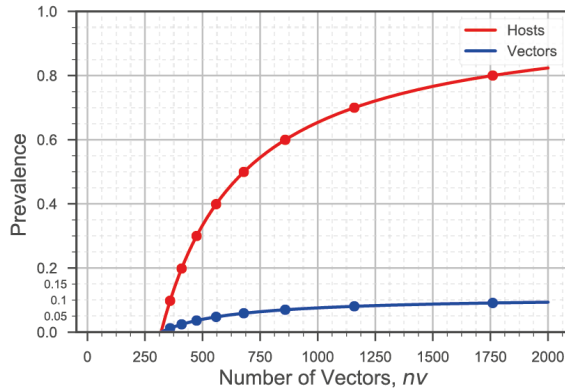


Figure 2.6: Varying the parameters of `forward-dream` to produce simulations with different equilibrium host parasite prevalence values. (a) Varying host clearance rate to produce different prevalence values. Prevalence in hosts indicated in red, vectors in blue. Dots mark host parasite prevalence values of 0.1 through 0.8. Note that host clearance rate dictates duration of infection, show in right plot. (b) Same as (a) but varying vector biting rate. Vector biting rate dictates the average time a vector takes between successive bites, show in right plot. (c) Same as (a), but varying the number of vectors.

the population⁷, including, between a pair of samples, the average fraction of the genome in IBD, the average number of of IBD tracks, and the average length of an IBD track. We note that these statistics are not all independent, rather many are significantly co-linear (Figure 2.7).

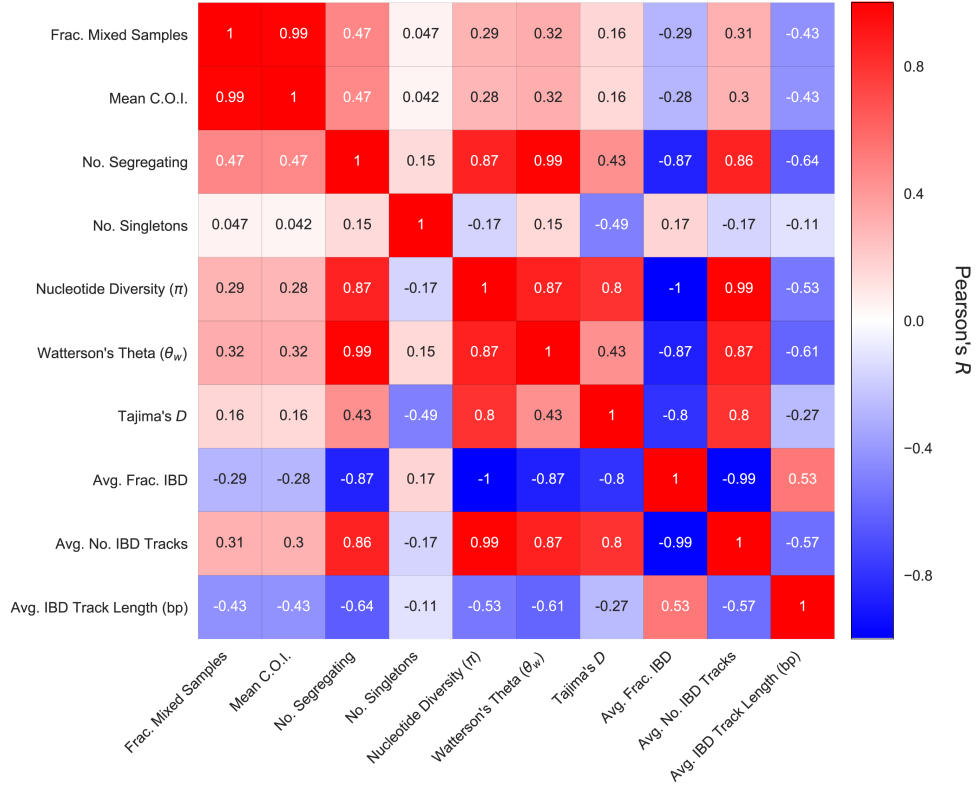


Figure 2.7: Correlation between different genetic diversity statistics across varying host prevalence levels.

We plotted the distributions of each of these ten statistics across different host parasite prevalence values separated by which epidemiological parameter was varied (Figure 2.8, Figure 2.9, and Figure 2.10). We used linear regression to determine the fraction of the variance in each genetic diversity statistic that could be explained by variance in host parasite prevalence (r^2). The results of this analysis are shown in Figure 2.11. Of the ten genetic diversity statistics, all but Tajima's D and the number

⁷For the purposes of the simulation, IBD is simply identity-by-state (IBS). Note that since mutation is reversible in the simulation, IBS can occur through recurrent mutation. This inflates IBD statistic values but preserves trends.

Statistic	Definition
Fraction Mixed Samples	Fraction of samples collected that contain > 1 distinct parasites
Mean Complexity of Infection (C.O.I.)	Mean number of distinct parasites per sample
Number of segregating sites (S_n)	Total number of polymorphic sites across all samples
Number of singletons	Number of polymorphic sites where the minor allele is present in only one sample
Nucleotide diversity (π)	Mean number of pairwise differences between two samples, normalized to the total number of sites
Watterson's Theta (θ_w)	Point estimate of $\theta = 2N_e\mu$ where μ is the mutation rate, given as $\theta_w = S_n / \sum_{i=1}^{n-1} \frac{1}{i}$
Tajima's D	A test statistic for the neutral mutation hypothesis (Tajima 1989)
Mean IBD	Mean fraction of sites identical between two parasites within the sample
Mean IBD track length	Mean length of an IBD track (≥ 1 consecutive identical sites) between two parasites within the sample
Mean number of IBD tracks	Mean number of IBD tracks between two parasites within the sample

Table 2.2: Genetic diversity statistics computed in `forward-dream`.

of singletons have a significant proportion of their variance ($> 10\%$) explained by host parasite prevalence. Furthermore, all of the genetic diversity statistics have a positive relationship with prevalence, except for the average fraction of IBD and average IBD track length, which decrease in higher host parasite prevalence regimes.

Interestingly, there are pronounced differences in the amount of variance explained by host parasite prevalence when different epidemiological parameters underlie the prevalence variation. In particular, varying host clearance rate (γ) results in a significantly higher variance explained for all of the genealogy and IBD statistics. For example, the variance in nucleotide diversity explained by host parasite prevalence drops from 64% to 10% when host parasite prevalence is modulated by the vector biting rate instead of the host clearance rate. In contrast,

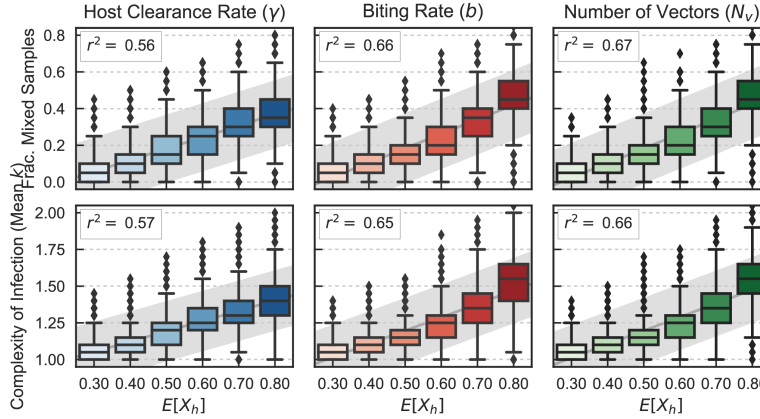


Figure 2.8: Distributions of mixed infection related genetic diversity statistics, plotted for equilibrium host parasite prevalence values tuned to between 0.3 and 0.8 in **forward-dream** simulations. Left, middle, and right panels show distributions when host parasite prevalence is varied as a function of the host clearance rate (γ , in blue), vector biting rate (b , in red) or number of vectors (N_v , in green). Each boxplot contains the result of 10 replicate experiments, where the parasite genomes within 20 randomly selected hosts are collected at every ten days for 100 years (resulting in 3650 samples per experiment) and are used to compute the genetic statistic of interest. The variance explained by ordinary least squares regression is given at top left, and line of best fit and confidence intervals indicated in grey.

the two statistics describing the mixed infections within the population have high r^2 values ($>50\%$) regardless of which of the epidemiological parameters underlies the host parasite prevalence change.

2.4 Genetic diversity and host parasite prevalence in times of change

An important application of this work will be to settings where malaria control interventions are actively occurring. In such cases, and also in cases with seasonal variation in host parasite prevalence, it is unlikely that the parasite population will be in equilibrium. Although in the previous section we established several relationships between genetic diversity statistics and host parasite prevalence, our analysis did not explore over what time scale these relationships come into existence, or how they are affected by a population leaving equilibrium. Thus, our second aim with **forward-dream** was to explore the dynamic relationships

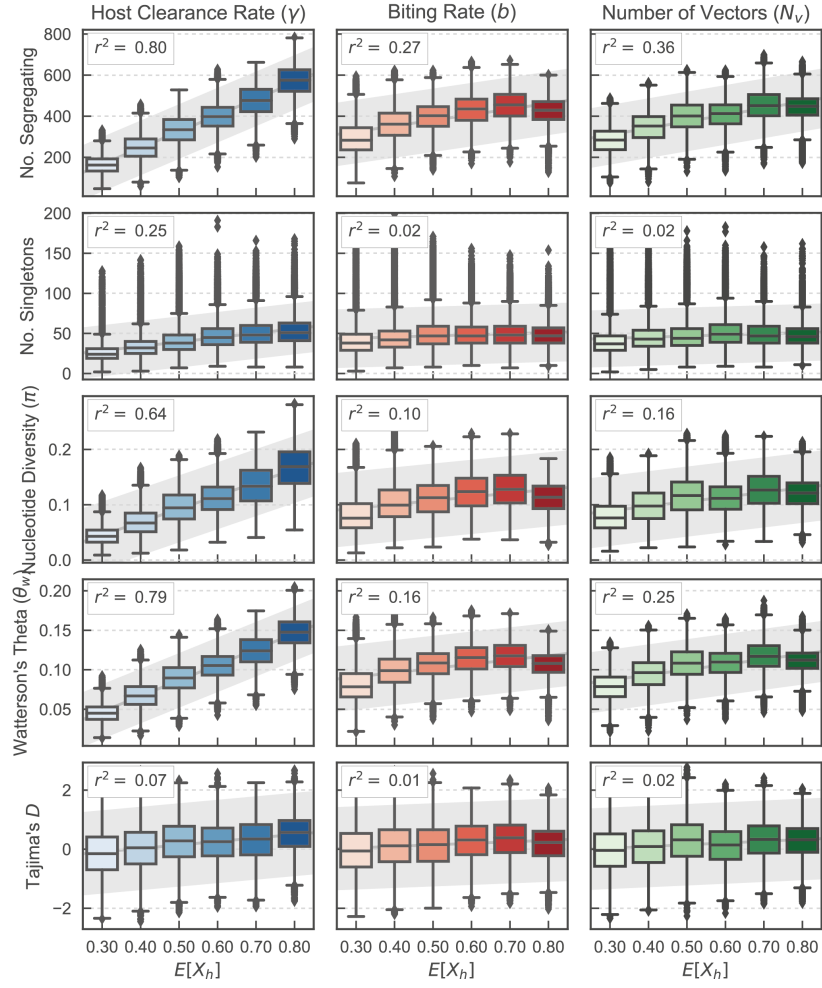


Figure 2.9: Distributions of genetic diversity statistics related to the shape and size of the sample genealogy, plotted for equilibrium host parasite prevalence values tuned to between 0.3 and 0.8 in **forward-dream** simulations. See 2.8 caption for details.

between host parasite prevalence and *P. falciparum* genetic diversity in cases where populations are outside of equilibrium.

2.4.1 Malaria control interventions

In order to understand how genetic diversity statistics relate to host parasite prevalence in scenarios where a malaria control intervention has been deployed, we developed a framework where individual **forward-dream** simulations had three distinct epochs (Figure 2.12). In the first epoch, simulations are run until equilibrium at a host parasite prevalence of 0.65, under the parameters listed in Table 2.2. At the start of the second epoch, one of either the host clearance rate (γ), the vector biting

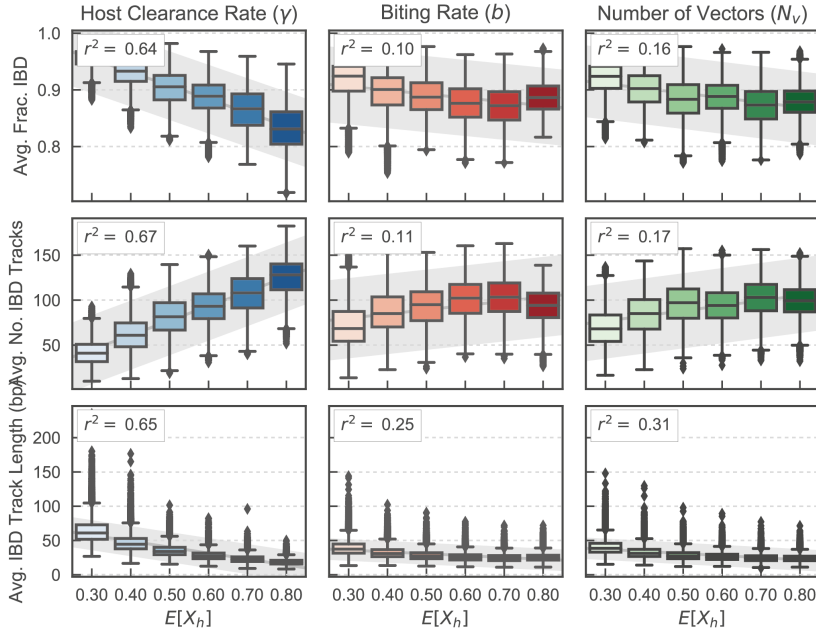


Figure 2.10: Distributions of genetic diversity statistics related to identity-by-descent, plotted for equilibrium host parasite prevalence values tuned to between 0.3 and 0.8 in *forward-dream* simulations. See 2.8 caption for details.

rate (b), or the number of vectors (N_v), is changed such that the new equilibrium host parasite prevalence is 0.2. The parameter change occurs incrementally over a period of thirty days following a logistic transition function, as to mimic the gradual introduction of a malaria control intervention. As a consequence, the simulation leaves equilibrium, with host and vector prevalence declining. The second epoch is allowed to continue until the population has regained equilibrium. At the start of the third epoch, we return the changed simulation parameter to its original value, again over thirty days following a logistic transition function. Again, this results in the simulation leaving equilibrium, with the host parasite prevalence increasing to 0.65. As with the second epoch, the third epoch continues until the population regains equilibrium. In summary, the three epoch model allows us to explore both a parasite population decline and rebound.

Throughout the three epochs parasite genomes are collected from twenty hosts selected at random every ten days, allowing the same suite of genetic diversity statistics discussed in Section 2.3 to be calculated. The trajectories of Tajima's D ,

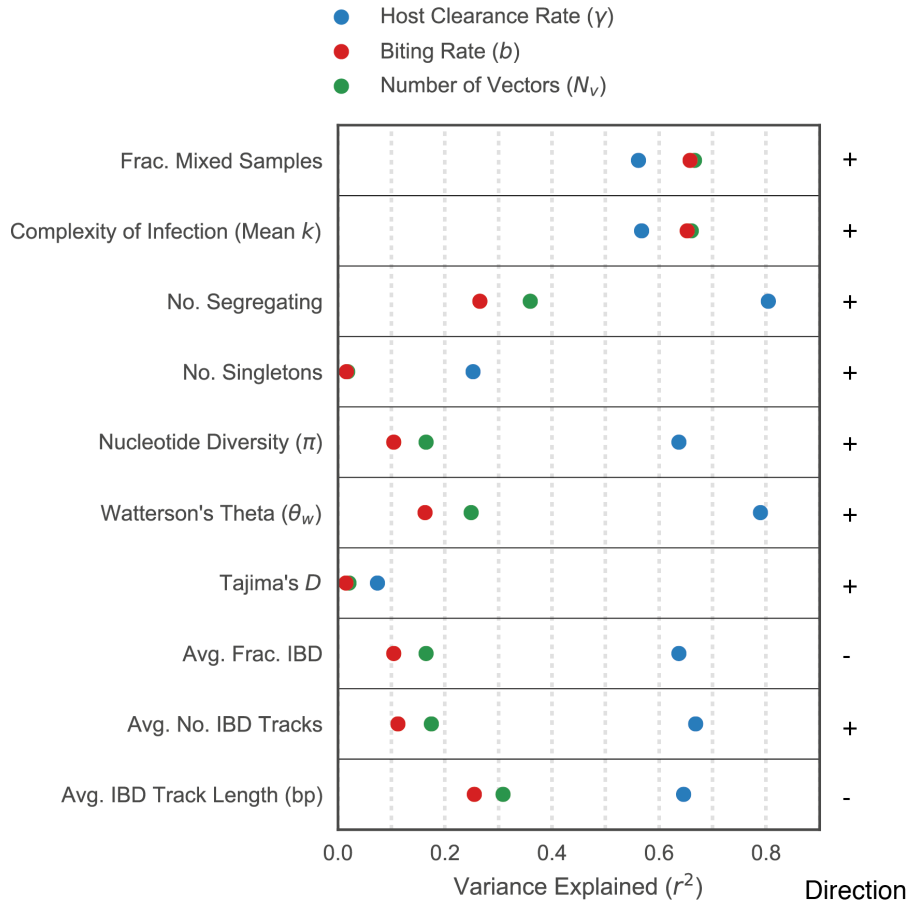


Figure 2.11: Summary of variance explained by host parasite prevalence across a suite of genetic diversity statistics. For each genetic diversity statistic, the variance explained by host parasite prevalence in an ordinary least squares regression is shown by three coloured points: blue, variance explained when host clearance rate is varied; red, variance explained when biting rate is varied; green; variance explained when the number of vectors is varied. At right, the direction of the line of best fit is indicated: +, increasing slope; -, decreasing slope.

mean C.O.I., nucleotide diversity, and average pairwise IBD for a single 3 epoch **forward-dream** simulation are shown in Figure 2.13. An immediate observation is that all four genetic diversity statistics have "noisy" trajectories, tending to fluctuate considerably through time. All genetic diversity statistics considered here (Table 2.1) were similarly noisy (data not shown). Several factors are likely contributing to this, including the stochastic nature of the epidemiological process itself (host and vector prevalence fluctuate through time around their equilibrium values), the stochastic nature of the genetic process (drift, mutation, and recombination all occur randomly) and variance introduced by sampling only 20 of 200 hosts.

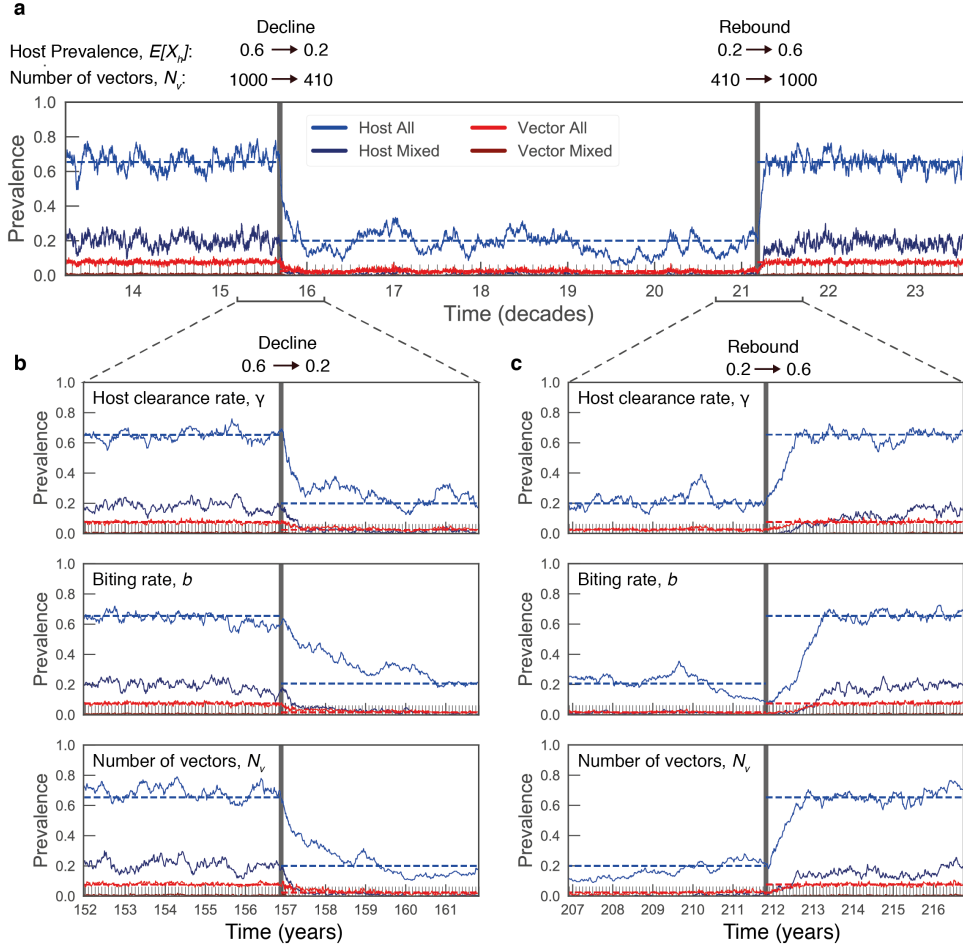


Figure 2.12: A three epoch framework for **forward-dream**. (a) Host and vector prevalence visualized across three epochs. Epochs are delineated by thick vertical bars. In the first epoch there are one-thousand vectors ($N_v = 1000$) and the equilibrium host parasite prevalence is 0.65 ($E[X_h] = 0.65$). In the second epoch, the number of vectors is reduced to 410 and the equilibrium host parasite prevalence becomes 0.2. In the third epoch, the number of vectors is returned to 1000. Blue and red lines give prevalence of infection across the host and vector populations, respectively. Similarly, dark blue and dark red lines give the prevalence of mixed infections (infections containing > 1 distinct parasite) across the host and vector populations, respectively. Horizontal blue and red dashed lines give equilibrium prevalence values predicted from the Ross-Macdonald model (Equations 2.1 and 2.2) (b) Same as (a), but focused on transition period between epochs 1 and 2, where of prevalence is in decline. From top to bottom, three panels show prevalence declining when either host clearance rate (γ), vector biting rate (b), or the number of vectors (N_v) is changed. Equilibrium host parasite prevalence changes from 0.65 to 0.2 in all cases. (c) Same as (b), but focused on transition period between epochs 2 and 3, where prevalence rebounds.

To better discern how different genetic diversity statistics respond to changes in host parasite prevalence, we averaged the trajectories of each statistic across

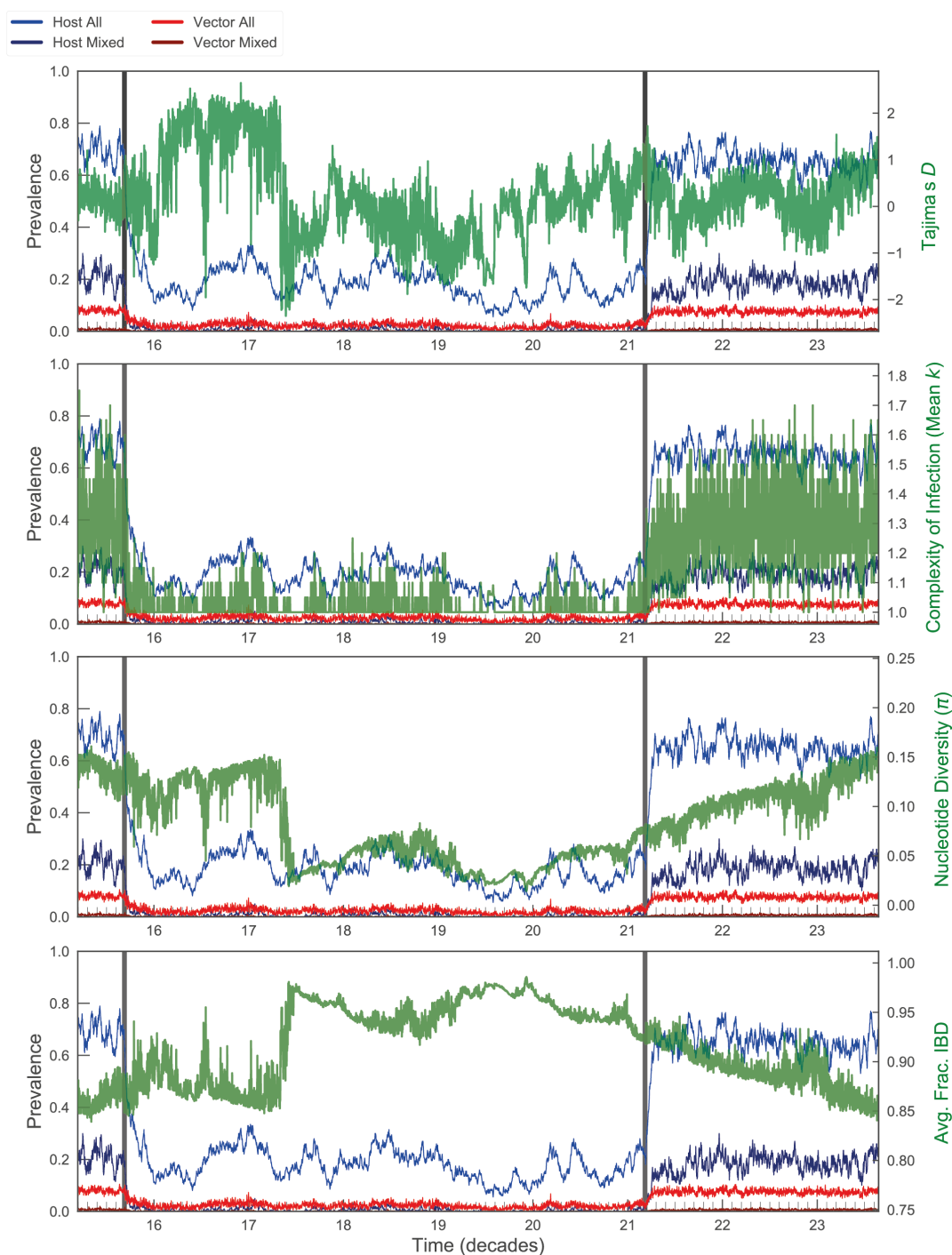


Figure 2.13: Behaviour of four genetic diversity statistics in a single 3 epoch forward-dream simulation. Each panel shows host and vector prevalence in blue and red, and the genetic diversity statistic in green on the second y-axis indicated at right. From top to bottom, panels show Tajima's D , the mean complexity of infection, nucleotide diversity, and the mean IBD. In the simulation shown, the number of vectors was changed to change host parasite prevalence.

100 replicate 3 epoch **forward-dream** simulations⁸. The value for a given genetic diversity statistic in each interval was then averaged across 100 replicate experiments. The averaged trajectories Tajima's D , mean C.O.I., nucleotide diversity, and average pairwise IBD are shown in Figure 2.14. Several observations emerge. First, consistent with theory, Tajima's D increases above zero in the period where host parasite prevalence is declining, indicative of a contracting effective population size (N_e) (Tajima 1989). Similarly, Tajima's D falls below zero in the period where host parasite prevalence is climbing, indicative of an expanding N_e . Second, there are clear differences in the rate at which different genetic diversity statistics respond to change. For example, the mean C.O.I. changes much faster than nucleotide diversity or the pairwise fraction of the genome in IBD (Figure 2.14). Finally, the rate at which a given genetic diversity statistics responds to a decline in host parasite prevalence may be different from the rate at which it responds to an increase in host parasite prevalence (Figure 2.14).

To better understand these differences, I developed two metrics to describe how a given genetic diversity statistic responds to a change in host parasite prevalence. The first is the "time until detection" (t_d), which describes the time it would take to detect a change in host parasite prevalence if that change was monitored for using a given genetic diversity statistic. In the three epoch simulations, the "time until detection" is computed by determining the first day after a host parasite prevalence change has been induced (i.e. after the beginning of the second and third epochs) that a genetic diversity statistic has a value outside of $\mu \pm 1.96\sigma^2$, where μ is the mean and σ^2 is the variance of the genetic diversity statistic in the previous (i.e. pre-intervention) epoch. The second measure is the "time until equilibrium" (t_e), which describes the time it takes for a given genetic diversity statistic to reach its new equilibrium value following a host parasite prevalence change. Both statistics capture the latency with which information about host parasite prevalence change is manifest in genetic diversity.

⁸Note that **forward-dream** operates in continuous-time. To average trajectories, time was first discretised into 25 day intervals. Time was discretised into smaller intervals, of 1 day in duration, at the beginning of each epoch in order to increase resolution.

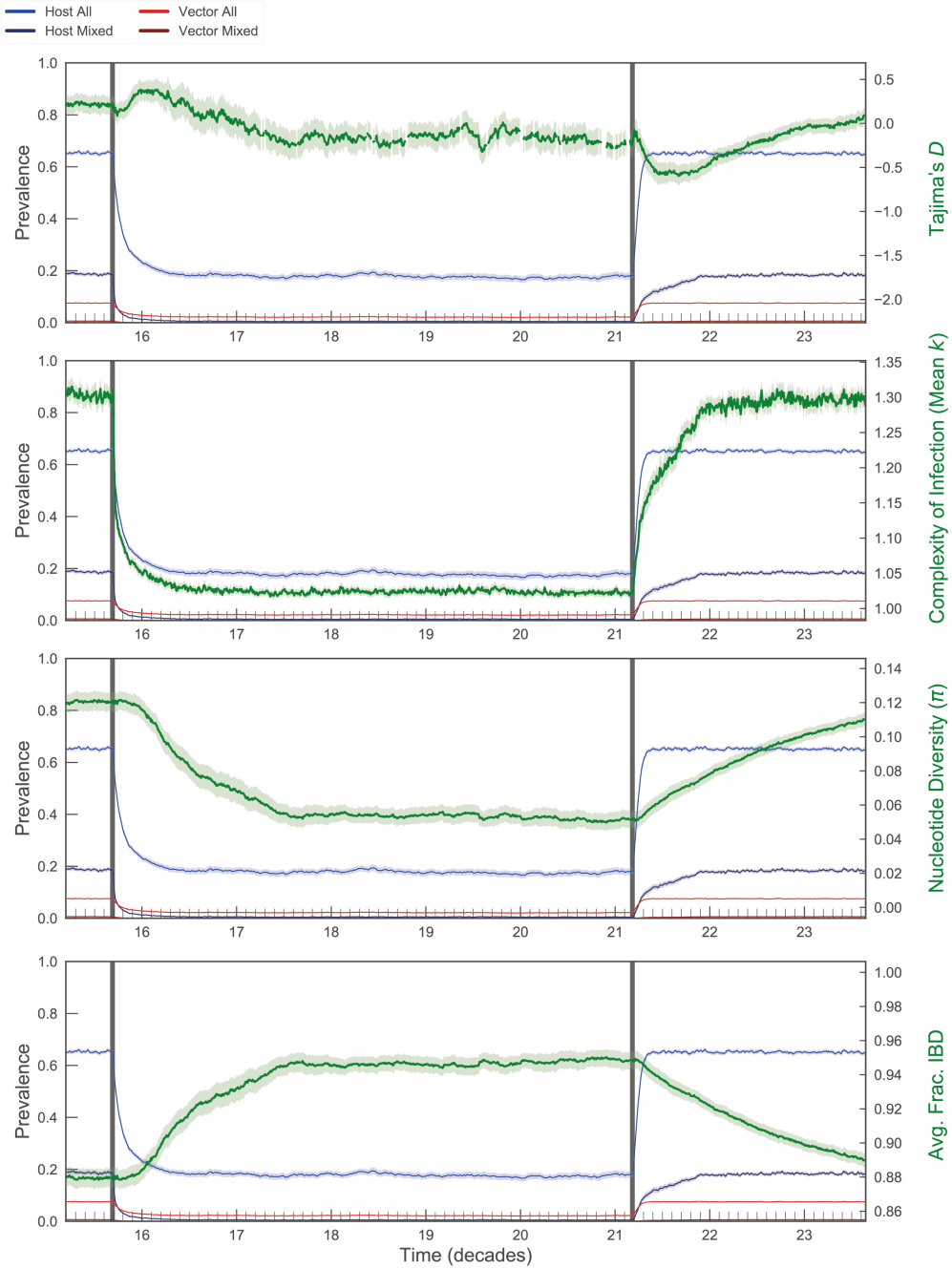


Figure 2.14: Average behaviour of four genetic diversity statistics across 100 replicate 3 epoch **forward-dream** simulations. Each panel shows host and vector prevalence in blue and red, and the genetic diversity statistic in green on the second y-axis indicated at right. Shaded area shows the 95% confidence intervals for the mean of the diversity statistic across the 100 replicate experiments.

Figure 2.15 shows both t_d and t_e for all considered genetic diversity statistics following the host parasite prevalence decline at the beginning of the second

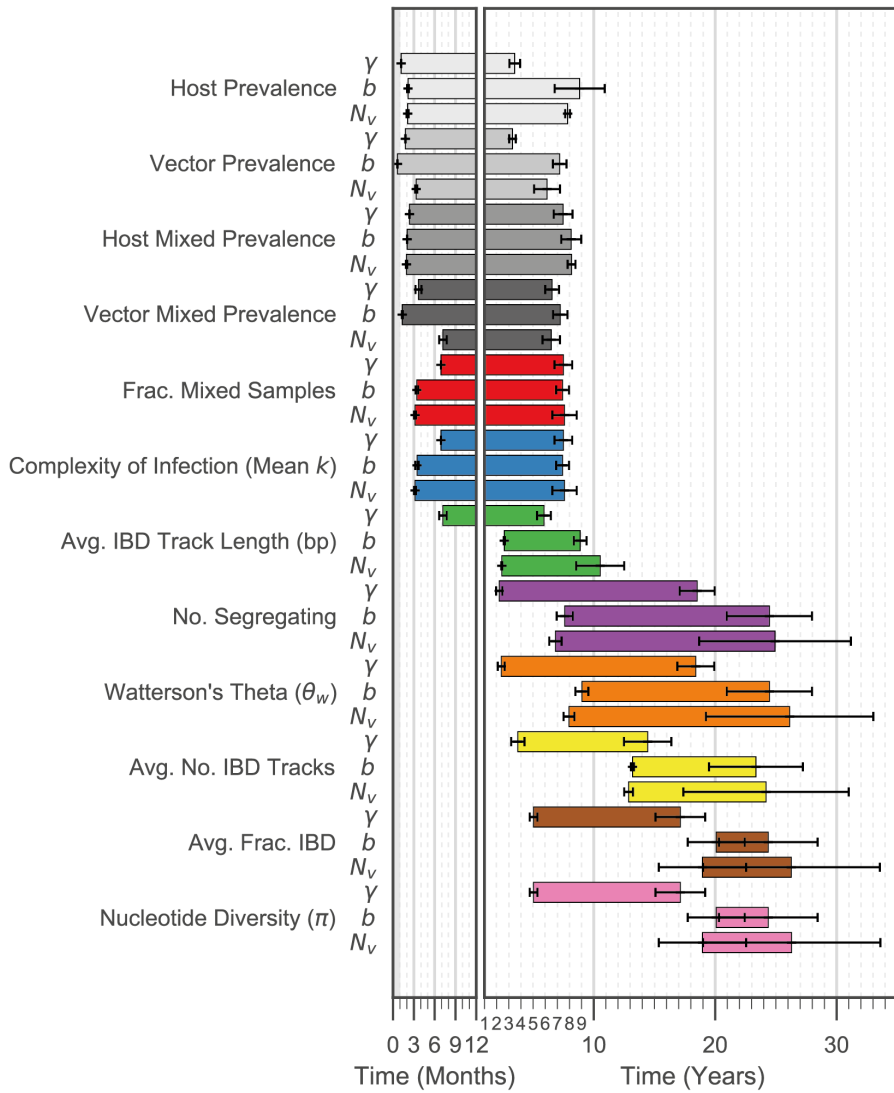


Figure 2.15: Temporal response of different genetic diversity statistics to a decline in host parasite prevalence, induced at the beginning of the second epoch in a 3 epoch **forward-dream** simulation. The x-axis gives time (with the first year expanded for clarity) and each individual horizontal bar presents the time until detection (left end) and time until equilibrium (right end), as well as their 95% confidence intervals across 100 replicate 3 epoch **forward-dream** simulations. Each genetic diversity statistic has its own colour and three associated bars giving the response times when different parameters are changed: top bar is host clearance rate was changed (γ), middle is vector biting rate (b), and bottom is number of vectors (N_v). Host and vector prevalence and rates of mixed infection (across the population) are also shown in greys. Statistics are ordered according to their mean time until detection across the three intervention types.

epoch. Additionally, the host and vector prevalence, as well as their rates of

mixed infection⁹ are shown in grey.

The fastest responding statistics are those related to mixed infections: both the fraction of mixed samples and mean C.O.I. have t_d values of approximately three months when either vector biting rate (b) or the number of vectors (N_v) is changed. Excluding these, many statistics have a very large t_d . For example, the t_d of nucleotide diversity is between 5 - 20 years, depending on intervention type. Interestingly, changing the host clearance rate (γ) increases the time until detection (t_d) for the mixed infection related statistics, but decreases it for all other statistics. For all statistics, there is a considerable delay between detection and equilibrium, such that $t_e - t_d$ is between 5 and 15 years, depending on the statistic and intervention.

In 3 epoch **forward-dream** simulations, both the time until detection (t_d) and time until equilibrium (t_e) tend to be greater for host parasite prevalence increases than for decreases (Figure 2.16). For example, for mixed infection related statistics, the t_d is 6 months rather than 3 months for prevalence increases induced by changing the vector biting rate or number of vectors; and a year rather than 6 months for those induced by changing the host parasite prevalence clearance rate. For many other statistics the increase in t_e is more pronounced. For example, the t_e for nucleotide diversity climbs from approximately 5-25 years to 25-85 years (depending on intervention type) when monitoring for a prevalence increase as opposed to a decrease.

2.4.2 Seasonality

Given the magnitude of detection and equilibrium times, I wanted to use **forward-dream** to explore whether any genetic diversity statistics would respond to seasonal changes in prevalence. To this end, I developed a simple seasonality framework in **forward-dream**, where the number of vectors oscillates between a peak of $N_v = 1000$ in the wet season, and $N_v = 10$ in the dry season. Vectors that die

⁹Note that host and vector prevalence also fluctuate stochastically and so t_d and t_e can be calculated. Note also that host and vector mixed infection prevalence refers to the entire population, whereas the genetic diversity statistic fraction mixed infection refers only to the sample.

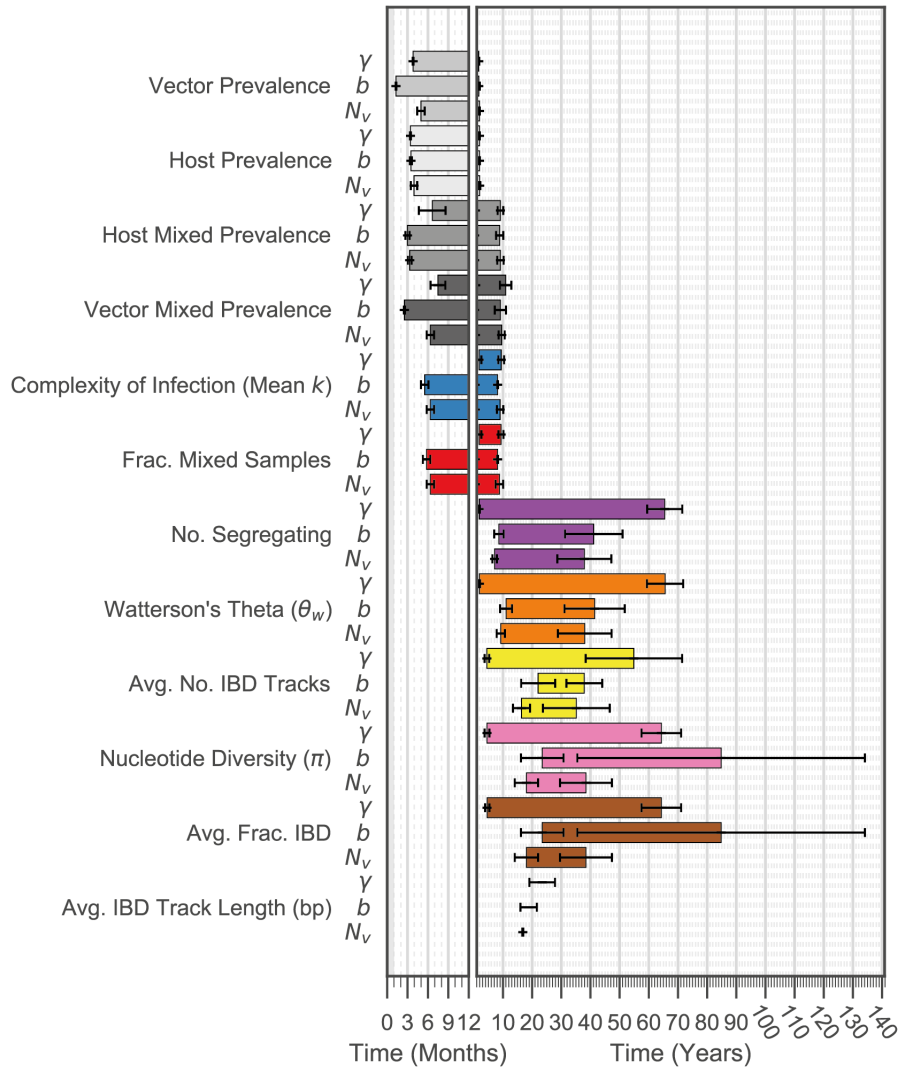


Figure 2.16: Temporal response of different genetic diversity statistics to a increase in host parasite prevalence, induced at the beginning of the third epoch in a 3 epoch forward-dream simulation. See Figure 2.15 for details.

entering the dry season are selected at random, with the dry season lasting 170 days and the wet season lasting 195 days¹⁰. In this simulation, only the fraction of mixed infections and mean complexity of infection show clear responses to seasonal change in prevalence (Figure 2.17).

¹⁰Guided by seasonality data from northern Ghana, courtesy of Cristina Ariani.

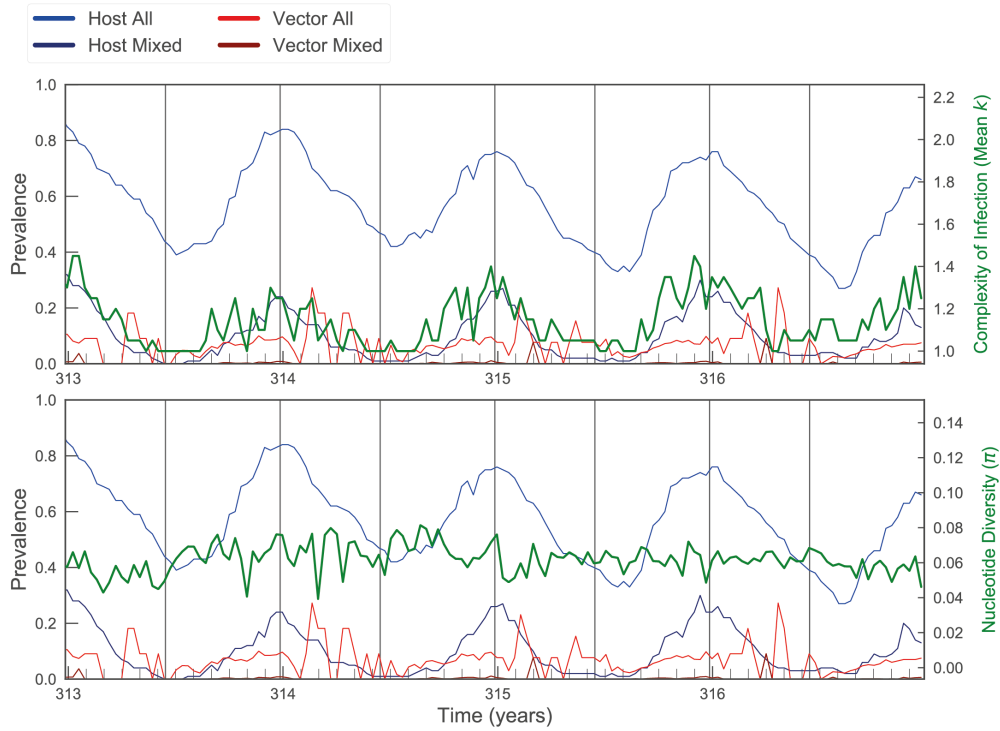


Figure 2.17: Simulating seasonality using **forward-dream**. An individual year has two seasons delineated by vertical grey lines: the dry season, where host parasite prevalence declines; and the wet season, where host parasite prevalence climbs. The decline in prevalence is induced by reducing the number of vectors in the dry season. The values of two genetic diversity statistics across seasons are indicated in green. Top panel, mean complexity of infection. Bottom panel, nucleotide diversity.

2.5 Conclusion

The relationships between statistics summarizing parasite genetic diversity and the fundamental parameters of malaria epidemiology are unknown. While classical models in population genetics are analytically tractable and can be used to make demographic inferences about populations, they do not readily apply to malaria. On the other hand, classical vector-based disease models do not explicitly incorporate the genetic process. Here, we merged the Ross-Macdonald model – capturing population-level transmission and equilibrium – with a continuous-time Moran model – capturing the intrahost evolution of malaria infections – thereby linking genetic processes with epidemiological ones. We implemented the resulting model as an agent-based simulation called **forward-dream**.

We used **forward-dream** to identify a suite of genetic diversity statistics that are correlated with host parasite prevalence, though to differing degrees. A striking observation was that the magnitude of many of the correlations with host parasite prevalence were sensitive to the epidemiology underlying the prevalence change. In particular, changes in host parasite prevalence caused by a faster average host clearance time (for example, due to quicker recourse to treatment) produced stronger correlations in statistics related to IBD or the size of the sample genealogy. One possible explanation is that a given parasite lineage oscillates through host and vector populations, with the host population typically being the smaller of the two, and therefore having a greater impact on the effective population size¹¹. Thus changing the host clearance rate will have a more pronounced effect on the overall effective population size – which influences statistics like the nucleotide diversity, number of segregating sites, and number of singletons – than changing parameters that influence the vectors. Such an observation is relevant, as in natural settings which epidemiological parameters underlie host parasite prevalence change will be unknown, and thus the statistics used for surveillance should be robust to different causes of change. To this point, we found that in **forward-dream** statistics related to mixed infections were responsive to host parasite prevalence change, regardless of the underlying epidemiological cause.

That a given genetic diversity statistics is correlated with prevalence in an equilibrium setting is a prerequisite for its utility in surveillance, but it does not capture how rapidly the statistic responds to change. To address this question, we ran a series of **forward-dream** simulations where host parasite prevalence was changed in a manner designed to mimic real-time interventions, and monitored a set of genetic diversity statistics summarizing the parasite population throughout. An important, if obvious, observation was that genetic diversity statistics can be highly variable through time. This can also be seen in the boxplots summarizing the distribution of genetic diversity statistics at fixed prevalence values – several of which have relatively large interquartile ranges (Figure 2.9). Though here we averaged

¹¹The effective population size across a set of populations is the harmonic mean of those population sizes.

across several replicates to determine the mean behaviour of different diversity statistics, in practice this will not be feasible. The overall variance in a given genetic diversity statistic can be partitioned into at least three contributions: (i) variance due to stochasticity in the epidemiological process; (ii) variance due to stochasticity in the genetic process; and (iii) variance due to random sampling. All statistics are subjected equally to variance in the epidemiological process – i.e. random fluctuation in host and vector prevalence at a given set of fixed epidemiological parameter values – but may be influenced by it to different degrees. Some genetic diversity statistics have notably higher variance as a consequence of randomness in the genetic process (drift, mutation, and recombination) than others. For example, in a neutral Wright-Fisher population, nucleotide diversity has a co-efficient of variation (CV)¹² of 0.7 while the number of segregating sites has a CV of 0.3 (Wakeley 2016). Consistent with this, the variance explained by the number of segregating sites was higher than that explained by nucleotide diversity in **forward-dream**. We note that the population and genome size simulated in **forward-dream** were both relatively small (200 hosts and 1000 SNPs) which may increase variance, but an average of 10% of the total population was sampled, which is likely higher than what would be achieved in a field setting. Determining the relative contributions of epidemiological, genetic and sampling variance to the total variance of a genetic diversity statistic, and exploring whether sampling a larger fraction of the genome or more individuals could lower this variance, is a point for future work.

Another important area of future work is to explore the behaviour of genetic diversity statistics in lower transmission settings. In the present analysis, host prevalence values below 30% were not included, as simulations with a low prevalence have a tendency to experience stochastic extinction of the parasite population. Particularly when the aim is to achieve both epidemiological and genetic equilibrium (which can take many decades), these extinction events hinder analysis. However, many regions around the globe (such as Southeast Asia and South America) have prevalence values entirely below 30% (Weiss et al. 2019). At least two

¹²The co-efficient of variation is equal to the ratio of the standard deviation to the mean, $CV = \sigma/\mu$.

improvements can be considered here. The first is to attempt to increase the size of the host and vector populations, which should decrease the likelihood of extinction. This will require further optimization of **forward-dream** to reduce run-times, as currently the simulation is computationally expensive even with only two-hundred hosts and a thousand vectors. Secondly, the extinction events experienced in **forward-dream** likely reflect what actually occurs in low prevalence regions: small-scale extinction and re-colonization through migration (for example see Reiner et al. (2015a)). Thus, incorporating a migration framework into **forward-dream** may both stabilize the simulation at low prevalence values, while also creating a more realistic representation of what occurs in many areas with unstable transmission.

A key observation from Section 2.4 was that some diversity statistics respond much more quickly to changes in prevalence than others. In particular, statistics related to mixed infections can be expected to respond to host parasite prevalence change on the order of months – whilst a statistic like nucleotide diversity may take 5 - 20 years to respond. An explanation is that the statistics that are influenced by the time to coalescence – i.e. N_e – take longer to respond. Nucleotide diversity, in particular, is strongly influenced by coalescent events that have occurred deeper in the history of the genealogy of a sample. An implication is that though these statistics may correlate with prevalence at equilibrium, they are unsuitable for monitoring host parasite prevalence change longitudinally over short time-scales. Such an interpretation is supported by genetic data collected from along the Thai-Burmese border, where across a ten-year period a pronounced decline in prevalence was associated with a decreasing number of mixed infections, but nucleotide diversity remaining constant (Nkhoma et al. 2013).

It is worth noting that **forward-dream** was designed to represent idealized conditions, in that we modelled no migration, no selection, no heterogeneous biting, no spatial or temporal variation in prevalence, and we enforced a perfectly uniform sampling procedure across all prevalence ranges and through time. As a consequence of this, we expect that our results represent something like a theoretical upper limit on the ability of differences in prevalence to explain genetics. Yet, many

of these phenomenon will occur in natural populations of *P. falciparum* and how robust our observations are to their occurrence is an important consideration. Malaria simulations including more complex epidemiological processes (although not explicitly modelling genetics) do exist (reviewed in N. R. Smith et al. (2018)). However, they have the associated difficulty of requiring more parameterisation, and data supporting specific choices of parameter values may be sparse. **forward-dream** has only 14 parameters which were carefully selected following a literature review, and its simplicity makes its assumptions more explicit.

Finally, only a small set of genetic diversity statistics were considered here, and the genetic material was represented by a single, 1000 SNP chromosome. Incorporating a more realistic genome model that would allow the study of statistics such as linkage disequilibrium, would be valuable future work.

3

Epidemiological inference from *P. falciparum* mixed infections

Contents

3.1	Introduction	63
3.1.1	Deconvoluting mixed infections	66
3.1.2	Incorporating relatedness between strains	71
3.2	IBD and malaria meiosis	72
3.2.1	Simulating malaria meiosis with <code>pf-meiosis</code>	72
3.2.2	Simulation validation	74
3.2.3	Analysis of oocysts and IBD	76
3.3	IBD and the classes of mixed infections	81
3.3.1	Simulating mixed infection chains with <code>pf-meiosis</code>	81
3.3.2	Classification of mixed infections across Africa and Asia	83
3.4	Mixed infection classification and $PfPR_{2-10}$	86
3.5	Conclusion	88

3.1 Introduction

In this chapter we explore malaria mixed infections, seeking to extract information about their composition and how it relates to the mechanisms by which they are created. Mixed infections (variously called multiple, complex, or polygenomic infections) are infections wherein more than one genetically distinct *P. falciparum* strain is present in the blood of a host (Figure 3.1a). *P. falciparum* is haploid within

the host, and so mixed infections can be identified by the presence of multiple alleles at a single locus. Historically, the loci under investigation were microsatellite repeat regions or highly variable surface antigen genes (typically *MSP1*, *MSP2* or *GLURP*). In these cases length variation, detectable by gel electrophoresis, indicated the presence of multiple alleles (Viriyakosol et al. 1995). More recently, work has focused on profiling polymorphic loci using 24- or 96-SNP genotyping arrays (deemed "molecular barcoding"), where heterozygous calls suggest a mixed infection (Daniels et al. 2008). In principle, whole-genome sequencing (WGS) allows identification of mixed infections through the analysis of any polymorphic region with adequate coverage, though to date the focus has been on high-quality biallelic SNPs (O'Brien et al. 2016; Zhu et al. 2018).

Regardless of the technology used for their identification, mixed infections have been found ubiquitously for as long as genetic data has been collected from malaria-containing blood samples. Paul et al. (1995) found that 72% of infections from Madang in Papua New Guinea contained multiple alleles of *MSP2*, *MSP1* or *GLURP*. Anderson et al. (2000) assayed microsatellite length variation in 465 infections collected from 9 global locations, and found that mixed infections comprised greater than 60% of all infections in Africa, with lower but still considerable percentages found in Asia and South America. The malaria genome was first sequenced in 2002 (Gardner et al. 2002), and later similarly high and geographically patterned rates of mixed infection were discovered through whole-genome sequencing, assessed by levels of within-host heterozygosity (Manske et al. 2012).

Beyond their sheer abundance, mixed infections have promising links to malaria epidemiology. Most directly, several authors have reasoned that the rate of mixed infection should be a function of transmission intensity, and observed variation in both space and time often supports this view (Anderson et al. 2000; Volkman et al. 2012). Anderson et al. (2000)'s observation of lower rates of mixed infection in Asia and South America relative to Africa is consistent with malaria prevalence being highest in Africa (MAP 2017). Where transmission intensity data has been unavailable, correlations between rates of mixed infection and altitude have been

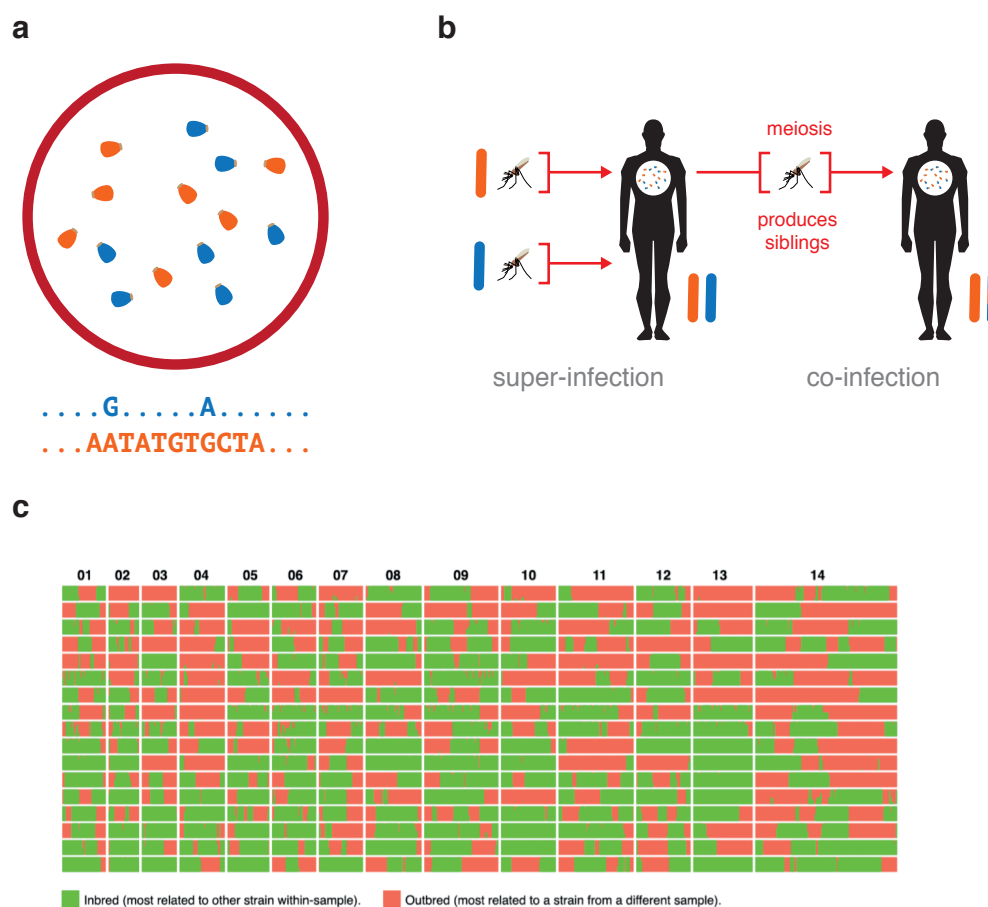


Figure 3.1: Mixed infection properties and mechanisms of creation (A) Two strains within a single host blood sample lead to heterozygous loci by SNP arrays or sequencing. (B) A $K = 2$ mixed infection can be generated by two independent mosquito bites, known as super-infection (left). Alternatively, a mosquito harbouring two strains can deliver both in a single bite, known as co-infection (right). Since meiosis occurs within the mosquito, these strains will be related and share tracts of IBD. (C) Extensive IBD observed in *P. falciparum* samples collected from Ghana as part of the Pf3k data resource (analysis by Joe Zhu and Jacob Almargo-Garcia, unpublished). Rows are individual samples and columns demarcate the 14 chromosomes of *P. falciparum*. Green indicates IBD.

observed (Boyce et al. 2018)¹. A temporal link between transmission and mixed infections was convincingly established in two longitudinal studies, the first along the border of Thailand and Burma, and the second in Thies, Senegal (Nkhoma et al. 2013; Daniels et al. 2013). On the Thai-Burmese border, clinical incidence of malaria fell ~ 12 -fold between 2000 and 2010, and data from over a thousand

¹Malaria prevalence tends to decline with higher altitude (Githeko et al. 2006).

malaria samples collected during that period showed a concomitant decline in the rate of mixed infections, from 63% to 14% (Nkhoma et al. 2013). Senegal’s National Malaria Control Program (NMCP) was redesigned 2006, and by 2009 it was reported that confirmed cases had decreased by 95% (Daniels et al. 2015). Over the same time period, molecular barcode data indicated a decline in the rate of mixed infection from 30% to 20% (Daniels et al. 2013).

3.1.1 Deconvoluting mixed infections

Though molecular barcoding and whole-genome sequencing have underpinned the large-scale collection of malaria genetic data (Nkhoma et al. 2013; Daniels et al. 2013; Manske et al. 2012), they have also created unique challenges for interpreting mixed infections. Formerly, with the genotyping of microsatellite repeats or surface antigen genes, an estimate of the number of strains in a mixed infection (referred to as the complexity or multiplicity of infection) could be obtained simply by counting the bands produced by gel electrophoresis. Neither molecular barcoding nor whole-genome sequencing admit such a direct approach, and their use convolutes much of the structure of a mixed infection. Consider that with either technology, data from a mixed infection will be derived from an unknown number of strains, each comprising an unknown proportion of the infection, and each with an unknown haplotype. Recovering this structure requires statistical modelling.

COIL

The first method developed to address this problem was COIL, which estimated the number of strains in a mixed infection detected by molecular barcoding (Galinsky et al. 2015). Imagining strains within a mixed infection as randomly sampled from a population, the number carrying the alternative allele at a given SNP will be $\text{Bin}(K, p)$ where K is the complexity of infection, and p is the population-level alternative allele frequency (PLAF) of the SNP. The probability a given SNP is observed to be heterozygous is then $1 - p^K - (1 - p)^K$, which climbs with K . COIL assumes all SNPs on the barcode are unlinked, such that the final likelihood is

the product of the individual SNP likelihoods. Folded into a Bayesian approach (though the prior on K is uniform), the software tests K from 1 to 5 and returns the *Maximum a posteriori* estimate. With COIL, the PLAF (p) for each SNP is estimated from clonal samples, which may be limited in number. Later, the same model (renamed THE REAL McCOIL) was incorporated into a Metropolis-Hasting algorithm which allowed for simultaneous inference of K and p across all the barcodes (clonal and mixed) collected in a region (Chang et al. 2017).

pfmix

Ultimately, the signal used by methods built around molecular barcoding is the observed number of heterozygous sites, which increases with the complexity of infection². Methods developed on whole-genome sequencing data have the advantage not only of more SNPs to scrutinize, but also the ability to convert categorical labelling of SNPs (i.e. heterozygous or homozygous) into a within-sample alternative allele frequency (WSAF), based on counting the number of reference and alternative calls in reads spanning a given loci (the WSAF at a given SNP i is $q_i := a_i / (a_i + r_i)$, where a_i , r_i are the counts of alternative and reference alleles, respectively).

The first method to take advantage of this was **pfmix** (O'Brien et al. 2016). In their manuscript, O'Brien et al. (2016) point out that the WSAF across all SNPs should be multimodal, whereby an infection comprised of K strains will have a WSAF distribution with 2^K modes. Each WSAF mode (indexed by r) contains the set of SNPs at which the same subset of the K strains carry the alternative allele. In other words, each WSAF mode results from a different underlying allelic configuration amongst the K strains. If those strains comprise proportions w_j of the infection, the WSAF mode will have an expected value equal to the sum of those proportions ($q_r = \sum_{j \in r} w_j$). Visualized along the genome, the multimodal nature of the WSAF yields "bands".

²In fact, if the SNPs across the barcode have the same PLAF, the number of heterozygous sites itself is $\text{Bin}(n, p_{het})$, with n being the number of SNPs on the barcode and $p_{het} = 1 - p^K - (1 - p)^K$ as before.

In `pfmix`, the probability that a SNP has a particular WSAF is given as a finite-mixture model with a component for each of the 2^K bands (O'Brien et al. 2016). If a SNP i has an allelic configuration corresponding to band r with WSAF q_r , it is assumed the read data (a_i, r_i) will follow a Beta-binomial distribution:

$$P(a_i, r_i | q_r, \nu) = \binom{a_i + r_i}{a_i} \frac{B(a_i + \nu q_r, r_i + \nu(1 - q_r))}{B(\nu q_r, \nu(1 - q_r))}$$

where O'Brien et al. (2016) have reparameterized such that $\alpha = \nu q_r$ and $\beta = \nu(1 - q_r)$. This fixes the expected WSAF at q_r , as required, but allows for the over-dispersion typical of whole-genome sequencing data, with the hyperparameter ν . For any given SNP, however, the allelic configuration of the strains within the mixed infection is not known, and so it is not known to what band r the SNP belongs. Therefore, it is modelled as a mixture across all possible bands:

$$f_{Q_i}(q_i) = \sum_{r=1}^{2^K} \lambda(r, p_i) P(a_i, r_i | q_r, \nu)$$

The weights in the mixture, which describe the probability of different allelic configurations for the SNP i , are, similarly to `COIL`, considered as resulting from Bernoulli sampling from the PLAF:

$$\lambda(r, p_i) = p_i^{C_r} (1 - p_i)^{K - C_r}$$

where C_r is the number of strains that carry the alternative allele in band r . Finally, SNPs are assumed unlinked, and an MCMC approach is used to construct the posterior distribution. The value-added from whole-genome sequencing is clear: the parameters inferred include the number of strains K , as well as their proportions w_j , and the proportion of unexplained mixture α^3 .

³O'Brien et al. (2016) model a mixed infection as a weighted mixture of K distinct strains, as well as a fraction α that is panmictic with the reference population. However, in my experience running `pfmix` on the Pf3k data set, α almost always goes to zero in early iterations of the MCMC.

DEploid

Both approaches described above imagine that in a mixed infection of K strains, the allelic configuration at each SNP is sampled from some reference population as K Bernoulli trials. This assumption holds if the SNPs considered are sufficiently distant to be in linkage equilibrium, and the population used to parameterize the Bernoulli distributions is panmictic with the mixed infection in question. Unfortunately, the first of these two requirements makes reconstructing haplotypes impossible. Linkage equilibrium by definition means there is no association between SNPs; a haplotype is an associated set of SNPs. Even if these methods specifically estimated the allelic configuration at each SNP (rather than computing a weighted average over all possible configurations) they would be unable to phase the configurations into haplotypes, ultimately leaving the structure of the mixed infection ambiguous. Notably, though linkage equilibrium is a passive consequence of the sparse genotyping arrays used for molecular barcoding, it is achieved through the active omission of SNPs for WGS data.

The first method to take advantage of the dense SNP data provided by WGS to estimate the haplotype structure of mixed infections was **DEploid** (Zhu et al. 2018). Given K strains with w_j proportions, **DEploid** employs a per-SNP likelihood very similar to **pfmix** – the only difference being it ignores the possibility of a panmictic fraction (α in **pfmix**) and introduces sequencing error. However, there are major changes in how it models (and samples) the underlying haplotypic structure. Critically, it replaces the mixture-model over all 2^K allelic configurations with K explicit, phased, haplotypes. The key innovation is a modified version of the Li and Stephens algorithm that proposes haplotypes within the mixed infection as an "imperfect mosaic" of a user-defined reference population (N. Li and Stephens 2003) *and* from the posterior distribution of the model, given K , w_j , and the observed sequencing data. Briefly, this is practiced by first removing one of the K strains (call this s) at random and computing a partial WSAF in its absence:

$$q_{i,-s} = \sum_{j \neq s} w_j h_{ij}$$

where h_{ij} is the allelic state ($\in [0, 1]$) of haplotype j at SNP i . If we were to replace this strain, the replacement allelic state at site i could be either reference or alternative $\epsilon_i \in [0, 1]$. This would create a new expected WSAF q_i :

$$\begin{aligned} q_{i,\epsilon_i=0} &= q_{i,-s} \\ q_{i,\epsilon_i=1} &= q_{i,-s} + w_s \end{aligned} \tag{3.1}$$

Now we imagine the excluded strain is replaced by a haplotype from the reference panel h_{iz} and allow for mis-copying at a fixed rate μ . The likelihood of our selected replacement haplotype is a mixture over the mis-copying process, and the probability of the data given resulting allelic state:

$$L(h_{iz}) = \begin{cases} (1 - \mu)P(a_i, r_i | q_{i,\epsilon_i=0}, \nu) + \mu P(a_i, r_i | q_{i,\epsilon_{ij}=1}, \nu) & \text{if } h_{iz} = 0 \\ \mu P(a_i, r_i | q_{i,\epsilon_i=0}, \nu) + (1 - \mu)P(a_i, r_i | q_{i,\epsilon_{ij}=1}, \nu) & \text{if } h_{iz} = 1 \end{cases} \tag{3.2}$$

These are now used as emission probabilities in the hidden Markov Model that underlies the Li and Stephens algorithm (N. Li and Stephens 2003). The forward matrix is initialized by setting all strains in the reference panel as equally likely:

$$F(z, 1) = \frac{1}{n} L(h_{1z})$$

Where n is the number of haplotypes in the set of reference haplotypes. Moving to the next site $i + 1$, recombination can occur with probability ρ , which depends only on the distance between sites. The forward matrix recursion is then:

$$F(z, i + 1) = \left[\frac{\rho}{n} \sum_{m=1}^n F(m, i) + (1 - \rho)F(z, i) \right] L(h_{(i+1)z})$$

Once the forward matrix is populated, a haplotype can be sampled as a path through the reference population, allowing mis-copying, and from the posterior of the model given a_i, r_i, K and w_j , as was desired. This process is embedded into an MCMC that iterates randomly between updating the proportions w_j and haplotypes. K is

fixed at some upper limit (typically 4) and at the end of the MCMC strains with proportions below some threshold (typically 0.05) are discarded.

This approach made **DEploid** the only method to completely deconvolute a mixed infection: recovering the number of strains, their proportions, and their haplotypes. Applied the Pf3k project, it converted 2393 samples into 3814 haploid genomes.

3.1.2 Incorporating relatedness between strains

All of the methods above have one assumption in common: that the strains within a mixed infection are independently drawn from the population, i.e., they are unrelated. Yet, the life cycle of *P. falciparum* guarantees that this assumption will be violated in at least some mixed infections. A mixed infection of K strains can be produced by either K independent infectious bites or by $j < K$ infectious bites (Figure 3.1b). The first case is commonly referred to as super-infection; each genetically distinct parasite is delivered by a separate vector. The second case entails co-infection; multiple parasite strains being delivered from a single vector. Critically, *P. falciparum* undergoes meiosis within the mosquito vector, a fact that suggests strains delivered through co-infection should be siblings and violate the unrelatedness assumption. Indeed, work by Jacob Almargo-Gracia and Joe Zhu of the McVean group identified extensive identity-by-descent (IBD) in a panel of mixed infections from Ghana (Figure 3.1c), suggesting that mixed infections harbouring related strains do occur.

These reflections motivated further development of **DEploid** to incorporate the possibility of relatedness between strains. The essential change was noting that the effect of IBD is to limit the number of distinct allelic configurations possible, in a way that depends on the pattern of IBD sharing. This restriction in turn limits the expected WSAF values. The new method was called **DEploidIBD** and inferred – in addition to the number of strains, their proportions, and haplotypes – an IBD profile for every pair of strains in the infection.

The work in this chapter was undertaken with the aim of interpreting the IBD patterns produced by **DEploidIBD**. In particular, I wanted to understand:

1. How IBD patterns within mixed infections relate to the meiotic process (Section 3.2);
2. If IBD patterns within mixed infections could be leveraged to learn about the infection process (i.e. super- vs. co-infection) (Section 3.3);
3. To what extent IBD patterns within mixed infections reflect regional transmission intensity (Section 3.4).

Practically, this involved first developing a simulation of malaria meiosis, called **pf-meiosis**. Second, using it to classify all $K = 2$ mixed infections present in the Pf3k data resource based on the most-likely mechanism by which they were generated. And third, comparing the classification results in different geographical regions with prevalence data from the Malaria Atlas Project (MAP 2017). The work was published in *ELife*, and Figures 3.4, 3.5, 3.8, 3.9 as well as portions of the text below are from that reference (Zhu et al. 2019).

3.2 IBD and malaria meiosis

3.2.1 Simulating malaria meiosis with **pf-meiosis**

In order to understand the expected levels of relatedness in a malaria mixed infection, I developed simulation tool called **pf-meiosis**. The tool allows for simulation of malaria meiosis given two parental strains, capturing unique aspects of the process that impact resultant levels of IBD. Malaria meiosis and the structure of the simulation are described briefly below.

When a mosquito vector bites an infected host, it may imbibe any number of gametocytes, the pre-sexual form of *P. falciparum* that exists in small percentages in the blood (Bousema and Drakeley 2011). These gametocytes are gendered, and if the host is carrying a mixed infection, may be of multiple distinct strains (Bousema and Drakeley 2011). Once inside the mosquito midgut, gametocytes differentiate into gametes (likely as a result of temperature change and the detection

of specific compounds), and mating between male and female forms can begin (Ghosh et al. 2000).

During mating two gametes fuse to form a zygote, and the ploidy of the cell doubles (Ghosh et al. 2000). The zygote then differentiates into a motile cell form known as an ookinete, which migrates towards and then penetrates through the midgut wall (Ghosh et al. 2000). Microscopic evidence suggests that during this stage the main steps of the meiotic process take place (Sinden and Hartley 1985). DNA replication occurs, producing pairs of sister chromatids joined along their length by the ring-shaped protein cohesin, and resulting in a cell ploidy of $4n$ (Janse et al. 1986). Programmed double-stranded breaks then allow for pairing of homologous chromosomes into a 4-chromosome structure known as a bivalent, which facilitates crossing-over and appropriate segregation later on (reviewed in Tsai and McKee (2011)). Finally, each of bivalent (one for each of the fourteen chromosomes of *P. falciparum*) undergoes segregation from $4n$ down to $1n$. At this point, the ookinete has begun differentiating into an oocyst, which exists on the outer-wall of the mosquito midgut, thus escaping the process of blood-meal digestion. Within the oocyst, the $1n$ meiotic progeny form their own cell walls and undergo exponential growth for a period of approximately ten days (Aly et al. 2009). Finally, the oocysts burst, releasing thousands of sporozoites into the hemolymph from which they eventually localize in the mosquito salivary glands. The emission of saliva as an anticoagulant in a subsequent biting event releases a sample of tens to hundreds of these sporozoites into a new host (Ghosh et al. 2000).

The simulation framework of **pf-meiosis** begins after gametes have been imbibed by the mosquito vector. In **pf-meiosis**, at most two unique types of gametes are simulated, and mating occurs randomly between all available types. As a consequence, if two unique gamete types (say A and B) are present, any of A, A , A, B or B, B mating events can occur. **pf-meiosis** models meiosis for all fourteen chromosomes of *P. falciparum*, which are represented as bit-arrays where each bit corresponds to a 1 kbp segment of DNA. Throughout this work, the values of bits are used to encode regions of IBD.

The number of recombination events per-chromosome is drawn from a zero-truncated Poisson with a rate of 13.5kbp/cM (Miles et al. 2016); zero-truncation enforcing at least one recombination event per bivalent, which is necessary for homologous pairing as described above. Recombination breakpoints are chosen uniformly from each chromosomes length, and the pairs of homologous chromosomes selected for crossing-over are selected at random. The result of this process is the conversion of the two parental haplotypes into four progeny haplotypes.

Dissection of mosquito vectors harbouring malaria have revealed the presence of multiple oocysts on the midgut wall, suggesting that meiosis can occur between multiple pairs of gametes in parallel within a single mosquito vector. This is modelled within **pf-meiosis**, where the number of oocysts can be explicitly set, or drawn from a distribution. Studies of wild-caught *Anopheles Gambiae* suggest that the distribution of oocysts is roughly geometric, with the majority of infected mosquitoes carrying only one oocyst (Beier et al. 1991; Collins et al. 1984). In Section 3.3, oocysts counts are drawn from a distribution calibrated to match Collins et al. (1984).

Finally, a mixed infection is created from a given vector through random sampling of the meiotic progeny generated from all oocysts. **pf-meiosis** simulates a mixed infection of K strains by drawing K progeny (without replacement, as replacement yields a $K - 1$ mixed infection) from the meiotic progeny simulated within the vector. The simulation is implemented in python.

3.2.2 Simulation validation

I first performed experiments to validate the behaviour of **pf-meiosis**. In studying relatedness, two informative summary statistics are the mean fraction of the genome found to be identical-by-descent between a pair of strains (hereafter IBD fraction, f), and the mean IBD segment length (l). Both of these statistics have expected relationships as a function of the number of meioses (m) that separate a parent-progeny pair, assuming that each intervening meiosis involves an unrelated individual. Namely, the mean IBD decays exponentially with the number of meioses, $E[f|m] = 2^{-m}$, whereas the mean IBD track length decreases according

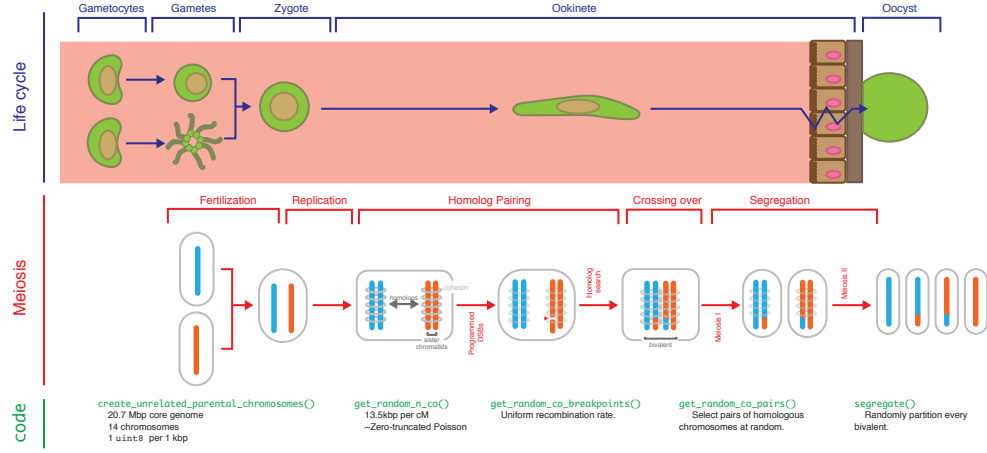


Figure 3.2: Schematic of **pf-meiosis** simulation based on malaria life cycle and meiosis. Top panel shows the life-cycle stages of the malaria parasite, starting upon entry to the mosquito midgut. Middle panel displays the meiotic steps corresponding with the top panel. Bottom panel displays some a selection of python functions that comprise **pf-meiosis**. Note that the whole process visualized here occurs multiple times in parallel for each oocyst that develops on the mosquito midgut.

to an inverse relationship, $E[l|m] = \rho/m$, where ρ is the recombination rate expressed as kbp/Morgan (S. R. Browning and B. L. Browning 2012)⁴. Using **pf-meiosis**, I simulated meiosis between two unrelated parental strains, and then simulated $m = [1, 2, \dots, 10]$ more rounds of meiosis with one of the progeny, separating it from the original parental pair by m meioses. This experiment was repeated 1000 times. Across these experiments, both the IBD fraction and average IBD segment length generated by **pf-meiosis** closely conformed to the expected relationship (Figure 3.3a).

As described above, the recombination rate in **pf-meiosis** is set to 13.5kbp per cM. This value was determined in laboratory cross experiments, where known parental strains were used to determine the number of recombination breakpoints across progeny after a single round of meiosis (Miles et al. 2016). I repeated this experiment *in silico*, using **pf-meiosis** to breed two unrelated parental strains. I then computed the number of observed recombination breakpoints in progeny based

⁴Note that the second expression accounts only for IBD tracks being broken by recombination and in practice needs to be adjusted for the number of chromosomes

on the haplotypes of the parental strains. In this way, I produced an empirical recombination rate distribution across 1000 replicate meioses. The mean of this distribution was 13.4kbp per cM, within 1% of the expected value, demonstrating that **pf-meiosis** generates a number of recombination events appropriate for the specified recombination rate.

3.2.3 Analysis of oocysts and IBD

The presence of multiple oocysts in a single mosquito vector, allowing for meiosis to occur several times in parallel between a single set of parental strains, is perhaps the most unique aspect of malaria meiosis. Following validation, I used **pf-meiosis** in combination with mathematical analysis to explore the relationship between oocysts and the expected IBD levels within mixed infections. I found that the amount of IBD observed in a mixed infection is a function of the number and type of oocysts present in the biting mosquito. This is demonstrated below.

First, let us briefly review the relevant fundamentals of malaria meiosis. In our case, we imagine a mosquito bites a human host containing two distinct malaria strains. Call these strains A and B . Some number of gametocytes of A and B are imbibed during the bite, differentiate into gametes, and undergo fertilization to produce zygotes (reviewed in Ghosh et al. 2000; Bennink et al. 2016). Some fraction of these zygotes succeed in establishing themselves as oocysts on the mosquito midgut (Ghosh et al. 2000). Three products of fertilization are possible, and thus the oocysts can be either: $A + A$ or $B + B$ (inbred oocysts, n_{ii}), or $A + B$ (outbred oocysts, n_{ij}). The oocyst state of a mosquito can be characterized by (n_{ij}, n_{ii}) . Which strain is maternal and paternal may vary from oocyst to oocyst, but this is of no consequence here.

A $K = 2$ mixed infection is established when two distinct sporozoites, produced from the oocysts of this mosquito, infect a host. Each oocyst produces thousands of sporozoites (Beier et al. 1991), of four types (discussed in McKenzie et al. 2001), which pool in the mosquito salivary glands (Ghosh et al. 2000). Imagine drawing a $K = 2$ mixed infection from a mosquito harbouring a single outbred oocyst ($n_{ii}=0$,

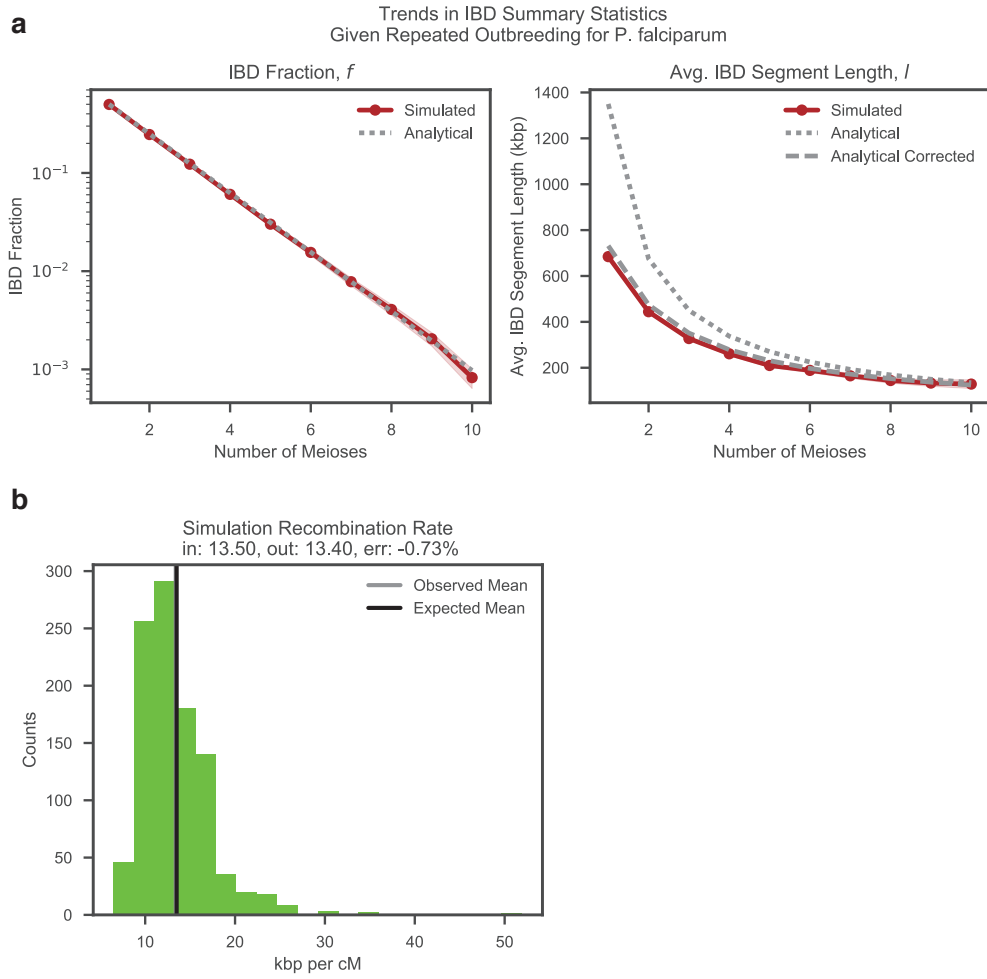


Figure 3.3: Validating the behaviour of **pf-meiosis**. (a) Behaviour of IBD statistics following repeated outbreeding. Two unrelated parental strains were crossed and then one of the progeny was repeatedly crossed with an unrelated strain for ten meioses. This offspring was compared to one of the parental strains to compute the fraction of the genome in IBD (left panel) and the average IBD segment length (right panel). Red line shows the mean result across 1000 replicate experiments, shading shows 95% confidence intervals. Dashed grey line gives analytical expectation. In the plot on the right, the expected average IBD segment length is also adjusted to account for *P. falciparum* having 14 chromosomes in a thicker dash. (b) Meiosis between two unrelated parents was performed 1000 times using **pf-meiosis**, with the recombination rate set to 13.50 kbp per cM. The number of recombination events in the progeny were calculated and used to determine the empirical recombination rate for each cross. The mean recombination rate from the crosses conforms closely to the expected rate.

$n_{ij} = 1$). Note that pairs are drawn without replacement because if sporozoites of only one type seed the infection, it will be $K = 1$. In such a mosquito there are two copies of each of the two strains (two sets of sister chromatids; A , A ,

B, B). Thus, ignoring recombination for the present, there are two pairs with an IBD fraction of 1 and, if our original strains are unrelated, the remainder of the $\binom{4}{2}$ pairs will have an IBD of 0. Thus a single n_{ij} oocyst has an expected IBD fraction of $E[f] = 2/\binom{4}{2} = 1/3$.

This somewhat surprising conclusion is consistent with the observations of (Wong et al. 2018). Importantly, neither recombination nor segregation affect the result, as they only shuffle how the total identity is distributed between pairs, rather than create or destroy identity (identity is created by DNA replication and destroyed by mutation); the expectation is taken over all pairs and is thus unaffected.

Computing the expected IBD fraction for a mosquito possessing n_{ij} outbred oocysts is an extension of the above. Again ignoring recombination, the expected IBD fraction $E[f|n_{ij}]$ is equal to the total number of pairs with an IBD of 1 (IBD pairs), over all possible pairs. In a mosquito with n_{ij} oocysts, we have $2n_{ij}$ copies of each parental strain, thus we have $\binom{2n_{ij}}{2}$ IBD pairs for each parental strain, thus $2\binom{2n_{ij}}{2}$ IBD pairs total. Dividing this by the total number of pairs amongst n_{ij} oocysts we have

$$E[f|n_{ij} > 0] = \frac{2\binom{2n_{ij}}{2}}{\binom{4n_{ij}}{2}} = \frac{2n_{ij} - 1}{4n_{ij} - 1}. \quad (3.3)$$

The above yields $1/3$ for $n_{ij} = 1$, approaching $1/2$ as n_{ij} grows. This result has been validated with **pf-meiosis** in Figure 3.4.

Including n_{ii} oocysts is somewhat involved, as some pairs (selected without replacement) may be identical (thus yielding $K = 1$) or completely unrelated (yielding $K = 2$, but effectively without having undergone meiosis or producing any detectable recombination breakpoints between parental strains). We are interested in the expected IBD produced as a result of meiosis between parental strains, and thus for the moment we exclude these pairs. In practice, this means the mosquito must have at least one outbred oocyst, and at least one of the infecting sporozoites must be from an outbred oocyst.

The derivation is as above: first ignoring recombination and segregation, then enumerating all IBD pairs (pairs with IBD fraction of 1) and dividing by the total

number of pairs to compute the expectation. Note that the additional IBD pairs possible between an outbred and inbred oocyst are given by the term $8n_{ij}n_{ii}$, and that you can no longer use all possible pairs drawn without replacement as the denominator, but must exclude the pairs described above.

$$E[f|n_{ij} > 0, n_{ii}] = \frac{2\binom{2n_{ij}}{2} + 8n_{ij}n_{ii}}{2\binom{2n_{ij}}{2} + 16n_{ij}n_{ii} + 4n_{ij}^2} = \frac{2(n_{ij} + 2n_{ii}) - 1}{4(n_{ij} + 2n_{ii}) - 1} \quad (3.4)$$

Which is of a similar form to above, but increases to $1/2$ quicker if more inbred oocysts are present. As before the equation is validated in Figure 3.5A.

The expression for $E[f|n_{ij} > 0, n_{ii}]$ can also be derived by recognizing that there are three *types* of pairs possible in a mosquito with a collection of n_{ij} and n_{ii} oocysts: (1) a pair can contain two strains from a single n_{ij} , ($n_{ij}^{o=1}$); (2) a pair can contain two strains from two different n_{ij} , ($n_{ij}^{o=2}$); or (3) a pair can contain one strain from an n_{ij} oocyst and one from an n_{ii} oocyst, $n_{ij,ii}^{o=2}$. Pair type (1) is unique to malaria and has an $E[f|n_{ij}^{o=1}] = 1/3$, as shown above; pair type (2) are standard siblings with $E[f|n_{ij}^{o=2}] = 1/2$; and pair type (3) represent a mother-daughter relationship, also with $E[f|n_{ij,ii}^{o=2}] = 1/2$. The full IBD fraction and IBD segment length distributions of these pairs were generated using `pf-meiosis` and can be seen in Figure 3.5B. We enumerate the number of each pair type given n_{ij} and n_{ii} , weighted by their expectation, to derive $E[f|n_{ij} > 0, n_{ii}]$:

$$\begin{aligned} E[f|n_{ij} > 0, n_{ii}] &= \frac{n_{ij}^{o=1} E[f|n_{ij}^{o=1}] + n_{ij}^{o=2} E[f|n_{ij}^{o=2}] + n_{ij,ii}^{o=2} E[f|n_{ij,ii}^{o=2}]}{n_{ij}^{o=1} + n_{ij}^{o=2} + n_{ij,ii}^{o=2}} \\ &= \frac{n_{ij} \binom{4}{2} 1/3 + 16 \binom{n_{ij}}{2} 1/2 + 16 n_{ij} n_{ii} 1/2}{n_{ij} \binom{4}{2} + 16 \binom{n_{ij}}{2} + 16 n_{ij} n_{ii}} \\ &= \frac{2(n_{ij} + 2n_{ii}) - 1}{4(n_{ij} + 2n_{ii}) - 1} \end{aligned} \quad (3.5)$$

As above.

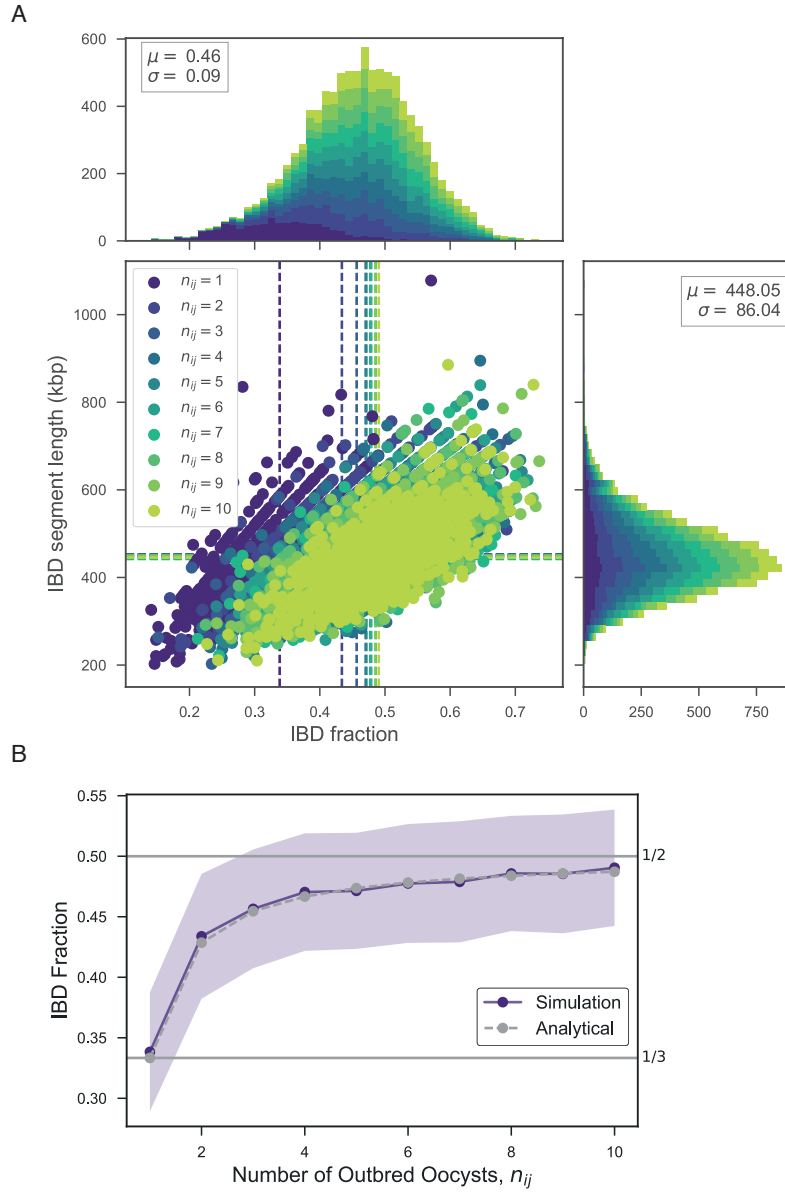


Figure 3.4: Exploring the relationship between number of outbred oocysts (n_{ij}) and IBD. (A) Joint IBD fraction and IBD segment length distributions for $k = 2$ mixed infections simulated from two unrelated strains and a fixed number of outbred oocysts n_{ij} , using `pf-meiosis`. Mean values for each distribution are indicated by same-color dashed lines. Each distribution is created from 1000 simulated mixed infections. (B) Validation of theoretical result given by equation 3.3. Line plot compares trend in expected IBD fraction with the number of outbred oocysts, n_{ij} , for infections simulated in panel A, and analytical expression 3.3. Shading gives ± 1.96 times the standard error of the mean

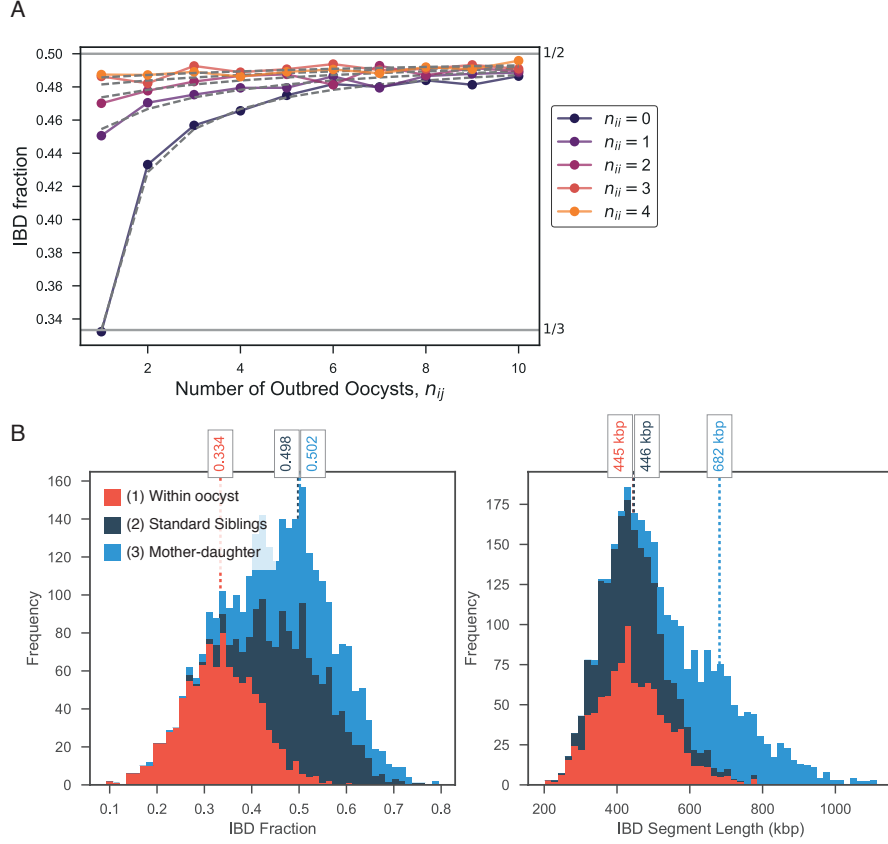


Figure 3.5: Exploring expected IBD allowing for outbred (n_{ij}) and inbred (n_{ii}) oocysts (A) Validation of expression for expected IBD fraction conditional on outbred n_{ij} and inbred n_{ii} oocysts. Line plot compares trend in expected IBD fraction with varying number of outbred (x-axis, n_{ij}) and inbred (line color, n_{ii}) oocysts and the analytical expression 3.4 (grey dashed lines). (B) Using `pf-meiosis` to simulate $K = 2$ mixed infections generated from (1) two strains from the same outbred oocyst from ($n_{ij}^{o=1}$, 'Within oocyst'); (2) two strains different outbred oocysts ($n_{ij}^{o=2}$, 'Standard Siblings'); (3) one strain from an outbred and one strain from an inbred oocyst ($n_{ij,ii}^{o=2}$, 'Mother-daughter').

3.3 IBD and the classes of mixed infections

3.3.1 Simulating mixed infection chains with `pf-meiosis`

In the previous section we learned that a $K = 2$ mixed infection generated from a single infection event will contain sibling strains, with an IBD fraction ranging between $1/3$ and $1/2$, depending on the number of oocysts in the mosquito vector. However, after deconvolution and IBD estimation with `DEploidIBD`, we observed that many of the mixed infections across the Pf3k project had IBD fractions significantly above $1/2$ (Figure 3.6). What could explain these higher levels

of relatedness?

A first possibility is low levels of genetic diversity resulting in IBD between different strains within a population. An IBD fraction in the range of $1/3$ to $1/2$ is predicated on meiosis occurring between parental strains sharing no IBD. However, there may be substantial IBD present between strains in separate infections, which augments the expected IBD in a mixed infection (generated either by super-infection or co-infection). Mathematically, the expected IBD fraction following meiosis between two parental strains, given the presence of background IBD (f_b) between those strains is:

$$E[f|f_b > 0] = f_b + (1 - f_b)E[f|f_b = 0] \quad (3.6)$$

Consistently, studies have found that malaria from countries in Asia have significantly lower genetic diversity than those in Africa (Anderson et al. 2000; Manske et al. 2012). This would be consistent with estimates of higher IBD observed within mixed infections produced by DEploidIBD (Figure 3.6).

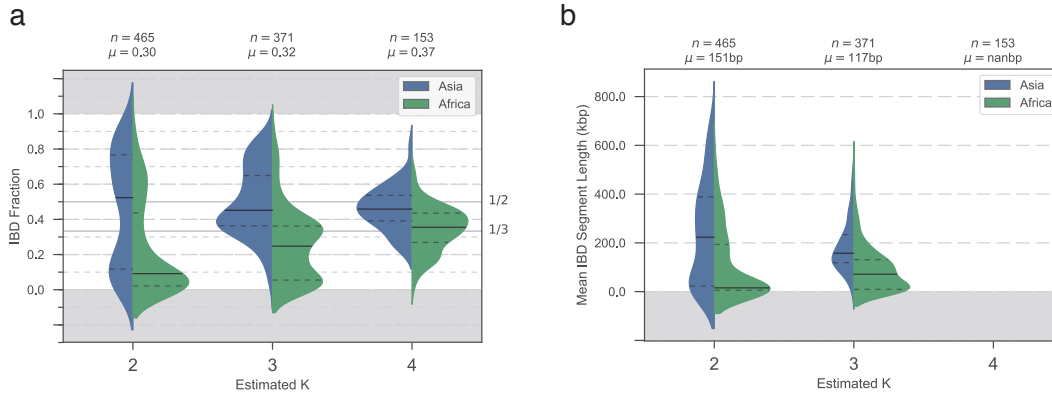


Figure 3.6: Distribution of IBD statistics in Pf3k estimated by DEploidIBD. (a) Distribution of IBD fraction across the Pf3k project as partitioned by K and continent-of-origin and represented as a violinplot. For each K , the mean (μ) and number of infections (n) are indicated above. Note grey shading of background indicates regions outside of IBD fraction parameter range; the violinplot fits a kernel density causing the distributions to spill outside. (b) Same as in (a) but for Mean IBD segment length. IBD tract lengths were not computed for mixed infections of $K = 4$.

A second possibility is what we here call "mixed infection chains". For this discussion we will limit our considerations to $K = 2$ mixed infection chains, though

higher K chains are possible. Consider that there are at least three outcomes for a $K = 2$ mixed infection generated by co-infection: (1) it may be cleared before being transmitted (either by time or treatment), (2) one of the $K = 2$ strains may be transmitted to another host, or (3) both of the $K = 2$ strains may be transmitted to another host. If both strains are transmitted to another host via a single mosquito vector, they will undergo a second round of meiosis producing progeny with a higher level of IBD as a consequence of inbreeding (Equation 3.6 applies). If this process is repeated, it produces a chain of serially co-transmitted mixed infections with an IBD fraction asymptotically increasing towards unity.

I explored the distribution of IBD statistics expected in such mixed infection chains using `pf-meiosis` (Figure 3.7). As expected, the mean IBD fraction within a mixed infection climbs from 41% after a single meiosis, up to 93% after five rounds of co-transmission and meiosis (Figure 3.7b). Segment length similarly increases. These results establish that mixed infection chains are a possible explanation for highly related ($f > 1/2$) mixed infections.

3.3.2 Classification of mixed infections across Africa and Asia

We hypothesized that both the mechanisms discussed above – background IBD and mixed infection chains – were making contributions to the striking levels of IBD observed in Figure 3.6. To explore this further, we sought to leverage *pf-meiosis* to classify all $K = 2$ mixed infections based on their patterns of IBD. In particular, we sought to classify each $K = 2$ mixed infection as being generated either by super-infection or co-infection, and if by co-infection, either as being the product of a single co-infection event or a mixed infection chain.

We used two summary statistics to perform the classification: mean IBD segment length and IBD fraction. We built empirical distributions for these two statistics for each country in Pf3k, by using `pf-meiosis` to simulate meiosis between pairs of clonal samples from that country. In this way, we effectively control for variation in the levels of background IBD by using clonal samples from each country. Starting

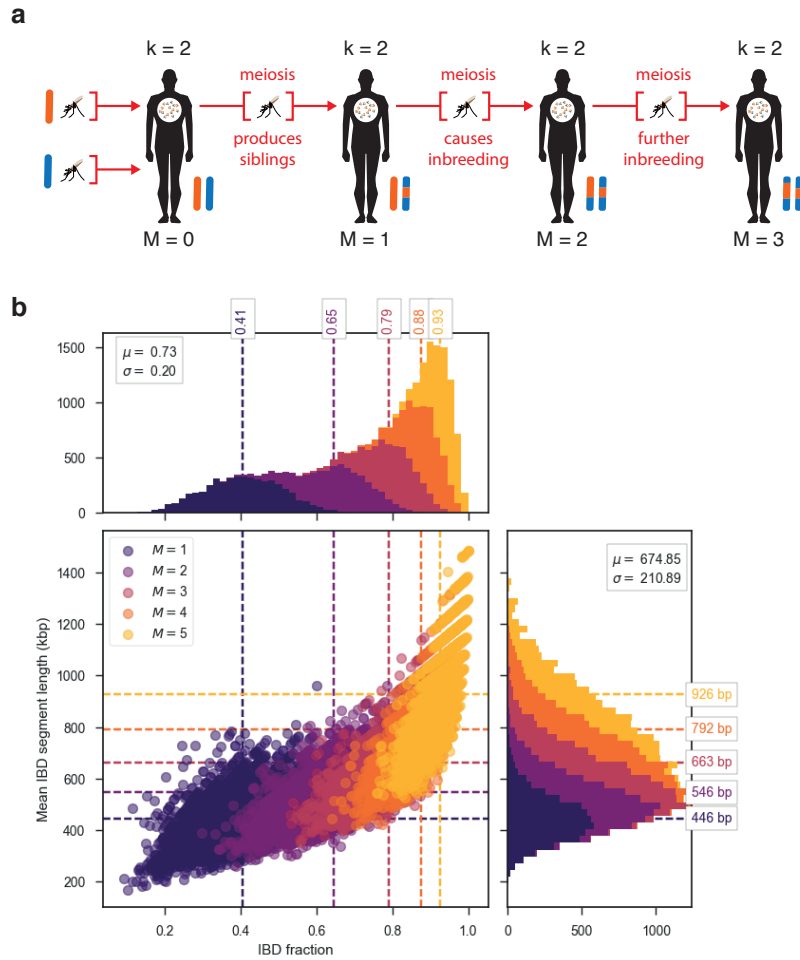


Figure 3.7: Serial co-transmission of $K = 2$ mixed infections simulated with **pf-meiosis**. (a) Schematic of serial co-transmission producing a mixed infection chain. At left, the first host receives two distinct strains via two independent mosquito bites, producing a $K = 2$ mixed infection by super-infection (labelled $M = 0$, where M indicates the number of meiosis that have occurred). Moving rightwards, the mixed infection is propagated by serial co-transmission and the IBD within each consecutive host climbs. (b) The process in (a) was simulated with **pf-meiosis** for mixed infection chains of length $M = 5$ a total of 10,000 times. The parental strains were unrelated resulting in a $M = 0$ class with no IBD, and so this class is not shown. At each M , the IBD fraction and mean IBD segment length within each mixed infection was calculated and is shown as a point on the scatterplot. Coloured horizontal and vertical dashed lines show the means of their respective M distribution. Marginal distribution legends indicate mean and variance for the IBD fraction and Mean IBD segment length statistics across all M .

from a pair of clonal samples ($M = 0$, where M indicates the number of meioses that have occurred), we simulated three successive rounds of meiosis ($M = 1, 2, 3$), representing the creation (by super-infection) and serial co-transmission of a mixed

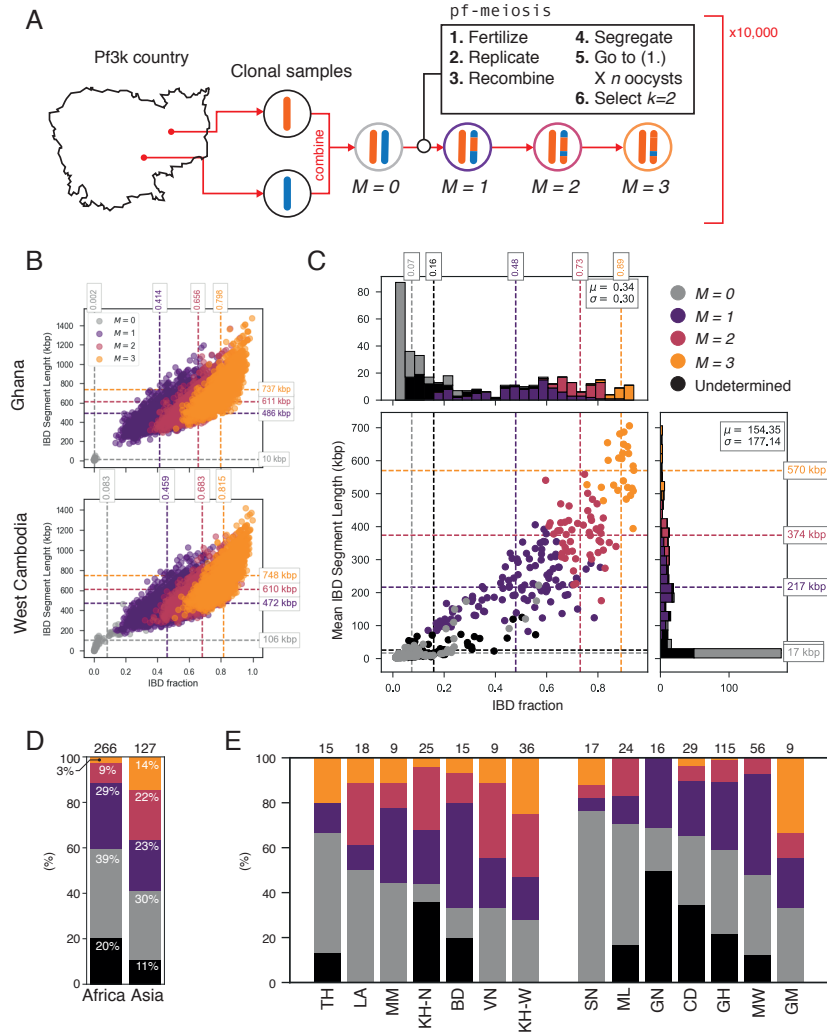


Figure 3.8: Identifying sibling strains within mixed infections. (A) Schematic showing how IBD fraction and IBD segment length distributions are created for $K = 2$ mixed infections using *pf-meiosis*. Two clonal samples from a given country are combined to create an unrelated ($M = 0$, where M is number of meioses that have occurred) mixed infection. The $M = 0$ infection is then passed through 3 rounds of *pf-meiosis* to generate $M = 1, 2, 3$ classes. (B) Simulated IBD distributions for $M = 0, 1, 2, 3$ for Ghana (top) and West Cambodia (bottom). A total of 10,000 mixed infections are simulated for each class, from 500 random pairs of clonal samples. (C) Classification results for 393 $K = 2$ mixed infections from 13 countries. Undetermined indicates mixed infections with IBD statistics that were never observed in simulation. (D) Breakdown of class percentage by continent. Total number of samples is given above bars. Colours as in panel C ($M = 0$, grey; $M = 1$, purple; $M = 2$, pink; $M = 3$, orange; Undetermined, black). (E) Same as (D), but by country. Abbreviations: SN-Senegal, GM-The Gambia, NG-Nigeria, GN-Guinea, CD-The Democratic Republic of Congo, ML-Mali, GH-Ghana, MW-Malawi, MM-Myanmar, TH-Thailand, VN-Vietnam, KH-Cambodia, LA-Laos, BD-Bangladesh.

infection (Figure 3.8A). The result is a mixed infection chain of length four. As in the previous section, each round of meiosis increases the amount of observed IBD. For example, in Ghana, the mean IBD fraction for $M = 0$ was 0.002, for $M = 1$ was 0.41, for $M = 2$ was 0.66, and for $M = 3$ was 0.80 (Figure 3.8B). West Cambodia, which has lower genetic diversity, had a mean IBD fraction of 0.08 for $M = 0$ and consequently, the mean IBD fractions for higher values of M were slightly increased, to 0.46, 0.68, 0.81 for $M = 1, 2$ and 3, respectively (Figure 3.8B).

With these simulated distributions, we used Naive Bayes to classify $K = 2$ mixed infections in Pf3k (Figure 3.8C). Of the 393 $K = 2$ samples containing only high-quality haplotypes (see Appendix), 325 (83%) had IBD statistics that fell within the range observed across all simulated M . Of these, nearly half (183, 47%) were classified as siblings ($M > 0$), suggesting the creation of mixed infections by co-infection is routine. Moreover, we observed strong evidence for the existence of mixed infection chains, particularly in Asia, where 38% of infections were classified as $M > 1$. Finally, clear geographical differences in the rate at which sibling and unrelated mixed infections occur were present. Notably, in Asia a greater fraction of all mixed infections contained siblings (59% vs. 41% in Africa), driven by a higher frequency of $M = 2$ and $M = 3$ mixed infections (Figure 3.8D).

3.4 Mixed infection classification and $PfPR_{2-10}$

To assess how characteristics of mixed infections relate to local transmission intensity, we obtained estimates of *P. falciparum* prevalence (standardised as $PfPR_{2-10}$, prevalence in the 2-to-10 year age range) from the Malaria Atlas Project MAP 2017. The country-level prevalence estimates range from 0.01% in Thailand to 55% in Ghana, with African countries having up to two orders of magnitude greater values than Asian ones (mean of 36% in Africa and 0.6% in Asia). However, seasonal and geographic fluctuations in prevalence mean that, conditional on sampling an individual with malaria, local prevalence may be much higher than the longer-term (and more geographically widespread) country-level average, hence we extracted the individual pixel-level estimate of prevalence (corresponding to a 5x5km region)

from MAP nearest to each genome collection point. We summarise mixed infection rates by the average effective number of strains ($K_{eff} = 1/\sum w_i^2$), which reflects both the number and proportion of strains present.

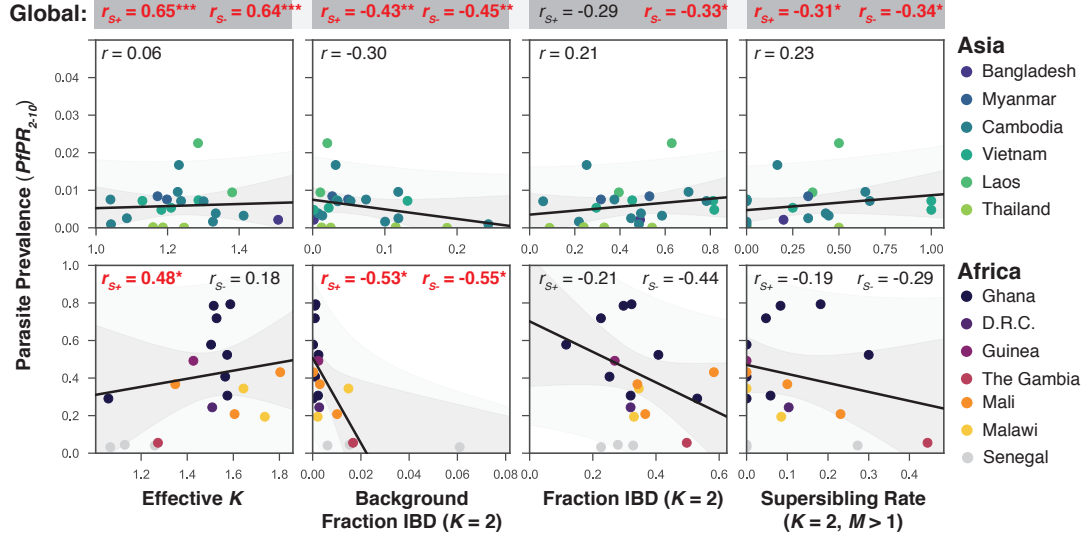


Figure 3.9: The relationship between *P. falciparum* prevalence and characteristics of mixed infection. Four mixed infection statistics are shown including the average effective number of strains (Effective k, first column), given by $k_e = (\sum w_i^2)^{-1}$, where w_i is the proportion of the i th strain; background IBD observed between clonal samples (Background Fraction IBD, second column); fraction IBD within $K = 2$ mixed infections (Fraction IBD, third column); and the rate of $K = 2$ mixed infections classified as having $M > 1$ (Supersibling Rate, fourth column). Each point relates to a row in Table 1 from different sampling locations and years. Pearson's r is computed globally (shown at top in a grey box for each statistic), across Asian countries (upper panel) and across African countries (lower panel). Globally and for Africa, the correlations were computed including Senegal (r_{S+}) and excluding Senegal (r_{S-}). The slope and confidence intervals for the regression line excluding Senegal are drawn. Significant correlations ($p < 0.05$) are highlighted in red and significance levels indicated by asterisks (* < 0.05 , ** < 0.01 , *** < 0.001)

Given that samples from Senegal were screened to be primarily single-genome (Daniels et al. (2015)), we computed all correlations with prevalence including (r_{S+}) and excluding them (r_{S-} ; Figure 3.9). We find that the effective number of strains is a significant predictor of $PfPR_{2-10}$ globally ($r_{S+} = 0.65, p < 10^{-5}$) and in African populations when Senegal is included ($r_{S+} = 0.48, p = 0.04$; $r_{S-} = 0.18, p = 0.51$), but is uncorrelated across Asia. Similarly, $PfPR_{2-10}$ is negatively correlated with background IBD globally ($r_{S+} = -0.43, p = 0.004$) and across Africa but

not in Asia. Surprisingly, the amount of IBD observed within $K = 2$ mixed infections was not correlated with prevalence in Africa or Asia. Similarly, the rate at which sibling infections occur ($M = 1$) was also not correlated with the parasite prevalence (Asia: $r_{S+} = 0.23, p = 0.2$, Africa: $r_{S+} = 0.16, p = 0.5$). However, the rate of supersiblings ($K = 2, M > 1$) is significantly correlated with $PfPR_{2-10}$ ($r_{S+} = -0.31, p = 0.04$) at the global scale, suggesting that serial co-transmission may occur more readily in low prevalence regions.

3.5 Conclusion

The past ten years have seen the widespread application of molecular barcoding and whole-genome sequencing to malaria, greatly expanding the amount of genetic data available for analysis. To realize the full value of this data, much of which is convoluted in mixed infections, methods for inferring the number of strains, their proportions and haplotypes have been developed. Yet, until the development of **DEploidIBD**, all of these methods made the simplifying assumption that the strains within a mixed infection were unrelated. Though this may be true in the case of super-infection, it is false for co-infection. By estimating the relatedness between strains in a mixed infection, **DEploidIBD** opened the door on a suite of new questions about how this relatedness is linked to malaria meiosis, infection mechanisms, and regional prevalence.

A key finding was that there is a relationship between the expected IBD in a $K = 2$ mixed infection and the number and type of oocysts in the mosquito vector. Uniquely, we have shown that a $K = 2$ mixed infection produced from a mosquito harbouring a single oocyst is expected to produce an infection with an IBD fraction of $1/3$. This observation, however, raises the question of how often mixed infections are created from single-oocyst mosquito vectors. Though there are studies demonstrating geometric distributions of oocysts (Beier et al. 1991; Collins et al. 1984), whether those oocysts are inbred or outbred, and how the distribution of oocysts relates (if at all) to the probability of mixed infection is unknown. It might be that more detailed classification of $K = 2$ mixed infections, based on the most likely

oocyst configuration in the infecting mosquito, is possible. The fact that the joint distribution of IBD fraction of IBD segment length is unique (though significantly overlapping) for different relationships between strains within a mixed infection (Figure 3.5B) is promising, though this has not been explored further. At present, it should be acknowledged that simply drawing oocysts from a geometric distribution is equivalent to assuming that all oocyst states are equal likely to produce a mixed infection. Investigating the validity of this assumption has been left to future work.

Exploring whether the IBD within a mixed infection could inform on how that mixed infection was created – through, for example, co-infection or super-infection – was one of the primary motivations of this work. Classifying all of the $K = 2$ mixed infections in the Pf3k project revealed that nearly 50% of them were generated via co-infection ($M > 0$). A high rate of co-infection has been previously suggested by Nkhoma et al. (2012), where individual haplotypes isolated from multiple infections by dilution cloning were found to share high levels of IBD, in both Malawi and Thailand. Firstly, this finding is relevant for the epidemiological modelling of malaria. In particular, assumptions made in 2.2 about the relative ease with which super-infections are established may need reconsidering. Secondly, it suggests that the bottleneck of transmission is relatively large; that is, a relatively large number of parasites are transmitted to a host. Insight into the size of the transmission bottleneck has been difficult to obtain by experimentation. The most direct studies have been able to quantify the number of sporozoites delivered during a biting event, e.g. by feeding infected vectors the ear of mice models and then quantifying sporozoites antigens (Medica and Sinnis 2005; Jin et al. 2007; Yamauchi et al. 2007) . Beyond questions of the suitability of this experimental model, there is the fact that it fails to determine the number of sporozoites that reach the liver, or establish themselves at an appreciable frequency in the blood. It is interesting to note that, using only genetic data, one can put a lower bound on the bottleneck size. In particular, the frequency at which more than one genetically distinct strain passes through the transmission bottleneck will be greater than the fraction of

all samples (clonal and mixed) that were generated by co-infection; across the Pf3k project, this fraction is ~ 0.12 .

It is worth noting that a non-trivial fraction (17%) of all mixed infections had patterns of IBD inconsistent with the simulations (typically with slightly higher IBD levels than background but lower than among siblings). In *Zhu et al.*, we proposed three possible explanations (Zhu et al. 2019). A first is that the unclassified samples result from the IBD profiles produced by DEploidIBD, in particular the overestimation of short IBD tracks, similar to the issue observed by Wong et al. 2018. Alternatively, our estimate of background IBD, generated by combining pairs of random clonal samples from a given country into an artificial $M = 0$ mixed infection, will underestimate true background IBD if there is very strong local population structure. Finally, we only simulated simple mixed infection transmission chains, at the exclusion of more complex transmission histories, such as those involving strains related at the level of cousins. The extent to which such complex histories can be inferred with certainty remains to be explored.

An important step in ensuring infection mechanism inferences were correct was to control for background IBD levels in different populations. With DEploidIBD we found, as expected, that background levels of IBD were higher in Asia than in Africa (Figure 3.6). Yet, even after accounting for this, there was still compelling evidence for a greater frequency of co-infection in Asia. Moreover, much of this co-infection was the result of mixed infection chains. Consistent with this, Nkhoma et al. (2012) found higher rates of "extreme relatedness within infections" in Thailand (21.1% of mixed infections) than in Malawi (15.2% of mixed infections). *Prima facie*, this could be explained by the lower prevalence of malaria in Asia. Consider that, the propagation of a mixed infection by co-infection requires (i) a host harbouring a $K = 2$ mixed infection is bitten by an *uninfected* vector and (ii) that vector subsequently bite an *uninfected* host. Such conditions are more likely in low transmission intensity regions. However, comparing mixed infection classifications to prevalence data extracted from the Malaria Atlas Project revealed only relatively weak and noisy correlations. For example, the rate of mixed infection chains

($M > 1$) was negatively correlated with prevalence globally ($r = -0.31$) and across Africa ($r = -0.19$) but had a slightly positive correlation in Asia ($r = 0.23$). Several factors could explain these results. Noting that prevalence is very low in Asia, both estimation uncertainty in prevalence and/or lack of sensitivity in the relationship between co-infection and prevalence are possible. Alternatively, sample collection across the Pf3k is opportunistic, which may be introducing sampling bias (e.g. passive vs. active collection) weakening relationships. Finally, summarizing prevalence around a given genome collection point requires judging over how large a region one should average: here, I simply took the nearest 5x5km pixel within the MAP prevalence raster. A more nuanced exploration of the spatial co-variation between prevalence and IBD statistics would likely improve results.

There is undoubtedly more to be uncovered about the relationships between mixed infections, IBD, and transmission intensity. For example, a recently published single-cell genomics study found extensive and complex patterns of IBD, indicative of co-infection, in a high-transmission region of Malawi (Nkhoma et al. 2020). This would argue against a simple inverse relationship between transmission intensity and the rate of co-infection. Further, Nkhoma et al. (2020) found a large number of higher order mixed infections ($K > 2$), and it is clear that further exploration of IBD patterns and their relationship to co-infection are needed in this context. However, as we've seen in this chapter, the tools are now in place to facilitate this exploration, and we can hope to learn both intimate details about malaria biology – the number of oocysts in mosquito vectors, or the size of the transmission bottleneck – as well as mechanisms of infection on the level of individual hosts. Whether these insights can be leveraged to learn about population-level epidemiology and, more importantly, to do so in a manner that could benefit malaria control efforts, largely remains to be seen.

4

Collecting *P. falciparum* genetic data from a mobile laboratory in Sub-Saharan Africa

Contents

4.1	Introduction	94
4.1.1	The evolution of antimalarial drug resistance	94
4.1.2	Monitoring for antimalarial drug resistance in Africa . .	100
4.1.3	The Land Rover Bursary	102
4.2	Developing a mobile sequencing laboratory	104
4.2.1	Laboratory Protocols	104
4.2.2	Computational Pipeline	113
4.2.3	Mobility, costs, and shipping	115
4.3	Deploying the laboratory in Sub-Saharan Africa	118
4.3.1	Sequencing metrics	120
4.3.2	Error rate analysis	122
4.3.3	Calling known antimalarial drug resistance mutations .	125
4.3.4	Downsampling analysis	127
4.4	Discussion	129
4.4.1	Why is artemisinin resistance not already in Africa? . .	133
4.5	Ethical Approval	135

4.1 Introduction

In previous chapters, we have seen how information about the regional prevalence of *P.falciparum* malaria is present within genetic data. While this information is valuable to malaria control programs, obtaining it requires statistical inference, and well-validated methods to infer prevalence from genetic data are not yet available. In recent decades, however, the causal mutations underlying many forms of antimalarial drug resistance – including to chloroquine, sulfadoxine, pyrimethamine, mefloquine, artemisinin and piperazine – have all been identified (Fidock et al. 2000; Brooks et al. 1994; Plowe et al. 1997; Price et al. 2004; Ariey et al. 2014; Witkowski et al. 2017). Information about the distribution of these forms of antimalarial drug resistance is highly valuable to malaria control programs, and genetic data can now directly provide this information. However, in order to be useful for malaria control, genetic data must be collected at an appropriate scale, frequency and cost. In this chapter, we explore one approach by which genetic data can be collected to monitor for the emergence and spread of antimalarial drug resistance.

4.1.1 The evolution of antimalarial drug resistance

The marked decline in malaria burden observed during the first fifteen years of the twenty-first century was largely effected by only three interventions: insecticide-treated bednets (ITNs), indoor residual spraying (IRS), and artemisinin combination therapy (ACT) (Bhatt et al. 2015). Although these tools have been highly effective, their continued efficacy is threatened by the emergence of resistance-conferring mutations; both ITNs and IRS to insecticide resistance mutations in *Anopheles* vectors, and ACTs to antimalarial drug resistance mutations in *Plasmodium falciparum*. Other technologies supporting malaria control – such as vaccination and gene-drive – are being actively developed and piloted, but they are not yet ready for widespread operational use (Coelho et al. 2017; James et al. 2018). Thus, a deterioration of the efficacy of ITNs, IRS, and ACT by resistance mutations would considerably diminish our ability to control malaria.

ACT contains a co-formulation of an artemisinin derivative and a partner drug, typically piperaquine, mefloquine, or lumefantrine (Bloland and Organization 2001). Artemisinin was discovered in 1972 by Tu Youyou, who, searching for antimalarial drugs amongst traditional Chinese medicines, isolated the compound from the shrub *Artemisia annua* (Tu 2011). For the next several decades, artesunate monotherapy was used to dramatically reduce malaria in China: from an estimated 2.3 million cases per annum in 1979 (Zhou 1981) to approximately 120 thousand cases per annum by 1990 (Tang et al. 1991). However, it was not until 2001 that ACT became the World Health Organization (WHO) recommended first-line treatment against uncomplicated malaria, and not until 2004 that it was widely used (WHO 2006). By that time, chloroquine, the former first-line treatment, was failing globally as a consequence of resistance in *P. falciparum*, and malaria deaths had been climbing steadily since the 1980s (Murray et al. 2012).

The adoption of ACT – along with increased funding following the Millenium Development Goals – aligned with a reduction in malaria deaths both in Africa and Asia (Murray et al. 2012; Bhatt et al. 2015). However, in 2008 and 2009, two studies reported the emergence of *P. falciparum* with partial resistance to artemisinin, both along the Thai-Cambodian border (Noedl et al. 2008; Dondorp et al. 2009). In particular, they observed a prolongation of parasite clearance times in a fraction of cases following administration of artesunate monotherapy. Artemisinin is a potent and fast-acting drug with a very short half-life (0.5-1.4hrs), eliminating the majority of parasites within a few days (Bloland and Organization 2001). In ACT, the partner drug is weaker and longer-lasting, designed to eliminate residual parasites over the following weeks (Bloland and Organization 2001). The 2009 study found that after treatment with artesunate, parasites from Wang Pha, in northern Thailand, had a median parasite clearance time of 48hrs (Dondorp et al. 2009). Those from Pailin, in western Cambodia, had a median clearance time of 84hrs (Dondorp et al. 2009). No demographic or pharmacokinetic factors could explain the discrepancy (Dondorp et al. 2009).

Over the next several years prolonged clearance times following treatment with artesunate monotherapy and ACT, originally confined to western Cambodia, were documented elsewhere. On the northwestern border of Thailand, a ten-year study involving 3202 patients showed a gradual increase in the percentage of patients with a prolonged parasite clearance half-life ($PC_{1/2}$)¹, from 0.6% in 2001 to 20% in 2010 (Phyo et al. 2012). Similar findings were reported in Vietnam and Myanmar (Ashley et al. 2014). Genotyping data suggested that increasing parasite clearance times had a genetic contribution. For example, the heritability of $PC_{1/2}$ rose from 32% in 2001-2004 to 58% in 2007-2010 (Phyo et al. 2012). Yet, the alleles responsible were unknown.

The genetic basis of artemisinin resistance

In a landmark 2014 paper, Ariey et al. (2014) discovered the underlying mutations causing prolonged parasite clearance. To select for these mutations, Ariey et al. (2014) cultured an artemisinin sensitive *P. falciparum* strain from Tanzania under a dose-escalating regimen of artemisinin for five years. Performing whole-genome sequencing (WGS) on the resultant strain, and comparing with WGS data from the same strain cultured in the absence of artemisinin, they identified eight novel mutations that arose specifically under artemisinin exposure. The first mutation to arise during the five years of culturing was in the propeller domain of the gene *Kelch13*. Moreover, amongst a set of 49 culture adapted strains collected from Cambodia in 2010 and 2011, only those harbouring *Kelch13* propeller-domain mutations exhibited increased survival in an *in vitro* ring-stage parasite survival assay (RSA): *Kelch13* wild-type parasites had a median RSA of 0.17%, whereas *Kelch13* mutant parasites has an RSA 18.8%.

Interestingly, four distinct non-synonymous mutations in *Kelch13* exhibited an increased RSA, C580Y, R539T, Y493H and I543T (Ariey et al. 2014). Looking across a larger set of 886 archived patient samples from Cambodia, 17 distinct

¹Defined as a parasite clearance half-life of greater than 6.2hrs. The authors describe the threshold as "stringent", as 119 patients from western Cambodia had a geometric mean parasite clearance half-life of 5.5hrs (Phyo et al. 2012).

non-synonymous mutations were found, with C580Y being at by-far the highest frequency (85% of all mutant alleles in 2011-2012). Furthermore, the authors found that as early as 2001, C580Y comprised over 40% of all *Kelch13* alleles in Pailin, where resistance was first identified (Ariey et al. 2014). Finally, by examining the *Kelch13* status of 163 patients from western Cambodia with varying $PC_{1/2}$ times, Ariey et al. (2014) associated three *Kelch13* propeller-domain mutations (C580Y, R539T, and Y439H) with increased parasite clearance times *in vivo*. Those patients with wild-type *Kelch13* had a median $PC_{1/2}$ of 3.3hrs; those with Y493H, 6.28hrs; those with R539T, 6.64hrs; and those with C580Y, 7.19hrs.

That same year, Ashley et al. (2014) examined the complete sequence of *Kelch13* for 1241 cases of uncomplicated *P. falciparum* malaria across Southeast Asia, documenting the presence of propeller domain mutations in Thailand, Laos, Vietnam, Myanmar, and Cambodia. Of the 144 African cases included in the study (from Nigeria and the Democratic Republic of the Congo), only three had non-synonymous *Kelch13*-propeller domain mutations, and none of those showed a prolonged $PC_{1/2}$ (Ashley et al. 2014). Importantly, the authors investigated the utility of *Kelch13* as a molecular marker, finding propeller-domain *Kelch13* mutations had a 91.8% sensitivity and 88.4% specificity in predicting a $PC_{1/2}$ greater than 5hrs (Ashley et al. 2014). The following year, Miotto et al. (2015) conducted a *P.falciparum* genome-wide association study (GWAS) across 928 patients from Southeast Asia, identifying non-synonymous mutations in five genes associated with an increased $PC_{1/2}$. The most significant association genome-wide was the C580Y mutation in *Kelch13* (Miotto et al. 2015). In addition, 20/25 other non-synonymous mutations discovered in the *Kelch13*-propeller and adjacent BTB/POZ domain were shown to increase parasite clearance time *in vivo* (Miotto et al. 2015).

Though many mutations in the propeller and BTB/POZ domain of *Kelch13* have been demonstrated to confer resistance to artemisinin in the form of an increased $PC_{1/2}$ in malaria patients, the only mutation at high frequency in nature is C580Y. One question is whether the existence of the C580Y mutation in multiple geographical locales has resulted from migration or independent emergence – i.e. is

the result of a soft- or hard-selective sweep. With samples from 2009-2013, Miotto et al. (2015) examined the genomic region flanking C580Y mutations and found a handful of distinct SNP combinations, often delineating geographical locations, suggesting the C580Y mutation emerged independently on different local genetic backgrounds. However, the following year this interpretation was complicated: by genotyping microsatellites flanking *Kelch13* in patient samples, Imwong et al. (2017) found that in 2014-2015 a single long C580Y haplotype (-50 to 31.5kbp) predominated in Pailin (14/14 patients), Thailand (65/88 patients), and southern Laos (86/111 patients). The authors interpreted this as a trans-national migration event originating from Pailin, and further suggested that parasite populations within Southeast Asia were undergoing a transition from a soft- to hard-selective sweep, with a single C580Y haplotype outcompeting other *Kelch13* alleles (Imwong et al. 2017).

ACT treatment failure

By including two antimalarial drugs with different mechanisms, the ACT co-formulation was designed to prevent the emergence of resistance (WHO 2010). Yet, widespread resistance to one drug effectively reduces a combination therapy to a monotherapy. In Cambodia, ACT was first introduced in 2000, as a combination of artemisinin and mefloquine. At the time, mefloquine monotherapy was already exhibiting failure rates of 7-30% (Denis et al. 2006); it was discovered in 2004 that this was driven by amplification of the gene *mdr1* (Price et al. 2004). The Cambodian National Centre for Parasitology, Entomology and Malaria Control (CNM) reported an overall treatment efficacy of artesunate mefloquine combination therapy of 95.8% across 14 sites from 2001-2004 (Denis et al. 2006). In retrospect, however, we can see that late treatment failure rates² in Pailin were already at 9.9% in 2004, by which time we now know the frequency of C580Y mutation exceeded 50% (Denis et al. 2006; Ariey et al. 2014). These results have been taken to suggest that some fraction of patients with prolonged $PC_{1/2}$ as a consequence

²Measured as patients with detectable parasitemia 28- or 42-days after treatment, with reinfection excluded by PCR of surface antigen genes.

of *Kelch13* mutations ultimately experience treatment failure – perhaps in cases where parasites also harbour amplifications of *mdr1*.

Between 2006 and 2008, a study of 150 patients recruited from southern Cambodia reported a late treatment failure rate of artesunate mefloquine of 18.8% (Rogers et al. 2009). In response to the increasing failure rates of artesunate mefloquine, the CNM switched to dihydroartemisinin piperaquine in 2008 (Leang et al. 2013). However, by 2013 it was clear that that dihydroartemisinin piperaquine was also failing: 16% (10/63) of patients in Preah Vihear and 46% (37/81) of patients in Pursat experienced treatment failure (Amaratunga et al. 2016). In 2016 it was determined that piperaquine resistance resulted from amplification of the *plasmepsin2-plasmepsin3* gene cluster, and moreover that *plasmepsin2-3* amplification predicted dihydroartemisinin piperaquine treatment failure with a 94% sensitivity and 77% specificity (Witkowski et al. 2017). A subset of the samples carrying the long *Kelch13* C580Y haplotype described above, and purportedly dispersing across Southeast Asia from a birthplace in western Cambodia, were found to also carry a *plasmepsin2-3* amplification (Imwong et al. 2017). Interestingly, a retrospective study looking at 464 patient samples collected between 2007-2013 found that an increasing *plasmepsin2-3* amplification frequency was associated with a declining frequency of *mdr1* amplifications, and that *plasmepsin2-3* and *mdr1* amplifications are rarely found within the same sample; suggestive of a negative genetic interaction (Amato et al. 2018). In 2016, the CNM switched back to artesunate mefloquine to try and militate against the further spread of dihydroartemisinin piperaquine resistance (Hamilton et al. 2019).

A recent summary of the state of artemisinin resistance was provided in "Status report on artemisinin resistance and ACT efficacy", published by WHO in 2018 (WHO 2018a). They report that of the five ACT currently recommended, four are experiencing high failure rates (defined as >10%) in Cambodia, three experiencing high failure rates in Thailand and Laos, two experiencing high failure rates in Vietnam, and one experiencing high failure rates in Myanmar (WHO 2018a). As of yet no ACT failure has been observed in Africa.

The migration of antimalarial drug resistance

The majority of deaths from malaria occur in Africa and there is considerable concern amongst the global health community that ACT treatment failure, mediated by resistant strains of *P.falciparum*, may migrate from Southeast Asia to Africa (discussed in Hanboonkunupakarn and White (2016)). Indeed, such a migration event could result in greater than one-hundred thousand additional deaths annually, as well as substantial economic costs (Lubell et al. 2014). Moreover, the scenario has historical precedent. Chloroquine resistance was first detected in 1957 in Thailand (Harinasuta et al. 1965). In 1978, it was detected in Kenya and Tanzania (Clarke 1979). Over the next two decades it spread across the continent (reviewed in Wernsdorfer and Payne (1991)), and the dominant allele (as assessed by the flanking microsatellites) was found to have originated in Southeast Asia (Wootton et al. 2002). Similarly, resistance to the drugs sulfadoxine and pyrimethamine, often given as a combination therapy under the trade name Fansidar, was deployed following increasing treatment failure to chloroquine. Resistance is now widespread in Africa, and the dominant resistance alleles have been found to have originated in Southeast Asia (Roper et al. 2004). The arrival of chloroquine and later sulfadoxine-pyrimethamine resistant parasites in Africa coincided with a increase in malaria-related deaths on the continent – from 995 thousand a year in 1980 to approximately 1.8 million a year in 2004 – and antimalarial drug resistance was found to be "...a prominent covariate and the likely driver of malaria mortality rise in this period." (Murray et al. 2012).

4.1.2 Monitoring for antimalarial drug resistance in Africa

It is clear that monitoring for antimalarial drug resistance in Africa, particularly for the arrival of ACT resistance mutations, would be beneficial. Traditionally, antimalarial drug resistance is assessed through therapeutic efficacy studies (TES). Although TES have produced the majority of data on antimalarial drug resistance that currently exists³, there are challenges associated with their implementation.

³For example, see <https://www.wwarn.org/>.

Assessment of treatment failure by TES requires repeated blood sampling and a long follow-up period (28 or 42 days, depending on the half-life of the drug), with microscopic analysis of blood films and genotyping to distinguish recrudescence from reinfection (WHO 2018a). Ultimately, the approach is time consuming and labour intensive, demanding on patients, and without considerable investment, relatively low-throughput. As a result, large scale and routine surveillance of antimalarial drug resistance is currently out of reach for most malaria control programs.

The identification of causal mutations underlying antimalarial drug resistance has made possible a new type of surveillance, mediated by genetic data. Here, a single dried blood spot (DBS), subjected to DNA sequencing, could be sufficient to determine the antimalarial drug resistance status of a patient – across multiple antimalarial drugs – and without the need for a protracted follow-up period and associated delay to result. The growing sequencing throughput and declining costs mean this approach is highly scalable, and it could facilitate cost-effective nationwide surveillance. Yet, using genetic data to monitor for antimalarial drug resistance also presents challenges. The tremendous expansion in DNA sequencing starting from the mid-2000s was driven by so-called "Second-generation" sequencing platforms, such as those from Illumina. Although these machines have a high-throughput and a low error rate, they are costly (>\$100,000 USD). In addition, they require a service contract for regular microscope alignment, calibration and maintenance, and take significant expertise to run. As a consequence, they largely remain in the preserve of industrial-scale sequencing facilities, none of which are located in malaria endemic Africa. This results in samples being shipped out of Africa to be sequenced elsewhere.

For several reasons, shipping samples out of their country-of-origin for sequencing and analysis should be avoided, unless absolutely necessary. Firstly, it reduces country ownership by disenfranchising local scientists and policy makers from engaging with and understanding the implications of the generated data. Secondly, from a practical perspective, it is generally becoming harder to get permission to export human tissue samples and, in some countries, e.g. Tanzania, it is near impossible to get the necessary approvals (correspondence with Dr. Dan

Bridges). Finally, it inevitably increases time-to-result, nullifying one of the important advantages a genetic approach has over a TES, and potentially delaying evidence-based policy making.

Nanopore sequencing, developed by Oxford Nanopore Technologies (ONT) in the early 2000s, is a third-generation sequencing platform designed to be portable and affordable. It was deployed during the Ebola and Zika outbreaks, in Guinea and Brazil, and has been used to sequence *P. falciparum* in a lab setting (Quick et al. 2016; Faria et al. 2017; Runtuwene et al. 2018). The sequencing device itself – called the MinION – is small (90g), requires no maintenance and costs under a thousand US dollars (Loman and Watson 2015; Jain et al. 2016). These features become significant advantages in resource constrained settings.

4.1.3 The Land Rover Bursary

In April 2019, George Busby, Isaac Ghinai, and myself, applied for the Royal Geographic Society (RGS) Land Rover Bursary⁴. The bursary provides £30,000 and the loan of a Land Rover to facilitate a scientific project. Our proposal was to develop a mobile laboratory to monitor for antimalarial drug resistance in Sub-Saharan Africa using nanopore sequencing. In particular, the project proposal had two primary aims. Firstly, we proposed to convert the Land Rover into a mobile laboratory and to trial in-country nanopore sequencing for the genetic surveillance of antimalarial drug resistance. Secondly, as a public engagement aim, we proposed to raise awareness about the challenges facing malaria scientists, policy makers, and healthcare providers working in malaria endemic Africa.

To achieve these aims, we established collaborations with a set of African scientists working across Sub-Saharan Africa, including Dr. Davis Mumbengegwi at the University of Namibia, in Windhoek, Namibia; Dr. Dan Bridges and Dr. Busiku Hamainza at the National Malaria Elimination Centre, in Lusaka, Zambia; Dr. Fredros Okumu at the Ifikara Health Institute, in Ifikara, Tanzania; and Dr. Eric Ochomo, Medical Entomologist at Kenya Medical Research Institute (KEMRI)

⁴ <https://www.rgs.org/in-the-field/in-the-field-grants/expedition-grants/the-land-rover-bursary/>

in Kisumu, Kenya. We proposed a route across Africa starting in Walvis Bay, Namibia and terminating in Mombasa, Kenya; from the Atlantic to Indian Ocean. The intervening transect, which we would drive in the mobile lab, would take us to each of the aforementioned collaborators, as well as across a diverse ecology of malaria-endemic Africa, to health clinics, universities, research institutions, and villages suffering from a high malaria burden. During the trip we would engage in several activities, including piloting the mobile sequencing laboratory with our collaborators, running sequencing and data analysis workshops, learning about the challenges of malaria control from African research scientists and policy makers, and documenting our journey throughout.

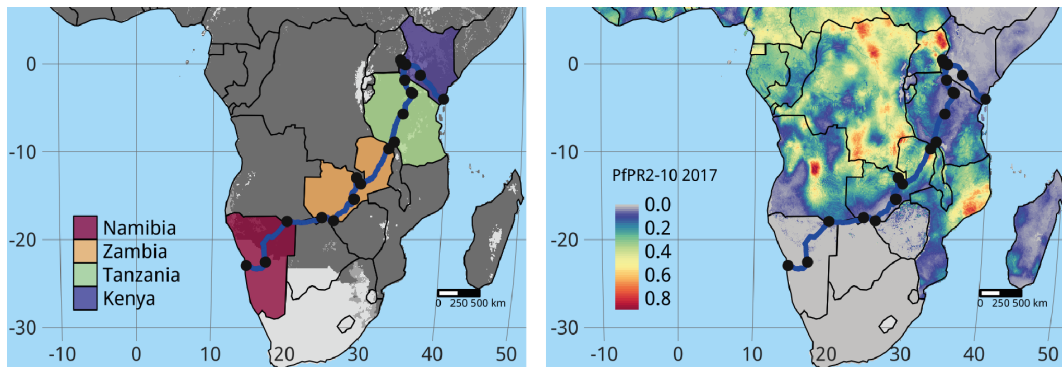


Figure 4.1: Route of The Mobile Malaria Project with stops labelled. Left panel, route overlaid on countries; right panel, route overlaid on malaria prevalence data (Bhatt et al. 2015). Images courtesy of George Busby.

We were awarded the grant in August 2018. The original departure date was February 2019, but due to delays in development of the Land Rover, departure occurred on March 22 2019. We reached Mombasa on May 7 2019. The route including all our planned stops is presented in Figure 4.1.

Dr. George Busby was the project lead and Dr. Isaac Ghinai was the project medic. My role was project scientist and as such I was responsible for developing the mobile sequencing laboratory. I had two main aims:

1. **Develop a mobile laboratory supporting nanopore sequencing in Africa (Section 4.2), which entailed developing:**

- (a) **The laboratory protocol (Section 4.2.1);**

(b) A computational pipeline for data analysis (Section 4.2.2).

2. Trial the laboratory in Sub-Saharan in Africa (Section 4.3).

The trip was documented by George Busby online at mobilemalaria.com.

4.2 Developing a mobile sequencing laboratory

4.2.1 Laboratory Protocols

I developed a laboratory protocol to support nanopore sequencing of malaria parasites in Africa. As dried-blood spots (DBS) are minimally invasive and relatively easy to collect, the protocol was developed with DBS as input. There are several commercially available kits that allow for DNA extraction from a DBS using a combination of proteinase K digestion and DNA adsorption onto silica membranes. For all DNA extractions described in this chapter the QIAamp DNA Investigator Kit (56504) from Qiagen was used.

A challenge to sequencing *P. falciparum* from host blood is obtaining sufficient parasite DNA in a context where the majority of DNA in an extraction – upwards of 95% by weight – derives from the human host. This problem recurs across other infectious organisms and there are three commonly employed approaches taken to enrich for pathogen DNA: (i) separation of host DNA from pathogen DNA (Venkatesan et al. 2012); (ii) amplification of pathogen DNA (Oyola et al. 2016); or (iii) degradation of host DNA (Oyola et al. 2013). For *P. falciparum*, a panel of primers have been developed that support selective whole-genome amplification (sWGA) of parasite DNA in the presence of human DNA (Oyola et al. 2016). The amplification of pathogen DNA as a means of enrichment has the auxiliary benefit of increasing the quantity of DNA available for sequencing, and so I chose to use sWGA to enrich for *P. falciparum* DNA extracted from DBS.

WGS with the Rapid Sequencing Kit

ONT currently offers two distinct library preparation approaches: (i) the Rapid Sequencing Kit (SQK-RAD004); or (ii) the Ligation Sequencing Kit (SQK-LSK109). The Rapid Sequencing Kit uses a transposase to simultaneously fragment DNA and add adaptor sequences and the protocol takes approximately ten minutes to complete. The Ligation Sequencing Kit uses DNA ligase to add adaptor sequences. First requiring DNA fragments be end-repaired and dA-tailed, it takes over an hour to complete. Given the intent to sequence from a mobile platform and in low-resource settings, I first attempted to develop a protocol using the Rapid Sequencing Kit.

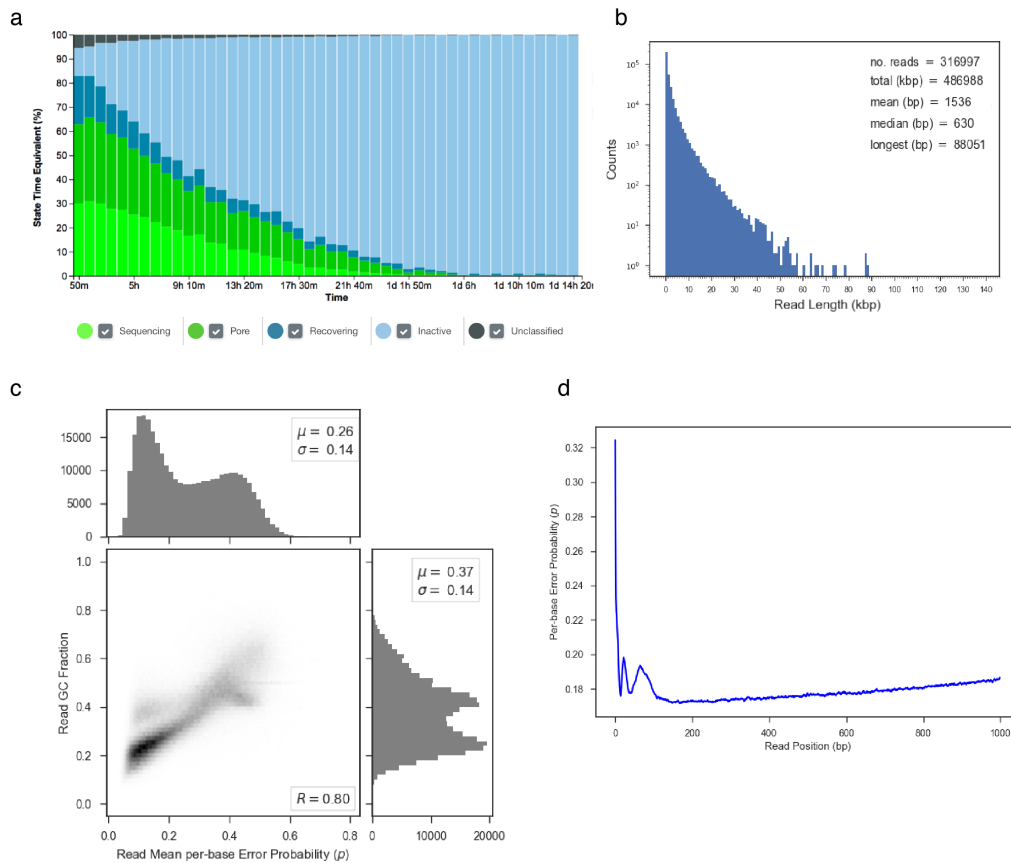


Figure 4.2: Summary of sequencing diagnostics for whole-genome sequencing of *P. falciparum* using sWGA and the ONT Rapid Sequencing Kit. (a) Pore occupancy diagnostic graph from the MinKNOW software. (b) Read length distribution. Note the y-axis is on a log scale. (c) Mean per-base error probability (computed from Q-scores) of each read vs. read G+C fraction. Legends in marginal plots give mean and standard deviation. (d) Per-base error probability (from Q-score) against base position within read, for the first one thousand bases of all sequenced reads. Note how error starts at nearly 33% before declining rapidly, and then climbing slowly after 200 bases.

I combined a mixture of human DNA derived from The International HapMap Project (HapMap 2003) with DNA from the *P.falciparum* lab strain GB4, in proportions similar to those expected in a patient sample (1% (w/w) parasite to human DNA at 2ng/uL). I performed sWGA following the protocol from Oyola et al. (2016). After sWGA, the DNA concentration was 204ng/uL (measured by Qubit), an approximate ten thousand-fold enrichment of parasite DNA. sWGA uses the ϕ 29 polymerase (M0269S) which has multiple strand displacement activity and produces branched DNA structures during amplification. ONT reports that branched structures can cause blockage of the protein nanopores on a R9.4.1 Flow Cell, and so following sWGA I treated the sample with T7 Endonuclease I (M0302S) for 15min at 37°C, which cleaves at DNA mismatches and at non- β DNA structures like those produced by DNA branching (Blanco and Salas 1984). As components of the sWGA reaction can interfere with T7 Endonuclease activity, I performed a DNA clean-up step using Agencourt AMPure XP beads (A63881) prior to and after digestion. Finally, I performed library preparation using the Rapid Sequencing Kit. Approximately 400ng of DNA was loaded onto a new R9.4.1 Flow Cell. The run was initiated 20/11/2018 and an initial pore occupancy of 30% fell to 0% in less than 24 hours of sequencing, as pores rapidly entered the inactive state (Figure 4.2a).

In total, 317 thousand reads totalling 486Mbp of sequence data were generated (Figure 4.2b). The read length distribution was roughly exponential with a mean read-length of 1536bp (Figure 4.2b). The average error rate across all reads, estimated from the Q-scores generated by MinKNOW, was 26%. However the error rate was bi-modal and strongly correlated with G+C content, suggesting the reads deriving from the human genome contained more error (Figure 4.2c). Furthermore, error was much higher at the beginning of reads, reaching a minimum after approximately two-hundred bases, and then slowly climbing (Figure 4.2d). Using *Minimap2* (H. Li 2018), I found that 38% of reads mapped to the *P. falciparum* reference genome resulting in a mean coverage across the genome of 9.5X (Figure 4.3b). However, as is typical of sWGA, coverage was highly variable across the genome with only 33% of the genome having greater than 10X coverage (Figure

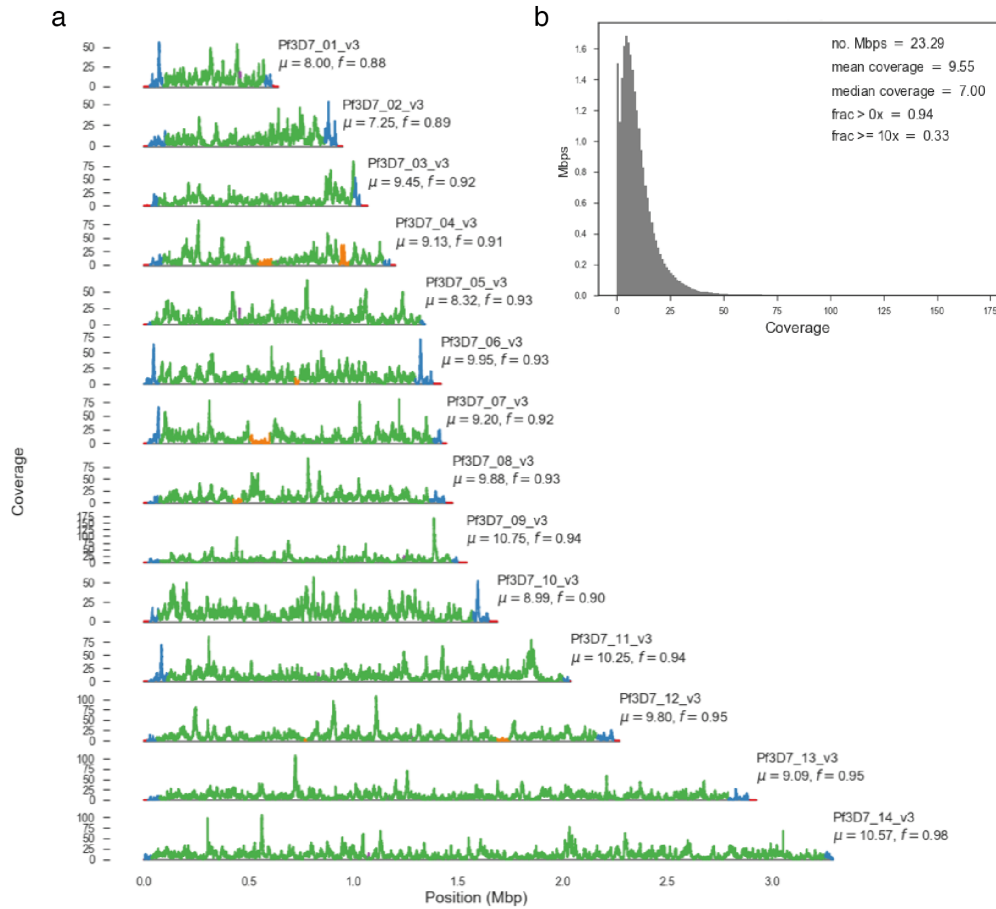


Figure 4.3: Sequencing coverage across the *P. falciparum* genome using sWGA and the ONT Rapid Sequencing Kit. (a) Coverage across the fourteen chromosomes of the *P. falciparum* reference genome. Green indicates core genome, orange centromere, and blue repetitive regions. For each chromosome, mean coverage (μ) and fraction of chromosome with >0X coverage (f) is given at right. (b) Coverage distribution across the entire genome.

4.3a). Focusing on a set of epidemiological relevant genes, I found that coverage was typically less than 10X (Figure 4.4).

The rapid decline in pore occupancy, and resultant low sequencing throughput, may have been caused by a variety of factors. One probable contribution was the relatively short read lengths (the median read length was 680bp). Short reads spend less time inside the protein nanopores on the R9.4.1 Flow Cell, and it is known that unoccupied nanopores experience faster rates of degradation during sequencing. As the Rapid Sequencing Kit cleaves DNA, ONT recommends a mean

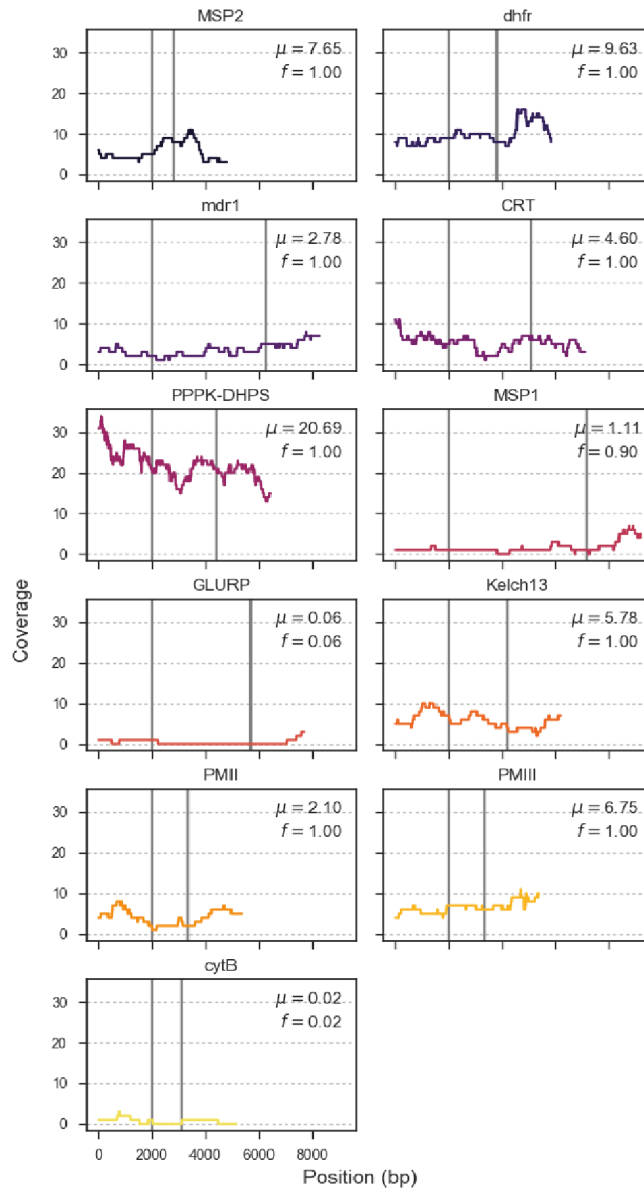


Figure 4.4: Coverage against a selected set of key antimalarial resistance and epidemiological *P.falciparum* genes using sWGA and the ONT Rapid Sequencing Kit. Vertical grey lines delineate coding sequence. μ , mean coverage across coding sequence. f , fraction of coding sequence with >0X coverage.

input DNA fragment length of 30kbp. Given the processivity of $\phi 29$, it is unlikely parasite DNA produced by sWGA would exceed ≥ 30 kbp (first assuming the extracted DNA exceeded ≥ 30 kbp) (Blanco and Salas 1984). Work quantifying the distribution of DNA fragment lengths at various steps during the library preparation is needed to explore this further.

Amplicon sequencing with the Ligation Sequencing Kit

Given the low sequencing throughput observed with the Rapid Ligation Sequencing Kit, I decided to switch to the Ligation Sequencing Kit, and from WGS to targeted, polymerase chain reaction (PCR)-based amplicon sequencing. The advantages of PCR-based amplicon sequencing include (i) being able to focus sequencing throughput on genes of interest, in this case the set of genes harbouring mutations known to confer antimalarial drug resistance; (ii) increasing the enrichment for parasite DNA, which both improves mapping statistics and makes the assay more robust to variability in parasitemia or DNA extraction efficiency; and (iii) control over DNA fragment length, which allows optimization of nanopore sequencing throughput. There are also significant disadvantages, including that PCR-based amplicon sequencing (i) requires developing a PCR for each gene of interest, and (ii) can introduce amplification bias and/or error. Finally, the Ligation Sequencing Kit is more labour intensive and expensive.

I first sought to develop PCRs for a set of *P. falciparum* genes known to harbour antimalarial resistance mutations. There are challenges particular to developing PCRs for *P. falciparum* and for nanopore sequencing. The genome of *P. falciparum* is very low complexity, with an A+T content of 80.6% (Gardner et al. 2002). Low complexity sequences limit the availability of suitable and unique primer binding locations, and decrease the efficiency of DNA polymerase, which can stall or experience slippage in homopolymeric or tandemly repeating regions (Lovett 2004). Nanopore sequencing throughput is higher with long DNA fragments (ideally in the range of 2 – 10kbp), whereas PCR is in general more efficient⁵ and robust the shorter the amplified template. Longer templates require more processivity⁶ from the DNA polymerase and increase the likelihood of secondary structures developing during thermocycling, both effects which reduce product yield.

I designed primers against a suite of antimalarial resistance genes, using the Primer3 software with template sequences extracted from release 39 of the *P.*

⁵Efficiency is the percent conversion of template to product per cycle.

⁶Processivity is the average number of nucleotides added before dissociation.

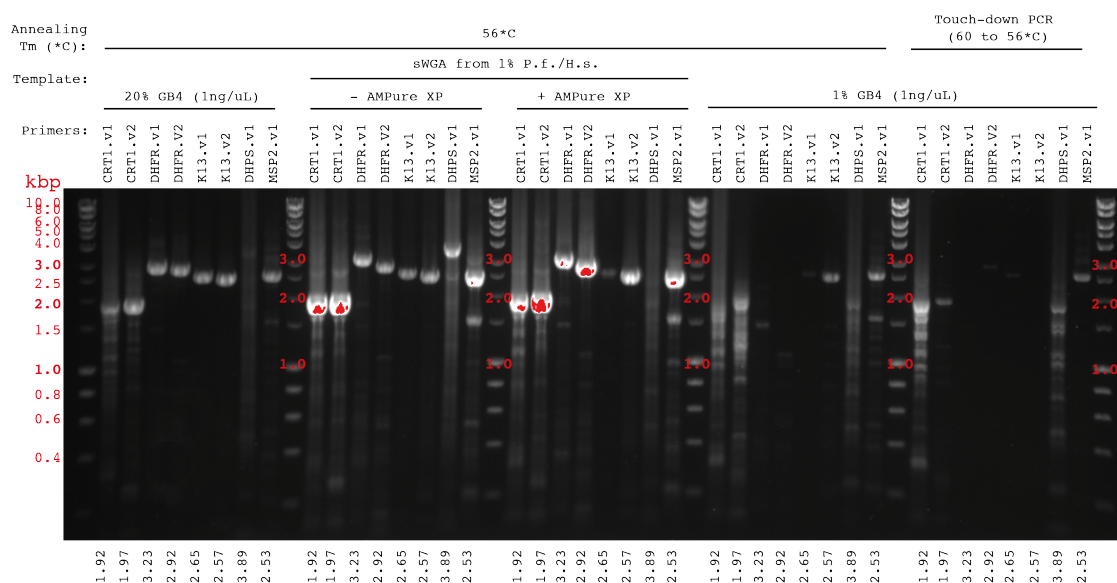


Figure 4.5: Using sWGA to facilitate PCR in low percentage *P. falciparum* DNA blood samples. The same eight PCRs are shown, two against against *CRT* (CRT.v1, CRT.v2), *DHFR* (DHFR.v1, DHFR.v2), and *Kelch13* (K13.v1, K13.v2), plus one against *DHPS* (DHPS.v1) and *MSP2* (MSP2.v1), across six different template DNA conditions. From left to right: 20% (w/w) GB4/HapMap DNA mixture at 1ng/uL, simulating expected output of sWGA reaction; next two panels, 1% (w/w) GB4/HapMap DNA mixture after sWGA, either with or without Agencourt AMPure XP DNA clean-up; 1% (w/w) GB4/HapMap DNA mixture before sWGA, notice nearly all PCRs fail; % (w/w) GB4/HapMap DNA mixture without sWGA, using a Touch-down PCR protocol in attempt to increase early cycle specificity. Again most reactions fail.

falciparum 3D7 reference genome (Untergasser et al. 2007). I adjusted the following settings in Primer3 to increase primer selection stringency: "Max T_m difference" to 2°C; "Max Self-Complementarity" to 3; "Max 3' Complementarity" to 0; "Concentration of divalent cations" to 3mM; "Concentration of dNTPs" to 1mM; the remaining settings were left as default. For every gene, I extracted the coding sequence plus 2kbp of flanking sequencing and selected primers that annealed >200bp from the start and end of the coding sequences when possible, in order to minimize sequencing error across the gene body. For *CRT*, which contains 13 exons and spans over 3kbp, I designed primers that covered the first half of the gene only; this spans the chloroquine-resistance conferring mutation K76T while making the PCR more robust.

I tested the primers using TaqMan polymerase (ID TK) and standard PCR conditions, with a DNA template of 100% GB4. Under these conditions, all the

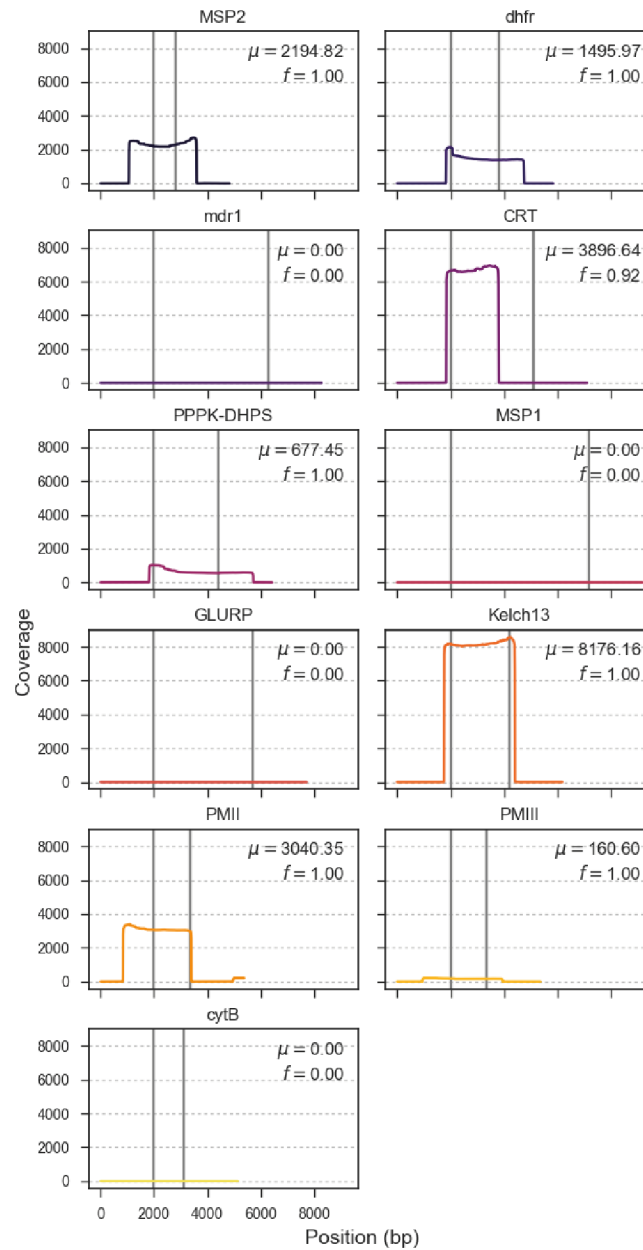


Figure 4.6: Coverage against a selected set of key antimalarial resistance and epidemiological *P.falciparum* genes using sWGA, PCR, and the Ligation Sequencing Kit. Vertical grey lines delineate coding sequence. μ , mean coverage across coding sequence. f , fraction of coding sequence with >0X coverage. Results shown for 3D7; results were similar for GB4. Note that PCR reactions were run for *MSP2*, *DHFR*, *CRT*, *DHPS*, *Kelch13*, *PMII* and *PMIII*; other resistance genes are only shown for consistency with Figure 4.4.

PCRs failed (data not shown). Work by Su et al. (1996) has demonstrated that PCR product yield in *P. falciparum* is often improved by reducing the extension temperature below 72°C (to 60-65°C), despite 72°C being the temperature at which

Step	T (°C)	Time
1	94	2m
2	94	15s
3	56	2m
4	60	1m
5	goto 2	29x
6	60	2m
7	4	10m

Table 4.1: KAPA Polymerase PCR Conditions.

Taq polymerase has the greatest activity. The A+T content of *P. falciparum* tends to be even higher in regions flanking the coding sequence, and it is believed that a low extension temperature improves amplification through these regions (Su et al. 1996). I repeated the PCR reactions with adjusted thermocycling conditions (Table 4.1). I also switched to the KAPA HiFi PCR Kit (KK2101), which contains a polymerase with a higher processivity and better performance in low complexity regions. On 100% GB4 template, all PCR reactions were successful (data not shown). I repeated the PCR reactions using template conditions meant to mimic those excepted from DNA material extracted from a DBS, either with or without sWGA before PCR (Figure 4.5). I found that PCR failed for most target genes when done directly from a DBS. After sWGA, however, robust product was observed for all target genes tested.

I conducted an experiment designed to replicate the entire laboratory protocol, from sWGA to nanopore sequencing, that would be conducted in Africa from the mobile lab. I created two 1% (w/w) parasite-human DNA mixtures, with parasite DNA from the lab strains 3D7 and GB4, and human DNA from the The International HapMap Project (HapMap 2003). I performed sWGA and PCR on these samples as described above. I targeted the following genes with PCR: *MSP2*, *DHFR*, *CRT*, *DHPS*, *Kelch13*, *plasmepsin2* and *plasmepsin3*. Following PCR, the amplicons for each sample were pooled at equimolar amounts. In preparation for the Ligation Sequencing Kit, samples were end-repaired and dA-tailed following the protocol from ONT. Individual samples were barcoded using the Native Barcoding Kit (EXP-NBD104), which I selected to avoid additional PCR steps. Finally, both

samples were pooled at equimolar amounts and the library was prepared using the Ligation Sequencing Kit. The run was initiated on 2019/02/11 and 223Mbp of data was collected; 104Mbp for 3D7 and 116Mbp for GB4. This resulted in deep coverage (160X-8170X) of the seven amplicons targeted (Figure 4.6). The read-level error rate was also lower, at 8% across 3D7 and GB4. For these reason, an amplicon approach using the Ligation Sequencing Kit was employed for our sequencing in Africa.

4.2.2 Computational Pipeline

Real-time basecalling

I developed a computational pipeline for analysing malaria genetic data generated by nanopore sequencing. My primary aims were to generate a pipeline that (i) would run in real-time, such that base-calling, adapter trimming, demultiplexing, mapping, and mutation identification occurred contemporaneous with sequencing; (ii) would run in a field-setting, in effect without the need for an internet connection or any compute larger than a laptop; and (iii) would be easy to install and run.

The ONT MinION is operated from the MinKNOW software which is provided with purchase of a MinION. MinKNOW produces "squiggle" files, in .fast5 format, which record information about current disruption detected by the application specific integrated circuit (ASIC) as reads thread through the protein nanopores arrayed on the flow cell. The user is free to set how many reads should be recorded before a single file is written to disk. I allowed for 4000 reads per .fast5 file (the default) which, for a typical run, resulted in .fast5 files being generated at a rate of approximately one every two minutes. ONT suggests the software is run with a 512Gb SSD HD in order to accommodate the rapid disk I/O and the potentially hundreds of gigabytes of .fast5 files that can be generated during a run.

In the context of nanopore sequencing, real-time basecalling entails converting these .fast5 files into .fastq files while sequencing is ongoing. The best basecalling method at present is Guppy, developed by ONT, which employs a recursive neural network to infer sequence composition from current disruption (Wick et al. 2019). Its performance is greatly enhanced if it is run on a GPU, which allows for parallelization

of the matrix multiplication operations that occur during a forward pass through a neural network. Guppy can be set to run real-time from the MinKNOW software interface, though in the absence of adequate compute, only a small fraction (1-5%) of reads will be called by the end of a run. To allow for real-time basecalling, I opted to purchase the MinIT – a small compute platform developed by ONT that contains a GPU and a 512Gb SSD. The MinIT is operated through a local laptop, which can connect to the MinIT through either an ethernet cable or bluetooth. Using the MinIT, basecalling keeps up with .fast5 file creation in all but the highest throughput runs. For example, our first run in Zambia produced nearly 20Gbps of sequence data, and the MinIT had called greater than 90% of these bases during sequencing, with the remaining being called shortly afterwards. An additional advantage of the MinIT is that it frees compute and drive space on the laptop that would otherwise be required to run MinKNOW; both the .fast5 files and .fastq files are processed on the MinIT GPU and stored on its SSD.

Real-time mapping and variant identification

I wrote a collection of python scripts to conduct adapter trimming, demultiplexing, mapping and variant identification for a targeted set of antimalarial resistance mutations. These were designed to run in real-time on the laptop connected to the MinIT.

Briefly, the script `run_pf-mapping.py` monitors the MinIT hard-drive and copies .fastq files to the associated laptop as they are generated. It then executes adapter trimming and demultiplexing (where reads containing different sample barcodes are separated into their own .fastq files) using `Porechop`⁷ under default settings. Once reads are adapter trimmed and demultiplexed, `run_pf-mapping.py` maps them to *P.falciparum* reference genome⁸ using `Minimap2`, generating .SAM files. `Minimap2` was designed with a specific parameter set (`-ax ont`) to facilitate the mapping of the long, indel-containing reads typical of nanopore sequencing (H. Li 2018). Subsequently, the script `run_sam-to-bam.py` sorts, indexes, and converts

⁷Available at <https://github.com/rrwick/Porechop>.

⁸Release 39 of laboratory strain 3D7, downloaded from <https://plasmodb.org/>

the .SAM files to binary format (.BAM). For a .fastq file containing 4000 reads, this entire process typically took around two minutes⁹.

The script `run_mutation-search.py` was used to search the resulting .BAM files for the presence of a known set of antimalarial resistance mutations. The user is required to create an .ini file for each gene of interest, giving the gene name and a list of the mutations of interest. Mutations are specified by their amino-acid change (e.g. K76T) and the script scans for all codons that would result in such a change (e.g. AAA to ACT, ACC, ACA or ACG). It does not compute a consensus sequence, but rather creates a read pileup across the gene using `samtools mpileup`, and computes the fraction of all codons at a given position that correspond to the listed mutation (a within-sample mutation frequency on the amino-acid level). This is done to accommodate the possibility of mixed infections, where taking a consensus sequence is inappropriate.

The entire pipeline is available on GitHub (github.com/JasonAHendry/mmp-pipelines) and dependencies can be installed using `conda` (<https://docs.conda.io/>). It was used for both amplicon sequencing experiments in Zambia, as described in Section 4.3.

4.2.3 Mobility, costs, and shipping

The laboratory protocol and computational pipeline were to be incorporated into a mobile laboratory, transported approximately 7500km across Africa inside a 2019 Land Rover Discovery, and operational from the Land Rover with no additional infrastructure. We consulted with Land Rover's Special Vehicle Operations (SVO) team to develop a set of modifications for the Land Rover that would accommodate the lab and facilitate the drive.

As part of the RGS Land Rover Bursary, we were loaned a 2019 Land Rover Discovery HSE TdS, which is a 7-seater SUV (Figure 4.7). After removing the back two seats, we had a loadspace of 940mm x 1100mm x 8000mm in which to store all the lab equipment, reagents, and other luggage required for the estimated

⁹On a 2018 2.9GHz Intel Core i9 Macbook Pro with 32Gb of RAM



Figure 4.7: 2019 Land Rover Discovery HSE TdS used to house lab.

two-month trip. Figure 4.8 shows a schematic of all of the lab equipment and reagents needed to perform nanopore sequencing. I purchased as-small-as-possible versions of all the essential lab equipment – including a minicentrifuge, vortex, heat block, and miniature PCR machines produced by miniPCR. The laboratory protocol also required a Qubit, for DNA quantification, a magnetic rack, for DNA clean-up steps, the MinION, MinIT and a laptop. The full equipment list and costs can be found in Table 4.2. Land Rover SVO added a roof-rack which allowed for storage of our tents, the majority of our personal luggage, and scientific plastic-ware (pipette tips, microcentrifuge tips, &c.) on the roof. The scientific equipment was stored in pelican cases with custom cut foam for support. We used three cases, one for primary equipment (Nanuk 945), one for consumables (Nanuk 945), and one with backup equipment, additional scientific plastic-ware, microcentrifuge racks, electrical cables, waste disposal jars, stationary, &c. (Nanuk 935). In a final large plastic container we stored exclusively pipette tips.

The majority of reagents required for sequencing need storage at 4°C or -20°C, and critical to the lab was maintaining a cold-chain across the continent. The SVO purchased and installed a fridge-freezer unit with two compartments, one which we kept at 4°C and the other at -20°C. The fridge-freezer was powered

Item	Serial No.	Use	Price (£)
Pipetman Classic 4 Pipette Kit	F167380	Throughout	458
Greiner Mini Vortex Mixer	845060	Throughout	93
Eppendorf TM MiniSpin TM Benchtop Centrifuge	12833252	Throughout	665
Fisherbrand TM Mini Digital Block Heater	12186560	DNA Extraction	245
miniPCR TM mini16 thermal cycler	QP-1016-01	sWGA and PCR	620
Qubit 4 Fluorometer	Q33226	DNA Quantitation	2,555
DynaMag-2 Magnet-1	12321D	DNA Clean-up	225
ONT MinION	NA	DNA Sequencing	820*
Oxford Nanopore MinIT	Q33226	Basecalling	1,872
LaCie 1TB Rugged Mini USB 3.0 Portable Hard Drive	LAC301558	Data Storage	66
Total (£)			7,617

Table 4.2: List of laboratory equipment purchased for mobile lab. Prices are from University of Oxford Oracle R12 purchasing system and rounded to the nearest pound. *MinION comes as part of the MinION Starter Pack, which also includes one Ligation or Rapid Sequencing Kit, two R9.4.1 Flow Cells, a Control Expansion Kit, and a Flow Cell Wash Kit.

by a 410Ah Li-ion leisure battery housed in an underseat loadspace of the Land Rover. The leisure battery was charged through the car engine, and a digital display on the fridge-freezer as well as a power inverter allowed us to monitor and maintain both the fridge-freezer temperature and car battery voltage. We stored high-density plastic microcentrifuge tube racks in the -20°C which allowed us to prepare temperature-sensitive reactions "on ice" when working in the field.

The car was transported on a container ship to the port of Walvis Bay, Namibia, by DB Shenker, and the majority of the lab equipment transported inside. The Qubit, MinION and MinIT were brought in our hand luggage. We made two -20°C shipments containing all our reagents: one to Windhoek, Namibia; and one to Kisumu, Kenya. In addition, three flow cells were transported at room

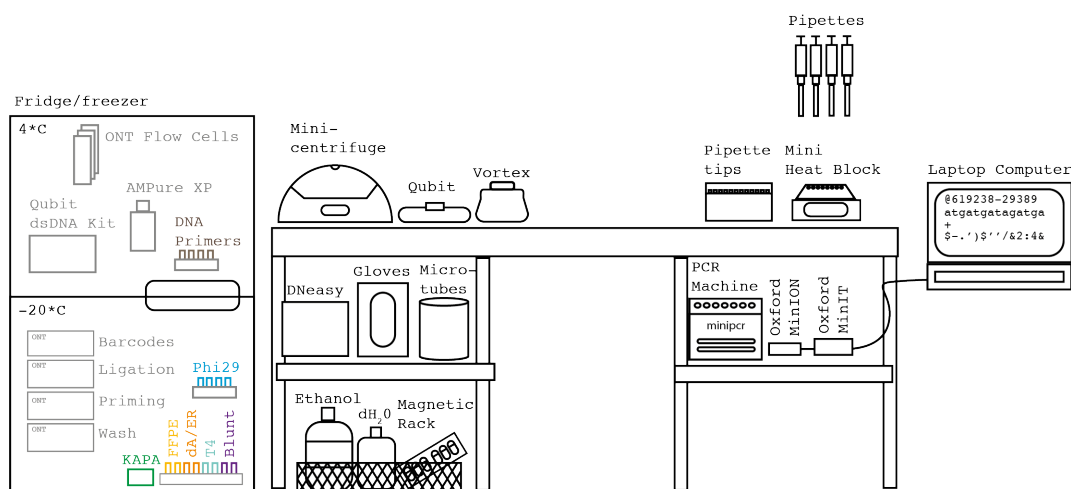


Figure 4.8: Schematic of mobile laboratory including hardware and reagents.

temperature in hand luggage. A plastic table was purchased in Windhoek on which lab work could be conducted.

4.3 Deploying the laboratory in Sub-Saharan Africa

With our collaborators, we sequenced *P. falciparum* DNA using the MinION three times in Zambia: twice in amplicon sequencing experiments, and once in a WGS experiment as part of a workshop. This section describes the two amplicon sequencing experiments, which are summarized in Table 4.3.

Our first sequencing experiment was conducted at The National Malaria Elimination Centre (NMEC) in Lusaka, Zambia (Figure 4.9). The NMEC is a directorship under the Ministry of Health and is centrally involved in guiding Zambia's National Malaria Strategic Plan. It also houses a molecular laboratory conducting a variety of assays on *Plasmodium* and their vectors. The facilities accommodate NMEC staff and partners, as well as hosting local university students and post-graduate interns. Our aim at NMEC was to pilot the nanopore sequencing assay for the first time in Africa, as well as run a two-day sequencing workshop.

We arrived at NMEC on the evening of April 8th and the workshop was to be held on the 11th and 12th. Upon arrival, the NMEC team provided us with four DBS from malaria patients in Zambia; two from the north and two from

Date	Location	Samples	Genes	Throughput
10/04/2019	NMEC Lusaka	3D7	<i>Kelch13</i>	19.84 Gbp
		IPC5202	<i>CRT</i>	
		4 Zambian	<i>DHFR</i>	
			<i>MSP2</i>	
17/04/2019	TDRC Ndola	3D7	<i>Kelch13</i>	0.89 Gbp
		IPC5202	<i>CRT</i>	
		6 Zambian	<i>DHFR</i>	
			<i>MSP2</i>	

Table 4.3: Summary of *P. falciparum* amplicon sequencing experiments conducted in Zambia.

the south of the country. In addition, they provided us with DNA from the *P. falciparum* reference strain, 3D7, and DNA from a known multi-drug resistant strain, IPC5202, from western Cambodia.

On the 9th and 10th of April I conducted the laboratory protocol as outlined in Section 4.2.1. We chose to focus on four target genes: *Kelch13*, *DHFR*, *CRT* and *MSP2*. Initially, all PCR reactions done with the KAPA HiFi PCR Kit (KK2101) shipped from Oxford, including those with the 3D7 and IPC5205 control, failed. These PCRs were successful when repeated with a TaqMan PCR Kit provided by the NMEC team, suggesting the KAPA HiFi PCR Kit was compromised in shipping. The laboratory protocol was completed and the library – containing four amplicons across six samples – was loaded on the evening of the 10th. Using the MinIT and real-time analysis pipeline described above, we were able to call the drug resistance status of the samples on the morning of the 11th.

A second experiment was performed on the 15th and 16th of April, near Ndola, the capital of Zambia’s Copperbelt Province. On the morning of the 15th, we visited a health clinic in Kaniki, which is located in a high malaria transmission zone approximately 100m from the border with the Democratic Republic of the Congo. Here, our collaborators collected DBS from volunteers who tested positive for *P. falciparum* malaria by RDT. Of 11 RDT positive individuals our collaborators met at the clinic, 7 agreed to contribute a DBS to our study.

The DBS were brought to the The Tropical Diseases Research Centre (TDRC) in Ndola, where DNA extraction was performed by members of NMEC and TDRC. Due to time limitations, I worked with members of NMEC to complete the remainder of the laboratory protocol in Aabiya Lodge, Ndola. One of the samples failed during DNA extraction, and a second sample had very low DNA concentration following sWGA but was included in the library; resulting in six Zambian samples total. As with the first experiment, we focused on generating four amplicons per sample: *Kelch13*, *DHFR*, *CRT*, and *MSP2*; and included 3D7 and IPC5205 as controls. Again due to time limitations, the library was loaded on the MinION on the morning of the 17th, in the back of the Land Rover Discovery. Sequencing took place during the drive between Ndola and Nakonde, and the drug resistance status of the samples was clear by that afternoon.

4.3.1 Sequencing metrics

Sequencing diagnostics for the run on 10/04/2019 are shown in Figure 4.10. Unfortunately, for the run on 17/04/2019, the MinIT came unplugged before the run completed and the sequencing diagnostic plot was lost. Throughput was significantly lower for the 17/04/2019 run (Table 4.3); it is unclear whether this is due to the run taking place in a moving vehicle, the run terminating prematurely, or a combination of these and other factors.

For the run on 10/04/2019, pore occupancy maintained at >80% for 24hrs. The run terminated as a consequence of the 512Gb MinIT hard-drive reaching capacity. A total of 19.84Gb of sequencing data was generated, 16.6Gb of which passed the MinKNOW software's quality control threshold.

Despite starting with a DBS, for both sequencing runs the vast majority of reads mapped to the 3D7 reference genome (>80%), suggesting the two rounds of amplification (sWGA and PCR) substantially enriched for *Plasmodium* DNA (Figure 4.11b). The mean read length was 2.2 kbp on 10/04/2019 and 1.7 kbp 17/04/2019 (Figure 4.11a). The read-length distributions showed pronounced peaks at read lengths corresponding to the size of each of the four amplicons.

4. Collecting *P. falciparum* genetic data from a mobile laboratory in Sub-Saharan Africa

121

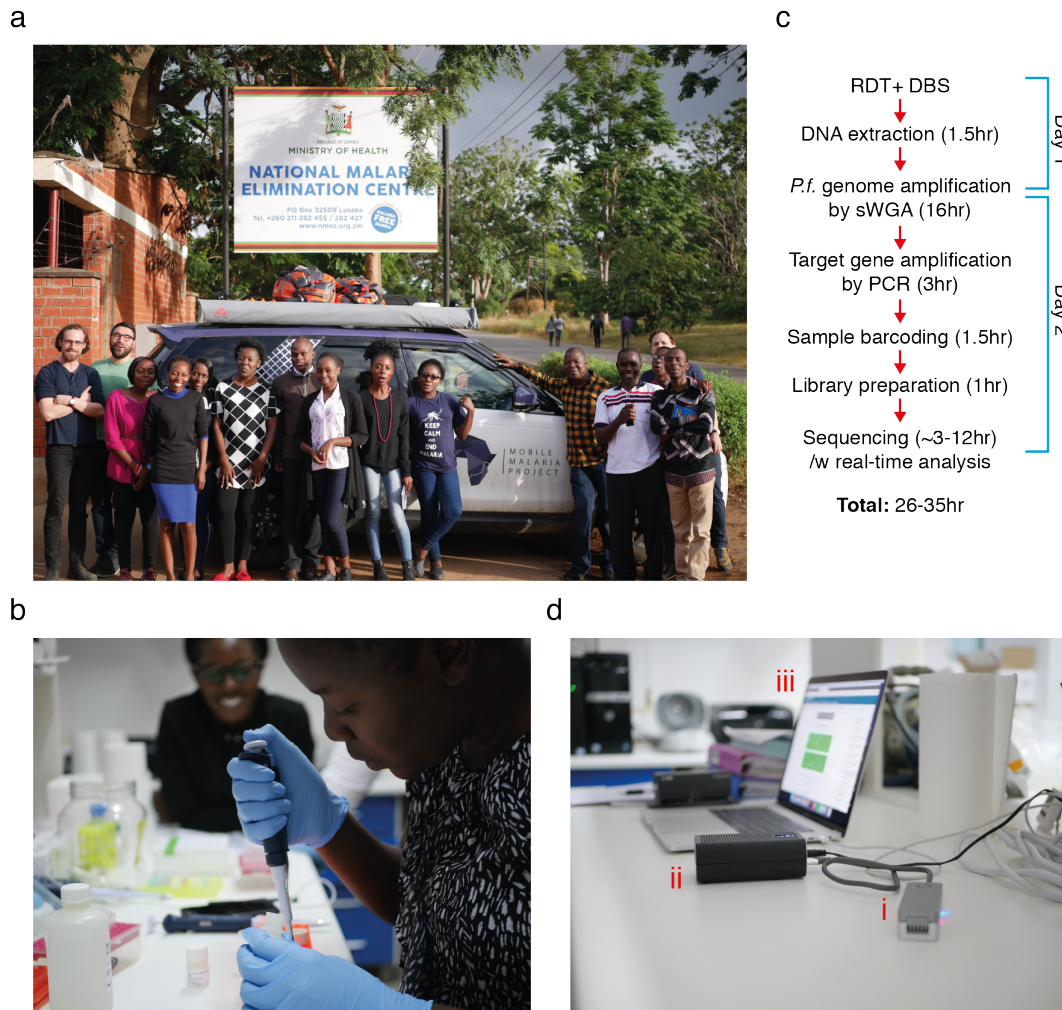


Figure 4.9: Nanopore sequencing of *P. falciparum* antimalarial resistance genes at The National Malaria Elimination Centre in Zambia. (a) The National Malaria Elimination Centre and Mobile Malaria Project team, outside NMEC before the workshop deployed in Zambia. (b) During a two-day workshop local students were trained on nanopore sequencing. (c) Schematic of the sequencing workflow. Dried blood-spots (DBS) from patients that are rapid diagnostic test positive (RDT+) undergo DNA extraction, enrichment of *P. falciparum* DNA by selective-whole genome amplification (sWGA) (Oyola et al. 2016), enrichment of antimalarial resistance genes by polymerase chain reaction (PCR), library preparation and sequencing. The entire process takes under two days. (d) Sequencing setup at NMEC, allowing for real-time sequencing and analysis. (i) Oxford Nanopore Technology (ONT) MinION DNA sequencer, (ii) ONT MinIT, a GPU-based compute-platform, (iii) 15" Macbook Pro laptop.

Interestingly, there were also minor peaks at lengths corresponding to the sum of two or more amplicons, suggesting that the pooling of different amplicons prior to the ligation steps that occur during library preparation can result in concatenation of amplicons. The presence of a significant number of chimera-mapped reads

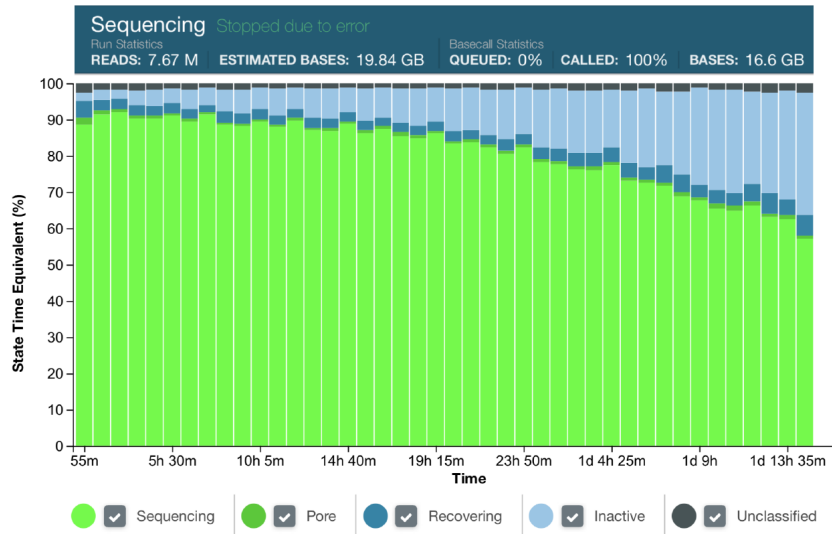


Figure 4.10: MinION sequencing throughput and pore occupancy for the amplicon sequencing experiment on 10/04/2019. Pore occupancy began at above 90% and finished at above 60% after approximately 38 hours of sequencing. 7.67M reads totalling 19.84GB of DNA sequence was generated. Shown are screenshots from the MinKNOW software.

corroborates this hypothesis.

4.3.2 Error rate analysis

Though declining with each new generation pore, nanopore sequencing has a significantly higher error rate than some other sequencing approaches. For example, a typical read-level error rate for nanopore sequencing is between 5-15%, whereas for Illumina sequencing it is often <0.1%. Low-complexity sequence, such as that found in *P. falciparum*, is expected to exacerbate this error.

The inclusion of 3D7, a *P. falciparum* laboratory strain and the reference genome, allowed for a thorough characterization of the error rate of our experiments in Zambia. After mapping with `Minimap2`, I generated read pileups across the coding sequences of each amplicon using `samtools mpileup`. I created a consensus sequence by taking the most frequently observed base, hereafter written as the "consensus basecall"¹⁰, across all reads spanning a given site. I compared the consensus basecall at each site in our 3D7 sample to the correct base given by the 3D7 reference genome (Figure

¹⁰For a site i , the consensus basecall is defined as $\operatorname{argmax}(A_i, T_i, C_i, G_i, del_i, ins_i)$, where $A_i, T_i, C_i, G_i, del_i, ins_i$ are the number of reads containing an A, T, C, G, deletion or insertion, respectively, at site i .

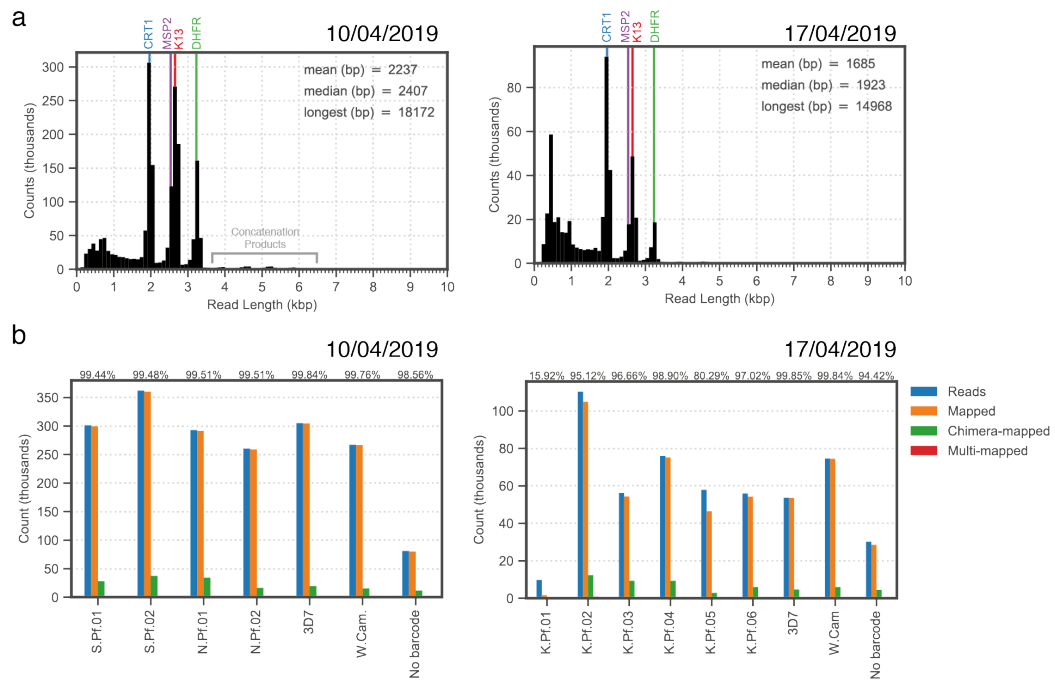
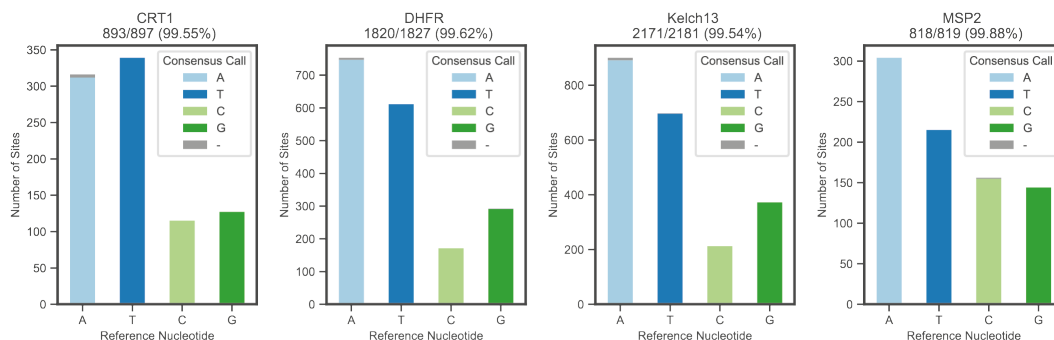


Figure 4.11: Read-length distributions and mapping statistics for both amplicon sequencing experiments. (a) Read-length distributions for the experiment on 10/04/2019 (left) and 17/04/2019 (right). The distributions show a multi-modal pattern, with the highest peaks corresponding to the size of the four amplicons (indicated by colored vertical lines). A series of smaller peaks correspond closely to combinations of amplicons, suggesting concatenation. (b) Mapping statistics for each barcoded sample and for those reads without a barcode, to release 39 of the *P. falciparum* reference genome. Mapping percentages are indicated for each sample at top, and are all greater than >80%, except for sample K.Pf.01 on 17/04/2019, which had a very low DNA concentration following sWGA. The presence of chimera-mapped reads again suggests concatenation of amplicons.

4.12). For all four genes and in both experiments, the accuracy of the consensus sequence was >99.3% (Figure 4.12). Moreover, all errors in the consensus sequence were deletions, and all of these deletions resided in homopolymer regions of length four or greater (Figure 4.13a). For all consensus errors, the correct basecall had the second highest frequency in the read pileup (Figure 4.13b).

I also explored the error rate at the level of individual reads, again by comparing basecalls in the 3D7 sequencing data to the correct base given by the 3D7 reference genome. The overall per-site read-level error rate distribution had a mean of 11.5% and 10.6%, for the experiments on 10/04/2019 and 17/04/2019, respectively (Figure 4.14). For both experiments, the majority of these errors were deletions, such that

10/04/2019



17/04/2019

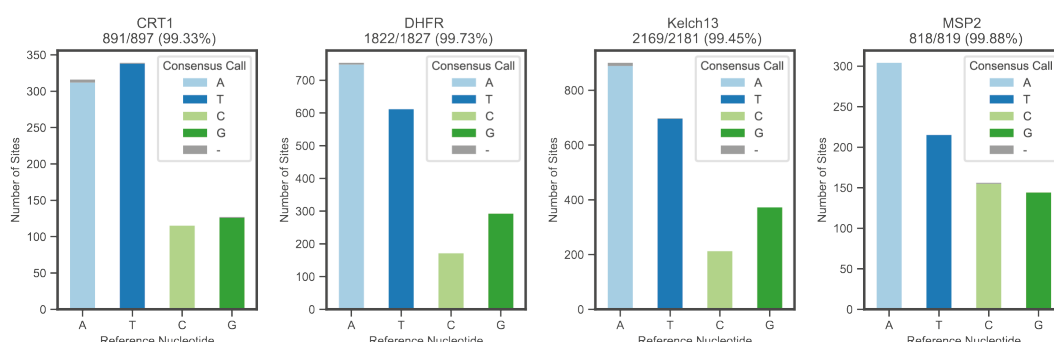


Figure 4.12: Consensus-level error analysis across both amplicon sequencing experiments. Top row shows the consensus sequence error for *CRT1*, *DHFR*, *Kelch13*, and *MSP2* for the amplicon experiment on 10/04/2019, from left to right. Bottom row is the same as top but for the experiment on 17/04/2019. Each plot gives the reference sequence base on the x-axis and represents the consensus sequence basecall in color, as a stacked barplot. The number of sites correct over all sites within the gene body is given at the top of each plot. Notice that the only errors in the consensus sequence are deletions, shown in grey (-).

the mean rate of per-site single nucleotide variant (SNV) error at the read-level was 2.6% and 2.3% for the 10/04/2019 and 17/04/2019 experiments, respectively. ONT believes that very high-throughput runs (> 15Gbp) may produce higher error rates in their later hours, due to the consumption of ATP within the sequencing buffer (correspondence with Divya Mirrington); which may explain the slightly higher read-level error on 10/04/2019.

Similar to the error bias observed at the consensus-level, deletion errors at the read-level were enriched in long homopolymer tracks (Figure 4.15a). In addition, I noted that the SNV errors had a purine-to-purine and pyrimidine-to-pyrimidine substitution bias (Figure 4.15b).

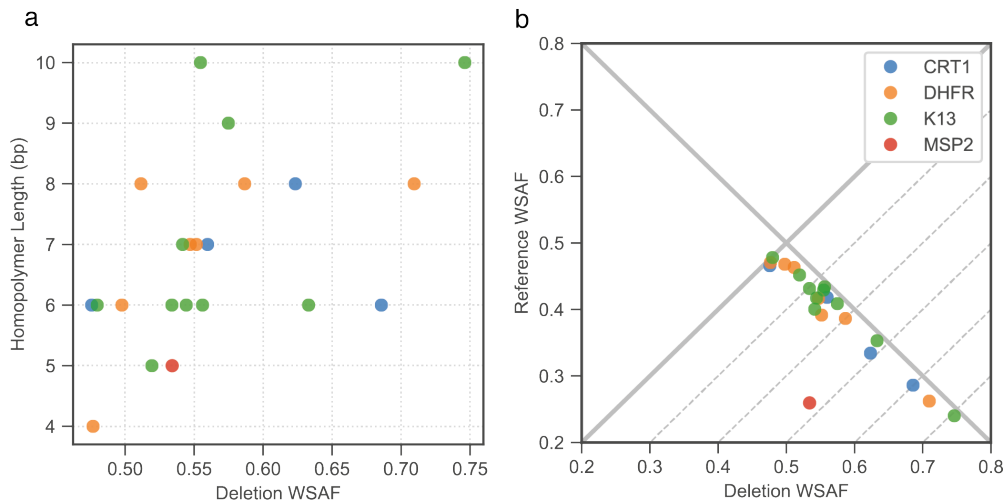


Figure 4.13: Consensus-level error bias. (a) Scatterplot visualizing the Deletion WSAF (within-sample allele frequency, fraction of reads harbouring a given allele, in this case a deletion, across the pileup) against homopolymer length for all consensus sequence errors from 10/04/2019 experiment. Note the y-axis begins at four. (b) Reference WSAF vs Deletion WSAF shows that the correct call had the second highest frequency for all errors in the consensus sequence. Aside from the deletion in MSP2, the deletion plus reference WSAF is very close to 1 (indicated by decreasing diagonal line). Results were similar for the experiment on 17/04/2019.

4.3.3 Calling known antimalarial drug resistance mutations

I used the computational pipeline described in section 4.2.2 to search for evidence of antimalarial drug resistance mutations in both experiments. The combined results are displayed in Figure 4.16.

The pipeline searched for a total for 31 mutations in *Kelch13* which have been associated with artemisinin resistance (WHO 2018a). All mutations were at a low frequency (<10%) in the ten Zambian samples and two 3D7 control samples¹¹. As expected, we identified the R539T mutation at a high frequency (>80%) in IPC5202, the multidrug resistance strain from western Cambodia. Three mutations in *DHFR*, associated with resistance to pyrimethamine, have spread across Africa since the early 1960s (Roper et al. 2004). We found all three of these mutations at

¹¹Except for sample *K.Pf.01*. This sample had a very low DNA concentration (<1ng/uL) after sWGA.

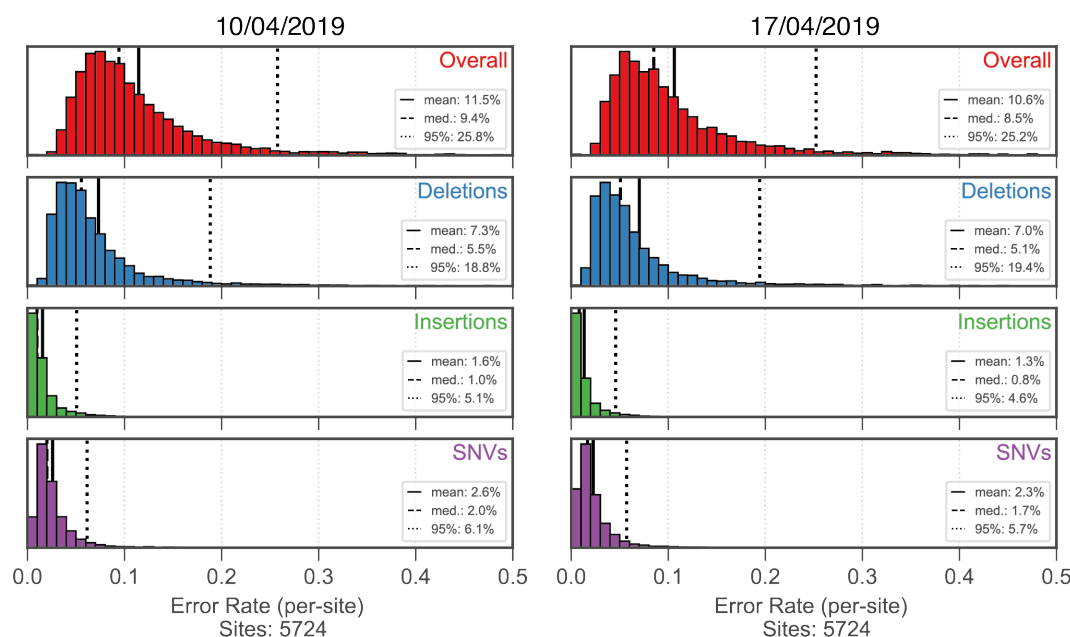


Figure 4.14: Read-level error analysis for both amplicon sequencing experiments. Distributions of overall, deletion, insertion, and SNV error, per-site at the read-level. Computed by comparison of 3D7 samples sequenced on the MinION with the 3D7 reference genome.

a high frequency in the Zambian samples and IPC5202 samples, but not in the 3D7 samples. Finally, we assayed for chloroquine resistance by looking for the K76T mutation in *CRT*. In both experiments, the IPC5202 sample displayed this mutation at a relatively high frequency (>50%). Interestingly, our data suggests that all of the Zambian samples are sensitive to chloroquine, as all display a low frequency of K76T. A study documenting the return of chloroquine sensitive *P.falciparum* to Zambia was published in 2016; 100% (302/302) of samples genotyped at the *CRT* K76 position were found to be wild-type (Mwanza et al. 2016).

It is notable that the mutation frequencies observed in the IPC5202 samples, at sites where IPC5202 is known to be mutant, are less than 100%. IPC5202 is monoclonal, and so deviation from 100% reflects only sequencing error. Error rates are expected to be higher on the amino acid level than the nucleotide level, as a correct codon call requires three correct nucleotide calls. It is also noteworthy that the two sites with the lowest mutation frequency (K76T in *CRT* and C59R in *DHFR*) both involve homopolymer sequences of length four¹². Nonetheless,

¹²In release 39 of the 3D7 reference genome, the sequence of K76 in *CRT* and the two flanking

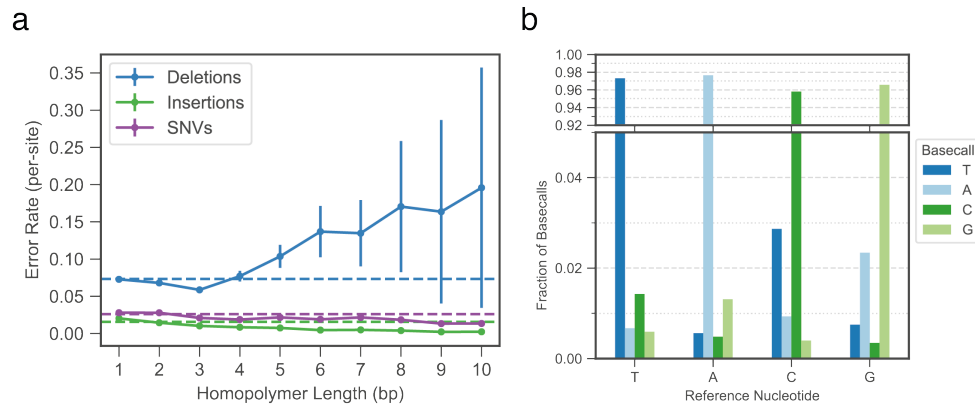


Figure 4.15: Read-level error bias for experiment on 10/04/2019. (a) Mean per-site error rate at the read-level for deletions, insertions, and SNVs, is shown as a function of the number of consecutive identical bases (homopolymer length). As homopolymers grow longer, sites tend to contain more deletion errors. Error bars give the 95% confidence interval. Horizontal dashed lines give mean across all sequenced sites. (b) SNV error is biased to incorrectly call purines as other purines, (e.g. A to G) or pyrimidines as other pyrimidines (e.g. C to T). Correct basecall is given on the x-axis and observed basecall, at the read-level, is given in color. Results are similar for the 17/04/2019 experiment.

the differences in mutation frequency between the IPC5202 samples and the 3D7 samples permitted the resistance status be called with confidence for all mutations we investigated in *Kelch13*, *DHFR*, and *CRT*.

4.3.4 Downsampling analysis

The two experiments described above involved six and eight samples, respectively. Estimates of the frequency of antimalarial drug resistance in a population are more accurate as more samples are collected. Moreover, the per-sample cost of sequencing declines if more samples are run on a single flow cell.

To explore the potential for increasing the number of samples per nanopore sequencing run, I performed a downsampling analysis with the data from the amplicon experiment on 10/04/2019 (Figure 4.17). The aim was to determine the minimum number of reads required for a high consensus sequence accuracy per sample. To achieve this, I randomly downsampled to a fixed number of reads

codons is AAT AAA ATT. The sequence of C59R in *DHFR* and the two flanking codons is TTT TGT GCA

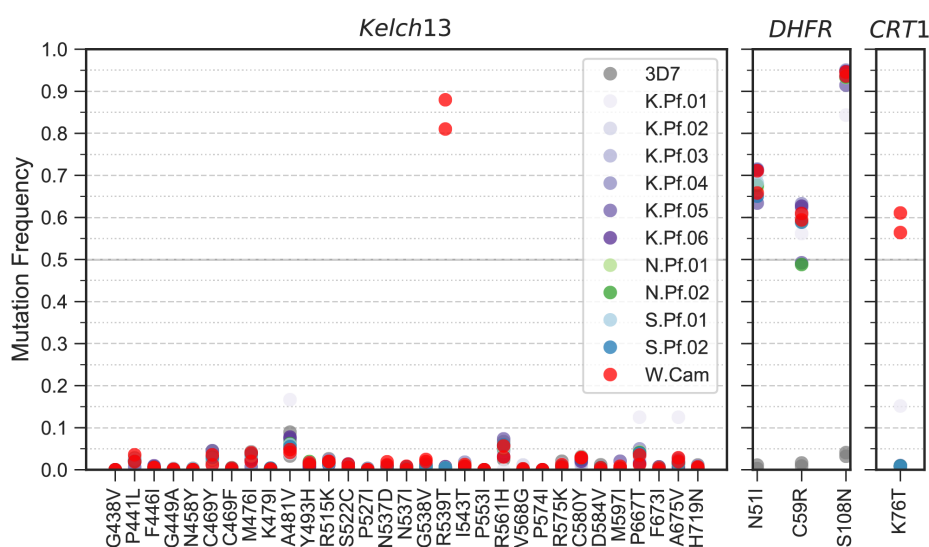


Figure 4.16: Assessing the resistance status of ten Zambian DBS samples to artemisinin, pyrimethamine and chloroquine. Left: frequency of a series of known artemisinin resistance mutations in the gene *Kelch13* for ten Zambian samples, a known-sensitive lab strain (3D7, grey) and a known resistant strain from Western Cambodia (W. Cam., red). Mutation frequency is calculated as the fraction of all reads at a given codon position which encode for the mutant amino acid. Center: Same as left, except querying for pyrimethamine resistance mutations in the gene *DHFR*. Right: Same as left, but querying for chloroquine resistance in *CRT*. Note data is pooled from 10/04/2019 and 17/04/2019 experiments. W.Cam is IPC5202.

(ranging from 10 to 2000, five times each), from the .SAM file containing all of the reads generated from the 3D7 control sample (Figure 4.17a). I then constructed a read pileup (using `samtools mpileup` as before) and computed consensus accuracy by comparing to the 3D7 reference genome. I found that with as few as 100 reads, a 100% consensus accuracy (excluding deletion errors) could be achieved across all four amplicons (Figure 4.17d). Given that both amplicon experiments produced in excess of millions of reads, this analysis suggests that a single R9.4.1 Flow Cell has the capacity to accurately sequence multiple amplicons across thousands of samples.

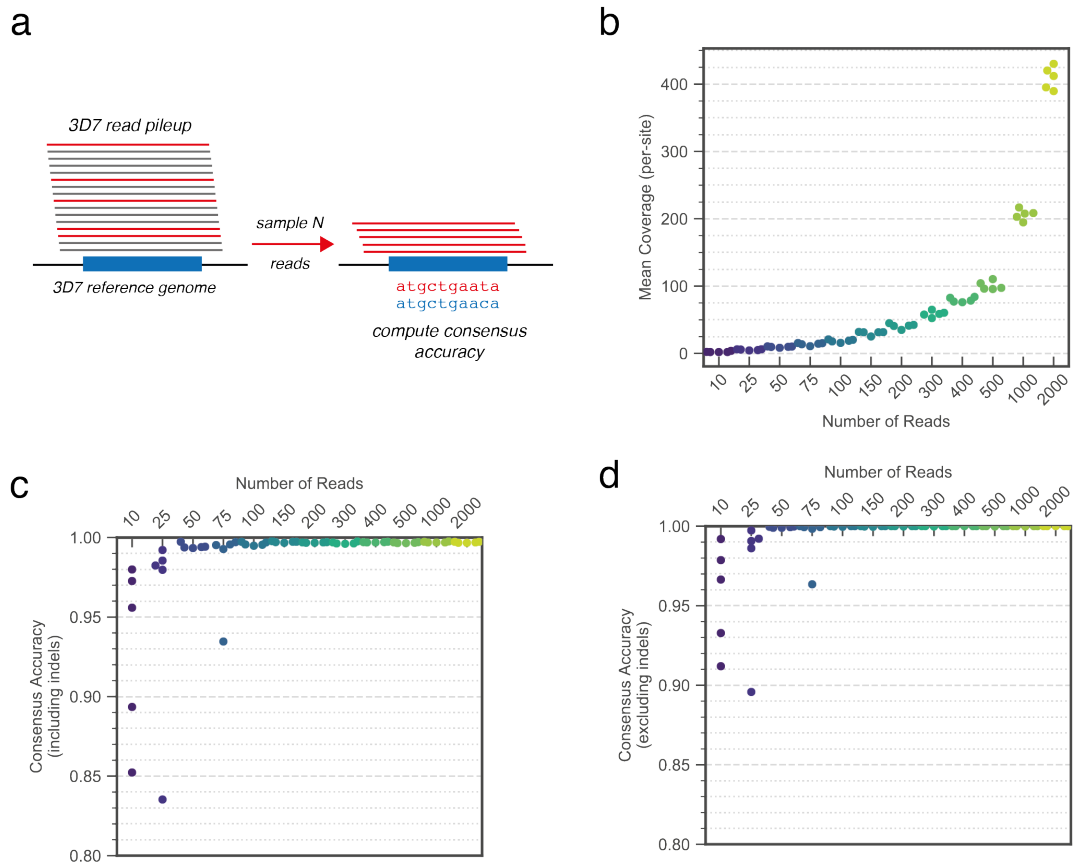


Figure 4.17: Determining throughput potential through downsampling (a) Schematic of experiment procedure. Reads found to map to 3D7 on the 10/04/2019 experiment were randomly downsampled, used to generate a pileup and compute consensus sequence accuracy. Note this was conducted for all four amplicons simultaneously. (b) Average per-site coverage after downsampling to a fixed number of reads. Swarmplot is coloured by number of reads in downsample; the five points with the same color represent randomly repeated downsamplings at a same number of reads. (c) Consensus accuracy following downsampling, including deletion errors. (d) Consensus accuracy following downsampling, excluding deletion errors.

4.4 Discussion

Artemisinin-resistance, resulting in high rates of ACT treatment failure, has spread across Southeast Asia. There is a possibility this resistance will migrate to Africa, as both chloroquine and sulfadoxine-pyrimethamine resistance historically began in Southeast Asia but are now pervasive across Africa (Wootton et al. 2002; Roper et al. 2004). To prepare for this possibility, large scale surveillance of antimalarial drug resistance is needed. Therapeutic-efficacy studies, however, are labour-intensive, demanding on patients, require a minimum 28-day follow-

up period, and are ultimately difficult to scale (WHO 2018a). On the other hand, genetic surveillance based on second-generation sequencing platforms often entails shipping samples out of their country-of-origin, limiting local ownership, adding logistical challenges, and increasing time-to-result. Here, we developed an assay for antimalarial drug resistance surveillance using nanopore sequencing. Working with scientists at the National Malaria Elimination Centre in Zambia, we deployed the assay twice, determining the drug resistance status of ten samples – to chloroquine, pyrimethamine, and artemisinin – in-country and in both cases in just under two days.

Nanopore sequencing has a higher error rate than second-generation sequencing platforms, but here we have demonstrated that it does not prevent antimalarial resistance detection. First, the inclusion of known resistant and sensitive samples allows for estimation of the error rate in each run, overall but also at specific genomic positions of interest. Second, though the overall read-level error rate was between 10.5-11.7%, the consensus accuracy of all genes exceeded 99.3%. Moreover, all errors at the consensus-level were deletions, and all consensus deletions occurred in long homopolymer tracts. There are well understood mechanistic reasons that cause nanopore sequencing to struggle with long homopolymer tracts (O'Donnell et al. 2013). Thus, despite a higher read-level error, the consensus error is low, predictable, and explicable; it has little impact on determining the drug resistance genotype of a sample.

The current computational pipeline searches for known drug resistance mutations at the read-level – computing frequencies of codon changes (such as K76T in *CRT*) within a read pileup. In several cases, as a consequence of mutations lying near homopolymer tracts, known mutant samples have mutation frequencies well below 100%. This approach was taken to account for the possibility of mixed infections, which make computing a single consensus sequence across a sample inappropriate. A better approach would be to leverage the long read data produced by nanopore sequencing to create a consensus sequence for each of the $K \geq 1$ strains present with the sample. In previous chapters, I have discussed how the distribution of

number of strains per sample – the complexity of infection (K) – is one of the most useful metrics for inferring regional prevalence. In the amplicon sequencing experiments described here, I included the hypervariable surface antigen gene *MSP2* to explore whether complexity of infection could be estimated with nanopore sequencing data. Indeed, hierarchical clustering of the reads mapping to the *MSP2* coding sequence based on sequence identity produced pairwise distance matrices suggestive of multiple strains in several of the Zambian samples, but not for the control samples (Figure 4.18). Determining complexity of infection from whole-genome sequencing data is possible, but methods often truncate the range of K due to a limited sequencing depth¹³. The greater sequencing depth afforded by an amplicon sequencing approach may support inferring higher values of K , allowing K distributions to be fit with more complex statistical models. Here, I have shown that a minimum of 100 reads is sufficient to generate a high-accuracy consensus sequence for a monoclonal sample (Figure 4.17). What depth of sequencing is required to infer the consensus sequences of K strains at high accuracy, given that some may be present at low proportions, is an important question for future research.

Malaria control is a resource-constrained activity¹⁴, and two factors impacting the practicality of using DNA sequencing for resistance surveillance are scalability and expense. First, scaling to sequence several hundreds of samples per R9.4.1 Flow Cell requires only developing additional barcodes, and optionally converting the individual laboratory steps into 96-well plate format. Second, the capital costs of nanopore sequencing are very low. As detailed in Table 4.2, our entire laboratory setup, sufficient for DNA extraction, library preparation, nanopore sequencing and real-time analysis cost well under £10,000. Reagent costs, however, are high – in the experiments conducted here they were approximately £60 per sample. These costs can be reduced with optimization. DNA extraction with the Qiagen DNA Investigator Kit costs about £5 per sample, but Chelex columns or phenol-chloroform

¹³For example, DEploid infers a maximum K of four (Zhu et al. 2018).

¹⁴The WHO Global Technical Strategy for Malaria 2016-2030 describes the greatest challenge facing malaria control as "...the lack of robust, predictable and sustained international and domestic financing." (WHO 2015)).

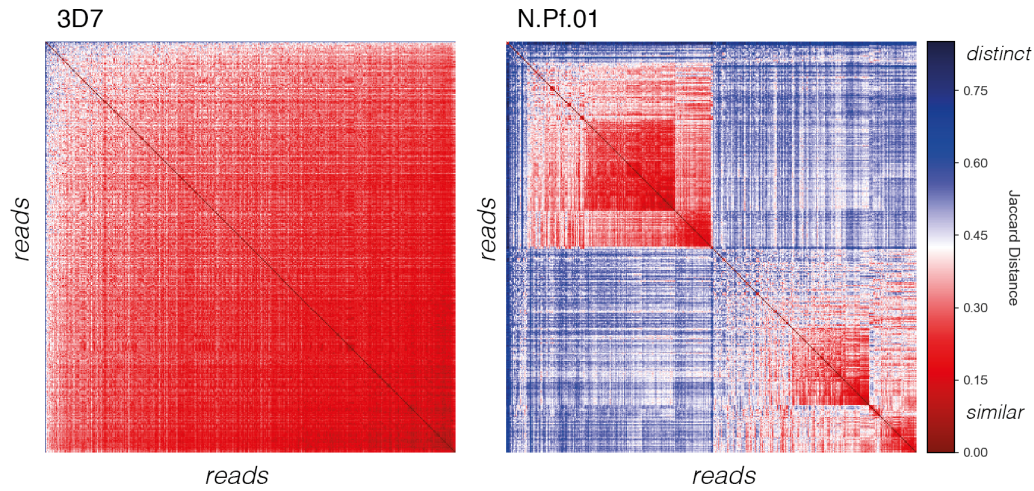


Figure 4.18: Pairwise similarity matrices for reads mapping to *MSP2* in the 3D7 and N.Pf.01 samples sequenced on 10/04/2019. Each matrix is calculated by taking a subset of 400 reads that span the entire coding sequencing of *MSP2* and clustering them based on pairwise Jaccard Distance; which here gives the fraction of non-identical sites between two sequences. Left matrix is for 3D7, and right matrix is for the Zambian sample N.Pf.01. Notice the structure in the N.Pf.01 matrix, suggestive of $K \geq 2$ strains.

solutions can be used to extract DNA at a much lower cost. PCR is in general not expensive ($<£1$ per reaction with the KAPA HiFi Kit), but multiplexing¹⁵ can be undertaken to further reduce costs and also labour requirements. In our pipeline, the enzymes used for ligation-based barcoding were the bulk of the expense (nearly £30 per sample), however with foreknowledge about the number of samples being prepared for a run, ligation reactions can be either downsized, or with optimization potentially supplanted with PCR-based barcoding. A single R9.4.1 Flow Cell costs approximately £500, but as demonstrated in Section 4.3.4, has the throughput to support the sequencing of thousands of samples. Moreover, the recently developed Nuclease Wash Kit (EXP-WSH003) permits multiple sequencing runs with a single R9.4.1 Flow Cell without the need for alternative sets of barcode sequences.

Though these optimizations would significantly reduce per-sample reagent costs, it may still be that they are not as low as those possible from industrial-scale sequencing facilities, which benefit from the economy of scale. An additional advantage of centralized sequencing is that it promotes large-scale standardization,

¹⁵Amplifying multiple target sequences in a single reaction, by including multiple primer sets.

which distributed and in-country sequencing may lack. However, a distributed approach to sequencing enables researchers and policy-makers living in malaria endemic countries, who have a deep knowledge of the local context – including its disease epidemiology, but also its economic, political and cultural landscape, to design and conduct suitable surveillance programs.

4.4.1 Why is artemisinin resistance not already in Africa?

A recent retrospective analysis identified 31 independent origins of artemisinin resistance in Southeast Asia, and yet beyond anecdotal reports, no resistance has been detected in Africa (Amato et al. 2018). At the same time, Africa accounted for 92% of the estimated 219 million cases of malaria that occurred globally in 2017 (WHO 2018b). Why has artemisinin resistance not already been detected in Africa?

In order for resistance to be detected at least two events must occur: (1) resistance mutations must arise and (2) they must experience selection sufficient to bring them to an appreciable frequency¹⁶. Are either of these events less likely in Africa than in Southeast Asia? The rate at which resistance mutations arise is proportional to the mutation rate multiplied by the effective population size; Africa has twenty-fold more malaria cases, and so its effective population size should be greater. Some have suggested that *P. falciparum* in Southeast Asia may have a higher mutation rate, but this hypothesis is unsupported by studies of substitution rates in field samples (Brown et al. 2015) and, in fact, *Kelch13* mutations have been detected at low-frequencies in Africa (Ashley et al. 2014; MalariaGEN 2016). If resistance mutations are arising regularly in Africa, then perhaps they experience weaker or less sustained selection. Africa has a much higher transmission intensity than Southeast Asia, which is expected to increase acquired immunity and the number of asymptomatic cases, thereby reducing treatment rates. It further causes higher rates of mixed infection, which induce competition between resistant and

¹⁶Drift alone is unlikely to result in antimalarial resistance mutations achieving high frequency, as currently documented resistance mutations often exhibit a deleterious fitness cost in the absence of drug pressure. Examples of this are the *CRT* K76T mutation disappearing from Zambia after chloroquine was discontinued, and *mdr1* amplifications declining in Cambodia after a switch to dihydroartemisinin piperaquine.

wild-type alleles, potentially in untreated or pre-treatment contexts. Though these effects would reduce selection, they can't unconditionally prohibit the spread of resistance, as chloroquine and sulfadoxine-pyrimethamine resistance swept across Africa when transmission rates were higher than they are today.

Other attempts at explanation evoke specific historical events or geographical features. Pailin has a busy gem-mining industry that draws non-immune migrant workers and high rates of malaria and malaria treatment (Packard 2014). In an attempt to increase treatment rates, chloroquine was introduced in low doses to cooking salt and distributed around Southeast Asia, but also Africa (Payne 1988). Artesunate monotherapy was available from private pharmacies in Southeast Asia, in an unregulated fashion, decades before it was recommended by WHO as a combination therapy (Dondorp et al. 2009). The first ACT combined artemisinin with mefloquine, to which resistance parasites already existed (Denis et al. 2006), and then artemisinin with piperaquine after artemisinin resistant parasites existed (Amaratunga et al. 2016).

The emergence of ACT treatment failure in one geography leads to the possibility of migration to another. Although chloroquine and sulfadoxine-pyrimethamine resistance migrated from Southeast Asia to Africa, the current landscape of malaria is very different than it was historically (Roper et al. 2004). Malaria transmission has declined globally (Bhatt et al. 2015). For example, there is much less malaria in India, a populous country which has a significant diaspora in East Africa¹⁷ (Kumar et al. 2007). At the same time, flourishing commercial air travel has vastly increased human mobility.

Ultimately, our uncertainty about how, when or whether artemisinin resistance will spread to Africa only underscores the importance of surveillance, as it points to our inability to forecast an event that could have dramatic public health consequences (Lubell et al. 2014). If a higher transmission intensity is insulating Africa against the emergence of resistance, then declining prevalence may result in artemisinin resistance outbreaks. Finally, it is worth noting that surveillance alone is not a

¹⁷Interestingly, chloroquine resistance was detected in India five years before it was detected in Africa (Kumar et al. 2007).

sufficient solution. Chloroquine resistance was detected on the eastern seaboard of Africa in 1978, and despite awareness of its presence, it spread across Africa over the next decade (Clarke 1979; Wernsdorfer and Payne 1991). A collaborative effort, where surveillance approaches are integrated with malaria control programs, and where new technologies are guided by local knowledge, is our best hope both for preventing the spread of ACT resistance and combating malaria more generally.

4.5 Ethical Approval

The Mobile Malaria Project had approval from the Oxford Tropical Research Ethics Committee (reference: 554-18) as well as the University of Zambia Biomedical Research Ethics Committee (reference: 008-01-19). We visited and were granted approval by the director of the Ndola District Health Office under the Ministry of Health for our work in Kaniki. All samples were collected with informed consent by our collaborators at NMEC by fingerprick. All data collected was given to our collaborators at NMEC.

5

Conclusions

Contents

5.1	Epidemiology and genetics	138
5.1.1	Model complexity and inference	139
5.2	Mixed infections and transmission	141
5.2.1	Statistical models of mixed infections	142
5.3	Collecting malaria genetic data	144
5.3.1	Data collection and collaboration in Africa	145

The overarching aim of this thesis has been to understand how malaria genetic data can contribute to malaria surveillance efforts – in particular, to the estimation of transmission intensity. In Chapter 2, we developed a genetic epidemiological model and implemented it as an agent-based simulation (**forward-dream**), allowing us to explore how malaria prevalence across hosts is correlated with the genetic diversity of the parasite population. We noted that statistics related to the frequency and complexity of mixed infections were strongly, robustly¹, and rapidly responsive to changes in prevalence. In Chapter 3, we analyzed mixed infections further, understanding how their IBD patterns relate to individual transmission events, again through the development of a simulation, this time of the process of malaria meiosis (**pf-meiosis**). In Chapter 4, we turned to the problem of how malaria genetic data

¹They responded regardless of the cause of prevalence change.

can be collected in a manner suitable to support malaria surveillance. We developed a nanopore sequencing approach for collecting malaria genetic data in-country and, as it is a pressing surveillance issue, used the approach for antimalarial drug resistance detection. In addition, in the discussion of Chapter 4 we provided preliminary data suggesting that our nanopore sequencing approach will also allow for the detection and analysis of mixed infections, thus linking to our previous two chapters.

To finish, we will discuss some of the more significant findings and outstanding problems that have arisen from this work. We also briefly describe a mathematical model of mixed infections and consider some of the barriers to increasing DNA sequencing capacity in Africa.

5.1 Epidemiology and genetics

Although inferring the demographic history of a population has been a longstanding aim of population genetics, the inference of epidemiological history from genetic data is a relatively new endeavour. As I have tried to illustrate, the latter problem is likely more difficult, as familiar models in population genetics fail to capture the complex life cycles typical of most pathogens (malaria included), and epidemiological models, such as the compartment based models, don't capture genetic processes. In Chapter 2 of this thesis we sought to bridge this gap using **forward-dream** – a merging of a continuous-time Moran model with a Ross-Macdonald epidemiological model. The major innovation was in limiting the size of parasite populations within individual hosts and vectors (to $k_{max}^h = 10$ and $k_{max}^v = 10$) – making the simulation of reasonable community sizes² feasible, while still supporting mixed infections and capturing the majority of intra-host genetic diversity that would be detected by sequencing. Simple Bernoulli sampling approaches to model transmission bottlenecks and permitting co- and super-infection were also important additions.

²Hundreds of hosts and thousands of vectors.

5.1.1 Model complexity and inference

As mentioned in the discussion, **forward-dream** involves considerable simplification, and omits many epidemiological phenomenon (e.g. acquired immunity and migration) that occur ubiquitously in natural populations. Moving forward, we face the age-old problem in statistics of choosing a suitable degree complexity for our models (in the spirit of Professor George Box's "All models are wrong, but some are useful." (Box 1976)). There are many examples of highly sophisticated epidemiological models of malaria – both at the population level, and also of some specific steps in the life cycle (Eckhoff 2011; Eckhoff 2012a; Eckhoff 2012b). In Daniels et al. (2015) – where estimates of R_0 were made from barcode data over an eight year period in Senegal – the approach taken was to incorporate a basic model of a genetic barcode into a "...simplification of the malaria-transmission dynamics in an existing agent-based model (Eckhoff 2012b)". Though the nature of the simplification is not described, the model in Eckhoff (2012b) is substantially more complicated than **forward-dream**, including features such as acquired immunity, an explicit representation of the mosquito and parasite life-cycle including various life-cycle stages, and environmental features such as rainfall and temperature. Although all of these process are important and moreover, reasonably well-characterized, their parameterisation involves a large number of assumptions and/or approximations. Even parameters with values derived from empirical studies may vary across environments. An important question arises: are simple models with more explicit assumptions and omissions sufficient, or does the problem at hand require with more complicated models, notwithstanding the difficulty of parameterisation?

Another major issue facing future work will be how to do inference from the models that are created. We were able to derive the expected time to coalescence in the **forward-dream** model, but we have not been able to derive a likelihood over parasite prevalence, or any other epidemiological parameter. Ultimately, it is clear that even for simple models analytical solutions will not be forthcoming. As a result, for most models computational approaches to inference will be either favoured or necessary. One popular approach, which does not require derivation of the

likelihood, is Approximate Bayesian Computation (ABC) (reviewed in Lintusaari et al. (2016)). ABC is a form of rejection algorithm where simulations are run under a variety of different parameter values, typically randomly drawn from a set of prior distributions. Those simulations that produce data sufficiently close to the observed data are used to construct posterior distributions over the parameters. Determining the distance between the simulated data and the observed data is done by computing the same set of summary statistics over both, constructing a composite distance metric from those summary statistics, and then requiring it be less than some constant value for acceptance. This approach was used in Daniels et al. (2015), and also to estimate R_0 and other transmission-related parameters of *Mycobacterium tuberculosis* from genetic data (Tanaka et al. 2006). Although clearly powerful, ABC is controversial due to several drawbacks, with critics pointing to the arbitrariness and information loss associated with using summary statistics, and the arbitrariness of defining the size of the acceptance region (Lintusaari et al. 2016). Summary statistics that are not sufficient do not reproduce the posterior distribution that would be generated by the full data, and there may be a lack of sufficient statistics for the parameters of interest (Csilléry et al. 2010). Moreover, the question arises of what parameters of the simulation model we should seek to fit using ABC: should we use genetic data to fit all the parameters in the model? If we seek to fit too many parameters, ABC becomes immensely computational expensive. Or should we set some selection of parameters to fixed constant values, where we again face the problem of parameterisation.

In Chapter 2, we choose not to attempt direct inference using **forward-dream**, but rather used our simulation to explore whether simple relationships existed that could justify and/or guide less complex inference approaches with real data. It is possible that many of the relationships we are interested in are well approximated as linear, in which case using sophisticated simulations for inference is unnecessary. Here, again, we used summary statistics of genetic diversity, which entails information loss. Moreover, we found that individual statistics were often noisy and lacked robustness. Multivariate models, using multiple genetic diversity statistics

to estimate prevalence, could be helpful; however we noted that many statistics are significantly co-linear. Of course we were not exhaustive in our selection of statistics, and many other statistics can be developed and employed³. Statistics related to linkage disequilibrium may be particularly informative, as the rate of recombination is linked to the rate of outcrossing, which is a function of the frequency of mixed infections (Anderson et al. 2000). We note that regardless of the what statistics are used, their variance, robustness, and timeliness will always have to be considered carefully, and we hope the approach presented in Chapter 2 can provide a framework for such investigations.

A final point regarding inference is that genetic data should not be used in isolation, but combined with other sources of information to improve transmission estimates. As an example of this, The Malaria Atlas Project combines clinical incidence data with a suite of co-variables in order to make prevalence estimates (Weiss et al. 2015; Bhatt et al. 2015). Future inference methods utilizing genetic data should also be able to incorporate other relevant sources of data, either as priors or as co-variables.

5.2 Mixed infections and transmission

One conclusion of this thesis has been that mixed infections may be especially informative for the genetic surveillance of malaria. Using **forward-dream**, I was able to show that amongst a suite of other genetic diversity statistics those related to mixed infections were strongly, robustly, and rapidly correlated with parasite prevalence. Notably, a variety of methods already exist to detect and estimate the complexity of malaria mixed infections (reviewed in Chapter 3); producing distributions of the complexity of infection from patient samples on a yearly basis is achievable now, and moreover does not require whole-genome sequencing. Newer methods such as **DEploidIBD** can determine the IBD structure of strains within a mixed infection, and another important finding was that IBD statistics can be used to infer the nature of individual transmission events. In particular, I was able to

³As was done by Daniels et al. (2015) in their ABC approach.

show that one can differentiate between mixed infections generated by co-infection or super-infection using IBD. Evidence for "mixed infection chains" was also present in these IBD patterns, and their more frequent occurrence in Southeast Asia may be related to the lower transmission intensity there.

Jointly considered, these observations suggest that inferring transmission intensity with genetic data could be fruitfully pursued by building statistical models that are focused on mixed infections. Below I present some preliminary work in this direction.

5.2.1 Statistical models of mixed infections

It is natural to model the infection of an individual host (or vector) as a continuous-time Markov process, where states correspond to the complexity of infection, K . In such a process, transitions between states that increase K represent infection events, and those that decrease K represent clearance events. Starting simply, we apply the following assumptions:

- The probability of a host becoming infected is a constant λ , independent of K
- A host can only be infected by a single strain simultaneously (i.e., co-infection does not occur)
- The probability of a host clearing an infection is a constant γ , independent of K
- A host clears all K parasite strains simultaneously (i.e., through complete cross-immunity or treatment)
- K has no maximum $K \in [0, 1, 2, 3, \dots]$

Such a chain is shown diagrammatically in Figure 5.1a. The embedded jump-chain has the following transition probabilities:

$$P_{K,K+1} = \lambda/(\lambda + \gamma)$$

$$P_{K,0} = \gamma/(\lambda + \gamma)$$

All other transitions have a probability of zero. This Markov process has several convenient properties. First, note that λ and γ are definitionally equivalent to the same parameters in the Ross-Macdonald model (see Section 2.2.1). Following from this, it can be shown that:

$$P_{K,K+1} = \lambda/(\lambda + \gamma) = p$$

where here p represents the host parasite prevalence in the Ross-Macdonald model (X_h in Chapter 2). Second, the stationary distribution of this Markov process is easily obtained:

$$P_K(K = k) = (1 - p)p^k$$

which is geometric with an expectation $E[K] = p/(1 - p)$. Thus, the model predicts that the mean complexity of infection climbs non-linearly with host parasite prevalence. Interestingly, in this model, the rate of mixed infection (amongst collected samples, i.e. individuals with malaria) is equal to the prevalence:

$$P(K > 1 | K > 0) = P(K > 1)/P(K > 0) = p$$

Importantly, the simplicity of this model (in particular the fact that the only parameter is the parasite prevalence) makes the maximum likelihood estimate of p easy to derive:

$$\hat{p} = \bar{k}/(\bar{k} + 1)$$

where $\hat{p}$ is the estimate of the prevalence and $\bar{k}$ is the mean complexity of infection in the observed data.

Although this model overly simplistic, it can be built from in a straightforward fashion. Figure 5.1 shows several other models, which incorporate co-infection or partial clearance of parasites. Developing these further and applying them to existing data sets like the Pf3k project will be part of my future work.

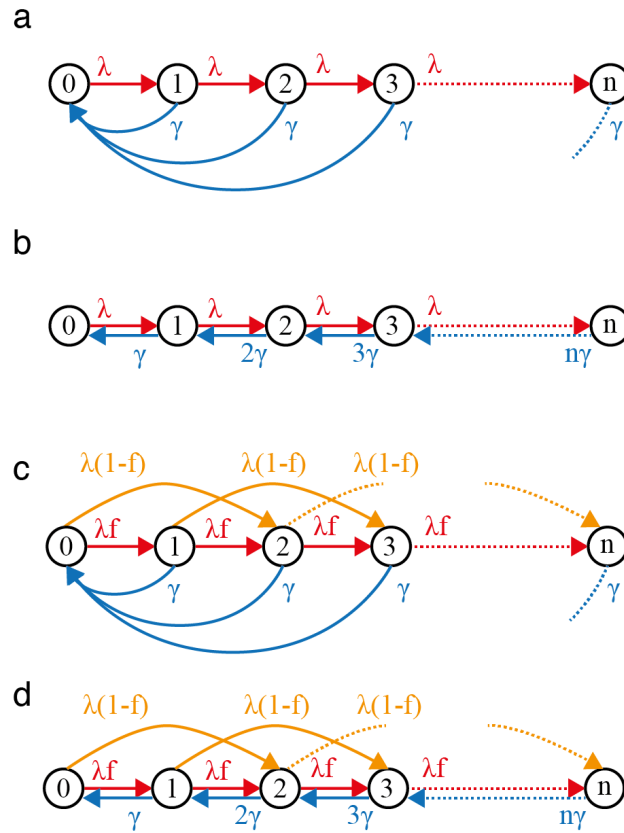


Figure 5.1: Markov processes can be used to model complexity of infection. (a-d) Schematics of a Markov processes representing the complexity of infection of an individual host (or vector) under differing assumptions. (a) No co-infection, parasite clearance is complete. (b) No co-infection, different parasites strains are cleared individually. (c) Co-infection from K to $K + 2$ is possible and occurs with probability $1 - f$, where $0 \leq f \leq 1$; parasite clearance is complete. (d) Same as (c) but different parasite strains are cleared individually.

5.3 Collecting malaria genetic data

There are at least two reasons for which those involved in developing statistical models for malaria surveillance should endeavour to understand the data collection process. First, without adequate data a statistical model cannot be used for inference and therefore cannot support malaria surveillance. Thus, considering the amount and type of data that can (or could) be realistically collected should be a critical step in the design of statistical models meant to be operationally useful. Second, a thorough understanding of the data generation process gives a

modeller insight into what signals are likely representative of underlying biology, and what signals may be technical artifacts⁴; ultimately supporting the development of models that are more robust. My involvement in the Mobile Malaria Project, which allowed me to participate in the data generation process, was motivated in part by these considerations.

5.3.1 Data collection and collaboration in Africa

Malaria no longer occurs in Europe or North America, and so malaria research conducted on those continents is inherently detached from the daily realities of malaria control. At present, the predominant paradigm for malaria genetic research involves the shipment of samples out of malaria endemic countries, a practice which is suboptimal for reasons I outlined in Chapter 4. A primary reason for this is that the capital and maintenance costs of Second-generation sequencing platforms creates a prohibitively high barrier-to-entry for African research institutions. This barrier was reduced with the launch of the MinION in 2015 (Loman and Watson 2015), but we have yet to witness the widespread use of nanopore sequencing for malaria research in Africa.

Another potential barrier is the expertise requirements associated with DNA sequencing. In both Zambia and Kenya, George Busby, Isaac Ghanai and myself ran multi-day nanopore sequencing workshops. The scientists we interacted with there were more than capable of understanding and executing all of the laboratory protocols associated with nanopore sequencing. Indeed, the NMEC team in Zambia ran a nanopore sequencing experiment independently only a few weeks after our departure, and Dr. Eric Ochomo's team is currently designing primers to support an *Anopheles* targeted nanopore sequencing assay in the near future. In contrast, teaching data analysis skills during our workshops was more challenging. In my experience this issue is not unique to the African context; universally, laboratory scientists struggle to learn computational approaches, typically because their

⁴As an extreme example, I have known a bioinformatician who puzzled about a "strange trend" repeating every 384 samples, oblivious to the fact that the assay used to generate the data was conducted on 384-well plates.

undergraduate education has not included courses with a statistical or computational focus. Online-learning tools, such as MIT OpenCourseware (ocw.mit.edu/index.htm) or codecademy (www.codecademy.com), allow for individuals to engage in the long-term learning necessary to gain expertise in computational analyses, without additional formal education. These online-learning tools are available in lower-income countries, however, the computational infrastructure required to actually conduct analyses is often lacking. For nanopore sequencing, a single laptop with a 512Gb SSD and 16GB of RAM is sufficient to run the MinION; amongst the approximately 30 students we worked with in Zambia and Kenya, none had a laptop that met this requirement. A simple way to address this is to ensure that research funding that is awarded to international collaborations includes provisions for purchasing computational infrastructure for collaborators based in Africa.

Scientists from malaria endemic countries possess a great understanding of local malaria biology, epidemiology, and the challenges facing malaria control; scientists from outside these countries typically have more experience with modern genetic technologies and data analysis. This creates a great opportunity for collaboration and knowledge exchange to the benefit of both malaria control and science.

References

- Aly, A. S., A. M. Vaughan, and S. H. Kappe (2009). “Malaria parasite development in the mosquito and infection of the mammalian host”. In: *Annu Rev Microbiol* 63, pp. 195–221.
- Amambua-Ngwa, A. et al. (2019). “Major subpopulations of Plasmodium falciparum in sub-Saharan Africa”. In: *Science* 365.6455, pp. 813–816.
- Amaratunga, C. et al. (2016). “Dihydroartemisinin-piperaquine resistance in Plasmodium falciparum malaria in Cambodia: a multisite prospective cohort study”. In: *Lancet Infect Dis* 16.3, pp. 357–65.
- Amato, R. et al. (2018). “Origins of the current outbreak of multidrug-resistant malaria in southeast Asia: a retrospective genetic study”. In: *Lancet Infect Dis* 18.3, pp. 337–345.
- Anderson, T. J. et al. (2000). “Microsatellite markers reveal a spectrum of population structures in the malaria parasite Plasmodium falciparum”. In: *Mol Biol Evol* 17.10, pp. 1467–82.
- Ariey, F. et al. (2014). “A molecular marker of artemisinin-resistant Plasmodium falciparum malaria”. In: *Nature* 505.7481, pp. 50–5.
- Ashley, E. A. et al. (2014). “Spread of artemisinin resistance in Plasmodium falciparum malaria”. In: *N Engl J Med* 371.5, pp. 411–23.
- Band, Gavin et al. (2019). “Insights into malaria susceptibility using genome-wide data on 17,000 individuals from Africa, Asia and Oceania”. In: *Nature Communications* 10.1, p. 5732. URL: <https://doi.org/10.1038/s41467-019-13480-z>.
- Bartoloni, Alessandro and Lorenzo Zammarchi (2012). “Clinical aspects of uncomplicated and severe malaria”. In: *Mediterranean journal of hematology and infectious diseases* 4.1, e2012026–e2012026. URL: <https://www.ncbi.nlm.nih.gov/pubmed/22708041><https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3375727/>.
- Bates, M. (1970). *The Natural History of Mosquitoes*. Peter Smith.
- Beier, J. C. (1998). “Malaria parasite development in mosquitoes”. In: *Annu Rev Entomol* 43, pp. 519–43.
- Beier, J. C. et al. (1991). “Quantitation of malaria sporozoites in the salivary glands of wild Afrotropical Anopheles”. In: *Med Vet Entomol* 5.1, pp. 63–70.
- Beier, J. C. et al. (1994). “Plasmodium falciparum incidence relative to entomologic inoculation rates at a site proposed for testing malaria vaccines in western Kenya”. In: *Am J Trop Med Hyg* 50.5, pp. 529–36.
- Bennink, S., M. J. Kiesow, and G. Pradel (2016). “The development of malaria parasites in the mosquito midgut”. In: *Cell Microbiol* 18.7, pp. 905–18.
- Bhatt, S. et al. (2015). “The effect of malaria control on Plasmodium falciparum in Africa between 2000 and 2015”. In: *Nature* 526.7572, pp. 207–211. URL: <https://www.ncbi.nlm.nih.gov/pubmed/26375008><https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4820050/>.

- Blanco, L. and M. Salas (1984). "Characterization and purification of a phage phi 29-encoded DNA polymerase required for the initiation of replication". In: *Proc Natl Acad Sci U S A* 81.17, pp. 5325–9.
- Bloand, P.B. and World Health Organization (2001). *Drug resistance in malaria*. Ed. by. URL: <https://apps.who.int/iris/handle/10665/66847>.
- Bolad, A. et al. (2005). "Distinct interethnic differences in immunoglobulin G class/subclass and immunoglobulin M antibody responses to malaria antigens but not in immunoglobulin G responses to nonmalarial antigens in sympatric tribes living in West Africa". In: *Scand J Immunol* 61.4, pp. 380–6.
- Bousema, T. and C. Drakeley (2011). "Epidemiology and infectivity of *Plasmodium falciparum* and *Plasmodium vivax* gametocytes in relation to malaria control and elimination". In: *Clin Microbiol Rev* 24.2, pp. 377–410.
- Box, George E. P. (1976). "Science and Statistics". In: *Journal of the American Statistical Association* 71.356, pp. 791–799. URL: <https://www.tandfonline.com/doi/abs/10.1080/01621459.1976.10480949>.
- Boyce, R. M. et al. (2018). "Reuse of malaria rapid diagnostic tests for amplicon deep sequencing to estimate *Plasmodium falciparum* transmission intensity in western Uganda". In: *Sci Rep* 8.1, p. 10159.
- Brooks, D. R. et al. (1994). "Sequence variation of the hydroxymethyldihydropterin pyrophosphokinase: dihydropteroate synthase gene in lines of the human malaria parasite, *Plasmodium falciparum*, with differing resistance to sulfadoxine". In: *Eur J Biochem* 224.2, pp. 397–405.
- Brown, T. S. et al. (2015). "*Plasmodium falciparum* field isolates from areas of repeated emergence of drug resistant malaria show no evidence of hypermutator phenotype". In: *Infect Genet Evol* 30, pp. 318–322.
- Browning, S. R. and B. L. Browning (2012). "Identity by descent between distant relatives: detection and applications". In: *Annu Rev Genet* 46, pp. 617–33.
- Chang, H. H. et al. (2013). "Malaria life cycle intensifies both natural selection and random genetic drift". In: *Proceedings of the National Academy of Sciences* 110.50, p. 20129. URL: <http://www.pnas.org/content/110/50/20129.abstract>.
- Chang, H. H. et al. (2017). "THE REAL McCOIL: A method for the concurrent estimation of the complexity of infection and SNP allele frequency for malaria parasites". In: *PLoS Comput Biol* 13.1, e1005348.
- Clarke, V. D. (1979). "Chloroquine-resistant malaria acquired in Kenya and Tanzania–Denmark, Georgia, New York". In: *Cent Afr J Med* 25.2, pp. 39–40.
- Coelho, Camila Henriques et al. (2017). "Advances in malaria vaccine development: report from the 2017 malaria vaccine symposium". In: *NPJ vaccines* 2, pp. 34–34. URL: <https://www.ncbi.nlm.nih.gov/pubmed/29522056><https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5709382/>.
- Collins, F. H. et al. (1984). "First field trial of an immunoradiometric assay for the detection of malaria sporozoites in mosquitoes". In: *Am J Trop Med Hyg* 33.4, pp. 538–43.
- Cook, J. et al. (2011). "Serological markers suggest heterogeneity of effectiveness of malaria control interventions on Bioko Island, equatorial Guinea". In: *PLoS One* 6.9, e25137.
- Corran, P. et al. (2007). "Serology: a robust indicator of malaria transmission intensity?" In: *Trends Parasitol* 23.12, pp. 575–82.

- Cowman, A. F., D. Berry, and J. Baum (2012). “The cellular and molecular basis for malaria parasite invasion of the human red blood cell”. In: *J Cell Biol* 198.6, pp. 961–71.
- Cowman, A. F. and B. S. Crabb (2006). “Invasion of red blood cells by malaria parasites”. In: *Cell* 124.4, pp. 755–66.
- Cox, Francis E. G. (2010). “History of the discovery of the malaria parasites and their vectors”. In: *Parasites Vectors* 3.1, p. 5. URL: <https://doi.org/10.1186/1756-3305-3-5>.
- Craig, M. H., R. W. Snow, and D. le Sueur (1999). “A climate-based distribution model of malaria transmission in sub-Saharan Africa”. In: *Parasitol Today* 15.3, pp. 105–11.
- Csilléry, Katalin et al. (2010). “Approximate Bayesian Computation (ABC) in practice”. In: *Trends in Ecology Evolution* 25.7, pp. 410–418. URL: <http://www.sciencedirect.com/science/article/pii/S0169534710000662>.
- Dalmat, Ronit et al. (2019). “Use cases for genetic epidemiology in malaria elimination”. In: *Malaria Journal* 18.1, p. 163. URL: <https://doi.org/10.1186/s12936-019-2784-0>.
- Daniels, R. et al. (2008). “A general SNP-based molecular barcode for *Plasmodium falciparum* identification and tracking”. In: *Malar J* 7, p. 223.
- Daniels, R. et al. (2013). “Genetic surveillance detects both clonal and epidemic transmission of malaria following enhanced intervention in Senegal”. In: *PLoS One* 8.4, e60780.
- Daniels, R. et al. (2015). “Modeling malaria genomics reveals transmission decline and rebound in Senegal”. In: *Proc Natl Acad Sci U S A* 112.22, pp. 7067–72.
- Davidson, G. (1954). “Estimation of the Survival-Rate of Anopheline Mosquitoes in Nature”. In: *Nature* 174.4434, pp. 792–793. URL: <https://doi.org/10.1038/174792a0>.
- Denis, M. B. et al. (2006). “Surveillance of the efficacy of artesunate and mefloquine combination for the treatment of uncomplicated *falciparum* malaria in Cambodia”. In: *Trop Med Int Health* 11.9, pp. 1360–6.
- Dietz, K. (1993). “The estimation of the basic reproduction number for infectious diseases”. In: *Stat Methods Med Res* 2.1, pp. 23–41.
- Dolo, A. et al. (2005). “Difference in susceptibility to malaria between two sympatric ethnic groups in Mali”. In: *Am J Trop Med Hyg* 72.3, pp. 243–8.
- Dondorp, Arjen M. et al. (2009). “Artemisinin Resistance in *Plasmodium falciparum* Malaria”. In: 361.5, pp. 455–467. URL: <https://www.nejm.org/doi/full/10.1056/NEJMoa0808859>.
- Drakeley, C. and J. Cook (2009). “Chapter 5. Potential contribution of sero-epidemiological analysis for monitoring malaria control and elimination: historical and current perspectives”. In: *Adv Parasitol* 69, pp. 299–352.
- Echeverry, Diego F. et al. (2013). “Long term persistence of clonal malaria parasite *Plasmodium falciparum* lineages in the Colombian Pacific region”. In: *BMC Genetics* 14.1, p. 2. URL: <https://doi.org/10.1186/1471-2156-14-2>.
- Eckhoff, P. A. (2011). “A malaria transmission-directed model of mosquito life cycle and ecology”. In: *Malar J* 10, p. 303.
- (2012a). “*P. falciparum* infection durations and infectiousness are shaped by antigenic variation and innate and adaptive host immunity in a mathematical model”. In: *PLoS One* 7.9, e44950.

- Eckhoff, Philip A. (2012b). “Malaria parasite diversity and transmission intensity affect development of parasitological immunity in a mathematical model”. In: *Malaria Journal* 11.1, p. 419. URL: <https://doi.org/10.1186/1475-2875-11-419>.
- Elandt-Johnson, R. C. (1975). “Definition of rates: some remarks on their use and misuse”. In: *Am J Epidemiol* 102.4, pp. 267–71.
- Escalante, A. A. and F. J. Ayala (1994). “Phylogeny of the malarial genus *Plasmodium*, derived from rRNA gene sequences”. In: *Proceedings of the National Academy of Sciences of the United States of America* 91.24, pp. 11373–11377. URL: <https://www.ncbi.nlm.nih.gov/pubmed/7972067><https://www.ncbi.nlm.nih.gov/pmc/articles/PMC45233/>.
- Faria, N. R. et al. (2017). “Establishment and cryptic transmission of Zika virus in Brazil and the Americas”. In: *Nature* 546.7658, pp. 406–410.
- Fidock, D. A. et al. (2000). “Mutations in the *P. falciparum* digestive vacuole transmembrane protein PfCRT and evidence for their role in chloroquine resistance”. In: *Mol Cell* 6.4, pp. 861–71.
- Francis, S. E., Jr. Sullivan D. J., and D. E. Goldberg (1997). “Hemoglobin metabolism in the malaria parasite *Plasmodium falciparum*”. In: *Annu Rev Microbiol* 51, pp. 97–123.
- Galinsky, K. et al. (2015). “COIL: a methodology for evaluating malarial complexity of infection using likelihood from single nucleotide polymorphism data”. In: *Malar J* 14, p. 4.
- Gardner, M. J. et al. (2002). “Genome sequence of the human malaria parasite *Plasmodium falciparum*”. In: *Nature* 419.6906, pp. 498–511.
- Gething, P. W., A. P. Patil, and S. I. Hay (2010). “Quantifying aggregated uncertainty in *Plasmodium falciparum* malaria prevalence and populations at risk via efficient space-time geostatistical joint simulation”. In: *PLoS Comput Biol* 6.4, e1000724.
- Gething, P. W. et al. (2011). “A new world malaria map: *Plasmodium falciparum* endemicity in 2010”. In: *Malar J* 10, p. 378.
- Ghosh, A., M. J. Edwards, and M. Jacobs-Lorena (2000). “The journey of the malaria parasite in the mosquito: hopes for the new century”. In: *Parasitol Today* 16.5, pp. 196–201.
- Gimnig, John E. et al. (2001). “Characteristics of Larval Anopheline (Diptera: Culicidae) Habitats in Western Kenya”. In: *Journal of Medical Entomology* 38.2, pp. 282–288. URL: <https://doi.org/10.1603/0022-2585-38.2.282>.
- Githeko, Andrew K. et al. (2006). “Topography and malaria transmission heterogeneity in western Kenya highlands: prospects for focal vector control”. In: *Malaria Journal* 5.1, p. 107. URL: <https://doi.org/10.1186/1475-2875-5-107>.
- Guerra, C. A. et al. (2007). “Assembling a global database of malaria parasite prevalence for the Malaria Atlas Project”. In: *Malar J* 6, p. 17.
- Gutenkunst, Ryan N. et al. (2009). “Inferring the Joint Demographic History of Multiple Populations from Multidimensional SNP Frequency Data”. In: *PLOS Genetics* 5.10, e1000695. URL: <https://doi.org/10.1371/journal.pgen.1000695>.
- Hamilton, W. L. et al. (2017). “Extreme mutation bias and high AT content in *Plasmodium falciparum*”. In: *Nucleic Acids Res* 45.4, pp. 1889–1901.
- Hamilton, W. L. et al. (2019). “Evolution and expansion of multidrug-resistant malaria in southeast Asia: a genomic epidemiology study”. In: *Lancet Infect Dis* 19.9, pp. 943–951.

- Hanboonkunupakarn, Borimas and Nicholas J. White (2016). "The threat of antimalarial drug resistance". In: *Tropical Diseases, Travel Medicine and Vaccines* 2.1, p. 10. URL: <https://doi.org/10.1186/s40794-016-0027-8>.
- HapMap (2003). "The International HapMap Project". In: *Nature* 426.6968, pp. 789–96.
- Harbach, R.E. (2013). *Genus Anopheles Meigen, 1818. Mosquito Taxonomic Inventory*. URL: <http://mosquito-taxonomic-inventory.info/genus-anopheles-meigen-1818> (visited on 01/10/2020).
- Harinasuta, T., P. Suntharasamai, and C. Viravan (1965). "Chloroquine-resistant falciparum malaria in Thailand". In: *Lancet* 2.7414, pp. 657–60.
- Hastings, Alan and Susan Harrison (1994). "METAPOPULATION DYNAMICS AND GENETICS". In: *Annual Review of Ecology and Systematics* 25.1, pp. 167–188. URL: <https://doi.org/10.1146/annurev.es.25.110194.001123>.
- Hay, S. I., D. L. Smith, and R. W. Snow (2008). "Measuring malaria endemicity from intense to interrupted transmission". In: *Lancet Infect Dis* 8.6, pp. 369–78.
- Hay, S. I. et al. (2000). "Annual Plasmodium falciparum entomological inoculation rates (EIR) across Africa: literature survey, Internet access and review". In: *Trans R Soc Trop Med Hyg* 94.2, pp. 113–27.
- Hay, S. I. et al. (2004). "The global distribution and population at risk of malaria: past, present, and future". In: *Lancet Infect Dis* 4.6, pp. 327–36.
- Hay, S. I. et al. (2009). "A world malaria map: Plasmodium falciparum endemicity in 2007". In: *PLoS Med* 6.3, e1000048.
- Hay, Simon I. et al. (2010). "Developing global maps of the dominant anopheles vectors of human malaria". In: *PLoS medicine* 7.2, e1000209–e1000209. URL: <https://www.ncbi.nlm.nih.gov/pubmed/20161718> <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2817710/>.
- Huber, John H. et al. (2016). "Quantitative, model-based estimates of variability in the generation and serial intervals of Plasmodium falciparum malaria". In: *Malaria Journal* 15.1, p. 490. URL: <https://doi.org/10.1186/s12936-016-1537-6>.
- Imwong, M. et al. (2017). "The spread of artemisinin-resistant Plasmodium falciparum in the Greater Mekong subregion: a molecular epidemiology observational study". In: *Lancet Infect Dis* 17.5, pp. 491–497.
- Ishengoma, D. S. et al. (2019). "Deployment and utilization of next-generation sequencing of Plasmodium falciparum to guide anti-malarial drug policy decisions in sub-Saharan Africa: opportunities and challenges". In: *Malar J* 18.1, p. 267.
- Jain, M. et al. (2016). "The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community". In: *Genome Biol* 17.1, p. 239.
- James, Stephanie et al. (2018). "Pathway to Deployment of Gene Drive Mosquitoes as a Potential Biocontrol Tool for Elimination of Malaria in Sub-Saharan Africa: Recommendations of a Scientific Working Group(†)". In: *The American journal of tropical medicine and hygiene* 98.6suppl, pp. 1–49. URL: <https://www.ncbi.nlm.nih.gov/pubmed/29882508> <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5993454/>.
- Janse, C. J. et al. (1986). "DNA synthesis in Plasmodium berghei during asexual and sexual development". In: *Mol Biochem Parasitol* 20.2, pp. 173–82.
- Jin, Yamei, Chahnaz Kebaier, and Jerome Vanderberg (2007). "Direct Microscopic Quantification of Dynamics of Plasmodium berghei Sporozoite

- Transmission from Mosquitoes to Mice”. In: *Infection and Immunity* 75.11, p. 5532. URL: <http://iai.asm.org/content/75/11/5532.abstract>.
- Kilama, Maxwell et al. (2014). “Estimating the annual entomological inoculation rate for *Plasmodium falciparum* transmitted by *Anopheles gambiae* s.l. using three sampling methods in three sites in Uganda”. In: *Malaria Journal* 13.1, p. 111. URL: <https://doi.org/10.1186/1475-2875-13-111>.
- Kingman, J. F. C. (1982). “The coalescent”. In: *Stochastic Processes and their Applications* 13.3, pp. 235–248. URL: <http://www.sciencedirect.com/science/article/pii/0304414982900114>.
- Krief, S. et al. (2010). “On the diversity of malaria parasites in African apes and the origin of *Plasmodium falciparum* from Bonobos”. In: *PLoS Pathog* 6.2, e1000765.
- Kumar, A. et al. (2007). “Burden of malaria in India: retrospective and prospective view”. In: *Am J Trop Med Hyg* 77.6 Suppl, pp. 69–78.
- Leang, R. et al. (2013). “Efficacy of dihydroartemisinin-piperaquine for treatment of uncomplicated *Plasmodium falciparum* and *Plasmodium vivax* in Cambodia, 2008 to 2010”. In: *Antimicrob Agents Chemother* 57.2, pp. 818–26.
- Li, H. (2018). “Minimap2: pairwise alignment for nucleotide sequences”. In: *Bioinformatics* 34.18, pp. 3094–3100.
- Li, Heng and Richard Durbin (2011). “Inference of human population history from individual whole-genome sequences”. In: *Nature* 475.7357, pp. 493–496. URL: <https://doi.org/10.1038/nature10231>.
- Li, N. and M. Stephens (2003). “Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data”. In: *Genetics* 165.4, pp. 2213–33.
- Lintusaari, Jarno et al. (2016). “Fundamentals and Recent Developments in Approximate Bayesian Computation”. In: *Systematic Biology* 66.1, e66–e82. URL: <https://doi.org/10.1093/sysbio/syw077>.
- Liu, Weimin et al. (2010). “Origin of the human malaria parasite *Plasmodium falciparum* in gorillas”. In: *Nature* 467.7314, pp. 420–425. URL: <https://www.ncbi.nlm.nih.gov/pubmed/20864995> <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2997044/>.
- Liu, Xiaoming and Yun-Xin Fu (2015). “Exploring population size changes using SNP frequency spectra”. In: *Nature Genetics* 47.5, pp. 555–559. URL: <https://doi.org/10.1038/ng.3254>.
- Loman, N. J. and M. Watson (2015). “Successful test launch for nanopore sequencing”. In: *Nat Methods* 12.4, pp. 303–4.
- Lovett, S. T. (2004). “Encoded errors: mutations and rearrangements mediated by misalignment at repetitive DNA sequences”. In: *Mol Microbiol* 52.5, pp. 1243–53.
- Lubell, Yoel et al. (2014). “Artemisinin resistance – modelling the potential human and economic costs”. In: *Malaria Journal* 13.1, p. 452. URL: <https://doi.org/10.1186/1475-2875-13-452>.
- Macdonald, G. (1952). “The analysis of equilibrium in malaria”. In: *Trop Dis Bull* 49.9, pp. 813–29.
- MalariaGEN, *Plasmodium falciparum* Community Project (2016). “Genomic epidemiology of artemisinin resistant malaria”. In: *eLife* 5, e08714. URL: <https://doi.org/10.7554/eLife.08714>.

- Mandal, Sandip, Ram Rup Sarkar, and Somdatta Sinha (2011). “Mathematical models of malaria - a review”. In: *Malaria Journal* 10.1, p. 202. URL: <https://doi.org/10.1186/1475-2875-10-202>.
- Manske, M. et al. (2012). “Analysis of Plasmodium falciparum diversity in natural infections by deep sequencing”. In: *Nature* 487.7407, pp. 375–9.
- MAP (2017). *The malaria atlas project: developing global maps of malaria risk*. Ed. by. Project Date: December 8, 2017. URL: <https://map.ox.ac.uk/>.
- McDew-White, M. et al. (2019). “Mode and Tempo of Microsatellite Length Change in a Malaria Parasite Mutation Accumulation Experiment”. In: *Genome Biol Evol* 11.7, pp. 1971–1985.
- McKenzie, F. E. et al. (2001). “Meiotic recombination, cross-reactivity, and persistence in Plasmodium falciparum”. In: *Evolution* 55.7, pp. 1299–307.
- Medica, Darcy L. and Photini Sinnis (2005). “Quantitative Dynamics of Plasmodium yoelii Sporozoite Transmission by Infected Anopheline Mosquitoes”. In: *Infection and Immunity* 73.7, p. 4363. URL: <http://iai.asm.org/content/73/7/4363.abstract>.
- Miles, Alistair et al. (2017). “Genetic diversity of the African malaria vector Anopheles gambiae”. In: *Nature* 552.7683, pp. 96–100. URL: <https://doi.org/10.1038/nature24995>.
- Miles, A. et al. (2016). “Indels, structural variation, and recombination drive genomic diversity in Plasmodium falciparum”. In: *Genome Res* 26.9, pp. 1288–99.
- Miller, L. H. et al. (1986). “Research toward malaria vaccines”. In: *Science* 234.4782, pp. 1349–56.
- Miotto, O. et al. (2015). “Genetic architecture of artemisinin-resistant Plasmodium falciparum”. In: *Nat Genet* 47.3, pp. 226–34.
- Moran, P. A. P. (1958). “Random processes in genetics”. In: *Mathematical Proceedings of the Cambridge Philosophical Society* 54.1, pp. 60–71. URL: <https://www.cambridge.org/core/article/random-processes-in-genetics/9EEED52D6AE22A026036F32D9B1CA07C>.
- Murray, C. J. et al. (2012). “Global malaria mortality between 1980 and 2010: a systematic analysis”. In: *Lancet* 379.9814, pp. 413–31.
- Mwanza, S. et al. (2016). “The return of chloroquine-susceptible Plasmodium falciparum malaria in Zambia”. In: *Malar J* 15.1, p. 584.
- Najera, J. A., M. Gonzalez-Silva, and P. L. Alonso (2011). “Some lessons for the future from the Global Malaria Eradication Programme (1955-1969)”. In: *PLoS Med* 8.1, e1000412.
- Ndenga, Bryson A. et al. (2011). “Productivity of Malaria Vectors from Different Habitat Types in the Western Kenya Highlands”. In: *PLOS ONE* 6.4, e19473. URL: <https://doi.org/10.1371/journal.pone.0019473>.
- Neghina, R. et al. (2010). “Malaria, a journey in time: in search of the lost myths and forgotten stories”. In: *Am J Med Sci* 340.6, pp. 492–8.
- Network, Malaria Genomic Epidemiology (2008). “A global network for investigating the genomic epidemiology of malaria”. In: *Nature* 456.7223, pp. 732–7.
- Nkhoma, S. C. et al. (2012). “Close kinship within multiple-genotype malaria parasite infections”. In: *Proc Biol Sci* 279.1738, pp. 2589–98.
- Nkhoma, S. C. et al. (2013). “Population genetic correlates of declining transmission in a human pathogen”. In: *Mol Ecol* 22.2, pp. 273–85.
- Nkhoma, S. C. et al. (2020). “Co-transmission of Related Malaria Parasite Lineages Shapes Within-Host Parasite Diversity”. In: *Cell Host Microbe* 27.1, 93–103.e4.

- Noedl, H. et al. (2008). "Evidence of artemisinin-resistant malaria in western Cambodia". In: *N Engl J Med* 359.24, pp. 2619–20.
- O'Brien, J. D. et al. (2016). "Inferring Strain Mixture within Clinical *Plasmodium falciparum* Isolates from Genomic Sequence Data". In: *PLoS Comput Biol* 12.6, e1004824.
- O'Donnell, C. R., H. Wang, and W. B. Dunbar (2013). "Error analysis of idealized nanopore sequencing". In: *Electrophoresis* 34.15, pp. 2137–44.
- O'Meara, W. P., W. E. Collins, and F. E. McKenzie (2007). "Parasite prevalence: a static measure of dynamic infections". In: *Am J Trop Med Hyg* 77.2, pp. 246–9.
- Onori, E. and B. Grab (1980). "Indicators for the forecasting of malaria epidemics". In: *Bull World Health Organ* 58.1, pp. 91–8.
- Overgaard, H. J. et al. (2012). "Light traps fail to estimate reliable malaria mosquito biting rates on Bioko Island, Equatorial Guinea". In: *Malar J* 11, p. 56.
- Oyola, S. O. et al. (2013). "Efficient depletion of host DNA contamination in malaria clinical sequencing". In: *J Clin Microbiol* 51.3, pp. 745–51.
- Oyola, S. O. et al. (2016). "Whole genome sequencing of *Plasmodium falciparum* from dried blood spots using selective whole genome amplification". In: *Malar J* 15.1, p. 597.
- Paaijmans, Krijn P. et al. (2007). "Unexpected High Losses of *Anopheles gambiae* Larvae Due to Rainfall". In: *PLOS ONE* 2.11, e1146. URL: <https://doi.org/10.1371/journal.pone.0001146>.
- Packard, R. M. (2014). "The origins of antimalarial-drug resistance". In: *N Engl J Med* 371.5, pp. 397–9.
- Paul, R. E. et al. (1995). "Mating patterns in malaria parasite populations of Papua New Guinea". In: *Science* 269.5231, pp. 1709–11.
- Payne, D. (1988). "Did medicated salt hasten the spread of chloroquine resistance in *Plasmodium falciparum*?" In: *Parasitol Today* 4.4, pp. 112–5.
- Pearson, Richard D. et al. (2016). "Genomic analysis of local variation and recent evolution in *Plasmodium vivax*". In: *Nature genetics* 48.8, pp. 959–964. URL: <https://www.ncbi.nlm.nih.gov/pubmed/27348299>
<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4966634/>.
- Phyo, A. P. et al. (2012). "Emergence of artemisinin-resistant malaria on the western border of Thailand: a longitudinal study". In: *Lancet* 379.9830, pp. 1960–6.
- Plowe, C. V. et al. (1997). "Mutations in *Plasmodium falciparum* dihydrofolate reductase and dihydropteroate synthase and epidemiologic patterns of pyrimethamine-sulfadoxine use and resistance". In: *J Infect Dis* 176.6, pp. 1590–6.
- Price, R. N. et al. (2004). "Mefloquine resistance in *Plasmodium falciparum* and increased *pfmdr1* gene copy number". In: *Lancet* 364.9432, pp. 438–447.
- Prudencio, M., A. Rodriguez, and M. M. Mota (2006). "The silent path to thousands of merozoites: the *Plasmodium* liver stage". In: *Nat Rev Microbiol* 4.11, pp. 849–56.
- Pybus, Oliver G., Andrew Rambaut, and Paul H. Harvey (2000). "An Integrated Framework for the Inference of Viral Population History From Reconstructed Genealogies". In: *Genetics* 155.3, p. 1429. URL: <http://www.genetics.org/content/155/3/1429.abstract>.
- Quick, J. et al. (2016). "Real-time, portable genome sequencing for Ebola surveillance". In: *Nature* 530.7589, pp. 228–232.
- Reiner, R. C. et al. (2015a). "Mapping residual transmission for malaria elimination". In: *eLife* 4, e09520. URL: <https://doi.org/10.7554/eLife.09520>.

- Reiner, R. C. et al. (2015b). "Seasonality of *Plasmodium falciparum* transmission: a systematic review". In: *Malaria Journal* 14.1, p. 343. URL: <https://doi.org/10.1186/s12936-015-0849-2>.
- Rejmnkov, Erika et al. (2013). "Ecology of Larval Habitats". In:
- Rich, S. M. et al. (2009). "The origin of malignant malaria". In: *Proc Natl Acad Sci U S A* 106.35, pp. 14902–7.
- Rogers, W. O. et al. (2009). "Failure of artesunate-mefloquine combination therapy for uncomplicated *Plasmodium falciparum* malaria in southern Cambodia". In: *Malar J* 8, p. 10.
- Roper, C. et al. (2004). "Intercontinental spread of pyrimethamine-resistant malaria". In: *Science* 305.5687, p. 1124.
- Runtuwene, L. R. et al. (2018). "Nanopore sequencing of drug-resistance-associated genes in malaria parasites, *Plasmodium falciparum*". In: *Sci Rep* 8.1, p. 8286.
- Sama, W., K. Dietz, and T. Smith (2006). "Distribution of survival times of deliberate *Plasmodium falciparum* infections in tertiary syphilis patients". In: *Trans R Soc Trop Med Hyg* 100.9, pp. 811–6.
- Shampo, Marc A. and Robert A. Kyle (1989). "Nei Chingx2014;Oldest Known Medical Book". In: *Mayo Clinic Proceedings* 64.1, p. 134. URL: [https://doi.org/10.1016/S0025-6196\(12\)65320-8](https://doi.org/10.1016/S0025-6196(12)65320-8).
- Sinden, R. E. and R. H. Hartley (1985). "Identification of the meiotic division of malarial parasites". In: *J Protozool* 32.4, pp. 742–4.
- Sinka, Marianne E. et al. (2012). "A global map of dominant malaria vectors". In: *Parasites Vectors* 5.1, p. 69. URL: <https://doi.org/10.1186/1756-3305-5-69>.
- Smith, D. L. et al. (2007a). "Revisiting the basic reproductive number for malaria and its implications for malaria control". In: *PLoS Biol* 5.3, e42.
- Smith, D. L. et al. (2007b). "Standardizing estimates of the *Plasmodium falciparum* parasite rate". In: *Malar J* 6, p. 131.
- Smith, David L. et al. (2012). "Ross, Macdonald, and a Theory for the Dynamics and Control of Mosquito-Transmitted Pathogens". In: *PLOS Pathogens* 8.4, e1002588. URL: <https://doi.org/10.1371/journal.ppat.1002588>.
- Smith, J. D. et al. (2013). "Malaria's deadly grip: cytoadhesion of *Plasmodium falciparum*-infected erythrocytes". In: *Cell Microbiol* 15.12, pp. 1976–83.
- Smith, Neal R. et al. (2018). "Agent-based models of malaria transmission: a systematic review". In: *Malaria Journal* 17.1, p. 299. URL: <https://doi.org/10.1186/s12936-018-2442-y>.
- Snow, R. W. and K. Marsh (2002). "The consequences of reducing transmission of *Plasmodium falciparum* in Africa". In: *Adv Parasitol* 52, pp. 235–64.
- St. Laurent, Brandyce et al. (2016). "Cow-baited tents are highly effective in sampling diverse *Anopheles* malaria vectors in Cambodia". In: *Malaria Journal* 15.1, p. 440. URL: <https://doi.org/10.1186/s12936-016-1488-y>.
- Stevenson, J. et al. (2012). "Novel vectors of malaria parasites in the western highlands of Kenya". In: *Emerg Infect Dis* 18.9, pp. 1547–9.
- Sturrock, H. J. W. et al. (2016). "Mapping Malaria Risk in Low Transmission Settings: Challenges and Opportunities". In: *Trends Parasitol* 32.8, pp. 635–645.
- Su, X. Z. et al. (1996). "Reduced extension temperatures required for PCR amplification of extremely A+T-rich DNA". In: *Nucleic Acids Res* 24.8, pp. 1574–5.

- Sundararaman, Sesh A. et al. (2016). “Genomes of cryptic chimpanzee *Plasmodium* species reveal key evolutionary events leading to human malaria”. In: *Nature Communications* 7.1, p. 11078. URL: <https://doi.org/10.1038/ncomms11078>.
- Tajima, F. (1989). “Statistical method for testing the neutral mutation hypothesis by DNA polymorphism”. In: *Genetics* 123.3, pp. 585–95.
- Tanaka, Mark M. et al. (2006). “Using approximate Bayesian computation to estimate tuberculosis transmission parameters from genotype data”. In: *Genetics* 173.3, pp. 1511–1520. URL: <https://www.ncbi.nlm.nih.gov/pubmed/16624908><https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1526704/>.
- Tang, L. H., H. L. Qian, and S. H. Xu (1991). “Malaria and its control in the People’s Republic of China”. In: *Southeast Asian J Trop Med Public Health* 22.4, pp. 467–76.
- Tessema, Sofonias et al. (2019a). “Using parasite genetic and human mobility data to infer local and cross-border malaria connectivity in Southern Africa”. In: *eLife* 8, e43510. URL: <https://doi.org/10.7554/eLife.43510>.
- Tessema, S. et al. (2019b). “Applying next-generation sequencing to track falciparum malaria in sub-Saharan Africa”. In: *Malar J* 18.1, p. 268.
- Tsai, J. H. and B. D. McKee (2011). “Homologous pairing and the role of pairing centers in meiosis”. In: *J Cell Sci* 124.Pt 12, pp. 1955–63.
- Tu, Y. (2011). “The discovery of artemisinin (qinghaosu) and gifts from Chinese medicine”. In: *Nat Med* 17.10, pp. 1217–20.
- Tusting, L. S. et al. (2014). “Measuring changes in *Plasmodium falciparum* transmission: precision, accuracy and costs of metrics”. In: *Adv Parasitol* 84, pp. 151–208.
- Untergasser, A. et al. (2007). “Primer3Plus, an enhanced web interface to Primer3”. In: *Nucleic Acids Res* 35.Web Server issue, W71–4.
- Vaughan, A. M., A. S. Aly, and S. H. Kappe (2008). “Malaria parasite pre-erythrocytic stage infection: gliding and hiding”. In: *Cell Host Microbe* 4.3, pp. 209–18.
- Venkatesan, M. et al. (2012). “Using CF11 cellulose columns to inexpensively and effectively remove human DNA from *Plasmodium falciparum*-infected whole blood samples”. In: *Malar J* 11, p. 41.
- Videvall, E. (2018). “*Plasmodium* parasites of birds have the most AT-rich genes of eukaryotes”. In: *Microb Genom* 4.2.
- Viriyakosol, S. et al. (1995). “Genotyping of *Plasmodium falciparum* isolates by the polymerase chain reaction and potential uses in epidemiological studies”. In: *Bull World Health Organ* 73.1, pp. 85–95.
- Volkman, S. K. et al. (2012). “Harnessing genomics and genome biology to understand malaria biology”. In: *Nat Rev Genet* 13.5, pp. 315–28.
- Wakeley, J. (2016). *Coalescent Theory: An Introduction*. Macmillan Learning. URL: <https://books.google.co.uk/books?id=x3ORAgAACAAJ>.
- Weiss, D. J. et al. (2015). “Re-examining environmental correlates of *Plasmodium falciparum* malaria endemicity: a data-intensive variable selection approach”. In: *Malar J* 14, p. 68.
- Weiss, D. J. et al. (2019). “Mapping the global prevalence, incidence, and mortality of *Plasmodium falciparum*, 2000–17: a spatial and temporal modelling study”. In: *Lancet* 394.10195, pp. 322–331.
- Wernsdorfer, W. H. and D. Payne (1991). “The dynamics of drug resistance in *Plasmodium falciparum*”. In: *Pharmacol Ther* 50.1, pp. 95–121.

- Wesolowski, Amy et al. (2018). “Mapping malaria by combining parasite genomic and epidemiologic data”. In: *BMC Medicine* 16.1, p. 190. URL: <https://doi.org/10.1186/s12916-018-1181-9>.
- White, Nicholas J. (2004). “Antimalarial drug resistance”. In: *The Journal of clinical investigation* 113.8, pp. 1084–1092. URL: <https://www.ncbi.nlm.nih.gov/pubmed/15085184><https://www.ncbi.nlm.nih.gov/pmc/articles/PMC385418/>.
- (2014). “43 - Malaria”. In: *Manson’s Tropical Infectious Diseases (Twenty-third Edition)*. Ed. by Jeremy Farrar et al. London: W.B. Saunders, 532–600.e1. URL: <http://www.sciencedirect.com/science/article/pii/B9780702051012000443>.
- WHO (2006). *Guidelines for the Treatment of Malaria, First Edition*. Ed. by World Health Organization. Geneva, Switzerland.
- (2010). *Guidelines for the Treatment of Malaria, Second Edition*. Ed. by World Health Organization. Geneva, Switzerland.
- (2015). *Global technical strategy for malaria 2016-2030*. Ed. by World Health Organization. Geneva, Switzerland.
- (2018a). *Artemisinin resistance and artemisinin-based combination therapy efficacy*. Ed. by World Health Organization. Project Date: August 2018. Geneva, Switzerland. URL: www.who.int/malaria/publications/atoz/artemisinin-resistance-august2018/en/.
- (2018b). *World Malaria Report*. Ed. by World Health Organization. Project Date: August 2018. Geneva, Switzerland. URL: <https://www.who.int/malaria/media/world-malaria-report-2018/en/>.
- (2019). *World Malaria Report*. Ed. by World Health Organization. Project Date: August 2019. Geneva, Switzerland. URL: <https://www.who.int/publications-detail/world-malaria-report-2019>.
- Wick, Ryan R., Louise M. Judd, and Kathryn E. Holt (2019). “Performance of neural network basecalling tools for Oxford Nanopore sequencing”. In: *Genome Biology* 20.1, p. 129. URL: <https://doi.org/10.1186/s13059-019-1727-y>.
- Witkowski, B. et al. (2017). “A surrogate marker of piperaquine-resistant *Plasmodium falciparum* malaria: a phenotype-genotype association study”. In: *Lancet Infect Dis* 17.2, pp. 174–183.
- Wong, Wesley et al. (2018). “Modeling the genetic relatedness of *Plasmodium falciparum* parasites following meiotic recombination and cotransmission”. In: *PLOS Computational Biology* 14.1, e1005923. URL: <https://doi.org/10.1371/journal.pcbi.1005923>.
- Wootton, J. C. et al. (2002). “Genetic diversity and chloroquine selective sweeps in *Plasmodium falciparum*”. In: *Nature* 418.6895, pp. 320–3.
- Yamauchi, L. M. et al. (2007). “*Plasmodium* sporozoites trickle out of the injection site”. In: *Cell Microbiol* 9.5, pp. 1215–22.
- Zhou, Z. J. (1981). “The malaria situation in the People’s Republic of China”. In: *Bull World Health Organ* 59.6, pp. 931–6.
- Zhu, S. J., J. Almagro-Garcia, and G. McVean (2018). “Deconvolution of multiple infections in *Plasmodium falciparum* from high throughput sequencing data”. In: *Bioinformatics* 34.1, pp. 9–15.
- Zhu, S. J. et al. (2019). “The origins and relatedness structure of mixed infections vary with local prevalence of *P. falciparum* malaria”. In: *Elife* 8.