Estimation of *Plasmodium falciparum* allele and multi-SNP haplotype and genotype frequencies

Aimee Rebecca Taylor

Supervised by Professor Chris Holmes,

Professor Philippe Guérin and Dr Jennifer Flegg

Department of Statistics

University of Oxford

This dissertation is submitted for the degree of

Doctor of Philosophy

St John's College July 2016

I would like to dedicate this thesis to Elizabeth, Ian and James Taylor.

Declaration

The content of this thesis is entirely my own and has not been submitted for the fulfillment of a degree elsewhere. The content of chapter 3 has been published [248]. In addition to advice from my supervisors, I received general guidance on the topic of malaria from Carol Sibley, Philip Rosenthal and Grant Dorsey in its editing. The remaining authors of the aforesaid paper, Samuel Nsobya, Adoke Yeka, Moses Kamya, contributed the data that feature in the paper and to the paper's editing. The data presented in this thesis are not my own. Those that feature in chapter 4 were sent to me by Melissa Conrad, Philip Rosenthal and Grant Dorsey. Those that feature in chapter 6 were sent to me by Lucas Amenga-Etego and Kirk Rockett.

Aimee Rebecca Taylor

July 2016

Acknowledgements

First and foremost, I would like to thank my supervisors, Chris Holmes, Philippe Guérin and Jennifer Flegg. I am extremely fortunate to have had three truly inspirational and knowledgable mentors. Transitioning from my undergraduate degree in Chemistry to the fields of both Statistics and Malaria has been both challenging and rewarding. I would not have been able to do it without Chris's statistical guidance, Philippe's mentorship within the malaria field and Jen's unerring support. I owe all that I have achieved to their supervision and cannot thank them sufficiently for it.

I am also indebted to the Systems Biology Doctoral Training Centre for the opportunity to embark upon the DPhil programme, to the EPSRC for funding, and to Philip Rosenthal, Grant Dorsey and Melissa Conrad, and to Lucas Amenga-Etego and Kirk Rockett for granting me permission to analyse the Ugandan and Ghanaian data, respectively.

I would like to thank the WorldWide Antimalarial Resistance Network (WWARN) for welcoming me into the WWARN effort. I have met many inspiring people through the network and cannot overstate how formative being a member of it has been. In addition to my supervisors at WWARN, Philippe Guérin and Jennifer Flegg, I am especially grateful for the mentorship of Carol Sibley and Christian Nsanzabana, as well as the camaraderie and assistance of fellow PhD student Prabin Dahal.

Many thanks also go to the Department of Statistics for fostering an intellectually thriving community of many brilliant and friendly people. In particular, I would like to acknowledge my contemporaries, Tristan Gray-Davies and James Watson for their friendship and support;

Christopher Yau for insightful discussions on the analyses of genetically diverse sample; George Nicholson for encouragement and assistance; and Pierre Jacob for his unfailing support, furtherance, countless captivating discussions and dancing.

I would also like to acknowledge my college, St John's, and the many friends I have met through the college and beyond. I have lived with awesome housemates and formed many cherished friendships during my time in Oxford. Thanks goes to all those who have made me laugh and kept me sane, especially James Watson and Jacob Bush, two very dear friends throughout.

Finally, my deepest gratitude goes to Pierre Jacob and my parents for their kindness, encouragement and immense patience. It's fair to say, I have been a bit difficult at times over the past few months. My parents and Pierre have been my lifeline.

I

Abstract

Malaria kills hundreds of thousands of people each year, yet is entirely curable given prompt treatment. Malaria parasites evolve resistance to antimalarial drugs, hence routine surveillance of antimalarial resistance is vital. The surveillance of parasite genetic markers of resistance provides an economical adjunct to clinical efficacy trials, and has the potential to resolve drug-specific resistance ahead of clinical failure. To monitor spatiotemporal changes using genetic markers, frequencies of alleles and/or haplotypes and genotypes spanning multiple single nucleotide polymorphisms (SNPs) are required. However, multiclonal infections complicate frequency estimation, especially in highly endemic regions.

With the aim of harnessing the full potential of genetic markers for the surveillance of antimalarial resistance, a statistical model to estimate frequencies is proposed. The model builds upon existing methods (reviewed in chapter 2), without reliance upon experimentally-derived estimates of the sample-wise multiplicities of infection (MOIs). Its ability to generate precise and accurate estimates within a Bayesian framework is documented in chapter 3. In chapter 4, the model is applied to data collected from a cohort of children enrolled in a longitudinal trial in Uganda, generating valuable insight into haplotype frequency trends. In chapter 5, the model is extended to investigate inter-child variability in the aforesaid cohort, revealing a small amount of inter-child variation. In chapter 6, the model is modified to enable the analysis of short-read sequencing data, with application to data from malaria patients in Northern Ghana, providing insight into the extent of within-host diversity and anti-folate resistance in the region.

In summary, this thesis documents the development, application, extension and modification

of a model designed to estimate population-level frequencies of *P. falciparum* alleles and multi-SNP haplotypes and genotypes within a Bayesian framework. It is hoped that the model and its proposed framework will provide a practical tool for surveillance of antimalarial resistance, as well as a foundation on which to develop further methods.

Table of contents

List of figures

List of tables

Nomenclature

Acronyms / Abbreviations

ACT Artemisinin combination therapy

AL Artemether-lumefantrine

AQ Amodiaquine

AS/AQ Artesunate-amodiaquine

COI Complexity of infection

CQ Chloroquine

CQ+SP Chloroquine plus sulfadoxine-pyrimethamine

DDT Dichloro-diphenyl-trichloroethane

DIC Deviance information criterion

DNA Deoxyribonucleic acid

DP Dihydroartemisinin-piperaquine

ELISA Enzyme-linked immunosorbent assay

EM Expectation maximisation

GMEP Global malaria eradication programme

GMS Greater Mekong subregion

IPTi Intermittent preventative treatment of infants

IPTp Intermittent preventative treatment of pregnant women

IRS Indoor residual spraying

ITN Insecticide treated net

MCMC Markov chain Monte Carlo

MDA Mass drug administration

MOI Multiplicity of infection

nSNP Non-synonymous single nucleotide polymorphism

PCR Polymerase chain reaction

PSRF Potential scale reduction factor

qPCR Quantitative polymerase chain reaction

RFLP Restriction fragment length polymorphism

SMC Seasonal malaria chemoprevention

SNP Single nucleotide polymorphism

SP Sulfadoxine-pyrimethamine

SSOP Sequence specific oligonucleotide probe

WHO World Health Organization

Chapter 1

Introduction

1.1 Malaria

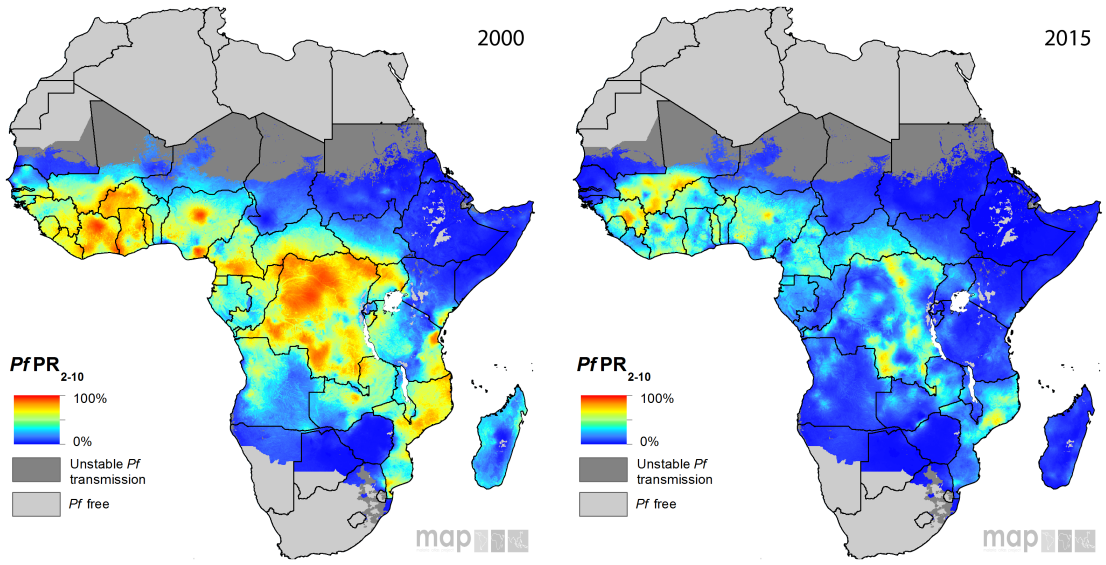
1.1.1 Morbidity and mortality

Malaria is thought to have killed between 150 and 300 million people in the 20th century alone [40]. Following the introduction of indoor residual spraying (IRS) with dichloro-diphenyl-trichloroethane (DDT) in the 1940s, the World Health Assembly launched the Global Malaria Eradication Programme (GMEP) in 1955 [171]. The mainstay of the GMEP was IRS with DDT and other insecticides; that is to say, vector control [171]. Vector control programmes brought success in Europe and North America, but in many endemic countries, where the goal of eradication was infeasible given the available resources, transmission remained high [40, 171]. Ultimately, the over-optimistic, uncompromising and top-down strategy of the GMEP failed [171]. Malaria eradication was abandoned in 1969, and international efforts to fight the disease diminished [171]. The decades that followed saw a resurgence in malaria [122, 175, 21], as well as the spread of resistance to the antimalarial drug chloroquine (CQ) [193]. CQ was the linchpin of malaria control in the latter half of the 20th century. In the 1980s and 1990s, resistance to the drug drove an escalation in mortality, especially in Africa [254, 231]. With

a need to establish a coordinated plan to control the re-emerging disease [175], international interest was renewed [147], and programmes such as the Roll Back Malaria campaign [165] were initiated. At the United Nations' summit, the Millennium Development Goal 'to have halted and begun to reverse the incidence of malaria' by 2015 was set [259], followed by the Global Malaria Action Plan [251].

Recent efforts to combat malaria have been met with considerable success [290, 27]. In the past 15 years, the World Health Organization (WHO) reports an estimated 48% reduction in mortality worldwide [290], while in Africa, Bhatt *et al.* estimate that 663 million clinical cases have been averted (range: 542–753 million), and that there has been a 40% reduction in endemicity, as measured by the *Plasmodium falciparum* (*P. falciparum*) positivity rate (figure 1.1) [27][1]. Despite progress, almost half of the world's population still lives at risk of malaria [290]. The disease continues to kill hundreds of thousands of people each year, most of whom are children in sub-Saharan Africa. According to the World Malaria Report 2015, there were an estimated 214 million cases of malaria in 2015 (range: 149–303 million), and 438,000 deaths (range: 236,000–635,000) [290]. Of these, an estimated 88% of cases and 90% of deaths were in Africa, with an estimated 292,000 deaths of children less than 5 years of age (range: 212,000–384,000) [290]. Furthermore, resistance to insecticides and antimalarial drugs threatens to exacerbate the burden and reverse the progress attained thus far [65, 290]. As we look towards the post-2015 era, sustained concerted political commitment and increased investment are needed to both avert a retrogression, and to move towards the current WHO goal of a 90% reduction in malaria morbidity and mortality by 2030 [286, 291].

[1]Reports of malaria morbidity and mortality are hampered by incomplete and poor quality surveillance data [147]. Consequently, estimates, rather than exact counts, are reported. Full details of the data sources and methods used to generate the reported estimates can be found in the respective references, [290] and [27].



(a) Year 2000 (b) Year 2015

Figure 1.1: The year 2000–2015 decline in malaria endemicity in Africa. Heat maps show the *P. falciparum* parasite rate for the 2 to 10 year old age range, $Pf\mathrm{PR}_{2-10}$, in Africa in years 2000 and 2015. Image sourced from [27].

1.1.2 Epidemiological context

Malaria is a potentially fatal mosquito-borne parasitic disease. Protozoan parasites from the genus *Plasmodium* are transmitted from human to human by female *Anopheles* mosquitoes. Five species cause malaria in humans: *P. falciparum*, *P. vivax*, *P. ovale*, *P. malariae*, and the monkey malaria parasite, *P. knowlesi*, which occasionally infects humans [28]. The most common species, *P. falciparum*, is responsible for the deadliest type of malaria, and prevails across sub-Saharan Africa (figure 1.1). The second most common species, *P. vivax*, is responsible for a type of malaria that causes relapses. In 2015, *P. vivax* accounted for an estimated 6% of cases worldwide, and 51% of cases outside of Africa [290]. This thesis will focus on *P. falciparum*, although the methods could be applied to other species.

Symptoms of what is sometimes termed uncomplicated malaria include fever, chills, aches and nausea [40]. If untreated, uncomplicated malaria caused by *P. falciparum* can progress to severe malaria, a term encompassing many life threatening conditions including severe

anaemia, respiratory distress and organ failure [285]. Repeated exposure to malaria leads to the development of naturally acquired immunity, but does not afford complete protection [197].

The lifecycle of *P. falciparum* is summarised in figure 1.2. The human stage of the life cycle is initiated upon the inoculation of the host with sporozoites (see ①, figure 1.2). Note that the number of sporozoites inoculated per bite is thought to be small and highly variable (estimates include median 8 (range 0–522) [208], and median 15 (range 0–978) [218]). The parasites are haploid throughout the human phases of their lifecycle [267]; that is to say, each parasite has a single copy of its 14 chromosomes [85]. The haploid parasites reproduce asexually, first in the liver (exo-erythrocytic cycle, Ⓐ, figure 1.2), and then over several 48-hour cycles in the blood (erythrocytic cycle, Ⓑ, figure 1.2). In the mature stages of the 48-hour cycle infected red blood cells withdraw from peripheral blood in a process known as sequestration [152], resulting in complex infection dynamics [77]. The mosquito phase of the lifecycle (sporogonic cycle, Ⓒ, figure 1.2) includes an obligate sexual replicative stage (⑨, figure 1.2).

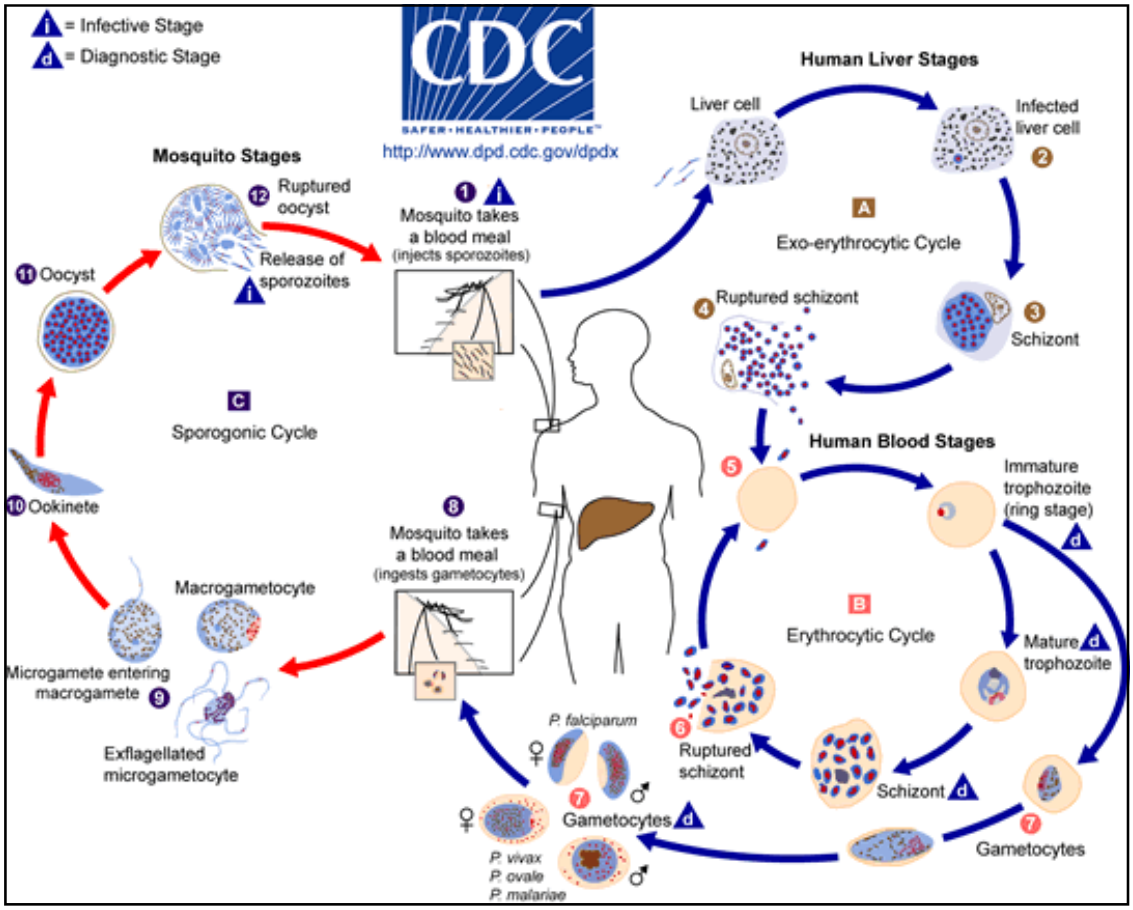
In endemic settings people can receive 100s of infective bites per year [99], and can consequently get infected with different species [220], and/or with genetically distinct clones of the same species (for example, [260]). Note that following convention, we use the term ‘clone’ to denote a collection of genetically identical parasites. Like humans, mosquitoes can harbour mixed infections comprising different species [24] and genetically distinct clones [184]. Hence, within-host diversity may result from co-transmission (a single inoculation from a mosquito harbouring a genetically mixed but related infection [52, 66, 238, 180]), as well as upon receipt of successive infectious bites with genetically distinct clones (a mechanism known as superinfection [174]). Traditionally, superinfection was thought to be the main generative mechanism of multiclonal infections in endemic settings; however, work detailing the genetic diversity of within-host infections, suggests the occurrence of co-transmission has been underestimated [66]. In reality, both mechanism likely play a role, leading to an abundance of multiclonal *P. falciparum* infections, especially across malaria endemic regions

[121, 17].

1.1.3 MOI

The number of genetically distinct clones in an infection is known either as the MOI or the complexity of infection (COI). The MOI cannot be measured directly. Typically, an estimate is determined experimentally by polymerase chain reaction (PCR) based genotyping of one or more established marker loci. The most commonly genotyped markers are variable number random repeats in the genes that encode merozoite surface proteins one and two (*msp1*, *msp2*, respectively) and the gene that encodes glutamine-rich protein, *glurp*, as recommended by the WHO [282]. Experimentally determined MOI estimates are normally based on the maximum number of distinct alleles detected per locus per sample [164]. The allelic diversity of markers at loci such as *msp1*, *msp2* and *glurp* needs to be high, to avoid two or more clones having the same allele. Despite two or more clones potentially having the same allele, an *in silico* study has shown experimentally-derived estimates to be a reasonable approximation of the average MOI, providing there are 20 plus alleles per marker locus and a number of additional conditions are met [219]. However, this *in silico* study does not take into account imperfect detectability [219].

Akin to two or more clones having the same allele, imperfect detectability causes the number of clones to be underestimated. Imperfect detectability may arise from the limited sensitivity of PCR-based genotyping for the detection of minority clones in the peripheral blood, and/or because of sequestration [219]. The sensitivity of the PCR-based method to minority clones depends on the procedure (ranging between 80% and 99%) [114, 58]. The extent of imperfect detectability due to sequestration can be probed by repeat sample collection [33, 120]. Based on repeat samples, Bretscher *et al.* estimate that at most, one can hope to detect 50% of the diversity using PCR-based methods at a single time point. Similar estimates have been obtained using models of infection and recovery (for example [221]). Higher detection



The malaria parasite life cycle involves two hosts. During a blood meal, a malaria-infected female *Anopheles* mosquito inoculates sporozoites into the human host (1). Sporozoites infect liver cells (2) and mature into schizonts (3), which rupture and release merozoites (4). (Of note, in *P. vivax* and *P. ovale* a dormant stage [hypnozoites] can persist in the liver and cause relapses by invading the bloodstream weeks, or even years later.) After this initial replication in the liver (exo-erythrocytic schizogony (A)), the parasites undergo asexual multiplication in the erythrocytes (erythrocytic schizogony (B)). Merozoites infect red blood cells (5). The ring stage trophozoites mature into schizonts, which rupture releasing merozoites (6). Some parasites differentiate into sexual erythrocytic stages (gametocytes) (7). Blood stage parasites are responsible for the clinical manifestations of the disease.

The gametocytes, male (microgametocytes) and female (macrogametocytes), are ingested by an *Anopheles* mosquito during a blood meal (8). The parasites' multiplication in the mosquito is known as the sporogonic cycle (C). While in the mosquito's stomach, the microgametes penetrate the macrogametes generating zygotes (9). The zygotes in turn become motile and elongated (ookinetes) (10) which invade the midgut wall of the mosquito where they develop into oocysts (11). The oocysts grow, rupture, and release sporozoites (12), which make their way to the mosquito's salivary glands. Inoculation of the sporozoites into a new human host perpetuates the malaria life cycle (1).

Figure 1.2: The lifecycle of *P. falciparum*. Both the illustration and legend were sourced from [60].

(70–80%) was reported using repeat samples over a 24-hour interval [120]. Most studies base MOI estimates on a sample taken at a single time point, hence report the number of surveyable clones in the peripheral blood at a given time, which is likely underestimating of the true MOI.

Typically, the number of genetically distinct clones surveyable in the peripheral blood at a single time point ranges from one to eight [189], where eight is the maximum resolution of the commonly used PCR-based method of analysing restriction fragment length polymorphisms (RFLP) [80]. The MOI distribution has been shown to vary with several epidemiological covariates including transmission intensity [121, 10, 221, 225, 258], age [183, 121, 230, 189, 157, 91] and clinical status [23, 79, 1, 209, 119, 112, 30].

Alternative methods for MOI estimation use microsatellite markers [10], genome-wide panels of SNPs [57, 84], or whole genome sequencing data [17, 186]. Averaging over the whole genome can generate more sensitive estimates, since it circumvents the problem of two or more clones having the same allele at a specific locus [17]. The scenario where two or more clones have the same allele can also be accounted for under a statistical model (as is done in the aforementioned methods applied to genome-wide data, [17, 84]). Statistical models for MOI estimation are discussed in chapter 2. Different marker loci and methods lead to different estimates, both experimentally and statistically [223], but most are of the same order as those based on *msp1*, *msp2* and *glurp*.

A related metric, the F_{ws} statistic, is a genome-wide population-normalised measure of within-host diversity, ranging from zero (when the within-host diversity is equivalent to that seen in the surrounding population) to one (when the infection is monoclonal) [19, 142]. The F_{ws} statistic can provide insight into population structure [153], has been shown to correlate with MOI estimates (especially those based on *msp1*) [19], and can be related to the MOI under a statistical model [186].

1.1.4 Prevention and treatment

Despite killing hundreds of thousand of people each year, malaria is an entirely preventable disease. Preventative measures that focus on vector control include insecticide treated nets (ITNs) and IRS. Over the past 15 years, ITNs and IRS are estimated to have contributed to 68% and 10%, respectively, of the estimated 663 million cases averted in Africa [27].

Preventative measures using antimalarial drugs aim to suppress infection in risk groups [290]. Risk groups in endemic regions include children who have not yet acquired immunity and pregnant women (especially primigravidae), who experience a temporary state of reduced immunity [32], due to the specific pathology of malaria in pregnancy [215]. Preventative measures include intermittent preventative treatment of infants (IPTi) and pregnant women (IPTp) with sulfadoxine-pyrimethamine (SP), and seasonal malaria chemoprevention (SMC) of children with amodiaquine (AQ) plus SP (AQ+SP) [287]. IPTi is recommended in regions of moderate-to-high transmission in Africa where *P. falciparum* SP resistance is not too high, while IPTp is recommended in all malaria-endemic regions in Africa, regardless of resistance to SP [287]. SMC is recommended in areas of highly seasonal transmission in sub-Sahel Africa where *P. falciparum* remains sensitive to both AQ and SP [287].

Work to develop a malaria vaccine has been underway for many years. The first and only vaccine to have been approved by the European Medicines Agency is RTS,S, developed by GlaxoSmithKline and the PATH Malaria Vaccine Initiative [104, 73]. It is only partially effective, however, and protection deteriorates over time [252, 271]. The WHO have therefore recommended pilot studies to assess the utility of the vaccine as a complementary tool for malaria control [292].

Uncomplicated malaria is curable if treated with efficacious antimalarial drugs. Artemisinin-based combination therapy (ACT) is the recommended first-line treatment for uncomplicated malaria caused by *P. falciparum* [287]. ACTs are comprised of a potent fast-acting artemisinin derivative (artemisinin, artesunate, artemether and dihydroartemisinin) with a slow-acting

partner drug [270]. The role of the artemisinin derivative is to rapidly kill the vast majority of parasites [272], leaving the partner drug to kill any that remain [270]. WHO-recommended partner drugs currently include SP, AQ, lumefantrine, mefloquine and piperaquine [287]. Alternative partner drugs include naphthoquinef and pyronaridine [287]. The profile of pyronaridine looks promising [55], but since there is insufficient evidence as to its comparative efficacy in young children, it is not as yet WHO-recommended [287]. CQ monotherapy is still used to treat uncomplicated malaria caused by *P. vivax*, *P. ovale*, *P. malariae* or *P. knowlesi* in regions where the parasites susceptible to CQ [287].

1.1.5 Antimalarial resistance

Antimalarial resistance is defined as the 'persistence or recurrence of malaria parasites after appropriate drug treatment' [65]; that is to say, the ability of the malaria parasites to withstand antimalarial drugs. It arises upon the emergence of a genetic variant in the parasite genome that confers a selective advantage in the presence of drug pressure [188], and results in partial or complete treatment failure. Efforts to combat malaria have long been thwarted by antimalarial resistance [198], and the emergence and spread of resistance to antimalarial drugs, such as CQ and SP, have had a significant impact on public health [254, 40, 147].

CQ resistance is thought to have emerged independently in both Colombia and on the Thai-Cambodia border in the late 1950s [193]. Confirmed cases were first reported in Colombia in 1961, followed by delayed reports of cases in Thailand in 1962 (see [193] and references therein). CQ resistance was detected in East Africa in the late 1970s [268]. Initially, the emergence of CQ resistance in Africa was thought to be an independent event [193], but genetic analyses later revealed that CQ resistance spread to East Africa from southeast Asia [280, 13]. In the decades that followed, CQ resistance spread across the African continent [193, 280], driving a surge in child mortality [254, 231].

In many countries, first-line treatment with CQ was superseded by SP [270]. SP is an

antimalarial comprising two synergistic, antifolate drugs, sulfadoxine and pyrimethamine, which both act upon the folate biosynthesis pathway [188]. Reports of the capacity of *P. falciparum* to spontaneously evolve *in vivo* resistance to pyrimethamine date back to studies in Tanzania in the 1950s [50]. Sulfadoxine was combined with pyrimethamine in a bid to find an effective treatment for parasites resistant to both CQ and pyrimethamine [47]. Reports of *in vivo* resistance to the fixed combination SP date back to 1979 on the Thai-Cambodia border [108], and 1981 in Colombia [71], not long after its widespread deployment as a replacement of CQ. Case reports of combined CQ and SP resistant infections were reported in Africa in the 1980s (for example [100]). Akin to resistance to CQ, genetic analyses showed that, in South America, parasites with mid to high resistance to SP share a common ancestor [53], while in Africa, parasites with a high level of resistance to pyrimethamine descend from those in Southeast Asia [216]. Genetic mutations conferring resistance to sulfadoxine were not reported until the 1990s in Africa, where they were detected against the prevailing backdrop of existing pyrimethamine resistance [168]. Five distinct lineages, all thought to be of African descent, were discerned, two conferring 'full resistance' to SP [195]. However, a subsequent study implies that highly SP-resistant parasites have emerged in only two locations (South America and Southeast Asia), and that two of the highly resistant lineages from Southeast Asia spread to East Africa [156], following the path of parasites resistant to pyrimethamine and CQ. SP resistance is now extensive, particularly in Southeast Asia, East Africa and across South America [168, 296]. Despite widespread resistance, SP monotherapy is still recommended for IPTp and IPTi in some regions of Africa [287].

Artemisinin derivatives were developed in China in the 1970s [275]. They were internationally disclosed in 1979, but universal uptake was slow [275]. Artemisinin and its derivative did not feature globally until the 1990s [65, 16]. Initially they were administered as monotherapies [275], but in a bid to impede the development of resistance, there was an urgent call for artemisinin and its derivatives to be used in combination with a partner drug that has an

independent mode of action [270, 281]. It was hoped that by using artemisinin in combination in this way, the probability of a parasite evolving resistance to both drugs would be curtailed [270]. Despite efforts to stave off the development of resistance and avert yet another disaster, artemisinin resistance was reported in 2008, yet again on the Thai-Cambodia border [181, 64]. Resistance has since been reported at additional sites in Cambodia, in Thailand, Myanmar, Vietnam, Laos and recently in China [7, 38, 201, 126, 101, 16, 288, 107].

The term 'artemisinin resistance' represents partial resistance to artemisinin and its derivatives [288]. It is defined by 'delayed parasite clearance following treatment with an artesunate monotherapy, or after treatment with an artemisinin-based combination therapy (ACT).' [288]. People infected with artemisinin resistant parasites are curable if treated with an effective partner drug [288]. Nevertheless, the emergence of artemisinin resistance is viewed as a potential public health disaster [249], since alternative treatments with equivalent tolerability and efficacy are currently unavailable [284]. The public health ramifications of artemisinin resistance in Africa would be catastrophic.

In 2011, the WHO launched a Global Plan for Artemisinin Resistance Containment [283], followed by an emergency response in 2013 [284] and the Strategy for Malaria Elimination in the Greater Mekong subregion (GMS) in 2015 [289]. The aim being to eliminate malaria in the GMS, and therefore contain artemisinin resistance. Containment strategies to date have focused largely on preventing the spread of resistance. However, recent studies suggests resistance has emerged independently many times, with two main foci in Southeast Asia, suggesting differing factors, such as the genetic background of the parasites, may play a role [242, 154, 150]. Indeed, both Takala-Harrison *et al.* and Miotto *et al.* attribute the spread of resistance to newly emerging mutations [154, 242], raising fears of further emergence *de novo*. Understanding the factors that promote the emergence of *de novo* resistant foci is thus critical.

Western Cambodia is the epicentre of antimalarial resistance (CQ, SP and artemisinin resistance were first reported on the Thai-Cambodia border). Several factors are thought to play

a role in Cambodia's pivotal involvement, including the genetic background of the parasites [153], the fact that Cambodia has a relatively low level of transmission [273], and its history of indirect mass drug administration (MDA) using medicated salt [194, 263]. That is to say, low parasite diversity and transmission result in less opportunity for outcrossing [226, 97], while low transmission also results in more symptomatic infections on account of people developing less immunity [273], and so, under strong drug pressure, there is more opportunity for the emergence and propagation of drug resistant parasites in low transmission settings (although this may be confounded by intra-host competition [97]). Of note, campaign-level drug pressure (such as that induced by SMC, mass screening and treatment, and MDA) could also expedite the propagation of resistance by decreasing transmission, diversity and therefore outcrossing. As such, enduring parasites are likely to be the most resistant and it is critical that eradication campaigns are fully realised [148].

Cambodia was the first country to adopt ACT as first-line policy in 2001 [64, 305], followed by a ban of all artemisinin-based monotherapies in 2009 [283]. However, unlike most countries, artemisinin-based monotherapies had been available in Cambodia for almost 30 years prior to the ban [275]. Moreover, until recently, artemisinin-based monotherapies were still widely available: in 2008, 78% of treatments were artemisinin-based monotherapy according to survey by Yeung *et al.* [304]. A more recent survey has reported a sharp decline; however, the availability of poor quality drugs that contain sub-therapeutic concentrations of the active ingredients remains a problem [305]. Similar results were reported in neighbouring Laos: a dramatic reduction in the availability of artemisinin-based monotherapies (from 22.9% to 4.8%), but a high percentage of substandard drugs (25.4%) [239].

According to the WHO's most recent status report [288], artemisinin resistance is suspected if 10% or more patients remain parasitemic on day three [288]. The recommended threshold in Africa, where more people have acquired partial immunity, is 5% [295]. However, resistance has not yet been detected in Africa [288, 150], In contrast to Southeast Asia, the selection

of resistance is potentially forestalled in high transmission regions of Africa because there is more parasite diversity, more opportunity for outcrossing and, in comparison with Southeast Asia, proportionally fewer parasites exposed to drug pressure, due to the large reservoir of asymptomatic infections [274].

As indicated above, people infected with artemisinin resistant parasites can recover if treated with an effective partner drug [288], however doing so exposes the partner drug as a monotherapy, promoting the selection of resistance to the partner drug, which, in turn, promotes the selection of artemisinin resistance [288]. Hence artemisinin resistance not only compromises the effectivity of artemisinin and its derivatives, it also jeopardises the effectivity of the partner drugs. There is thus an urgent call for effective and timely surveillance of resistance to both artemisinin and its derivatives, and to partner drugs [72, 286].

1.2 Monitoring antimalarial resistance

While *in vivo* therapeutic studies are the 'gold standard' for measuring the clinical efficacy of an antimalarial drug [106], treatment failure is a late indicator of drug resistance, and is complicated by host factors [206]. Moreover, clinical trials are costly and time consuming, especially in resource poor settings [106]. Surveillance of genetic markers by blood sample survey can provide a complementary approach, which is both inexpensive and efficient by comparison [206]. Furthermore, since genetic markers are generally drug specific [188], in contrast to surveillance by *in vivo* therapeutic studies, genetic surveillance is often able to differentiate between resistance to the different components of combination therapies [217]. Had markers of artemisinin resistance been identified before reports of failing artesunate-mefloquine on Thai-Cambodia border [279], the role of artemisinin resistance, which was otherwise unresolved at the time because artesunate was used in combination, might have been identified earlier.

The research community has invested much effort into the identification of resistance

markers [198]. These efforts have been met with considerable success, and many markers have been used to retrospectively map the spread of resistance [216, 195, 166–168]. Until recently, most markers were identified long after resistance had spread, thus precluding their use for real-time genetic surveillance [217]. The identification of markers of artemisinin resistance ahead of widespread clinical failure marks a 'golden opportunity' for evidence based policy [217]. We are already seeing near real-time surveillance of the artemisinin markers identified thus far (see [16, 257, 297]).

1.2.1 Genetic markers of resistance

Markers of resistance are genetic variants in the parasite's haploid genome that have been associated with increased tolerance to antimalarial drugs *in vivo* and/or *in vitro*. They are the key determinants of the resistance phenotype.

Many markers of resistance comprise alleles at one or more non-synonymous SNPs (nSNPs) within genes that encode antimalarial drug targets. For example, PfCRT:76**T**, an amino acid encoded for by a marker of CQ resistance [82, 61, 62], denotes a amino acid change from lysine (K) to threonine (**T**) at position 76 in the protein PfCRT [82], which is the consequence of the nSNP in codon 76 of the gene *pfcrt*. PfCRT is a transporter protein found in the membrane of the parasite's digestive vacuole, where CQ is thought to act [69]. The PfCRT:K76**T** amino acid change has been shown to mediate the efflux of CQ from the digestive vacuole [69]. The mutant residue, PfCRT:76**T** has also been associated with decreased sensitivity to AQ [105], while the wild type, PfCRT:K76, has been associated with decreased sensitivity to lumefantrine [162], demonstrating an anticorrelated drug response.

Anticorrelated drug responses are also associated with copy number variants (CNVs) of the multidrug resistance 1 gene *pfmdr1*. More specifically, high *pfmdr1* CNVs are associated with decreased resistance to CQ and increased resistance to MQ, halofantrine, quinine and artemisinin [278, 12]. CNVs of *pfghc1*, the gene that encodes the first enzyme in the pathway

upon which SP acts, are also associated with drug resistance. It is not known if *ghc1* CNVs are the direct consequence of drug pressure or a compensatory mechanism, but they have been shown to be prevalent in Thailand, where SP was once the first-line antimalarial and the frequency of SP resistance markers is high, and low in Laos, where SP was seldom used and the frequency of SP resistance markers is low [169].

SP resistance markers include mutant alleles at multiple nSNPs in the genes *pfdhfr* and *pfdhps*, which encode dihydrofolate reductase (PfDHFR) and dihydropteroate synthase (PfDHPS), respectively (reviewed in [188]). As mentioned previously, SP is a fixed combination comprising two synergistic drugs, sulfadoxine and pyrimethamine. Both drugs act upon the folate biosynthesis pathway: sulfadoxine inhibits PfDHPS, whereas pyrimethamine inhibits PfDHFR. Haplotypes linking several mutant alleles at nSNPs in *pfdhfr* confer comparatively high level resistance to pyrimethamine [200, 54], while haplotypes linking several mutant alleles at nSNPs in *pfdhps* confer comparatively high-level resistance to sulfadoxine [34, 255]. Genotypes linking haplotypes across both genes on the haploid genome confer high levels of resistance to SP, both *in vitro* (reviewed in [228]) and *in vivo* [62, 264, 204, 205, 123, 243, 202].

Note that throughout this thesis, bold font is used to distinguish amino acid residues encoded for by mutant type alleles from those encoded for by wild type alleles. The term ‘SNP’ is used when referring to single nucleotide polymorphisms in general, while ‘nSNP’ is used only when referring to a specific non-synonymous mutation. The term ‘marker’ is used interchangeably to refer to variants associated with resistance, including alleles and multi-SNP haplotypes and genotypes, and their associated amino acid residues (such as PfCRT:76**T**), where multi-SNP haplotype refers to a sequence of alleles at multiple SNPs spanning a single chromosome (for example, markers of pyrimethamine or sulfadoxine resistance), and multi-SNP genotype refers to a sequence of alleles at multiple SNPs spanning two or more chromosomes (for example, markers of SP resistance).

The emergence of artemisinin resistance spawned an intensive effort to elucidate the

genetic basis of artemisinin resistance [249]. Many early studies identified putative markers on chromosome 13 [241, 43, 29, 153]. In 2014, Ariery *et al.* [14] identified several mutations in the *P. falciparum* propeller domain of a kelch gene on chromosome 13, *pfkelch13*, associated with resistance *in vivo* (delayed parasite clearance) and *in vitro* (delayed parasite survival rates). The association between *pfkelch13* mutations and artemisinin resistance has since been corroborated [154, 44, 242], and a plethora of *pfkelch13* mutations has been reported, including many mutations in Africa [288, 150]. However, only 13 mutations have been associated with resistance *in vitro* and/or *in vivo* thus far, and, of those associated with resistance, all are confined to Southeast Asia and China [154, 242, 288, 150, 297].

Natural *pfkelch13* variation obscures markers of resistance, presenting a major challenge for surveillance; effects associated with the genetic background also complicate matters. For instance, of the validated resistance-associated markers, the *in vitro* resistance of the most prevalent one in Cambodia depends upon its genetic background [236]. Furthermore, this marker is not the most resistant, suggesting that transmission potential and/or fitness may play a role [236], a hypothesis that is corroborated by the fact that, despite the many mutations identified, the probability of observing two or more on the same genome is extremely low [16, 265, 257, 150].

1.2.2 Blood sample surveys and prevalence data

As mentioned above, genetic surveillance of antimalarial resistance by blood sample survey can provide a timely adjunct to clinical efficacy trials. The objective of most blood sample surveys is to estimate the incidence of resistance in the within-host parasite population using blood samples collected from many infected individuals. Typically, a single filter paper blood sample is collected from each individual in a group belonging either to an asymptomatic or symptomatic cross section of the population (for example, [22]), or to a cohort enrolled in a clinical trial (for example [51]). Blood collection by filter paper is ideal for routine surveillance

in the field [203]. Filter paper blood spots are quick and simple to obtain (capillary blood is obtained by finger prick and spotted onto filter paper, which is air dried), storable (at 4°C, or −20°C for more than 90 days) and easy to transport (ideally, in a cooler) [299]. Having collected many samples, deoxyribonucleic acid (DNA) is extracted from each filter paper blood sample in preparation for genotyping. Typically, extraction is performed using one of two methods:

1. extraction by QIAamp® DNA Mini Kit (red blood cells are lysed enzymatically in a protease solution, which is then filtered, leaving purified DNA) [300];
2. extraction by heating in a suspension of Chelex® 100 (an ion chelator that protects the DNA from being destroyed by enzymes and heavy metals) [293].

Both of these methods are cheap and efficient. In general, they generate DNA suitable for PCR-based genotyping methods.

There are many different methods for PCR-based genotyping [76, 6, 145, 75, 115]. They all rely on PCR to amplify DNA fragments, but differ with respect to their products and the detection thereof. One common method, RFLP PCR, involves PCR amplification, followed by enzymatic digestion of the PCR products, then separation and identification of the digests by electrophoresis ([78, 294]). Providing the mutation creates (or destroys) a restriction enzyme digestion site, this method works because enzymatic digestion is allele-specific, thus digests have different lengths, depending on the presence or absence of the mutation. Different length digests carry different electric changes, so can be separated by electrophoresis, typically using a gel or capillary electrophoresis, which is more sensitive [92]. One can compare the electrophoretic results to a control, and thus identify the different digests. If electrophoresis is performed using a gel, the digest bands might vary in intensity. However, the intensity of the band does not necessarily correspond with the quantity of DNA [77]. Another commonly used method is SSOP-ELISA (see [22], for example), where sequence specific oligonucleotide probes (SSOP) hybridise to PCR products, and are detected by the enzyme-linked immunosorbent

assay (ELISA) [6]. As with the intensity of bands on a gel, the intensity of the ELISA output does not necessarily correlate with the quantity of DNA, since PCR amplification may be disproportional. Quantitative PCR (qPCR), on the contrary, does generate quantitative results [45], but qPCR is not optimised for high-throughput surveillance [240]. In summary, for a given SNP, PCR-based methods used in routine surveillance simply identify whether or not the blood sample tests positive for a given allele. Such data are known as prevalence data, since they capture allele prevalence at genotyped SNPs.

Typically, in a blood sample survey, the parasite DNA is genotyped at one or more SNPs associated with resistance to one or more antimalarial drugs. Investigators might also genotype MOI marker loci, we shall focus on markers of drug resistance, however. Recall that the parasites are haploid throughout the human phase of their lifecycle [267]. Consequently, upon genotyping DNA from a monoclonal infection, all genotyping outcomes ought to be unambiguous. However, in areas of high endemicity, multiclonal samples are commonplace [183, 23, 256, 140, 225]. If two or more clones differ at a genotyped SNP within a multiclonal sample, the corresponding genotyping outcome will be heteroallelic (providing both clones are detected); that is to say, differing alleles will be detected. An example of a small hypothetical dataset comprising six samples is shown in table 1.1. Three of the six samples are discernibly multiclonal based on one or more heteroallelic genotyping outcomes. When faced with data such as these, investigators customarily report allele-wise sample prevalence: the proportion of blood samples that test positive for specified alleles. According to table 1.1, the sample prevalences of alleles corresponding the amino acids PfDHFR:51**I**, PfDHFR:59**R** and PfDHFR:108**N** are $^4/_6$, $^4/_6$ and $^3/_6$, respectively. Allele prevalence is a population summary at the level of the infection. It is a measure of the probability of being infected with one or more clones carrying a specified allele [96]. It is a potentially informative metric of resistance, especially with a view to predicting treatment failure, since a single clone characterised by a single allele could potentially contribute to the clinical outcome. However, in order to monitor

Blood sample	PfDHFR:51	PfDHFR:59	PfDHFR:108
1	N	C	S
2	N	**R**	S
3	**I**	**R**	**N**
4	N & **I**	C	S
5	**I**	C & **R**	S & **N**
6	N & **I**	C & **R**	S & **N**

Table 1.1: A hypothetical dataset based on six blood samples genotyped at three nSNPs in codons 51, 59 and 108 of *pfdhfr* corresponding with amino acid changes PfDHFR:N51**I**, PfDHFR:C59**R** and PfDHFR:S108**N**, respectively, which are associated with resistance to pyrimethamine (reviewed in [205]). Amino acids encoded for by mutant alleles are highlighted in bold, while those encoded for by wild type alleles are not. Amino acid code: C, cysteine; I, isoleucine; N, asparagine; R, arginine; S, serine.

trends in parasite resistance across time and space, and/or to monitor resistance when its genetic basis is a haplotype or genotype, the metric of resistance must be defined at the level of the parasite. Frequency provides such a metric, whereas single allele prevalence does not. We explain why below, focusing first on the need for frequency when monitoring trends in parasite resistance (section 1.2.3), and then on the problem of reconstructing haplotypes and genotypes (section 1.2.4).

1.2.3 Frequency versus prevalence

Following Hastings *et al.* [96], frequency and prevalence are defined as follows,

Frequency The proportion of parasite clones in the parasite population that carry the marker.

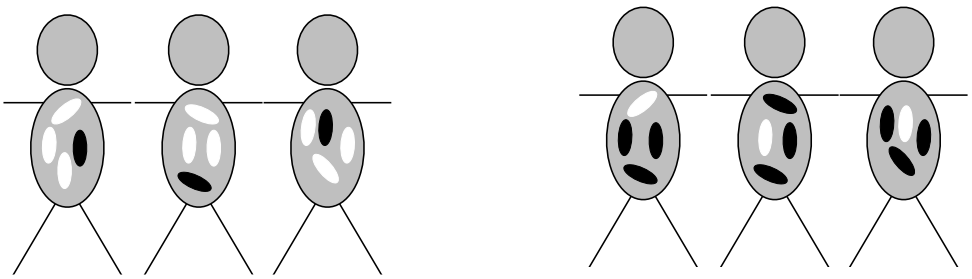
Prevalence The proportion of infections in the host population that test positive for the marker.

Marker prevalence is a summary at the level of the infection; it does not capture resistance at the level of the parasite clones. Moreover, prevalence is a function of both frequency and the MOI [95], hence may vary with changes in the average MOI, rendering it unsuitable as a comparable metric of parasite resistance across space and time. To illustrate these two points, let us consider two hypothetical examples. In the first, let us consider a scenario where the

prevalence is fixed while the level of resistance in the parasite population increases. In the second example, let us consider a scenario where, based on prevalence we incorrectly conclude that an intervention has decreased the level resistance, when, in fact, the level of resistance, as indicated by frequency, remains the same. For the sake of illustration, in both examples, let us assume the marker of resistance is a single allele, hence we have no problems associated with the reconstruction of haplotypes nor genotypes (addressed in section 1.2.4 below).

Example 1: Let us compare data collected from three people in a village before (figure 1.3a) and after (figure 1.3b) a change in treatment policy. Three of twelve clones are characterised by the resistant marker before the intervention, while nine of twelve clones are characterised by the resistant marker after. Hence, there has been a three-fold increase in the frequency of the resistant marker, from $3/12 = 0.25$ before the change in treatment policy to $9/12 = 0.75$ after. Based on prevalence we are unable to discern the impact of the change in treatment policy: all blood samples are positive for the resistant marker before and after the change; prevalence remains 1.00.

Example 2: Let us compare data collected from three people in a village before (figure 1.4a) and after (figure 1.4b) an intervention that lowers the transmission intensity, and therefore the average MOI. Three of twelve clones are characterised by the resistant marker before the intervention, while one in four is characterised by the resistant marker after. Hence, the frequency of the resistant marker before and after the intervention is $3/12 = 1/4 = 0.25$. Before the intervention, all blood samples are positive for the resistant marker, while only one is positive for the resistant marker after. Thus, the prevalence of the resistant marker decreases from 1.00 before the intervention to 0.33 after. Based on the decrease in prevalence, we incorrectly conclude that the intervention has an impact upon resistance. However, the change in prevalence is merely due to a drop in transmission, resulting in a



(a) Malaria infected people before a change in treatment policy.

(b) Malaria infected people after a change in treatment policy.

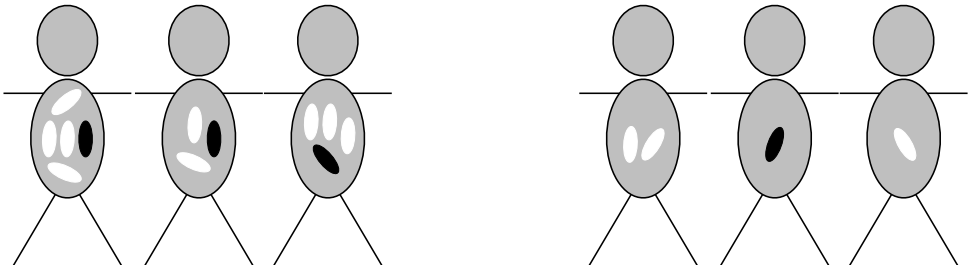
Figure 1.3: A cartoon of malaria infected people before and after a hypothetical switch in treatment policy. Each ellipse represents a unobserved malaria clone: black ellipses represent clones characterised by a marker of resistance, while white ellipses represent sensitive clones. The switch results in a three-fold increase in the frequency of the resistant marker from 0.25 before (1.4a) to 0.75 after (1.4b).

decrease in the average MOI, from 4.00 before the intervention to 1.33 after. The intervention had no impact of the proportion of resistant parasites. As noted by Alifrangis *et al.*, a scenario such as this might explain the increase in prevalence of the wild type markers observed in a village in Tanzania after the introduction of bed nets [5].

These two examples illustrate why prevalence should not be used to monitor spatial or temporal changes in parasite resistance [95, 96, 276]. Of course, sample estimates of allele prevalence can be reported alongside frequency, but to monitor parasite resistance across space and time, frequency should be reported. Estimates of frequency are not directly calculable, however, since the parasite clones (ellipses in figure 1.4) are not directly observed. Several solutions have been proposed to address this problem (outlined in section 1.3). First, let us consider the problem of reconstructing haplotypes and genotypes.

1.2.4 Reconstructing haplotypes and genotypes

Sometimes a sequence of alleles on the parasite genome is necessary for the elaboration of the resistant phenotype [228]. For example, parasites with the triple mutant residue marker



(a) Malaria infected people before a transmission-reducing intervention. The average MOI is $^{(5+3+4)}/_{3} = 4$.

(b) Malaria infected people after a transmission-reducing intervention. The average MOI is $^{(2+1+1)}/_{3} = 1.33$.

Figure 1.4: A cartoon of malaria infected people before and after a hypothetical transmission-reducing intervention. Each ellipse represents a unobserved malaria clone: black ellipses represent clones characterised by a marker of resistance, while white ellipses represent sensitive clones. The intervention results in a three-fold reduction in the average MOI from 4.00 (1.4a) to 1.33 (1.4b).

IRN (corresponding to amino acid residues encoded for by nSNPs in codons 51, 59 and 108 in *pfdhfr*) have been shown to be 16 to 32-fold more resistant than those with the single mutant residue marker NC**N** [199]. If the determinant of resistance is an allelic sequence, as in the aforesaid example, we need to reconstruct the haplotypes at the level of the parasites in order to investigate resistance.

Because parasites are haploid throughout the human phase of their lifecycle [267], reconstructing the sequences of alleles on the genomes of parasites residing in monoclonal infections is trivial. All one need do is genotype the SNPs of interest and reconstruct the sequence from the series of observed alleles. For example, the amino acid sequences encoded for by the haplotypes of the clones in samples one, two and three in the hypothetical dataset shown in table 1.1 are NCS, N**R**S and **IRN**, respectively, while sample four must contain clones with haplotypes encoding NCS and **I**CS, since only the outcome corresponding to the nSNP in codon 51 is heteroallelic (as indicated by the detection of both alleles encoding both amino acid residues). Determining the haplotypes when the constituent clones differ at two or more genotyped SNPs is non-trivial. Sample five, for example, could contain clones with haplotypes encoding **I**CS and **IRN**, or **I**C**N** and **IR**S, or a mixture of three, or a mixture of all four. All

possible haplotypes are compatible with sample six; it could contain clones with haplotypes encoding NCS and **IRN**, haplotypes encoding **I**CS and N**RN**, haplotypes encoding NC**N** and **IR**S, haplotypes encoding **I**C**N** and N**R**S, or a mixture of three, up to a mixture of all eight.

Given the high level of resistance associated with the triple mutant residue marker **IRN** [199], if this example were true, we would likely be interested in the prevalence and/or frequency of the haplotype encoding **IRN**. Some authors report the sample prevalence of the combination of markers, for example the proportion of samples that test positive for alleles encoding **I**, **R** and **N** (see [22], for an example with markers of SP resistance). But the prevalence of the alleles encoding a combination of markers **I**, **R** and **N** is not the same as the prevalence of the haplotype encoding **IRN**, since samples that test positive for alleles encoding **I**, **R** and **N** do not necessarily contain haplotypes encoding **IRN** (as explained above). Reporting the prevalence of the combination of markers can thus lead to confusion. For example, if one noted a 100% prevalence of the combination of alleles encoding **I**, **R** and **N** and zero incidences of clinical failure, one might conclude that this is evidence that **IRN** is not correlated with resistance *in vivo*, when in fact, the haplotype encoding **IRN** might not be present in the sample. Moreover to monitor changes in resistance in the parasite population across space or time, we need to estimate the frequencies of the haplotypes and genotypes (as explained in section 1.2.3 above).

1.3 Estimating haplotype and genotype frequencies

In order to monitor parasite resistance, we would like to estimate frequencies and be able to reconstruct haplotypes and genotypes. Frequency is a measure of resistance defined at the level of the parasite clones, many of which reside in multiclonal infections, particularly in high transmission settings. Using standard experimental methods, we cannot observe the individual clones within multiclonal infections (see section 1.1.3). To address the problem, several experimental protocols, counting methods and statistical models have been proposed; these are discussed in more detail below and in chapter 2.

1.3.1 Experimental methods

Blood collection by filter paper is the mainstay or routine genetic surveillance of antimalarial resistance. Most filter paper blood samples are analysed using non-quantitative PCR-based genotyping methods, generating prevalence data, which do not capture the clonal proportions within multiclonal infections. Methods that generate allele frequencies from multiclonal infections include real-time qPCR, pyrosequencing and high-throughput Illumina sequencing.

Real-time qPCR involves monitoring the amplification of the PCR products using clone specific primers. It has been shown to accurately resolve allelic proportions in biclonal infections of *Plasmodium chabaudi chabaudi* (a malaria parasite that infects mice) [45], and to identify major and minor alleles in artificial clonal mixtures of *P. falciparum* [57]. The qPCR method is highly sensitive [111], but not optimised for widespread surveillance of highly multiclonal infections. Pyrosequencing is a sequencing-by-synthesis method for genotyping short DNA fragments surrounding known genetic variants. It involves real-time monitoring of the sequencing reaction by tracking the intensity of a light signal that is emitted upon synthesis [240]. If calibrated, pyrosequencing can generate accurate allele frequency estimates from multiclonal *P. falciparum* mixtures, but a haplotype-reconstruction algorithm (described in detail in chapter 2) is required to reconstruct allelic sequences [240]. High-throughput Illumina sequencing generates millions of short template reads from which allele frequencies are directly attainable (see chapter 6 for more details). Despite rapidly decreasing costs [269], Illumina sequencing remains prohibitively expensive for large scale surveillance. Moreover, Illumina sequencing generally requires a comparatively large quantity of uncontaminated parasite DNA [37]. Obtaining a large quantity of uncontaminated parasite DNA requires a comparatively high-volume of infected blood, thereby necessitating blood collection by venipuncture [111], which is not readily deployable for routine surveillance, and technical expertise to decontaminate the blood. In a bid to reduce overheads, pooling protocols (whereby samples from different individuals in the survey are pooled) have been proposed [247]. Pooling provides low-cost

single-allele frequencies. However, like traditional PCR-based methods, short-read sequencing methods cannot resolve long-range allelic sequences, hence are not directly amenable for the surveillance of haplotypes and genotypes.

Experimental methods capable of distinguishing haplotypes and genotypes within multiclonal mixtures typically work by separating the clones. Clonal aggregates can be separated and the haplotypes ascertained using PCR cloning [70], yeast transformation [41] or single cell/molecular genotyping [170]. PCR cloning can be difficult and is labour intensive [180], while the method of transformation requires optimisation [41]. Moreover, both methods are prone to amplification bias [41, 98], impeding frequency estimation. Although technically feasible and potentially groundbreaking, single-genomics for malaria is extremely challenging at present [170]. Consequently, none of the clonal-separations methods currently available are amenable for routine surveillance. An alternative strategy, amplicon sequencing, has been used to resolve *P. falciparum* haplotypes spanning the C-terminal of the gene that encodes circumsporozoite protein in multiclonal infections [177]. Amplicon sequencing involves PCR amplification of targeted regions using customised probes. The amplicons are then tagged with uniquely identifiable indexes prior to sequencing [110]. This method is very powerful, and could be used more widely. Amplicon sequencing, and alternative methods generating increasingly long reads [127, 211], will ultimately render statistical haplotype frequency estimations methods redundant, although statistical methods will remain necessary for dealing with the errors generated by these methods and for estimating genotype frequencies where genotypes span genes on different chromosomes.

1.3.2 Counting methods

Prevalence is not the same as frequency (see section 1.2.3), but the two are often conflated [96]. Not only can this lead to wholly inaccurate frequency estimates (for example, in the figure 1.3a, the difference between the prevalence and frequency of the resistant marker is four-fold),

it can also lead to misleading assertions regarding haplotypes and genotypes. For example, the prevalence of the combination **I**, **R**, and **N** in table 1.1 is $^3/_6$. This does not mean to say the prevalence of the haplotype encoding **IRN** is $^3/_6$ (samples five and six need not contain haplotype **IRN**).

If the intensity of detection is graduated, investigators sometimes discount the minority allele at heteroallelic SNPs [9, 11, 86, 138, 37, 140]. Even if the genotyping method is quantitative, reconstructing sequences by discounting minority alleles can sometimes lead to spurious haplotype and genotype reconstruction. Consider the following example. Suppose that a person is infected with three clones with *pfdhfr*-51-59-108 haplotypes encoding NC**N**, N**R**S and **I**CS. The minority residues at positions 51, 59 and 108 are **I**, **R** and **N**, respectively. Discarding the minority residues thus leaves N, C and S, respectively. We would therefore score this sample as having a majority clone with haplotype encoding NCS, despite not one of the clones actually having the NCS marker. Discounting minority alleles can also lead to divergent trends between prevalence and frequency at a single allele. For example, in figure 1.4, we would discount the resistant clones in all three blood samples before the intervention (figure 1.4a) and count the resistant clone once after the intervention (figure 1.4b), leading to an increase in estimated frequency from 0.00 to 0.33, despite a decrease in prevalence from 1.00 to 0.33 and a true frequency of 0.25.

Another frequently documented approach is to discard samples that are discernibly multiclonal [96]. Doing so circumvents the problem of haplotype and genotype reconstruction, but leads to large losses of data in highly endemic regions, and is disposed to generate inaccurate frequency estimates, since rare variants are likely to be underrepresented in the remaining sample.

In summary, all of the conventional counting methods result in inaccurate and potentially biased estimates of frequency, they are also liable to generate spurious haplotypes and genotypes [96].

1.3.3 Statistical methods

To address the problems associated with counting methods and overcome the challenge presented by multiclonal infections, various statistical methods for estimating *P. falciparum* allele, and haplotype and genotype frequencies have been proposed [39, 102, 224, 129, 95, 276, 125, 223]. Statistical methods are desirable, since they make use of all the information in the data, while taking into account the uncertainty in the output, with no additional experimental cost. However, every model is built upon a set of assumptions and is restricted in its application by an inherent set of limitations. Indeed, as George Box famously said, 'all models are wrong, some are useful.' [31]. The merits and limitations of the existing statistical methods are discussed in detail in chapter 2.

1.4 Objective and outline

With a view to harnessing the full potential of genetic markers for the surveillance of antimalarial resistance in high transmission settings where multiclonal samples are common place, we sought to develop a statistical model that complements and builds upon existing methods for *P. falciparum* genetic marker frequency estimation.

The proposed model uses prevalence data and a single estimate of the average MOI to generate a posterior density estimate of single-allele and multi-SNP haplotype and genotype frequencies. It is presented in chapter 3. First, we present an exhaustive review of the statistical methods capable of generating population-level allele or multi-SNP haplotype and genotype frequency estimates that have been documented within the field of malaria (chapter 2). In chapters 4, 5 and 6, we demonstrate, respectively, the application of the model, its extension to account for repeat sampling within a cohort of children and its modification to analyse short-read sequencing data. The thesis ends with a discussion of the progress made and avenues for further work.

Chapter 2

Literature review

2.1 Statistical estimation of *P. falciparum* allele, haplotype and genotype frequencies using prevalence data

The difficulty in estimating *P. falciparum* allele, haplotype and genotype frequencies is due to the occurrence of multiclonal infections [276]. Statistical modelling of multiclonal malaria infections began in the 1950s, prior to the advent of molecular genetics. The British epidemiologist George Macdonald believed that people could be 'superinfected' (and thus harbour multiclonal infections) upon receipt of successive infectious bites ([137] described in [174]). Macdonald was interested in the effect of superinfection on rates of infection and recovery. Following Macdonald, the phenomena of superinfection was incorporated into several models of infection and recovery rates, using, for example, a Poisson process [178], a binomial distribution [173], or a Poisson distribution [174] to model the MOI. To the best of our knowledge, the first model designed to estimate the average MOI was also the first to generate *P. falciparum* allele frequencies [39]; it was published in 1973, before the confirmation of multiclonal infections, and is described in more detail below. Only a small number of statistical methods capable of generating *P. falciparum* allele, haplotype and genotype frequencies using prevalence data

have been described since; they are outlined in detail below.

2.1.1 Carter and McGregor, 1973

In 1973, Carter and McGregor published a paper featuring a model designed to estimate the average MOI and the frequencies of alleles at a single biallelic locus [39]. Carter and McGregor used prevalence data derived from the analyses of bivariate enzymes with electrophoretically distinct variants to study the genetic diversity of the within-host *P. falciparum* population in the Gambia. Note that under the model of Carter and McGregor, an enzyme with two variants is equivalent to a biallelic SNP.

Under the model of Carter and McGregor, the expected prevalence of a given allele is equal to one minus the probability of a blood sample containing zero clones characterised by the given allele, normalised by the probability of a zero valued MOI. The probabilities of a blood sample containing zero clones characterised by a given allele and of a zero-valued MOI are calculated assuming that the MOI is distributed according to a Poisson distribution, and that the variant of each clone within the blood sample is a Bernoulli variable with probability equal to its frequency. By solving equations for expected prevalence (described above) equal to observed prevalence for alleles one and two simultaneously, estimates of the allele frequencies and the average MOI are retrieved.

The pioneering approach of Carter and McGregor is limited to a single biallelic locus. Nevertheless, its publication paved the way for subsequent models of *P. falciparum* allele, haplotype and genotype frequencies.

2.1.2 Hill and Babiker, 1995

Hill and Babiker extended the model of Carter and McGregor to two multiallelic loci [102]. Hill and Babiker were interested in the rate of recombination in the *P. falciparum* population. The rate of recombination depends upon the likelihood of a mosquito ingesting multiple genetically

district clones, which, in turn, depends upon the average MOI [102]. Hence, the model of Hill and Babiker was presented with a view to estimating the average MOI. It is, however, capable of generating two-SNP haplotype frequency estimates.

Under the model of Hill and Babiker, the assumption that MOIs are either distributed according to a Poisson or negative binomial distribution is adopted. Zero is removed from the support of the MOI distribution if uninfected blood samples are excluded from the sampling procedure. Note that henceforth, we shall refer to distributions from which the zero valued support as been removed as non-zero conditioned. The likelihood is calculated as a function of the expected frequencies of the sample-wise genotyping outcomes and their respective counts. For a given MOI, an expression for the expected sample-wise genotyping outcome frequency is derived assuming the allele or haplotype of each clone within the corresponding blood sample is a Bernoulli variable with probability equal to its frequency. When two loci are considered, the haplotype frequencies are either modelled assuming linkage equilibrium (independence between loci) or disequilibrium, whereby each sequence of two alleles is attributed its own frequency. Since the MOI is unknown, the expected frequency is calculated by summing over the distribution on the MOI. Maximum likelihood estimates of the parameters of interest (the parameters of the distribution over the MOI and the allele or haplotype frequencies) are either available in closed form or found by iteration using the numerical method of Nelder and Mead [179][1].

Hill and Babiker applied the model to two datasets, each with two bi or triallelic loci based on prevalence data derived from the analyses of enzymes with electrophoretically distinct variants (akin to Carter and McGregor [39]). Using the standard log-likelihood ratio test, they found that the negative binomial distribution provided the best fit to data derived from the analysis of a single triallelic locus. They did not find evidence against linkage equilibrium between the alleles of the merozoite surface proteins 1 and 2, however, nor between

[1] The numerical method of Nelder and Mead is a downhill simplex routine for estimating the minimum of a multivariate function [179], such as a multivariate log-likelihood.

alleles of glucose phosphate isomerase and lactate dehydrogenase. The statistical estimates were experimentally corroborated using data derived from diploid *P. falciparum* oocysts. Statistical estimates of the average MOI and allele frequencies were in agreement with their experimentally-derived counterparts.

The major difference between the model of Hill and Babiker and its predecessor ([39]) is the ability to reconstruct the sequences of two-SNP haplotypes. The model by Hill and Babiker was the first to address *P. falciparum* haplotype frequency estimation. It requires an expression for each sample-wise genotyping outcome, limiting its application to a small number of loci, since both the number of expressions and their complexity increase exponentially with the number of loci.

2.1.3 Schneider *et al.*, 2002

In 2002, Schneider *et al.* published a paper featuring a Bayesian approach to *P. falciparum* allele frequency estimation [224]. The aim of the study was to ascertain the association between CQ resistance *in vivo* and markers of resistance at nSNPs in codons *pfcrt*-76 and *pfmdr1*-86.

Under the model proposed by Schneider *et al.*, the likelihood is calculated as a function of the expected frequencies of the sample-wise genotyping outcomes and their respective counts, akin to that of Hill and Babiker [102]. For a known MOI (equal to an experimentally-derived estimate), the expected frequency of a sample-wise genotyping outcome is calculated assuming that the allele of each clone within the blood sample is a Bernoulli variable with probability equal to its frequency. Since the two allele frequencies sum to unity, we need only estimate one. The model is constructed within a Bayesian framework, hence requires the specification of a prior distribution on the allele frequency. A uniform prior between zero and one is opted for. The posterior distribution of the allele frequency given the data is sampled using a Markov chain Monte Carlo (MCMC) sampler (see [90] for more details).

To the best of our knowledge, the model developed by Schneider *et al.* was the first

constructed within a Bayesian framework using an MCMC sampler. The Bayesian framework allows for full and formal treatment of the uncertainty in the estimates, while the MCMC algorithm generates a sample that approximates the posterior distribution and thus contains a wealth of information compared to a single point estimate. The approach of Schneider *et al.* was therefore innovative in the field of *P. falciparum* allele frequency estimation, despite the specified model being limited to data from a single biallelic SNP with known MOI.

2.1.4 Li *et al.*, 2007

In 2007, Li *et al.* [129] published details of the first fully statistical model designed specifically to estimate multi-SNP haplotype frequencies. The model is an extension of an earlier model by Excoffier and Slatkin [74]. The model of Excoffier and Slatkin was designed to estimate haplotype frequencies in a diploid population. Estimating haplotype frequencies in a diploid population is analogous to estimating *P. falciparum* haplotype frequencies from a population of individuals each infected with exactly two clones. Li *et al.* extended the model by Excoffier and Slatkin by allowing individuals to be infected with a variable number of clones.

Under the model by Li *et al.*, each sample-wise genotyping outcome is modelled by a haplotype count summarising the haplotypes of the unobserved clones within the infection. The haplotype count vector is modelled as a realisation of size equal to an unobserved MOI from a multinomial distribution. The probability vector of the multinomial distribution is set equal to the haplotype frequencies. The unknown MOIs are modelled according to either an unconditioned or non-zero conditioned Poisson distribution, depending on whether or not samples from non-infected individuals are included in the sampling design of the data, respectively.

Maximum likelihood estimates of the haplotype frequencies are found using the expectation maximisation (EM) algorithm (formalised by Dempster *et al.* [59]). The EM algorithm is commonly used to estimate parameters in models involving unobserved variables (such as the

haplotype counts and MOIs) [63]. It breaks the problem down by iterating over two steps: first, the distribution over the unobserved variables given the current estimate of the parameters is approximated by calculating expected probability densities; second, the parameter estimate is updated by maximising the log-likelihood given the expected probability densities from the previous step. Under the model of Li *et al.*, solutions to the step-wise calculations are either available in closed form or found using the Newton-Raphson method (an iterative method for maximising a function [159]). Confidence intervals are constructed using an estimate of the asymptotic variance-covariance matrix of the haplotype frequencies, which is computed within the EM framework. Given the variance-covariance matrix, one can also test hypotheses. Li *et al.*, for example, tested the hypothesis that frequency estimates for a given haplotype are the same in Uganda, Cameroon and Sudan.

The model by Li *et al.* was embedded within a generalised linear model designed to test associations between haplotypes and clinical traits [130]. It was later incorporated into an R package called `malaria.em` [131], which, unfortunately, is no-longer available. Ross *et al.* demonstrated the use of `malaria.em`, using it to obtain unbiased estimates of the average MOI from simulated data on a single highly polymorphic locus [219].

The publication of the model by Li *et al.* marked an important development in *P. falciparum* haplotype frequency estimation. The major advantage of the method is its ability to analyse multiple SNPs (up to ten demonstrated in [129]) and/or highly multi-allelic SNPs (see [219]) in a computationally efficient manner. The efficiency follows from the use of the EM algorithm to iteratively average over latent variables. Moreover, the likelihood is guaranteed to monotonically increase under the EM algorithm. However, this also means the algorithm is in danger of getting stuck on a local maximum. The danger of getting stuck can be mitigated and assessed by running the algorithm multiple times from different starting points. The use of an alternative algorithm, such as an MCMC sampler, would also mitigate the prospect of getting stuck, since the likelihood does not monotonically increase under an MCMC sampler. A MCMC sampler

does not negate the need to run the algorithm multiple times from different starting positions to check convergence, however.

2.1.5 Hastings and Smith, 2008

As early as 1998, Hastings and Smith recognised the need for an open-source software for malaria haplotype frequency estimation and so began the development of a program, MalHaploFreq, which was launched in 2008 (see [95], additional file 1). In contrast to the preceding methods [39, 102, 224, 129], Hastings and Smith put onus on the incorporation of an indicator to account for imperfect detectability of minority clones [95].

The default model underpinning MalHaploFreq assumes prevalence data are accompanied by sample-wise MOI estimates, akin to the scenario considered by Schneider *et al.* [224]). For all samples with a given MOI, the vector of observed sample-wise genotyping outcome counts is modelled as a realisation from a multinomial distribution of size equal to the number of samples with the specified MOI value. The probability vector of the multinomial distribution is equal to the expected frequencies of the sample-wise genotyping outcomes. For a given sample-wise genotyping outcome, the expected frequency is calculated by summing over all compatible haplotype count vectors given the specified MOI[2]. The probability of an unobserved haplotype count vector is captured by two terms. First, a probability density assuming the haplotype count vector is a realisation of size equal to the specified MOI from a multinomial distribution with probability vector equal to the haplotype frequencies. Second, an indicator variable, which is equal to one if the haplotype combination is compatible with the sample-wise genotyping outcome given a user-specified threshold for detectability, and zero otherwise. The likelihood of the entire dataset is calculated by summing over the multinomial densities of the sample-wise genotyping outcome counts for each observed MOI level.

[2]Note that by summing over all compatible haplotype count vectors given all possible MOIs, it is theoretically possible to fit a modification of the model without knowing the MOIs [95]. Under the modified model the MOIs are modelled according to either a negative binomial, Poisson or zero-conditioned Poisson distribution. Since the analyses using said modification have not been validated, we proceed assuming a specified MOI.

MalHaploFreq generates maximum likelihood estimates of the haplotype frequencies using a hill climbing routine, which proceeds as follows. First the probability of the dataset under the model given an initial guess at the haplotype frequencies is calculated. The initial guess is then perturbed and the probability recalculated. If the dataset is more probable given the new frequencies, the frequencies are retained, otherwise the initial guess is considered the current best estimate. The process is reiterated until the algorithm finds the set of haplotype frequencies that produces the greatest probability of observing the dataset. Confidence intervals are constructed using the profile log-likelihood.

Two key distinctions set MalHaploFreq apart from the existing methods [39, 102, 224, 129], first its ability to deal with imperfect detectability, and second its availability as an open-source software. The computational expense of summing over all possible haplotype count vectors limits its application to three or fewer SNPs, however, while the hill climbing routine is computationally inefficient.

2.1.6 Smith and Penny, 2008

In the current version of the MalHaploFreq user manual ([95], additional file 1)[3], an alternative model courtesy of Smith and Penny is described. Under the model of Smith and Penny, the likelihood is calculated as a function of the expected frequencies of the sample-wise genotyping outcomes and their respective counts. The expected frequency of a given sample-wise genotyping outcome is an algebraic expression derived under the assumption that the allele or haplotype of each clone within an infection is a Bernoulli trial with probability equal to its frequency, akin to Carter and McGregor [39], Hill and Babiker [102] and Schneider *et al.* [224], but with extension to three SNPs. The algorithm by Smith and Penny was not incorporated into MalHaploFreq, however, as it is cumbersome to implement for more than two loci, for the same reasons noted by Hill and Babiker [102].

[3] At the time of writing, the MalHaploFreq user manual ([95], additional file 1) corresponds with MalHaploFreq version 1.1.1.

2.1.7 Wigger *et al.*, 2013

In early 2013, Wigger *et al.* published a Bayesian method specifically constructed for *P. falciparum* multi-SNP haplotype frequency estimation using prevalence data [276]. The model of Wigger *et al.* was principally designed for prevalence data derived from microarray SNP assays. As such, Wigger *et al.* put onus on the inclusion of a random error term representing the probability of miscalls (when one allele is misinterpreted for another), which is set equal to a fixed quantity based on experimental estimates. The authors also sought to address the triple-SNP limitation of MalHaploFreq, using a computationally efficient MCMC algorithm. Wigger *et al.* model the sample-wise genotyping outcome data given an experimentally-derived estimate of the MOI, which is treated as fixed (as in [95] and [224]).

The model of Wigger *et al.* is described in terms of a multinomial mixture model. Each infection is modelled as a clonal conglomerate characterised by a vector of counts describing a collection of potentially error-ridden haplotypes. The probability of an error-ridden haplotype is calculated assuming that it is an imperfect copy of an error-free haplotype, where each allele copied is a Bernoulli variable with probability equal to the probability of a miscall. In the parlance of mixture models, the error-free haplotypes are the mixture components, while the density of the component distribution is equal to the probability of the error-ridden haplotype given the error rate and the error-free haplotype. For a given sample-wise genotyping outcome, the vector of error-free haplotype counts is modelled as a realisation, of size equal to the specified MOI, from a multinomial distribution, whose probability vector is equal to the haplotype frequencies. A uniform Dirichlet prior is placed over the error-free haplotype frequencies, taking advantage of the fact that the Dirichlet is conjugate to the multinomial distribution.

An MCMC sampler is used to sample from the posterior density of the error-free haplotype frequencies in an computationally efficient manner. More specifically, a Gibbs sampler is used, a type of MCMC algorithm whereby samples are drawn from the target distribution

by iteratively updating the parameters in blocks [90]. In each block, the updated parameters are sampled from their full conditional distribution given all parameters outside the block at their current values. Under the sampler of Wigger *et al.*, there are three blocks: error-ridden haplotypes, error-free haplotypes and haplotype frequencies. In the first block, for each sample, a collection of error-ridden haplotypes of size equal to the experimentally-derived MOI is drawn from the set of haplotypes compatible with the observed sample-wise genotyping outcome. Doing so does not ensure the collection complies with the observed data (for example, all haplotypes are theoretically compatible with an entirely heteroallelic sample-wise genotyping outcome, but a random draw of two identical haplotypes does not comply with heteroallelic genotyping outcomes); thus, incompatible draws are resampled. In the second block, for each sample, the vector of error-free haplotype counts is updated conditional on the newly updated error-ridden haplotypes and the existing error-free haplotype frequencies. In the third block, the error-free haplotype frequencies are updated by drawing from a Dirichlet distribution whose parameter vector is based upon the updated error-free haplotype counts.

The Gibbs sampler generates an MCMC sample that approximates the full posterior distribution of the frequencies and haplotypes (both error-ridden and error-free). Wigger *et al.* compared analyses of simulated data under the model with and without error. Inclusion of the random error was shown to be favourable for errors that exceed 1–2%. The results of the models with and without random error applied to real data were not statistically different unless a large number of samples (in excess of 500) were analysed.

The main advantages of the model by Wigger *et al.* are its ability to deal with up to seven SNPs, the inclusion of a random error term (albeit fixed and equal for all SNPs), the computationally efficient implementation, and its construction within a Bayesian framework, allowing for full posterior summaries, in contrast to singular point estimates. Its main limitation is the dependency on SNP-wise MOI estimates, which are not always readily available.

2.1.8 Kum *et al.*, 2013

In late 2013, Kum *et al.* published a model from which *P. falciparum* multi-SNP haplotype frequencies can be derived [125]. The model was designed with a specific application in mind: to compare triple-SNP *pfdhfr* haplotype prevalence (the probability of being infected with one or more clones characterised by a particular haplotype) before and after treatment with two different antimalarial drugs in two different clinical trial sites. Although single allele prevalence is easy to estimate (by simply calculating the proportion of blood samples that test positive for a given allele), haplotype prevalence is not, because of the problem of reconstructing the allelic sequences of the haplotypes. Haplotype prevalence estimates are therefore generated under a model. However, estimates of frequencies and the average MOI (referred to as the expected number of infection times by Kum *et al.*) can be obtained as byproducts of the haplotype prevalence model.

Akin to Carter and McGregor [39], Hill and Babiker [102], Schneider *et al.* [224] and Hastings and Smith [95], under the model by Kum *et al.*, the likelihood is calculated as a function of the expected frequencies of the sample-wise genotyping outcomes and their respective counts. Unlike the aforementioned models, for a given sample-wise genotyping outcome, the expected frequency is calculated as a function of haplotype prevalence, not frequency. The functions are derived from combinatorial expressions, by first assuming the probability of *not* being infected with a clone characterised by a specific haplotype is one minus the haplotype prevalence. Kum *et al.* showed that, by assuming a Poisson distribution over the MOI, one can equate the probability of not being infected with a clone characterised by a specific haplotype with the density of a zero-valued realisation from a Poisson distribution with a haplotype-specific parameter; algebraic expressions for the haplotype frequencies and the average MOI follow.

Maximum likelihood estimates of the haplotype prevalences are generated using the method of Nelder and Mead [179]. The model does not require sample-wise MOI estimates, however,

akin to the aforesaid methods [39, 102, 224, 95], it is limited to three SNPs.

2.1.9 Schneider and Escalante, 2014

Akin to Carter and McGregor and Hill and Babiker [102], Schneider and Escalante [223] proposed a model with a view to estimating the average MOI. The model by Schneider and Escalante is an extension of that of Hill and Babiker in the direction of the number of alleles at a single multiallelic locus, rather than multiple loci. It was not designed to analyse multiple loci simultaneously, hence does not generate multi-SNP haplotype frequencies.

Akin to the majority of the foregoing models [39, 102, 224, 95, 125], under the model of Schneider and Escalante the likelihood is a function of the expected frequencies of the sample-wise genotyping outcomes and their respective counts. An expression for the expected frequency is derived by first assuming the expected frequency it is equal to a summation over the distribution of unobserved allele counts and MOIs compatible with the observed allele prevalence. Schneider and Escalante assume a non-zero conditioned Poisson distribution over the MOIs, while describing how the framework might be generalised under different assumptions. They assume the allele counts are realisations, of size equal to the MOI, from a multinomial distribution whose probability vector is equal to the vector of allele frequencies. Maximum likelihood estimates and confidence intervals are found using Newton-Raphson's method (the EM algorithm and least-squares were proposed as alternatives).

Schneider and Escalante applied their model to three datasets on MOI markers from Kenya, Cameroon and Venezuela. Confidence intervals were used to test the statistical significance of differences between MOI estimates based on different marker loci. Allele frequencies were not reported since they were treated as nuisance parameters (the objective being to estimate the MOI). In summary, this method is capable of analysing highly polymorphic SNPs, but is limited to only a single locus.

2.1.10 Overview

In total, we have identified eight published articles and nine methods either designed for or capable of generating population-level *P. falciparum* frequency estimates using prevalence data [39, 102, 224, 129, 95, 276, 125, 223]. Three of these methods are restricted to prevalence data from a single locus [39, 224, 223], while the remaining six address the problem of phasing haplotypes and genotypes [102, 129, 95, 276, 125].

Of the three methods restricted to prevalence data restricted to a single locus [39, 224, 223], two were designed with a view to generating estimates of the average MOI [39, 223]. The most recent, [223], is capable of analysing a highly-polymorphic locus.

Of the six models that address the problem of phasing, two are capable of analysing data from more than three SNPs in a computationally efficient manner [129, 276]; they are summarised in table 2.1. Both of these methods approach the problem as one of data augmentation, using latent variables to model unobserved haplotype counts. The number of haplotype count vectors compatible with a given sample-wise genotyping outcome grows exponentially with the number of SNPs. Consequently, summing over the unobserved haplotype counts is computationally expensive beyond three SNPs. To avoid explicit summation, both algorithms iteratively average over the latent variables, either by iteratively recalculating their expected probability densities [129] or by recursive sampling [276]. With the exception of MalHaploFreq, the remaining methods [39, 102, 224, 125, 223] derive an algebraic expression for the expected frequency of each sample-wise genotyping outcome under an assumed model. The likelihood is then a function of the expected frequencies of the sample-wise genotyping outcomes and their respective counts. The number of sample-wise genotyping outcomes grows exponentially with the number of SNPs, hence the aforesaid methods are limited to three of fewer SNPs.

The likelihood underpinning MalHaploFreq (also summarised in table 2.1) is also a function of the expected frequencies of the sample-wise genotyping outcomes and their observed counts. However, given a specified MOI, the expected frequencies of the sample-wise genotyping

	Li *et al.* 2007 [129]	Hastings and Smith 2008 [95]	Wigger *et al.* 2013 [276]
Input	Genotyping outcomes derived from prevalence data on ≈ 10 or fewer multiallelic SNPs.	Genotyping outcomes derived from prevalence data on ≤ 3 biallelic SNPs plus MOIs.	Genotyping outcomes derived from prevalence data on ≈ 7 of fewer biallelic SNPs plus MOIs.
Missing data (genotyping failures) and errors	No mention of either.	Samples with genotyping failures ignored. Option to account for imperfect detectability of minority clones given a user-defined detectability threshold.	No mention of genotyping failures. SNP miscalls are modelled using a fixed error.
Likelihood	The observed sample-wise genotyping outcomes are modelled given unobserved haplotype counts, which in turn are modelled as realisations from a multinomial distribution with size equal to an unobserved MOI and probability vector equal to the haplotype frequencies.	For a given MOI value, the observed genotyping outcome counts are modelled as realisations from a multinomial distribution whose size is equal to the number of samples with MOI equal to the specified value, and whose probability vector is calculated by summing over the probabilities of the haplotypes counts possible given the specified MOI value. The haplotype counts are modelled as realisations of a multinomial distribution with size is equal to the MOI and probability vector equal to the haplotypes frequencies, multiplied by an indicator variable that accounts for imperfect detectability.	The sample-wise genotyping outcomes are modelled given error-ridden haplotypes, which in turn are modelled as imperfect copies of error-free haplotypes given a fixed error. Counts of the error-free haplotypes are modelled as realisations of a multinomial distribution with size equal to the MOI and probability vector equal to the error-free haplotype frequencies.
Model on MOIs	Several options. Of those relevant to malaria, the MOIs are either assumed to be realisations from a Poisson distribution or from a non-zero conditioned Poisson distribution if non-infected individuals do not feature in the dataset.	Assumed given and fixed (theoretically, it is possible to sum over unknown MOIs, in which case the MOIs are modelled as realisations from either a negative binomial, Poisson or zero-conditioned Poisson distribution, but this option has not been validated).	Assumed given and fixed.
Model on frequencies	Frequencies are treated as parameters of a multinomial distribution.		A uniform Dirichlet prior is used to model the frequencies, which in turn are treated as parameters of a multinomial distribution.
Implementation	EM algorithm with solutions to the step-wise calculations available either in closed form or via the Newton-Raphson method.	Hill climbing algorithm that evaluates the log likelihood upon perturbing the frequencies.	Gibbs sampler with rejection sampling.
Output	Maximum likelihood estimates of the haplotype counts and frequencies, and the parameters of the model over the MOIs. Confidence intervals are constructed using an estimate of the asymptotic variance-covariance matrix of the haplotype frequencies, which is calculated within the EM framework.	Maximum likelihood estimates of the frequencies with approximate confidence intervals constructed using the profile log-likelihood.	Posterior density estimates of the sample-wise haplotypes (with and without errors) and frequencies.

Table 2.1: A table summarising the three most competitive statistical multi-SNP haplotype and genotype frequency estimation methods published in the field of malaria.

outcomes are calculated by marginalising out all possible haplotype count vectors compatible with the observed data. Hence, MalHaploFreq treats unobserved haplotype counts as latent, akin to Li *et al.* and Wigger *et al.* [129, 276], but does not avoid explicit summation, rendering MalHaploFreq computationally less efficient. Despite being computationally less efficient, MalHaploFreq is the only method with functioning open-source software, and it has been used by numerous investigators to estimate haplotype and allele frequencies (for example, [256, 144, 182, 246, 51]).

All the models described above either generate posterior density estimates [224, 276] or maximum likelihood estimates [39, 102, 129, 95, 125, 223]. Maximum likelihood estimates are point estimates [277]. Typically, confidence intervals surrounding the maximum likelihood estimates are derived using either the profile-likelihood or asymptotic assumptions of normality. The paper by Schneider and Escalante includes detailed derivations and proofs for expressions relating to the maximum likelihood estimates (proving both the existence and uniqueness of the estimates under the non-zero conditioned Poisson distribution), confidence intervals (both profile-likelihood and asymptotic), as well as three different hypothesis tests (one based on the profile-likelihood confidence intervals, and two based on the asymptotic confidence intervals).

The methods of Schneider *et al.* [224] and Wigger *et al.* [276] output posterior density estimates using an MCMC sampler. Posterior density estimates approximate distributions, hence capture more information than a singular point estimate. They are generated within a Bayesian framework, the construction of which provides a straightforward yet comprehensive treatment of uncertainty [88]. The Bayesian framework also enables the incorporation of specialist knowledge [88]. Typically, MCMC samplers are used to sample from target distributions that do not belong to standard families of distributions [90]. As we have seen for the model by Wigger *et al.*, recursive sampling within the MCMC scheme efficiently averages over latent variables. Posterior summaries, including point estimates and credible intervals (which can be used to assess statistical significance), are readily available given the MCMC sample set.

2.2 Related work

2.2.1 Statistical MOI estimation

As we have seen already, *P. falciparum* allele and multi-SNP haplotype and genotype frequency estimation is inherently linked to *P. falciparum* MOI estimation [219]. In fact, all of the aforesaid models that do not rely on experimentally-derived MOI estimates, were either designed specifically to estimate the average MOI [39, 102, 223] or are capable of generating average MOI estimates as a byproduct of frequency estimation [129, 125]. In general, standalone MOI estimation is less complex than multi-SNP haplotype frequency estimation, because allelic sequences do not require phasing [219]. Allelic sequences do not require phasing because MOI estimates are normally based on either one highly polymorphic locus (see, for example, [223]), or on a panel of independent SNPs (see, for example, [84]). Statistical models designed specifically to estimate within-sample MOIs include estMOI [17] and COIL [84]. Both estMOI and COIL are designed to estimate sample-wise MOIs using genome-wide data. The model underpinning estMOI generates a genome-wide average based on local multi-SNP haplotype phasing. It is therefore related to the models for multi-SNP haplotype phasing in malaria. It does not generate frequency estimates, however, and requires whole genome sequencing data with a high density of SNPs. The model underpinning COIL does not address the problem of phasing. Instead, MOI estimates are based on data from a panel of 96 independent biallelic SNPs, ideally, with minority allele frequencies equal to 0.4. If the frequencies are unknown (if the data are prevalence data, for example), one can model them using an uninformative prior. The implementation of COIL therefore has the capacity to generate *a posteriori* sample-wise single allele frequency estimates. This is not the intended purpose of the model, however, since there is very little information in the prevalence of a single SNP in a single sample. Neither estMOI nor COIL take into imperfect detection due to complex infection dynamics. MOI estimates that do take into account detectability can sometimes be derived from models of

within-host infection dynamics (for example, [229, 221, 222]). Realistic models of within-host dynamics are comparatively complex [46], however, and they do not lend themselves directly to allele, haplotype and genotype frequency estimation.

2.2.2 Statistical within-sample haplotype frequency estimation

Another closely related area of research is within-sample *P. falciparum* haplotype frequency reconstruction. In fact, to the best of our knowledge, the first high-throughput *P. falciparum* 'haplotype-estimating algorithm' was designed to reconstruct within-sample *P. falciparum* six-SNP haplotypes in multiclonal samples collected from a vaccine-testing site in Mali. It was published by Takala *et al.* in 2006 [240]. We do not include it above because it relies heavily on an experimental protocol requiring pyrosequencing data. More recently, O'Brien *et al.* published a model designed to infer the identity of a set of haplotypes in a population, their joint phylogeny and their within-sample frequencies using short-read deep sequencing data derived from multiple metagenomic samples [185]. The model was applied to data derived from *P. falciparum* plastids extracted from samples collected in Ghana, as well as data from green sulfur bacteria and Neisseria meningitidis bacteria. It is broken down into two components, a phylogenetic component (details of which are beyond the scope of this chapter), and within-sample haplotype frequency estimation. The method of Takala *et al.* and the haplotype frequency estimation of O'Brien *et al.* are described in more detail below.

Takala *et al.*, 2006

As indicated above, the method of Takala *et al.* [240] relies heavily upon pyrosequencing data, but is partly based upon a statistical model. In summary, for a given blood sample, the likelihood of a proposed vector of haplotype counts is calculated given the observed data and a set of reference haplotypes. The reference set is determined by PCR cloning of blood samples in which only one haplotype sequence has been detected. The likelihood is based

on the difference between the observed and expected allele frequencies, where the former are determined by pyrosequencing and the latter are computed given the proposed haplotype counts. The smaller the difference, the bigger the likelihood that the proposed vector of haplotype counts represents the true collection of clones in the blood sample. Because the accuracy of pyrosequencing varies with both the allele's identity and its frequency, the raw-observed frequencies are calibrated using standard curves. Standard curves are constructed using clones from the reference haplotypes. To account for residual variation in the calibration, the difference is normalised by the mean residual error of the standard curve. The likelihood function is equal to the sum over the normalised differences minus the probability density of the MOI of the proposed haplotype count vector, which is calculated assuming a Poisson distribution over the MOI minus one (to account for non-zero realisations). The maximum likelihood is found by cycling over all haplotype count vectors compatible with the observed data. For a given sample, the proposed haplotype count vector corresponding to the maximum likelihood represents the most probable collection of clones given the observed data and the reference set. Multiple proposals may have the same likelihood given the observed data and the reference set. In other words, there is sometimes a problem of identifiability. Takala *et al.* validated their model experimentally using artificial clonal mixtures with known combinations of haplotypes. The method works well given an MOI less than or equal to three, but breaks down thereafter on account of the problem of identifiability.

This method lends itself to accurate within-sample haplotype reconstruction given experimentally-derived allele frequencies and an MOI less than or equal to three. Its main advantage over alternative methods lies in its thorough treatment of experimental error, namely the use of standard curves to correct for variation in the accuracy of the experimentally-derived allele frequencies. Its use for population-level frequency estimates is not optimal: it does not borrow information across samples. Borrowing information across samples would help address the identifiability problem. To generate a reference set of haplotype clones, this method requires

PCR cloning, which is expensive and time consuming. Moreover, the reference set is generated using blood samples in which only one haplotype is detected. In a high transmission setting, the construction of a reference set in this way might overlook rare haplotypes, since rare haplotypes are unlikely to be found in isolation, especially if the sample-wise MOI exceeds one.

O'Brien *et al.*, 2014

Under the model of O'Brien *et al.* [185], each multiclonal malaria infection is modelled as a meta-genomic sample. For a given sample, the data comprise wild and mutant read counts for a set of biallelic SNPs. For a given SNP within a sample, the mutant read count is modelled as a realisation, of size equal to the total read count, from a binomial distribution with probability equal to an error-prone mutant allele frequency. The error-prone mutant allele frequency is calculated in two stages as follows. First, the error-free allele frequency is calculated by summing over the allele states of the haplotypes at the specified SNP weighted by their corresponding sample frequencies. The error-prone frequency is then calculated assuming the probability of a miscalled read and the error-free frequency. Conditional upon the allele sequences of the haplotypes, O'Brien *et al.* assume independence between samples. The allele sequences of the haplotypes are modelled using a phylogenetic tree with constant mutation rate and noise (to account for recent mutation and rare variants, for example). Inference is performed within a Bayesian framework. The within-sample haplotype frequencies are modelled using a uniform Dirichlet prior. Samples are drawn from the full posterior distribution using an MCMC algorithm. The model is fit assuming a fixed number of haplotypes, then refit assuming an alternative number. The number that provides the best fit to the data is chosen using a model selection criterion.

Compared with the models described in this chapter hitherto, the model by O'Brien *et al.* is relatively complex due to the phylogenetic component. However, the noise in the phylogenetic tree allows O'Brien *et al.* to analyse 1000s of SNPs assuming a small number (≤ 20) of

haplotypes. Moreover, the noise in the tree allows variation due to recent mutation in the taxa. This is especially important for multi-genomic samples containing organisms that mutate on a timescale close to that of the duration of infection (see the section on tumour diversity and quasispecies spectrum reconstruction below).

2.2.3 Statistical methods beyond malaria

The problem of estimating of *P. falciparum* multi-SNP haplotype frequencies for antimalarial surveillance is related to several other fields of research including haplotype assembly (reviewed in [36]), characterising tumour diversity (for example, [237, 109, 3]), quasispecies spectrum reconstruction (reviewed in [18]) and metagenomics (for example, [185]). All of these fields are united in their aim to estimate the frequencies and/or reconstruct the sequences of genetically distinct components within a mixture. There is no general purpose method, however, since each has its own set of challenges.

All of the aforementioned fields of research collectively differ to that of estimating frequencies for routine antimalarial surveillance on account of the data. In the field of antimalarial surveillance, the usual aim is to characterise diversity using prevalence data derived from blood sample surveys. There is little information in prevalence data from a single sample to inform within sample frequencies, but, by borrowing information across multiple filter-paper blood samples, inference is possible at the level of the population. Typically, the aim in the aforesaid related fields is to estimate diversity within a sample of genetically distinct components using next generation sequencing data. In fact, it is the ability to rapidly generate lots of reads from a single sample at low cost that has made the analysis of within-sample diversity possible [18]. Unfortunately, next generation sequencing requires a comparatively large volume of uncontaminated DNA, rendering it impractical for routine surveillance of antimalarial resistance.

Haplotype assembly

The problem of reconstructing haplotypes is a standalone field known as 'phasing' or haplotype assembly [36]. Typically, the goal of haplotype assembly is to reconstruct long-range haplotypes for a given organism with fixed ploidy (for example, [26]). The problem of phasing samples derived from multiple multiclonal *P. falciparum* infections is equivalent to that of phasing multiple polyploid organisms with unknown variable ploidy. The fact that the ploidy is known and fixed in most haplotype assembly applications, renders the problem more tractable. Another major difference is the number of SNPs analysed. The goal of most phasing methods is to reconstruct haplotypes across hundreds of thousands of SNPs or the entire genome, necessitating highly scalable methods, that borrow information across large sample sets [36]. In the field of antimalarial resistance surveillance, we are primarily interested in the frequency of a marker corresponding to a single SNP (for example, PfCRT:76**T**) or multi-SNP haplotypes and genotypes (for example, markers of resistance to sulfadoxine-pyrimethamine (SP) [168]). In fact, the ideal genetic marker of antimalarial resistance would be a single allele at a decisive SNP [61].

Tumour and viral quasispecies spectrum reconstruction

A viral quasispecies refers to a family of closely related viral variants within an infected host [18]. Viral diversity evolves within the host resulting in a family of related offspring, some members of which may withstand antiviral therapy or contribute to vaccine failure [18]. In oncology, diversity also evolves within the individual, sometimes leading to cell lines that are resistant to anticancer agents. The problem of reconstructing tumour diversity and the quasispecies spectrum is that of reconstructing the genomic sequences of the related variants and estimating their relative abundance using next generation sequencing data, the objective being to understand within-patient diversity, in a bid to provide efficacious personalised treatment. The major difference compared to haplotype frequency estimation for antimalarial surveillance

is the focus on patient-specific diversity. Sharing information across individuals is not fruitful in tumour and quasispecies spectrum reconstruction due to individuality following from high mutation rates (as high as one per 1000 base pairs per replication cycle for viruses [18], compared with one per 10^9 base pairs per replication cycle for malaria [216]). In contrast, most methods for antimalarial resistance surveillance assume the mutation rate is low compared with the time of infection [223]. Nevertheless, following [185], it would be interesting to incorporate mutation rates into prior distributions over haplotypes of interest in antimalarial resistance surveillance, especially in view of recent evidence highlighting the importance of *de novo* mutations in the spread of artemisinin resistance [242, 154].

Metagenomics

The term 'metagenomics' is broad, encompassing many methods designed to reconstruct biodiversity within genetically diverse mixtures using next generation sequencing data [134]. Samples of genetically diverse mixtures are often referred to as 'pools' in the metagenomic literature (see [185], for example). Pools include naturally diverse samples (ecological samples [261], for example) as wells samples that have been intentionally pooled [247, 44]. The problem of reconstructing within-host *P. falciparum* haplotype frequencies thus falls within the scope of metagenomics. Many methods labelled as metagenomic, however, aim to resolve species in environmental pools, rather than clones within species [117]. Methods that are capable of reconstructing within-host *P. falciparum* haplotype frequencies are detailed above (section 2.2.2).

Chapter 3

Frequency estimation using prevalence data

3.1 Background

Considerable progress has been made in the fight against malaria in the past 15 years, manifest in a reduction in transmission in many regions [290, 27]. Nevertheless, there is a continued need for genetic surveillance of antimalarial resistance [286], and the changing landscape of transmission renders comparable measures of antimalarial resistance more important than ever. To assess spatiotemporal trends, we require estimates of the frequencies of the key determinants of parasite resistance [95, 96, 276], including alleles at individual SNPs and haplotypes and genotypes spanning multiple SNPs (for example, the quintuple mutant genotype associated with clinical resistance to SP [123]).

In areas of high endemicity people are often infected with multiple parasite clones [225, 258]. Multiclonal infections pose an analytical challenge, since standard analyses of blood sample surveys cannot resolve the clones within multiclonal infections, nor reconstruct the allelic sequences that comprise the haplotypes and genotypes of the constituent clones [276]. Instead, most blood sample surveys generate prevalence data, summaries at the level of the

blood samples [96].

Several statistical models have been designed to overcome the challenge of multiclonal infections and estimate allele, haplotype and genotype frequencies using prevalence data [39, 102, 224, 129, 95, 276, 125, 223]. They are described in detail in chapter 1. With the aim of harnessing the full potential of genetic markers for the surveillance of antimalarial resistance, we present a model that complements and builds upon the existing methods. Several differences set our model apart from existing methods of malaria haplotype and genotype frequency estimation using prevalence data [102, 224, 129, 95, 276, 125]. First, in contrast to all previously published methods, the model makes use of all available data, including those that are incomplete due to unsuccessful genotyping outcomes or study design (see, for example, [83]). Second, in contrast to the Bayesian method by Wigger *et al.* [276] and the model underpinning the freely available online software MalHaploFreq [95], our model is not reliant upon experimentally-derived estimates of sample-wise MOIs. Third, in contrast to most existing approaches [39, 102, 224, 95, 125, 223], it enables rapid analysis of data from three or more SNPs.

Akin to Wigger *et al.* [276], we construct our model within a Bayesian framework. Construction as such provides a straightforward yet comprehensive treatment of uncertainty [88]. We model each infection as an unobserved clonal conglomerate, while taking into account the sample-wise uncertainty in the population-level frequency estimates. Inference within a Bayesian framework also allows the incorporation of specialist knowledge [88], enabling the MOIs to be modelled as random variables whose prior distributions are centred about a reported average.

Also similar to Wigger *et al.* [276], we use a Markov chain Monte Carlo (MCMC) algorithm to sample from the posterior distribution of the haplotype frequencies conditional on the prevalence data. MCMC samplers are of great consequence because they allow sampling from distributions that do not belong to standard families of distributions [90]. Recursive sampling

within the MCMC scheme enables the sampler to efficiently average over the unobserved clonal conglomerates, allowing the analysis of prevalence data for more than three SNPs. Unlike methods designed to generate point estimates, the MCMC method generates a sample set that approximates the full posterior distribution. Posterior summaries, including mean estimates and credible intervals, are readily available given the MCMC sample.

The aim of this chapter is to provide full details of the model and its implementation. For exposition, we focus on results based on simulated data (which also feature in Additional file 2 of [248]). A full demonstration of the model's utility using data from the field can be found in [248] and in chapter 4. In the next section we introduce our notation. In section 3.2.2 we provide a hypothetical example to illustrate the challenge associated with multiclonal infections. We provide full details of the model and its implementation in sections 3.2.3 and 3.2.4, respectively, followed by details of the simulated data (section 3.2.5), convergence (section 3.2.6) and sensitivity analyses (section 3.2.7). Results can be found in section 3.3. The chapter ends with a discussion (section 3.4).

3.2 Methods

3.2.1 Notation

Suppose that I *P. falciparum* positive blood samples, each derived from an independent episode of malaria, are genotyped at J SNPs associated with antimalarial resistance. Due to the multiclonal nature of malaria, when the ith blood sample is genotyped at the jth SNP, the observed datum, y_{ij}, is a summary of all the alleles at the jth SNP belonging to all ≥ 1 clones within the sample. Let $y_{ij} = w$ denote the detection of wild type alleles only, $y_{ij} = m$ denote the detection of mutant type alleles only, $y_{ij} = h$ denote the detection of both wild and mutant type alleles (a *h*eteroallelic SNP) and $y_{ij} =?$ represent a missing genotyping outcome (due to assay failure, for example). For example, $\boldsymbol{y}_i = (h, ?, w)$ denotes the detection of both wild and mutant

I is the number of blood samples in the dataset (each from a distinct episode of malaria).
J is the number of SNPs genotyped.
$R \leq 2^J$ is the number of haplotypes compatible with the data.
$\boldsymbol{\pi} = (\pi_1, \cdots, \pi_R)$ is a vector of haplotype frequencies. $\boldsymbol{\pi} \in \mathbb{S}^R$, where $\mathbb{S}^R$ denotes the R dimensional simplex, and hence $\sum_{r=1}^{R} \pi_r = 1$.
$\boldsymbol{y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_I)^T$ is the collection of data for the $i = 1, \ldots, I$ blood samples, where $\boldsymbol{y}_i = (y_{i1}, \cdots, y_{iJ})$ is the vector of sample-wise genotyping outcomes for the ith blood sample and y_{ij} is the genotyping outcome of the ith blood sample at the jth SNP, where $y_{ij} \in \{w, m, h, ?\}\ \forall\ i = 1, \ldots I$ and $j = 1, \ldots J$.
m_{max} is the global maximum MOI set by the user.
$m_{i\text{min}}$ is the minimum MOI possible for the ith blood sample.
$\boldsymbol{a} = (\boldsymbol{a}_1, \cdots, \boldsymbol{a}_I)^T$ is the collection of unobserved haplotype counts for the $i = 1, \ldots, I$ blood samples, where $\boldsymbol{a}_i = (a_{i1}, \ldots, a_{iR})$ is the vector of haplotype counts for the ith blood sample, and a_{ir} denotes the number of clones in the ith blood sample characterised by the rth haplotype, where $a_{ir} \in \{0, \ldots, m_{\text{max}}\}\ \forall\ i = 1, \ldots, I,\ r = 1, \ldots R$.
$\boldsymbol{m} = (m_1, \ldots, m_I)$ is the collection of unobserved MOIs for $i = 1, \ldots, I$ blood samples, where the sample-wise MOI, $m_i \in \{1, \cdots, m_{max}\}\ \forall\ i = 1, \cdots, I$, is the total number of clones in the ith blood sample, $m_i = \sum_{r=1}^{R} a_{ir}$.
$\boldsymbol{H}$ is a $R \times J$ matrix summarising the allelic sequences of the R haplotypes over the J SNPs.
p_{ij} is the proportion of mutant type alleles at the jth SNP in the ith blood sample.
$\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_R)$ is the hyperparameter of the Dirichlet prior on the haplotype frequencies, $\boldsymbol{\pi}$.
λ is the hyperparameter for the prior on the MOI, m_i.
ϕ is an additional hyperparameter for the prior on the MOI, m_i.

Table 3.1: Model notation.

type alleles at the first SNP in the ith blood sample, a missing genotype outcome at the second SNP and two or more wild type alleles at the third (since the first SNP is heteroallelic, the ith blood sample must comprise two or more clones). If $J > 1$ (as in the above example), the vector $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{iJ})$ is a summary of the all the haplotypes or genotypes of the clones within the clonal conglomerate. Recall that the term haplotype applies if the SNPs are in the same gene, whereas the term genotype applies if the SNPs belong in multiple genes. Henceforth, haplotypes are referred to exclusively, noting that the same methods apply for genotypes. Let a_{ir} denote the unobserved haplotype count for the rth haplotype (the number of clones characterised by the rth haplotype) in the ith blood sample and $\boldsymbol{a}_i = (a_{i1}, \ldots, a_{iR})$ denote the vector of R haplotype counts for the ith blood sample, where R is the total number of haplotypes compatible across the dataset. Note that $\sum_{r=1}^{R} a_{ir} = m_i$ is the total number of clones in the ith blood sample, henceforth referred to as the MOI. Finally, let $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_R)$ denote the vector of R haplotype frequencies (the proportions of parasite clones in the *P. falciparum* population characterised by haplotypes 1 to R). For reference, a full list of model notation can be found in table 3.1. To illustrate the structure of the data, let us consider the hypothetical example outlined below.

3.2.2 Running example

Suppose parasite DNA extracted from $I = 5$ blood samples is genotyped at $J = 3$ SNPs generating prevalence data (table 3.2) which can be represented by the 5×3 matrix $\boldsymbol{y}$,

Blood sample	SNP 1	SNP 2	SNP 3
1	w	w	w
2	w	m	m
3	h	h	h
4	w	h	?
5	w	h	h

Table 3.2: A hypothetical prevalence dataset based on five samples genotyped at three SNPs. For a given blood sample and SNP, w denotes the detection of wild type alleles only, m denotes the detection of mutant type alleles only, h denotes the detection of both wild and mutant type alleles and ? indicates the genotyping outcome is missing.

$$\boldsymbol{y} = \begin{array}{c} \\ i=1 \\ \vdots \\ \\ \\ i=5 \end{array} \begin{array}{c} \begin{array}{ccc} j=1 & \ldots & j=3 \end{array} \\ \begin{pmatrix} w & w & w \\ w & m & m \\ h & h & h \\ w & h & ? \\ w & h & h \end{pmatrix} \end{array}. \tag{3.1}$$

Since both wild and mutant type alleles are detected at all three SNPs (see equation (3.1)), $R = 2^3 = 8$ haplotypes are compatible with the observed data. The allele sequences of the R

haplotypes are stored in the rows of the $R \times J$ matrix $\boldsymbol{H}$,

$$\boldsymbol{H} = \begin{array}{c} \\ r=1 \\ \vdots \\ \\ \\ \\ \\ \\ r=8 \end{array} \begin{array}{c} \begin{array}{ccc} j=1 & \ldots & j=3 \end{array} \\ \begin{pmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix} \end{array}, \tag{3.2}$$

where, arbitrarily, '0' denotes a wild type allele and '1' denotes a mutant type allele. Note that under the model, $R \leq 2^J$, because only haplotypes that are possible given the data are considered. Haplotypes that are impossible given the data are assigned zero frequency and do not feature in the $\boldsymbol{H}$ matrix. For example, if the hypothetical data are instead $\boldsymbol{y}_1 = (w, w, h)$ and $\boldsymbol{y}_2 = (m, w, m)$, the matrix of haplotypes would be,

$$\boldsymbol{H} = \begin{array}{c} \\ r=1 \\ \vdots \\ \\ r=4 \end{array} \begin{array}{c} \begin{array}{ccc} j=1 & \ldots & j=3 \end{array} \\ \begin{pmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 1 \end{pmatrix} \end{array}, \tag{3.3}$$

since the second SNP ($j = 2$) is homoallelic ($y_{12} = y_{22} = w$). Returning to the data in table 3.2 and equation (3.1), suppose the unobserved underlying haplotype count vectors for blood

samples $i = 1, \ldots, 5$ are

$$\boldsymbol{a} = \begin{array}{c} \\ i=1 \\ \vdots \\ \\ i=5 \end{array} \begin{array}{c} \begin{array}{cccccccc} 000_{r=1} & 100 & 010 & 001 & 110 & 101 & 011 & 111_{r=8} \end{array} \\ \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 4 & 0 \\ 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 3 & 1 & 0 & 0 & 1 & 0 \\ 2 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \end{pmatrix} \end{array}, \tag{3.4}$$

where a_{ir} denotes the number of clones with the rth haplotype in the ith sample. For reference, the allele sequences of haplotypes $r = 1, \ldots, R$ are shown as column headings. For example, $\boldsymbol{a}_5 = (2,0,0,0,0,0,1,0)$ indicates that the 5th blood sample contains three clones, two with haplotype '000' and one with haplotype '011', such that when the first SNP is genotyped pure wild type alleles are detected and when the second and third SNPs are genotyped both wild and mutant type alleles are detected, giving rise to $\boldsymbol{y}_5 = (w,h,h)$, as seen in equation (3.1). Given $\boldsymbol{a}$, the empirical sample haplotype frequencies are directly calculable,

$$\begin{aligned} \boldsymbol{\pi} &= (\pi_1, \ldots, \pi_R) \equiv (\pi_{000}, \pi_{100}, \pi_{010}, \pi_{001}, \pi_{110}, \pi_{101}, \pi_{011}, \pi_{111}), \\ &= \frac{\sum_{i=1}^{I} \boldsymbol{a}_i}{\sum_{i=1}^{I} \sum_{r=1}^{R} a_{ir}}, \\ &= \frac{\sum_{i=1}^{I} \boldsymbol{a}_i}{\sum_{i=1}^{I} m_i}, \\ &= \left(\frac{3}{16}, \frac{1}{16}, \frac{4}{16}, \frac{2}{16}, \frac{0}{16}, \frac{0}{16}, \frac{6}{16}, \frac{0}{16} \right). \end{aligned}$$

In this hypothetical example, the vector of haplotype frequencies is trivial to estimate because the sample haplotype counts are known. In reality, the haplotype counts are not observed. To estimate the vector of haplotype frequencies, the following model is proposed.

3.2.3 The model

We propose the following model (figure 3.1) to estimate the vector of haplotype frequencies, $\boldsymbol{\pi}$, conditional on prevalence data. A number of simplifying assumptions are made in its construction:

1. blood samples are independently distributed;
2. clones are independently distributed (for example, the probability of being infected with two clones with allelic sequences '000' and '011' is $\pi_{000} \times \pi_{011}$);
3. perfect detection (for example, if a person is infected with ten clones, nine of which are characterised by '000' and one by '100', the mutant allele is detected);
4. alleles are error-free (for example '0' is correctly identified as '0' and not as '1').

The terms in which these assumptions are introduced are indicated below; their implications are discussed in section 3.4. Latent variables include the MOIs, m_i for $i = 1,\ldots,I$, and the haplotype count vectors, $\boldsymbol{a}_i$ for $i = 1,\ldots I$. The prevalence data for the ith blood sample, y_{ij} for $j = 1,\ldots,J$, are modelled directly upon the unobserved haplotype count vector, $\boldsymbol{a}_i$. Since the model is constructed within a Bayesian framework, we put prior distributions on $\boldsymbol{\pi}$, $\boldsymbol{a}_i$ and m_i for $i = 1,\ldots I$. The priors are specified according to the dependencies in figure 3.1. The joint posterior density is,

$$\begin{aligned}\rho\left(\boldsymbol{\pi},\boldsymbol{a},\boldsymbol{m} \mid \boldsymbol{y}\right) &= \frac{\rho\left(\boldsymbol{y}|\boldsymbol{a}\right)\rho\left(\boldsymbol{a}|\boldsymbol{m},\boldsymbol{\pi}\right)\rho\left(\boldsymbol{m}\right)\rho\left(\boldsymbol{\pi}\right)}{\rho\left(\boldsymbol{y}\right)} \\ &\propto \prod_{i=1}^{I}\left\{\prod_{j=1}^{J}\left\{\rho\left(y_{ij}|\boldsymbol{a}_i\right)\right\}\rho\left(\boldsymbol{a}_i|m_i,\boldsymbol{\pi}\right)\rho\left(m_i\right)\right\}\rho\left(\boldsymbol{\pi}\right), \qquad (3.5)\end{aligned}$$

where the product over $i = 1,\ldots,I$ results from the assumptions of independence between blood samples, and the product over $j = 1,\ldots,J$ results from an assumption of conditional independence between SNPs within the ith blood sample given the haplotype counts, $\boldsymbol{a}_i$. The

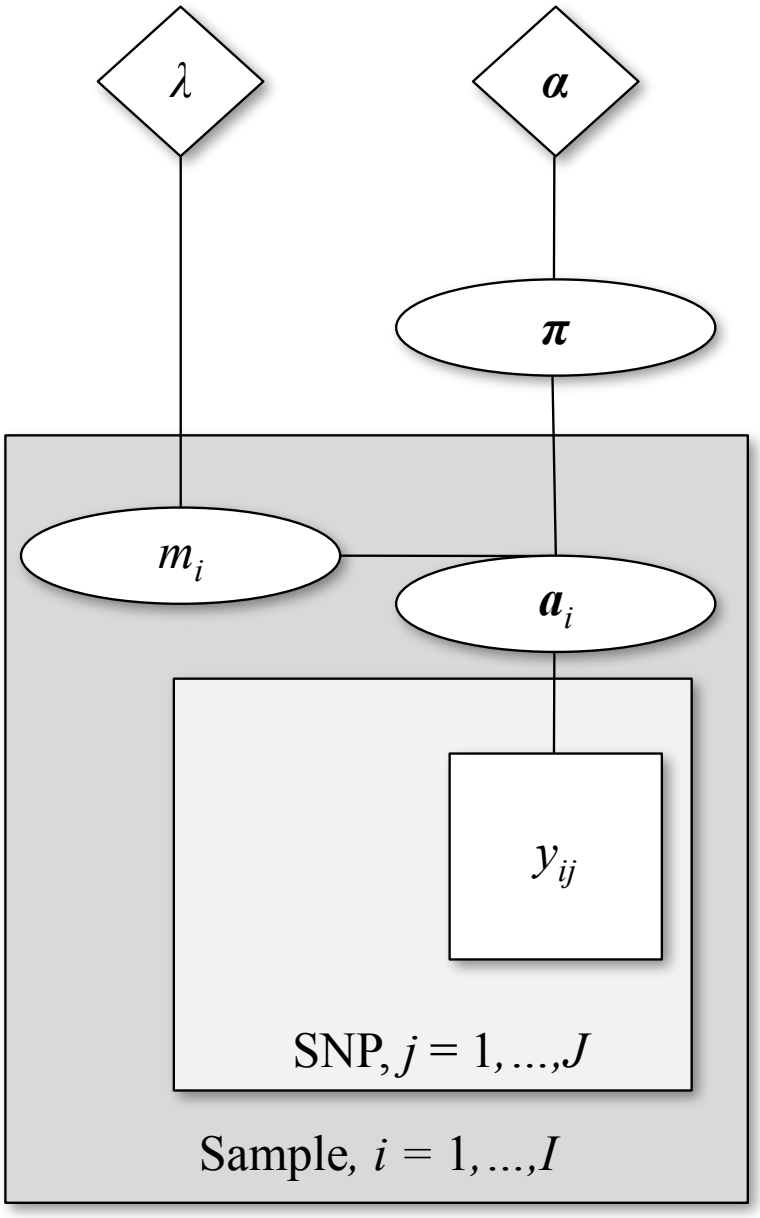


Figure 3.1: Haplotype frequency estimation model for prevalence data. The graph shows the model quantities and their conditional dependencies. The data y_{ij} for $I = 1, \ldots, I$ and $j = 1 \ldots, J$ are represented by a square. Ellipses denote unobserved variables: the vectors of haplotype counts, $\boldsymbol{a}_i$ for $I = 1, \ldots, I$; the MOI, m_i for $I = 1, \ldots, I$; and the haplotype frequency vector, $\boldsymbol{\pi}$. The diamonds represent the hyperparameters of the priors on m_i and $\boldsymbol{\pi}$ (λ and $\boldsymbol{\alpha}$, respectively). The density of the joint distribution is $\rho(\boldsymbol{\pi}, \boldsymbol{a}, \boldsymbol{m}, \boldsymbol{y}) = \rho(\boldsymbol{y} \mid \boldsymbol{a})\, \rho(\boldsymbol{a} \mid \boldsymbol{m}, \boldsymbol{\pi})\, \rho(\boldsymbol{m})\, \rho(\boldsymbol{\pi})$.

likelihood, $\rho\left(y_{ij}|\boldsymbol{a}_i\right)$, is specified as follows,

$$\rho\left(y_{ij}|\boldsymbol{a}_i\right) = \begin{cases} 1 \text{ if } y_{ij} = w \text{ and } p_{ij} = 0, \\ 1 \text{ if } y_{ij} = m \text{ and } p_{ij} = 1, \\ 1 \text{ if } y_{ij} = h \text{ and } 0 < p_{ij} < 1, \\ 1 \text{ if } y_{ij} \text{ is missing}, \\ 0 \text{ otherwise}, \end{cases} \tag{3.6}$$

where p_{ij} is the proportion of mutant type alleles at the jth SNP in the ith blood sample. That is, the number of haplotypes with a mutant type allele at the jth SNP, $\boldsymbol{a}_i \cdot \boldsymbol{h}_j$, where $\boldsymbol{h}_j$ is the jth column vector of the matrix $\boldsymbol{H}$, enlisting all the allele states (mutant '1' or wild type '0') of the R possible haplotypes at the jth SNP, normalised by the total number of haplotypes in the ith blood sample ($m_i = \sum_{r=1}^{R} a_{ir}$),

$$p_{ij} = \frac{\boldsymbol{a}_i \cdot \boldsymbol{h}_j}{\sum_{r=1}^{R} a_{ir}}. \tag{3.7}$$

Note that in equation (3.7) above, 100% detectability of the minority allele is assumed. For example, if the ith blood sample is infected with ten clones, nine with haplotype '000' and one with haplotype '100', $p_{i1} > 0$. That is, the clone with haplotype '100' is detected despite constituting only 10% of the total haplotype count for the ith blood sample. The correct identification of each allele is also assumed. For example, the alleles of both haplotypes at the second and third SNPs are identified correctly as '0' not '1', hence $p_{i\,2} = p_{i\,3} = 0$. As mentioned above, we discuss the implications of these assumptions in section 3.4.

The priors on $\boldsymbol{a}_i$ and $\boldsymbol{\pi}$ are specified as follows,

$$\rho\left(\boldsymbol{a}_i|m_i, \boldsymbol{\pi}\right) = \mathscr{M}\text{ultinomial}(\boldsymbol{a}_i \mid m_i, \boldsymbol{\pi}), \tag{3.8}$$

which implies independence between clones, and

$$\rho(\boldsymbol{\pi}) = \mathscr{D}\text{irichlet}(\boldsymbol{\pi} \mid \boldsymbol{\alpha}), \tag{3.9}$$

where $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_R)$ is the hyperparameter of the prior on $\boldsymbol{\pi}$. We set $\alpha_r = 1$ for $r = 1, \ldots, R$ with the effect that all R haplotypes are regarded *a priori* as biologically feasible and equally probable. Doing so provides an objective basis against which the validity of the results can be compared. Alternatively, one can incorporate prior knowledge about the viability of the different haplotypes by varying the elements of $\boldsymbol{\alpha}$. Instead of specifying one definitive prior for the MOI, we propose four options (equations (3.10) to (3.13)), with a view to selecting the one that provides the best fit to the data. All four are probability distributions (which assume independence between clones) on the set of integers $\{m_{i\min}, \ldots, m_{\max}\}$, where $m_{i\min} = 2$ if the ith blood sample is discernibly multiclonal (if for some $j \in \{1, \ldots, J\}$, $y_{ij} = h$) and 1 otherwise, while $m_{\max}$ is a global maximum set by the user and based on auxiliary data where available. All four follow classic distributions,

1. $$\rho(m_i) = \mathscr{U}\text{niform}(m_i \mid m_{i\min}, m_{\max}); \tag{3.10}$$

2. $$\begin{aligned}\rho(m_i) &= \mathscr{P}\text{oisson}_{\text{truncated}}(m_i \mid \lambda, m_{i\min}, m_{\max}),\\ &= \frac{\mathscr{P}\text{oisson}(m_i \mid \lambda)}{\left(\sum_{m_i=m_{i\min}}^{m_{\max}} \mathscr{P}\text{oisson}(m_i \mid \lambda)\right)};\end{aligned} \tag{3.11}$$

3. $$\begin{aligned}\rho(m_i) &= \mathscr{G}\text{eometric}_{\text{truncated}}(m_i \mid \lambda, m_{i\min}, m_{\max}),\\ &= \frac{\mathscr{G}\text{eomtric}(m_i \mid \lambda)}{\left(\sum_{m_i=m_{i\min}}^{m_{\max}} \mathscr{G}\text{eometric}(m_i \mid \lambda)\right)};\end{aligned} \tag{3.12}$$

4. $$\begin{aligned}\rho(m_i) &= \mathscr{N}\text{egative } \mathscr{B}\text{inomial}_{\text{truncated}}(m_i \mid \lambda, \phi, m_{i\min}, m_{\max}),\\ &= \frac{\mathscr{N}\text{egative } \mathscr{B}\text{inomial}(m_i \mid \lambda, \phi)}{\left(\sum_{m_i=m_{i\min}}^{m_{\max}} \mathscr{N}\text{egative } \mathscr{B}\text{inomial}(m_i \mid \lambda, \phi)\right)};\end{aligned} \tag{3.13}$$

where λ denotes the mean of the Poisson, geometric and negative binomial distributions and can be interpreted as an approximate *a priori* average MOI; and ϕ denotes the dispersion

parameter of the negative binomial distribution whose density is parameterised as follows,

$$\mathscr{N}\text{egative }\mathscr{B}\text{inomial}(m_i \mid \lambda, \phi) = \frac{\Gamma(m_i+\phi)}{\Gamma(\phi)\, m_i!}\left(\frac{\phi}{\lambda+\phi}\right)^{\phi}\left(1-\frac{\phi}{\lambda+\phi}\right)^{m_i} \tag{3.14}$$

The parameter λ is set by the user and based on auxiliary data where available. In this chapter, the Poisson prior is used for the sensitivity analyses based on simulated data. Upon analysing real data, the prior providing the best fit is selected using posterior predictive checks (see section A.2 for example).

3.2.4 The sampler

One cannot evaluate the posterior density given by equation (3.5) directly: it does not belong to a standard family of distributions and evaluating the normalising constant, $\rho(\boldsymbol{y})$, requires integrating over all possible $\boldsymbol{a}_i$, m_i and $\boldsymbol{\pi}$. Nevertheless, it is a distribution whose density can be evaluated pointwise up to a normalising constant. We therefore use a MCMC algorithm to sample from the distribution whose density is given by equation (3.5). More specifically, we use a Gibbs sampler. The Gibbs sampler works by breaking the problem down into blocks of variables which are iteratively sampled. For a given block, the variables within it are sampled from their full conditional distribution given the data and all variables outside the block at their current values. The variables under our model are $\boldsymbol{\pi}$, $\boldsymbol{a}_i$ and $m_i \;\; \forall\, i = 1,\ldots,I$; the target density is $\rho(\boldsymbol{\pi}, \boldsymbol{a}, \boldsymbol{m} \mid \boldsymbol{y})$ (equation (3.5)); and the blocks are

1. $\boldsymbol{a}_i$ and m_i given $\boldsymbol{\pi}$ and $\boldsymbol{y}$ for each $i = 1,\ldots,I$ independently,
2. $\boldsymbol{\pi}$ given $\boldsymbol{a}, \boldsymbol{m}$ and $\boldsymbol{y}$.

In other words, on each iteration of the sampler, for each $i = 1,\ldots,I$, we update the MOI and haplotype count vector, $\boldsymbol{a}_i$ and m_i, conditional on the current estimate of the haplotype frequency vector, $\boldsymbol{\pi}$, and the data, $\boldsymbol{y}$. Second, we update the haplotype frequency vector, $\boldsymbol{\pi}$,

given the collection of all the haplotype count vectors, $\boldsymbol{a}$, the collection of all the MOIs, $\boldsymbol{m}$, and the data, $\boldsymbol{y}$. The full conditional distribution of the variables in the first block, $\boldsymbol{a}_i$ and m_i, does not belong to a standard family of distributions, hence we use a Metropolis-Hastings step to update $\boldsymbol{a}_i$ and m_i. Due to conjugacy, the full conditional distribution of $\boldsymbol{\pi}$ is a Dirichlet distribution whose parameter vector is based on the haplotype count vectors. Hence we can Gibbs sample haplotype frequency vectors exactly. The mathematical details of the updates within both blocks are outlined in detail below. Starting at iteration $t = 0$, initial estimates, $\boldsymbol{\pi}^{(t)}$, $\boldsymbol{m}^{(t)}$ and $\boldsymbol{a}^{(t)}$, are either drawn from their respective priors (see section 3.2.3) or set equal to some specified values. For $t > 0$, the sampler proceeds as follows.

Update MOI and haplotype counts

The density of the joint conditional distribution of $\boldsymbol{a}_i$ and m_i is given by

$$\rho\left(\boldsymbol{a}_i, m_i \mid \boldsymbol{\pi}^{(t-1)}, \boldsymbol{y}\right) \propto \prod_{j=1}^{J} \left\{\rho\left(y_{ij} \mid \boldsymbol{a}_i\right)\right\} \rho\left(\boldsymbol{a}_i \mid m_i, \boldsymbol{\pi}^{(t-1)}\right) \rho\left(m_i\right) \;\forall\, i = 1, \ldots, I, \tag{3.15}$$

where $\rho\left(y_{ij} \mid \boldsymbol{a}_i\right)$ and $\rho\left(\boldsymbol{a}_i \mid m_i, \boldsymbol{\pi}^{(t-1)}\right)$ are given by equations (3.6) and (3.8), respectively, and $\rho\left(m_i\right)$ depends on the choice of MOI prior (equations (3.10) to (3.13)). Regardless of the MOI prior choice, the joint conditional distribution (equation (3.15)) does not belong to a standard family of distributions, hence cannot be sampled directly. Instead we use a Metropolis-Hastings update to sample from the joint conditional distribution whose density is given by equation (3.15). The Metropolis-Hastings step relies on the availability of a proposal distribution whose density can be evaluated. To ensure the MCMC algorithm only explores space compatible with the observed data, we use a proposal, q, conditioned upon the current vector of haplotype counts and the current MOI,

$$\left(\boldsymbol{a}_i^{\star}, m_i^{\star}\right) \sim q\left(\cdot \mid \boldsymbol{a}_i^{(t-1)}, m_i^{(t-1)}\right). \tag{3.16}$$

The joint proposal (equation (3.16)) is broken down into two stages, $m_i^\star \sim q_m\left(\cdot \mid \boldsymbol{a}_i^{(t-1)}, m_i^{(t-1)}\right)$ and $\boldsymbol{a}_i^\star \sim q_a\left(\cdot \mid m_i^\star, \boldsymbol{a}_i^{(t-1)}, m_i^{(t-1)}\right)$, described in detail below.

Propose a new MOI: The proposal q_m is implemented as follows. For, $i = 1, \ldots, I$, $m_i^\star$ is generated by either adding or subtracting a clone to the existing MOI,

$$m_i^\star \quad = \quad \begin{cases} m_i^{(t-1)} \pm 1 \text{ with probability } = 1/2 & \text{if } m_{\text{masked}\,i}^{(t-1)} > 0 \text{ and } m_i^{(t-1)} < m_{\max}, \\ m_i^{(t-1)} + 1 \text{ with probability } = 1 & \text{if } m_{\text{masked}\,i}^{(t-1)} = 0, \\ m_i^{(t-1)} - 1 \text{ with probability } = 1 & \text{if } m_i{}^{(t-1)} = m_{\max}, \end{cases} \tag{3.17}$$

where

$$m_{\text{masked}\,i}^{(t-1)} = \sum_{r=1}^{R} a_{\text{masked}\,ir}^{(t-1)} \text{ and} \tag{3.18}$$

$$\boldsymbol{a}_{\text{masked}\,i}^{(t-1)} = f(\boldsymbol{a}_i^{(t-1)} \mid \boldsymbol{y}_i). \tag{3.19}$$

The function, $f : \boldsymbol{a}_i^{(t-1)} \to \boldsymbol{a}_{\text{masked}}{}_i^{(t-1)}$ conditional on $\boldsymbol{y}_i$, ensures that the proposed MOI, $m_i^\star$, is compatible with $\boldsymbol{y}_i$. Essentially, $\boldsymbol{a}_{\text{masked}}{}_i^{(t-1)}$ is a template of $\boldsymbol{a}_i^{(t-1)}$, but with all counts whose removal would render $\boldsymbol{a}_i^{(t-1)}$ incompatible with $\boldsymbol{y}_i$ set equal to zero, thus 'masked', preventing their removal. The counts whose removal would render $\boldsymbol{a}_i^{(t-1)}$ incompatible with $\boldsymbol{y}_i$ include those that contribute either a solitary mutant allele or a solitary wild type allele. The function f is determined algorithmically as follows.

> First assign $\boldsymbol{a}_{\text{masked}}{}_i^{(t-1)} \leftarrow \boldsymbol{a}_i^{(t-1)}$. Second, if $\boldsymbol{a}_i^{(t-1)} \cdot \mathbf{h}_j = 1$, locate the solitary mutant count (r for which $a_{ir}^{(t-1)} \times h_{rj} = 1$) and, if $y_{ij} \neq ?$, set $a_{\text{masked}}{}_{ir}^{(t-1)} \leftarrow 0$ (see footnote[1]). Third, if $\boldsymbol{a}_i^{(t-1)} \cdot \mathbf{h}_j = \left(\sum_{r=1}^{R} a_{ir}^{(t-1)}\right) - 1$, locate the solitary wild type count (r for which $a_{ir}^{(t-1)} > 0$ and $a_{ir}^{(t-1)} \times h_{rj} = 0$) and, if $y_{ij} \neq ?$, set

[1] As an aside, in section 6.2.4 we encounter an application where $m_{i\min} = 2$ for all $i = 1, \ldots, I$. In this case, if $m_i^{(t-1)} = 2$, $a_{\text{masked}}{}_{ir}^{(t-1)} \leftarrow 0$ for all r corresponding to $a_{ir}^{(t-1)} > 0$.

$$a_{\text{masked}_{ir}}^{(t-1)} \leftarrow 0.$$

For example, if the hypothetical $\boldsymbol{a}$ given by equation (3.4) were our estimate at iteration $t-1$, $\boldsymbol{a}_{\text{masked}}^{(t-1)}$ would be given by

$$\boldsymbol{a}_{\text{masked}}^{(t-1)} = \begin{array}{c} \\ i=1 \\ \vdots \\ \\ \\ i=5 \end{array} \begin{array}{c} \begin{array}{cccccccc} 000_{r=1} & 100 & 010 & 001 & 110 & 101 & 011 & 111_{r=8} \end{array} \\ \begin{pmatrix} \mathbf{0} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 4 & 0 \\ 0 & \mathbf{0} & \mathbf{0} & \mathbf{0} & 0 & 0 & 0 & 0 \\ 0 & 0 & 3 & \mathbf{0} & 0 & 0 & 1 & 0 \\ 2 & 0 & 0 & 0 & 0 & 0 & \mathbf{0} & 0 \end{pmatrix} \end{array}, \tag{3.20}$$

where the elements that have been 'masked' are highlighted in bold, leading to,

$$\boldsymbol{m}_{\text{masked}}^{(t-1)} = \sum_{r=1}^{R} \boldsymbol{a}_{\text{masked}r}^{(t-1)} = \begin{array}{c} i=1 \\ \vdots \\ \\ \\ i=5 \end{array} \begin{pmatrix} 0 \\ 4 \\ 0 \\ 4 \\ 2 \end{pmatrix}. \tag{3.21}$$

For any $m_i^{\star}$, $m_i^{(t-1)}$ and $\boldsymbol{a}_i^{(t-1)}$, we can calculate $q_m\left(m_i^{\star} \mid m_i^{(t-1)}, \boldsymbol{a}_i^{(t-1)}\right)$, which is equal to $1/2$ or 1 as described by equation (3.17) and equation (3.22) below,

$$q_m\left(m_i^{\star} \mid m_i^{(t-1)}, \boldsymbol{a}_i^{(t-1)}\right) = \begin{cases} 1/2 & \text{if } m_{\text{masked}_i}^{(t-1)} > 0 \text{ and } m_i^{(t-1)} < m_{\max}, \\ 1 & \text{if } m_{\text{masked}_i}^{(t-1)} = 0, \\ 1 & \text{if } m_i{}^{(t-1)} = m_{\max}. \end{cases} \tag{3.22}$$

The probability density of the reverse step, $q_m\left(m_i^{(t-1)} \mid m_i^\star, \boldsymbol{a}_i^\star\right)$, is given by

$$q_m\left(m_i^{(t-1)} \mid m_i^\star, \boldsymbol{a}_i^\star\right) = \begin{cases} 1/2 & \text{if } m^\star_{\text{masked}\,i} > 0 \text{ and } m_i^\star < m_{\max}, \\ 1 & \text{if } m^\star_{\text{masked}\,i} = 0, \\ 1 & \text{if } m_i{}^\star = m_{\max}, \end{cases} \tag{3.23}$$

where $m^\star_{\text{masked}\,i}$ is derived from $\boldsymbol{a}_i^\star$ following equations (3.17) and (3.19). Note that the probability density of the forward step does not necessarily equal the probability of the reverse step, see for example figure 3.2, where $q_m\left(m_i^\star \mid m_i^{(t-1)}, \boldsymbol{a}_i^{(t-1)}\right) = 1$, but $q_m\left(m_i^{(t-1)} \mid m_i^\star, \boldsymbol{a}_i^\star\right) = 1/2$.

Propose a new haplotype count The proposal q_a is implemented as follows. For $i = 1, \ldots I$, the newly proposed haplotype count vector, $\boldsymbol{a}_i^\star$, is generated by either adding or subtracting a haplotype count vector representing a single clone, $\boldsymbol{a}_{\text{single clone}}$, to or from the current haplotype count vector, $\boldsymbol{a}_i^{(t-1)}$, conditional upon $m_i^\star$:

$$\boldsymbol{a}_i^\star = \begin{cases} \boldsymbol{a}_i^{(t-1)} - \boldsymbol{a}_{\text{single clone}}, \text{ where } \boldsymbol{a}_{\text{single clone}} \sim \mathscr{M}\text{ultinomial}\left(1, \boldsymbol{p}_{\text{sub}\,i}{}^{(t-1)}\right) \text{ if } m_i^\star = m_i^{(t-1)} - 1, \\ \boldsymbol{a}_i^{(t-1)} + \boldsymbol{a}_{\text{single clone}}, \text{ where } \boldsymbol{a}_{\text{single clone}} \sim \mathscr{M}\text{ultinomial}\left(1, \boldsymbol{p}_{\text{add}\,i}\right) \text{ if } m_i^\star = m_i^{(t-1)} + 1. \end{cases} \tag{3.24}$$

The probability vectors $\boldsymbol{p}_{\text{sub}\,i}{}^{(t-1)}$ and $\boldsymbol{p}_{\text{add}\,i}$ are calculated as follows,

$$\boldsymbol{p}_{\text{sub}\,i}{}^{(t-1)} = \frac{\boldsymbol{a}^{(t-1)}_{\text{masked}\,i}}{\sum_{r=1}^{R} a^{(t-1)}_{\text{masked}\,ir}} \text{ and } \boldsymbol{p}_{\text{add}\,i} = \frac{\boldsymbol{a}_{\text{compatible}_i}}{\sum_{r=1}^{R} a_{\text{compatible}_{ir}}}, \tag{3.25}$$

where $\boldsymbol{a}^{(t-1)}_{\text{masked}\,i}$ is given by equation (3.19) above. The vector $\boldsymbol{a}_{\text{compatible}_i}$ is the ith row of the look up matrix $\boldsymbol{a}_{\text{compatible}}$ in which the compatibilities of the $r = 1, \ldots, R$ haplotypes with $\boldsymbol{y}_i$ are

recorded. For example, for the hypothetical dataset (equation (3.1)),

$$\boldsymbol{a}_{\text{compatible}} = \begin{array}{r} \\ \boldsymbol{y}_1 = (w,w,w) \\ \boldsymbol{y}_2 = (w,m,m) \\ \boldsymbol{y}_3 = (h,h,h) \\ \boldsymbol{y}_4 = (w,h,?) \\ \boldsymbol{y}_5 = (w,h,h) \end{array} \begin{array}{c} \begin{array}{cccccccc} 000_{r=1} & 100 & 010 & 001 & 110 & 101 & 011 & 111_{r=8} \end{array} \\ \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 1 & 0 \end{pmatrix} \end{array} \tag{3.26}$$

where '1' denotes compatible and '0' denotes incompatible, and compatibility is defined as follows. If only wild or mutant type alleles are detected at the jth SNP of the ith sample ($y_{ij} = w$ or $y_{ij} = m$, respectively) only allele sequences with wild or mutant type alleles at the jth SNP are compatible, respectively; whereas if both wild and mutant type alleles are detected at the jth SNP of the ith sample, or if the datum is missing ($y_{ij} = h$ or $y_{ij} = ?$), allele sequences with both wild and mutant types alleles at the jth SNP are compatible. Akin to the dependence of $\boldsymbol{p}_{\text{sub}_i}^{(t-1)}$ upon $\boldsymbol{a}_{\text{masked}_i}^{(t-1)}$, the dependence of $\boldsymbol{p}_{\text{add}_i}$ upon $\boldsymbol{a}_{\text{compatible}_i}$ ensures the compatibility of $\boldsymbol{a}_i^\star$ with $\boldsymbol{y}_i$.

Note that given any $\boldsymbol{a}_i^\star$, $m_i^\star$, $\boldsymbol{a}_i^{(t-1)}$ and $m_i^{(t-1)}$, we can compute $q_a\left(\boldsymbol{a}_i^\star \mid m_i^\star, \boldsymbol{a}_i^{(t-1)}, m_i^{(t-1)}\right)$ following the multinomial distribution described above (equation (3.24)). That is to say,

$$q_a(\boldsymbol{a}_i^\star \mid m_i^\star, \boldsymbol{a}_i^{(t-1)}, m_i^{(t-1)}) = \begin{cases} p_{\text{add}_{ir}} \text{ if } m_i^\star = m_i^{(t-1)} + 1, \\ p_{\text{sub}_{ir}}^{(t-1)} \text{ if } m_i^\star = m_i^{(t-1)} - 1, \end{cases} \tag{3.27}$$

where the r specifies the rth element corresponding to $a_{\text{single clone}_r} = 1$ (the only element of $\boldsymbol{a}_{\text{single clone}}$ not equal to zero). The probability density of the reverse step, $q_a\left(\boldsymbol{a}_i^{(t-1)} \mid m_i^{(t-1)}, \boldsymbol{a}_i^\star, m_i^\star\right)$,

is also governed by equation (3.24), leading to

$$q_a(\boldsymbol{a}_i^{(t-1)} \mid m_i^{(t-1)}, \boldsymbol{a}_i^\star, m_i^\star) = \begin{cases} p_{\text{add}ir} \text{ if } m_i^{(t-1)} = m_i^\star + 1, \\ p^\star_{\text{sub}ir} \text{ if } m_i^{(t-1)} = m_i^\star - 1. \end{cases} \tag{3.28}$$

Acceptance probability: Having generated $m_i^\star$ and $\boldsymbol{a}_i^\star$, the newly proposed parameters are either rejected, in which case $\left(m_i^{(t)}, \boldsymbol{a}_i^{(t)}\right) \leftarrow \left(m_i^{(t-1)}, \boldsymbol{a}_i^{(t-1)}\right)$, or accepted with probability,

$$\mathbb{P}\left(\left(\boldsymbol{a}_i^{(t)}, m_i^{(t)}\right) \leftarrow (\boldsymbol{a}_i^\star, m_i^\star)\right) = \min\left\{1, \frac{\rho\left(\boldsymbol{a}_i^\star, m_i^\star \mid \boldsymbol{\pi}, \boldsymbol{y}_i\right)}{\rho\left(\boldsymbol{a}_i^{(t-1)}, m_i^{(t-1)} \mid \boldsymbol{\pi}, \boldsymbol{y}_i\right)} \frac{q\left(\boldsymbol{a}_i^{(t-1)}, m_i^{(t-1)} \mid \boldsymbol{a}_i^\star, m_i^\star\right)}{q\left(\boldsymbol{a}_i^\star, m_i^\star \mid \boldsymbol{a}_i^{(t-1)}, m_i^{(t-1)}\right)}\right\}, \tag{3.29}$$

where

$$\frac{\rho\left(\boldsymbol{a}_i^\star, m_i^\star \mid \boldsymbol{\pi}, \boldsymbol{y}_i\right)}{\rho\left(\boldsymbol{a}_i^{(t-1)}, m_i^{(t-1)} \mid \boldsymbol{\pi}, \boldsymbol{y}_i\right)} = \frac{\rho\left(\boldsymbol{a}_i^\star \mid m_i^\star, \boldsymbol{\pi}\right)}{\rho\left(\boldsymbol{a}_i^{(t-1)} \mid m_i^{(t-1)}, \boldsymbol{\pi}\right)} \frac{\rho\left(m_i^\star\right)}{\rho\left(m_i^{(t-1)}\right)}, \tag{3.30}$$

since $\prod_{j=1}^J \left\{\rho\left(y_{ij} \mid \boldsymbol{a}_i^\star\right)\right\} = \prod_{j=1}^J \left\{\rho\left(y_{ij} \mid \boldsymbol{a}_i^{(t-1)}\right)\right\} = 1$ by construction, and

$$\frac{q\left(\boldsymbol{a}_i^{(t-1)}, m_i^{(t-1)} \mid \boldsymbol{a}_i^\star, m_i^\star\right)}{q\left(\boldsymbol{a}_i^\star, m_i^\star \mid \boldsymbol{a}_i^{(t-1)}, m_i^{(t-1)}\right)} = \frac{q_m\left(m_i^{(t-1)} \mid \boldsymbol{a}_i^\star, m_i^\star\right)}{q_m\left(m_i^\star \mid \boldsymbol{a}_i^{(t-1)}, m_i^{(t-1)}\right)} \frac{q_a\left(\boldsymbol{a}_i^{(t-1)} \mid m_i^{(t-1)}, \boldsymbol{a}_i^\star, m_i^\star\right)}{q_a\left(\boldsymbol{a}_i^\star \mid m_i^\star, \boldsymbol{a}_i^{(t-1)}, m_i^{(t-1)}\right)}. \tag{3.31}$$

Each term on the right hand sides of equations (3.30) and (3.31) can be computed: $\rho\left(\boldsymbol{a}_i^\star \mid m_i^\star, \boldsymbol{\pi}\right)$ is a multinomial distribution (equation (3.8)), $\rho\left(m_i^\star\right)$ is one of the four prior distributions on the MOI (equations (3.10) to (3.13)), $q_m\left(m_i^{(t-1)} \mid \boldsymbol{a}_i^\star, m_i^\star\right)$ is equal to 1 or $1/2$ according to equation (3.23), $q_a\left(\boldsymbol{a}_i^{(t-1)} \mid m_i^{(t-1)}, \boldsymbol{a}_i^\star, m_i^\star\right)$ is equal to $p_{\text{add}ir}$ or $p^\star_{\text{sub}ir}$ according to equation (3.28), and likewise for the terms in the denominators (see, for example, figure 3.2). Note that although the notation does not make it explicit, the joint proposal (equation 3.16) is parameterised by $\boldsymbol{y}_i$, $m_{\max}$, $\boldsymbol{H}$ and $m_{i\min}$. These values do not feature in the notation, however, because they are fixed. Also note that there are no tuning parameters in the stage-wise proposals (equations

(3.17) and (3.24)), hence, under the Gibbs sampler described above, the acceptance rate of the update cannot be adjusted.

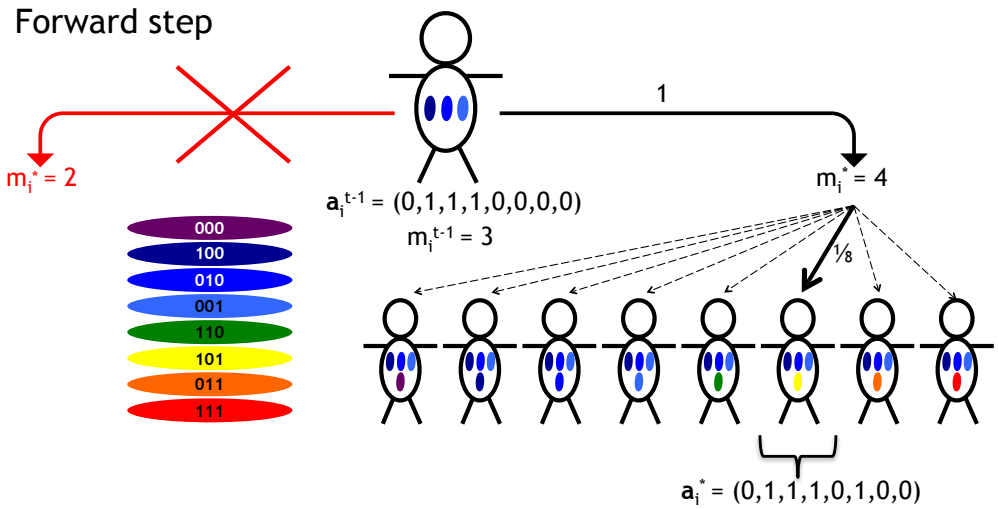
Update the vector of haplotype frequencies:

$\boldsymbol{\pi}$ is updated by sampling from its full conditional distribution with density,

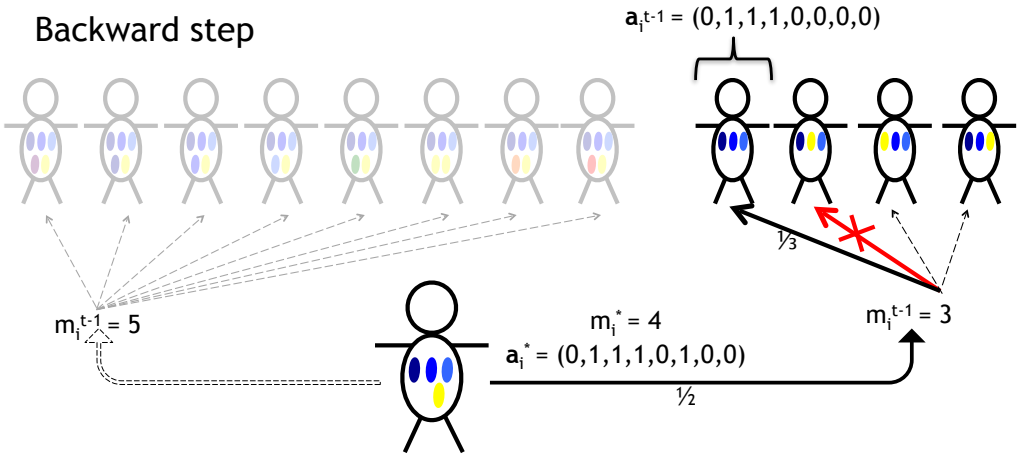
$$\begin{aligned}\rho(\boldsymbol{\pi} \mid \boldsymbol{a}^{(t)}, \boldsymbol{m}^{(t)}, \boldsymbol{y}) &\propto \prod_{i=1}^{\mathrm{I}} \left\{ \rho\left(\boldsymbol{a}_i^{(t)} \mid m_i^{(t)}, \boldsymbol{\pi}\right) \right\} \rho(\boldsymbol{\pi} \mid \boldsymbol{\alpha}), \\ &= \prod_{i=1}^{\mathrm{I}} \left\{ \mathscr{M}\text{ultinomial}\left(\boldsymbol{a}_i^{(t)} \mid m_i^{(t)}, \boldsymbol{\pi}\right) \right\} \mathscr{D}\text{irichlet}(\boldsymbol{\pi} \mid \boldsymbol{\alpha}), \\ &= \mathscr{D}\text{irichlet}\left(\boldsymbol{\pi} \mid \alpha_1 + \sum_{i=1}^{\mathrm{I}} a_{i1}^{(t)}, \ldots, \alpha_R + \sum_{i=1}^{\mathrm{I}} a_{iR}^{(t)}\right).\end{aligned}$$

Overview

The sampler is run for T iterations until convergence (see section 3.2.6). To spare computer memory, we sometimes set a thinning interval, meaning that only traces for each multiple of the thinning interval are retained. Following the general recommendation of Gelman *et al.* [88], we discard the first $t = 1, \ldots, T/2$ traces as burnin, leaving the MCMC sample, $\{\boldsymbol{\pi}^n, \boldsymbol{a}^n, \boldsymbol{m}^n\}_{n=1}^N$, which approximates the joint posterior (equation (3.5)), where N is the size of the MCMC sample post burnin and thinning. The sample $\{\boldsymbol{\pi}^n\}_{n=1}^N$, which approximates $\rho(\boldsymbol{\pi} \mid \boldsymbol{y})$, is obtained by discarding $\boldsymbol{a}^n$ and $\boldsymbol{m}^n$ from the joint sample, $\{\boldsymbol{\pi}^n, \boldsymbol{a}^n, \boldsymbol{m}^n\}_{n=1}^N$. Note that the inverse of the likelihood function (equation (3.6)) maps $\{0,1\}$ onto w, m or h conditional on $p_{ij} = (\boldsymbol{a}_i \cdot \boldsymbol{h}_j) / \sum_{r=1}^R a_{ir}$. By specifying an initial estimate of $\boldsymbol{a}_i^{(t)}$ at $t = 0$ for $i = 1, \ldots, I$, $p_{ij}^{(t)}$ is also specified for $j = 1, \ldots, J$, and thus each missing datum is assigned an initial estimate, $\hat{y}_{ij}^{(t)}$. Each time a new set of haplotypes is sampled for $t = 1, \ldots, T$, new imputed values for the missing data, $\hat{y}_{ij}^t$, are assigned. Imputation in this way assumes that the probability that a datum is missing does not depend on its value and that the missing mechanism is 'ignorable'. In other words, the parameters governing the missingness mechanism (for example, DNA quantity) are



(a) Forward step. At iteration $t-1$ of the sampler, an individual with sample data $\boldsymbol{y}_i = (h,h,h)$ is thought to be infected with $m_i^{(t-1)} = 3$ clones with haplotypes '100', '010' and '001', hence $\boldsymbol{a}_i^{(t-1)} = (0_{000}, 1_{100}, 1_{010}, 1_{001}, 0_{110}, 0_{101}, 0_{011}, 0_{111})$. The proposal involves the addition of a clone with probability one (subtraction is prohibited since $m_{\text{masked}i}^{(t-1)} = 0$, see $i=3$ equation (3.21)), hence $m_i^\star = 4$. All eight haplotypes are compatible with the observed data, hence $\boldsymbol{p}_{\text{add}\,i} = (1/8, \ldots, 1/8)$, again see $i=3$ equation (3.26)). We propose haplotype '101' (yellow ellipse), resulting in $\boldsymbol{a}_i^\star = (0_{000}, 1_{100}, 1_{010}, 1_{001}, 0_{110}, 1_{101}, 0_{011}, 0_{111},)$.



(b) Backward step. The reversal of the above proposal involves removal of the clone with haplotype '101' to recover $\boldsymbol{a}_i^{(t-1)}$. Following equations (3.18), (3.19) and (3.25), we have $\boldsymbol{a}_{\text{masked}i}^\star = (0_{000}, 1_{100}, \mathbf{0}_{010}, 1_{001}, 0_{110}, 1_{101}, 0_{011}, 0_{111})$, $m_{\text{masked}i}^\star = \sum_{r=1}^R a_{\text{masked}ir}^\star = 3$ and $\boldsymbol{p}_{\text{sub}i}^\star = \boldsymbol{a}_{\text{masked}i}^\star / \sum_{r=1}^R a_{\text{masked}ir}^\star = (0, 1/3, 0, 1/3, 0, 1/3, 0, 0)$. In other words, the probability of removing a clone is $1/2$ since addition is also possible (we could go down the lefthand branch resulting in $m_i^\star = 5$, since $m_{\text{masked}i}^\star > 0$ and $m_i^\star < m_{\max}$). We want to go down the righthand branch resulting in $m_i^{(t-1)} = 3$. Having chosen to remove a clone, the probability that we remove the clone with haplotype '101' (the yellow ellipse) is $1/3$, since we could also remove the clone with haplotype '100' (darkest blue ellipse) or the clone with haplotype '001' (lightest blue ellipse) without invalidating the compatibility of the ensuing vector of haplotype counts with the observed data.

Figure 3.2: A schematic of the proposal for the MOIs and haplotype counts. Malaria clones are represented by ellipses, colour-coded by haplotype (see stacked ellipse legend, subplot 3.2a). Branches with zero probability are depicted in red. They have zero probability because their outcomes violate compatibility with the observed data $\boldsymbol{y}_i = (h,h,h)$. Proposed branches are depicted by solid black lines. Alternative branches, that the proposal could have but did not take, are depicted by dashed lines. Each proposed branch is labeled by its probability, $\mathbb{P}(\text{proposed branch}) = 1/\text{number of available branches}$. The probabilities are equivalent to the terms in the proposal ratio, $q_m\left(m_i^{(t-1)} \mid \boldsymbol{a}_i^\star, m_i^\star\right) / q_m\left(m_i^\star \mid \boldsymbol{a}_i^{(t-1)}, m_i^{(t-1)}\right) \times q_a\left(\boldsymbol{a}_i^{(t-1)} \mid m_i^{(t-1)}, \boldsymbol{a}_i^\star, m_i^\star\right) / q_a\left(\boldsymbol{a}_i^\star \mid m_i^\star, \boldsymbol{a}_i^{(t-1)}, m_i^{(t-1)}\right) = 1/2 / 1 \times 1/3 / 1/8$.

not related to the parameters of interest (the frequencies) [132]. This assumption, does not hold if genotyping fails because of an unanticipated allele. Since sequencing is often used to identify *de novo* mutations before genotyping, the ignorable assumption is likely to hold. However, if the proposed model is used to analyse genotyping data in which failed assay attempts are likely due to unanticipated alleles, samples with missing data should be discarded.

3.2.5 Simulated data

Data are simulated to enable assessment of model performance. A haplotype frequency vector is drawn from a uniform Dirichlet distribution. A stated number of blood samples per dataset are then generated as follows. Unless otherwise stated, for each blood sample, a MOI is drawn from a non-zero conditioned Poisson distribution with $\lambda = 3$. For each blood sample, the haplotype count vector is drawn from a multinomial distribution with size equal to the MOI and probability vector equal to the vector of haplotype frequencies. The haplotype frequencies in the simulated dataset are calculated. Unless otherwise stated, for each blood sample, an observation is generated assuming 100% detectability using the inverse of the likelihood function (equation (3.6)).

3.2.6 Convergence

For every run of the sampler, log-posterior and frequency trace plots are visually inspected to monitor convergence. In addition to habitual visual inspection, a preliminary study to assess the number of iterations required for convergence is performed using 50 simulated datasets comprising one to five SNPs and 100 blood samples. For each dataset, the sampler is for 10,000, 20,000 and 50,000 iterations ($50 \times 3 = 150$ analyses in total). For each analysis, the within and between sequence variances of three parallel chains, initialised at different initial frequency vectors, are compared. Initial frequency vectors are generated by setting all but one of the initial frequencies (selected at random) to 0.02. The remaining frequency is fixed such that the

frequencies sum to unity. For datasets with only one SNP, one of the chains is initialised from a frequency vector equal to (0.5, 0.5). Comparison is based on the potential scale reduction factor (PSRF), a metric of convergence recommended by Gelman *et al.* [88]. The PSRF is an indicator of the factor by which the discrepancy in variation might be reduced if the current chains are continued for an infinite number of iterations. A value close to one supports the conjecture that the chain has converged. Gelman *et al.* advise running the chain long enough such that every PSRF < 1.1, with higher precision for final analyses. The PSRF values reported in this chapter are calculated for each haplotype frequency according to the equations on pages 303 and 304 of [88]. In total, 50,000 iterations are found to be sufficient, taking approximately five minutes to analyse a dataset comprising 100 blood samples and five SNPs.

3.2.7 Sensitivity analyses

Model performance is assessed using a series of simulated datasets, investigating the precision and accuracy of the frequency point estimates as a function of the data. For each dataset, frequency point estimates are defined by the medians of the MCMC sample. Their 95% credible intervals range from the 2.5th to the 97.5th percentiles of the MCMC sample. Accuracy is defined as the absolute error between the point estimate and the true frequency in the simulated sample, while precision is defined by the standard deviation of the marginal MCMC sample. Note that this is non-standard counterintuitive (in that lower values correspond to more accurate and precise estimates), and that both accuracy and precision decrease with the number of SNPs because the frequency mass is shared over a greater number of haplotypes. We also investigate the sensitivity of the frequency estimates to missing data, their initial values, the MOI prior and the assumption of perfect detectability. It is important to note that the tabulated results in the following section are averaged over the frequency estimates within each analysis, as well as across the analyses of ten different datasets for each combination of variables investigated. Doing so accounts for variation in the haplotype frequencies and datasets, but may also mask

haplotype specific effects. To see how average results translate into specific estimates, for each dataset we plot the frequency point estimates and their 95% credible intervals (see for example figure 3.3). Additional details of the specific sensitivity analyses are outlined below.

Precision and accuracy as a function of the data: In total, 150 simulated datasets varying in both width (one to five SNPs) and height (50, 100 and 1000 blood samples) are analysed and the average frequency and precision of the point estimates calculated as outlined above. For comparison, the datasets are also analysed using an approximate method: all blood samples with one or more heteroallelic SNPs are discarded, leaving a dataset with no discernibly multiclonal blood samples from which frequencies could be directly calculated using proportions. The frequencies of any unobserved sequences are set to zero to ensure accuracy is averaged over the same number of haplotypes as under the model.

The sensitivity of the frequency point estimates to missing data: From each of the 50 datasets used to assess convergence (section 3.2.6) data are erased from 0, 25, 50 and 75 of the blood samples selected at random. The number of genotyping outcomes erased per blood sample is selected at random, so too are the outcomes erased. Given each level of erosion, the datasets are analysed twice: first opting to impute missing data and second opting to discard blood samples with incomplete data.

The sensitivity of the frequency point estimates to their initial values: Each of the 50 datasets used to assess convergence (section 3.2.6) are reanalysed. For datasets with only one SNP, results generated post running three parallel chains with initial frequency vectors equal to (0.02, 0.98), (0.08, 0.98) and (0.5, 0.5) are compared. For datasets with two to five SNPs, results from five different chains are compared. The initial frequency vectors are selected at random from a set of frequency vectors containing a vector of uniform frequencies and all vectors generated by setting all but one of the frequencies to 0.02.

The sensitivity of the frequency point estimates to the MOI prior specification: Each of the 50 datasets used to assess convergence (section 3.2.6) are reanalysed another three times: first incorrectly assuming the distribution over the MOI is uniform; second, incorrectly assuming it is a truncated negative binomial (with $\lambda = 3$ and $\phi = 0.5$); and third, incorrectly assuming it is truncated geometric (with $\lambda = 3$). The same 50 datasets are further reanalysed twice, this time correctly assuming a Poisson prior, but with $\lambda = 1$, and then $\lambda = 5$, instead of $\lambda = 3$.

The sensitivity of frequency point estimates to the assumption that all clones are detected equally: For one to five SNPs, ten cohorts of 100 blood samples are generated as outlined above (section 3.2.5) but with parameter λ equal to one, three, five and seven. Observations are then calculated: first assuming 100% detectability; second, assuming 90% detectability (minority alleles that contributed less than 10% to a given SNP are ignored); and finally assuming 70% detectability (minority alleles that contributed less than 30% to a given SNP are ignored). All the datasets ($5 \times 10 \times 4 \times 3 = 600$ in total) are analysed assuming 100% detectability.

3.3 Results

Precision and accuracy as a function of the data: As one would hope from a valid model and functioning sampler, for a given number of SNPs, precision and accuracy increase with the number of samples in the dataset (table 3.3). Importantly, for any given dataset, the accuracies of the estimates generated under the statistical model are superior to those generated by discarding multiclonal samples (table 3.3).

The sensitivity of the frequency point estimates to missing data: Unsurprisingly, for a given number of SNPs, the impact of missing data on the mean accuracy and precision of

Number of SNPs	Number of blood samples	Statistical model		Approximate method
		Precision	Accuracy	Accuracy
1	50	0.034	0.015	0.075
	100	0.027	0.016	0.101
	1000	0.010	0.006	0.083
2	50	0.038	0.022	0.092
	100	0.034	0.022	0.058
	1000	0.011	0.007	0.046
3	50	0.039	0.035	0.074
	100	0.029	0.022	0.047
	1000	0.010	0.007	0.026
4	50	0.032	0.024	0.053
	100	0.024	0.017	0.040
	1000	0.008	0.007	0.014
5	50	0.021	0.017	0.040
	100	0.017	0.013	0.028
	1000	0.007	0.006	0.010

Table 3.3: Precision and accuracy as a function of the width (number of SNPs) and height (number of samples) of the simulated datasets. Lower values are indicative of higher accuracy and precision. Note that due to the way accuracy and precision are defined (see the introductory paragraph to section 3.2.7), neither accuracy nor precision is comparable across different numbers of SNPs.

Accuracy	Number of blood samples with incomplete data			
Number of SNPs	0	25	50	75
1	0.16	0.20 (0.20)	0.34 (0.33)	0.42 (0.43)
2	0.23	0.28 (0.28)	0.32 (0.33)	0.40 (0.42)
3	0.23	0.26 (0.26)	0.24 (0.28)	0.37 (0.46)
4	0.17	0.20 (0.20)	0.22 (0.25)	0.31 (0.30)
5	0.13	0.14 (0.15)	0.15 (0.16)	0.17 (0.19)

Table 3.4: The impact of incomplete data upon the mean accuracy of the frequency estimates. Lower values indicate higher accuracy. For those datasets with missing data, the mean accuracy obtained from analyses based on only the blood samples with complete data are included in parentheses.

Precision	Number of blood samples with incomplete data			
Number of SNPs	0	25	50	75
1	0.27	0.31 (0.31)	0.38 (0.38)	0.54 (0.54)
2	0.35	0.39 (0.40)	0.42 (0.45)	0.54 (0.64)
3	0.29	0.32 (0.33)	0.36 (0.39)	0.45 (0.53)
4	0.24	0.26 (0.27)	0.29 (0.31)	0.33 (0.38)
5	0.17	0.18 (0.18)	0.19 (0.20)	0.21 (0.23)

Table 3.5: The impact of incomplete data upon the mean precision of the frequency estimates. Lower values indicate higher precision. For datasets with missing data, the mean precision obtained from analyses based on only the blood samples with complete data are included in parentheses.

the frequency point estimates is unfavourable (compare column two with columns three, four and five, tables 3.4 and 3.5, respectively). However, in general, estimates are more accurate and precise upon imputation (compare the numbers within and outside the parenthesis). In summary, imputation enables use of all available data, whereas, for datasets with two or more SNPs, partial data are squandered when blood samples with incomplete data are discarded.

The sensitivity of the frequency point estimates to their initial values: Frequency point estimates are robust to their initial values. The mean difference between estimates obtained from chains initiated at different values is < 0.01, while the maximum is 0.02 (haplotype 00000, dataset B, figure 3.3).

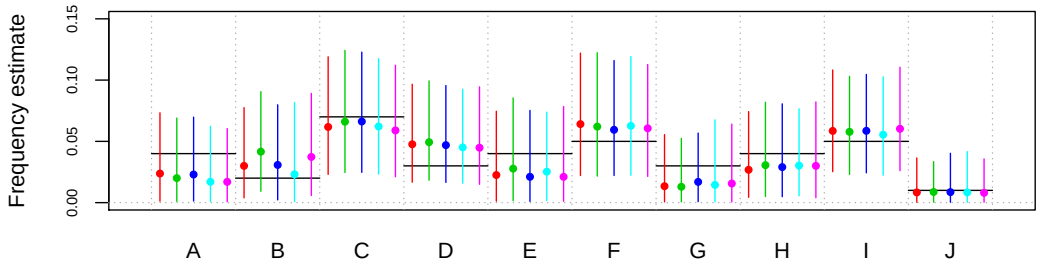


Figure 3.3: The sensitivity of a frequency point estimate to different initial values. The plot shows frequency point estimates (dots) with 95% credible intervals (vertical lines) for the haplotype with the allele sequence 00000, generated by five chains (five different colours) initiated at a different initial frequency vectors) for ten different datasets (A–J). For each dataset, the frequency of the haplotype with the allele sequence 00000 in the simulated sample is depicted by a black horizontal bar.

The sensitivity of the frequency point estimates to the MOI prior distribution: Unsurprisingly, the model with the correctly specified Poisson distribution over the MOI gives rise to the most accurate haplotype frequencies on average (table 3.6). At the level of the individual point estimates, the differences between estimates generated under the uniform, Poisson and negative binomial distributions are relatively small and the prior had little to no effect on precision (for example, see figure 3.4). Likewise, on average, the correct λ parameter specification gives rise to the most accurate frequency estimates (table 3.7). The detrimental effect of overestimating λ appears to be slightly less than that of underestimating it, but the range tested is small. In fact, it seems that no λ specification is preferable to misspecification (compare values in columns three and four of table 3.7 to values under the uniform prior in table 3.6). Overestimation has a spuriously favourable effect on precision, probably because overestimation augmented the number of clones per blood sample, thus leading to a greater number of haplotype assignments on which to base the haplotype frequencies. Sensitivity of the model to the parameter λ motivates the repeat analysis of field data, each time varying λ in order to establish the sensitivity of the results (for example, see section 4.2.3). Based on the accuracy when λ is unspecified compared with misspecified, if the *a priori* average MOI is unknown, a uniform prior is to be worth investigating.

Number of SNPs	MOI prior distribution			
	Uniform	Poisson	N. Binomial	Geometric
1	0.051	**0.016**	**0.016**	0.024
2	0.032	**0.023**	0.027	0.028
3	0.027	**0.022**	0.023	0.023
4	**0.017**	0.018	**0.017**	0.018
5	**0.013**	**0.013**	**0.013**	**0.013**

Table 3.6: The impact of MOI prior misspecification on the mean accuracy of the frequency estimates. All data are generated under a model with a MOI Poisson prior with $\lambda = 3$. Note that N. Binomial refers to a negative binomial distribution. Lower values (highlighted in bold) indicate higher accuracy.

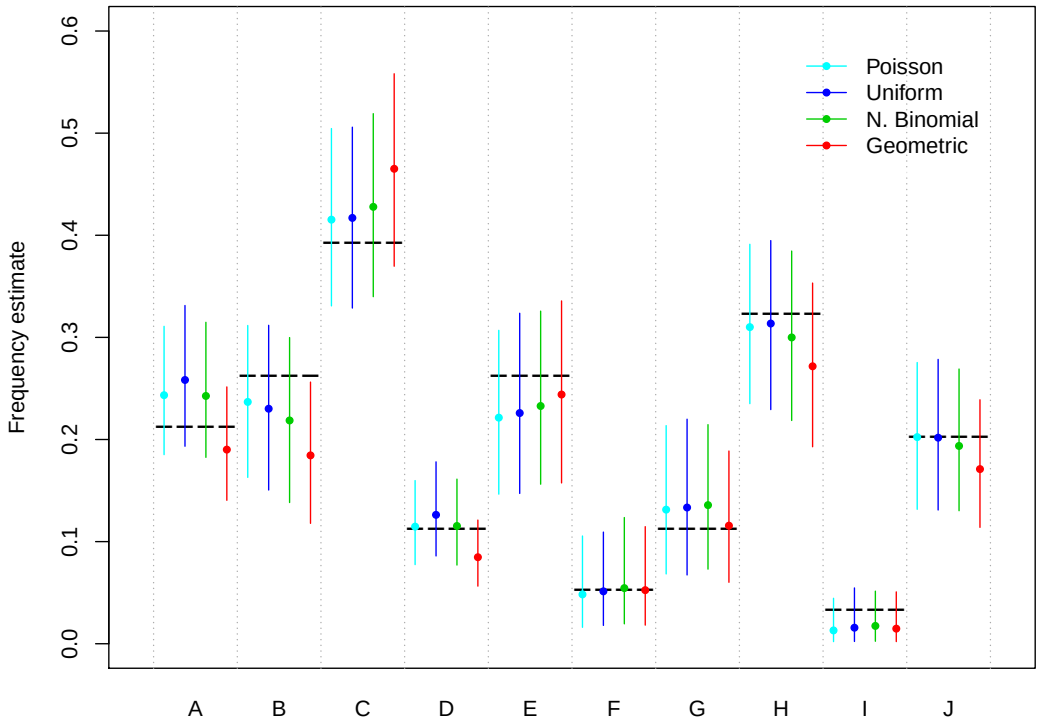


Figure 3.4: The impact of MOI prior misspecification on the frequency estimates of a single haplotype. The plot shows frequency point estimates (dots) and their 95% credible intervals (vertical lines) for the haplotype with allele sequence 11, colour-coded by the MOI prior distribution, across ten different datasets (A–J). The frequency in the simulated data set is denoted by the black horizontal bar. Note that N. Binomial refers to a negative binomial distribution. The data are simulated under the Poisson prior.

Number of SNPs	MOI prior parameter λ		
	Correct ($\lambda = 3$)	Underestimate ($\lambda = 1$)	Overestimate ($\lambda = 5$)
1	**0.016**	0.060	0.058
2	**0.023**	0.044	0.032
3	**0.023**	0.028	0.026
4	**0.017**	0.021	**0.017**
5	**0.013**	0.014	**0.013**

Table 3.7: The impact of the MOI prior parameter misspecification on the mean accuracy of the frequency estimates. Lower values indicate higher accuracy.

The sensitivity of frequency point estimates to the assumption that all clones are detected equally: Estimates are robust to suboptimal detectability when data are generated using MOI parameter, λ, of one or three (for example see figure 3.5). For data generated using $\lambda \geq 5$, estimates are robust to 90% detectability, but the accuracy decreases when the detectability drops to 70%, (figure 3.5). Unsurprisingly, the detrimental effect of suboptimal detectability has more impact upon datasets generated under a comparatively large MOI parameter, λ, since blood samples with a large number of clones are more likely to qualify as blood samples in which alleles might be in a minority. Suboptimal detectability appears to have a small spuriously positive effect on precision, seemingly due to the relative decline in the number of heteroallelic alleles. Suboptimal detectability primarily affects datasets comprised of three or fewer SNPs (table 3.8). In addition to the method used to define accuracy (see introductory paragraph to section 3.2.7), this may, in part, be due to the method used to generate the data, explained as follows. The data are generated using a frequency vector drawn from a uniform Dirichlet distribution. Since the number of possible haplotypes increases exponentially with the number of SNPs, haplotypes frequencies tend to be more uniform in datasets comprised of four or more SNPs. Since the likelihood that a single SNP is dominated by a single allele is smaller in a dataset comprising a large number of haplotypes over which mass is evenly distributed, suboptimal detectability primarily affects datasets comprised of three or fewer SNPs, especially

Number of SNPs	MOI parameter, λ	Limit of detection		
		100%	90%	70%
1	1	0.011	0.011	0.012
	3	0.015	0.015	0.051
	5	0.012	0.012	0.107
	7	0.033	0.029	0.160
2	1	0.017	0.017	0.017
	3	0.018	0.018	0.041
	5	0.028	0.028	0.081
	7	0.035	0.031	0.123
3	1	0.011	0.011	0.011
	3	0.019	0.019	0.025
	5	0.028	0.025	0.049
	7	0.039	0.04	0.065
4	1	0.010	0.010	0.010
	3	0.016	0.016	0.017
	5	0.023	0.023	0.028
	7	0.033	0.033	0.031
5	1	0.009	0.009	0.009
	3	0.013	0.014	0.014
	5	0.017	0.017	0.016
	7	0.019	0.019	0.018

Table 3.8: The impact of suboptimal detectability on the accuracy of the frequency estimates. For one to five SNPs, ten datasets are generated given detectability equal to 100%, 90% and 70%. The datasets are analysed assuming optimal detectability (100%). Lower values indicate higher accuracy.

when mass is unevenly distributed (for example, see the estimate for haplotype 010, cohort 26, $\lambda = 5$, figure 3.5).

3.4 Discussion

In this chapter, we present a statistical model designed to estimate population-level frequencies of *P. falciparum* allele and multi-SNP haplotype and genotype frequencies using prevalence data from malaria endemic regions where multiclonal infections are commonplace. Multiclonal

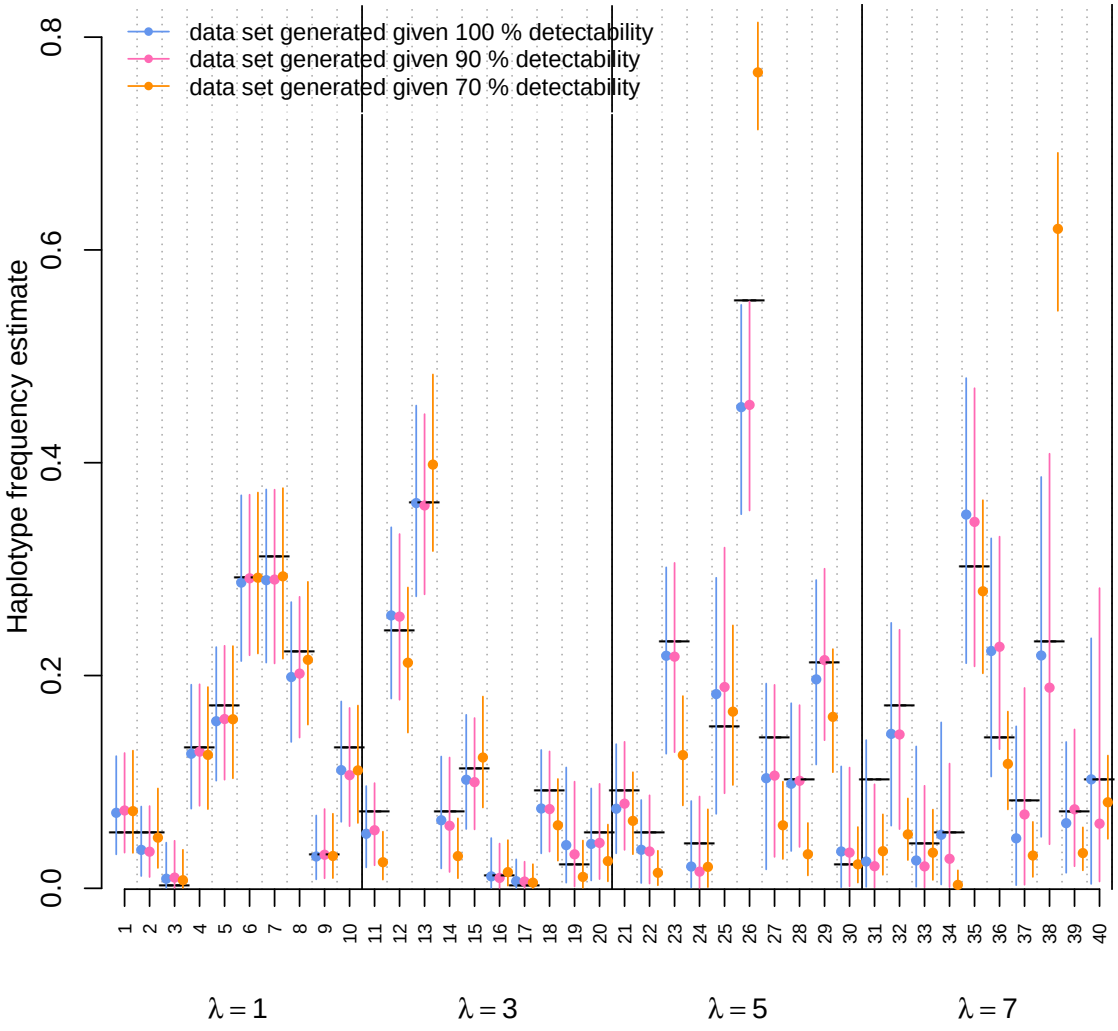


Figure 3.5: The impact of suboptimal detectability on the frequency estimates for a single haplotype with allele sequence 010. The plot shows haplotype frequency estimates (points) and their 95% credible intervals (vertical lines), given suboptimal detectability. Ten cohorts of one hundred infected blood samples are generated per specified MOI parameter, λ (40 cohorts in total: 1–10, $\lambda = 1$; 11–20, $\lambda = 3$; 21–30, $\lambda = 5$; 31–40, $\lambda = 7$). To enable comparison between results given different limits of detectability, three datasets are generated per cohort: one given 100% detectability (blue), another given 90% detectability (pink), and another given 70% detectability (orange). Each of the three datasets generated from a common cohort have the same frequencies in the simulated datasets (black horizontal bar), since the haplotypes in the infected blood samples remained the same despite changing the detectability. The datasets are all analysed assuming 100% detectability.

infections hamper the genetic surveillance of antimalarial resistance. The model is designed to overcome the problems associated with multiclonal infections. Its application generates comparable frequency estimates, allowing markers of resistance to be tracked in malaria endemic regions, yielding important information on the dynamics of resistance.

Importantly, the model does not require measurements of the sample-wise MOIs. Instead, it uses an *a priori* estimate of the average MOI (based on auxiliary data where available), an initial estimate of the vector of haplotype frequencies, and all available prevalence data to infer the haplotypes of the unobserved clones within individual blood samples. The initial frequency estimates are then revised based on clonal assignments. The implementation algorithm cycles over the aforementioned steps thousands of times until convergence. Application of the model reconstructs haplotypes within samples, provides a consistent method of frequency estimation, and avoids the loss of information that results from the usual adjustments made for multiclonal blood samples and unsuccessful genotyping outcomes.

To assess the impact of various model choices a suite of sensitivity analyses is performed using simulated data. The simulation study demonstrates that the frequencies estimated using the model are more accurate than estimates based on simply calculating proportions after discarding discernibly multiclonal blood samples. The model is robust to changes in the initial frequency estimates, but sensitive to deviations in the prior on the MOI. In light of these results, we recommend an investigation to find the MOI prior distribution that provides the best fit to the data (see section A.2 for an example), followed by repeat analyses of the data, each time varying the MOI prior parameter value within a reasonable range (such as the limits of its 95% confidence interval), to establish the sensitivity of the results to its value (see section 4.2.3 in the following chapter).

A number of simplifying assumptions are made in the construction of the model; they are listed below. As with any model, it is important to note that although the assumptions are likely to be violated in practice, the model may still be useful, as famously remarked upon by George

Box [31]. The assumptions include

1. blood samples are independently distributed;
2. clones are independently distributed (for example, the probability of being infected with two clones with allelic sequences '000' and '011' is $\pi_{000} \times \pi_{011}$);
3. perfect detection (for example, if a person is infected with ten clones, nine of which are characterised by '000' and one by '100', the mutant allele is detected);
4. alleles are error-free (for example '0' is correctly identified as '0' and not as '1').

The first two assumptions are common to all of the existing statistical methods of *P. falciparum* haplotype frequency estimation [102, 224, 129, 95, 276, 125], while more realistic assumptions regarding detectability and SNP miscalls are incorporated into alternative models ([95] and [276], respectively). We now discuss each assumption in turn.

Depending on the study design, the assumption of independence between samples is a valid one. For example, blood samples surveyed in a cross sectional study (such as [22]) should be independent. On the contrary, repeat sampling from the same child (for an example, see [51]) might lead to dependence. In chapter 4, we analyse the data from [51] under the assumption of independence. In chapter 5, we relax the assumption by adding an extension to our model.

The assumption that clones are independent depends on the manner in which multiclonal infections are acquired. An individual infected with clones obtained from multiple successive bites in a high transmission setting is likely to harbour independent clones, whereas the assumption is unlikely to hold for a person infected with multiple clones following a single inoculation from a mosquito harbouring a multiclonal infection [102]. Since both mechanisms are likely to occur, especially in high transmission settings, the assumption that clones are independently distributed is questionable. Reasons as to why the assumption might not harm inference are discussed in length by Hill and Babiker [102]. Perhaps the most compelling argument put forward by Hill and Babiker is the agreement between the experimentally-derived within-vector diversity (based on diploid oocysts from dissected mosquitoes collected in

the same village as the prevalence data), and the within-host diversity estimated under the assumption of independence. We agree with Hill and Babiker that the assumption is tenuous but pragmatic, noting (as do they) that there is not enough information in the data to support a model that distinguishes between inoculation with recombinant and independent clones [102]. In addition, we note that dependence between haplotypes will unlikely harm average estimates, since correlation typically leads to over-dispersed but unbiased realisations (see binomial example in [84]).

The validity of the assumption that all SNPs are correctly identified depends on the technology used to generate the data and differs for different SNPs. For example, concordances between calls based on microarray technology and calls based on RFLP analyses ranging from 63.9% (for *pdfhfr*-51) to 100% (for *pfmdr1*-86 and *pfdhps*-581) have been reported [145]. The model by Wigger *et al.*, includes an error probability term, which is fixed and equal for all SNPs [276] (see chapter 2 for a full description). Based on simulated data, Wigger *et al.* conclude that the error model is beneficial if the miscall rate exceeds 1–2%. Following Wigger *et al.*, it would be interesting to incorporate an error term into our model to account for miscalled SNPs. It is noted, however, that when analysing microarray data from the field (in which the probability of an error is thought to be 0.05 based on SNP-wise comparison with RFLP base calls [145]), the frequency estimates are statistically indifferent unless a large number of samples (> 500) are analysed [276]. If a large number of samples are analysed, omission of a fixed random error is likely to cause the model to overfit noise, hence underestimate dominant frequencies and overestimate rare frequencies [276]. A simple way to avoid overfitting noise without adding an error term, is to analyse the data twice, setting rare frequencies to zero in the second analysis [276].

The assumption that all clones in the blood sample are perfectly detected is almost certainly violated, especially when analysing data generated by PCR based methods [96]. Detection limits are thought to range between 80% and 99% [133, 113, 114, 58]. To assess the impact

of this assumption, simulated data are generated under imperfect detectability. Our model is robust to imperfect detection, providing the MOI prior parameter, λ, is less than or equal to three, or the limit of detectability is 90% or more.

Further to the problem of imperfect detection, is the fact that the blood sample itself might not contain a representative sample of the infection. This might occur because of low parasite numbers and/or because *P. falciparum* infections undergo complex cycles of sequestration [77]. Even when imperfect detection due to experimental procedures is taken into account (see [95], for example), what actually is estimated is the proportion of accessible parasite clones among the within-host parasite population. However, if the parameters of sequestration are independent to the frequencies of interest, which we assume they are, the inaccessible clones are ignorable and the estimates based on the accessible clones should be accurate; that is to say, collectively, the blood samples should equitably represent the host-infecting parasite population.

Following Hastings *et al.* [95, 96], we define frequency in terms of parasite clones (recall that, following convention, we use the word clone to denote a collection of genetically identical parasites). Alternatively, one could define frequency in terms of the proportion of parasites. In fact, one could see the former definition as an approximation of the latter, assuming clones represent populations of equal size. Unfortunately, there is not enough information in prevalence data to support a model that accommodates estimates in terms of biomass. Hence, all models that generate *P. falciparum* allele, haplotype and genotype frequencies based on prevalence data either define frequency in terms of parasite clones, or assume clones represent clones of equal size [39, 102, 224, 129, 95, 276, 125, 223].

As outlined in the introduction, several differences set our model apart from existing methods of *P. falciparum* allele, haplotype and genotype estimation using prevalence data. In contrast to preceding methods, the model presented here is able to analyse prevalence data for more than three SNPs, using all available data, including those that are incomplete due

to unsuccessful genotyping assays, without reliance upon experimentally-derived estimates of the sample-wise MOI, within a Bayesian framework, thus providing a readily extendable framework in which uncertainty is captured in a straightforward yet comprehensive manner. However, superior assumptions regarding detectability and experimental error are incorporated into alternative models [95, 276] (see above). It is especially important to take into account the suboptimal detectability of minority clones, addressed by Hastings *et al.* [95], when the experimentally-derived MOI estimates are regarded as fixed [96]. The latter is not the case in the current model (patient-level MOIs are treated as unobserved random variables), perhaps explaining why our model is comparatively robust to imperfect detectability.

In summary, genetic monitoring of *P. falciparum* plays an important role in the timely surveillance of antimalarial drug resistance. However, multiclonal infections present an analytic challenge, especially in areas of high transmission. We provide a full description of a model designed to overcome the challenge of multiclonal infections and estimate the frequencies of *P. falciparum* allele, multi-SNP haplotypes and genotypes. Its validity is demonstrated using a suite of sensitivity analyses, while its utility is demonstrated elsewhere using prevalence data for markers of resistance to SP [248]. Its applicability, however, extends beyond markers of SP resistance, as demonstrated in the following chapter. To the best of our knowledge, this is the first model that combines rapid analysis of three or more SNPs, using all available data without reliance upon measurements of the MOI in individual blood samples.

Chapter 4

Frequency trends in Uganda

4.1 Background

Artemether-lumefantrine (AL) is the recommended first-line treatment for uncomplicated malaria in Uganda, the country estimated to have the fourth highest burden of malaria in the world [290]. National first-line treatment policy switched to AL in 2004 [139], following poor efficacy of the preceding first-line treatment, CQ plus SP (CQ+SP) [234, 301]. The policy was not launched until 2006, however, and its implementation was thwarted by frequent AL stock outs, low public confidence and prohibitively high private sector costs [172]. Efforts to increase coverage have improved uptake in recent years [302].

AL is an ACT, comprising artemether, a fast acting artemisinin derivative, and lumefantrine, a slower acting partner drug [89]. Resistance to artemisinin has been reported in Southeast Asia, but not yet in Africa [288]. Concordantly, the efficacy of AL is high in Uganda [250, 266, 116]. That said, a recent study by Yeka *et al.* reports that, in contrast to previous findings, the rate of recurrent infections following treatment with AL is now higher than that after treatment with artesunate-amodiaquine (AS/AQ) [303], possibly reflecting decreased parasite sensitivity to lumefantrine. Effective surveillance of both artemisinin and partner-drug resistance is therefore critical to ensure the provision of efficacious treatment in Uganda, and has been heralded as a

mainstay of malaria control in the country [244].

Definitive markers of lumefantrine resistance are yet to be identified. High-level, but unstable resistance following selection *in vitro* has been associated with the differential expression of many genes, while markers of reduced sensitivity identified in field isolates include nSNPs in the genes *pfmdr1* and *pfcrt* [163]. More precisely, wild type amino acid residues PfMDR1:N86 and PfCRT:K76 have been associated with decreased sensitivity to lumefantrine, contrarily to the mutant type amino acid residues associated with reduced sensitivity to CQ, PfMDR1:86**Y** and PfCRT:76**T** [162]. (Note that throughout this thesis a bold font is used to distinguish amino acids encoded for by mutant type alleles.) Observations *in vivo* also attest to PfMDR1 and PfCRT markers of decreased sensitivity to lumefantrine. A recent study in Uganda by Conrad *et al.*, for example, reported an increase in wild type markers PfMDR1:N86, PfMDR1:184**F**, PfMDR1:D1246 and PfMDR1:K76 following treatment with AL, similar to observations reported elsewhere in Africa (see [51], and references therein). In addition, Conrad *et al.* report a yearly decrease in the prevalence and frequency of PfMRP1:I876 associated with AL, while a study in Tanzania found AL selected for PfMRP1:I876 [56],

Numerous studies have reported trends in the prevalence of PfMDR1 and PfCRT markers in Uganda [298]. Most recently, Mbogo *et al.* published a study summarising trends based on prevalence data collected between 2003–2012 in Tororo [149], a region of very high transmission intensity [302]. A gradual increase in the prevalence of several markers including PfMDR1:N86, PfMDR1:184**F** and PfMDR1:D1264 was observed, as well as a sudden increase in PfCRT:K76 in 2012, consistent with national treatment policy switching from CQ to AL. The trends are also consistent with yearly trends reported previously by Conrad *et al.* [51], and are based, in part, on the same data.

Conrad *et al.* [51] reported changes in the prevalence and frequency of single alleles using data collected from a longitudinal trial of AL versus dihydroartemisinin-piperaquine (DP) in Tororo [15, 266]. Like AL, DP is an ACT and has been shown to be highly efficacious in

Uganda [250, 266]. The DP partner drug, piperaquine, has a longer half-life than lumefantrine (estimates range from 23–80 days for piperaquine and 3–5 days for lumefantrine[89]), which protects against recurrent infection [250, 266]. Definitive markers of piperaquine resistance have yet to be identified [42]. Resistance following selection *in vitro* has been associated with copy number variations on chromosome five, as well as a previously unreported amino acid change, PfCRT:C101**F** [68]. Reduced sensitivity to piperaquine based on field isolates analysed *in vitro* has been associated with parasites characterised by PfCRT:76**T** [161], with a novel PfCRT amino acid change, PfCRT:C350**R** [196], and in regions where parasites with multiple *pfmdr1* copies abound, with clones that carry a single copy of *pfmdr1* [42]. Conrad *et al.* observed a positive correlation between immediacy since last treatment with DP and PfMDR1:86**Y**, PfMDR1:Y184, PfMDR1:1246**Y** and PfCRT:76**T**, opposite to trends following treatment with AL [51]. The inverse trends associated with immediacy since last treatment with AL and DP seem to explain why yearly trends consistent with national AL policy are less marked in the DP arm of the study.

In the study by Conrad *et al.*, single nSNPs are analysed separately since there is a high proportion of heteroallelic genotyping outcomes, precluding haplotype assignment [51]. Single SNP analyses overlook important synergistic effects between alleles on the same genome. The aim of this chapter is to build upon the study by Conrad *et al.* [51], by investigating changes in the frequencies of haplotypes, rather than single alleles, thereby enhancing our understanding of resistance. To this end, we seek to estimate frequencies using the model introduced in chapter 3, and further investigate any frequencies displaying notable trends. To the best of our knowledge, this is the first study to characterise trends in *P. falciparum* haplotype frequencies in Uganda.

The outline of this chapter is as follows. The following section includes a brief description of the previously published data (subsection 4.2.1), a description of how the data are partitioned in order to estimate frequencies (subsection 4.2.1), an outline of the study-specific details

regarding frequency estimation under the previously described model (subsection 4.2.3) and a description of the regression model (subsection 4.2.4). In the results section, we focus first on the frequency estimates (subsection 4.3.1) and then on the results of the regression (subsection 4.3.3). The results are discussed in (subsection 4.4).

4.2 Methods

4.2.1 Previously published data

In this chapter, haplotype frequencies are estimated using previously published prevalence data [51], which also feature in [149]. The data are derived from *P. falciparum* positive filter paper blood spots collected between 2007–2012 from a cohort of children enrolled in a longitudinal trial of AL versus DP in Tororo, Uganda [15, 266]. In total, 312 children, aged between 4 and 12 months, were randomised to either AL or DP when first diagnosed with uncomplicated malaria. They were followed for five years and treated according to their initial randomisation for all episodes of uncomplicated malaria. A subset of episodes was selected (as described below) for genetic analysis, leading to a total of 291 children contributing one or more samples. To illustrate this sampling framework, for each child, the selected episodes are plotted with respect to time and colour-coded by drug arm (figure 4.1). Since only the episodes selected for genetic analysis are plotted, the number of days since prior treatment (hence episode) is indicated by the saturation of the coloured points, which decreases with time (catgorised by days), as does residual drug pressure. The subset of episodes was selected as follows. Of the 312 episodes of uncomplicated malaria first diagnosed, 50 were randomly selected from each drug arm for baseline genetic analysis (crosses, figure 4.1). To generate longitudinal data, 50 episodes per quarter from January 2008 to December 2012 were randomly selected from each drug arm (in the final quarter of 2012 there are only 39 episodes in DP arm, hence all were selected), amounting to 1889 longitudinal episodes (coloured circles, figure 4.1), each preceded

by at least one prior episode 4 or more days ago. Of the 1889 episodes, 17 were treated with quinine due to either treatment failure or severe malaria (black dots, figure 4.1). To generate data, a blood spot was collected from each of 100 baseline, plus 1889 longitudinal, episodes selected for genetic analysis. *P. falciparum* DNA was extracted from each and genotyped at eight nSNPs in codons 86, 184, 1034, 1042 and 1246 in *pfmdr1*, codons 876 and 1466 in *pfmrp1*, and codon 76 in *pfcrt*. Genotyping outcomes were classified as either pure wild type if only wild type alleles are detected, pure mutant type if only wild type alleles are detected, heteroallelic if both wild and mutant type alleles are detected, or missing if the assay failed. The outcomes for all 1989 samples are depicted in figure 4.2. Successful outcomes at codons 1034 and 1042 in *pfmdr1* are exclusively pure wild type (figure 4.2) hence do not feature in this chapter hereafter. To estimate the MOI, Conrad *et al.* also genotyped ten blood spots per quarter, per drug arm (plus 11 baseline blood spots upon request) at *pfmsp1* and *pfmsp2* [51]. Per sample MOI estimates were defined as the maximum number of alleles at either gene. They range from 1 to 7 (see figure A.1 in section A.1).

4.2.2 Partitioning the data

To estimate baseline frequencies and investigate the impact of drug pressure upon *pfcrt* allele and *pfmdr1* and *pfmrp1* haplotype frequencies in different drug arms, the data (figure 4.2) are partitioned as follows. First the data are separated by gene (panel A, figure 4.3) and the data from nSNPs in codons *pfmdr1*-1034 and *pfmdr1*-1042 are removed. Data derived from baseline samples are then set apart from the data derived from longitudinal samples (panel B, figure 4.3). Data obtained from children randomised to the DP arm are separated from those derived from children randomised to AL arm (panel C, figure 4.3). To investigate yearly

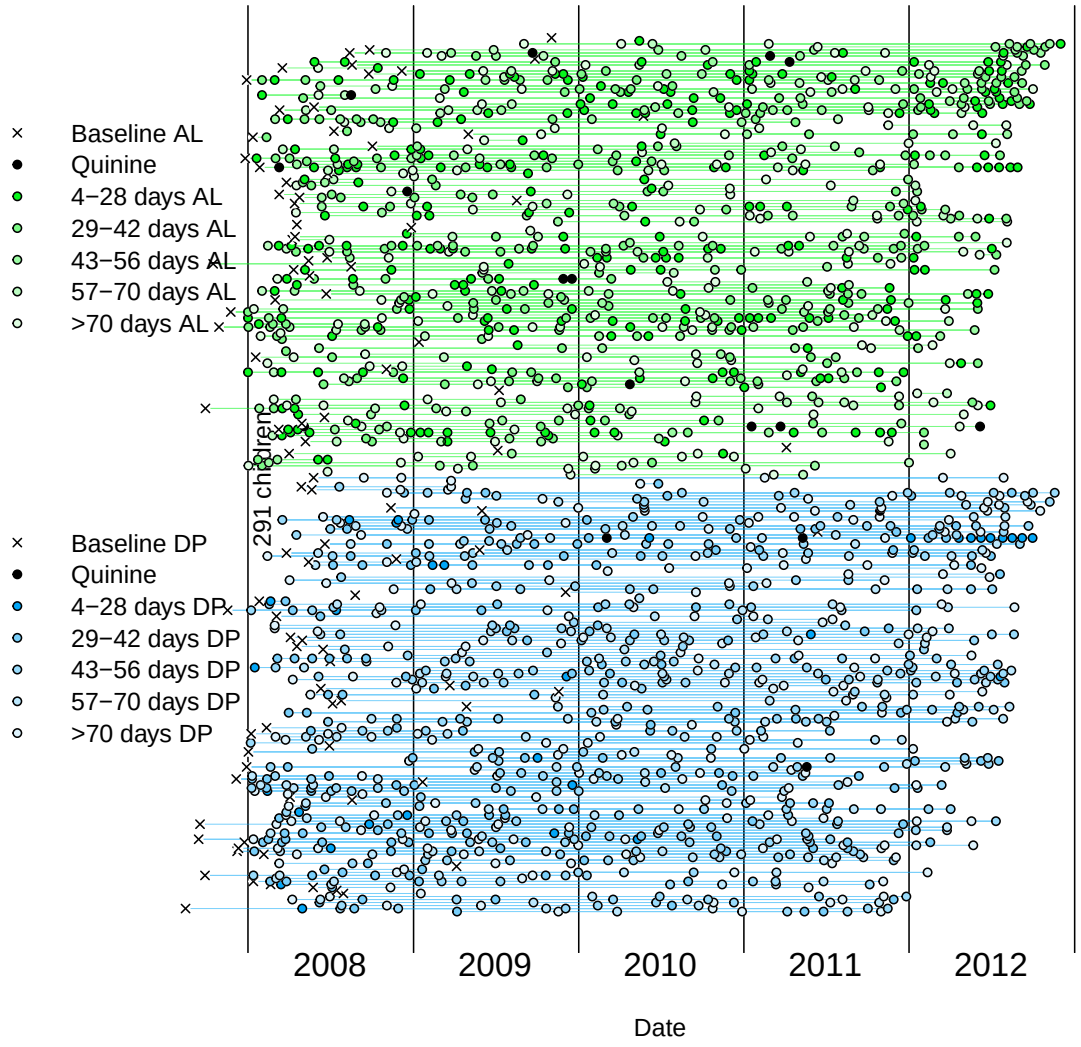


Figure 4.1: A graphical summary of the malaria episodes selected for genetic analysis. The plot shows 1989 of 5564 malaria episodes selected for genetic analysis over the course of a clinical trial [51, 266]. A single blood sample is collected per malaria episode. In total, samples are obtained from 291 children, 259 of whom provided two or more samples. Green horizontal lines represent children randomised to the AL drug arm who provided two or more samples. Blue horizontal lines represent children randomised to the DP drug arm who provided two or more samples. Malaria episodes are plotted with respect to time on the horizontal axis. Vertical lines delineate years. The episodes are either categorised as episodes treated with quinine, either because of treatment failure or severe malaria (black dots); baseline episodes, for which no prior treatment is given (crosses); or longitudinal episodes (episodes following one or more episodes in the same child). Longitudinal episodes are colour coded by drug arm (DP blue, AL green), and by days since last treatment (saturation of the coloured points — see legend).

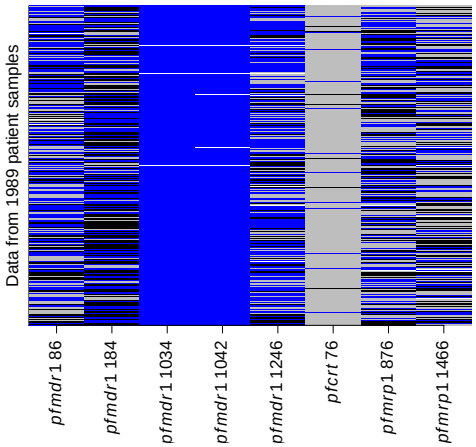


Figure 4.2: A graphical representation of the prevalence data. The coloured grid shows genotyping outcomes for 1989 samples genotyped at eight nSNPs. The data for each sample are stored on the rows (one row per sample). The outcomes for each nSNP genotyped are stored in the columns (one column per nSNP). For a given sample and nSNP, blue represents the detection of wild type alleles only, grey represents the detection of mutant alleles only, black represents the detection of both wild and mutant type alleles, and white represents missing data due to a failed assay.

changes in haplotype frequencies, longitudinal data are grouped by year (panel D, figure 4.3). To investigate changes in haplotype frequencies associated with immediacy since treatment, longitudinal data are categorised by the number of days since last treatment (panel E, figure 4.3). In total, partitioning results in 66 subdivisions of the data: 6 corresponding to baseline samples (coloured endpoints, panel C, figure 4.3), each comprising 50 samples; and 60 corresponding to longitudinal samples (coloured endpoints, panels E and D, figure 4.3), comprising different numbers of samples depending on the division (table 4.1). Data derived from episodes treated with quinine (black dots, figure 4.1) are included in both yearly subdivisions and subdivisions partitioned by days since last treatment. In hindsight, data points associated with quinine should have been discarded since they do not exert the same drug pressure as DP or AL. Their inclusion is unlikely to bias the results, however, since they constitute $< 1\%$ of the longitudinal episodes.

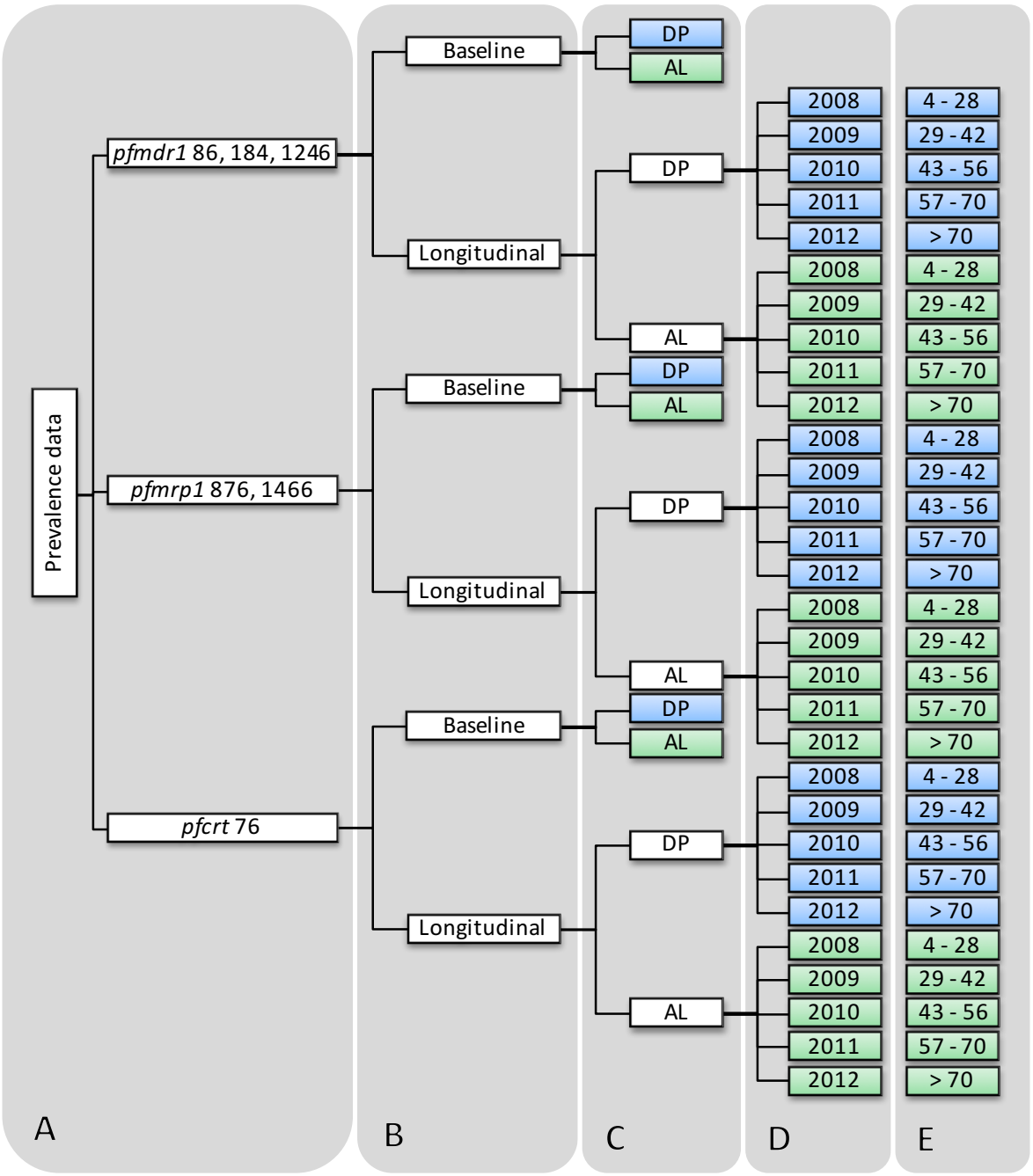


Figure 4.3: A schematic summarising how the prevalence data are partitioned. Panel A, the marker data are separated by gene. Panel B, data from baseline and longitudinal samples are separated. Panel C, samples are segregated by drug arm. Panel D, longitudinal samples are categorised by year (2008–2012). Panel E, longitudinal samples are categorised by days since last treatment (>70, 57–70, 43–56, 29–42 and 4–28 days). In total, the 66 colour-coded subdivisions (panels C, D and E) are analysed separately using the haplotype frequency estimation model (section 3.2.3).

Drug arm	Days since last treatment				
	>70	57–70	43–56	29–42	4–28
DP	252	163	310	181	33
AL	200	79	150	236	285
Drug arm	**Year**				
	2008	2009	2010	2011	2012
DP	200	200	200	200	139
AL	200	200	200	200	150

Table 4.1: Numbers of blood samples in data subdivisions based on days since last treatment and year.

4.2.3 Estimating frequencies

Vectors of *pfcrt* allele frequencies encoding amino acid residues $\boldsymbol{\pi}_{\text{PfCRT}} = (\pi_{\text{K}}, \pi_{\mathbf{T}})$, and vectors of *pfmrp1* and *pfmdr1* haplotype frequencies encoding amino acid sequences,

$$\boldsymbol{\pi}_{\text{PfMRP1}} = (\pi_{\text{IK}}, \pi_{\mathbf{IR}}, \pi_{\mathbf{VK}}, \pi_{\mathbf{VR}}) \text{ and}$$

$$\boldsymbol{\pi}_{\text{PfMDR1}} = (\pi_{\mathbf{Y}\text{YD}}, \pi_{\mathbf{YF}\text{D}}, \pi_{\mathbf{Y}\text{Y}\mathbf{Y}}, \pi_{\mathbf{YFY}}, \pi_{\text{NYD}}, \pi_{\text{NF}\mathbf{D}}, \pi_{\text{NY}\mathbf{Y}}, \pi_{\text{NF}\mathbf{Y}},),$$

respectively, are estimated using the haplotype frequency estimation model (equation (3.5)), described previously (chapter 3). In estimating frequencies under the aforesaid model, the simplifying assumption that all malaria episodes within a given subdivision are independent is made. In other words, structure due to repeat infections from multiple children is ignored. Note that the impact of this assumption is explored in chapter 5. The model is fit to each of the 66 subdivisions of the data (coloured end points, panels C, D and E, figure 4.3) in turn. The data are analysed using a truncated geometric prior on the MOI (equation (3.12)). The choice of prior is based on a study of auxiliary MOI data (see section A.2). The minimum MOI is two if the sample-wise data are discernibly multiclonal (included one or more heteroallelic genotyping outcomes) and one otherwise. The maximum MOI is eight, based on auxiliary data (figure A.1). The MOI prior parameter, λ, is equal to the mean of the auxiliary MOI data (2.94).

To assess the sensitivity of the results to λ, the data are reanalysed as follows using values derived from the study of the auxiliary MOI data (section A.1). All 60 longitudinal subdivisions are reanalysed with $\lambda = 2.80$ and 3.08 (the lower and upper bounds of the 95% confidence interval of the overall mean, 2.94). Yearly subdivisions are further reanalysed with λ equal to the yearly average MOIs per drug arm (table A.1), while baseline subdivisions are further reanalysed using the baseline mean (3.30) and the bounds of the 95% confidence interval on the baseline mean (2.62–3.98).

The model is implemented using the previously described Gibbs sampler (section 3.2.4). The sampler is initialised by a uniform frequency vector. Initial haplotype count vectors and MOIs are drawn from their respective priors (equations (3.8) and (3.12), respectively). Missing data (white cells, figure 4.2) are imputed. The sampler is run for 50,000 iterations, found to be sufficient for convergence[1], with thinning interval equal to 10. Each run of the sampler generates a MCMC sample of frequency vectors. The first 40% are discarded as burnin. For a given subdivision of data, $\boldsymbol{y}$, the MCMC sample post thinning and removal of burnin, $\{\boldsymbol{\pi}^n\}_{n=1}^{3000}$, approximates the posterior with density $\rho(\boldsymbol{\pi} \mid \boldsymbol{y})$. For each of the 66 subdivisions, the marginal posterior density estimates generated under the model with $\lambda = 2.94$ are plotted (figures 4.4 to 4.7). The significance of differences between marginal density estimates is based on a comparison of the 2.5th to 97.5th percentile intervals: if the intervals do not overlap, the marginal posterior density estimates are deemed significantly different from one another at the 5% level.

4.2.4 Fitting a regression model to estimated frequencies

Notable trends are observed for *pfmdr1* haplotypes encoding **Y**YD, **YYY**, NYD and N**F**D (figures 4.7a to 4.7p). To further investigate *pfmdr1* haplotype frequency trends, yearly frequencies

[1] To check convergence, three preliminary chains initiated at different initial frequency vectors were run on each subdivision of the data. Trace plots of the frequencies were visually inspected and π_r PSRFs calculated [88]. After 50,000 iterations, the trace plots and π_r PSRFs indicated convergence for all datasets.

are regressed onto categorical covariates for drug arm and year, while frequencies categorised by days since last treatment are regressed onto categorical covariates for drug arm and the number of days since last treatment. Henceforth we refer to both regression exercises (that for the yearly frequencies, and that for frequencies categorised by days since last treatment) as regression onto drug type (AL and DP arm) and pressure (categorical variables for years and for days since last treatment). A separate regression is fit to each of the eight *pfmdr1* haplotypes in turn.

The following subsections include a description of the regression model, a subsection regarding prior specification, a subsection about propagating uncertainty and a description of the sampler used to implement the model.

Notation

For the rth haplotype, let $\boldsymbol{\pi}_r = (\pi_{r1}, \ldots, \pi_{rK})^T$ denote the vector of frequencies that feature in a given regression exercise, where there are two regression exercises per haplotype (one for yearly frequencies and another for frequencies categorised by days since last treatment) and $\pi_{rk} \in (0,1)$ for $k = 1, \ldots, K$. For example, for the yearly estimates of the haplotype encoding **Y**YD,

$$\boldsymbol{\pi}_r \equiv \boldsymbol{\pi}_{\mathbf{Y}\text{YD}} = (\pi_{\mathbf{Y}\text{YD DP 2008}}, \ldots, \pi_{\mathbf{Y}\text{YD DP 2012}}, \pi_{\mathbf{Y}\text{YD AL 2008}}, \ldots, \pi_{\mathbf{Y}\text{YD AL 2012}})^T .$$

Note that $\boldsymbol{\pi}_r \neq \boldsymbol{\pi} = (\pi_1, \ldots, \pi_R)^T$, the latter being a vector of frequencies for the different haplotypes $r = 1, \ldots R$ (see section 3.2.3). Let $\boldsymbol{\theta}_r = (\theta_{r1}, \ldots, \theta_{rK})^T$ denote the vector $\boldsymbol{\pi}_r$ mapped onto the real line,

$$\theta_{rk} = \log\left(\frac{\pi_{rk}}{1-\pi_{rk}}\right) \text{ for } k = 1, \ldots K. \tag{4.1}$$

Regression

The vector $\boldsymbol{\theta}_r$ is regressed onto correlates of drug type and pressure following,

$$\boldsymbol{\theta}_r = \boldsymbol{X}\boldsymbol{\beta}_r + \boldsymbol{\varepsilon}_r, \tag{4.2}$$

where $\boldsymbol{X}$ denotes the $K \times P$ design matrix, where $P = 4$ (see below); $\boldsymbol{\beta}_r = (\beta_1, \cdots, \beta_P)^T$ is a vector of regression coefficients[2]; and $\boldsymbol{\varepsilon}_r = (\varepsilon_{r1}, \ldots, \varepsilon_{rK})^T$ a vector of errors. The errors are assumed to be independently and identically distributed according to a normal distribution with mean zero and variance σ^2:

$$\varepsilon_{rk} \sim \mathscr{N}\text{ormal}(0, \sigma_r^2) \text{ for } k = 1, \cdots, K. \tag{4.3}$$

[2]On the logit scale, β_1 is the expected haplotype frequency for children in the DP arm with zero drug pressure; β_2 is the expected difference in the frequency between children in the DP and AL arm with zero drug pressure; β_3 is the expected difference in the frequency between children in the DP arm who vary by a unit change in drug pressure (the slope of the DP arm); and β_4 is the expected difference between the slope of the DP arm (β_3) and the slope of the AL arm ($\beta_3 + \beta_4$).

Two inputs (drug type and drug pressure) and $P = 4$ predictors (intercept, drug type, drug pressure, and an interaction term) are encoded in the $R \times P$ design matrix,

$$\boldsymbol{X} = \begin{array}{c} \begin{array}{cccc} \text{Intercept} & \text{Drug type} & \text{Drug Pressure} & \text{Interaction} \end{array} \\ \left(\begin{array}{cccc} 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 2 & 0 \\ 1 & 0 & 3 & 0 \\ 1 & 0 & 4 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 2 & 2 \\ 1 & 1 & 3 & 3 \\ 1 & 1 & 4 & 4 \end{array} \right) \end{array}. \tag{4.4}$$

where for $k = 1, \ldots K$, $x_{k1} = 1$ allows estimation of the intercept; x_{k2} is an indicator variable for drug type (0 for randomisation to the DP arm, 1 for randomisation to the AL arm); x_{k3} is an ordered categorical variable for drug pressure (0–4 for years 2008–2012, respectively, for the regression with yearly frequencies; or 0–4 for treatment >70 days ago to 4–28 days ago, respectively, for the regression with frequencies categorised by days since last treatment)[3]; and $x_{k4} = x_{k2} \times x_{k3}$ allows estimation of the interaction between drug type and drug pressure, permitting the effect of drug pressure to differ depending on the drug arm.

[3]We count backwards in terms of days since treatment more that 70 days ago, since 70 days ago is thought to exert 'zero' drug pressure, while treatment 4–28 days ago is thought to exert maximum drug pressure.

Bayesian inference

The regression model is fit within a Bayesian framework,

$$\rho(\boldsymbol{\beta}_r, \sigma_r^2 | \boldsymbol{\theta}_r) \propto \rho(\boldsymbol{\theta}_r | \boldsymbol{\beta}_r, \sigma_r^2) \rho(\boldsymbol{\beta}_r, \sigma_r^2). \tag{4.5}$$

where $\boldsymbol{\beta}_r$ and σ_r^2 are the regressions parameters and $\boldsymbol{\theta}_r$ is the 'data' vector. The term 'data' is quoted since the elements of $\boldsymbol{\theta}_r$ are not, in actuality, observed. Instead they are inferred using the haplotype frequency estimation model (see section 4.2.3). Moreover, the haplotype frequency estimation model generates MCMC samples, not point estimates (although point estimates are obtainable from the MCMC sample). The manner in which we reconcile $\boldsymbol{\theta}_r$ and the MCMC sample is described below, in the subsection entitled 'Propagating uncertainty'. Note that in equation (4.5), we treat the design matrix, $\boldsymbol{X}$, as non-stochastic, as is typically the case when the parameters governing $\boldsymbol{X}$ are thought to be independent to those of interest [88]. Following equations (4.2) and (4.3),

$$\rho(\boldsymbol{\theta}_r \mid \boldsymbol{\beta}_r, \sigma_r^2) = \mathcal{N}\text{ormal}_K(\boldsymbol{\theta}_r \mid \boldsymbol{X}\boldsymbol{\beta}_r, \sigma_r^2 \boldsymbol{I}_K), \tag{4.6}$$

where $\boldsymbol{I}_K$ is the $K \times K$ identity matrix. To assess the robustness of the results to the prior on the regression parameters, $\rho(\boldsymbol{\beta}_r, \sigma_r^2)$, four conjugate priors are used, each giving rise to a normal inverse gamma posterior with density

$$\rho\left(\boldsymbol{\beta}_r, \sigma_r^2 | \boldsymbol{\theta}_r\right) = \mathcal{N}\text{ormal}_K(\boldsymbol{\beta}_r \mid \boldsymbol{\mu}_r, \sigma_r^2 \boldsymbol{V}_r) \times \mathcal{I}\text{nverse } \mathcal{G}\text{amma}(\sigma_r^2 \mid a_r, b_r) \tag{4.7}$$

where $\boldsymbol{\mu}_r$ is the mean, $\sigma_r^2 \boldsymbol{V}_r$ is the covariance matrix of the conditional posterior on $\boldsymbol{\beta}_r$ given σ_r^2, and a_r and b_r are the scale and shape parameters, respectively, of the marginal posterior on σ_r^2. The analytical expressions for $\boldsymbol{\mu}_r$, $\boldsymbol{V}_r$, a_r and b_r under the four different priors are listed in the appendix (section A.3).

Prior specification

As previously mentioned, four conjugate priors are fit, including a standard improper flat prior, Zellner's g prior and two normal inverse gamma priors. The two normal inverse gamma priors differ only in terms of a single hyperparameter. All four priors are described below.

1. A standard improper flat prior (see [4], page 206), with density

$$\rho(\boldsymbol{\beta}_r, \sigma_r^2) \propto \frac{1}{\sigma_r^2}. \tag{4.8}$$

2. Zellner's g prior (see [4], page 217), with density

$$\rho(\boldsymbol{\beta}_r, \sigma_r^2) \propto \mathcal{N}\text{ormal}_K\left(\boldsymbol{\beta}_r \mid \boldsymbol{\mu}_0, c\sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}\right) \times \frac{1}{\sigma_r^2}, \tag{4.9}$$

   where

   - $c = K$, such that prior information has equal weight to that of a single θ_{rk}; and
   - $\boldsymbol{\mu}_0 = (\log(1/7), 0, 0, 0)^T$, such that $\exp(\boldsymbol{x}_k\boldsymbol{\mu}_0)/(\exp(\boldsymbol{x}_k\boldsymbol{\mu}_0)+1) = 1/8$ for $k = 1, \ldots K$ (in other words, all eight haplotypes are thought to be equally likely *a priori*).

3. A normal inverse gamma prior (see [187], page 246), with density

$$\rho(\boldsymbol{\beta}_r, \sigma_r^2) = \mathcal{N}\text{ormal}_P(\boldsymbol{\beta}_r \mid \mu_0, \sigma^2\boldsymbol{V}_0) \times \mathcal{I}\text{nverse }\mathcal{G}\text{amma}(\sigma_r^2 \mid a_0, b_0), \tag{4.10}$$

   where

   - $\boldsymbol{\mu}_0 = (\log(1/7), 0, 0, 0)^T$, as for Zellner's g prior above;
   - $\boldsymbol{V}_0 = 10 \times \boldsymbol{I}_P$, where $\boldsymbol{I}_P$ is the $P \times P$ identity matrix, representing a scenario in which confidence in the prior mean $\boldsymbol{\mu}_0$ is small compared with $\hat{\boldsymbol{\beta}}_r$; and
   - $a_0 = 2$ and $b_0 = 1/3000 \sum_{n=1}^{3000} \{S_r^n/(K-p)\}$ where

$$S_r^n = \left(\boldsymbol{\theta}_r^n - \boldsymbol{X}\hat{\boldsymbol{\beta}}_r^n\right)^T \left(\boldsymbol{\theta}_r^n - \boldsymbol{X}\hat{\boldsymbol{\beta}}_r^n\right) \text{ and } \hat{\boldsymbol{\beta}}_r^n = \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{\theta}_r^n, \tag{4.11}$$

such that $\mathbb{E}[\sigma_r^2] = {}^{b_0}/_{(a_0-1)}$ is equal to the residual variance of the MCMC sample averaged over $n = 1, \ldots, 3000$.

4. Another normal inverse gamma prior with density identical to that above (equation (4.10)) but with

$$\boldsymbol{V}_0 = \left\{ 16.59^2 \times \frac{a_0}{b_0},\ 4.15^2 \times \frac{a_0}{b_0},\ 4.15^2 \times \frac{a_0}{b_0},\ 4.15^2 \times \frac{a_0}{b_0} \right\} \times I_P, \tag{4.12}$$

chosen such that 70% of the prior mass for the non-intercepts is between 0.01 and 0.99 (on the frequency/probability scale) based on a recommendation for a weakly informative default prior distribution for logistic regression [87]. Also following the recommendations of [87], the binary input (drug type) is standardised such that it has mean equal to zero and differs by one in its upper and lower conditions, while drug pressure (encoded 0 to 4 in equation (4.4)) is standardised such that it has mean equal to zero and standard deviation equal to 0.5.

Propagating uncertainty

Since the haplotype frequency estimation model (chapter 3) is fit to each of the 20 *pfmdr1* data subdivisions (top ten coloured boxes, panels D and E, figure 4.3) separately, we have 20 MCMC samples, each approximating a different posterior distribution given a specified *pfmdr1* data subdivision. We now want to perform a regression of *pfmdr1* haplotype frequencies. We could regress posterior point estimates; however, doing so would ignore uncertainty in the haplotype frequency estimation. To take full account of the uncertainty, one ought to fit the joint model that both estimates *pfmdr1* haplotype frequencies and regression coefficients. The joint model is likely to be computationally complex, due to the joint distribution over the *pfmdr1* haplotypes (we cannot estimate frequencies for each haplotype separately, because estimation relies upon information borrowed across the haplotypes). Using the meta-analytic

approach proposed by Lunn *et al.* [136], the complexity can be circumvented, in part, by fitting the regression to each haplotype separately. The approach of Lunn *et al.* is essentially a two-stage approximation of the joint model. In our case, frequency estimation is the first stage, while regression is the second stage. We propagate the uncertainty from the first stage into the second stage by iteratively drawing frequencies from the MCMC sample generated following haplotype frequency estimation. Following [136], we use a Gibbs sampler to iteratively update the regression parameters given the frequencies, then the frequencies given the regression parameters and the data. The frequency update requires a Metropolis-Hastings step. The idea of Lunn *et al.* is to use the posterior from the first stage as a proposal in the Metropolis-Hastings step and to use the MCMC samples to approximate it [136] (see equation (4.13), below).

The sampler

For the rth haplotype and the kth subdivision, let $\left\{\bar{\theta}_{rk}^{n}\right\}_{n=1}^{3000}$ denote the posterior summary of frequencies mapped onto the real line according to equation (4.1). Note that in section 4.2.3, we use $\boldsymbol{y}$ to denote any given subdivision of the data, whereas here we use $\boldsymbol{y}_k$ to particularise the kth subdivision, despite $\boldsymbol{y}_k$ and $\boldsymbol{y}$ being the same by definition. At $t = 0$, for $k = 1, \ldots, K$, we sample $\theta_{rk}^{(t)}$ uniformly at random from $\left\{\bar{\theta}_{rk}^{n}\right\}_{n=1}^{3000}$. We generate $\boldsymbol{\beta}_r^{(t)}$ and $\sigma_r^{2(t)}$ given $\boldsymbol{\theta}_r{}^{(t)} = \left(\theta_{r1}^{(t)}, \ldots, \theta_{rK}^{(t)}\right)$ by sampling directly from the normal inverse gamma posterior whose density is given by equation (4.7). For $t = 1, \ldots, T$, the sampler then proceeds as follows.

1. For $k = 1, \ldots K$, generate $\theta_{rk}^{(t)}$ given $\boldsymbol{\beta}_r^{(t-1)}, \sigma^{2}{}_r^{(t-1)}$ and $\boldsymbol{y}_k$ as follows. First sample $\theta_{rk}{}^{\star}$ with replacement uniformly at random from $\left\{\bar{\theta}_{rk}^{n}\right\}_{n=1}^{3000}$. Second, accept $\theta_{rk}{}^{\star}$ ($\theta_{rk}{}^{(t)} \leftarrow \theta_{rk}{}^{\star}$), or reject $\theta_{rk}^{\star}$ ($\theta_{rk}^{(t)} \leftarrow \theta_{rk}^{(t-1)}$), using a Metropolis-Hastings update with acceptance probability given by equation (4.13). Since $\left\{\bar{\theta}_{rk}^{n}\right\}_{n=1}^{3000} \approx \rho(\theta_{rk} \mid \boldsymbol{y}_k)$, calculate

the acceptance probability assuming $\mathscr{P}\text{roposal}(\theta_{rk}) \approx \rho\left(\theta_{rk} \mid \boldsymbol{y}_k\right)$,

$$\begin{aligned}
\mathbb{P}(\theta_{rk}^{(t)} \leftarrow \theta_{rk}^{\star}) &= \min\left(1, \frac{\mathscr{T}\text{arget}(\theta_{rk}^{\star})}{\mathscr{T}\text{arget}(\theta_{rk}^{(t-1)})} \frac{\mathscr{P}\text{roposal}(\theta_{rk}^{(t-1)})}{\mathscr{P}\text{roposal}(\theta_{rk}^{\star})}\right) \\
&\approx \min\left(1, \frac{\rho(\theta_{rk}^{\star} \mid \boldsymbol{\beta}_r^{(t-1)}, \sigma^{2(t-1)}_r, \boldsymbol{y}_k)}{\rho(\theta_{rk}^{(t-1)} \mid \boldsymbol{\beta}_r^{(t-1)}, \sigma^{2(t-1)}_r, \boldsymbol{y}_k)} \frac{\rho(\theta_{rk}^{(t-1)} \mid \boldsymbol{y}_k)}{\rho(\theta_{rk}^{\star} \mid \boldsymbol{y}_k)}\right) \\
&= \min\left(1, \frac{\rho(\boldsymbol{y}_k \mid \theta_{rk}^{\star})\rho(\theta_{rk}^{\star} \mid \boldsymbol{\beta}_r^{(t-1)}, \sigma^{2(t-1)}_r)}{\rho(\boldsymbol{y}_k \mid \theta_{rk}^{(t-1)})\rho(\theta_{rk}^{(t-1)} \mid \boldsymbol{\beta}_r^{(t-1)}, \sigma^{2(t-1)}_r)} \frac{\rho(\boldsymbol{y}_k \mid \theta_{rk}^{(t-1)})\rho(\theta_{rk}^{(t-1)})}{\rho(\boldsymbol{y}_k \mid \theta_{rk}^{\star})\rho(\theta_{rk}^{\star})}\right) \\
&= \min\left(1, \frac{\rho(\theta_{rk}^{\star} \mid \boldsymbol{\beta}_r^{(t-1)}, \sigma^{2(t-1)}_r)}{\rho(\theta_{rk}^{(t-1)} \mid \boldsymbol{\beta}_r^{(t-1)}, \sigma^{2(t-1)}_r)} \frac{\rho(\theta_{rk}^{(t-1)})}{\rho(\theta_{rk}^{\star})}\right) \\
&= \min\left(1, \frac{\rho(\theta_{rk}^{\star} \mid \boldsymbol{\beta}_r^{(t-1)}, \sigma^{2(t-1)}_r)}{\rho(\theta_{rk}^{(t-1)} \mid \boldsymbol{\beta}_r^{(t-1)}, \sigma^{2(t-1)}_r)} \frac{\rho(\pi_{rk}^{(t-1)}) \mid \mathrm{d}\pi_{rk}^{(t-1)}/\mathrm{d}\theta_{rk}^{(t-1)} \mid}{\rho(\pi_{rk}^{\star}) \mid \mathrm{d}\pi_{rk}^{\star}/\mathrm{d}\theta_{rk}^{\star} \mid}\right),
\end{aligned} \tag{4.13}$$

where $\rho(\theta_{rk} \mid \boldsymbol{\beta}_r, \sigma_r^2)$ is given by equation (4.6); $\pi_{rk} = {\exp(\theta_{rk})}/{(1+\exp(\theta_{rk}))}$ (the inverse of the logit function in equation (4.1)); $\rho(\pi_{rk}) = \mathscr{B}\text{eta}(1, R-1)$, because $\rho(\pi_{rk})$ is the marginal of a R dimensional uniform Dirichlet distribution (see section 3.2.3); and ${\mathrm{d}\pi_{rk}}/{\mathrm{d}\theta_{rk}} = {\exp(\theta_{rk})}/{(1+\exp(\theta_{rk}))^2}$.

2. Generate $\boldsymbol{\beta}_r^{(t)}$ and $\sigma_r^{2(t)}$ given $\boldsymbol{\theta}_r^{(t)} = \left(\theta_{r1}^{(t)}, \ldots, \theta_{rK}^{(t)}\right)$ by sampling directly from the normal inverse gamma posterior whose density is summarised by equation (4.7).

Four chains of the sampler are run for $T = 10{,}000$ iterations for each regression. The first 50% per chain is removed as burnin. The output of the sampler post removal of burnin approximates the joint posterior with density $\rho\left(\boldsymbol{\beta}_r, \sigma_r^2, \boldsymbol{\theta}_r | \boldsymbol{y}_1, \ldots, \boldsymbol{y}_K\right)$. In total, the regression is performed 64 times: two regression exercises (one for yearly frequencies, and another for frequencies categorised by days since last treatment) for each of the eight *pfmdr1* haplotypes under four different priors. Standard diagnostic plots[4] are used to assess the performance

[4]Standard diagnostic plots include trace plots, plots of PSRFs and autocorrelation plots to assess the mixing, convergence and autocorrelation, respectively, of $\boldsymbol{\beta}_r, \sigma_r^2$ and the resampled $\boldsymbol{\theta}_r$; percentile percentile plots to assess the difference between $\boldsymbol{\theta}_r$ samples before and after resampling; and density plots to assess the posterior samples of $\boldsymbol{\beta}_r, \sigma_r^2$, as well as the overlap between $\boldsymbol{\theta}_r$ samples before and after resampling.

of the sampler. To assess difference between trends, posterior summaries of the slopes are compared. The significance of a trend is based on the 2.5th to 97.5th percentile interval of the posterior distribution over the slope: if the interval does not contain zero, the trend is deemed significantly non-zero at the 5% level.

4.3 Results

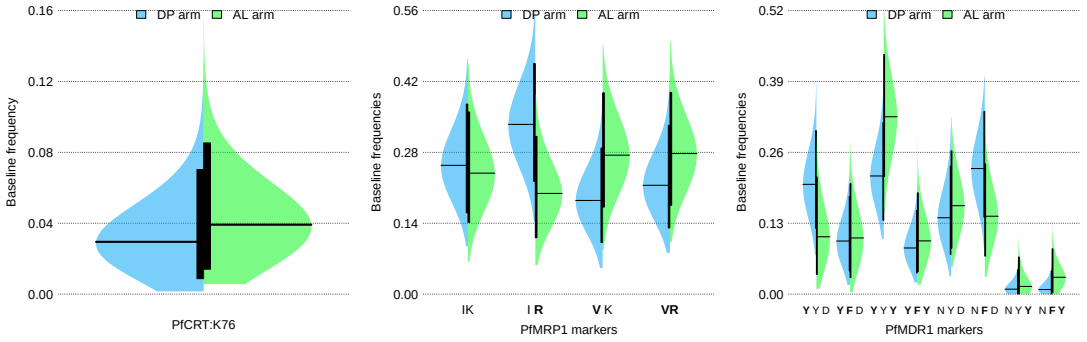
4.3.1 Frequency estimates

Sensitivity to the mean MOI

The difference between haplotype frequencies estimated using $\lambda = 2.94, 2.80$ and 3.08 is less than 0.01. The maximum difference between baseline and yearly haplotype frequencies re-estimated using the mean baseline MOI ($\lambda = 3.30$) and yearly MOIs (table A.1), respectively, is less than 0.01. Hence, we conclude that the frequency estimates are robust over the λ range based on the observed MOI data.

4.3.2 Baseline frequencies

As controls for comparison, baseline frequencies are estimated. There are no notable differences between the frequencies associated with PfCRT:K76 in the different drug arms at baseline (figure 4.4a). The baseline frequencies of all but one of the *pfmrp1* haplotypes (represented by their corresponding amino acid sequences) differ notably between the two drug arms (figure 4.4b), but the differences are not significantly different from zero. The posterior frequencies of the *pfmdr1* haplotypes range from 0.0 to 0.5 (figure 4.4c). Markers NY**Y** and N**F**Y are comparatively rare, while markers **Y**YD, **YYY**, and N**F**D have relatively high frequencies that differed between the two drug arms; the differences are not significantly different from zero, however.



(a) PfCRT:K76 marker frequencies (b) PfMRP1 marker frequencies (c) PfMDR1 marker frequencies

Figure 4.4: Marginal posterior density estimates of baseline frequencies colour coded by drug arm (DP blue, AL green). Black horizontal lines denote the median posterior estimates. Black vertical lines denote the 95% credible intervals ranging from the 2.5th percentile to the 97.5th percentile.

Longitudinal *pfcrt* allele frequencies

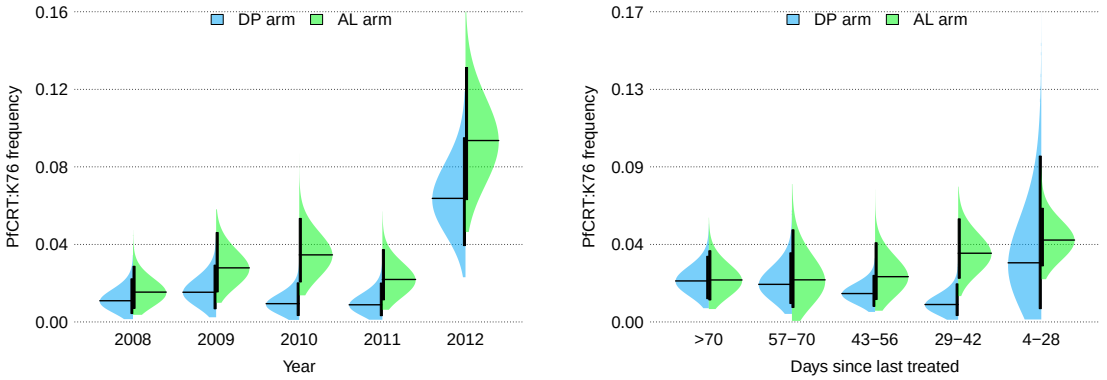
Estimated frequencies associated with PfCRT:K76 frequencies vary little from 2008 to 2011 (figure 4.5a). There is a notable increase in 2012, accompanied by a concomitant drop in the frequency associated with PfCRT:76**T** (result not shown). The increase is more marked in the AL arm, but the difference between the drug arms is not significantly different from zero. There does not appear to be any notable trends with respect to immediacy since last treatment (figure 4.5b).

Longitudinal *pfmrp1* haplotype frequencies

There are no notable trends in *pfmrp1* haplotype frequencies (figure 4.6). The variability between drug arms is similar to that seen at baseline (figure 4.4b).

Longitudinal *pfmdr1* haplotype frequencies

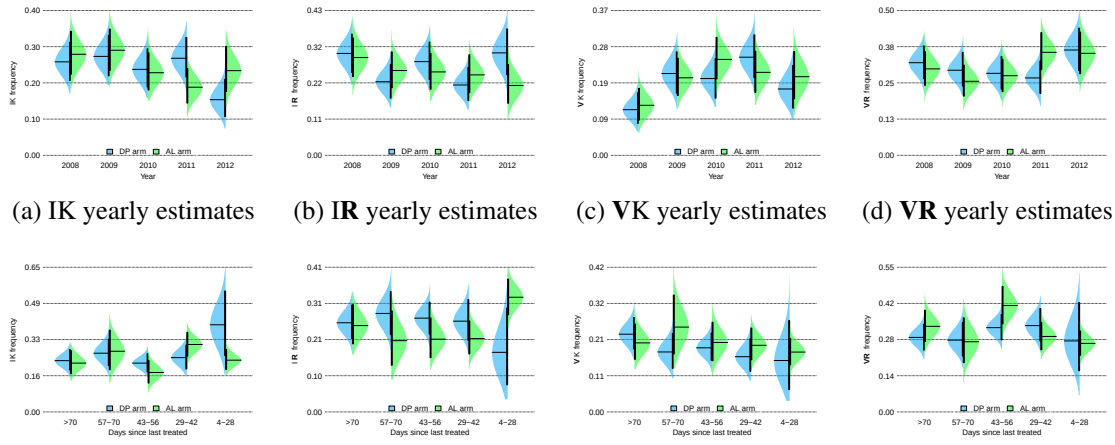
Four of the eight *pfmdr1* haplotypes (encoding **Y**YD, **YYY**, NYD and N**F**D) have relatively large frequencies that change over time (figures 4.7a to 4.7d, respectively) and with immediacy since last treatment (figures 4.7e to 4.7h, respectively). Haplotypes encoding **Y**FD, **YFY**, NY**Y**



(a) Yearly estimates

(b) Immediacy estimates

Figure 4.5: Marginal posterior density estimates of longitudinal frequencies encoding PfCRT:K76 colour coded by drug arm (DP blue, AL green). Black horizontal lines denote the median posterior estimates. Black vertical lines denote the 95% credible intervals ranging from the 2.5th percentile to the 97.5th percentile.



(a) IK yearly estimates

(b) I**R** yearly estimates

(c) **V**K yearly estimates

(d) **VR** yearly estimates

(e) IK Immediacy estimates

(f) I**R** Immediacy estimates

(g) **V**K Immediacy estimates

(h) **VR** Immediacy estimates

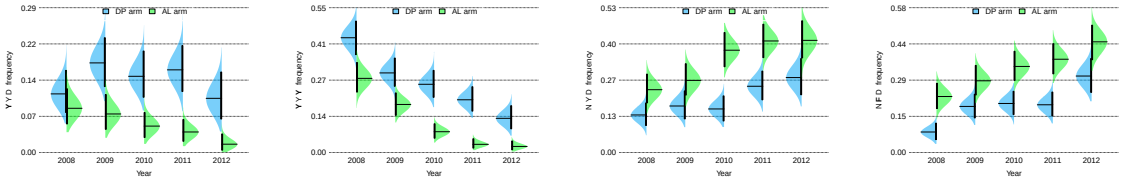
Figure 4.6: Marginal posterior density estimates of longitudinal *pfmrp1* haplotype frequencies (represented by their corresponding amino acid sequences) colour coded by drug arm (DP blue, AL green). Black horizontal lines denote the median posterior estimates. Black vertical lines denote the 95% credible intervals ranging from the 2.5th percentile to the 97.5th percentile.

and N**F**Y (figures 4.7i to 4.7l, respectively, and 4.7m to 4.7p, respectively) are comparatively rare (posterior median estimates all less than 0.1), and there are no notable trends associated with the latter two. There are clear differences between drug arms: the frequencies of haplotypes encoding **Y**YD, **YYY** and **YF**D, and all but one frequency encoding **YFY** (figures 4.7a, 4.7b, 4.7i, 4.7j, respectively, and 4.7e, 4.7f, 4.7m, 4.7n, respectively) are consistently higher in the DP arm; whereas, the frequencies of haplotypes encoding NYD and N**F**D are consistently higher in the AL arm (figures 4.7c and 4.7d, respectively, and 4.7g and 4.7h, respectively). The frequencies of haplotypes encoding NY**Y** and N**FY** look to be roughly the same across the two drug arms (figures 4.7k and 4.7l, respectively, and 4.7o and 4.7p, respectively) relative to baseline variation (figure 4.4c).

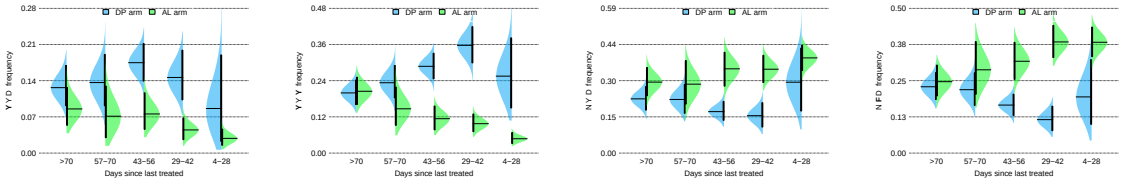
On the whole, yearly trends appear to be consistent across the two drug arms (figures 4.7a to 4.7d and 4.7i to 4.7l). There appear to be positive yearly trends associated with haplotypes encoding NYD and N**F**D (figures 4.7c and 4.7d, respectively) and negative yearly trends with haplotypes encoding **YYY** and **YFY** (figures 4.7b and 4.7j, respectively), and haplotypes encoding **Y**YD and **YF**D in the AL (but not the DP) arm (figures 4.7a and 4.7i, respectively). With respect to immediacy since last treatment, trends are consistent with AL selection for haplotypes encoding N**F**D (figure 4.7h) and NYD (figure 4.7g) and against haplotypes encoding **Y**YD (figure 4.7e), **YYY** (figure 4.7h) and **YF**D (figure 4.7m). DP trends, on the contrary, are seemingly consistent with selection for **YYY** (figure 4.7f) and against N**F**D (figure 4.7h). To further evaluate the trends, yearly *pfmdr1* haplotype frequencies are regressed onto yearly covariates, while frequencies categorised by days since last treatment are regressed onto covariates representing categories of days since last treatment.

4.3.3 Regression

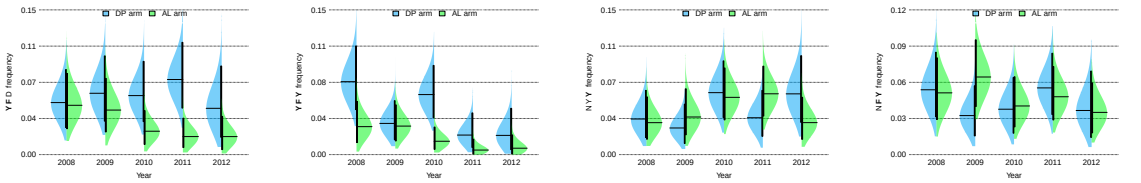
Before examining the fitted regression trends, let us consider the performance of the regression sampler under the different priors and the sensitivity of the trends to the prior.



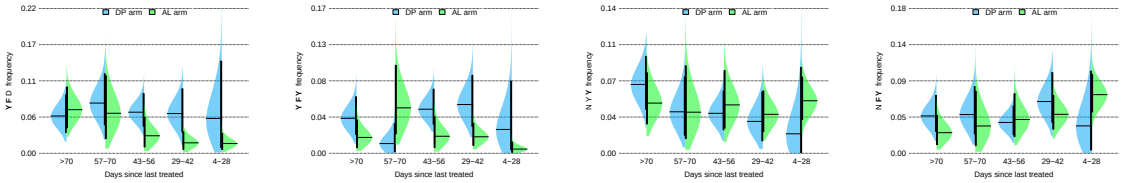
(a) **Y**YD yearly estimates (b) **YYY** yearly estimates (c) NYD yearly estimates (d) N**F**D yearly estimates



(e) **Y**YD Immediacy estimates (f) **YYY** Immediacy estimates (g) NYD Immediacy estimates (h) N**F**D Immediacy estimates



(i) **YF**D yearly estimates (j) **YFY** yearly estimates (k) NY**Y** yearly estimates (l) N**FY** yearly estimates



(m) **YF**D Immediacy estimates (n) **YFY** Immediacy estimates (o) NY**Y** Immediacy estimates (p) N**FY** Immediacy estimates

Figure 4.7: Marginal posterior density estimates of longitudinal *pfmdr1* haplotype frequencies (represented by their corresponding amino acid sequences) colour coded by drug arm (DP blue, AL green). Black horizontal lines denote the median posterior estimates. Black vertical lines denote the 95% credible intervals ranging from the 2.5th percentile to the 97.5th percentile.

Performance of the regression sampler under the different priors

In general, the diagnostic plots for the regression parameters $\boldsymbol{\beta}_r$ and σ_r^2 indicate convergence. Under all four prior models, there are examples of misalignment between the density estimates of the samples of θ_{rk} before and after resampling (for example, figure 4.8). Misalignment of θ_{rk} is coupled with relatively low θ_{rk} acceptance rate. Low acceptance is also paired with non-normality of $\left\{\bar{\theta}_{rk}^n\right\}_{n=1}^{3000}$, detected by normal quantile-quantile plots. Non-normality is not always paired with low acceptance, however. Misalignment is an inherent problem due to the discrete proposal. The extent of the problem varies with the different priors. Under Zellner's g prior there are only $^{11}/_{160}$ instances of θ_{rk} misalignment[5]. The worst example of misalignment under Zellner's g prior (figure 4.8) is coupled with a relatively high acceptance rate of approximately 40%. There are more instances of misalignment under the normal inverse gamma priors, coupled with lower θ_{rk} acceptance rates. Nevertheless, the diagnostic plots of $\boldsymbol{\beta}_r$ and σ_r^2 indicate convergence. The sampler does not converge under the flat prior and there are numerous instances of misalignment. In hindsight, difficulty to converge under the flat prior is inevitable because the trace-wise variance is entirely dependent upon the data (equations (A.8) and (A.9)). Under the informative priors (especially Zellner's g prior), incorporation of prior information leads to more diffuse marginal posterior distributions on σ_r^2. Consequently, if $\theta_{rk}^\star$ is not close to $\boldsymbol{x}_k\boldsymbol{\beta}_{rk}^{(t)}$, $\theta_{rk}^\star$ is more likely to be rejected under the model with the flat prior on $\rho\left(\boldsymbol{\beta}_r, \sigma_2^2\right)$ compared with the model with an informative prior.

Sensitivity of the regression trends to the prior

To assess the sensitivity of the reported trends to the prior on the regression parameters, posterior summaries of the slopes under the different priors are compared (figures 4.9 and 4.10). There is a notable difference in precision, mostly due to the flat prior (under which the sampler does

[5]There are 160 sets of $\{\bar{\theta}_{rk}^n\}_{n=1}^{3000}$ ($K = 10$ haplotype frequencies per regression, two regressions exercises per haplotype (one corresponding to yearly frequencies, another corresponding to frequencies categorised by days since last treatment) and $R = 8$ haplotypes).

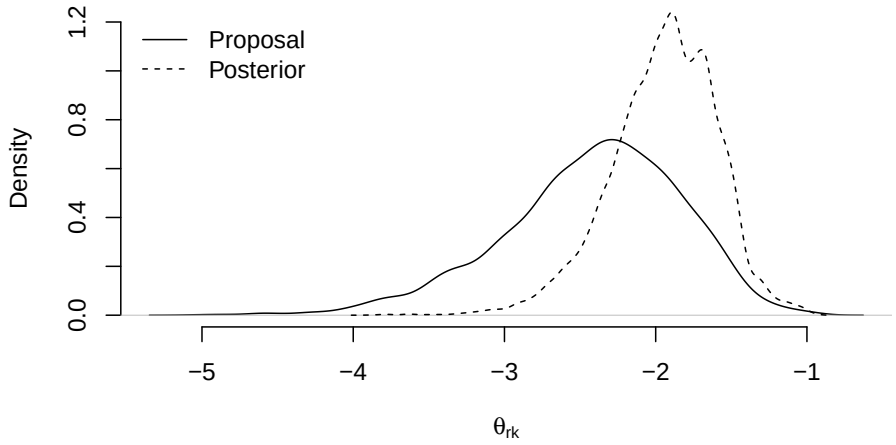


Figure 4.8: Density estimates of θ_{rk} before (proposal) and after (posterior) resampling under the regression model using Zellner's g prior. The plot shows estimates for the haplotype encoding **Y**YD ($r = 1$) and 4–28 days ($k = 5$).

not converge), but also due to Zellner's g prior (under which the sampler does converge). There are some small differences in the significance of the trends (summarised below), mainly due to the difference in precision. Despite small differences, the scientific interpretation is consistent: there are no differences that lead to a different interpretation of the direction of trend.

Summary of trends based on regression

The trends under all four priors are summarised in figures 4.9 and 4.10 on the logit scale. For the sake of illustration, the results on the probability scale generated under Zellner's g prior are depicted in figures 4.11 and 4.12. As 4.9 and 4.10 suggest, the results under Zellner's g prior are almost identical to those under the alternative priors (figures A.5 to A.4).

Yearly trends (figures 4.9 and 4.11). Under all four priors, regression supports a significantly non-zero negative trend for haplotypes encoding **Y**YD in the AL arm and **YYY** in both drug arms, and a non-zero positive trend for the haplotype encoding N**F**D in the DP arm. Under all but Zellner's g prior, there is evidence of a non-zero positive trend for haplotype encoding NYD and a non-zero negative trend for that encoding **YFY** in both drug arms. Furthermore, in

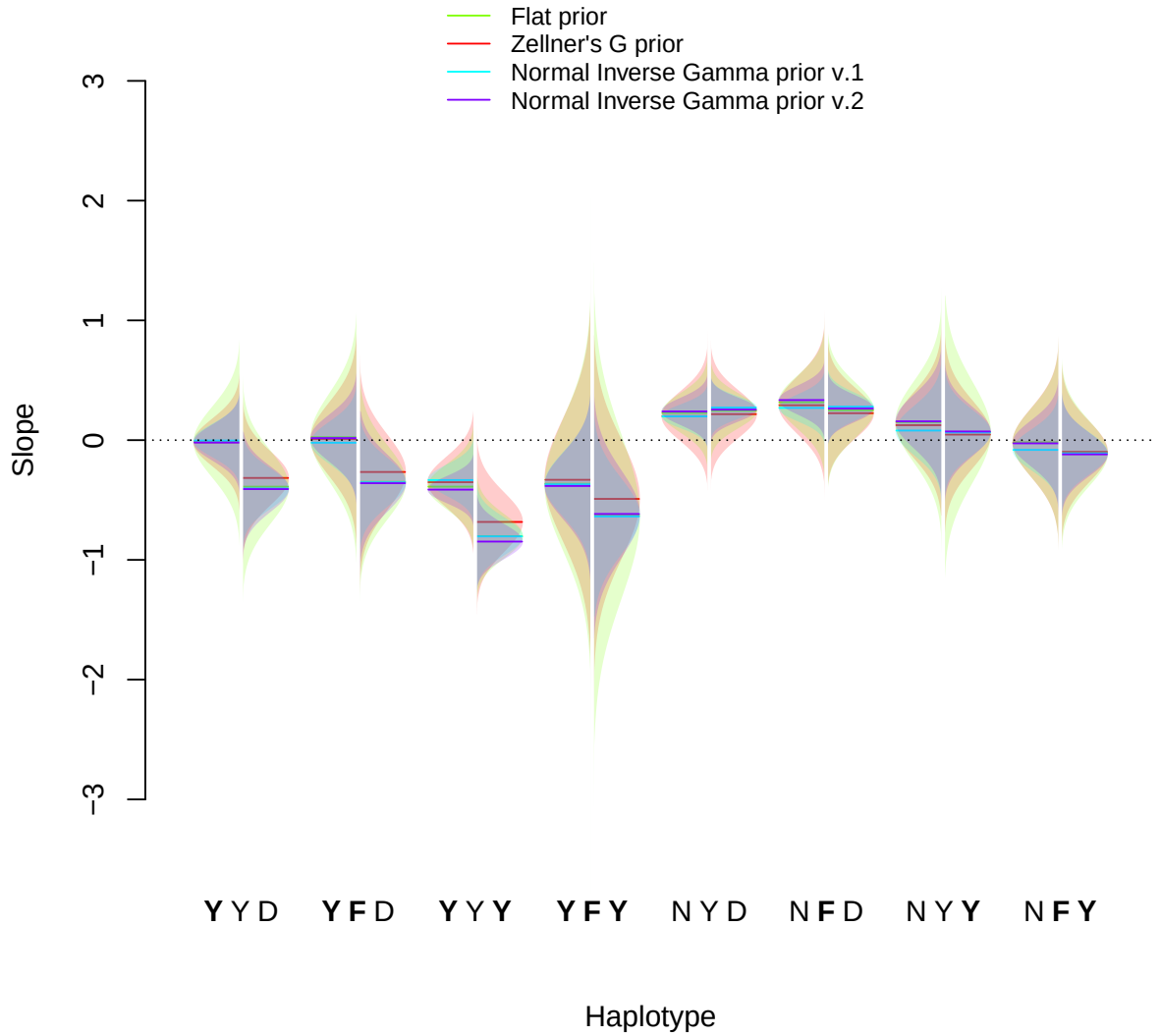


Figure 4.9: Sensitivity of the yearly trend (slope) to the prior. Haplotypes are denoted by their corresponding amino acid sequences. The sampler is run for 10,000 iterations using the informative priors and 50,000 iterations for the flat prior. The former converges but the latter does not. For each violin motif, the results corresponding to the DP drug arm are on the left hand side, while the results corresponding to the AL drug arm are on the right hand side. 'Normal Inverse Gamma prior v.1' refers to the first of the two normal inverse gamma priors listed in section 4.2.4 (equation (4.10), with prior matrix $\boldsymbol{V}_0 = 10 \times \boldsymbol{I}_p$), while 'Normal Inverse Gamma prior v.2' refers to the second of the two normal inverse gamma priors listed in section 4.2.4 (equation (4.10), where the prior matrix $\boldsymbol{V}_0$ is given by equation (4.12)). In order to generate a plot on a common scale, in the case of Normal Inverse Gamma prior v.2, the slope is divided by twice the standard deviation of the second input (the third column of the unnormalised design matrix, equation (4.4)).

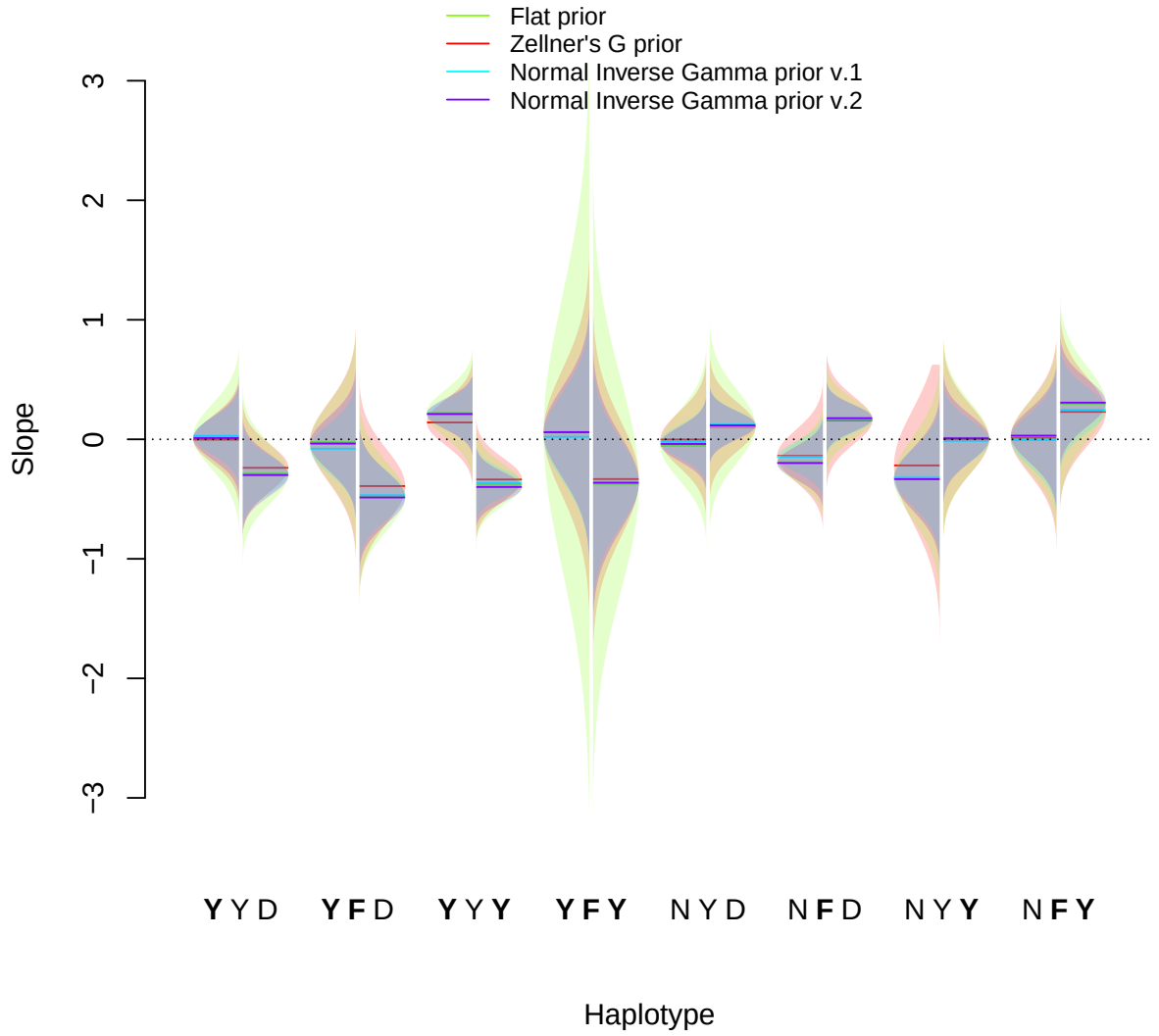


Figure 4.10: Sensitivity of the immediacy since last treatment trend to the prior. Haplotypes are denoted by their corresponding amino acid sequences. The sampler is run for 10,000 iterations using the informative priors and 50,000 iterations for the flat prior. The former converges but the latter does not. For each violin motif, the results corresponding to the DP drug arm are on the left hand side, while the results corresponding to the AL drug arm are on the right hand side. Normal Inverse Gamma prior v.1 refers to the first of the two normal inverse gamma priors listed in section 4.2.4 (equation (4.10), with prior matrix $\boldsymbol{V}_0 = 10 \times \boldsymbol{I}_p$), while Normal Inverse Gamma prior v.2 refers to the second of the two normal inverse gamma priors listed in section 4.2.4 (equation (4.10), where the prior matrix $\boldsymbol{V}_0$ is given by equation (4.12)). In order to generate a plot on a common scale, in the case of Normal Inverse Gamma prior v.2, the slope is divided by twice the standard deviation of the second input (the third column of the unnormalised design matrix, equation (4.4)).

the AL drug arm, there is evidence of a non-zero positive trend for the haplotype encoding N**F**D and a non-zero negative trend for that encoding **YF**D. There appear to be no net changes year on year for haplotypes encoding **Y**YD, **YF**D and N**FY** in the DP arm, nor for the haplotype encoding NY**Y** in the AL arm.

Trends with respect to immediacy since last treatment (figures 4.10 and 4.12). Under all four priors, regression supports a non-zero negative trend with AL treatment for the haplotype encoding **YYY**. Under all but Zellner's g prior, there is evidence for an inverse trend with DP treatment. A trend reversal is seen for that encoding N**F**D (both AL and DP trends are significantly non-zero under all but Zellner's g prior). Regression also suggests significantly non-zero negative trends with immediacy since last treatment with AL for haplotypes encoding **Y**YD and **YF**D. The frequency of the haplotype encoding NYD increases with immediacy since last treatment with AL (although the trend is not significantly different from zero). There is also partial evidence for a non-zero trend with immediacy since last treatment with DP for the haplotype encoding NY**Y** and with AL for the haplotype encoding N**FY**. Both haplotypes encoding N**FY** and NY**Y** are comparatively rare throughout, however. Median point estimates of at least six trends (**Y**YD, **YF**D, **YFY**, N**FY** and NYD in the DP arm, and NY**Y** in the AL arm) are zero (figure 4.12).

4.4 Discussion

In this chapter we report estimates of *pfcrt* allele and *pfmrp1* and *pfmdr1* haplotype frequencies (represented by the corresponding amino acid sequencies) to investigate trends over the course of clinical trial in Uganda. Estimates are based on previously published data [51], collected between 2007–2012 from children enrolled in a longitudinal trial of AL versus DP in Tororo, Uganda [15, 266]. This chapter builds upon the study by Conrad *et al.* [51], by investigating haplotype frequencies as well as allele frequencies. The estimates imply that there are notable

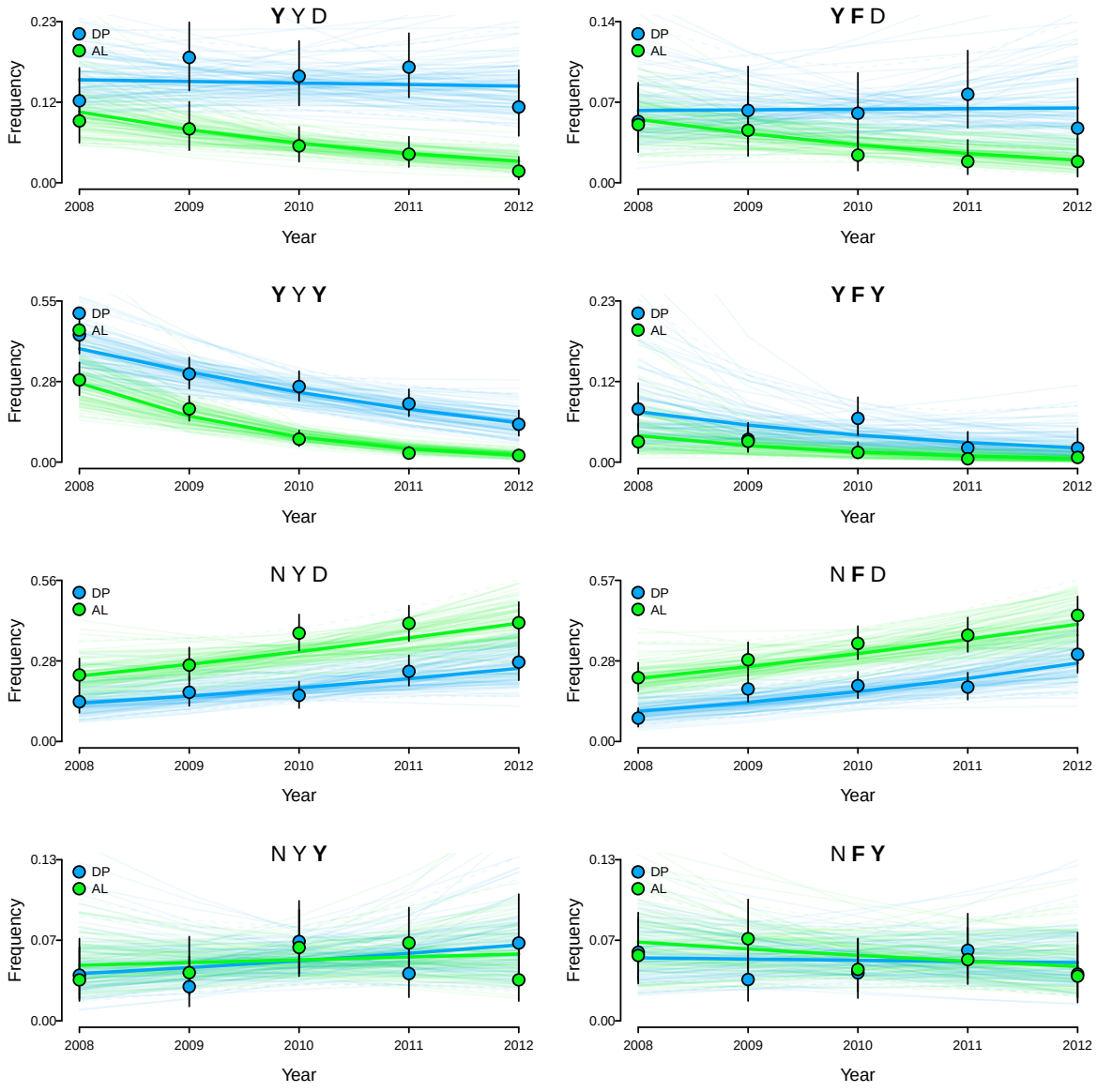


Figure 4.11: *Pfmdr1* haplotype frequency trends with year. The plots show *pfmdr1* haplotype frequencies (represented by the their corresponding amino acid sequences) categorised by drug arm and year regressed onto covariates of drug arm and year. Dots denote the MCMC sample estimates of the posterior median haplotype frequencies (DP blue, AL green) before resampling. Vertical black lines denote 95% credible intervals, ranging from the 2.5th percentile to the 97.5th percentile of the MCMC sample before resampling. The regression is performed using Zellner's g prior on the regression parameters. The thick blue and green lines (DP and AL, drug arms respectively) denote the trends constructed using the posterior median estimates of the regression coefficients. The thin blue and green lines represent trends based on 100 traces from the MCMC sample of regression coefficients.

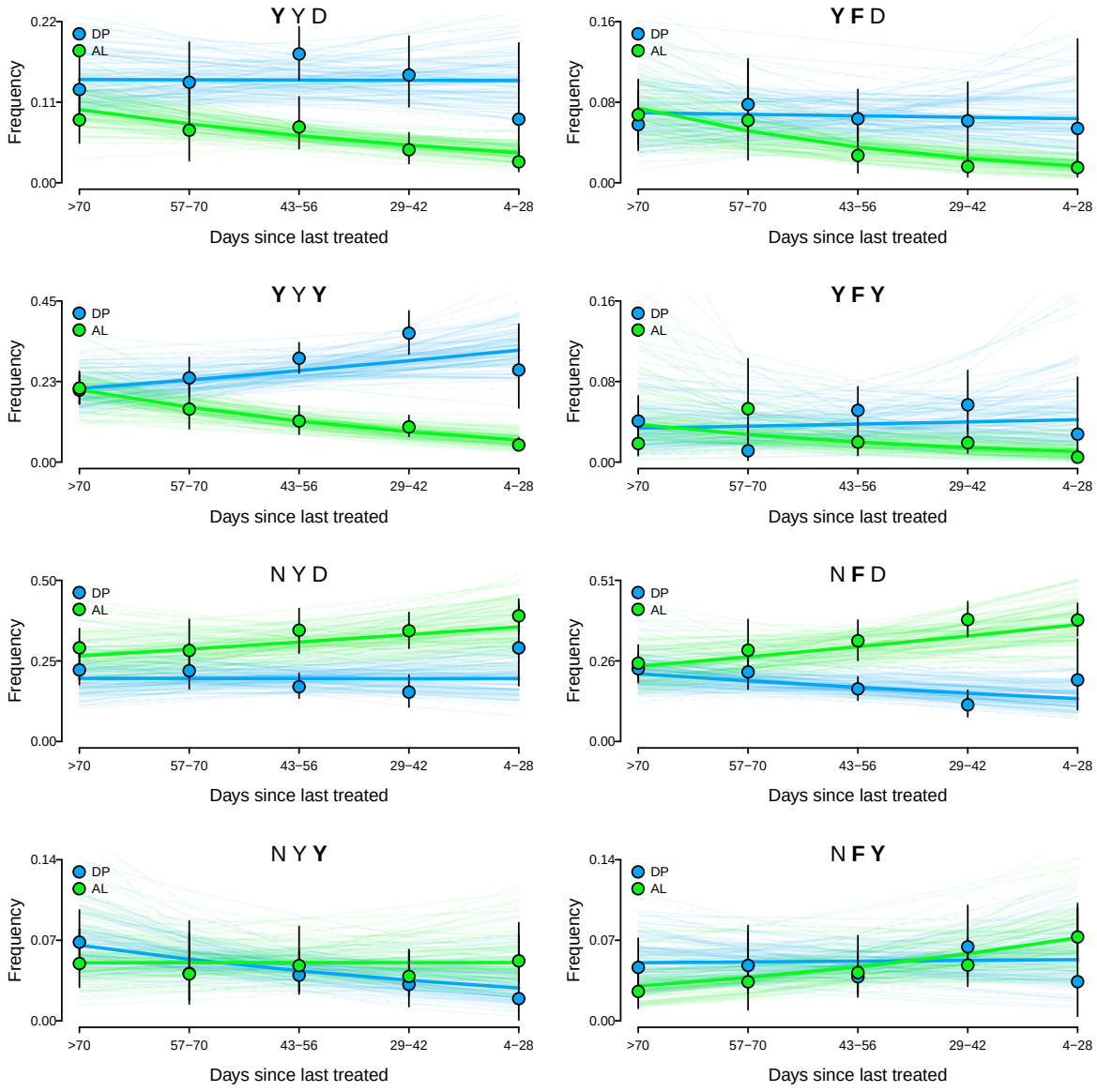


Figure 4.12: *Pfmdr1* haplotype frequency trends with immediacy since last treatment. The plot shows *pfmdr1* haplotype frequencies (represented by their corresponding amino acid sequences) categorised by drug arm and days since last treatment regressed onto covariates of drug arm and duration since last treatment. Dots denote the MCMC sample estimates of the posterior median haplotype frequencies (DP blue, AL green) before resampling. Vertical black lines denote 95% credible intervals, ranging from the 2.5th percentile to the 97.5th percentile of the MCMC sample before resampling. The regression is performed using Zellner's g prior on the regression parameters. The thick blue and green lines (DP and AL, drug arms respectively) denote the trends constructed using the posterior median estimates of the regression coefficients. The thin blue and green lines represent trends based on 100 traces from the MCMC sample of regression coefficients.

trends associated with the *pfmdr1* haplotype frequencies. To further examine the trends, *pfmdr1* haplotype frequencies are regressed onto drug pressure and drug type, using a meta-analysis approach (see [136]) to propagate the uncertainty in the frequency estimates into the regression. To the best of out knowledge, this is the first study to report *P. falciparum* multi-nSNP haplotype frequency trends in Uganda.

We observe an increase in the frequency of PfCRT:K76 in 2012 (figure 4.5a), consistent with prior reports based on the same data [51, 149]. Since PfCRT:K76 is associated with decreased sensitivity to lumefantrine [162], while PfCRT:76**T** is associated with CQ resistance [162], an increase in the frequency of PfCRT:K76 is consistent with an uptake in the use of AL, as well as a decrease in CQ use. Increases in PfCRT:K76 have been seen upon cessation of CQ use in many countries including Malawi [155, 124], Kenya [118, 135], Tanzania [158, 141], Senegal [176], and Mozambique [253]. Given effective national policy switched from CQ+SP to AL in 2006 in Uganda, an increase in PfCRT:K76 prior to 2012 might have been expected. Mbogo *et al.* posit that the delay might reflect recent use of CQ [149].

We detect no notable *pfmrp1* haplotype frequency trends (figure 4.6), while Conrad *et al.* report a yearly decrease in the prevalence and frequency of PfMRP1:I876 in the AL arm [51], and Dahlström *et al.* report an increase in PfMRP1:I876 following treatment with AL [56]. It seems that more work is required to better understand the *pfmrp1* haplotype frequency trends. With the availability of more data, we might hope to elucidate more subtle trends.

We detect notable *pfmdr1* haplotype frequencies trends, including trends consistent with those based on single allele trends [51, 149]. More specifically, the haplotype encoding N**F**D increases year on year and is more prevalent in the AL arm, while the marker **Y**Y**Y** decreases year on year and is more prevalent in the DP arm (figure 4.11). With respect to prior treatment, trends consistent with AL selecting for N**F**D and against **Y**Y**Y**, and DP selecting for **Y**Y**Y** and against N**F**D are observed (figure 4.12).

In addition to the *pfmdr1* haplotype frequencies trends summarised above, trends that cannot

be extrapolated directly from single allele trends are observed. For example, the frequencies of haplotypes encoding **Y**YD, **YF**D and **YFY** decrease year on year (the former two in the AL arm only), while the frequency of that encoding NYD increases year on year (figure 4.11). Note that the increase in the NYD associated frequency seemingly explains why the previously reported trends in the prevalence and frequency of the single alleles at codon 184 are not as marked as those for alleles at codons 86 and 1246 [51]. We also see trends compatible with AL treatment selecting for NYD (although the trend is not significantly different from zero) and against **Y**YD, **YF**D and **YFY** (although the trend for the latter is not significantly different from zero) (figure 4.12).

In summary, DP treatment appears to select for **YYY** and against N**F**D, whereas almost all haplotype frequencies vary with immediacy since last treatment with AL, with selection acting against all those with PfMDR1:86**Y** (figure 4.12). The fact that numerous haplotypes are found to vary with immediacy since last treatment with AL, seemingly explains why we see yearly trends in haplotypes beyond those predicted from nSNP-wise allele trends (figure 4.11). It also seemingly explains why previously reported single allele trends for codon 86 are more marked than those for codons 1246 and 184 [51, 149]. The fact that trends are consistent with DP selection primarily upon **YYY** and N**F**D (figure 4.12), suggests that the sequence of specific alleles on the haploid genome might be critical for determining piperaquine sensitivity, perhaps explaining why some studies do not see trends at the level of the single allele [232].

In generating the aforementioned results, several simplifying assumptions are made. First, frequencies are estimated using a model that assumes data are derived from independent malaria episodes. This is thought to be true for almost all episodes due to a low estimated frequency of recrudescence [51]. Nevertheless, in assuming independence we ignore inter-child variability. An extension to accommodate inter-child variability is proposed in the following chapter. Based on results reported in chapter 5, we conclude that the assumption of independence will likely not affect the trends reported here.

To enable regression of *pfmdr1* haplotypes frequencies, errors are assumed to be independently and identically distributed according to a normal distribution, no attempt is made to model the dependence between the haplotypes and a two stage approximation is used to propagate the error from the frequency estimates to the regression. Residual plots suggest the independently and identically distributed normal assumption is satisfactory, while summation summaries suggest that overlooking dependence between haplotypes leads to only small deviations from summation to unity (results not shown).

To fully capture the uncertainty in the frequencies in the regression, one ought to fit a model that jointly estimates a matrix of frequencies and a matrix of regression coefficients. The joint model would require the specification of prior distributions over matrices: a matrix normal prior over the matrix of regression coefficients and an inverse Wishart prior over a covariance matrix capturing the dependence between the *pfmdr1* haplotypes. Given our experience fitting variations of the current model using Gibbs sampling (see upcoming chapters 5 and 6), fitting a model with matrix-valued random variables is likely to be computationally challenging. Moreover, it is uncertain whether the inferential gain will merit the additional complexity. The complexity is primarily due to the joint distribution over the *pfmdr1* haplotypes. It can be circumvented, in part, by fitting the regression to each haplotype separately using the two-stage approximate approach proposed by Lunn *et al.* (see [136]). The two stage approximation is not without its limitations, however. If the posterior on the haplotype frequencies post regression is far away in parameter space from the discrete proposal, the MCMC sample post regression will not represent the true posterior on the frequencies. It is easy to identify cases where the proposal and posterior may be misaligned (for example, figure 4.8), but not easy to discern whether of not the posterior is sufficiently supported. Hence we proceed with caution, noting that the least number of cases of misalignment are seen under Zellner's g prior, and that the results are consistent across the specified priors.

In summary, despite a number of model assumptions described above, the results reported

in this chapter support previous findings based on single alleles, namely that AL and DP exert inverse selection pressure [51]. The results reported here also build upon those based on single alleles, providing insight into haplotype trends. They suggest AL and DP drug pressures are unequal due to haplotypic effects, which has potentially important implications for any policy hoping to exploit the inverse pressure for averting widespread resistance.

Chapter 5

Accounting for inter-child variability

5.1 Background

In the previous chapter (chapter 4), the haplotype-frequency estimation model (equation (3.5)) is fit to marker prevalence data collected from a cohort study in Uganda (see section 4.2.1 for a detailed description of the data). Before applying the model, the data are heavily partitioned. More precisely, genotyping outcomes are partitioned by gene (panel A, figure 4.3), baseline samples are separated from longitudinal samples (panel B, figure 4.3), samples are divided by drug arm (panel C, figure 4.3), and then either by year (panel D, figure 4.3) or by days since last treatment (panel E, figure 4.3). With the exception of baseline data, the subdivisions contain repeat samples from children who suffered multiple episodes of malaria (table 5.1). However, when the model is fit to a given subdivision, the samples within it are treated as independent observations, discarding the fact that they are indexed by children (note that we henceforth refer to the treatment of samples as independent observations as the independence assumption). The frequencies are then plotted against correlates of time in order to examine trends (for example, figure 4.5). The aim of this chapter is to investigate the impact of the independence assumption on the frequencies, and therefore trends, via extension of the original haplotype-frequency estimation model (equation (3.5)).

Subdivision	Number of samples that feature in the genetic analysis per child													
	1	2	3	4	5	6	7	8	9	10	11	12	13	14
2008 DP	28	31	12	9	4	3	0	0	0	0	0	0	0	0
2009 DP	42	30	17	5	4	1	0	0	0	0	0	0	0	0
2010 DP	48	23	23	4	3	1	0	0	0	0	0	0	0	0
2011 DP	48	33	15	8	2	0	0	0	0	0	0	0	0	0
2012 DP	21	25	9	5	2	0	0	0	0	0	1	0	0	0
2008 AL	27	22	11	10	6	2	2	0	0	0	0	0	0	0
2009 AL	39	26	15	7	5	2	0	0	0	0	0	0	0	0
2010 AL	36	28	15	10	2	1	1	0	0	0	0	0	0	0
2011 AL	31	28	17	6	5	2	0	0	0	0	0	0	0	0
2012 AL	21	15	9	8	3	3	1	0	0	0	0	0	0	0
4–28 days DP	17	3	0	0	0	0	0	0	0	1	0	0	0	0
29–42 days DP	33	13	17	8	2	1	2	0	1	0	0	0	0	0
43–56 days DP	26	22	14	13	11	8	5	1	0	0	0	0	0	0
57–70 days DP	38	26	10	5	2	1	1	0	0	0	0	0	0	0
$>$ 70 days DP	45	37	23	8	4	2	0	0	0	0	0	0	0	0
4–28 days AL	31	19	13	5	4	4	6	2	3	0	0	0	0	2
29–42 days AL	35	23	14	7	9	4	0	2	0	0	0	0	0	0
43–56 days AL	44	18	15	5	1	0	0	0	0	0	0	0	0	0
57–70 days AL	42	12	3	1	0	0	0	0	0	0	0	0	0	0
$>$ 70 days AL	46	36	17	5	1	1	0	0	0	0	0	0	0	0

Table 5.1: The table entries list the number of children, partitioned by the numbers of samples that feature in the genetic analysis per child (columns) and subdivision (rows). For example, 28 children each provided one sample per child in the DP 2008 subdivision (top, leftmost entry), while 31 children each provided 2 samples per child in the DP 2008 subdivision (top, second-left entry). Subdivisions are separated by drug arm (DP and AL), year or by days since last treatment category.

Before embarking upon the model extension, a preliminary study (section B.1) was undertaken to investigate evidence of inter-child variation. In the preliminary study, inter-child variation was defined in terms of the propensity to present with parasites with different levels of 'mutatedness' at the level of individual nSNPs. By fitting a model with a random effect for each child, an estimate of the inter-child variability was obtained. Statistically significant inter-child variability was detected for two of the six nSNPs investigated: *pfmdr-86* and *pfmrp-876*. In both instances, the estimate of inter-child variance was small (0.12 and 0.08, respectively) and statistical significance modest (p-values 0.02 and 0.05, respectively). Nevertheless, the detection of inter-child variance reinforced the call for an investigation into the impact of the independence assumption on the frequency estimates.

From a practical viewpoint, one of major advantages of a Bayesian framework is the natural manner in which hierarchical structure can be built into it [88]. The obvious extension to account for repeat samples from multiple children, therefore, is to add a level of hierarchy to the original haplotype-frequency estimation model (equation (3.5)) to accommodate structure at the level of the child.

The outline of this chapter is as follows. In the following section, we present details of the extended model (section 5.2) and its accompanying sampler (section 5.3), followed by a study based on simulated data (section 5.4), before focusing on the Ugandan data (section 5.5.2). The chapter concludes with a discussion (section 4.4).

5.2 The extended model

The framework of the extended model (figure 5.1) is based on the original model (figure 3.1), but with an additional level of hierarchy (see box encapsulating $\boldsymbol{\pi}_{\text{child}}$, figure 5.1). The notation of the extended model follows that of the original model (see section 3.2.1 and table 3.1), but with the addition of the subscript 'child', which particularises child specific variables. For example, $y_{ij\,\text{child}}$ represents the genotyping outcome at the jth SNP for the

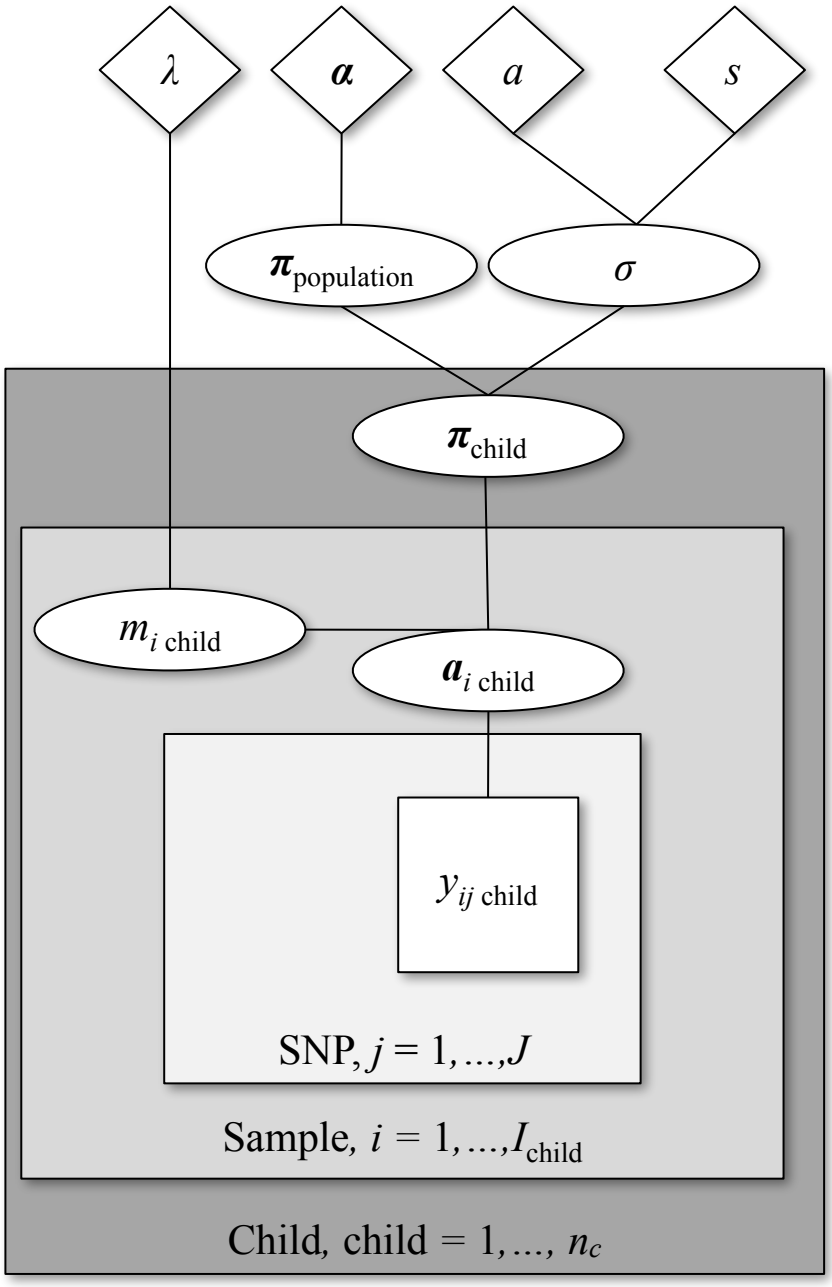


Figure 5.1: Haplotype frequency estimation model with inter-child variation. The graph shows the dependence structure of a haplotype frequency estimation model that accommodates structure at the level of the child. Diamonds depict fixed parameters, ellipses depict inferred parameters and the square depicts the data.

ith malaria episode experienced by the specified child; $m_{i\,\text{child}}$ represents the MOI of the ith episode experienced by the specified child; while $\boldsymbol{a}_{i\,\text{child}}$ represents the haplotype count vector of the clones that constitute the ith episode experienced by the specified child. The variable $\boldsymbol{\pi}_{\text{child}} = (\pi_{\text{child}\,1}, \ldots, \pi_{\text{child}\,R})$, where $\boldsymbol{\pi}_{\text{child}} \in \mathbb{S}^R$, is unique to the extended model. It denotes a child specific frequency vector. Under the extended model, child specific frequency vectors are conditionally independent given the population-level frequency vector $\boldsymbol{\pi}_{\text{population}}$ (which is equivalent to $\boldsymbol{\pi}$ under the original model), and another new parameter $\sigma \in (0, \infty)$, which captures the relatedness between $\boldsymbol{\pi}_{\text{child}}$ and $\boldsymbol{\pi}_{\text{population}}$. More precisely, when σ is large, $\boldsymbol{\pi}_{\text{child}}$ for child $= 1, \ldots, n_c$ (where n_c denotes the number of children) are closely related to $\boldsymbol{\pi}_{\text{population}}$. When σ is small, $\boldsymbol{\pi}_{\text{child}}$ for child $= 1, \ldots, n_c$, are highly dispersed and non-uniform, but with expectation equal to $\boldsymbol{\pi}_{\text{population}}$. The posterior density under the extended model is given by

$$\begin{aligned}\rho\left(\boldsymbol{\pi}_{\text{population}}, \boldsymbol{\pi}_{\text{children}}, \boldsymbol{a}, \boldsymbol{m} | \boldsymbol{y}\right) \propto \prod_{\text{child}=1}^{n_c} \Bigg\{ \prod_{i=1}^{I_{\text{child}}} \Bigg\{ \prod_{j=1}^{J} \left\{\rho\left(y_{ij\,\text{child}} | \boldsymbol{a}_{i\,\text{child}}\right)\right\} \\ \times \rho\left(\boldsymbol{a}_{i\,\text{child}} | m_{i\,\text{child}}, \boldsymbol{\pi}_{\text{child}}\right) \rho\left(m_{i\,\text{child}}\right) \Bigg\} \\ \times \rho\left(\boldsymbol{\pi}_{\text{child}} \mid \sigma, \boldsymbol{\pi}_{\text{population}}\right) \Bigg\} \\ \times \rho\left(\boldsymbol{\pi}_{\text{population}}\right) \rho\left(\sigma\right), \qquad (5.1)\end{aligned}$$

where,

- $\boldsymbol{\pi}_{\text{children}}$ denotes the collection of $\boldsymbol{\pi}_{\text{child}}$ for child $= 1, \ldots, n_c$;
- $\boldsymbol{a}$ and $\boldsymbol{m}$ denote the collections of $\boldsymbol{a}_{i\,\text{child}}$ and $m_{i\,\text{child}}$, respectively, for child $= 1, \ldots, n_c$ and $i = 1, \ldots, I_{\text{child}}$;
- I_{child} denotes the number of malaria episodes for a specified child;
- The densities $\rho\left(y_{ij\,\text{child}} \mid \boldsymbol{a}_{i\,\text{child}}\right)$, $\rho\left(\boldsymbol{a}_{i\,\text{child}} \mid m_{i\,\text{child}}, \boldsymbol{\pi}_{\text{child}}\right)$, $\rho\left(m_{i\,\text{child}}\right)$ and $\rho\left(\boldsymbol{\pi}_{\text{population}}\right)$ are equivalent to those defined previously for $\rho\left(\boldsymbol{y}_{ij} \mid \boldsymbol{a}_i\right)$, $\rho\left(\boldsymbol{a}_i \mid m_i, \boldsymbol{\pi}\right)$, $\rho\left(m_i\right)$ and $\rho\left(\boldsymbol{\pi}\right)$ (equations (3.6), (3.8), (3.10) to (3.13), and (3.9), respectively).

In addition,

$$\rho\left(\boldsymbol{\pi}_{\text{child}} \mid \sigma, \boldsymbol{\pi}_{\text{population}}\right) = \mathscr{D}\text{irichlet}(\boldsymbol{\pi}_{\text{child}} \mid \sigma \times \boldsymbol{\pi}_{\text{population}}), \tag{5.2}$$

$$\rho\left(\sigma\right) = \mathscr{G}\text{amma}(\sigma \mid a, s), \tag{5.3}$$

where a is the shape and s is the scale parameter of a gamma prior parameterised as follows,

$$\rho(\sigma) = \frac{1}{\Gamma(a)\, s^a} \cdot \frac{\sigma^{a-1}}{\exp(\sigma/s)} \text{ for } \sigma > 0, \text{ and } a, s > 0, \tag{5.4}$$

where Γ denotes the gamma function. Since preliminary evidence of a inter-child variation was found for some nSNPs but not all (section B.1), values of a and s (2 and 100, respectively), are chosen such that the prior on σ extends over a wide range, such that $\mathbb{P}(\sigma < 600) = 0.98$.

5.3 The sampler

The proposed model (equation (5.1), figure 5.1) is implemented iteratively using a Gibbs sampler. At iteration $t = 0$, initial parameter values, $\boldsymbol{\pi}^{(t)}_{\text{population}}$, $\boldsymbol{\pi}^{(t)}_{\text{children}}$, $\boldsymbol{a}^{(t)}$, $\boldsymbol{m}^{(t)}$, are sampled from their respective priors. Thereafter, within each iteration of the sampler, the parameters are successively updated in blocks by sampling from their full conditional distributions with density $\rho(\text{param}_{\text{update}} \mid \text{param}_{\text{other}}, \boldsymbol{y})$, where $\text{param}_{\text{update}}$ denotes the parameters in the block to update and $\text{param}_{\text{other}}$ denotes all the parameters outside the block at their current values. For a given $t \geq 1$, the sampler proceeds as follows.

Update genotype counts and MOIs

For malaria episode $i = 1, \ldots, I_{\text{child}}$ and child $= 1, \ldots, n_c$, a Metropolis-Hastings update is used to sample from the full conditional distribution on both $\boldsymbol{a}_{i\,\text{child}}$ and $m_{i\,\text{child}}$ with density,

$$\rho(\boldsymbol{a}_{i\,\text{child}}, m_{i\,\text{child}} \mid \text{param}_{\text{other}}, \boldsymbol{y}) \propto \prod_{j=1}^{J} \left\{\rho(y_{ij\,\text{child}} \mid \boldsymbol{a}_{i\,\text{child}})\right\} \rho\left(\boldsymbol{a}_{i\,\text{child}} \mid m_{i\,\text{child}}, \boldsymbol{\pi}_{\text{child}}^{(t-1)}\right) \rho(m_{i\,\text{child}}).$$

The Metropolis-Hastings update is equivalent to that documented in section 3.2.4, but with $\boldsymbol{a}_{i\,\text{child}}$ replacing $\boldsymbol{a}_i$, $m_{i\,\text{child}}$ replacing m_i and $\boldsymbol{\pi}_{\text{child}}^{(t-1)}$ replacing $\boldsymbol{\pi}^{(t-1)}$.

Update child-specific frequencies

For child $= 1, \ldots, n_c$, $\boldsymbol{\pi}_{\text{child}}$ is updated by sampling directly from its full conditional distribution with density,

$$\begin{aligned}
\rho(\boldsymbol{\pi}_{\text{child}} \mid \text{param}_{\text{other}}, \boldsymbol{y}) &\propto \prod_{i=1}^{I_{\text{child}}} \left\{\rho\left(\boldsymbol{a}_{i\,\text{child}}^{(t)} \mid m_{i\,\text{child}}^{(t)}, \boldsymbol{\pi}_{\text{child}}\right)\right\} \rho(\boldsymbol{\pi}_{\text{child}} \mid \sigma^{(t-1)}, \boldsymbol{\pi}_{\text{population}}^{(t-1)}), \\
&= \prod_{i=1}^{I_{\text{child}}} \left\{\mathscr{M}\text{ultinomial}\left(\boldsymbol{a}_{i\,\text{child}}^{(t)} \mid m_{i\,\text{child}}^{(t)}, \boldsymbol{\pi}_{\text{child}}\right)\right\} \mathscr{D}\text{irichlet}(\boldsymbol{\pi}_{\text{child}} \mid \sigma^{(t-1)} \times \boldsymbol{\pi}_{\text{population}}^{(t-1)}), \\
&= \mathscr{D}\text{irichlet}\left(\sigma^{(t-1)}\left\{\pi_{\text{population}}{}_1^{(t-1)} + \sum_{i=1}^{I_{\text{child}}} a_{i\,\text{child}}{}_1^{(t)}, \ldots, \pi_{\text{population}}{}_R^{(t-1)} + \sum_{i=1}^{I_{\text{child}}} a_{i\,\text{child}}{}_R^{(t)}\right\}\right).
\end{aligned}$$

Update population-level frequencies

A Metropolis-Hastings update is used to sample from the full conditional distribution on $\boldsymbol{\pi}_{\text{population}}$ with density,

$$\begin{aligned}
\rho(\boldsymbol{\pi}_{\text{population}} \mid \text{param}_{\text{other}}, \boldsymbol{y}) &\propto \prod_{\text{child}=1}^{n_c} \left\{\rho\left(\boldsymbol{\pi}_{\text{child}}^{(t)} \mid \sigma^{(t-1)}, \boldsymbol{\pi}_{\text{population}}\right)\right\} \rho(\boldsymbol{\pi}_{\text{population}}), \\
&= \prod_{\text{child}=1}^{n_c} \left\{\mathscr{D}\text{irichlet}(\boldsymbol{\pi}_{\text{child}}^{(t)} \mid \sigma^{(t-1)} \times \boldsymbol{\pi}_{\text{population}})\right\} \mathscr{D}\text{irichlet}(\boldsymbol{\pi}_{\text{population}} \mid \boldsymbol{\alpha}).
\end{aligned}$$

First, a proposed update, $\boldsymbol{\pi}_{\text{population}}^{(\star)}$, is sampled from a proposal distribution. Two alternatives are considered: a proposal distribution on the simplex (equation (5.5)) and a proposal

distribution on the real line (equation (5.6)),

$$\boldsymbol{\pi}^{(\star)}_{\text{population}} \sim \mathscr{D}\text{irichlet}\left(\cdot \mid \xi \times \boldsymbol{\pi}^{(t-1)}_{\text{population}}\right), \tag{5.5}$$

$$\boldsymbol{\theta}^{(\star)}_{\text{population}} \sim \mathscr{N}\text{ormal}_{R-1}\left(\cdot \mid \boldsymbol{\theta}^{(t-1)}_{\text{population}}, \xi \times \boldsymbol{\Sigma}\right), \tag{5.6}$$

where ξ is a tuning parameter set such that the acceptance rate of $\boldsymbol{\pi}^{(\star)}_{\text{population}}$ is between 0.05 and 0.30; $\boldsymbol{\theta}^{(\star)}_{\text{population}}$ is equal to the first $R-1$ elements of $\boldsymbol{\pi}^{(\star)}_{\text{population}}$ mapped onto the real line,

$$\begin{aligned}
\boldsymbol{\theta}^{(\star)}_{\text{population}} &= \boldsymbol{f}\left(\boldsymbol{\pi}^{(\star)}_{\text{population}\,1:R-1}\right), \\
&= \left(f_1\left(\boldsymbol{\pi}^{(\star)}_{\text{population}\,1:R-1}\right), \ldots, f_{R-1}\left(\boldsymbol{\pi}^{(\star)}_{\text{population}\,1:R-1}\right)\right), \text{ where} \\
f_r(\boldsymbol{\pi}^{(\star)}_{\text{population}\,1:R-1}) &= \log\left(\frac{\pi^{(\star)}_{\text{population}_r}}{1-\sum_{i=1}^{R-1}\pi^{(\star)}_{\text{population}_i}}\right), \text{ for } r = 1,\ldots R-1;
\end{aligned} \tag{5.7}$$

and $\boldsymbol{\Sigma}$ is a covariance matrix computed using 10,000 draws from the prior on $\boldsymbol{\pi}_{\text{population}}$ (equivalent of equation (3.9)), mapped onto the real line using equation (5.7). When the proposal on the real line is used (equation (5.6)), $\boldsymbol{\theta}^{(\star)}_{\text{population}}$ is subsequently mapped back to $\boldsymbol{\pi}^{(\star)}_{\text{population}}$ following,

$$\begin{aligned}
\boldsymbol{\pi}^{(\star)}_{\text{population}} &= \left(\boldsymbol{f}^{-1}(\boldsymbol{\theta}^{(\star)}_{\text{population}}), 1-\sum_{r=1}^{R-1} f_r^{-1}\left(\boldsymbol{\theta}^{(\star)}_{\text{population}}\right)\right), \text{ where} \\
\boldsymbol{f}^{-1}(\boldsymbol{\theta}^{(\star)}_{\text{population}}) &= \left(f_1^{-1}(\boldsymbol{\theta}^{(\star)}_{\text{population}}), \ldots, f_{R-1}^{-1}(\boldsymbol{\theta}^{(\star)}_{\text{population}})\right), \text{ and} \\
f_r^{-1}(\boldsymbol{\theta}^{(\star)}_{\text{population}}) &= \frac{\exp\left(\theta^{(\star)}_{\text{population}_r}\right)}{1+\sum_{i=1}^{R-1}\exp\left(\theta^{(\star)}_{\text{population}_i}\right)} \text{ for } r = 1,\ldots,R-1.
\end{aligned} \tag{5.8}$$

Having generated $\boldsymbol{\pi}^{(\star)}_{\text{population}}$, the proposal is either rejected ($\boldsymbol{\pi}^{(t)}_{\text{population}} \leftarrow \boldsymbol{\pi}^{(t-1)}_{\text{population}}$) or accepted with probability,

$$\mathbb{P}(\boldsymbol{\pi}^{(t)}_{\text{population}} \leftarrow \boldsymbol{\pi}^{(\star)}_{\text{population}}) = \min\left\{1, \frac{\mathscr{T}\text{arget}\left(\boldsymbol{\pi}^{(\star)}_{\text{population}}\right)}{\mathscr{T}\text{arget}\left(\boldsymbol{\pi}^{(t-1)}_{\text{population}}\right)} \cdot \frac{\mathscr{P}\text{roposal}\left(\boldsymbol{\pi}^{(t-1)}_{\text{population}}\right)}{\mathscr{P}\text{roposal}\left(\boldsymbol{\pi}^{(\star)}_{\text{population}}\right)}\right\},$$

where

$$\frac{\mathscr{T}\text{arget}\left(\boldsymbol{\pi}^{(\star)}_{\text{population}}\right)}{\mathscr{T}\text{arget}\left(\boldsymbol{\pi}^{(t-1)}_{\text{population}}\right)} = \frac{\prod_{\text{child}=1}^{n_c}\left\{\mathscr{D}\text{irichlet}\left(\boldsymbol{\pi}^{(t)}_{\text{child}} \mid \sigma^{(t-1)} \times \boldsymbol{\pi}^{(\star)}_{\text{population}}\right)\right\}\mathscr{D}\text{irichlet}\left(\boldsymbol{\pi}^{(\star)}_{\text{population}} \mid \boldsymbol{\alpha}\right)}{\prod_{\text{child}=1}^{n_c}\left\{\mathscr{D}\text{irichlet}\left(\boldsymbol{\pi}^{(t)}_{\text{child}} \mid \sigma^{(t-1)} \times \boldsymbol{\pi}^{(t-1)}_{\text{population}}\right)\right\}\mathscr{D}\text{irichlet}\left(\boldsymbol{\pi}^{(t-1)}_{\text{population}} \mid \boldsymbol{\alpha}\right)},$$

and

$$\frac{\mathscr{P}\text{roposal}\left(\boldsymbol{\pi}^{(t-1)}_{\text{population}}\right)}{\mathscr{P}\text{roposal}\left(\boldsymbol{\pi}^{(\star)}_{\text{population}}\right)} = \begin{cases} \dfrac{\mathscr{D}\text{irichlet}\left(\boldsymbol{\pi}^{(t-1)}_{\text{population}} \mid \xi \times \boldsymbol{\pi}^{(\star)}_{\text{population}}\right)}{\mathscr{D}\text{irichlet}\left(\boldsymbol{\pi}^{(\star)}_{\text{population}} \mid \xi \times \boldsymbol{\pi}^{(t-1)}_{\text{population}}\right)} & \text{if the proposal is on the simplex,} \\ \dfrac{\mathscr{N}\text{ormal}_{R-1}\left(\boldsymbol{\theta}^{(t-1)}_{\text{population}} \mid \boldsymbol{\theta}^{(\star)}_{\text{population}}, \xi \times \boldsymbol{\Sigma}\right)}{\mathscr{N}\text{ormal}_{R-1}\left(\boldsymbol{\theta}^{(\star)}_{\text{population}} \mid \boldsymbol{\theta}^{(t-1)}_{\text{population}}, \xi \times \boldsymbol{\Sigma}\right)} \dfrac{\mid \det\left(\mathbf{J}\left(\boldsymbol{\theta}^{(\star)}_{\text{population}}\right)\right) \mid}{\mid \det\left(\mathbf{J}\left(\boldsymbol{\theta}^{(t-1)}_{\text{population}}\right)\right) \mid} & \text{otherwise.} \end{cases}$$

Here $\mathbf{J}(\cdot)$ denotes the Jacobian:

$$\mathrm{J}(\boldsymbol{\theta}) = \begin{bmatrix} \dfrac{\partial\, \pi_1}{\partial\, \theta_1} & \cdots & \dfrac{\partial\, \pi_1}{\partial\, \theta_{R-1}} \\ \vdots & \ddots & \vdots \\ \dfrac{\partial\, \pi_{R-1}}{\partial\, \theta_1} & \cdots & \dfrac{\partial\, \pi_{R-1}}{\partial\, \theta_{R-1}} \end{bmatrix}.$$

Update the relatedness parameter

Finally, the following Metropolis-Hastings update is used to sample from the full conditional distribution on σ, with density

$$\rho(\sigma \mid \text{param}_{\text{other}}, \boldsymbol{y}) \propto \prod_{\text{child}=1}^{n_c}\left\{\rho\left(\boldsymbol{\pi}^{(t)}_{\text{child}} \mid \sigma, \boldsymbol{\pi}^{(t)}_{\text{population}}\right)\right\}\rho(\sigma), \tag{5.9}$$

$$= \prod_{\text{child}=1}^{n_c}\left\{\mathscr{D}\text{irichlet}\left(\boldsymbol{\pi}^{(t)}_{\text{child}} \mid \sigma \times \boldsymbol{\pi}^{(t)}_{\text{population}}\right)\right\}\mathscr{G}\text{amma}(\sigma \mid a, s). \tag{5.10}$$

First, $\varepsilon^{(\star)}$ is sampled from a univariate normal distribution with mean equal to $\varepsilon^{(t-1)} = \log(\sigma^{(t-1)})$ and variance equal to one (chosen such that the acceptance rate of $\sigma^{(\star)} = \exp(\varepsilon^{(\star)})$ is approximately 0.15 discounting burnin.), then mapped onto $\mathbb{R}^+$: $\exp(\varepsilon^{(\star)}) \to \sigma^{(\star)}$. Second, $\sigma^{(\star)}$ is either rejected ($\sigma^{(t)} \leftarrow \sigma^{(t-1)}$) or accepted with probability,

$$\mathbb{P}\left(\sigma^{(t)} \leftarrow \sigma^{(\star)}\right) = \min\left\{1, \frac{\mathscr{T}\text{arget}\left(\sigma^{(\star)}\right)}{\mathscr{T}\text{arget}\left(\sigma^{(t-1)}\right)}\frac{\mathscr{P}\text{roposal}\left(\sigma^{(t-1)}\right)}{\mathscr{P}\text{roposal}\left(\sigma^{(\star)}\right)}\right\},$$

where

$$\frac{\mathscr{T}\text{arget}\left(\sigma^{(\star)}\right)}{\mathscr{T}\text{arget}\left(\sigma^{(t-1)}\right)} = \frac{\prod_{\text{child}=1}^{n_c}\left\{\mathscr{D}\text{irichlet}\left(\boldsymbol{\pi}_{\text{child}}^{(t)} \mid \sigma^{(\star)} \times \boldsymbol{\pi}_{\text{population}}^{(t)}\right)\right\}\mathscr{G}\text{amma}\left(\sigma^{(\star)} \mid a,s\right)}{\prod_{\text{child}=1}^{n_c}\left\{\mathscr{D}\text{irichlet}\left(\boldsymbol{\pi}_{\text{child}}^{(t)} \mid \sigma^{(t-1)} \times \boldsymbol{\pi}_{\text{population}}^{(t)}\right)\right\}\mathscr{G}\text{amma}\left(\sigma^{(t-1)} \mid a,s\right)},$$

and

$$\frac{\mathscr{P}\text{roposal}\left(\sigma^{(t-1)}\right)}{\mathscr{P}\text{roposal}\left(\sigma^{(\star)}\right)} = \frac{\mathscr{N}\text{ormal}\left(\varepsilon^{(t-1)} \mid \varepsilon^{(\star)},1\right)}{\mathscr{N}\text{ormal}\left(\varepsilon^{(\star)} \mid \varepsilon^{(t-1)},1\right)} \cdot \frac{\left|\det\left(\mathrm{d}\varepsilon^{(t-1)}/\mathrm{d}\sigma^{(t-1)}\right)\right|}{\left|\det\left(\mathrm{d}\varepsilon^{(\star)}/\mathrm{d}\sigma^{(\star)}\right)\right|} = \frac{\sigma^{(\star)}}{\sigma^{(t-1)}}.$$

Performance indicators

To assess convergence and general performance of the sampler, multiple chains of the sampler are run per analysis and numerous diagnostic plots generated for visual inspection. Diagnostic plots included

- chain-wise trace, density, and autocorrelation plots of $\pi_{\text{population}_r}$ for $r = 1,\ldots R$, and σ;
- chain-wise log-likelihood and log-posterior trace plots;
- chain-wise trace plots of the acceptance rates for all Metropolis-Hastings updates;
- chain-wise trace and density plots of a randomly sampled $\boldsymbol{\pi}_{\text{child}}$;
- child-wise trace and density plots of 16 randomly sampled $\boldsymbol{\pi}_{\text{child}}$;
- trace plots and plots of posterior mass estimates of the discrete latent variables, $\boldsymbol{a}$ and $\boldsymbol{m}$.

In addition, the respective potential scale reduction factors (PSRFs) of $\pi_{\text{population}_r}$ for $r = 1,\ldots R$, and σ are calculated to further assess convergence [88].

5.4 Simulated data study

To assess the performance of the extended model (equation (5.1)) and its sampler (section 5.3), four datasets are simulated (figures 5.2a, 5.2b, 5.2c and 5.2d) and analysed as follows.

5.4.1 Methods

Data generation

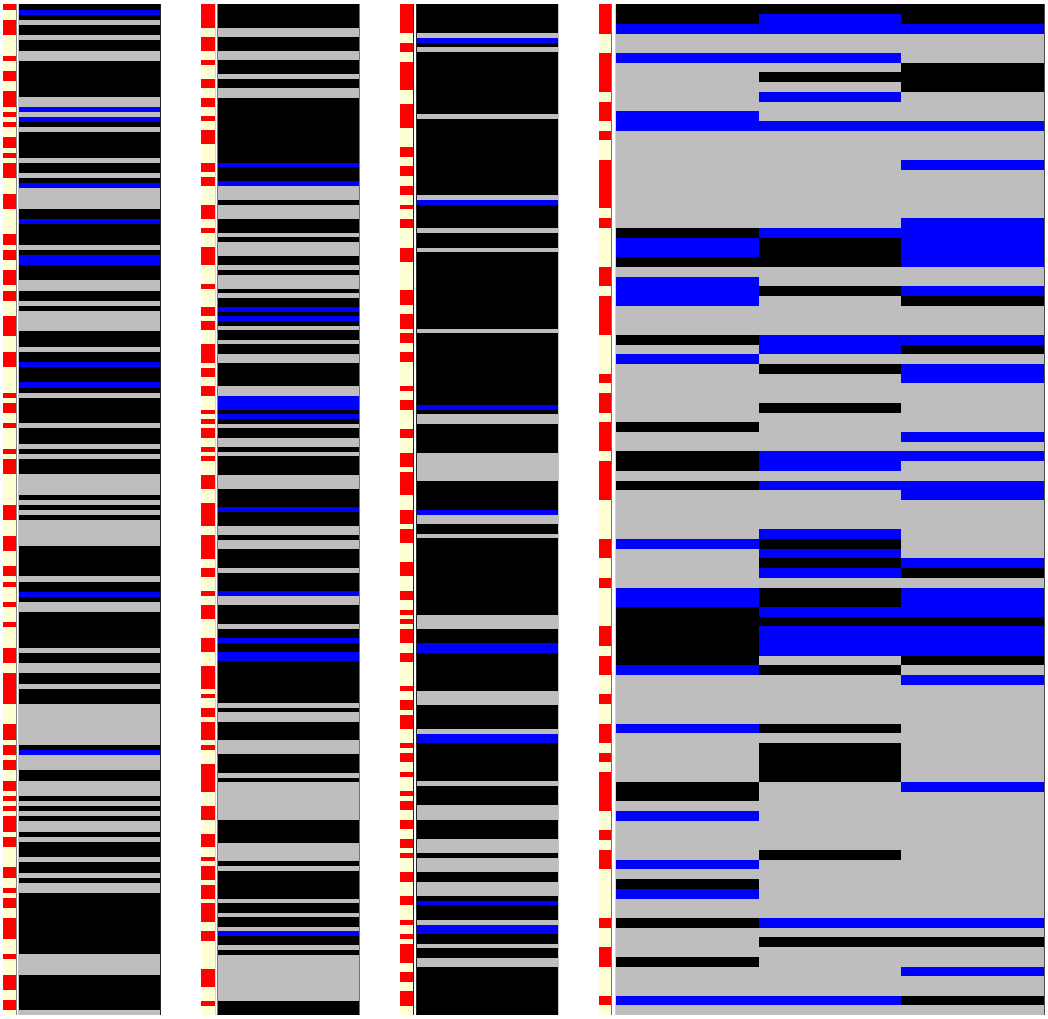
Four simulated datasets are simulated, including three individual-SNP datasets with $\sigma = 200$, 20 and 2 (figures 5.2a, 5.2b and 5.2c, respectively), and one triple-SNP dataset with $\sigma = 681$ (5.2d). Each individual-SNP dataset has $n_c = 90$ children, whereas the triple-SNP dataset has $n_c = 50$ children. The number of episodes, and therefore samples, per child is drawn from a non-zero conditioned Poisson distribution with parameter equal to two, leading to datasets with variable numbers of samples in total (198, 217, 212 and 104 for datasets depicted in figures 5.2a, 5.2b, 5.2c and 5.2d, respectively). Given the specified values of σ and numbers of episodes per child, the genotyping outcomes are generated under the extended model (equation (5.1)), with a uniform Dirichlet prior on the population vector of haplotype frequencies and a non-zero conditioned Poisson prior with parameter equal to three on each MOI.

Data analysis

The model (equation (5.1)) is fit to the simulated data by running the sampler for T iterations (the first $t = 1, \ldots, {}^{T}/_{2}$ are discarded as burnin), with hyperparameters assigned as follows: $a = 2$, $s = 100$, $\boldsymbol{\alpha} = (1_1, \ldots, 1_R)$, $\lambda = 3$ and $m_{\max} = 8$. To assess convergence, multiple chains are run per analysis. Results generated using the $\boldsymbol{\pi}^{(\star)}_{\text{population}}$ proposal on the simplex (equation (5.5)) and the $\boldsymbol{\pi}^{(\star)}_{\text{population}}$ proposal on real line (equation (5.6)) are compared.

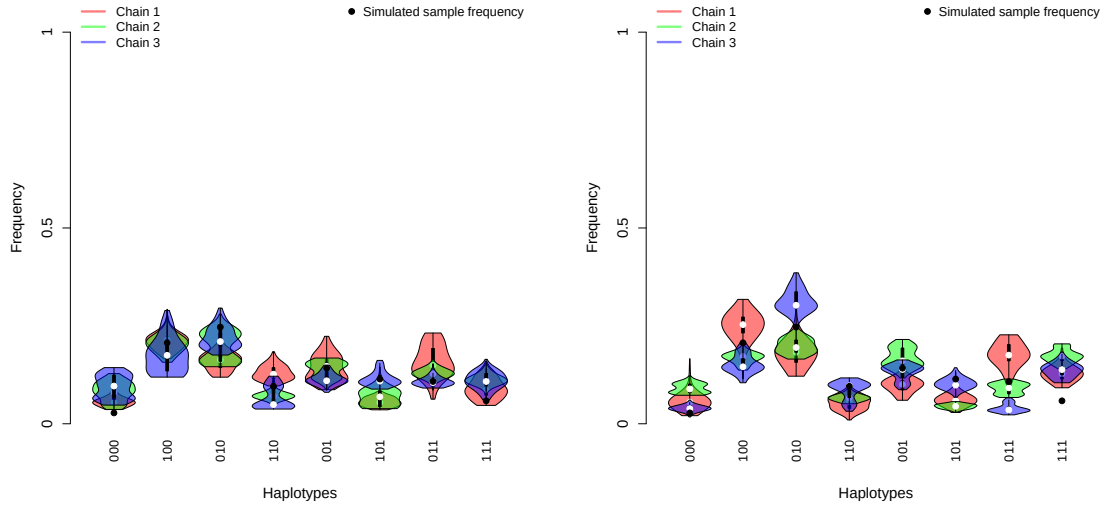
5.4.2 Results

In summary, computation is efficient for individual-SNP data but prohibitively slow for triple-SNP data. The results are described in more detail below, focusing first on the analysis of the triple-SNP dataset.



(a) $\sigma = 200$ (b) $\sigma = 20$ (c) $\sigma = 2$ (d) $\sigma = 681$

Figure 5.2: A graphical representation of data simulated under the model with inter-child variation. The figure shows four simulated datasets (subplots 5.2a, 5.2b 5.2c and 5.2d) with repeat samples from multiple children. For a given subplot, each row of equal height represents a sample, where samples are grouped by children. The child indexes are depicted in red (odd indices) and off-white (even indices) on the left of the main rectangle. A contiguous block of red or white spanning multiple rows depicts multiple samples from the same child. The prevalence data are depicted in blue (detection of wild type markers only), grey (detection of mutant type markers only) and black (detection of both wild and mutant type markers). There are no missing genotyping outcomes. Figures 5.2a, 5.2b and 5.2c, each represent individual-SNP datasets with 90 children and variable numbers of samples (198, 217, 212, respectively.) Figure 5.2d represents a triple-SNP dataset with 50 children and 104 samples in total.



(a) $\boldsymbol{\pi}_{\text{population}}$ proposal on the real line

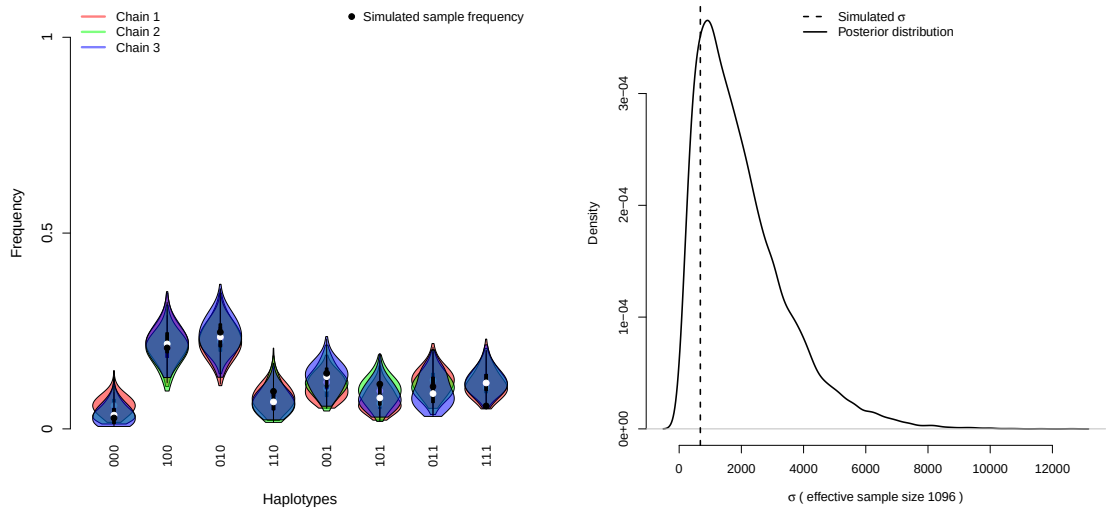
(b) $\boldsymbol{\pi}_{\text{population}}$ proposal on the simplex

Figure 5.3: Frequency estimates generated using the proposal on the simplex versus that on the real line. Plots show chain-wise haplotype frequency posterior density estimates, colour-coded by chain. Black dots mark the haplotype frequencies in the simulated sample.

Triple-SNP data

The Gibbs sampler is extremely slow to converge. After 100,000 iterations, the sampler with the $\boldsymbol{\pi}_{\text{population}}$ proposal on the real line (equation (5.6)) appears to perform marginally better than that with the proposal on the simplex (equation (5.5)) (compare figure 5.3a and figure 5.3b, respectively), hence use of the proposal on the simplex is discontinued. After 1,000,000 iterations the results for $\boldsymbol{\pi}_{\text{population}}$ and σ are accurate (figures 5.4a and 5.4b, respectively), but the chains have not formally converged (see, for example, the trace of haplotype 000, which has the highest PSRF (1.26), figure 5.5).

As a quality control check, the model is fit to a set of entirely missing data. As anticipated, the model returns the priors for both $m_{i\text{child}}$ (figure 5.6a) and σ (figure 5.6b), and looks to be venturing towards the prior on $\boldsymbol{\pi}_{\text{population}}$ (figure 5.6c) despite not having converged. Thus, the algorithm appears to sample from the correct posterior, but at a prohibitively slow rate, calling for a more efficient sampler for the analysis of triple-SNP data (see section 4.4).



(a) Chain-wise posterior density estimates of the marginal haplotype frequencies, colour-coded by chain. Black dots mark the haplotype frequencies in the simulated sample.

(b) Posterior density estimate of the relatedness parameter (σ) based on all chains combined. The vertical dotted line marks the location of the relatedness parameter used to simulate the data.

Figure 5.4: Posterior density estimates based on triple-SNP simulated data. The density estimates are generated using the sampler with the $\boldsymbol{\pi}_{\text{population}}$ proposal on the real line run for 1,000,000 iterations.

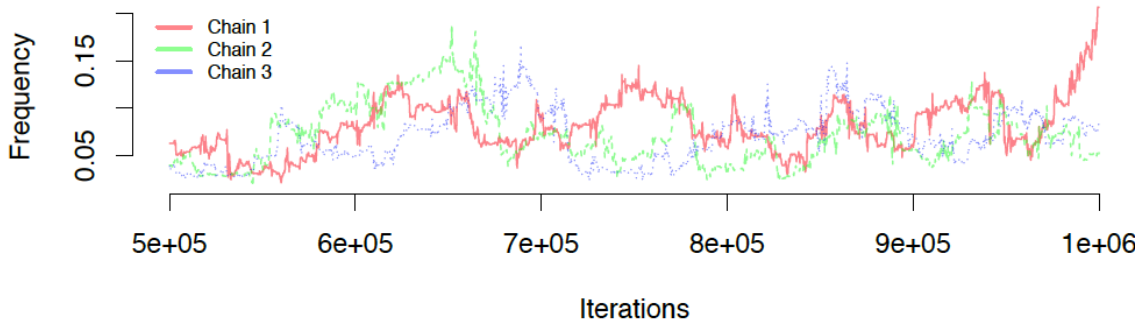
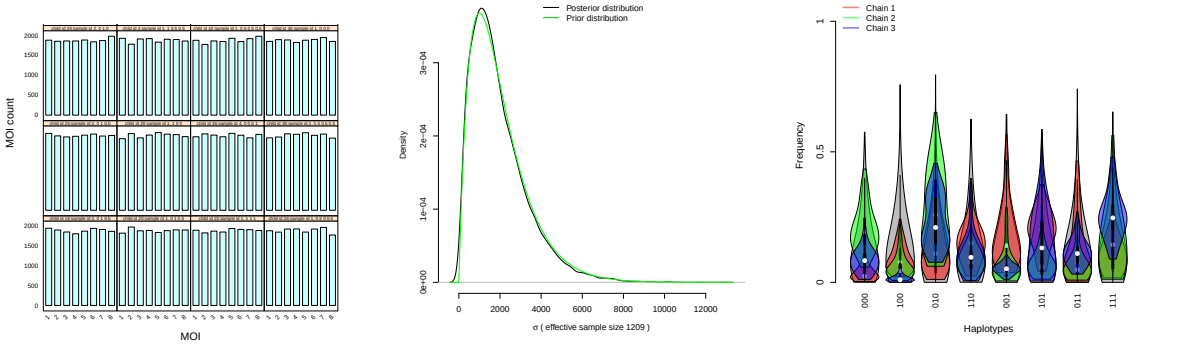


Figure 5.5: Chain-wise frequency trace plot for the haplotype 000, which has the highest PCRF (1.26). Traces are colour-coded by chain.



(a) Posterior MOI density estimates for 12 randomly selected samples from the dataset. The density is based on all chains combined.

(b) Posterior (black) and prior (green) density estimates of the relatedness parameter, σ. The density is based on all the chains combined.

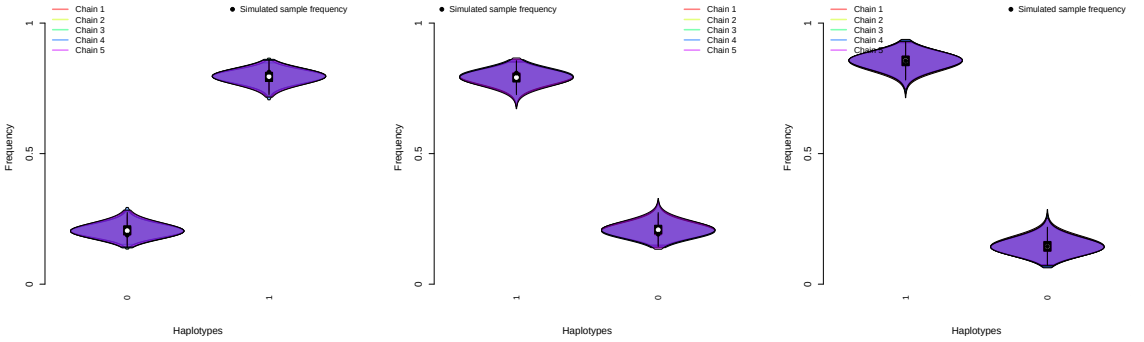
(c) Chain-wise haplotype frequency posterior density estimates, colour-coded by chain. Marginal density estimates of the uniform Dirichlet prior are plotted in grey.

Figure 5.6: Quality control check: posterior density estimates when the model is fit to a set of entirely missing data. The plots show posterior density estimates obtained when the model is fit to triple-SNP data that have been erased (all the data are replaced by missing genotyping outcomes). The model is fit using the sampler with the $\boldsymbol{\pi}_{\text{population}}$ proposal on the real line run for 1,000,000 iterations.

Individual-SNP data

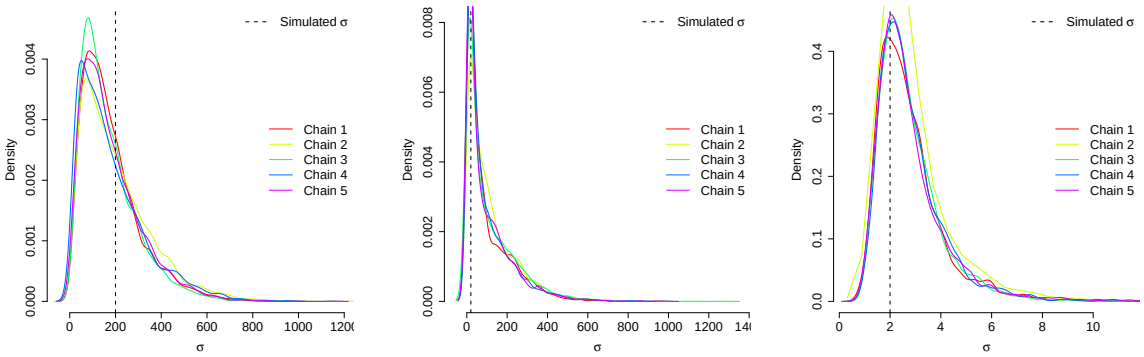
Based on the $\pi_{\text{population}_r}$ trace plots and PSRFs, all chains converge after 100,000 iterations when run on individual-SNP data (log-posterior and $\pi_{\text{population}_r}$ trace plots are stable and mixing well, and $\pi_{\text{population}_r}$ PSRFs are < 1.01). The overlap of the $\pi_{\text{population}_r}$ chain-wise marginal posterior density estimates is almost exact and the estimates are accurate (figure 5.7). Based on the σ trace plots and corresponding PSRFs, the chains run on the dataset with $\sigma = 200$ and 20 converge (PSRFs < 1.02), but not those run on the dataset with $\sigma = 2$ (PSRF point estimate 1.27), seemingly on account of the second chain (yellow line, figure 5.8c). These results demonstrate the viability of the sampler for individual-SNP data. They also highlight the need to check convergence for each parameter of interest.

To assess the results when the same data are analysed assuming no inter-child variation, the individual-SNP datasets are analysed under the original model (equation (3.5)). Only when σ is low ($\sigma = 2$) does the assumption have a bearing on the estimated frequency (left violin motif, figure 5.9). For $\sigma = 2$, the model without a random effect underestimates the minor



(a) $\sigma_{\text{sample}} = 200$ (b) $\sigma_{\text{sample}} = 20$ (c) $\sigma_{\text{sample}} = 2$

Figure 5.7: Frequency estimates based on individual-SNP simulated data. The plots show chain-wise marginal posterior density estimates of the allele frequencies (colour-coded by chain), obtained by running the sampler with the $\boldsymbol{\pi}_{\text{population}}$ proposal on the real line for 100,000 iterations.



(a) $\sigma = 200$ (b) $\sigma = 20$ (c) $\sigma = 2$

Figure 5.8: Estimates of relatedness parameters based on individual-SNP simulated data. The plots show chain-wise posterior density estimates (colour-coded by chain) of the relatedness parameter, σ, for three simulated datasets with $\sigma = 200, 20$ and 2, after for 100,000 iterations using the $\boldsymbol{\pi}_{\text{population}}$ proposal on the real line. The vertical, dashed black line marks the value of σ used to simulate the data.

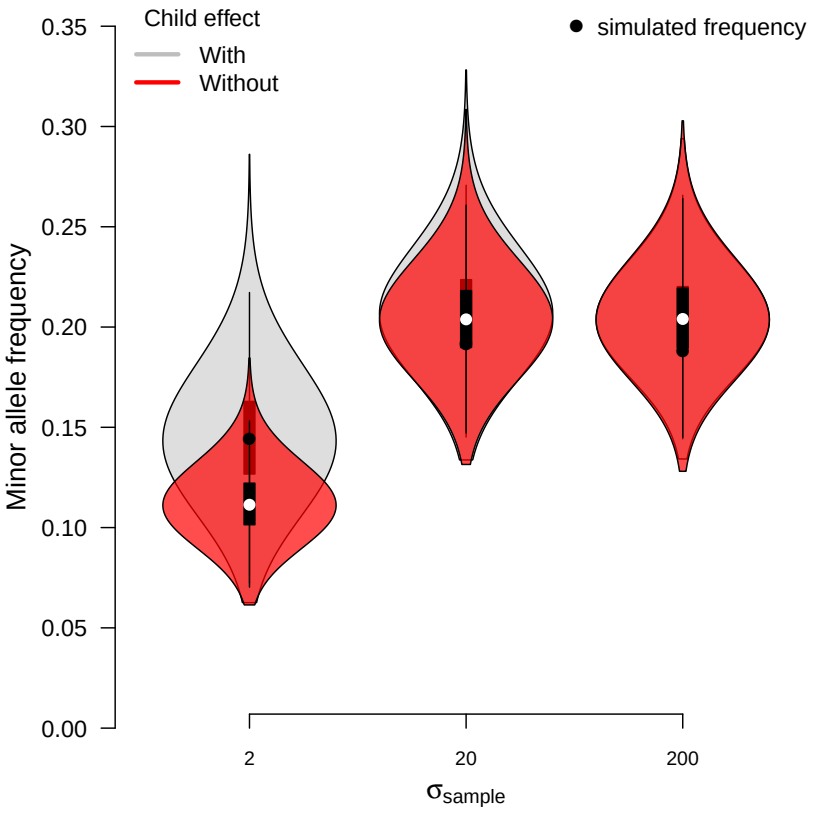


Figure 5.9: Allele frequency estimates under models with and without a chid effect. The plots show the marginal posterior density estimates of the minor allele frequencies of three simulated datasets with varying σ values, each obtained by fitting the model with (equation (5.1)) and without (equation (3.5)) inter-child variation.

allele frequency. Moreover, the variance of the marginal posterior density estimate generated under the model without the child effect is comparatively small, namely because the model without a child effect ignores intra-child dependence.

5.5 Ugandan data study

5.5.1 Methods

Outline of data analysed

To investigate the impact of the independence assumption on the frequencies reported in the previous chapter, the original aim was to re-estimate frequencies for all 60 gene-wise longitudinal subdivisions of the Ugandan data (see panels D and E, figure 4.3) under the extended model (equation (5.1)), and then compare the results with those generated under the original model (equation (3.5)). However, the sampler proposed for the extended model (section 5.3) does not formally converge when run on triple-SNP data (see section 5.4.2). As such, we are able to analyse all the subdivisions pertaining to *pfcrt* (since they comprise a single nSNP only), but not those pertaining to *pfmrp1* and *pfmdr1*. To enable indirect consideration of the *pfmrp1* and *pfmdr1* haplotype frequency trends, the longitudinal *pfmrp1* and *pfmdr1* subdivisions are further divided by nSNP, leading to 100 individual-SNP *pfmrp1* and *pfmdr1* subdivisions (5 nSNPs $\times$ 5 years $\times$ 2 drug arms) plus (5 nSNPs $\times$ 5 days since last treatment $\times$ 2 drug arms, AL and DP). Instead of reanalysing all *pfmrp1* and *pfmdr1* individual-SNP subdivisions, a subset is selected on prior evidence of a child effect. Prior evidence of a child-effect is based on auxiliary study (section B.2), in which subdivisions are ranked by the p-value representing the statistical significance of inter-child variation. Nine *pfmrp1* and *pfmdr1* single-nSNP subdivisions with p-values < 0.15 are selected; in order of increasing p-values, these are

1. *pfmdr1*-86, 29–42 days since last treatment, AL (p-value 0.01),

2. *pfmdr1*-86, 2010, AL (p-value 0.01),
3. *pfmrp1*-876, 29–42 days since last treatment, DP (p-value = 0.03),
4. *pfmdr1*-86, 2009, AL (p-value 0.03),
5. *pfmdr1*-1246, 2012, AL (p-value = 0.08),
6. *pfmdr1*-184, 57–70 days since last treatment, AL (p-value = 0.08),
7. *pfmdr1*-86, 2008, DP (p-value = 0.08),
8. *pfmdr1*-1246, 2009, AL (p-value = 0.10),
9. *pfmdr1*-1246, 2008, AL (p-value = 0.12),

Hence, in total, 29 individual-SNP subdivisions are analysed under the extended model: 20 pertaining to *pfcrt* (see panels D and E, figure 4.3) and nine pertaining to *pfmrp1* and *pfmdr1* (listed above). Of the 20 pertaining to *pfcrt*, three are found to have statistically significant inter-child variation based on the auxiliary study (section B.2). In order of increasing p-values, these are

1. *pfcrt*-76, 29–42 days since last treatment, AL (p-value = 0.06),
2. *pfcrt*-76, 43–56 days since last treatment, AL (p-value = 0.06),
3. *pfcrt*-76, 2011, AL (p-value = 0.13).

Data analysis

For each of the 29 individual-SNP subdivisions selected for analysis, five chains are run for 10,000 iterations using the $\boldsymbol{\pi}_{\text{population}}$ proposal on the real line, since it is found to be marginally more efficient than that on the simplex (see section 5.4.2). The hyperparameters are $a = 2$, $s = 100$, $\boldsymbol{\alpha} = (1,1)$, $\lambda = 2.94$ and $m_{\max} = 8$. The prior on $m_{i\,\text{child}}$ is a truncated geometric prior (identical to that used to analyse the Ugandan data in the previous chapter).

5.5.2 Results

Based on trace plots and PSRFs, the analyses of all 29 individual-SNP Ugandan data subdivisions converge. In general, the diagnostic plots are exemplary. Overlap between the chain-wise posterior density estimates is virtually-perfect for all but three subdivisions, for which overlap is near-perfect. There are numerous cases of high $\pi_{\text{population}_r}$ and σ auto-correlation, however. The worse case scenarios for *pfcrt* and *pfmrp1* are mentioned in the text below. For *pfmdr1*, very high $\pi_{\text{population}_r}$ autocorrelation is observed for *pfmdr1*-184, AL, 57–70 days since last treatment; relatively high $\pi_{\text{population}_r}$ and σ autocorrelation for *pfmdr1*-86, DP, 2008; and medium $\pi_{\text{population}_r}$ autocorrelation for *pfmdr1*-1246, AL, 2012. In addition, it is noted that, for all but five of the *pfcrt*-76 subdivisions (AL, 2012; DP, 2012; and DP 43–56, 57–70 and $>$ 70 days since last treatment), the log-posterior densities are bottom heavy with high spikes (ranging 100s of log-posterior units), suggesting the posterior might be highly concentrated but with heavy tails.

pfcrt subdivisions

A moderate child effect is detected in one of the 20 subdivisions *pfcrt* subdivisions: AL, 29–42 days since last treatment (orange dashed line, figure 5.10). The child effect also appears to impact the posterior density estimate of the PfCRT:K76 frequency (grey violin motif, 29–42 days, top left subplot, figure 5.11). However, comparatively high $\pi_{\text{population}_r}$ and σ auto-correlation caution against too much confidence in the result. In any case, the estimate for AL, 29–42 days has little bearing on the overall trend (figure 5.11a). In fact, the independence assumption appears to have little bearing on any of the trends (figure 5.11).

pfmdr1 individual-SNP subdivisions

Based on posterior support for low values of σ (figure 5.12a), there is evidence of a child effect for five of the eight *pfmdr1* data subdivisions analysed:

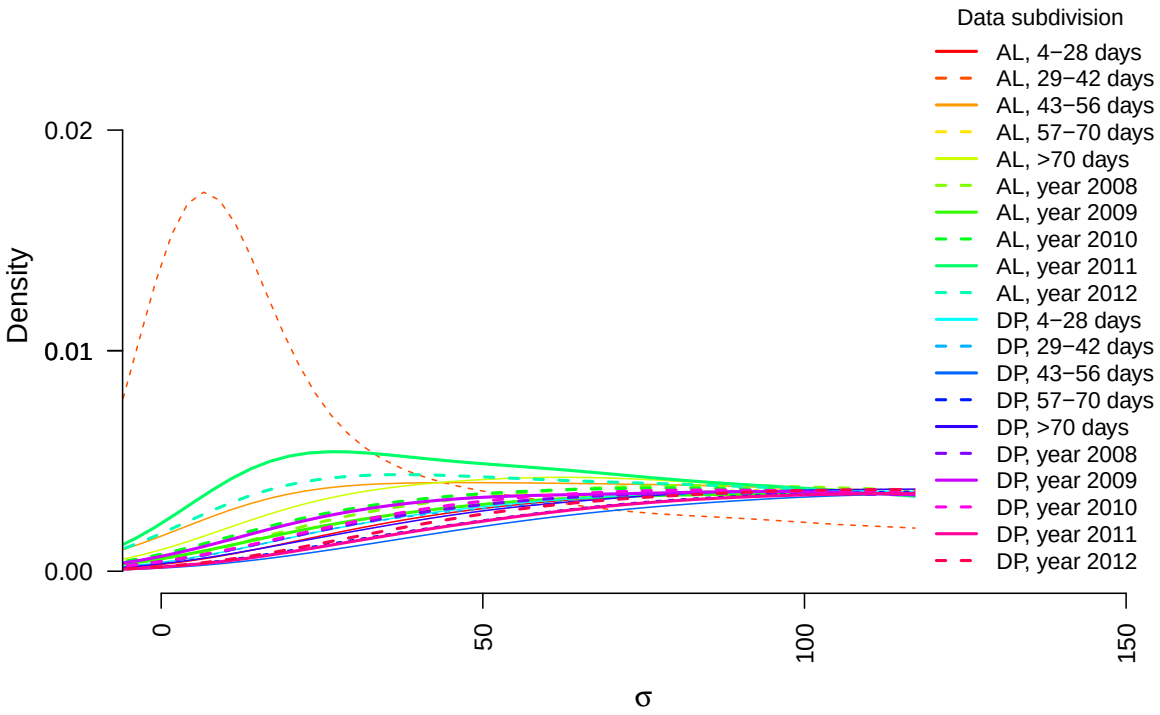
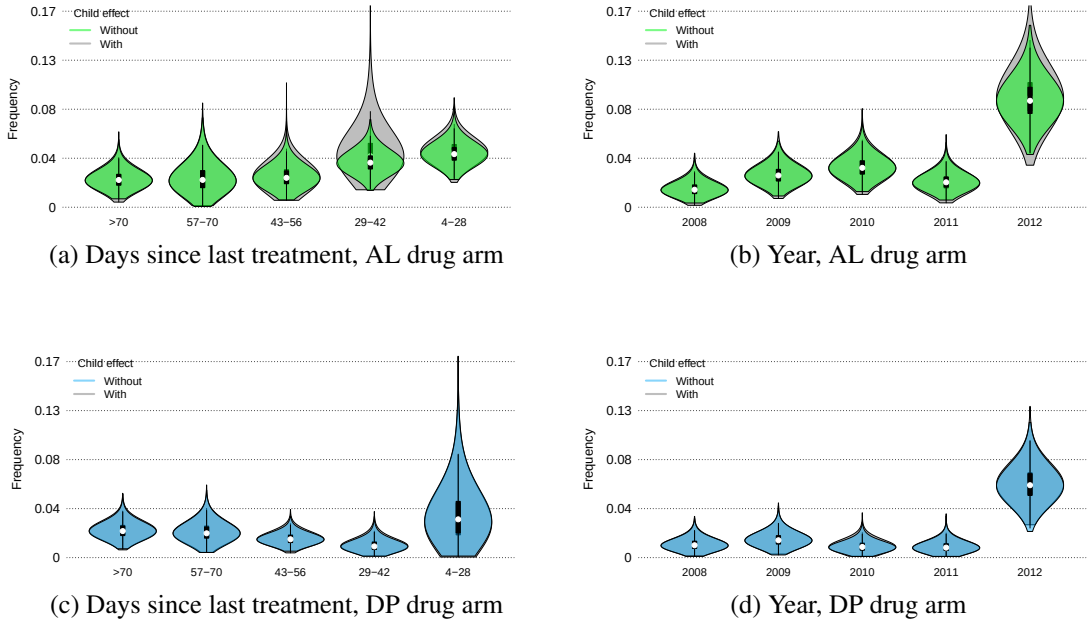


Figure 5.10: Posterior density estimate of the relatedness parameter, σ, for each of the 20 *pfcrt* subdivisions. The lines depict posterior density estimates colour-coded by subdivisions.



(a) Days since last treatment, AL drug arm

(b) Year, AL drug arm

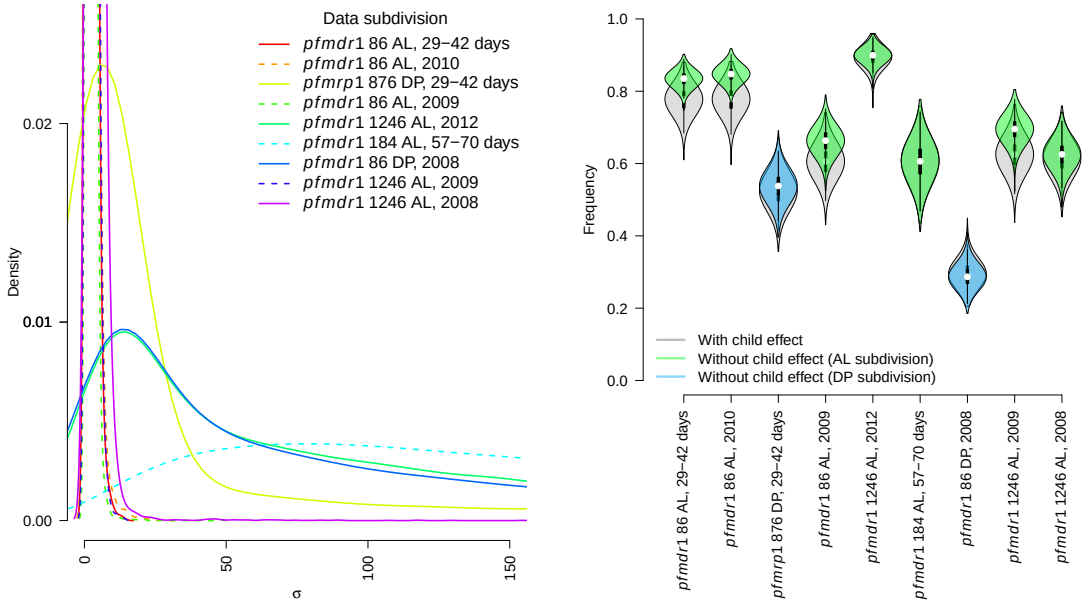
(c) Days since last treatment, DP drug arm

(d) Year, DP drug arm

Figure 5.11: Posterior density estimates of the PfCRT:K76 marker frequencies estimated under the models with (equation (5.1)) and without (equation (3.5)) the child effect taken into account plotted against correlates of time.

1. *pfmdr1*-86, 29–42 days since last treatment, AL,
2. *pfmdr1*-86, 2010, AL,
3. *pfmdr1*-86, 2009, AL,
4. *pfmdr1*-1246, 2009, AL,
5. *pfmdr1*-1246, 2008, AL,

four of which are also associated with a shift in the corresponding marginal posterior densities (figure 5.12b). As anticipated, under the independence assumption, the posterior density estimates of the majority allele frequencies are more precise and less uniform (closer to one). The child effect for *pfmdr1*-1246, AL, 2008, has no impact upon the allele frequency (far right, figure 5.12b), perhaps because the σ posterior (purple solid line, figure 5.12a) is shifted rightwards just enough to render the child effect undetectable with respect to the frequency estimates.



(a) Posterior density estimates of σ, colour-coded in order of the respective p-values.

(b) Posterior density estimates of allele frequencies generated under models with (equation (5.1)) and without (equation (3.5)) the child effect taken in account. Horizontally arranged in order of increasing p-values.

Figure 5.12: Posterior density estimates based on selected individual-SNP *pfmdr1* and *pfmrp1* subdivisions. The data subdivisions are ordered with respect to increasing p-values (see list above, section 5.5.1)

pfmrp1 individual-SNP subdivision

The yellow, solid line in figure 5.12 suggests there is a relatively moderate child effect for the one *pfmrp1* subdivision analysed. However, high σ auto-correlation cautions against too much confidence in the result. In any case, the child effect has no impact upon the marginal posterior density estimate of the minor allele frequency (leftmost blue violin motif, figure 5.12b).

5.6 Discussion

The aim of this chapter is to investigate an assumption made in the previous chapter, namely that all malaria episodes within a given subdivision of the data are independent, with a view to assessing of how it might shape our scientific interpretation of the results. To assess the impact of the assumption, results generated under the original haplotype-frequency model (equation (3.5)) are compared with results generated under an extended model (equation (5.1)). Implementation of the extended model is viable for individual-SNP data, enabling direct assessment of the allele frequency trends corresponding to PfCRT:K76, but prohibitively slow for triple-SNP data. Consequently, nSNP-wise *pfmdr1* and *pfmrp1* data are analysed, allowing for indirect assessment of the impact of the assumption on the *pfmdr1* and *pfmrp1* haplotype frequencies. The results suggest the independence assumption is reasonable with regards to the *pfcrt*-K76 frequency trends. For the *pfmdr1* haplotype frequency trends, the assumption is likely to have a small effect, but unlikely to change the overall trend. The results are discussed in more detail below.

The impact of the independence assumption upon the allele frequencies corresponding to PfCRT:K76 is either negligible or small, with only one exception (29–42 days, top subplot, figure 5.11), which has no bearing on the overall trend. Similarly, the assumption has little impact upon the *pfmrp1*-876 allele frequencies of the one *pfmrp1* subdivision analysed (leftmost blue violin motif, figure 5.12). On the contrary, the independence assumption has a notable

impact upon the population-level allele frequencies of four of the eight *pfmdr1* subdivisions analysed (figure 5.12). These are

1. *pfmdr1*-86, 29–42 days since last treatment, AL,
2. *pfmdr1*-86, 2010, AL,
3. *pfmdr1*-86, 2009, AL,
4. *pfmdr1*-1246, 2009, AL.

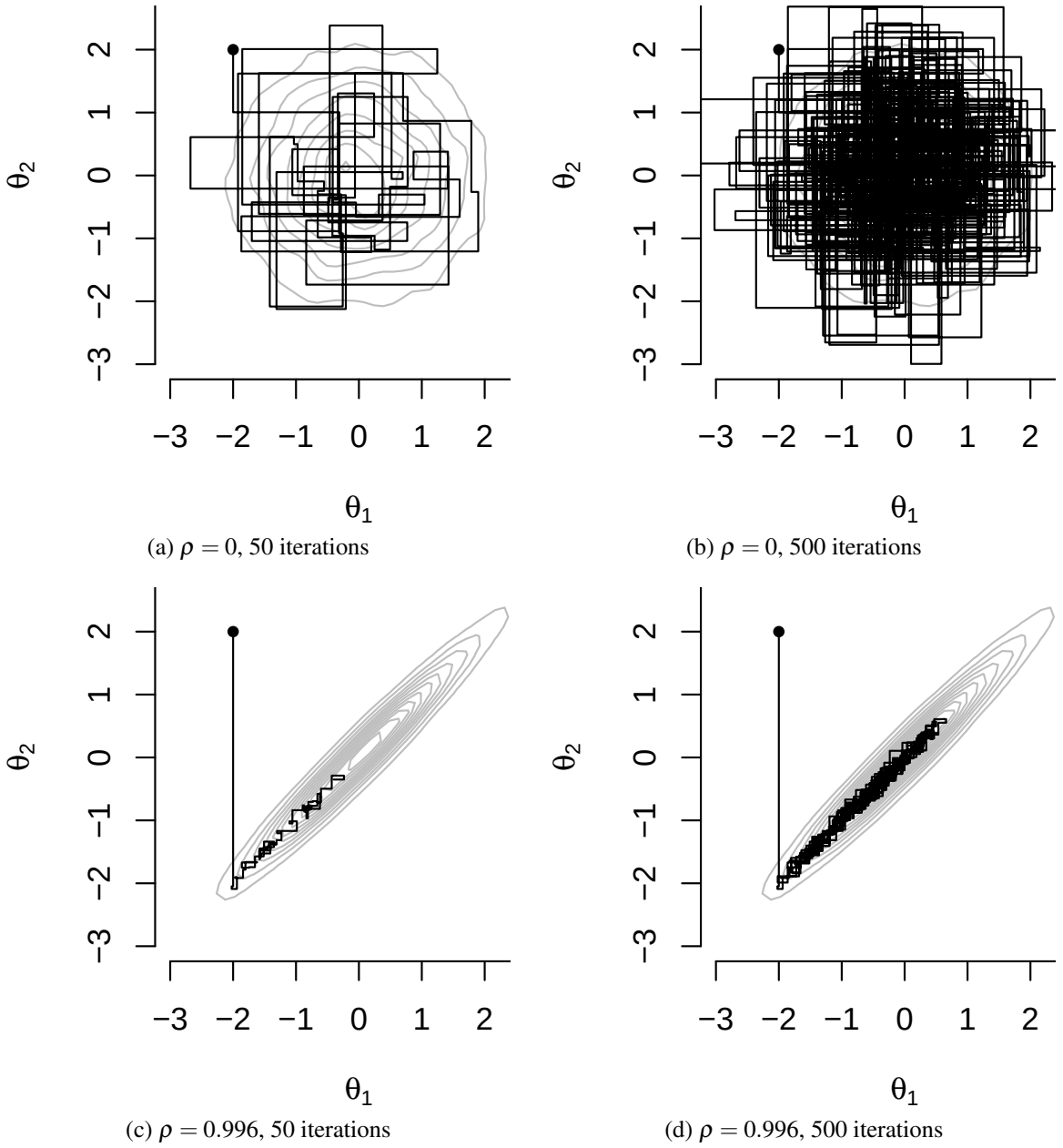
Since *pfmdr1* subdivisions pertaining to AL, 2009, 2010, and 29–42 days since last treatment are impacted by the independence assumption at the level of a single nSNP (figure 5.12), it is posited that the same three subdivisions will be impacted at the level of the haplotype. More specifically, given the allele frequencies (figure 5.12) and the prior impact of σ on $\boldsymbol{\pi}_{\text{child}}$, the *pfmdr1* AL, 2009 and 2010 and 29–42 days since last treatment haplotype frequencies (figures 4.12 and 4.12, respectively) are likely to be spuriously precise, and some may be underestimated. The impact upon the haplotype frequencies is expected to be less than that on allele frequencies, however, since for AL, 2010 and 29–42 days since last treatment, only one of three *pfmdr1* nSNP (codon 86) is affected, while for AL 2009, two of three (codons 86 and 1246) are affected (figure 5.12). Of course, without testing each and every subdivision, the impact on the analyses of the remaining *pfmrp1* and *pfmdr1* subdivisions cannot be established with certainty. There is little ancillary evidence of a child effect in the remaining subdivisions (see table B.2), however.

The above conjecture assumes that the ancillary study has enough power to detect all subdivisions for which a child effect is likely to have an impact upon the population-level allele frequencies. From the ordering of individual-SNP subdivisions (in order of their respective p-values) in figures 5.12a and 5.12b, it is clear that evidence of a child effect based on the posterior support of σ is associated with an impact on the posterior frequency estimate, but does not agree entirely with evidence derived from the ancillary study. To test the power assumption,

each and every subdivision could be analysed. Ideally, however, one would focus on improving the implementation of the model to enable direct analysis of triple-nSNP data.

As indicated above, the failure to converge for triple-SNP data calls for a more efficient sampler. It is well known that posterior correlation between parameters restricts the mobility of the Gibbs sampler [213], because the sampler takes smaller steps and therefore more time to explore the posterior distribution as demonstrated in Figure 5.13, which is adapted from figures 11.2 and 11.3 in [88]. As such, a measure that reduces correlation (for example, a parameter transformation or the addition of an auxiliary variable) could accelerate convergence [245, 192, 88]. Similarly, a joint update (of $\boldsymbol{\pi}_{\text{child}}$ and $\boldsymbol{\pi}_{\text{population}}$, say) could improve efficiency [213]. With regards to the existing Metropolis-Hastings update, one could replace $\boldsymbol{\Sigma}$ (equation (5.6)) with a scaled estimate of the covariance at the mode of the target density (which can be updated along the run of the algorithm as in [93]), or utilise the Langevin algorithm to manoeuvre the proposal in a favourable direction (as dictated by the gradient of the log target density) [214]. Alternatively, a more advanced sampler could be employed. Parallel tempering, for example, makes use of auxiliary chains with dampened modes to improve the mobility of the target chain [146], whereas Hamiltonian Monte Carlo uses a 'momentum' variable to propel the sampler through the parameter space [67] (also see [35]). A variation of the Hamiltonian Monte Carlo algorithm [103] can be implemented in STAN [235], a probabilistic programming language. However, given the spiky posteriors seen for the *pfcrt* subdivisions, parallel tempering might be the sensible next option.

Returning to the results obtained for *pfmdr1*, there is clear evidence of a child effect for *pfmdr1*-86, AL, 29–42 days since last treatment; *pfmdr1*-86, AL, 2010 and 2009; and *pfmdr1*-1246, AL, 2009 and 2008 (figure 5.12). It is tempting to suggest that there may be some intrinsic relationship between the host immune system and the drug resistant nSNPs *pfmdr1*-86 and *pfmdr1*-1246. However, to the best of our knowledge, there is no evidence to support this. Prevailing wisdom appears to endorse the view that, "The immune response kills parasites



(a) $\rho = 0$, 50 iterations

(b) $\rho = 0$, 500 iterations

(c) $\rho = 0.996$, 50 iterations

(d) $\rho = 0.996$, 500 iterations

Figure 5.13: Posterior correlation and the Gibbs sampler. The plots show two different Gibbs samplers (black lines), both simulating from bivariate unit normal distributions (grey contours), one with zero correlation, $\rho = 0$ (5.13a and 5.13b) and one with $\rho = 0.996$ (5.13c and 5.13d) after 50 (5.13a and 5.13c) and 500 (5.13b and 5.13d) iterations. Both samplers start at 2 (black points). Neither have converged after 50 iterations, but the sampler simulating from the distribution with zero correlation has explored more of its target distribution (5.13a). After 500 iterations the sampler simulating from the distribution with zero correlation is nearer to convergence (5.13b), while the sampler simulating from the distribution with high correlation is yet to fully explore its target distribution (5.13d).

irrespective of their sensitivity to antimalarial drugs." [273]. Moreover, if an innate interaction between the drug resistant nSNPs and the host immune system does exist, one might expect all *pfmdr1*-86 and *pfmdr1*-1246 subdivisions to show evidence of a child effect, regardless of the drug arm, year or immediacy since last treatment. Nonetheless, inter-individual variation in immunity may play a role. Imagine, for example, a child who suffers more frequent bouts of malaria than the average child. He or she will receive more frequent treatment, thus spend more time with sub-therapeutic residual drug in his or her blood. Sub-therapeutic residual drug acts as a filter upon reinfection, selecting resistant parasites [273]. Consequently, the child who is more frequently sick is more likely to present with the resistant parasites upon reinfection. As such, we might expect biases in favour of *pfmdr1*-N86 and *pfmdr1*-D1246 in the AL arm under the model that does not take the child effect into account, consistent with the results observed (figure 5.12b). Of course, other factors that lead to inter-child diversity might also play a role (for example, drug metabolism [94], sickle cell trait [227] or sleeping under a bed net [49]). Recrudescence is also possible, although less that 1% of infections are classified as recrudescent in the Ugandan study [51]. It can therefore be safely assumed that almost all recurrent episodes are due to reinfection.

Finally, it is noted that, in the previous chapter, the *pfmdr1* haplotype frequencies are regressed onto correlates of time. The obvious next step would be to repeat the regression but with frequencies estimated under the extended model (equation (5.1)). In the previous chapter, a separate regression is fit for each of the R haplotypes, and conditional independence is assumed between the frequencies, $\pi_{r1}, \ldots, \pi_{rK}$, that featured in the regression (equation (4.2)). Yet within a given drug arm, children not only feature multiple times within a subdivision, they also feature multiple times across the K subdivisions (figure 4.1). At the population level, one could model dependence between the $\pi_{r1}, \ldots, \pi_{rK}$ using a $K \times K$ covariance matrix that is not proportional to the identity matrix. This does not capture dependence between the R haplotypes, however, nor does it take into account dependence between the K subdivisions at the level of

the child (between $\boldsymbol{\pi}_{\text{childAL,2008}}$ and $\boldsymbol{\pi}_{\text{childAL,2009}}$ say). To capture dependence between the haplotypes and at the level of the child, a joint model that both estimates frequencies and regression coefficients is required.

Chapter 6

Frequency estimation using short-read sequencing data

6.1 Background

The data that have featured in this thesis thus far are prevalence data (categorical variables indicating whether or not wild and/or mutant type alleles have been detected at genotyped SNPs (see section 1.2.2 for more details)). Prevalence data are routinely generated for surveillance of antimalarial resistance, hence most models capable of generating estimates of malaria haplotype frequencies have been designed to analyse prevalence data [39, 102, 224, 129, 95, 276, 125, 223, 248]. Other types of data include sequencing data [143]. Sequencing methods are numerous, including many next generation technologies capable of generating an unprecedented volume of highly resolved genomic data [143]. For reasons explained below, sequencing is not yet amenable for routine surveillance of antimalarial resistance. Nevertheless, with a view to what lies ahead, we present a modification of our original model to accommodate Illumina short-read sequencing data, a common type of next generation sequencing data [151]. This chapter is intended as an exploratory study with the expectation of future elaboration.

In the field of malaria, sequencing data are commonly used in genome-wide association stud-

ies [14, 44, 154, 242] and to study population genetics [262, 142]. Despite rapidly decreasing costs [269], they remain prohibitively expensive for routine surveillance. Moreover, sequencing typically requires a comparatively large quantity of uncontaminated parasite DNA, which, in turn, requires a comparatively high-volume of blood. High volumes of blood necessitate blood collection by venipuncture [111], which is not readily deployable for routine surveillance in the field. Furthermore, high volumes of blood necessitate thorough decontamination, requiring technical expertise [37]. Despite the challenges facing sequencing for routine surveillance, sequencing plays a critical role in the surveillance of artemisinin resistance [217]. This is because the list of *pfkelch13* mutations associated with prolonged parasite clearance is still evolving, and sequencing can detect unanticipated mutations, whereas genotyping generally cannot [217]. Given the wealth of information in next generation sequencing data, innovative methods to sequence DNA extracted from filter paper blood samples are being developed [177]. Positive rapid diagnostic tests have also proven to be valuable source of *P. falciparum* DNA for monitoring resistance [191], their retention for sequencing has been proposed [217].

Illumina sequencing of *P. falciparum* DNA extracted from a blood sample generates millions of reads of the genetic template [151], which are typically mapped onto a reference genome such as that of clone 3D7 (see figure 6.1 for an example of mapped reads). Since there is more than one genetically distinct template in a multiclonal sample, reads from a multiclonal blood sample will differ at heteroallellic SNPs (for example, positions 2, 4, 8 and 11, figure 6.2). For a given heteroallelic SNP, the within-sample frequency (termed probability in figure 6.2) of a particular allele is equal to the allele read count divided by the total read count. For example, in figure 6.2, the adenine frequency at position 2 is between 0.3 and 0.4.

Although allele frequencies can be obtained directly from short-read sequencing data, long-range haplotype frequencies cannot [247]. Many statistical models have been proposed to estimate haplotype frequencies from short-read data generated by sequencing metagenomic pools [306, 134, 117, 185]. Most aim to estimate within-pool haplotype frequencies. O'Brien

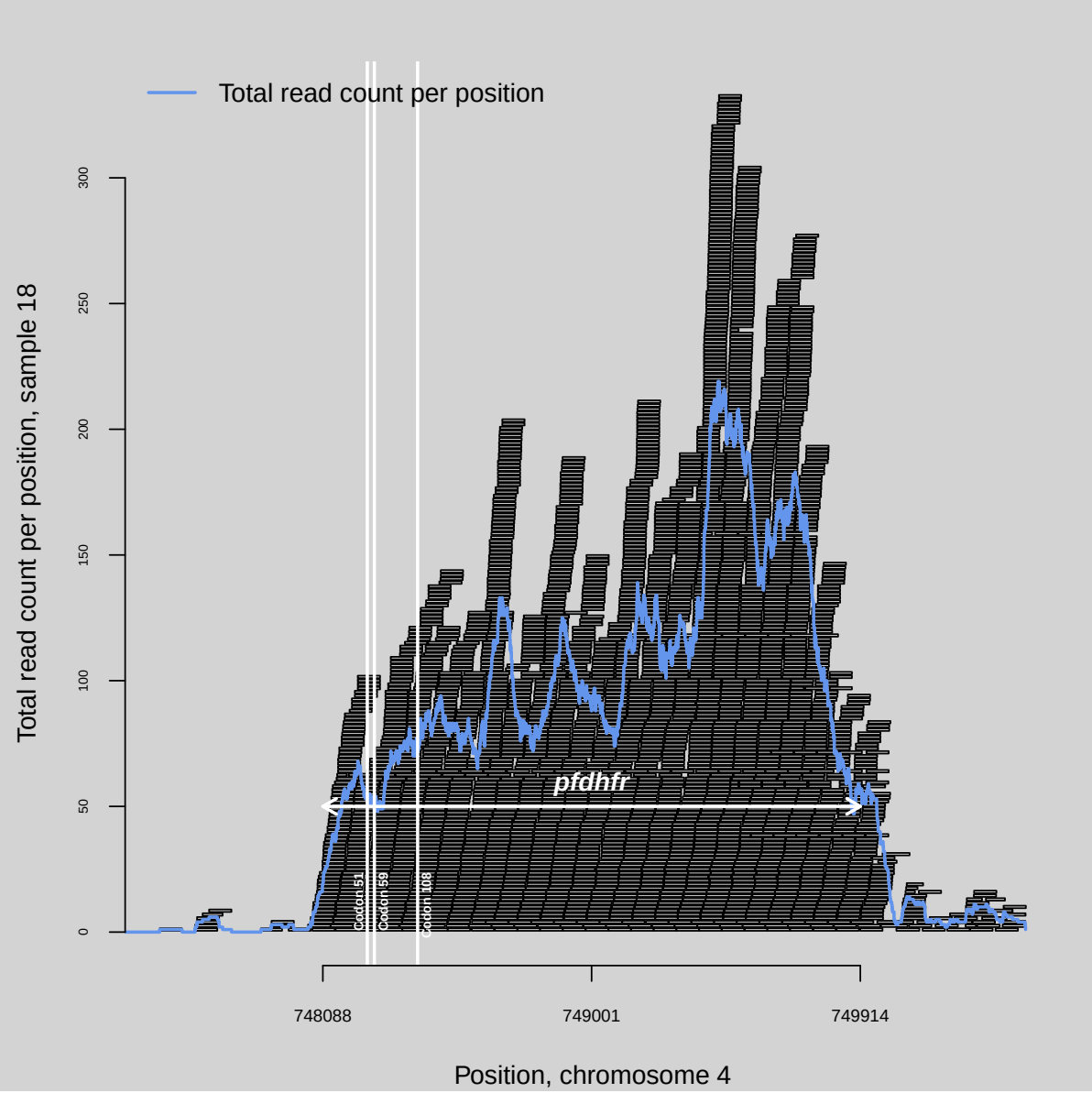


Figure 6.1: An example of the total read count per position in the genomic region of *pfdhfr*. The total read count per position is depicted by a blue line. It is calculated using the function `coverage` from the R package `IRanges` [128]. It is based on thousands of short-reads (represented by grey/black stacks) mapped to the canonical *P. falciparum* genome 3D7 [85]. The reads are generated by Illumina sequencing of *P. falciparum* DNA extracted from a blood sample collected from a malaria patient in Northern Ghana (see [8] for details). They are plotted using the function `plotRanges`, which can be found in the vignette of the R `IRanges` package [128]. The coding region of *pfdhfr* is indicated by a white horizontal line. The central positions of codons 51, 59 and 108 are denoted by white vertical lines.

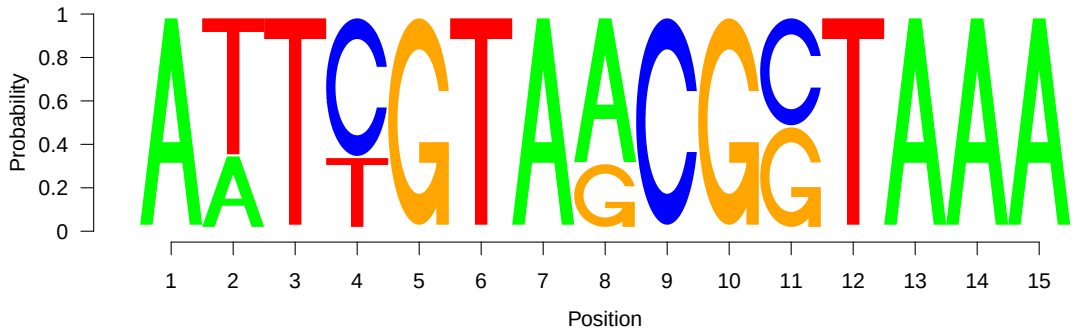


Figure 6.2: Nucleotide assignment and allele frequencies determined by Illumina sequencing of *P. falciparum* DNA extracted from a blood sample collected from a malaria patient in Northern Ghana (see [8] for details). The genomic positions correspond to position 1 to 3 of codons 51, 59, 108 in *pfdhfr* and codons 437 and 540 in *pfdhps*. Letters represent nucleotides (Adenine, Guanine, Cytosine and Thymine). At each position, the height of a given letter is equal to the corresponding nucleotide allele frequency (labelled probability on the vertical axis). The allele frequency is equal to the allele read count dived by the total read count. The figure is generated in R [210] using the function `seqLogo` from the package `seqLogo` [25].

et al. [185], for example, infer sample-wise haplotype frequencies for numerous samples, each of which is treated as a pool. O'Brien *et al.* model read counts using a binomial likelihood. Following [185], we replace the original likelihood function of our original model (equation (3.6)) with a binomial likelihood, thereby accommodating short-read data.

To demonstrate the utility of the modified model, pre-existing short-read sequencing data from Ghana are analysed. The data are described in detail elsewhere [8]. In summary, they were generated by Illumina sequencing of *P. falciparum* DNA purified from blood samples collected between August 2009 and November 2011 from all-age patients presenting with uncomplicated malaria at the Kologo Health Clinic in the Kassena-Nankana District (KND) of Northern Ghana [8]. To assess the extent of anti-folate resistance in the KND, the genomes were sequenced in the regions of *pfdhfr* and *pfdhps* (positions 748088 to 749914 on chromosome 4, and positions 548200 to 550616 on chromosome 8, respectively, of version 3.0 of *Pf*3D7) (see PlasmoDB [20]). The reads were mapped to the canonical reference, 3D7 [85] (see figure 6.1 for an example of mapped reads).

The modified model proposed here (figure 6.3) has practically the same framework as

that of the original model (figure 3.1), including the prior on the MOI with parameter λ and maximum MOI $m_{\max}$. In previous chapters, λ and $m_{\max}$ are set equal to estimates of the average and maximum MOI, respectively, based on auxiliary MOI data. Auxiliary MOI data are not available for the 50 Ghanaian samples. Instead, Amenga-Etego found high within-host genetic diversity based on F_{ws} scores [8], population-normalised measures of the genome-wide proportions of heteroallelic SNPs [142]. Due to the reportedly high within-host genetic diversity, MOI estimation by conventional PCR methods was deemed to be extremely challenging [8]. As such, MOI data are not available, leading us to base values of λ and $m_{\max}$ on literature derived estimates. We identified five studies concerning estimates of MOI in the KND [189, 221, 75, 33, 81]. All five relate to subsets of the data described in [221], which are derived from samples collected between July 2000 and May 2001 from all-aged, asymptomatic *P. falciparum* positive individuals. Based on the full dataset, estimates of the average and maximum MOIs are 3.16 and 8, respectively, the latter being the upper limit of the experimental method [221]. That said, due to complex infection dynamics, at any given point in time, the fraction of detectable clones circulating in an infected individual is thought to be $\approx 50\%$ of the total [33], suggesting the true all-age average is approximately $2 \times 3.16 = 6.32$. Moreover, when MOI estimates are stratified by age, sample means vary considerably, and inferred estimates are thought to reach as high as 19 [81]. Given the wide range of possible values for λ and $m_{\max}$, in this chapter the data are analysed under three different hyperparameter assignments (see section 6.2.4). The fit is compared using model checks based on deviance (outlined in section 6.2.5).

The aim of this chapter is to expand the statistical toolbox for the genetic surveillance of antimalarial resistance. More specifically, we plan to modify the model documented in chapter 3 to enable analysis of short-reading sequencing data. In addition, we aim to assess the performance of the model using simulated data, and to apply the modified model to previously published data from Ghana, thereby demonstrating the capacity for model checking under the

modified model, while garnering insight into the level of anti-folate resistance in Ghana.

This chapter is outlined as follows. In the next section we describe the likelihood modification (subsection 6.2.1), the sampler used to implement the modified model (6.2.2), the data simulated to assess model performance (6.2.3), the formatting of the Ghanaian data required to fit the modified model (6.2.4) and a suite of model checks (6.2.5). In the results section (section 6.3), we concentrate first on the simulated study (6.3.1), then on the Ghanaian results (6.3.2). The chapter ends with a discussion (section 6.4)

6.2 Methods

6.2.1 The modified model

The model designed to analyse short-read data is a modification of the original model (equation 3.1), the only difference being the likelihood. Details of the original model can be found in chapter 3. Herein we focus on the modified likelihood (equation (6.2)), followed by a summary of the posterior decomposition in light of the modified likelihood (equation (6.4)).

Let y_{ij} denote the read count of the mutant type allele in the ith blood sample at the jth SNP and z_{ij} denote the total read count. Treating z_{ij} as a non-stochastic observation and y_{ij} as a datum, the framework of the modified model (figure 6.3) is almost identical to that of the original model (figure 3.1), the only difference being the dependence upon z_{ij}. To capture the dependence upon z_{ij}, a binomial likelihood following [185] is proposed,

$$\rho(\boldsymbol{y}_i \mid \boldsymbol{a}_i) = \prod_{j=1}^{J} \left\{ \rho\left(y_{ij} \mid z_{ij}, p_{ij}\right) \right\} \tag{6.1}$$

$$= \prod_{j=1}^{J} \left\{ \mathscr{B}\text{inomial}\left(y_{ij} \mid z_{ij}, p_{ij}\right) \right\}, \tag{6.2}$$

where $\boldsymbol{a}_i$ is the vector of haplotype counts for the ith sample (see table 3.1 for a full list of notation), and where p_{ij} is the proportion of a mutant alleles. That is, the number of haplotypes

in the ith blood sample with the mutant type allele at the jth SNP ($\boldsymbol{a}_i \cdot \boldsymbol{h}_j$, where $\boldsymbol{h}_j$ is a column vector enlisting the allele states of the R possible haplotypes at the jth SNP), normalised by the total number of haplotypes in the ith blood sample ($m_i = \sum_{r=1}^{R} a_{ir}$),

$$p_{ij} = \frac{\boldsymbol{a}_i \cdot \boldsymbol{h}_j}{\sum_{r=1}^{R} a_{ir}}. \tag{6.3}$$

Note that equation (6.3) is the same as equation (3.7), hence we retain the simplifying assumption that all haplotypes are unequivocally detected (see the note under equation (3.7) for an illustrative example). Also note that under the proposed likelihood (equation (6.2)), y_{ij} are conditionally independent given z_{ij} and p_{ij}, hence any information captured on reads spanning multiple SNPs (for example, reads spanning nSNPs in codons 51 and 59, figure 6.1) will be lost. The joint posterior density of the modified model is still captured by equation (3.5), but with likelihood given by equation (6.2) above. For example, the joint posterior density of the modified model with a truncated Poisson prior on the MOI is[1],

$$\begin{aligned}\rho\left(\boldsymbol{\pi}, \boldsymbol{a}, \boldsymbol{m} \mid \boldsymbol{y}\right) \propto \prod_{i=1}^{I} \Bigg\{ \prod_{j=1}^{J} \left\{ \mathscr{B}\text{inomial}\left(y_{ij} \mid z_{ij}, p_{ij}\right) \right\} \mathscr{M}\text{ultinomial}\left(\boldsymbol{a}_i \mid m_i, \boldsymbol{\pi}\right) \\ \times \mathscr{P}\text{oisson}_{\text{trunctated}}\left(m_i \mid \lambda, m_{i\min}, m_{\max}\right) \Bigg\} \mathscr{D}\text{irichlet}\left(\boldsymbol{\pi} \mid \boldsymbol{\alpha}\right),\end{aligned} \tag{6.4}$$

where $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_R)$ is the vector of R population-level haplotype frequencies; $\boldsymbol{a}$ denotes the collection of sample-wise haplotype count vectors, $\boldsymbol{a}_i$ for $i = 1, \ldots I$; $\boldsymbol{m}$ denotes the collection of sample-wise MOIs, m_i for $i = 1, \ldots I$; $m_{i\min}$ and $m_{\max}$ denote the lower and upper bounds of the prior on the MOI; and λ and $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_R)$ are hyperparameters of the distributions on m_i and $\boldsymbol{\pi}$, respectively (see section 3.2 for a full list of notation). Note that we could specify a different distribution on the MOI (see equation (6.6) for example).

[1] Note that z_{ij} for $i = 1, \ldots I$ and $j = 1, \ldots, J$ do not feature in the notation $\rho\left(\boldsymbol{\pi}, \boldsymbol{a}, \boldsymbol{m} \mid \boldsymbol{y}\right)$ since they are treated as non-stochastic observations.

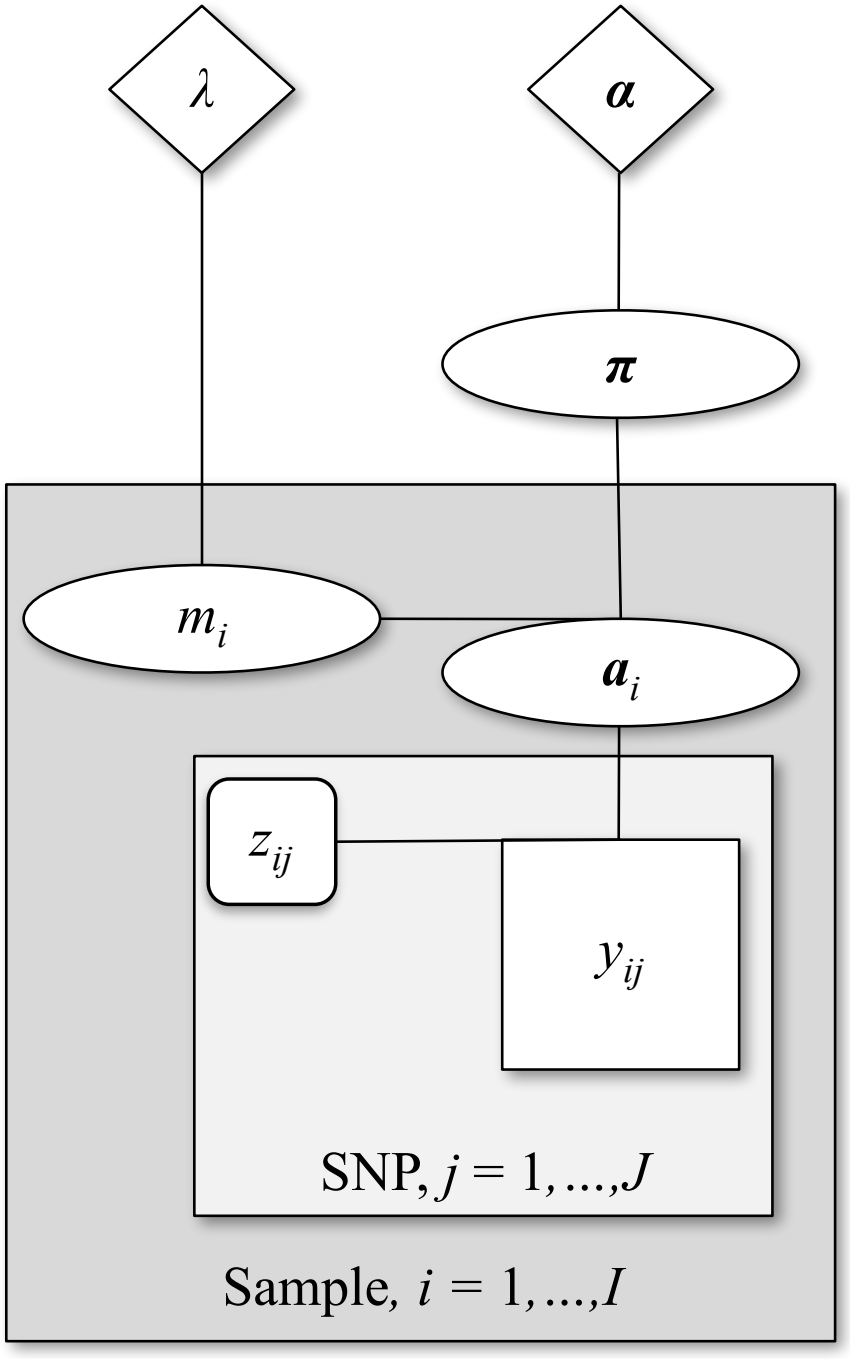


Figure 6.3: Haplotype frequency estimation model for short-read sequencing data. The graph shows the model quantities and their conditional dependencies. The data, y_{ij} for $I = 1,\ldots,I$ and $j = 1\ldots,J$, are represented by a square with hard edges. The total read counts, z_{ij} for $I = 1,\ldots,I$ and $j = 1\ldots,J$, are represented by a square with round edges. Ellipses denote unobserved parameters: the vectors of haplotype counts, $\boldsymbol{a}_i$ for $I = 1,\ldots,I$; the MOIs, m_i for $I = 1,\ldots,I$; and the vector of haplotype frequencies, $\boldsymbol{\pi}$. The diamonds represent the hyperparameters on m_i and $\boldsymbol{\pi}$ (λ and $\boldsymbol{\alpha}$, respectively). The density of the full model factorises as follows: $\rho(\boldsymbol{\pi},\boldsymbol{a},\boldsymbol{m},\boldsymbol{y}) = \rho(\boldsymbol{y} \mid \boldsymbol{a})\,\rho(\boldsymbol{a} \mid \boldsymbol{m},\boldsymbol{\pi})\,\rho(\boldsymbol{m})\,\rho(\boldsymbol{\pi})$. Note that z_{ij} for $i = 1,\ldots I$ and $j = 1,\ldots,J$ do not feature in the notation since they are treated as non-stochastic observations.

6.2.2 The sampler

The sampler described previously (section 3.2.4) is used to sample from the full posterior of the modified model (equation (6.4)), the only difference being $\mathscr{B}\text{inomial}\left(\boldsymbol{y}_{ij} \mid z_{ij}, p_{ij}\right)$ instead of equation (3.6) is plugged into equation (3.15). To assess convergence, five chains of the sampler are run for each analysis and trace plots visually inspected. In addition, PSRFs are calculated using the function `gelman.diag` from the package `CODA` [207]. Each chain is run for T iterations per analysis and the first half discarded as burnin. Initially the sampler was run for $T = 10,000$ iterations, but the sampler did not converge, hence, herein, the sampler is rerun for $T > 10,000$ with different thinning intervals (see section 6.3 for final values of T). A full suite of diagnostic plots is generated and examined for each MCMC run.

6.2.3 Simulated data

For one to five SNPs, three different datasets are simulated, each with 100 blood samples. A vector of haplotype frequencies is drawn from a uniform Dirichlet prior. MOIs are drawn from a non-zero conditioned Poisson distribution with mean equal to three. Total read counts are based on those from the Ghanaian dataset sampled with replacement with added Poisson noise. Haplotype counts vectors are drawn from the multinomial prior (equation (3.8)), mutant allele proportions calculated according to equation (6.3), and mutant read counts drawn from the binomial likelihood (equation (6.2)). All 15 simulated datasets are analysed under the modified model (equation (6.4)), as well as the original haplotype-frequency estimation model (equation (3.5)), with $\boldsymbol{\alpha} = (1,\ldots,1)$, $\lambda = 3$, and $m_{\text{max}} = 8$. For the analyses under the original model

(equation (3.5)), the short-read sequencing data are summarised as prevalence data as follows,

$$f\left(\frac{y_{ij}}{z_{ij}}\right) = \begin{cases} 0 \text{ if } {}^{y_{ij}}/_{z_{ij}} = 0 \\ 1 \text{ if } {}^{y_{ij}}/_{z_{ij}} = 1 \\ 0.5 \text{ otherwise.} \end{cases} \tag{6.5}$$

6.2.4 Formatting the Ghanaian data

To demonstrate the utility of the modified model (equation (6.4)), data derived from $I = 50$ Ghanaian samples are analysed (see [8] for full details of the data). A brief overview of the Ghanaian data can be found in section 6.1. Here we focus on how the data are formatted in preparation for fitting the model. The data are formatted in R [210] using a suite of functions from the packages `Rsamtools` [160], `GenomicAlignments` [128], `IRanges` [128], `GenomicRanges` [128] and `Biostrings` [190]. In summary, the mapped reads are extracted from the `BAM` files in which they are stored (one file per blood sample). A pile-up matrix of reads per nucleotide base per position is generated for each genomic region per blood sample. Base calls with mapping qualities less than 30 are ignored when summarising the pile-up matrix, as are reads with mapping qualities less than 30. The minority read threshold is set equal to 2. Reads corresponding to nSNPs associated with amino acid changes N51**I**, C59**R** and S108**N** in PfDHFR and amino acid changes A437**G** and K540**E** in PfDHPS are extracted. (Note that throughout this thesis bold font is used to distinguish mutant type alleles from wild type alleles.) Read counts are characterised as either mutant type or wild type according to the amino acids they encode. There are no unidentified base calls nor missing data. For the ith blood sample at the jth nSNP, the mutant allele read count is stored in y_{ij} and the total read count stored in z_{ij}. The Ghanaian data are analysed under the modified model (equation (6.4)). To determine the best distribution for the prior on m_i, four different distributions are considered:

λ	$m_{\max}$	Reason for choice of hyperparameter values
3.16	8	Sample estimates reported by Sama *et al.* [221]
6.32	16	On account of the 50% detectability reported by Bretscher *et al.* [33]
15	$20^{\dagger}$	λ based on model inference reported by Felger *et al.* [81]

Table 6.1: Three hyperparameter assignments (one per row). The reason for the choice of λ and $m_{\max}$ is stated. †Standard errors are not reported in [81], hence the upper bound is somewhat arbitrary.

equations (3.10) to (3.13), summarised below,

$$\rho(m_i) = \begin{cases} \mathscr{U}\text{niform}(m_i \mid m_{i\min}, m_{\max}), \\ \mathscr{P}\text{oisson}_{\text{truncated}}(m_i \mid \lambda, m_{i\min}, m_{\max}), \\ \mathscr{G}\text{eometric}_{\text{truncated}}(m_i \mid \lambda, m_{i\min}, m_{\max}) \text{ and} \\ \mathscr{N}\text{egative } \mathscr{B}\text{inomial}_{\text{truncated}}(m_i \mid \lambda, \phi = 10, m_{i\min}, m_{\max}). \end{cases} \tag{6.6}$$

Of 50 samples, 26 are heteroallelic at one or more of the $J = 5$ nSNPs listed above, but all 50 samples have at least one heteroallelic loci across the genomic regions sequenced, hence $m_{i\min} = 2$ for $i = 1, \ldots 50$. Given the wide range of possible values for λ and $m_{\max}$ (see section 6.1), three different sets of λ and $m_{\max}$ are considered (table 6.1). In total, the Ghanaian dataset is thus analysed under 12 different prior combinations (once for each of the three sets of hyperparameter assignments summarised in table 6.1, under each of the four distributions listed in equation (6.6)). Posterior predictive checks (section 6.2.5) are used to assess fit of each prior combination to the data.

6.2.5 Model checking

Following [88], posterior predictive checks are used to assess the adequacy of the model fit. More specifically, replicate data are compared to observed data (replicate data that resemble the observed data are indicative of model fit, since they imply that data that resemble the observed data are feasible under the model). To assess the similarity between observed and replicate

data, graphical summaries are plotted side by side. The discrepancy between the replicate and observed data is measured quantitatively using an *a posteriori* diagnostic, the deviance at the posterior means of the haplotype count vectors. Deviance can also be incorporated into an estimator of model selection, the deviance information criterion (DIC) [233]. The DIC is a measure of the tradeoff between the fit of the model and its complexity. It is easily calculable given the MCMC output and deviance at the posterior means [233], and therefore is a natural complement to posterior predictive checks based on deviance. Replicate data, their graphical summaries, the deviance test statistic and the DIC are outlined in detail below.

Replicate data: For each of the 12 prior combinations, ten thousand replicate datasets, $\{\boldsymbol{y}^{\text{rep}_l}\}_{l=1}^{10,000}$, are generated as follows. Recall that $\boldsymbol{a}$ denotes a collection of haplotype vectors, $\boldsymbol{a}_i$ for $i = 1,\ldots,I$, where $I = 50$ in the Ghanaian dataset. Ten thousand $\boldsymbol{a}$ are drawn uniformly at random from the MCMC samples of all five chains combined, $\{\boldsymbol{\pi}^n, \boldsymbol{a}^n, \boldsymbol{m}^n\}_{n=1}^{25,000}$. Let $\{\boldsymbol{a}^l\}_{l=1}^{10,000}$ denote the sample of ten thousand $\boldsymbol{a}$. For each $\boldsymbol{a}_i^l \in \boldsymbol{a}^l$, $p_{ij}^l = (\boldsymbol{a}_i^l \cdot \boldsymbol{h}_j)/\sum_{j=1}^J a_{ij}^l$ is calculated. The elements of the lth replicate dataset, $y_{ij}^{\text{rep}_l}$, are sampled according to,

$$y_{ij}^{\text{rep}_l} \sim \mathscr{B}\text{inomial}\left(\cdot \mid z_{ij}, p_{ij}^l\right) \text{ for } j = 1,\ldots 5 \text{ and } i = 1,\ldots 50. \tag{6.7}$$

Graphical summaries of replicate data: To visually compare the replicate and observed data, a graphical summary of the observed dataset is plotted alongside graphical summaries of the replicate datasets. Under each prior combination, the replicate datasets are sampled uniformly at random from the 10,000 generated. To highlight any patterns in the data, graphical summaries are constructed as follows. First, the mutant read counts are normalised by the total read counts, y_{ij}/z_{ij}, for $j = 1,\ldots 5$ and $i = 1,\ldots 50$. The columns and rows of the 50×5 matrix $\boldsymbol{y}/\boldsymbol{z}$ are then arranged in order of decreasing column and row sums. The ordered matrix is plotted as a heat map with greyscale ranging from white to black proportional to $y_{ij}/z_{ij} \in [0,1]$. Heat maps of normalised replicate datasets are then plotted alongside the observed data. The

columns and rows of the heat maps of the normalised replicate datasets are arranged in the same order as the columns and rows of the heat map of the observed dataset. Hence, cells across multiple heat maps are comparable to one another.

The deviance at the posterior means of the haplotype count vectors: A quantitive measure of the difference between the observed and replicate data is based on the deviance, D, at the posterior means of the haplotype count vectors, $\hat{\boldsymbol{a}}_i = \dfrac{1}{25,000}\sum_{n=1}^{25,000} \boldsymbol{a}_i^n$ for $i = 1,\ldots,50$. Deviance is a measure of prediction error, equal to twice the negative log-likelihood (page 185 [88]). For $l = 1,\ldots,10,000$, the deviance of the replicate datasets at the posterior means, $\hat{\boldsymbol{a}}_i$ for $i = 1,\ldots,50$, is calculated as follows,

$$\begin{aligned} D(\boldsymbol{y}^{\text{rep}_l},\hat{\boldsymbol{a}}) &= -2\log\left\{\rho(\boldsymbol{y}^{\text{rep}_l},\hat{\boldsymbol{a}})\right\}, \\ &= -2\,\log\left\{\prod_{i=1}^{I}\left\{\prod_{j=1}^{J}\left\{\rho(y_{ij}^{\text{rep}_l}\mid\hat{\boldsymbol{a}}_i)\right\}\right\}\right\}, \\ &= -2\,\log\left\{\prod_{i=1}^{I}\left\{\prod_{j=1}^{J}\left\{\mathscr{B}\text{inomial}\left(y_{ij}^{\text{rep}_l}\mid z_{ij},\hat{p}_{ij}\right)\right\}\right\}\right\} \end{aligned} \tag{6.8}$$

where $\hat{p}_{ij} = {}^{(\hat{\boldsymbol{a}}_i\cdot\boldsymbol{h}_j)}/_{\sum_{j=1}^{J}\hat{a}_{ij}}$. The deviance of the replicate data at the posterior means (equation (6.8)) is compared with the deviance of the observed data at the posterior means,

$$D(\boldsymbol{y},\hat{\boldsymbol{a}}) = -2\,\log\left\{\prod_{i=1}^{I}\left\{\prod_{j=1}^{J}\left\{\mathscr{B}\text{inomial}\left(y_{ij}\mid z_{ij},\hat{p}_{ij}\right)\right\}\right\}\right\}, \tag{6.9}$$

by plotting a density estimate of the distribution over $\left\{D(\boldsymbol{y}^{\text{rep}\,l},\hat{\boldsymbol{a}})\right\}_{l=1}^{10,000}$ and a vertical line at the point $D(\boldsymbol{y},\hat{\boldsymbol{a}})$ for each of the 12 prior combinations considered. If the vertical line falls in the tail of the density estimate, the check implies the probability of a replicate dataset having a deviance as extreme as the observed data is low under the model, hence inadequate fit with respect to the deviance at the posterior means of the haplotype count vectors.

DIC: Model selection is based on the DIC,

$$\text{DIC} = \mathbb{E}[D(\boldsymbol{y}, \boldsymbol{a}^l)] + p_D. \tag{6.10}$$

The DIC measures the tradeoff between model fit and model complexity [233, 212]. Under the DIC, model fit is measured by expected deviance averaged over the entire MCMC sample,

$$\begin{aligned}
\mathbb{E}[D(\boldsymbol{y}, \boldsymbol{a}^n)] &= \frac{1}{25,000} \sum_{n=1}^{25,000} \{-2 \log\{\rho(\boldsymbol{y} \mid \boldsymbol{a}^n)\}\} \\
&= \frac{1}{25,000} \sum_{n=1}^{25,000} \left\{ -2 \log \left\{ \prod_{i=1}^{I} \left\{ \prod_{j=1}^{J} \{\rho(y_{ij} \mid \boldsymbol{a}_i^n)\} \right\} \right\} \right\} \\
&= \frac{1}{25,000} \sum_{n=1}^{25,000} \left\{ -2 \log \left\{ \prod_{i=1}^{I} \left\{ \prod_{j=1}^{J} \{\mathscr{B}\text{inomial}(y_{ij} \mid z_{ij}, p_{ij}^n)\} \right\} \right\} \right\},
\end{aligned} \tag{6.11}$$

where $p_{ij}^n = {}^{(\boldsymbol{a}_i^n \cdot \boldsymbol{h}_j)}/_{\sum_{j=1}^{J} a_{ij}^n}$. Model complexity is measured by the difference between the expected deviance (equation (6.11)) and the deviance of the observed data at the posterior means (equation (6.9)),

$$p_D = \mathbb{E}[D(\boldsymbol{y}, \boldsymbol{a}^n)] - D(\boldsymbol{y}, \hat{\boldsymbol{a}}). \tag{6.12}$$

The model with the lowest expected deviance can be seen to provide the best fit to the data, whereas the model with the lowest DIC is considered the best model in light of model complexity [233, 212].

6.3 Results

6.3.1 Simulated data

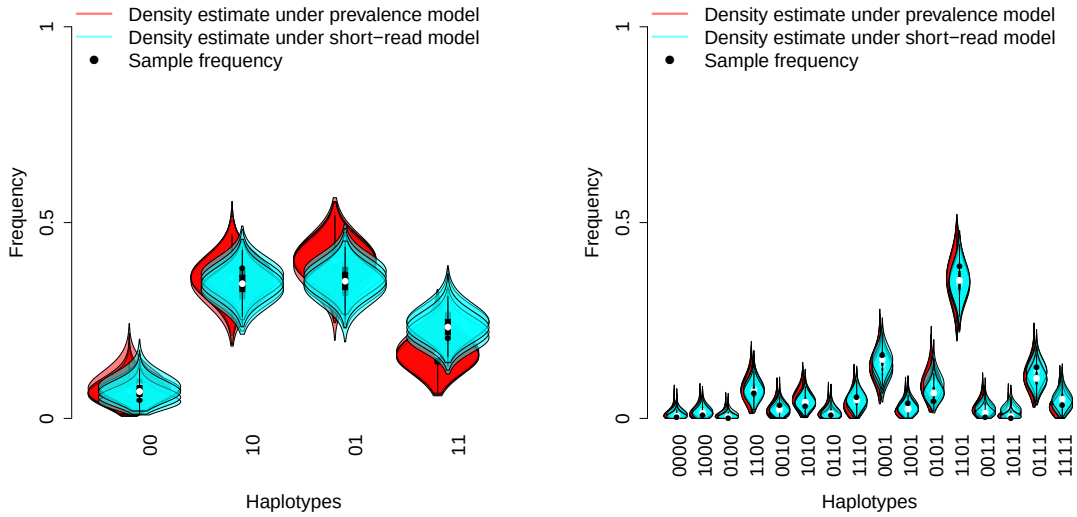
Based on both visual inspection of trace plots and PSRFs being ≤ 1.1, the sampler converges for datasets with $J = 1$ and for two of the three datasets with $J = 2$. For $J = 1$ the diagnostic plots are exemplary, aside from some autocorrelated m_i trace plots, and the posterior density estimates

of the frequencies, π_r, are highly accurate (blue violin motifs, figure 6.4a). As J increases, so too does π_r autocorrelation, leading to autocorrelated π_r trace plots. Based on visual inspection of log-posterior trace plots and PSRFs, the sampler fails to converge for $J > 2$. Trace plots of the log-likelihood are stable, however. The stability of the log-likelihood versus the instability of the log-posterior suggests the sampler explores regions of differing probability under the prior but equal probability under the likelihood. Hence, despite not having formally converged, the π_r density estimates (blue violin motifs, figure 6.4b) concentrate around the simulated π_r values (black dots, figure 6.4b). Plots of the posterior MOI samples suggest the posterior MOI mass is highly concentrated on the simulated values, generating autocorrelated MCMC samples but relatively accurate m_i point estimates (figure 6.5a). When the log-likelihood is set equal to zero and the sampler run on datasets with entirely missing genotype outcomes for $J \in \{1,\ldots,5\}$ the model returns both the prior on m_i and π_r as anticipated. In conclusion, the sampler appears to be drawing from the correct distribution but inefficiently. When the original model designed for prevalence data (equation (3.5)) is fit to the short-read data mapped to prevalence data, the corresponding sampler (see section 3.2.4) converges after $\leq$50,000 iterations. Aside from moderate π_r autocorrelation for $J \in \{2,3,4,5\}$, the diagnostic plots are exemplary. Under the model designed for prevalence data, the posterior density estimates of π_r are practically the same as those generated under the modified model (compare blue and red violin motifs, figure 6.4), while m_i posterior point estimates are improved (compare figure 6.5a with figure 6.5b).

6.3.2 Ghanaian data

First the data are analysed under the model with

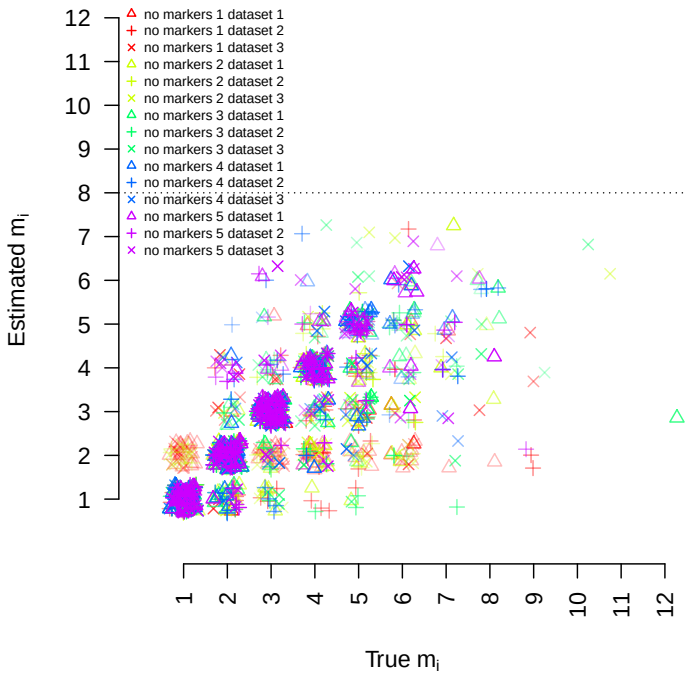
$$\rho(m_i) = \mathscr{P}\text{oisson}_{\text{truncated}}(m_i \mid \lambda = 3.16, m_{\min} = 2, m_{\max} = 8).$$



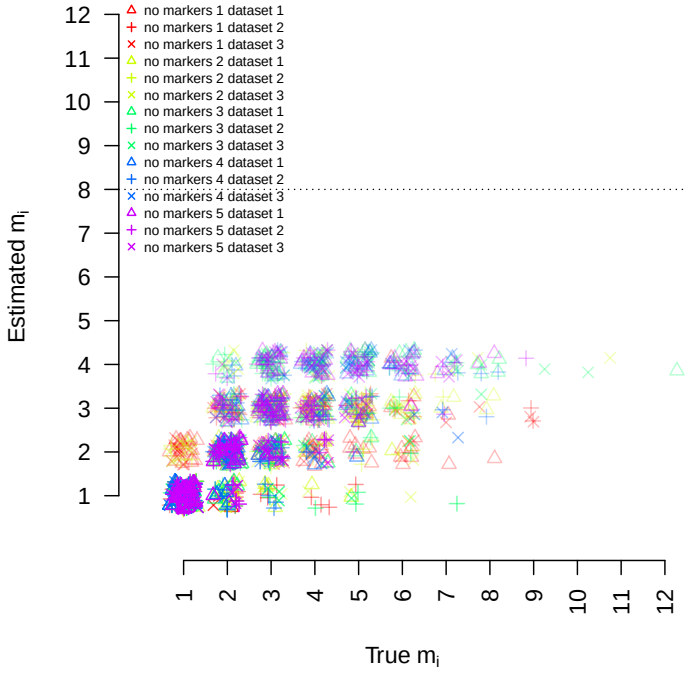
(a) Data simulated with one SNP

(b) Data simulated with three SNPs

Figure 6.4: Frequency estimates based on simulated data analysed under the original haplotype frequency estimation model (the prevalence model, equation (3.5)) and the model modified to accommodate short-read sequencing data (equation (6.4)). The plots depict chain-wise marginal posterior density estimates. The density estimates generated under the prevalence model (depicted in red) are shifted to the left to aid visibility. Under the prevalence model (equation (3.5)), the short-read data are summarised as prevalence data using equation (6.5). Since the data are simulated, the sample frequencies (black dots) are known.



(a) Data analysed under short-read model (equation (6.4))



(b) Data analysed under prevalence model (equation (3.5))

Figure 6.5: MOI estimates based on simulated data analysed under the original model designed for prevalence data and the model modified to accommodate short-read sequencing data. The plots show the 'True' simulated MOI, m_i, versus the posterior modal estimates for 15 datasets, each with $i = 1, \ldots, 100$ blood samples hence MOIs. The MOIs are jittered to aid visibility. The transparency of the points is inversely proportional to the posterior mass. Both plots are based on the same data. Before applying the prevalence model (equation (3.5)), the short-read data are mapped to prevalence data using equation (6.5).

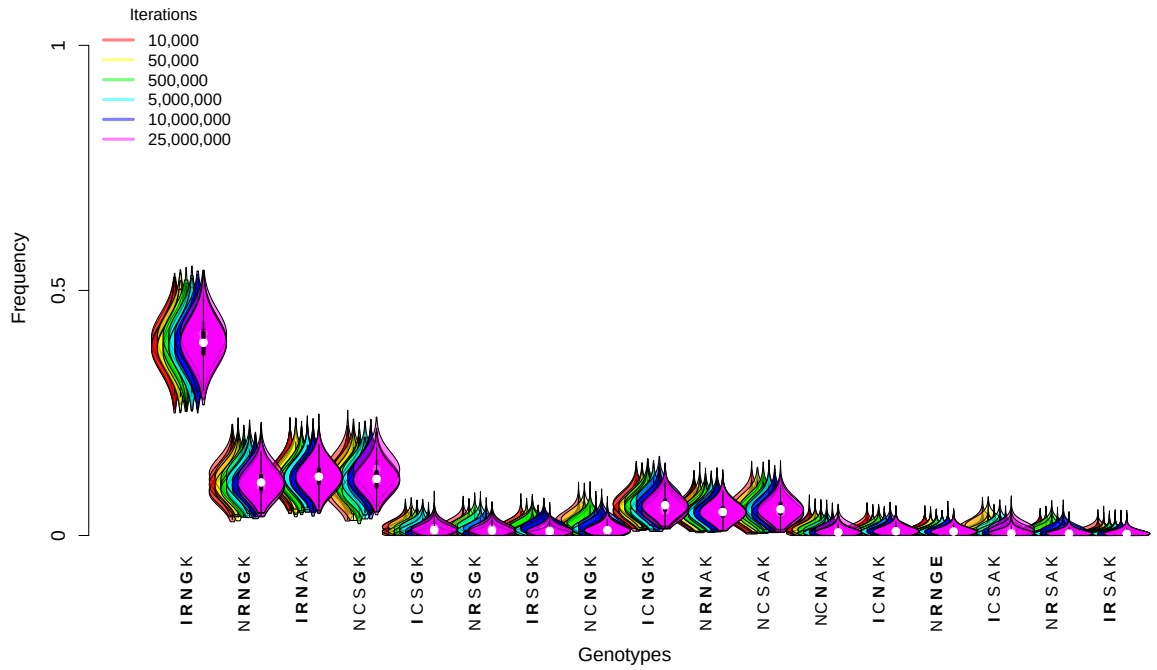


Figure 6.6: Frequency estimates for the Ghanaian data analysed using the sampler run for different numbers of iterations. The plot shows chain-wise marginal posterior density estimates of the genotype frequencies, π_r for $r = 1, \ldots, 17$, generated under the short-read model with a truncated Poisson prior on the MOI with $\lambda = 3.16$, $m_{\text{min}} = 2$ and $m_{\text{max}} = 8$. Genotypes are denoted by their corresponding amino acid sequences.

The sampler is run for 10,000 to 25,000,000 iterations yet fails to converge based on log-posterior trace plots and PSRFs (4 of 17 π_r PSRFs are between 1.1 and 1.5, when, ideally, they should be ≤ 1.1 [88]), similar to results based on simulated data (section 6.3.1, above). Also similar to results based on simulated data, visual inspection reveals that the log-likelihoods are stable for 10,000 iterations plus, explaining why posterior density estimates of π_r (figure 6.6), posterior modal estimates of the MOI (figure 6.7) and deviance summaries (figure 6.8) are stable. In summary, it seems that the sampler rapidly converges upon a set of parameter values that have high explanatory power with regards to the data, while proceeding to explore space that has higher probability under the prior. Consequently, the sampler appears not to have converged when, for the purposes of frequency estimation, it has. Based on this understanding, for the 12 analyses under the different prior combinations proposed for comparison (see section 6.2.4), five chains of the sampler are run for 10,000 iterations each.

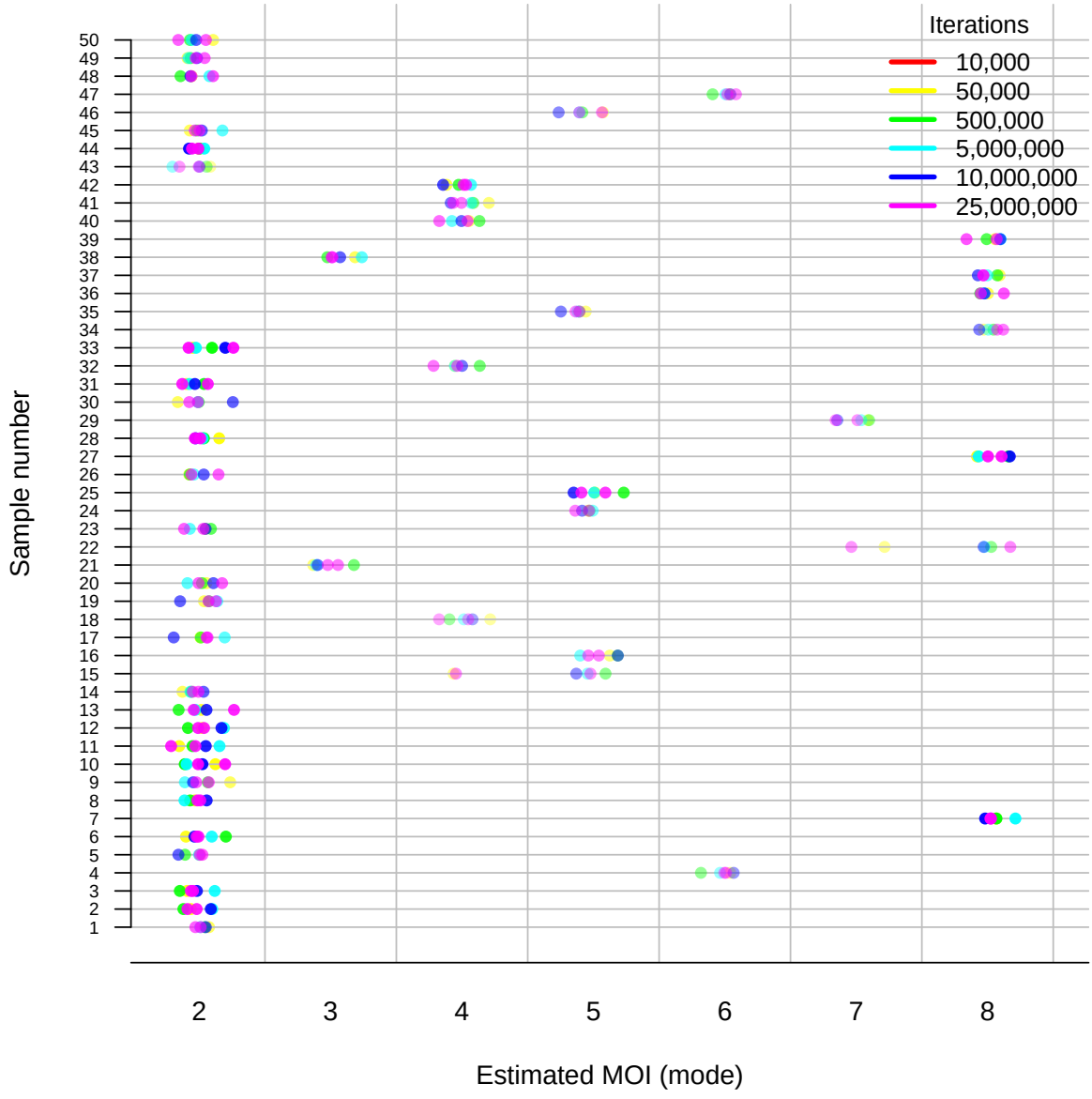
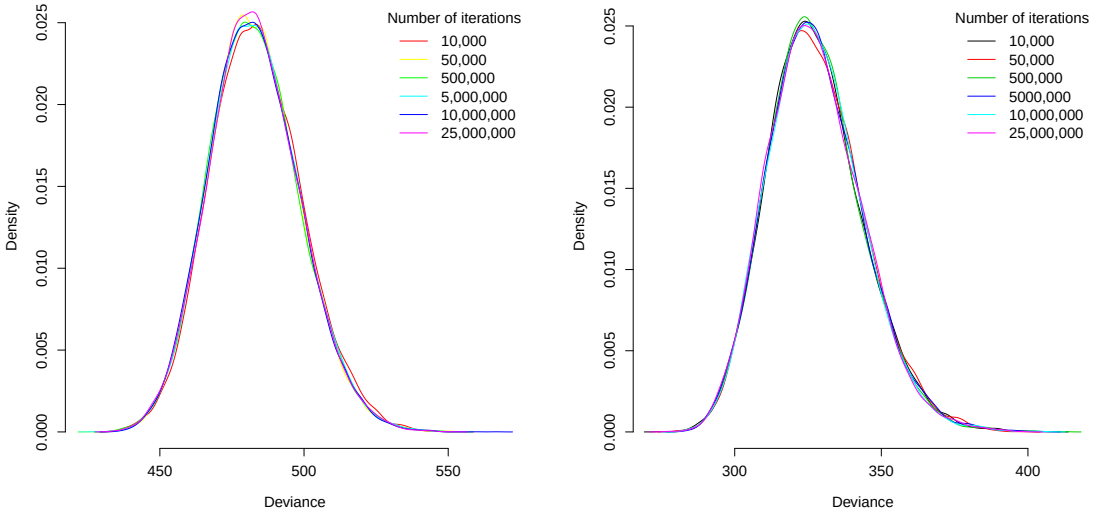


Figure 6.7: MOI estimates for the Ghanaian data analysed using the sampler run for different numbers of iterations. For each sample, the plot shows the posterior mode generated under the short-read model (equation (6.4)) with a truncated Poisson prior on the MOI with $\lambda = 3.16$, $m_{\text{min}} = 2$ and $m_{\text{max}} = 8$. The transparency of the points is inversely proportional to the posterior mass. The MOI estimates are jittered slightly for visibility.



(a) Density estimate of deviance distributed over posterior haplotype count vectors

(b) Density estimate evaluated at the posterior mean of the haplotype count vectors distributed over replicate data

Figure 6.8: Deviance summaries versus iterations. The deviance summaries are generated by fitting the short-read model with a truncated Poisson prior on the MOI with $\lambda = 3.16$, $m_{\text{min}} = 2$ and $m_{\text{max}} = 8$ to the Ghanaian data. The model is fit using the sampler run for different numbers of iterations.

As anticipated, not one of the 12 analyses proposed for comparison formally converges after 10,000 iterations, but the log-likelihood is stable and the π_r traces mix well, leading to near-perfect overlap of the marginal posterior density estimates of π_r. Many of the m_i trace plots are strongly autocorrelated (some acceptance rates approach zero), suggesting the posterior MOI distributions might be highly concentrated. Moreover, numerous posterior modes equal m_{max}, suggesting m_{max} might be too restrictive. In fact, of the 12 analyses performed, the prior combination

$$\rho(m_i) = \mathscr{P}\text{oisson}_{\text{truncated}}(m_i \mid \lambda = 15, m_{\text{min}} = 2, m_{\text{max}} = 20)$$

provided the best fit the data in terms of posterior predictive checks based on deviance evaluated at the point estimates of the haplotype count vectors (figure 6.9), and most preferable in terms of the DIC (table 6.2). To check if the result holds after more iterations, the data

Prior distribution	Hyperparameter assignment		
	$\lambda = 3.16, m_{\max} = 8$	$\lambda = 6.32, m_{\max} = 16$	$\lambda = 15, m_{\max} = 20$
Truncated Poisson	518 (34)	427 (35)	381 (33)
Truncated negative Binomial	510 (34)	422 (35)	395 (35)
Truncated Geometric	506 (33)	425 (36)	413 (36)
Uniform	489 (32)	412 (36)	409 (36)

Table 6.2: DIC with complexity (p_D) in parentheses. The table shows the DIC and p_D for twelve models differing with respect to four different prior distributions on the MOI and three different hyperparameter assignments. Each model is fit to the Ghanaian dataset using the sampler run for 10,000 iterations.

are reanalysed under the model with prior combination $\rho(m_i) = \mathscr{P}\text{oisson}_{\text{truncated}}(m_i \mid \lambda = 15, m_{\min} = 2, m_{\max} = 20)$, by running the sampler for 25,000,000 iterations. The analysis is also repeated under the model with prior combination $\rho(m_i) = \mathscr{U}\text{niform}\,(m_{\min} = 2, m_{\max} = 20)$. Both samples under the prior combinations with $m_{\max} = 20$ formally converge after 25,000,000 iterations (that is to say, they converge based on the log-posterior traces plots and PSRFs, as well as the log-likelihood trace plots), and the diagnostic plots are exemplary (the π_r traces mixed well and autocorrelation is low). The prior combination $\mathscr{P}\text{oisson}_{\text{truncated}}(m_i \mid \lambda = 15, m_{\min} = 2, m_{\max} = 20)$ remains the best fitting option and most preferable in terms of the DIC. Graphical summaries of replicate data (figure 6.10) show little difference between the model with a uniform prior with $m_{\max} = 20$ and a Poisson prior with $\lambda = 15$ and $m_{\max} = 20$.

To assess the sensitivity of the posterior estimates to the hyperparameter assignment, the results under the best and worst fitting models are compared (figure 6.11). Unsurprisingly given the vastly different prior support for the MOI, the π_r posterior density estimates differ. Note, however, that the difference in the marginal estimates is statistically insignificant, in that there is considerable overlap between the 95% credible intervals of the marginal density estimates. With regards to the posterior point estimates of the sample-wise MOIs (figure 6.12), some of the point estimates under the Poisson prior with $m_{\max} = 8$ (green points) are equal to $m_{\max} = 8$ with high modal mass (depicted by low transparency), suggesting that the upper bound may be too low. Under the Poisson prior with $m_{\max} = 20$, posterior mass encompasses most, if not all, available space (plot not shown), but the modal estimates (blue dots, figure 6.12) are all less

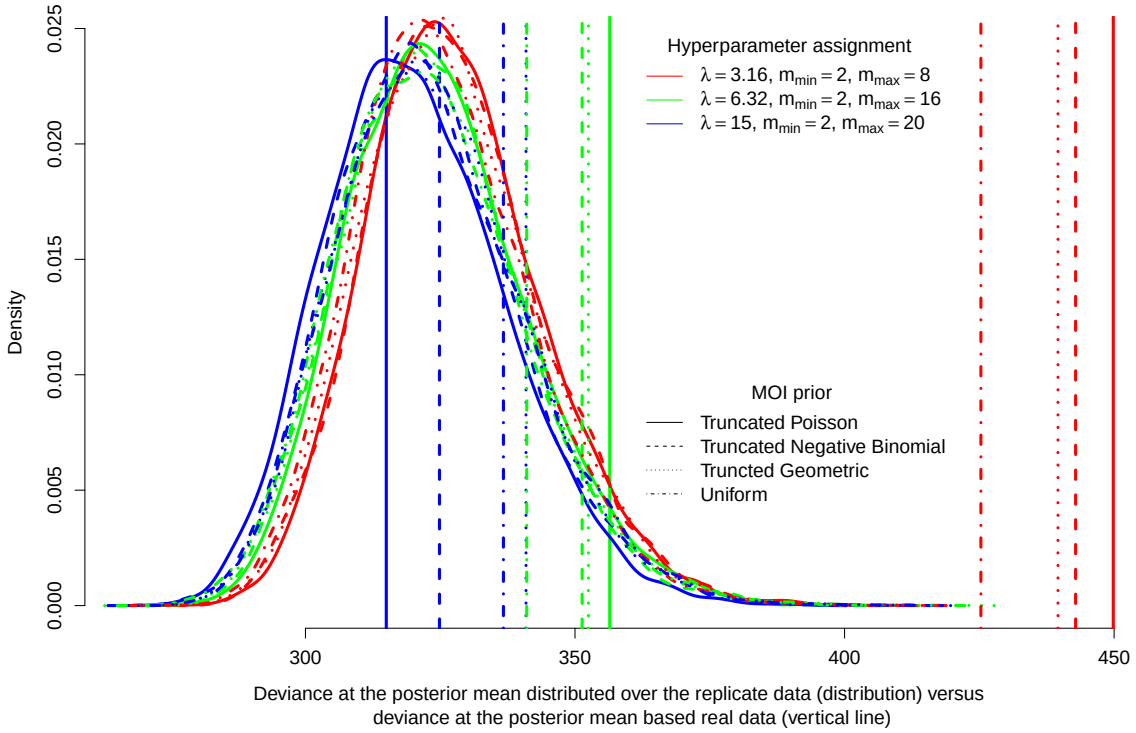
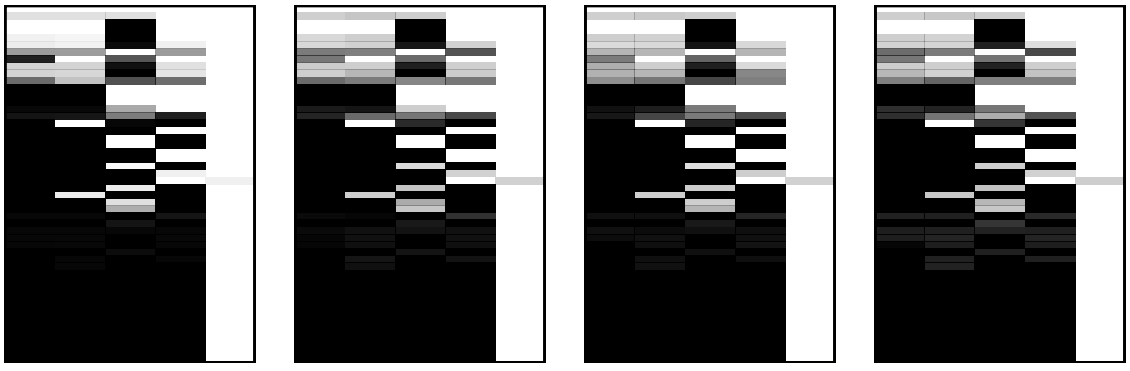


Figure 6.9: Density estimates of the deviance at the posterior means of the haplotype count vectors distributed over replicate data (curve) compared with the deviance of the observed data at the posterior means of the haplotype count vectors (vertical line) for twelve prior combinations, differing with respect to the prior on the MOI.



(a) Observed data, y/z

(b) Replicate data, y^{rep}/z; Uniform MOI prior with $m_{\text{min}} = 2$, $m_{\text{max}} = 20$

(c) Replicate data, y^{rep}/z; truncated Poisson prior with $\lambda = 15$, $m_{\text{max}} = 20$

(d) Replicate data, y^{rep}/z; truncated Poisson prior with $\lambda = 3.16$, $m_{\text{max}} = 8$

Figure 6.10: Visual posterior predictive check: graphical representations of the Ghanaian data (left-most plot), versus randomly selected replicate datasets generated under three different prior combinations (see subplot captions), each with $m_{\text{min}} = 2$. The replicate datasets are generated from random traces drawn from the MCMC sample having run five chains of the sampler each for 25,000,000 iterations on the Ghanaian data.

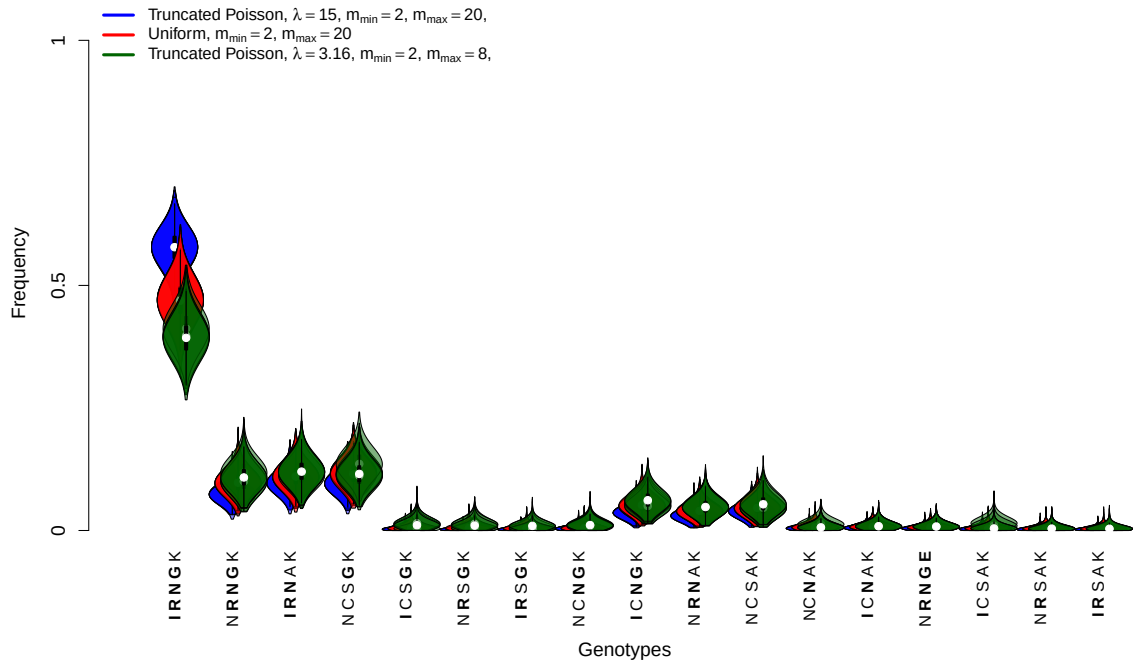


Figure 6.11: Ghanaian genotype frequency estimates. The plot shows chain-wise marginal posterior density estimates of the genotype frequencies, colour coded by model. Each model is fit by running five chains of the sampler for 25,000,000 iterations. Genotypes are denoted by their corresponding amino acid sequences.

than the upper bound of 20, with many equal to eight. Under the uniform prior, the posterior mass is concentrated upon lower MOIs compared to those generated under the Poisson prior with $m_{\max} = 20$ (plot not shown), with many modal estimates (red dots, figure 6.12) at the lower bound ($m_{\min} = 2$), but some as high as 14.

As an aside, to assess the impact of further increasing the mean and maximum MOI, additional models with unrealistically high mean and maximum MOIs are run. Under the uniform distribution, increasing $m_{\max}$ beyond 20 has little effect on the results nor fit (the extra support being superfluous). Under the Poisson distribution, increasing λ beyond 30 and $m_{\max}$ beyond 40 has a detrimental effect of the fit (the fit for $\lambda = 30$ and $m_{\max} = 40$ being approximately the same as that for $\lambda = 15$ and $m_{\max} = 20$).

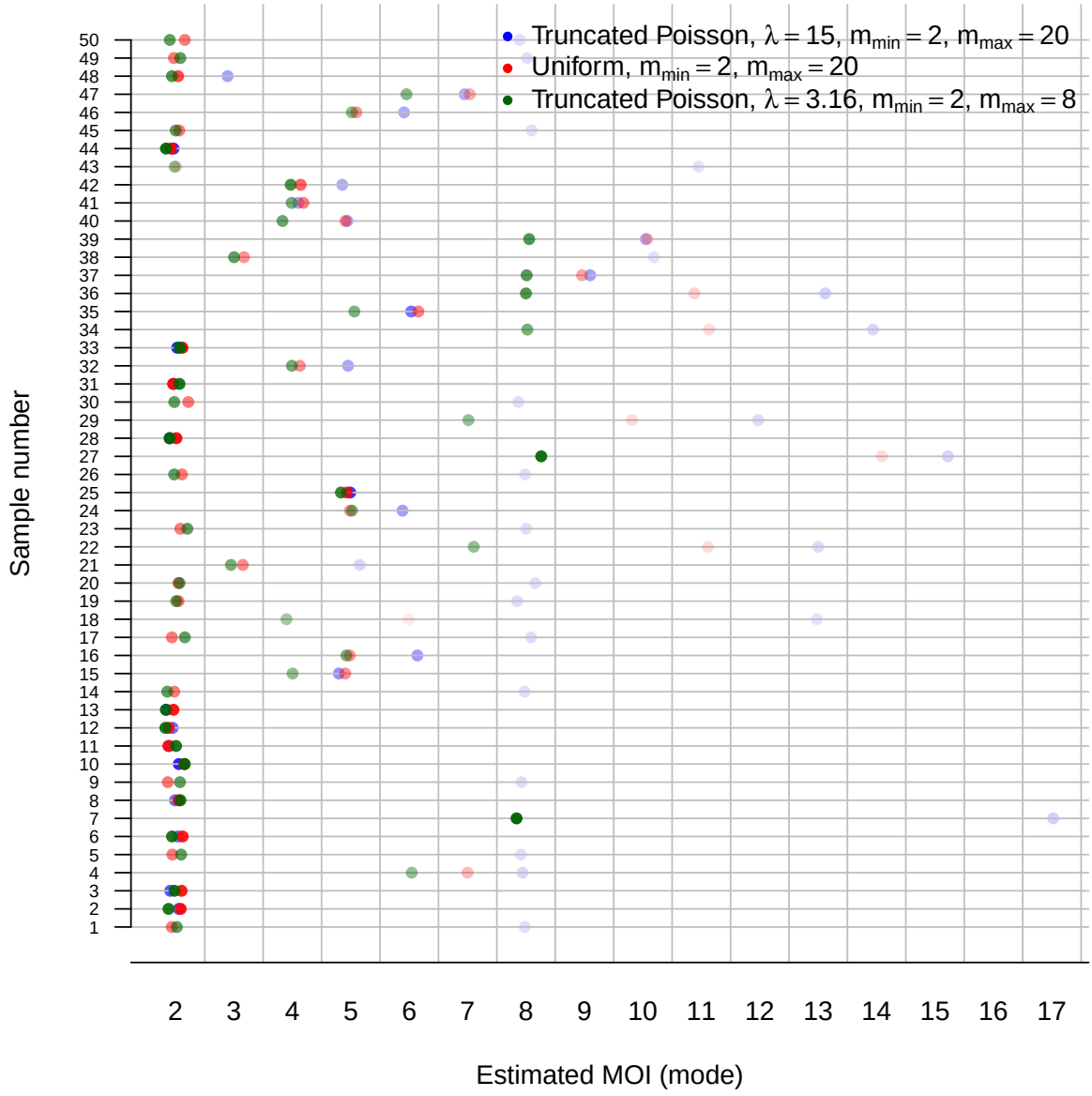


Figure 6.12: Ghanaian MOI estimates. For each sample, the plot shows the posterior mode, colour coded by prior combination (see legend). Each model is fit by running five chains of the sampler for 25,000,000 iterations. The transparency of the points is inversely proportional to the posterior mass. The MOI mode estimates are jittered slightly for visibility.

6.4 Discussion

In this chapter we present a modification to the likelihood of our original model (chapter 3), to enable the analysis of short-read sequencing data. It is an exploratory chapter, with a view to developing methods for analysing anticipated *P. falciparum* sequencing data from samples collected in regions of high transmission, where multiclonal samples are ubiquitous. Formal convergence of the sampler is not achieved, but the log-likelihood and posterior estimates are stable, hence convergence is deemed sufficient for practical purposes. The modification to the likelihood has little impact upon the posterior density estimates of the haplotype frequencies, but does affect the posterior estimates of the MOIs. Under the modified model, we are able to demonstrate the added capacity for model checks, and garner potential insights into the haplotype frequencies and MOIs of samples collected in Ghana. These findings are explored in more detail below.

Formal convergence (based on visual inspection of the log-posterior trace plots and PSRFs of the frequency estimates) proves prohibitively slow for the sampler run both on simulated data with three or more SNPs, and on the Ghanaian data under the model with prior support for MOIs less than or equal to eight. Nevertheless, the log-likelihood trace plots are stable throughout, as are the results for the Ghanaian data (figures 6.6 and 6.7). It seems that the sampler finds regions of the parameter space that are equal in terms of the likelihood, but not when the prior is factored in. A more sophisticated sampler is needed to understand and improve formal convergence. The existing sampler is considered adequate for the exploratory investigation.

The likelihood modification has little impact upon the posterior density estimates of the haplotype frequencies, suggesting there is little information to gain from the read counts over the prevalence data, and that the haplotype frequency estimates are robust to the likelihood modification. It is worth noting that information regarding multiple variants on the same read (for example, numerous reads overlap both nSNPs at codons 51 and 59 in *dhfr* in sample 18

— see figure 6.1) is ignored. Assuming each count is derived from an independent read is computationally convenient, but neglects valuable information in the data. As longer reads become more feasible with advances in sequencing technologies, exploiting such information will become paramount [185].

Despite no striking differences, density estimates under the modified model are more precise (figure 6.4). Under the binomial likelihood (equation (6.2)) one would expect a gain in precision as the total read count per position (z_{ij}) increases. Increased coverage does not necessarily concord with increased confidence in the biological makeup of the infection, however, since variation in the number of reads might represent an artefact of the data generating process. A natural extension to the model framework (figure 6.3) would be to build a probability model over z_{ij}, incorporating errors in the data generating process. A probability model over z_{ij} would also enable the analysis of datasets with missing z_{ij} (although, this is not a problem in our current application). A more realistic model would also incorporate errors in y_{ij} (see, for instance, the phylogenetic model proposed by O'Brien *et al.* [185], which is described in section 2.2.2).

In contrast to the frequency estimates, the impact of the likelihood modification upon the MOI estimates is notable. There appears to be more information in short-read sequencing data compared with prevalence data to support posterior MOI estimates (compare figures 6.5a and 6.5b). Because of the paucity of information in the prevalence data, the MOI is treated as a nuisance parameter in previous chapters. Nonetheless, the MOI is itself the subject of much research. It is an important measure in the study of population structure [102, 19], and has been associated with transmission intensity [183, 121, 221, 225, 258], age [183, 121, 230, 189, 157, 91] and clinical status [23, 79, 1, 209, 119, 112, 30]. Consequently, there have been numerous efforts to statistically estimate the MOI [102, 129, 219, 17, 223, 84, 186] (a number of which are integrally linked to haplotype frequency estimation [102, 129, 17]). Most recently, a model by Galinsky *et al.* [84] set out to estimate the MOI from prevalence data derived from a panel

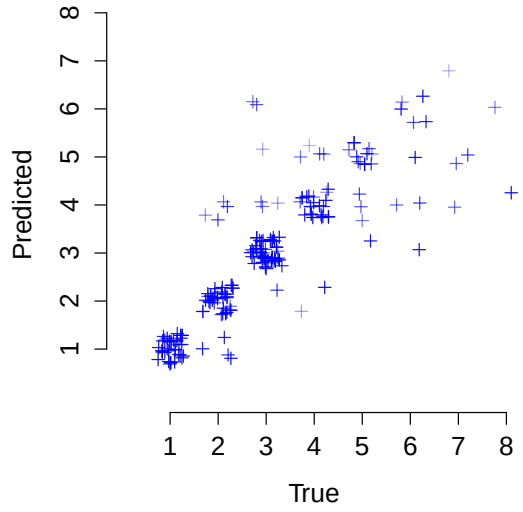


Figure 6.13: MOI estimates based on simulated quintuple-SNP data analysed under the modified model. The plot shows simulated 'True' MOIs (horizontal axis) versus posterior modal estimates (vertical axis). Results for three datasets, each with five SNPs and $i = 1, \ldots, 100$ samples are depicted, hence 300 MOI estimates in total. The MOIs are jittered to aid visibility. The transparency of the points is inversely proportional to the posterior mass.

of 96 independent SNPs. The results presented here compare favourably with those reported by Galinsky *et al.* (compare figure 6.13 with figure 2 in [84]), noting that the results reported here are based on simulated read data from only five nSNPs associated with drug resistance. Our results suggest preferential estimates could be obtained if MOI estimation models are adapted to capture information in short-read sequencing data.

Expanding upon the discussion of the MOI, we turn to the analysis of the Ghanaian data. Similar to the analyses based on simulated data, the analyses of the Ghanaian data under models with prior support for MOIs of eight or less do not formally converge, but given the stability of the log-likelihood trace plots, frequency estimates and MOI estimates, convergence is deemed sufficient for practical purposes. The model with the support for high MOIs afforded the best fit to the data in terms of deviance, in support of high MOI estimates reported in the literature [81], as well as the high level of genetic diversity reported in [8]. In light of the age stratified results reported in [81], it would be interesting to see if the Ghanaian samples with the highest

MOI estimates correspond with the age brackets with the highest inferred MOI in [81] (5–9 years and $>$ 60 years) and with the F_{ws} scores reported in [8].

The analyses of the Ghanaian data also demonstrate the added potential for model checks under the binomial likelihood (equation (6.2)) compared with the likelihood under the model designed to analyse prevalence data (equation (3.6)). We base model checks on deviance, however deviance is not the only model check [88], hence additional checks to corroborate these results would be valuable. The frequency estimates are sensitive to λ and $m_{\max}$ (figure 6.11). The genotype frequency estimates based on our model framework could be anywhere in the range depicted in figure 6.11, although model checks based on deviance favour the results depicted in blue. In any case, the conclusion remains: of the $2^5 = 32$ possible genotypes, there is evidence in the data to support 17, and of those 17, the frequency of the quadruple genotype **IRNG**K dominates. There is no evidence in this dataset to support the quintuple mutant genotype **IRNGE**.

The aim of this chapter is to modify the framework of the original model with a view to analysing anticipated sequencing data. More specifically, the likelihood of the original model is modified to accommodate short-read Illumina sequencing data. Formal convergence is imperfect, but deemed sufficient for practical purposes. Results based on simulated data imply that, compared to prevalence data, there is little to gain from short-read sequencing with regards to frequency estimation, but some information to inform MOI estimates. Using the model, we are able to make interesting insights regarding the Ghanaian data, including the estimation of sample-wise MOIs, previously deemed to be too challenging by conventional PCR methods [8]. Much work remains to develop the model beyond the exploratory stage, however, the first step being to investigate the barriers to formal convergence. One could also build complexity (base call and alignment errors, for example) into the model to make it more realistic and to use all available information.

To conclude, based on the results of this exploratory study, we believe the analysis of

short-read sequencing data summarised as prevalence data (see [2]) using existing methods designed for prevalence data [39, 102, 224, 129, 95, 276, 125, 223, 248] is apt for frequency estimation. On the contrary, our results suggest models designed to estimate MOIs should exploit short-read sequencing data where available, yielding more accurate MOI estimates, with important implications for our understanding of the complex dynamics of multiclonal infections.

Chapter 7

Conclusion

7.1 Thesis overview

In this thesis the development, application, extension and modification of a model designed for *P. falciparum* allele and multi-SNP haplotype and frequency estimation is documented, with a view to harnessing the full potential of data for the genetic surveillance of antimalarial resistance. More precisely, in chapter 2, existing statistical methods are reviewed. To the best of our knowledge, this is the first exhaustive review of statistical methods for *P. falciparum* allele and multi-SNP haplotype and frequency estimation. In chapter 3 the development of a model that complements and builds upon the existing methods is documented. Its capacity to generate accurate and precise frequency estimates is demonstrated, and assumptions and limitations discussed. To the best of our knowledge, this is the first model, constructed within a Bayesian framework, capable of generating multi-SNP haplotype and genotype frequency estimates, using all available data and without reliance upon experimentally-derived MOI estimates. In chapter 4, the utility of the model is illustrated, revealing informative trends in haplotype frequencies, which are consistent with national policy in Uganda. To the best of our knowledge, this is the first study of haplotype frequency trends in Uganda. In chapter 5 the model documented in chapter 3 is extended to investigate inter-child variation. Potential

evidence of small and selective variation was identified. To the best of knowledge, this is the first study to investigate inter-child variation within the field of statistical estimation of *P. falciparum* allele and haplotype frequencies for the genetic surveillance of antimalarial resistance. In chapter 6 the model introduced in chapter 3 is modified to accommodate short-read sequencing. The modified model is applied to data from Ghana, yielding insight into within-host diversity and anti-folate resistance. The modified model marks a first step towards the estimation of *P. falciparum* multi-SNP haplotype and genotype frequencies for routine genetic surveillance of antimalarial resistance using anticipated short-read sequencing data.

7.2 Future work

Ultimately, it is hoped that the model documented in chapter 3 will provide a practical tool for monitoring antimalarial resistance, and that the proposed framework (the versatility of which is demonstrated in chapters 5 and 6) will provide a foundation for further development. With a view to providing a practical tool for the malaria community, we hope to integrate the model into a user-friendly, open-source program, using the online platform `Shiny` by RStudio. Regarding further development, there are many attributes of the model and its proposed framework that could be improved. Those that are believed to be top-priority are outlined below.

Under the proposed model perfect detectability of minority clones and correct SNP identification is assumed. More realistic assumptions are incorporated into alternative models of *P. falciparum* allele, haplotype and genotype frequencies [95, 276]. More precisely, Hastings *et al.* model imperfect detectability using an indicator variable and a user-defined detectability threshold [95], while Wigger *et al.* model SNP miscalls, by incorporating error-ridden haplotype assignments into the hierarchy of the Bayesian framework [276]. Following Wigger *et al.*, the hierarchical framework could be extended to accommodate undetected minority clones and error-ridden haplotypes, akin to what is done in chapter 5 to investigate inter-child variation. The addition of further hierarchy will likely exhaust the computational feasibility of the current

sampler, however.

The computational feasibility of the inclusion of error-ridden haplotypes in the method of Wigger *et al.* is partly because the method does not capture uncertainty in MOI estimates. Instead it relies on experimentally-derived MOI estimates, which are considered fixed observations. Adaptation of our model to allow the incorporation of experimentally-derived sample-wise MOI estimates is an important consideration. However, in contrast to the aforesaid methods, which regard the MOI as a fixed quantity [95, 276], any extension of our model would preserve the current treatment of MOIs as random variables, perhaps using experimentally-derived sample-wise MOI estimates to inform sample-wise MOI prior distributions.

In chapter 4, allele and haplotype frequencies were estimated using the model documented in chapter 3, then regressed onto correlates of drug pressure and type using the meta-analysis approach proposed by Lunn et al. [136]. This approach is a two stage approximation of a joint model that both estimates frequencies and performs regression. As indicated in chapter 4, a joint model would potentially be preferable but complex, requiring prior distributions over matrices. Given our experience of fitting the extended and modified models documented in chapters 5 and 6, respectively, a more sophisticated sampler would likely be required to fit the joint model.

As discussed in section 5.6, more advanced samplers include parallel tempering [146], Hamiltonian Monte Carlo [67] (also see [35]), and variations thereof [103]. Parallel tempering works by switching between the target chain and parallel chains with diminished modes. Given the spiky posteriors reported in chapter 5, parallel tempering might be the sensible next option to explore. More investigation needs to be done to better understand the behaviour of the sampler documented in chapter 6 before an appropriate sampler can be identified for the modified model.

The model modification documented in chapter 6 describes an exploratory investigation into the adaptation of the original model with a view to analysing anticipated sequencing

data. Much could be done to improve the model, including the incorporation of miscall error rates as described above, as well as specific characteristics pertaining to short-read sequencing data, such as the information captured on reads spanning multiple SNPs. One could also incorporate mutations rates, borrowing from related fields, such as viral quasi-species spectrum reconstruction. Doing so would most certainly call for more a sophisticated implementation, as described above.

7.3 Closing remark

The identification of markers of artemisinin resistance ahead of widespread clinical failure marks a 'golden opportunity' for evidence-based policy using real-time genetic surveillance [217]. Markers of artemisinin resistance have thus far been documented in regions of low transmission. In high transmission settings, analyses of markers of resistance are comparatively complex, due to the abundance of multiclonal infections. Large amounts of valuable data have either been squandered or not used to their full potential in the past. It is imperative that the analytical toolbox is ready to avert a similar scenario if (or when) artemisinin resistance spreads to high transmission settings. The development of statistical methods that generate reliable and comparable frequency estimates provides a means to harness the full potential of current and prospective markers of antimalarial resistance. Here we present one such method, which builds upon existing methods and will hopefully provide a foundation for further development.

References

[1] I. E. A-Elbasit, G. ElGhazali, T. M. E. A-Elgadir, A. A. Hamad, H. A. Babiker, M. I. Elbashir, and H. A. Giha. Allelic polymorphism of MSP2 gene in severe P. falciparum malaria in an area of low and seasonal transmission. *Parasitology Research*, 102:29–34, 2007.

[2] A. O. Achieng, P. Muiruri, L. A. Ingasia, B. H. Opot, D. W. Juma, R. Yeda, B. S. Ngalah, B. R. Ogutu, B. Andagalu, H. M. Akala, and E. Kamau. Temporal trends in prevalence of Plasmodium falciparum molecular markers selected for by artemether–lumefantrine treatment in pre-ACT and post-ACT parasites in western Kenya. *International Journal for Parasitology: Drugs and Drug Resistance*, 5:92–99, 2015.

[3] D. Aguiar, W. S. W. Wong, and S. Istrail. Tumor haplotype assembly algorithms for cancer genomics. *Pacific Symposium on Biocomputing*, 2014.

[4] J. Albert. *Bayesian Computation with R*. Springer, 2 edition, 2009.

[5] M. Alifrangis, M. M. Lemnge, A. M. Rønn, M. D. Segeja, S. M. Magesa, I. F. Khalil, and I. C. Bygbjerg. Increasing prevalence of wildtypes in the dihydrofolate reductase gene of Plasmodium falciparum in an area with high levels of sulfadoxine/pyrimethamine resistance after introduction of treated bed nets. *The American journal of tropical medicine and hygiene*, 69(3):238–43, 2003.

[6] M. Alifrangis, S. Enosse, R. Pearce, C. Drakeley, C. Roper, I. F. Khalil, W. M. Nkya, A. M. Rønn, T. G. Theander, and I. C. Bygbjerg. A simple, high-throughput method to detect Plasmodium falciparum single nucleotide polymorphisms in the dihydrofolate reductase, dihydropteroate synthase, and P. falciparum chloroquine resistance transporter genes using polymerase chain reaction- and enzy. *The American journal of tropical medicine and hygiene*, 72(2):155–62, 2005.

[7] C. Amaratunga, S. Sreng, S. Suon, E. S. Phelps, K. Stepniewska, P. Lim, C. Zhou, S. Mao, J. M. Anderson, N. Lindegardh, H. Jiang, J. Song, X. Z. Su, N. J. White, A. M. Dondorp, T. J. C. Anderson, M. P. Fay, J. Mu, S. Duong, and R. M. Fairhurst. Artemisinin-resistant Plasmodium falciparum in Pursat province, western Cambodia: A parasite clearance rate study. *The Lancet Infectious Diseases*, 12:851–858, 2012.

[8] L. N. Amenga-Etego. *Plasmodium falciparum population genetics in northern Ghana*. PhD thesis, University of Oxford, 2012.

[9] T. J. C. Anderson, X. Z. Su, M. Bockarie, M. Lagog, and K. P. Day. Twelve microsatellite markers for characterization of Plasmodium falciparum from finger-prick blood samples. *Parasitology*, 119(02):113–25, 1999.

[10] T. J. Anderson, B. Haubold, J. T. Williams, J. G. Estrada-Franco, L. Richardson, R. Mollinedo, M. Bockarie, J. Mokili, S. Mharakurwa, N. French, J. Whitworth, I. D. Velez, A. H. Brockman, F. Nosten, M. U. Ferreira, and K. P. Day. Microsatellite markers reveal a spectrum of population structures in the malaria parasite Plasmodium falciparum. *Molecular biology and evolution*, 17(10):1467–82, 2000.

[11] T. J. C. Anderson, S. Nair, D. Sudimack, J. T. Williams, M. Mayxay, P. N. Newton, J.-P. Guthmann, F. M. Smithuis, T. Tinh Hien, I. V. F. van den Broek, N. J. White, and F. Nosten. Geographical distribution of selected and putatively neutral SNPs in Southeast Asian malaria parasites. *Molecular biology and evolution*, 22(12):2362–74, 2005.

[12] T. J. C. Anderson, J. Patel, and M. T. Ferdig. Gene copy number and malaria biology. *Trends in parasitology*, 25(7):336–343, 2009.

[13] F. Ariey, T. Fandeur, R. Durand, M. Randrianarivelojosia, R. Jambou, E. Legrand, M. T. Ekala, C. Bouchier, S. Cojean, J. B. Duchemin, V. Robert, J. Le Bras, and O. Mercereau-Puijalon. Invasion of Africa by a single pfcrt allele of South East Asian type. *Malaria Journal*, 5:34–39, 2006.

[14] F. Ariey, B. Witkowski, C. Amaratunga, J. Beghain, A.-C. Langlois, N. Khim, S. Kim, V. Duru, C. Bouchier, L. Ma, P. Lim, R. Leang, S. Duong, S. Sreng, S. Suon, C. M. Chuor, D. M. Bout, S. Ménard, W. O. Rogers, B. Genton, T. Fandeur, O. Miotto, P. Ringwald, J. Le Bras, A. Berry, J.-C. Barale, R. M. Fairhurst, F. Benoit-Vical, O. Mercereau-Puijalon, and D. Ménard. A molecular marker of artemisinin-resistant Plasmodium falciparum malaria. *Nature*, 505:50–55, 2014.

[15] E. Arinaitwe, T. G. Sandison, H. Wanzira, A. Kakuru, J. Homsy, J. Kalamya, M. R. Kamya, N. Vora, B. Greenhouse, P. J. Rosenthal, J. Tappero, and G. Dorsey. Artemether-lumefantrine versus dihydroartemisinin-piperaquine for falciparum malaria: a longitudinal, randomized trial in young Ugandan children. *Clinical infectious diseases : an official publication of the Infectious Diseases Society of America*, 49:1629–37, 2009.

[16] E. A. Ashley, M. Dhorda, R. M. Fairhurst, C. Amaratunga, P. Lim, S. Suon, S. Sreng, J. M. Anderson, S. Mao, B. Sam, C. Sopha, C. M. Chuor, C. Nguon, S. Sovannaroth, S. Pukrittayakamee, P. Jittamala, K. Chotivanich, K. Chutasmit, C. Suchatsoonthorn, R. Runcharoen, T. T. Hien, N. T. Thuy-Nhien, N. V. Thanh, N. H. Phu, Y. Htut, K.-T. Han, K. H. Aye, O. A. Mokuolu, R. R. Olaosebikan, O. O. Folaranmi, M. Mayxay, M. Khanthavong, B. Hongvanthong, P. N. Newton, M. A. Onyamboko, C. I. Fanello, A. K. Tshefu, N. Mishra, N. Valecha, A. P. Phyo, F. Nosten, P. Yi, R. Tripura, S. Borrmann, M. Bashraheil, J. Peshu, M. A. Faiz, A. Ghose, M. A. Hossain, R. Samad, M. R. Rahman, M. M. Hasan, A. Islam, O. Miotto, R. Amato, B. MacInnis, J. Stalker, D. P. Kwiatkowski, Z. Bozdech, A. Jeeyapant, P. Y. Cheah, T. Sakulthaew, J. Chalk, B. Intharabut, K. Silamut, S. J. Lee, B. Vihokhern, C. Kunasol, M. Imwong, J. Tarning, W. J. Taylor, S. Yeung, C. J. Woodrow, J. A. Flegg, D. Das, J. Smith, M. Venkatesan, C. V. Plowe, K. Stepniewska, P. J. Guerin, A. M. Dondorp, N. P. Day, and N. J. White. Spread of Artemisinin Resistance in Plasmodium falciparum Malaria. *New England Journal of Medicine*, 371:411–423, 2014.

[17] S. A. Assefa, M. D. Preston, S. Campino, H. Ocholla, C. J. Sutherland, and T. G. Clark. estMOI: Estimating multiplicity of infection using parasite deep sequencing data. *Bioinformatics*, 30(9):1292–1294, 2014.

[18] I. Astrovskaya, N. Manusco, B. Tork, S. Mangul, A. Artyomenko, P. Skums, L. Ganova-Raeva, I. Mandoiu, and A. Zelikovsky. Inferring viral quasispecies spectra from shortgun and aplicon next-generation sequencing reads. In M. S. Poptsova, editor, *Genome analysis: current procedures and applications*, chapter 12. Caister Academic Press, 2014.

[19] S. Auburn, S. Campino, O. Miotto, A. A. Djimde, I. Zongo, M. Manske, G. Maslen, V. Mangano, D. Alcock, B. MacInnis, K. A. Rockett, T. G. Clark, O. K. Doumbo, J. B. Ouédraogo, and D. P. Kwiatkowski. Characterization of within-host Plasmodium falciparum diversity using next-generation sequence data. *PloS one*, 7(2):e32891, 2012.

[20] C. Aurrecoechea, J. Brestelli, B. P. Brunk, J. Dommer, S. Fischer, B. Gajria, X. Gao, A. Gingle, G. Grant, O. S. Harb, M. Heiges, F. Innamorato, J. Iodice, J. C. Kissinger, E. Kraemer, W. Li, J. A. Miller, V. Nayak, C. Pennington, D. F. Pinney, D. S. Roos, C. Ross, C. J. Stoeckert Jr, C. Treatman, and H. Wang. PlasmoDB: a functional genomic database for malaria parasites. *Nucleic acids research*, 37(suppl 1), 2008.

[21] J. K. Baird. Resurgent malaria at the millennium: control strategies in crisis. *Drugs*, 59(4):719–43, 2000.

[22] V. Baraka, D. S. Ishengoma, F. Fransis, D. T. R. Minja, R. A. Madebe, D. Ngatunga, and J.-P. Van Geertruyden. High-level Plasmodium falciparum sulfadoxine-pyrimethamine resistance with the concomitant occurrence of septuple haplotype in Tanzania. *Malaria Journal*, 14:439, 2015.

[23] H. P. Beck, I. Felger, W. Huber, S. Steiger, T. Smith, N. Weiss, P. Alonso, and M. Tanner. Analysis of multiple Plasmodium falciparum infections in Tanzanian children during the phase III trial of the malaria vaccine SPf66. *The Journal of infectious diseases*, 175:921–6, 1997.

[24] M. S. Beier, I. K. Schwartz, J. C. Beier, P. V. Perkins, F. Onyango, J. K. Koros, G. H. Campbell, P. M. Andrysiak, and A. D. Brandling-Bennett. Identification of malaria species by ELISA in sporozoite and oocyst infected Anopheles from western Kenya. *The American Journal of Tropical Medicine and Hygiene*, 39(4):323–327, 1988.

[25] O. Bembom. seqLogo: Sequence logos for DNA sequence alignments. *R package version 1.36.0*.

[26] E. Berger, D. Yorukoglu, J. Peng, and B. Berger. HapTree: a novel Bayesian framework for single individual polyplotyping using NGS data. *PLoS computational biology*, 10(3):e1003502, 2014.

[27] S. Bhatt, D. J. Weiss, E. Cameron, D. Bisanzio, B. Mappin, U. Dalrymple, K. E. Battle, C. L. Moyes, A. Henry, M. A. Penny, T. A. Smith, A. Bennett, J. Yukich, T. P. Eisele, P. A. Eckhoff, E. A. Wenger, O. Brie, J. T. Griffin, C. A. Fergus, M. Lynch, F. Lindgren, J. M. Cohen, C. L. J. Murray, D. L. Smith, S. I. Hay, R. E. Cibulskis, and P. W. Gething. The effect of malaria control on Plasmodium falciparum in Africa between 2000 and 2015. *Nature*, 526, 2015.

[28] Y. L. Boo, H. T. Lim, P. W. Chin, S. Y. Lim, and F. K. Hoo. A case of severe Plasmodium knowlesi in a splenectomized patient. *Parasitology international*, 65(1):55–57, 2015.

[29] S. Borrmann, J. Straimer, L. Mwai, A. Abdi, A. Rippert, J. Okombo, S. Muriithi, P. Sasi, M. M. Kortok, B. Lowe, S. Campino, S. Assefa, S. Auburn, M. Manske, G. Maslen, N. Peshu, D. P. Kwiatkowski, K. Marsh, A. Nzila, and T. G. Clark. Genome-wide screen identifies new candidate genes associated with artemisinin susceptibility in Plasmodium falciparum in Kenya. *Scientific reports*, 3:3318, 2013.

[30] M. K. Bouyou-Akotet, N. P. M'Bondoukwé, and D. P. Mawili-Mboumba. Genetic polymorphism of merozoite surface protein-1 in Plasmodium falciparum isolates from patients with mild to severe malaria in Libreville, Gabon. *Parasite*, 22(12), 2015.

[31] G. E. P. Box and N. R. Draper. *Empirical Model-Building and Response Surfaces*. Wiley New York, 1987.

[32] B. J. Brabin. An analysis of malaria in pregnancy in Africa. *Bulletin of the World Health Organization*, 61(6):1005–1016, 1983.

[33] M. T. Bretscher, F. Valsangiacomo, S. Owusu-Agyei, M. A. Penny, I. Felger, and T. Smith. Detectability of Plasmodium falciparum clones. *Malaria Journal*, 9:234, 2010.

[34] D. R. Brooks, P. Wang, M. Read, W. M. Watkins, P. F. Sims, and J. E. Hyde. Sequence variation of the hydroxymethyldihydropterin pyrophosphokinase: dihydropteroate synthase gene in lines of the human malaria parasite, Plasmodium falciparum, with differing resistance to sulfadoxine. *European journal of biochemistry*, 224(2):397–405, 1994.

[35] S. Brooks, A. Gelman, G. Jones, and X. Meng, editors. *Handbook of Markov Chain Monte Carlo*. CRC press, 2011.

[36] S. R. Browning and B. L. Browning. Haplotype phasing: existing methods and new developments. *Nature reviews. Genetics*, 12(10):703–714, 2011.

[37] S. Campino, S. Auburn, K. Kivinen, I. Zongo, J.-B. Ouedraogo, V. Mangano, A. Djimde, O. K. Doumbo, S. M. Kiara, A. Nzila, S. Borrmann, K. Marsh, P. Michon, I. Mueller, P. Siba, H. Jiang, X.-Z. Su, C. Amaratunga, D. Socheat, R. M. Fairhurst, M. Imwong, T. Anderson, F. Nosten, N. J. White, R. Gwilliam, P. Deloukas, B. MacInnis, C. I. Newbold, K. Rockett, T. G. Clark, and D. P. Kwiatkowski. Population genetic analysis of Plasmodium falciparum parasites using a customized Illumina GoldenGate genotyping assay. *PLoS ONE*, 6(6):e20251, 2011.

[38] V. I. Carrara, J. Zwang, E. A. Ashley, R. N. Price, K. Stepniewska, M. Barends, A. Brockman, T. Anderson, R. McGready, L. Phaiphun, S. Proux, M. van Vugt, R. Hutagalung, K. M. Lwin, A. P. Phyo, P. Preechapornkul, M. Imwong, S. Pukrittayakamee, P. Singhasivanon, N. J. White, and F. Nosten. Changes in the treatment responses to artesunate-mefloquine on the northwestern border of Thailand during 13 years of continuous deployment. *PloS one*, 4(2):e4551, 2009.

[39] R. Carter and I. A. Mcgregor. Enzyme Variation in Plasmodium Falciparum in the Gambia. *Trans R Soc Trop Med Hyg*, 67(6):830–837, 1973.

[40] R. Carter and K. N. Mendis. Evolutionary and historical aspects of the burden of malaria. *Clinical microbiology reviews*, 15(4):564–594, 2002.

[41] L. K. Certain and C. H. Sibley. Plasmodium falciparum: a novel method for analyzing haplotypes in mixed infections. *Experimental parasitology*, 115(3):233–241, 2007.

[42] S. Chaorattanakawee, D. L. Saunders, D. Sea, N. Chanarat, K. Yingyuen, S. Sundrakes, P. Saingam, N. Buathong, S. Sriwichai, S. Chann, Y. Se, Y. Yom, T. K. Heng, N. Kong, W. Kuntawunginn, K. Tangthongchaiwiriya, C. Jacob, S. Takala-Harrison, C. Plowe, J. T. Lin, C. M. Chuor, S. Prom, S. D. Tyner, P. Gosi, P. Teja-Isavadharm, C. Lon, and C. A. Lanteri. Ex vivo drug susceptibility testing and molecular profiling of clinical Plasmodium falciparum isolates from Cambodia from 2008 to 2013 suggest emerging piperaquine resistance. *Antimicrobial Agents and Chemotherapy*, 59(8):4631–4643, 2015.

[43] I. H. Cheeseman, B. A. Miller, S. Nair, S. Nkhoma, A. Tan, J. C. Tan, S. Al Saai, A. P. Phyo, C. L. Moo, K. M. Lwin, R. McGready, E. Ashley, M. Imwong, K. Stepniewska, P. Yi, A. M. Dondorp, M. Mayxay, P. N. Newton, N. J. White, F. Nosten, M. T. Ferdig, and T. J. C. Anderson. A major genome region underlying artemisinin resistance in malaria. *Science*, 336:79–82, 2012.

[44] I. H. Cheeseman, M. McDew-White, A. P. Phyo, K. Sriprawat, F. Nosten, and T. J. Anderson. Pooled sequencing and rare variant association tests for identifying the determinants of emerging drug resistance in malaria parasites. *Mol Biol Evol*, 32(4):1080–1090, 2015.

[45] S. J. Cheesman, J. C. de Roode, A. F. Read, and R. Carter. Real-time quantitative PCR for analysis of genetically mixed infections of malaria parasites: technique validation and applications. *Molecular and Biochemical Parasitology*, 131:83–91, 2003.

[46] L. M. Childs and C. O. Buckee. Dissecting the determinants of malaria chronicity: why within-host models struggle to reproduce infection dynamics. *Journal of the Royal Society, Interface*, 12(104), 2015.

[47] W. Chin, P. G. Contacos, G. R. Coatney, and H. K. King. The evaluation of sulfonamides, alone or in combination with pyrimethamine, in the treatment of multi-resistant falciparum malaria. *American Journal of Tropical Medicine and Hygiene*, 15(6), 1966.

[48] R. H. B. Christensen. ordinal—Regression Models for Ordinal Data. *R package version 2015.6-28. http://www.cran.r-project.org/package=ordinal/*, 2015.

[49] T. D. Clark, B. Greenhouse, D. Njama-Meya, B. Nzarubara, C. Maiteki-Sebuguzi, S. G. Staedke, E. Seto, M. R. Kamya, P. J. Rosenthal, and G. Dorsey. Factors determining the heterogeneity of malaria incidence in children in Kampala, Uganda. *The Journal of infectious diseases*, 198(3):393–400, 2008.

[50] D. Clyde and G. Shute. Resistance of Plasmodium falciparum in Tanganyika to pyrimethamine administered at weekly intervals. *Transactions of the Royal Society of Tropical Medicine and Hygiene*, 51(6):505–513, 1957.

[51] M. D. Conrad, N. Leclair, E. Arinaitwe, H. Wanzira, A. Kakuru, V. Bigira, M. Muhindo, M. R. Kamya, J. W. Tappero, B. Greenhouse, G. Dorsey, and P. J. Rosenthal. Comparative impacts over 5 years of artemisinin-based combination therapies on P.falciparum polymorphisms that modulate drug sensitivity in Ugandan children. *The Journal of infectious diseases*, 210(3):344–353, 2014.

[52] D. J. Conway, B. M. Greenwood, and J. S. McBride. The epidemiology of multiple-clone Plasmodium falciparum infections in Gambian patients. *Parasitology*, 130(01):1–5, 1991.

[53] J. F. Cortese, A. Caraballo, C. E. Contreras, and C. V. Plowe. Origin and dissemination of Plasmodium falciparum drug-resistance mutations in South America. *The Journal of infectious diseases*, 186(7):999–1006, 2002.

[54] A. F. Cowman, M. J. Morry, B. A. Biggs, G. A. Cross, and S. J. Foote. Amino acid changes linked to pyrimethamine resistance in the dihydrofolate reductase-thymidylate synthase gene of Plasmodium falciparum. *Proceedings of the National Academy of Sciences*, 85(23):9109–9113, 1988.

[55] S. L. Croft, S. Duparc, S. J. Arbe-Barnes, J. C. Craft, C.-S. Shin, L. Fleckenstein, I. Borghini-Fuhrer, and H.-J. Rim. Review of pyronaridine anti-malarial properties and product characteristics. *Malaria Journal*, 11(270), 2012.

[56] S. Dahlström, P. E. Ferreira, M. I. Veiga, N. Sedighi, L. Wiklund, A. Mårtensson, A. Färnert, C. Sisowath, L. Osório, H. Darban, B. Andersson, A. Kaneko, G. Conseil, A. Björkman, and J. P. Gil. Plasmodium falciparum multidrug resistance protein 1 and artemisinin-based combination therapy in Africa. *The Journal of infectious diseases*, 200(9):1456–64, 2009.

[57] R. Daniels, S. K. Volkman, D. A. Milner, N. Mahesh, D. E. Neafsey, D. J. Park, D. Rosen, E. Angelino, P. C. Sabeti, D. F. Wirth, and R. C. Wiegand. A general SNP-based molecular barcode for Plasmodium falciparum identification and tracking. *Malaria Journal*, 7(223), 2008.

[58] R. Daniels, E. J. Hamilton, K. Durfee, D. Ndiaye, D. F. Wirth, D. L. Hartl, and S. K. Volkman. Methods to Increase the Sensitivity of High Resolution Melting Single Nucleotide Polymorphism Genotyping in Malaria. *Journal of visualized experiments : JoVE*, (105):1–8, 2015.

[59] A. A. Dempster, N. N. Laird, and D. D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society Series B Methodological*, 39(1):1–38, 1977.

[60] Division of Parasitic Diseases and Malaria. Laboratory diagnosis of malaria: Plasmodium spp. Life Cycle of Plasmodium spp. *http://www.cdc.gov/dpdx/resources/pdf/benchAids/malaria/Parasitemia_and_LifeCycle.pdf*, 2013.

[61] A. Djimdé, O. K. Doumbo, J. F. Cortese, K. Kayentao, S. Doumbo, Y. Diourte, A. Dicko, X.-Z. Su, T. Nomura, D. A. Fidock, T. E. Wellems, and C. V. Plowe. A molecular marker for chloroquine-resistant falciparum malaria. *New England Journal of Medicine*, 344(4):257–263, 2001.

[62] A. A. Djimde, A. Dolo, A. Ouattara, S. Diakite, C. V. Plowe, and O. K. Doumbo. Molecular diagnosis of resistance to antimalarial drugs during epidemics and in war zones. *The Journal of infectious diseases*, 190(4):853–855, 2004.

[63] C. B. Do and S. Batzoglou. What is the expectation maximization algorithm? *Nature biotechnology*, 26(8):897–899, 2008.

[64] A. M. Dondorp, F. Nosten, P. Yi, D. Das, A. P. Phyo, J. Tarning, K. M. Lwin, F. Ariey, W. Hanpithakpong, S. J. Lee, Others, P. Ringwald, K. Silamut, M. Imwong, K. Chotivanich, P. Lim, T. Herdman, S. S. An, S. Yeung, P. Singhasivanon, N. P. J. Day, N. Lindegardh, D. Socheat, and N. J. White. Artemisinin resistance in Plasmodium falciparum malaria. *The New England journal of medicine*, 361(5):455–67, 2009.

[65] A. M. Dondorp, R. M. Fairhurst, L. Slutsker, J. R. MacArthur, J. G. Breman, P. J. Guerin, T. E. Wellems, P. Ringwald, R. D. Newman, and C. V. Plowe. The threat of artemisinin-resistant malaria. *New England Journal of Medicine*, 365(12):1073–1075, 2011.

[66] P. Druilhe, P. Daubersies, J. Patarapotikul, C. Gentil, L. Chene, T. Chongsuphajaisiddhi, S. Mellouk, and G. Langsley. A primary malarial infection is composed of a very wide range of genetically diverse but related parasites. *Journal of Clinical Investigation*, 101(9):2008–2016, 1998.

[67] S. Duane, A. Kennedy, B. J. Pendleton, and D. Roweth. Hybrid Monte Carlo. *Physics Letters B*, 195(2):216–222, 1987.

[68] R. T. Eastman, N. V. Dharia, E. A. Winzeler, and D. A. Fidock. Piperaquine resistance is associated with a copy number variation on chromosome 5 in drug-pressured Plasmodium falciparum parasites. *Antimicrobial Agents and Chemotherapy*, 55(8):3908–3916, 2011.

[69] A. Ecker, A. M. Lehane, J. Clain, and D. A. Fidock. PfCRT and its role in antimalarial drug resistance. *Trends in parasitology*, 28(11):504–514, 2012.

[70] A. A. Escalante, H. M. Grebert, S. C. Chaiyaroj, M. Magris, S. Biswas, B. L. Nahlen, and A. A. Lal. Polymorphism in the gene encoding the apical membrane antigen-1 (AMA-1) of Plasmodium falciparum. X. Asembo Bay Cohort Project. *Molecular and biochemical parasitology*, 113(2):279–87, 2001.

[71] C. A. Espinal, L. M. Uribe, A. Eslava, and M. E. Rodriguez. Resistencia del Plasmodium falciparum a la combinacion sulfa-primetamina. *Biomedica*, 1(4), 1981.

[72] Europe PMC Funders Group. Mitigating the threat of artemisinin resistance in Africa: improvement of drug-resistance surveillance and response systems. *Lancet Infectious Diseases*, 12(11):888–896, 2013.

[73] European Medicines Agency. First malaria vaccine receives positive scientific opinion from EMA. *Press release*, 2015.

[74] L. Excoffier and M. Slatkin. Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. *Mol. Biol. Evol.*, 12(5):921–927, 1995.

[75] N. Falk, N. Maire, W. Sama, S. Owusu-Agyei, T. Smith, H.-P. Beck, and I. Felger. Comparison of PCR-RFLP and Genescan-based genotyping for analyzing infection dynamics of Plasmodium falciparum. *The American journal of tropical medicine and hygiene*, 74(6):944–50, 2006.

[76] D. R. V. Färnert A, Arez AP, Babiker HA, Beck HP, Benito A, Björkman A, Bruce MC, Conway DJ, Day KP, Henning L, Mercereau-Puijalon O, Ranford-Cartwright LC, Rubio JM, Snounou G, Walliker D, Zwetyenga J. Genotyping multicentre of Plasmodium study

falciparum infections by PCR: a comparative multicentre study. *Transactions of the Royal Society of Tropical Medicine & Hygiene*, 95:225–232, 2001.

[77] A. Farnert, G. Snounou, I. Rooth, and A. Bjorkman. Daily dynamics of Plasmodium falciparum subpopulations in asymptomatic children in a holoendemic area. *American Journal of Tropical Medicine and Hygiene*, 56(5):538–547, 1997.

[78] I. Felger and H.-P. Beck. Genotyping of Plasmodium falciparum. PCR-RFLP analysis. *Methods in Molecular Medicin*, 72:117–129, 2002.

[79] I. Felger, T. Smith, D. Edoh, A. Kitua, P. Alonso, M. Tanner, and H. P. Beck. Multiple plasmodium falciparum infections in Tanzanian infants. *Transactions of the Royal Society of Tropical Medicine and Hygiene*, 93(suppl 1):29–34, 1999.

[80] I. Felger, A. Irion, S. Steiger, and H. P. Beck. Epidemiology of multiple Plasmodium falciparum infections. *Transactions of the Royal Society of Tropical Medicine & Hygiene*, 100, 1999.

[81] I. Felger, M. Maire, M. T. Bretscher, N. Falk, A. Tiaden, W. Sama, H.-P. P. Beck, S. Owusu-Agyei, and T. A. Smith. The Dynamics of Natural Plasmodium falciparum Infections. *PLoS ONE*, 7(9):e45542, 2012.

[82] D. A. Fidock, T. Nomura, A. K. Talley, R. A. Cooper, S. M. Dzekunov, M. T. Ferdig, L. M. Ursos, A. B. Sidhu, B. Naudé, K. W. Deitsch, X. Z. Su, J. C. Wootton, P. D. Roepe, and T. E. Wellems. Mutations in the P. falciparum digestive vacuole transmembrane protein PfCRT and evidence for their role in chloroquine resistance. *Molecular cell*, 6(4):861–71, 2000.

[83] D. Francis, S. L. Nsobya, A. Talisuna, A. Yeka, M. R. Kamya, R. Machekano, C. Dokomajilar, P. J. Rosenthal, and G. Dorsey. Geographic differences in antimalarial drug efficacy in Uganda are explained by differences in endemicity and not by known molecular markers of drug resistance. *The Journal of infectious diseases*, 193(7):978–86, 2006.

[84] K. Galinsky, C. Valim, A. Salmier, B. de Thoisy, L. Musset, E. Legrand, A. Faust, M. Baniecki, D. Ndiaye, R. F. Daniels, D. L. Hartl, P. C. Sabeti, D. F. Wirth, S. K. Volkman, and D. E. Neafsey. COIL: a methodology for evaluating malarial complexity of infection using likelihood from single nucleotide polymorphism data. *Malaria Journal*, 14(1):4, 2015.

[85] M. Gardner, N. Hall, E. Fung, O. White, M. Berriman, R. Hyman, J. Carlton, A. Pain, K. Nelson, S. Bowman, I. Paulsen, K. James, J. Eisen, K. Rutherford, S. Salzberg, A. Craig, S. Kyes, M. Chan, V. Nene, S. J. Shallom, B. Suh, J. Peterson, S. Angiuoli, M. Pertea, J. Allen, J. Selengut, D. Haft, M. W. Mather, A. B. Vaidya, D. M. A. Martin, A. H. Fairlamb, M. J. Fraunholz, D. S. Roos, S. A. Ralph, G. I. McFadden, L. M. Cummings, G. M. Subramanian, C. Mungall, J. C. Venter, D. J. Carucci, S. L. Hoffman, C. Newbold, R. W. Davis, C. M. Fraser, and B. Barrell. Genome sequence of the human malaria parasite Plasmodium falciparum. *Nature*, 419:498–511, 2002.

[86] W. Gatei, S. Kariuki, W. Hawley, F. ter Kuile, D. Terlouw, P. Phillips-Howard, B. Nahlen, J. Gimnig, K. Lindblade, E. Walker, M. Hamel, S. Crawford, J. Williamson, L. Slutsker, and Y. P. Shi. Effects of transmission reduction by insecticide-treated bed nets (ITNs) on parasite genetics population structure: I. The genetic diversity of Plasmodium falciparum parasites by microsatellite markers in western Kenya. *Malaria Journal*, 9(1):353, 2010.

[87] A. Gelman, A. Jakulin, M. G. Pittau, and Y.-S. Su. A weakly informative default prior distribution for logistic and other regression models. *The Annals of Applied Statistics*, 2(4):1360–1383, 2008.

[88] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin. *Bayesian Data Analysis*. Chapman & Hall/CRC Press, second edition, 2009.

[89] P. I. German and F. T. Aweeka. Clinical pharmacology of artemisinin-based combination therapies. *Clinical pharmacokinetics*, 47(2):91–102, 2008.

[90] W. R. Gilks, S. Richardson, and D. J. Spiegelhalter, editors. *Markov Chain Monte Carlo in Practice*. Chapman & Hall, first edition, 1996.

[91] A. Guerra-Neira, J. M. Rubio, J. R. Royo, J. C. Ortega, A. S. Auñón, P. B. Diaz, and A. B. Llanes. Plasmodium diversity in non-malaria individuals from the Bioko Island in Equatorial Guinea (West Central-Africa). *International journal of health geographics*, 5:27, 2006.

[92] V. Gupta, G. Dorsey, A. E. Hubbard, P. J. Rosenthal, and B. Greenhouse. Gel versus capillary electrophoresis genotyping for categorizing treatment outcomes in two anti-malarial trials in Uganda. *Malaria Journal*, 9:19, 2010.

[93] H. Haario, E. Saksman, and J. Tamminen. An adaptive Metropolis algorithm. *Bernoulli*, 7(2):223–242, 2001.

[94] J. A. Hasler, I. Johansson, and C. M. Masimirembwa. Inhibitory effects of antiparasitic drugs on cytochrome P450 2D6. *European journal of clinical pharmacology*, 48(1):35–38, 1995.

[95] I. M. Hastings and T. A. Smith. MalHaploFreq: A computer programme for estimating malaria haplotype frequencies from blood samples. *Malaria Journal*, 7(130), 2008.

[96] I. M. Hastings, C. Nsanzabana, and T. A. Smith. A comparison of methods to detect and quantify the markers of antimalarial drug resistance. *The American journal of tropical medicine and hygiene*, 83(3):489–95, 2010.

[97] I. M. Hastings. Malaria control and the evolution of drug resistance: an intriguing link. *Trends in parasitology*, 19(2):70–73, 2003.

[98] T. Havryliuk, P. Orjuela-Sánchez, and M. U. Ferreira. Plasmodium vivax: microsatellite analysis of multiple-clone infections. *Experimental parasitology*, 120(4):330–6, 2008.

[99] S. I. Hay, D. J. Rogers, J. F. Toomer, and R. W. Snow. Annual Plasmodium falciparum entomological inoculation rates (EIR) across Africa: literature survey, Internet access and review. *Transactions of the Royal Society of Tropical Medicine and Hygiene*, 94(2):113–127, 2011.

[100] U. Hess, P. M. Timmermans, and M. Jones. Combined chloroquine/Fansidar-resistant falciparum malaria appears in East Africa. *American Journal of Tropical Medicine and Hygiene*, 32(2):217–220, 1983.

[101] T. T. Hien, N. T. Thuy-Nhien, N. H. Phu, M. F. Boni, N. V. Thanh, N. T. Nha-Ca, L. H. Thai, C. Q. Thai, P. V. Toi, P. D. Thuan, L. T. Long, L. T. Dong, L. Merson, C. Dolecek, K. Stepniewska, P. Ringwald, N. J. White, J. Farrar, and M. Wolbers. In vivo susceptibility of Plasmodium falciparum to artesunate in Binh Phuoc Province, Vietnam. *Malaria Journal*, 11(355), 2012.

[102] W. G. Hill and H. A. Babiker. Estimation of numbers of malaria clones in blood samples. *Proceedings of the Royal Society of London B: Biological sciences*, 262(1365):249–57, 1995.

[103] M. Hoffman and A. Gelman. The no-U-turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, 15(2008):1–31, 2014.

[104] S. L. Hoffman, J. Vekemans, T. L. Richie, and P. E. Duffy. The March Toward Malaria Vaccines. *American Journal of Preventive Medicine*, 49(6):S319–S333, 2015.

[105] G. Holmgren, J. P. Gil, P. M. Ferreira, M. I. Veiga, C. O. Obonyo, and A. Björkman. Amodiaquine resistant Plasmodium falciparum malaria in vivo is associated with selection of pfcrt 76T and pfmdr1 86Y. *Infection, Genetics and Evolution*, 6(4):309–314, 2006.

[106] C. Hopkins Sibley, P. J. Guerin, and R. Pascal. Monitoring antimalarial resistance: launching a cooperative effort. *Trends in parasitology*, 26(5):221–224, 2010.

[107] F. Huang, S. Takala-Harrison, C. G. Jacob, H. Liu, X. Sun, H. Yang, M. M. Nyunt, M. Adams, S. Zhou, Z. Xia, P. Ringwald, M. D. Bustos, L. Tang, and C. V. Plower. A single mutation in K13 predominates in Southern China and is associated with delayed clearance of Plasmodium falciparum following artemisinin treatment. *Journal of Infectious Diseases*, 2015.

[108] E. S. Hurwitz, D. Johnson, and C. C. Campbell. Resistance of Plasmodium falciparum malaria to sulfadoxine-pyrimethamine ('Fansidar') in a refugee camp in Thailand. *The Lancet*, 317(8229):1068–1070, 1981.

[109] A. Iliadis, D. Anastassiou, and X. Wang. A sequential Monte Carlo framework for haplotype inference in CNV/SNP genotype data. *EURASIP journal on bioinformatics & systems biology*, 2014(1):7, 2014.

[110] Illumina. An Introduction to Next-Generation Sequencing Technology Table of Contents. Technical Report (Illumina, 2015.

[111] M. Imwong, S. Hanchana, B. Malleret, L. Rénia, N. P. J. Day, A. Dondorp, F. Nosten, G. Snounou, and N. J. White. High throughput ultra-sensitive molecular techniques to quantify low density malaria parasitaemias. *Journal of clinical microbiology*, 52(9):3303–3309, 2014.

[112] H. A. Ismail, U. Ribacke, L. Reiling, J. Normark, T. Egwang, F. Kironde, J. G. Beeson, M. Wahlgren, and K. E. M. Persson. Acquired antibodies to merozoite antigens in children from Uganda with uncomplicated or severe plasmodium falciparum malaria. *Clinical and Vaccine Immunology*, 20(8):1170–1180, 2013.

[113] J. J. Juliano, J. J. Kwiek, K. Cappell, V. Mwapasa, and S. R. Meshnick. Minority-variant pfcrt K76T mutations and chloroquine resistance, Malawi. *Emerging Infectious Diseases*, 13(6):872–827, 2007.

[114] J. J. Juliano, P. Trottman, V. Mwapasa, and R. S. Meshnick. Short Report: Detection of the Dihydrofolate Reductase–164L Mutation in Plasmodium falciparum Infections from Malawi by Heteroduplex Tracking Assay Jonathan. *American Journal of Tropical Medicine and Hygiene*, 78(6):892–894, 2008.

[115] J. J. Juliano, M. Randrianarivelojosia, B. Ramarosandratana, F. Ariey, V. Mwapasa, and S. R. Meshnick. Nonradioactive heteroduplex tracking assay for the detection of minority-variant chloroquine-resistant Plasmodium falciparum in Madagascar. *Malaria Journal*, 8(47), 2009.

[116] J. Kapisi, V. Bigira, T. Clark, S. Kinara, F. Mwangwa, J. Achan, M. Kamya, S. Soremekun, and G. Dorsey. Efficacy and safety of artemether-lumefantrine for the treatment of uncomplicated malaria in the setting of three different chemopreventive regimens. *Malaria Journal*, 14(1):53, 2015.

[117] D. Kessner, T. L. Turner, and J. Novembre. Maximum likelihood estimation of frequencies of known haplotypes from pooled sequence data. *Molecular biology and evolution*, 30(5):1145–58, 2013.

[118] W. C. Kiarie, L. Wangai, E. Agola, F. T. Kimani, and C. Hungu. Chloroquine sensitivity: diminished prevalence of chloroquine-resistant gene marker pfcrt-76 13 years after cessation of chloroquine use in Msambweni, Kenya. *Malaria Journal*, 14(328), 2015.

[119] M. S. Kiwuwa, U. Ribacke, K. Moll, J. Byarugaba, K. Lundblom, A. Färnert, K. Fred, and M. Wahlgren. Genetic diversity of Plasmodium falciparum infections in mild and severe malaria of children from Kampala, Uganda. *Parasitology Research*, 112(4):1691–1700, 2013.

[120] C. Koepfli, S. Schoepflin, M. Bretscher, E. Lin, B. Kiniboro, P. A. Zimmerman, P. Siba, T. A. Smith, I. Mueller, and I. Felger. How much remains undetected? Probability of molecular detection of human Plasmodia in the field. *PloS one*, 6(4):e19010, 2011.

[121] L. Konate, J. Zwetyenga, C. Rogier, E. Bischoff, D. Fontenille, A. Tall, A. Spiegel, J.-F. Trape, and O. Mercereau-Puijalon. Variation of Plasmodium falciparum mspl block 2 and msp2 allele prevalence and of infection complexity in two neighbouring Senegalese villages with different transmission conditions. *Transactions of the Royal Society of Tropical Medicine and Hygiene*, 93(suppl 1):S1/21–S1/28, 1999.

[122] D. J. Krogstad. Malaria as a Reemerging Disease. *Epidemiologic Reviews*, 18(1):77–89, 1996.

[123] J. G. Kublin, F. K. Dzinjalamala, D. D. Kamwendo, E. M. Malkin, J. F. Cortese, L. M. Martino, R. A. G. Mukadam, S. J. Rogerson, A. G. Lescano, M. E. Molyneux, P. A. Winstanley, P. Chimpeni, T. E. Taylor, and C. V. Plowe. Molecular markers for failure of sulfadoxine-pyrimethamine and chlorproguanil-dapsone treatment of Plasmodium falciparum malaria. *The Journal of infectious diseases*, 185(3):380–388, 2002.

[124] J. G. Kublin, J. F. Cortese, E. M. Njunju, R. A. G. Mukadam, J. J. Wirima, P. N. Kazembe, A. A. Djimdé, B. Kouriba, T. E. Taylor, and C. V. Plowe. Reemergence of chloroquine-sensitive Plasmodium falciparum malaria after cessation of chloroquine use in Malawi. *The Journal of infectious diseases*, 187(12):1870–5, 2003.

[125] C. K. Kum, D. Thorburn, G. Ghilagaber, P. Gil, and A. Björkman. On the effects of malaria treatment on parasite drug resistance–probability modelling of genotyped malaria infections. *The international journal of biostatistics*, 9(1):135–148, 2013.

[126] M. P. Kyaw, M. H. Nyunt, K. Chit, M. M. Aye, K. H. Aye, N. Lindegardh, J. Tarning, M. Imwong, C. G. Jacob, C. Rasmussen, J. Perin, P. Ringwald, and M. M. Nyunt. Reduced Susceptibility of Plasmodium falciparum to Artesunate in Southern Myanmar. *PloS one*, 8(3):e57689, 2013.

[127] T. Laver, J. Harrison, P. A. O'Neill, K. Moore, A. Farbos, K. Paszkiewicz, and D. J. Studholme. Assessing the performance of the Oxford Nanopore Technologies MinION. *Biomolecular Detection and Quantification*, 3:1–8, 2015.

[128] M. Lawrence, W. Huber, H. Pages, P. Aboyoun, M. Carlson, R. Gentleman, M. T. Morgan, and V. J. Carey. Software for Computing and Annotating Genomic Ranges. *PLoS Computational Biology*, 9(8):1–10, 2013.

[129] X. Li, A. S. Foulkes, R. M. Yucel, and S. M. Rich. An expectation maximization approach to estimate malaria haplotype frequencies in multiply infected children. *Statistical applications in genetics and molecular biology*, 6(1), 2007.

[130] X. Li, B. N. Thomas, S. M. Rich, D. Ecker, J. K. Tumwine, and A. S. Foulkes. Estimating and testing haplotype-trait associations in non-diploid populations. *Journal of the Royal Statistical Society. Series C, Applied statistics*, 58(5):663–678, 2009.

[131] X. Li. EM Estimation of Malaria Haplotype Probabilities from Multiply Infected Human Blood Samples. *https://cran.r-project.org/src/contrib/Archive/malaria.em/*, R package, 2012.

[132] J. Little, Roderick and B. Rubin, Donald. *Statistical Analysis with Missing Data*. John Wiley & Sons, second edition, 2014.

[133] S. Liu, J. Mu, H. Jiang, and X.-z. Su. Effects of Plasmodium falciparum mixed infections on in vitro antimalarial drug tests and genotyping. *The American journal of tropical medicine and hygiene*, 79(2):178–84, 2008.

[134] Q. Long, D. C. Jeffares, Q. Zhang, K. Ye, V. Nizhynska, Z. Ning, C. Tyler-Smith, and M. Nordborg. PoolHap: inferring haplotype frequencies from pooled samples by next generation sequencing. *PloS one*, 6(1):e15292, 2011.

[135] N. W. Lucchi, F. Komino, S. A. Okoth, I. Goldman, P. Onyona, R. E. Wiegand, E. Juma, Y. P. Shi, J. W. Barnwell, V. Udhayakumar, and S. Kariuki. In vitro and molecular surveillance for antimalarial drug resistance in Plasmodium falciparum parasites in western Kenya reveals sustained artemisinin sensitivity and increased chloroquine sensitivity. *Antimicrobial Agents and Chemotherapy*, 59(12):7540–7547, 2015.

[136] D. Lunn, J. Barrett, M. Sweeting, and S. Thompson. Fully Bayesian hierarchical modelling in two stages, with application to meta-analysis. *Journal of the Royal Statistical Society: Series C (Applied statistics)*, 62(4):551–572, 2013.

[137] G. Macdonald. The analysis of infection rates in diseases in which superinfection occurs. *Tropical diseases bulletin*, 47(10):907–15, 1950.

[138] O. Maïga-Ascofaré, J. Le Bras, R. Mazmouz, E. Renard, S. Falcão, E. Broussier, D. Bustos, M. Randrianarivelojosia, S. A. Omar, A. Aubouy, J.-F. Lepère, V. Jean-François, A. A. Djimdé, and J. Clain. Adaptive differentiation of Plasmodium falciparum populations inferred from single-nucleotide polymorphisms (SNPs) conferring drug resistance and from neutral SNPs. *The Journal of infectious diseases*, 202(7):1095–1103, 2010.

[139] Malaria Control Programme Ministry of Health. Uganda Malaria Control Strategic Plan. Technical report, 2005.

[140] A. Malisa, R. Pearce, B. Mutayoba, S. Abdullah, H. Mshinda, P. Kachur, P. Bloland, and C. Roper. Quantification of markers of antimalarial drug resistance from an area of high malaria transmission: Comparing frequency with prevalence. *African Journal of Biotechnology*, 11(69):13250–13260, 2012.

[141] M. Malmberg, B. Ngasala, P. E. Ferreira, E. Larsson, I. Jovel, A. Hjalmarsson, M. Petzold, Z. Premji, J. P. Gil, A. Björkman, and A. Mårtensson. Temporal trends of molecular markers associated with artemether-lumefantrine tolerance/resistance in Bagamoyo district, Tanzania. *Malaria Journal*, 12(1):103, 2013.

[142] M. Manske, O. Miotto, S. Campino, S. Auburn, J. Almagro-Garcia, G. Maslen, J. O'Brien, A. Djimde, O. Doumbo, I. Zongo, J.-B. Ouedraogo, P. Michon, I. Mueller, P. Siba, A. Nzila, S. Borrmann, S. M. Kiara, K. Marsh, H. Jiang, X.-Z. Su, C. Amaratunga, R. Fairhurst, D. Socheat, F. Nosten, M. Imwong, N. J. White, M. Sanders, E. Anastasi, D. Alcock, E. Drury, S. Oyola, M. A. Quail, D. J. Turner, V. Ruano-Rubio, D. Jyothi, L. Amenga-Etego, C. Hubbart, A. Jeffreys, K. Rowlands, C. Sutherland, C. Roper, V. Mangano, D. Modiano, J. C. Tan, M. T. Ferdig, A. Amambua-Ngwa, D. J. Conway, S. Takala-Harrison, C. V. Plowe, J. C. Rayner, K. A. Rockett, T. G. Clark, C. I. Newbold, M. Berriman, B. MacInnis, and D. P. Kwiatkowski. Analysis of Plasmodium falciparum diversity in natural infections by deep sequencing. *Nature*, 487:375–379, 2012.

[143] E. R. Mardis. Next-Generation DNA Sequencing Methods. *Annual Review of Genomics and Human Genetics*, 9(1):387–402, 2008.

[144] J. Marfurt, T. A. Smith, I. M. Hastings, I. Müller, A. Sie, O. Oa, M. Baisor, J. C. Reeder, H.-P. Beck, and B. Genton. Plasmodium falciparum resistance to anti-malarial drugs in Papua New Guinea: evaluation of a community-based approach for the molecular monitoring of resistance. *Malaria Journal*, 9:8, 2010.

[145] J. Marfurt. *Drug resistant malaria in Papua New Guinea and molecular monitoring of parasite resistance*. PhD thesis, 2006.

[146] E. Marinari and G. Parisi. Simulated Tempering: A New Monte Carlo Scheme. *EPL (Europhysics Letters)*, 19(6):451–458, 1992.

[147] K. Marsh. Malaria disaster in Africa. *The Lancet*, 352(9132):924, 1998.

[148] R. J. Maude, W. Pontavornpinyo, S. Saralamba, R. Aguas, S. Yeung, A. M. Dondorp, N. P. J. Day, N. J. White, and L. J. White. The last man standing is the most resistant: eliminating artemisinin-resistant malaria in Cambodia. *Malaria Journal*, 8:31, 2009.

[149] G. W. Mbogo, S. Nankoberanyi, S. Tukwasibwe, F. N. Baliraine, S. L. Nsobya, M. D. Conrad, E. Arinaitwe, M. Kamya, J. Tappero, S. G. Staedke, G. Dorsey, B. Greenhouse, and P. J. Rosenthal. Temporal Changes in Prevalence of Molecular Markers Mediating Antimalarial Drug Resistance in a High Malaria Transmission Setting in Uganda. *The American journal of tropical medicine and hygiene*, 2014.

[150] D. Ménard, N. Khim, J. Beghain, A. A. Adegnika, M. Shafiul-Alam, O. Amodu, G. Rahim-Awab, C. Barnadas, A. Berry, Y. Boum, M. D. Bustos, J. Cao, J.-H. Chen, L. Collet, L. Cui, G.-D. Thakur, A. Dieye, D. Djallé, M. A. Dorkenoo, C. E. Eboumbou-Moukoko, F.-E.-C. J. Espino, T. Fandeur, M.-F. Ferreira-da Cruz, A. A. Fola, H.-P. Fuehrer, A. M. Hassan, S. Herrera, B. Hongvanthong, S. Houzé, M. L. Ibrahim, M. Jahirul-Karim, L. Jiang, S. Kano, W. Ali-Khan, M. Khanthavong, P. G. Kremsner, M. Lacerda, R. Leang, M. Leelawong, M. Li, K. Lin, J.-B. Mazarati, S. Ménard, I. Morlais, H. Muhindo-Mavoko, L. Musset, K. Na-Bangchang, M. Nambozi, K. Niaré, H. Noedl, J.-B. Ouédraogo, D. R. Pillai, B. Pradines, B. Quang-Phuc, M. Ramharter, M. Randrianarivelojosia, J. Sattabongkot, A. Sheikh-Omar, K. D. Silué, S. B. Sirima, C. Sutherland, D. Syafruddin, R. Tahar, L.-H. Tang, O. A. Touré, P. Tshibangu-wa Tshibangu, I. Vigan-Womas, M. Warsame, L. Wini, S. Zakeri, S. Kim, R. Eam, L. Berne, C. Khean, S. Chy, M. Ken, K. Loch, L. Canier, V. Duru, E. Legrand, J.-C. Barale, B. Stokes, J. Straimer, B. Witkowski, D. A. Fidock, C. Rogier, P. Ringwald, F. Ariey, and O. Mercereau-Puijalon. A Worldwide Map of Plasmodium falciparum K13-Propeller Polymorphisms. *The New England journal of medicine*, 374(25):2453–2464, 2016.

[151] M. L. Metzker. Sequencing technologies - the next generation. *Nature reviews. Genetics*, 11(1):31–46, 2010.

[152] L. H. Miller. Distribution of mature trophozoites and schizonts of Plasmodium falciparum in the organs of Aotus trivirgatus, the night monkey. *American Journal of Tropical Medicine and Hygiene*, 18(6):860–865, 1969.

[153] O. Miotto, J. Almagro-Garcia, M. Manske, B. Macinnis, S. Campino, K. A. Rockett, C. Amaratunga, P. Lim, S. Suon, S. Sreng, J. M. Anderson, S. Duong, C. Nguon, C. M. Chuor, D. Saunders, Y. Se, C. Lon, M. M. Fukuda, L. Amenga-Etego, A. V. O. Hodgson, V. Asoala, M. Imwong, S. Takala-Harrison, F. Nosten, X.-Z. Su, P. Ringwald, F. Ariey, C. Dolecek, T. T. Hien, M. F. Boni, C. Q. Thai, A. Amambua-Ngwa, D. J. Conway, A. A. Djimdé, O. K. Doumbo, I. Zongo, J.-B. Ouedraogo, D. Alcock, E. Drury, S. Auburn, O. Koch, M. Sanders, C. Hubbart, G. Maslen, V. Ruano-Rubio, D. Jyothi, A. Miles, J. O'Brien, C. Gamble, S. O. Oyola, J. C. Rayner, C. I. Newbold, M. Berriman, C. C. A. Spencer, G. McVean, N. P. Day, N. J. White, D. Bethell, A. M. Dondorp, C. V. Plowe, R. M. Fairhurst, and D. P. Kwiatkowski. Multiple populations of artemisinin-resistant Plasmodium falciparum in Cambodia. *Nature Genetics*, 45(April):648–55, 2013.

[154] O. Miotto, R. Amato, E. A. Ashley, B. MacInnis, J. Almagro-Garcia, C. Amaratunga, P. Lim, D. Mead, S. O. Oyola, M. Dhorda, M. Imwong, C. Woodrow, M. Manske, J. Stalker,

E. Drury, S. Campino, L. Amenga-Etego, T.-N. N. Thanh, H. T. Tran, P. Ringwald, D. Bethell, F. Nosten, A. P. Phyo, S. Pukrittayakamee, K. Chotivanich, C. M. Chuor, C. Nguon, S. Suon, S. Sreng, P. N. Newton, M. Mayxay, M. Khanthavong, B. Hongvanthong, Y. Htut, K. T. Han, M. P. Kyaw, M. A. Faiz, C. I. Fanello, M. Onyamboko, O. A. Mokuolu, C. G. Jacob, S. Takala-Harrison, C. V. Plowe, N. P. Day, A. M. Dondorp, C. C. A. Spencer, G. McVean, R. M. Fairhurst, N. J. White, and D. P. Kwiatkowski. Genetic architecture of artemisinin-resistant Plasmodium falciparum. *Nature genetics*, 47(3):226–34, 2015.

[155] T. Mita, A. Kaneko, J. K. Lum, B. Bwijo, M. Takechi, I. L. Zungu, T. Tsukahara, K. Tanabe, T. Kobayakawa, and A. Björkman. Recovery of chloroquine sensitivity and low prevalence of the Plasmodium falciparum chloroquine resistance transporter gene mutation K76T following the discontinuance of chloroquine use in Malawi. *American Journal of Tropical Medicine and Hygiene*, 68(4):413–415, 2003.

[156] T. Mita, M. Venkatesan, J. Ohashi, R. Culleton, N. Takahashi, T. Tsukahara, M. Ndounga, L. Dysoley, H. Endo, F. Hombhanje, M. U. Ferreira, C. V. Plowe, and K. Tanabe. Limited geographical origin and global spread of sulfadoxine-resistant dhps alleles in plasmodium falciparum populations. *Journal of Infectious Diseases*, 204(12):1980–1988, 2011.

[157] F. P. Mockenhaupt, S. Ehrhardt, R. Otchwemah, T. A. Eggelte, S. D. Anemana, K. Stark, U. Bienzle, and E. Kohne. Limited influence of haemoglobin variants on Plasmodium falciparum msp1 and msp2 alleles in symptomatic malaria. *Transactions of the Royal Society of Tropical Medicine and Hygiene*, 98(5):302–310, 2004.

[158] A. Mohammed, A. Ndaro, A. Kalinga, A. Manjurano, J. F. Mosha, D. F. Mosha, M. van Zwetselaar, J. B. Koenderink, F. W. Mosha, M. Alifrangis, H. Reyburn, C. Roper, and R. A. Kavishe. Trends in chloroquine resistance marker, Pfcrt-K76T mutation ten years after chloroquine withdrawal in Tanzania. *Malaria Journal*, 12:415, 2013.

[159] J. F. Monahan. *Numerical methods of statistics*. Cambridge University Press, 2011.

[160] M. Morgan, H. Pagès, V. Obenchain, and N. Hayden. Rsamtools: Binary alignment (BAM), FASTA, variant call (BCF), and tabix file import. *http://bioconductor.org/packages/release/bioc/html/Rsamtools.html*.

[161] S. Muangnoicharoen, D. J. Johnson, S. Looareesuwan, S. Krudsood, and S. A. Ward. Role of known molecular markers of resistance in the antimalarial potency of piperaquine and dihydroartemisinin in vitro. *Antimicrobial Agents and Chemotherapy*, 53(4):1362–1366, 2009.

[162] L. Mwai, S. M. Kiara, A. Abdirahman, L. Pole, A. Rippert, A. Diriye, P. Bull, K. Marsh, S. Borrmann, and A. Nzila. In Vitro Activities of Piperaquine, Lumefantrine, and Dihydroartemisinin in Kenyan Plasmodium falciparum Isolates and Polymorphisms in pfcrt and pfmdr1. *Antimicrobial agents and chemotherapy*, 53(12):5069–5073, 2009.

[163] L. Mwai, A. Diriye, V. Masseno, S. Muriithi, T. Feltwell, J. Musyoki, J. Lemieux, A. Feller, G. R. Mair, K. Marsh, C. Newbold, A. Nzila, and C. K. Carret. Genome Wide Adaptations of Plasmodium falciparum in Response to Lumefantrine Selective Drug Pressure. *PloS one*, 7(2), 2012.

[164] F. Mwingira, G. Nkwengulila, S. Schoepflin, D. Sumari, H.-P. Beck, G. Snounou, I. Felger, P. Olliaro, and K. Mugittu. Plasmodium falciparum msp1, msp2 and glurp allele frequency and diversity in sub-Saharan Africa. *Malaria Journal*, 10(79), 2011.

[165] D. N. Nabarro and E. M. Tayler. The "Roll Back Malaria" Campaign. *Science*, 280(5372):2067–2068, 1998.

[166] I. Naidoo and C. Roper. Following the path of most resistance: dhps K540E dispersal in African Plasmodium falciparum. *Trends in parasitology*, 26(9):447–456, 2010.

[167] I. Naidoo and C. Roper. Drug resistance maps to guide intermittent preventive treatment of malaria in African infants. *Parasitology*, 138:1469–1479, 2011.

[168] I. Naidoo and C. Roper. Mapping 'partially resistant', 'fully resistant', and 'super resistant' malaria. *Trends in parasitology*, 29(10):505–15, 2013.

[169] S. Nair, B. Miller, M. Barends, A. Jaidee, J. Patel, M. Mayxay, P. Newton, F. Nosten, M. T. Ferdig, and T. J. C. Anderson. Adaptive copy number evolution in malaria parasites. *PLoS Genetics*, 4(10), 2008.

[170] S. Nair, S. C. Nkhoma, D. Serre, P. A. Zimmerman, K. Gorena, B. J. Daniel, F. Nosten, T. J. C. Anderson, and I. H. Cheeseman. Single-cell genomics for dissection of complex malaria infections. *Genome Research*, 24(6):1028–1038, 2014.

[171] J. A. Nájera, M. González-Silva, and P. L. Alonso. Some lessons for the future from the Global Malaria Eradication Programme (1955-1969). *PLoS medicine*, 8(1):e1000412, 2011.

[172] M. Nanyunja, J. Nabyonga Orem, F. Kato, M. Kaggwa, C. Katureebe, and J. Saweka. Malaria Treatment Policy Change and Implementation: The Case of Uganda. *Malaria Research and Treatment*, 2011:1–14, 2011.

[173] I. Nasell. *Hybrid Models of Tropical Infections*. Springer, first edition, 1985.

[174] I. Nåsell. On superinfection in malaria. *IMA journal of mathematics applied in medicine and biology*, 3(3):211–27, 1986.

[175] T. C. Nchinda. Malaria: A reemerging disease in Africa. *Emerging Infectious Diseases*, 4(3):398–403, 1998.

[176] M. Ndiaye, B. Faye, R. Tine, J. L. Ndiaye, A. Lo, A. Abiola, Y. Dieng, D. Ndiaye, R. Hallett, M. Alifrangis, and O. Gaye. Assessment of the molecular marker of Plasmodium falciparum chloroquine resistance (Pfcrt) in Senegal after several years of chloroquine withdrawal. *American Journal of Tropical Medicine and Hygiene*, 87(4):640–645, 2012.

[177] D. E. Neafsey, M. Juraska, T. Bedford, D. Benkeser, C. Valim, A. Griggs, M. Lievens, S. Abdulla, S. Adjei, T. Agbenyega, S. T. Agnandji, P. Aide, S. Anderson, D. Ansong, J. J. Aponte, K. P. Asante, P. Bejon, A. J. Birkett, M. Bruls, K. M. Connolly, U. D'Alessandro, C. Dobaño, S. Gesase, B. Greenwood, J. Grimsby, H. Tinto, M. J. Hamel, I. Hoffman, P. Kamthunzi, S. Kariuki, P. G. Kremsner, A. Leach, B. Lell, N. J. Lennon, J. Lusingu, K. Marsh, F. Martinson, J. T. Molel, E. L. Moss, P. Njuguna, C. F. Ockenhouse, B. R. Ogutu, W. Otieno, L. Otieno, K. Otieno, S. Owusu-Agyei, D. J. Park, K. Pellé, D. Robbins, C. Russ, E. M. Ryan, J. Sacarlal, B. Sogoloff, H. Sorgho, M. Tanner, T. Theander, I. Valea, S. K.

Volkman, Q. Yu, D. Lapierre, B. W. Birren, P. B. Gilbert, and D. F. Wirth. Genetic Diversity and Protective Efficacy of the RTS,S/AS01 Malaria Vaccine. *The New England journal of medicine*, 373(21):2025–37, 2015.

[178] J. Nedelman. Estimation for a model of multiple malaria infections. *Biometrics*, 41(2):447–53, 1985.

[179] J. A. Nelder and R. Mead. A simplex method for function minimization. *The Computer Journal*, 7(4):308–313, 1965.

[180] S. C. Nkhoma, S. Nair, I. H. Cheeseman, C. Rohr-Allegrini, S. Singlam, F. Nosten, and T. J. C. Anderson. Close kinship within multiple-genotype malaria parasite infections. *Proceedings of the Royal Society of London B: Biological Sciences*, 279(1738):2589–98, 2012.

[181] H. Noedl, Y. Se, K. Schaecher, B. L. Smith, D. Socheat, and M. M. Fukuda. Evidence of artemisinin-resistant malaria in western Cambodia. *The New England journal of medicine*, 359(24):2619–20, 2008.

[182] C. Nsanzabana, I. M. Hastings, J. Marfurt, I. Müller, K. Baea, L. Rare, A. Schapira, I. Felger, B. Betschart, T. A. Smith, H.-P. Beck, and B. Genton. Quantifying the evolution and impact of antimalarial drug resistance: drug use, spread of resistance, and drug failure over a 12-year period in Papua New Guinea. *The Journal of infectious diseases*, 201(3):435–43, 2010.

[183] F. Ntoumi, H. Contamin, C. Rogier, S. Bonnefoy, J. F. Trape, O. Mercereau-Puijalon, and O. Mercereau- Puijalon. Age-dependent carriage of multiple Plasmodium falciparum merozoite surface antigen-2 alleles in asymptomatic malaria infections. *The American journal of tropical medicine and hygiene*, 52(1):81–8, 1995.

[184] D. Nwakanma, A. Kheir, M. Sowa, S. Dunyo, M. Jawara, M. Pinder, P. Milligan, D. Walliker, and H. A. Babiker. High gametocyte complexity and mosquito infectivity of Plasmodium falciparum in the Gambia. *International Journal for Parasitology*, 38(2):219–227, 2008.

[185] J. D. O'Brien, X. Didelot, Z. Iqbal, L. Amenga-Etego, B. Ahiska, and D. Falush. A Bayesian Approach to Inferring the Phylogenetic Structure of Communities from Metagenomic Data. *Genetics*, pages 1–27, 2014.

[186] J. D. O'Brien, Z. Iqbal, and L. Amenga-Etego. An integrative statistical model for inferring strain admixture within clinical Plasmodium falciparum isolates. *arXiv:1505.08171v1*, pages 1–20, 2015.

[187] A. O'Hagan. *Kendall's advanced theory of statistics, volume 2B: Bayesian inference*. Arnold, 1994.

[188] P. Olliaro. Mode of action and mechanisms of resistance for antimalarial drugs. *Pharmacology & therapeutics*, 89(2):207–19, 2001.

[189] S. Owusu-Agyei, T. Smith, H. P. Beck, L. Amenga-Etego, and I. Felger. Molecular epidemiology of Plasmodium falciparum infections among asymptomatic inhabitants of a holoendemic malarious area in northern Ghana. *Tropical Medicine and International Health*, 7(5):421–428, 2002.

[190] H. Pages, P. Aboyoun, R. Gentleman, and S. DebRoy. Biostrings: String objects representing biological sequences, and matching algorithms.

[191] N. Papa Mze, Y. D. Ndiaye, C. K. Diedhiou, S. Rahamatou, B. Dieye, R. F. Daniels, E. J. Hamilton, M. Diallo, A. K. Bei, D. F. Wirth, S. Mboup, S. K. Volkman, A. D. Ahouidi, and D. Ndiaye. RDTs as a source of DNA to study Plasmodium falciparum drug resistance in isolates from Senegal and the Comoros Islands. *Malaria Journal*, 14:373, 2015.

[192] O. Papaspiliopoulos, G. O. Roberts, and M. Sköld. A general framework for the parametrization of hierarchical models. *Statistical Science*, 22(1):59–73, 2007.

[193] D. Payne. Spread of chloroquine resistance in Plasmodium falciparum. *Parasitology Today*, 3(8):241–246, 1987.

[194] D. Payne. Did medicated salt hasten the spread of chloroquine resistance in Plasmodium falciparum? *Parasitology Today*, 4(4):112–115, 1988.

[195] R. J. Pearce, H. Pota, M.-S. Evehe, E.-H. Bâ, G. Mombo-Ngoma, A. L. Malisa, R. Ord, W. Inojosa, A. Matondo, D. A. Diallo, W. Mbacham, d. B. Van, T. D. Swarthout, A. Getachew, S. Dejene, M. P. Grobusch, F. Njie, S. Dunyo, M. Kweku, S. Owusu-Agyei, D. Chandramohan, M. Bonnet, J.-P. Guthmann, S. Clarke, K. I. Barnes, E. Streat, S. T. Katokele, P. Uusiku, C. O. Agboghoroma, O. Y. Elegba, B. Cissé, I. A-Elbasit, H. A. Giha, S. P. Kachur, C. Lynch, J. B. Rwakimari, P. Chanda, M. Hawela, B. Sharp, I. Naidoo, and C. Roper. Multiple Origins and Regional Dispersal of Resistant dhps in African Plasmodium falciparum Malaria. *PLoS Med*, 6(4):e1000055, 2009.

[196] S. Pelleau, E. L. Moss, S. K. Dhingra, B. Volney, J. Casteras, and S. J. Gabryszewski. Adaptive evolution of malaria parasites in French Guiana : Reversal of chloroquine resistance by acquisition of a mutation in pfcrt. *Proceedings of the National Academy of Sciences*, 112(37), 2015.

[197] P. Perlmann and M. Troye-blomberg. Malaria and the Immune System in Humans. *Malaria Immunity in Humans*, 80:229–235, 2002.

[198] I. Petersen, R. Eastman, and M. Lanzer. Drug-resistant malaria: molecular mechanisms and implications for public health. *FEBS letters*, 585(11):1551–62, 2011.

[199] D. S. Peterson, D. Walliker, and T. E. Wellems. Evidence that a point mutation in dihydrofolate reductase-thymidylate synthase confers resistance to pyrimethamine in falciparum malaria. *Proceedings of the National Academy of Sciences of the United States of America*, 85(23):9114–8, 1988.

[200] D. S. Peterson, W. K. Milhous, and T. E. Wellems. Molecular basis of differential resistance to cycloguanil and pyrimethamine in Plasmodium falciparum malaria. *Proceedings of the National Academy of Sciences of the United States of America*, 87(8):3018–22, 1990.

[201] A. P. Phyo, S. Nkhoma, K. Stepniewska, E. A. Ashley, S. Nair, R. McGready, C. ler Moo, S. Al-Saai, A. M. Dondorp, K. M. Lwin, P. Singhasivanon, N. P. Day, N. J. White, T. J. C. Anderson, and F. Nosten. Emergence of artemisinin-resistant malaria on the western border of Thailand: a longitudinal study. *The Lancet*, 379(9830):1960–1966, 2012.

[202] S. Picot, P. Olliaro, F. de Monbrison, A.-L. Bienvenu, R. N. Price, and P. Ringwald. A systematic review and meta-analysis of evidence for correlation between molecular markers of parasite resistance and treatment outcome in falciparum malaria. *Malaria Journal*, 8:89, 2009.

[203] C. V. Plowe, A. Djimde, M. Bouare, O. Doumbo, and T. E. Wellems. Pyrimethamine and proguanil resistance-conferring mutations in Plasmodium falciparum dihydrofolate reductase: polymerase chain reaction methods for surveillance in Africa. *The American journal of tropical medicine and hygiene*, 52(6):565–8, 1995.

[204] C. V. Plowe, J. F. Cortese, A. Djimde, O. C. Nwanyanwu, W. M. Watkins, P. A. Winstanley, J. G. Estrada-Franco, R. E. Mollinedo, J. C. Avila, J. L. Cespedes, D. Carter, and O. K. Doumbo. Mutations in Plasmodium falciparum dihydrofolate reductase and dihydropteroate synthase and epidemiologic patterns of pyrimethamine-sulfadoxine use and resistance. *The Journal of infectious diseases*, 176:1590–6, 1997.

[205] C. V. Plowe, J. G. Kublin, and O. K. Doumbo. P. falciparum dihydrofolate reductase and dihydropteroate synthase mutations: epidemiology and role in clinical resistance to antifolates. *Drug Resistance Updates*, 1(6):389–396, 1998.

[206] C. V. Plowe. Monitoring antimalarial drug resistance: making the most of the tools at hand. *Journal of Experimental Biology*, 206(21):3745–3752, 2003.

[207] M. Plummer, N. Best, K. Cowles, and K. Vines. CODA: convergence diagnosis and output analysis for MCMC. *R News*, 6(1):7–11, 2006.

[208] T. Ponnudurai, A. H. W. Lensen, G. J. A. van Gemert, M. G. Bolmer, and J. H. E. Th. Meuwissen. Feeding behaviour and sporozoite ejection by infected Anopheles stephensi. *Transactions of the Royal Society of Tropical Medicine and Hygiene*, 85(2):175–180, 1991.

[209] S. Portugal, H. Drakesmith, and M. M. Mota. Superinfection in malaria: Plasmodium shows its iron will. *EMBO reports*, 12(12):1233–1242, 2011.

[210] R Core Team. R: A language and environment for statistical computing. *https://www.R-project.org/*, 2013.

[211] A. Rhoads and K. F. Au. PacBio Sequencing and Its Applications. *Genomics, Proteomics and Bioinformatics*, 13(5):278–289, 2015.

[212] C. P. Robert. *The Bayesian Choice*. Springer, second edition, 2007.

[213] G. O. Roberts and S. K. Sahu. Updating Schemes, Correlation Structure, Blocking and Parameterization for the Gibbs Sampler. *Journal of the Royal Statistical Society. Series B (Methodological)*, 59(2):291–317, 1997.

[214] G. O. Roberts and R. L. Tweedie. Exponential convergence of Langevin distributions and their discrete approximations. *Bernoulli*, 2(4):341–363, 1996.

[215] S. J. Rogerson, L. Hviid, P. E. Duffy, R. F. Leke, and D. W. Taylor. Malaria in pregnancy: pathogenesis and immunity. *Lancet Infectious Diseases*, 7(2):105–117, 2007.

[216] C. Roper, R. Pearce, S. Nair, B. Sharp, F. Nosten, and T. J. C. Anderson. Intercontinental spread of pyrimethamine-resistant malaria. *Science*, 305(5687):1124, 2004.

[217] C. Roper, M. Alifrangis, F. Ariey, A. Talisuna, D. Menard, O. Mercereau-Puijalon, and P. Ringwald. Molecular surveillance for artemisinin resistance in Africa. *The Lancet. Infectious diseases*, 14(8):668–70, 2014.

[218] R. Rosenberg, R. A. Wirtz, I. Schneider, and R. Burge. An estimation of the number of malaria sporozoites ejected by a feeding mosquito. *Transactions of the Royal Society of Tropical Medicine and Hygiene*, 84(2):209–212, 1990.

[219] A. Ross, C. Koepfli, X. Li, S. Schoepflin, P. Siba, I. Mueller, I. Felger, and T. Smith. Estimating the numbers of malaria infections in blood samples using high-resolution genotyping data. *PloS one*, 7(8):e42496, 2012.

[220] J. M. Rubio, A. Benito, J. Roche, P. J. Berzosa, M. L. García, M. Micó, M. Edú, J. Alvar, M. L. Garcia, M. Mico, and M. Edu. Semi-nested, multiplex polymerase chain reaction for detection of human malaria parasites and evidence of Plasmodium vivax infection in Equatorial Guinea. *American Journal of Tropical Medicine and Hygiene*, 60(2):183–187, 1999.

[221] W. Sama, S. Owusu-Agyei, I. Felger, P. Vounatsou, and T. Smith. An immigration-death model to estimate the duration of malaria infection when detectability of the parasite is imperfect. *Statistics in Medicine*, 24(21):3269–3288, 2005.

[222] W. Sama, S. Owusu-Agyei, I. Felger, K. Dietz, and T. Smith. Age and seasonal variation in the transition rates and detectability of Plasmodium falciparum malaria. *Parasitology*, 132(Pt 1):13–21, 2006.

[223] K. A. Schneider and A. A. Escalante. A likelihood approach to estimate the number of co-infections. *PloS one*, 9(7):e97899, 2014.

[224] A. G. Schneider, Z. Premji, I. Felger, T. Smith, S. Abdulla, H.-P. Beck, and H. Mshinda. A point mutation in codon 76 of pfcrt of P. falciparum is positively selected for by Chloroquine treatment in Tanzania. *Infection, genetics and evolution : journal of molecular epidemiology and evolutionary genetics in infectious diseases*, 1(3):183–9, 2002.

[225] S. Schoepflin, F. Valsangiacomo, E. Lin, B. Kiniboro, I. Mueller, and I. Felger. Comparison of Plasmodium falciparum allelic frequency distribution in different endemic settings by high-resolution genotyping. *Malaria Journal*, 8:250, 2009.

[226] K. F. Scmidt. Inbred parasites may spur resistance (News). *Science*, 269(22 September):1670, 1995.

[227] J. M. G. Shelton, P. Corran, P. Risley, N. Silva, C. Hubbart, A. Jeffreys, K. Rowlands, R. Craik, V. Cornelius, M. Hensmann, S. Molloy, N. Sepulveda, T. G. Clark, G. Band, G. M. Clarke, C. C. A. Spencer, A. Kerasidou, S. Campino, S. Auburn, A. Tall, A. B. Ly, O. Mercereau-Puijalon, A. Sakuntabhai, A. Djimde, B. Maiga, O. Toure, O. K. Doumbo,

A. Dolo, M. Troye-Blomberg, V. D. Mangano, F. Verra, D. Modiano, E. Bougouma, S. B. Sirima, M. Ibrahim, A. Hussain, N. Eid, A. Elzein, H. Mohammed, A. Elhassan, I. Elhassan, T. N. Williams, C. Ndila, A. Macharia, K. Marsh, A. Manjurano, H. Reyburn, M. Lemnge, D. Ishengoma, R. Carter, N. Karunaweera, D. Fernando, R. Dewasurendra, C. J. Drakeley, E. M. Riley, D. P. Kwiatkowski, K. A. Rockett, and MalariaGen Consortium. Genetic determinants of anti-malarial acquired immunity in a large multi-centre study. *Malaria Journal*, 14(333):1–18, 2015.

[228] C. H. Sibley, J. E. Hyde, P. F. Sims, C. V. Plowe, J. G. Kublin, E. K. Mberu, A. F. Cowman, P. A. Winstanley, W. M. Watkins, and A. M. Nzila. Pyrimethamine-sulfadoxine resistance in Plasmodium falciparum: what next? *Trends in parasitology*, 17(12):582–8, 2001.

[229] T. Smith and P. Vounatsou. Estimation of infection and recovery rates for highly polymorphic parasites when detectability is imperfect, using hidden Markov models. *Statistics in medicine*, 22(10):1709–24, 2003.

[230] T. Smith, H. P. Beck, A. Kitua, S. Mwankusye, I. Felger, N. Fraser-Hurt, A. Irion, P. Alonso, T. Teuscher, and M. Tanner. Age dependence of the multiplicity of Plasmodium falciparum infections and of other malariological indices in an area of high endemicity. *Transactions of the Royal Society of Tropical Medicine and Hygiene*, 93(suppl 1):15–20, 1999.

[231] R. W. Snow, J. F. Trape, and K. Marsh. The past, present and future of childhood malaria mortality in Africa. *Trends in parasitology*, 17(12):593–597, 2001.

[232] A. F. Somé, Y. Y. Séré, C. Dokomajilar, I. Zongo, N. Rouamba, B. Greenhouse, J. B. Ouédraogo, and P. J. Rosenthal. Selection of known plasmodium falciparum resistance-mediating polymorphisms by artemether-lumefantrine and amodiaquine-sulfadoxine-pyrimethamine but not dihydroartemisinin-piperaquine in Burkina Faso. *Antimicrobial Agents and Chemotherapy*, 54(5):1949–1954, 2010.

[233] D. J. Spiegelhalter, N. G. Best, B. P. Carlin, and A. Van Der Linde. Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(4):583–639, 2002.

[234] S. G. Staedke, A. Mpimbaza, M. R. Kamya, B. K. Nzarubara, G. Dorsey, and P. J. Rosenthal. Combination treatments for uncomplicated falciparum malaria in Kampala, Uganda: randomised clinical trial. *The Lancet*, 364:1950–1957, 2004.

[235] Stan Development Team. Stan Modeling Language: User's Guide and Reference Manual. *http://mc-stan.org/*, Stan Versi, 2015.

[236] J. Straimer, N. F. Gnädig, B. Witkowski, C. Amaratunga, V. Duru, A. P. Ramadani, M. Dacheux, N. Khim, L. Zhang, S. Lam, P. D. Gregory, F. D. Urnov, O. Mercereau-Puijalon, F. Benoit-Vical, R. M. Fairhurst, D. Ménard, and D. A. Fidock. K13-propeller mutations confer artemisinin resistance in Plasmodium falciparum clinical isolates. *Science*, 2624(1985):428–431, 2015.

[237] S.-Y. Su, J. E. Asher, M.-R. Jarvelin, P. Froguel, A. I. F. Blakemore, D. J. Balding, and L. J. M. Coin. Inferring combined CNV/SNP haplotypes from genotype data. *Bioinformatics*, 26(11):1437–45, 2010.

[238] P. L. Sutton, V. Neyra, J. N. Hernandez, and O. H. Branch. Plasmodium falciparum and Plasmodium vivax infections in the Peruvian Amazon: Propagation of complex, multiple allele-type infections without super-infection. *American Journal of Tropical Medicine and Hygiene*, 81(6):950–960, 2009.

[239] P. Tabernero, M. Mayxay, M. J. Culzoni, P. Dwivedi, I. Swamidoss, E. L. Allan, M. Khanthavong, C. Phonlavong, C. Vilayhong, S. Yeuchaixiong, C. Sichanh, S. Sengaloundeth, H. Kaur, F. M. Fernandez, M. D. Green, and P. N. Newton. A Repeat Random Survey of the Prevalence of Falsified and Substandard Antimalarials in the Lao PDR: A Change for the Better. *American Journal of Tropical Medicine and Hygiene*, 92(suppl 6):95–104, 2015.

[240] S. L. Takala, D. L. Smith, O. C. Stine, D. Coulibaly, M. A. Thera, O. K. Doumbo, and C. V. Plowe. A high-throughput method for quantifying alleles and haplotypes of the malaria vaccine candidate Plasmodium falciparum merozoite surface protein-1 19 kDa. *Malaria Journal*, 5(31), 2006.

[241] S. Takala-Harrison, T. G. Clark, C. G. Jacob, M. P. Cummings, O. Miotto, A. M. Dondorp, M. M. Fukuda, F. Nosten, H. Noedl, M. Imwong, D. Bethell, Y. Se, C. Lon, S. D. Tyner, D. L. Saunders, D. Socheat, F. Ariey, A. P. Phyo, P. Starzengruber, H.-P. Fuehrer, P. Swoboda, K. Stepniewska, J. Flegg, C. Arze, G. C. Cerqueira, J. C. Silva, S. M. Ricklefs, S. F. Porcella, R. M. Stephens, M. Adams, L. J. Kenefic, S. Campino, S. Auburn, B. Macinnis, D. P. Kwiatkowski, X.-Z. Su, N. J. White, P. Ringwald, and C. V. Plowe. Genetic loci associated with delayed clearance of Plasmodium falciparum following artemisinin treatment in Southeast Asia. *Proceedings of the National Academy of Sciences of the United States of America*, 110(1):240–245, 2012.

[242] S. Takala-Harrison, C. G. Jacob, C. Arze, M. P. Cummings, J. C. Silva, A. M. Dondorp, M. M. Fukuda, T. T. Hien, M. Mayxay, H. Noedl, F. Nosten, M. P. Kyaw, N. T. T. Nhien, M. Imwong, D. Bethell, Y. Se, C. Lon, S. D. Tyner, D. L. Saunders, F. Ariey, O. Mercereau-Puijalon, D. Menard, P. N. Newton, M. Khanthavong, B. Hongvanthong, P. Starzengruber, H.-P. Fuehrer, P. Swoboda, W. A. Khan, A. P. Phyo, M. M. Nyunt, M. H. Nyunt, T. S. Brown, M. Adams, C. S. Pepin, J. Bailey, J. C. Tan, M. T. Ferdig, T. G. Clark, O. Miotto, B. MacInnis, D. P. Kwiatkowski, N. J. White, P. Ringwald, and C. V. Plowe. Independent Emergence of Artemisinin Resistance Mutations Among Plasmodium falciparum in Southeast Asia. *Journal of Infectious Diseases*, 211(5):670–679, 2015.

[243] A. O. Talisuna, A. Nalunkuma-Kazibwe, N. Bakyaita, P. Langi, T. K. Mutabingwa, W. W. Watkins, E. Van Marck, U. D'Alessandro, and T. G. Egwang. Efficacy of sulphadoxine-pyrimethamine alone or combined with amodiaquine or chloroquine for the treatment of uncomplicated falciparum malaria in Ugandan children. *Tropical medicine & international health*, 9(2):222–9, 2004.

[244] A. Talisuna, S. Adibaku, G. Dorsey, M. R. Kamya, and P. J. Rosenthal. Malaria in Uganda: challenges to control on the long road to elimination. II. The path forward. *Acta tropica*, 121(3):196–201, 2012.

[245] M. A. Tanner and W. H. Wong. The Calculation of Posterior Distributions by Data Augmentation. *Journal of the American Statistical Association*, 82(398):528–540, 1987.

[246] S. M. Taylor, A. Antonia, G. Feng, V. Mwapasa, E. Chaluluka, M. Molyneux, F. O. ter Kuile, S. J. Rogerson, and S. R. Meshnick. Adaptive evolution and fixation of drug-resistant

Plasmodium falciparum genotypes in pregnancy-associated malaria: 9-year results from the QuEERPAM study. *Infection, genetics and evolution*, 12(2):282–90, 2012.

[247] S. M. Taylor, C. M. Parobek, N. Aragam, B. E. Ngasala, A. Mårtensson, S. R. Meshnick, and J. J. Juliano. Pooled deep sequencing of Plasmodium falciparum isolates: an efficient and scalable tool to quantify prevailing malaria drug-resistance genotypes. *The Journal of infectious diseases*, 208(12):1998–2006, 2013.

[248] A. R. Taylor, J. A. Flegg, S. L. Nsobya, A. Yeka, M. R. Kamya, P. J. Rosenthal, G. Dorsey, C. H. Sibley, P. J. Guerin, and C. C. Holmes. Estimation of malaria haplotype and genotype frequencies: a statistical approach to overcome the challenge associated with multiclonal infections. *Malaria Journal*, 13(1):102, 2014.

[249] A. Taylor Bright and E. A. Winzeler. Resistance mapping in malaria. *Nature*, 498:446–447, 2013.

[250] The Four Artemisinin-Based Combinations Study group. A head-to-head comparison of four artemisinin-based combinations for treating uncomplicated malaria in african children: A randomized trial. *PLoS Medicine*, 8(11), 2011.

[251] The Roll Back Malaria (RBM) Partnership. Executive Summary: The Global Malaria Action Plan. Technical report, 2008.

[252] The RTSS Clinical Trials Partnership. A Phase 3 Trial of RTS,S/AS01 Malaria Vaccine in African Infants. *New England Journal of Medicine*, 367(24):2284–2295, 2012.

[253] T. T. Thomsen, L. B. Madsen, H. H. Hansson, E. V. E. Tomás, D. Charlwood, I. C. Bygbjerg, and M. Alifrangis. Rapid Selection of Plasmodium falciparum Chloroquine Resistance Transporter Gene and Multidrug Resistance Gene-1 Haplotypes Associated with Past Chloroquine and Present Artemether-Lumefantrine Use in Inhambane District, Southern Mozambique. *The American journal of tropical medicine and hygiene*, 88(3):536–41, 2013.

[254] J. F. Trape, G. Pison, M. P. Preziosi, C. Enel, A. Desgrées du Loû, V. Delaunay, B. Samb, E. Lagarde, J. F. Molez, and F. Simondon. Impact of chloroquine resistance on malaria mortality. *Medical Sciences*, 321(8):689–97, 1998.

[255] T. Triglia, P. Wang, P. F. Sims, J. E. Hyde, and A. F. Cowman. Allelic exchange at the endogenous genomic locus in Plasmodium falciparum proves the role of dihydropteroate synthase in sulfadoxine-resistant malaria. *The EMBO journal*, 17(14):3807–15, 1998.

[256] Y. Tsumori, M. Ndounga, T. Sunahara, N. Hayashida, M. Inoue, S. Nakazawa, P. Casimiro, R. Isozumi, H. Uemura, K. Tanabe, O. Kaneko, and R. Culleton. Plasmodium falciparum: Differential Selection of Drug Resistance Alleles in Contiguous Urban and Peri-Urban Areas of Brazzaville, Republic of Congo. *PloS one*, 6(8):e23430, 2011.

[257] K. M. Tun, M. Imwong, K. M. Lwin, A. A. Win, T. M. Hlaing, T. Hlaing, K. Lin, M. P. Kyaw, K. Plewes, M. A. Faiz, M. Dhorda, P. Y. Cheah, S. Pukrittayakamee, E. A. Ashley, T. J. C. Anderson, S. Nair, M. McDew-White, J. A. Flegg, E. P. M. Grist, P. Guerin, R. J. Maude, F. Smithuis, A. M. Dondorp, N. P. J. Day, F. Nosten, N. J. White, and C. J. Woodrow. Spread of artemisinin-resistant Plasmodium falciparum in Myanmar: a cross-sectional survey of the K13 molecular marker. *The Lancet Infectious Diseases*, 15(4):415–421, 2015.

[258] L. S. Tusting, T. Bousema, D. L. Smith, and C. Drakeley. Measuring Changes in Plasmodium falciparum Transmission: Precision, Accuracy and Costs of Metrics. In *Advances in parasitology*, volume 84, pages 151–208. Elsevier Ltd., first edition, 2014.

[259] United Nations General Assembly. United Nations Millennium Declaration. Technical report, 2000.

[260] M. Vafa, M. Troye-Blomberg, J. Anchang, A. Garcia, and F. Migot-Nabias. Multiplicity of Plasmodium falciparum infection in asymptomatic children in Senegal: relation to transmission, age and erythrocyte variants. *Malaria Journal*, 7:17, 2008.

[261] J. C. Venter, K. Remington, J. F. Heidelberg, A. L. Halpern, D. Rusch, J. A. Eisen, D. Wu, I. Paulsen, K. E. Nelson, W. Nelson, D. E. Fouts, S. Levy, A. H. Knap, M. W. Lomas, K. Nealson, O. White, J. Peterson, J. Hoffman, R. Parsons, H. Baden-Tillson, C. Pfannkoch, Y.-H. Rogers, and H. O. Smith. Environmental genome shotgun sequencing of the Sargasso Sea. *Science*, 304(5667):66–74, 2004.

[262] S. K. Volkman, P. C. Sabeti, D. Decaprio, D. E. Neafsey, S. F. Schaffner, D. A. Milner, J. P. Daily, O. Sarr, D. Ndiaye, O. Ndir, S. Mboup, M. T. Duraisingh, A. Lukens, A. Derr, N. Stange-thomann, S. Waggoner, R. Onofrio, L. Ziaugra, E. Mauceli, S. Gnerre, D. B. Jaffe, J. Zainoun, R. C. Wiegand, B. W. Birren, D. L. Hartl, J. E. Galagan, E. S. Lander, and D. F. Wirth. A genome-wide map of diversity in Plasmodium falciparum. *Nature genetics*, 39(1):113–119, 2007.

[263] L. Von Seidlein and B. M. Greenwood. Mass administrations of antimalarial drugs. *Trends in Parasitology*, 19(10):452–460, 2003.

[264] P. Wang, C. S. Lee, R. Bayoumi, A. Djimde, O. Doumbo, G. Swedberg, L. D. Dao, H. Mshinda, M. Tanner, W. M. Watkins, P. F. Sims, and J. E. Hyde. Resistance to antifolates in Plasmodium falciparum monitored by sequence analysis of dihydropteroate synthetase and dihydrofolate reductase alleles in a large number of field samples of diverse origins. *Molecular and biochemical parasitology*, 89(2):161–77, 1997.

[265] Z. Wang, S. Shrestha, X. Li, J. Miao, L. Yuan, M. Cabrera, C. Grube, Z. Yang, and L. Cui. Prevalence of K13-propeller polymorphisms in Plasmodium falciparum from China-Myanmar border in 2007-2012. *Malaria Journal*, 14(1):168, 2015.

[266] H. Wanzira, A. Kakuru, E. Arinaitwe, V. Bigira, M. K. Muhindo, M. Conrad, P. J. Rosenthal, M. R. Kamya, J. W. Tappero, and G. Dorsey. Longitudinal Outcomes in a Cohort of Ugandan Children Randomized to Artemether-lumefantrine Verus Dihydroartemisinin-piperaquine for the Treatment of Malaria. *Clinical infectious diseases*, 59(4):509–516, 2014.

[267] D. A. Warrell and H. M. Gilles. *Essential malariology*, volume 1. ARNOLD, London, fourth edition, 2002.

[268] T. E. Wellems and C. V. Plowe. Chloroquine-resistant malaria. *The Journal of infectious diseases*, 184(6):770–6, 2001.

[269] K. A. Wetterstrand. DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP). *http://www.genome.gov/sequencingcosts/*, 2016.

[270] N. J. White, F. Nosten, S. Looareesuwan, W. M. Watkins, K. Marsh, R. W. Snow, G. Kokwaro, J. Ouma, T. T. Hien, M. E. Molyneux, T. E. Taylor, C. I. Newbold, T. K. Ruebush, M. Danis, B. M. Greenwood, R. M. Anderson, and P. Olliaro. Averting a malaria disaster. *Lancet*, 353:1965–1967, 1999.

[271] M. T. White, R. Verity, J. T. Griffin, K. P. Asante, S. Owusu-Agyei, B. Greenwood, C. Drakeley, S. Gesase, J. Lusingu, D. Ansong, S. Adjei, T. Agbenyega, B. Ogutu, L. Otieno, W. Otieno, S. T. Agnandji, B. Lell, P. Kremsner, I. Hoffman, F. Martinson, P. Kamthunzu, H. Tinto, I. Valea, H. Sorgho, M. Oneko, K. Otieno, M. J. Hamel, N. Salim, A. Mtoro, S. Abdulla, P. Aide, J. Sacarlal, J. J. Aponte, P. Njuguna, K. Marsh, P. Bejon, E. M. Riley, and A. C. Ghani. Immunogenicity of the RTS,S/AS01 malaria vaccine and implications for duration of vaccine efficacy: secondary analysis of data from a phase 3 randomised controlled trial. *The Lancet Infectious Diseases*, 3099(15):1–9, 2015.

[272] N. J. White. Assessment of the pharmacodynamic properties of antimalarial drugs in vivo. *Antimicrobial Agents and Chemotherapy*, 41(7):1413–1422, 1997.

[273] N. J. White. Antimalarial drug resistance. *Trends in Parasitology*, 113(8):1084–1092, 2004.

[274] N. J. White. Antimalarial drug resistance. *Trends in Parasitology*, 113(8):1084–1092, 2004.

[275] N. J. White. Qinghaosu (Artemisinin): The Price of Success. *Science*, 320:330–334, 2008.

[276] L. Wigger, J. E. Vogt, and V. Roth. Malaria haplotype frequency estimation. *Statistics in medicine*, 32(21):3737–3751, 2013.

[277] B. G. Williams and C. Dye. Maximum likelihood for parasitologists. *Parasitology Today*, 10(12):489–93, 1994.

[278] C. M. Wilson, S. K. Volkman, S. Thaithong, R. K. Martin, D. E. Kyle, W. K. Milhous, and D. F. Wirth. Amplification of pfmdr1 associated with mefloquine and halofantrine resistance in Plasmodium falciparum from Thailand. *Molecular and Biochemical Parasitology*, 57(1):151–160, 1993.

[279] C. Wongsrichanalai and S. R. Meshnick. Declining artesunate-mefloquine efficacy against falciparum malaria on the Cambodia-Thailand border. *Emerging infectious diseases*, 14(5):716–9, 2008.

[280] J. C. Wootton, X. Feng, M. T. Ferdig, R. A. Cooper, J. Mu, D. I. Baruch, A. J. Magill, and X.-Z. Su. Genetic diversity and chloroquine selective sweeps in Plasmodium falciparum. *Nature*, 418(6895):320–323, 2002.

[281] World Health Organization. Antimalarial drug combination therapy: report of a WHO technical consultation. Technical report, Geneva, 2001.

[282] World Health Organization. Recommended Genotyping Procedures (RGPs) to identify parasite populations. Technical report, Amsterdam, 2007.

[283] World Health Organization. Global plan for artemisinin resistance containment. Technical report, Geneva, 2011.

[284] World Health Organization. Emergency Response to Artemisinin Resistance in the Greater Mekong Subregion. Technical report, Geneva, 2013.

[285] World Health Organization. Severe malaria. *Tropical Medicine and International Health*, 19(suppl 1):7–131, 2014.

[286] World Health Organization. Global technical strategy for malaria 2016-2030. Technical report, Geneva, 2015.

[287] World Health Organization. Guidelines for the treatment of malaria: third edition. Technical report, World Health Organization, Geneva, 2015.

[288] World Health Organization. Status report on artemisinin and ACT resistance. Technical report, Geneva, 2015.

[289] World Health Organization. Strategy for Malaria Elimination in the Greater Mekong Subregion (2015-2030). Technical report, Geneva, 2015.

[290] World Health Organization. World Malaria Report 2015. Technical report, Geneva, 2015.

[291] World Health Organization. Action and investment to defeat malaria 2016-2030: for a malaria-free World. Technical report, Geneva, 2016.

[292] World Health Organization. Weekly epidemiological record: relevé épidémiologique hebdomadaire. *World Health Organization*, 91(4):33–52, 2016.

[293] WWARN. Molecular Testing for Malaria Standard Operating Procedure (SOP) DNA Extraction by Chelex.

[294] WWARN. PCR-RFLP for genotyping candidate P. falciparum artemisinin Procedure Molecular Module. *http://www.wwarn.org/tools-resources/procedures*.

[295] WWARN Artemisinin based Combination Therapy (ACT) Africa Baseline Study Group. Clinical determinants of early parasitological response to ACTs in African patients with uncomplicated falciparum malaria: a literature review and meta-analysis of individual patient data. *BMC Medicine*, 13(212), 2015.

[296] WWARN. Molecular surveyor dhfr & dhps. *http://www.wwarn.org/tracking-resistance/molecular-surveyor-dhfr-dhps*, 2016.

[297] WWARN. Molecular surveyor K13. *http://www.wwarn.org/molecular-surveyor-k13*, 2016.

[298] WWARN. Molecular surveyor pfmdr1 & pfcrt. *http://www.wwarn.org/tracking-resistance/molecular-surveyor-pfmdr1-pfcrt*, 2016.

[299] WWARN. Molecular Testing for Malaria Standard Operating Procedure (SOP): collection of blood on filterpaper. *http://www.wwarn.org/tools-resources/procedures*, 2016.

[300] WWARN. Molecular Testing for Malaria Standard Operating Procedure (SOP) DNA Extraction by QIAamp DNA Mini Kit. *http://www.wwarn.org/tools-resources/procedures*, 2016.

[301] A. Yeka, K. Banek, N. Bakyaita, S. G. Staedke, M. R. Kamya, A. Talisuna, F. Kironde, S. L. Nsobya, A. Kilian, M. Slater, and Others. Artemisinin versus nonartemisinin combination therapy for uncomplicated malaria: randomized clinical trials from four sites in Uganda. *PLoS medicine*, 2(7):e190–e190, 2005.

[302] A. Yeka, A. Gasasira, A. Mpimbaza, J. Achan, J. Nankabirwa, S. Nsobya, S. G. Staedke, M. J. Donnelly, F. Wabwire-Mangen, A. Talisuna, G. Dorsey, M. R. Kamya, and P. J. Rosenthal. Malaria in Uganda: challenges to control on the long road to elimination: I. Epidemiology and current control efforts. *Acta tropica*, 121(3):184–95, 2012.

[303] A. Yeka, R. Kigozi, M. D. Conrad, M. Lugemwa, P. Okui, C. Katureebe, K. Belay, B. K. Kapella, M. A. Chang, M. R. Kamya, S. G. Staedke, G. Dorsey, and P. J. Rosenthal. Artesunate/amodiaquine versus artemether/lumefantrine for the treatment of uncomplicated malaria in Uganda: a randomized trial. *Journal of Infectious Diseases*, pages 1–9, 2015.

[304] S. Yeung, W. Van Damme, D. Socheat, N. J. White, and A. Mills. Access to artemisinin combination therapy for malaria in remote areas of Cambodia. *Malaria Journal*, 7(96), 2008.

[305] S. Yeung, H. L. S. Lawford, P. Tabernero, C. Nguon, A. van Wyk, N. Malik, M. DeSousa, O. Rada, M. Boravann, P. Dwivedi, D. M. Hostetler, I. Swamidoss, M. D. Green, F. M. Fernandez, and H. Kaur. Quality of Antimalarials at the Epicenter of Antimalarial Drug Resistance: Results from an Overt and Mystery Client Survey in Cambodia. *The American journal of tropical medicine and hygiene*, 92(suppl 6):39–50, 2015.

[306] O. Zagordi, L. Geyrhofer, V. Roth, and N. Beerenwinkel. Deep sequencing of a genetically heterogeneous sample: local haplotype reconstruction and read error correction. *Journal of computational biology*, 17(3):417–28, 2010.

Appendix A

Frequency trends in Uganda

A.1 Auxiliary study of the MOI

To ascertain the appropriate summaries of the MOIs for the sensitivity analyses described in section 4.2.3, a brief exploratory study of auxiliary MOI data is performed. Note that the data were generated by Conrad *et al.*, and are described in detail elsewhere [51]. In summary, to estimate average MOIs across two drug arms of a longitudinal drug trial, Conrad *et al.* genotyped ten blood spots per quarter, per drug arm at *msp1* and *msp2* using capillary electrophoresis [51]. Per sample MOI estimates were based on the maximum number of alleles at either gene. To estimate baseline MOIs, ten baseline blood spots were also genotyped at *msp1* and *msp2*.

All 380 longitudinal MOI estimates are depicted in figure A.1. They range from 1 to 7 with mean 2.94 (black horizontal line, figure A.1). The confidence intervals depicted in figure A.1 and summarised below are constructed using the student t distribution with $n-1$ degrees of freedom, while average MOI summaries are compared using the Kruskall Wallis test.

There is little difference between the average MOI in the AL drug arm (2.80, green horizontal line, figure A.1) and the average MOI in the DP drug arm (3.07, blue horizontal line, figure A.1) (Kruskal-Wallis $\chi^2 = 3.37$, 1 degree of freedom, p-value 0.07). In fact, the

	Year				
	2008	2009	2010	2011	2012
DP	3.50 (3.05–3.95)	2.98 (2.54–3.41)	2.85 (2.41–3.29)	3.17 (2.71–3.64)	2.80 (2.25–3.35)
AL	3.08 (2.60–3.55)	3.00 (2.60–3.40)	2.80 (2.29–3.31)	2.48 (2.14–2.81)	2.60 (2.14–3.06)

Table A.1: Experimentally-derived yearly average MOI estimates per drug arm (95% confidence interval).

averages per drug arm are almost identical to the 95% confidence bounds (2.80 and 3.08, black dashed lines, figure A.1) of the overall mean (2.94, black solid line, figure A.1). There is weak evidence of a difference between the yearly average MOIs (black bars, figure A.1, Kruskal-Wallis $\chi^2 = 9.52$, 4 degrees of freedom, p-value = 0.05), but not when the data are categorised by drug arm (table A.1 and coloured bars, figure A.1, Kruskal-Wallis $\chi^2 = 9.02$, 5 degrees of freedom, p-value = 0.11, AL arm; Kruskal-Wallis $\chi^2 = 10.81$, 5 degrees of freedom, p-value = 0.06, DP arm). There is a significant difference between the MOI estimates associated with the AL and DP arms in 2011 (Kruskal-Wallis $\chi^2 = 4.74$, 1 degree of freedom, p-value = 0.03). The baseline mean is 3.30 (95% confidence interval, 2.62–3.94).

A.2 Determining of the prior on the MOI

The choice of prior distribution on the MOI described in section 4.2.3 is based on a preliminary study using posterior predictive (see [88] for more details regarding posterior predictive checking). The prior that provides the best fit to the data is used in the final analyses.

The haplotype frequency estimation model (section 3.2.3) is fit to each of the 66 subdivisions of the data (coloured endpoints, panels C, D and E, figure 4.3) four times using four different

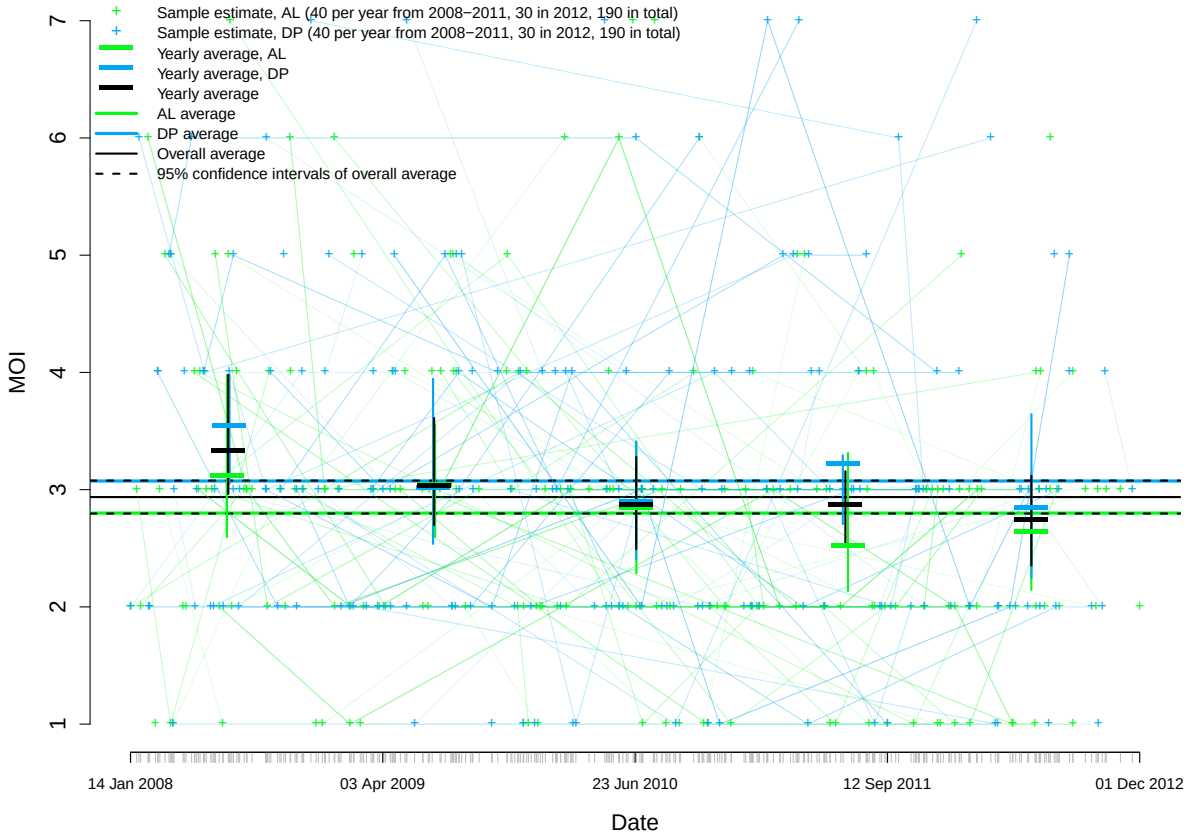


Figure A.1: Experimentally-derived MOI estimates. Crosses (+) represent sample-wise estimates, colour-coded by drug arm (blue DP, green AL). Each sample was obtained from a single episode of malaria. However, not all were derived from different children. Multiple samples obtained from the same child are joined by diagonal lines colour-coded by drug arm (blue DP, green AL). The overall average MOI (2.94) is depicted by a black horizontal line. The bounds of the 95% confidence interval surrounding the overall average (2.80 and 3.08) are depicted by dashed horizontal black lines. The average MOI per drug arm (2.80 AL, 3.07 DP) is depicted by a coloured horizontal line (blue DP, green AL). Yearly average MOIs (table A.1) are depicted by black bars, while yearly average MOIs per drug arm are depicted by coloured bars (blue DP, green AL).

priors on the MOI,

$$\rho(m_i) = \begin{cases} \mathscr{P}\text{oisson}_{\text{truncated}}\left(m_i \mid \lambda, m_{i\min}, m_{\max}\right), \\ \mathscr{U}\text{niform}_{\text{truncated}}\left(m_i \mid m_{i\min}, m_{\max}\right), \\ \mathscr{G}\text{eometric}_{\text{truncated}}\left(m_i \mid \lambda, m_{i\min}, m_{\max}\right), \\ \mathscr{N}\text{egative }\mathscr{B}\text{inomial}_{\text{truncated}}\left(m_i \mid \lambda, \phi, m_{i\min}, m_{\max}\right), \end{cases} \tag{A.1}$$

where λ denotes the mean of the Poisson, geometric and negative binomial distributions, $m_{i\min}$ and $m_{\max}$ denote the minima and maxima of the prior support and ϕ is the dispersion parameter of the negative binomial prior (parameterised according to equation 3.14). Under all four priors, $\lambda = 2.94$, $m_{i\min} = 2$ if $y_{ij} = h$ for some $j \in \{1,\ldots,J\}$ and 1 otherwise, and $m_{\max} = 8$. Under the negative binomial prior, $\phi = 0.5$.

As described in section 4.2.3, the MCMC sampler is run for 50,000 iterations with thinning interval equal to 10 and the first 40% discarded as burnin, generating an MCMC sample, $\{\boldsymbol{\pi}^n, \boldsymbol{a}^n, \boldsymbol{m}^n\}_{n=1}^{3000}$, which approximates the joint posterior with density $\rho\left(\boldsymbol{\pi}, \boldsymbol{a}, \boldsymbol{m} \mid \boldsymbol{y}\right)$. Since the model is fit to each of the 66 subdivisions of the data four times using each of the priors listed above, we have $66 \times 4 = 264$ MCMC sample sets $\{\boldsymbol{\pi}^n, \boldsymbol{a}^n, \boldsymbol{m}^n\}_{n=1}^{3000}$, one for each data subdivision under each prior. Posterior predictive checks based on 3000 replicate datasets, $\{\boldsymbol{y}^{\text{rep}_n}\}_{n=1}^{3000}$, generated from each of the 264 MCMC samples are used to assess model fit. For each MCMC sample set, the nth replicate dataset, $\boldsymbol{y}^{\text{rep}_n}$, is generated by drawing replicate vectors of haplotype counts, $\boldsymbol{a}_i^{\text{rep}_n}$, one patient at a time, from a multinomial distribution,

$$\boldsymbol{a}_i^{\text{rep}_n} \sim \mathscr{M}\text{ultinomial}\left(m_i^n, \boldsymbol{\pi}^n\right) \text{ for } i = 1,\ldots,I. \tag{A.2}$$

(A.3)

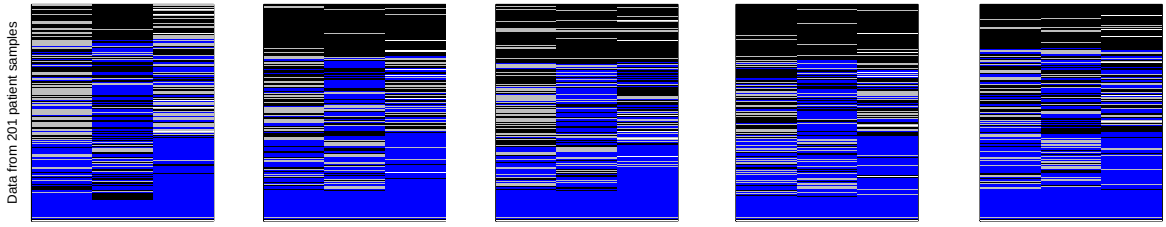
The proportion of mutant alleles at jth SNP per blood sample is then calculated,

$$p_{ij}^{\text{rep}_n} = \frac{\boldsymbol{a}_i^{\text{rep}_n} \cdot \boldsymbol{h}_j}{\sum_{r=1}^{R} a_{ir}^{\text{rep}_n}} \tag{A.4}$$

where $\boldsymbol{h}_j$ is a column vector enlisting the alleles at of the R haplotypes at the jth SNP, and the replicate data generated according to,

$$y_{ij}^{\text{rep}_n} = \begin{cases} w \text{ if } p_{ij}^{\text{rep}_n} = 0 \\ m \text{ if } p_{ij}^{\text{rep}_n} = 1 \\ h \text{ otherwise.} \end{cases} \tag{A.5}$$

To assess the similarity between real and replicate data, for each of the 66 data subdivisions, one replicate dataset per prior is selected at random and plotted alongside the real data (for example, figure A.2). The visual checks suggest replicate data that resemble the real data are probable under all four priors. To quantitatively compare real and replicate data for a given MCMC sample a test statistic is calculated as follows. For a given MCMC sample, we calculate the fraction of replicate datasets for which the proportion of discernibly multiclonal samples (samples with one or more heteroallelic genotyping outcome) exceeds the proportion of discernibly multiclonal samples in the real data. A fraction close to zero or one suggests that data that resemble the real data (with respect to the fraction of discernibly multiclonal samples) are improbable under the model, while a fraction close to 0.5 is indicative of model fit [88]. We generated 264 fractions in total (one for each of the MCMC samples), plotted in figure A.3. The truncated negative binomial and truncated geometric appear to provide the best fit to the data. The average distances from 0.5 per gene are summarised in table A.2. On the whole, the truncated geometric prior provides the best fit for the data.



(a) Real data (b) Replicate data, uniform MOI prior (c) Replicate data, truncated Poisson MOI prior (d) Replicate data, truncated geometric MOI prior (e) Replicate data, truncated negative binomial MOI prior

Figure A.2: An example of a visual posterior predictive check for a single data subdivision (*pfmdr1*, year 2011, DP drug arm) analysed under four different priors on the MOI. The real data are depicted in the left-most panel. Panels two to five (left to right) correspond to replicate datasets generated under models with four different prior distributions on the MOI (equation (A.1)). For each dataset, be it real or replicate, data for a given sample are stored on a given row (see vertical axis of left-most panel for the total number of samples per subdivision). The genotyping outcomes for codons 86, 184 and 1246 in *pfmdr1* are stored in the columns from left to right respectively. For a given sample and codon, blue represents the detection of wild type alleles only, grey represents the detection of mutant type alleles only, black represents the concurrent detection of both wild and mutant type alleles and white represents a missing genotyping outcome. To aid visibility of any patterns in the data, the samples are ordered such that the samples with the most heteroallelic genotyping outcomes are at the top, while the samples with the most pure wild type genotyping outcomes are at the bottom.

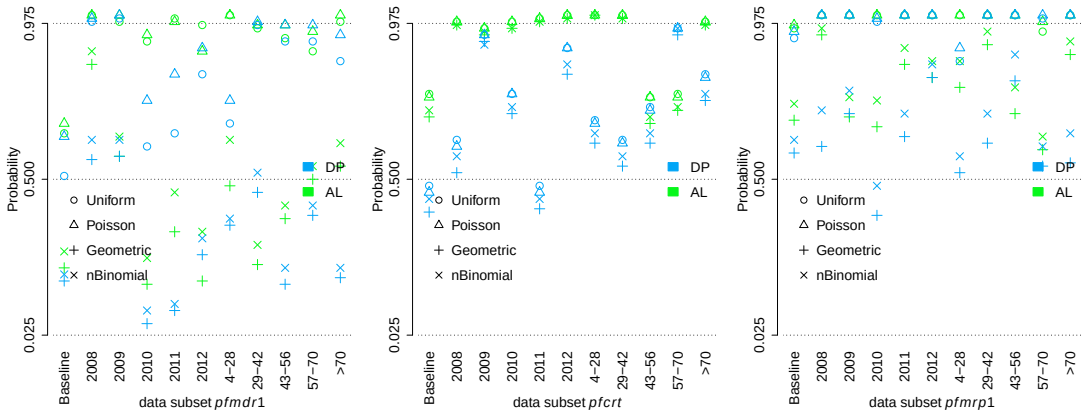


Figure A.3: Graphical summaries of the test statistic (based on the fraction of discernibly multiclonal samples) for all 66 subdivisions fit using four different priors on the MOIs: a uniform, a truncated Poisson, a truncated geometric and a truncated negative binomial (equation A.1), colour-coded by drug arm (DP blue, AL green). Note that for brevity, the term truncated has been dropped from the legends, and nBinomial is shorthand for negative binomial. Test statistics for all 22 subdivisions per gene are represented within a single plot. Different points represent the different priors. Values close to 0.5 are indicative of model fit.

Gene	Uniform	Truncated Poisson	Truncated Geometric	Truncated Negative Binomial
Pfmdr1	0.37	0.41	0.20	**0.18**
Pfmrp1	0.48	0.49	**0.20**	0.26
Pfcrt	0.33	0.33	**0.29**	0.30

Table A.2: Average absolute difference between the test statistic (based on the fraction of discernibly multiclonal samples) and 0.5. A small difference is indicative of model fit with respect to the attributes of the test statistic. The lowest difference per gene is highlighted bold.

A.3 Analytical expressions for posterior quantities

In section 4.2.4 *pfmdr1* haplotype frequencies are regressed onto correlates of drug pressure and drug type. The regression is fit within a Bayesian framework (equation 4.5). Four different prior models are considered, one standard flat prior (equation 4.8), Zellner's g prior (equation 4.9) and two models both based on the normal inverse gamma prior (equation 4.10). Hence, despite there being four prior models, there are only three named distributions, each of which gives rise to a normal inverse gamma posterior with density,

$$\rho\left(\boldsymbol{\beta}_r, \sigma_r^2 | \boldsymbol{\theta}_r\right) = \mathcal{N}\text{ormal}_K(\boldsymbol{\beta}_r \mid \boldsymbol{\mu_r}, \sigma_r^2 \boldsymbol{V}_r) \times \mathcal{I}\text{nverse } \mathcal{G}\text{amma}(\sigma_r^2 \mid a_r, b_r)$$

where $\boldsymbol{\mu_r}$ is the mean and $\sigma_r^2 \times \boldsymbol{V}_r$ is the covariance matrix of the posterior on $\boldsymbol{\beta}_r$ given σ_r^2, and a_r and b_r are the scale and shape parameters, respectively, of the marginal posterior on σ_r^2. The analytical expressions for $\boldsymbol{\mu_r}$, $\boldsymbol{V}_r$, a_r and b_r under the three different distributions are listed below.

1. Under the standard improper flat prior with density $\rho\left(\boldsymbol{\beta}_r, \sigma_r^2\right) \propto \frac{1}{\sigma_r^2}$ (see [4], page 206),

$$\boldsymbol{\mu}_r = \hat{\boldsymbol{\beta}}_r = \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{\theta}_r, \tag{A.6}$$

$$\boldsymbol{V}_r = (\boldsymbol{X}^T\boldsymbol{X})^{-1}, \tag{A.7}$$

$$a_r = \frac{K-P}{2}, \tag{A.8}$$

$$b_r = \frac{S_r}{2}, \text{ where} \tag{A.9}$$

$$S_r = (\boldsymbol{\theta}_r - X\hat{\boldsymbol{\beta}}_r)^T(\boldsymbol{\theta}_r - X\hat{\boldsymbol{\beta}}_r) \tag{A.10}$$

2. Under Zellner's g prior with density $\rho(\boldsymbol{\beta}_r, \sigma_r^2) \propto \mathcal{N}\text{ormal}_K\left(\boldsymbol{\beta}_r \mid \boldsymbol{\mu}_0, c\,\sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}\right) \times \frac{1}{\sigma_r^2}$ (see [4], page 217) ,

$$\boldsymbol{\mu}_r = \frac{1}{(c+1)}\left(\boldsymbol{\mu}_0 + c\hat{\boldsymbol{\beta}}_r\right), \tag{A.11}$$

$$\boldsymbol{V}_r = \frac{c}{(c+1)}(\boldsymbol{X}^T\boldsymbol{X})^{-1}, \tag{A.12}$$

$$a_r = \frac{K}{2}, \tag{A.13}$$

$$b_r = \frac{S_r}{2} + \frac{1}{2}\frac{1}{(c+1)}(\boldsymbol{\mu}_0 - \hat{\boldsymbol{\beta}}_r)^T(\boldsymbol{X}^T\boldsymbol{X})(\boldsymbol{\beta}_0 - \hat{\boldsymbol{\beta}}_r), \tag{A.14}$$

where $\hat{\boldsymbol{\beta}}_r$ and S_r are defined above (equations (A.6) and (A.10) respectively), $c \leftarrow K$ and $\boldsymbol{\mu}_0 \leftarrow (\log(1/7), 0, 0, 0)^T$ for reasons explained in section 4.2.4.

3. Under the normal inverse gamma priors with density $\rho(\boldsymbol{\beta}_r, \sigma_r^2) = \mathcal{N}\text{ormal}_P(\boldsymbol{\beta}_r \mid \mu_0, \sigma^2\boldsymbol{V}_0) \times$

$\mathcal{I}$nverse $\mathcal{G}$amma$(\sigma_r^2 \mid a_0, b_0)$ (see page 246 [187]),

$$\boldsymbol{\mu}_r = \left(X^T X + \boldsymbol{V}_0^{-1}\right)^{-1} \left(X^T \boldsymbol{\theta}_r + \boldsymbol{V}_0^{-1} \boldsymbol{\mu}_0\right), \tag{A.15}$$

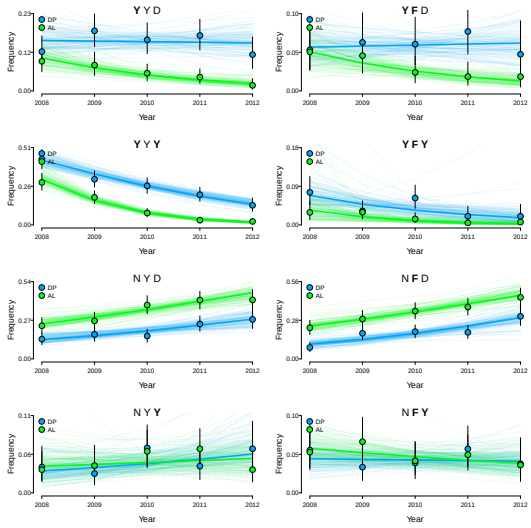
$$\boldsymbol{V}_r = \left(X^T X + \boldsymbol{V}_0^{-1}\right)^{-1}, \tag{A.16}$$

$$a_r = a_0 + \frac{K}{2}, \tag{A.17}$$

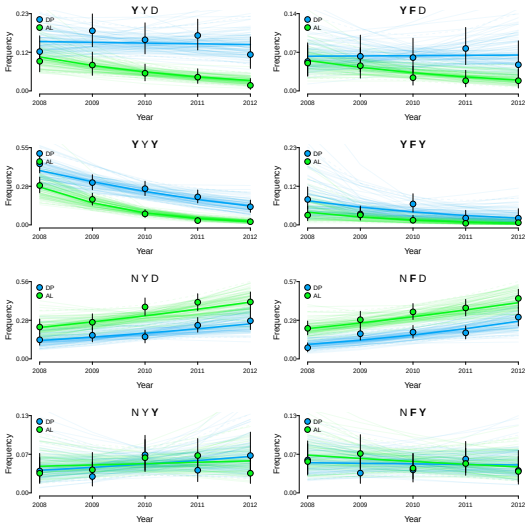
$$b_r = b_0 + \frac{1}{2}\left(\boldsymbol{\theta}_r^T \boldsymbol{\theta}_r + \boldsymbol{\mu}_0^T \boldsymbol{V}_0^{-1} \boldsymbol{\mu}_0 - \mu_r^T \boldsymbol{V}_r^{-1} \boldsymbol{\mu}_r\right). \tag{A.18}$$

where $\boldsymbol{\mu}_0 \leftarrow (\log(1/7), 0, 0, 0)^T$ as above, and $\boldsymbol{V}_0$, a_0 and b_0 are additional hyperparameters also set by the user (see section 4.2.4 for details).

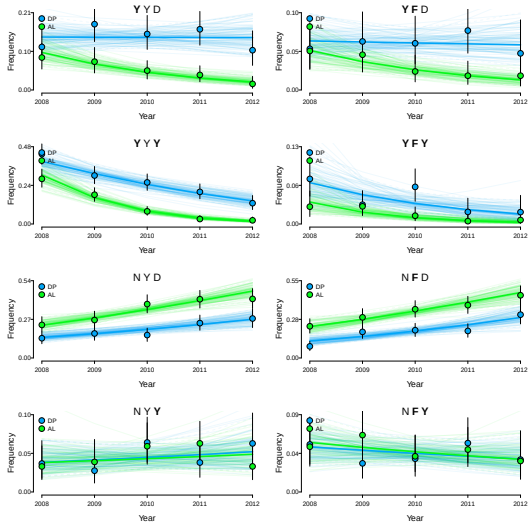
A.4 Plots of trends under different priors



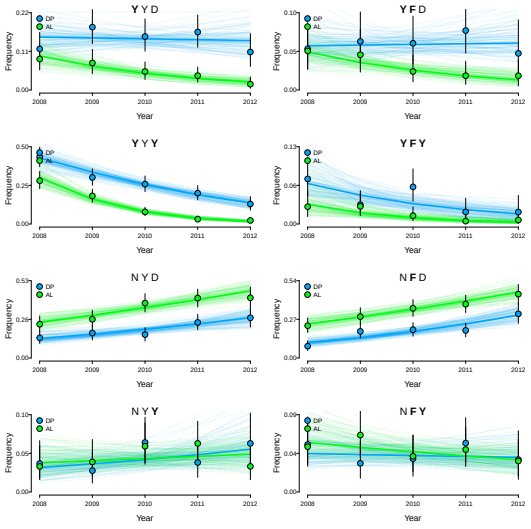
(a) Flat prior



(b) Zellner's g prior

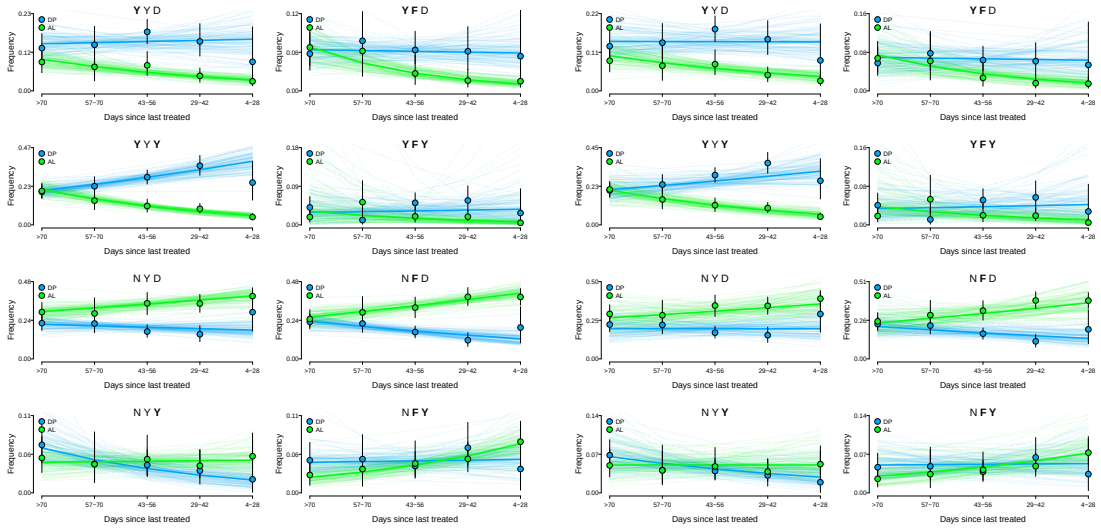


(c) Informative prior with $\boldsymbol{V}_0 = 10 \times I_P$



(d) Informative prior based on recommendation in [87]

Figure A.4: *Pfmdr1* haplotype frequencies (represented by their corresponding amino acid sequences) categorised by drug arm and year regressed onto covariates of drug arm and year under four different priors. Dots denote the MCMC sample estimates of the posterior median haplotype frequencies (DP blue, AL green) before resampling. Vertical black lines denote 95% credible intervals, ranging from the 2.5th percentile to the 97.5th percentile of the MCMC sample before resampling. The regression is performed using four different priors on the regression parameters (see subplot caption). The thick blue and green lines denote the trends constructed using the posterior median estimates of the regression coefficients. The thin blue and green lines represent trends based on 100 traces selected at random from the MCMC sample of regression coefficients.



(a) Standard noninformative prior

(b) Zellner's g prior

(c) Informative prior with $\boldsymbol{V}_0 = 10 \times I_P$

(d) Informative prior based on recommendation in [87]

Figure A.5: *Pfmdr1* haplotype frequencies (represented by their corresponding amino acid sequences) categorised by drug arm and days since last treatment regressed onto covariates of drug arm and days since last treatment under four different priors. Dots denote the MCMC sample estimates of the posterior median haplotype frequencies (DP blue, AL green) before resampling. Vertical black lines denote 95% credible intervals, ranging from the 2.5th percentile to the 97.5th percentile of the MCMC sample before resampling. The regression is performed using four different priors on the regression parameter (see subplot captions). The thick blue and green lines denote the trends constructed using the posterior median estimates of the regression coefficients. The thin blue and green lines represent trends based on 100 traces selected at random from the MCMC sample of regression coefficients.

Appendix B

Model extension

B.1 Evidence of inter-child variation

Introduction

To investigate inter-child variation in the propensity to be infected with different levels of 'mutatedness', cumulative link mixed models are fit to prevalence data collected from a cohort of children in Uganda (see section 4.2.1). By fitting a model with a random effect for each child, an estimate of the inter-child variability is obtained.

Methods

Data: For a given SNP, the genotyping outcome can either be fully wild type (w), partially wild and partially mutant type (h), or fully mutant type (m) (missing data are excluded). Hence, the genotyping outcome for the ith malaria episode, y_i, can be modelled as a multivariate ordinal response: $y_i = 1, 2$ or 3 corresponding to not mutated (w), partially mutated (h) or fully mutated (m), respectively.

Model: The genotyping outcomes, y_i for $i = 1, \ldots, I$, where $I = 1889$, are regressed onto covariates of drug arm and year using the following cumulative link mixed model,

$$\text{logit}\{\mathbb{P}(y_i \leq j)\} = \theta_j - \beta_{\text{drug}} \times \text{drug}_i - \beta_{\text{year}} \times \text{year}_i - \beta_{\text{drug:year}} \times \text{drug}_i \times \text{year}_i - \mu_{\text{child}} \times \text{child}_i,$$
$$\text{for } i = 1, \ldots n \text{ and } j = 1, \ldots, J-1. \tag{B.1}$$

where

- j denotes the 'mutatedness' category: 1 for not at all mutated, 2 for partially mutated and $J = 3$ for fully mutated.
- θ_j is a baseline threshold parameter for the jth mutatedness category. For the average child (a child for whom $\mu_{\text{child}} = 0$) in the dihydroartemisinin-piperaquine (DP) drug arm in 2008:
  - $\mathbb{P}(\text{not mutated}) = \text{logit}^{-1}(\theta_1)$
  - $\mathbb{P}(\text{partially mutated}) = \text{logit}^{-1}(\theta_2) - \text{logit}^{-1}(\theta_1)$
  - $\mathbb{P}(\text{fully mutated}) = 1 - \text{logit}^{-1}(\theta_1) - \text{logit}^{-1}(\theta_2)$.
- The β parameters are regression coefficients, each representing the expected change in the outcome for the average child upon a unit increase in the corresponding covariate.
- μ_{child} is a random effect associated with each child, $\mu_{\text{child}} \sim \mathcal{N}\text{ormal}(0, \sigma^2_{\text{child}})$, where σ^2_{child} represents the inter-child variability.
- drug_i, year_i and child_i are indicator variables for the drug arm, year and child of the ith malaria episode.

Fitting the model: The above model (equation B.1) is fit in R [210], using the `clmm2` function from the R package `ordinal` [48]. Based on the condition number of the Hessian (a measure of identifiability, which, when $> 10^4$, suggests a model is ill-defined and can be simplified),

SNP	θ_1	θ_2	β_{drug}	β_{year}	$\beta_{\text{drug: year}}$	σ^2_{child}
pfmdr1-86	-1.27 (0.12)	-0.16 (0.11)	-0.73 (0.16)	-0.27 (0.05)	-0.38 (0.07)	0.12
pfmdr1-184	0.15 (0.09)	2.14 (0.10)	0.42 (0.09)	0.16 (0.03)	-	-
pfmdr1-1246	-0.79 (0.09)	0.10 (0.09)	-0.75 (0.09)	-0.32 (0.03)	-	-
pfmrp1-876	-0.30 (0.09)	0.86 (0.09)	-	0.10 (0.03)	-	0.08
pfmrp1-1466	-0.87 (0.05)	0.33 (0.05)	-	-	-	-
pfcrt-76	-4.78 (0.29)	-4.18 (0.28)	-0.69 (0.22)	-0.42 (0.08)	-	0.24

Table B.1: Parameter estimates (with standard errors in parentheses) of cumulative link mixed models based on equation (B.1). All β and σ^2_{child} parameters included in the models are statistically significant at the 5% level aside from σ^2_{child} for *pfcrt*-76 (p-value $= 0.13$).

the full model does not fit the data for all but one nSNP. Consequently, small and statistically insignificant parameters are dropped. The log-likelihood ratio test is used to assess the statistical significance at the usual 5% level.

Results and conclusion

The maximum likelihood estimates of the parameters of the selected models are summarised in table B.1. Intra-child variability is found to be statistically significant at the 5% level for two of the six nSNPs investigated: *pfmdr1*-86 (p-value = 0.02) and *pfmrp1*-876 (p-value = 0.05). For three nSNPs (*pfmdr*-184, *pfmdr*-1246 and *pfmrp1*-1466), inter-child variability is negligible ($\hat{\sigma}^2_{\text{child}} < 0.01$) and statistically insignificant (p-values ≥ 0.30), thus dropped from the fitted models. For *pfcrt*-76, intra-child variability was comparatively large ($\hat{\sigma}^2_{\text{child}} = 0.23$), but statistically insignificant at the 5% level (p-value = 0.13). In conclusion, inter-child variability, and therefore inter-child variation in the propensity to be infected with different levels of 'mutatedness', is detected, but is either negligible or small.

B.2 Selection of individual SNP subdivisions

The aim of this exercise is to select a set of approximately ten individual-SNP data subdivisions based on evidence of intra-child variability. The individual-SNP data subdivisions are based on

the gene-wise data subdivisions depicted in panels D and E of figure 4.3, further subdivided by SNP. There are 120 subdivisions in total (see row and column headings, table B.2). Following the preliminary study described above (section B.1), evidence is based on fitting a model with a random effect for each child, and assessing the statistical significance of the inter-child variability. More specifically in this exercise, the following two models, one with a random effect for each child (equation B.2) and another without (equation B.3),

$$\text{logit}\{\mathbb{P}(y_j \leq j)\} = \theta_j - \mu_{\text{child}} \times \text{child}_i \text{ for } i = 1,\ldots,I \text{ and } j = 1,\ldots,J-1, \tag{B.2}$$

$$\text{logit}\{\mathbb{P}(y_i \leq j)\} = \theta_j \text{ for } i = 1,\ldots,I \text{ and } j = 1,\ldots,J-1, \tag{B.3}$$

(where y_i, j and J, θ_j, μ_{child} and child_i are defined above in section B.1, and I is the number of episodes per data subdivision), are fit to each of the 120 individual-SNP subdivisions and the fit compared using the standard log-likelihood ratio test. Off all 120 subdivisions, 12 had statistically significant inter-child variability estimates at the 15% level (p-values highlighted in bold, table B.2).

Subdivision	*pfmdr*-86	*pfmdr*-184	*pfmdr*-1246	*pfmrp1*-876	*pfmrp1*-1466	*pfcrt*-76
2008 DP	**0.08 (0.75)**	~~1.00 (0.00)~~	0.57 (0.19)	~~1.00 (0.00)~~	~~1.00 (0.00)~~	~~1.00 (0.00)~~
2009 DP	~~1.00 (0.00)~~	~~1.00 (0.00)~~	~~1.00 (0.00)~~	0.94 (0.02)	~~1.00 (0.00)~~	~~1.00 (0.00)~~
2010 DP	0.59 (0.18)	~~1.00 (0.00)~~	~~1.00 (0.00)~~	~~1.00 (0.00)~~	~~1.00 (0.00)~~	~~1.00 (0.00)~~
2011 DP	~~1.00 (0.00)~~	0.30 (0.39)	0.68 (0.13)	~~1.00 (0.00)~~	0.73 (0.09)	~~1.00 (0.00)~~
2012 DP	0.42 (0.35)	~~1.00 (0.00)~~	0.55 (0.28)	~~1.00 (0.00)~~	~~1.00 (0.00)~~	~~1.00 (0.00)~~
2008 AL	~~1.00 (0.00)~~	0.25 (0.32)	**0.12 (0.39)**	0.17 (0.41)	~~1.00 (0.00)~~	0.67 (0.54)
2009 AL	**0.03 (0.88)**	0.75 (0.07)	**0.10 (0.62)**	~~1.00 (0.00)~~	~~0.98 (0.01)~~	~~1.00 (0.00)~~
2010 AL	**0.01 (1.42)**	0.33 (0.32)	~~1.00 (0.00)~~	0.24 (0.36)	~~1.00 (0.00)~~	0.79 (0.30)
2011 AL	~~1.00 (0.00)~~	~~1.00 (0.00)~~	~~1.00 (0.00)~~	0.46 (0.21)	0.75 (0.09)	**0.13 (2.31)**
2012 AL	~~1.00 (0.00)~~	0.91 (0.04)	**0.08 (1.15)**	0.68 (0.12)	~~1.00 (0.00)~~	0.61 (0.26)
$>$ 70 days DP	~~1.00 (0.00)~~	~~1.00 (0.00)~~	0.60 (0.13)	~~1.00 (0.00)~~	0.84 (0.05)	~~1.00 (0.00)~~
57–70 days DP	~~1.00 (0.00)~~	0.93 (0.03)	0.82 (0.06)	~~1.00 (0.00)~~	0.31 (0.42)	~~1.00 (0.00)~~
43–56 days DP	~~1.00 (0.00)~~	~~1.00 (0.00)~~	~~1.00 (0.00)~~	0.60 (0.11)	0.89 (0.02)	~~1.00 (0.00)~~
29–42 days DP	0.84 (0.05)	0.40 (0.33)	0.96 (0.01)	**0.03 (0.93)**	~~1.00 (0.00)~~	~~1.00 (0.00)~~
4–28 days DP	0.55 (0.82)	0.69 (0.41)	0.55 (0.89)	~~1.00 (0.00)~~	0.75 (0.21)	0.93 (0.13)
$>$ 70 AL	~~0.96 (0.01)~~	0.42 (0.26)	~~0.96 (0.01)~~	0.34 (0.31)	0.55 (0.19)	0.57 (0.74)
57–70 days AL	~~1.00 (0.00)~~	**0.08 (3.57)**	~~1.00 (0.00)~~	0.33 (1.11)	~~1.00 (0.00)~~	~~1.00 (0.00)~~
43–56 days AL	0.94 (0.03)	~~1.00 (0.00)~~	~~1.00 (0.00)~~	0.87 (0.06)	~~1.00 (0.00)~~	**0.06 (6.47)**
29–42 days AL	**0.01 (1.10)**	~~1.00 (0.00)~~	~~1.00 (0.00)~~	0.16 (0.31)	~~1.00 (0.00)~~	**0.06 (2.18)**
4–28 days AL	0.33 (0.28)	0.91 (0.02)	0.21 (0.30)	0.41 (0.16)	~~1.00 (0.00)~~	~~1.00 (0.00)~~

Table B.2: Tail probabilities of cumulative link mixed models fit to nSNP-wise subdivisions. More specifically, the p-values (and inter-child variability estimates, $\hat{\sigma}^2_{\text{child}}$, in parentheses), derived from the log-likelihood ratio test comparing the model with (equation B.2) and without (equation B.3) a random effect fit to each of the 120 individual-SNP subdivisions of the Ugandan data, where subdivisions are defined by either the year, or the number of days since last treatment, and the drug arm (DP, dihydroartemisinin-piperaquine; AL, artemether-lumefantrine). Values corresponding to ill defined random effect models (models with Hessian condition number $> 10^4$) are crossed out. There were no ill defined intercept only models (the maximum condition number was 25). Values for which the p-value < 0.15 are highlighted bold.