Global and local patterns of population structure and their role in the evolution and demography of *Plasmodium falciparum*

Jacob Almagro-Garcia

Green Templeton College

University of Oxford

Thesis submitted in partial fulfilment of the requirements for the degree of

Doctor of Philosophy

Michaelmas Term 2016

Abstract

Global and local patterns of population structure and their role in the evolution and demography of *Plasmodium falciparum*

Jacob Almagro-Garcia, Green Templeton College, *D.Phil.* Michaelmas 2016

In this thesis, I study the role of genetic population structure in the evolution and demography of *Plasmodium falciparum* by focusing on the recent onset of artemisinin resistance in Southeast Asia, an alarming event for global public health. I describe the population structure of *Plasmodium falciparum* in the Thai-Cambodian border region, characterizing sympatric but differentiated subpopulations associated with artemisinin resistance. I show evidence that they are the product of recent founder events and seem the primary force spreading resistance. Next, I study a superset of the *kelch13* mutations associated with artemisinin resistance, assessing their relationship with population structure and recent founder effects. Each resistant subpopulation possesses a distinct *kelch13* allele that, in conjunction with a particular genetic background, seem to have driven recent founder effects. I examine the demography of these resistance alleles using patterns of haplotype sharing and show that the primary mode of spread consists of independent mutational events, with limited gene flow within countries in East Southeast Asia. Subsequently, I assess the origin of *kelch13* mutations observed in African isolates, concluding that they are indigenous and have originated independently. These observations undermine localized resistance containment as a strategy for malaria control and suggest that population structure and founder effects may predate and facilitate the emergence of resistance. Therefore, monitoring these phenomena could warn about the development of resistance before phenotypic evidence materializes. Next, given the importance of demographic inference to inform malaria control programs and the advent of large genomic datasets, I develop a fast and scalable method to build the ancestral haplotype graph. I show that this data structure, composed of a collection of local haplotype trees, is informative about the recent genealogical history of the sequences and can be used to summarize and study shared haplotype patterns along the genome. I describe a set of algorithms with quasilinear time complexity as a first step in the development of scalable demographic inferential methods that can be applied to several thousands of sequences. I also evaluate how mixed infections affect the analysis of deep sequencing data and review the F_{WS} statistic, a relative measure of inbreeding and complexity of infection. In doing so, I show that the original F_{WS} estimator discards the diversity encoded by rare variants and provide an alternative estimator without such bias that is simpler, more intuitive and has a better resolution.

Statement of authorship

This thesis represents my original work. However, Chapters 2 and 4 show results published in large collaborative projects. Although these chapters focus on my contributions, to help the scientific narrative, some results for which I am not responsible or that are the product of collaborations with other colleagues have been included. I clarify my contributions at the beginning and end of each chapter so they can be adequately quantified. As a general summary, I also list here my individual contributions in the chapters that include results.

- Chapter 2: My work involved the identification and characterization of population structure, its relationship with drug resistance, and the elucidation of plausible demographic scenarios. I specifically worked with the first author of the study (Olivo Miotto) during the exploratory data analysis phase, characterized population structure with model-based methods on my own and, again, collaborated with Olivo during the evaluation of plausible demographic scenarios.

- Chapter 3: All the material represents my original work.

- Chapter 4: In the first part, my contributions were limited to the study and characterization of population structure, the relationship of founder populations with artemisinin resistance and the assessment of the most plausible demographic scenario that could explain the observed geographical spread of *kelch13* mutations. Specifically, I characterized population structure on my own and collaborated with Olivo Miotto to generate our conservative classification of subpopulations, discerning their relationship with drug resistance. I performed the demography assessment of *kelch13* mutations, using patterns of haplotype sharing. In the second part, my contribution exclusively consisted on assessing the putative origin of the *kelch13* mutations observed in African samples.

- Chapter 5: All the material represents my original work.

Acknowledgements

This thesis is dedicated to my mom, Maria del Carmen Almagro-Garcia, and to all the people that work to alleviate the hardship of human disease.

I would like to thank my supervisors, Dominic Kwiatkowski and Gilean McVean, for their guidance, support, infinite patience and, most of all, for trusting me with a spectacular opportunity for doing science. I am also indebted to the whole malaria group for their unconditional support at all levels. Among others, I am particularly grateful to the admin team (Vikki, Claire, Christa) for always saving me from paperwork hell and, science-wise, to Olivo Miotto and Roberto Amato for uncountable hours of stimulating work and discussion.

Finally, I need to thank all my friends for helping me to survive this long and exhausting ride. Especially Carla, who took the worst of it, and Luiz, with whom I shared many sleepless nights of work.

Contents

List of Figures

List of Tables

List of Algorithms

Acronyms

ACT	Artemisinin combination therapy
GWAS	Genome-wide association study
MoI	Multiplicity of infection
CoI	Complexity of infection
WHO	World Health Organization
LD	Linkage disequilibrium
PCA	Principal component analysis
PC	Principal component
PCoA	Principal coordinate analysis
NJ	Neighbor joining
HMM	Hidden Markov model
MCMC	Markov chain Monte Carlo
AFS	Allele frequency spectrum
MAF	Minor allele frequency[1]
IBD	Identity by descent

[1] Refers to the population-wide minor allele frequency unless specified otherwise.

IBS	Identity by state
HL	Half-life
PLAF	Population-level allele frequency
WSAF	Within-sample allele frequency
RMSE	Root mean squared error
TRAC	Tracking Resistance to Artemisinin Collaboration
MRCA	Most recent common ancestor
SEA	Southeast Asia
WSEA	West Southeast Asia
ESEA	East Southeast Asia
NNH	Nearest neighbor haplotype
AHG	Ancestral haplotype graph
ARG	Ancestral recombination graph
SHG	Subhaplotype graph
PBWT	Positional Burrows-Wheeler transform

Chapter 1

Introduction

Contents

1.1 Introduction

Why has malaria not been eradicated or brought under control in many countries? The disease predates modern humans [Su et al., 2003] and yet, despite enormous global efforts during the last century, is still an important public health problem. A myriad of factors makes malaria hard to confront. Political, economic and social aspects play a crucial role in countries where the disease is endemic. It is not a coincidence that malaria has become associated with poverty, and it is perceived nowadays as "a disease of the poor" [Worrall et al., 2005], [Gallup and Sachs, 2001]. Poverty imposes a heavy burden regarding access to resources, which translates into severe logistic problems for control programs and interventions, for example in remote rural areas [Sachs and Malaney, 2002]. Although socioeconomic forces are essential to understanding the prevalence of malaria and the difficulty of controlling it, we also need to highlight the role of its complex biology. The disease has a complicated life cycle that prompts an evolutionary arms race between parasites, mosquitoes, and human hosts. The quick evolutionary response to drugs from parasites and to insecticides from vectors adds an extra level of complexity to the problem.

Nowadays, important advances in sequencing technology and computational methods allow researchers to run large-scale analysis of genetic variation in natural populations of parasites [Manske et al., 2012]. Studies with a substantial number of samples distributed across different geographical regions are fundamental for tracking the evolution and demography of the parasite, specifically regarding the surveillance of drug resistance at global and regional scale [Volkman et al., 2012].

In this thesis, we study the role of genetic population structure in the evolution and demography of *Plasmodium falciparum*. In particular, we focus on the recent onset of artemisinin resistance detected in Southeast Asia, an event of severe gravity for global public health [Dondorp et al., 2009]. First, following reports of delayed clearance for artemisinin derivatives, we study patterns of genetic variation for *P. falciparum* in the Thai-Cambodian border region (Chapter 2), an area of historical importance in the emergence of resistance [Mita et al., 2009], [Wongsrichanalai et al., 2002], [Takala-Harrison et al., 2015]. There, we examine several sympatric but very differentiated subpopulations of parasites associated with artemisinin resistance that were, most likely, the product of founder effects.

Later, we extend this line of research and study a superset of the mutations associated with artemisinin resistance (Chapter 4). We provide a genetic epidemiological assessment of the spread of artemisinin resistance in Southeast Asia, including the roles played by population structure and founder effects, the genetic makeup that tends to promote the emergence of resistance, and the demographic events that may have modulated the spread of resistance alleles. Furthermore, we assess the origin of the resistance mutations observed in African isolates. Next, given the importance of demographic inference to inform malaria control programs and the advent of large genomic datasets, we develop a fast and scalable method to build the ancestral haplotype graph (Chapter 5). This data structure, composed of a collection of local haplotype trees, is informative about the recent genealogical history of the sequences and can be used to summarize and study haplotype patterns along the genome. Additionally, we also consider how complexity of infection affected the analysis of deep sequencing data and reviewed the F_{WS} statistic, showing that the original estimator discards the diversity encoded by rare variants and introducing an alternative estimator that do not have such bias, is simpler, more intuitive and has a better resolution (Chapter 3). Finally, we discuss how all the findings presented in this thesis relate to each other and lay out a tentative roadmap for further research (Chapter 6).

For the majority of this thesis, I have chosen to use the editorial *we* over the first person pronoun *I*. The main reason is the collaborative nature of the large-scale analyses included here. Some of the findings presented in Chapters 2 and 4 are the product of the interaction and discussion with many scientists, including my supervisors. I believe that the use of *I* could make the reading tiring and dull, as I would have to give specific attribution for each result or hide the agent in the sentence using the passive voice. Instead, I have included a section in each chapter precisely delineating the limits of my work, so my actual contributions can be quantified. To preserve the narrative coherence, I have also used the editorial *we* in Chapters 3 and 5, despite the work presented there being exclusively my own.

This introductory chapter reviews the main aspects of human malaria including its biology and historical context. We pay particular attention to *P. falciparum* drug resistance and population structure, and presents an overview of the technical background material required to follow this thesis.

1.2 Human malaria

Malaria is a disease caused by parasites that are transmitted by female *Anopheles* mosquitoes when they feed on blood meals. Four species of protozoan parasites of the genus *Plasmodium* (*Plasmodium malariae*, *Plasmodium ovale*, *Plasmodium vivax* and *Plasmodium falciparum*) are responsible for causing human malaria [Carter and Mendis, 2002]. Although *Plasmodium knowlesi* has been also found in human infections, it mainly infects primates and the mechanism underlying transmission to humans still remains unclear [Singh et al., 2004].

Plasmodium falciparum is the most dangerous species, and it is responsible for the majority of deaths associated with malaria [WHO, 2015b]. It goes through a very complex life cycle that comprises diploid and haploid stages. *P. falciparum* has a nuclear genome of 24Mb, formed by 14 linear chromosomes. Two additional circular chromosomes correspond to the mitochondrial and apicoplast genomes [Gardner et al., 2002]. Sexual reproduction of the parasite occurs in the *Anopheles* mosquito. Within human hosts, the parasite takes a haploid form and undergoes asexual reproduction. This process of clonal expansion is what leads to clinical manifestations of the disease. Characteristic symptoms include rigors, chills and perspiration due to periodic fever bouts, usually accompanied by headaches, abdominal discomfort, mild anemia and general weakness [WHO, 2015b].

1.2.1 The life cycle of *Plasmodium falciparum*

This section is a brief summary of the material found in [Greenwood et al., 2008], [Bousema et al., 2014], [Galizi et al., 2014], [Cowman et al., 2012] and [Schlagenhauf-Lawlor, 2007]. *P. falciparum* parasites undergo a complex multistage life cycle (Figure 1.1), with the asexual phase taking place on the human host and the sexual stage occurring in the mosquito vector. The cycle starts with an infected female *Anopheles* mosquito taking a blood meal and injecting sporozoites[1], tens to a few hundreds, into the bloodstream of the human host. These are the infectious form of the parasite. The sporozoites travel to the liver within a few hours and invade hepatocytes (i.e. liver cells). Some species, like *P. vivax* also produce dormant hypnozoites that remain in the liver and can trigger relapses weeks or even months later.

[1] From the Greek Sporos, meaning seeds.

In the liver, each sporozoite develops into a schizont that contains between 10,000 and 30,000 merozoites, generated by asexual replication. The term merozoite[2] refers to the form of the parasite that invades erythrocytes (i.e. red blood cells or RBCs). The merozoites enter the bloodstream and begin repeated cycles of invasion, replication and release. Invaded RBCs transition through different phases: ring stage, trophozoite[3] and mature erythrocytic schizont, which ruptures and releases between 8 and 32 new merozoites that start the cycle again. This replication cycle occurs every 48 hours in the case of *P. falciparum*, causing periodic fevers and chills. The population of parasites can grow in this fashion to more than 10^{13} parasites per host. A fraction of the merozoites exit the asexual cycle and undergo gametocytogenesis. Only mature gametocytes circulate in peripheral blood, as immature counterparts are sequestered in the bone marrow.

Gametocytes are not pathogenic, and it is the only form of the parasite that can be transmitted from humans to mosquitoes, where the sexual stage of the life-cycle is completed. In the mosquito, gametocytes develop into female or male gametes and travel into the midgut. Gametes fuse forming diploid zygotes that develop into motile ookinetes. In this form, the parasites penetrate the mosquito midgut wall and mature into oocysts on the other side. The oocysts grow and burst after 8-15 days, releasing thousands of haploid sporozoites that will travel to the salivary glands of the mosquito. The mosquito can then infect humans when taking a blood meal, starting the cycle again.

1.2.2 Vectors of human malaria

There are around 70 species of mosquitoes that act as vectors for human malaria, all of them belonging to the genus *Anopheles*. From them, 41 species are considered to be dominant malaria vectors [Sinka et al., 2012]. Only female mosquitoes can transmit malaria as they feed on human blood during egg production. Rainfall, temperature, and humidity are among the major factors influencing transmission and maturation of malaria parasites within the mosquito, with warmer temperatures being associated with higher transmission [Blanford et al., 2013]. A limiting factor for transmission is the lifespan of the infected mosquito, as parasites need to undergo sexual reproduction and mature after inoculation of the

[2]From the Greek Meros, meaning piece.
[3]From the Greek Trophes, meaning nourishment.

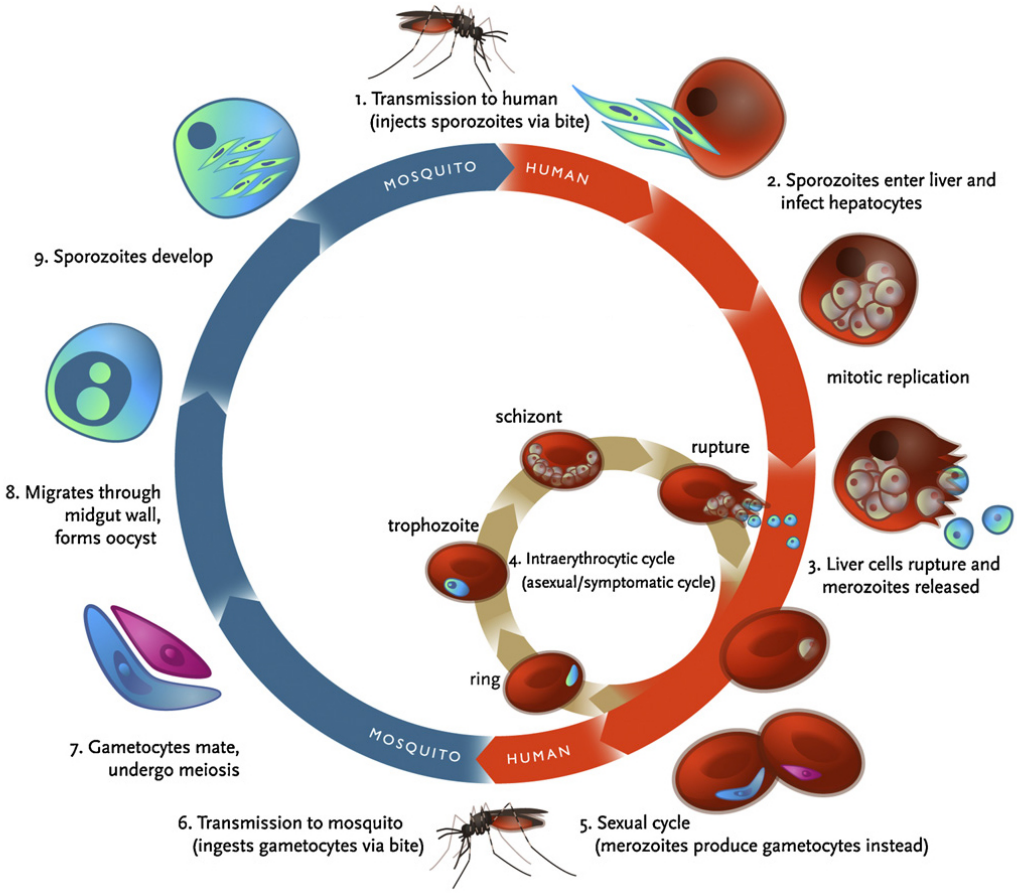

Figure 1.1: Diagram showing the life cycle of *Plasmodium falciparum*. Source: Klein EY. Antimalarial drug resistance: a review of the biology and strategies to delay emergence and spread. Int Antimicrob Agents (2013), `http://dx.doi.org/10.1016/j.ijantimicag.2012.12.007`

vector [Smith and McKenzie, 2004].

Mosquitoes do not benefit from being infected and it has been shown they are able to develop a memory-like immune response when facing repeated infection [Rodrigues et al., 2010]. However, there is evidence indicating they have better survival during starvation because of an increased storage of energy resources [Zhao et al., 2012]. It has also been shown in mice that malaria infection makes the host more attractive to *Anopheles* mosquitoes [De Moraes et al., 2014].

Anopheles mosquitos can be found anywhere on the globe except in Antarctica, including regions where malaria has been eliminated. Most *Anopheles* species are active at dusk/dawn or during the night, but they differ in their feeding/resting preferences. Mosquito behavior influences malaria transmission and control strategies. Species that prefer to feed indoors (endophagic) during the night (nocturnal) can be controlled by using insecticide-treated bed nets (ITNs). Likewise, mosquitoes that prefer to rest indoors (endophilic) can be controlled by indoor residual spraying[4] of insecticides. For species that feed or rest outside, the most effective control measures involve larval source management (e.g. eliminating breeding sites). Another important factor that modulates transmission is the preference of some species to feed on humans (anthropophily) instead of cattle or wild animals (zoophily). *Anopheles* species cover almost the whole spectrum between antropophily and zoophily. Nonetheless, the most efficient malaria vectors (such as *A. Gambiae* and *A. Funestus* in Africa) prefer to feed on humans almost exclusively [Elliott, 1972], [Muriu et al., 2008], [Tirados et al., 2006] and [Githeko et al., 1996].

1.2.3 Complexity of infection

Malaria infections are often triggered by a set of genetically distinct parasites acquired from several mosquito bites, the bite of a single mosquito infected with multiple strains or a combination of both scenarios. The number of different strains infecting a human host during a single infection is termed complexity of infection (COI) or multiplicity of infection (MOI) in the literature. Here we consider strains as genetically distinct[5] parasites and use

[4]Indoor residual spraying (IRS) consists of applying insecticide on walls or any surface that may act as resting place for mosquitoes.

[5]The meaning of genetically distinct is open to debate, as even the clonal population expansion of parasites within the human host produces genetic differences in the form of mutations or structural rearrangements, for instance regarding antigenic variation [Mendis et al., 1991], [Claessens et al., 2014].

indistinctly the terms multiple or mixed infection to refer to malaria infections with an MOI/COI > 1.

Since complexity of infection is positively correlated[6] with local transmission intensities [Arnot, 1998], [Nkhoma et al., 2013], its prevalence across populations is informative about differences in transmission regimes, underlying epidemiological factors or the success of an ongoing intervention [Volkman et al., 2012]. Despite evidence supporting the hypothesis that acquired immunity is specific to particular parasite strains [Laishram et al., 2012], the impact of COI in terms of disease development is not clear, with different studies presenting conflicting evidence [Branch et al., 2001], [de Roode et al., 2005], [Müller et al., 2001]. Furthermore, it has been suggested that within-host competition between co-infecting strains could accelerate the spread of drug resistance [de Roode et al., 2004], [Hastings and D' Alessandro, 2000].

The presence of polymorphic genotypes (indicating a COI > 1) in field samples complicates the study of parasite DNA sequences, making fine-scale analysis of genetic variation challenging and confounding methods that rely on haplotype data to detect selection, infer genealogical relationships or study recent demographic events among other phenomena. In this thesis, we use different strategies to deal with this difficulty and assess the potential biases introduced by them (Chapter 3).

1.2.4 Immunity

The risk of death, the severity of the symptoms and the number of parasites in blood are modulated by the immune response of the host. Acquired immunity requires frequent infections to develop, and it disappears after a short time without being exposed malaria-infected mosquitoes [Doolan et al., 2009]. The fact that immunity depends on the amounthods of exposure to malaria helps to explain the rise of epidemics in regions of unstable malaria endemicity, where individuals develop very low immunity or none at all [Romi et al., 2002], [de Zulueta, 1988], [Dowling et al., 1951]. It also explains why children under the age of four -and thus without time to acquire immunity- account for most deaths in regions of stable malaria [Doolan et al., 2009], [Carter and Mendis, 2002], [Snow et al., 1997].

[6] With higher transmission rates being associated with higher prevalence of mixed infections.

P. falciparum goes through programmed antigenic variation[7] to avoid the immune response. The erythrocyte membrane protein 1 (PfEMP1)[8] changes in 2-3% of the parasite population in each asexual cycle, causing a significant fraction of parasites to share the same antigenic phenotype in successive cycles [Kyes et al., 2001], [White, 2004].

In the context of immunity, J. B. S. Haldane proposed in 1948 the malaria hypothesis [Haldane, 1949]. Haldane suggested that a set of genetic polymorphisms, in particular those affecting red blood cells, had been selected because of their protective effects against the disease. Some of these genetic polymorphisms include thalassemias (defects in the genes responsible for hemoglobin), G6PD deficiency, sickle cell trait, ovalocytosis or red blood cell Duffy negativity (makes carriers refractory to *Plasmodium vivax* infections). Several studies have confirmed the malaria hypothesis by correlating the geographical dispersion and frequency of the polymorphisms with malaria endemicity. In some cases there is also evidence supporting convergent evolution, with the same mutation arising independently several times [Hedrick, 2011], [Kwiatkowski, 2005], [Flint et al., 1998], [Cavalli-Sforza et al., 1994].

1.2.5 The burden of malaria

Malaria imposes a heavy toll at many levels, not only in terms of human lives. The presence of the disease has been regarded as a key factor in the stagnation and decline of economies throughout history and still poses a significant economic burden on many endemic countries. There is a strong correlation between poverty and the presence of malaria, and different studies have shown that economic growth is slower in countries where the disease is endemic [Sachs and Malaney, 2002], [Gallup and Sachs, 2001].

The mortality associated with the disease revolves around 0.1% for uncomplicated cases when adequate treatment is administered. However, it rises up to 15-20% in cases of severe malaria. Deaths usually occur within 48 hours of presenting symptoms, during the first asexual cycle of the parasite [White, 2004]. If malaria infection is not treated, it leads to a characteristic enlargement of the spleen and disrupts the function of vital organs.

The disease can also leave serious sequels, such as mental impairment in the case of

[7] Of the epitopes expressed on the red blood cell surface.

[8] This kind of proteins is encoded by a family of approximately 60 var genes, located mainly in subtelomeric regions.

cerebral malaria [Carter and Mendis, 2002], [Carme et al., 1993], [Holding and Snow, 2001], which adds to its burden.

1.2.5.1 Historical context

The oldest written reference to malaria seems to appear in a Chinese medicine text known as "Nei Ching" (The Canon of Medicine), around 2700 BC [Neghina et al., 2010]. The descriptions of the fever episodes may suggest the presence of *Plasmodium vivax* or *P. malariae* [Needham, 1960]. Other references to the disease are mentioned in the old texts of many civilizations, such as Sumerian, Egyptian, Roman or Greek [Carter and Mendis, 2002], [Neghina et al., 2010]. Historians agree on malaria being a common disease in Rome due mainly to its favorable ecological conditions for the breeding of mosquitoes [Sallares, 2002]. The disease spread from the Mediterranean territories during the 16^{th} century, reaching the north of Europe. In the 16^{th} and 17^{th} centuries, *Plasmodium falciparum*, *P. vivax* and maybe *P. malariae* were brought to the American continent by the European colonizers and the trade of African slaves [Boyd, 1941]. The term malaria derives from the Italian expression "mal'aria" that refers to bad or corrupted air. Francesco Puccinotti used the modern form "malaria" for the first time in a book published in 1838 [Neghina et al., 2010], [Sallares, 2002], [Snowden, 2008].

The disease reached global expansion in the 19^{th} century and put more than half of the world population at risk, exhibiting high mortality rates (10% of infected individuals) [Carter and Mendis, 2002]. In the 20^{th} century, malaria was responsible for the death of 150-300 million people, accounting for 2-5% of all worldwide deaths. In the second half of that century, the disease receded at a global scale outside Africa, with the risk of dying from malaria plummeting to 1% of the risk present at the beginning of the century. Nevertheless, in Africa the odds remained roughly stable, reaching one million deaths during the decade of the 90's [WHO, 1999].

1.2.5.2 Eradication and control efforts

The Global Malaria Eradication Program began in 1955 following local eradications campaigns that run during the previous decade. The potential of the DDT[9] insecticide and the

[9] Dichlorodiphenyltrichloroethane

efficacy of the antimalarial drug chloroquine raised the expectations of tackling malaria over large geographical areas [Litsios, 1996]. The campaign had some success in eliminating malaria from particular regions, such as Europe, North America and some parts of Asia and South America, but it failed in sub-Saharan Africa and other malaria-endemic territories.

The effort faced many difficulties, including logistic and technical problems in underdeveloped regions, the spread of insecticide resistance, the displacement of large masses of people due to armed conflicts and the emergence of chloroquine resistance in Southeast Asia and South America. All these factors caused the Global Malaria Eradication program to be abandoned in 1969 [Nájera et al., 2011]. Despite the failure of the campaign in Africa, as a net result, 27 countries were declared malaria-free by the end of the operations [Mills et al., 2008]. The global eradication plan was replaced with malaria control strategies in the 1970s. During that decade the situation aggravated: resistance to chloroquine spread while vectors became insensitive to DDT [Nájera et al., 2011], [Hay et al., 2004].

Starting with the launch of The Roll Back Malaria initiative by the WHO, the 1990s saw a renewed global effort on malaria control [WHO, 2007] that has been maintained till today. Nowadays, the gravity of the current wave of artemisinin resistance in Southeast Asia has forced the WHO to set up a plan for the elimination of resistant parasites in the Great Mekong region [WHO, 2015a].

1.3 Drug Resistance

The core of malaria control programs rests on a tiny set of tools. Control of *Anopheles* mosquitoes relies nowadays on indoor insecticide spraying, bed nets, and removal of breeding sites. Despite progress in recent years, the promise of genetically modified mosquitoes for vector control still has to materialize [WHO, 2015b], [Galizi et al., 2014], [Okumu et al., 2011], [Marshall and Taylor, 2009], [Knols et al., 2007].

Treatment and prevention of disease make use of antimalarial drugs and rapid diagnostic tools, where prevention includes strategies like chemoprophylaxis or mass drug administration [WHO, 2015b]. Given that the only vaccine that could be available soon (RTS, S) only offers modest protection[10] in children [RTS-SCTP, 2015], the role played by antimalarial

[10]Efficacy ranges from 26% to 50% in infants and young children. A pilot implementation trial is due to start in three African countries in 2018.

drugs is crucial. This dependency makes prediction and surveillance of the emergence and spread of drug resistance indispensable for malaria control and future elimination.

1.3.1 Drug resistance in *P. falciparum*

Plasmodium falciparum has a proven track record of evolving drug resistance. Chloroquine, which used to be the gold standard for treating uncomplicated malaria, is nowadays useless in almost all regions due to resistance. The first cases of resistance to chloroquine were detected in Thailand in 1957. Resistance spread through Asia and by the 1970s reached Africa and South America. Other agents had the same fate. In fact, the parasite has developed resistance to nearly all drugs that are available, sometimes within just a few years -or even months- after they were introduced [Cui et al., 2015], [Packard, 2014], [Petersen et al., 2011]. Figure 1.2 summarizes how long resistance took to emerge after the most common anti-malarial drugs were introduced during the 20^{th} century.

After the global spread of chloroquine resistance, new synthetic drugs were adopted, such as mefloquine and sulfadoxine-pyrimethamine. These drugs were more expensive and sometimes caused adverse side effects; besides, resistance emerged relatively quickly [Packard, 2007]. In this context of despair, the search for new anti-malarial drugs led to the rediscovery of artemisinin in China. The plant that produces the compound had been used as a natural remedy for centuries and came to the attention of Chinese researchers during the 1970s. A team lead by the phytochemist Tu Youyou screened thousands of herbal remedies and obtained a list of candidates that included an extract from the plant *Artemisia annua*. Tu revisited the ancient literature of herbal remedies, identified how it had been used to treat malaria and was finally able to extract the active component, showing it could be used as a very effective anti-malarial drug. Tu Youyou was awarded half of the 2015 Nobel Prize in Physiology or Medicine for this work. It took more than 20 years for artemisinin derivatives to be of common use in other countries [Tu, 2011].

The WHO recommended using artemisinin-based drugs only when accompanied by a partner drug (termed as artemisinin combination therapies or ACTs), in order to prevent the parasite from developing resistance. As an additional measure, the WHO advised to use it only for treating severe malaria and resorting to older drugs when facing uncomplicated cases. International pressure made the WHO change its official guidelines in 2005, with many

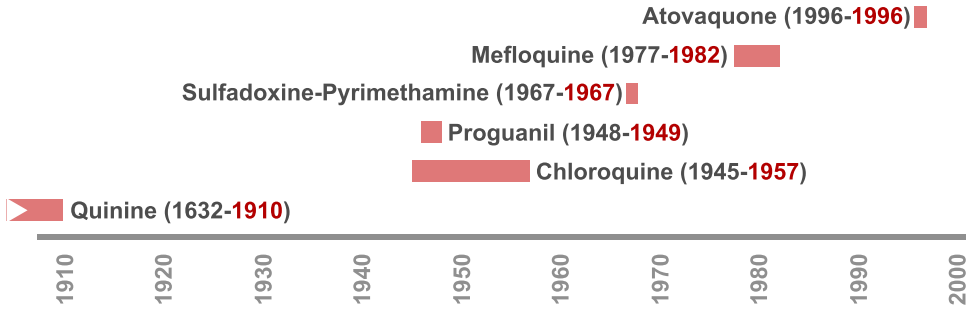

Figure 1.2: Timeline showing the time elapsed between the introduction (black) and first documented case of resistance (red) for the most popular antimalarial-drugs used in the 20th century.

countries adopting ACTs as first-line treatment, even for treating less severe malaria [Packard, 2014], [WHO, 2007]. In ACTs, the artemisinin component has a very short half-life[11] (0.5 - 1.4 h) but clears most parasites in a pulse-like fashion. The partner drug, having a longer half-life, takes care of eliminating the remaining parasites and limits the pressure on selecting for artemisinin resistance [Bloland et al., 2001].

First signs of resistance to artemisinin derivatives[12] appeared in the Thai-Cambodian border region in 2008; when a decline in the efficacy of the artesunate/mefloquine combination and delayed clearance of parasites were documented [Phyo et al., 2012], [Wongsrichanalai and Meshnick, 2008], [Noedl et al., 2008], [Dondorp et al., 2009]. The availability of cheap counterfeit drugs containing suboptimal doses of artemisinin and the use of artemisinin-based drugs as monotherapy are thought to be among the main factors leading the emergence of drug resistance [Dondorp et al., 2010], [Newton et al., 2003]. Several studies have focused on the spread of resistance in recent years. Ashley and colleagues documented the prevalence of delayed parasite clearance across mainland Southeast Asia [Ashley et al., 2014] whereas Ariey and colleagues found a set of molecular markers (a set of mutations in the *kelch13* gene) that can be used for surveillance [Ariey et al., 2014]. As the spread of artemisinin resistance is one of the central themes in this thesis, I will postpone a more detailed exposition to subsequent chapters where I will also disclose our specific contributions to the field.

[11] Period of time required for the concentration of the drug to be reduced by one half.

[12] As of March of 2017, the WHO has declared the presence of artemisinin resistance in 5 countries: Cambodia, Laos, Myanmar, Thailand and Vietnam.

1.3.2 Drug resistance and transmission rates

It is a long-standing hypothesis that malaria transmission rates, and their relationship to acquired immunity, are another major factor regulating the emergence of drug resistance. One of the observations that support this line of reasoning is the ratio of symptomatic versus asymptomatic infections found in different transmission regimes. In areas with low transmission rates the likelihood of developing acquired immunity is very low and this results in the majority of infections being symptomatic and therefore treated. When this is the case, selection for resistance is pushed by the specific drugs used during treatment. Conversely, the probability of an infected person becoming symptomatic decreases with age in regions with high transmission rates of malaria. The cause is that immunity is built up over the years due to continuous exposure to the disease. So, in these regimes, the probability of large numbers of parasites facing a specific drug in a context of low immunity is significantly smaller[13] [Petersen et al., 2011], [Hastings, 2004], [White and Pongtavornpinyo, 2003].

The spread of resistance requires that resistant parasites produce sufficient gametocytes to allow for transmission by mosquito bites. In asymptomatic individuals, this means that the resistant parasites need to evade the host immune system and outcompete their drug-sensitive counterparts. As sensitive parasites usually have a fitness advantage in the absence of the drug, the result is, again, a lower chance of emergence and spread of drug resistance in regions with high transmission rates. There are also historical observations supporting this argument, such as chloroquine resistance emerging in low transmission areas or antifolate sensitivity decreasing faster under this transmission regime [White, 2004]. However, recent studies comparing the presence of sub-microscopy malaria infections in high and low transmission regions challenge this hypothesis. In high transmission settings, it has been found that 20% of infections are detectable only at sub-microscopic level and are therefore asymptomatic. Surprisingly, prevalence increases to more than 70% in regions with low transmission rates, implying that not all infections develop into symptomatic infections under low transmission [Okell et al., 2012], [Bousema et al., 2014]. These findings also present further complications for elimination strategies since a large fraction of the

[13] Unless the host is a child and did not have time to develop immunity.

parasite reservoir would not be detected by classic methods [Harris et al., 2010]. In addition, it has wider implications for the deployment and assessment of malaria interventions.

1.3.3 Genetic basis of drug resistance

Resistance is caused by polymorphisms or changes in the copy number of genes that encode a drug's target or are related to the transportation of the drug's components. Resistant parasites have a fitness advantage in the presence of the drug. Resistant infections recrudesce more often and also take more time to be cleared by treatment, which translates into a greater probability of transmission [White, 2004]. Sometimes, resistance requires multiple unlinked events (epistasis) to develop. For instance, if a mutation that reduces drug sensitivity is associated with a fitness reduction, the parasite may require of additional compensatory polymorphisms to compete with non-resistant parasites when the drug is not present [Brown et al., 2010], [Petersen et al., 2015].

Many factors are important for the emergence of *P. falciparum* resistance to anti-malarial drugs, such as the parasite's mutation rate, the fitness cost of adaptive polymorphisms, the virulence[14] of the infection, the selection pressure imposed by the drugs and the level of adherence to treatment. In addition, host acquired immunity and transmission rates also regulate the probability of emergence in non-trivial ways [Petersen et al., 2011], [White, 2004].

Selection of resistance is triggered by suboptimal concentrations of the treatment drug. When that occurs, the drug constrains the number of sensitive parasites without inhibiting resistant ones. Evidence suggests that acute infections are the primary source of de novo mutations conferring resistance [White, 2004]. However, the long half-life of some drugs provides a strong selective regime for the spread of resistance acquired from re-infections, as these new parasites will encounter sub-optimal concentrations of the drug [Hastings, 2004]. Following this reasoning, some researchers encourage the use of mass drug administration programs for malaria control since effective drugs under strict administration surveillance would not contribute to the development and spread of resistance [von Seidlein and Greenwood, 2003].

Different anti-malarial drugs are used in combination to decrease the probability of

[14] In malaria, virulence is denoted by parasitemia, a measure parasite load in the bloodstream.

resistance emerging. The assumption here is that the development of resistance to two drugs with different mechanisms of action requires of different and independent genetic events. Given that these events are very rare, the probability of *P. falciparum* developing resistance to two drugs spontaneously and at the same time is remarkably lower than when drugs are used in monotherapy (assuming that the drugs are always used in combination) [White, 1999], [Petersen et al., 2011].

The following sections briefly highlight the genetic basis of resistance to the principal antimalarial drugs. Given the depth of the literature, this summary is by no means comprehensive; the only goal is for the reader to be familiar with the key findings and the relevant studies in this field.

1.3.3.1 Resistance caused by changes in transporter genes

Mutations in transporter genes can boost the removal of specific drugs from the parasite, therefore, controlling drug sensitivity [Borges-Walmsley et al., 2003].

Polymorphisms in the *Pfmdr1* gene (*P. falciparum multidrug resistance-1*) regulate sensitivity to different antimalarial drugs. The function of the protein encoded by the gene is still unknown, but its localization suggests that it works as a drug transporter. The association between drug resistance and mutations in this gene is complex, with some polymorphisms increasing resistance to some drugs while increasing sensitivity to others [Koenderink et al., 2010]. An amplification of the gene (i.e. increase in copy number) has been associated with the use of mefloquine (mainly in Southeast Asia) [Price et al., 2004]. There is some evidence implying that it is also associated with low sensitivity to lumefantrine, quinine and artemisinin [Sidhu et al., 2006]. It has been suggested that polymorphisms in this gene could induce reduced sensitivity to some of the partner drugs used in ACTs [Cui et al., 2015].

The *Pfcrt* (*P. falciparum chloroquine resistance transporter*) is also predicted to be a transporter gene [Martin and Kirk, 2004] and a single mutation, K76T, is the primary facilitator of resistance to chloroquine. The cited polymorphism seems to be always surrounded by a cohort of other mutations that are believed to compensate a fitness disadvantage in field parasites [Ecker et al., 2012], [Lakshmanan et al., 2005]. Some studies have shown that the presence of K76T also increases the sensitivity to mefloquine and

artemisinin derivatives [Sidhu et al., 2002]. Likewise, mutations in PfMDR1 seem to modulate the level of chloroquine resistance when PfCRT mutations are present [Plowe, 2003].

A set of polymorphisms in *Pfmrp1* (*P. falciparum multidrug resistance protein-1*) has been associated with reduced sensitivity to quinine, chloroquine, mefloquine, pyronaridine and lumefantrine. It has been hypothesized that it works in an epistatic fashion with other transporter genes [Raj et al., 2009].

1.3.3.2 Resistance to antifolates

These drugs target the enzymes DHFR[15] and DHPS[16]. DHFR reduces dihydrofolic acid to tetrahydrofolic acid, which is indispensable for cell growth since it regulates the synthesis of purines and thymidylates [Schnell et al., 2004]. DPHS produces dihydropteroate, crucial for de novo folate synthesis [Volpe et al., 1992]. Pyrimethamine and sulfadoxine are administered in single doses [Cui et al., 2015]. Introduced in the 1970s, the pyrimethamine-sulfadoxine drug combination was inexpensive, well-tolerated and, until the emergence of resistance, very effective [Petersen et al., 2011]. Resistance is associated with the sequential acquisition of mutations in the *Pfdhfr* and *Pfdhps* genes, which code for the target enzymes. In this case, each mutation reduces drug sensitivity in a stepwise manner [Plowe et al., 1997], [Cowman et al., 1988], [Peterson et al., 1988]. Molecular epidemiological studies suggest that resistance to pyrimethamine moved from Asia to Africa and South America via migration, most likely before the use of the drug combination in Africa [Roper et al., 2004]. Nonetheless, there is evidence that indicates an independent origin of the mutations that lead to sulfadoxine resistance in Africa and Asia [Alifrangis et al., 2014].

1.3.3.3 Resistance to atovaquone

Atovaquone[17] is usually used in combination with proguanil. The drug combination is principally aimed at tourists and used as a prophylactic (although it can be also used for treatment). Mutations associated with treatment failure arise in the *Pfcytb* gene although other mechanisms of resistance are still undetermined [Fivelman et al., 2002], [Wichmann

[15] Dihydrofolate reductase
[16] dihydropteroate synthase
[17] Impedes cytochrome electron transport.

et al., 2004].

1.3.3.4 Resistance to artemisinin derivatives

Several studies have assessed the extent of artemisinin resistance, detailing its presence in parts of Cambodia, Thailand, Myanmar and Vietnam [Ashley et al., 2014], [Bosman et al., 2014], [Tun et al., 2015]. The situation is aggravating in Southeast Asia, as treatment failure with ACTs has been already documented. Failure is due to the prevalence of parasites resistant to the partner drugs although it is exacerbated by delayed clearance in artemisinin derivatives. In Thailand and Cambodia, failure in artesunate-mefloquine treatment has been observed and attributed to the presence of mefloquine resistance [WHO, 2014a], [Wongsrichanalai and Meshnick, 2008]. In Cambodia, failure has been documented when administering dihydroartemisinin-piperaquine, most likely because of the presence of piperaquine resistance [Amaratunga et al., 2016], [Spring et al., 2015].

Due to the disastrous consequences that artemisinin resistance could have if it spreads to Africa, the international community is joining efforts to eliminate resistant parasites from the Greater Mekong subregion [WHO, 2015a]. Ariey and colleagues identified a set of mutations associated with delayed parasite clearance in the propeller domain of the *kelch13* gene (PF3D7_1343700) [Ariey et al., 2014]. They discovered these mutations by analyzing laboratory-adapted parasite clones from Tanzania that were subjected to high doses of artemisinin derivatives during 5 years. They confirmed their findings with in-vivo phenotypic data and showed how the frequency of mutant parasites increased in the first decade of the XXI century in regions where resistance was acknowledged. The in-vitro procedure identified four mutant alleles in the *kelch13 propeller* domain: Y493H, R539T, I543T and C580Y. Additionally, another set of 17 *kelch13* mutations were identified in field samples from Cambodia. Additional studies, including some of the articles this thesis has contributed to, provided compelling evidence that non-synonymous mutations in the *kelch13 propeller* and BTB-POZ domains provoke indeed delayed parasite clearance when using artemisinin derivatives [Ghorbal et al., 2014], [Miotto et al., 2015], [Straimer et al., 2015], [Takala-Harrison et al., 2015]. Perhaps the most pressing issue in malaria control is to avoid artemisinin resistance spreading to or emerging independently in Africa, as this would increase childhood mortality dramatically [Dondorp and Ringwald, 2013].

Regarding the molecular mechanisms behind resistance, it has been shown that parasites with resistant *kelch13* mutations have an enhanced protein response and also exhibit slower growth rates during the first part of the erythrocytic cycle, perhaps protecting against the oxidative effect of artemisinin [Dogovski et al., 2015], [Mot et al., 2015]. Furthermore, another study has shown that *kelch13* tags for degradation a protein (PI3K) targeted by artemisinin compounds. Resistant mutations in the *kelch13* gene disrupt this binding, lessening the effect of artemisinin [Mbengue et al., 2015].

As previously noted, the evolution of these mutations and their relationship with the dynamics of population structure constitutes one of the core themes in this thesis. Understanding the evolutionary, demographic and epidemiological factors shaping resistance is critical to design sensible strategies to avoid its spread. We study the relationship of population structure with artemisinin resistance in Chapter 2 and continue with this line of research in Chapter 4, this time looking in detail at the global geographical distribution of *kelch13* mutations.

1.4 Population structure

In this section we introduce the concept of genetic population structure and review its relevance in malaria. We also describe the most popular methods used for characterizing structured populations.

It is helpful to start with the idealized definition of a population given by the Wright-Fisher model [Wright, 1931], [Fisher, 1930] to develop the intuition behind the concept of population structure. The Wright-Fisher model is a simplified version of the genetic process, and it is used to model stochastic changes in allele frequencies between generations (i.e. genetic drift). A population in the Wright-Fisher model consists of a set of individuals that mate randomly without any restriction. This kind of population is said to be panmictic. In this idealized population, and given the other assumptions of the model are also met[18], all individuals are expected to share approximately the same degree of relatedness, having a similar distribution[19] of alleles at each locus (or allele frequency spectrum) [Hamilton,

[18]These are discrete and non-overlapping generations, constant population size and the absence of natural selection.

[19]Assuming a homogeneous mutation rate across individuals.

2011], [Hartl and Clark, 2007].

In the real world, however, random mating is rarely observed, as organisms tend to have specific mating preferences and mate with counterparts that are geographically close. In fact, Wright coined the term isolation by distance to refer to the spatial structure of mating, i.e. the decrease in the probability of mating as geographic distance increases [Wright, 1943]. Nonetheless, these factors are usually self-averaging for geographically confined populations, making the random mating assumption appropriate in terms of modeling [Lawson, 2012]. If we impose mating structure, such as systematic mating preferences or barriers that limit mating between subsets of the population, individuals will no longer be genetically related to each other in the same manner, and different subgroups will share a distinctive distribution of alleles. A population like this is said to be genetically structured or to have population structure.

A more formal definition of population structure relies on the presence of limited gene flow between subpopulations. Gene flow refers to the transfer of alleles between gene pools, which is achieved by the movement of gametes or the migration of individuals. In this setting, the term subpopulation denotes a subset of individuals in which gene flow from other parts of the population is somehow restricted. It follows that limited gene flow would allow the subpopulations to evolve, to some degree, independently of each other. Population structure can then be defined as the variation in the allele frequency spectrum, induced by restricted gene flow, across a population [Hamilton, 2011].

The previous definitions, though useful, are not very satisfactory because they treat populations as discrete entities. In reality, many demographic processes shape population structure in a continuous fashion. For instance, recent admixture events, the process in which distinct populations merge and breed, tend to create a spectrum of genetic relatedness (and allele frequency variation) among individuals. Such continuous range of variation is problematic to accommodate within a discrete classification [Lawson, 2012]. Conceptually, a speciation event could be seen as an extreme case of population structure where gene flow has been limited for so long between populations, either by geographical barriers, natural selection or other forces; that genetic changes have made inter-population breeding impossible [Slatkin, 1987].

1.4.1 The importance of population structure in malaria

Characterizing and understanding parasite population structure, gene flow and associated demographic processes has become a fundamental piece in the efforts for controlling malaria. These features lay at the core of many critical applications, including surveillance, tracking of migration patterns and the identification of loci under selection or associated with particular phenotypes [Anderson et al., 2011], [Escalante et al., 2004], [Volkman et al., 2012].

When looking for signatures of selection in the genome, a detailed understanding of population structure and demographic processes is needed to distinguish between selective sweeps and neutral events. For example, population expansions or bottlenecks can alter allele frequencies or reduce diversity, confounding the apparent significance of selection tests [Sabeti et al., 2006]. Likewise, knowledge of population structure is crucial when performing genome-wide association studies (GWAS). In this case, population stratification can produce false positives and inflate associations if it is not corrected for [Sabeti et al., 2006], [Altshuler et al., 2008], [Knowler et al., 1988].

The presence of selective pressures imposed by control interventions, the human immune system or the use of anti-malarial drugs is constantly forcing *Plasmodium falciparum* to adapt. As the parasites respond to these pressures, the evolutionary process leaves a genetic imprint at the population level. For instance, elimination interventions are supposed to decrease transmission rates over time, reducing genetic diversity and favoring inbreeding. One of the outcomes of this process would be a reduction in the number of multiple infections (COI > 1) and a transition towards more inbred populations of parasites. Similarly, linkage disequilibrium (LD) patterns, the non-random association of alleles, provide clues about diversity and transmission history in the parasite. Low LD is usually associated with high rates of transmission and elevated levels of complexity of infection. Furthermore, the emergence of drug-resistant strains can trigger founder effects followed by recolonization events as resistant parasites take over and outcompete sensitive ones under selective pressure. All these signals can be recognized in populations of isolates by the analysis of sequencing data [Volkman et al., 2012]. It is important to highlight here the inherent difficulty of inferring the nature of the processes causing a specific set of genetic patterns. These

patterns are generated by an overlapping sequence of genetic processes and demographic events that interact in non-trivial ways during the history of a population. In addition, different processes can leave similar patterns in the data, confounding the interpretation of the signal [McVean, 2009]. For all these reasons, it is very challenging to reconstruct the past evolutionary history of an organism solely from sequencing data.

Nonetheless, methods for characterizing and tracking changes in population structure have the potential to become central tools for monitoring interventions and identifying ongoing demographic events [Volkman et al., 2012]. An analogous approach has proven successful in tracking the evolution of simpler pathogens, such as influenza [Ghedin et al., 2005] or HIV [Lukashov et al., 1998], [Rai et al., 2010]. Moreover, the characterization of parasite populations and the identification of key bio-markers, like highly differentiated or fixated alleles, are the building blocks for real-time control systems. These methods could allow researchers to track the migration of different lineages of parasites or to ascertain the source of new infections, making possible to distinguish between reinfection and recrudescence in the field [Daniels et al., 2008], [Campino et al., 2011].

1.4.2 Population structure and drug resistance

As mentioned in the previous section, population structure allows populations to evolve with a certain degree of independence from each other and, therefore, to diversify and adapt to their local environment. In species where extinction and recolonization events are frequent, such as parasites, gene flow[20] between subdivided populations might play a central role in evolutionary terms. The reasoning is that gene flow could spread local adaptations that are beneficial when selective pressures are present [Slatkin, 1987], [Lenormand, 2002]. Conversely, sustained gene flow might act as a force against divergence, disallowing local adaptations by homogenizing the gene pool [Abbott et al., 2013], [Lenormand, 2002].

These observations connect with the long-standing debate about the emergence of antimalarial drug resistance in *P. falciparum* and its relation with low-intensity transmission settings. This type of transmission regime effectively restricts gene flow, favoring parasite inbreeding and structured populations. In contrast, populations in high transmission areas, like sub-Saharan Africa, appear panmictic over large geographical regions [Anderson et al.,

[20]Notice that recolonization is effectively just another form of gene flow.

2000], [Dye and Williams, 1997], [Manske et al., 2012]. This demographic difference, the presence of population structure within small geographic territories, might also play a decisive role in the emergence of antimalarial resistance in Southeast Asia, supporting the spread of local novel adaptations under strong drug pressure.

In this work, we explore this hypothesis and argue that genomic signals of population structure can be used to monitor the emergence and spread of resistance.

1.4.3 Characterizing population structure

There is a myriad of methods available for identifying population structure from sequence data and for characterizing how subpopulations relate to each other [Lawson, 2012]. In the following sections, we briefly introduce the most popular techniques used in the literature, paying special attention to the approaches we adopt in this thesis.

1.4.3.1 Non-parametric methods

In this section, we introduce some of the most widely used non-parametric (also termed model-free) methods. In this context, non-parametric means that the method does not model the data or the data generation process explicitly. Nonetheless, these methods can make implicit assumptions or impose a given criterion. For example, neighbor joining may be interpreted as a greedy optimization algorithm that attempts to build a tree according to the minimum evolution criterion [Gascuel and Steel, 2006]. The fundamental disadvantage of these methods is the difficulty of interpreting results, as they cannot be related to an underlying model. Despite this, non-parametric techniques are computationally fast and scale well with sample size, making them excellent tools for exploratory data analysis.

1.4.3.1.1 Principal Component Analysis and spectral methods Principal Component Analysis (PCA) has become a standard tool for identifying genetic variation structure. PCA is a dimension reduction technique; it projects the observations onto perpendicular axes (or principal components) via linear transformations. These transformations are defined in such a way that the first component accounts for the largest fraction of variance in the data. Each subsequent component also maximizes the fraction of variance it explains, conditioned on being orthogonal to all previous components. A technical description can

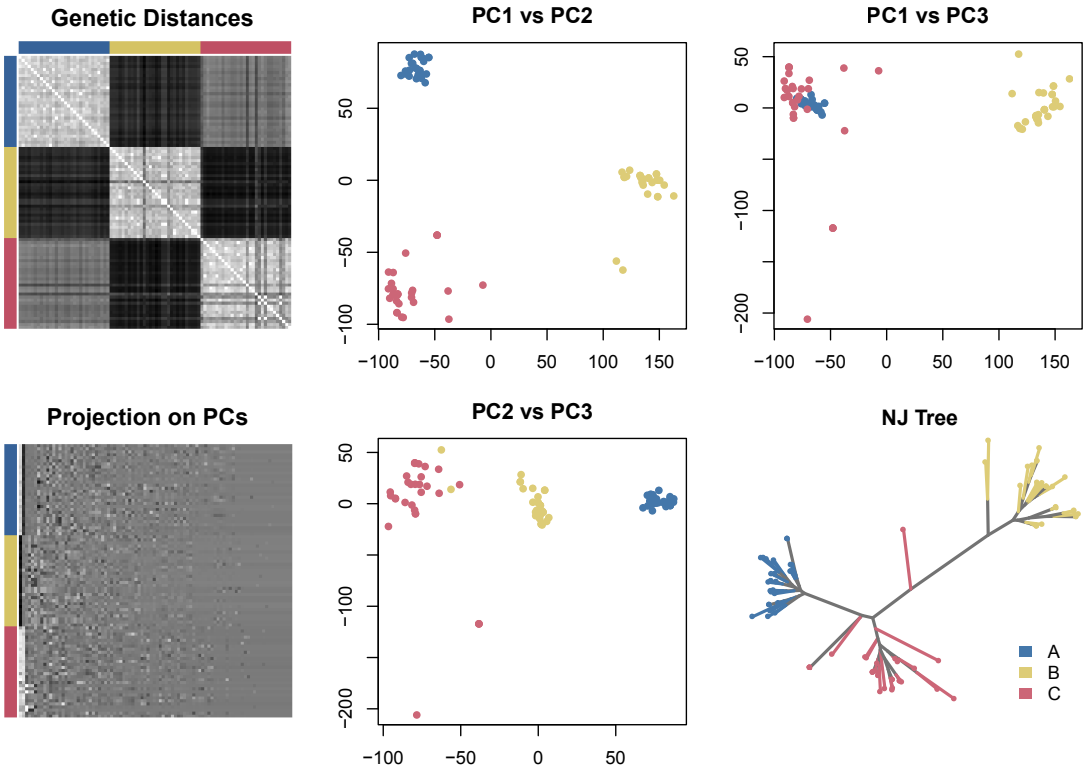

Figure 1.3: PCA and NJ analysis on simulated data for three populations with low levels of migration. Top left panel shows a heatmap of genetic distances (average number of differences per nucleotide) for 3 populations (blue, yellow and red) with 30 samples each. Sequences of 2.5Mb in length were simulated with the SCRM package [Staab et al., 2015] using $\mu = 1 \times 10^{-7}$, $r = 1 \times 10^{-8}$, $N_e = 50000$ and a symmetric migration rate of 4×10^{-6}. The top middle panel shows how the three populations can be clearly identified when projected onto the first two principal components. In addition, PC1 indicates that the blue (A) and red (C) populations are more related among them than to the orange group (B). This is consistent with the patterns observed in the NJ tree (bottom right) and the distance matrix. The bottom left panel shows the projection of the sequences (rows) onto the principal components (columns), with the greatest variance (and structure) of the data being captured by the first few components.

be found elsewhere [Jolliffe, 2002].

Projecting the genetic observations onto the first two or three principal components provides an intuitive visual representation that depicts the genetic relatedness of the samples. Individuals that share many mutations tend to be close in higher components[21] whereas mutations that do not exhibit any pattern, usually considered sampling noise, dominate the lower components. Individuals in a population are expected to share the same allele frequency spectrum and tend to be placed close together in PC space [Lawson, 2012].

Figure 1.3 presents PCA results for simulated data consisting of three populations with

[21]The components that explain the highest fraction of variance.

low rates of migration (i.e. gene flow). Following the previous reasoning, clusters can be interpreted as genetically different populations while admixture is frequently recognized as bands of individuals connecting two groups[22] [Novembre et al., 2008]. Figure 1.4 shows PCA results for a partial admixture event (simulated data), where the aforementioned band can be recognized. Patterson and colleagues developed formal significance tests for detecting population differentiation via PCA [Patterson et al., 2006]. However, accurate interpretation of PCA is challenging. Technical artifacts can appear due to uneven sampling, and different historical demographic scenarios can produce similar projections of the data [McVean, 2009], [Novembre and Stephens, 2008], [François et al., 2010].

Additional features of PCA include its computational speed and the circumstance that principal components of variation can be used to correct for population stratification when performing GWAS [Price et al., 2006], [Marchini et al., 2004]. PCA belongs to a broader class of techniques denominated spectral methods whose overall purpose is to detect variation in the data. Another method included in this class is multidimensional scaling (MDS), also known as principal coordinate analysis (PCoA), an approach intimately related to PCA that we use in some of our analyses. In the case of PCoA, the principal components are built from a pairwise distance matrix that summarizes the raw genetic data. PCoA also uses a linear mapping for projecting the data onto the ordination space and tries to maximize the amount of variance explained by the projections [Ramette, 2007].

PCA and PCoA have similar time complexity, since PCA via singular value decomposition (SVD) has a time complexity of $O(\min\{mn^2, m^2n\})$, for a $m \times n$ matrix, and PCoA requires to compute a pairwise distance, an operation that has complexity $O(m^2n)$. Because the original raw data has been summarized, PCoA does not indicate which variants are most influential. Because we use the pairwise distribution of genetic distances as a starting point for different exploratory analyses, we prefer the use of PCoA. To identify influential variants, we follow a population genetics approach. For instance, using statistics that measure the degree of differentiation of a SNP across populations, instead of relying on PCA results. For exploratory data analysis and data visualization, differences between PCoA and PCA results are negligible given the nature of our data. Figure 1.5 shows the equivalent PCoA results for the simulated data used in Figure 1.3 and Figure 1.4, Pearson

[22]Notice that sometimes the groups do not need to be present in the analysis for the signal to appear.

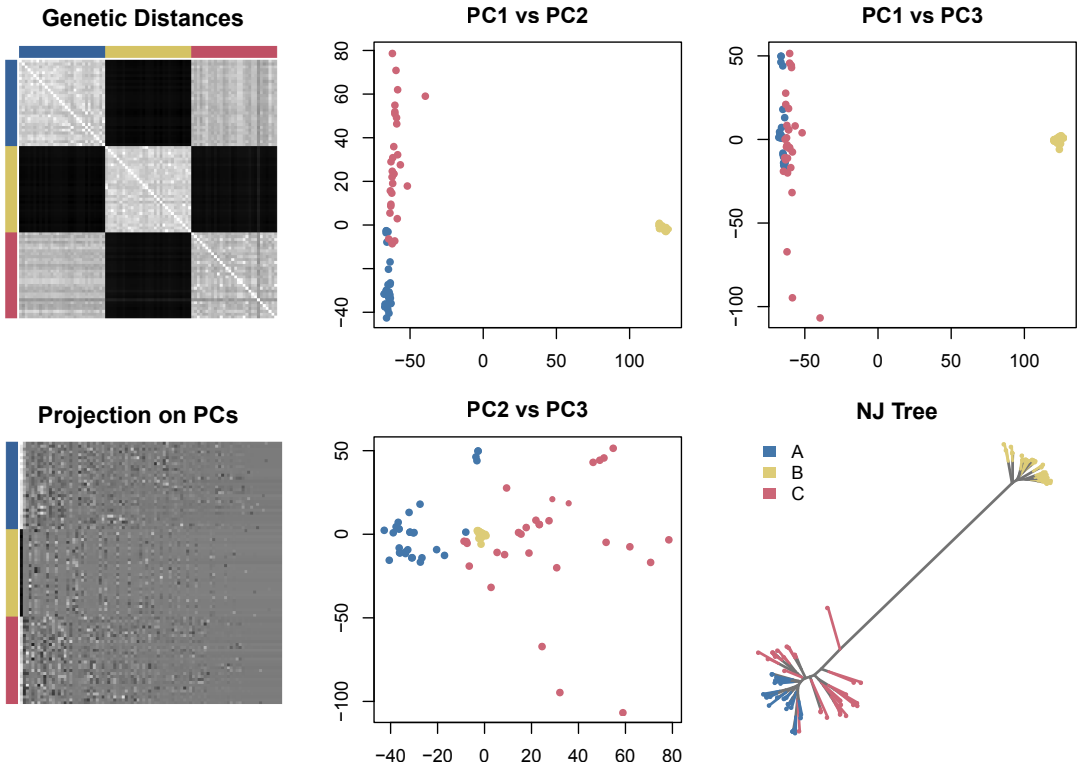

Figure 1.4: PCA and NJ analysis on simulated data for three populations undergoing a partial admixture event. Top left panel shows a heatmap of genetic distances (average number of differences per nucleotide) for 3 populations (blue, yellow and red) with 30 samples each. Sequences of 2.5Mb in length were simulated with the SCRM package using $\mu = 1 \times 10^{-7}$, $r = 1 \times 10^{-8}$, $N_e = 50000$ and a symmetric migration rate of 4×10^{-6}. 500 generations in the past, the migration rate from population C (red) to population A (blue) was increased to 1×10^{-4}. The top middle panel shows a band of samples connecting populations A (blue) and C (red) in PC2, characteristic of admixture events. The NJ tree (bottom right) does not accommodate this demographic scenario so well. A fraction of samples from population C (red) clustered within the clade of population A (blue) whereas the rest of group C formed its own clade. The bottom left panel shows the projection of the sequences (rows) onto the principal components (columns), with the greatest variance (and structure) of the data being captured by the first few components.

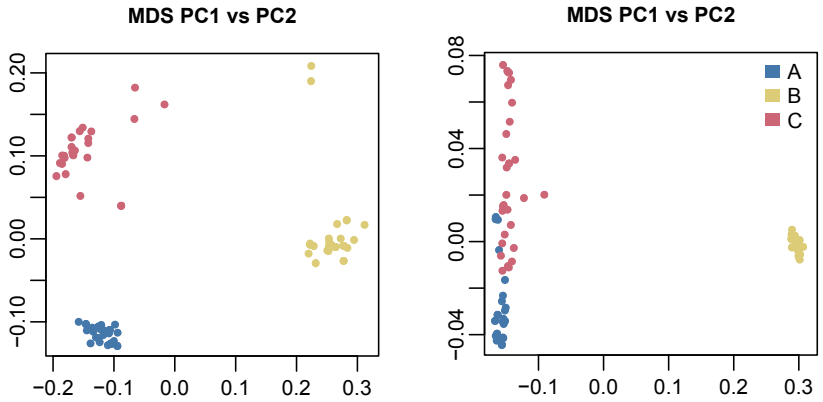

Figure 1.5: Projection of the simulated data presented in Figure 1.3 (left) and Figure 1.4 (right) on the first two principal components computed via PCoA (MDS). In both cases, Pearson correlation between PCA and PCoA projections was greater than 0.99 for the three first PCs.

correlation between the first three PC projections in PCA and PCoA was greater than 0.99.

1.4.3.1.2 Clustering and tree building methods Clustering approaches can be used on similarity or distance matrices to identify putative populations. A similarity matrix is constructed by computing a pairwise similarity measure from the genetic data, such as edit distance[23], covariance distance[24] or cosine distance[25]. It can also be the output of a model-based method, for instance, the coancestry matrix produced by chromosome painting [Lawson et al., 2012]. These matrices are not required to be symmetric since some methods compute similarity taking into account the whole sample composition instead of using each pair of individuals in isolation [Lawson and Falush, 2012].

Clustering methods can be applied to the similarity matrix, capitalizing on the fact that individuals within the same population are expected to be more similar than individuals from different populations. Variations on k-means [MacQueen et al., 1967] or centroid-based clustering [Kaufman and Rousseeuw, 1987], [Nock and Nielsen, 2006] are popular. Hierarchical clustering procedures, such as UPGMA [Sokal, 1958], build a hierarchical tree relating the individuals and are also frequently used in the literature. These methods depend on parameters that are not trivially estimated from the data and need to be provided by the researcher (e.g. the number of clusters for k-means or the cutting tree height for

[23] $\sum_l |y_{il} - y_{jl}|$, where y_{il} refers to the genotype of the i-th individual at locus l.
[24] $\sum_l (y_{il} - f_l)(y_{jl} - f_l)$, where f_l refers to the allele frequency at locus l.
[25] $(Y_i Y_j)/(|Y_i||Y_j|)$ where Y_i refers to the genotype vector of the i-th individual (binary).

hiearchical clustering methods).

It is also common to reduce the dimensionality of the data before performing clustering. For instance, reducing the raw genetic types or the similarity matrix to a set of principal components via PCA. The dimensionality reduction attempts to remove noise while conserving the most important features of the data. Some clustering techniques, like spectral clustering algorithms [Ng et al., 2002], exploit this idea.

Neighbor joining (NJ) [Saitou and Nei, 1987] is a fast distance-based tree building method that has its origins in phylogenetic analysis. It is widely used to explore the relatedness of different taxa or populations of samples. NJ starts with a start-like topology that is iteratively refined by a greedy algorithm. The choice of the operation to perform at each step is guided by a heuristic. The method tries to build an optimal tree according to the minimum evolution criteria[26] [Gascuel and Steel, 2006]. In evolutionary terms, it does not assume a molecular clock; meaning that different lineages can evolve at different rates. NJ does not identify populations *per se*, but the resulting tree offers a striking visual summary of the genetic structure present in the data. For example, the length of the branches separating clades tends to be indicative of their level of differentiation, and complex structure usually produces intricate branching patterns. Figure 1.3 (bottom right panel) shows a neighbour-joining tree for three simulated populations with very limited gene flow. The tree correctly recovers the structure of the data, placing each population in a different clade. It also exposes the higher degree of relatedness between population A (blue) and C (red) with a lower mean patristic[27] distance among these groups than between any of them and population B (yellow). Nonetheless, as is the case with PCA, NJ has severe limitations for identifying the genetic processes and demographic events that shaped genetic variation. For instance, admixture episodes cannot be well accommodated by a tree structure. It is also difficult to assess the degree of confidence we should assign to each clade in the tree[28] and results might be dependent on the type of distance used [Felsenstein, 2004], [Kalinowski, 2009], [Lawson, 2012]. Figure 1.4 (bottom right panel) illustrates how the NJ tree has problems accommodating a partial admixture event on simulated data.

[26] Minimizing the total branch length of the tree.
[27] Total branch length separating two tips.
[28] Methods like bootstrapping can be applied to palliate this drawback.

1.4.3.2 Model-based methods

Several statistical methods are available for the inference of global or local population ancestry. In these models, a population is defined as a set of individuals that are statistically indistinguishable[29] from the data [Lawson, 2012]. Once a number of distinct populations have been identified, individuals can be expressed as a mixture of ancestral proportions: the fraction of the genome whose ancestral origin can be attributed to each population.

Global ancestry inference refers to methods that estimate the proportion of ancestry from each putative population as an average over the genome of an individual. Conversely, local ancestry inference relates to methods that can determine ancestry proportions for specific regions of the genome[30]. Some methods estimate ancestry proportions as the parameters of an imposed statistical model [Alexander et al., 2009] whereas others build a pairwise similarity matrix from which structure is inferred [Lawson et al., 2012]. Below we review the most popular models.

1.4.3.2.1 Methods for unlinked data STRUCTURE [Pritchard et al., 2000], [Falush et al., 2003] is a very popular model-based clustering implementation. It uses the allele frequency spectrum to model how individuals vary within a population. The model assumes that there are K populations characterized by the allele frequencies at each locus. Using a Bayesian clustering approach, it assigns individuals to each cluster probabilistically, estimating the population allele frequency at the same time. The two main assumptions of this model are that all loci are unlinked (i.e. not correlated) and at Hardy-Weinberg equilibrium within each population. To accommodate the first assumption, correlated markers are filtered out before proceeding with the analysis. STRUCTURE estimates the global ancestral proportions of each individual or assigns them to a single cluster when admixture is not allowed. A Markov Chain Monte Carlo (MCMC) algorithm approximates the multivariate posterior distribution from which the *maximum a posteriori* (MAP) estimates are obtained.

STRUCTURE is computationally intensive, making infeasible to work with the big

[29] A set of random variables are statistically indistinguishable when their distribution is identical almost everywhere (i.e. the statistical distance among them is negligible). In the context of this thesis, we refer to the distribution of alleles or haplotypes in a set of individuals (i.e. populations).

[30] These can be obviously aggregated to give a global (i.e. genome-wide) estimate of ancestry.

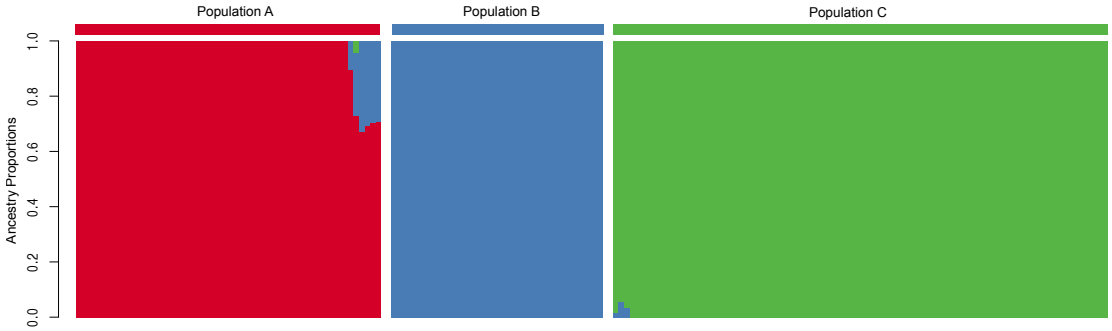

Figure 1.6: Ancestry proportions (or fractions) inferred by ADMIXTURE ($K = 3$) for a group of samples (simulated data). Each vertical bar represents a sample and the color reveals the putative ancestral origin for that fraction of the genome. Samples have been clustered into groups (i.e. populations) based on ancestry proportions. A bunch of samples in population A seem to be harbor a significant portion of ancestry (20-40%) from population B, suggesting that these particular individuals could be admixed.

datasets that are currently being produced by large-scale analysis projects. Recently, Raj and colleagues released FastSTRUCTURE, in which they replaced the original MCMC inference framework with a variational Bayes approach, dramatically reducing the execution time [Raj et al., 2014]. An alternative to STRUCTURE is ADMIXTURE [Alexander et al., 2009], a method that implements the same model but uses a likelihood optimization approach. This implementation makes the inference procedure incredibly fast[31] when compared to the original STRUCTURE software and yet provides results that are comparable in accuracy and resolution.

A major drawback of the mentioned methods is that the researcher must provide the number of putative populations (K). Heuristic procedures exist to guide the choice of this parameter. For instance, the authors of ADMIXTURE suggest studying how the cross-validation error varies with K while others evaluate the second order rate of change of the likelihood function on K [Evanno et al., 2005]. Figure 1.6 illustrates the output produced by these programs when admixture is allowed. In the example, samples have been clustered into population by running a clustering algorithm (k-means) on the inferred ancestry proportions.

1.4.3.2.2 Chromosome painting Both STRUCTURE and ADMIXTURE assume that markers (SNPs) are independent of each other. However, patterns of linkage dise-

[31] Although FastSTRUCTURE can compete with ADMIXTURE regarding execution time.

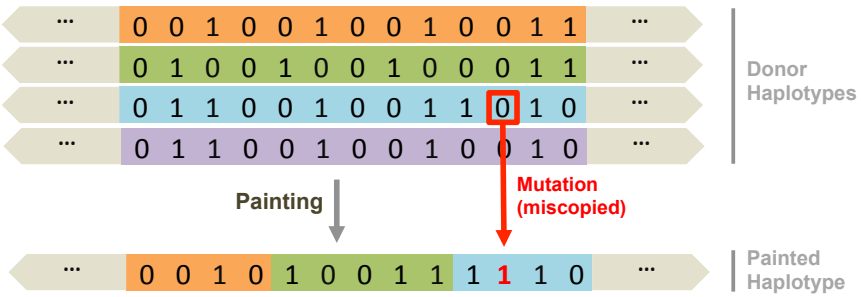

Figure 1.7: Illustration of how the painting process works. The haplotype at the bottom is reconstructed (probabilistically) from segments donated by the other haplotypes. A mutation is emitted by the HMM when copying the last chunk (cyan)

quilibrium violate this assumption and also encode relevant information about the shared genealogical history of individuals. Chromosome painting [Lawson et al., 2012] is a method that acknowledges linkage information and is based on the Li and Stephens model[32] [Li and Stephens, 2003].

The model reconstructs haplotypes with DNA segments of similar individuals in the dataset. Segments are donated and copied to and from individuals as if they were bound by recombination, creating a mosaic (Figure 1.7 illustrates this idea). The probability of an individual donating a segment has a simple interpretation in genealogical terms. It can be seen as the probability of the donor individual sharing the most recent common ancestor, for that segment of DNA, with the haplotype being reconstructed. Consequently, when painting an individual, switches between donor haplotypes can be assumed to be changes in the closest genealogical neighbor that are caused by recombination. This method estimates local ancestry as it assesses the putative origin of each locus probabilistically, once a set of populations has been inferred (i.e. samples are labeled as belonging to a given population and the painting is recoded in terms of ancestry segments).

Mathematically the painting process is implemented as a hidden Markov model (HMM) whose hidden states are associated, at each locus, with the set of potential donor haplotypes. These hidden states can emit mutations (i.e. differences between the observed genotype and the one being copied) with a given miscopying probability. The HMM is also parameterized with a recombination map from which transition (*switching*) probabilities are derived.

Here we adopt the formulation and notation introduced by Lawson and colleagues

[32]We note there that chromosome is a modification of the original model proposed by Li and Stephens.

[Lawson et al., 2012]. Let $\vec{Y} = \{Y_1, \ldots, Y_L\}$ represent the sequence of hidden states, with Y_l referring to the sequence from which the current haplotype, h_*, copies from at locus l. The transition probabilities between loci l and $l+1$ are given by

$$P(Y_{l+1} = y_{l+1} | Y_l = y_l) = \begin{cases} \exp(-\rho_l) + \big(1 - \exp(-\rho_l)\big)\frac{1}{n} & \text{if } y_{l+1} = y_l; \\ \big(1 - \exp(-\rho_l)\big)\frac{1}{n} & \text{otherwise,} \end{cases} \tag{1.1}$$

where $\rho_l = N_e g_l$, with N_e being a scaling parameter related to the effective population size and g_l being the genetic distance between loci l and $l+1$ measured in Morgans. Likewise, the emission probabilities at locus l are given by

$$P(h_{*l} = o | Y_l = y) = \begin{cases} 1 - \theta & h_{yl} = o; \\ \theta & h_{yl} \neq o. \end{cases} \tag{1.2}$$

where h_{*l} refers to the haplotype being painted, o to the observed allelic state, h_{yl} to the allelic state in the donor haplotype and θ corresponds to the mutation probability per nucleotide, usually set to the modifed Watterson's estimate used by Li and Stephens [Watterson, 1975], [Li and Stephens, 2003]. It is worth noting that the values of θ and N_e are usually established by using an E-M procedure.

The output of the painting process (after running the forward-backward algorithm [Rabiner, 1989]) is a matrix of posterior copying probabilities for each reconstructed haplotype. These matrices are combined into a news matrix, termed the *coancestry* matrix [Lawson et al., 2012], matrix that describes the number of expected chunks donated from and to each chromosome. The coancestry matrix is indeed a similarity matrix, and it is used to infer putative populations with other methods. The coancestry matrix does not need to be symmetric (and usually it is not) as the painting process takes into account all individuals instead of pairs in isolation. Figure 1.8 illustrates the painting process on simulated data.

1.4.3.2.3 FineSTRUCTURE FineSTRUCTURE [Lawson et al., 2012] is a software package that relies on the coancestry matrix generated by the chromosome painting procedure to cluster individuals into populations. The underlying model tries to partition

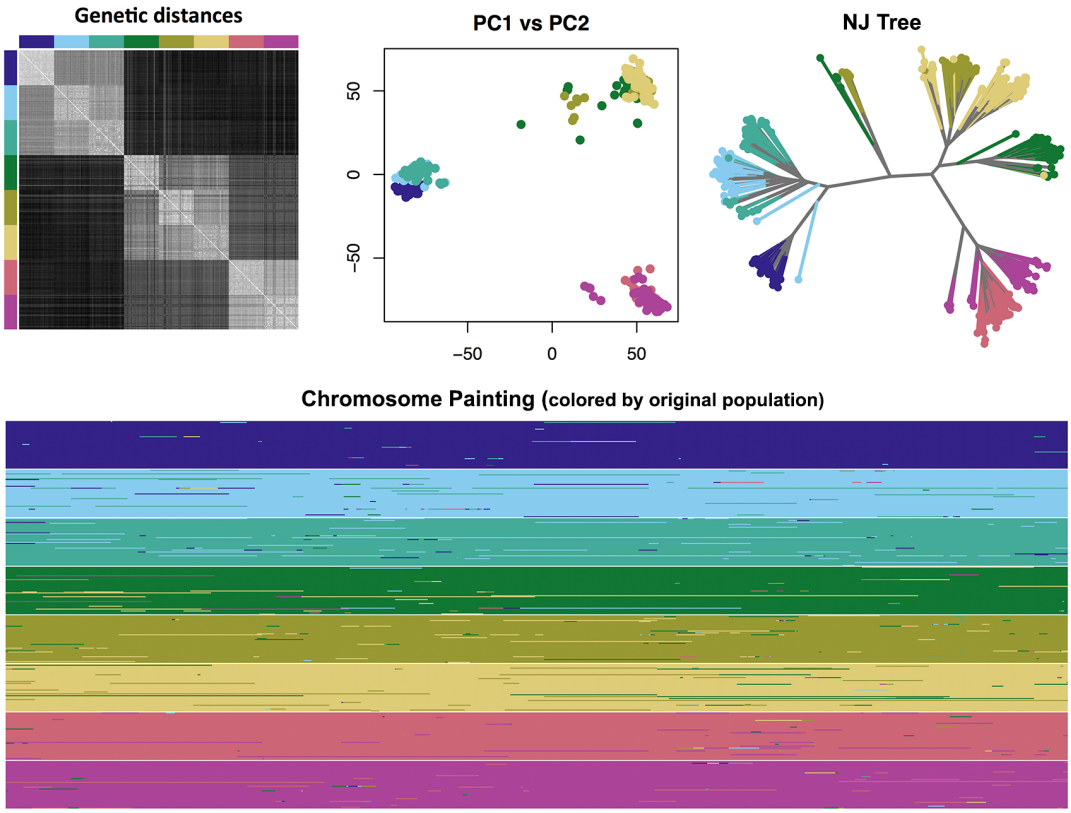

Figure 1.8: Chromosome painting on simulated data. We simulated 8 populations of 50 samples with the SCRM package (sequences of 2.5Mb in length were simulated using $\mu = 1 \times 10^{-7}$, $r = 1 \times 10^{-8}$, $N_e = 50000$) and different migration rates among demes. The top row shows the heatmap of genetic distances, the projection of the sequences on the first top PCs and the neighbor joining tree built from the distance matrix. The bottom row shows the result of performing chromosome painting. Each row represents a sample in which each locus has been colored with the (color of the) most likely donor population. Populations are easily distinguishable from the painting although some *noise* can be appreciated due to some populations interchanging segments. Notice that, in this case, we have discarded the uncertainty associated with the painting to make the figure easier to visualize.

the dataset into K clusters that have indistinguishable genetic ancestry and can be interpreted as putative populations. The model uses a Bayesian approach and explores different values for K by employing a reversible-jump MCMC algorithm. The method uses the MAP state of the sampled posterior to classify individuals into populations and then builds a hierarchical tree (by merging clusters) that relates populations. The tree building procedure follows a heuristic and is not based on a model of population differentiation. Therefore, this tree should not be interpreted as an account of population history but as a summary of population similarity. In this setting, the hierarchical nature of the tree and the discretization of populations makes it difficult to accommodate admixed populations that represent a gradient of relatedness between different groups [Lawson, 2012]. Although FineSTRUCTURE is a method designed specifically to take advantage of chromosome painting, other methods that only require a similarity matrix, can be used to perform inference on the coancestry matrix as well. For instance, PCA can be applied to the coancestry matrix to explore the relatedness of the individuals when projected on the higher principal components.

1.4.3.3 Statistics for describing populations

There is a collection of statistics derived from population genetics theory that is extensively used to assess relationships between populations and plausible demographic hypothesis. For instance, the shape of the allele frequency spectrum can be informative about a recent population expansion or contraction. Statistics that quantify the level of differentiation among populations and estimates of gene flow are also standard when characterizing population structure. As we have seen before, gene flow and population differentiation are intimately related. It is the case that many estimators of gene flow are derived from observed levels of genetic differentiation [Mallet, 2001], with the most popular statistic being F_{ST}.

F_{ST} was originally introduced by Wright and Malécot [Wright, 1949], [Malécot, 1948] as a measure of structure in populations but it is also used to quantify the level of population differentiation. For a given locus, Cockerham defined F_{ST} in terms of population allele

frequencies [Cockerham, 1969],

$$F_{ST_i} = \frac{\sigma_i^2(p_i)}{p(1-p)}, \tag{1.3}$$

where p_i is the allele frequency in subpopulation i, $\sigma_i^2(p_i)$ represents the variance over populations and p is the allele frequency in the whole population (for a given locus). Likewise, F_{ST} can be defined regarding the probabilities of sampling two different alleles (i.e. heterozygosity) within a subpopulation and in the aggregated population

$$F_{ST} = 1 - \frac{H_S}{H_T}, \tag{1.4}$$

where H_S refers to the probability when sampling from the same population and H_T to the probability when sampling from the whole population. For sequence data, this can also be expressed in terms of pairwise differences, as given by Hudson [Hudson et al., 1992]

$$F_{ST} = 1 - \frac{\Pi_S}{\Pi_T}, \tag{1.5}$$

where Π_S refers to the average[33] pairwise differences within subpopulations and Π_S refers to the population as a whole.

There are many definitions of F_{ST} in the literature, and some estimators can give different answers or are sensitive to certain features of the data, such as uneven sampling, diversity or the set of chosen loci [Bhatia et al., 2013]. Despite some controversy, and the confusion generated by alternative estimators, F_{ST} and related statistics are broadly used in the literature [Jost, 2008], [Ryman and Leimar, 2009].

1.5 Patterns of haplotype sharing

Due to the nature of the genetic process, demographic events such as bottlenecks or the occurrence of positive selection regimes shape the way in which genetic variation is structured among individuals in a population. This leaves a recognisable genomic imprint that is informative about past or current events and can be exploited accordingly. Limited by the technology available, population genetics has relied heavily on the study of the allele frequency spectrum (AFS) as the main source of demographic information [Charlesworth

[33] Depends on how averages are computed.

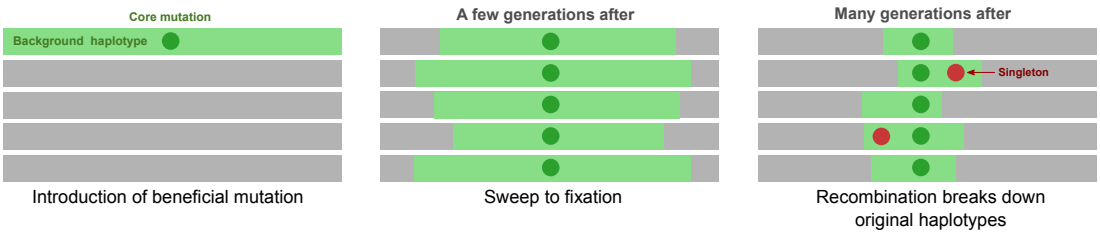

Figure 1.9: Sequence of cartoons illustrating the effects of recombination on the ancestral haplotype harbouring a beneficial mutation during a hard selective sweep. (Left) The beneficial mutation arises on a given haplotype. (Middle) The mutation sweeps to fixation along with a surrounding segment of the ancestral haplotype due to linkage. (Right) Recombination breaks downs the ancestral haplotype after many generations, reducing the length of the haplotypes shared by mutants.

and Charlesworth, 2016], [Marth et al., 2004], [Voight et al., 2005], [Keinan et al., 2007], [Gutenkunst et al., 2009], [Hartl and Clark, 2007]. However, methods that focus on the AFS ignore most of the information encoded by recombination as linkage disequilibrium and haplotype sharing patterns.

With the advent of whole-genome sequencing, several studies have revealed that these signals are useful for demographic inference, characterizing population structure or detecting selection [Gutenkunst et al., 2009], [Sabeti et al., 2006], [Lawson et al., 2012]. In particular, the distribution of shared haplotype lengths[34] has been shown to be very informative about recent demographic history, admixture or selective regimes [Pool and Nielsen, 2009], [Palamara et al., 2012]. Since recombination events break down ancestral haplotypes, sharing patterns can be used for dating, in the same fashion the accumulation of mutations is used as a clock (the *molecular clock*) [Kumar, 2005], [Hasegawa et al., 1985], [Schierup and Hein, 2000]. Furthermore, shared haplotypes have also been used to date rare variants [Mathieson and McVean, 2014], [Gandolfo et al., 2014]. Figure 1.9 summarizes how recombination shapes the distribution of shared haplotype lengths in the canonical example of a hard selective sweep.

1.5.1 Identity by state and identity by descent

The notion of identity by descent (IBD) entails the inheritance of a genomic region by a set of individuals from a common ancestor. It also requires that such region has not been affected by recombination [Browning, 2008]. In contrast, identity by state (IBS) regions

[34]The distribution of lengths of the haplotypes shared by two or more individuals in a population.

only require the set of DNA segments to match. Technically, IBD segments can harbour mutations but usually the timescale (in number of generations) for long IBD segments is so short that they tend to be IBS as well. This has motivated heuristic procedures that use IBS as an approximation for detecting IBD regions.

For historical reasons, the utilization of IBD segments used to be limited to pedigree experiments, IBD was then defined as the region inherited from a common ancestor *at most* a number of generations in the past [Chapman and Thompson, 2003]. However, for the work presented here we will follow the definition of Palamara and colleagues [Palamara et al., 2012] and only require the segment to be transmitted without recombination from any time in the past[35]. There are several methods available for IBD detection, including GERMLINE [Gusev et al., 2009] and fastIBD [Browning and Browning, 2011] among others [Browning and Browning, 2012]. Many of these methods try to infer the underlying IBD distribution for a set of individuals by applying an HMM along the genome in a pairwise manner.

Because of the rich genealogical and demographic information encoded on shared haplotypes, we explore how to efficiently build the ancestral haplotype graph in Chapter 5.1, a data structure that is informative about the recent genealogical history of the sequences and can be used to summarize and study shared haplotype patterns along the genome.

1.6 Conclusions

In this chapter, we briefly reviewed the historical context of malaria and introduced its most remarkable biological aspects, discussing its complex life cycle and the interactions between host, parasite, and vector. We also explored the major topics concerning drug resistance in *P. falciparum* and examined the concept of genetic population structure, introducing the associated technical background required to follow this thesis. We finished by commenting on the use of shared haplotypes for inferring recent demographic history.

In Chapter 2, we start our study of population structure and its relationship with the emergence of resistance. We continue this thread of research in Chapter 4, investigating the set of mutations associated with artemisinin resistance. In Chapter 3, we take a brief detour

[35]They also imposed an arbitrary length threshold (in centiMorgans) but we relax that requirement.

to assess how the presence of mixed infections influences our analysis of deep sequencing data and review a statistic widely used to characterize complexity of infection. Finally, in Chapter 5, we develop a scalable method to summarize patterns of haplotype sharing using a data structure that can be exploited for learning about recent genealogical history.

Chapter 2

Population structure and drug resistance in Cambodia

Contents

2.1 Introduction

As we highlighted in the introductory chapter, the emergence and spread of artemisinin resistance in Southeast Asia portrays a worrying prospect for current and future malaria control efforts[1]. The WHO launched the Global Plan for Artemisinin Resistance Containment (GPARC) [WHO, 2011] in 2011, a control campaign with the goal of containing the emergence and spread of resistance in Southeast Asia. However, given the gravity of the situation, the WHO switched in 2014 to an elimination campaign in the Greater Mekong sub-region [WHO, 2015a]. Here, we investigate the role that structured populations of parasites might play in the emergence of drug resistance. In particular, we focus on the Thai-Cambodian border area, within the cited context of artemisinin resistance. This chapter present the findings described in [Miotto et al., 2013]. We remind the reader that at the time of this work artemisinin resistance was described as delayed parasite clearance and that the candidate molecular marker had not been identified yet.

This study was a big collaborative project, my contributions were primarily on the identification and characterization of population structure, its relationship with drug resistance, and the elucidation of plausible demographic scenarios. I disclose my specific contributions and interactions with other scientists[2] in the last section of the chapter (Section 2.7).

2.2 Rationale

Following reports of delayed clearance for artemisinin derivatives [Phyo et al., 2012], [Wongsrichanalai and Meshnick, 2008], [Noedl et al., 2008], [Dondorp et al., 2009], [Dondorp et al., 2010], we studied patterns of genetic variation for *P. falciparum* in the Thai-Cambodian border region. As noted earlier, this is a region of historical importance for the emergence of antimalarial resistance, with some evidence suggesting that parasites from Southeast Asia have a predisposition to develop drug resistance. In particular, emergence of resistance has been documented in this region for chloroquine, sulfadoxine, pyrimethamine, quinine, mefloquine and artemisinin [Packard, 2014], [White, 2004], [White, 2010], [Rathod

[1] See Section 1.3.3.4.
[2] Other than my supervisors.

Region	Country	Location	Sample size
West Africa	Burkina Faso	Bobo-Dioulasso	48
	The Gambia	Banjul	67
	Ghana	Navrongo	243
	Mali	Bamako	56
Southeast Asia	Thailand	Mae Sot	106
	Vietnam	Binh Phuoc Province	12
	Cambodia (east)	Ratanakiri	50
	Cambodia (west)	Pailin	49
		Tasanh	47
		Pursat	147

Table 2.1: Geographic distribution of samples for the [Miotto et al., 2013] study.

et al., 1997]. The rationale behind our study was to evaluate if the genetic epidemiology of *P. falciparum* in Southeast Asia was informative about the emergence of artemisinin resistance. The findings presented here were published in 2013 in Nature Genetics [Miotto et al., 2013]. My role in this publication, in which I am second author, was to assess and characterize the unusual patterns of highly structured sympatric populations found in western Cambodia, and to provide evidence for plausible demographic scenarios. As three subpopulations were associated with clinical resistance to artemisinin, we elaborated on the probable evolutionary origins of these populations and on the role that population structure might have played in the emergence and spread of drug resistance.

2.3 Data and scope of the study

We analyzed 825 *P. falciparum* samples from ten different locations in West Africa and Southeast Asia, including Ghana, Mali, Burkina Faso, The Gambia, Thailand, Vietnam, and four sites in Cambodia (Table 2.1). Genotype calling was performed using a set of validated procedures described at length in [Manske et al., 2012][3]. We refined the dataset by applying a sequence of quality filters, with the aim of removing errors and artifacts, and obtained a subset of 86,158 high-quality SNPs. We review sequencing and genotyping in Section 2.6.1.

[3]Please notice that although I am an author in the cited article, my contributions there have no role in this thesis.

2.4 Results

We conducted an initial exploratory data analysis of population structure with non-parametric methods, such as building a neighbor joining tree and performing PCoA. After observing patterns of interest, we confirmed and refined our findings by applying model-based methods. In particular, we used ADMIXTURE and chromosome painting. Based on the trends identified and on further characterization of the subpopulations, we discussed plausible demographic scenarios and their implications for the emergence and potential spread of drug resistance in Southeast Asia.

2.4.1 Exploratory data analysis

We started by building a neighbor joining tree [Saitou and Nei, 1987] for the 825 samples in the dataset (Figure 2.1). We computed a matrix of pairwise genetic distances that served as input to the tree building algorithm.

The resulting tree clearly separated samples at continental level (West Africa vs. Southeast Asia). Within the West African clade, there was not any discernible structural pattern, with the subtree following a star-like topology. However, we observed complex branching arrangements on the Southeast Asian part of the tree, with Cambodian and Vietnamese samples separating into several clades and Thai samples being apparently detached from these populations. The very short length of the terminal branches of many sub-clades containing western Cambodian samples suggested that some sequences were very similar or practically identical. As mentioned in the introduction of this thesis, neighbor joining is a very crude method for discerning fine scale patterns of genetic variation. We considered the tree just a visual summary of genome-wide variation and, therefore, did not pursue any further refinement, such as computing clade support values via bootstrapping [Felsenstein, 2004].

We continued our exploratory data analysis by performing PCoA on all samples. The results (Figure 2.2) revealed an unexpected pattern of population structure. The first principal component (PC1) separated samples at continental level while PC2 and PC3 captured genetic variation present, primarily, within western Cambodia alone. Although some degree of structure was expected given the lower malaria transmission rates present

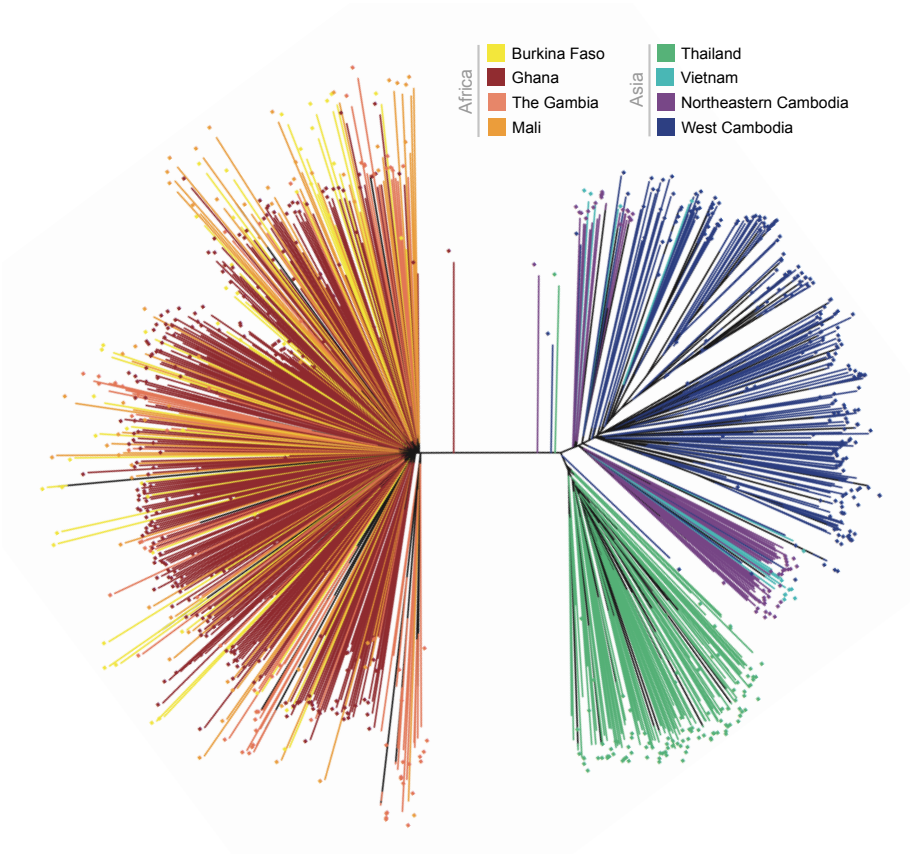

Figure 2.1: Neighbor joining tree of all 825 samples, based on a pairwise genetic distance matrix. The tree clearly separates samples at the continental level. Furthermore, the complex branching pattern within West Cambodia, with many samples clustering in small clades, suggests the presence of substantial population structure.

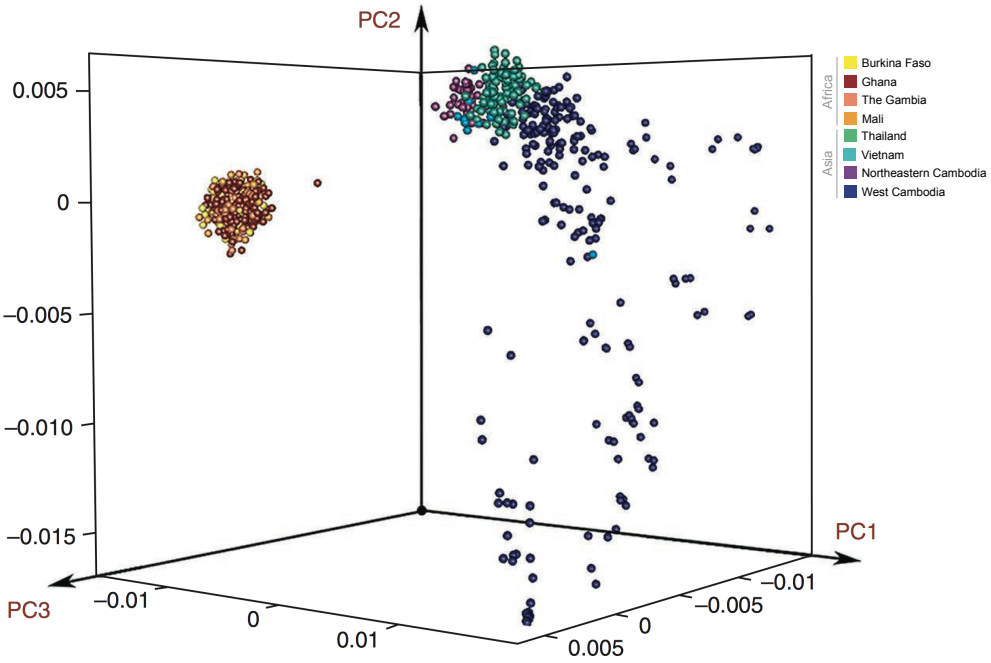

Figure 2.2: 3D PCoA plot showing the first 3 principal components (PC1, PC2 and PC3) for all 825 samples in the dataset. PC1 separates samples at continental level. PC2 and PC3, however, mainly capture variation present only in Cambodia.

in Southeast Asia [Hay et al., 2009], the severity of the structure exposed in western Cambodia was remarkable when compared to the other populations. This finding could not be explained by the isolation by distance model, as western Cambodian samples were collected within a small geographical area. Sampling sites in this region were separated by a maximum of 200 Km. In contrast, sampling locations in other parts of Asia were situated up to 1,000 Km apart. In West Africa, distance between sampling sites reached 1,500 Km. We performed PCoA after applying a subsampling scheme[4] to discard that the signal was a product of uneven sampling. Results were robust and consistent with the original analysis.

Since we were interested in the patterns found in Southeast Asian samples, we removed the West African populations and repeated the analysis (Figure 2.3). In this case, the first three principal components were all driven by the variability in western Cambodia. A subset of samples from Cambodia (western and northeastern) and Vietnam formed a regular or *core* group (named KH1), close to the Thai population. Three different outlier groups of samples from western Cambodia could be identified visually (termed KH2, KH3, and KH4). Furthermore, we observed distinct bands of samples (termed KHA) spreading from

[4]See Section 2.6.

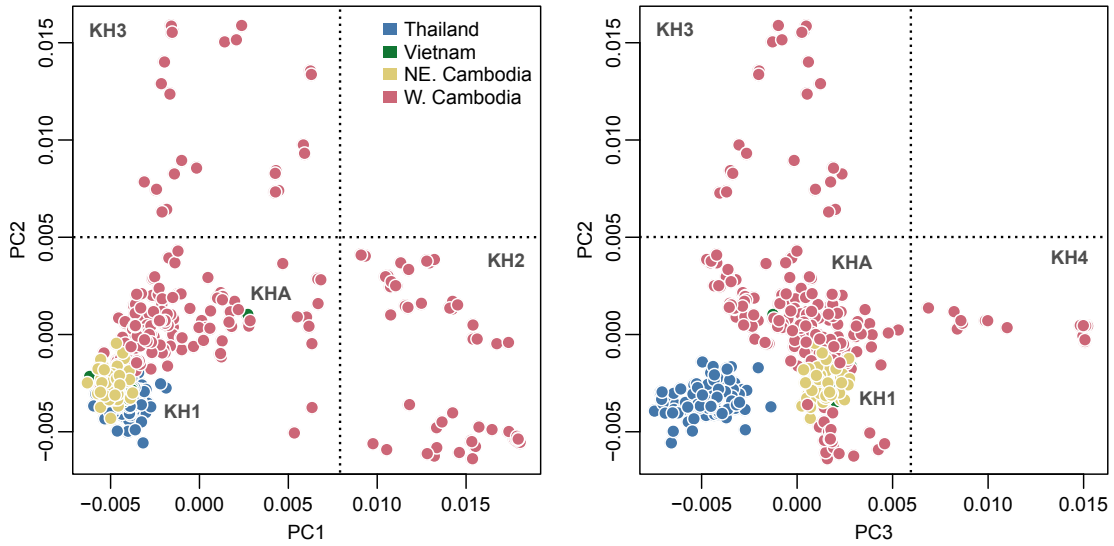

Figure 2.3: PCoA plots for Southeast Asian samples. The plots show the initial classification of Cambodian samples based on PC coordinates. We identified a regular or core group (KH1), three outlier groups (KH2, KH3 and KH4) and a set of admixed parasites (KHA).

the central core cluster to the outlier groups, consistent with the signals that admixture events tend to produce in principal component projections. Subsequent components did not show any other obvious grouping.

From this exploratory analysis, we derived a provisional classification of samples within Cambodia. We classified samples based on PC coordinates (quadrants), as shown in Figure 2.3. As indicated, we termed KH1[5] the group of samples that clustered with samples from Vietnam and Thailand, serving as a proxy for the ancestral type[6]. Outlier groups were labeled as KH2, KH3 and KH4, whereas the putatively admixed samples were labeled as KHA[7].

To conclude our exploratory analysis, we performed PCoA on the outlier groups and projected the KH1 and KHA samples on the first two principal components. As shown in Figure 2.4, the KH1 group clustered very tightly between KH3 and KH4, sligthly separated from KH2. However, the KHA samples were widely spread within the space separating the three outlier clusters, agreeing with all our previous observations.

[5]We use KH as prefix since it is the ISO code for Cambodia.
[6]What we considered to be the population that is most related to the ancestral population of the region.
[7]Unsurprisingly, it stands for KH-Admixed group.

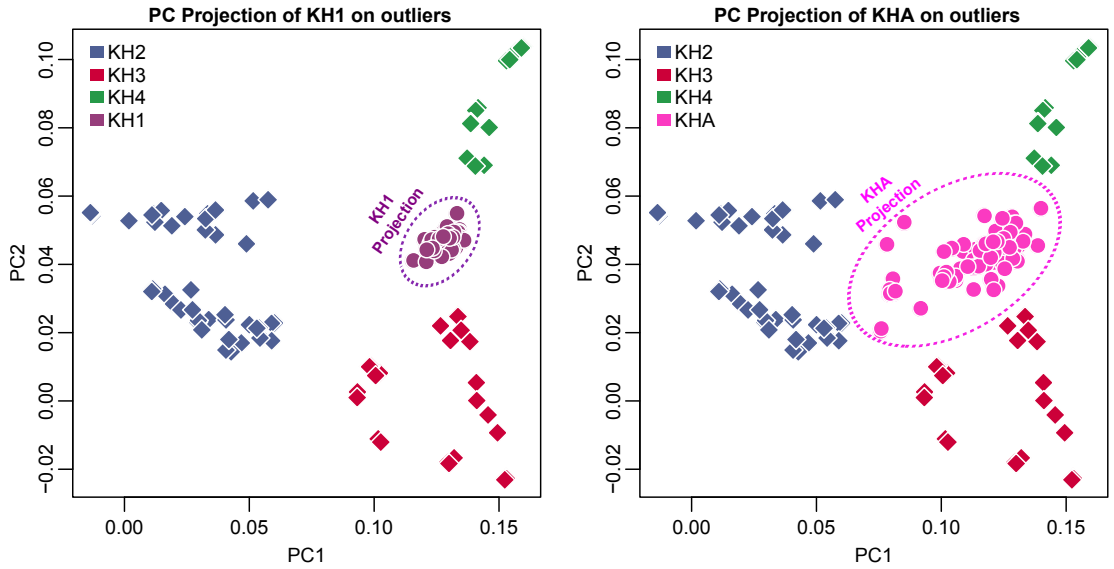

Figure 2.4: PCoA projections for KH1 and KHA. (Left) Plot showing the projection of the KH1 group on the first two principal components computed using the outlier groups (KH2, KH3 and KH4). (Right) Equivalent plot showing the projection of the KHA group. The large spread of the KHA samples suggests this group could be a set of admixed samples.

2.4.2 Model-based analysis of population structure

To assess and refine these initial discoveries, we performed chromosome painting [Lawson et al., 2012] on the Cambodian samples. As pointed out in Section 1.4.3.2.2, chromosome painting reconstructs haplotypes probabilistically by copying segments from other chromosomes (i.e. donors) in the dataset, thus taking into account linkage disequilibrium and producing copying patterns that are informative for the inference of population structure.

Based on our preliminary findings, we painted all chromosomes restricting the set of potential donors to individuals naively classified as belonging to the KH1, KH2, KH3 and KH4 groups, therefore assuming that the KHA group was an admixing population. From the output of the painting process we computed the coancestry matrix, a genome-wide estimate of the expected number of markers, or chunks, copied by each individual from all other chromosomes. We aggregated this matrix by using our initial group classification in a supervised manner, obtaining the equivalent of the global ancestry proportions computed by STRUCTURE [Pritchard et al., 2000] or ADMIXTURE [Alexander et al., 2009] for a given value of K[8]. From these proportions (Figure 2.5), the three outlier clusters could be clearly distinguished and complex admixture patterns were apparent in most of the

[8] $K = 4$ in this case

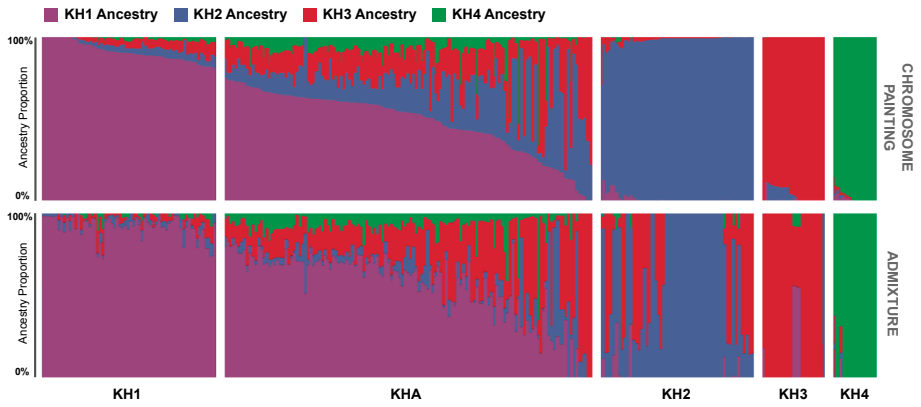

Figure 2.5: (Top) Ancestry analysis of the 293 Cambodian samples, based on the results of chromosome painting. Each vertical bar represents a sample and it is colored according to the proportion of the genome that was determined to have originated in each of the four KH1-KH4 clusters, as defined from PCoA coordinates. Samples were classified according to the 80% rule (see main text). (Bottom) Ancestry proportions produced by ADMIXTURE for $k = 4$. The samples are presented in the same order in both plots. The two methods give consistent results, with marginal differences in how proportions are attributed to KH2 and KH3. We note here that, with both methods, KH3 has a larger presence in KHA samples than the other outlier groups.

KHA parasites. We used the ancestry proportions to refine our sample classification. We assigned a sample to a group if it had at least 80% of ancestry originating in that group. In the same fashion, we labeled any sample with less than 80% of ancestry coming from any single group as admixed, and therefore belonging to KHA.

Using these definitions 63 samples were classified as KH1, representing the core group that clustered with samples from neighboring countries. Likewise, 55 samples were labeled as KH2, 23 as KH3 and 17 as KH4, representing the outlier subpopulations. Finally, 135 samples were labeled as KHA, exhibiting a varying degree of admixture. In fact, the ancestry plot exposed KHA as a continuum instead of as a distinct discrete subpopulation. This group displayed a gradient in the proportions of ancestry coming from the core (KH1) and the outlier groups (KH2, KH3, and KH4). The continuous nature of KHA rendered problematic with classical clustering algorithms or tree-building approaches and was the principal factor justifying our bespoke classification. Generic clustering procedures, such as K-means or spectral clustering, rendered very similar results[9] to the 80% rule but tend to split KHA, assigning samples on the extremes of the continuum to other groups. We found a similar problem when using FineSTRUCTURE [Lawson et al., 2012], a model-based clustering method designed to perform inference on the output of the chromosome painting

[9] for $K = 5$

technique, as the resulting hierarchical tree could not accommodate admixed samples within its structure and split KHA into several subpopulations.

We used the posterior copying probabilities, derived from the chromosome painting process, to build a genome-wide painting for all samples, aggregating the probability of copying a marker or chunk from each of the subpopulations. Figure 2.6 summarizes the distribution of putative ancestry along the genome for different samples and subpopulations. It also reveals the level of uncertainty faced by the painting process, with dimmer colors (i.e. closer to white) indicating more uncertainty about the origin of the chunk. We observed that segments exogenous to a group (i.e. whose most likely source is a different group) tended to be shorter in KH1, longer in the outlier groups (KH2, KH3, and KH4) and to cover a wider range in the admixed samples (KHA). These observations are compatible with KHA being the result of an ongoing admixture event between the ancestral population (KH1) and the outlier groups (KH2, KH3, and KH4).

We performed an additional model-based analysis of population structure using ADMIXTURE [Alexander et al., 2009]. Since the ADMIXTURE model assumes perfect linkage equilibrium between markers (i.e. they are independent of each other), we excluded SNP pairs that appeared to be linked using a filtering step. We classified samples using the same set of rules (i.e. at least 80% of ancestry to be labeled as belonging to a given group) on the results obtained for four putative ancestral populations ($K = 4$). Figure 2.5 compares the two ancestry plots. There was 82% concordance between the classification of samples derived from ADMIXTURE and the chromosome painting method. The majority of mismatches concerned KHA samples, although the gradient signal seen in KHA samples was still apparent. We favor the chromosome painting classification as this method can detect finer patterns but acknowledge that the current labeling should be viewed as a first approximation since the observed population structure is clearly complex. Further epidemiological sampling will be required to understand these patterns in detail.

2.4.3 Geographic location of subpopulations

All samples were collected in four independent studies conducted in a timespan of 5 years (2007 to 2011). To check that population structure was not an artifact produced by sampling, we stratified samples by study, sampling location and year (Tables 2.2, 2.3, and 2.4). We

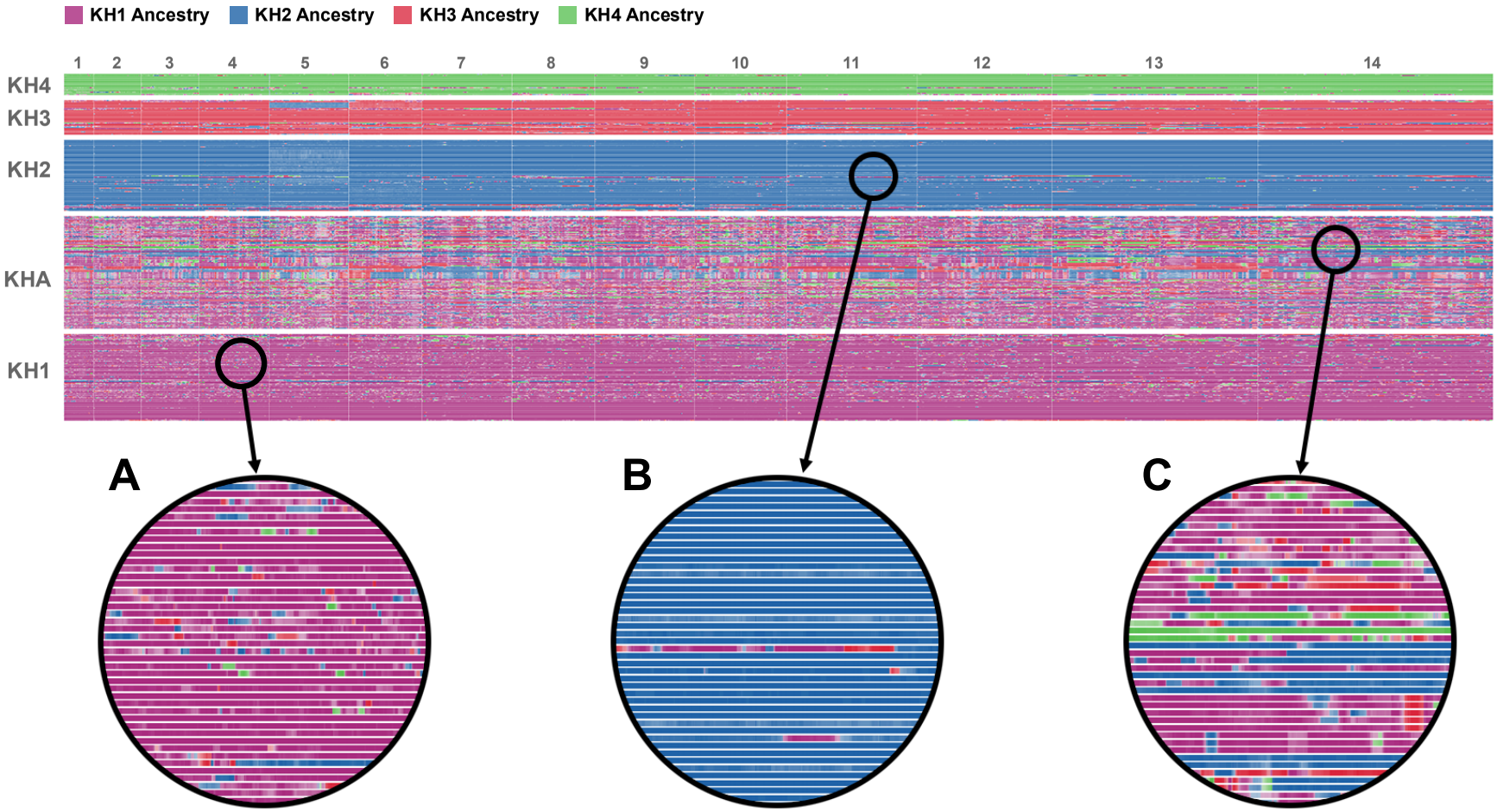

Figure 2.6: (Top) High-level view of chromosome painting across all 14 chromosomes, each row represents a sample. (A-C) Callouts show details for some features of the painting plots. (A) KH1 samples contain many short segments from the other groups, consistent with KH1 representing the ancestral population. (B) KH2, KH3 and KH4 samples comprise a few but very long segments labeled as belonging to other populations, suggesting recent admixture. (C) KHA samples contain a larger number of both short and long segments, consistent with the hypothesis that it is in fact an admixed population.

Location	KH1	KH2	KH3	KH4	KHA	Total
Pailin	4 (1)	14 (5)	5 (3)	0	26 (21)	49 (30)
Tasanh	1 (1)	9 (9)	3 (3)	0	34 (34)	47 (47)
Pursat	10 (4)	32 (21)	15 (12)	17 (12)	73 (40)	147 (89)
Ratanakiri	48 (44)	0	0	0	2 (2)	50 (46)
Total	63 (50)	55 (35)	23 (18)	17 (12)	135 (97)	293 (212)

Table 2.2: Distribution of the Cambodian cluster samples by geographic location. Counts of samples for which a clinical artemisinin phenotype was available are shown in parentheses.

found clear evidence of population structure in each study, which was not limited to a particular sampling location or year. KH1 was predominant in northeastern Cambodia (Ratanakiri) while all groups were present in western Cambodia. We concluded that all subpopulations were coexisting within a small geographical region in western Cambodia (i.e. were sympatric) despite being clearly distinct at the genetic level.

Year	KH1	KH2	KH3	KH4	KHA	Total
2007	-	-	-	5	25	30
2008	5	16	7	-	55	83
2009	1	8	3	-	10	22
2010	10	18	5	8	18	59
2011	47	13	8	4	27	99
Total	63	55	23	17	135	293

Table 2.3: Distribution of the Cambodian cluster samples by sampling year. Samples from all clusters but KH4 were found in 2008, 2009, 2010, and 2011. Samples from KH4 were missing in 2008 and 2009 but present in 2007, indicating continuity. These data suggest that the subpopulations are not transient.

2.4.4 Association with resistance

We investigated the relationship that these sympatric subpopulations could have with artemisinin resistance. To do that, we estimated (*in vivo*) parasite clearance half-life rates in patients with severe malaria after artesunate treatment [Dondorp et al., 2009], [Amaratunga et al., 2012].

Clearance data were available for 212 samples. After stratifying by subpopulation, we

Year	KH1	KH2	KH3	KH4	KHA	Total
ARC3	2	15	6	-	55	78
PF11	-	-	-	5	22	27
PF12	-	-	-	-	3	3
PFV2	58	32	15	12	53	170
Other	3	8	2	-	2	15
Total	63	55	23	17	135	293

Table 2.4: Distribution of the Cambodian cluster samples by study. Cluster presence was heterogeneous across studies, with all clusters being represented in at least two independent studies, suggesting subpopulations are not artifacts linked to specific studies.

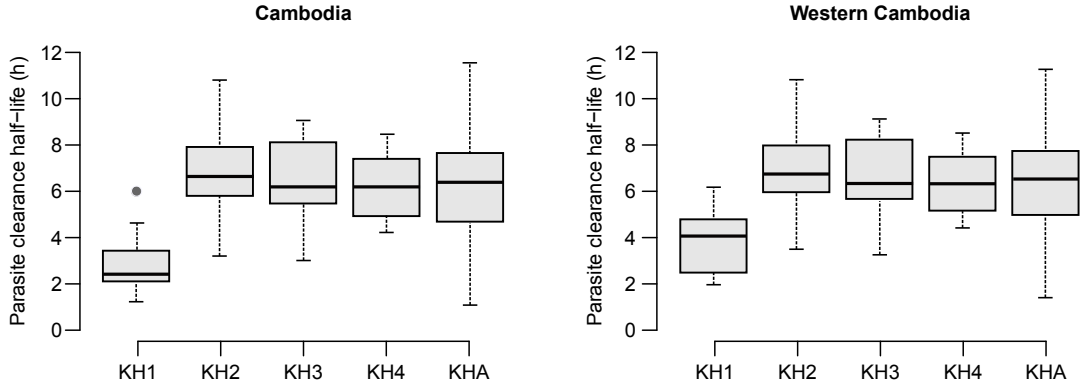

Figure 2.7: Box plots showing the distribution of artemisinin half-life phenotypes for Cambodian samples ($n = 212$) (left) and western Cambodian ($n = 166$) samples (right). Each box represents the interquartile range of values, split at the median; whiskers extend to the furthest points that are within 1.5 times the interquartile range.

identified a significant prolongation of parasite clearance half-life in the outlier groups (KH2, KH3, and KH4) when compared to the KH1 subpopulation (Figure 2.7).

The median half-life ranged from 6.4h to 6.8h in the outlier groups and had a value of 2.7h in the KH1 group. In the samples classified as KHA (admixed) we also observed longer parasite clearance half-life (median of 6.6h) than in the KH1 group. KHA phenotypes had the largest variance, a fact that was consistent with the hypothesis of the KHA group representing an admixture of artemisinin-resistance and artemisinin-sensitive parasites.

We did not observe any correlation between admixture proportions and parasite clearance half-life (Figure 2.8). The lack of correlation suggested that artemisinin-resistant phenotypes should depend on inheriting a limited set of DNA segments from resistant parasites. These findings held when we analyzed the subset of samples present in western Cambodia (166 of

	Cambodia			W. Cambodia		
Group	N	**Half-life**	P	N	**Half-life**	P
KH1	50	2.7 h		6	4.1 h	
KH2	35	6.8 h	2×10^{-14}	35	6.8 h	2×10^{-4}
KH3	18	6.4 h	7×10^{-9}	18	6.4 h	6×10^{-3}
KH4	12	6.4 h	1×10^{-7}	12	6.4 h	2×10^{-3}
KHA	97	6.6 h	7×10^{-16}	95	6.7 h	4×10^{-3}

Table 2.5: Artemisinin half-life phenotypes for Cambodian samples. We indicate the number of samples (N), the median parasite clearance half-life and the P value of a Mann-Whitney test comparing the half-life distribution for the subpopulation against that of KH1. We performed the analysis on all Cambodian samples and also using samples from western Cambodia alone (Pailin, Tasanh and Pursat sites). Clearance half-lives are significantly longer in the KH2, KH3 and KH4 subpopulations and in the KHA admixed population.

212 samples). Table 2.5 summarizes the comparison between KH1 and all other populations when focusing on all Cambodian samples or only samples from western Cambodia.

Different clinical studies have characterized faster parasite clearance rates in northeastern Cambodia [Lim et al., 2013][10] and in Vietnam [Hien et al., 2012] than in western Cambodia. This is consistent with our observation of sensitive KH1 parasites being predominant in northeastern Cambodia and the presence of the KH2, KH3, KH4 and KHA subpopulations in the western region.

2.4.5 Subpopulations as product of strong founder effects

The most noticeable characteristics of the artemisinin-resistant subpopulations (KH2, KH3 and KH4) were their strong genetic differentiation (relative to KH1 and each other) and their reduced haplotype diversity (indicating high levels of inbreeding). These features were visible in our exploratory analysis of population structure. The neighbor joining tree showed small clades with very short branches (Figure 2.1) whereas PCoA exposed many samples collapsing onto the same three first PC coordinates (Figure 2.2 and Figure 2.3), revealing in both cases the almost clonal nature of some sets of parasites. The fact that the first three principal components were driven by the presence of the outlier groups suggested strong genetic differentiation of these clusters. Moreover, when a neighbor joining tree was

[10]Unpublished data at the time of the publication of our article.

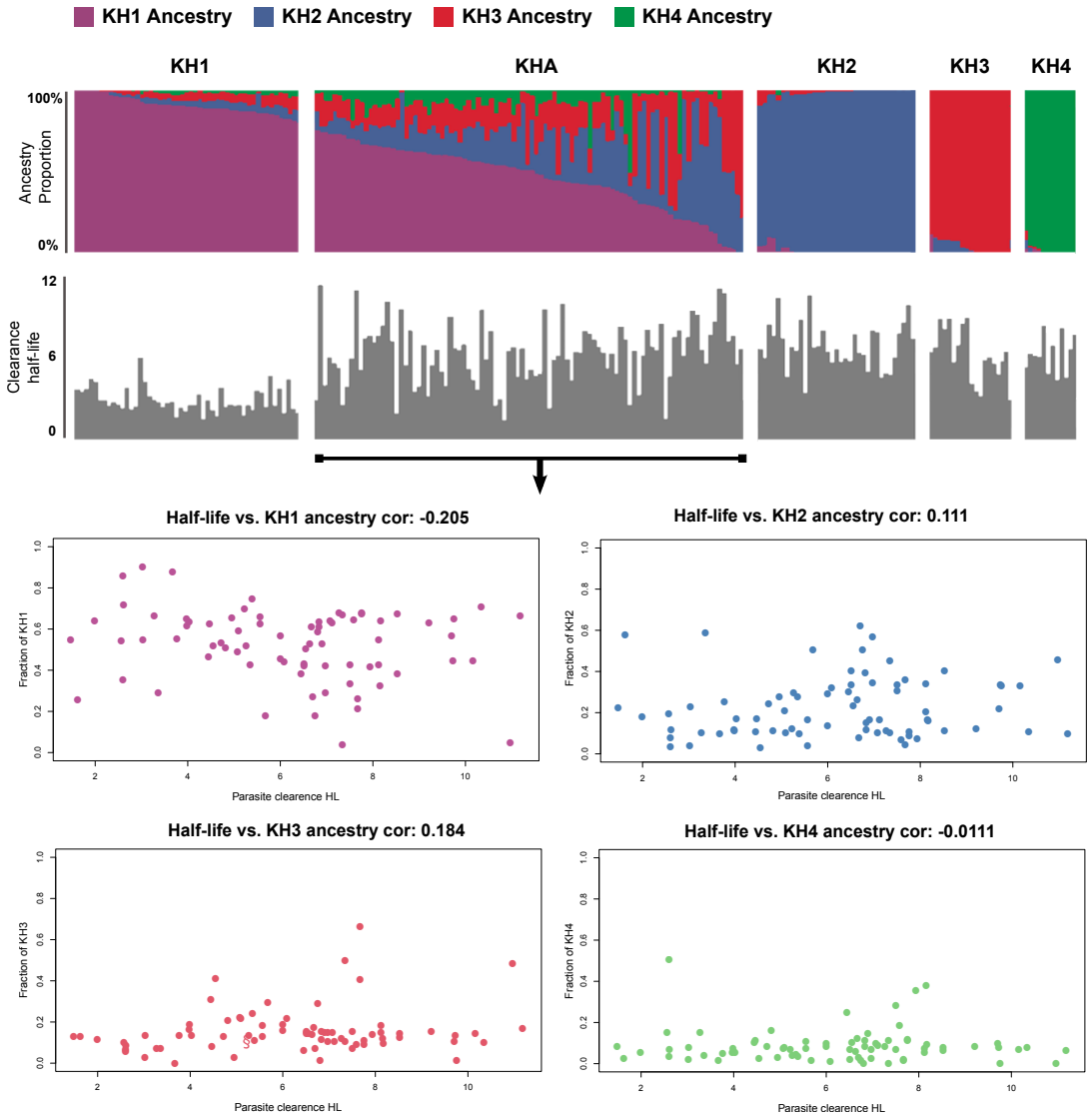

Figure 2.8: (Top) Plots showing the ancestry proportions derived from chromosome painting (above) and the associated parasite clearance half-life for the 212 Cambodian samples with a phenotype (below). (Bottom) Plots showing the relationship between the ancestry proportion ascribed to different groups and phenotypes for KHA samples. We do not observe any strong correlation at genome level between admixture proportions and half-life within the KHA group. Furthermore, regressing half-life against cluster proportions rendered non-significant results.

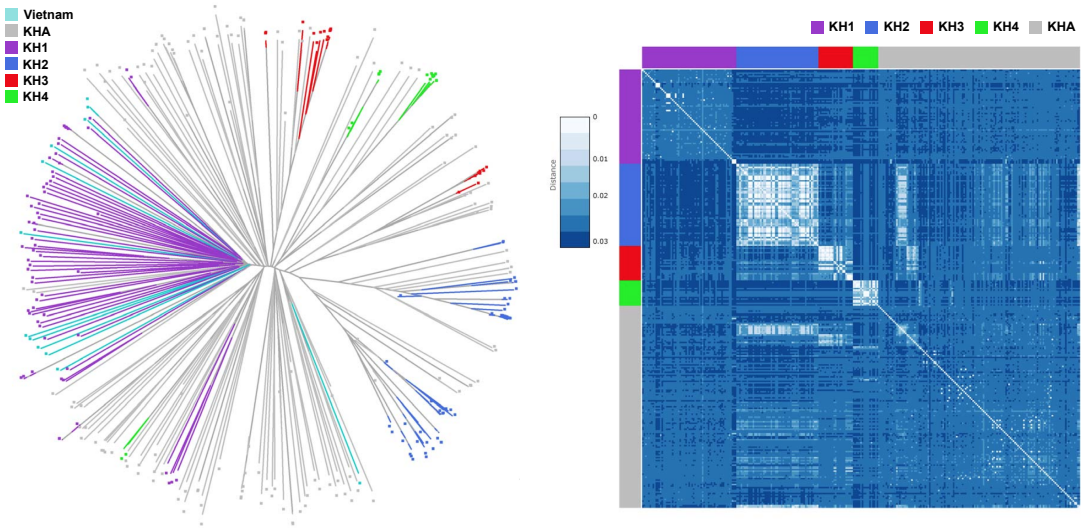

Figure 2.9: (Left) Neighbor joining tree of Cambodian and Vietnamese samples. Samples belonging to the Cambodian artemisinin-resistant subpopulations (KH2, KH3 and KH4) tend to form small clusters, distinct from the main KH1/Vietnam core group. (Right) Heatmap visualization of the parwise distance matrix computed using all Cambodian samples. It is clear that samples within each outlier subpopulation (KH2, KH3 and KH4) are strongly interrelated.

built for Cambodian and Vietnamese samples alone (Figure 2.9, left), the outlier groups tended to cluster tightly and separated from the main KH1/Vietnam core cluster. The cited features can also be spotted in the pairwise genetic distance matrix (Figure 2.9, right) that served as input for the neighbor joining and PCoA methods.

The model-based methods corroborated these observations. Chromosome painting revealed that individuals from KH2, KH3 and KH4 could be reconstructed almost exclusively from samples belonging to the same subpopulation with very little uncertainty (Figure 2.6), consistent with the groups being very differentiated and inbred. To further assess the degree of genetic differentiation, we computed all pairwise genome-wide estimates of F_{ST} between groups (Table 2.6). We included the Cambodian subpopulations and the populations of Thailand, Vietnam, Ghana and The Gambia. The pairwise comparison showed an extreme degree of differentiation between outlier subpopulations. For instance, the genome-wide F_{ST} value when comparing KH2 against KH4 (0.38) was higher than the value obtained when comparing Thailand and Ghana (0.16). As expected from these results, we also found an excess of SNPs with high F_{ST} when comparing any pair of outlier groups (Table 2.7). These results uncovered a greater level of relative[11] differentiation between

[11]Since F_{ST} is relative to the total heterozygosity in each case.

	GH	GM	TH	VN	KH1	KH2	KH3	KH4	KHA
GH	-	0.01	0.16	0.17	0.15	0.3	0.26	0.3	0.17
GM	0.01	-	0.15	0.16	0.14	0.27	0.24	0.27	0.16
TH	0.16	0.15	-	0.06	0.04	0.18	0.15	0.21	0.03
VN	0.17	0.16	0.06	-	0.04	0.21	0.18	0.23	0.05
KH1	0.15	0.14	0.04	0.04	-	0.2	0.16	0.22	0.03
KH2	0.3	0.27	0.18	0.21	0.2	-	0.29	0.38	0.14
KH3	0.26	0.24	0.15	0.18	0.16	0.29	-	0.33	0.12
KH4	0.3	0.27	0.21	0.23	0.22	0.38	0.33	-	0.19
KHA	0.17	0.16	0.03	0.05	0.03	0.14	0.12	0.19	-

Table 2.6: Mean genome-wide pairwise F_{ST} values, computed by using a bootstrapping procedure [Efron, 1992] for 1,000 iterations (mean). Outlier subpopulations are very differentiated from each other and the rest of populations. Values ≥ 0.20 are indicated in red.

subpopulations in western Cambodia than between representative parasite populations of Africa and Southeast Asia, due to the low genetic diversity of the subpopulations.

Several genetic features suggested that the extreme differentiation of the KH2, KH3 and KH4 subpopulations was due to a very recent and strong founder effect. Firstly, the allele frequency spectrum of these groups showed a deficit of low-frequency variants when compared to any other population in the study (Figure 2.10). Secondly, we did not find any significant difference in the ratio of nonsynonymous to synonymous polymorphisms when comparing the sets of highly ($F_{ST} > 0.9$) and slightly differentiated SNPs ($F_{ST} < 0.2$) of the three Cambodian subpopulations. Overall, pN/pS was 2.1 and 2.4 respectively, suggesting that differentiation was not caused by selection[12] but rather by a clonal population expansion.

High levels of inbreeding were also evident from a substantial reduction in haplotype diversity within the KH2, KH3 and KH4 subpopulations. We measured haplotype homozygosity and diversity in a sliding window along the genome and found higher levels of homozygosity and reduced diversity in the KH2, KH3, and KH4 populations when compared with KH1 (Figure 2.11, top and bottom left). In some cases, we found that a single haplotype was present across large regions. For instance, in the KH4 subpopulation, all samples shared basically a unique haplotype across most of chromosome 9, whereas in KH2 a single haplotype spanned half of chromosome 13 (1.3-3.4 Mb). Accordingly, the outlier clusters showed higher levels of linkage disequilibrium when compared to KH1 and other populations (Figure 2.11, bottom left).

The analysis of within-host diversity (F_{WS}) was also consistent with the high levels of

[12]Usually a higher pN/pS ratio is expected for alleles at high frequencies as a result of the cumulative effects of selection through time [Landry and Aubin-Horth, 2013].

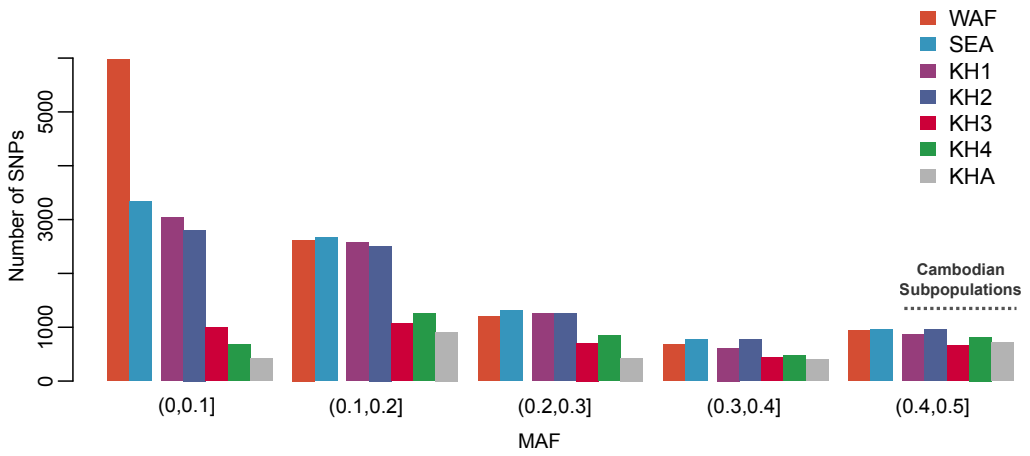

Figure 2.10: Minor allele frequency (MAF) spectra for seven populations. The lower boundary is excluded in each bin. A subsampling procedure was used to minimize confounding by differences in sample size (we used the mean frequency).

	GH	**GM**	**TH**	**VN**	**KHA**	**KH1**	**KH2**	**KH3**	**KH4**
GH	-	0	562	521	586	510	1386	1128	1390
GM	0	-	527	515	556	498	1346	1080	1368
TH	562	527	-	7	0	11	382	219	495
VN	521	515	7	-	1	0	480	263	541
KHA	586	556	0	1	-	5	119	84	358
KH1	510	498	11	0	5	-	482	279	529
KH2	1386	1346	382	480	119	482	-	729	1200
KH3	1128	1080	219	263	84	279	729	-	989
KH4	1390	1368	495	541	358	529	1200	989	-

Table 2.7: Number of highly differentiated SNPs ($F_{ST} > 0.8$) found when comparing population pairs. This table confirms the genome-wide characterization presented in Table 2.6. Values ≥ 700 are indicated in red.

inbreeding observed in these populations (see Section 2.4.7). Most samples in the KH2, KH3 and KH4 subpopulations looked effectively clonal ($F_{WS} \simeq 1$), as did many samples in the admixed (KHA) group (Figure 2.12). In contrast, samples that did not belong to the artemisinin-resistant groups, with the exception of Vietnam[13], were less likely to be clonal, indicating that recombination between different genetic types is less constrained in the ancestral type populations.

Based on all the evidence presented here, we conclude that the KH2, KH3 and KH4 subpopulations in western Cambodia, each of which is artemisinin-resistant, were the result of a recent founder effect. These groups coexist (i.e. are sympatric) with a more diverse

[13]This may be due, among other things, to study sampling bias, lower transmission rates or an expression of Vietnamese parasites being genuinely more inbred.

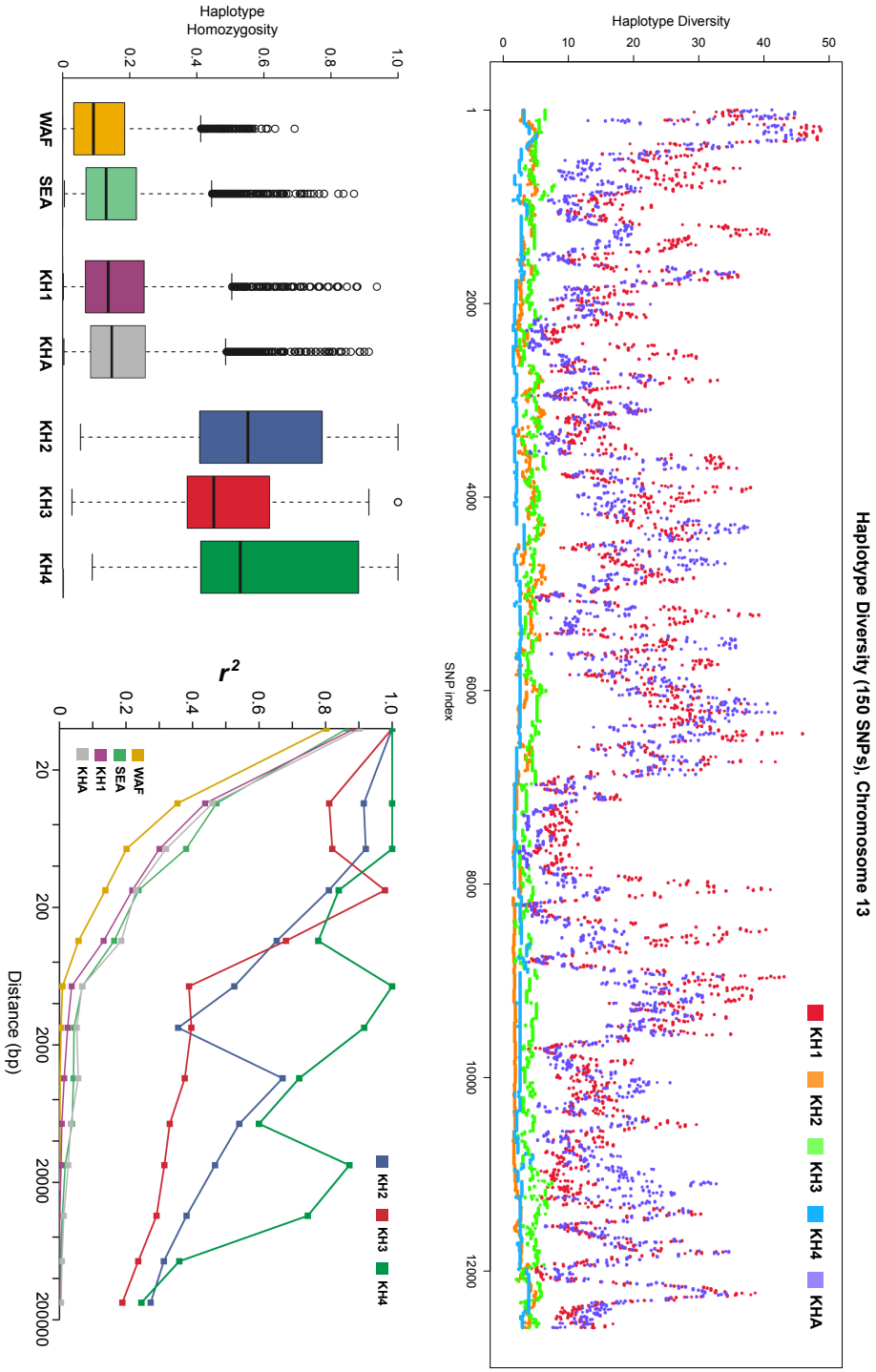

Figure 2.11: (Top) Haplotype diversity (i.e. number of different haplotypes) along chromosome 13 (overlapping windows of 150 SNPs), showing the almost clonal nature of the founder subpopulations (KH2, KH3 and KH4) for large genomic regions (legend colors are different for this plot). (Bottom Left) Genome-wide distribution of haplotype homozygosity, computed using a sliding window of 201 SNPs. (Bottom Right) Genome-wide decay of LD, expressed as r^2. SNPs with MAF > 0.15 were organized into equally-sized frequency bins. Each bin was analyzed independently and was corrected for the effects of sample size and population structure before bin results were aggregated.

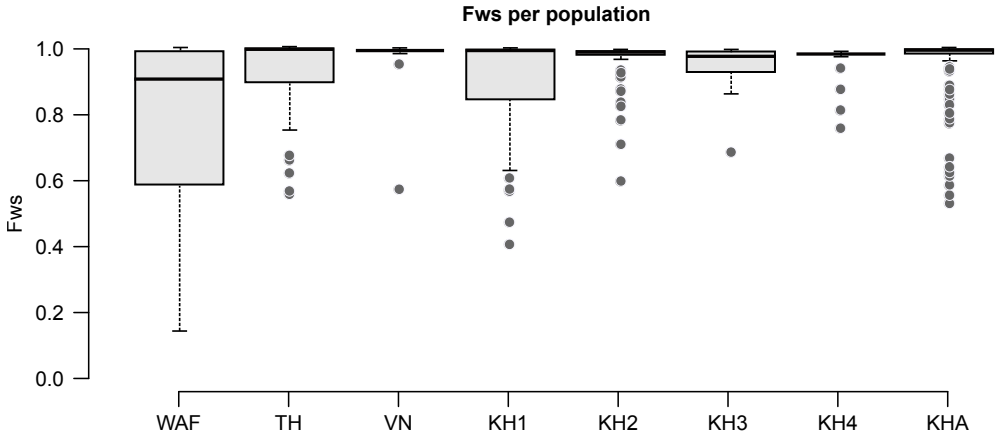

Figure 2.12: Distribution of F_{WS}. Clonal infections are common in Southeast Asian populations (TH, VN, KH1, KH2, KH3, KH4 and KHA) whereas higher levels of within-sample diversity characterize samples from West Africa (WAF). The founder populations (KH2-KH4), the admixed samples (KHA) and samples from Vietnam (VN) present very low levels of within-sample diversity, in agreement with the high levels of inbreeding observed in these groups. Each box represents the interquartile range of values, split at the median; whiskers extend to the furthest points that are within 1.5 times the interquartile range.

and widely distributed population (KH1) to which they seem to be ancestrally related. The analysis of population structure, (in particular chromosome painting), suggest that there is some degree of gene flow between these groups, with KHA being an admixed mosaic of the other populations (Figure 2.6). The wide range of copied chunk lengths in KHA samples indicates that the admixture event could be relatively recent and, most likely, still undergoing. However, in this study we do not try to impose a temporal scale or to depict the precise sequence of demographic and evolutionary events that shaped these populations. Future studies may be able to elucidate the nature and the timescale of the processes responsible for these observations.

2.4.6 Differentiated genetic markers

We provided a comprehensive list of differentiated genetic markers for the artemisinin-resistant founder populations to facilitate the monitoring of their spread outside western Cambodia. Within such populations, we also identified several antimalarial resistance alleles and transporter genes with highly differentiated sequences. Finally, we highlighted a set of mutations that might confer hypermutability to parasites from the KH2 subpopulation. All these data can be found in the supplementary material[14] of the original article [Miotto

[14]Not included in this thesis because of the sheer size of the tables

et al., 2013].

2.4.7 Biases caused by complexity of infection

Because haplotypic artifacts (i.e. the creation of chimeric haplotypes due to multiplicity of infection) can affect haplotype-based analyses, we assessed the degree of within-sample diversity for all the samples in the study. We relied on the F_{WS} statistic [Manske et al., 2012], which can be interpreted as a relative inbreeding coefficient. We remind the reader that F_{WS} summarizes the relationship between population and within-sample heterozygosity. Clonal infections are characterized by an F_{WS}value close to 1 (an exact value of 1 means that there is no detectable within-sample diversity). Conversely, F_{WS} values close to 0 represent infections with a degree of diversity that is comparable to the one found at the population level. We observed that the vast majority of Southeast Asian samples represented clonal or almost clonal infections (Figure 2.12), including the majority of admixed samples classified as KHA, indicating very high levels of inbreeding. Therefore, we concluded that complexity of infection did not affect any of the key findings of the study in a substantial way.

2.5 Discussion

Population structure in *P. falciparum* is usually related to the transmission regime of the area, with more structured and inbred populations arising in regions with low intensity of transmission [Anderson et al., 2000], [Dye and Williams, 1997]. However, this fact could not explain the extreme patterns[15] of population structure found in western Cambodia, where sympatric subpopulations of parasites coexist within a small geographical area and have exceptional levels of genetic differentiation. Three of these subpopulations present a deficit of low-frequency variants and very reduced haplotype diversity, suggesting recent founder effects. As these populations are associated with clinical artemisinin-resistance, it is important to assess the biological relevance of their association with resistance and to address the possible origins of the founder effects.

We hypothesize that the founder effects represent the recent expansion of resistant

[15]When compared to the genetic structure observed in other parasite populations, see for instance [Manske et al., 2012].

parasite lineages whose fitness depends on a specific genetic background[16] (i.e. combination of alleles). Our reasoning is based on two assumptions. Firstly, that parasites under drug pressure will eventually acquire genetic variants that endow reduced drug sensitivity. Secondly, that many potential resistance-conferring alleles will be associated with a biological fitness cost in the absence of the drug[17]. In this setting, we expect that sexual recombination would shuffle variants, occasionally producing parasites with a particular set of alleles able to confer drug resistance and also compensate for any fitness disadvantage. We postulate that the selective advantage of such a genetic background would cause the progeny of the selected lineage to expand, triggering a founder effect. It follows that, under this scenario, outcrossing would not be favored since it could disturb the genetic background required to sustain the parasite's resistance and optimal fitness.

If this hypothesis is largely correct, other factors will play a role in the emergence and spread of resistance. For instance, low transmission rates and other aspects that favor inbreeding, such as physical isolation in remote areas or adaptation to different *Anopheles* species [Sinka et al., 2011], could help the conservation and propagation of the required genetic background once it has emerged.

In light of these observations, we briefly revisit the historical circumstances of the emergence of antimalarial drug resistance in western Cambodia. We have noticed that several haplotypes associated with resistance to other antimalarial drugs are present in the subpopulations of this region. This might well be the product of a similar sequence of evolutionary and demographic events that happened under elevated drug pressure in the late 1950s and early 1960s, when mass administration of chloroquine and pyrimethamine occurred in Pailin [Payne, 1988], [Verdrager, 1986]. In this context, human demographic factors in the area might have had a fundamental part. The restricted migration during the period of Khmer Rouge resistance (1979-1998) and the isolation of rural settlements in forested mountain regions due to poor infrastructure could have favored the inbreeding of parasites.

The discovery of multiple and genetically different subpopulations of resistant parasites in western Cambodia may imply that different types of resistance need to be monitored

[16]Here we refer to genetic background as a set of alleles at multiple loci that confer resistance and potentially compensate for any fitness penalty in the absence of the drug.

[17]See 1.3.3 for a summary of the views on the genetic basis of drug resistance.

and controlled to avoid spread to other countries. The definition of genetic markers for these subpopulations will help in this endeavor, but further genetic epidemiological studies are needed to understand the dynamics provoking the emergence of resistance.

Finally, the strong association between artemisinin resistance and founder populations, leads us to hypothesize that population structure and founder effects may predate and facilitate the emergence of resistance in Southeast Asia.

2.6 Materials and methods

In this section, we review and summarize the methods that are relevant to the work described in this chapter. A complete account can be found in the online methods section and the supplementary information of the original publication [Miotto et al., 2013].

2.6.1 Sequencing and genotyping

For 670 (out of 825) samples, parasite-infected RBCs where obtained from blood samples. We performed leukocyte depletion to remove the majority of human DNA without culturing. To amplify DNA quantities, we cultured *in vitro*[18] a subset of samples from Cambodia (51) and Thailand (104). After DNA extraction, we determined the amount of human and *Plasmodium* DNA by PicoGreen analysis and quantitative real-time PCR. We used the Illumina Genome Analyzer II to sequence samples with more than 1 μ g of DNA and less than 60% human DNA contamination, with read length ranging from 37 to 105 (of paired-end reads). We obtained a median of 2Gb of read data per sample. Short reads were aligned to the *P. falciparum* 3D7 reference genome v2.1.5[19]. We produced a candidate list of 1,313,570 SNPs by analyzing an initial set of 425 samples with the Burrows-Wheeler Aligner (BWA) [Li and Durbin, 2009].

We removed spurious SNPs by running the SNP-o-matic algorithm [Manske and Kwiatkowski, 2009] as a stringent filtering step, only allowing polymorphisms present in the set of candidate SNPs. We refined the dataset by applying a sequence of quality filters with the aim of removing errors and artifacts, finally obtaining a subset of 86,158 high-quality SNPs. These SNPs were genotyped based on the number of reads assigned to

[18] Culturing time was in the order of weeks.

[19] `ftp://ftp.sanger.ac.uk/pub/pathogens/Plasmodium/falciparum/3D7/3D7.version2.1.5/`

each allele (reference or alternative). We required a minimum of five reads to make a call; otherwise, the genotype was set as missing. If due to the presence of multiple infections reads for more than one allele were present, we called the allele with the majority of reads (*majority calling* heuristic). In case of a draw, we sampled one of the alleles randomly. This approach approximated the dominant strain in an infection with multiple genetic types that have not diverge substantially from each other (i.e. F_{WS} close to 1).

2.6.2 Distance computation, NJ and PCoA

Pairwise distances[20] were computed as the normalized number of mismatches (i.e. genotype differences) between each pair of samples, discarding comparisons with missing calls. To build the tree, we feed this 825×825 distance matrix to the `nj` implementation of the `R ape` package [Paradis et al., 2004].

We performed principal coordinate analysis (PCoA) [Gower, 1966] instead of PCA. The method is very similar to PCA but uses a pairwise distance or dissimilarity matrix as input. PCoA also uses a linear mapping for projecting the data onto the ordination space and tries to maximize the amount of variance explained by the projections [Ramette, 2007]. The main advantage of PCoA is that it is more efficiently computed for large genetic datasets, as it only requires a summary pairwise similarity matrix[21]. Given the binary nature of our data, differences in the projections were negligible for our study. We opted for the more efficient PCoA method, as we did not require an account of variable (i.e. SNP) contribution to each component. To highlight significant differentiating markers we used well-defined population genetic statistics, which were easier to interpret. We used the `R` language `cmdscale` implementation for performing PCoA.

We applied a subsampling strategy to test if the PCoA signal was driven by unequal sample sizes. We considered the three most represented countries in our dataset (Ghana, Cambodia and Thailand). For the subsampling procedure, we set $n = 97$, the sample size of the smallest group (Thailand), and randomly picked that number of samples from the other populations. We performed PCoA in this reduced dataset and found similar results, concluding that differences in sample sizes could not account for the structure present in

[20] We scale our genetic distances by 0.7, as our genotypes covers approximately 70% of the coding genome. However this scaling factor is irrelevant for the analyses presented here.

[21] We reduced the input matrix from $825 \times 86,158$ to 825×825.

Cambodia.

2.6.3 Allele frequency analysis

We estimated non-reference allele frequency (NRAF) as the proportion of genotyped samples in the population that did not possess the reference allele. We excluded undetermined genotypes (i.e. missing calls) from the computation. For calculating the minor allele frequency (MAF), we used the proportion of genotyped samples carrying the least common genotype in the population. After discretizing the MAF into five equally sized bins, we used a subsampling procedure (averaged over 1,000 iterations) to minimize confounding due to uneven sampling of populations.

When the derived allele frequency (DAF) was required, we determined the putative ancestral and derived genotypes at each SNP by comparing the *P. falciparum* 3D7 reference sequence with the sequence of *Plasmodium reichenowi* [Manske and Kwiatkowski, 2009]. For delimiting gene reading frames and exon boundaries we relied on the PlasmoDB 5.5 annotations of the 3D7 genome [Aurrecoechea et al., 2009]. Synonymous and nonsynonymous amino acid changes were inferred by substituting the reference allele with the non-reference allele at that SNP in the 3D7 genome without other changes.

2.6.4 Chromosome painting

We used the `ChromoPainter` package available at the time (v1.0, `www.paintmychromosomes.com`). The value of the parameters N_e (effective population size) and μ (mutation or miscopying probability) were estimated in Cambodian samples by running an expectation-maximization (*E-M*) algorithm during 40 iterations, for each sample and chromosome, maximizing over copying proportions. Due to some particularities of the *E-M* implementation, we excluded a small number of samples from this procedure. The aim was to ensure that no two samples in the dataset were more than 99.5% similar, as recommended[22] by the authors of the package. The final values for the parameters ($N_e = 8824$, $\mu = 0.000492$) were computed as weighted means, where weights were assigned by chromosome length. Here we emphasized the role of these parameters solely as a scaling factor and a miscopying probability, respectively, for the underlying HMM model. By no means these values should

[22] Personal communication.

be interpreted, in the context of population genetics theory, as characterizing an idealized Wright-Fisher population representative of our dataset. N_e scales the recombination rate, thus affecting the probability of switching to a different donor individual when performing the painting. An infinite value would force markers to be independent of each other (i.e. discarding linkage information). μ denotes the probability of the model emitting a mutation when copying from a donor, therefore copying a marker that is different from the one in the haplotype being painted. We assumed a uniform recombination map with a recombination rate of 10^{-8} per base pair per generation[23].

2.6.4.1 Genome-wide painting

The output of the painting process is a matrix of posterior copying probabilities for each locus and individual. Each entry, (i, l), in this matrix specifies the probability that we copy locus l from individual i when reconstructing the haplotype of the chromosome being painted. This matrix has dimensions $N \times L$, where N is the number of individuals and L the number of loci in the chromosome. To produce the genome-wide painting, we aggregated these probabilities for each putative population and assigned, to each locus and individual, the color of the population with the highest copying probability (i.e. we use the maximum a posteriori estimate regarding populations).

To express uncertainty, we dimmed colors according to the information entropy of the aggregated copying probabilities. We computed information entropy following $H(p) = -\sum p_i \log_K(p_i)$ [MacKay, 2003], where p is a vector of length K containing the proportion of chunks copied from population i and $\sum p_i = 1$. Maximum entropy is achieved when the probability of copying from any population is the same ($1/K$, where K refers to the number of populations) thus dimming the assigned color to pure white.

2.6.5 ADMIXTURE

To remove dependencies due to linkage disequilibrium, we trimmed the SNP set according to the observed correlation coefficients. Using the `PLINK` toolset [Purcell et al., 2007], we scanned the genome with a sliding window of 100 SNPs in size, advanced in steps of 10

[23] Due to the lack of reliable recombination maps for *falciparum*.After being scaled by N_e, the recombination was in the range of 15-30 Kbs per cM, in agreement with the rates observed in the literature [Mu et al., 2005].

SNPs, and removed any SNP with a correlation coefficient $r^2 \geq 0.02$ with any other SNP within the window. Additionally, we removed all SNPs with extremely low minor allele frequency (MAF ≤ 0.01), as these SNPs are less informative in the inference process. The 5,484 SNPs that remained after this filtering step were used to run ADMIXTURE with 5-fold cross-validation, 1,000 replicates for bootstrapping and K values (number of putative populations) ranging from 2 to 5. We assessed the value of K by using the cross-validation error produced by ADMIXTURE and found $K = 4$ to be a good candidate (and comparable with our chromosome painting results) using the heuristic elbow rule suggested by the authors of the software.

2.6.6 Clinical Phenotypes

We estimated (*in vivo*) parasite clearance half-life rates in patients with severe malaria after artesunate treatment [Dondorp et al., 2009], [Amaratunga et al., 2012]. We computed these rates from frequent blood parasite counts, using the slope of the linear portion (between the lag and tail phases) of the parasite clearance curve [Flegg et al., 2011]. For this task, we used the Parasite Clearance Estimator developed by the Worldwide Antimalarial Resistance Network (WWARN). The details of the studies that collected samples were registered at `http://clinicaltrials.gov` for Pursat (NCT00341003) and Ratanakiri (NCT01240603). Similarly, the study collecting samples in Pailin was recorded at ISRCTN (ISRCTN15351875).

2.6.7 F_{ST} estimation

We used a definition of F_{ST} equivalent to that of Hudson (see Section 1.4.3.3). For a given locus, l, we estimated F_{ST} following

$$F_{ST} = 1 - \frac{\hat{\Pi}_S^l}{\hat{\Pi}_T^l}, \tag{2.1}$$

where

$$\begin{aligned} \hat{\Pi}_S^l &= \frac{2p_1^l(1-p_1^l) + 2p_2^l(1-p_2^l)}{2}, \\ \hat{\Pi}_T^l &= (p_1^l + p_2^l)(1 - \frac{p_1^l + p_2^l}{2}). \end{aligned} \tag{2.2}$$

In our estimator, p_1^l and p_2^l are the frequencies of the allele observed at locus l in the two populations being compared. For genome-wide F_{ST} estimates between two populations (F_{ST}^{GW}), we averaged across L loci following

$$F_{ST}^{GW} = 1 - \frac{\sum_i^L \hat{\Pi}_S^i}{\sum_i^L \hat{\Pi}_T^i}. \tag{2.3}$$

We computed 90% confidence intervals for these estimates by using the non-parametric bootstrap [Efron, 1992] with 1000 iterations.

2.6.8 Haplotype diversity and LD analysis

We imputed missing calls[24]. (i.e. coverage was < 5 reads) in order to reconstruct haplotypes. For a given missing call, we used the most frequent allele in the population that had the same flanking alleles than the missing call. If no valid genotypes were found on the flanks, we just used the most frequent allele in the population. We analyzed haplotypes by using a sliding window of 201 SNPs, advanced in steps of 20 overlapping SNPs. No haplotype contained SNPs separated by more than 20kbp. We computed the frequency (p_i) of all N haplotypes present in a population within each window, recording the haplotype homozygosity $H_0 = \sum_i^N p_i^2$. We also produced genome-wide scans of haplotype diversity (i.e. the number of different haplotypes), sliding a window of 150 SNPs (one SNP at a time). For analyzing LD decay we relied on the procedures described at length in [Manske et al., 2012], correcting for the effects of sample size and population structure.

2.6.9 Ethical approval

Local ethics committee approved sample collection at each location. Blood samples were collected with informed consent from patients[25] that had been diagnosed with *P. falciparum* malaria.

[24]The majority of samples presented very low *missingess* with just a few samples reaching values below 20%. The distribution of missing calls appeared random and sparse along the genome.

[25]Or a parent or a guardian.

2.7 Individual contributions

My contributions to this study were primarily on the identification and characterization of population structure, its relationship with drug resistance, and the elucidation of plausible demographic scenarios, which have been described at length in this chapter. I specifically worked with the first author (Olivo Miotto) during the exploratory data analysis phase, studied population structure with model-based methods on my own and, again, worked with Olivo and others during the assessment of plausible demographic scenarios (i.e. founder effects).

I did not play any role in the collection of the blood samples, sample preparation, sequencing, phenotype collection or genotyping efforts. Similarly, I did not contribute to the extensive selection of differentiated and candidate markers in the founder populations, the analysis of LD decay or the characterization of amino-acid changes.

Chapter 3

Complexity of infection on deep sequencing data

Contents

3.1 Introduction

In this chapter, we take a brief detour to study how complexity of infection, the presence of multiple genetic types within a sample, can distort the results of our analyses when using deep sequencing data. Although *Plasmodium falciparum* is a haploid organism, when multiple strains coexist in a sample deep sequencing data render heterozygous genotypes, with each allele being supported by a different number of reads. We focus on the biases this phenomenon introduces in the estimation of genetic distances and downstream analyses that are dependent on these, such as building neighbor joining trees. We also review the F_{WS} statistic, used to characterize mixed infections, uncovering some inherent biases of its original formulation and introducing an alternative estimator that is better behaved. Although the results presented here have not been published *per se*, the insights gained from this investigation have been incorporated into subsequent publications. We start by commenting on the influence that *majority calling*, a heuristic method for calling heterozygous genotypes, have on the estimation of genetic distances and propose a simple correction scheme based on read count data. Next, we review the F_{WS} statistic and suggest a resampling procedure that takes into account uncertainty in the data. Finally, we propose an alternative estimator for F_{WS} that allows a higher degree of resolution and overcomes some of the limitations of its predecessor. We also study the robustness of this new estimator in situations of strong ascertainment bias. We remind the reader that we use the terms complexity of infection (COI), multiplicity of infection (MOI), multiple infection, mixed infection and mixed sample interchangeably. In the same spirit, we refer to heterozygous genotypes within a sample as mixed calls, mixed genotypes or *het* calls.

I claim full authorship for the work presented in this chapter. Nonetheless, I acknowledge my interactions and discussions with other colleagues in the last section of the chapter (Section 3.4.10).

3.2 Data

In the first part of this chapter, we used 1,723 *Plasmodium falciparum* samples from Africa, Asia, South America and Oceania, provided by the MalariaGEN community project. Sequences in this dataset have been produced following the same protocols and criterions

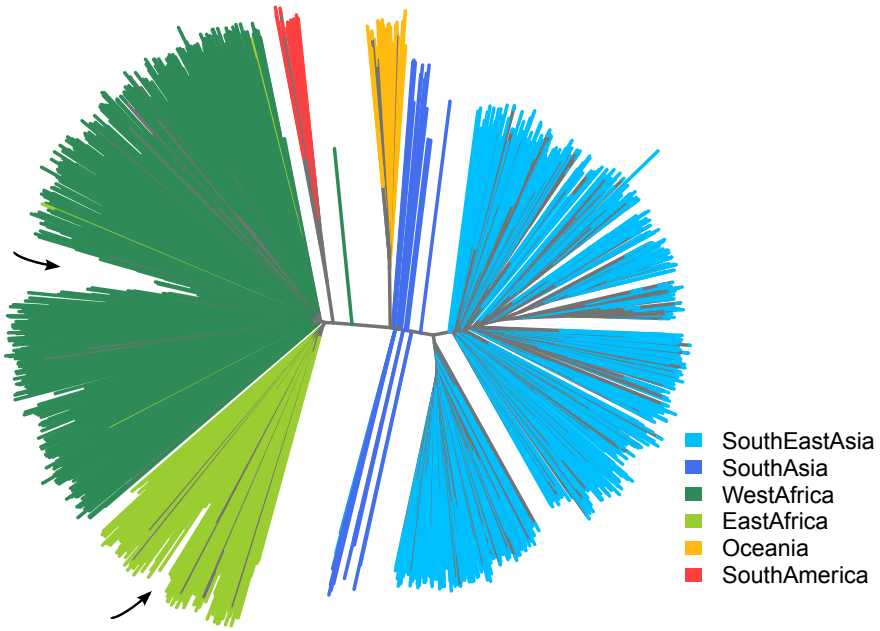

Figure 3.1: Genome-wide neighbor joining tree for 1,723 *P. falciparum* samples from Asia, Africa, Oceania and South America. Arrows highlight some of the sections of the tree where a group of samples seems to be considerably less diverse (i.e. shorter branches) than the other samples in the clade.

described at length in Chapter 2. For assessing the F_{WS} statistic, we relied on 2,500 *P. falciparum* samples from Africa and Asia offered by the Pf3k project [Pf3k, 2016], specifically in its pilot data release (v5.1).

3.3 Effects of majority calling

Majority calling is a heuristic procedure for calling mixed haploid genotypes (in a sample) from deep sequencing data. It merely consists of calling the haploid genotype supported by the major number of reads at a given locus. Assuming that a single strain dominates a mixed infection, the procedure approximates the genome of such dominant genetic type [Manske et al., 2012], [Miotto et al., 2013]. However, when this assumption is not met, and strains are significantly different from each other, artifacts are bound to introduce biases during analysis. For instance, haplotype reconstruction would produce haplotypes that are a mosaic of all the strains within the sample and haplotype diversity would be significantly overestimated. We focus on the effect that majority calling can have in the computation of genetic distances.

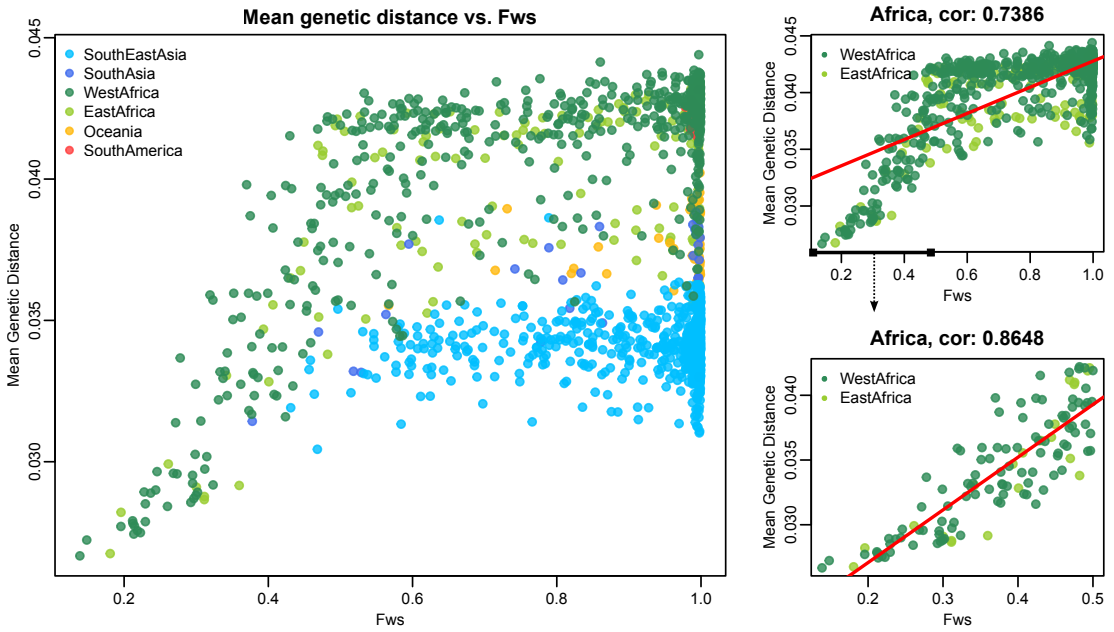

Figure 3.2: (Left) Mean genetic distance against F_{WS}. Distances decay fast with F_{WS} for samples with $F_{WS} \leq 0.5$. (Right, top) Linear regression (red line) and Pearson correlation (0.73) between mean genetic distance and F_{WS} for African samples. (Right, bottom) Same analysis as in the previous panel (Pearson correlation: 0.86) for African samples with $F_{WS} \leq 0.5$.

This assessment was triggered by the examination of surprising branch length patterns in genome-wide neighbor joining trees. We observed noteworthy differences on branch lengths within star-like clades (see Figure 3.1 and Figure 2.1).

Here, following [Manske et al., 2012] and [Miotto et al., 2013], we estimated genetic distance as the fraction of genotype mismatches[1], computed over genotype calls obtained by using the majority calling heuristic. Closer inspection of the genetic distance matrix used to build the tree revealed an obvious correlation between mean genetic distance (i.e. the average distance to any other sample in the dataset) and F_{WS}.

Figure 3.2 summarizes this finding. Each sub-continental population clusters around a particular distance (y axis) that is informative about the underlying amount of genetic diversity within the group. Genetic distances tend to cluster around these values for clonal samples but decay quickly with F_{WS} once the statistic goes below 0.5. This association was more apparent when focusing on African samples alone, as they tend to be more mixed and present a higher number of mixed calls. The Pearson correlation between genetic distances and F_{WS} reached 0.73 for all African samples and 0.86 for the subset of African samples

[1]Or number of nucleotide differences per site, also know as p-distance in the literature [Lemey, 2009].

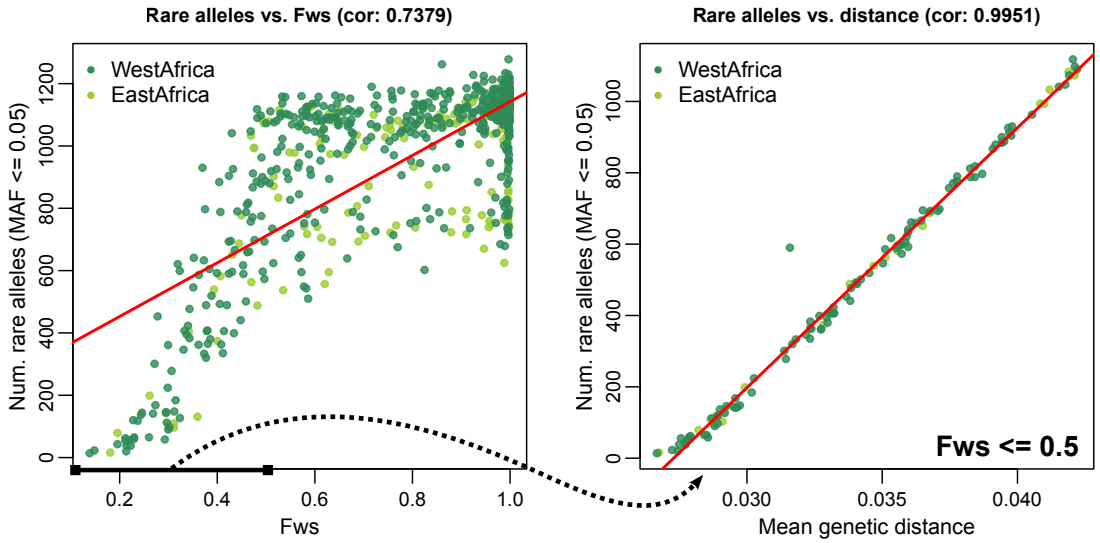

Figure 3.3: (Left) Number of rare variants (MAF ≤ 0.05) versus F_{WS} for African samples (linear regression in red, Pearson correlation: 0.73). (Right) Number of rare variants versus mean genetic distance, focusing on samples with $F_{WS} \leq 0.5$ (Pearson correlation: 0.99).

with $F_{WS} \leq 0.5$ (i.e. very mixed samples). These results did not fit our expectations. Since the vast majority of variants were rare (below 5% frequency), we anticipated that mixed infections would cause just a subtle reduction in genetic distances, as more mixed samples tend to aggregate more diverse genetic types and a fraction of this diversity would be lost when using majority calling.

Given the very strong negative correlation between the number of mixed calls and F_{WS} (see Figure 3.10), we postulated that the majority calling procedure was reducing diversity (and hence, genetic distances) by effectively removing rare variants from mixed calls. As the number of within-sample strains and mixed calls grows (i.e. F_{WS} decreases), a higher number of low-frequency variants are replaced by the most common allele (population-wise) when using majority calling. Because the probability of a strain harboring a specific allele is proportional to its frequency at the population level, in mixed samples rare alleles tend to accumulate a minority of supporting reads (i.e. usually only one of the within-sample strains will carry the rare allele whereas all the others will carry the alternative). We explored this hypothesis by counting the number of rare variants (MAF ≤ 0.05) present in each sample and measuring its correlation with F_{WS}.

In agreement with our reasoning, we observed a very quick decay in the number of rare alleles for very mixed samples ($F_{WS} \leq 0.5$). We also found a strong positive correlation

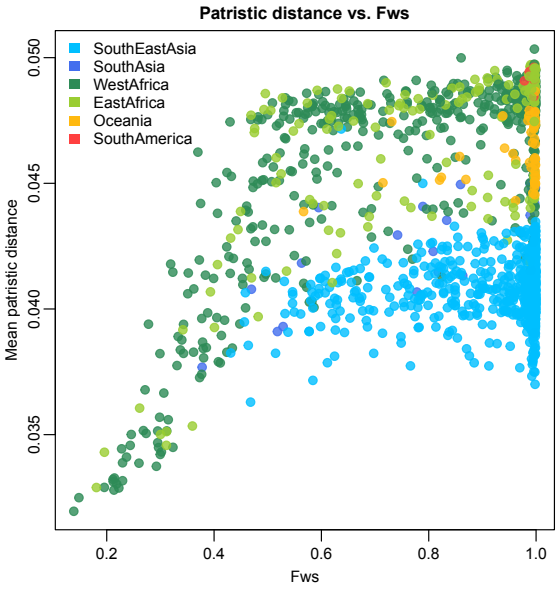

Figure 3.4: Plot showing the average patristic distance (from the neighbor joining tree of Figure 3.1) versus F_{WS}. Once again, we observe the same pattern, with a fast decay in distances for very mixed samples (F_{WS} below 0.5). The samples with low mean patristic distance (bottom left) correspond to the branches that are substantially shorter in Figure 3.1.

between the number of rare variants and both F_{WS} and mean genetic distances (Figure 3.3). The shorter branches observed within the African clades in Figure 3.1 correspond to these mixed samples, as shown in Figure 3.4, where we plot mean patristic distances[2] in the neighbor joining tree against F_{WS} values.

For estimating genetic distances, we propose a simple estimator based on the number of reads supporting each allele that does not require the use of majority calling. The genetic distance between two samples i and j can be estimated as

$$d_{ij} = \frac{\sum_{l=1}^{L} p_{il}(1-p_{jl}) + p_{jl}(1-p_{il})}{L}, \tag{3.1}$$

where p_{il} is the within-sample non-reference allele frequency for sample i at locus l, p_{jl} is the within-sample non-reference allele frequency for sample j at locus l and L is the total number of SNPs where samples i and j could be both genotyped.This can be interpreted as the probability of the genotypes at locus l being different in samples i and j when randomly sampling from their reads covering that position. Within-sample non-reference allele frequencies are computed as

$$p_{il} = \frac{a_{il}}{r_{il} + a_{il}}, \tag{3.2}$$

[2]The average branch distance between a tip (sample) and all other tips in the tree.

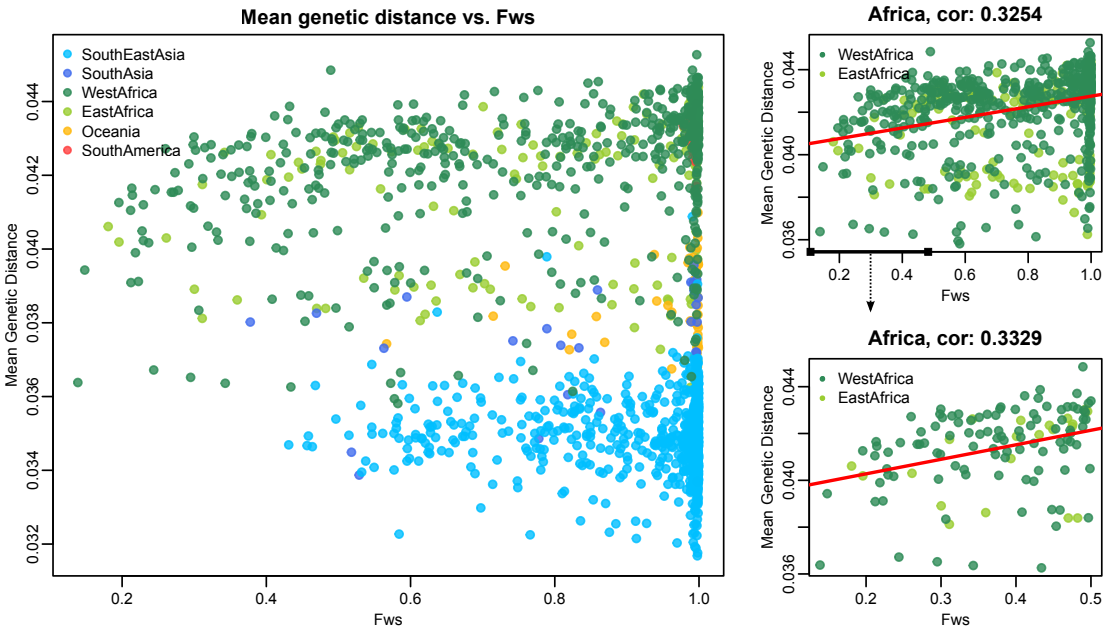

Figure 3.5: (Left) Scatterplot showing the same analysis as in Figure 3.2, mean genetic distance versus F_{WS}, but this time using corrected distances. It is evident from the figure that the strong positive correlation between F_{WS} and genetic distances has disappeared. (Right, top) Linear regression (red line) and Pearson correlation (0.32) between mean genetic distance and F_{WS} for African samples. (Right, bottom) Same analysis as in the previous panel (Pearson correlation: 0.33) for African samples with $F_{WS} \leq 0.5$.

where a_{il} and r_{il} are the number of reads for the alternative and reference alleles, respectively, observed in sample i at locus l. Figure 3.5 shows the relationship between the corrected distances and F_{WS}. It demonstrates that the downward trend in genetic distances for very mixed samples has disappeared. The correlation between distances and F_{WS} for this type of samples (in Africa) has been reduced from 0.86 to 0.33. When only African samples were taken into account, we found a similar correlation (0.32). Figure 3.6 compares the NJ trees built from raw and corrected distances. When using corrected distances, the set of very short African branches revert to the average branch length of the clade.

3.4 F_{WS} as a relative measure of inbreeding and CoI

Miotto and Kwiatkowski developed the F_{WS} statistic [Manske et al., 2012], [Auburn et al., 2012] to measure the degree of inbreeding of a mixed infection relative to that of the local population. Inspired by Wright's inbreeding coefficient (F_{IS}) [Wright, 1949], their insight was to treat a mixed sample as a subpopulation of parasites. Comparing the expected

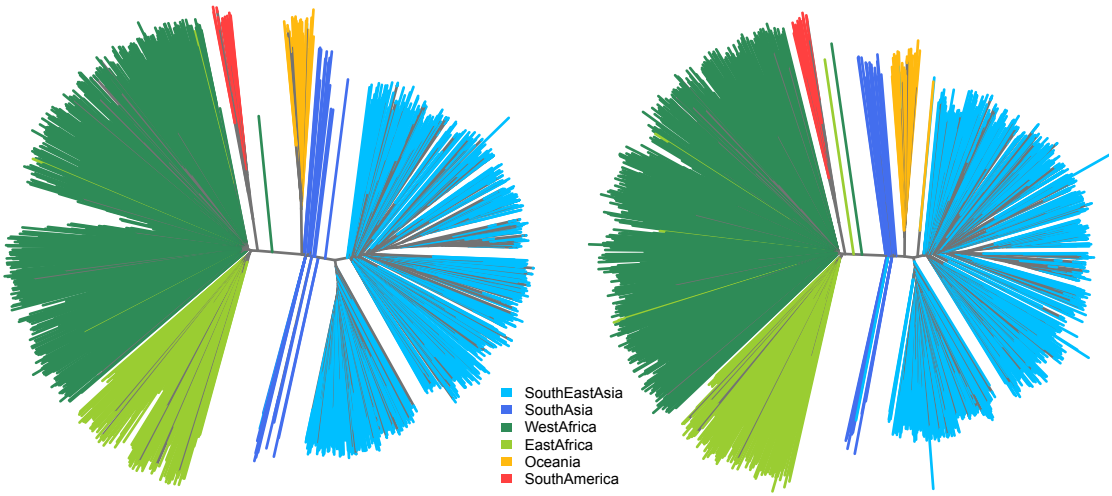

Figure 3.6: Comparison of neighbor joining trees built from raw (left) and corrected (right) genetic distances. It can be seen how the very short branches in Africa (very mixed samples for which rare and very rare variants were being discarded) have reverted to the average branch length of their clades.

heterozygosity within a mixed sample with the expected heterozygosity at the population level, they were able to estimate how closely the mixed infection represented, in average, the genetic diversity observed in the local population. They define the F_{WS} statistic as

$$F_{WS} = 1 - \frac{H_W}{H_S}, \tag{3.3}$$

where H_W denotes the expected within-sample heterozygosity and H_S refers to the expected heterozygosity in the population. Their estimator of F_{WS} is a bit involved[3], so we present it here as a sequence of steps. They start by estimating population-level allele frequencies (PLAF) from the total fraction of reads supporting each allele

$$\text{PLAF}_l = \frac{A_l}{(R_l + A_l)}, \tag{3.4}$$

where A_l and R_l refer, respectively, to the total number of reads supporting the alternative (i.e. non-reference) and reference alleles in the population at locus l. Next, they bin SNPs into ten equally sized buckets according to their minor allele frequency, $\text{MAF}_l = \min(\text{PLAF}_l, 1-\text{PLAF}_l)$, and compute the average expected heterozygosity at the population

[3]Later on we address the reasons behind this complex formulation.

level ($\bar{H}_S$) within each bin following

$$\bar{H}_S = \frac{1}{L_b} \sum_{l}^{L_b} 2\,\mathrm{PLAF}_l\,(1 - \mathrm{PLAF}_l), \tag{3.5}$$

where L_b is the total number of SNPs in bin b. They repeat the procedure for the sample of interest (whose index is denoted here by s), this time computing the within-sample allele frequency (WSAF) using the observed number of reads

$$\mathrm{WSAF}_l = \frac{a_{sl}}{(r_{sl} + a_{sl})}, \tag{3.6}$$

where a_{sl} and r_{sl} refer, respectively, to the number of reads supporting the alternative and reference alleles in sample s at locus l. Afterwards, they bin each SNP according to the frequency determined at the population level and proceed to compute the within-sample average expected heterozygosity ($\bar{H}_W$) for each bin

$$\bar{H}_W = \frac{1}{L_b} \sum_{l}^{L_b} 2\,\mathrm{WSAF}_l\,(1 - \mathrm{WSAF}_l). \tag{3.7}$$

Finally, they fit a linear regression between ($\bar{H}_W$) and ($\bar{H}_S$), constraining the intercept to be 0. The slope of the fitted line (i.e. the β coefficient) represents the linear relationship between sample and population average expected heterozygosity. F_{WS} is then computed as

$$F_{WS} = 1 - \beta. \tag{3.8}$$

Therefore, the F_{WS} statistic takes a value of one for samples that are strictly clonal and approaches zero when the mixed sample mimics the level of heterozygosity seen at the population level. This statistic is significantly correlated with COI estimates derived from PCR-based genotyping approaches[4] [Auburn et al., 2012].

The estimator presented above uses linear regression as a way of limiting the effect that sampling variants can have on the statistic under a scenario of uneaven coverage. We will see in this chapter that there exists an alternative estimator for F_{WS} that is better behaved and does not require so much complexity.

[4]For instance by genotyping highly diversed loci such as *msp-1* and *msp-2*.

3.4.1 Sensitivity of the PLAF estimator to coverage fluctuations

As indicated in the previous section, the population allele frequency at a locus is estimated by computing the fraction of total alternative read counts (observed for all samples); Equation (3.4). Here we argue that this estimator is very sensitive to variations in coverage across samples. As an illustration, imagine a study with 10 clonal samples (i.e. free of mixed infections) in which 9 of them carry the alternative allele and the remaining sample carries the reference allele for a given locus. Because the samples are clonal, we know that a reasonable point estimate for the PLAF would be 0.9. Consider now using the total read counts estimator. If all samples followed the same coverage profile[5], the estimator would produce a point estimate close[6] to our previous guess (0.9). Conversely, imagine that half of the samples have a different coverage profile; for instance, because of an improved sequencing protocol. We can make the example concrete by setting the observed number of reads to 10 in half of the samples and 10 times this quantity (100) in the other half. In this scenario, there are 10 possible allele configurations (i.e. which sample carries the reference allele) and $\binom{10}{5}$ possible sample partitions, but there are only two possible PLAF values according to the total read counts estimator: 0.98 and 0.81. These estimates deviate almost 10% from 0.9, solely due to differences in coverage across samples. The toy example embodies the potential difficulties of using the PLAF estimator in data with heterogeneous coverage profiles. We have observed this problem when working with samples sequenced at different points in time or coming from unrelated studies. To address this shortcoming, we recommend using a different estimator for the PLAF, given by

$$\mathrm{PLAF}_l = \frac{1}{n}\sum_{s}^{n}\frac{a_{sl}}{r_{sl}+a_{sl}}, \tag{3.9}$$

where a_{sl} and r_{sl} are, correspondingly, the number of reads observed for the alternative and reference alleles in sample s at locus l. This estimator is just the mean population WSAF, which tends to be better behaved in a situation of coverage heterogeneity. Since our concern here is the assessment of mixed infections, we studied how both estimators

[5]With coverage profile we refer informally to the distribution of read counts for the reference and alternative alleles in a sample.

[6]The differences observed in read counts due to the spread of the underlying coverage distribution is expected to be self-averaging if all samples follow the same coverage distribution, being of little concern with large sample sizes.

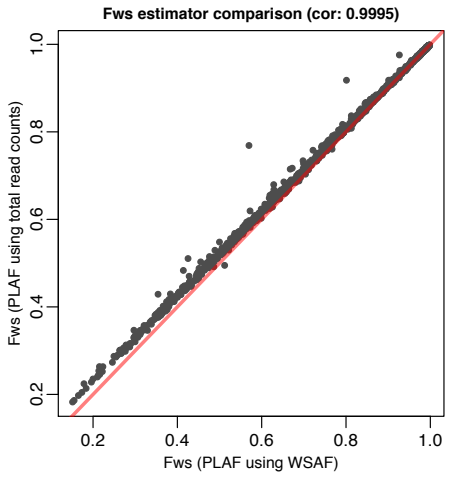

Figure 3.7: Comparison of F_{WS} estimates for 2,500 samples from the Pf3k project when using both PLAF estimators. Differences between the two are marginal, with a Pearson correlation close to 1, although the use of the total read counts PLAF estimator tends to produce higher F_{WS} estimates, characterizing samples as slightly more clonal.

influenced the original F_{WS} statistic by analyzing 2,500 samples from the Pf3k project [Pf3k, 2016]. Figure 3.7 reveals that the use of the total read counts PLAF estimator tends to produce higher F_{WS} estimates, characterizing samples as slightly more clonal than when using our mean WSAF estimator. Although a few outliers can be seen over the diagonal, caused by samples with very different coverage profiles, the correlation between F_{WS} estimates was outstanding (0.99). We ascribe the little influence of coverage fluctuations to the smoothing introduced by the binning step used in the F_{WS} estimator.

3.4.2 Incorporating uncertainty in the F_{WS} estimator

One of the assumptions of the F_{WS} estimator given by Miotto and Kwiatkowski is that within-sample allele frequencies can be estimated accurately from read counts. However, read data is still far from perfect and can be influenced by sequencing errors, alignment artifacts and coverage fluctuations along regions difficult to map to the reference genome. To minimize the effects of low-quality data, Miotto and Kwiatkowski rely on a stringent filtering process [Manske et al., 2012]. Here we study how robust is the F_{WS} statistic when used on low-coverage data. To simplify exposition, we assume that the result of the sequencing process for a locus is analogous to a binomial sampling procedure and that read counts for the alternative allele (a_l) are binomially distributed with probability π_l, i.e. $a_l \sim$ Binom(π_l, t_l), where π_l is the real within-sample allele frequency for the alternative allele at locus l, and t_l is the total number of reads observed at that locus. In a scenario of

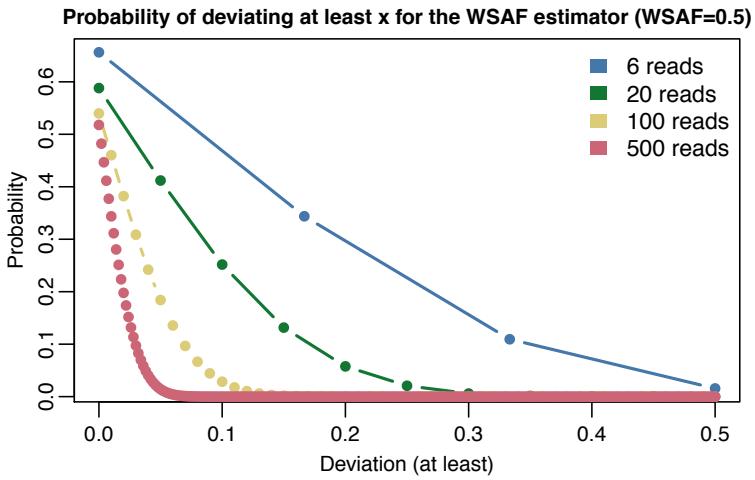

Figure 3.8: Probability of deviating (at least) a given amount in the estimation of the within-sample allele frequency for different total coverage values, assuming the real WSAF is 0.5. The figure shows how uncertainty in the estimator (fraction of reads for the alternative allele) decays quickly as the total coverage increases. In low-coverage situations (i.e. less than 20 reads), the uncertainty in the estimator is not negligible.

low coverage, there would be a lot of uncertainty in the estimation of $\hat{\pi}_l = a_l/t_l$, Equation (3.6). To appreciate this, imagine a locus in which a total of 6 reads are observed and whose underlying alternative allele frequency is 0.5. In this example, the probability of estimating $\hat{\pi}_l$ perfectly is equivalent to the probability of sampling 3 alternative reads, $P(a_l = 3 \mid \pi_l = 0.5, t_l = 6)$, which is only 0.31. Figure 3.8 compares the probability of deviating (at least) a given amount in the estimation of the within-sample allele frequency (i.e. $|\pi_l - \hat{\pi}_l|$) for different coverage values.

In order to incorporate the uncertainty associated with the estimation of WSAF frequencies into the F_{WS} statistic, we propose a sampling scheme akin to bootstrapping[7] [Efron, 1992]. That is, we sample alternative read counts from Binom$(\hat{\pi}_l, t_l)$ and re-compute $\hat{\pi}_l$ from the sampled number of alternative reads. We repeatedly perform this procedure along all loci and compute F_{WS} using the estimator introduced in Section 3.4 (but using the PLAF estimator described in Equation (3.9) to avoid artifacts due to fluctuations in coverage across samples). We use the mean of the bootstrapped F_{WS} values as the final estimate and rely on the empirical quantiles of this distribution to construct confidence intervals.

[7]We refer informally to this estimator as the F_{WS} *bootstrapping estimator* although F_{WS} *Monte Carlo estimator* might have been a more adequate name.

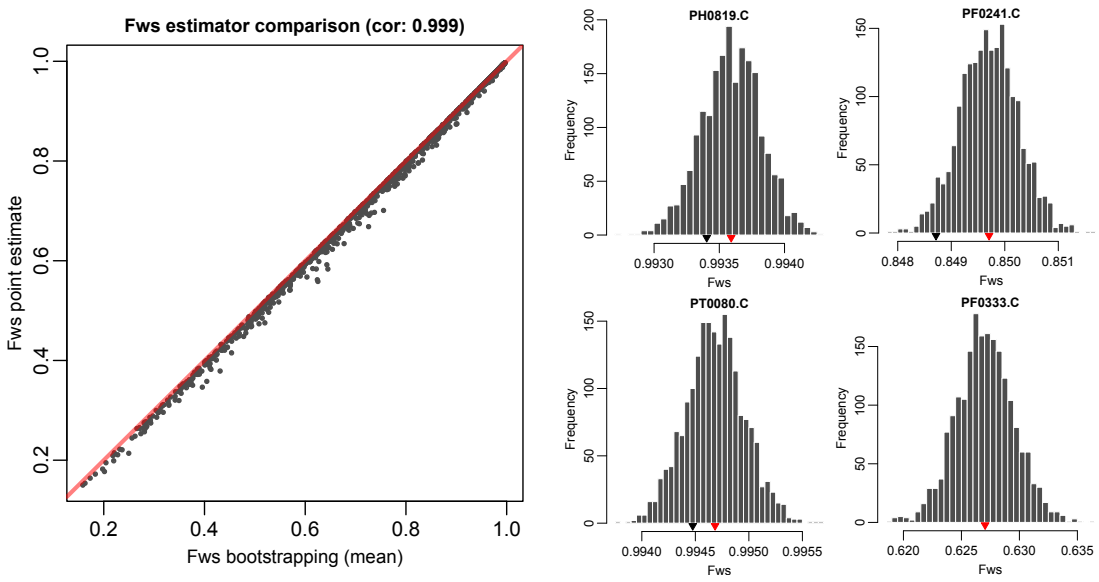

Figure 3.9: (Left) Comparison of point and bootstrapping (mean) F_{WS} estimates for 2,500 samples from the Pf3k project. There are only marginal differences between the two (Pearson correlation close to 1) although the bootstrapping estimates tend to characterize samples as more clonal. (Right) Distribution of bootstrapped F_{WS} estimates for four different samples. The red triangle indicates the mean of the distribution whereas the black triangle refers to the original F_{WS} point estimate. In the last histogram (bottom-right), the mean of the distribution and the point estimate are the same.

We emphasize that this method is only advantageous when dealing with low-coverage data since in the limit, as the total coverage of mixed calls increases, the original and (mean) bootstrapped F_{WS} estimates converge and become indistinguishable from each other. We tested our bootstrapped F_{WS} estimator with the Pf3k samples and found our results to be very close to the original[8] point estimates, with very narrow confidence intervals (95%). This can be explained by the quality and deep coverage of the data but also by the binning step used by the F_{WS} estimator (as it lessens the influence of low-quality sites). Despite both estimates being only marginally different, we found many cases in which the original point estimate was not included in the bootstrapping confidence interval. As Figure 3.9 shows, the (mean) bootstrapping estimator tends to characterize samples as slightly more clonal (higher F_{WS}), which is justified by the original estimator being overconfident when faced with low read counts in mixed calls.

[8]We also use the PLAF estimator introduced by Equation (3.9) in the original F_{WS} formulation in order to make the two estimators comparable.

3.4.3 The original F_{WS} estimator is biased

A multiple infection can be completely characterized by three features: the number of strains or different genetic types present in the sample, the abundance or relative proportions of these strains and their associated haplotypes. In contrast, F_{WS} compares the expected level of inbreeding in the sample mixture with that of the local population but provides no direct information about genetic diversity, the number of strains or their relative abundance. In fact, F_{WS} is mainly driven by the number of mixed calls observed in a sample. Figure 3.10 (top-left) shows a very strong negative correlation (-0.96) between the number of mixed calls observed and the value of F_{WS}. The within-sample frequencies in these heterozygous sites modulate the value of the statistic, causing the spread in F_{WS} estimates for a given number of mixed calls. Intuitively, we would expect a sample with a F_{WS} value close to zero to approximate well the average heterozygosity seen at the population level and, consequently, to present as mixed calls a substantial fraction of the sites that are polymorphic in the population[9].

In light of these observations, we expected to see a pattern similar to that of Figure 3.10 (top-left) when plotting the relationship between F_{WS} and the fraction of polymorphic sites (in the population) that appear as mixed calls in a sample. However, we observed a systematic, fast decay of F_{WS} before even reaching 20% of polymorphic sites in most samples (Figure 3.10, top-right). The same pattern persisted when excluding singletons from the population variants (Figure 3.10, bottom-left). Only when considering common variants (MAF ≥ 0.05), we uncovered the pattern of F_{WS} decay (Figure 3.10, bottom-right) that met our linear expectation.

For completeness, and to get a better feeling for the ability of the original F_{WS} estimator to capture the relationship between the average expected heterozygosity seen within a mixed infection and that of the local population, we computed the root-mean-squared-error between both heterozygosity measures $\left(\text{RMSE}_H = \frac{\sqrt{\sum_l (H_W - Hs)^2}}{L}\right)$. Intuitively, when $\text{RMSE}_H \simeq 0$, the F_{WS} statistic should also be close to zero (as the sample captures almost perfectly the average expected heterozygosity of the local population). For the opposite case,

[9]This is just a rough guide since, as mentioned before, F_{WS} is regulated by the distribution of within-sample heterozygosity, and that could compensate for a multiple infection not representing a fraction of variant sites.

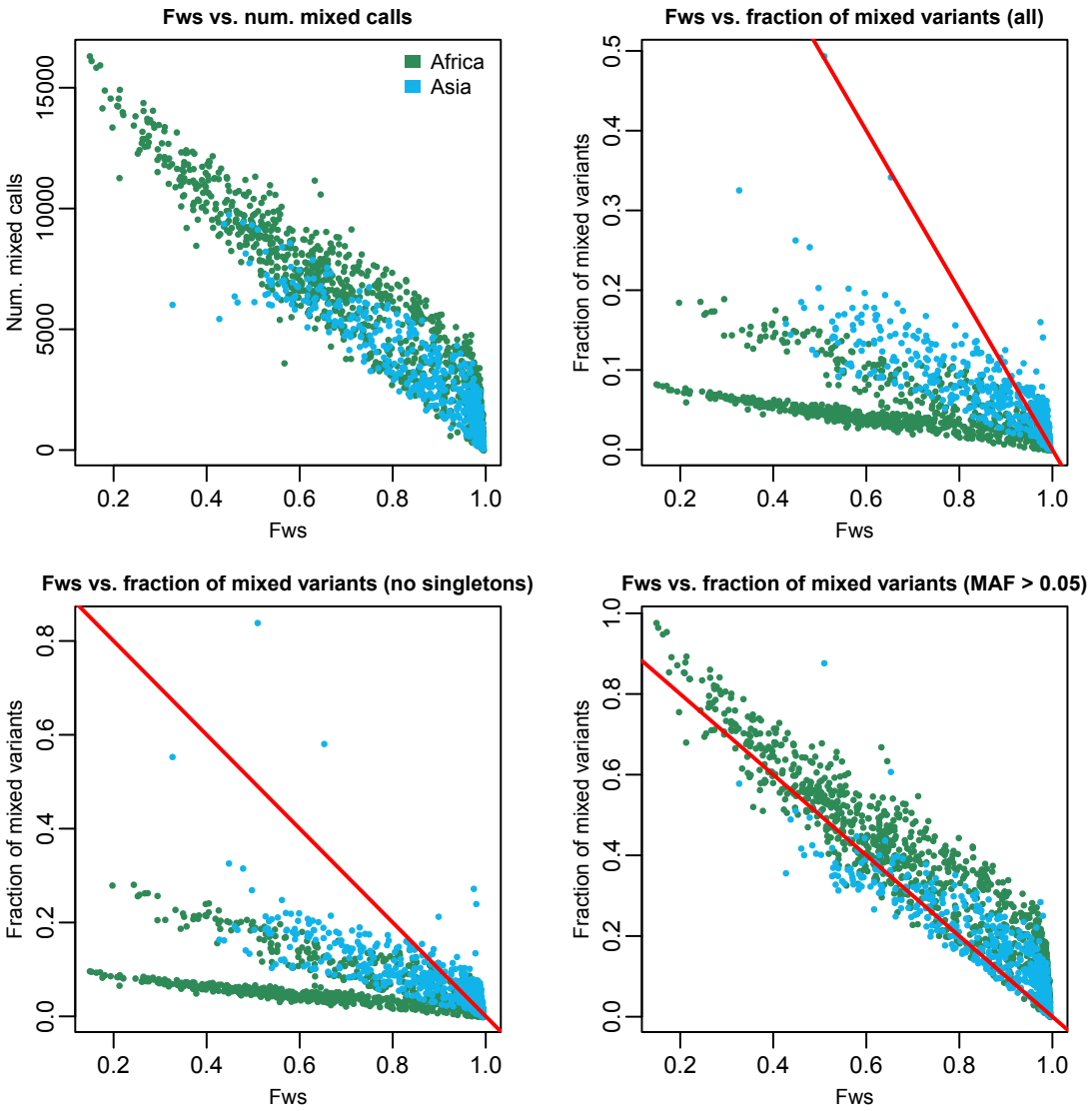

Figure 3.10: Relationship between F_{WS} and number of mixed calls in 2500 samples from the Pf3k project. (Top-left) F_{WS} vs. number of mixed calls observed in each sample. (Top-right) F_{WS} vs. fraction of polymorphic sites in the population that appear as mixed calls in each sample. F_{WS} decays to very low levels before even reaching 20% of polymorphic sites (i.e. fraction of mixed variants). The two clouds of points primarily distinguish samples from different continents. (Bottom-left) Same analysis as in the previous panel but excluding singletons, still the same fast decaying trend is evident. (Bottom-right) Same analysis for common variants at population level (MAF ≥ 0.05, population-wise). In this case, we recover the pattern seen in the top-left panel, with very mixed samples (low F_{WS}) corresponding to those in which the majority of common variants appear as mixed calls. When shown, the red line represents the identity line.

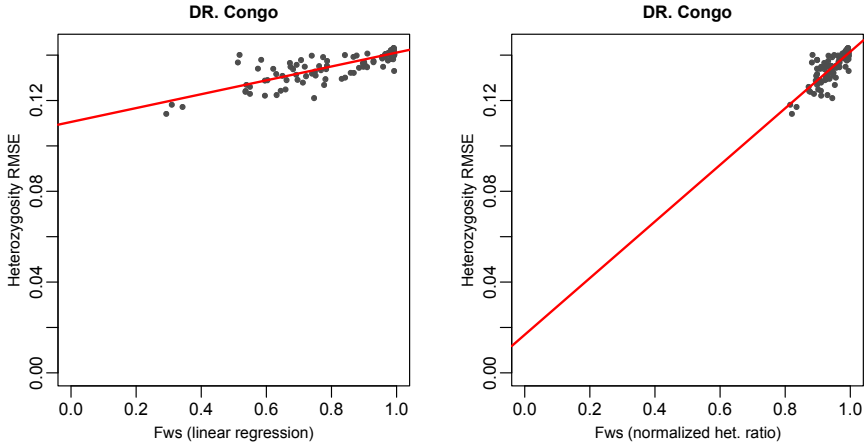

Figure 3.11: (Left) Relationship between the original F_{WS} estimator and the RMSE of heterozygosity differences ($H_W - H_s$) for samples from Congo. The red line corresponds to the fit of a linear regression. (Right) Same analysis but using the alternative F_{WS} estimator (Section 3.4.4).

when the RMSE_H value is close to the average expected heterozygosity of the population[10], we expect $F_{WS} \simeq 1$. However, we found that the original F_{WS} estimator produced very low F_{WS} estimates for samples that had large RMSE_H values, indicating that the estimator is indeed biased and does not take into account a substantial amount of variation present at the population level (i.e. it underestimates F_{WS} when samples are not clonal). Figure 3.11 shows the relationship between RMSE_H and F_{WS} for samples of the Democratic Republic of Congo. The left panel uses the original F_{WS} estimator whereas the right panel assesses this relationship for an alternative estimator that we introduce in Section 3.4.4. We have fitted a linear regression to both distributions to appraise the projected RMSE_H value when F_{WS} reaches zero.

These analyses indicate that the original F_{WS} estimator has no sensitivity for rare variants, which are predominant in many populations. We attribute this bias to the way in which the estimator is constructed, in particular to the imposition of the regression intercept being zero and the use of a frequency-binning step. Hence, we conclude that the original F_{WS} estimator is able to identify clonal or near clonal samples but greatly overestimates the degree in which a mixed sample mimics the level of heterozygosity or inbreeding seen at the population level.

[10]Each population is a characterized by a different value that is contingent on genetic diversity and population structure.

3.4.4 An alternative F_{WS} estimator

Given the intrinsic bias detected in the original F_{WS} estimator, here we develop an alternative that does not have such a bias, has a better resolution and a simpler formulation. We remind the reader the intuition behind the F_{WS} statistic: treating a mixed infection as a subpopulation of parasites and relating the expected heterozygosity of this set to that of the local population. Following this reasoning, we define F_{WS} as the coefficient[11] that scales the total expected heterozygosity within-sample accumulated for L loci, (H_W^L), to match that of the local population, (H_S^L). Therefore

$$\begin{aligned} F_{WS} &= 1 - \frac{H_W^L}{H_S^L} \\ &= 1 - \frac{\sum_l^L 2\,\mathrm{WSAF}_l\,(1-\mathrm{WSAF}_l)}{\sum_l^L 2\,\mathrm{PLAF}_l\,(1-\mathrm{PLAF}_l)}, \end{aligned} \tag{3.10}$$

where WSAF_l is given by Equation 3.6 and PLAF_l follows Equation 3.9. For a single locus l, $\left(1 - \frac{H_W^l}{H_S^l}\right) \in (-\infty, 1]$ and the positive part of the range has the same interpretation as the original F_{WS} statistic. If l is not a mixed call in the sample we are considering, $H_W^l = 0$ and the expression takes a value of 1, indicating the sample possesses a homozygous call and cannot explain any of the diversity observed at the population level. If conversely l is a mixed call and $H_W^l = H_S^l$, the expression takes a value of 0, signifying that the mixed call perfectly explains the population expected heterozygosity at that locus. Finally, if $H_W^l > H_S^l$, the expression takes a negative value, denoting that the mixed call shows a higher degree of heterozygosity than that of the population (e.g. imagine a singleton that is only observed in a mixed call with a within-sample frequency of 50%). The genome-wide statistic is simply the ratio of the total expected heterozygosity at sample and population level, we will refer to this formulation as the F_{WS} heterozygosity ratio estimator. The main disadvantage of this alternative formulation is its unbounded range (i.e. F_{WS} values can take an arbitrary large negative value when samples harbour more diversity than the local population). Nonetheless, it is easy to modify the new estimator to provide a normalized

[11] We actually follow the convention of using $1-F_{WS}$, and the nomenclature (H_W,H_S) adopted in the original formulation.

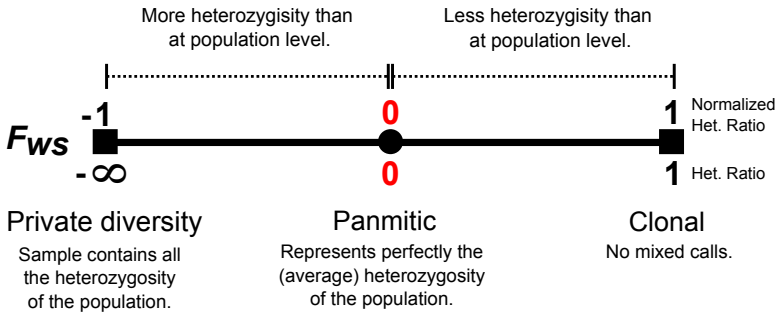

Figure 3.12: Diagram that illustrates how the interpret the range of values of the alternative F_{WS} estimators. Field samples map on the positive interval of the range while the negative part accommodates special cases such as artificial mixtures.

F_{WS} value that lies in the $[-1, 1]$ range. We propose

$$\begin{aligned} F_{WS} &= \frac{H_S^L - H_W^L}{H_S^L + H_W^L} \\ &= \frac{\sum_l^L 2\,\mathrm{WSAF}_l\,(1-\mathrm{WSAF}_l) - \sum_l^L 2\,\mathrm{PLAF}_l\,(1-\mathrm{PLAF}_l)}{\sum_l^L 2\,\mathrm{WSAF}_l\,(1-\mathrm{WSAF}_l) + \sum_l^L 2\,\mathrm{PLAF}_l\,(1-\mathrm{PLAF}_l)}, \end{aligned} \tag{3.11}$$

where, again, WSAF_l is given by Equation 3.6 and PLAF_l follows Equation 3.9. This estimator has a more intuitive interpretation, F_{WS} takes a minimum value of -1 when all the heterozygosity of the population is contained in the sample being assessed. We refer to this alternative formulation as the normalized F_{WS} heterozygosity ratio estimator.

Figure 3.12 summarizes how to interpret the statistic when using these different formulations. Natural infections map to the right half of the range $[0, 1]$ whereas special cases (e.g. artificially constructed mixtures) map to the left half $[-1, 0)$ or $(-\infty, 0)$. We termed the mid-point of the F_{WS} range (0) the panmitic limit since it represents the value of the statistic that characterizes samples with the same level of total expected heterozygosity observed in the population

We prefer the more intuitive interpretation of the normalized F_{WS} estimator. However, as natural infections map on the positive side of the F_{WS} range, what formulation to use may be considered a matter of taste. In Figure 3.13 we compare how they relate to each other. The plots show clearly that the normalized versiod produce lower F_{WS} estimates for values that are not close to the extremes. Figure 3.14 depicts the surface of the normalized F_{WS} estimator.

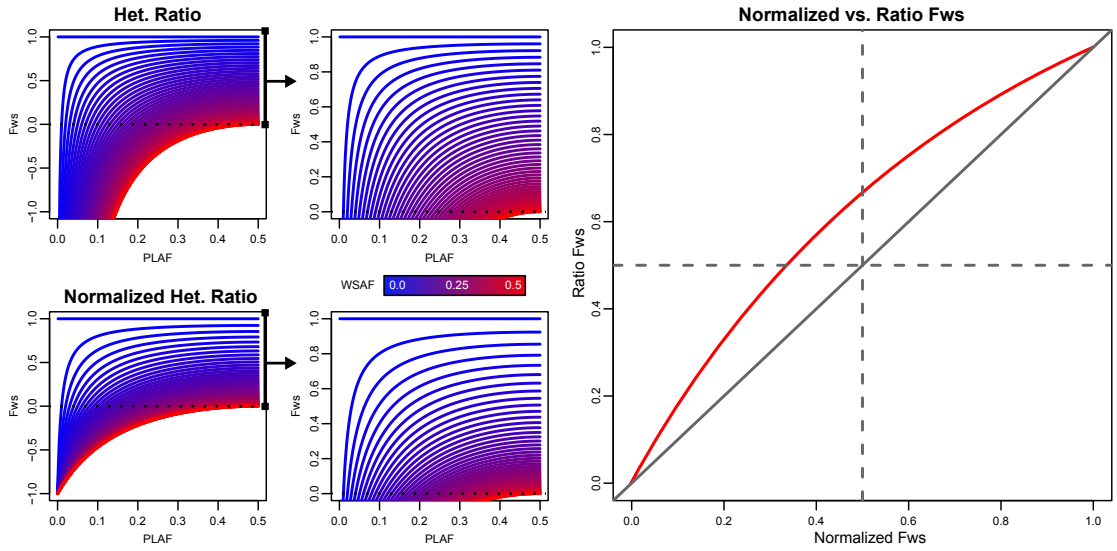

Figure 3.13: Plots comparing the alternative F_{WS} formulations. (Left) Slices of the Fws surfaces for the heterozygosity ratio and the normalized heterozygosity ratio estimators. PLAF is specified on the x-axis whereas WSAF is enconded using a color gradient. Notice that the het. ratio estimator is unbounded and produce F_{WS} values less than -1 when WSAF $\gg$ PLAF (not shown). (Right) Comparison of both estimators within the $[0, 1]$ range. The continuous gray line represents the identity line.

3.4.5 Populations have characteristic F_{WS} decay profiles

Following the intuition behind the F_{WS} statistic, we can interpret a multiple infection as a random set of parasites sampled from the local population. As the size of this set grows (i.e. the mixed infection captures more diversity), its associated F_{WS} value converges to the panmitic limit ($F_{WS} \to 0$). The F_{WS} decay curve defined in this fashion is contingent on the genetic structure of the samples[12] and, consequently, is different for each population. Figure 3.15 illustrates this concept for three of the countries present in the Pf3K dataset, using the normalized F_{WS} heterozygosity ratio estimator (Equation 3.11).

Because of this population-level dependency, one must be careful when comparing individual F_{WS} estimates from different groups or when drawing conclusions regarding the multiplicity of infection of a set of samples. If the populations involved have a very distinct genetic structure, individual F_{WS} values might not be directly comparable without contextualizing how the populations compare to each other. For instance, a multiple infection in a very diverse population might present a higher F_{WS} (i.e. more clonal) than a mixture in an inbred population, even if the former contains more strains than the latter.

[12] We note that it can also be affected by differences in sequencing protocols.

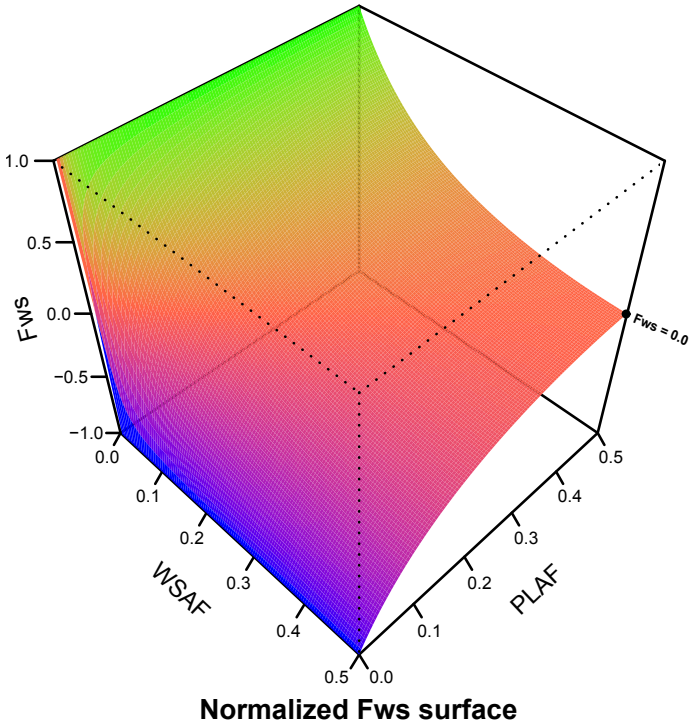

Figure 3.14: Surface for the normalized F_{WS} heterozygosity ratio estimator. The color gradient indicates the F_{WS} value (z-axis) for a particular combination of WSAF/PLAF, ranging from blue ($F_{WS} \simeq -1$) to green ($F_{WS} \simeq 1$). Shades of red indicate the panmitic limit (i.e. when F_{WS} takes a value close to 0). We only show F_{WS} for WSAF/PLAF values in the range $[0, 0.5]$ since the distribution is symmetrical for the other half of the range ($[0.5, 1]$).

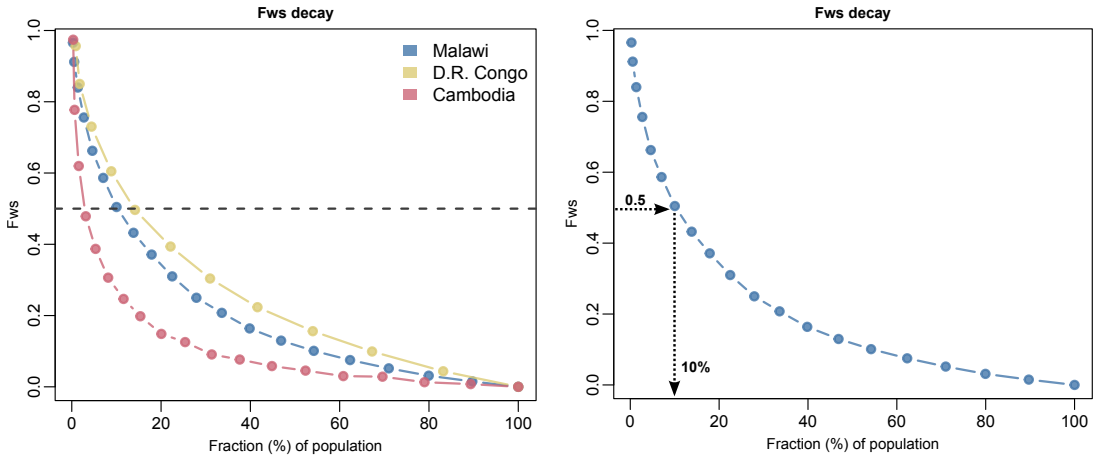

Figure 3.15: (Left) F_{WS} decay curves for Malawi ($n = 369$), Cambodia ($n = 570$) and the Democratic Republic of Congo ($n = 113$). Each point represents the average F_{WS} value when creating artificial mixtures by randomly selecting a subset of samples of size k, without replacement, from the population (100 replicates). The first point represents the average sample F_{WS} in the population ($k = 1$). The x-axis has been normalized, dividing k by the number of samples in the corresponding population and represents the fraction of population samples included in the artificial mixture. F_{WS} decays at different rates due to differences in diversity and the genetic structure of the populations. The horizontal line indicates the point at which F_{WS} value decays to 0.5. All F_{WS} estimates computed using the normalized heterozygosity ratio estimator, Equation 3.11. (Right) Schematic view of how we derive the average number of samples required for F_{WS} to decay to 0.5 for the Malawian population.

We can appreciate this in Figure 3.15. In Cambodia, we reach an F_{WS} of 0.5 by aggregating (in average) only 3% of the samples in the population. However, for the Democratic Republic of Congo we need to aggregate 14% of the samples to reach the same F_{WS} value. As a consequence, individual F_{WS} comparisons across populations are bound to be misleading unless differences in population diversity are taken into account.

3.4.6 Comparison with the original F_{WS} estimator

In this section we compare the original linear regression F_{WS} estimator with the normalized version of the heterozygosity ratio estimator (Equation 3.11). We found that the majority of samples characterized as very mixed ($F_{WS} \leq 0.5$) by the original estimator in the Pf3k dataset were substantially more clonal according to our proposed alternative (Figure 3.16). Again, this difference highlights the fact that the linear regression estimator is insensitive to the population diversity encoded by low-frequency variants whereas our alternative takes all variants into account.

Our alternative F_{WS} estimates need to be interpreted with this new scale and resolution

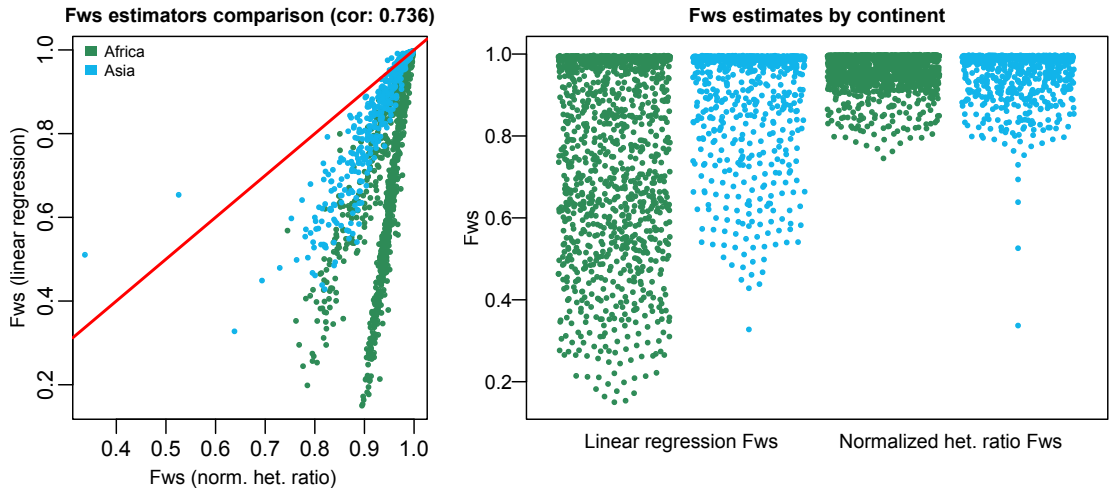

Figure 3.16: (Left) Comparison of the alternative (x-axis) and original (y-axis) F_{WS} estimators for the Pf3k samples. The clusters of points are caused by the differences in the genetic structure of the local populations, mainly distinguishable at continental level. (Right) Distribution of F_{WS} values according to both estimators. The plot shows how the original estimator inflates the F_{WS} estimate for African samples due to the rare variants bias. Median values for the original estimator: Africa, 0.92; Asia, 0.99. Median values for the alternative estimator: Africa, 0.98; Asia, 0.99.

in mind. Figure 3.17 shows the relationship between F_{WS} (normalized heterozygosity ratio estimator) and the fraction of variants in the local population that appear as mixed in a multiple infection. The auxiliary panels display the spread of the estimates for F_{WS} values close to one.

As with the original estimator, we recommend using the same sampling procedure presented in Section 3.4.2 to incorporate uncertainty from the read count data only when facing scenarios of low coverage. We performed the mentioned *bootstrapping* analysis on the Pf3k dataset and found the estimates distribution to be very narrow for most samples (95% confidence intervals within ± 0.01 units of the distribution mean).

3.4.7 F_{WS} and ascertainment bias

The F_{WS} statistic was devised to be used with whole-genome deep sequencing data. However, variability among protocols, sample quality and genome accessibility can introduce a strong ascertainment bias. In this section, we study the accuracy of our alternative estimators when the set of observed variants is a very small sample of loci, either distributed randomly along the genome or constrained to lie within a particular allele frequency range. Instead of relying on a point estimate, we make use of the non-parametric Bayesian bootstrap [Rubin et al., 1981], [Alfaro et al., 2003] to approximate posterior credible intervals for our parameter. In

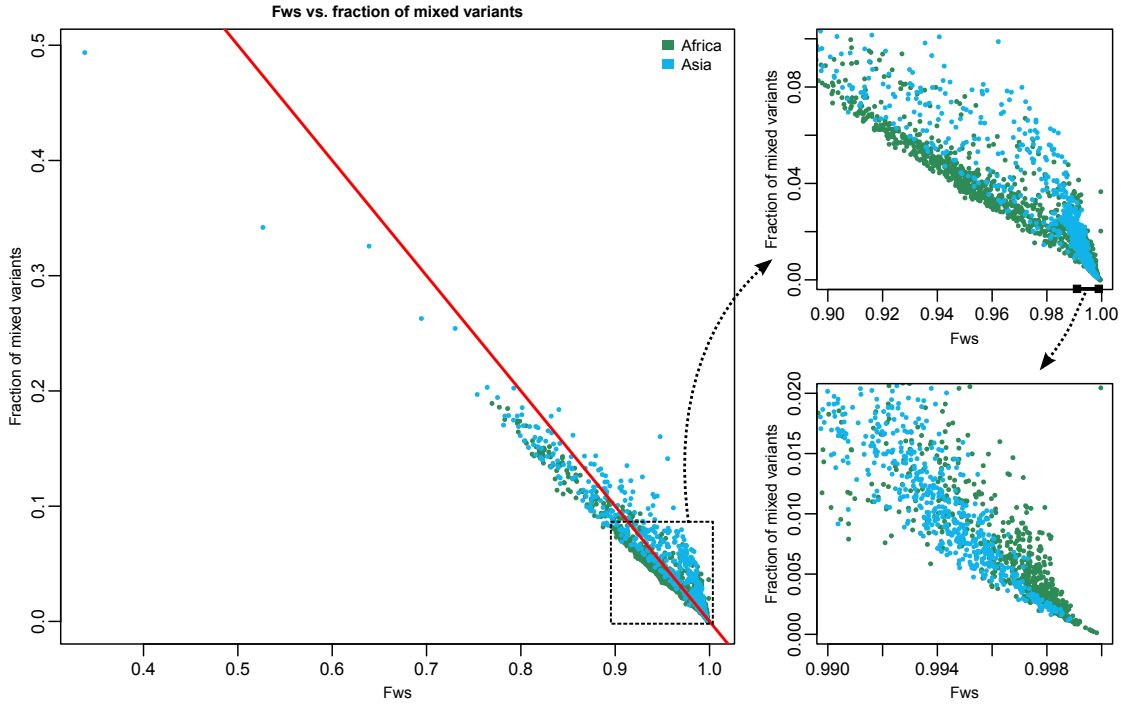

Figure 3.17: Relationship between F_{WS} (normalized heterozygosity ratio estimator) and fraction of polymorphic sites that appear mixed in a sample. (Left) There is a strong negative correlation between the two (-0.96) and most samples map to a narrow interval, between 0.85 and 1. (Right) Closer view of the intervals that contain most of the samples.

all our *in silico* experiments, we consider the F_{WS} value computed using all polymorphic sites as the underlying parameter we are trying to estimate using the reduced data set and used data from the Pf3k project [Pf3k, 2016].

In Figure 3.18 we show the results of an experiment in which we repeatedly sampled a subset of 100 SNPs[13] from a Ghanaian isolate and approximated 95% credible intervals using the Bayesian bootstrap for the underlying genome-wide F_{WS} statistic. This exploratory analysis revealed good coverage[14] ($\sim$ 95%) for both estimators despite using a dataset that is orders of magnitude smaller than the genome-wide equivalent.

We continued with this line of investigation and repeated the analysis for 200 isolates, this time using a fixed panel of 100 SNPs randomly sampled within the $[0.1, 0.25]$ MAF bin (Figure 3.19). As expected, we found that the coverage of the intervals drop substantially for both estimators (to 82% and 83%) when very clonal samples were included. This is due to the difficulty of sampling the sparse mixed calls present in these clonal samples when using only 100 SNPs (i.e. we estimated an F_{WS} of 1 whereas the genome-wide value

[13]Each SNP was sampled with a probability proportional to its MAF.
[14]Proportion of times that the interval contains the true value we are trying to estimate.

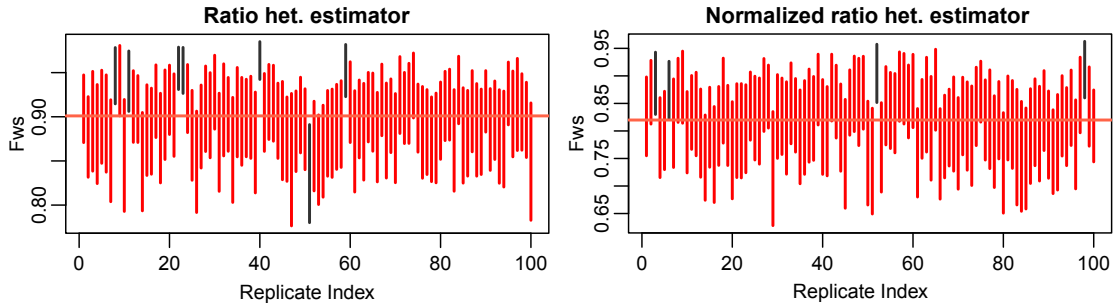

Figure 3.18: (Left) Credible intervals (95%) for the F_{WS} heterozygosity ratio estimator when using a random sample of 100 SNPs (sampling probability proportional to MAF) on a mixed sample, 100 independent replicates. The horizontal red line indicates the genome-wide value, vertical red lines designate the credible intervals that contain the genome-wide estimate. (Right) Same analysis for the normalized version of the estimator, colors have the same meaning. Credible intervales for F_{WS} in both analyses were approximated by using the Bayesian bootstrap with 500 replicates.

was effectively $1 - \epsilon$, where ϵ is a very small number due to noise or very sparse mixed calls). However, as these samples were effectively clonal, we regard these differences as non-significant. Besides, we found the average bias of both estimators to be marginal (-0.009 and -0.002) when compared to the range of plausible F_{WS} values.

Finally, we explored how constraining the SNP sampling to different MAF frequency bins affected F_{WS} estimates. In this case we sampled 100 SNPs from five equally sized MAF bins (500 replicates) and computed F_{WS}, on the same sample, using our proposed estimators. We found both estimators to be well calibrated, relative to the genome-wide estimate, for all but for the $(0, 0.1]$ MAF bin, which presented large variance. In Figure 3.20, we illustrate the results of this analysis for three samples with different F_{WS} levels (very mixed, almost clonal and mixed). In agreement with our previous observations, we found that we tend to slightly overestimate F_{WS} for samples that are almost clonal (Figure 3.20, middle panels). In light of these results, we conclude that both estimators are reasonably robust to ascertainment bias if the SNPs used are common in the population (i.e. MAF > 0.05), even when the number of SNPs used is small (e.g. 100).

3.4.8 F_{WS} and complexity of infection

In this last section, we have a look at the relationship between CoI (provided by the Pf3k project for each sample, [Pf3k, 2016]) and the F_{WS} values produced by the normalized F_{WS} heterozygosity ratio estimator. Figure 3.21 (left and middle panels) shows how the absolute

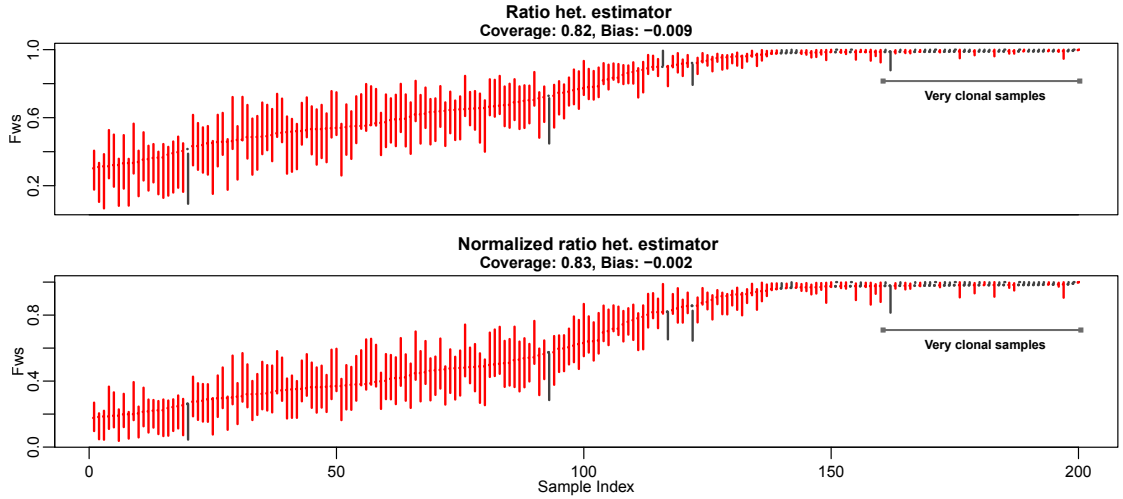

Figure 3.19: (Top) Credible intervals (95%) for the F_{WS} heterozygosity ratio estimator on 200 samples from the Pf3k project. We generated a panel of 100 SNPs by randomly sampling from the $[0.1, 0.25]$ MAF bin. Credible intervals were computed by using the Bayesian bootstrap. We have highlighted the set of very clonal samples that caused a substantial drop of interval coverage. Intervals containing the genome-wide F_{WS} estimate have been colored in red. (Bottom) Same analysis when using the normalized estimator, colors have the same meaning. Credible intervales for F_{WS} in both analyses were approximated by using the Bayesian bootstrap with 500 replicates.

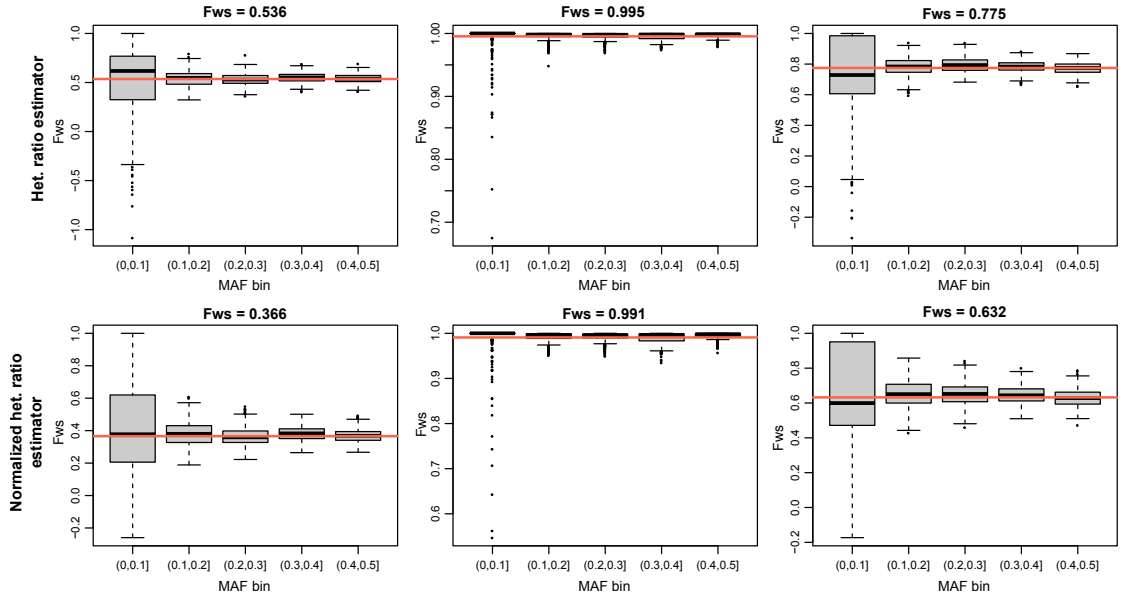

Figure 3.20: Boxplots summarizing the F_{WS} estimates computed on sets of 100 SNPs (500 replicates) sampled from different MAF bins on the same sample. The top row shows the results for the heterozygosity ratio estimator whereas the bottom row does so for its normalized version. Each column represents a sample with a different mixture level: very mixed (left), almost clonal (middle) and mixed (right). The genome-wide F_{WS} estimate is shown in top of the plots and also as a horizontal red line. We observed that both estimators were well calibrated but when the SNP set was constrained to lie within the $(0, 0.1]$ bin.

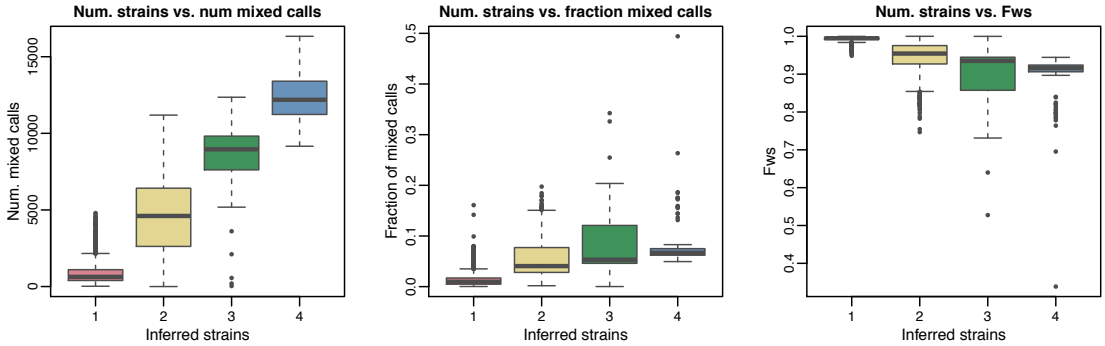

Figure 3.21: (Left) Boxplot showing absolute number of mixed calls aggregated by CoI. (Middle) Same analysis using the fraction of polymorphisms (in the local population) that appear as mixed calls. (Right) Boxplot showing F_{WS} estimates aggregated by CoI.

and relative (i.e. fraction of polymorphisms in the population) number of mixed calls relates to the inferred number of strains. A clear positive correlation between number of strains and absolute number of mixed calls is evident, but the pattern does not seem so striking when the fraction of mixed calls is considered instead. Likewise, the negative correlation between F_{WS} and number of strains (Figure 3.21, right) seems mild in comparison. As discussed earlier, this has to do with population differences in diversity and genetic structure.

Figure 3.22 displays the relationship between F_{WS} and number of inferred strains for three different populations. The first row aggregates F_{WS} estimates by number of inferred strains. The second row relates the entropy of the relative abundance of the strains present in a sample with F_{WS}. We computed information entropy following $H(p) = -\sum p_i \log_5(p_i)$ [MacKay, 2003], where p is a vector of length 5 containing the inferred strain proportions (i.e. relative abundance) and $\sum p_i = 1$.

The number of strains within a sample is derived from the vector of proportions by counting the components where $p_i \geq 0.01$. As expected, samples with a higher number of strains tend to have a lower F_{WS} value. Likewise, samples that share the same number of co-occurring strains but present more balanced proportions (i.e. higher abundance entropy) also tend to have a lower F_{WS} value.

3.4.9 Conclusions

In this chapter, we have looked into some of the biases that could be introduced by the presence of multiple infections when working with deep sequencing data. We advised against the use of the majority calling heuristic unless samples are clonal or nearly clonal, as it

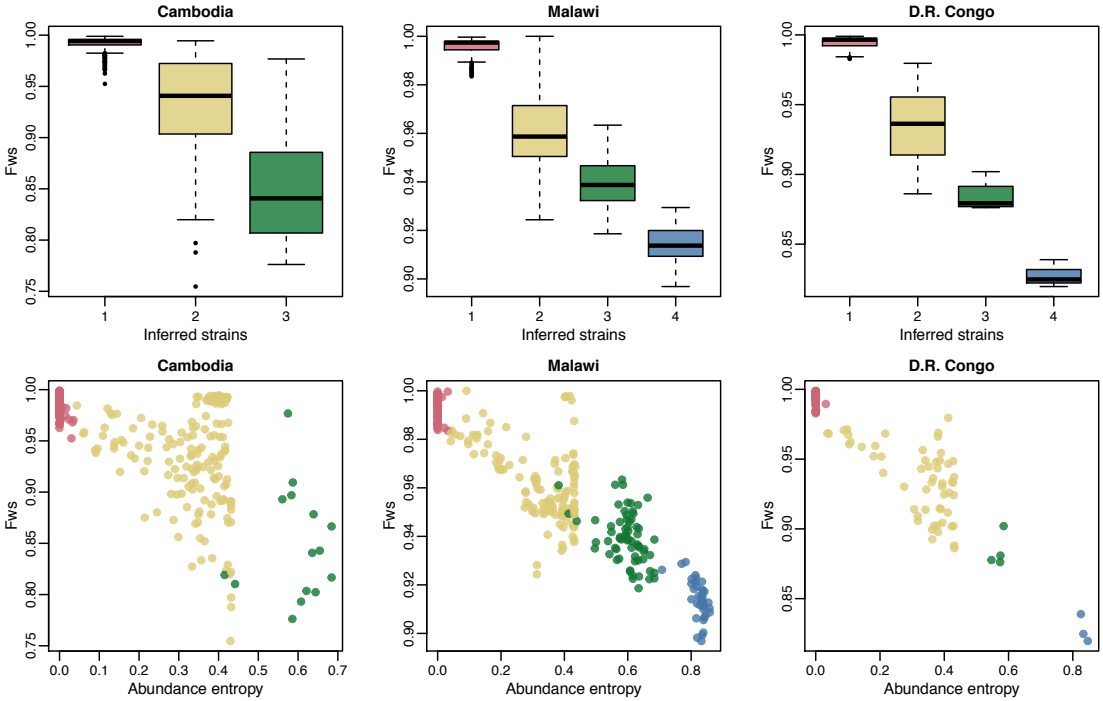

Figure 3.22: (Top row) Boxplots for F_{WS} (normalized heterozygosity ratio) estimates when aggregated by number of strains, for three different populations (Cambodia, Malawi and Democratic Republic of the Congo). (Bottom row) Plots showing the entropy of relative abundance against F_{WS}. Points are colored by the number of inferred strains (1=red, 2=orange, 3=green, 4=blue).

tends to remove rare variants. For the particular case of estimating genetic distances, we proposed a simple estimator based on the fraction of allele read counts that is robust to the presence of multiple strains. We also explored the limitations of the original F_{WS} estimator. In doing so, we proposed an alternative estimator for population-level allele frequency that is better behaved in situations of coverage heterogeneity. We also proposed a sampling mechanism whose merit is to incorporate uncertainty from read count data into the F_{WS} estimate. Our suggestion is to build confidence intervals, using this procedure, only in scenarios of low coverage.

We showed that the original F_{WS} estimator is biased and does not acknowledge the diversity encoded by rare variants, overestimating the fraction of the population expected heterozygosity that is represented within a mixed infection. To address this shortcoming, we introduced an alternative estimator (and its normalized version) for F_{WS} that does not suffer from this bias, offers a better resolution and has a simpler formulation. We studied how this estimator can be affected by ascertainment biases and found it to be reliable when the set of variants used are not rare (MAF > 0.05) in the population of origin. We also

provided further evidence that the F_{WS} statistic is contingent on the diversity and genetic structure of the local populations. Because of this, we warn against comparing individual F_{WS} estimates across populations without contextualizing population differences.

3.4.10 Individual contributions

Although I claim full authorship for the work presented in this chapter, I would like to acknowledge my interactions with other colleagues. In the first part of the chapter, I performed all analyses leading to the identification of the majority calling bias. However, I have to share the credit for building the read counts estimator for genetic distances with Olivo Miotto and Roberto Amato. Likewise, I claim full authorship of the work done regarding the study of the F_{WS} statistic, but I need to acknowledge Olivo Miotto for guiding me through the formulation of the original estimator and subsequent discussions of my results. I would also like to thank Joe Zhu for providing the inferred number of strains (CoI) and relative abundance of mixed samples from the Pf3K project.

Chapter 4

Genetic architecture and genomic epidemiology of artemisinin-resistant *Plasmodium falciparum*

Contents

4.1 Introduction

In this chapter, we study a superset of the mutations associated with resistance that were found by Ariey and colleagues [Ariey et al., 2014]. Following our previous line of research, we focus on the relationship of these mutations with population structure. We have divided the chapter into two parts. The first part studies the genetic architecture of artemisinin-resistant parasites and focuses on Southeast Asia; here we characterize the role and distribution of *kelch13* mutations and their relationship with population structure. We also discern the most likely demographic scenario (i.e. independent origins or gene flow) that explains the spread of resistance. The second part investigates the origin of artemisinin resistance mutations in Africa. The results presented here were published in [Miotto et al., 2015] and [MalariaGEN-Pf-Community-Project, 2016]. Since I am an author in both articles (fifth and fourth position in the author list, respectively), I present a brief overview of all the findings but concentrate primarily on my contributions. As usual, I detail my contributions at the end of each part (Sections 4.2.6 and 4.3.4) since these were two large collaborative studies.

4.2 Genetic architecture of artemisinin-resistant *P. falciparum*

In 2014, Ariey and colleagues[1] [Ariey et al., 2014] found multiple mutations in a gene located on chromosome 13 (PF3D7_1343700, termed *kelch13* in this work) to be associated with artemisinin resistance in *Plasmodium falciparum*. These mutations are on the β-propeller domain of a kelch-like protein. This discovery predated the execution of our study, a large genome-wide association study (GWAS) across 15 locations in Southeast Asia and Africa, whose ultimate goal was to find candidate molecular markers for artemisinin resistance. Nevertheless, we (1) confirmed the previous association results, (2) exposed a common genetic background associated with resistant populations, (3) assessed the geographical distribution of *kelch13* mutations, (4) examined the association between these mutations and population structure and, finally, (5) articulated the type of demographic and evolutionary events most likely to have triggered the rise and spread of *kelch13* mutants. The next four sections give an account of these findings. Here, we specifically focus on points 3-5 since

[1] See also Section 1.3.3.4

Contributor	Years	Country (region)	Code	Location	Samples
TRAC	2011-2013	Bangladesh	BD	Ramu	50
		Myanmar	MM	Bago division	59
		Thailand (western)	WTM	Mae Sot	103
		Thailand (southern)	STH	Ranong	20
		Thailand (eastern)	ETH	Sisakhet	21
		Vietnam	VN	Binh Phuoc	97
		Laos	LA	Attapeu	77
		Cambodia (western)	WKH	Pursat, Pailin	185
		Cambodia (northern)	NKH	Preah Vihear	106
		Cambodia (northeastern)	NEKH	Ratanakiri	95
		D.R. of the Congo	CD	Kinshasa	112
		Nigeria	NG	Ilorin	3
NIAID/NIH	2009-2010	Cambodia (western)	WKH	Pursat	89
		Cambodia (northeastern)	NEKH	Ratanakiri	46
Total					1063

Table 4.1: Geographical distribution of the samples used in the GWAS (only samples for which genotypes and phenotypes were available).

they comprise my contributions to this study and pursue the research questions introduced in this thesis. These results were published in Nature Genetics in 2015 [Miotto et al., 2015].

4.2.1 Data

This study was based on 1,612 clinical samples collected from patients infected with *falciparum* malaria in 15 different locations. After performing whole-genome sequencing and subjecting the data to a stringent quaility filtering process, we built a variation catalog of more than 600,000 SNPs. The GWAS was conducted on a subset of samples (1,063; 13 different locations) for which clinical phenotypic data was available. The Tracking Resistance to Artemisinin Collaboration (TRAC) initiative provided 928 samples from infected individuals recruited during 2011-2013. The National Institute of Allergy and Infectious Disease (US, National Institute of Health) provided 135 samples of individuals infected during 2009-2010. Table 4.1 summarizes the geographical distribution and sampling year of the samples used in the GWAS. To assist downstream analysis, we included additional samples (549) from multiple projects working in Southeast Asia. Table 4.2 discloses the origin and sampling year of this additional set of samples[2]. We note that the two major contributing studies were clinical studies whose goal was to assess the extent of artemisinin resistance in Southeast Asia.

[2]Contributed by various studies participating in the MalariaGEN *Plasmodium falciparum* Community Project (for further details, see `http://www.malariagen.net/projects/parasite/pf`).

Contributor	Years	Country (Region)	Code	Locations	Samples
MORU/SMRU	2001-08	Thailand (West)	WTH	Mae Sot	108
MORU	2007	Cambodia (West)	WKH	Pailin	3
ARC3	2008-09	Bangladesh	BD	Bandarban	29
		Cambodia (West)	WKH	Pailin	35
				Tasanh	49
		Thailand (West)	WTH	Mae Sot	3
UMD/HHMI	2009	Cambodia (West)	WKH	Pailin	15
OUCRU	2009-10	Vietnam	VN	Binh Phuoc	19
ARCE	2010-11	Laos	LA	Xepon	36
		Myanmar (South)	SMM	Kawthaung	50
		Vietnam	VN	Binh Phuoc	108
NIAID/NIH (*)	2008-10	Cambodia (West)	WKH	Pursat	67
		Cambodia (Northeast)	NEKH	Ratanakiri	4
TRAC (*)	2011-13	Myanmar (Central)	CMM	Bago Division	1
		Laos	LA	Attapeu	8
		Thailand (West)	WTH	Mae Sot	3
		Cambodia (West)	WKH	Pursat	3
Total					541

Table 4.2: Samples included in the population structure analysis but not in the GWAS. The abbreviations in the code column refer to geographical regions. Samples marked with a star (*) were not included in the GWAS because phenotypes were not available.

4.2.1.1 Phenotypes

We used parasite clearance half-life (PC $t_{1/2}$), the time taken for artesunate (an artemisinin derivative) to reduce parasite density by half, since this characterizes *P. falciparum* response to treatment. Clearance half-life was estimated from the log-linear decay in parasite density in blood samples collected every 6 hours after admission. The findings regarding the distribution of phenotypes were published in [Amaratunga et al., 2012] and [Ashley et al., 2014].

4.2.1.2 Sequencing and genotyping

Parasites from pre-treatment blood samples were sequenced using an Illumina sequencing platform after DNA enrichment by leukocyte depletion. We used the *P. falciparum* 3D7 reference genome to align reads. A collection of worldwide samples was added to discover polymorphisms and a quality control assessment was performed based on coverage and other quality metrics, yielding 681,687 high-quality exonic SNPs. We used a sequencing and genotyping process similar to that described in Chapter 2.

4.2.2 GWAS confirmed artemisinin resistance candidate markers

The GWAS results are fundamental for articulating our subsequent findings, but the following exposition is brief as I did not contribute to this piece of work. The strongest association signal found by the GWAS was located on the k13-C580Y mutation (a nonsynonymous mutation in the propeller domain of the *kelch13* gene), with a P-value of 4×10^{-26}, which corresponds to the most common variant that Ariey and colleagues found to be associated with artemisinin resistance [Ariey et al., 2014]. It was also the most frequent *kelch13* mutation according to the assessment performed by the TRAC study [Ashley et al., 2014]. The GWAS also identified another eight independent candidate loci ($P < 1 \times 10^{-7}$) after applying a conservative significance threshold (when Bonferroni correction is applied to $<$ 20,000 SNP tests). Table 4.3 summarizes these findings.

To conduct the GWAS, we relied on a subset of 18,322 SNPs with good coverage for all the samples (1,063) and a minor allele frequency greater than 0.01. We treated the parasite clearance half-life as a continuous dependent variable and studied the association between genotypes and phenotypes using the FaST-LMM software [Listgarten et al., 2012] . This package implements a linear regression mixed-model and corrects for the effect of population structure by treating genetic similarity as a random effect (this reduced the genomic inflation factor from 14.254 to 1.003). At each locus, missing or mixed genotype calls (i.e. caused by a multiple infections) were excluded.

Only one of the SNPs present in the *kelch13* gene displayed a significant association with clearance half-life but we were conscious that our analysis would have difficulties for detecting signals driven by low-frequency variants in the presence of allelic heterogeneity. Therefore, we tested individually each of the 33 nonsynonymous SNPs we could genotype in the *kelch13* gene. A total of 25 SNPs were located in the highly conserved BTP/POZ and propeller domains [Ariey et al., 2014] and 20 of them rendered a clear association with prolonged clearance half-life. In contrast, none of the variants outside these domains showed a significant association (Figure 4.1).

As 13 of these 20 significant mutations were observed previously in Cambodian samples [Ariey et al., 2014], we classified samples into those with and without *kelch13* resistance alleles. We defined as a resistant *kelch13* mutation any nonsynonymous change in the

Locus	Chr.	Position	Gene ID	Gene description	N/S	Alteration	*P*-value
13-01 (*kelch*)	13	1,725,259	PF3D7_1343700	Kelch protein, putative	N	p.Cys580Tyr	4×10^{-26}
14-01 (*arps10*)	14	2,481,070	PF3D7_1460900.1	Apicoplast ribosomal protein S10 precursor, putative	N	p.Val127Met	1×10^{-20}
13-02 (*fd*)	13	748,395	PF3D7_1318100	Ferredoxin, putative	N	p.Asp193Tyr	3×10^{-17}
14-02	14	2,098,642	PF3D7_1451200	Conserved *Plasmodium* protein, unknown function	S	p.71Asn	3×10^{-12}
14-03 (*mdr2*)	14	1,956,225	PF3D7_1447900	Multidrug resistance protein 2+ (heavy metal transport family) (MDR2)	N	p.Thr484Ile	2×10^{-10}
07-02	7	896,660	PF3D7_0720700	Phosphoinositide-binding protein, putative	N	p.Cys1484Phe	4×10^{-10}
07-01 (*crt*)	7	405,600	PF3D7_0709000	Chloroquine resistance transporter (CRT)	N	p.Ile356Thr	7×10^{-10}
13-03	13	958,469	PF3D7_1322700	Conserved *Plasmodium* protein, unknown function	N	p.Thr236Ile	7×10^{-8}
10-01 (*pph*)	10	490,720	PF3D7_1012700	Protein phosphatase, putative	N	p.Val1157Leu	8×10^{-8}

Table 4.3: Loci most strongly associated with parasite clearance half-life, according to the GWAS.

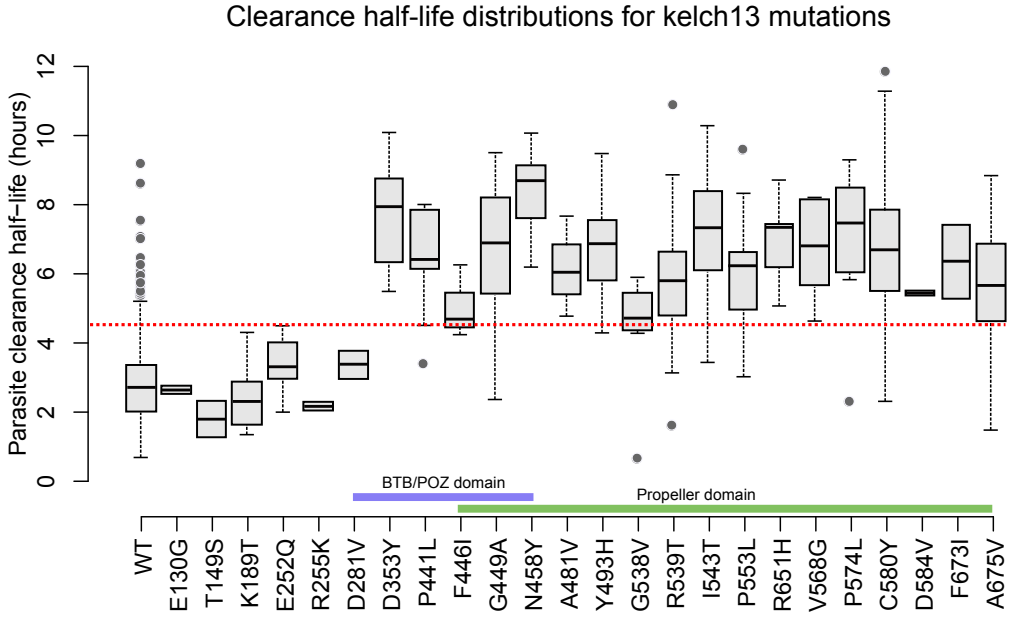

Figure 4.1: Parasite clearance half-life for *kelch13* mutants ($n = 469$) and samples carrying the wild-type (WT) alleles ($n = 630$). We excluded samples with missing or mixed calls ($n = 190$) and *kelch13* singletons ($n = 8$). The red line indicates a speculative threshold (4.5 h) to classify phenotypes as sensitive or resistant. The colored stripes at the bottom of the plot distinguish the mutations occurring in the BTB/POZ and propeller domains. All mutations were observed in isolation (i.e. no sample carried more than one *kelch13* mutation.

BTB/POZ and propeller domains when compared to the 3D7 reference genome. Samples carrying a homozygous *kelch13* resistance allele (i.e. not observed in a mixed call) had a median half-life of 6.5h (interquartile range of $5.4 - 7.8$h) whereas samples without any of such alleles had a median half-life of 2.6h ($2.0 - 3.3$h). For samples carrying a resistance allele in a mixed call, the median half-life was 5.9h ($4.4 - 7.1$h). We interrogated the African samples present in this dataset and found that 8 samples were heterozygous carriers of a *kelch13* resistance allele (i.e. within a mixed call). However, they were not associated with elevated clearance half-life and showed a median of 1.9h ($1.0 - 4.6$h).

We concluded that mutations in the BTB/POZ and propeller domains were the strongest predictors of prolonged clearance half-life across the genome, confirming the association unveiled by Ariey and colleagues. We also concluded that the other candidate loci were associated with clearance half-life because of their population genetics relationship to *kelch13* resistance alleles. Table 4.4 summarizes the geographical distribution of all *kelch13* mutations observed in our dataset.

4.2.3 Population structure was associated with *kelch13* resistance alleles and background mutations

In this section, we investigate the relationship of *kelch13* resistance alleles and population structure. Our goal here is threefold: (1) to identify plausible founder populations (i.e. sympatric subpopulations which, most likely, arose as a result of a recent founder effect), (2) to delimit the set of representative wild-type parasites in each geographical area, and (3) to assess how resistance alleles are related to each of these groups of parasites.

We characterized population structure using a supervised iterative method that produced a conservative classification of samples[3]. We started by applying the model-based approach implemented in the ADMIXTURE (v1.23) package [Alexander et al., 2009] to estimate ancestry proportions for samples in Vietnam and Cambodia. We imputed sparse missing genotypes (as described in Section 2.6) and, since the majority of samples were practically clonal (as characterized by their F_{WS} estimates[4]), we used majority calling to approximate

[3]To minimize ascertainment biases, we only used samples that were included in the SNP discovery phase.

[4]Data not shown. We direct the reader to Chapter 2 as this type of analysis was performed for a substantial subset of Southeast Asian samples

Mutation	CD	NG	BD	MM	TH	LA	KH	VN	Total	Phenotypes	Mean HL
WT	64	2	67	47	118	118	292	138	846	630	2.77
K92N	1								1	1	2.40
E130G							2		2	2	2.60
T149S	2								2	2	1.74
K189T	10	1	5	3					19	18	2.39
E252Q				11	8				19	15	3.38
R255K	2								2	2	2.12
E270K							1		1	1	4.59
D281V				2			1		3	2	3.33
D353Y								5	5	5	7.70
F395Y							1		1	1	3.78
K438N				1					1	1	1.38
P441L				5	6				11	10	6.34
P443S					1				1		-
F446I				3					3	3	5.02
G449A				2	3		2		7	7	6.55
N458Y					6				6	6	8.38
A481V					1		2		3	3	6.13
Y493H							45	4	49	37	6.76
N525D					1				1	1	4.68
N537I					1				1	1	5.02
G538V					8				8	8	4.42
R539T					13	2	25	4	44	38	5.70
I543T							2	22	24	23	7.07
P553L					2			9	11	9	6.03
R561H				2	5				7	5	6.93
V568G								5	5	5	6.67
P574L				6	1				7	7	6.85
C580Y				11	21		241	9	282	246	6.72
D584V							2		2	1	5.41
F614L					1				1	1	2.50
F673I				2					2	2	6.32
A675V				2	11				13	13	5.60
H719N							1		1	1	5.80
Total	79	4	72	97	207	120	617	196	1392	1107	

Table 4.4: Geographical distribution of the samples that carry any of the 33 *kelch13* nonsynonymous mutations (samples with missing or mixed calls have been excuded). The table also shows the mean half-life computed from the available phenotypic data. Mutations in red are located in the *kelch13* downstream region (BTB/POZ and propeller domain), mutations in black are located in the *kelch13* upstream region (with WT refering to wild-type alleles). The major *kelch13* GWAS hit (C580Y) is shown underlined.

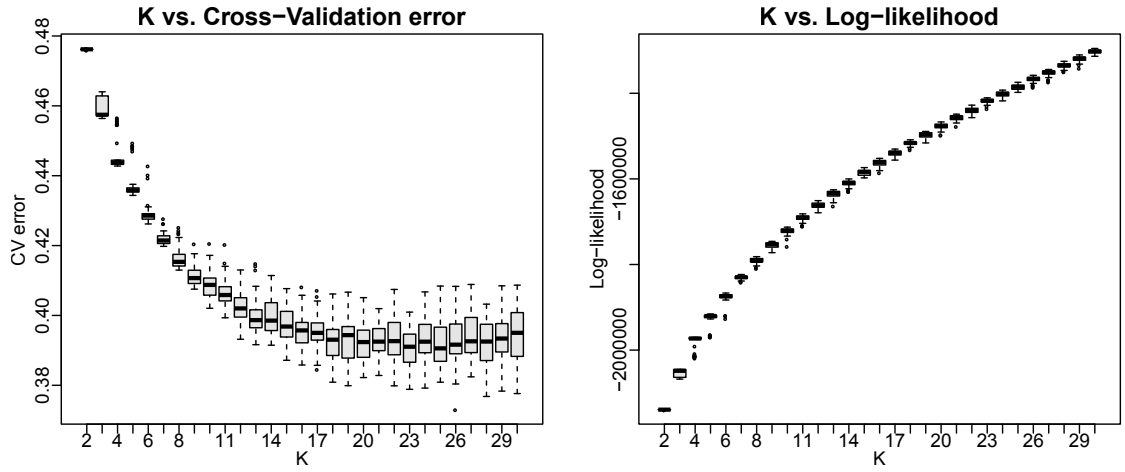

Figure 4.2: Plots showing how the cross-validation and the log-likelihoods vary with respect to K (50 replicates per K with different random seeds). (Left) Boxplot showing the CV-errors for different K values. (Right) Boxplot showing how the log-likelihood of the ancestry proportions assigned by ADMIXTURE vary with K.

the dominant strain in mixed infections. SNPs with extremely low population-wide MAF (MAF ≤ 0.01) were discarded, as they are not very informative for the inference procedure performed by ADMIXTURE. Because the underlying model assumes that all loci are independent, we also excluded SNPs in short-range linkage disequilibrium. We pruned SNPs by using the PLINK toolset [Purcell et al., 2007]. For each regional dataset, we scanned each chromosome using a sliding window of 100 SNPs advanced in steps of 10 SNPs and removed any SNP that had a correlation $r^2 \geq 0.02$ with any other SNP within the window.

We ran ADMIXTURE using 5-fold cross-validation for different values of K (i.e. number of putative populations) in each geographical region. As ADMIXTURE uses a likelihood optimization approach that is dependent on initial conditions, we executed the analysis 50 times for each value of K with different random seeds. We assessed the distribution of cross-validation errors and log-likelihoods for all sets of runs. When all samples were considered together, the cross-validation error distribution presented a large plateau where solutions showed only marginal improvement as K grew (as can be seen in Figure 4.2), accompanied by an increase in variance. This problem has been reported elsewhere when dealing with complex population structure and a substantial number of populations [Porras-Hurtado et al., 2013], [Decker et al., 2014], [Evanno et al., 2005]. This fact made problematic choosing an optimum value for K as we could not apply the "elbow rule" heuristic recommended by

the authors of ADMIXTURE.

To overcome this difficulty, we followed a conservative iterative process for assigning samples to populations in each geographical region. First, we identified the value of K that represented the uppermost level of structure by evaluating the second-order rate of change of the likelihood function on K [Evanno et al., 2005]. Starting with this value, we gradually increased K to capture structure at a finer resolution. For each K, we chose the solution with the lowest cross-validation error. From the proportions estimated in this solution, we assigned each sample to one of K groups corresponding to the putative ancestral populations by using a heuristic. A sample was allocated to a group if the proportion estimated for the corresponding ancestral component was greater than 0.5 and at least four times higher than the second highest proportion. Samples not meeting these criteria were labeled as "unclassified". The objective of this heuristic was to uncover strong signals of population structure by identifying representative samples of "pure" ancestral components (since we were only interested in discovering founder populations, we did not need to retain admixed samples). We sought clusters of samples that consistently grouped together as K increased. Within each group, we allowed one classification mismatch (at most) between consecutive K values. The value of K was incremented until newly identified clusters were deemed to be too small ($n < 5$) or unstable (where clusters separated as an independent group and then merged with a different group at a higher value of K). In addition, any sample that was labeled as unclassified for more than a single value of K was not assigned to any primary cluster but aggregated into the unclassified group. After obtaining a set of reliable clusters for each geographical region, we proceeded to identify core (wild-type representative) groups and potential founder populations. To determine the core populations, we capitalized on the wild-type groups found in Cambodia and Vietnam in our previous study [Miotto et al., 2013]. We evaluated how these previously analyzed samples clustered with the new samples added in this work and assessed their level of genetic similarity via PCoA. Putative founder clusters were characterized by a pairwise count of highly differentiated SNPs ($F_{ST} \geq 0.5$) with respect to the local core populations. The rationale behind this comparison is that the number of highly differentiated SNPs is expected to be much higher in a population that originated from a recent selective founder effect and went through a subsequent population expansion than in a population that diverged from the ancestral group mainly by genetic

drift.

4.2.3.1 Results

We proceeded by dividing samples into two geographical regions[5] to reduce the complexity of the analysis. The first group comprised 288 samples and included 205 samples from Vietnam (VN), 35 from Laos (LA), and 48 from northeast Cambodia (NKH). We included parasites from Laos and Cambodia since they were genetically similar to Vietnamese samples and, in general, non-resistant to artemisinin. After running ADMIXTURE with several K values in the range [2,16] we applied the iterative procedure described in the previous section. We started at $K = 3$, as the ΔK statistic [Evanno et al., 2005] suggested this was the uppermost level of structure, and applied our classification heuristic for increasing values of K. We stop at $K = 8$ since higher K values produced very erratic classifications. After discarding clusters with less than 5 samples and unstable groups, we obtained a total of 5 robust clusters.

We labeled one of these groups VN-C (standing for Vietnam core population) after observing that it consistently clustered with the LA and NKH samples. The other four clusters were named VN-F01 to VN-F04 (indicating that these were candidate founder populations). Figure 4.3 depicts the iterative process followed to identify these groups. As a measure of how likely the candidate founder populations were indeed the result of recent founder effects, we counted how many well-differentiated SNPs ($F_{ST} \geq 0.5$) were present in each group when compared to the core cluster (VN-C). We found large numbers of differentiated SNPs for all putative founder clusters (Table 4.5) although VN-F02 exhibited a substantially lower count (253 SNPs). We also found the aggregated group of unclassified samples (VN-U) to be very close to the core population, presenting no highly differentiated SNPs when compared with wild-type parasites. In addition, we confirmed that all founder populations could be clearly identified as cluster outliers on PC projections when PCoA was performed (Figure 4.4).

The second geographical region included only samples from Cambodia. We expected population structure to be complex since at least four sympatric populations were previously

[5]This decision was supported by the separation of these geographical group in preliminary exploratory data analysis with PCoA and NJ trees.

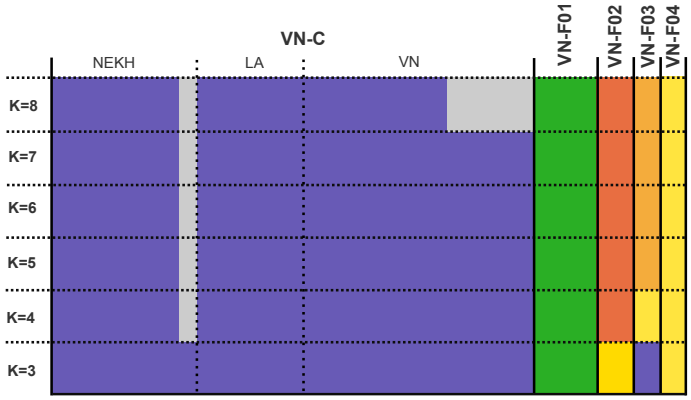

Figure 4.3: Analysis of population structure for VN samples based on ADMIXTURE results. Columns represent samples and each color corresponds to one of the identified populations, with grey signaling samples that failed to meet the clustering criterion at a given K. Samples that clustered together consistently across all K values were assigned to the same group. Samples labeled as unclassified and clusters with less than 5 samples have been omitted for clarity. The core VN population (VN-C) was identified based on its similarity to the samples from LA and NEKH, as they regularly clustered together (only a small fraction of samples from NEKH presented an unstable clustering, colored in gray in the figure). A small set of samples from VN-C clustered with other groups at $K = 8$ (gray band under the VN-C/VN group) but we retained them as they showed robust grouping from $K = 3$ to $K = 7$. All founder populations (in particular VN-F01) showed a very robust clustering pattern across K, indicating they were well differentiated.

	VN-C	**VN-F01**	**VN-F02**	**VN-F03**	**VN-F04**	**VN-U**
VN-C	-	1673	253	718	754	0
VN-F01	1673	-	2298	3093	3259	1479
VN-F02	253	2298	-	1804	1909	216
VN-F03	718	3093	1804	-	2651	717
VN-F04	754	3259	1909	2651	-	682
VN-U	0	1479	216	717	682	-

Table 4.5: Differentiation of SNPs between *P. falciparum* subpopulations in Vietnam. The table shows the number of SNPs with a F_{ST} higher than 0.5 between each pair of populations. As a baseline measure, VN-C and VN-U present no differentiated SNPs when compared. Besides, founder subpopulations show a higher number of differentiated SNPs when compared with each other than when compared with the core population (VN-C).

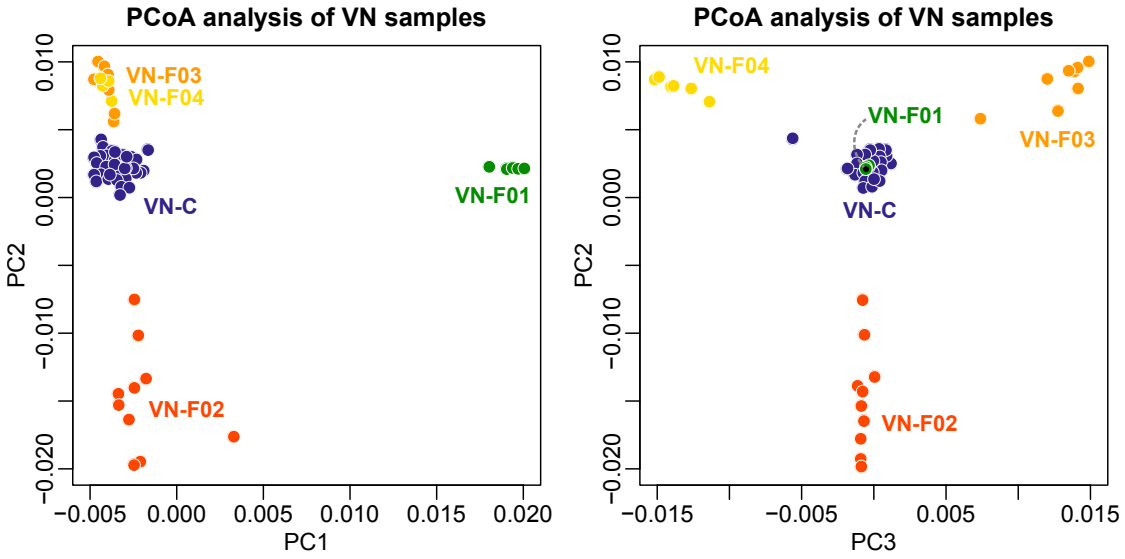

Figure 4.4: PCoA plots for VN samples showing the projections on different PCs. Samples have been colored with respect to their assigned population (VN-U samples have been omitted for clarity). The plots show a high correspondence between outlier clusters and the populations identified by the ancestry iterative ADMIXTURE analysis.

identified in west and northeast Cambodia [Miotto et al., 2013]. In this study, we analyzed all Cambodian samples from that previous study and included additional samples from the Preah Vihear Province (north Cambodia).

After running our iterative procedure for K ranging from 2 to 16, we found that when K was greater than 12, all new groups were either segregating from the putative wild-type population (without differentiating in the PCoA projections) or were too small ($n < 5$) for the aims of our analysis.

We classified samples using the ADMIXTURE results obtained for K in the range [4,11] since the ΔK statistic [Evanno et al., 2005] suggested $K = 4$ to be the uppermost level of structure. We identified a Cambodian core population (KH-C), mainly containing samples from northeast Cambodia that were classified as part of the wild-type population in our previous study [Miotto et al., 2013]. We evaluated an initial set of 10 candidate founder populations (7 in west Cambodia and 3 in north Cambodia). We removed any population with less than 5 samples and assessed the number of highly differentiated SNPs (when compared to the KH-C population) present in each group (Table 4.6).

We also explored how differentiated these populations rendered from the core cluster in the first 10 PCoA projections (Figure 4.6). After this combined analysis, we retained 7

populations, 4 in west Cambodia (termed WKHF01 to WKHF04) and 3 in north Cambodia (named NKHF01 to NKHF03).

A summary of the iterative classification for the identified populations can be seen in Figure 4.5. For both geographical regions, we aggregated all samples that did not end up belonging to the core or founder groups into VN-U and KH-U (standing for Vietnam/Cambodia unclassified).

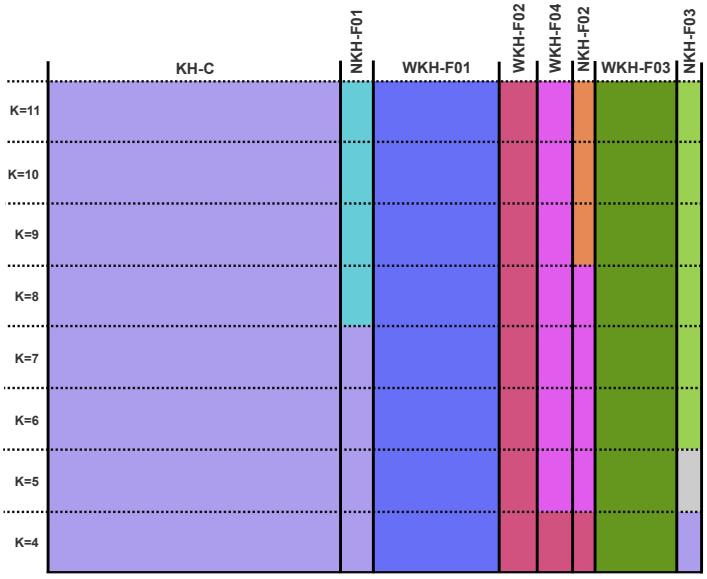

Figure 4.5: Analysis of population structure for Cambodian samples based on ADMIXTURE results. Columns represent samples and each color corresponds to one of the identified populations, with grey signaling samples that fail to meet the clustering criterion at a given K. Samples that clustered together consistently across all K values were assigned to the same group. Samples labeled as unclassified and clusters with less than 5 samples or judged to lack evidence supporting a founder effect have been omitted for clarity. The core KH population (KH-C) contained the majority of samples classified as wild type in a previous study [Miotto et al., 2013]. All founder populations had very stable clustering patterns, suggesting they were very well differentiated.

4.2.3.2 Founder populations were associated with artemisinin resistance

Having outlined a conservative set of founder populations (11 in total), we quantified their association with artemisinin resistance. We found that seven of these founder populations (5 in Cambodia and 2 in Vietnam) were strongly associated with artemisinin resistance (Figure 4.7, and Tables 4.7 and 4.8).

We performed a genome-wide F_{ST} analysis to identify SNPs that were differentiated from the core population of each country (VN-C and KH-C) and were common to all

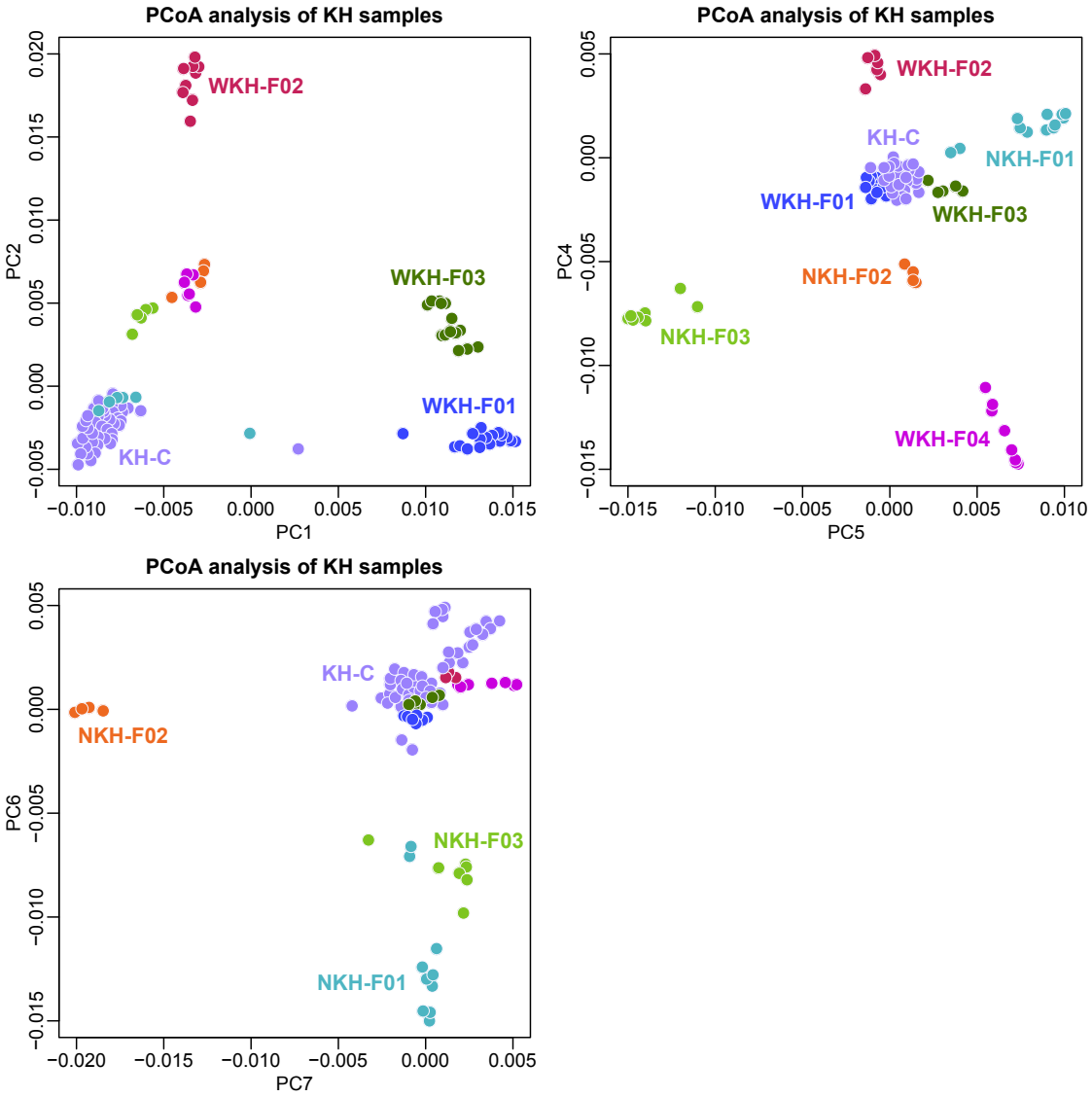

Figure 4.6: PCoA plots for Cambodian samples showing the projections on different PCs. Samples have been colored with respect to their assigned population (KH-U samples have been omitted for clarity). The plots show a clear correspondence between outlier clusters and the populations identified by the iterative ADMIXTURE analysis.

	KH-C	WKH-F01	WKH-F02	WKH-F03	WKH-F04	NKH-F01	NKH-F02	NKH-F03	KH-U
KH-C	-	1431	1412	1192	654	492	1438	1276	4
WKH-F01	1431	-	3098	1490	2113	2428	3038	3329	644
WKH-F02	1412	3098	-	2568	2174	2385	2997	3239	947
WKH-F03	1192	1490	2568	-	1772	2073	2516	2698	631
WKH-F04	654	2113	2174	1772	-	1596	2145	2364	464
NKH-F01	492	2428	2385	2073	1596	-	2461	2592	498
NKH-F02	1438	3038	2997	2516	2145	2461	-	3270	1172
NKH-F03	1276	3329	3239	2698	2364	2592	3270	-	1202
KH-U	4	644	947	631	646	498	1172	1202	-

Table 4.6: Number of SNPs that are highly differentiated (as defined by $F_{ST} \geq 0.5$) between each pair of populations in Cambodia. As in Table 4.5 we treat the comparison between KH-C and KH-U (4) as a baseline measure.

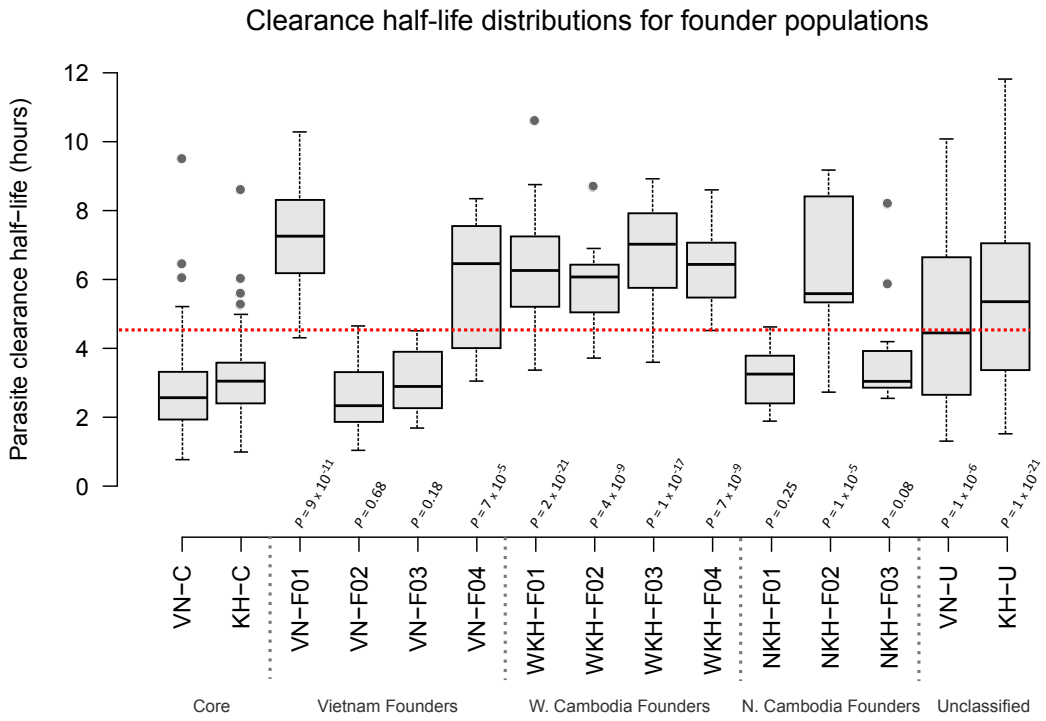

Figure 4.7: Boxplot showing the distribution of parasite clearance half-life in all populations identified by our population structure analysis. The red line indicates a speculative threshold (4.5 h) to classify phenotypes as sensitive or resistant. Two founder populations in Vietnam (VN-F02, VN-F03) and another two in Northern Cambodia (NKH-F01, NKH-F03) rendered as sensitive whereas the rest of founder populations were resistant. In agreement with our expectations, the groups aggregating unclassified samples (VN-U and KH-U) appear to be a mixture of resistant and sensitive parasites. At the bottom we show the P-values that assess how much the distribution of phenotypes in each of the subpopulations differ from that of the core group (see Table 4.8 and Table 4.6).

Population	**Samples**	**Median HL (IQR)**	***N***	***P***
VN-C	76	2.5 (1.9-3.3)	64	
VN-F01	21	7.2 (6.2-8.3)	20	9×10^{-11}
VN-F02	12	2.3 (1.8-3.2)	12	0.68
VN-F03	9	2.8 (2.2-3.9)	8	0.18
VN-F04	8	6.4 (4.0-7.5)	8	7×10^{-5}
VN-U	79	4.4 (2.6-6.6)	72	1×10^{-6}
Total	205		184	

Table 4.7: Association between subpopulations identified in Vietnam and parasite clearance half-life (N refer to the number of samples with phenotype information). VN-C represents the core or wild-type population, VN-F01 to VN-F04 are differentiated founder subpopulations, and VN-U contains samples that could not be classified by our conservative iterative procedure. We assessed how much the distribution of phenotypes in each of the subpopulations differ from that of the core group (P-value computed by two-sample Wilcoxon test).

Population	Samples				Median HL (IQR)	*N*	*P*
	NEKH	**NKH**	**WKH**	**Total**			
KH-C	120	1	3	124	3.0 (2.4-3.5)	119	
WKH-F01			54	54	6.2 (5.2-7.2)	47	2×10^{-21}
WKH-F02		2	14	16	6.0 (5.0-6.4)	15	4×10^{-9}
WKH-F03			34	34	7.0 (5.8-7.9)	33	1×10^{-17}
WKH-F04			15	15	6.4 (5.5-7.1)	13	7×10^{-9}
NKH-F01		14		14	3.2 (2.3-3.7)	14	0.25
NKH-F02		9		9	5.5 (5.3-8.4)	9	1×10^{-5}
NKH-F03		11		11	3.0 (2.8-3.9)	11	0.08
KH-U	11	48	179	238	5.3 (3.4-7.0)	206	1×10^{-21}
Total	131	85	299	515		467	

Table 4.8: Association between subpopulations identified in Cambodia and parasite clearance half-life (N refer to the number of samples with phenotype information). KH-C represents the core or wild-type population, WKH-F01 to WKH-F04 are differentiated founder subpopulations in west Cambodia, NKH-F01 to NKH-F03 are differentiated founder subpopulations in north Cambodia, and VN-U contains samples that could not be classified by our conservative iterative procedure. We stratified the sample count by region (NEKH: northeast, NKH: north, and WKH: west). We assessed how much the distribution of phenotypes in each of the subpopulations differ from that of the core group (P-value computed by two-sample Wilcoxon test).

founders. We observed that the most differentiated SNPs that were common to all founders were the ones identified by our GWAS analysis (Section 4.2.2) with the exception of *kelch13* resistance alleles, mainly due to their relatively low frequency and allelic heterogeneity. The top SNP associated with resistant founder populations was fd-D193Y, followed by crt-I356T, crt-N326S, arps10-V127M and mdr2-T484I. When we aggregated SNPs by gene and repeated the analysis, we found that the *kelch13* gene was the top hit. We observed that 5 SNPs within this gene were close to 100% frequency in at least one of the resistant founder populations and absent (close to 0% frequency) in the core groups (Table 4.9). All the resistant founder populations identified were associated with a unique *kelch13* resistance allele: three carried the common k13-C580Y variant, and the remaining four had the k13-R539T, k13-Y493H, k13-I542T and k13-P553L variants.

In summary, with our analysis of population structure we found that (1) each resistant founder population was associated with a specific *kelch13* resistance mutation, and (2) the majority of the founder populations (both in Cambodia an Vietnam) shared the same mutations (*fd*, *crt*, *mdr2*, *aprs10* and other loci) identified by the GWAS analysis.

These findings suggest that *kelch13* resistance alleles might act as driver mutations in the emergence of artemisinin-resistant founder populations. As these independent *kelch13* mutations appear very often in combination with a specific set of mutations at other loci, we termed this secondary set of alleles the genetic background (or background alleles) of the resistant populations.

Locus (Chr.)	Position	Alteration	GH	CD	BD	MM	TH	LA	VN-C	KH-C	VN-F01	VN-F04	WKH-F01	WKH-F02	WKH-F03	WKH-F04	NKH-F02
kelch (13)	1725259	p.Cys580Tyr				0.11	0.14					0.14	**0.96**		**1.00**		**0.98**
kelch (13)	1725340	p.Prp553Leu					0.01		0.03			**0.66**					
kelch (13)	1725370	p.Ile543Thr									**1.00**						
kelch (13)	1725382	p.Arg539Thr					0.06	0.02						**1.00**			
kelch (13)	1725521	p.Tyr493His							0.01							**1.00**	
fd (13)	748395	p.Asp193Tyr		0.01	0.02	**0.65**	**0.90**	0.02	0.06	0.02	**1.00**	**0.82**	**0.98**	**1.00**	**1.00**	**1.00**	**0.98**
mdr2 (14)	1956225	p.Thr484Ile			0.06	**0.79**	**0.79**	0.22	0.28	0.22	**1.00**	**0.89**	**1.00**	**1.00**	**1.00**	**1.00**	**0.97**
arps10 (14)	2481070	p.Val127Met				0.49	**0.70**	0.12	0.13	0.08	**1.00**	0.07	**1.00**	**1.00**	**1.00**	**1.00**	**0.98**
crt (7)	405362	p.Asn326Ser	0.01		0.31	**1.00**	**1.00**	0.12	0.14	0.06	**1.00**		**1.00**	**1.00**	**1.00**	**1.00**	**0.98**
crt (7)	405600	p.Ile356Thr	0.02	0.29	**0.84**	**0.99**	**0.99**	0.13	0.15	0.05	**1.00**		**1.00**	**1.00**	**1.00**	**1.00**	**0.98**
pph (10)	490720	p.Val1157Leu			0.01	0.32	0.22	0.04	0.31	0.09	**1.00**	**0.92**	**1.00**	**1.00**	**1.00**	**0.80**	**0.99**

Table 4.9: Allele frequency of the mutations that are highly associated with artemisinin resistance in the seven resistant founder populations (columns on the right), core populations of Vietnam (VN-C) and Cambodia (KH-C), and in other sensitive populations (GH, Ghana; CD, Democratic Republic of the Congo; BD, Bangladesh; MM, Myanmar; TH, Thailand; LA, Laos). Blank cells represent a frequency of 0.

4.2.4 Geographical distribution of *kelch13* mutations

In this section, we explore how artemisinin resistance is spreading through Southeast Asia by mapping how *kelch13* resistance alleles are distributed across countries and along sampling years. Figure 4.8 displays the distribution of resistance alleles using a treemap. It is clear that k13-C580Y is the most common resistance allele across countries and that there is also a significant fraction of mutations observed in mixed calls (suggesting that the samples are a mixture of sensitive and resistant parasites).

Figure 4.9 describes the distribution of resistance mutations by sample collection year and shows that the distribution of mutations is very heterogeneous. Two crude patterns can be appreciated in this figure. Firstly, that the number of different resistance mutations observed has been increasing over the years and, secondly, that k13-C580Y has been the most common mutation since 2009[6].

[6]Notice that in 2007 we observed k13-Y493H to be the most common mutation but since then k13-C580Y

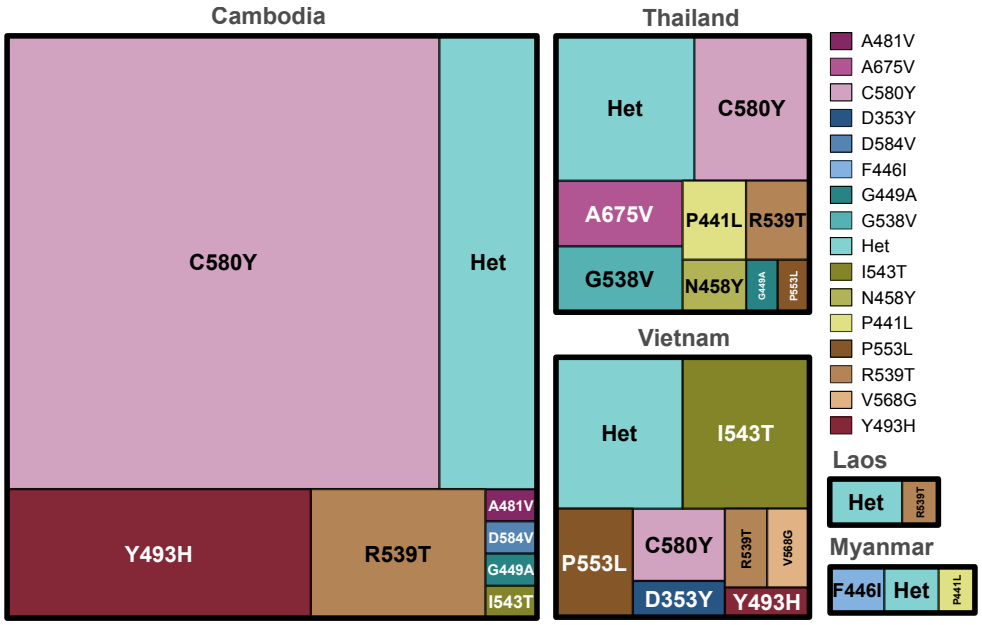

Figure 4.8: Treemap showing the distribution of resistance alleles in Southeast Asia. The area of each rectangle is proportional to the number of samples carrying the associated mutation. Samples that present a resistance mutation in a mixed call are aggregated in the Het category. Only showing data for mutations observed in more than 2 samples as a homozygous call ($n = 552$)

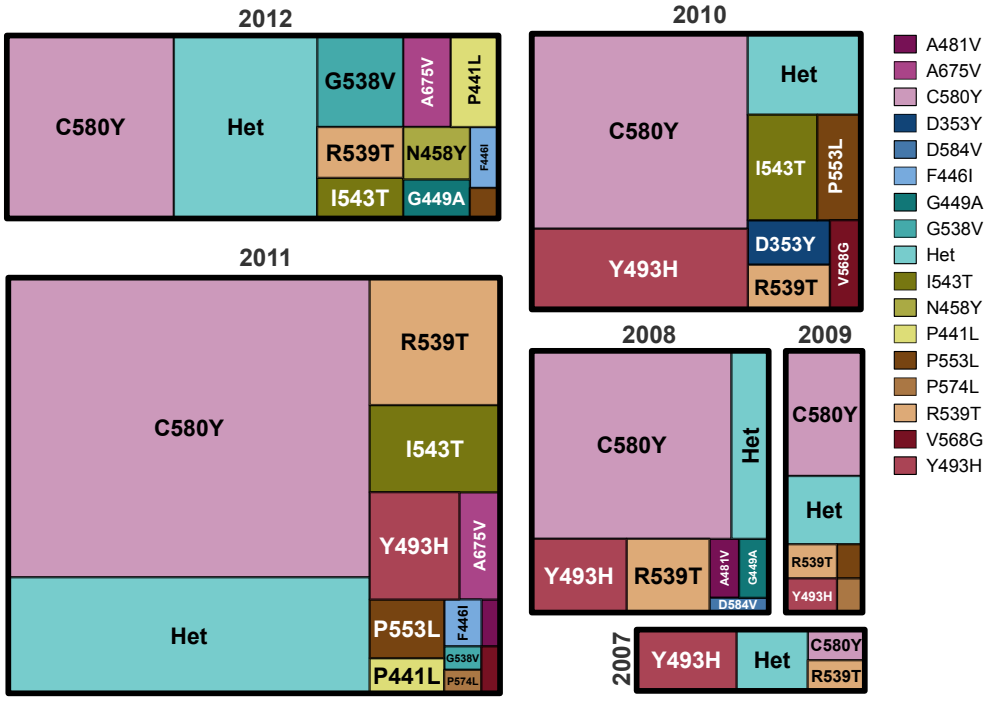

Figure 4.9: Treemap showing the distribution of resistance alleles in Southeast Asia according to sample collection year. The area of each rectangle is proportional to the number of samples carrying the associated mutation. Samples that present a resistance mutation in a mixed call are aggregated in the Het category. Only showing data for mutations observed in more than 2 samples (as an homozygous call) and for which collection year was available ($n = 546$)

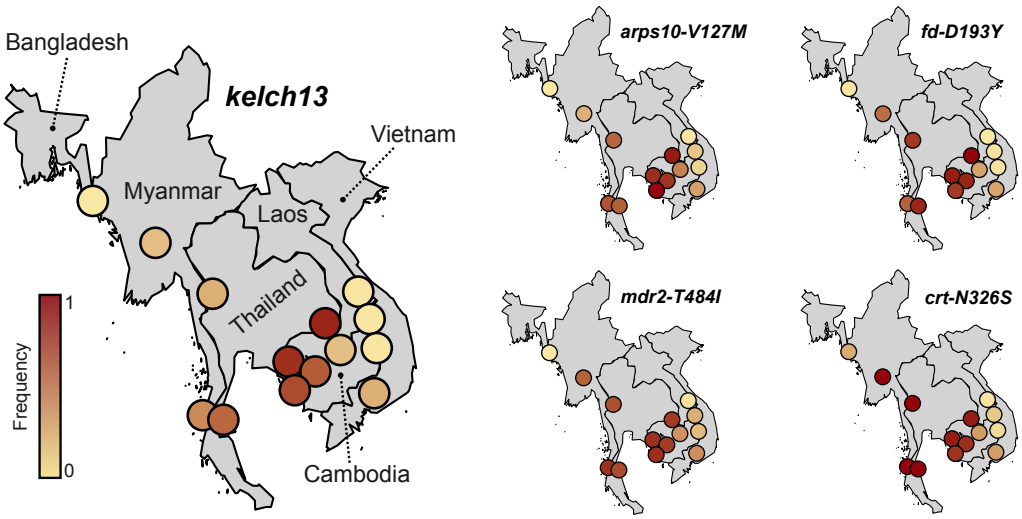

Figure 4.10: (Left) Distribution of *kelch13* resistance mutations in Southeast Asia, each circle represents a sampling site with color proportional to frequency. (Rigth) Distribution of four potential background mutations on the same geographical region (only artemisinin-resistant parasites).

We also composed a map of resistance allele frequencies across all sampling locations (Figure 4.10, left; and Table 4.10). The pattern emerging from this picture is consistent with a hotspot of resistance in western Cambodia. We also observed intermediate-frequency of resistance alleles in Vietnam, Thailand and Myanmar, and very low frequency in Laos and Bangladesh. Nonetheless, the spatial distribution of frequencies appears to be partitioned, with a sharp discontinuity in the frequency of resistance alleles at the junction between Cambodia, Thailand and Laos. We investigated this matter by building a neighbor-joining tree (Figure 4.11) from genome-wide genetic distances[7]. This tree split the samples into three major clades that geographically corresponded to western Southeast Asia (WSEA), Eastern Southeast Asia (ESEA) and Bangladesh[8] (BD). The WSEA clade contained samples from Myanmar and western Thailand whereas the ESEA clade included samples from Cambodia, Vietnam, Laos and eastern Thailand. In a broad sense, this exploratory analysis supports the view that two genetically distinct sets of parasites are separated by the malaria-free corridor that runs through the center of Thailand (Figure 4.11, left).

These results prompted us to investigate the differences between populations of parasites

has dominated the pool of mutations.

[7]We computed genetic distances following the method introduced in Section 3.3

[8]Here we note that the BD group appear to be in fact more related to the WSEA group.

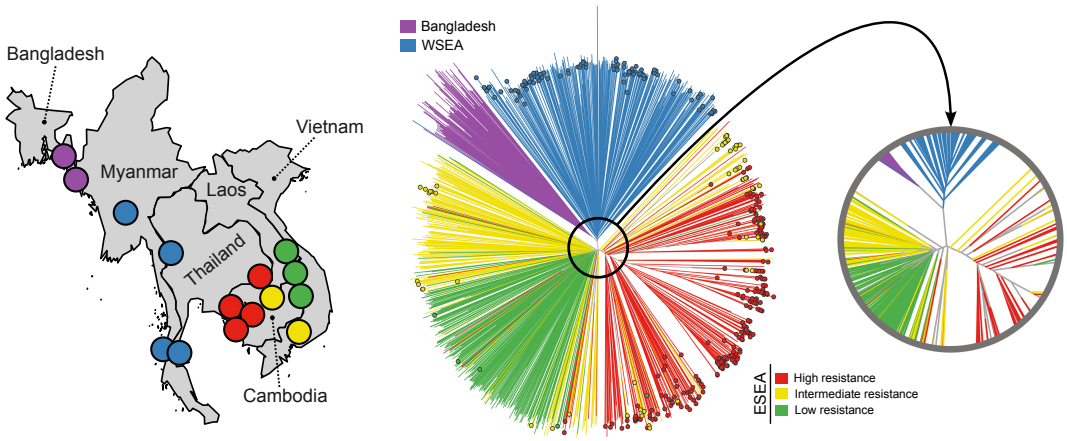

Figure 4.11: (Left) Geographic distribution of samples. (Right) Neighbor joining tree in which colored tips indicate *kelch13* carriers (samples with a mixed infections for *kelch13* were omitted). The inset shows the branching pattern near the tree root. The tree separates samples located in Bagladesh and WSEA (Thailand-Myanmar border) from the ESEA (lower Mekong region) region. Besides, the tree stratifies samples within ESEA according to their degree of resistance (high/low), with samples with intermediate phenotypes being split between the two.

Compartment	**Country (Region)**	**Location**	**Frequency**
BD	Bangladesh	Bandarban	0.0%
		Ramu	0.0%
WSEA	Myanmar (Central)	Bago	21.8%
	Thailand (West)	Mae Sot	29.2%
	Myanmar (South)	Kawthaung	47.8%
	Thailand (South)	Ranong	66.7%
ESEA-HR	Thailand (East)	Sisakhet	94.4%
	Cambodia (West)	Pailin	91.2%
		Tasanh	79.2%
		Pursat	70.9%
ESEA-VR	Vietnam	Binh Phuoc	29.4%
	Cambodia (North)	Preah Vihear	20.6%
ESEA-LR	Cambodia (Northeast)	Ratanakiri	2.1%
	Laos	Attapeu	2.4%
		Xepon	0.0%

Table 4.10: Frequencies of *kelch13* mutants at the 15 Asian sites surveyed, stratified by country and reistance comparments (as defined in main text).

Chr	Pos	Gene Id	Gene Description	N/S	Mutation	F_{ST}		
						WSEA-BD	**HR-LR**	**Mean**
13	748395	PF3D7_1318100	ferredoxin, putative	N	D193Y	0.64	0.85	0.75
13	754133	PF3D7_1318300	conserved Plasmodium protein, unknown function	N	T75I	0.46	0.81	0.63
7	405362	PF3D7_0709000	chloroquine resistance transporter (CRT)	N	N326S	0.53	0.71	0.62
14	2481070	PF3D7_1460900.1	apicoplast ribosomal protein S10 precursor, putative	N	V127M	0.44	0.64	0.54
14	1956225	PF3D7_1447900	multidrug resistance protein 2+(MDR2)	N	T484I	0.53	0.44	0.48
12	1010085	PF3D7_1224800	conserved Plasmodium protein, unknown function	N	N266I	0.31	0.54	0.42
14	2096133	PF3D7_1451200	conserved Plasmodium protein, unknown function	N	G908R	0.43	0.39	0.41
14	1481740	PF3D7_1436300	translocon component PTEX150 (PTEX150)	N	D655A	0.30	0.51	0.41
10	497461	PF3D7_1012900	conserved Plasmodium protein, unknown function	N	T38I	0.15	0.64	0.39
7	405600	PF3D7_0709000	chloroquine resistance transporter (CRT)	N	I356T	0.08	0.70	0.39

Table 4.11: Highly differentiated SNPs between population compartments (WSEA vs. Bangladesh, and high vs. low resistance).

residing in low and high resistance areas. Guided by the topology of the neighbor joining tree, we framed this question by performing a genome-wide F_{ST} scan that compared samples from WSEA and BD and also parasites belonging to the high and low resistance areas of ESEA (Figure 4.11 and Table 4.11). As a result of this analysis, we found four SNPs that clearly marked the geographical boundaries of resistance (fd-D193Y, mdr2-T484I, crt-N326S and arps10-V127M). Besides, we found that the frequency of these mutations mimicked that of *kelch13* resistance alleles (Figure 4.10, right) and that three of these mutations (arps10-V127M, fd-D193Y and mdr2-T484I) were either rare of absent in the African populations represented in this dataset[9]. We note that these findings are coherent with selection acting on these loci within Southeast Asia (Table 4.9).

The WSEA and ESEA populations were genetically and geographically distinct. In ESEA we found that background alleles were strongly associated with the presence of *kelch13* alleles in individual samples, in part due to the presence of strong founder effects. In WSEA, however, there was less evidence suggesting strong founder effects and background alleles were present at high frequency but weakly associated with resistance alleles in individual parasites (Table 4.12). We speculate that the higher malaria transmission rates observed in WSEA may be a determinant factor as they tend to increase recombination among different strains and, because of this reduced inbreeding, to decouple resistance alleles from their original genetic background.

[9]D.R. Congo ($n = 113$) and Ghana ($n = 475$).

Mutation	BD	WSEA	ESEA		
			HR	IR	LR
kelch13	0%	33%	79%	27%	2%
arps10 V127M	0%	61%	92%	42%	12%
fd D193Y	2%	81%	95%	35%	3%
mdr2 T484I	6%	78%	88%	46%	23%
crt N326S	31%	100%	94%	38%	10%

Table 4.12: Frequency of mutations associated with artemisinin resistance (*kelch13* and background) in different population compartments.

4.2.5 Independent origin of *kelch13* mutations

A fundamental matter in terms of policy making and malaria control efforts is to understand how artemisinin resistance is spreading across Southeast Asia. Two broad scenarios can be hypothesized, either spread due to the geographical migration of resistant parasites or multiple independent origins of resistance in different locations. The difference might appear subtle at first hand, once resistance has spread but it has very different implications for approaching containment [WHO, 2011].

In this study, we observed a large number of *kelch13* resistance alleles (20) the majority of which seemed to be geographically localized, suggesting a scenario of independent origins. Nonetheless, we also observed some resistance alleles present in multiple sites and countries. For instance, the k13-C580Y allele was observed at 3 locations in WSEA and 7 locations in ESEA. To ascertain if the geographical dispersion of the k13-C580Y allele was due to migration (i.e. gene flow) or to parallel evolution, we reconstructed the haplotypes extending 100 kb on both flanks of the mutation. This exploratory analysis showed a set of different background haplotypes (Figure 4.12, top), signaling the multiple origins scenario as the most likely explanation.

We extended our exploratory data analysis to all resistance mutations in the dataset and found that most were associated with one or more unique haplotypes, again suggesting that, most likely, they have originated from independent mutational events. In some cases, different haplotypes harbored the same resistance alleles indicating that some mutations were recurrent (i.e. have emerged independently multiple times), as recently suggested by another analysis [Takala-Harrison et al., 2015].

We performed a more detailed demographic analysis to investigate if parasite migration

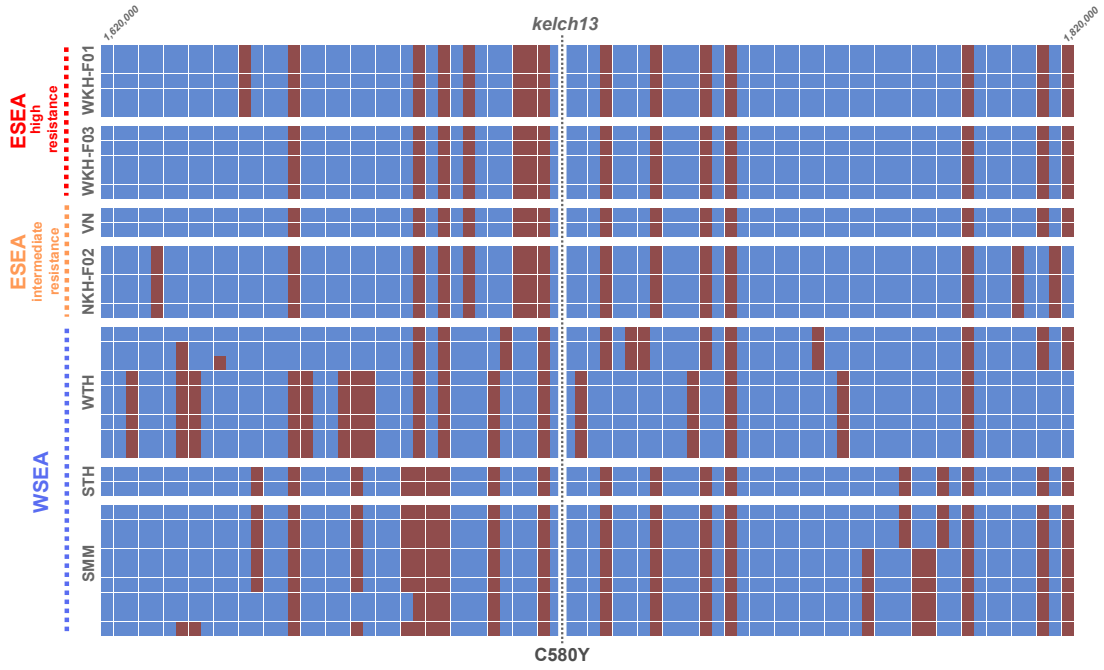

Figure 4.12: Haplotypes for the region surrounding the *kelch13* gene (100kb on each flank) for a selection of resistant k13-C580Y mutants. Columns represent SNPs (MAF ≥ 0.1 in our dataset), with rows corresponding to samples. Light blue represents the reference allele whereas red corresponds to the alternative. We have stratified samples by geographical region, founder population and level of resistance (see labels).

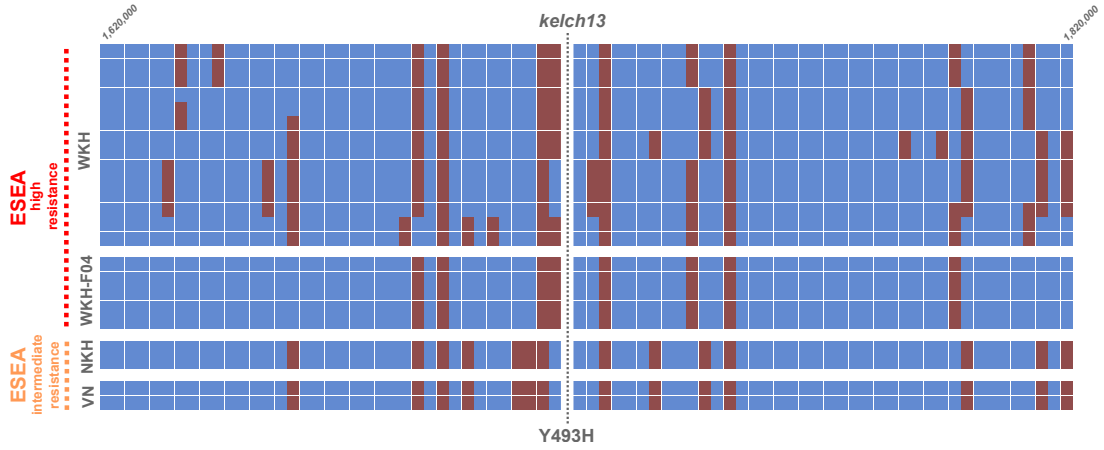

Figure 4.13: Same analysis as in Figure 4.12 for a selection of k13-Y493H mutants (100kb on each flank surrounding the *kelch13* gene). Again, we have stratified samples by geographical region, founder population and level of resistance.

was also a major contributor to the spread of resistance. Our analysis was based on summarizing patterns of haplotype sharing and it is motivated by the fact that sequences sharing longer haplotypes are also more likely to share the same recent genealogical history. In this regard, longer haplotypes can be interpreted in terms of the time to the MRCA of both sequences, with the length of the shared haplotype interpreted as a manifestation of the recombination clock. In technical terms, what we mean here with the length of a shared haplotype is the length of identity by descent (or IBD) tracts, as discussed in Section 1.5.1.

Since IBD segments are not directly observable but need to be inferred from sequence data, we relied on identity by state (IBS) tracts as an approximation. More formally, we approximated the length of the underlying IBD segment surrounding a locus (l) by computing the total length of the IBS tracts on both flanks. We extended the IBS tract in both strand directions, starting from the mutation of interest, till finding a breakpoint (a difference in the genotypes of both sequences). The distance, in base pairs, between two breakpoints is the length of the IBS tract for the pair of sequences. To minimize the influence of possible artifacts due to the IBS approximation, we heuristically called a breakpoint only if the mismatching allele had a frequency $> 5\%$ among samples carrying the same *kelch13* mutation. Variants arising from the same recent evolutionary event will be embedded within almost identical long haplotypes, whereas mutations originating separately will share significantly shorter haplotypes[10]. Besides, the length of the IBD tract decreases as the time to the MRCA increases due to the effects of recombination. Hence, samples that share a very recent genealogical history are expected to group together when clustered by IBS tract lengths.

Our method to summarize patterns of haplotype sharing consists, firstly, on constructing a pairwise matrix relating each pair of samples via their reciprocal IBS tract length. From this matrix, we build a tree using a standard hierarchical clustering method (`hclust`) implemented in the R `stats` package[11]. Figure 4.14 sketches the steps involved in building this summary tree. We pursue a more formal and efficient approach to summarizing and visualizing genome-wide patterns of haplotype diversity in Chapter 5.1, expanding on some of the ideas introduced here.

[10] Assuming that recombination rates are comparable among populations.
[11] We used the Ward's minimum variance criterion

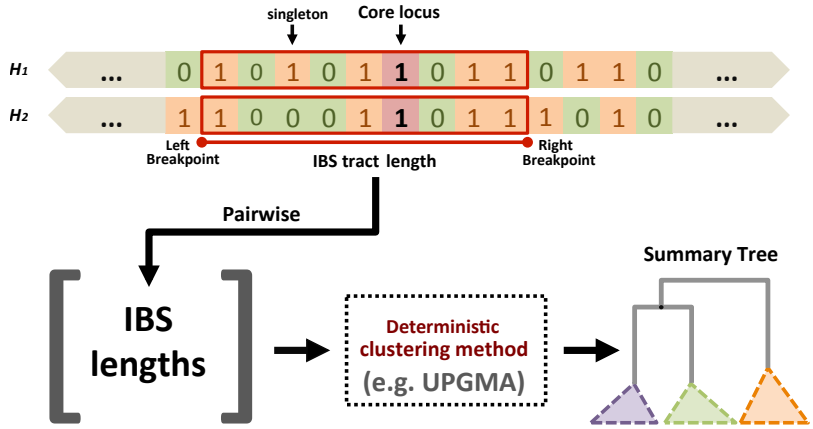

Figure 4.14: Diagram representing our tree-based method to summarize patterns of haplotype sharing (see main text). Notice that to build the tree we use the reciprocal of IBS tract lengths.

Our IBS approximation is justified by the fact that different subpopulations of resistant parasites shared very long haplotypes around the *kelch13* region (in some cases spanning the whole chromosome), a pattern consistent with a hard sweep acting on resistance mutations[12]. Since our goal here is not to date demographic events but to clarify the role played by gene-flow or parallel evolution in the spread of resistance, we believe the extreme patterns of haplotype sharing exposed by the IBS[13] approximation has enough resolution to address this question.

We used the method introduced above to build a tree summarizing haplotype sharing among all *kelch13* mutants. To account for heteroallelism, we collapsed all *kelch13* mutations into a single polymorphism and used the k13-C580Y locus as the notional position of the mutation of interest. The resulting tree[14] summarized the haplotype sharing distribution of the mutants and clustered samples that were likely to share the same recent demographic history. Figure 4.15 shows the summary tree for the 11 most common resistance alleles, stratifying mutations by country at the bottom. Furthermore, Figure 4.16 highlights the distribution of the four most common resistance alleles (k13-C580Y, k13-Y493Y, k13-R539T and k13-I543T) within the same tree.

[12]Notice that due to the nature of the *kelch13* gene, exhibiting allelic exclusion, the sweep would present itself as *soft* if mutations were not aggregated.

[13]On a technical note, in the article we termed the IBS segment lengths as the longest common haplotype length or (LCHL) but here we revert back to the more standard *IBS tract lengths* as it avoids confusion and can be easily related to the literature.

[14]As we were only interested in the topology of such tree (i.e. the clustering) we disregarded branch lengths.

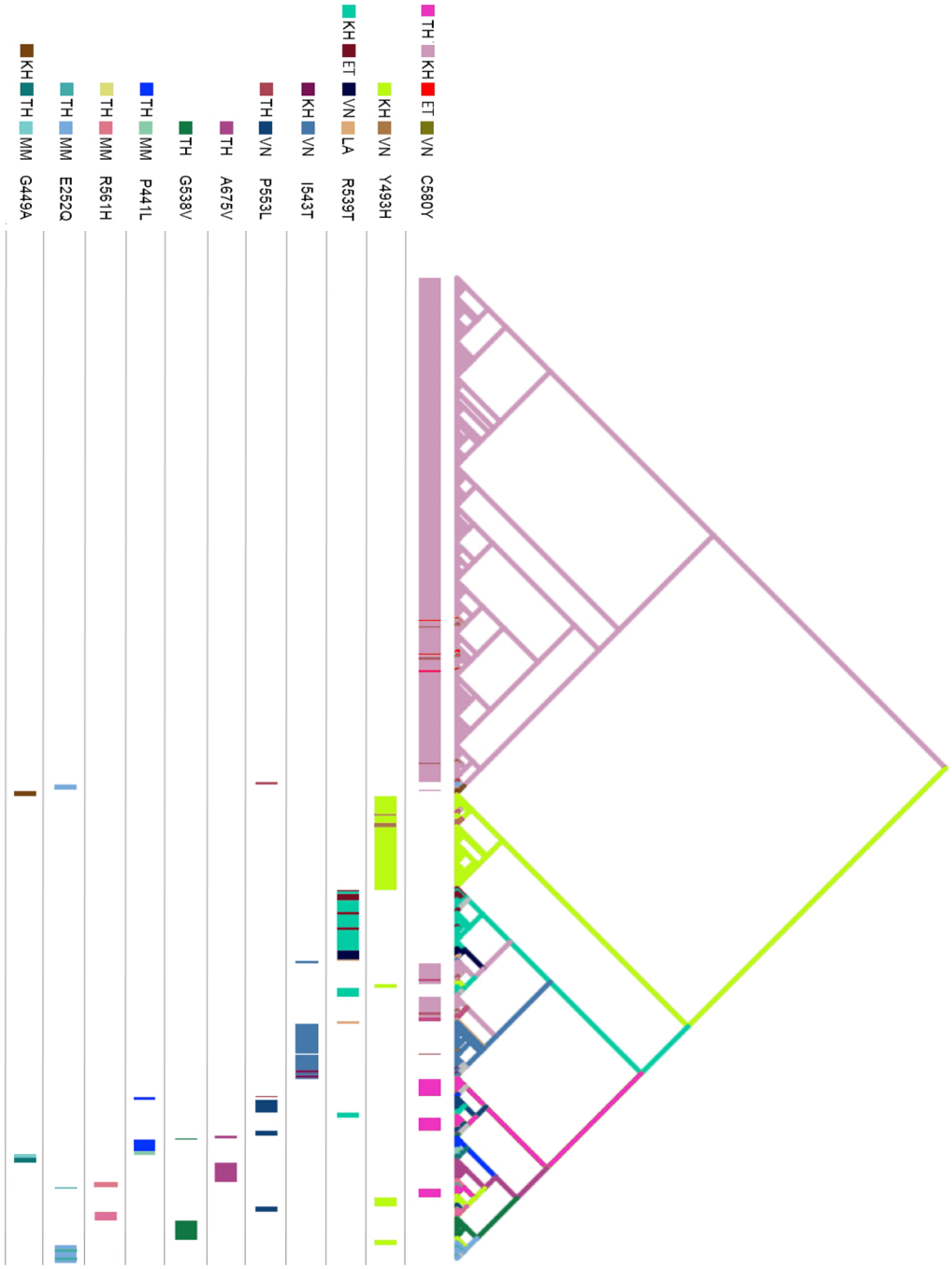

Figure 4.15: Summary tree for the pariwise distribution of IBS tract lengths. Tips represent samples and have been colored by mutation and country of origin (KH, Cambodia; VN, Vietnam; LA, Laos; TH, western Thailand; ET, eastern Thailand; MM, Myanmar). Internal branches have been colored by the most frequent mutation present in the pending subtrees. The bars at the bottom offer visual aid for tracking how different mutations cluster and segregate across the tree. Branch lengths have no meaning as we are only interested in tree topology.

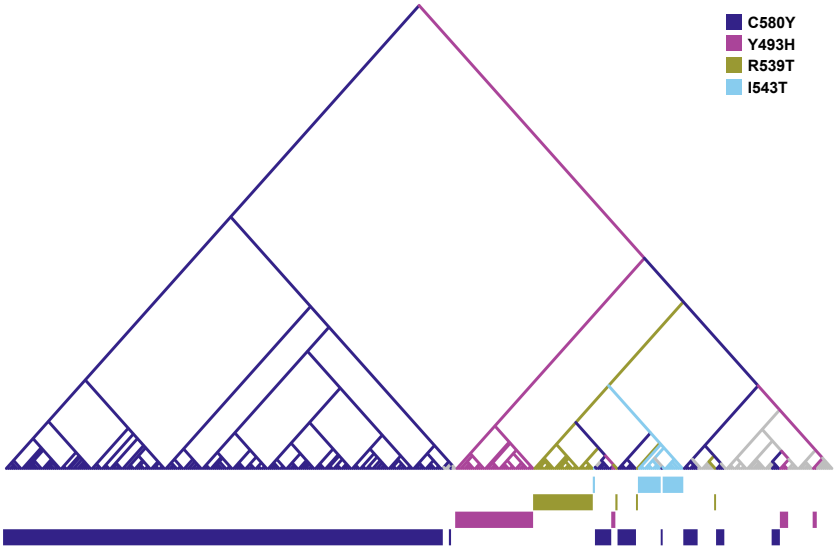

Figure 4.16: Same summary tree as in figure 4.15 but highlighting the four most common *kelch13* resistance mutations. Tips represent samples and have been colored by mutation (gray for other mutations), internal banches colored by the most frequentn mutations present in the pending subtrees. Branch lengths have no meaning as we are only interested in tree topology.

We found that samples carrying *kelch13* mutations that were not common tended to cluster by allele, in agreement with each of these groups originating from a different but recent evolutionary event. The majority of samples carrying the most common resistance allele (k13-C580Y) grouped within a large clade that encompassed several clusters, consistent with a common origin of the k13-C580Y alleles shared by different subpopulations in Cambodia. Nonetheless, we observed two separate clusters of Cambodian k13-C580Y mutants whose flanking haplotypes were identical to those of parasites carrying other *kelch13* mutations (k13-R529T and k13-I543H), suggesting that this allele might have emerged[15] independently multiple times in ESEA.

Furthermore, we found that parasites carrying the k13-C580Y mutation in western Thailand occupied a separate clade from those in ESEA and shared longer haplotypes with other WSEA mutants, suggesting an independent mutational event, a conclusion supported by the short length of the haplotype shared by ESEA and WSEA parasites. Figure 4.17 (panel A) shows this in detail by mapping shared haplotype decay. Clusters in ESEA present extremely long haplotypes (sometimes chromosome-wide), consistent with founder effects. These clusters share a sizeable core haplotype, suggesting a common origin of

[15]We cannot discard the occurrence of gene conversion.

their C580Y mutations. By comparison, the haplotype shared with C580Y mutants from western Thailand is significantly shorter, suggesting an independent origin of the mutation in the two regions. Figure 4.18 compares graphically the length of the core IBS region found in each country for all C580Y mutants. In addition, the heatmap representation of the pairwise IBS tract length distribution (Figure 4.17, D) also reveals several subsets of parasites sharing extremely long haplotypes.

We also found evidence that suggested the migration of resistant parasites for some of the most common alleles (k13-C580Y, k13-I543T, k13-Y493H and k13-R539T). In particular, we observed clusters containing a mixture of ESEA parasites from more than one country (Cambodia, Vietnam and eastern Thailand), suggesting that mutants have crossed international borders, at least within this region (Figure 4.15).

However, we found no evidence that k13-C580Y mutants from ESEA might have migrated to the WSEA region or vice versa, consistent with the observation that k13-C580Y mutants are genetically more similar to samples from their own geographical region than to k13-C580Y mutants in other regions as revealed by our exploratory and population structure analysis.

4.2.5.1 Discussion

In this work we identified 20 distinct mutations in the *kelch13* gene that were associated with artemisinin resistance, all of them located within the BTB/POZ and propeller domains. Our demographic analysis, based on patterns of haplotype sharing, in combination with our characterization of population structure indicated that most resistance alleles have emerged independently and are geographically confined into small regions. We also found evidence that the geographical distribution of the most common resistance allele (k13-C580Y) is the product of recurrent mutations, arising multiple times at different locations. Finally, for some of the most common resistance alleles (k13-C580Y, k13-I543T, k13-Y493H and k13-R539T), we also found evidence supporting the migration of resistant parasites across countries but only within the ESEA region . Thus, we conclude that the main factor involved in the increase of artemisinin resistance was the independent emergence of kelch13 resistance alleles.

The independent introduction of *kelch13* resistance alleles has left a recognizable signal

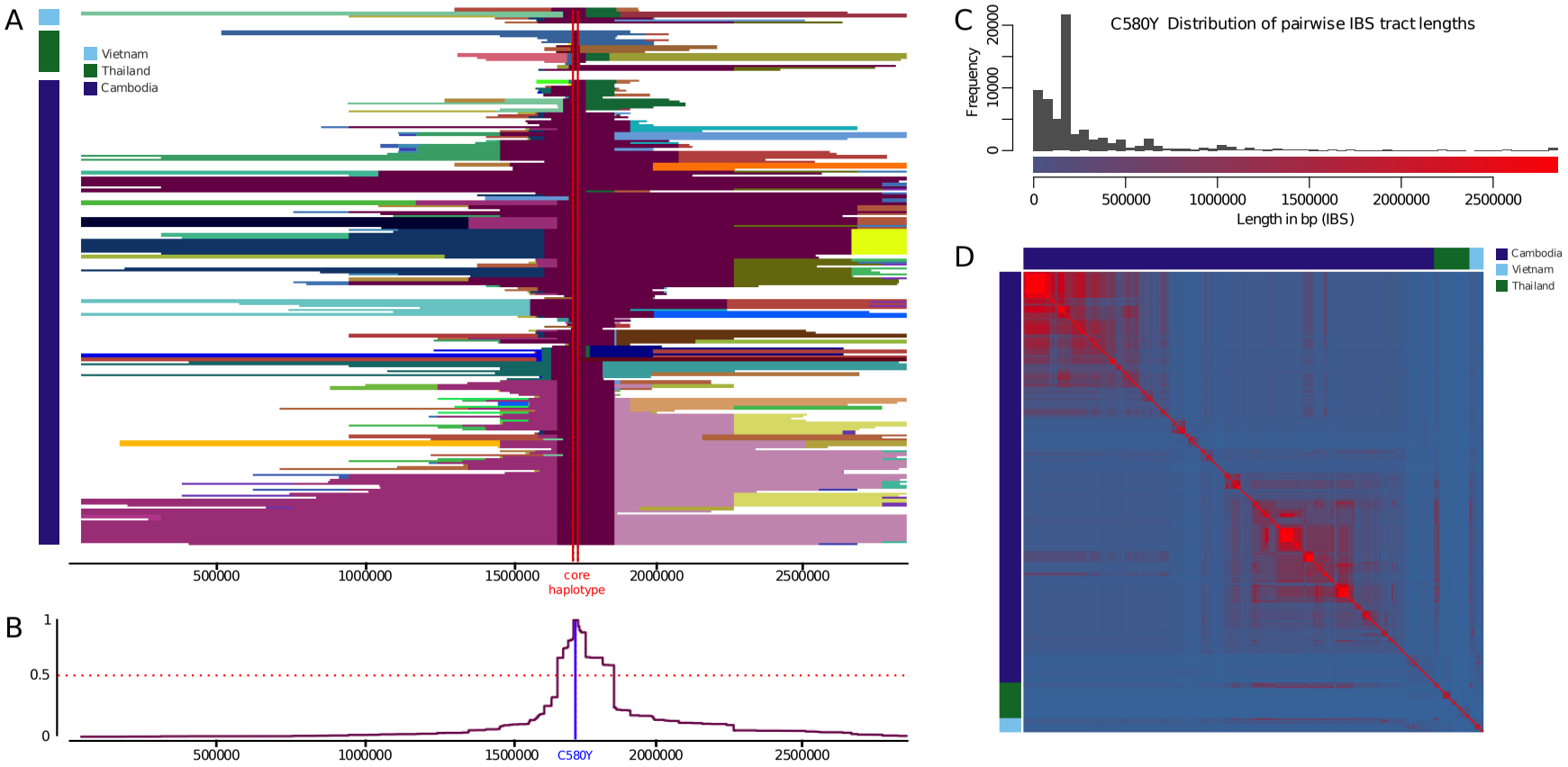

Figure 4.17: (A) Shared haplotype decay. Each horizontal line represents a haplotype, and each color is used to indicate a shared haplotype (we used different colors for each flank). Two vertical red lines mark a 17kb core region identical in all samples.(B) Haplotype homozygosity decay, the blue line highlights the position of the C580Y locus. (C) Histogram showing the pariwise distribution of IBS tract lengths for the k13-C580Y mutation. (D) Representation of the pairwise IBS distribution as a heatmap. Red-like blocks indicate sets of parasites sharing very long IBS tracts.

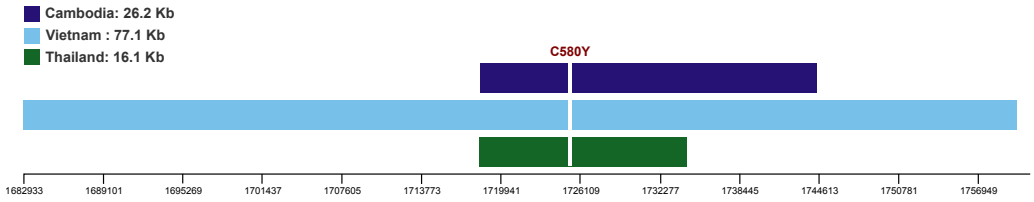

Figure 4.18: Comparison of core haplotypes (i.e. IBS tract length shared by all samples) for the k13-C580Y mutation when stratified by country of origin.

in the genome of *P. falciparum*, in the form of soft selective sweeps, exhibiting high levels of haplotype homozygosity over large regions surrounding most mutations [Hermisson and Pennings, 2005], [Pennings and Hermisson, 2006], [Messer and Petrov, 2013]. There is evidence that this type of sweeps have played a fundamental role in the emergence of resistance to previous drugs [Nair et al., 2007], [Salgueiro et al., 2010].

From an epidemiological perspective, understanding what contributes to the emergence of new *kelch13* mutations is fundamental to inform plans of malaria control. In this regard, we observed that *kelch13* resistance alleles were frequent in Cambodia, Vietnam, Thailand and Myanmar, but present at very low frequency or nonexistent in other Asian countries and in the African samples used in this study. Malaria transmission intensity, the prevalence of different vector species or the use of specific of antimalarial drugs are just a few of the vast array of epidemiological factors that could trigger or modulate the emergence of resistance alleles in Southeast Asia. However, the fact that resistance has emerged independently in countries with different treatment protocols and transmission rates suggests that these features are unlikely to be the sole cause prompting the emergence of resistance.

From the inspection of genomic data, we noticed that a common set of mutations consistently escorted *kelch13* resistance alleles. As a result of our GWAS, and subsequent population structure analysis, we identified and characterized a collection of markers for this genetic background. We observed that (1) the frequency of resistance alleles was correlated with the frequency of the background markers; (2) these markers showed high levels of genetic differentiation between regions of high and low resistance and were absent or present at lower frequency in others parts of the world, mimicking the distribution of *kelch13* mutations; (3) different resistant founder populations, dominated by different *kelch13* alleles, tended to share the same background mutations; and (4) these markers were strongly

associated with slow parasite clearance but their association was seriously diminished[16] when *kelch13* mutations were treated as a covariate. This collection of background markers included non-synonymous variants for *arps10* (chromosome 14), *fd* (chromosome 13), *mdr2* (chromosome 14) and *crt* (chromosome 7). Based on these observations, we came to the conclusion that the presence of this set of background markers can be used as a warning sign that a particular strain will acquire *kelch13* resistance alleles.

Although the specific function of this genetic background is yet to be clarified, we find plausible that this set of alleles has either a compensatory or resistance-enhancing effect, or perhaps a mix of the two, with different parts of the background playing singular roles. The compensatory hypothesis is motivated by the observation that the *kelch13* propeller domains are highly conserved across *Plasmodium* species. It is not unlikely that resistance alleles carry a fitness cost and this burden is somehow reduced via compensatory mutations. We evaluated the possibility that the background alleles had been selected for resistance to any of the ACT companion drugs but judged it unlikely because the background is present in several countries that use at least three different partner drugs.

Nonetheless, since the background mutations appear to have arisen long time before *kelch13* resistance alleles[17] and we also observed that most parasites were resistant[18] to chloroquine and pyrimethamin, another possibility is that their presence is a manifestation of Southeast Asian parasites adapting to a new niche in which most parasites are already resistant to multiple drugs.

As resistance to other drugs, such as chloroquine, sulfadoxine-pyrimethamine and mefloquine [Wongsrichanalai et al., 2002], emerged in this region in the past, we recollect here the set of factors that have been proposed as responsible for the predisposition of Southeast Asian parasites to develop resistance: the high level of inbreeding caused by low transmission rates [Miotto et al., 2013], an erratic history of public health interventions [Payne, 1988], the widespread availability of counterfeit antimalarial drugs [Tabernero et al., 2014] and, finally, the possibility of these parasites being equipped with a genetic makeup [Rathod et al., 1997] that promotes hypermutability, a claim backed up by some experimental evidence [Beez et al., 2011], [Sen and Ferdig, 2004].

[16] Therefore they have a small effect on delaying parasite clearance half-life.
[17] Base on allele frequency and the length distribution of shared haplotypes.
[18] The majority of parasites possessed both the *crt* (p.Lys76Thr) and *dhfr* (p.Ser108Asn) mutations.

In this study, we provide a clear genetic epidemiological assessment of the spread of artemisinin resistance in Southeast Asia, including the role of population structure and founder effects, the genetic makeup that tends to promote the emergence of resistance and the demographic events that may have modulated the spread of resistance alleles. After assessing how artemisinin resistance is spreading across Asia, we anticipate that, unless immediate efforts are concentrated on eliminating malaria from the greater Mekong subregion, further independent mutations or opportunistic gene flow events may lead to high-levels of resistance for both artemisinin derivatives and ATC partner drugs. These concerns are not solely limited to Southeast Asia. As the parasite keeps evolving resistant mechanisms, a solid understanding of the multiple genetic and epidemiological factors involved in this process will render vital in the battle to prevent its spread.

4.2.6 Individual contributions

In this collaborative project, my contributions were limited to the study and characterization of population structure, the relationship of founder populations with artemisinin resistance and the assessment of the most plausible demographic scenario that could explain the observed geographical spread of *kelch13* mutations. Specifically, I characterized population structure on my own and collaborated with Olivo Miotto to generate our conservative classification of populations, discerning their relationship with drug resistance. I performed the demography assessment of *kelch13* mutations (based on patterns of haplotype sharing) but want to acknowledge helpful discussions with Roberto Amato. I played no role in any other task related to this study but collaborating on articulating the final discussion.

Type	Locus	AFR (N=1648)	SEA (N=1599)	SAS (N=75)	OCE (N=62)	SAM (N=27)
Non-synonymous	KPBD	26	34	1	1	0
	Upstream region	42	16	2	1	1
Synonymous	KPBD	38	9	1	0	0
	Upstream region	22	3	1	0	0

Table 4.13: Distribution of *kelch13* mutations stratified by type and locus (i.e. BTB-POZ domain (KPBD) vs. upstream region).

4.3 Presence of *kelch13* mutations in Africa

As we highlighted in the previous section, the most pressing issue for malaria control is to prevent artemisinin resistance emerging in Africa (either by the occurrence of gene flow or due to an independent mutation event). Several studies have reported an increasing number of *kelch13* mutations in African parasites. However, they rendered as sensitive when faced with artemisinin treatment [Ashley et al., 2014], [Sibley, 2014], [Kamau et al., 2014], [Takala-Harrison et al., 2015]. As a consequence, assessment of resistance in Africa cannot depend on genetic markers alone but requires phenotypic data, which in turn makes difficult monitoring the epidemic of artemisinin resistance.

In a follow-up study [MalariaGEN-Pf-Community-Project, 2016], we addressed the questions of (1) what is the frequency and origin of resistance *kelch13* mutations in Africa and (2) what are the selective differences that prompt the emergence and spread of resistance in Southeast Asia but not yet in Africa. In this section, we focus exclusively on the first point, as it relates to my contribution to this piece of work (assessing the origin of the *kelch13* mutations observed in Africa). Nonetheless, it is worthy to note that, with regard to the second question, the study concluded that the major factor halting the independent emergence and spread of artemisinin resistance in Africa, despite its reservoir of resistance alleles, was the lack of strong selective forces.

4.3.1 Data

This analysis included genome data for 3,411 *P. falciparum* samples from 43 different locations in 23 countries (Table 4.14), with 1648 samples from Africa and 1599 samples from Southeast Asia. These data were part of the MalariaGEN *Plasmodium falciparum* Community Project[19] and followed the same procedures we described in Section 4.2.1.2 for

[19] `www.malariagen.net/projects/parasite/pf`

Region	Code	Samples	Country	Code	Samples
West Africa	WAF	957	Burkina Faso	BF	56
			Cameroon	CM	134
			Ghana	GH	478
			The Gambia	GM	73
			Guinea	GN	124
			Mali	ML	87
			Nigeria	NG	5
East Africa	EAF	412	Kenya	KE	52
			Madagascar	MG	18
			Malawi	MW	262
			Tanzania	TZ	68
			Uganda	UG	12
Central Africa	CAF	279	D.R.Congo	CD	279
South America	SAM	27	Colombia	CO	16
			Peru	PE	11
South Asia	SAS	75	Bangladesh	BD	75
West Southeast Asia	WSEA	497	Myanmar	MM	111
			Thailand	TH	386
East Southeast Asia	ESEA	1102	Cambodia	KH	762
			Laos	LA	120
			Vietnam	VN	220
Oceania	OCE	62	Indonesia (Papua)	ID	17
			Papua New Guinea	PG	45

Table 4.14: Origin of the samples used in the study.

genome sequencing, assembly, genotyping and quality filtering. From an initial set of over 4 million SNPs, we retained 935,601 high-quality coding SNPs that could be genotyped with confidence in most samples.

We observed 128 distinct *kelch13* mutations in Africa and 62 in Southeast Asia (Table 4.13), with a total of 155 different SNPs. We found 46 potential resistance alleles[20] of which 26 were present in Africa, 34 in Southeast Asia and 14 were found in both regions (Table 4.13). In particular, seven of these African mutations were associated with artemisinin resistance in Asia, including the most frequent resistant allele, C580Y (see Table 4.15 and Section 4.2.2).

4.3.2 African *kelch13* mutations appear to be indigenous

Having observed a wide repertoire of *kelch13* polymorphisms, we assessed if the most likely origin of these mutations was due to migration of Southeast Asian parasites or if they had emerged independently. A first exploratory genome-wide analysis based on genetic

[20]Non-synonymous mutations in the propeller or BPZ domains.

Mutation	Locus	AFR (N=1648)	SEA (N=1599)	SAS (N=75)	OCE (N=62)	SAM (N=27)	Observed resistance?
D353Y	1725941	-	4	-	-	-	**Yes**
F395Y	1725814	-	1	-	-	-	No
I416V	1725752	1	1	-	-	-	
I416M	1725750	1	-	-	-	-	
K438N	1725684	-	1	-	-	-	No
P441L	1725676	-	27	-	-	-	**Yes**
P443S	1725671	-	1	-	-	-	
F446I	1725662	-	7	-	-	-	**Yes**
G449A	1725652	-	7	-	-	-	**Yes**
S459L	1725622	2	2	-	-	-	
A481V	1725556	-	4	-	-	-	**Yes**
S485N	1725544	-	1	-	-	-	
Y493H	1725521	1	76	-	-	-	**Yes**
V520I	1725440	1	-	-	-	-	
S522C	1725434	2	1	-	-	-	**Yes**
P527H	1725418	1	5	-	-	-	
C532S	1725404	1	-	-	-	-	
V534L	1725398	2	-	-	-	-	
N537I	1725388	1	1	-	-	-	No
G538V	1725385	-	19	-	-	-	**Yes**
R539T	1725382	-	63	-	-	-	**Yes**
I543T	1725370	-	34	-	-	-	**Yes**
P553L	1725340	2	24	-	-	-	**Yes**
A557S	1725329	1	-	-	-	-	
R561H	1725316	1	24	-	-	-	**Yes**
V568G	1725295	-	6	-	-	-	**Yes**
T573S	1725280	2	-	-	-	-	
P574L	1725277	-	12	-	-	-	**Yes**
R575K	1725274	-	3	-	-	-	
A578S	1725266	18	-	-	-	-	No
C580Y	1725259	2	423	-	1	-	**Yes**
D584V	1725247	-	3	-	-	-	**Yes**
V589I	1725233	2	-	-	-	-	
T593S	1725221	1	-	-	-	-	
E612D	1725162	1	-	-	-	-	
Q613E	1725161	5	1	-	-	-	
Q613L	1725160	1	-	-	-	-	No
F614L	1725158	-	1	-	-	-	No
Y630F	1725109	2	1	-	-	-	
V637I	1725089	2	-	-	-	-	
P667A	1724999	-	2	-	-	-	
P667L	1724998	-	2	1	-	-	
F673I	1724981	-	3	-	-	-	**Yes**
A675V	1724974	1	18	-	-	-	**Yes**
A676S	1724972	2	3	-	-	-	
H719N	1724843	1	8	-	-	-	**Yes**

Table 4.15: Non-synonymous mutations found in the *kelch13* propeller and BTB-POZ domains (KPBD). We indicate if the mutation has been previously observed in patients with a prolonged parasite clearance half-life (> 5h) [Miotto et al., 2015], [Ashley et al., 2014].

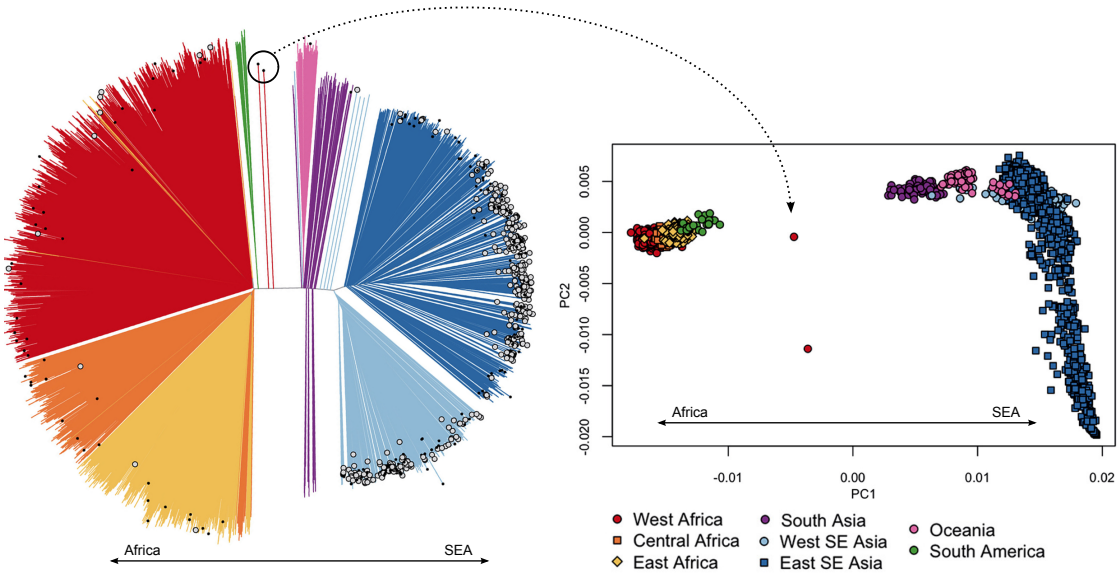

Figure 4.19: Neighbor joining tree (left) and PCoA (right) analysis based on genome-wide genetic distances for all the samples in the study. Samples that are *kelch13* carriers are highlighted in the tree via white circles (homozygous) or black dots (mixed call) at the tips. A clear separation between Africa and SEA samples is evident in both analysis, with only a pair of African samples placed halfway both continental groups.

distances revealed that African and Southeast Asian samples formed two distinct clusters, as can be appreciated in the neighbor joining tree and PCoA (Figure 4.19) results. Despite of this, we noticed that two African samples (carriers of *kelch13* mutations) were placed fairly distant from the African cluster. This observation suggested that parasites from different regions have a characteristic and localized genetic diversity, in agreement with previous studies [Manske et al., 2012], [Miotto et al., 2013]. The fact that we did not find African parasites clustering within the Southeast Asian clade leads to the belief that African *kelch13* mutant parasites are not migrants that originated in Asia.

Nonetheless, it is possible that the current African *kelch13* mutants are the product of a past gene flow event and that the original Asian haplotypes have been assimilated, through recombination, into the African characteristic genome. This might well be the case for the region surrounding *kelch13*. As genome-wide exploratory analyses such as building a neigbor joining tree or performing PCoA do not have the resolution to detect these kind of events, we rely on chromosome painting [Lawson et al., 2012] to reconstruct the most likely source (geographically) of the haplotypes that carry *kelch13* mutants in African parasites. We have discussed chromosome painting several times in this thesis

(Section 1.4.3.2.2) but here we remind the reader of the intuition behind the painting process. The probability of a recipient haplotype (h_r) being painted at locus l by a segment copied from a donor haplotype (h_d), can be interpreted as the probability of h_d being the closest genealogical neighbor of h_r in the underlying genealogy at locus l. As we know the geographical origin of each sample, we can assess the probability of different regions being the source of the haplotypes found in African *kelch13* regions. In this analysis, we prefered chromosome painting to the summary haplotype trees introduced in Section 4.2.5 as the painting provides a probabilistic framework that allows the assessment of uncertainty and can be easily modified to accomodate sparse mixed calls.

We performed chromosome painting across the 250kb flanking regions on each side of the *kelch13* gene[21], for all African and Southeast Asian samples (Figure 4.20). For each sample, the method assigned a posterior copying probability (with respect to all other samples, i.e. unrestricted painting) to the segments defined by consecutive SNPs (chunks). We aggregated these probabilities according to the geographical origin of the donor samples, effectively obtaining a posterior distribution on the putative geographical origin of each chunk. These probabilities were then combined to obtain the expected number of chunks that each sample copied from each geographical region (Figure 4.20, bottom).

In this analysis, we assumed a mutation rate per base per generation of 3.9×10^{-9} [Claessens et al., 2014] and a uniform recombination map. Since the two populations differed substantially in effective recombination rates [Mu et al., 2005], we assumed a recombination rate of 30 kbp/cM. The scaling parameter[22], N_e, was initially set to 10,000 and optimized with an expectation-maximization procedure for 10 iterations [Lawson et al., 2012]. In order to accommodate the presence of sparse mixed calls in samples containing mixed infections, we modified the matrix of emission probabilities in the underlying HMM (Section 1.4.3.2.2, Equation 1.2) by introducing a new parameter epsilon (ϵ) that represents the probability of emitting a mixed call. We also tested several alternatives for estimating the emission probability of each allele once a mixed call is observed, these included using allele frequency at each locus, using the fraction of read counts supporting each allele or simply using the same probability of emission for both alleles. Mixed calls were so sparse that none of

[21] We collapsed all mutants since they present allelic exclusion and masked the *kelch13* gene since it is very conserved across all samples.

[22] Effective population size.

Figure 4.20: (Top) Chromosome painting of the 52 African *kelch13* mutants across the ∼90kb flanking regions of the *kelch13* gene. Each genome chunk is coloured according to the aggregated posterior probabilities that it originated in (i.e. *copied from*) the African (red) or SEA (blue) population. (Bottom) Detail of the flanking regions over a span of approximately 15kb. Country and proportion of African chunks are indicated on the left, the *kelch13* mutation carried is shown on the right.

the alternatives produced significant differences in our analysis. Therefore, we used the simplest formulation. Equation 4.1 describes the stochastic matrix defining the emission probabilities to be used in the underlying HMM, row and column labels indicate genotypes (0: reference, 1: alternative, 2: mixed call), θ corresponds to the miscopying parameter and ϵ to the probability of emitting a mixed call. With this modification, we allowed the presence of mixed calls at the cost of introducing noise in the painting process[23]. We tested how a wide range of values affected the output of our analysis for this particular dataset, finding only negligible differences that did not alter our conclusions[24].

$$\begin{array}{c} \\ 0 \\ 1 \\ 2 \end{array} \begin{array}{c} \begin{array}{ccc} 0 & 1 & 2 \end{array} \\ \begin{pmatrix} (1-\theta)(1-\epsilon) & \theta(1-\epsilon) & \epsilon \\ \theta(1-\epsilon) & (1-\theta)(1-\epsilon) & \epsilon \\ (1-\epsilon)/2 & (1-\epsilon)/2 & \epsilon \end{pmatrix} \end{array} \tag{4.1}$$

The vast majority of African *kelch13* mutants showed no evidence of flanking haplotypes being imported from Southeast Asia, displaying over 90% probability of their origin being local. This suggests that most mutations observed in Africa do not share a common origin with those in Asia and are very likely to have arisen independently on different haplotypic backgrounds.

Only a minority of samples (5 out of 56) exhibited a non marginal probability of having a flanking haplotype of Southeast Asian origin. A finer stratification of the painting results (Figure 4.21) revealed that only two of these samples (carrying a C580Y and Y493H mutations) had a significant probability of possessing an imported haplotype observed in Asian mutants. These results aligned with our previous remark of two African samples clustering apart from the main African cluster (Figure 4.19). A closer inspection of these two samples showed that they were very mixed ($F_{WS} < 0.4$), with a substantial amount of mixed calls randomly distributed across the genome. Such observation would be consistent with the isolate carrying a mix of Asian and African parasites and, therefore, representing the first evidence of a *kelch13* migration between the two regions. However, since these

[23]This does not represent an issue if mixed calls are sparse.

[24]We repeated the whole analysis with different parameter configurations but found the results to be qualitatively robust. In the end, we set epsilon to 10^{-8}.

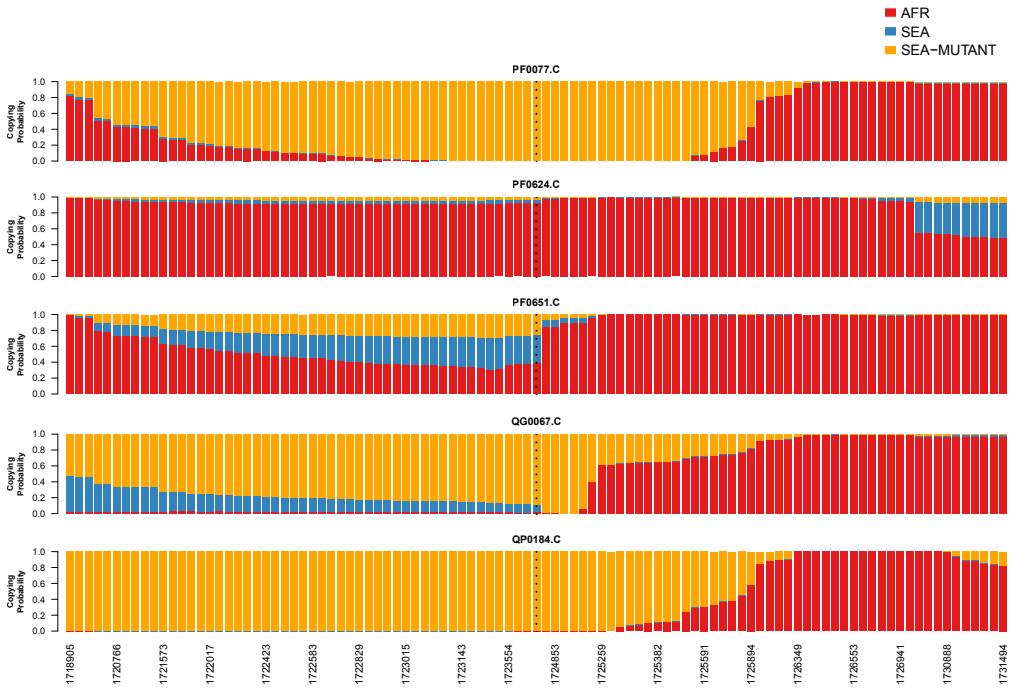

Figure 4.21: Finer analysis of chromosome painting in the 15kb region around the *kelch13* gene (as in Figure 4.20, bottom) for five samples for which *kelch13* was flanked by chunks with $> 50\%$ probability of having Asian origin. The SEA population was divided into two groups: *kelch13* wild type (blue) and mutants (yellow), to estimate the probability that the haplotype originated in SEA *kelch13* mutants. Strong evidence of origin from SEA mutants was only found in two samples: a Y483H mutant from Ghana (top row), and a C580Y mutant from Cameroon (bottom row).

samples have passed through several laboratories and we only observed this phenomenon in two samples, we consider the evidence too weak to support this scenario. We cannot rule out the possibility that this is just an artificial mixture due to contamination occurring during sample preparation and processing[25]. These anecdotal findings would need to be confirmed by re-sequencing the original samples or by assessing the African samples collected during subsequent malaria seasons.

4.3.3 Discussion

We assessed the origin of *kelch13* mutations observed in African samples. We used, first, a genome-wide exploratory data analysis based on performind PCoA and building a NJ tree and, subsequently, a more refined probabilistic examination, based on chromosome painting. Our conclusion is that the reservoir of *kelch13* mutations observed in Africa appears indigenous and, therefore, have most likely emerged as a consequence of independent

[25]We also note that sample contamination could have happened at the Wellcome Trust Sanger Institute facilities, where sample preparation and sequencing was performed by our team.

Mutation	Chr	Pos	AFR	SAS	SEA	PNG	SAM
arps10-V127M	14	2481070	0.0%	0.0%	59.4%	0.0%	0.0%
fd-D193Y	13	748395	0.1%	2.2%	62.8%	23.9%	0.0%
mdr2-T484I	14	1956225	0.1%	5.7%	64.2%	0.4%	0.0%
crt-N326S	7	405362	0.8%	28.2%	68.6%	0.1%	0.0%

Table 4.16: Frequency of the genetic background alleles identified in [Miotto et al., 2015] for each geographical region.

mutational events. However, we remark that two African samples (from Ghana and Cameroon, respectively), carriers of *kelch13* resistance alleles, challenge these conclusions and exhibit a non-negligible probability of possessing haplotypes that originated in Southeast Asia and carried the C580Y and Y493H mutations. Since these observations seem anecdotal (2 out of 54 *kelch13* African mutant samples) and both samples presented compelling evidence of being mixed infections, we are cautious about concluding these are genuine gene flow events instead of an artificial mixture produced in the preparation and processing stages of different laboratories.

These findings have implications for the prevention and control of artemisinin resistance in the African continent. As the presence of *kelch13* mutations in Africa is well documented, resistance could emerge without the necessity of gene flow from Southeast Asia if the selective pressure of artemisinin increases. Several factors might be behind this differential selection pressure. Artemisinin has been extensively used in Southeast Asia for more than twenty years whereas only 20% of African infections are treated with frontline ACTs [WHO, 2014b]. In addition, as we detailed in the introduction (Chapter 1) there are larger numbers of parasites that are not exposed to antimalarial drugs since individuals develop partial immunity that render infections asymptomatic. Furthermore, the emergence or spread of resistance to ACT partner drugs, which is rising in Southeast Asia [Saunders et al., 2014], can also intensify the selective pressure for artemisinin in Africa.

An open question is if there is a specific genetic background that prompts the development of artemisinin resistance, as suggested by our previous study [Miotto et al., 2015]. Table 4.16 displays the geographical allele frequency for the markers of the genetic background associated with resistance *kelch13* alleles, showing that the background is effectively absent from Africa. We conclude that it is plausible that *kelch13* mutations impose a fitness cost

too high to be tolerated if not in the presence of sustained drug pressure. This would explain the limited spread of individual *kelch13* mutations and would also support the case of a genetic background composed of compensatory mutations. Selective pressure can be substantially increased if African countries reach levels of relative low transmission and therefore would be vital to monitor the effects of interventions, in particular tracking the occurrence of founder effects as they seem to play a crucial role in the emergence of resistance.

4.3.4 Individual contributions

In this work, my contribution consisted on assessing the putative origin of the *kelch13* mutations observed in African samples by adapting the chromosome painting model introduced by Lawson and colleagues [Lawson et al., 2012].

Chapter 5

A fast and scalable method for building the ancestral haplotype graph

Contents

5.1 Introduction

As we have seen in previous chapters, demographic inference is of capital importance to inform malaria control programs. In the introduction of this thesis, we commented on how patterns of shared haplotypes can be used to learn about the recent genealogical history of sequences (Section 1.5.1). We exploited that very same idea when assessing the origin of *kelch13* mutations (Chapter 4). The advent of large and very large datasets (see for instance the Pf3k effort, [Pf3k, 2016]) requires the development of scalable methods that can capitalize on haplotype data. In this chapter, we examine how to describe the structure and diversity present in the haplotypes surrounding a locus. An ideal representation should be informative, summarizing the main patterns of interest, and easy to interpret visually, supporting the comparison of different loci. Another desirable characteristic is scalability, permitting its use at genome-wide level. This last requirement demands an efficient generative procedure and also an effective way of combining local representations into a compact high-level description of the genome. To achieve these goals, we propose a tree data structure for summarizing haplotype diversity. We call this construction the nearest neighbor haplotype tree (NNH tree). It can be used as a visualization tool, describing localized haplotype diversity and structure, and also scrutinized at genome-wide scale by comparing the properties of each local tree, characterizing how haplotype composition changes along the genome. NNH trees are intuitively assembled in a bottom-up agglomerative fashion from the pairwise distribution of IBS (or IBD) tracts surrounding a locus. Building on this, we define the ancestral haplotype graph (AHG) as the collection (or forest) of sequential NNH trees that describe the haplotypic structure of each locus in the genome of a set of sequences[1]. Because regular agglomerative methods are too slow[2] for dealing with large data sets, we also propose an alternative building procedure that scales in quasilinear time for the average case, and can generate the AHG for datasets with several thousands of sequences.

This chapter has two major parts. We first introduce a simple bottom-up approach for

[1] We avoided the term ancestral recombination graph (ARG) since our only goal here is to compactly summarize genome-wide haplotype diversity. Nonetheless, the AHG can be interpreted as an approximation of the ARG as it is meant to be very informative about the recent history of a set of sequences and, among other phenomena, the ocurrence of natural selection and specific demographic events.

[2] Time complexity is $O(n^2 \log n)$ for the general case [Maimon and Rokach, 2005]

building NNH trees from the pairwise IBS tract distribution and comment on the merits of such representation (Section 5.2). The rest of the chapter is dedicated to find an equivalent but more efficient building method (Section 5.3). We finish by discussing future directions of research. Concerning my contributions, I claim full autorship of the material presented in this chapter.

5.2 NNH trees

In this section, we describe the initial method we use to compute the pairwise IBS distribution from a set of sequences and how NNH trees can be built on a bottom-up fashion from it. An IBS tract, for a given locus and a set of sequences, refers in this context to the stretch of identical DNA spanning the locus of interest that is shared by two or more sequences, disregarding differences caused by singletons along the IBS segment or by mutations occurring at such locus. Notice that we use the terms *shared haplotype*, *haplotype block* and *IBS tract* as synonymous. Likewise, we sometimes refer to sequences as *samples*. We define the pairwise IBS distribution for a given locus (also termed core locus, core mutation or focal mutation) as the collection of IBS tracts associated with each pair of sequences. We assume all polymorphisms are biallelic and we encode genotypes using 0/1 for the reference/alternative[3] alleles.

5.2.1 Computing the pairwise IBS distribution

To compute the pairwise IBS distribution, the procedure starts at the locus of interest (l) and extends the IBS tract on both flanks, skipping singletons, until finding a genotype mismatch (breakpoint) between each pair of sequences. We store the coordinates of the breakpoints at both flanks and compute the physical IBS tract length. We omit singletons, as they are more likely to be genotyping errors or recent mutations on terminal branches of the underlying genealogy that generated the observed haplotype pattern. Figure 5.1 depicts this procedure for an arbitrary pair of sequences.

For a given locus, we obtain three matrices representing the pairwise distributions of flank breakpoints, IBS tract lengths, and number of mismatches within IBS tracts. Due

[3]Or the ancestral/derived alleles when the allelic status is known.

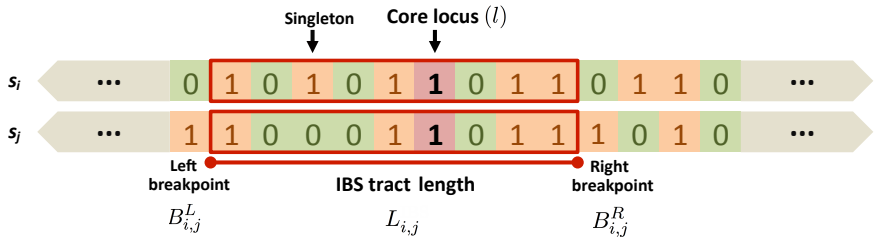

Figure 5.1: Diagram illustrating the procedure we use to detect IBS tracts for a pair of sequences, highlighting the notation we use in this chapter. We encode genotypes as 0/1 for the reference (green) and alternative (orange) alleles. The locus of interest appears with a red background.

to computational reasons, the lower and upper triangular parts of these matrices refer, respectively, to the left and right flanks surrounding the locus, with the diagonal values having no meaning. However, to simplify our notation we will assume that B^L and B^R are symmetric matrices containing the pairwise distribution of left and right breakpoints, that L is a symmetric matrix representing the pairwise distribution of IBS tract lengths (notice that $L = B^R - B^L - 1$), and that M is also a symmetric matrix that stores the number of differences or mutations present in the IBS tracts shared by each pair of samples. The shared haplotype for sequences i and j would then correspond to the loci interval $(B^L_{i,j}, B^R_{i,j})$, have length $L_{i,j}$ and contain $M_{i,j}$ mismatches.

We stress that, by default, we do not take into account genotype differences at the locus of interest when extending IBS tracts because the method seeks to characterize the haplotypic background where mutations arise[4]

5.2.2 Hierarchical relationship between IBS tracts

Although the pairwise distribution of IBS tract lengths depicts the underlying haplotypic diversity surrounding a locus, direct graphical representations of L, such as heatmaps or histograms, are difficult to interpret and do not provide an intuitive outlook of the relationship among sequences. Alternative graph representations[5] are even harder to assess visually but for trivial cases. We argue that the hierarchical nature of IBS tracts makes trees a very intuitive and informative representation for these type of data. Given the

[4]We can easily accomodate a strict infinite sites model by forcing all sequences with the same genotype at the locus of interest to coalesce with each other before joining other sequences at the root of the tree, see Section 5.2.7.

[5]Graphs where samples are represented by nodes and the thickness of connecting edges are proportional to the IBS tract lengths.

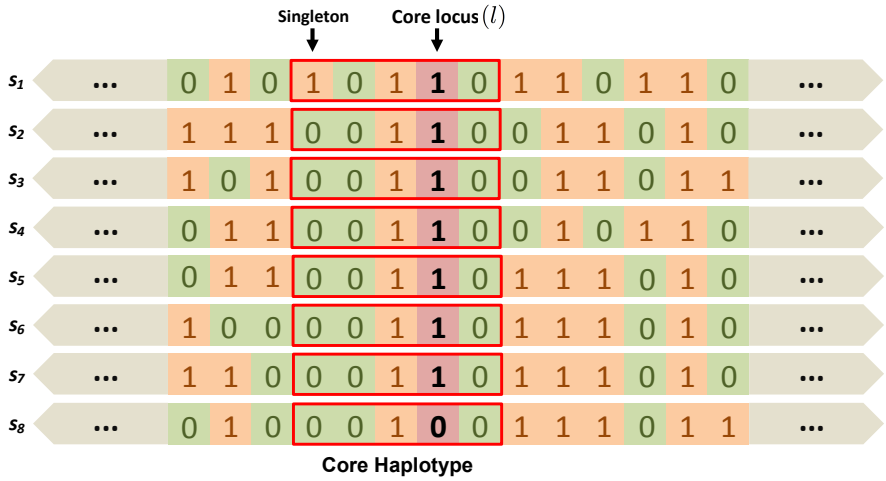

Figure 5.2: Diagram showing the core (IBS) haplotype for a set of five sequences, notice that singletons and differences at the locus of interest are discarded when calling breakpoints.

pairwise distribution of IBS breakpoints, we can compute the left breakpoint (b^L) of the IBS tract shared by three arbitrary sequences as

$$b^L_{\{i,j,k\}} = \max\big(B^L_{i,j}, B^L_{i,k}, B^L_{j,k}\big), \tag{5.1}$$

where i, j, k represent sequence indices. Likewise, the right breakpoint (b^R) is given by

$$b^R_{\{i,j,k\}} = \min\big(B^R_{i,j}, B^R_{i,k}, B^R_{j,k}\big). \tag{5.2}$$

We can extend this reasoning to an arbitrary number of sequences in a set (Q) by following

$$b^L_Q = \max\big(B^L_{i,j}\big), \qquad \forall\, i, j \in Q, i \neq j, \tag{5.3}$$

$$b^R_Q = \min\big(B^R_{i,j}\big), \qquad \forall\, i, j \in Q, i \neq j. \tag{5.4}$$

This property induces a hierarchical relationship between IBS tracts. Let H_Q be the IBS tract (or haplotype) shared among all sequencies in the set Q for a particular locus. It follows from Equations 5.3 and 5.4 that $H_Q \sqsubseteq H_{Q\setminus\{k\}}$, meaning that the IBS tract shared by all sequences in Q is a subhaplotype of, or equal to, the shared haplotype obtained when sequence k is removed. This means that H_Q is a substring[6] of $H_{Q\setminus\{k\}}$ that spans the core

[6]Discarding singletons and differences at the core locus.

locus. In this setting it is easy to define concepts such as the core haplotype for a group of samples (i.e. maximal IBS tract that is shared by all sequences, see Figure 5.2).

In genealogical terms, we can think of the IBS tract shared by a set of sequences as rough proxies for the ancestral haplotype they shared in the past. In this regard, the better the IBS tracts approximate the true IBD tracts, the more accurate these ancestral haplotypes will be.

5.2.3 An agglomerative algorithm for building NNH trees

We propose a deterministic agglomerative algorithm (Algorithm 1), similar in spirit to hierarchical clustering [Johnson, 1967], for building a binary tree that summarizes the pairwise IBS distribution. The algorithm takes as input the number of sequences (n), both breakpoint matrices (B^L,B^R), the matrix of IBS tract lengths (L), the matrix of singletons/mutations within IBS tracts (M), and a collection of lists with the singleton coordinates of each sequence (S^*). At each step, we create an ancestor (i.e. inner node of the tree) by merging the pair of sequences/ancestors that share the longest IBS tract. We update the breakpoint matrices, regarding the newly created ancestor, by using Equations 5.1 and 5.2. The lengths of IBS tracts are also updated using $L = B^R - B^L - 1$. To resolve ties, we rely on the number of mutations within IBS tracts, choosing the segment with fewer mutations[7]. Because M refers to the count of singletons within an IBS tract, we set all counts to zero when relating any two ancestors as we assume these mutations always occur on the terminal branches of the underlying genealogy. The number of singletons between an ancestor and any extant sequence can be computed by counting the number of mutations within their IBS tract[8]. Each ancestral node is annotated with the length and breakpoint coordinates of the IBS tract shared by all the sequences that are its descendants. Figure 5.3 illustrates each coalescing step performed by the algorithm when executed on the example data presented in Figure 5.2. Figure 5.4 shows the final tree in terms of shared haplotypes.

The algorithm returns an annotated binary tree that has the core haplotype at the root, and whose inner nodes represent shared IBS tracts for different subsets of sequences.

[7]When a tie cannot be resolved in this fashion, we choose tha pair that contains the sequence with the lowest index.

[8]This can be done efficiently by saving the position of the singletons present in each sequence and performing binary search when required.

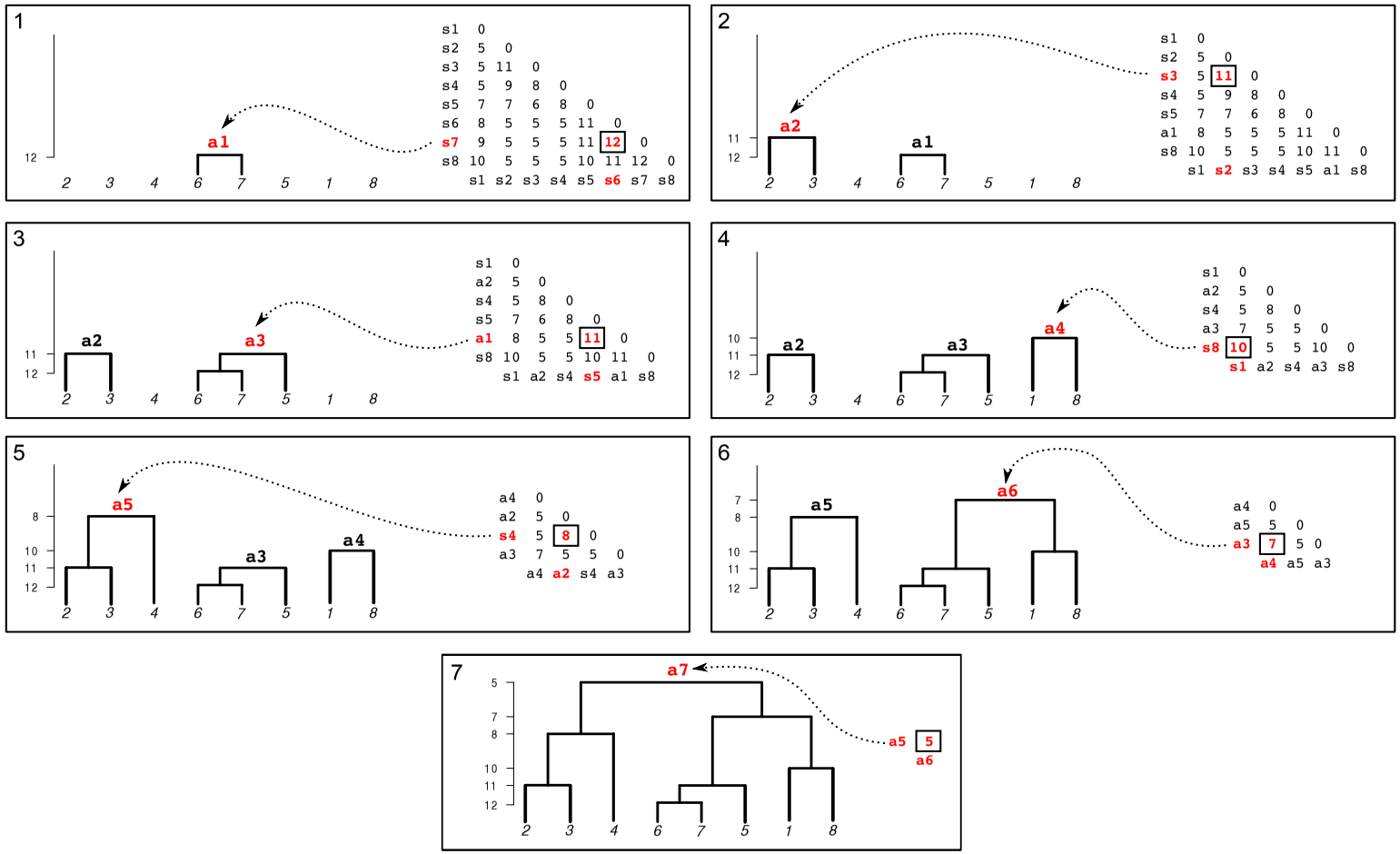

Figure 5.3: Diagram showing a trace of the NNH algorithm when executed on the set of sequences presented in Figure 5.2. To improve readability, we only show how the matrix L is updated, replacing any coalescing pair with its ancestral node. Singletons have been discarded for simplicity. The y-axis shows the IBS length (we assume marker loci are consecutively numbered) of each ancestral haplotype. The pair chosen to coalesce is highlighted in red. We refer to sequences (i.e. leaves) with the s prefix, while any ancestral node has a as prefix. Ancestral nodes are numbered sequentially following the order in which they are created (from a_1 to a_{n-1}).

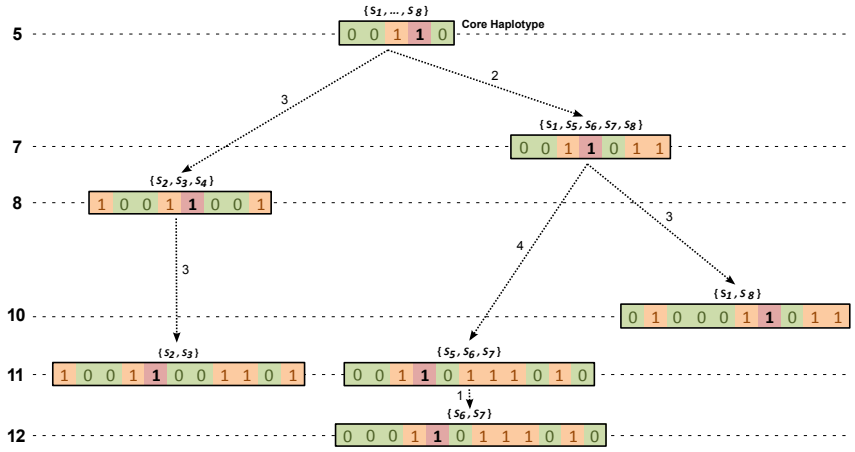

Figure 5.4: An alternative view of the tree built in Figure 5.3, this time showing the shared haplotypes. Notice again that, for simplicity, we have discarded singletons in this example. We have placed the difference in IBS tract length of ancestral haplotypes over the branches. We encode genotypes as 0/1 for the reference (green) and alternative (orange) alleles. The locus of interest appears with a red background.

We encode the differences in physical length of these IBS tracts on the branch lengths of the tree, which define the height of all inner nodes (i.e. the length of the IBS tracts they represent). This annotated data structure can be used as a graphical summary or can be further analyzed for other purposes (e.g. computing distances between trees, finding the tree that best represent a given gene or scanning how tree balancedness changes along the genome). We have branded this representation as the nearest neighbor haplotype tree (NNH tree) because, in the merging step, we choose the pair of haplotypes that are more likely to be closest genealogical neighbors.

5.2.4 Visual representation and scale

NNH trees are not ultrametric. Because they summarize the distribution of shared IBS tracts, by convention, leaves are connected to inner nodes by branches of length zero. However, to make the trees easier to visualize, we prefer to render them as ultrametric, placing all leaves at the same height as the longest shared haplotype tract. This enhancement provides an easier visualization and does not alter the interpretation of the tree, as only inner nodes have meaning. Visualization is also improved by transforming the scale of the

Algorithm 1 Builds an NNH tree from the pairwise IBS tract distribution.

Input: n, B^L, B^R, L, M, S^*

Output: T ▷ Start with an empty tree.

$T = \emptyset$

while $n > 1$ **do**

  $c = \big\{\{i,j\} \in \max_{i,j}(L)\big\}$ ▷ Longest IBS pair(s).

  if $|c| > 1$ **then**

    $c = c \cap \big\{\{i,j\} \in \min_{i,j}\big(M\big)\big\}$ ▷ Pair with least mutations.

  end if

  $\texttt{pi} \leftarrow c_{1,1}$

  $\texttt{pj} \leftarrow c_{1,2}$

  $\texttt{bl} \leftarrow B^L_{\texttt{pi},\texttt{pj}}$ ▷ Ancestor left breakpoint.

  $\texttt{br} \leftarrow B^R_{\texttt{pi},\texttt{pj}}$ ▷ Ancestor right breakpoint.

  $\texttt{l} \leftarrow \texttt{br} - \texttt{bl} - 1$ ▷ IBS tract length.

  $\texttt{m} \leftarrow M_{\texttt{pi},\texttt{pj}}$ ▷ Mutations.

  $\texttt{children} = \{\texttt{pi}, \texttt{pj}\}$ ▷ Children.

  $T = T \cup \{\texttt{bl}, \texttt{br}, \texttt{l}, \texttt{m}, \texttt{children}\}$ ▷ Add ancestral node to tree.

  for $k = 1$ **to** n **do** ▷ Update matrices (ancestor replaces `pi`).

    $B^L_{\texttt{pi},k} \leftarrow B^L_{k,\texttt{pi}} \leftarrow \max\big(B^L_{\texttt{pi},k}, B^L_{\texttt{pj},k}, \texttt{bl}\big)$

    $B^L_{\texttt{pj},k} \leftarrow B^L_{k,\texttt{pj}} \leftarrow -\infty$

    $B^R_{\texttt{pi},k} \leftarrow B^R_{k,\texttt{pi}} \leftarrow \min\big(B^R_{\texttt{pi},k}, B^R_{\texttt{pj},k}, \texttt{br}\big)$

    $B^R_{\texttt{pj},k} \leftarrow B^R_{k,\texttt{pj}} \leftarrow \infty$

    $L_{\texttt{pi},k} \leftarrow L^{\text{IBS}}_{k,\texttt{pi}} \leftarrow B^R_{\texttt{pi},k} - B^L_{\texttt{pi},k}$

    $L^{\text{IBS}}_{\texttt{pj},k} \leftarrow L_{k,\texttt{pj}} \leftarrow -\infty$

    $M_{\texttt{pj},k} \leftarrow M_{k,\texttt{pj}} \leftarrow \infty$

    $M_{\texttt{pi},k} \leftarrow M_{k,\texttt{pi}} \leftarrow \text{Singletons}\big(B^L_{\texttt{pi},k}, B^R_{\texttt{pi},k}, \texttt{pi}, k, S^*\big)$ ▷ Compute singletons

  end for

  $n \leftarrow n - 1$

end while

tree. The scale of a NNH tree always correspond to the interval

$$\big[\min_{i,j}(L),\ \max_{i,j}(L)\big], \qquad \forall\, i,j \in S, i \neq j,$$

where S represents the set of all sequences. The two extremes are given by the length of the core and longest shared haplotype, respectively. Let $l_{a\dashrightarrow b}$ be the length of the path connecting two arbitrary inner nodes a and b, with a being and ancestor of b, and whose corresponding IBS tracts differ in $\Delta l_{b,a}^{\text{IBS}}$ base pairs[9]. Here we define the following tree scales

- Reciprocal:

$$l_{a\dashrightarrow b} = 1/\Delta l_{b,a}^{\text{IBS}}. \tag{5.5}$$

- Reciprocal-log:

$$l_{a\dashrightarrow b} = 1/\log(\Delta l_{b,a}^{\text{IBS}}). \tag{5.6}$$

- Log-linear:

$$l_{a\dashrightarrow b} = \log(\Delta l_{b,a}^{\text{IBS}}). \tag{5.7}$$

- Linear (original):

$$l_{a\dashrightarrow b} = \Delta l_{b,a}^{\text{IBS}}. \tag{5.8}$$

Figure 5.5 compares the four scales for an NNH tree built over 50 simulated sequences. The reciprocal scale (bottom-left panel), where branch lengths are inversely proportional to the difference in length of shared haplotypes, may be the most familiar to researchers because of its resemblance to coalescent trees. This interpretation is helped by the fact that, if IBS tracts are considered as a proxy for IBD segments, sequences sharing longer IBS tracts are more likely to have a more recent common ancestor than sequences sharing shorter ones, translating this scale into relative coalescing times[10]. Nonetheless, we find the reciprocal-log and log-linear scales to be the most useful for visualization purposes as they reveal the internal structure of the clades, even when sequences share a very well conserved haplotypic background.

[9] Notice that the minimum value for $\Delta l_{b,a}^{\text{IBS}}$ is 1.

[10] We stress here that our goal here is only to describe the haplotypic structure of the data and not to infer the underlying genealogy although, again, NNH trees can be regarded as a very crude approximation of the genealogy

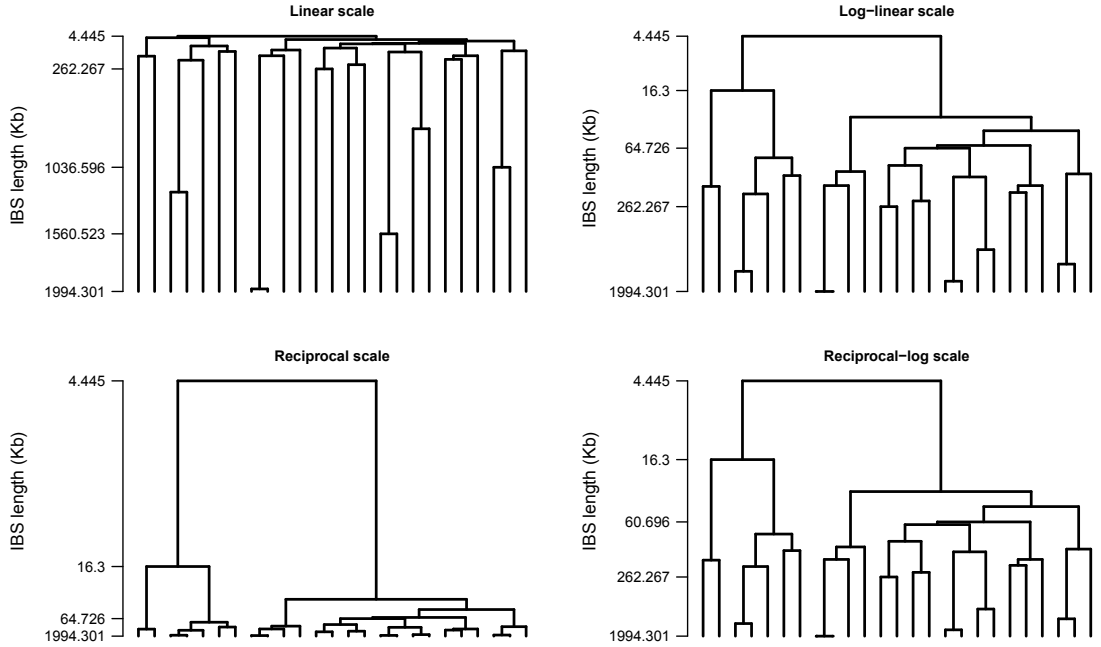

Figure 5.5: Figure showing different scales for the same NNH tree. Top-left: linear, top-right: log-linear, bottom-left: reciprocal and bottom-right: reciprocal-log. NNH tree built for the median locus of a chromosome (50 sequences) simulated under the neutral coalescent with the SCRM package [Staab et al., 2015], using $\mu = 1 \times 10^{-7}$, $r = 1 \times 10^{-8}$, $N_e = 50000$.

5.2.5 Bounds for the length of shared haplotypes on NNH trees

We can consider the process of building an NNH tree as a kind of dimensionality reduction on the pairwise IBS length distribution. We transform $\binom{n}{2}$ data points, where n is the number of sequences, into the $n-1$ inner nodes of a binary tree. Although we cannot recover perfectly the initial pairwise IBS length distribution by navigating the tree, one of the nice properties of this data structure is that we can easily derive bounds for the length of IBS tracts shared by any pair of sequences.

Given an arbitrary pair of sequences, (i, j), the NNH tree provides their exact IBS tract length ($L_{i,j}$) when they coalesce with each other before coalescing with any other sequence. When they join other sequences first, the IBS tract length associated with their MRCA ($l_{\text{MRCA}_{i,j}}$) serves as a lower bound for $L_{i,j}$ (denoted here as $\downarrow L_{i,j}$). We have

$$\downarrow L_{i,j} = l_{\text{MRCA}_{i,j}}. \tag{5.9}$$

This follows from the property $H_Q \subseteq H_{Q\setminus\{k\}}$ (see Section 5.2.2). Thus, any ancestor of this

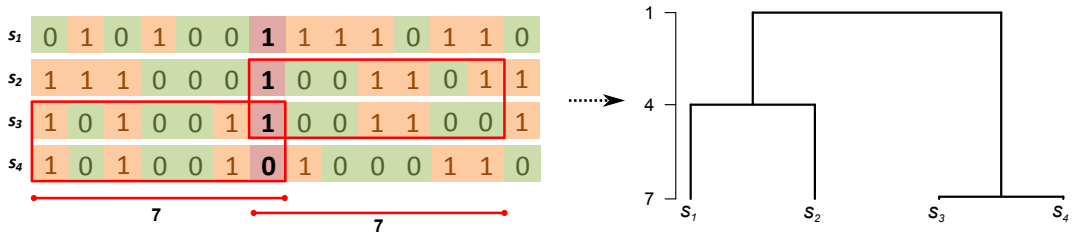

Figure 5.6: Figure illustrating the reasoning behind the computation of $\uparrow L_{i,j}$, see main text for details.

MRCA will exhibit a shorter (or at most equal) IBS tract length. As an extreme example, the lower IBS tract length bound for a pair of sequences that find their MRCA at the root of an NNH tree will be given by the length of the core haplotype.

Likewise, the upper bound for the IBS tract length of an arbitrary pair of sequences ($\uparrow L_{i,j}$) is given by the length of the longest ancestral haplotype associated with i or j within the subtree whose root is $\text{MRCA}_{i,j}$. Let $c^1_{i,*}$ and $c^1_{j,*}$ be the tree nodes where i and j first coalesce with other sequence(s). We have

$$\uparrow L_{i,j} = \max\big(\, l_{c^1_{i,*}},\, l_{c^1_{j,*}} \big). \tag{5.10}$$

To develop the intuition behind the upper bound, consider the set of sequences presented in Figure 5.6. In this example there is a tie for the longest shared haplotype, this is $L_{s_2,s_3} = L_{s_3,s_4} = 7$. Assume the algorithm chooses, arbitrarily, the pair (s_3, s_4) to resolve the tie. This coalescing event makes s_2 the nearest neighbour of s_1, since now $L_{s_1,s_2} > L_{s_2,a_1}$, where $a_1 = \{s_3, s_4\}$. As a result, the bounds for L_{s_2,s_3} that can be derived from the tree are $[1, 7]$. The fact that we lost any information about tie resolution is what motivates Equation 5.10. When the algorithm chooses to merge sequence i with another sequence (or ancestral node) k instead of j, the only thing that we can conlude is that $L_{i,k} \geq L_{i,j}$.

The interval $[\downarrow L_{i,j}, \uparrow L_{i,j}]$ gets very wide when considering sequences from distant clades. However, the true tract lengths are not uniformly distributed within this interval. The lower bound is, in general, a better estimator for $L_{i,j}$ than the midpoint of the interval, $(\downarrow L_{i,j} + \uparrow L_{i,j})/2$. The reason is that, unless faced with cases of very unbalanced flanks like the example in Figure 5.6, NNH trees usually provide a very good clustering for shared haplotype lengths. This renders the ancestral haplotype of the MRCA of a set of

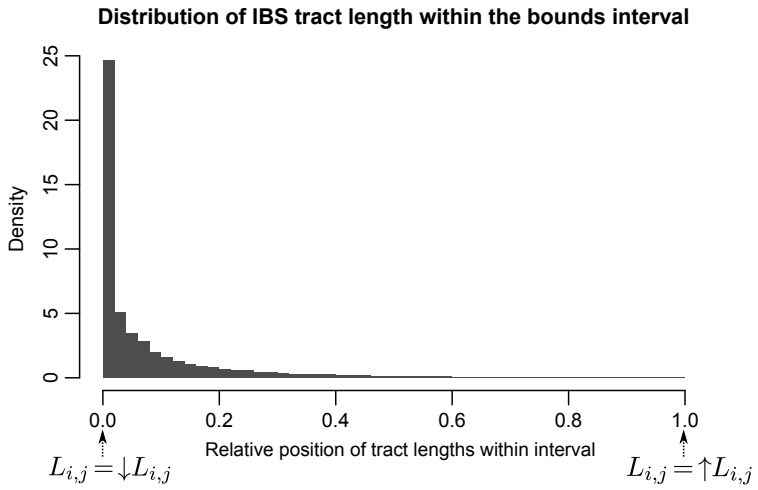

Figure 5.7: Distribution of the relative placement of the actual IBS tract length within the bounds interval, i.e. $(L_{i,j} - \downarrow L_{i,j})/(\uparrow L_{i,j} - \downarrow L_{i,j})$, for 1000 NNH trees generated over 25 sequences (randomly sampled loci). Sequences simulated under the neutral coalescent with the SCRM package [Staab et al., 2015], using $\mu = 1 \times 10^{-7}$, $r = 1 \times 10^{-8}$, $N_e = 50000$.

extant sequences as a very good proxy for the haplotype they share. Figure 5.7 shows the distribution of $L_{i,j}$ within the bounds interval for 1000 NNH trees, with the vast majority of IBS tract lengths being very close to the lower bound. In light of this, we recommend the following estimator for reconstructing the pairwise IBS length distribution from an NNH tree

$$\widehat{L_{i,j}} = l_{\text{MRCA}_{i,j}}. \tag{5.11}$$

We finish by noting that Equations 5.9 and 5.10 can be easily extended to accomodate an arbitrary number of sequences.

5.2.6 NNH trees as data summary

Following on our previous dicussion, here we evaluate how well NNH trees work as summary of the pairwise IBS tract length distribution. To this end, we compare NNH trees with the widely used neighbor joining (NJ) trees [Saitou and Nei, 1987], that tend to work well even when distances are not additive [Felsenstein, 2004].

We rely on the sum of squared errors to assess how good is each type of tree as a summary of the data, $\text{RSS} = \sum_{i,j} w_{i,j}(L_{i,j} - \widehat{L_{i,j}})^2$. This is similar to the way goodness of fit is usually measure for distance-based phylogenetic methods [Lemey, 2009], we set $w_{i,j} = 1$ so we penalize equally deviations on pairs sharing short or long IBS tracts. To

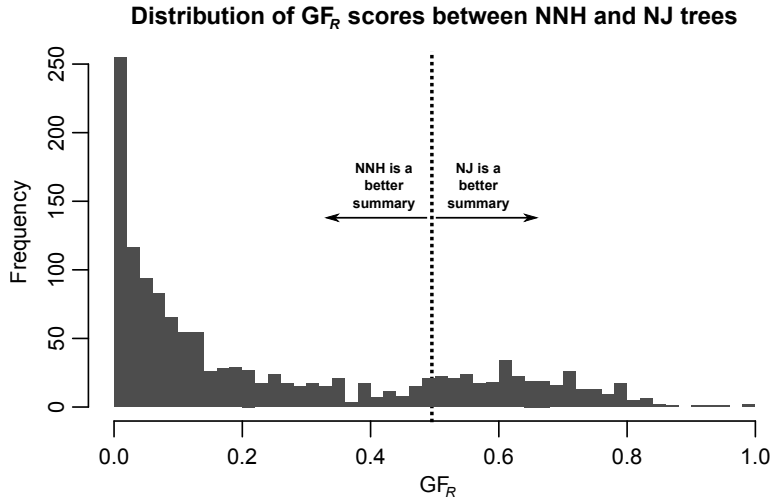

Figure 5.8: Distribution of GF_R comparing 1000 NNH and NJ trees (same data as in Figure5.7). A value of $GF_R = 0.5$ indicates that both trees summarize the pairwise IBS tract length distribution equally well, below this value NNH trees produce a smaller RSS than NJ. We obtained an average GF_R score of 0.24.

build NJ trees we use $d = 1/L$ as distance matrix and the reciprocal of the pairwise patristic distance[11] as $\widehat{L_{i,j}}$. For NNH trees we employ Equation 5.11. Since we are only interested in the relative merits of each tree method, we compute a relative goodness of fit (GF_R) score given by

$$GF_R = \frac{\mathrm{RSS_{NNH}}}{\mathrm{RSS_{NNH}} + \mathrm{RSS_{NJ}}}, \tag{5.12}$$

where $\mathrm{RSS_{NNH}}$ and $\mathrm{RSS_{NJ}}$ are, respectively, the sum of squared errors for the NNH and NJ trees. Figure 5.8 shows the results of this analysis, indicating than NNH trees to perform better than NJ trees as summaries of the pairwise IBS length distribution. Besides, NJ trees are difficult to interpret in terms of shared haploytes and can have branches with negative length. In Figure 5.9 we compare an NNH tree with its NJ equivalent.

5.2.7 Imposing a strict infinite sites model

If NNH are interpreted as genealogical histories, they allow the occurrence of recurrent mutations at the locus of interest. Nonehteless, it is easy to impose a strict infinite sites model in which a derived mutation arises only once and back-mutations are not allowed. To do this, we stratify the sequences by the allelic status of the core locus. Next, for each of these two subsets of sequences (remember that our data is biallelic), we build the

[11]Length of the path (i.e. sum of branch lengths) between two leaves in the tree.

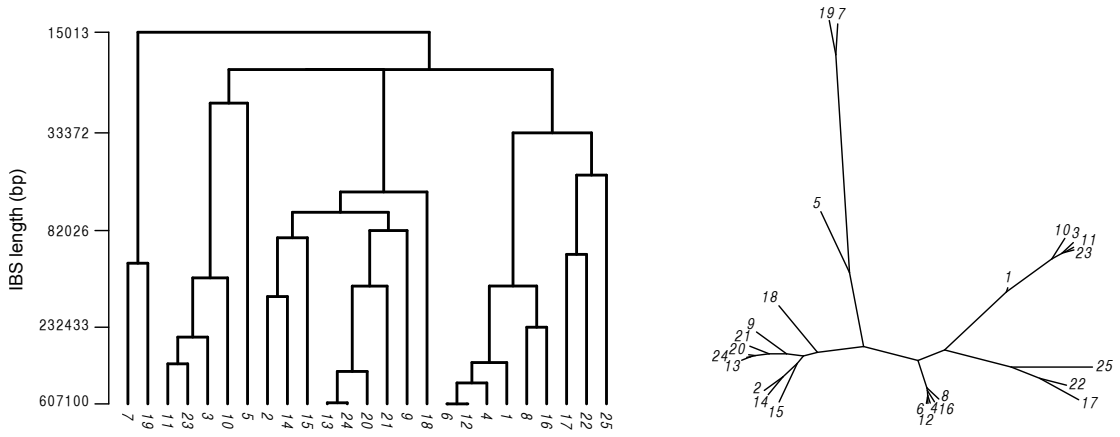

Figure 5.9: Comparison of an NNH tree (left panel) with its equivalent NJ tree (right panel) over 25 sequences simulated under the neutral coalescent (same generation procedure as in Figure 5.7.

corresponding NNH tree. Finally, we join both trees by coalescing their roots into the core haplotype for all sequences. NNH trees built in such manner present always two clades that exclusively contain sequences with the same allele at core locus.

5.2.8 Overall time complexity

The heuristic method for detecting IBS tracts has a time complexity of $O(n^2\, L_m)$ in terms of genotype comparisons, where n refers to the number of samples and L_m to the number of genetic markers of the longest shared haplotype. An additional scan of the region (i.e. chromosome) is needed for identifying singletons, but this has no real impact on performance, as it does not need to be recomputed for each locus.

Like the complete-link clustering algorithm, the tree building process has a time complexity of $O(n^2 \log n)$ [Day and Edelsbrunner, 1984] when a priority queue is used for tracking merging candidates. Further reduction of time complexity can be achieved by using approximated search strategies, trading computation speed for accuracy when dealing with very large sample sizes. The fact that each locus can be processed independently makes the task of constructing the AHG (i.e. building an NNH tree for each locus) embarrassingly parallel and, therefore, scalable when clusters with many CPUs are available.

These methods are able to cope with moderate datasets (hundreds of sequences) but would require large computational resources for genome-wide use in large or very large datasets (in the order of thousands of sequences).

5.3 A fast and scalable method to build the AHG

After having introduced NNH trees and some of their merits, we dedicate the rest of the chapter to develop an equivalent but more efficient tree building method, making feasible the processing of large datasets. We start by studying the subhaplotype graph (SHG) associated with a locus and how the agglomerative NNH algorithm implicitly explores and trims this graph. Next, we show that NNH trees can be constructed with a simple procedure from a specific topological order of the nodes of the SHG. Because the efficiency of this method depends on the order (i.e. the number of nodes) of the SHG, we find an upper bound and evaluate, via simulations, the order of the SHG for the average case. To do this, we study how the SHG of a locus is related to the subhaplotype graphs of its flanks and prove that the set of nodes of the SHG can be assembled from the NNH trees of the left and right flanks. Finally, we provide a very efficient method for generating NNH flank trees via the positional Burrows-Wheeler transform (PBWT) [Durbin, 2014] that scales linearly with the number of sequences and markers. The overall time complexity of this alternative approach renders as quasilinear on n (the number of sequences) for the average case, when set intersections are considered elementary operations.

5.3.1 The subhaplotype graph of a locus

In this section, we introduce the subhaplotype graph of a locus (SHG_l). The SHG_l has as nodes all maximal shared haplotypes (i.e. IBS tract blocks) that span the locus of interest (l). In this graph, there exists a directed edge between any pair of nodes whenever they satisfy the binary relation *is-a-subhaplotype-of* (denoted by $\sqsubseteq$). From now on, we drop the l symbol noting that we always refer to the SHG of a particular locus in the genome. Each node x in the SHG is associated with a haplotype (h_x) and the set of sequences that share that haplotype (s_x). We denote as H the set of haplotypes in the SHG, whereas S refers to the collection of sequence sets. The *is-a-subhaplotype-of* relation induces a partial order over H since $\sqsubseteq$ is reflexive, antisymmetric and transitive. In effect, $(\sqsubseteq, H)$ is a partially ordered set (or poset). This fact guarantees that the SHG does not contain any cycle (i.e. it is a directed acyclic graph or DAG). If instead of H we consider S together with the *is-a-superset-of* relation ($\supseteq$), we also obtain a partially ordered set over the sets

of sequences.

Lemma 5.3.1. *The partial orders defined by* $(\supseteq, S)$ *and* $(\sqsubseteq, H)$ *are equivalent on the SHG.*

Proof. For the sake of contradiction, assume that there is a pair of nodes (a, b) in the SHG such that $h_a \sqsubseteq h_b \wedge s_a \not\supseteq s_b$. From $s_a \not\supseteq s_b$, it follows that

$$(s_a \cap s_b) \subseteq (s_b \setminus \{k\}), \qquad k \in s_b \wedge k \notin s_a.$$

Let $b^L_{h_a}$ and $b^R_{h_a}$ be, respectively, the left and right breakpoints of h_a. Because $h_a \sqsubseteq h_b$ we know that the segment $(b^L_{h_a}, b^R_{h_a})$ is IBS for any sequence $z \in (s_a \cup s_b)$ and $s_a = s_a \cup \{z\}$ holds. Without loss of generality, let $z = k$. We have that

$$(s_a \cup \{k\}) \cap s_b \subseteq (s_b \setminus \{k\}), \qquad k \in s_b \wedge k \notin s_a,$$

which leads to

$$(s_a \cap s_b) \cup \{k\} \subseteq (s_b \setminus \{k\}),$$

that is a contradiction since $k \neq \emptyset$ by definition.

For the converse direction, assume now that there is a pair of nodes (a, b) in the SHG such that $h_a \not\sqsubseteq h_b \wedge s_a \supseteq s_b$. Because $h_a \not\sqsubseteq h_b$, we know that h_a and h_b must have, at least, an exclusive sequence not present in each other. Let x and y two of these exclusive sequences for h_a and h_b, respectively. We have that

$$(s_a \cap s_b) \cap \{x, y\} = \emptyset, \qquad x \in s_a,\ y \in s_b.$$

However, from the second clause we derive $s_a \cap s_b = s_b$. Substituting we obtain

$$s_b \cap \{x, y\} = \emptyset, \qquad x \in s_a,\ y \in s_b.$$

which is clearly a contradiction since, by definition, $s_b \neq \emptyset$. □

The core haplotype (h_c) and its associated set of sequences (s_c) are the least[12] elements of the posets defined by $(\sqsubseteq, H)$ and $(\supseteq, S)$, respectively. Figure 5.10 shows the SHG for

[12] l is the least element in a poset $(\leq, P)$ if for every element $e \in P$, $l \leq e$.

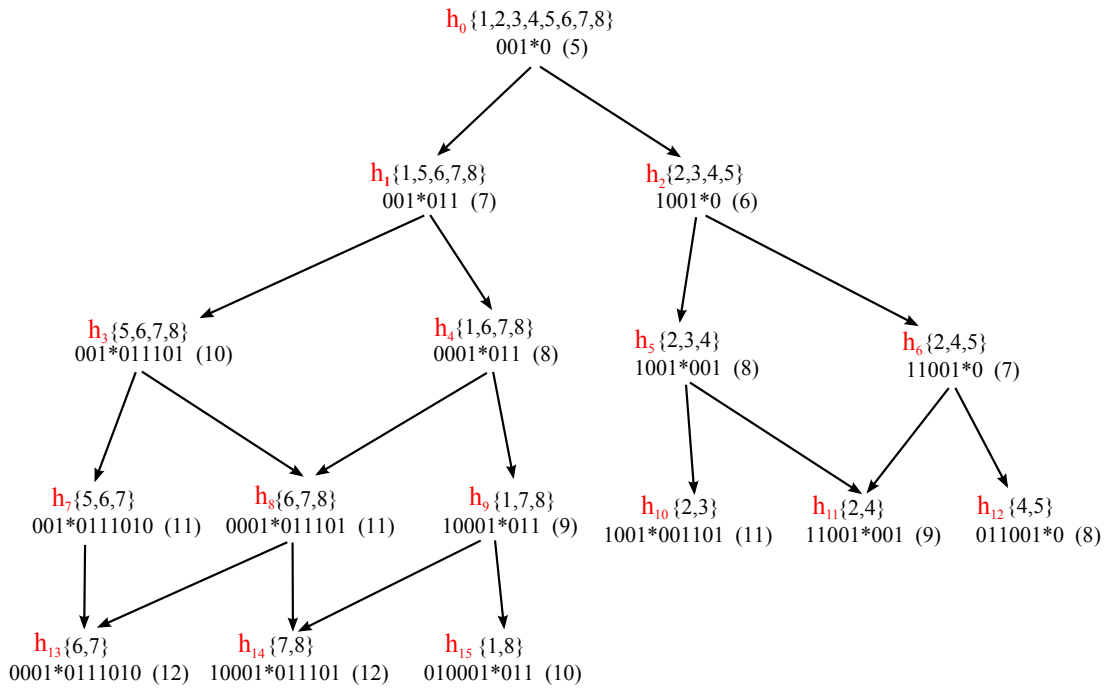

Figure 5.10: Transitive reduction of the subhaplotype graph (SHG) for the data shown in Figure 5.2. Each node has been labeled with its haplotype (core locus masked with a star), set of associated sequences (on top) and tract length (between curly brackets). h_0 represents the core haplotype. Edges represent the two equivalent binary relations defined in the main text ($\sqsubseteq$ and $\supseteq$).

the sequences presented in Figure 5.2. Notice that, for compactness, we always represent the SHG as its transitive reduction (as is typical for posets). The transitive reduction of a directed graph is another graph that contains the same set of nodes but the smallest subset of edges that maintain the same reachability relation of the original graph [Aho et al., 1972].

5.3.2 The NNH algorithm explores a subgraph of the SHG

We show now how the NNH tree building algorithm (Algorithm 1) implicitly explores a subgraph of the SHG. Instead of providing a rigurous proof, we develop the intuition of how the algorithm works by tracing the NNH tree construction for the data presented in Figure 5.2, and mapping the algorithm actions into operations performed on the SHG (see Figure 5.11). Because the algorithm implicitly trims the SHG during its execution, we always refer to the *current* SHG, meaning the updated version of the graph.

The NNH algorithm builds the tree in a bottom-up greedy fashion, it considers the longest IBS tract at each step. In line 3 of Algorithm 1 ($c = \{\{i, j\} \in \max_{i,j}(L)\}$) we

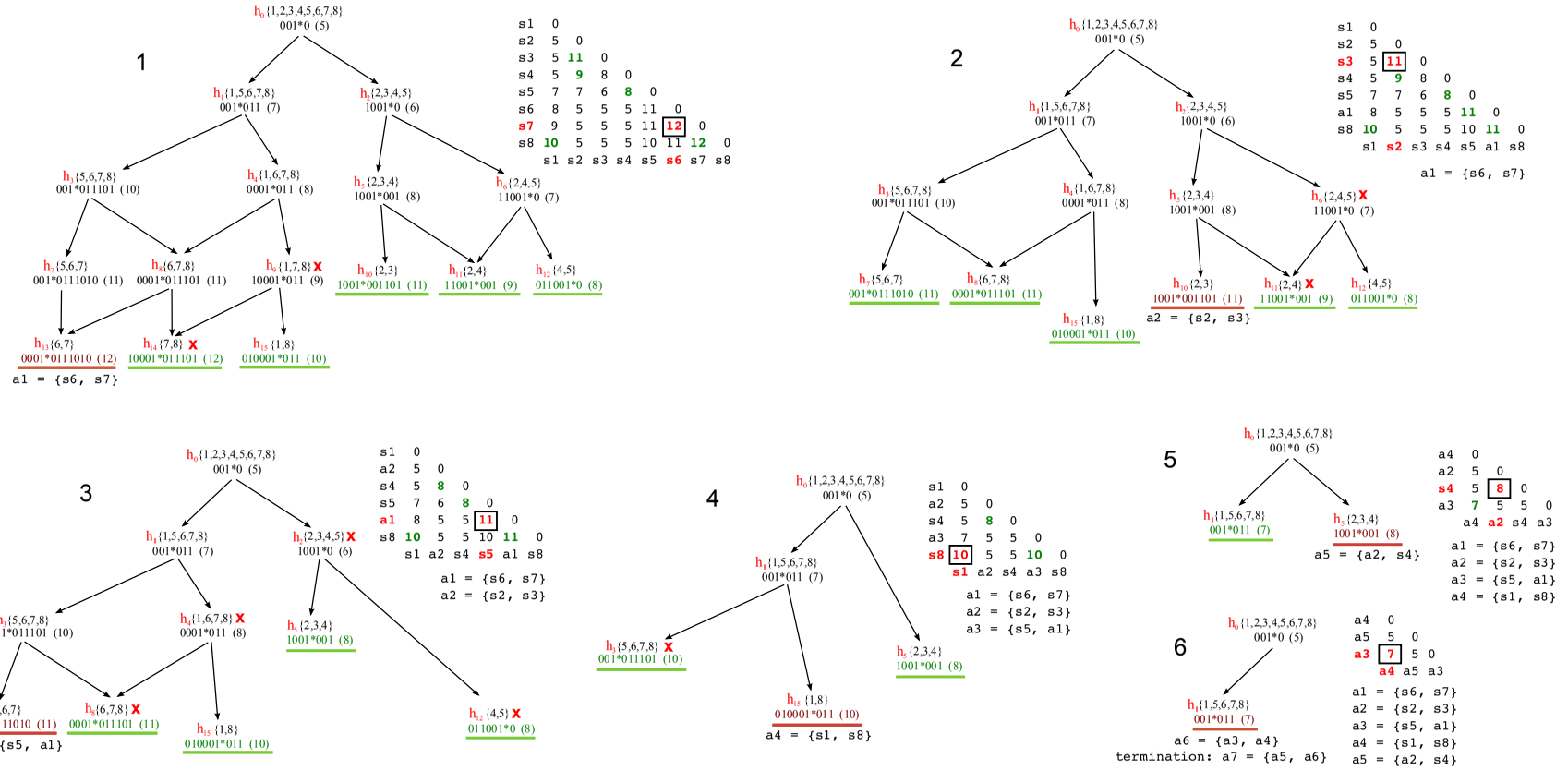

Figure 5.11: Trace of the execution of Algorithm 1 on the data presented in Figure 5.2 and how it implicitly explores the SHG. At each step, we have highlighted in green the candidate haplotypes considered by the algorithm, and in red the candidate chosen. Graph nodes that become incompatible with the tree at each step have been marked with a red cross. For simplicity, we only present the updated L matrix at each step. Sequences are prefixed with s, while tree inner nodes (i.e. ancestors) are prefixed with a. We stress that we display the transitive reduction of the SHG and not the original graph.

choose the pair of sequences/ancestors (i, j) with the longest IBS tract[13]. This operation is akin to searching for the longest IBS tract among the terminal nodes of the current SHG. In reality, the algorithm explores many spurious pairs[14] that do not correspond to maximal haplotype blocks and, therefore, do not have a corresponding node in the SHG. In this regard, the algorithm is *blind* to the structure of the SHG.

Figure 5.11 describes the execution of the algorithm and how it explores and modifies the SHG. Panel 1 shows the beginning of the process, nodes highlighted in green are the terminal nodes considered by the algorithm (also in green in the pairwise distribution of IBS lengths (L), shown on the right). Once the algorithm chooses a pair (highlighted in red in the figure), it merges the two subtrees they represent (either two leaves, two ancestral nodes or a leaf and an ancestral node). This merging operation modifies the breakpoint and IBS length matrices, implicitly trimming from the SHG the chosen node and any other node incompatible with the new topology of the tree being built. An SHG node is compatible with the NNH tree under construction if any of the following statements is true:

1. It covers at least two inner nodes that have no ancestors.

2. It covers at least an inner node (that has no ancestors) and a leaf that still has to coalesce.

3. It covers, exclusively, a set of leaves that still have to coalesce.

We say an SHG node *covers* an inner node of the tree if it contains all the leaves that are its descendants. Likewise, an SHG node covers a leaf of the tree if it contains it. Figure 5.12 exemplifies the notion of compatibility. Before the NNH algorithm starts, all nodes in the SHG are compatible (i.e. each node covers a set of leaves still to coalesce). In figure 5.11, we have indicated with a red cross the SHG nodes that become incompatible with the NNH tree after choosing a particular merging node (again, indicated in red at each step). For example, after choosing $\{h_{13}, \{6, 7\}\}$ as merging node (first iteration, panel 1), the nodes $\{h_{14}, \{7, 8\}\}$ and $\{h_9, \{1, 7, 8\}\}$ become incompatible and are removed from the SHG. We stress that trimming operations are not limited to the set of current terminal nodes

[13]For simplicity, we disregard the resolution of ties via number of singletons. We always resolve ties by simply choosing the pair that contains the sequence with the smallest index.

[14]$\binom{k}{2}$ pairs, where $k = l + a$ is the sum of the number of leaves that remain to join the tree (l) and ancestral nodes (a).

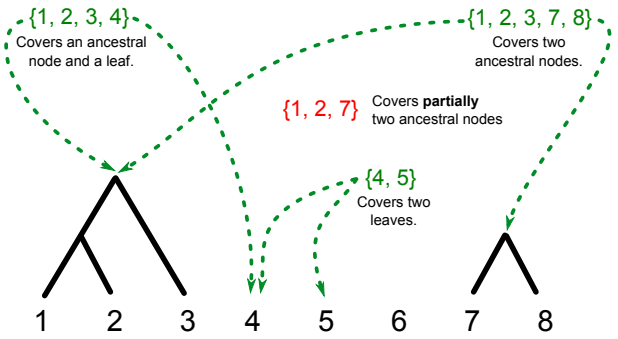

Figure 5.12: Diagram illustrating the notion of compatibility for SHG nodes. At the bottom we show the current state of the tree being built by the NNH algorithm. Above, we list four SHG nodes, highlighting in green the ones that are compatible with the current tree topology and in red the instance that is incompatible. Green arrows indicate the merging operation that would be induced by each compatible SHG node.

in the graph. The matrix updates performed by the algorithm remove any incompatible node in the graph, independently of its location, as can be seen in panel 2. After choosing $\{h_{10}, \{2, 3\}\}$, the non-terminal node $\{h_6, \{2, 4, 5\}\}$ is trimmed since now it only covers partially a tree inner node and two leaves.

Removing non-terminal nodes from the SHG never makes the graph disconnected. To see this, remember that for simplicity we are visualizing the transitive reduction of the SHG and that the core haplotype is connected to any other node in the graph (i.e. only the removal of the core haplotype would create different connected components).

During each subsequent iteration of the algorithm, a candidate is chosen from the set of terminal nodes of the current SHG, the associated merging operations are executed on the tree, and any graph nodes incompatible with the updated tree are trimmed. An SHG node can encode more than one merging operation, for instance if a haplotype block does not terminate (i.e. does not break down before reaching the beginning/end of the chromosome). In this case, the NNH algorithm just proceeds to merge all the nodes/leaves involved in a iterative fashion[15]. The process finishes when only the core haplotype remains compatible (panel 6 in Figure 5.11) and we merge the set of tree nodes/leaves that remain to coalesce.

As one can observe in Figure 5.11, the algorithm must resolve ties, and differet choices could lead to different NNH trees. Recall that the algorithm uses the number of singletons to resolve ties and a deterministic choice if the tie persists. However, here we have skipped this step for the sake of simplicity and resolved ties by sorting candidates by the leaves they contain. Because we use a greedy strategy to choose the merging node, there is no guarantee that we will obtain the tree (or one of the trees) that maximizes the total branch

[15]Notice the branches joining these nodes will have length zero.

length or any optimality criterium. Nonetheless, trees built in this fashion are easier to interpret as a summary of the observed haplotypic diversity on extant sequences as pairs sharing longer IBS tracts are guaranteed to coalesce first.

5.3.3 The NNH algorithm uses a reverse topological order

Now we focus on the observation that the NNH algorithm always finds the longest haplotype block on a terminal node of the current SHG. This property is the result of the SHG being a partially ordered set as given by $(\supseteq, S)$ and the fact that this partial order also respects the increasing order of IBS tract lengths. Let L be in this context the set of IBS tract lengths associated with the nodes of the SHG. We can show that the binary relation $\leq$ over L is a linear extension, or topological sort, of $(\supseteq, S)$, meaning that $(\leq, L)$ is a total order that is compatible with $(\supseteq, S)$ on the SHG.

Lemma 5.3.2. *The binary relation $\leq$ (over L) is a linear extension, or topological sort, of $\supseteq$ (over S) on the SHG.*

Proof. To prove this claim, we need to show that $(\leq, L)$ is a total order and that for every pair of nodes (a, b) in the SHG, $s_a \supseteq s_b \Rightarrow l_a \leq l_b$.[16] Proving that $(\leq, L)$ is a total order is trivial since it is equivalent to $(\leq, \mathbb{N})$, which is already a total order. For proving the second part, assume for the sake of contradiction that there exist a pair of nodes (a, b) in the SHG such that $s_a \supseteq s_b \wedge l_a \not\leq l_b$. The first part of the clause is informative about the relationship of the IBS tract lengths associated with s_a and s_b, specifically

$$s_a \supseteq s_b \Rightarrow h_a \sqsubseteq h_b \Rightarrow l_a \leq l_b.$$

Replacing with the consequent of this chain of reasoning leads to

$$l_a \leq l_b \wedge l_a \not\leq l_b,$$

which is a contradiction. □

Since we have proved that $(\leq, L)$ is a linear extension of $(\supseteq, S)$, we have also proved that $(\leq, L)$ is a linear extension of $(\sqsubseteq, H)$ as both partial orders are equivalent (Lemma 5.3.1).

[16] As expected, l_a and l_b refers to the IBS tract lengths associated with nodes a and b.

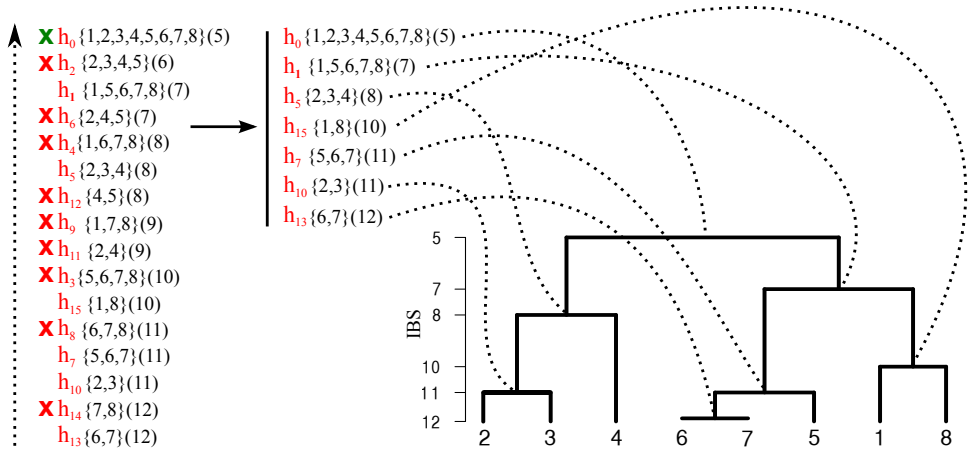

Figure 5.13: The first column shows the reverse order (following $(\geq, L)$, from the bottom up) in which the SHG nodes are explored by the NNH algorithm for the data presented in Figure 5.2 and Figure 5.11. Nodes that became incompatible have been marked with a red cross (the green cross identifies the core haplotype). The right side of the figure maps the merging operations onto the NNH tree topology.

Because $(\geq, L)$ is the inverse of the $(\leq, L)$ total order, $(\geq, L)$ is also a total order[17]. As a consequence, the total order $(\leq, L)$ is a topological sorting of the SHG nodes that preserves the $(\supseteq, S)$ and $(\sqsubseteq, H)$ partial orders. The key observation is that the NNH algorithm uses the $(\geq, L)$ total order to find merging candidates, the inverse of $(\leq, L)$, and thus explores the SHG using this topological sorting in reverse, assesing at each step the set of terminal nodes of the current SHG that are associated with the longest shared haplotypes.

In figure 5.13 we illustrate the precise topological order in which the NNH tree explores the SHG (in reverse) for our previous example (see Figure 5.11). Notice this topological order respects both $(\supseteq, S)$ and $(\leq, L)$. We have highlighted the SHG nodes that become incompatible during the merging operations with a red cross. The remaining nodes enconde the operations required to build the NNH tree (i.e. they are the merging nodes selected by the algorithm).

5.3.4 Building NNH trees directly from the SHG

Capitalizing on our previous observations, we derive a very simple algorithm for building an NNH tree from the nodes of the SHG. We emphasize that for this algorithm to work, only the nodes of the SHG (V^{SHG}) are required as we make no use of the edges of the graph.

[17] $(\geq, L)$ is also a linear extension of the partial orders $(\sqsupseteq, H)$ and $(\subseteq, S)$, which are the inverse of the partial orders $(\sqsubseteq, H)$ and $(\supseteq, S)$, respectively.

The idea is to generate a reverse topological ordering of the SHG that follows $(\geq, L)$. To achieve this, we simply sort nodes by their associated IBS tract length. After obtaining the topological order, we just scan all SHG nodes sequentially. For each node, we compute its coverage (i.e. how many tree nodes and leaves it covers) and apply the rules presented in Section 5.3.2 to assess its compatibility. If the node renders incompatible, we discard it; otherwise we merge the tree nodes and leaves it covers. The algorithm terminates after reaching the core haplotype (the last node in the topological order since it is the least element in $(\leq, L)$), assuring that all tree nodes and leaves that remain to coalesce at this point will join at the root of the tree. The tie resolution via singletons can be easily achieved by annotating the nodes with the total singleton load of their sequences, and using these counts to resolve ties during the topological ordering.

Algorithm 2 Builds an NNH tree from set of nodes of the SHG.

Input: V^{SHG}
Output: T

1: $T = \emptyset$ ▷ Start with an empty tree.
2: $\texttt{tv} \leftarrow \text{TopologicalSort}(V^{\text{SHG}})$ ▷ Sort nodes following $(\geq, L)$
3: **for** $k = 1$ **to** $|V^{\text{SHG}}|$ **do** ▷ Process SHG nodes in order.
4: $\texttt{v} \leftarrow \texttt{tv}[k]$ ▷ Current SHG node.
5: **if** IsCompatible(v, T) **then** ▷ Check compatibility with T.
6: $T = T \cup \text{MergeNodes(v,T)}$ ▷ Update tree by merging covered nodes.
7: **end if**
8: **end for**

The pseudocode for this procedure is presented in Algorithm 2. Regarding time complexity, let $v = |V^{\text{SHG}}|$, this algorithm needs to sort all nodes, which can be done in $O(v \log v)$. Next, it has to process all SHG nodes sequentially, which gives $O(v)$. The coverage check can be performed with very little overhead by the simple scan of the vector that represents all the sequences associated with each SHG node. If we consider checking if a sequence is covered by a node of the SHG as the elemental operation, processing all nodes has a time complexity of $\Theta(\sum_{x \in V^{\text{SHG}}} |s_x|)$, where s_x is the set of sequences associated with node x. The scalability of this algorithm depends on the order of the SHG graph. To characterize the number of nodes of the SHG, in the next section we study the number of maximal haplotype blocks present in the flanks of the locus of interest and introduce a

naïve algorithm that is able to produce the SHG from these sets of shared haplotypes.

5.3.5 Number of maximal haplotype blocks on a flank

Due to the nature of our data (biallelic) and the fact that haplotype blocks get either reduced or split by breakpoints, the maximum number of maximal haplotype blocks we can find on a single flank is $n-1$, where n referes to te number of sequences. With a flank we refer in this context to the genomic region on the right or left of the core locus (but including it).

Lemma 5.3.3. *For n sequences with biallelic markers, the maximum number of maximal haplotype blocks present on a flank of the locus of interest is $n-1$.*

Proof. We start by enumerating all possible ways in which a breakpoint can affect the structure of a haplotype block. For simplicity, we assume that all haplotype blocks terminate (i.e. find a breakpoint) before reaching the beginning/end of the chromosome. When a breakpoint is called at a locus, there are four possible scenarios:

1. The breakpoint terminates a haplotype block. This can only happen when a block is composed of two sequences and the breakpoint affects only one of them. After the breakpoint, both sequences are discarded (cannot join any other haplotype block).

2. The breakpoint splits a haplotype block in two new blocks. This happens when a breakpoint affects at least two sequences within a given block. Notice that the same breakpoint can split several blocks at the same time.

3. The breakpoint removes a sequence from a haplotype block (i.e. reduces the block). When the breakpoint affects a single sequence within a haplotype block, the sequence is removed from the block and discarded.

4. The breakpoint occurs in sequences that do not form part of any haplotype block at that locus and, therefore, has no effect.

It is important to highlight that scenarios (1-3) are not exclusive and that the same breakpoint can affect multiple haplotype blocks. Cases (2) and (3) effectively create

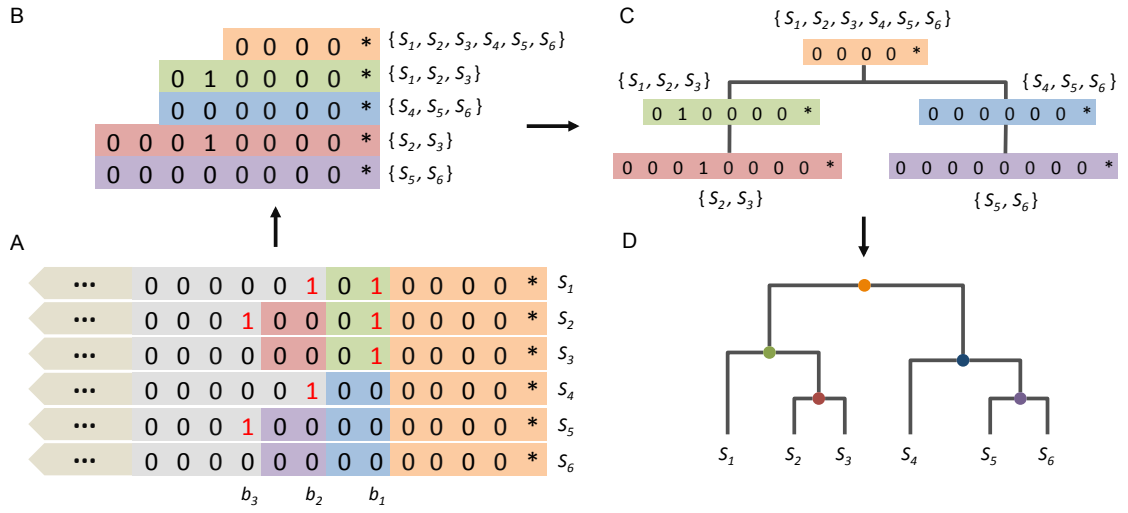

Figure 5.14: (A) Example showing haplotype blocks for six sequences on the left flank of the locus of interest (marked with '*'). Breakpoints are highlighted in red and named at the bottom. The first breakpoint (b_1) splits the core haplotype (orange) into two (green and blue). The second breakpoint (b_2) reduces the green (discarding sequence S_1) and the blue (discarding sequence S_4) haplotype blocks, creating two new blocks (red and purple). The last breakpoint (b_3) terminates the red and purple shared haplotypes. (B) Collection of shared haplotypes and set of sequences associated with each one (on the right). (C) Tree representing the relationship between blocks (ancestors are subhaplotypes), with the core haplotype at the root (orange). (D) Binary tree induced over sequences, inner nodes represent haplotype blocks (corresponding color).

superhaplotypes[18], composed by subsets of the sequences that belong to the blocks being split or reduced. We illustrate all scenarios in Figure 5.14 (see figure caption).

We can show that this process always produces a complete binary tree by associating haplotype blocks with tree inner nodes and mapping each of the previous breakpoint scenarios to tree buiding operations.

1. Termination of a haplotype block. In this scenario we join the inner node represented by the block with two leaves (the two sequences in the block).

2. Splitting a haplotype block into two new sub-blocks. This is equivalent to joining the nodes representing the original block being split (parent) with the new sub-blocks (children).

3. Reducing a block (removing a sequence). In this case, we join the inner node representing the original block (parent) with a leaf (the sequence being removed) and the node that represents the reduced block.

[18]The haplotype of the original block is a subhaplotype of the new blocks induced by the breakpoint.

4. No modification on blocks. It does not translate into any tree operation.

Because we assumed that all blocks terminate and, by definition, there is always a block common to all sequences at the core locus (the root), it is guaranteed that this process will build a complete binary tree (with sequences as leaves) when the operations are performed in the reverse order in which breakpoints are encountered (away from the core locus). Because a complete binary tree has always $n-1$ internal nodes (where n is the number of leaves), we have proved that the maximum number of maximal haplotype blocks on a single flank is indeed $n-1$. □

It is easy to relax the condition that all blocks must terminate before reaching the beginning/end of the chromosome. In that case, we just coalesce all sequences belonging to any non-terminated block (leaves) with the node representing such block (parent). With this relaxation, we are not guaranteed to build a complete binary tree (as the resulting tree may have polytomies at some level above the leaves) but it does not change our previous conclusion since any multiforcating trees will have less than $n-1$ inner nodes.

5.3.6 Upper bound for the order of the SHG

Here we show that the set of haplotype blocks in the SHG is contained in the pairwise intersection between the blocks of the left and right flanks, giving the order of the SHG an upper bound of $|V^{\text{SHG}}| \leq n^2$ nodes. Let S^L and S^R be, respectively, the collection of sequence sets associated with the blocks of the left and right flanks while S conserves its usual meaning (collection of sequence sets of the SHG).

Lemma 5.3.4. *The set of haplotype blocks of the SHG is contained in the pairwise intersection of the maximal haplotype blocks present on the flanks. In terms of the set of sequences associated with each block, we have* $S \subseteq \{s_l \cap s_r\},\ \forall s_l \in S^L, \forall s_r \in S^R$.

Proof. Recall that, because of Lemma 5.3.1, $s_a = s_b \Leftrightarrow h_a = h_b$. Once again, for the sake of contradiction, assume that there exists a node x in the SHG that is not included in the pairwise intersection of flank blocks

$$s_x \notin \{s_l \cap s_r\}, \qquad \exists s_x \in S,\ \forall s_l \in S^L,\ \forall s_r \in S^R.$$

Because s_x belongs by definition to the SHG, it must represent a maximal haplotype block. It follows that at least a block from each flank must be a subhaplotype of the haplotype associated with s_x (h_x) and, equivalently, be associated with a superset of s_x. Let l and r be arbitrary blocks from the right and left flanks. We have, $h_l \sqsubseteq h_x \wedge h_r \sqsubseteq h_x \Leftrightarrow s_l \supseteq s_x \wedge s_r \supseteq s_x$. And from the consequent, we know that

$$s_l \supseteq s_x \Rightarrow s_l \cap s_x = s_x,$$
$$s_r \supseteq s_x \Rightarrow s_r \cap s_x = s_x.$$

However, this makes $s_l \cap s_r = s_x$, replacing in the original statement we obtain $s_x \notin \{s_x\}$, which leads to a contradiction. □

Lemma 5.3.4 provides a very simple, although inefficient, way of building the SHG. We compute the pairwise intersection of the sequence sets associated with the maximal haplotype blocks found at the flanks, and remove any intersection with cardinality < 2, since a shared haplotype requires by definition at least 2 sequences. Next, we add a directed edge between any pair of nodes (a, b) that satisfy $s_a \supseteq s_b$. This procedure has a time complexity of $O(n^2)$, if we consider set intersections as elemental operations, for computing the set of nodes of the SHG. Adding edges requires comparing all pairs of nodes in the SHG, giving an overall complexity of $O(n^4)$, in terms of set operations.

5.3.7 Order of the SHG for the average case

In Section 5.3.6, we showed that the theoretical maximum number of nodes in the SHG is n^2, where n refers to the number of sequences. This resulted from the pairwise intersection of S^L and S^R, the set of sequences associated with the maximal haplotype blocks on each flank (left and right). However, here we argue that the actual order of the SHG for the average case is much smaller. Our reasoning relies on the following points:

1. Many pairwise intersections would produce an empty set or a set with cardinality one, which are discarded and do not form part of the SHG.

2. Many intersections would render sets already present in the SHG, thus not incresing its order.

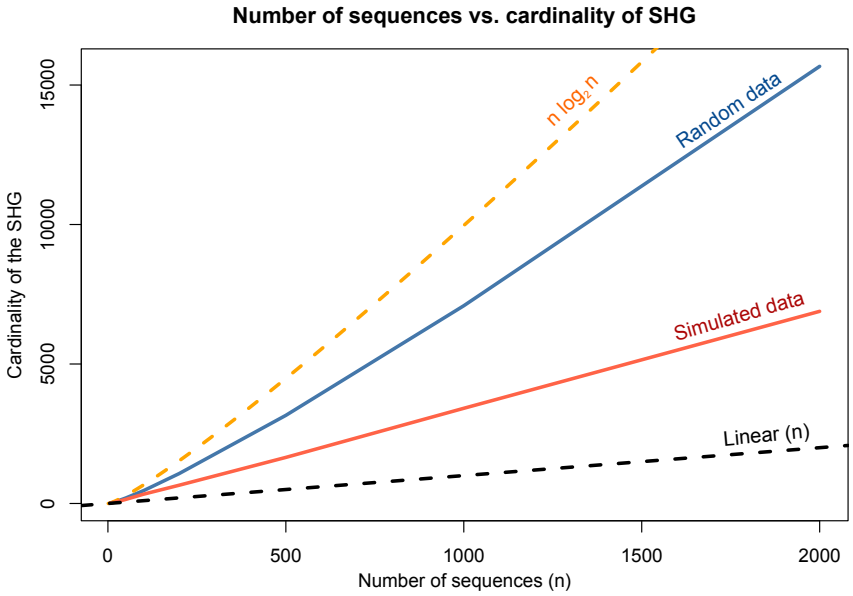

Figure 5.15: Plot showing how the order of the SHG varies for different values of n, for random data (blue line) and sequences simulated under the neutral coalescent (red line). We computed the average order over 50 replicates for $n \in \{5, 10, 25, 50, 100, 200, 500, 1000, 2000\}$. For simulating sequences we used the SCRM package [Staab et al., 2015] using $\mu = 1 \times 10^{-7}$, $r = 1 \times 10^{-8}$, $N_e = 50000$. The figure indicates that the order of the SHG is quasilinear, closer to being linear for simulated data.

3. An SHG with many nodes is the product of very dissimilar haplotype structure on the flanks. However, sequences are not independent of each other, they are correlated via their common genealogical history. This provokes a reduction in the variability of the haplotypic structure seen on the flanks relative to that of random data.

To test our hypothesis, we computed the average order of the SHG for different samples sizes when using random and simulated data. Figure 5.15 shows the results of this analysis. We found that, in average, the number of nodes of the SHG is quasilinear[19] regarding n, and much closer to being linear, $O(n)$, for simulated than for random data.

5.3.8 A branch and bound strategy to find the nodes of the SHG

In the next sections, we provide a quasilinear algorithm (in terms of set intersections) to compute the set of nodes of the SHG. To do this, we start by characterizing the flank SHG and its relation to NNH flank trees. By showing that NNH flank trees are equivalent to the

[19] $O(n^{1+\epsilon})$ with $0 < \epsilon < 1$.

transitive reduction of the underlying flank SHG, we are able to devise a strategy, akin to *branch and bound*, that reduces substantially the number of set intersections required to generate the nodes of the SHG.

5.3.8.1 An NNH flank tree is the transitive reduction of the flank SHG

Given all the haplotype blocks present on a flank of the core locus, we can build the flank SHG by adding a directed edge between any pair of nodes that satisfy either the $\sqsubseteq$ or $\supseteq$ relations (over H and S, respectively). Here we show that the transitive reduction of this graph is the tree (when leaves are removed) that we described in Section 5.3.5.

Theorem 5.3.5. *The flank tree, when leaves are removed, that results from the procedure described in Lemma 5.3.3 is a transitive reduction of the corresponding flank SHG.*

Proof. For the flank tree to be a transitive reduction of the SHG it has to contain the same set of nodes as the graph and preserve the same reachability relation. The first condition is trivial to prove since, by contruction, the flank tree contains as inner nodes all the maximal haplotype blocks present in the flank (see Lemma 5.3.3). Let a^{SHG} and b^{SHG} be two arbitrary nodes in the SHG, and a^T and b^T be the corresponding nodes in the flank tree. Because a^T refers to the same haplotype block as a^{SHG}, we know they are associated with the same set of sequences ($s_{a^T} = s_{a^{\text{SHG}}}$) and use just s_a to simplify our notation. The second condition requires us to prove

$$a^{\text{SHG}} \dashrightarrow b^{\text{SHG}} \;\Rightarrow\; a^T \dashrightarrow b^T,$$

meaning that, if there is a directed path between a and b in the SHG, there must exist a path among the same nodes in the flank tree. Suppose, for the sake of contradiction, that $a^{\text{SHG}} \dashrightarrow b^{\text{SHG}} \wedge a^T \not\dashrightarrow b^T$. Because a^T and b^T are not reachable from each other, and form part of a tree, we know that they are not ancestors and must reside in different clades of the tree.

Let m^T be the MRCA of both nodes. Since m^T is an ancestor of a^T and b^T, we know that $s_m \supseteq s_a$ and $s_m \supseteq s_b$. The edges of each inner node of the tree produce a bipartition[20] over the set of sequences or leaves.

[20] Polytomies can oly occur on inner nodes whose children are all leaves, see Section 5.3.5.

Without loss of generality, let $\triangle s_m^a$ be the subset of s_m associated with the clade where a^T resides. Likewise, let $\triangle s_m^b$ be the subset of s_m associated with the clade that contains b^T. Since the outgoing edges of m^T induce a bipartition on s_m, we know that $\triangle s_m^b = (\triangle s_m^a)^{\complement}$. Set intersection is associative, thus

$$s_a \cap (s_b \cap s_m) = (s_a \cap s_b) \cap s_m.$$

For the left side we have

$$(s_b \cap s_m) \subseteq (\triangle s_m^a)^{\complement} \wedge s_a \subseteq \triangle s_m^a \;\Rightarrow\; s_a \cap (s_b \cap s_m) = \emptyset.$$

Since a and b are connected in the SHG, by construction $s_b \subseteq s_a$, for the right side we have

$$(s_a \cap s_b) = s_b \;\Rightarrow\; (s_a \cap s_b) \cap s_m = s_b.$$

Substituting on both sides we obtain

$$\emptyset = s_b,$$

which is a contradiction, as b is by definition a node of the SHG and thus s_b cannot be empty. □

Because the transitive reduction of the flank SHG is a tree, the original NNH tree algorithm will build an NNH tree with the same topology than the reduction when applied on a single flank. Thus, NNH flank trees represent the transitive reduction[21] of the underlying flank SHG. Algorithm 1 can be easily adapted to build flank trees by fixing one of the breakpoints at the locus of interest (l). Equation 5.3 and 5.4 can be rewritten for the left flank as

$$b_Q^L = \max\big(B_{i,j}^L\big),\; b_Q^R = l, \qquad \forall\, i, j \in Q, i \neq j, \tag{5.13}$$

and for the right flank as

$$b_Q^L = l,\; b_Q^R = \min\big(B_{i,j}^R\big), \qquad \forall\, i, j \in Q, i \neq j. \tag{5.14}$$

[21] Expanded with the leaves covered by their terminal nodes.

Algorithm 2 does not need to be adapted as it only requires the nodes of the flank SHG. For this particular case, however, no node in the graph will be rejected as incompatbile (all nodes will induce coalescing events in the NNH tree).

5.3.8.2 The pairwise IBS distribution can be derived from NNH flank trees

Corollary 5.3.5.1. *The maximal haplotype shared by any pair of sequences on a flank is the haplotype associated with the MRCA of both sequences in the corresponding NNH flank tree.*

It follows from Theorem 5.3.5 that an NNH flank tree has the same topology[22] than the transitive reduction of the flank SHG. Therefore, the MRCA of the sequences in the NNH tree will be associated with the same maximal haplotype block.

This property makes feasible to reconstruct pefectly the pairwise IBS length distribution for a flank from its NNH flank tree. We just need to find the IBS tract length associated with the MRCA of every pair of sequences, which can be done on a single top-down scan of the tree. When both NNH flank trees are available, the original pairwise IBS distribution can be easily computed as $L = L^L + L^R - 1$. Where L, L^L, and L^R represent the pairwise IBS length distribution for both flanks, the left flank and the right flank, respectively.

5.3.8.3 Finding the nodes of the SHG with NNH flank trees

In Section 5.3.4 we presented an algorithm (Algorithm 2) for building NNH trees directly from the ordered collection of nodes of the SHG. In light of the results of Section 5.3.7 (see Figure 5.15), this algorithm is, for the average case, quasilinear on the number of sequences (n).

Here, we introduce a simple algorithm that exploits the structure of NNH flank trees to find the nodes that belongs to the SHG without performing n^2 set intersections. Recall that we related the set of nodes of the SHG to the pairwise intersection between the nodes of the left and right flank SHG in Lemma 5.3.4. Because NNH flank trees represent the transitive reduction of the flank SHG (Theorem 5.3.5), we can make use of their structure to avoid processing spurious intersections that will render invalid nodes (i.e. either empty or containing a single sequence). The intuition behind the algorithm is to traverse the NNH

[22] When leaves are added to the terminal nodes.

tree of the right flank with each node of the left NNH tree, discarding the traversal of a subtree whenever the intersection between a pair of nodes is invalid, reducing the number of set intersections to be performed.

Algorithm 3 Finds the nodes of the SHG from the nodes of the NNH flank trees.

Input: T^L, T^R
Output: V^{SHG}

1: $V^{\text{SHG}} = \emptyset$ ▷ Start with an empty SHG.
2: $\texttt{lt} \leftarrow \text{DepthFirstSearchOrder}(T^L)$ ▷ DFS order for the inner nodes of T^L.
3: **for** $i = 1$ **to** $|T^L|$ **do** ▷ Process T^L nodes in DFS order.
4: $\quad \texttt{l} \leftarrow \texttt{lt}[i]$ ▷ Current node of T^L.
5: $\quad V^{\text{SHG}} = V^{\text{SHG}} \cup \text{FindValidIntersections}(\texttt{l}, T^R)$ ▷ Find intersections.
6: **end for**

Algorithm 3 sketches this process. The key piece is the call to the function that finds valid intersections (line 5). This function traverses T^R efficiently, by stopping to explore any subtree whose root produces an invalid intersection with the node from the left tree. To speed up the computation of intersections, we update the traversing node when going down a subtree with the valid intersection of such node and the root of the subtree we are exploring. Figure 5.16 illustrates this procedure with an example.

We repeated the same analysis of Figure 5.15, this time counting the number of set operations performed when computing the set of SHG nodes using Algorithm 3. Figure 5.17 shows that the number of set intersections is, for the average case, quasilinear. We found that simulated data requires more intersections than the random data precisely because the NNH flank trees tend to be more correlated, and similar clades need to be explored to very deep levels (although this adds little to the number of nodes in the SHG).

5.3.9 Building NNH flank trees in linear time

With the introduction of Algorithm 3, the major bottleneck for building NNH trees resides now in the computation of the NNH flank trees, as this still has time complexity $O(n^2 \log n)$ by using our adapted version of Algorithm 1 (see Section 5.3.8.1). Here we show an alternative method for generating NNH flank trees that has linear time complexity. We capitalize on the connection between how haplotype blocks are structured within a flank and the positional Burrows-Wheeler transform [Durbin, 2014].

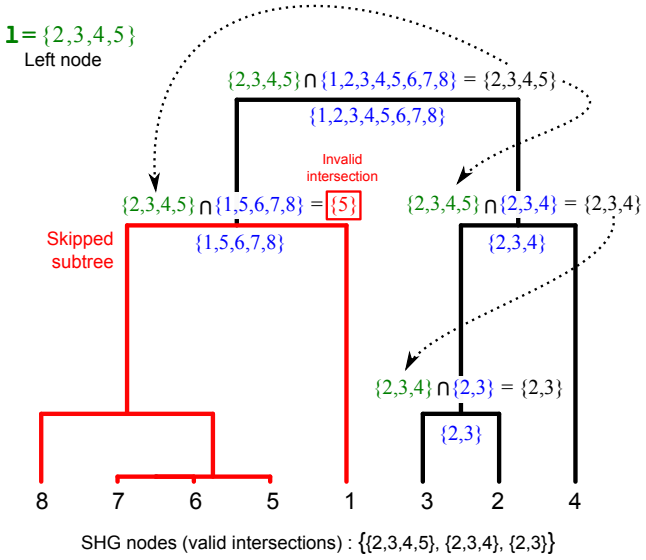

Figure 5.16: Example of the procedure that finds SHG nodes by traversing the right NNH tree with the nodes of the left NNH tree. The left node under consideration is $\{2, 3, 4, 5\}$, in green. We render the nodes of the right tree in blue. Valid intersections are colored black, whereas invalid intersections are shown in red. Arrows indicate how the traversing left node is updated when going down the tree. The procedure skips the left subtree after finding an invalid intersection ($\{5\}$).

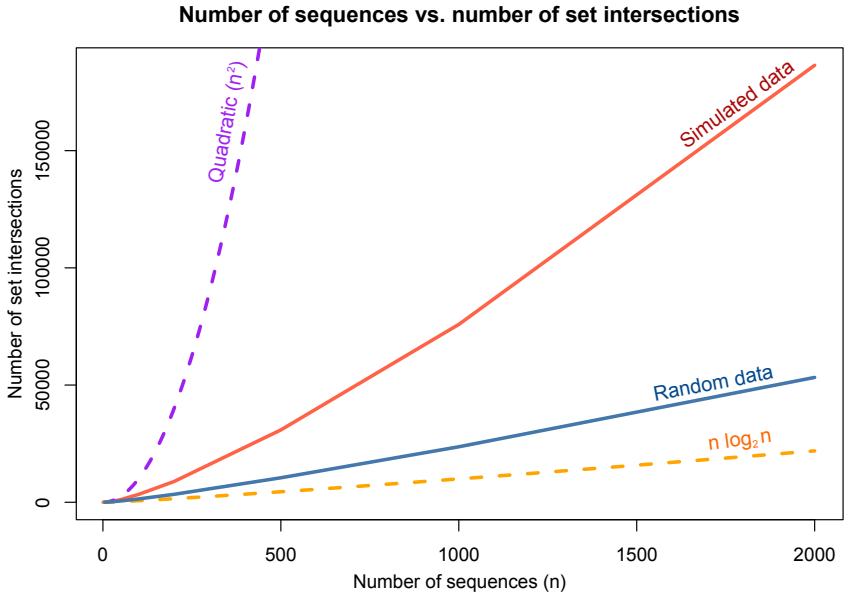

Figure 5.17: Plot showing how the number of set operations performed to find the nodes of the SHG varies as a function of n. Analysis executed on random data (blue line) and sequences simulated under the neutral coalescent (red line). We computed the average number of set operations over 50 replicates, same data as in Figure 5.15. The figure indicates that the number of set operations required to find the nodes of the SHG is quasilinear on n.

The PBWT is a simple permutation process that sequentially transforms the input haplotype data into a representation that is easy to index, query and compress. The PBWT for locus l is computed from the transformed data in $l-1$. There are three properties of the PBWT that we can exploit for our purposes:

1. Maximally overlapping haplotypes (starting at locus $k-1$, where k is our locus of interest) are adjacent in the PBWT represention.

2. The overlapping length of any two haplotypes (i, j) is given by the minimum overlap within the sequence of all haplotypes ocurring between i and j in the transformed order.

3. The locus at which the overlap between two adjacent haplotypes in the PBWT starts can be computed at no extra cost.

The PBWT produces two vectors for each processed locus, the positional prefix array (the permutation of the sequences), and the divergence array (the starting locus for the match between adjacent haplotypes). Because an NNH flank tree is constructed by merging pairs of maximally overlapping haplotypes, and no merging operation renders another node of the underlying SHG as incompatible, the PBWT output already encodes the sequence of merging operations rquired to build the tree. In light of this, we propose a simple algorithm (Algorithm 4) to build NNH flank trees from the PBWT on linear time. The algorithm receives the positional prefix array (`ppa`), the divergence array (`da`), the number of sequences (n) and the locus of interest (k). This algorithm works for the left flank of the locus of interest, it is straightforward to adapt the algorithm for right flanks or we can just reverse the sequences. Notice that we temporarily remove singletons via encoding before applying the PBWT (so they do not cause breakpoints).

5.3.10 Overall time complexity

In Figure 5.18 we detail our alternative procedure to generate NNH trees. The first step consists of building all left and flank trees via the PBWT (Algorithm 4). Since this approach avoids computing the pairwise IBS length distribution at each locus, building all NNH flank trees for a chromosome has a time complexity of $O(n\,L)$ for the PBWT (with n referring

Algorithm 4 Builds an NNH-L tree from the PBWT for locus k.

Input: $\texttt{ppa}, \texttt{da}, n, k$

Output: T

1: $T = \emptyset$
2: $\texttt{inode} \leftarrow n + 1$ ▷ Labels for inner nodes.
3: $\texttt{p} \leftarrow 2$ ▷ Cursor, 1-based indexing.
4: **while** $\texttt{length}(\texttt{da}) > 1$ **do**
5: **while** $\texttt{da}[\texttt{p}] \geq \texttt{da}[\texttt{p} + 1] \wedge \texttt{p} < \texttt{length}(\texttt{da})$ **do**
6: $\texttt{p} \leftarrow \texttt{p} + 1$
7: **end while**
8: $\texttt{children} = \{\texttt{ppa}[\texttt{p}], \texttt{ppa}[\texttt{p} - 1]\}$ ▷ Merging adjacent haplotypes.
9: $\texttt{breaks} = \{\texttt{da}[\texttt{p}], k\}$ ▷ Merging adjacent haplotypes.
10: $T = T \cup \{\texttt{inode}, \texttt{breaks}, \texttt{children}\}$ ▷ Save inner node and breakpoints.
11: $\texttt{remove}(\texttt{da}, \texttt{p})$
12: $\texttt{remove}(\texttt{ppa}, \texttt{p})$
13: **if** $\texttt{p} > 1$ **then**
14: $\texttt{p} \leftarrow \texttt{p} - 1$
15: **end if**
16: $\texttt{ppa}[\texttt{p}] \leftarrow \texttt{inode}$ ▷ Rename with the merged node.
17: $\texttt{inode} \leftarrow \texttt{inode} + 1$
18: **end while**

to the number of sequences and L being the number of loci), and $O(n)$ for each individual tree. Building all the trees in the chromosome has therefore a time complexity of $O(n\,L)$. The PBWT needs to be computed sequentially in each direction (left and right) but NNH flank trees can be constructed in parallel once PBWT results are available, which makes practical the analysis of very large datasets. The second step uses the NNH flank trees of a locus to compute the set of SHG nodes (Algorithm 3). This branch and bound strategy has a quasilinear time complexity ($O(n^{1+\epsilon})$ with $0 < \epsilon < 1$) for the average case, in terms of set intersections. Finally, we build NNH trees by applying Algorithm 2 to the set of SHG nodes. This algorithm is linear in the number of nodes of the SHG, which is quasilinear regarding n. The time complexity of building an NNH tree is dominated by the second step, thus making the whole procedure quasilinear in terms of set intersections. We considered set intersections as elementary operations because there are many efficient implementations available, for instance based on Bloom filters [Broder and Mitzenmacher, 2004]. Once the left and right flank trees have been computed, constructing the AHG (i.e. building an NNH tree for each locus) is still embarrassingly parallel and subject to a trivial parallelization when many CPUs are accessible. An efficient implementation of the algorithms presented in this section can cope with large datasets (in the order of several thousands of sequences) but would not scale well with very large datasets.

5.4 The AHG is informative about recent genalogical history

In this section, we informally show that the AHG is informative about recent genealogical history. To do this, we simulated a chromosome for 50 sequences segregated into two different populations that interchange individuals at a very low migration rate (see Figure 5.19). Because the SCRM package provides the genealogies responsible for the simulated sequences, we could interrogate their actual genealogical relationship.

For our purposes, we scanned through all loci identifying the closest genealogical neighbor of each sequence (i.e. the sequence that can be reached through the shortest path in the tree starting at the sequence of interest). This procedure induces a collection of non-symmetric boolean[23] pairwise matrices representing the *is-closest-neighbor* relationship.

[23] In the case of a tie, we split the weight between the neighbors. $1/k$, where k is the number of closest neighbors.

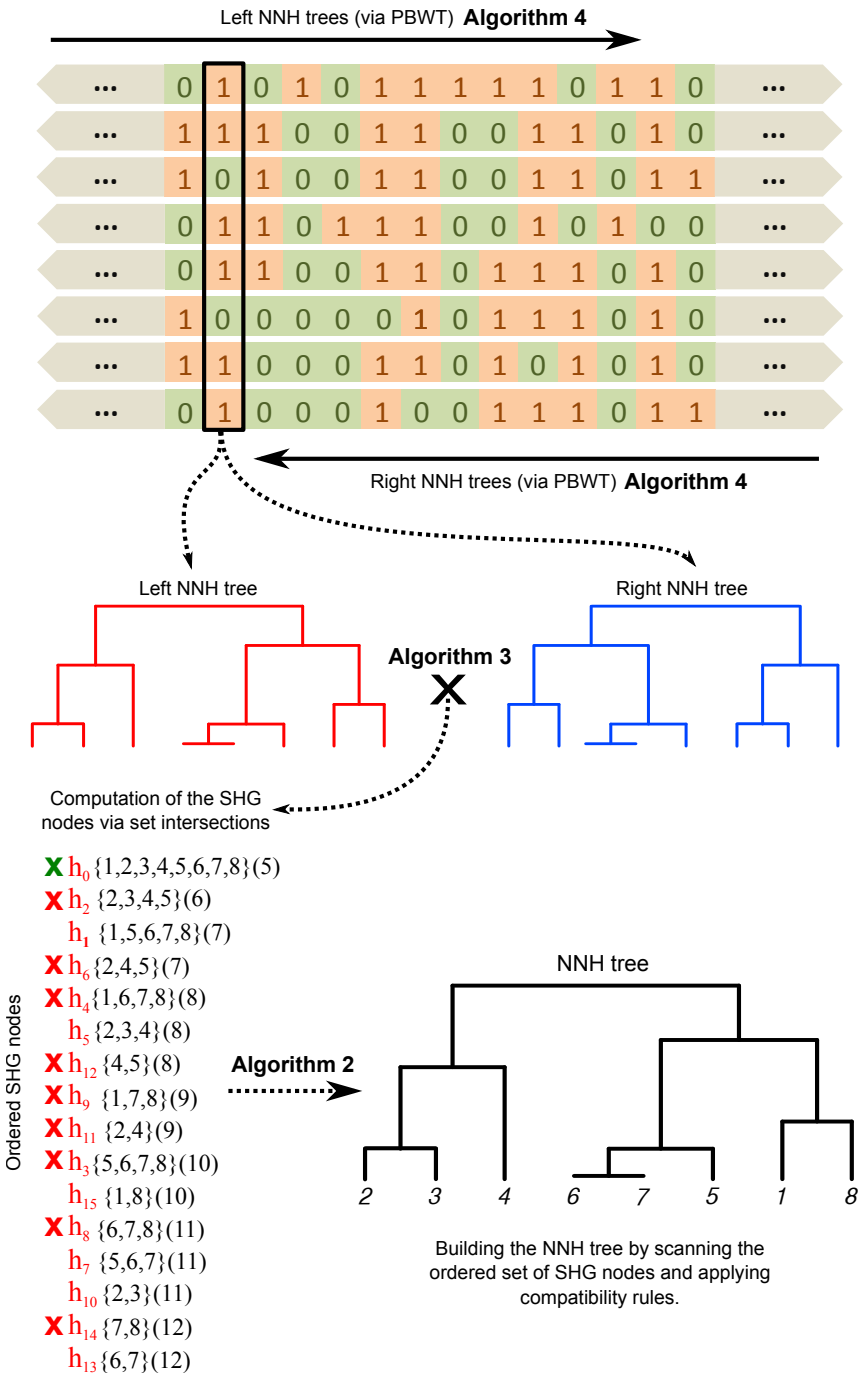

Figure 5.18: Cartoon detailing the alternative procedure for generating NNH trees. We have labeled each step with its associated algorithm.

Averaging all these matrices we obtain a summary matrix of weights (W) that indicates which sequences tend to be, in average, closest genealogical neighbors. This is exactly what the chromosome painting model [Lawson et al., 2012] approximates via the coancestry matrix (Section 1.4.3.2.2). If these weights are normalized, they can be interpreted as the expected probability of a sequence being the closest genealogical neighbor of another when sampling a random marker in the genome.

Figure 5.19 shows the results of our simulation. We can see how the coancestry matrix is a very good approximation of the truth (i.e. W computed on the simulated genealogies). In this little example, we compute W on the collection of local trees that compose the AHG. We obtain, as shown in Figure 5.19, a distribution of weights that is very similar to the truth and practically identical to the matrix obtained via chromosome painting[24]. This demonstrates that indeed the AHG contains rich information about the recent genealogical history shared by the sequences, and also suggests that, in this regard, provide a faster alternative[25] to more sophisticated methods.

5.5 Code availability

The code for all the algorithms presented in this chapter is freely available at `https://github.com/mcveanlab/NNH`. For the time being, we provide a prototype that consists of R scripts, with some of the most demanding functions implemented directly in C++. This is, nonetheless, just an early software prototype that can be highly optimized. Our plan is to release an R package (named `NNH`) for the final version of the software.

5.6 Limitations

We have focused on providing a fast and scalable method to produce the IBS ancestral haplotype graph for a set of genomes. If our goal is to approximate the underlying genealogical history, the major drawback of this approach is its sensitivity to the presence of genotyping errors and recurrent mutations. Although assumed to be rare, these events will shorten or lengthen the detected IBS tracts, and therefore confound any subsequent

[24] In reality, chromosome painting deviates less from the truth, but the difference is marginal

[25] We are aware that genotyping errors and recurrent mutations would make our approach less accurate, see Section 5.6.

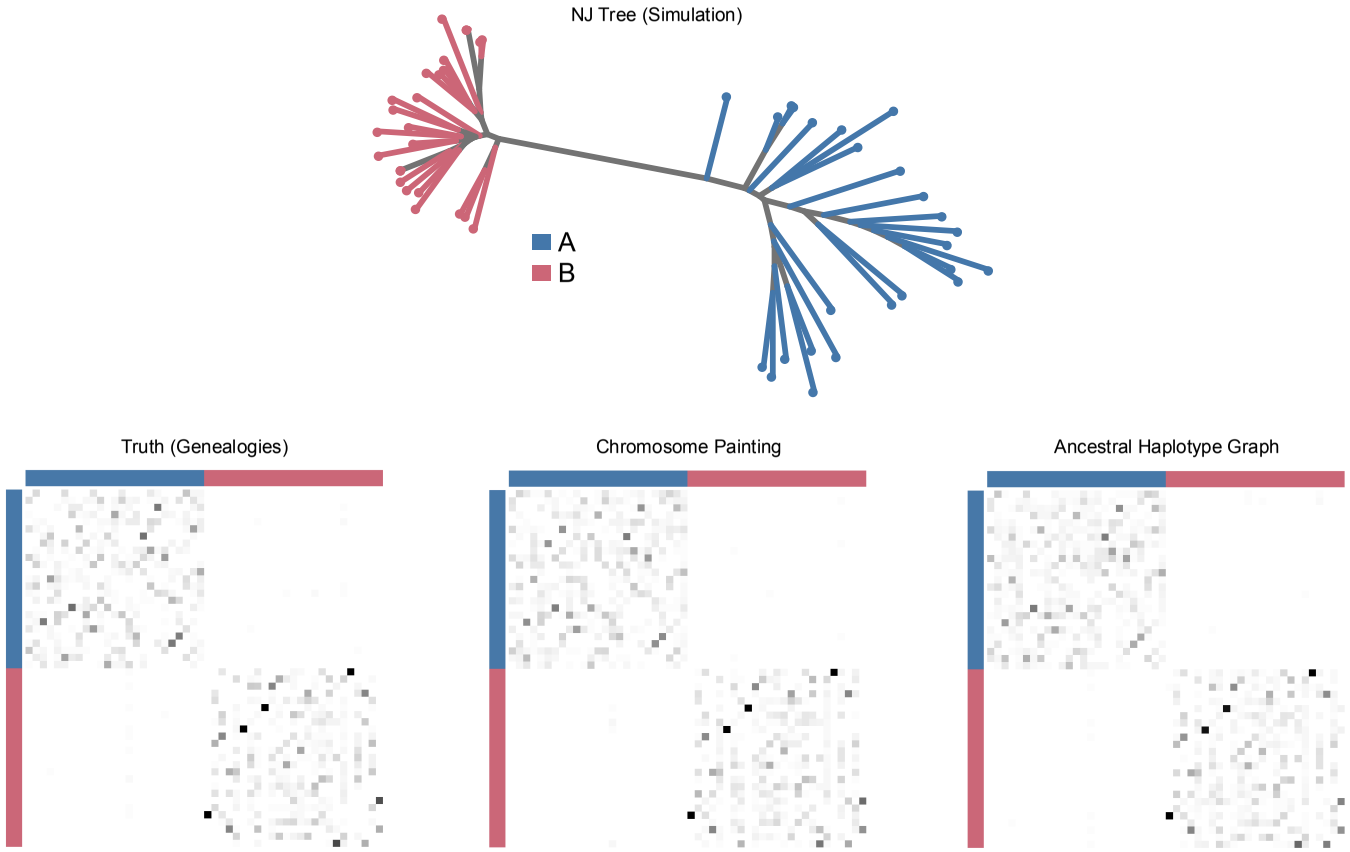

Figure 5.19: Illustrative example showing that the AHG is informative about recent genealogical history and suggesting that it could compete with more sophisticated methods like chromosome painting. (Top) A neighbor joining tree summarizing the simulated data. (Bottom-left) The true distribution of average weights for closest neighbors accross the simulated region. (Bottom-middle) The distribution of weights obtained via chromosome painting. (Bottom-right) The same matrix of weights obtained from the AHG. Sequences of 2.5Mb in length were simulated with the SCRM package [Staab et al., 2015] using $\mu = 1 \times 10^{-7}$, $r = 1 \times 10^{-8}$, $N_e = 50000$ and a symmetric migration rate of 4×10^{-6}.

inferential method. There are several alternatives to lessen the effect of this problem, for instance using more sophisticated heuristics whose aim is to detect the presence of recombination (e.g. the four-gamete test [Hudson and Kaplan, 1985]), but they would be difficult to reconcile with our particular use of the PBWT. A different approach would entail implementing a resampling procedure akin to bootstrapping, building the AHG many times after resampling genotypes. Our insight is that, if we resample genotypes according to a probabilistic model[26] that approximates the recombination process, genotype errors and recurrent mutations will be more likely to be *sampled out* than polymorphisms representing the true ancestral haplotype shared by a set of sequences. If successful, this methodology could form the basis of an ARG sampler. Assessing the merits of this approach, and its robustness requires further investigation. The other main limitation of our approach is that it can not accommodate mixed samples. However, the recent development of methods for the statistical deconvolution of mixed infections [Zhu et al., 2017] permit to infer the individual haplotypes of the strains contained in a mixed isolate and allow the build the AHG in the presence of mixed infections.

5.7 Conclusions and further work

In this chapter, we introduced NNH trees as a tool to compactly describe the haplotype diversity along the genome of a set of sequences. NNH trees are easy to interpret and visualize and do not require the use of sliding windows. We studied the properties of NNH trees in detail and shown that they perform better[27] than the widely used NJ trees. Furthermore, the collection of genome-wide NNH trees (what we call the AHG) can be analyzed to learn about the recent shared genealogical history of the sequences since the AHG can be seen as a crude approximation of the ARG.

We have focused our efforts on devising a scalable method for building the AHG on large datasets (in the order of thousands of sequences). Our approach has a quasilinear time complexity, and it is embarrassingly parallel. Nonetheless, further work is required to make it scalable to very large datasets. An obvious improvement is to find an alternative to Algorithm 3 that can either generate the nodes of the SHG more efficiently or find a subset

[26] The chromosome painting model [Lawson et al., 2012] would be a good candidate for this task.
[27] At summarizing the pairwise IBS tract distribution.

that contains all the nodes considered by the NNH algorithm. The other self-evident line of action is to develop a set of metrics and procedures to analyze the AHG and learn about phenomena that are important in genomic epidemiology, such as gene flow, selection or the relative ages of mutations. Finally, we consider worthy evaluating the ideas outlined in Section 5.6 as a way of devising an ARG sampler or demographic inferential methods.

5.8 Individual contributions

As this chapter was not part of any collaborative project, it is completely the product of my individual work.

Chapter 6

General discussion

Contents

6.1 Introduction

This thesis considers the role of genetic population structure in the evolution and demography of *P. falciparum*. Because of its gravity, concerning global public health, I have focused on the recent emergence of artemisinin resistance in Southeast Asia (Chapter 2 and 4). I investigated the relationship between population structure and resistance, and how the occurrence of founder effects shaped the demography of the parasite. Next, motivated by the importance of demographic inference to inform malaria control programs and the advent of large genomic datasets, I developed a fast and scalable method to build the ancestral haplotype graph (Chapter 5). This data structure, composed of a collection of local trees, can be used to summarize and analyze haplotype patterns along the genome, and it is informative about the recent genealogical history shared by the sequences. Following a secondary thread, I also examined how the presence of mixed infections can affect the analysis of deep sequencing data (Chapter 3). In doing so, I also evaluated the F_{WS} statistic, introducing an alternative estimator that does not have the shortcomings of the original formulation. In this last chapter, I finish by discussing the main conclusions and avenues for further research that can be derived from this work.

6.2 The role of population structure

In this work, I found that genetic population structure and the occurrence of founder effects had a prominent role in the development and spread of artemisinin resistance. In Chapter 2, following reports of delayed clearance for artemisinin derivatives, I assessed several sympatric subpopulations of parasites coexisting within a small geographical area that had exceptional levels of genetic differentiation in the Thai-Cambodian border region. Three of these subpopulations presented a deficit of low-frequency variants and very reduced haplotype diversity, suggesting recent founder effects. In this setting, the most likely scenario is that the founder effects represented the recent expansion of resistant parasite lineages whose fitness depended on a particular genetic background, and seemed the primary force spreading resistance.

Next, in the first part of Chapter 4, I collaborated in studying a superset of the *kelch13* mutations associated with artemisinin resistance, providing a genetic epidemiological

assessment of the spread of artemisinin resistance in Southeast Asia. Specifically, I examined the roles of population structure and founder effects, and the demographic events that may have modulated the spread of resistance alleles. Seven founder populations (5 in Cambodia and 2 in Vietnam) were strongly associated with artemisinin resistance, each of them being linked to a specific *kelch13* resistance mutation and sharing a distinct genetic background[1]. These findings suggested that *kelch13* resistance alleles might act, when escorted by the appropiated genetic background, as driver mutations in the emergence of artemisinin-resistant founder populations. It is also plausible that the genetic background that escorted *kelch13* mutations had either a compensatory or resistance-enhancing effect (perhaps a mix of the two), with different parts of the background playing singular roles. My demographic analysis, based on patterns of haplotype sharing, in combination with the characterization of population structure, indicated that most resistance alleles had emerged independently and were geographically confined to small regions. This also revealed that the primary determinant in the geographical distribution of artemisinin resistance was the independent emergence of *kelch13* resistance alleles, providing evidence that the spread of the most common resistance allele (k13-C580Y) was the product of recurrent mutations, arising multiple times at different locations. Nonetheless, for some of the most common resistance alleles (k13-C580Y, k13-I543T, k13-Y493H and k13-R539T), I also presented evidence supporting the migration of resistant parasites across countries but only within East Southeast Asia. These observations undermine localized resistance containment as a strategy for malaria control and suggest that population structure and founder effects may predate and facilitate the emergence of resistance. Therefore, monitoring these phenomena could warn about the development of resistance before phenotypic evidence materializes.

In the second part of Chapter 4, I investigated the origin of *kelch13* mutations observed in African samples. According to my statistical analysis, the reservoir of *kelch13* mutations found in Africa seemed to have emerged as a consequence of independent mutational events. I found anecdotal evidence (2 out of 54 *kelch13* African mutant samples) of gene flow between Southeast Asia and Africa. However, both samples presented high complexity of infection and were, most likely, the product of an artificial mixture produced in the preparation and processing stages of different laboratories. Despite discarding the migration

[1] A set of mutations also identified by a GWAS performed by others, see Section 4.2.2.

of parasites across continents, it is clear[2] that artemisinin resistance could emerge without the necessity of gene flow from Southeast Asia if the selective pressure of artemisinin increased. In this regard, it is probable that a particular set of accompanying mutations is required to prompt the development of resistance, as suggested by the results presented in the first part of Chapter 4 and the fact that the genetic background observed in resistant samples from Southeast Asia is absent in Africa. This observation supports the belief that *kelch13* mutations impose a fitness cost too high to be tolerated if not in the presence of sustained drug pressure, explains the limited spread of individual *kelch13* mutations, and strengthens the case for a genetic background composed of compensatory mutations.

6.2.1 Further research

In light of the results presented in this thesis, I argue that tracking genetic population structure and the occurrence of founder effects in the field could become a critical tool for malaria control. Monitoring protocols, for instance in the form of genetic observatories that regularly sample and sequence natural populations of parasites, could be employed to warn about the imminent emergence and spread of resistance before phenotypic evidence materializes. This is particularly relevant in Africa, where selective pressure can be substantially increased if countries reach levels of relatively low transmission. Further development of these ideas could be applied to the design and evaluation of health policy programs and interventions. However, a set of practical scientific questions stem from this proposal. How early and reliably could an observatory detect the signs that predict the advent of resistance? Once detected, how much time in advance do researchers in the field need to change the course of an ongoing intervention and avoid its spread? Is this possible at all? How many markers and samples are sufficient? Research programs that evaluate the implementation of genetic observatories may be able to quantify their usefulness among the new arsenal of genomic methods that can contribute to the control and eradication of malaria.

Although useful, my results provide a very rudimentary and simplified understanding of the mechanisms that prompt the emergence of drug resistance in the field. By focusing on the contribution of population structure, I have neglected other demographic and life-

[2]Since the presence of *kelch13* mutations in Africa is well documented.

cycle aspects of the parasite. It is still an open question how the unobserved reservoir of parasites that exist within asymptomatic patients influences the distribution of genetic diversity within populations and how this may modify the probability of resistance emerging under sustained drug pressure [Bousema et al., 2014], [Harris et al., 2010]. Likewise, I have not examined the role that complexity of infection can have in promoting resistance. For instance, facilitating evolutionary bottlenecks via within-host strain competition in the presence of antimalarial drugs [de Roode et al., 2004], [Hastings and D' Alessandro, 2000]. Addressing these kinds of questions jointly, instead of in isolation, would be key for understanding the biological reality that prompts the emergence of resistance in the field.

6.3 NNH trees and the ancestral haplotype graph

Motivated by the study of *kelch13* mutations, in Chapter 4 I exploited patterns of haplotype sharing to learn about their recent demographic history. Given the importance of demographic inference to inform malaria control programs and the advent of large genomic datasets, in Chapter 5 I developed a tree data structure for summarizing haplotype diversity. I called this construction the nearest neighbor haplotype tree (NNH tree). NNH trees can be used as a visualization tool, describing localized haplotype diversity and structure. They can also be scrutinized at genome-wide scale by comparing the properties of each local tree, characterizing how haplotype composition changes along the genome. I shown that the collection of genome-wide NNH trees (what I call the ancestral haplotype graph or AHG) can be analyzed to learn about the recent shared genealogical history of the sequences[3]. I described a method for building the AHG that has a quasilinear time complexity and it is embarrassingly parallel. Notice that my aim here was to develop a fast and scalable method that could be applied to thousands of sequences. I consider this a first step towards creating scalable inferential methods that can be constructed on top of this data structure.

6.3.1 Further research

An immediate research agenda stems from this piece of work. On the one hand, devising metrics and statistics that, when computed on the AHG, would allow the rapid detection of

[3]In this regard, the ancestral haplotype graph can be seen as a crude approximation of the ancestral recombination graph.

features of interest, such as the occurrence of selective sweeps or introgression events. On the other hand, improving the building method to make it robust to the presence of genotyping errors and recurrent mutations (its major limitation), and building probabilistic inference methods on top of the data structure. In Section 5.6, I sketched some ideas, similar in spirit to bootstrapping, that, in principle, could make the procedure robust to errors/recurrent mutations and also render an efficient mechanism to generate samples of the underlying ancestral recombination graph (ARG). I think this line of research, although very ambitious, is worthy of investigation. An efficient ARG sampler would allow researchers to address many significant problems in the genomic epidemiology of malaria, such as assessing gene flow or the relative dating of all mutations along the genome. Current methods for ARG inference are only able to cope with small datasets [Rasmussen et al., 2014], [Camara et al., 2015]. Nonetheless, because performing inference on the ARG is an extremely hard task[4], building a sampler following the bootstrapping idea could easily lead to a blind alley or be too slow to be useful for researchers. Regarding efficiency and scalability, I remark that reducing the time complexity of the building procedure (i.e. making it closer to linear) would allow researches to study very large datasets (in the order of hundreds of thousands of sequences). This is obviously another line of research worth pursuing.

6.4 Effects of complexity of infection on deep sequencing data

Because mixed infections makes problematic the analysis of deep sequencing data, in Chapter 3 I took a brief detour to study how their presence can introduce biases[5] on the estimation of genetic distances and downstream analyses that are dependent on these. As a result of this assessment, I advised against the use of the majority calling heuristic as it tends to remove rare variants, and advocated for a simple estimator of genetic distantes that is based on the fraction of allele read counts and is robust to the presence of multiple strains. I also explored the limitations of the original F_{WS} estimator and proposed a sampling mechanism whose merit is to incorporate uncertainty from read count data. I advised to use this procedure only in scenarios of low coverage. Besides, I showed that

[4]It is indeed one of the canonical problems in population genetics. An efficient ARG sampler would simplify immensely many other problems in the field [Gusfield, 2014].

[5]This investigation was motivated by the presence of strange patterns on genome-wide NJ trees.

the original F_{WS} estimator does not acknowledge the diversity encoded by rare variants, overestimating the fraction of the population expected heterozygosity that is represented within a mixed infection. I addressed this problem by introducing an alternative estimator (and its normalized version) that does not suffer from this bias, offers a better resolution and has a simpler formulation. I studied how this estimator can be affected by ascertainment biases and found it to be reliable when the set of variants used are not rare (MAF > 0.05) within the population of origin. I also provided further evidence on the F_{WS} statistic being contingent on the diversity and genetic structure of the local population. Because of this, I warned against comparing individual F_{WS} estimates across populations without contextualizing population differences.

6.4.1 Further research

Since its introduction [Manske et al., 2012], F_{WS} has been widely used to portray complexity of infection [Mobegi et al., 2014], [Murray et al., 2016], [Assefa et al., 2015]. The work presented here may falicitate the use and interpretation of the statistic, avoiding common pitfalls such as comparing estimates that originated in populations with a very different genetic structure.

The presence of mixed infections has been a major roadblock for applying genomic methods to parasite sequencing data, with many researchers opting for using only clonal or almost clonal samples for some type of analysis (for instance, analysis that depend on haplotype data). Nonetheless, recent advances in statistical genetics now permit to fully characterize mixed infections on deep sequencing data by providing estimates for the number of strains, their relative abundance, and their individual haplotypes [Zhu et al., 2017][6]. These new methods will allow to study mixed infections at very fine scale. Connecting with some of my remarks in Section 6.2.1, this direction of research will shed light on issues like the dynamics governing infection history or the role of within-host strain competition in promoting drug resistance, also facilitating the elucidation of more accurate epidemiological models for malaria transmission.

[6] Although I'm a author in this publication, the work presented there has no relationship with this thesis.

6.5 Final remarks

There is a common thread underpinning all the work I have discussed: the relevance of genomics tools for the control and elimination of malaria. In a way, we are witnessing how the field of genomic epidemiology comes of age. Statistical methods that interrogate sequencing data allow researchers to query the past (e.g. where did a mutation originate?) and also prepare for the future (e.g. detecting signs of an incipient founder effect). Furthermore, recent methodological advances and the collection of larger datasets will soon make feasible to address some of the open questions I have mentioned here and that have remained elusive during decades (e.g. determining infection history or the dynamics of within-host strain competition). Despite the looming risk of a severe epidemic of malaria occurring in Africa due to the emergence or spread of artemisinin resistance, I finish on an optimistic note. It would be naïve to think that we are anywhere close to eradicating malaria, but I am confident that genomics and the joint collaborative endeavor of the scientific community will play a pivotal role in paving the way for reducing the burden of the disease.

Bibliography

[Abbott et al., 2013] Abbott, R., Albach, D., Ansell, S., Arntzen, J., Baird, S., Bierne, N., Boughman, J., Brelsford, A., Buerkle, C., Buggs, R., et al. (2013). Hybridization and speciation. *Journal of Evolutionary Biology*, 26(2):229–246.

[Aho et al., 1972] Aho, A. V., Garey, M. R., and Ullman, J. D. (1972). The transitive reduction of a directed graph. *SIAM Journal on Computing*, 1(2):131–137.

[Alexander et al., 2009] Alexander, D. H., Novembre, J., and Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. *Genome Research*, 19(9):1655–1664.

[Alfaro et al., 2003] Alfaro, M. E., Zoller, S., and Lutzoni, F. (2003). Bayes or bootstrap? A simulation study comparing the performance of Bayesian Markov chain Monte Carlo sampling and bootstrapping in assessing phylogenetic confidence. *Molecular Biology and Evolution*, 20(2):255–266.

[Alifrangis et al., 2014] Alifrangis, M., Nag, S., Schousboe, M. L., Ishengoma, D., Lusingu, J., Pota, H., Kavishe, R. A., Pearce, R., Ord, R., Lynch, C., et al. (2014). Independent origin of *Plasmodium falciparum* antifolate super-resistance, Uganda, Tanzania, and Ethiopia. *Emerging Infectious Diseases*, 20(8):1280.

[Altshuler et al., 2008] Altshuler, D., Daly, M. J., and Lander, E. S. (2008). Genetic mapping in human disease. *Science*, 322(5903):881–888.

[Amaratunga et al., 2016] Amaratunga, C., Lim, P., Suon, S., Sreng, S., Mao, S., Sopha, C., Sam, B., Dek, D., Try, V., Amato, R., et al. (2016). Dihydroartemisinin–piperaquine resistance in *Plasmodium falciparum* malaria in Cambodia: a multisite prospective cohort study. *The Lancet infectious diseases*, 16(3):357–365.

[Amaratunga et al., 2012] Amaratunga, C., Sreng, S., Suon, S., Phelps, E. S., Stepniewska, K., Lim, P., Zhou, C., Mao, S., Anderson, J. M., Lindegardh, N., et al. (2012). Artemisinin-resistant

Plasmodium falciparum in Pursat province, western Cambodia: a parasite clearance rate study. *The Lancet Infectious Diseases*, 12(11):851–858.

[Anderson et al., 2011] Anderson, T., Nkhoma, S., Ecker, A., and Fidock, D. (2011). How can we identify parasite genes that underlie antimalarial drug resistance? *Pharmacogenomics*, 12(1):59–85.

[Anderson et al., 2000] Anderson, T. J., Haubold, B., Williams, J. T., Estrada-Franco, J. G., Richardson, L., Mollinedo, R., Bockarie, M., Mokili, J., Mharakurwa, S., French, N., et al. (2000). Microsatellite markers reveal a spectrum of population structures in the malaria parasite *Plasmodium falciparum*. *Molecular Biology and Evolution*, 17(10):1467–1482.

[Ariey et al., 2014] Ariey, F., Witkowski, B., Amaratunga, C., Beghain, J., Langlois, A.-C., Khim, N., Kim, S., Duru, V., Bouchier, C., Ma, L., et al. (2014). A molecular marker of artemisinin-resistant *Plasmodium falciparum* malaria. *Nature*, 505(7481):50–55.

[Arnot, 1998] Arnot, D. (1998). Unstable malaria in Sudan: the influence of the dry season: clone multiplicity of *Plasmodium falciparum* infections in individuals exposed to variable levels of disease transmission. *Transactions of the Royal Society of Tropical Medicine and Hygiene*, 92(6):580–585.

[Ashley et al., 2014] Ashley, E. A., Dhorda, M., Fairhurst, R. M., Amaratunga, C., Lim, P., Suon, S., Sreng, S., Anderson, J. M., Mao, S., Sam, B., et al. (2014). Spread of artemisinin resistance in *Plasmodium falciparum* malaria. *New England Journal of Medicine*, 371(5):411–423.

[Assefa et al., 2015] Assefa, S., Lim, C., Preston, M. D., Duffy, C. W., Nair, M. B., Adroub, S. A., Kadir, K. A., Goldberg, J. M., Neafsey, D. E., Divis, P., et al. (2015). Population genomic structure and adaptation in the zoonotic malaria parasite *Plasmodium knowlesi*. *Proceedings of the National Academy of Sciences*, 112(42):13027–13032.

[Auburn et al., 2012] Auburn, S., Campino, S., Miotto, O., Djimde, A. A., Zongo, I., Manske, M., Maslen, G., Mangano, V., Alcock, D., MacInnis, B., et al. (2012). Characterization of within-host *Plasmodium falciparum* diversity using next-generation sequence data. *PLoS One*, 7(2):e32891.

[Aurrecoechea et al., 2009] Aurrecoechea, C., Brestelli, J., Brunk, B. P., Dommer, J., Fischer, S., Gajria, B., Gao, X., Gingle, A., Grant, G., Harb, O. S., et al. (2009). PlasmoDB: a functional genomic database for malaria parasites. *Nucleic Acids Research*, 37(suppl 1):D539–D543.

[Beez et al., 2011] Beez, D., Sanchez, C. P., Stein, W. D., and Lanzer, M. (2011). Genetic predisposition favors the acquisition of stable artemisinin resistance in malaria parasites. *Antimicrobial Agents and Chemotherapy*, 55(1):50–55.

[Bhatia et al., 2013] Bhatia, G., Patterson, N., Sankararaman, S., and Price, A. L. (2013). Estimating and interpreting FST: the impact of rare variants. *Genome Research*, 23(9):1514–1521.

[Blanford et al., 2013] Blanford, J. I., Blanford, S., Crane, R. G., Mann, M. E., Paaijmans, K. P., Schreiber, K. V., and Thomas, M. B. (2013). Implications of temperature variation for malaria parasite development across Africa. *Scientific Reports*, 3:1300.

[Bloland et al., 2001] Bloland, P. B., Organization, W. H., et al. (2001). *Drug resistance in malaria*. World Health Organization Geneva.

[Borges-Walmsley et al., 2003] Borges-Walmsley, M. I., McKeegan, K. S., and Walmsley, A. R. (2003). Structure and function of efflux pumps that confer resistance to drugs. *Biochemical Journal*, 376(2):313–338.

[Bosman et al., 2014] Bosman, P., Stassijns, J., Nackers, F., Canier, L., Kim, N., Khim, S., Alipon, S. C., Char, M. C., Chea, N., Dysoley, L., et al. (2014). *Plasmodium* prevalence and artemisinin-resistant *falciparum* malaria in Preah Vihear Province, Cambodia: a cross-sectional population-based study. *Malaria Journal*, 13:394.

[Bousema et al., 2014] Bousema, T., Okell, L., Felger, I., and Drakeley, C. (2014). Asymptomatic malaria infections: detectability, transmissibility and public health relevance. *Nature Reviews Microbiology*, 12(12):833–840.

[Boyd, 1941] Boyd, M. F. (1941). An historical sketch of the prevalence of malaria in North America. *The American Journal of Tropical Medicine and Hygiene*, 1(2):223–244.

[Branch et al., 2001] Branch, O. H., Takala, S., Kariuki, S., Nahlen, B. L., Kolczak, M., Hawley, W., and Lal, A. A. (2001). *Plasmodium falciparum* genotypes, low complexity of infection, and resistance to subsequent malaria in participants in the Asembo Bay Cohort Project. *Infection and Immunity*, 69(12):7783–7792.

[Broder and Mitzenmacher, 2004] Broder, A. and Mitzenmacher, M. (2004). Network applications of bloom filters: A survey. *Internet Mathematics*, 1(4):485–509.

[Brown et al., 2010] Brown, K. M., Costanzo, M. S., Xu, W., Roy, S., Lozovsky, E. R., and Hartl, D. L. (2010). Compensatory mutations restore fitness during the evolution of dihydrofolate reductase. *Molecular Biology and Evolution*, 27(12):2682–2690.

[Browning and Browning, 2011] Browning, B. L. and Browning, S. R. (2011). A fast, powerful method for detecting identity by descent. *The American Journal of Human Genetics*, 88(2):173–182.

[Browning, 2008] Browning, S. R. (2008). Estimation of pairwise identity by descent from dense genetic marker data in a population sample of haplotypes. *Genetics*, 178(4):2123–2132.

[Browning and Browning, 2012] Browning, S. R. and Browning, B. L. (2012). Identity by descent between distant relatives: detection and applications. *Annual Review of Genetics*, 46:617–633.

[Camara et al., 2015] Camara, P. G., Levine, A. J., and Rabadan, R. (2015). Inference of ancestral recombination graphs through topological data analysis. *arXiv preprint arXiv:1505.05815*.

[Campino et al., 2011] Campino, S., Auburn, S., Kivinen, K., Zongo, I., Ouedraogo, J.-B., Mangano, V., Djimde, A., Doumbo, O. K., Kiara, S. M., Nzila, A., et al. (2011). Population genetic analysis of *Plasmodium falciparum* parasites using a customized Illumina GoldenGate genotyping assay. *PLoS One*, 6(6):e20251.

[Carme et al., 1993] Carme, B., Bouquety, J., and Plassart, H. (1993). Mortality and sequelae due to cerebral malaria in African children in Brazzaville, Congo. *The American Journal of Tropical Medicine and Hygiene*, 48(2):216–221.

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

[Cavalli-Sforza et al., 1994] Cavalli-Sforza, L. L., Menozzi, P., and Piazza, A. (1994). *The history and geography of human genes*. Princeton university press.

[Chapman and Thompson, 2003] Chapman, N. and Thompson, E. (2003). A model for the length of tracts of identity by descent in finite random mating populations. *Theoretical Population Biology*, 64(2):141–150.

[Charlesworth and Charlesworth, 2016] Charlesworth, B. and Charlesworth, D. (2016). Population genetics from 1966 to 2016. *Heredity*, 118(February):1–8.

[Claessens et al., 2014] Claessens, A., Hamilton, W. L., Kekre, M., Otto, T. D., Faizullabhoy, A., Rayner, J. C., and Kwiatkowski, D. (2014). Generation of antigenic diversity in *Plasmodium falciparum* by structured rearrangement of *Var* genes during mitosis. *PLoS Genet*, 10(12):e1004812.

[Cockerham, 1969] Cockerham, C. C. (1969). Variance of gene frequencies. *Evolution*, 23(1):72–84.

[Cowman et al., 2012] Cowman, A. F., Berry, D., and Baum, J. (2012). The cellular and molecular basis for malaria parasite invasion of the human red blood cell. *The Journal of Cell Biology*, 198(6):961–971.

[Cowman et al., 1988] Cowman, A. F., Morry, M. J., Biggs, B. A., Cross, G., and Foote, S. J. (1988). 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.

[Cui et al., 2015] Cui, L., Mharakurwa, S., Ndiaye, D., Rathod, P. K., and Rosenthal, P. J. (2015). Antimalarial Drug Resistance: Literature Review and Activities and Findings of the ICEMR Network. *The American Journal of Tropical Medicine and Hygiene*, 93(3_Suppl):57–68.

[Daniels et al., 2008] Daniels, R., Volkman, S. K., Milner, D. A., Mahesh, N., Neafsey, D. E., Park, D. J., Rosen, D., Angelino, E., Sabeti, P. C., Wirth, D. F., et al. (2008). A general SNP-based molecular barcode for *Plasmodium falciparum* identification and tracking. *Malaria Journal*, 7(1):1–11.

[Day and Edelsbrunner, 1984] Day, W. H. and Edelsbrunner, H. (1984). Efficient algorithms for agglomerative hierarchical clustering methods. *Journal of Classification*, 1(1):7–24.

[De Moraes et al., 2014] De Moraes, C. M., Stanczyk, N. M., Betz, H. S., Pulido, H., Sim, D. G., Read, A. F., and Mescher, M. C. (2014). Malaria-induced changes in host odors enhance mosquito attraction. *Proceedings of the National Academy of Sciences*, 111(30):11079–11084.

[de Roode et al., 2004] de Roode, J. C., Culleton, R., Cheesman, S. J., Carter, R., and Read, A. F. (2004). Host heterogeneity is a determinant of competitive exclusion or coexistence in genetically diverse malaria infections. *Proceedings of the Royal Society of London, Series B: Biological Sciences*, 271(1543):1073–1080.

[de Roode et al., 2005] de Roode, J. C., Pansini, R., Cheesman, S. J., Helinski, M. E., Huijben, S., Wargo, A. R., Bell, A. S., Chan, B. H., Walliker, D., and Read, A. F. (2005). Virulence and competitive ability in genetically diverse malaria infections. *Proceedings of the National Academy of Sciences of the United States of America*, 102(21):7624–7628.

[de Zulueta, 1988] de Zulueta, J. (1988). Report on a field mission in Madagascar, July 28–Sept 13, 1988. *World Health Organization, Geneva, Switzerland*.

[Decker et al., 2014] Decker, J. E., McKay, S. D., Rolf, M. M., Kim, J., Alcalá, A. M., Sonstegard, T. S., Hanotte, O., Götherström, A., Seabury, C. M., Praharani, L., et al. (2014). Worldwide patterns of ancestry, divergence, and admixture in domesticated cattle. *PLoS Genet*, 10(3):e1004254.

[Dogovski et al., 2015] Dogovski, C., Xie, S. C., Burgio, G., Bridgford, J., Mok, S., McCaw, J. M., Chotivanich, K., Kenny, S., Gnädig, N., Straimer, J., et al. (2015). Targeting the cell stress response of *Plasmodium falciparum* to overcome artemisinin resistance. *PLoS Biol*, 13(4):e1002132.

[Dondorp et al., 2009] Dondorp, A. M., Nosten, F., Yi, P., Das, D., Phyo, A. P., Tarning, J., Lwin, K. M., Ariey, F., Hanpithakpong, W., Lee, S. J., et al. (2009). Artemisinin resistance in *Plasmodium falciparum* malaria. *New England Journal of Medicine*, 361(5):455–467.

[Dondorp and Ringwald, 2013] Dondorp, A. M. and Ringwald, P. (2013). Artemisinin resistance is a clear and present danger. *Trends in Parasitology*, 29(8):359–360.

[Dondorp et al., 2010] Dondorp, A. M., Yeung, S., White, L., Nguon, C., Day, N. P., Socheat, D., and von Seidlein, L. (2010). Artemisinin resistance: current status and scenarios for containment. *Nature Reviews Microbiology*, 8(4):272–280.

[Doolan et al., 2009] Doolan, D. L., Dobaño, C., and Baird, J. K. (2009). Acquired immunity to malaria. *Clinical Microbiology Reviews*, 22(1):13–36.

[Dowling et al., 1951] Dowling, M. et al. (1951). The malaria eradication scheme in Mauritius. *British Medical Bulletin*, 8(1):72–75.

[Durbin, 2014] Durbin, R. (2014). Efficient haplotype matching and storage using the positional Burrows–Wheeler transform (PBWT). *Bioinformatics*, 30(9):1266–1272.

[Dye and Williams, 1997] Dye, C. and Williams, B. G. (1997). Multigenic drug resistance among inbred malaria parasites. *Proceedings of the Royal Society of London B: Biological Sciences*, 264(1378):61–67.

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

[Efron, 1992] Efron, B. (1992). *Bootstrap methods: another look at the jackknife*. Springer.

[Elliott, 1972] Elliott, R. (1972). The influence of vector behavior on malaria transmission. *The American Journal of Tropical Medicine and Hygiene*, 21(5 Suppl):755–763.

[Escalante et al., 2004] Escalante, A. A., Cornejo, O. E., Rojas, A., Udhayakumar, V., and Lal, A. A. (2004). Assessing the effect of natural selection in malaria parasites. *Trends in Parasitology*, 20(8):388–395.

[Evanno et al., 2005] Evanno, G., Regnaut, S., and Goudet, J. (2005). Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Molecular Ecology*, 14(8):2611–2620.

[Falush et al., 2003] Falush, D., Stephens, M., and Pritchard, J. K. (2003). Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics*, 164(4):1567–1587.

[Felsenstein, 2004] Felsenstein, J. (2004). *Inferring phylogenies*. Sinauer Associates Sunderland.

[Fisher, 1930] Fisher, R. A. (1930). *The genetical theory of natural selection: a complete variorum edition*. Oxford University Press.

[Fivelman et al., 2002] Fivelman, Q. L., Butcher, G. A., Adagu, I. S., Warhurst, D. C., and Pasvol, G. (2002). Malarone treatment failure and in vitro confirmation of resistance of *Plasmodium falciparum* isolate from Lagos, Nigeria. *Malaria Journal*, 1(1):1–4.

[Flegg et al., 2011] Flegg, J. A., Guerin, P. J., White, N. J., and Stepniewska, K. (2011). Standardizing the measurement of parasite clearance in *falciparum* malaria: the parasite clearance estimator. *Malaria Journal*, 10(1):1.

[Flint et al., 1998] Flint, J., Harding, R. M., Boyce, A. J., and Clegg, J. B. (1998). The population genetics of the haemoglobinopathies. *Baillière's Clinical Haematology*, 11(1):1–51.

[François et al., 2010] François, O., Currat, M., Ray, N., Han, E., Excoffier, L., and Novembre, J. (2010). Principal component analysis under population genetic models of range expansion and admixture. *Molecular Biology and Evolution*, 27(6):1257–1268.

[Galizi et al., 2014] Galizi, R., Doyle, L. A., Menichelli, M., Bernardini, F., Deredec, A., Burt, A., Stoddard, B. L., Windbichler, N., and Crisanti, A. (2014). A synthetic sex ratio distortion system for the control of the human malaria mosquito. *Nature Communications*, 5.

[Gallup and Sachs, 2001] Gallup, J. L. and Sachs, J. D. (2001). The economic burden of malaria. *The American Journal of Tropical Medicine and Hygiene*, 64(1 suppl):85–96.

[Gandolfo et al., 2014] Gandolfo, L. C., Bahlo, M., and Speed, T. P. (2014). Dating rare mutations from small samples with dense marker data. *Genetics*, 197(4):1315–1327.

[Gardner et al., 2002] Gardner, M. J., Hall, N., Fung, E., White, O., Berriman, M., Hyman, R. W., Carlton, J. M., Pain, A., Nelson, K. E., Bowman, S., et al. (2002). Genome sequence of the human malaria parasite *Plasmodium falciparum*. *Nature*, 419(6906):498–511.

[Gascuel and Steel, 2006] Gascuel, O. and Steel, M. (2006). Neighbor-joining revealed. *Molecular Biology and Evolution*, 23(11):1997–2000.

[Ghedin et al., 2005] Ghedin, E., Sengamalay, N. A., Shumway, M., Zaborsky, J., Feldblyum, T., Subbu, V., Spiro, D. J., Sitz, J., Koo, H., Bolotov, P., et al. (2005). Large-scale sequencing of human influenza reveals the dynamic nature of viral genome evolution. *Nature*, 437(7062):1162–1166.

[Ghorbal et al., 2014] Ghorbal, M., Gorman, M., Macpherson, C. R., Martins, R. M., Scherf, A., and Lopez-Rubio, J.-J. (2014). Genome editing in the human malaria parasite Plasmodium falciparum using the CRISPR-Cas9 system. *Nature Biotechnology*, 32(8):819–821.

[Githeko et al., 1996] Githeko, A. K., Adungo, N. I., Karanja, D. M., Hawley, W. A., Vulule, J. M., Seroney, I. K., Ofulla, A. V., Atieli, F. K., Ondijo, S. O., Genga, I. O., et al. (1996). Some observations on the biting behavior of *Anopheles gambiae ss*, *Anopheles arabiensis*, and *Anopheles funestus* and their implications for malaria control. *Experimental Parasitology*, 82(3):306–315.

[Gower, 1966] Gower, J. C. (1966). Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika*, 53(3-4):325–338.

[Greenwood et al., 2008] Greenwood, B. M., Fidock, D. A., Kyle, D. E., Kappe, S. H., Alonso, P. L., Collins, F. H., and Duffy, P. E. (2008). Malaria: progress, perils, and prospects for eradication. *The Journal of Clinical Investigation*, 118(4):1266–1276.

[Gusev et al., 2009] Gusev, A., Lowe, J. K., Stoffel, M., Daly, M. J., Altshuler, D., Breslow, J. L., Friedman, J. M., and Pe' er, I. (2009). Whole population, genome-wide mapping of hidden relatedness. *Genome Research*, 19(2):318–326.

[Gusfield, 2014] Gusfield, D. (2014). *ReCombinatorics: the algorithmics of ancestral recombination graphs and explicit phylogenetic networks*. MIT Press.

[Gutenkunst et al., 2009] Gutenkunst, R. N., Hernandez, R. D., Williamson, S. H., and Bustamante, C. D. (2009). Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLoS Genet*, 5(10):e1000695.

[Haldane, 1949] Haldane, J. S. (1949). The rate of mutation of human genes. *Hereditas*, 35(S1):267–273.

[Hamilton, 2011] Hamilton, M. (2011). *Population genetics*. John Wiley & Sons.

[Harris et al., 2010] Harris, I., Sharrock, W. W., Bain, L. M., Gray, K.-A., Bobogare, A., Boaz, L., Lilley, K., Krause, D., Vallely, A., Johnson, M.-L., et al. (2010). A large proportion of asymptomatic *Plasmodium* infections with low and sub-microscopic parasite densities in the low transmission setting of Temotu Province, Solomon Islands: challenges for malaria diagnostics in an elimination setting. *Malaria Journal*, 9(254):10–1186.

[Hartl and Clark, 2007] Hartl, D. L. and Clark, A. G. (2007). *Principles of Population Genetics*. Sinauer.

[Hasegawa et al., 1985] Hasegawa, M., Kishino, H., and Yano, T.-a. (1985). Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *Journal of Molecular Evolution*, 22(2):160–174.

[Hastings and D' Alessandro, 2000] Hastings, I. and D' Alessandro, U. (2000). Modelling a predictable disaster: The rise and spread of drug-resistant malaria. *Parasitology Today*, 16(8):340–347.

[Hastings, 2004] Hastings, I. M. (2004). The origins of antimalarial drug resistance. *Trends in Parasitology*, 20(11):512–518.

[Hay et al., 2009] Hay, S. I., Guerra, C. A., Gething, P. W., Patil, A. P., Tatem, A. J., Noor, A. M., Kabaria, C. W., Manh, B. H., Elyazar, I. R., Brooker, S., et al. (2009). A world malaria map: *Plasmodium falciparum* endemicity in 2007. *PLoS Med*, 6(3):e1000048.

[Hay et al., 2004] Hay, S. I., Guerra, C. A., Tatem, A. J., Noor, A. M., and Snow, R. W. (2004). The global distribution and population at risk of malaria: past, present, and future. *The Lancet Infectious Diseases*, 4(6):327–336.

[Hedrick, 2011] Hedrick, P. (2011). Population genetics of malaria resistance in humans. *Heredity*, 107(4):283–304.

[Hermisson and Pennings, 2005] Hermisson, J. and Pennings, P. S. (2005). Soft sweeps molecular population genetics of adaptation from standing genetic variation. *Genetics*, 169(4):2335–2352.

[Hien et al., 2012] Hien, T. T., Thuy-Nhien, N. T., Phu, N. H., Boni, M. F., Thanh, N. V., Nha-Ca, N. T., Thai, L. H., Thai, C. Q., Toi, P. V., Thuan, P. D., et al. (2012). In vivo susceptibility of *Plasmodium falciparum* to artesunate in Binh Phuoc Province, Vietnam. *Malaria Journal*, 11(1):1–11.

[Holding and Snow, 2001] Holding, P. A. and Snow, R. W. (2001). Impact of *Plasmodium falciparum* malaria on performance and learning: review of the evidence. *The American Journal of Tropical Medicine and Hygiene*, 64(1 suppl):68–75.

[Hudson and Kaplan, 1985] Hudson, R. R. and Kaplan, N. L. (1985). Statistical properties of the number of recombination events in the history of a sample of DNA sequences. *Genetics*, 111(1):147–164.

[Hudson et al., 1992] Hudson, R. R., Slatkin, M., and Maddison, W. (1992). Estimation of levels of gene flow from DNA sequence data. *Genetics*, 132:583–589.

[Johnson, 1967] Johnson, S. C. (1967). Hierarchical clustering schemes. *Psychometrika*, 32(3):241–254.

[Jolliffe, 2002] Jolliffe, I. (2002). *Principal component analysis*. Wiley Online Library.

[Jost, 2008] Jost, L. (2008). GST and its relatives do not measure differentiation. *Molecular Ecology*, 17(18):4015–4026.

[Kalinowski, 2009] Kalinowski, S. (2009). How well do evolutionary trees describe genetic relationships among populations. *Heredity*, 102(5):506–513.

[Kamau et al., 2014] Kamau, E., Campino, S., Amenga-Etego, L., Drury, E., Ishengoma, D., Johnson, K., Mumba, D., Kekre, M., William, Y., Mead, D., et al. (2014). K13-propeller polymorphisms in *Plasmodium falciparum* parasites from sub-Saharan Africa. *Journal of Infectious Diseases*, page 608.

[Kaufman and Rousseeuw, 1987] Kaufman, L. and Rousseeuw, P. (1987). *Clustering by means of medoids*. North-Holland.

[Keinan et al., 2007] Keinan, A., Mullikin, J. C., Patterson, N., and Reich, D. (2007). Measurement of the human allele frequency spectrum demonstrates greater genetic drift in East Asians than in Europeans. *Nature Genetics*, 39(10):1251–1255.

[Knols et al., 2007] Knols, B. G., Bossin, H. C., Mukabana, W. R., and Robinson, A. S. (2007). Transgenic mosquitoes and the fight against malaria: managing technology push in a turbulent GMO world. *The American Journal of Tropical Medicine and Hygiene*, 77(6 Suppl):232–242.

[Knowler et al., 1988] Knowler, W. C., Williams, R., Pettitt, D., and Steinberg, A. G. (1988). Gm3; 5, 13, 14 and type 2 diabetes mellitus: an association in American Indians with genetic admixture. *American Journal of Human Genetics*, 43(4):520.

[Koenderink et al., 2010] Koenderink, J. B., Kavishe, R. A., Rijpma, S. R., and Russel, F. G. (2010). The ABCs of multidrug resistance in malaria. *Trends in Parasitology*, 26(9):440–446.

[Kumar, 2005] Kumar, S. (2005). Molecular clocks: four decades of evolution. *Nature Reviews Genetics*, 6(8):654–662.

[Kwiatkowski, 2005] Kwiatkowski, D. P. (2005). How malaria has affected the human genome and what human genetics can teach us about malaria. *The American Journal of Human Genetics*, 77(2):171–192.

[Kyes et al., 2001] Kyes, S., Horrocks, P., and Newbold, C. (2001). Antigenic variation at the infected red cell surface in malaria. *Annual Reviews in Microbiology*, 55(1):673–707.

[Laishram et al., 2012] Laishram, D. D., Sutton, P. L., Nanda, N., Sharma, V. L., Sobti, R. C., Carlton, J. M., and Joshi, H. (2012). The complexities of malaria disease manifestations with a focus on asymptomatic malaria. *Malaria Journal*, 11(1):1.

[Lakshmanan et al., 2005] Lakshmanan, V., Bray, P. G., Verdier-Pinard, D., Johnson, D. J., Horrocks, P., Muhle, R. A., Alakpa, G. E., Hughes, R. H., Ward, S. A., Krogstad, D. J., et al. (2005). A critical role for PfCRT K76T in *Plasmodium falciparum* verapamil-reversible chloroquine resistance. *The EMBO Journal*, 24(13):2294–2305.

[Landry and Aubin-Horth, 2013] Landry, C. R. and Aubin-Horth, N. (2013). *Ecological Genomics: Ecology and the Evolution of Genes and Genomes*, volume 781. Springer Science & Business Media.

[Lawson and Falush, 2012] Lawson, D. and Falush, D. (2012). Population identification using genetic data. *Annual Review of Genomics and Human Genetics*, 13:337–361.

[Lawson, 2012] Lawson, D. J. (2012). Populations in statistical genetic modelling and inference. *arXiv preprint arXiv:1306.0701*.

[Lawson et al., 2012] Lawson, D. J., Hellenthal, G., Myers, S., and Falush, D. (2012). Inference of population structure using dense haplotype data. *PLoS Genet*, 8(1):e1002453.

[Lemey, 2009] Lemey, P. (2009). *The phylogenetic handbook: a practical approach to phylogenetic analysis and hypothesis testing*. Cambridge University Press.

[Lenormand, 2002] Lenormand, T. (2002). Gene flow and the limits to natural selection. *Trends in Ecology & Evolution*, 17(4):183–189.

[Li and Durbin, 2009] Li, H. and Durbin, R. (2009). Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics*, 25(14):1754–1760.

[Li and Stephens, 2003] Li, N. and Stephens, M. (2003). Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. *Genetics*, 165(4):2213–2233.

[Lim et al., 2013] Lim, P., Dek, D., Try, V., Eastman, R. T., Chy, S., Sreng, S., Suon, S., Mao, S., Sopha, C., Sam, B., et al. (2013). Ex vivo susceptibility of *Plasmodium falciparum* to antimalarial drugs in western, northern, and eastern Cambodia, 2011-2012: association with molecular markers. *Antimicrobial Agents and Chemotherapy*, 57(11):5277–5283.

[Listgarten et al., 2012] Listgarten, J., Lippert, C., Kadie, C. M., Davidson, R. I., Eskin, E., and Heckerman, D. (2012). Improved linear mixed models for genome-wide association studies. *Nature Methods*, 9(6):525–526.

[Litsios, 1996] Litsios, S. (1996). *The tomorrow of malaria.* Pacific Press.

[Lukashov et al., 1998] Lukashov, V., Karamov, E., Eremin, V., Titov, L., and Goudsmit, J. (1998). Extreme founder effect in an HIV type 1 subtype A epidemic among drug users in Svetlogorsk, Belarus. *AIDS research and human retroviruses*.

[MacKay, 2003] MacKay, D. J. (2003). *Information theory, inference and learning algorithms*. Cambridge university press.

[MacQueen et al., 1967] MacQueen, J. et al. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281–297. Oakland, CA, USA.

[Maimon and Rokach, 2005] Maimon, O. and Rokach, L. (2005). *Data mining and knowledge discovery handbook*, volume 2. Springer.

[MalariaGEN-Pf-Community-Project, 2016] MalariaGEN-Pf-Community-Project (2016). Genomic epidemiology of artemisinin resistant malaria. *eLife*, 5:e08714.

[Malécot, 1948] Malécot, G. (1948). Mathématiques de l'hérédit.

[Mallet, 2001] Mallet, J. (2001). *Gene flow in Insect movement: mechanisms and consequences*. CABI.

[Manske and Kwiatkowski, 2009] Manske, H. M. and Kwiatkowski, D. P. (2009). SNP-o-matic. *Bioinformatics*, 25(18):2434–2435.

[Manske et al., 2012] Manske, M., Miotto, O., Campino, S., Auburn, S., Almagro-Garcia, J., Maslen, G., OÃŢBrien, J., Djimde, A., Doumbo, O., Zongo, I., et al. (2012). Analysis of *Plasmodium falciparum* diversity in natural infections by deep sequencing. *Nature*, 487(7407):375–379.

[Marchini et al., 2004] Marchini, J., Cardon, L. R., Phillips, M. S., and Donnelly, P. (2004). The effects of human population structure on large genetic association studies. *Nature Genetics*, 36(5):512–517.

[Marshall and Taylor, 2009] Marshall, J. M. and Taylor, C. E. (2009). Malaria control with transgenic mosquitoes. *PLoS Med*, 6(2):e1000020.

[Marth et al., 2004] Marth, G. T., Czabarka, E., Murvai, J., and Sherry, S. T. (2004). The allele frequency spectrum in genome-wide human variation data reveals signals of differential demographic history in three large world populations. *Genetics*, 166(1):351–372.

[Martin and Kirk, 2004] Martin, R. E. and Kirk, K. (2004). The malaria parasiteś chloroquine resistance transporter is a member of the drug/metabolite transporter superfamily. *Molecular biology and evolution*, 21(10):1938–1949.

[Mathieson and McVean, 2014] Mathieson, I. and McVean, G. (2014). Demography and the age of rare variants. *PLoS Genet*, 10(8):e1004528.

[Mbengue et al., 2015] Mbengue, A., Bhattacharjee, S., Pandharkar, T., Liu, H., Estiu, G., Stahelin, R. V., Rizk, S. S., Njimoh, D. L., Ryan, Y., Chotivanich, K., et al. (2015). A molecular mechanism of artemisinin resistance in *Plasmodium falciparum* malaria. *Nature*, 520(7549):683–687.

[McVean, 2009] McVean, G. (2009). A genealogical interpretation of principal components analysis. *PLoS Genet*, 5(10):e1000686.

[Mendis et al., 1991] Mendis, K. N., David, P. H., and Carter, R. (1991). Antigenic polymorphism in malaria: Is it an important mechanism for immune evasion? *Immunology Today*, 12(3):A34–A37.

[Messer and Petrov, 2013] Messer, P. W. and Petrov, D. A. (2013). Population genomics of rapid adaptation by soft selective sweeps. *Trends in Ecology & Evolution*, 28(11):659–669.

[Mills et al., 2008] Mills, A., Lubell, Y., and Hanson, K. (2008). Malaria eradication: the economic, financial and institutional challenge. *Malaria Journal*, 7(1):1.

[Miotto et al., 2013] Miotto, O., Almagro-Garcia, J., Manske, M., MacInnis, B., Campino, S., Rockett, K. A., Amaratunga, C., Lim, P., Suon, S., Sreng, S., et al. (2013). Multiple populations of artemisinin-resistant *Plasmodium falciparum* in Cambodia. *Nature Genetics*, 45(6):648–655.

[Miotto et al., 2015] Miotto, O., Amato, R., Ashley, E. A., MacInnis, B., Almagro-Garcia, J., Amaratunga, C., Lim, P., Mead, D., Oyola, S. O., Dhorda, M., et al. (2015). Genetic architecture of artemisinin-resistant *Plasmodium falciparum*. *Nature Genetics*, 47(3):226–234.

[Mita et al., 2009] Mita, T., Tanabe, K., and Kita, K. (2009). Spread and evolution of *Plasmodium falciparum* drug resistance. *Parasitology International*, 58(3):201–209.

[Mobegi et al., 2014] Mobegi, V. A., Duffy, C. W., Amambua-Ngwa, A., Loua, K. M., Laman, E., Nwakanma, D. C., MacInnis, B., Aspeling-Jones, H., Murray, L., Clark, T. G., et al. (2014). Genome-wide analysis of selection on the malaria parasite *Plasmodium falciparum* in West African populations of differing infection endemicity. *Molecular Biology and Evolution*, 31(6):1490–1499.

[Mot et al., 2015] Mot, A. C., Bischin, C., Damian, G., and Silaghi-Dumitrescu, R. (2015). Antioxidant activity evaluation involving hemoglobin-related free radical reactivity. *Advanced Protocols in Oxidative Stress*, 3:247–255.

[Mu et al., 2005] Mu, J., Awadalla, P., Duan, J., McGee, K. M., Joy, D. A., McVean, G. A., and Su, X.-z. (2005). Recombination hotspots and population structure in *Plasmodium falciparum*. *PLoS Biol*, 3(10):e335.

[Müller et al., 2001] Müller, D., Charlwood, J., Felger, I., Ferreira, C., Do Rosario, V., and Smith, T. (2001). Prospective risk of morbidity in relation to multiplicity of infection with *Plasmodium falciparum* in Sao Tome. *Acta tropica*, 78(2):155–162.

[Muriu et al., 2008] Muriu, S. M., Muturi, E. J., Shililu, J. I., Mbogo, C. M., Mwangangi, J. M., Jacob, B. G., Irungu, L. W., Mukabana, R. W., Githure, J. I., and Novak, R. J. (2008). Host choice and multiple blood feeding behaviour of malaria vectors and other anophelines in Mwea rice scheme, Kenya. *Malaria Journal*, 7(Suppl 1):S4.

[Murray et al., 2016] Murray, L., Mobegi, V. A., Duffy, C. W., Assefa, S. A., Kwiatkowski, D. P., Laman, E., Loua, K. M., and Conway, D. J. (2016). Microsatellite genotyping and genome-wide single nucleotide polymorphism-based indices of *Plasmodium falciparum* diversity within clinical infections. *Malaria Journal*, 15(1):1.

[Nair et al., 2007] Nair, S., Nash, D., Sudimack, D., Jaidee, A., Barends, M., Uhlemann, A.-C., Krishna, S., Nosten, F., and Anderson, T. J. (2007). Recurrent gene amplification and soft selective sweeps during evolution of multidrug resistance in malaria parasites. *Molecular Biology and Evolution*, 24(2):562–573.

[Nájera et al., 2011] Nájera, J. A., González-Silva, M., and Alonso, P. L. (2011). Some lessons for the future from the Global Malaria Eradication Programme (1955–1969). *PLoS Med*, 8(1).

[Needham, 1960] Needham, J. (1960). Parasites and parasitic infections in early medicine and science. *Medical history*, 4(3):262.

[Neghina et al., 2010] Neghina, R., Neghina, A. M., Marincu, I., and Iacobiciu, I. (2010). Malaria, a journey in time: in search of the lost myths and forgotten stories. *The American Journal of the Medical Sciences*, 340(6):492–498.

[Newton et al., 2003] Newton, P. N., Dondorp, A., Green, M., Mayxay, M., and White, N. J. (2003). Counterfeit artesunate antimalarials in southeast Asia. *The Lancet*, 362(9378):169.

[Ng et al., 2002] Ng, A. Y., Jordan, M. I., Weiss, Y., et al. (2002). On spectral clustering: Analysis and an algorithm. *Advances in Neural Information Processing Systems*, 2:849–856.

[Nkhoma et al., 2013] Nkhoma, S. C., Nair, S., Al-Saai, S., Ashley, E., McGready, R., Phyo, A. P., Nosten, F., and Anderson, T. J. (2013). Population genetic correlates of declining transmission in a human pathogen. *Molecular Ecology*, 22(2):273–285.

[Nock and Nielsen, 2006] Nock, R. and Nielsen, F. (2006). On weighting clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(8):1223–1235.

[Noedl et al., 2008] Noedl, H., Se, Y., Schaecher, K., Smith, B. L., Socheat, D., and Fukuda, M. M. (2008). Evidence of artemisinin-resistant malaria in western Cambodia. *New England Journal of Medicine*, 359(24):2619–2620.

[Novembre et al., 2008] Novembre, J., Johnson, T., Bryc, K., Kutalik, Z., Boyko, A. R., Auton, A., Indap, A., King, K. S., Bergmann, S., Nelson, M. R., et al. (2008). Genes mirror geography within europe. *Nature*, 456(7218):98–101.

[Novembre and Stephens, 2008] Novembre, J. and Stephens, M. (2008). Interpreting principal component analyses of spatial population genetic variation. *Nature Genetics*, 40(5):646–649.

[Okell et al., 2012] Okell, L. C., Bousema, T., Griffin, J. T., Ouédraogo, A. L., Ghani, A. C., and Drakeley, C. J. (2012). Factors determining the occurrence of submicroscopic malaria infections and their relevance for control. *Nature Communications*, 3:1237.

[Okumu et al., 2011] Okumu, F. O., Moore, S. J., Okumu, F., and Moore, S. (2011). Combining indoor residual spraying and insecticide-treated nets for malaria control in Africa: a review of possible outcomes and an outline of suggestions for the future. *Malaria Journal*, 10(1):208.

[Packard, 2007] Packard, R. M. (2007). *The making of a tropical disease: a short history of malaria*. JHU Press.

[Packard, 2014] Packard, R. M. (2014). The origins of antimalarial-drug resistance. *New England Journal of Medicine*, 371(5):397–399.

[Palamara et al., 2012] Palamara, P. F., Lencz, T., Darvasi, A., and Peér, I. (2012). Length distributions of identity by descent reveal fine-scale demographic history. *The American Journal of Human Genetics*, 91(5):809–822.

[Paradis et al., 2004] Paradis, E., Claude, J., and Strimmer, K. (2004). APE: analyses of phylogenetics and evolution in R language. *Bioinformatics*, 20(2):289–290.

[Patterson et al., 2006] Patterson, N., Price, A. L., and Reich, D. (2006). Population structure and eigenanalysis. *PLoS genet*, 2(12):e190.

[Payne, 1988] Payne, D. (1988). Did medicated salt hasten the spread of chloroquine resistance in *Plasmodium falciparum*? *Parasitology Today*, 4(4):112–115.

[Pennings and Hermisson, 2006] Pennings, P. S. and Hermisson, J. (2006). Soft sweeps II : molecular population genetics of adaptation from recurrent mutation or migration. *Molecular Biology and Evolution*, 23(5):1076–1084.

[Petersen et al., 2011] Petersen, I., Eastman, R., and Lanzer, M. (2011). Drug-resistant malaria: molecular mechanisms and implications for public health. *FEBS Letters*, 585(11):1551–1562.

[Petersen et al., 2015] Petersen, I., Gabryszewski, S. J., Johnston, G. L., Dhingra, S. K., Ecker, A., Lewis, R. E., Almeida, M. J., Straimer, J., Henrich, P. P., Palatulan, E., et al. (2015). Balancing drug resistance and growth rates via compensatory mutations in the *Plasmodium falciparum* chloroquine resistance transporter. *Molecular Microbiology*, 97(2):381–395.

[Peterson et al., 1988] Peterson, D. S., Walliker, D., and Wellems, T. E. (1988). Evidence that a point mutation in dihydrofolate reductase-thymidylate synthase confers resistance to pyrimethamine in *falciparum* malaria. *Proceedings of the National Academy of Sciences*, 85(23):9114–9118.

[Pf3k, 2016] Pf3k (2016). The pf3k project: pilot data release 5. `http://www.malariagen.net/data/pf3kÂŋ5`.

[Phyo et al., 2012] Phyo, A. P., Nkhoma, S., Stepniewska, K., Ashley, E. A., Nair, S., McGready, R., ler Moo, C., Al-Saai, S., Dondorp, A. M., Lwin, K. M., et al. (2012). Emergence of artemisinin-resistant malaria on the western border of Thailand: a longitudinal study. *The Lancet*, 379(9830):1960–1966.

[Plowe, 2003] Plowe, C. V. (2003). Monitoring antimalarial drug resistance: making the most of the tools at hand. *Journal of Experimental Biology*, 206(21):3745–3752.

[Plowe et al., 1997] Plowe, C. V., Cortese, J. F., Djimde, A., Nwanyanwu, O. C., Watkins, W. M., Winstanley, P. A., Franco, J. G. E., Mollinedo, R. E., Avila, J. C., Cespedes, J. L., et al. (1997). Mutations in *Plasmodium falciparum* dihydrofolate reductase and dihydropteroate synthase and epidemiologic patterns of pyrimethamine-sulfadoxine use and resistance. *Journal of Infectious Diseases*, 176(6):1590–1596.

[Pool and Nielsen, 2009] Pool, J. E. and Nielsen, R. (2009). Inference of historical changes in migration rate from the lengths of migrant tracts. *Genetics*, 181(2):711–719.

[Porras-Hurtado et al., 2013] Porras-Hurtado, L., Ruiz, Y., Santos, C., Phillips, C., Carracedo, A., and Lareu, M. V. (2013). An overview of STRUCTURE: applications, parameter settings, and supporting software. *Front Genet*, 4:98.

[Price et al., 2006] Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., and Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*, 38(8):904–909.

[Price et al., 2004] Price, R. N., Uhlemann, A.-C., Brockman, A., McGready, R., Ashley, E., Phaipun, L., Patel, R., Laing, K., Looareesuwan, S., White, N. J., et al. (2004). Mefloquine resistance in *Plasmodium falciparum* and increased *pfmdr1* gene copy number. *The Lancet*, 364(9432):438–447.

[Pritchard et al., 2000] Pritchard, J. K., Stephens, M., and Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics*, 155(2):945–959.

[Purcell et al., 2007] Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., Maller, J., Sklar, P., De Bakker, P. I., Daly, M. J., et al. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics*, 81(3):559–575.

[Rabiner, 1989] Rabiner, L. R. (1989). A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286.

[Rai et al., 2010] Rai, M. A., Nerurkar, V. R., Khoja, S., Khan, S., Yanagihara, R., Rehman, A., Kazmi, S. U., and Ali, S. H. (2010). Evidence for a founder effect among HIV-infected injection drug users (IDUs) in Pakistan. *BMC Infectious Diseases*, 10(1):7.

[Raj et al., 2014] Raj, A., Stephens, M., and Pritchard, J. K. (2014). fastSTRUCTURE: variational inference of population structure in large SNP data sets. *Genetics*, 197(2):573–589.

[Raj et al., 2009] Raj, D. K., Mu, J., Jiang, H., Kabat, J., Singh, S., Sullivan, M., Fay, M. P., McCutchan, T. F., and Su, X.-z. (2009). Disruption of a *Plasmodium falciparum* multidrug resistance-associated protein (PfMRP) alters its fitness and transport of antimalarial drugs and glutathione. *Journal of Biological Chemistry*, 284(12):7687–7696.

[Ramette, 2007] Ramette, A. (2007). Multivariate analyses in microbial ecology. *FEMS Microbiology Ecology*, 62(2):142–160.

[Rasmussen et al., 2014] Rasmussen, M. D., Hubisz, M. J., Gronau, I., and Siepel, A. (2014). Genome-wide inference of ancestral recombination graphs. *PLoS Genet*, 10(5):e1004342.

[Rathod et al., 1997] Rathod, P. K., McErlean, T., and Lee, P.-C. (1997). Variations in frequencies of drug resistance in *Plasmodium falciparum*. *Proceedings of the National Academy of Sciences*, 94(17):9389–9393.

[Rodrigues et al., 2010] Rodrigues, J., Brayner, F. A., Alves, L. C., Dixit, R., and Barillas-Mury, C. (2010). Hemocyte differentiation mediates innate immune memory in textitAnopheles gambiae mosquitoes. *Science*, 329(5997):1353–1355.

[Romi et al., 2002] Romi, R., Razaiarimanga, M., Raharimanga, R., Rakotondraibe, E., Ranaivo, L., Pietra, V., Raveloson, A., and Majori, G. (2002). Impact of the malaria control campaign (1993-1998) in the highlands of Madagascar: parasitological and entomological data. *The American Journal of Tropical Medicine and Hygiene*, 66(1):2–6.

[Roper et al., 2004] Roper, C., Pearce, R., Nair, S., Sharp, B., Nosten, F., and Anderson, T. (2004). Intercontinental spread of pyrimethamine-resistant malaria. *Science*, 305(5687):1124–1124.

[RTS-SCTP, 2015] RTS-SCTP (2015). Efficacy and safety of RTS, S/AS01 malaria vaccine with or without a booster dose in infants and children in Africa: final results of a phase 3, individually randomised, controlled trial. *The Lancet*, 386(9988):31–45.

[Rubin et al., 1981] Rubin, D. B. et al. (1981). The Bayesian bootstrap. *The Annals of Statistics*, 9(1):130–134.

[Ryman and Leimar, 2009] Ryman, N. and Leimar, O. (2009). GST is still a useful measure of genetic differentiation, comment on Jostś D. *Molecular Ecology*, 18(10):2084–2087.

[Sabeti et al., 2006] Sabeti, P., Schaffner, S., Fry, B., Lohmueller, J., Varilly, P., Shamovsky, O., Palma, A., Mikkelsen, T., Altshuler, D., and Lander, E. (2006). Positive natural selection in the human lineage. *Science*, 312(5780):1614–1620.

[Sachs and Malaney, 2002] Sachs, J. and Malaney, P. (2002). The economic and social burden of malaria. *Nature*, 415(6872):680–685.

[Saitou and Nei, 1987] Saitou, N. and Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*, 4(4):406–425.

[Salgueiro et al., 2010] Salgueiro, P., Vicente, J. L., Ferreira, C., Teófilo, V., Galvão, A., Rosário, V. E., Cravo, P., and Pinto, J. (2010). Tracing the origins and signatures of selection of antifolate resistance in island populations of *Plasmodium falciparum*. *BMC Infectious Diseases*, 10(1):1.

[Sallares, 2002] Sallares, R. (2002). *Malaria and Rome: a history of malaria in ancient Italy*. Oxford University Press.

[Saunders et al., 2014] Saunders, D. L., Vanachayangkul, P., and Lon, C. (2014). Dihydroartemisinin–piperaquine failure in Cambodia. *New England Journal of Medicine*, 371(5):484–485.

[Schierup and Hein, 2000] Schierup, M. H. and Hein, J. (2000). Recombination and the molecular clock. *Molecular Biology and Evolution*, 17(10):1578–1579.

[Schlagenhauf-Lawlor, 2007] Schlagenhauf-Lawlor, P. (2007). *Travelersḿalaria*. PMPH-USA.

[Schnell et al., 2004] Schnell, J. R., Dyson, H. J., and Wright, P. E. (2004). Structure, dynamics, and catalytic function of dihydrofolate reductase. *Annu. Rev. Biophys. Biomol. Struct.*, 33:119–140.

[Sen and Ferdig, 2004] Sen, S. and Ferdig, M. (2004). QTL analysis for discovery of genes involved in drug responses. *Current Drug Targets-Infectious Disorders*, 4(1):53–63.

[Sibley, 2014] Sibley, C. H. (2014). Artemisinin resistance: the more we know, the more complicated it appears. *Journal of Infectious Diseases*, page jiu469.

[Sidhu et al., 2006] Sidhu, A. B. S., Uhlemann, A.-C., Valderramos, S. G., Valderramos, J.-C., Krishna, S., and Fidock, D. A. (2006). Decreasing *pfmdr1* copy number in *Plasmodium falciparum* malaria heightens susceptibility to mefloquine, lumefantrine, halofantrine, quinine, and artemisinin. *Journal of Infectious Diseases*, 194(4):528–535.

[Sidhu et al., 2002] Sidhu, A. B. S., Verdier-Pinard, D., and Fidock, D. A. (2002). Chloroquine resistance in *Plasmodium falciparum* malaria parasites conferred by *pfcrt* mutations. *Science*, 298(5591):210–213.

[Singh et al., 2004] Singh, B., Sung, L. K., Matusop, A., Radhakrishnan, A., Shamsul, S. S., Cox-Singh, J., Thomas, A., and Conway, D. J. (2004). A large focus of naturally acquired *Plasmodium knowlesi* infections in human beings. *The Lancet*, 363(9414):1017–1024.

[Sinka et al., 2011] Sinka, M. E., Bangs, M. J., Manguin, S., Chareonviriyaphap, T., Patil, A. P., Temperley, W. H., Gething, P. W., Elyazar, I., Kabaria, C. W., Harbach, R. E., et al. (2011). The dominant *Anopheles* vectors of human malaria in the Asia-Pacific region: occurrence data, distribution maps and bionomic précis. *Parasite Vectors*, 4(1):89.

[Sinka et al., 2012] Sinka, M. E., Bangs, M. J., Manguin, S., Rubio-Palis, Y., Chareonviriyaphap, T., Coetzee, M., Mbogo, C. M., Hemingway, J., Patil, A. P., Temperley, W. H., et al. (2012). A global map of dominant malaria vectors. *Parasite Vectors*, 5(1):69.

[Slatkin, 1987] Slatkin, M. (1987). Gene flow and the geographic structure of natural populations. *Science*, 236(4803):787–792.

[Smith and McKenzie, 2004] Smith, D. L. and McKenzie, F. E. (2004). Statics and dynamics of malaria infection in *Anopheles* mosquitoes. *Malaria Journal*, 3(1):13.

[Snow et al., 1997] Snow, R. W., Omumbo, J. A., Lowe, B., Molyneux, C. S., Obiero, J.-O., Palmer, A., Weber, M. W., Pinder, M., Nahlen, B., Obonyo, C., et al. (1997). Relation between severe malaria morbidity in children and level of *Plasmodium falciparum* transmission in Africa. *The Lancet*, 349(9066):1650–1654.

[Snowden, 2008] Snowden, F. (2008). *The conquest of malaria: Italy, 1900-1962*. Yale University Press.

[Sokal, 1958] Sokal, R. R. (1958). A statistical method for evaluating systematic relationships. *Univ Kans Sci Bull*, 38:1409–1438.

[Spring et al., 2015] Spring, M. D., Lin, J. T., Manning, J. E., Vanachayangkul, P., Somethy, S., Bun, R., Se, Y., Chann, S., Ittiverakul, M., Sia-ngam, P., et al. (2015). Dihydroartemisinin-piperaquine failure associated with a triple mutant including *kelch13* C580Y in Cambodia: an observational cohort study. *The Lancet Infectious Diseases*, 15(6):683–691.

[Staab et al., 2015] Staab, P. R., Zhu, S., Metzler, D., and Lunter, G. (2015). scrm: efficiently simulating long sequences using the approximated coalescent with recombination. *Bioinformatics*, 31(10):1680–1682.

[Straimer et al., 2015] Straimer, J., Gnädig, N. F., Witkowski, B., Amaratunga, C., Duru, V., Ramadani, A. P., Dacheux, M., Khim, N., Zhang, L., Lam, S., et al. (2015). K13-propeller mutations confer artemisinin resistance in *Plasmodium falciparum* clinical isolates. *Science*, 347(6220):428–431.

[Su et al., 2003] Su, X.-Z., Mu, J., and Joy, D. A. (2003). The Malariaś Eve hypothesis and the debate concerning the origin of the human malaria parasite *Plasmodium falciparum*. *Microbes and Infection*, 5(10):891–896.

[Tabernero et al., 2014] Tabernero, P., Fernández, F. M., Green, M., Guerin, P. J., and Newton, P. N. (2014). Mind the gaps–the epidemiology of poor-quality anti-malarials in the malarious world–analysis of the WorldWide Antimalarial Resistance Network database. *Malaria Journal*, 13(139):10–1186.

[Takala-Harrison et al., 2015] Takala-Harrison, S., Jacob, C. G., Arze, C., Cummings, M. P., Silva, J. C., Dondorp, A. M., Fukuda, M. M., Hien, T. T., Mayxay, M., Noedl, H., et al. (2015). Independent emergence of artemisinin resistance mutations among *Plasmodium falciparum* in Southeast Asia. *Journal of Infectious Diseases*, 211(5):670–679.

[Tirados et al., 2006] Tirados, I., Costantini, C., Gibson, G., and Torr, S. J. (2006). Blood-feeding behaviour of the malarial mosquito *Anopheles arabiensis*: implications for vector control. *Medical and Veterinary Entomology*, 20(4):425–437.

[Tu, 2011] Tu, Y. (2011). The discovery of artemisinin (qinghaosu) and gifts from Chinese medicine. *Nature Medicine*, 17(10):1217–1220.

[Tun et al., 2015] Tun, K. M., Imwong, M., Lwin, K. M., Win, A. A., Hlaing, T. M., Hlaing, T., Lin, K., Kyaw, M. P., Plewes, K., Faiz, M. A., et al. (2015). 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.

[Verdrager, 1986] Verdrager, J. (1986). Epidemiology of the emergence and spread of drug-resistant falciparum malaria in south-east asia and australasia. *The Journal of Tropical Medicine and Hygiene*, 89(6):277–289.

[Voight et al., 2005] Voight, B. F., Adams, A. M., Frisse, L. A., Qian, Y., Hudson, R. R., and Di Rienzo, A. (2005). Interrogating multiple aspects of variation in a full resequencing data set to infer human population size changes. *Proceedings of the National Academy of Sciences of the United States of America*, 102(51):18508–18513.

[Volkman et al., 2012] Volkman, S. K., Neafsey, D. E., Schaffner, S. F., Park, D. J., and Wirth, D. F. (2012). Harnessing genomics and genome biology to understand malaria biology. *Nature Reviews Genetics*, 13(5):315–328.

[Volpe et al., 1992] Volpe, F., Dyer, M., Scaife, J. G., Darby, G., Stammers, D. K., and Delves, C. J. (1992). The multifunctional folic acid sythesis fas gene of *Pneumocystis carinii* appears to encode dihydropteroate synthase and hydroxymethyldihydropterin pyrophosphokinase. *Gene*, 112(2):213–218.

[von Seidlein and Greenwood, 2003] von Seidlein, L. and Greenwood, B. M. (2003). Mass administrations of antimalarial drugs. *Trends in Parasitology*, 19(10):452–460.

[Watterson, 1975] Watterson, G. (1975). On the number of segregating sites in genetical models without recombination. *Theoretical Population Biology*, 7(2):256–276.

[White, 1999] White, N. (1999). Antimalarial drug resistance and combination chemotherapy. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 354(1384):739–749.

[White and Pongtavornpinyo, 2003] White, N. and Pongtavornpinyo, W. (2003). The de novo selection of drug–resistant malaria parasites. *Proceedings of the Royal Society of London B: Biological Sciences*, 270(1514):545–554.

[White, 2004] White, N. J. (2004). Antimalarial drug resistance. *Journal of Clinical Investigation*, 113(8):1084.

[White, 2010] White, N. J. (2010). Artemisinin resistance, the clock is ticking. *The Lancet*, 376(9758):2051–2052.

[WHO, 1999] WHO (1999). The World Health Report, 1999. *World Health Organization, Geneva, Switzerland*, pages Rolling back malaria, 49–63.

[WHO, 2007] WHO (2007). The Abuja declaration and the plan of action (extract of African Summit on Roll Back Malaria). *World Health Organization, Geneva, Switzerland*.

[WHO, 2011] WHO (2011). Global plan for artemisinin resistance containment (GPARC). *World Health Organization, Geneva, Switzerland*.

[WHO, 2014a] WHO (2014a). Status report on artemisinin resistance. *World Health Organization, Geneva, Switzerland*.

[WHO, 2014b] WHO (2014b). World Malaria Report 2014. *World Health Organization, Geneva, Switzerland*.

[WHO, 2015a] WHO (2015a). Malaria elimination strategy. *World Health Organization, Geneva, Switzerland*.

[WHO, 2015b] WHO (2015b). Malaria World Report, 2015. *World Health Organization, Geneva, Switzerland*.

[Wichmann et al., 2004] Wichmann, O., Muehlen, M., Gruss, H., Mockenhaupt, F. P., Suttorp, N., and Jelinek, T. (2004). Malarone treatment failure not associated with previously described mutations in the *cytochromeb* gene. *Malaria Journal*, 3(1):14.

[Wongsrichanalai and Meshnick, 2008] Wongsrichanalai, C. and Meshnick, S. R. (2008). Declining artesunate-mefloquine efficacy against *falciparum* malaria, Cambodia-Thailand border. *Emerging Infectious Diseases*, 14(5):716.

[Wongsrichanalai et al., 2002] Wongsrichanalai, C., Pickard, A. L., Wernsdorfer, W. H., and Meshnick, S. R. (2002). Epidemiology of drug-resistant malaria. *The Lancet Infectious Diseases*, 2(4):209–218.

[Worrall et al., 2005] Worrall, E., Basu, S., and Hanson, K. (2005). Is malaria a disease of poverty? A review of the literature. *Tropical Medicine & International Health*, 10(10):1047–1059.

[Wright, 1931] Wright, S. (1931). Evolution in mendelian populations. *Genetics*, 16(2):97–159.

[Wright, 1943] Wright, S. (1943). Isolation by distance. *Genetics*, 28(2):114.

[Wright, 1949] Wright, S. (1949). The genetical structure of populations. *Annals of Eugenics*, 15(1):323–354.

[Zhao et al., 2012] Zhao, Y. O., Kurscheid, S., Zhang, Y., Liu, L., Zhang, L., Loeliger, K., and Fikrig, E. (2012). Enhanced survival of *Plasmodium*-infected mosquitoes during starvation. *PloS One*, 7(7):e40556.

[Zhu et al., 2017] Zhu, S. J., Almagro-Garcia, J., and McVean, G. (2017). Deconvoluting multiple infections in *Plasmodium falciparum* from high throughput sequencing data. *bioRxiv*.