DIGITAL ACCESS TO SCHOLARSHIP AT HARVARD
DASH.HARVARD.EDU

HARVARD LIBRARY
Office for Scholarly Communication

Evolutionary Adaptation and Antimalarial Resistance in Plasmodium falciparum

Citation
Park, Daniel John. 2013. Evolutionary Adaptation and Antimalarial Resistance in Plasmodium falciparum. Doctoral dissertation, Harvard University.

Permanent link
http://nrs.harvard.edu/urn-3:HUL.InstRepos:11169775

Terms of Use
This article was downloaded from Harvard University's DASH repository, and is made available under the terms and conditions applicable to Other Posted Material, as set forth at http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#LAA

Share Your Story

The Harvard community has made this article openly available.
Please share how this access benefits you. Submit a story .

Accessibility

Evolutionary Adaptation and Antimalarial Resistance in Plasmodium falciparum

A DISSERTATION PRESENTED
BY
DANIEL JOHN PARK
TO
THE DEPARTMENT OF ORGANISMIC AND EVOLUTIONARY BIOLOGY

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY
IN THE SUBJECT OF
BIOLOGY

HARVARD UNIVERSITY
CAMBRIDGE, MASSACHUSETTS
JULY 2013

© 2013 - Daniel John Park
All rights reserved.

Dissertation advisor: Professor Pardis C. Sabeti Daniel John Park

Evolutionary Adaptation and Antimalarial Resistance in Plasmodium falciparum

Abstract

The malaria parasite, *Plasmodium falciparum*, has a demonstrated history of adaptation to antimalarials and host immune pressure. This ability unraveled global eradication programs fifty years ago and seriously threatens renewed efforts today. Despite the magnitude of the global health problem, little is known about the genetic mechanisms by which the parasite evades control efforts. Population genomic methods provide a new way to identify the mutations and genes responsible for drug resistance and other clinically important traits.

In this thesis, I set out to develop and apply novel approaches to studying parasite adaptation in three parts. First, I carry out a global, genomic survey of *P. falciparum* diversity and perform a genome-wide association study (GWAS) to identify a number of candidate markers of drug resistance. Subsequent validation shows a causal relationship between one of these candidates and parasite drug response.

Second, I further pursue the GWAS approach by sampling a single population more deeply and moving from array-based genotypes to whole-genome sequence data. I demonstrate the deficiences of array-based GWAS in low linkage disequilibrium (LD) populations and argue for a complete transition to sequence-based GWAS for small, low-LD genomes like *P. falciparum*. I additionally develop the use of a long-haplotype natural selection test to detect associations with adaptive traits.

Finally, I exploit the parasite's short generation time to detect temporal signatures of selection in progress from samples collected over several years. I develop and evaluate new genome-wide statistics for this test and find that it identifies coding variants, often in surface proteins subject to balancing selection. This approach is complementary to existing selection tests and is a timely addition to the genomic toolkit available to malaria eradication efforts.

This research contributes numerous novel approaches to the problem of the rapidly evolving *P. falciparum* parasite and significantly advances the field's ability to provide the tools and knowledge required for current global eradication campaigns.

Contents

1 Using Genomics to Inform Malaria Control and Eradication Efforts **1**
- 1.1 Background . . . 3
- 1.2 Technological Advancements . . . 4
- 1.3 Population Structure and Linkage Disequilibrium . . . 5
- 1.4 Signatures of Selection . . . 8
- 1.5 Finding the Genetic Basis for Specific Phenotypes . . . 13
- 1.6 Conclusions and Dissertation Overview . . . 18

2 Genome-wide Association Studies Identify the Antimalarial Resistance Locus *PF10_0355* **20**
- 2.1 Introduction . . . 21
- 2.2 Results . . . 22
- 2.3 Discussion . . . 33
- 2.4 Methods . . . 38
- 2.5 Acknowledgements . . . 42

3 Sequence-based Association and Selection-association Scans **43**
- 3.1 Introduction . . . 44
- 3.2 Results . . . 45
- 3.3 Discussion . . . 51
- 3.4 Methods . . . 56
- 3.5 Acknowledgements . . . 60

4 Temporal Signatures of Selection **61**
- 4.1 Introduction . . . 62
- 4.2 Results . . . 63
- 4.3 Discussion . . . 67
- 4.4 Methods . . . 71
- 4.5 Acknowledgements . . . 74

A Secondary Publications **76**
- A.1 Published manuscripts, Sept 2010 to July 2013 . . . 76

A.2 Manuscripts in preparation or review 80

B Ascertainment Bias Corrections to Selection Studies in *Plasmodium falciparum* **82**
B.1 Introduction 83
B.2 Results and Discussion 84
B.3 Methods 87
B.4 Acknowledgements 89

C Supplemental Material for Chapter 2 **90**
C.1 Author Contributions to Supplemental Material 90
C.2 Supplemental Methods 91
C.3 Supplemental Data Files 98
C.4 Supplemental Figures 99
C.5 Supplemental Tables 111
C.6 Figures and Tables Supporting Supplemental Methods 119

D Supplemental Material for Chapter 3 **124**
D.1 Author Contributions to Supplemental Material 124
D.2 Supplemental Data Files 125
D.3 Supplemental Results 125
D.4 Supplemental Methods 127
D.5 Supplemental Figures 130

E Supplemental Material for Chapter 4 **135**
E.1 Author Contributions to Supplemental Material 135
E.2 Supplemental Figures 136
E.3 Null models of the selection coefficient 138
E.4 Long-haplotype tests 142
E.5 Ex vivo GWAS tests 143

References **145**

Listing of figures

1.1 Signatures of balancing selection 9
1.2 Signatures of directional selection 12
1.3 Identifying drug resistance loci using GWASs and functional studies 17
2.1 Parasite global population structure and genetic diversity vs. divergence 23
2.2 Genome-wide association study (GWAS) results 27
2.3 Overexpression of *PF10_0355* decreases parasite susceptibility to halofantrine (HFN) and related antimalarials 30
2.4 Correlations between antimalarial drugs tested 32
2.5 Copy number variation at *PF10_0355* is associated with HFN resistance 34
3.1 Simulated *P. falciparum* arrays are unable to tag SNPs not present on the array 46
3.2 Association signals around *pfmdr1* (array vs. sequence) 49
3.3 Significant signals of drug-associated selection (XP-EHH) . . . 50
3.4 Localizing the pyrimethamine-associated selection signal on chr12 54
4.1 Two test statistics match null distributions. 64
B.1 Diversity vs. Divergence in Senegal, uncorrected 84
B.2 Watterson's $\hat{\theta}_W$ vs. Tajima's $\hat{\theta}_T$, uncorrected and corrected 86
B.3 F_{ST} vs. Tajima's D, uncorrected and corrected 87
B.4 Null simulations of Tajima's D 88
C.1 Population structure (PCA) within each continent 99
C.2 Linkage disequilibrium decay by distance 100
C.3 Functional (GO category) enrichment in nucleotide diversity . . 101
C.4 SNP diversity and divergence by translation consequence 102
C.5 Raw REHH scores across genome and by allele frequency 103
C.6 Long-range haplotype signals across the genome 103
C.7 GWAS *P*-value distributions for Fisher's exact test, permuted Fisher's exact test, and Cochran-Mantel-Haenszel (CMH) tests . 104
C.8 GWAS results for the EMMA test 105

C.9 GWAS *P*-value distributions for HLR tests for association to drug resistance . . . 106
C.10 GWAS *P*-value distributions for HLR tests for association to drug sensitivity . . . 107
C.11 *PF10_0355* copy number variation measured by Affymetrix hybridization intensity . . . 108
C.12 *PF10_0355* copy number variation measured by Southern blotting 109
C.13 Drug resistance phenotype classification for sweep and GWAS analyses . . . 110
C.14 Genic/intergenic effect on array performance . . . 119
C.15 Effect of SNP discovery minor allele count on array performance 119
C.16 Effect of GC composition on array performance . . . 120
C.17 Effect of non-unique flanking sequence on array performance . . 120
C.18 Array marker spacing . . . 121
C.19 Final concordant marker density across genome . . . 123
C.20 Distribution of markers per gene . . . 123
D.1 Drug response distributions . . . 130
D.2 Drug response correlation heat map . . . 131
D.3 EMMA GWAS plots (sequence data, 45 samples) . . . 132
D.4 XP-EHH GWAS plots (sequence data, 45 samples) . . . 133
D.5 EMMA GWAS plots (array data, 24 samples) . . . 134
E.1 DAF spectra . . . 136
E.2 DAF over time for major drug loci . . . 137
E.3 HMM statistics are biased at extreme DAF . . . 138
E.4 HMM statistics at different drift strengths and time spans . . . 139
E.5 Volcano plot: P vs. s . . . 140
E.6 Double-zero variants lead to inflated statistics . . . 141
E.7 Long-haplotype selection tests in W. Africa . . . 142
E.8 Ex vivo GWAS: EMMA . . . 143
E.9 Ex vivo GWAS: XP-EHH . . . 144

Listing of tables

1.1 Speed of drug resistance evolution in *P. falciparum* 2
2.1 Eleven genome-wide significant associations with antimalarial drug resistance 28
4.1 Top candidates for strong selection 66
C.1 Parasite list 112
C.2 Array tagging ability 113
C.3 Long Range Haplotype (LRH) hits 114
C.4 IC_{50} drug resistance phenotype data 115
C.5 Parasites used in GWAS 116
C.6 *PF10_0355* copy number summary for 38 parasites 117
C.7 Annotation and GeneID Information for identified genes in Figure 2.1B. 118
C.8 Statistics on marker spacing by chromosome 122

For my son Isaac, my greatest "accomplishment" in graduate school, and most of all, for my wife Susan, who bore the greater burden of my academic career indulgence.

Acknowledgments

The work reported in this dissertation was performed under the supervision of Professor Pardis Sabeti. She deserves credit, not just for excellent mentorship and support during my time as a student, but for successfully recruiting me to pursue doctoral studies in the first place. She has been an advocate and advisor in my career path these past several years. My dissertation committee, comprised of Professors Sabeti, Dan Hartl, John Wakeley, and Dyann Wirth, provided valuable feedback and guidance throughout the course of this research and I have enjoyed working with all of them during my time at Harvard and the Broad.

This research was made possible by, and performed in the in the context of, the malaria genetics group, a collaboration between labs at Harvard GSAS, Harvard SPH, and the Broad Institute. In addition to the labs represented in the dissertation committee, I would like to thank Dan Neafsey's team at the Broad, the Sequencing team at the Broad, and Sarah Volkman at the School of Public Health, for their critical roles in the work described here. Additionally, Steve Schaffner (Broad), Clarissa Valim (HSPH), and Hsiao-Han Chang (OEB) frequently provided help in thinking through analyses and methods.

The members of the Sabeti Lab provide a special kind of environment: one that is fun, inquisitive, and sharpening. I've learned much from simply sitting with fellow researchers regardless of how similar or different our methods, organisms, or training are. A number of the important methods I utilize in this dissertation started with ideas tossed around in conversation with fellow labmates.

My research built on many lessons learned over the past decade applying computational tools to biological problems. From my Master's thesis work with Prof. Amy Keating at MIT a dozen years ago, to the informatic analysis and statistical consulting work at MGH under Glenn Short and Mason Freeman, to the software development group at the Broad under Dave DeCaprio, Phil Montgomery, and Mike Koehrsen, these previous groups, mentors, and experiences continued to play an important role for how I approached my work throughout my time at Harvard.

I am indebted to my parents for all that they have provided for me, and for their encouragement to pursue this program. I value their love and support and for the ways my relationships with them and with my siblings, Andy and Janice, have evolved significantly over these past several years.

I am thankful for my son, Isaac, and especially for my wife, Susan. Susan has encouraged and supported my academic pursuits, despite the costs to our finances, stability, and most of all, our time together. I am grateful for her love and patience. She gets to enroll in the next degree program. We are both thankful for our community and friends in Boston and at Highrock Church, who have kept us sane, encouraged, loved, and fed, throughout the busiest seasons (particularly when Isaac was born).

My graduate studies are funded by a Graduate Research Fellowship from the National Science Foundation, which supported my entire three-year period at Harvard. Professor Sabeti is supported by fellowships from the Burroughs Wellcome and Packard Foundations.

The research described in this dissertation is funded by the Bill and Melinda Gates Foundation, National Institutes of Health (Grant: 1R01AI075080-01A1), Ellison Medical Foundation, ExxonMobil Foundation, NIH Fogarty International Center, NIAID, Harvard Malaria Initiative, and Broad SPARC.

Author List

Chapter 1: Using Genomics to Inform Malaria Control and Eradication Efforts

Chapter 1 is heavily based on a review article to which the following authors contributed: Sarah K. Volkman, Daniel E. Neafsey, Stephen F. Schaffner, Daniel J. Park, and Dyann F. Wirth. In addition, the staff at Nature Reviews Genetics made editorial contributions to the text and aesthetic modifications to the figures.

Most of the text in this chapter is largely duplicated from the original review article. However, many sections and visuals that were not relevant to this dissertation were removed in an effort to keep this chapter focused on introducing concepts and approaches used later in this thesis (assessment of interventions, complexity of infection, molecular barcoding, proteomics, Fig 3, Fig 4, Box 1, Box 2). Also, Fig 1 was split into two figures (1.1 and 1.2) for space and clarity.

Most of the introductory and concluding paragraphs have largely been replaced or rewritten. The first four and last four paragraphs of the chapter are original to this thesis and written by me. Table 1.1 was not in the review article, but was compiled by DEN separately for [112].

A number of sections have been revised for this dissertation. In particular, discussions of technologies and recent literature have been updated to reflect the advances and new publications in the past year since the review article was originally published.

For the original review article, I wrote the GWAS section (Section 1.5, SFS and

DEN wrote on population structure, LD, and selection (Sections 1.3 and 1.4) and SKV outlined the overall structure and wrote intro, conclusion, and transitional text throughout. After the first draft, all authors heavily edited all sections, and I made contributions throughout Sections 1.3 and 1.4 as well. Other sections that I had minimal involvement in were omitted from this version of the text.

Illustrative visuals for this review article were conceived by SKV and DEN and executed by the staff at Nature Reviews Genetics. In most cases, these figures re-utilize images or data previously published by our group and are noted as such in the figure captions (Figures 1.1 and 1.2).

Chapter 2: Genome-wide Association Studies Identify the Antimalarial Resistance Locus *PF10_0355*

The following authors contributed to Chapter 2: Daria Van Tyne*, Daniel J. Park*, Stephen F. Schaffner*, Daniel E. Neafsey*, Elaine Angelino*, Joseph F. Cortese, Kayla G. Barnes, David M. Rosen, Amanda K. Lukens, Rachel F. Daniels, Danny A. Milner, Jr, Charles A. Johnson, Ilya Shlyakhter, Sharon R. Grossman, Justin S. Becker, Daniel Yamins, Elinor K. Karlsson, Daouda Ndiaye, Ousmane Sarr, Souleymane Mboup, Christian Happi, Nicholas A. Furlotte, Eleazar Eskin, Hyun Min Kang, Daniel L. Hartl, Bruce W. Birren, Roger C. Wiegand, Eric S. Lander, Dyann F. Wirth*, Sarah K. Volkman*, and Pardis C. Sabeti*. The first five and last three authors of Chapter 2 are noted as sharing equal contribution.

In short, I was responsible for the front half of this study (chip design, analysis and GWAS) and DVT was responsible for the back half (functional follow up with the *PF10_0355* locus). EA was responsible for the long range haplotype selection scans shown in supplemental figures. DEN and SFS were responsible for analyses of LD, population structure, and natural selection. I designed the genotyping array based on previous experience with a pilot array [113] and analyzed and processed array data using a conservative validation approach. I performed all genome-wide association analyses and performed all corrections necessary for population structure and multiple testing.

Chapter 3: Sequence-based Association and Selection-association Scans

The following authors contributed to Chapter 3: Daniel J. Park, Amanda K. Lukens, Daniel E. Neafsey, Stephen F. Schaffner, Hsiao-Han Chang, Clarissa Valim, Ulf Ribacke, Daria Van Tyne, Kevin Galinsky, Meghan Galligan, Justin S. Becker, Daouda Ndiaye, Souleymane Mboup, Roger C. Wiegand, Daniel L. Hartl*, Pardis C. Sabeti*, Dyann F. Wirth*, and Sarah K. Volkman*. The last four authors contributed equally to Chapter 3.

I performed nearly all computational analyses for this chapter. AKL was responsible for drug assays and parasite culturing. HHC performed PCA analyses. I analyzed all sequence data, performed association and selection tests, and analyzed results.

Chapter 4: Temporal Signatures of Selection

The project described in Chapter 4 is not yet in preparation for manuscript submission, so a formal author list has not yet been prepared. However, a short list of individuals that have contributed to aspects of this project are mentioned in its acknowledgements section. These include: Diana Miao, Daria Van Tyne, Hsiao-Han Chang, Clarissa Valim, Hilary Finucane, Stephen F. Schaffner, Daniel E. Neafsey, Eli L. Moss, and Pardis C. Sabeti.

I performed all computational analyses described in this chapter. DVT inspired the project idea when observing trends at a few candidate genes. DM is performing PCR validation of the top hits this summer on a larger sample set. HHC, CV, HF, SFS, and DEN advised on modeling and methods throughout. ELM performed final sequence alignment and genotype calls based on protocols I first developed.

This chapter is based on material originally published in Volkman, Neafsey, Schaffner, Park, and Wirth, Harnessing genomics and genome biology to understand malaria biology, Nature Reviews Genetics (13), 2012 [176]. See page xiii for details on author contributions.

doi://10.1038/nrg3187

1

Using Genomics to Inform Malaria Control and Eradication Efforts

Global health organizations are in the midst of a renewed push towards the eradication and eventual extinction of the malaria parasites, the most lethal of which is *Plasmodium falciparum*. These efforts rely on a number of approaches applied in parallel, from indoor spraying and bed nets to improved local health infrastructure. Ultimately, one of the most vital tools in this arsenal remains the effective use of antimalarial drugs [94]. *P. falciparum*, however, has demonstrated an adaptability to numerous drugs (Table 1.1). For example, widespread chloroquine resistance was a major factor in the abandonment of the Global Malaria Eradication Programme in 1969 [111].

The eventual emergence of resistance to more modern drugs is a significant concern, as the current global eradication efforts rely heavily on a single class of drugs:

artemisinin and its derivatives. Although fully artemisinin-resistant forms have not yet been described, recent work has shown that the alarming trend of slower parasite clearance rates in artemisinin-treated patients is increasing over time and appears to be a genetically heritable trait of the parasite [10, 121]. This underscores an urgent need for the research community to be able to detect and understand newly emerging resistance adaptations.

Among the major diseases of greatest human impact, malaria is unique in that it is caused by a sexual eukaryote, allowing it to adopt beneficial combinations of mutations more quickly than asexual pathogens [66]. While this amplifies the challenges for global disease control, it is actually a favorable feature for genomic inquiry, as it allows us to apply many of the modern tools developed in human genomics that require meiotic recombination such as association studies and selection scans. This allows for the realistic expectation that understanding emerging drug resistance in *P. falciparum* is feasible in the time frame required for global eradication.

This introductory chapter describes the genomic tools now available to tackle these questions, the many challenges involved, the current state of the field in applying these approaches to the *P. falciparum* malaria parasite, and the lessons learned so far about parasite evolution.

Table 1.1: Speed of drug resistance evolution in *P. falciparum*

Drug	Introduction	*In vivo* resistance	Origin
Chloroquine	1945	Early 1960s [131]	SE Asia, S America
Sulfadoxine + Pyrimethamine	1967	Late 1960s [61]	SE Asia, S America
Mefloquine	1985	Early 1990s [137]	SE Asia
Atovaquone + Proguanil	2000	2002 [55]	Africa
Artemether + Lumefantrine	2001	2008 [47, 124]	SE Asia

1.1 Background

The WHO estimates that close to one million children die from malaria every year [126], with the highest mortality among African children. Increased control efforts have reduced the malaria burden in some areas, but evidence of rebounding malaria [166] brings these general trends into question and confirms that we have much to learn to defeat this important human pathogen. It is now more critical than ever to understand key aspects of malaria biology and transmission, identify targets vulnerable to intervention strategies, and create tools to interpret the changing landscape of infection. One powerful approach uses population biology-based investigations to provide critical insight about the causes and spread of disease. This strategy aids biological discovery by using population structure and genetic diversity to identify loci under selection or associated with clinical phenotypes, and for developing tools to monitor and evaluate interventions.

P. falciparum is a eukaryotic pathogen with a complex lifecycle, spending part of its lifespan in its definitive host, the anopheles mosquito, as mostly a diploid organism, and the remainder in its human host as a haploid organism where it gives rise to numerous clinical manifestations from mild to life-threatening illness. The 24 Mb genome of the parasite is distributed among 14 linear chromosomes and the parasite contains two extra-nuclear circular chromosomes that comprise the apicoplast and mitochondrial genomes. The full *P. falciparum* genome sequence was published in 2002 [57] and was followed by publication of other *Plasmodium* genome sequences, including that of *Plasmodium vivax* [23], which causes significant human malaria. These data have allowed elucidation of basic genome architecture and identification of key structural elements, common metabolic and biosynthesis pathways and unique aspects shared among several *Plasmodium* parasites [22, 23, 57, 64].

The *P. falciparum* genome is evolving in response to natural selection pressures of the human host immune system, the mosquito vector and various environmental factors including drug treatment and changes in transmission intensity due to specific interventions [12, 50, 178]. The data imply that parasites can escape both natural and artificial selection pressures through evolution. Understanding these

adaptations in the *P. falciparum* population can allow us to identify and circumvent survival strategies used by the parasite, guiding the development of new drugs and vaccines. Indeed, similar approaches have been applied to much simpler viral genomes for tracking influenza outbreaks and for developing effective influenza vaccines [59].

1.2 Technological Advancements

New genomic technologies are what drive the recent advances in our understanding of the history and evolution of the malaria parasite. Population genomic studies in malaria started with the low-coverage sequencing of a few dozen global parasites [74, 106, 174] and accelerated with the development of various whole genome genotyping arrays that allowed researchers to cheaply genotype a hundreds to hundreds of thousands of markers across a large population of parasites. Depending on the technology, these methods can be used to characterize genomic variation at different levels, including SNPs, microsatellite variation (MSVs), insertions or deletions (indels), and copy number variants (CNVs). However, array-based methods need to be custom-built for each organism, requiring specific tools for *P. falciparum* [30, 45, 75, 83, 106, 113, 161, 170] and *P. vivax* [180] to be constructed.

More recently, dramatic decreases in the per base-pair cost of next-generation sequencing technologies has led to a shift in the economics of studying the

---

Single Nucleotide Polymorphism (SNP)—A single base pair in the genome where the allele varies within a population. This is the most common form of genetic variation in populations and provides the data for most modern genomic studies.

Microsatellite—A class of repetitive DNA sequence that is made up of repeats that are 2–8 nucleotides in length. They can be highly polymorphic and are frequently used as molecular markers in population genetics studies.

Indel—A position in the genome that varies in the population where the alleles are not all of the same length. Microsatellites are a type of indel.

Copy number variant (CNV)—This is a segment of DNA (e.g. a gene or a chunk of chromosome) that is duplicated elsewhere in the genome, and individuals in the population vary as to how many copies they have.

modestly-sized parasite genome. Although some arrays remain valuable, for example to identify CNVs, whole-genome sequencing is now similar in cost to array-based genotyping, while providing information about a significantly larger number of genomic markers without the species-specific customization required of arrays. Consequently, additional *P. falciparum* genomes have been sequenced at greater depth to identify essentially all genetic variation in those genomes, thus allowing delineation of the composition and relative proportions of parasite types within a human infection [7, 13, 28, 95, 98, 102, 127–129].

These practical advancements are enhanced by several key bioinformatics analysis improvements, many of which were developed for the analysis of other organisms and driven by international collaborative efforts such as the Human Genome Project [86]. New computational strategies to identify SNPs and other variants including MSVs, indels and CNVs from sequencing data are being developed and applied to *P. falciparum*, providing additional markers that are potentially associated with drug resistance. Availability of genome sequences for closely related species, including *Plasmodium reichenowi* [74] (which infects non-human primates [91]), will greatly advance our ability to make population genomic inferences by identifying derived alleles for selection analysis.

Once obtained, this rich genomic information can yield important insights about the biology of the malaria parasite, as described in the following sections. Applications include gaining insights into parasite population structure and interventions, and identifying important regions of the genome—either those that show evidence of evolutionary selection or specific loci identified by genome-wide association studies (GWASs) or mutant screens—that are responsible for particular parasite phenotypes.

1.3 Population Structure and Linkage Disequilibrium

Genetic variation in the parasite reveals the exposure history of a given parasite or parasite population to selective pressures. Central to our understanding of genetic variation in *P. falciparum* is determining its current population structure, including how allele frequencies vary between different populations within the

species, and the degree to which alleles at neighboring variant sites are correlated by linkage disequilibrium (LD). Data used to inform our current understanding of population structure and LD is derived from genome-wide array-based methods [6, 31, 45, 107, 113, 160, 161, 170], from SNP information provided by older sequencing data from a few dozen published genomes [74, 106, 174], and from recent next-generation sequencing studies [28, 95, 102]. Genomic structure among isolates has been delineated using principal components analysis (PCA), from which one can infer the relatedness of samples.

1.3.1 Global Genomic Diversity

Advances in genomic technologies have enabled analysis of many different parasites derived from distinct geographic locations. Large-scale population structure in *P. falciparum* follows continental lines, with major branches in Africa, South and Central America, and South and East Asia (extending to Papua New Guinea) [105, 113]. The picture within each group depends on the region. Within Africa, population differences between countries have ranged from undetectable (Uganda vs. Congo [8], Cameroon vs. Congo [141]) or very small (Zimbabwe vs. Uganda and Congo [8]), to modest over the longest distances (Nigeria vs. Sudan vs. South Africa [35]). By contrast, in Southeast Asia and the western hemisphere, local population structure is pronounced even within a single country [105, 174]. The overall picture is consistent with a recent geographic spread from a source population in Africa, which has remained by far the largest population. Estimates of genetic diversity measurements present the same picture, with the highest values consistently occurring within Africa and the lowest in the Americas. From these types of analyses, we can identify mutations that are fixed in one parasite population but are distinct from other populations, and thus may be useful to identify spe-

Linkage Disequilibrium (LD)—In population genetics, linkage disequilibrium is the nonrandom association of alleles. For example, alleles of SNPs that reside near one another on a chromosome often occur in nonrandom combinations owing to infrequent recombination.

Principal components analysis (PCA)—A statistical method used to simplify data sets by transforming a series of correlated variables into a smaller number of uncorrelated factors. It is commonly used to correct for stratification in genome-wide association studies.

cific parasites. Knowledge of the genetic characteristics of geographically distinct populations is important for controlling for population stratification in GWASs, tracking persistent parasite types as interventions are applied and localizing new sources of infection to maximize the effectiveness of control measures.

1.3.2 Explanations for Observed LD Structure

Levels of genomic diversity in different populations are reflected in the observed patterns of LD. Very little LD is seen in Africa, and what is seen extends less than 1 kb [34, 105, 129, 174]. LD is slightly higher in Southeast Asia (mean $r^2 = 0.3$ for markers less than 1 kb apart) and more so in South America (mean $r^2 = 0.5$, for markers less than 1 kb apart), where it spans about 10 kb. This difference could stem from demographic history. For example, population bottlenecks in the non-African populations could have eliminated many allele combinations, leaving strong correlations, thus high LD, in the remaining parasites. It could also stem from a smaller effective population size outside of Africa, limiting these populations to fewer allele combinations, which contributes to higher levels of LD. Detailed understanding of recombination and of the demographic history of different populations is needed to distinguish the two causes.

LD patterns also reflect the transmission history of strains within that population. The sexual phase of the *P. falciparum* life cycle occurs in the midgut of female Anopheles mosquitoes, following consumption of a blood meal containing male and female *P. falciparum* gametocytes. Because female Anopheles mosquitoes typically bite only a single human host during each egg-brooding cycle, the gametocyte pool available for sexual union in the midgut matches the gametocyte composition in individual infected human hosts. In geographical regions with high levels of malaria transmission, a high complexity of infection (COI) is thought to

---

Population stratification—The presence of multiple population subgroups that show limited interbreeding. When such subgroups differ in both allele frequency and disease prevalence, this can lead to erroneous results in association studies.

Effective population size (N_e)—The number of individuals in a population genetic model that most closely matches the allele frequency distributions in the data. This can differ from actual population size owing to historical demographic events, such as population bottlenecks, migration and other factors.

be produced through 'super-infection' (multiple bites from distinct, *P. falciparum*-infected mosquitoes), although new evidence supports a model of co-infection of mixed infections [122]. This makes recombination (outcrossing) possible between genetically distinct *P. falciparum* gametocytes during the sexual phase (provided genetic diversity is high enough to ensure genetically distinct parasites), and results in short blocks of LD. By contrast, if only a single *P. falciparum* strain is present, the gametocytes will be identical, and the lack of recombination will result in longer blocks of LD.

Ultimately, as parasite population sizes get extremely small, one would anticipate that LD should become extended and theoretically approach a value of one. Thus, tools to measure changes in parasite population structure have the potential to inform reductions in malaria transmission as outcrossing rates are reduced to the point where selfing among parasites occurs in a given population, such as might be expected during a successful intervention strategy [40].

1.4 Signatures of Selection

In the course of its life cycle, the parasite faces numerous and intense selective pressures. These include exposure to antimalarial drugs and immune challenges from both the human host and mosquito vector. Evidence abounds from recent studies of parasite diversity for two broad classes of strong natural selection in the parasite genome.

1.4.1 Balancing Selection

Genotyping and sequence analysis indicate that an unusually large fraction of the *P. falciparum* genome exhibits the polymorphism profile of immune-mediated balancing selection: a high density of high frequency polymorphisms is seen in hundreds of antigenic genes [7, 50, 178]. Balancing selection maintains polymorphisms with the potential to encode alternate immunological identities and keeps

Balancing selection—Selection that favors the maintenance of more than one polymorphic allele in a population by mechanisms such as frequency-dependent selection or heterozygote advantage.

them at an intermediate population frequency indefinitely (Figure 1.1A)

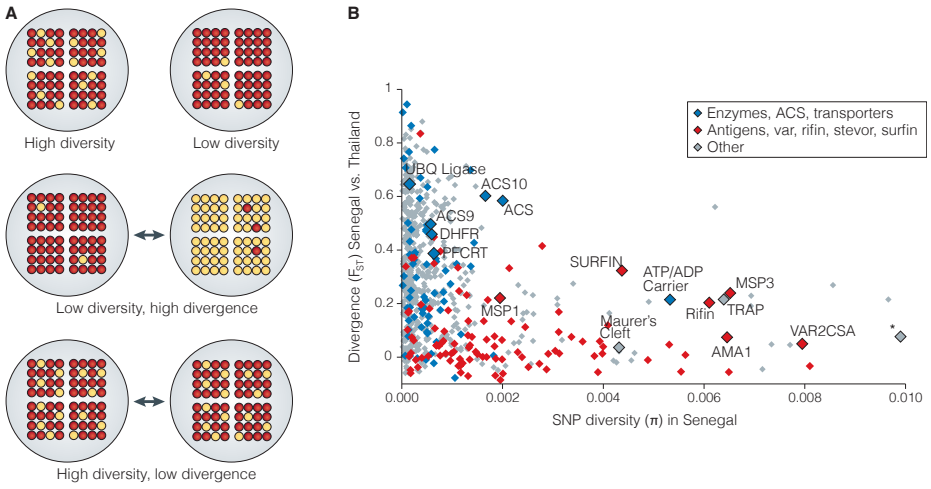


Figure 1.1: Signatures of balancing selection. **(A)** In a given population (large grey circle) there are numerous individuals (square matrix of circles) each containing alleles (red or yellow circles) across their genomes. Diversity refers to the amount of allelic variation among individuals in a population, whereas divergence refers to the amount of allelic variation between different populations. Under balancing selection, one would expect to find high diversity at a locus under selection, but low divergence between populations. **(B)** Distribution of loci based upon both within population differences (diversity, as measured by π) and between population differences (divergence, as measured by F_{ST} between parasites from Senegal and Thailand) is shown. Loci classified as transporters or enzymes including the acyl-CoA-synthetase (ACS) genes are shown as blue diamonds; loci classified as antigens including var, rifin, stevor, and surfin molecules are shown as red diamonds; and all other loci are shown as grey diamonds. Molecules along the x-axis (high diversity and low divergence) are under diversifying selection and these include a number of known antigens. In contrast, molecules along the y-axis (low diversity and high divergence) are more likely under directional selection and these include a number of known drug resistant molecules. Panel (B) was previously published in Van Tyne et al. [170].

When parasite populations are geographically separated, genes subject to balancing selection are unlikely to diverge as rapidly as other genes, because the selection prevents differences from differentially fixing in the populations. Genome-wide comparisons of diversity (within a population, measured by π) and divergence (between populations, measured by F_{ST}) identify genetic loci that are more likely to be affected by this diversifying selection [170] in that they exhibit elevated diversity with relatively low divergence. (Figure 1.1B shows an example diver-

sity and divergence analysis between parasites from Senegal and Thailand [170].) From these analyses known antigens and vaccine candidates are identified, as well as novel genetic loci that encode putative antigens that trigger the human immune response. This prediction was validated when several highly polymorphic genes were expressed and recognized by human immune sera, including seven previously unknown antigens [106]. This result suggests that diverse genomic regions may encode antigenic loci useful for vaccine approaches. However, a number of vaccine studies suggest that the ability to successfully target a polymorphic locus such as merozoite surface protein 1 (MSP1) [159] or apical membrane antigen 1 (AMA1) [163] may be undermined by the parasite's ability to survive the elicitation of a locus-specific immune response. Thus, strategies to use a combination of non-variant yet immunogenic vaccine targets may be warranted.

Diversity and divergence analyses can identify loci that, conversely, diverge between parasite populations ($F_{ST} > 0.4$). Divergent loci between populations from Senegal and Thailand [170] encode proteins that are proposed to have various cellular functions including DNA replication (e.g. *PF10_0165*; *PF14_0278*; *PF14_0316*), lipid metabolism (e.g. *PFB0695c*; *PFE1250w*; *PFB0685c*; *PFC0050c*); gametocytogenesis or sexual development molecules (e.g. *PF13_0248*; *PFC0640w*) and transporters (e.g. *PFL1125w*; *PF14_0342*; *PF14_0455*) (Figure 1.1B). The reasons for the divergence are currently unclear and require further investigation, but may be a consequence of differences in vector populations or other distinct selective pressures between Senegal and Thailand.

1.4.2 Directional Selection

Directional selection in the context of a traditional 'selective sweep' [155] leaves a distinctive genomic imprint consisting of depleted polymorphism and enhanced LD (Figure 1.2AB). This genomic signature is detectable via 'haplotype-based' tests of natural selection, such as the long-range haplotype (LRH) test. In response to strong selective pressures, long haplotype signals resulting from the rapid rise of variants linked to flanking mutations are easily detected as they stand out from the normally short LD of the genomic background [150] (Figure 1.2B). It is important to note that these signals may be absent if directional selection began on common or standing variation [150]. Equally important is the difficulty in identifying a clear demarcation between selective sweeps and neutral processes without a detailed understanding of demographic history and recombination rate variation, knowledge that is lacking for *P. falciparum*. Nevertheless, in genome-wide scans for selective sweeps a number of loci show strong evidence for recent directional selection, and they all point to a single, recent evolutionary pressure: drugs. Loci known to confer resistance to formerly effective anti-malarial drugs, including the chloroquine resistance transporter (*pfcrt*) for chloroquine [184] and bifunctional dihydrofolate reductase–thymidylate synthase (*dhfr-ts*) for pyrimethamine [109] show all the signs: a local desert of diversity and strong LD between those SNPs found in the swept region (Figure 1.2A). Other key modifier genes, including the *P. falciparum* multiple drug resistance gene (*pfmdr1*) [56, 183] and the GTP cyclohydrolase gene (*gch1*) [83, 110], have been implicated in some drug responses, generally through adaptive changes in their copy number.

A plethora of additional genes show weaker evidence for recent positive selec-

Directional selection—Selection that favors one allele over all other alleles of a gene. The frequency of this beneficial allele can rise or can be held in check by recurrent mutation.

Selective sweep—Describes the reduced levels of genetic diversity observed around a selected locus. A selective sweep (also referred to as genetic hitch-hiking) arises because positive selection for an advantageous allele increases the frequency of not only that allele but also other alleles contained within the same haplotypes.

Long-range haplotype (LRH) test —A test for haplotypes with significantly longer than expected ranges of linkage disequilibrium; this indicates a lack of recombination between genetic markers caused by natural selection.

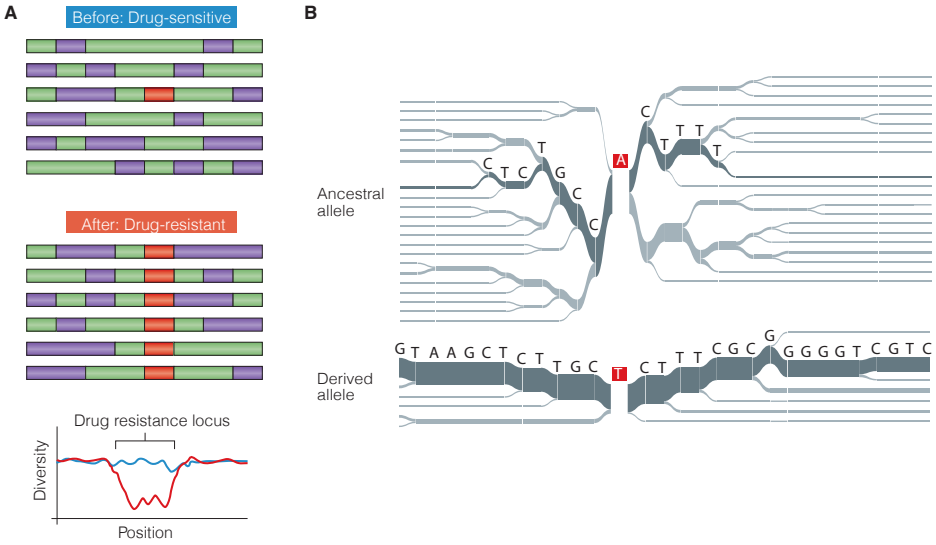


Figure 1.2: Signatures of directional selection. **(A)** Selective sweep as a consequence of selection for drug resistant parasites results in the reduction of diversity (red line) from average genome values (blue line) as a consequence of selection for an allele (red box) that confers survival under drug pressure. Neighboring alleles are maintained along with the advantageous allele, resulting in a relatively large area of the genome with reduced diversity. Identification of genomic regions with reduced diversity in drug resistant parasites as a consequence of directional selection reveals candidate drug resistant genes. **(B)** A haplotype bifurcation diagram [150] visualizes long-range associations for a given SNP. The thickness of the line represents the relative frequency of each allele in the population under study. Although the long-range associations between the ancestral (A) allele have been whittled away by recombination, the derived (T) allele maintains long-range associations with other SNPs, suggesting it arose recently and that insufficient time has passed for recombination to substantially break down these associations. Panel (B) was previously published in Volkman et al. [174].

tion [6, 28, 31, 95, 102, 107, 160, 170, 185], raising the possibility they are also associated with drug responses. Their products have various putative functional roles [170], including cell-surface adhesion, membrane transporters, genome maintenance, transcriptional regulation, metabolism and post-translational modification such as ubiquitination. Evidence for sweeps at multiple genes in a single pathway suggest that selection has been involved. For example, several genes in the ubiquitination pathway [133] are under positive selection both in worldwide populations and in a deep population analysis of parasites isolated recently in Senegal. Similarly, proteins in the fatty acid and lipid metabolism pathway have among the highest signals of selection, implicating the human or mosquito physiological state as strong selective forces on parasite survival and propagation. Key to the success of these approaches is functional characterization of candidate loci and demonstrating their involvement in conferring important clinical phenotypes such as drug resistance [170]. Preliminary functional data suggests that members of the cytoadherence-linked asexual protein or *clag* family can modulate parasite drug responses [116], consistent with the observation that some surface molecules are under positive selection [107, 170].

1.5 Finding the Genetic Basis for Specific Phenotypes

1.5.1 Parasite Phenotypes

A key challenge for identifying causal loci from genomic or functional studies is classification and quantification of robust and reliable phenotypes. Clinical phenotypes related to pathogenesis (anemia or severe disease), immunity or parasite clearance rates (associated with drug resistance) are most informative and reliable when they are assessed in parasites that are taken directly from the patient during a natural infection. However, the human host unavoidably complicates interpretation of these traits, and their assessment can often only be obtained once. A thorough phenotypic assessment would therefore necessitate large sample numbers and thoughtful study design to account for variation both within the human and parasite populations. Thus far, phenotyping has been mainly carried out

on culture-adapted [165] parasites that have been isolated from patients. These can then be tested for various in vitro phenotypes including drug response, invasion types, cytoadherence properties, the ability to produce gametocytes, or their metabolic profiles.

This section focuses on the use of genomics to identify mutations associated with altered drug responses, but these approaches could be applied to other biologically important phenotypes.

1.5.2 Linkage Analysis in *P. falciparum*

Linkage mapping in *P. falciparum* has been accomplished using laboratory genetic crosses to correlate segregation patterns in the progeny associated with specific phenotypes, including drug response [51, 65, 123, 148, 152, 179], pathogenesis [65], or mosquito infectivity [168]. Originally MSVs were used as genomic markers [157], but these have now been augmented with variants determined using whole-genome methods (array- or sequencing-based). Linkage studies leverage the reasonably high level of recombination in *P. falciparum* [76] to map the genetic determinants for specific traits: one round of recombination between parents and progeny results in large haplotype blocks that require fewer markers to identify [12]. In some geographic regions, where low diversity means that recombination rarely results in the reassortment of haplotype blocks, it may even be possible to carry out similar analyses using field isolates [46, 62]. Combining linkage analysis with other independent tests, such as association mapping, provides a potentially powerful means of prioritizing candidate genes responsible for a given phenotype.

1.5.3 Challenges of GWASs in *P. falciparum*

The molecular and genetic mechanisms of many phenotypic traits that are most relevant to elimination and eradication, such as variability in parasite responsiveness to drugs, are poorly understood. Because GWASs do not require prior knowledge of gene functions or trait mechanisms, they are useful for identifying important genetic variants in organisms such as *P. falciparum* that have many genetic loci with no known functional homologues. Although these candidate variants

require functional validation, the use of GWAS as a hypothesis-generating experiment provides a powerful starting point for identifying traits and is one of the most effective approaches available in our modern genomic toolkit.

Undertaking GWASs in *P. falciparum* requires overcoming various challenges including: identifying heritable traits, coping with low LD, using appropriate sample collection or other methods to deal with population stratification, and functionally validating associations. These challenges are described below, followed by examples of successful GWASs.

When surveying the *P. falciparum* genome for genotype-phenotype associations, only phenotypes with a strong genetic basis (those with high heritability) will be detected by a GWAS. Heritability of *P. falciparum* traits, such as drug resistance, can be variable: recent studies of parasite clearance rates in Southeast Asia found that the heritability of this phenotype depends on when and where samples are collected [10, 11]. Confounding this complication, anti-malarial responses can be quantified in various ways, including *in vitro* based metrics, such as IC_{50}, or clinically derived metrics, such as *in vivo* parasite clearance rates.

The short blocks of LD in *P. falciparum*, particularly in African populations [113, 170], are an important consideration for study design. Traditional GWASs rely on the genotyped markers being correlated to causal mutations through high LD [42]. In a population with low LD, an array-based GWAS may not have sufficient detection power unless the causal mutation is present on the array. However, when a signal is found, short LD makes localizing the signal to a single gene much easier [140]. Loss of detection power due to limited LD can therefore be circumvented by utilizing whole-genome sequencing to identify all variants in the genome. Use of sequence data for GWASs reveals stronger association signals, provides more supporting markers in areas of high LD and can detect candidate loci in areas of low LD that were previously missed by array-based GWASs. Sequencing-based

Heritability—The proportion of the total phenotypic variation in a trait that can be attributed to genetic effects.

IC_{50}—Also known as the half maximal inhibitory concentration, this is the concentration of a small molecule that results in 50% inhibition of a molecular target or cellular process. In the case of drug resistance studies, this is the concentration of drug that kills half of the parasites.

approaches are thus a promising avenue for future GWASs in malaria.

Population demography—particularly in the form of population stratification—can hinder GWAS analyses if not appropriately controlled for. The presence of closely related individuals in the dataset or, conversely, broad genetic differences between groups of samples due to differing population histories can erroneously inflate associations and produce false positives [5]. The ideal GWAS would eliminate such confounders by choosing samples with a broad range of values for each strongly heritable phenotype while sampling entirely from a single, non-stratified population. However, this is not always possible and many approaches have been developed to eliminate false positives [44, 135, 139] while analysing stratified data sets. In particular, mixed-model approaches [79, 80, 136, 164] have been used successfully to control for population stratification in malaria studies [129, 160, 170].

For studies with relatively small sample sizes, which include all malaria-based GWASs to date, gains in study power can be achieved by using multi-marker or haplotype-based association approaches [31, 42, 129, 160, 170] instead of standard single-marker tests. Positively selected variants typically lie on long haplotypes [150] that are more easily detectable by multi-marker tests, such as the LRH test.

1.5.4 GWAS Approaches to Identify Drug Resistance Loci

Although the GWAS approach is promising, we are still in the early days of applying this methodology for loci discovery in the malaria parasite, with the few GWASs primarily investigating novel variants associated with drug resistance. These studies [31, 107, 129, 160, 170, 185] generate long lists of loci hypothesized to be associated with specific phenotypes, and a current challenge is winnowing the list to the most likely candidates. One strategy involves combining results of independent tests to identify the most likely candidate genes for functional follow-up (Figure 1.3).

Mixed-model approaches—A class of linear model in which quantitative phenotypes are modelled using both fixed, genetic effects (including SNP genotypes and the principal components of population structure) and random effects (which explicitly models both the heritable and non-heritable components of phenotypic variation).

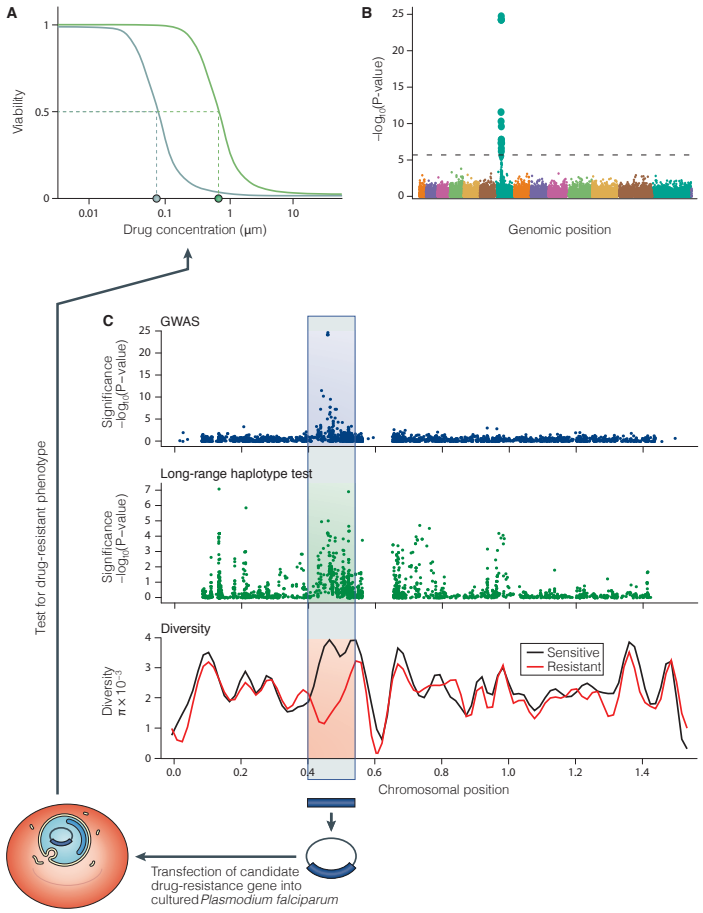


Figure 1.3: Identifying drug resistance loci using GWASs and functional studies. **(A)** Schematic representing data from a drug assay, measured as an inhibitory concentration of 50 percent (IC_{50}) that represents the drug concentration at which parasite viability is decreased by half. Such assays can be used to classify parasite responses to anti-malarial compounds, thus distinguishing a drug-resistant parasite (green line) from a drug-sensitive parasite (grey line). **(B)** Genome-wide association study (GWAS) analysis results shown as a Manhattan plot where P-values for variants across the fourteen chromosomes (represented by different colors across the x-axis) are shown. The Bonferroni level for genome-wide significance is shown as a dotted line, and genetic variants that rise above this level are associated with the drug phenotype observed in panel A. **(C)** Multiple independent analysis approaches can be combined to improve the power of locus identification for functional follow up. For example, alleles identified by GWASs, long-range haplotype (LRH) tests, or diversity and divergence analyses are identified, and a gene expression construct is created comprising the genetic variant that is associated with drug resistance. To test whether each gene variant is necessary and sufficient to confer the observed drug resistant phenotype, each putative drug-resistance variant is introduced into a drug sensitive parasite by transfection, followed by testing for drug resistance.

Several previously known drug targets and one novel GWAS result have now been functionally validated, lending support to the idea of using GWAS to find resistance loci. The GWAS by Van Tyne et al. [170] identified a highly polymorphic locus, *PF10_0355*, as being associated with halofantrine resistance based on a relatively small set of globally diverse parasites. *PF10_0355* was classified as a member of the *msp3* gene family [125]. When a variant of *PF10_0355* from a drug resistant parasite was introduced into a drug sensitive parasite through transfection, the parasite was rendered resistant not only to halofantrine but also to chemically similar drugs (mefloquine and lumefantrine), but not to chemically distinct compounds (chloroquine, artemisinin, atovaquone). This is the first functional demonstration that a potential drug resistance locus identified by a GWAS confers a drug resistance phenotype.

This study employed a modest number of parasites sampled from many populations using a limited marker set. More recent studies now assay larger parasite numbers from single populations using sequencing data to capture essentially all genetic variation and demonstrating significant increases in power [129].

1.6 Conclusions and Dissertation Overview

Genomic analysis is providing rich and unique insight into the biology and population history of *P. falciparum*. The strongest selective forces that leave visible imprints on the parasite genome are the signals of balancing selection on surface proteins due to human immune pressure, as well as the signals of directional selection due to anti-malarial drugs. The goal of contemporary disease elimination efforts should be to prevent selective sweeps from occurring in response to current treatment strategies. This can be achieved through careful genomic surveillance and flexible, combinatorial application of drug treatments.

In the decade since the publication of the *P. falciparum* genome sequence [57], modern genomics has advanced our understanding of the parasite. However, we are just beginning to leverage population genetic information to discover the variants responsible for clinically important phenotypes. The work in this dissertation plays an important role in the development of these tools and methods.

In Chapter 2, I perform one of the first GWASs for antimalarial drug resistance. As the only such study to date that includes a functional validation, it demonstrates the utility of the GWAS approach to find causal markers for poorly understood phenotypes. It also influenced the methodologies used by others in the malaria genomics field, with later drug studies adopting the linear model [31] and mixed model [160] association methods first successfully demonstrated here.

In Chapter 3, I demonstrate the use of whole-genome sequence data for association studies and the clear deficiencies of array-based association studies. Other groups in the field are now moving from array to next generation sequence data as well [7, 95, 102]. Importantly, I also demonstrate the novel use of the cross-population extended haplotype homozygosity selection test (XP-EHH, a type of LRH test, [151]) in the context of a GWAS. Similar approaches that combine XP-EHH and association tests can be seen soon afterwards in work by others [160].

In Chapter 4, I explore the use of time course data to directly infer the strength of directional selection throughout the genome. Such approaches have not yet been shown on a genome-wide scale, nor have they been applied in the malaria parasite previously. The results provide an additional source of insight into the evolutionary forces shaping the *P. falciparum* genome. Here I demonstrate a proof of concept study, and outline the parameters for properly designing and conducting a whole-genome time-course selection study.

The research described in this dissertation has the potential to provide substantial new insights into the recent evolution of the malaria parasite, particularly as it relates to drug resistance. The results highlight a number of candidate genes and mutations that can lead to biological insights and epidemiological surveillance tools. But more importantly, this thesis develops and provides new genomic methods that can be used by others in conducting similar types of research. As such, it makes significant advancements to the field of malaria genomics by improving the way research is conducted globally.

This chapter was originally published as Van Tyne, Park*, et al., Identification and functional validation of the novel antimalarial resistance locus PF10_0355 in Plasmodium falciparum, PLoS Genetics (7), 2011 [170]. This work was also presented at the American Society of Tropical Medicine and Hygiene (Nov 2010, Atlanta, GA). See page xiv for details on author contributions.*

doi://10.1371/journal.pgen.1001383

2

Genome-wide Association Studies Identify the Antimalarial Resistance Locus *PF10_0355*

The *Plasmodium falciparum* parasite's ability to adapt to environmental pressures, such as the human immune system and antimalarial drugs, makes malaria an enduring burden to public health. Understanding the genetic basis of these adaptations is critical to intervening successfully against malaria. To that end, we created a high-density genotyping array that assays over 17,000 single nucleotide polymorphisms (~1 SNP/kb), and applied it to 57 culture-adapted parasites from three continents. We characterized genome-wide genetic diversity within and between populations and identified numerous loci with signals of natural selection, suggesting their role in recent

adaptation. In addition, we performed a genome-wide association study (GWAS), searching for loci correlated with resistance to thirteen antimalarials; we detected both known and novel resistance loci, including a new halofantrine resistance locus, *PF10_0355*. Through functional testing we demonstrated that *PF10_0355* overexpression decreases sensitivity to halofantrine, mefloquine and lumefantrine but not to structurally unrelated antimalarials, and that increased gene copy number mediates resistance. Our GWAS and follow-on functional validation demonstrate the potential of genome-wide studies to elucidate functionally important loci in the malaria parasite genome.

2.1 Introduction

Plasmodium falciparum malaria is a major public health challenge that contributes significantly to global morbidity and mortality. Efforts to control and eliminate malaria combine antimalarial drugs, bed nets and indoor residual spraying, with vaccine development a longer-term goal. Genetic variation in the parasite population threatens to undermine these efforts, as the parasite evolves rapidly to evade host immune systems, drugs and vaccines. Studying genetic variation in parasite populations will expand our understanding of basic parasite biology and its ability to adapt, and will allow us to track parasites as they respond to intervention efforts. Such understanding is increasingly important as countries move towards reducing disease burden and the ultimate elimination of malaria.

Given the potential impact of rapid evolution of *P. falciparum* in response to control and eradication strategies, discovery and characterization of *P. falciparum* genetic diversity has accelerated in recent years. Since the first malaria genome was sequenced in 2002 [57], over 60,000 unique SNPs have been identified by concerted sequencing efforts [74, 106, 174], and several genomic tiling arrays [24, 45, 75, 83, 161] and low-density SNP arrays [107, 113] have been developed to query this genetic variation. Recently the first malaria GWAS was published [107], in which 189 drug-phenotyped parasites from Asia, Africa and the Americas were genotyped using a low-density array (3,257 SNPs); that study identified loci under positive selection and found several novel drug resistance candidates.

For our study, we adopted two strategies for identifying genes involved in the malaria parasite's adaptive response: searching for signals of recent or ongoing natural selection, and searching for loci associated with one important clinical adaptation—resistance to antimalarial drugs. To make these searches possible, we began by sequencing 9 geographically diverse strains of *P. falciparum* to identify novel variation, thereby doubling the number of publicly available SNPs to 111,536 (all accessible at `http://plasmodb.org/`), and used these SNPs to develop a high-density genotyping array assaying 17,582 validated markers. After characterizing linkage disequilibrium and population structure in our samples, we used the arrays to search for evidence of both ongoing balancing selection and recent positive selection, and to carry out a GWAS that sought loci associated with resistance to thirteen antimalarial agents. We then followed up one of the novel loci associated with drug resistance in order to verify that variation there was biologically involved in modulating drug response.

2.2 Results

2.2.1 Genetic Diversity

We identified global population structure among malaria parasites using principal components analysis (PCA) of 57 genotyped culture-adapted parasite samples (Figure 2.1A, Table C.1, Figure C.1). African, American and Asian samples form three distinct clusters, reflecting the likely independent introduction of *P. falciparum* from Africa into Asia and the Americas. There was little evidence for structure within Africa, suggesting high gene flow throughout the region (Figure C.1). Asian and American parasites however both show substantial internal structure.

There are also dramatic differences in linkage disequilibrium (LD) between populations, with substantial LD extending less than 1 kb in Senegal, 10 kb in Thailand, and 100 kb in Brazil (Figure C.2). These observations are consistent with previous findings, which showed that LD decays more rapidly in Africa, due either to founder effects in other continents [8] or to elevated outcrossing frequencies in

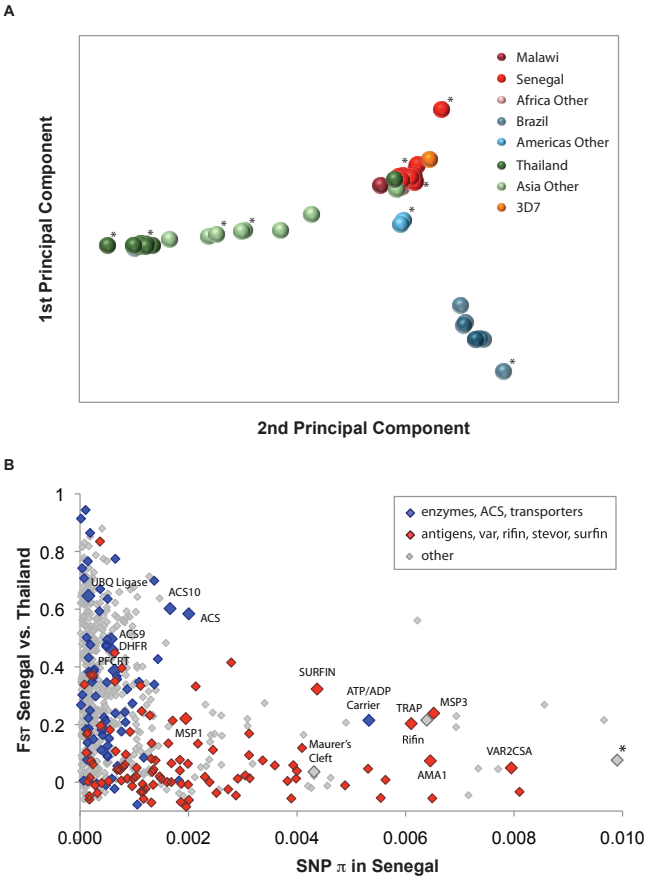


Figure 2.1: Parasite global population structure and genetic diversity vs. divergence. **(A)** Population structure is visualized using the first two principal components of genetic variation for 57 parasites. Solid circles represent individual parasites, with colors assigned by reported origin: Africa in red, America in blue, and Asia in green. The nine strains used for ascertainment sequencing are indicated with (*). **(B)** Genetic diversity (SNP π) in Senegal versus divergence (F_{ST}) between Senegal and Thailand is reported for 688 genes containing > 3 successfully genotyped SNPs. Blue diamonds: enzymes, acyl-CoA synthetases (ACS) or transporters; red diamonds: antigens, vars, rifins, stevors or surfins; gray diamonds: all other genes. Gene IDs (`http://plasmodb.org/`) for highlighted genes are listed in Table C.7. A gene with unknown function is flagged with (*) to indicate that SNP π is off-scale (0.014).

Africa [8, 105], where higher transmission intensity leads to a greater likelihood of sexual outcrossing rather than selfing within the mid-gut of vector mosquitoes.

The short LD in malaria, driven by high levels of recombination, means that a high density of markers is required to identify candidate loci in association studies, since causal variants not on the array can seldom be tagged by neighboring alleles (Table C.2). On the other hand, short LD can aid in fine-mapping candidate associations and greatly accelerates the search for causal genes. Short LD also aids in identifying genomic regions under recent positive selection with recombination-based methods, since the increased LD in selected regions should stand out against the short-LD background.

2.2.2 Detecting Signals of Natural Selection

We expect that many parasite proteins that interact with the host immune system will be under balancing selection, because they will be under selective pressure to maintain high levels of diversity. Indeed, previous studies have shown that regions of the *P. falciparum* genome that are highly polymorphic and appear to be under balancing selection encode antigens that are recognized by the human immune system [106]. We examined evidence for balancing selection in our data by searching for regions with high nucleotide diversity (as measured by SNP π) and low population divergence (as measured by F_{ST}) (Figure 2.1B). When we examined the loci lying in this region of the graph (Figure C.3), we found a number of known antigens and vaccine candidates. Loci in the same region with unknown function are thus potential novel antigens that trigger human immune response to malaria, and may prove useful as biomarkers or as candidate vaccine molecules.

We carried out a similar search for loci under positive selection by identifying regions with both low nucleotide diversity within Senegal and Thailand and high population divergence between the two populations (Figure 2.1B). We observed throughout the genome that nucleotide diversity was lower for nonsynonymous SNPs than for intergenic SNPs (Figure C.4), a characteristic result of widespread purifying selection. At the same time, nonsynonymous SNPs exhibited significantly greater divergence than intergenic SNPs in all pairwise population com-

parisons, suggesting the effect of positive selection in local *P. falciparum* populations. Nonsynonymous SNPs with low diversity within a population and high divergence between the two populations studied may represent polymorphisms responsible for adaptive evolution.

We also carried out a genome-wide scan for recent positive selection using the long-range haplotype (LRH) test [150], which identifies common variants that have recently spread to high prevalence using recombination as a clock. Approximately 15 genes were identified as having undergone recent positive selection by this approach, including known drug resistance loci (*pfcrt* and *dhfr*) as well as multiple members of the acyl-CoA synthetase (ACS) and ubiquitin protein ligase families (Figure C.5 and C.6); these latter genes also exhibit high divergence between Senegal and Thailand (Figure 2.1B), evidence for selection that is recent and population-specific. Taken as a group, the genes identified by the LRH test show a significant enrichment for higher than average population divergence (as measured by F_{ST}, Mann-Whitney $U = 1583$, $P = 0.0071$). All of these loci (Table C.3, Dataset 1), which include genes in the folate metabolism, lipid biosynthesis and ubiquitin pathways, should be viewed as candidates for functional follow-up and further characterization.

2.2.3 Genome-wide Associations with Drug Resistance

In order to directly assess the genetic basis for one important response to antimalarial intervention, we carried out a GWAS to identify loci associated with drug resistance in parasites. This same approach can potentially be applied to many other clinically relevant malaria phenotypes, e.g. host response, invasion, and gametocyte formation. Our first step was to measure drug resistance (IC_{50} values) to 13 antimalarial drugs (amodiaquine, artemether, artesunate, artemisinin, atovaquone, chloroquine, dihydroartemisinin, halofuginone, halofantrine, lumefantrine, mefloquine, piperaquine and quinine) in 50 culture-adapted parasites using a high-throughput assay (Tables C.4 and C.5, Appendix C, Dataset 1).

We performed the genome-wide association analysis using two statistical tests: efficient mixed-model association (EMMA) and a haplotype likelihood ratio

(HLR) test (Figures C.7, C.8, C.9, C.10, Appendix C). EMMA identifies quantitative trait associations in individuals with complex population structure and hidden relatedness; it has previously been shown to outperform both PCA-based and λ_{GC}-based correction approaches in highly inbred and structured mouse, maize, and Arabidopsis populations [79]. EMMA is particularly applicable for small and structured sample sets: one of its first applications was in a study of 24 diploid mouse strains [79], essentially the same sample size as in our study (50 haploid strains). The HLR test is a multi-marker test designed to detect the association of a single haplotype with a phenotype, and is particularly powerful when the associated haplotype experienced recent strong selection (and is therefore long) and occurs on a low-LD background [90]; it is therefore particularly appropriate for this study. We addressed the effect of population structure in the HLR test using population-specific permutation (Section 2.4). When used together, these two complementary approaches provide a highly sensitive screen for association signals within the *P. falciparum* genome.

The well-characterized chloroquine resistance locus, *pfcrt*, served as a positive control for our GWAS methods (Figure 2.2A and 2.2C, Table C.2), an important test given our small sample size and the limited LD present in *P. falciparum*. As expected, we found evidence for association with resistance to chloroquine using both tests, consistent with previous studies [107]; EMMA yielded evidence for association with genome-wide signficance, while the signal from the HLR test fell just short of genome-wide significance (Figure 2.2C).

Applying the same tests to the other drug phenotypes, we detected numerous novel loci showing significant associations with drug resistance (Figure 2.2A and 2.2D, Table 2.1). Quantile quantile plots for each test demonstrate that we were able to effectively control for population structure (Figure 2.2B). Despite our small sample size and the low LD in *P. falciparum*, in total eleven loci achieved genome-wide significance for association with resistance to five different drugs: amodiaquine, artemisinin, atovaquone, chloroquine and halofantrine. In most cases, the short extent of LD allowed localization to individual genes. Among the loci identified were various transporters and membrane proteins, as well as five conserved genes with unknown function (Table 2.1, Dataset 1).

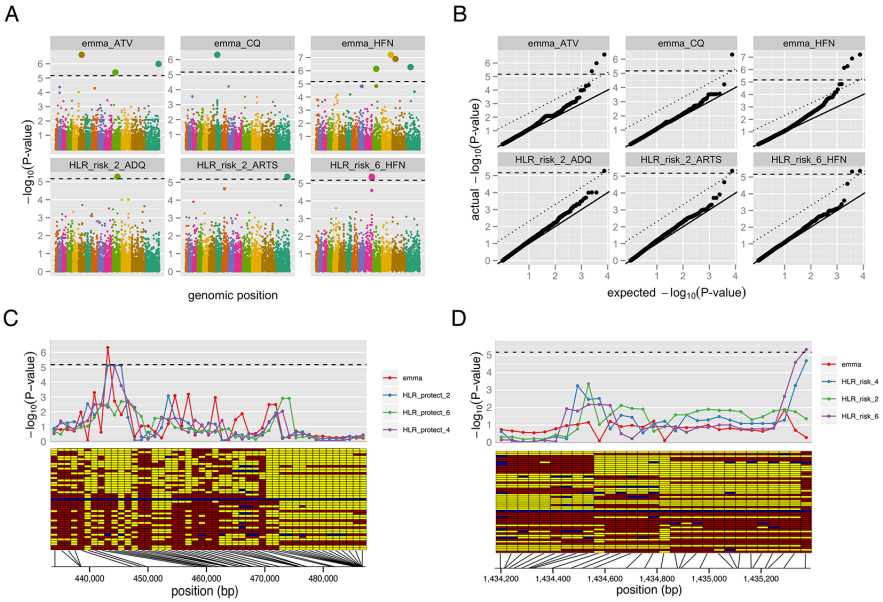


Figure 2.2: Genome-wide association study (GWAS) results. **(A)** Genome-wide significant associations were found for five antimalarials (out of thirteen tested) using EMMA and HLR tests. They include *pfcrt* (chromosome 7) associated with chloroquine resistance and eleven novel associations with resistance to several drugs, listed in Table 2.1. **(B)** Quantile-quantile plots for the P-value distributions in (A) show no significant confounding from population structure. Bonferroni-corrected genome-wide significance is marked with a dashed line; Benjamini-Hochberg significance is marked with a dotted line. **(C-D)** Close-ups are shown of the GWAS signal (top) and haplotypes (bottom) for resistance to **(C)** chloroquine (CQ) around the gene *pfcrt* and **(D)** halofantrine (HFN) around the gene *PF10_0355*. Yellow: sensitive allele; red: resistant allele; Blue: no data. Isolates are ordered by IC_{50}, with the highest IC_{50} on the bottom.

Table 2.1: Eleven genome-wide significant associations with antimalarial drug resistance. Positions are given with respect to the PlasmoDB 5.0 reference assembly of 3D7. Drug abbreviations are ATV: atovaquone; CQ: chloroquine; HFN: halofantrine; ADQ: amodiaquine; ARTS: artemisinin. The HLR test for CQ-*pfcrt* association is just below the genome-wide significance threshold and is omitted here, but is shown in Figure 2.2C.

chr	SNPs	test	drug	P-value	genes	PlasmoDB description
6	674,154	EMMA	ATV	2.36×10^{-7}	PFF0785w	Ndc80 homologue, putative
7	459,787	EMMA	CQ	4.72×10^{-7}	MAL7P1_27	chloroquine resistance transporter
10	1,435,226, 1,435,286, 1,435,370, 1,437,695, 1,437,718, 1,441,590, 1,444,868	HLR-risk-6 (2 overlapping hits)	HFN	4.71×10^{-6}, 4.25×10^{-6}	PF10_0355, PF10_0356	erythrocyte membrane protein putative, liver stage antigen 1
11	657,349	EMMA	ATV	4.01×10^{-6}	PF11_0178	conserved unknown
11	738,407	EMMA	HFN	7.20×10^{-7}	PF11_0203	peptidase, putative
11	1,123,028, 1,124,030	HLR-risk-2	ADQ	5.26×10^{-6}	PF11_0302	conserved unknown
12	1,964,935	EMMA	HFN	6.15×10^{-8}	PFL2285c	conserved unknown
13	757,689	EMMA	HFN	1.28×10^{-7}	PF13_0101	conserved unknown
14	1,233,470	EMMA	HFN	5.32×10^{-7}	PF14_0293	conserved unknown
14	2,814,793, 2,815,714	HLR-risk-2	ARTS	4.90×10^{-6}	PF14_0654	aminophospholipid transporter, putative
14	3,130,449	EMMA	ATV	1.03×10^{-6}	PF14_0729	early transcribed membrane protein 14.2

2.2.4 Functional Validation of a Novel Resistance Candidate

Demonstrating that a signal of association actually reflects a causal molecular process requires functional testing and validation of the candidate locus, both because of concerns about power and reproducibility of genetic association tests, and because even a robust statistical correlation need not imply biological causation. To confirm the ability of GWAS to identify functionally relevant candidates, we investigated one of our association findings, *PF10_0355*, in greater depth. This gene contains multiple SNPs associated with halofantrine resistance (Figure 2.2D), and encodes a putative erythrocyte membrane protein (http://plasmodb.org/) characterized by high genetic diversity.

We set out to determine the role of *PF10_0355* in halofantrine resistance by transfecting halofantrine-sensitive Dd2 parasites with episomal plasmids containing the *PF10_0355* gene from a halofantrine-resistant parasite (SenP08.04), a technique that is used routinely for stable transgene expression [36]. Two independent transfectants overexpressing the *PF10_0355* gene from SenP08.04 both showed reduced susceptibility to halofantrine when compared with the Dd2 parent or a transfection control (Figure 2.3A), suggesting that this gene is indeed involved in modulating parasite drug response.

Two independent transfectants overexpressing the endogenous *PF10_0355* gene from halofantrine-sensitive Dd2 also showed reduced susceptibility to halofantrine (Figure 2.3A), however, pointing to a role of overexpression in the observed resistance. Because *PF10_0355* is annotated as a putative erythrocyte membrane protein and belongs to the merozoite surface protein 3/6 family, we tested the hypothesis that the observed effect was the by product of a growth or invasion-related process, rather than resistance due to a direct interaction with the antimalarial itself. To that end, we expanded our drug testing in the transfectant lines to include other antimalarials, some structurally related and some unrelated to halofantrine.

Overexpression of *PF10_0355* from either the Dd2 or the SenP08.04 parent caused increased resistance to the structurally related antimalarials mefloquine and lumefantrine (Figure 2.3B and 2.3C), but had no effect on parasite sus-

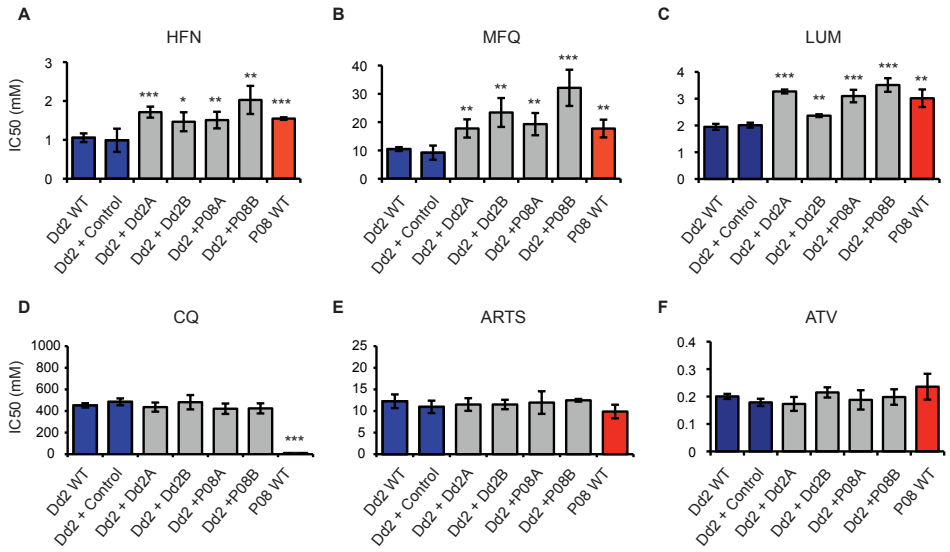


Figure 2.3: Overexpression of *PF10_0355* decreases parasite susceptibility to halofantrine (HFN) and related antimalarials. Parasite susceptibility to six antimalarials was measured by 3H-hypoxanthine incorporation. Comparisons were made between Dd2 (HFN-sensitive strain) and SenP08.04 (HFN-resistant strain), as well as 4 transfected lines. "Dd2+Dd2": Dd2 parasites overexpressing *PF10_0355* from Dd2; "Dd2+P08": Dd2 parasites overexpressing *PF10_0355* from SenP08.04. Overexpression of *PF10_0355* decreases parasite susceptibility to **(A)** HFN and structurally related **(B)** mefloquine (MFQ) and **(C)** lumefantrine (LUM). Overexpression of *PF10_0355* does not alter parasite susceptibility to **(D)** chloroquine (CQ), **(E)** artemisinin (ARTS) or **(F)** atovaquone (ATV). Mean $IC_{50} \pm$ standard error is shown. Significance levels: *: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$.

ceptibility to the structurally unrelated antimalarials chloroquine, artemisinin or atovaquone (Figure 2.3D and 2.3E). Indeed, we found evidence of cross-resistance between halofantrine and both mefloquine and lumefantrine (Figure 2.4). We also observed cross-resistance between halofantrine and artemisinin, which is expected as cross-resistance between aminoquinolines and artemisinin compounds has been previously demonstrated [107, 134] and resistance to all these drugs has been shown to be mediated by changes in *pfmdr1* copy number [29, 153]. Overexpression of *PF10_0355*, however, alters parasite susceptibility to the aminoquinolines but not to artemisinin, suggesting that this effect is specific for that set of structurally related compounds and distinct from the effect of *pfmdr1*, which seems to exert a global effect of resistance to unrelated compounds (i.e. both aminoquinolines and artemisinins). Using the Dd2 parasite line, which has amplified *pfmdr1* copy number, as a background for *PF10_0355* overexpression allowed us to distinguish between cross-resistance to a structurally related class of compounds (mediated by *PF10_0355* overexpression) and pan-resistance to multiple classes of drugs.

Given that overexpression of the *PF10_0355* gene both from a halofantrine-resistant and from a sensitive parasite conferred resistance to halofantrine-related drugs, we investigated whether gene amplification might be driving the observed resistance, as it often does for antimalarial drugs [9, 56, 110, 137, 146, 183]. We quantified *PF10_0355* copy number in our transfectants and found that the transfectant with the highest IC_{50} for all three drugs (Dd2+P08B) also had the highest *PF10_0355* copy number, as measured by quantitative PCR (qPCR) (Figure 2.5A). Furthermore, when we examined the *PF10_0355* gene on our SNP array, we detected a substantial increase in hybridization intensity at the *PF10_0355* locus compared to the genome average, suggesting that this gene is amplified in some parasites (Figure 2.5B). The amplified region appears only to contain the *PF10_0355* gene itself and not surrounding loci. We observed a similar pattern at *pfmdr1* on chromosome 5, where copy number variation is well established (Figure C.11). Follow-up qPCR analysis of 38 parasite lines confirmed that parasites with amplified *PF10_0355* have a greater mean halofantrine IC_{50}. (Figure 2.5C, Table C.11, Dataset 1). Copy number variation was further confirmed in a num-

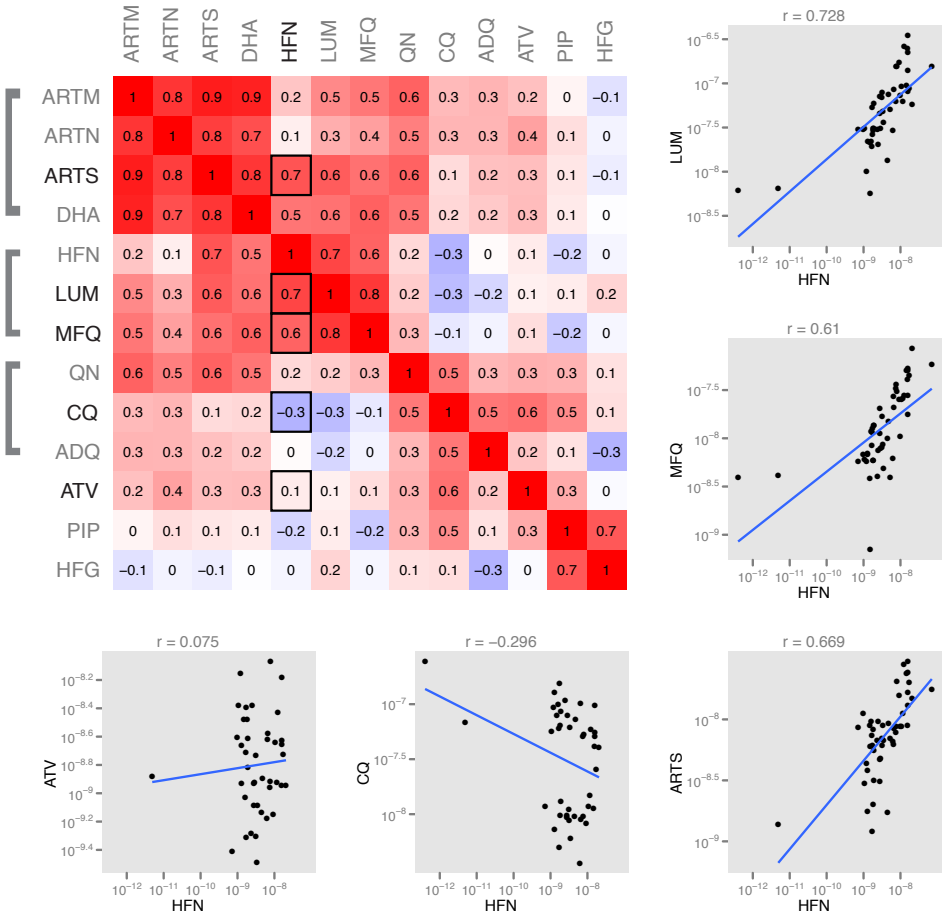


Figure 2.4: Correlations between antimalarial drugs tested. **(A)** Pearson correlation values (r) between $\log_{10}(IC_{50})$ values are rendered as a color in a symmetric correlation matrix (red: correlated; white-uncorrelated, blue: inversely correlated). Thirteen antimalarials are measured: artemether (ARTM), artesunate (ARTN), artemisinin (ARTS), dihydroartemisinin (DHA), halofantrine (HFN), lumefantrine (LUM), mefloquine (MFQ), quinine (QN), chloroquine (CQ), amodiaquine (ADQ), atovaquone (ATV), piperaquine (PIP), and halofuginone (HFG). Drugs are grouped by structural relatedness. **(B-F)** Correlation plots are given with a linear regression line for HFN compared to the 5 other drugs tested for antimalarial resistance with *PF10_0355* overexpression: **(B)** LUM, **(C)** MFQ, **(D)** ATV, **(E)** CQ, and **(F)** ARTS.

ber of parasites by quantitative Southern blotting (Figure C.12).

2.3 Discussion

In this study we used natural selection and genome-wide association methods to probe the genetic basis of adaptation in *P. falciparum*. These approaches are complementary: scanning for selected loci permits an unbiased search for unknown adaptive changes, but provides little information about the processes at work, while GWAS gives a focused look at one easily identified (and clinically critical) adaptive phenotype. Results from both approaches open up new avenues for study, as we seek to understand the biological significance of the findings.

The specifics of our strategy were designed to cope with two potential limitations in applying genome-wide population genetic approaches to malaria: small sample sizes, due to the difficulty in adapting parasites to culture and assessing drug and other phenotypes; and a lack of correlation (LD) between nearby variants in the parasite genome, which limits our ability to infer untyped SNPs from genotyped markers. The second limitation we addressed by developing a high-density genotyping array (based on new sequencing), to increase the fraction of genetic variation that we could directly interrogate, while the effect of the first was mitigated by the phenotype we targeted in our GWAS.

Drug resistance is a phenotype well-suited for GWAS because it is expected to be caused by common alleles of large effect at few genomic loci [65]. If this is the case, associations will be much easier to detect than in a typical human GWAS, in which the phenotype is caused by alleles at many loci that are either rare or of small effect. Additionally, the haploid nature of the intra-erythrocytic stage of *P. falciparum* further heightens GWAS power by eliminating the issue of allelic dominance. Finally, the increased LD caused by recent selection for drug resistance counteracts the loss of power that comes from short LD, small sample size, and the temporal and geographic stratification of the parasite population that we examined. Thus, despite the potential limitations, we were able to detect a known drug resistance locus (*pfcrt*), observed little P-value inflation in our GWAS data (Figures C.8, C.9, C.10), and identified a number of genome-wide significant loci

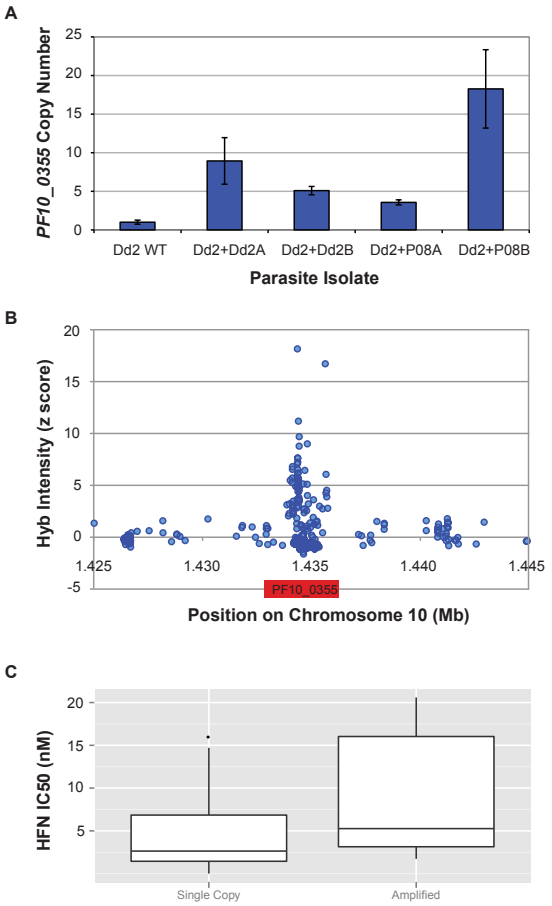


Figure 2.5: Copy number variation at *PF10_0355* is associated with HFN resistance. **(A)** Mean *PF10_0355* copy number ($\pm$ standard deviation for three replicates) in the parent Dd2 and transfected lines from qPCR analysis. Dd2+Dd2: Dd2 parasites overexpressing *PF10_0355* from HFN-sensitive Dd2; Dd2+P08: Dd2 parasites overexpressing *PF10_0355* from HFN-resistant SenP08.04. Copy number was compared to the reference locus *PF07_0076*. **(B)** Increased hybridization intensity at *PF10_0355* on the high-density SNP array, measured by Z-scores for normalized and background-corrected data, for the HFN-resistant isolate SenP19.04. **(C)** Strains with increased copy number of *PF10_0355* (as measured by qPCR $> 1.2\times$ 3D7) show a significantly higher resistance to HFN ($p = 0.02$, Student t-test).

associated with drug resistance. Part of this success was likely due to specific tests we used to account for population structure.

Going beyond these statistical tests, we went on to functionally validate one of these loci, demonstrating that increased *PF10_0355* copy number confer resistance to three structurally related antimalarial drugs. This demonstrates the feasibility of coupling GWAS and functional testing in the malaria parasite for identifying and validating novel drug resistance loci and illustrates the power of GWAS to find functionally important alleles.

Comparing our results to the recent GWAS described by Mu et al. [107], which was also directed at finding drug-resistance loci, we see that, beyond the well-known pfcrt locus, there was no overlap between the associations identified by each study. Differing sets of drugs tested and analytical methods explain much of the disagreement. Of the eleven candidate associations in Table 2.1, one (that with *pfcrt*) was found by both studies, eight were associations with drugs not assayed in Mu et al. (atovaquone and halofantrine), and two were found only with a haplotype-based test, an approach not used by Mu et al.. Our candidate locus at *PF10_0355*, in fact, would not have been detectable in the Mu et al. study because it was identified only by the multi-marker HLR test, because it involved an association with halofantrine, and because the Mu et al. genotyping array lacked markers within 4 kb of the gene (`http://plasmodb.org/`).

Different parasite populations and marker sets probably explain many of the dihydroartemisinin, mefloquine and quinine associations identified by Mu et al. but not seen in our data set. The studies used different parasite population sets—theirs was weighted toward southeast Asian strains, and ours toward African strains—and selection pressures and selected alleles can both vary between populations. Our smaller sample size also means that we might lack power to identify some associations accessible to Mu et al.. These difficulties are reflected in human GWAS studies as well, where the ability to replicate associations using multiple tests and in different sample sets has also been challenging to achieve [92].

Ultimately, the disparities in loci identified point to the role of population analysis as a tool for candidate gene discovery and not as a definitive study. Even

within each study, there is little overlap between the signals observed with different methods—our study detects only one gene (*pfcrt*) by both GWAS tests (EMMA and HLR), while Mu et al. detected only two genes (unknowns, not *pfcrt*) by both of their GWAS tests (Eigensoft and PLINK). Even a well-designed GWAS serves only as a hypothesis-generating experiment, and it is vital to empirically validate candidate loci associated with a phenotype of interest. Especially given the small sample sizes and relatively sparse marker density used in both malaria GWAS studies to date, functional validation of candidates is necessary to address concerns about false positive results.

Our functional result, that increased *PF10_0355* copy number confers decreased susceptibility to halofantrine, mefloquine and lumefantrine, raises additional questions for study. Further work will be needed to determine the precise contributions of copy number variation and gene mutation to the parasite's response to these drugs. The biological function of this gene's product is unknown, but previous work indicates putative localization to the parasite surface [154], as well as it being a potential target of host immunity and balancing selection [125]. While the protein itself does not appear to be a transporter, it is possible that it directly binds drug or perhaps couples with transport proteins to modulate drug susceptibility; interaction between membrane transporters and non-channel proteins has been demonstrated in cancer, plant and yeast systems [17, 58, 100]. Additional experiments are certainly required to determine the precise role of *PF10_0355* in modulating parasite response to this class of compounds, including assessing its relevance to resistance in natural populations, but it is clear that alteration of this locus can mediate drug resistance in *P. falciparum*.

Although halofantrine, mefloquine and lumefantrine are not commonly used as primary interventions, widespread halofantrine use has recently been documented in West Africa. Notably, halofantrine was used to treat nearly 18 million patients between 1988 and 2005 [20, 60], and it remains in production and use today. Use of halofantrine, mefloquine or lumefantrine as monotherapy may further explain how mutations and copy number variation in the *PF10_0355* gene were selected. Lumefantrine is also currently used as a partner drug in the artemisinin-based combination therapy (ACT) Coartem. The shorter half-life of artemether

allows lumefantrine to be present as monotherapy, making it vulnerable to selection of drug resistant mutants. As genetic loci associated with drug responses are identified and validated, these provide new molecular biomarkers to evaluate drug use and response in malaria endemic settings. Thus, our findings have implications for defining molecular biomarkers for monitoring partner drug responses as intervention strategies, such as ACTs, are applied.

Beyond identifying a novel drug resistance locus, this study illustrates the general utility of a GWAS approach for the discovery of gene function in *P. falciparum*. Even with a small and geographically heterogeneous sample of parasites, we identified a number of new loci associated with drug response and validated one of them. Larger samples from a single population will have much greater power to detect additional loci, including those where multiple and low frequency alleles contribute to resistance. Future GWAS have the potential both to provide greater insights into basic parasite biology and to identify biomarkers for drug resistance and other clinically relevant phenotypes like acquired protection, pathogenesis, and placental malaria.

Future GWAS will be able to counteract the loss of power caused by low LD, either by focusing on parasite populations with reduced outcrossing rates, or by studying cases of very strong selective pressure. This issue will soon become moot, however, as the declining cost of whole genome sequencing makes it practical to assay every nucleotide in the genome on a routine basis. Culture-adapted parasites are amenable to robust and reproducible phenotypic characterization, but their limitations—the potential for artifactual mutations during adaptation and for a biased selection of clones within a given infection—mean that genetic changes identified using them require both functional validation and demonstration that the changes are important during natural infection. As direct sequencing of clinical isolates with demonstrable clinical phenotypes such as ex vivo drug response or invasion properties becomes increasingly feasible, sequencing will enable us to directly identify genetic changes in the parasite associated with clinically relevant phenotypes. In the years ahead, genome analysis of *P. falciparum* has the potential to identify genetic loci associated with many phenotypes, enhance our understanding of the biology of this important human pathogen, and inform the devel-

opment of diagnostic and surveillance tools for malaria eradication.

2.4 Methods

2.4.1 Parasites, Drug Testing, and DNA Isolation

Parasite samples and origins are detailed in Appendix C and Table C.1. Parasites were maintained by standard methods [165] and were tested for their response to amodiaquine, artemether, artesunate, artemisinin, atovaquone, chloroquine, dihydroartemisinin, halofuginone, halofantrine, lumefantrine, mefloquine, piperaquine and quinine according to the methods outlined by Baniecki et al. [16] (Table C.4, Figure C.13, Appendix C). Follow-up drug testing was done by measuring uptake of ^{3}H-hypoxanthine [177]. Nucleic acids were obtained from parasite cultures using Qiagen genomic-tips (Qiagen, USA). All DNA samples were evaluated by molecular barcode [39].

2.4.2 Array Genotyping

We sequenced nine geographically diverse parasite isolates to 1.25× coverage, nearly doubling the number of publicly available SNPs to 111,536 (Appendix C). These parasites had been previously sequenced to 0.25× coverage [174] and the deeper sequencing allowed for more thorough SNP discovery. Using this combined marker set, we created a high-density Affymetrix-based SNP array for *P. falciparum* containing 74,656 markers. Arrays were hybridized to 57 independent parasite samples (Table C.1), including 17 previously sequenced strains used as a validation set. Genotype calls were produced using the BRLMM-P algorithm [1]. Markers that did not demonstrate perfect concordance between sequence and array data for the 17 strains were removed (Appendix C). The remaining 17,582 SNPs constituted the high-confidence marker set used throughout this study (median marker spacing 444 bp, mean spacing 1,316 bp). All genomic positions and translation consequences are listed with respect to the PlasmoDB 5.0 assembly and annotation. SNP genotype data are publicly available on `http://plasmodb.org/`

(release 6.0, July 2009) and dbSNP (Build B134, May 2011), accessible by searching for submission batches Pf_0002 (sequencing of nine isolates) and Pf_0003 (genotyping of 57 isolates) from submitter BROAD-GENOMEBIO. Genotype data is also available as Dataset 2.

2.4.3 Principal Component Analyses

Principal components analysis (PCA) was performed using the program SmartPCA [130]. All single-infection samples were used for the analysis in Figure 2.1. Samples that tightly clustered with the wrong continental population (A4, Malayan Camp and T2_C6) represented likely cases of contamination and were thus omitted from all other analyses.

2.4.4 Diversity/Divergence Analysis

We measured diversity using a statistic we term 'SNP π,' which quantifies the average number of pair-wise differences among samples from a given population at assayed SNPs. Population divergence was measured using F_{ST}, calculated using the method of Hudson et al. [69]. Statistical evaluation of the significance of differences in SNP π and F_{ST} among populations was performed using a bootstrapping approach, where the SNP set was re-sampled with replacement and each statistic recomputed 1000 times.

2.4.5 Linkage Disequilibrium (LD) Analysis

The statistic r^2 was calculated within each population for all pairs of SNPs sharing the same chromosome [67]; pairs were binned by distance and averaged within each bin. The level of LD between unlinked markers was estimated by calculating r^2 between all pairs of SNPs on different chromosomes. To determine the bias caused by small sample size, the unlinked calculation was repeated, with the change that for each pair of SNPs, the genotype for one was taken from one strain while the genotype for the second was taken from another strain. This background value of r^2 was calculated separately for the possible pairs of different strains and

then averaged. Only single infections, as assessed by molecular barcode, were used.

2.4.6 Long Range Haplotype (LRH) Analysis

Because of the small number of samples, LRH results for individual continental populations had a high level of variance. Thus, we pooled together samples from Africa ($n = 26$) and Asia ($n = 18$, excluding India), as suggested by our PCA analysis. SNPs included in the analysis had a minor allele frequency of at least 0.05 and a call rate of at least 0.8; missing genotypes were imputed using PHASE. LRH analysis was performed using Sweep. Each SNP defined two core alleles, one base pair in length. We calculated relative extended haplotype homozygosity (REHH) for each core allele, to its left and right [150], yielding up to four REHH scores per SNP locus. We standardized the REHH scores as a function of core allele frequency, defined on a discrete grid from 0.05 to 0.95 with even spaces of 0.025. This yielded a normally-distributed set of Z-scores for which we calculated corresponding P-values and Q-values.

2.4.7 Genome Wide Association Study (GWAS)

We performed a GWAS for drug resistance to thirteen antimalarials across 50 of our genotyped samples. 7,437 SNPs that had a minor allele count of five samples as well as an 80% call rate under every phenotype condition were used for GWAS. A Bonferroni significance threshold of $\log_{10}(\text{P-value}) > 5.17$ was used for all tests. See Appendix C for more details on GWAS methods.

The Efficient Mixed-Model Association (EMMA) test [79] models quantitative trait associations to a data set with complex population structure and hidden relatedness. It calculates a genotype similarity matrix instead of discrete categories and does not require *a priori* specification of populations. The resulting P-value distributions demonstrate little remaining effect from population structure (Figure C.8) while retaining power to find a number of associations at genome-wide significance (Figures C.8, 2.2A, Table 2.1).

The Haplotype Likelihood Ratio (HLR) test [90] models the likelihood that a single, resistant haplotype rose to dominance while all other haplotypes proportionally decreased. PLINK [142] is used to produce sliding window haplotypes across the genome and calculate haplotype frequencies for input to the HLR test. We produced input for all 2-, 4- and 6-marker windows. The LOD scores generated by the HLR test were converted to empirical pointwise P-values by performing approximately 370,000 permutations of the null model for each test condition, allowing us to calculate empirical P-values up to a significance of $10^{-5.6}$. We preserved population-specific phenotype frequencies by permuting only within each of three populations defined by our PCA analysis (Table C.1). Resulting P-value distributions fit expectations well for the vast majority of test conditions (Figures C.9, C.10) and the test demonstrates power to detect a number of loci at genome-wide significance (Figure 2.2A, Table 2.1).

2.4.8 Copy Number Variation (CNV)

Copy number was assessed by evaluating the hybridization intensity at the *PF10_0355* locus on the high-density SNP array (Appendix C). Follow-up analyses were done by quantitative real-time PCR (qPCR) of the *PF10_0355* locus using the Delta Delta Ct method [52] *PF10_0355* was compared to the reference locus *PF07_0076* and 3D7 was used as a reference strain. A summary of *PF10_0355* copy number for all parasite strains tested is provided in Table C.6. Select resistant strains that were found to have multiple copies of *PF10_0355* were further analyzed by quantitative Southern blotting and *PF10_0355* copy number was compared to the *dhps* gene from the 3D7 strain [167].

2.4.9 *PF10_0355* Overexpression

The full length ORF of *PF10_0355* was amplified from either the Dd2 (HFN sensitive) or SenP08.04 (HFN resistant) parasite isolate and cloned into the pBIC009 plasmid under the expression of the *Hsp86* promoter. Plasmid DNA was isolated, tranfected into the Dd2 parasite strain and stable transfectants were selected with 2.5nM WR99210 [53]. Parasites from two independent experiments for each vec-

tor type (Dd2+Dd2 and Dd2+SenP08.04) were isolated and successful transfection was confirmed by plasmid rescue as well as episome-specific PCR and sequencing. Additionally, a vector control strain was made by transfecting Dd2 parasites with the pBIC009 plasmid containing the firefly luciferase gene (EC 1.13.12.7).

2.5 Acknowledgements

We gratefully acknowledge B. Coleman, J. Dvorin, M.T. Duraisingh, U. Ribacke and C. Valim for help with overexpression vectors and useful discussions. T. Burke, N. Mahesh, G. Ramirez, and N. Senaratne provided technical help. Parasites lines or samples were provided by: J. Barnwell, A.P. Dash, C.E. Chitnis, K. Day, A. Djimde, C. Plowe, A.M. Katzin, D. Kyle, S. Thaithong, S.d.L. Moraes, J. Smith; and X. Su. Malaria Research and Reagent Resource Repository provided parasites deposited by: W. E. Collins, D.E. Kyle, L. H. Miller, D. Baruch, W. Trager, D. Walliker, U. Certa, R. Reber-Liske, T.E. Wellems, and Y. Wu (Appendix C).

This chapter was originally published as Park, et al., Sequence-based association and selection scans identiy drug resistance loci in the Plasmodium falciparum malaria parasite, PNAS (109), 2012 [129]. This work was also presented at the American Society of Tropical Medicine and Hygiene (Dec 2011, Philadelphia, PA). See page xv for details on author contributions.

doi://10.1073/pnas.1210585109

3

Sequence-based Association and Selection-association Scans

Through rapid genetic adaptation and natural selection, the *Plasmodium falciparum* parasite—the deadliest of those that cause malaria—is able to develop resistance to antimalarial drugs, thwarting present efforts to control it. Genome-wide association studies (GWAS) provide a critical hypothesis-generating tool for understanding how this occurs. However, in *P. falciparum*, the limited amount of linkage disequilibrium (LD) hinders the power of traditional, array-based GWAS. Here, we demonstrate the feasibility and power improvements gained by utilizing whole-genome sequencing for association studies. We analyze data from 45 Senegalese parasites and identify genetic changes associated with the parasites' *in vitro* response to twelve different antimalarials. To further increase statistical power, we adapt a common test for natural selection, XP-EHH, and utilize it to identify genomic regions

associated with resistance to drugs. Using this sequence-based approach and the combination of association and selection-based tests, we detect several loci associated with drug resistance. These include the previously known signals at *pfcrt*, *dhfr*, and *pfmdr1*, as well as many genes not previously implicated in drug resistance roles, including genes in the ubiquitination pathway. Based on the success of the analysis presented in this study, and on the demonstrated shortcomings of array-based approaches, we argue for a complete transition to sequence-based GWAS for small, low-LD genomes like that of *P. falciparum*.

3.1 Introduction

The malaria parasite *Plasmodium falciparum* imposes a tremendous disease burden on human societies and is responsible for 1.2 million deaths annually [108]. Current efforts to eradicate malaria depend on the continued success of antimalarial drugs [94]; however, the emergence of drug resistant parasites threatens to hamper global health efforts to control and eliminate the disease. Understanding the genetic basis of these adaptations will be necessary to maintain effective global health policies in the face of an ever-changing pathogen.

A key to elucidating the genetic basis of drug resistance is identifying the specific genes associated with the phenotype. In human studies of this kind, the genome-wide association study (GWAS) has overtaken the classic candidate gene approach, made affordable by the use of genotyping arrays (or SNP arrays) that measure only a subset of variants in the genome [5]. This optimization is only possible because of the extensive correlation between genetic markers (called "linkage disequilibrium" or LD) in the human genome, which allows the subset of SNPs on an array to act as proxies for other markers not present; this process is known as "tagging" [42].

In *P. falciparum*, however, array-based GWAS is severely limited by the relatively short extent of LD [107, 170, 176, 185]. Lacking that correlation between genetic markers, genotyping arrays usually cannot detect associations with untyped markers, effectively limiting inferences to markers actually present on the array; even the highest density *P. falciparum* array reported to date found that LD between adja-

cent markers on the array was too weak for tagging in African populations [170]. Consequently, current *P. falciparum* arrays cannot confidently capture all causal variants for important phenotypes.

The rapidly decreasing cost of whole-genome sequencing offers a promising solution. In principle, working with whole genome sequence allows one to directly assay all mutations segregating in the population, obviating the detection problems associated with short LD. Discovering mutations directly also avoids the ascertainment bias inherent to arrays—bias that is exacerbated when SNP discovery and genotyping are performed in different populations [2]. Additionally, the small size of the *P. falciparum* genome (23Mb, roughly the size of a human exome), makes it potentially a hundred fold cheaper than whole-genome sequencing in humans. As malaria sequencing projects become cost-competitive with genotyping arrays, whole-genome sequencing has the potential to become the most effective approach to performing association studies in malaria.

Here, we test the hypothesis that whole-genome sequencing will identify SNP associations not detected by classic array-based approaches. We apply this method to identify loci in the *P. falciparum* genome that are associated with antimalarial drug resistance and compare the approach to a standard array-based GWAS. We improve the statistical power of this analysis by adapting a commonly used selection test, the cross-population extended haplotype homozygosity (XP-EHH) test [151], and utilize it as an association test for positively selected phenotypes. These approaches identify a number of candidate loci associated with anti-malarial drug resistance, including genes in the ubiquitination pathway, suggesting that alteration of the parasite's ability to modulate stress may contribute to evasion of drug pressure and development of resistance in *P. falciparum*.

3.2 Results

3.2.1 45 Parasite Genomes and the Absence of LD

We chose a population in a West African region near Dakar, Senegal and culture-adapted 45 *Plasmodium falciparum* parasites recently isolated from malaria

infected patients. This population is particularly relevant for these studies as it has recently been exposed to multiple, changing drug regimens as clinical resistance to traditional drugs has emerged [104]. We obtained whole-genome sequence data and generated high-quality consensus base calls for an average of 83% of each genome. This produces 225,623 segregating single nucleotide polymorphisms (SNPs), of which 25,757 met our call rate and minor allele frequency criteria for further study (see Section 3.4). Sequence-based SNP calling in *P. falciparum* is technically challenging due to its extremely AT-rich genome [98, 127]. In light of this, we validated our sequence-based approach against array-based methods by using a previously described SNP array [170] to genotype 24 of the 45 isolates. Of the 74,656 SNPs assayed by the array, 4,653 meet our call rate and minor allele frequency criteria. We observe nearly perfect concordance between Affymetrix genotypes and sequence genotypes (see Section 3.4).

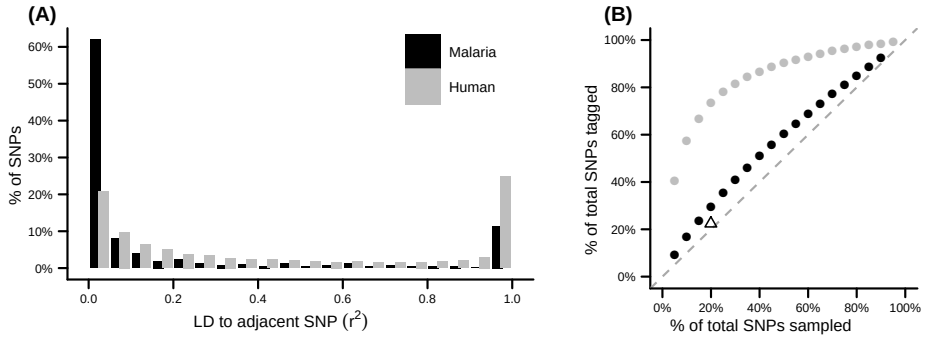


Figure 3.1: Simulated *P. falciparum* arrays are unable to tag SNPs not present on the array. **(A)** A histogram of LD between adjacent SNPs from sequenced *P. falciparum* (black). The vast majority of markers have little to no LD with their neighbors (62% of SNPs have $r^2 \leq 0.05$, 76% have $r^2 \leq 0.2$, and 87% have $r^2 \leq 0.8$). This contrasts with human studies where much more of the genome shows moderate to strong LD between neighboring SNPs (gray). **(B)** Simulated genotyping marker sets of various sizes are plotted against the percentage of the entire sequenced marker set that they are able to tag (with $r^2 \geq 0.8$). The dashed, identity line depicts the theoretical scenario where all SNPs are in complete linkage equilibrium and no SNP tags another. Since this is true of 87% of SNPs in the malaria sequence data, the increase is almost linear (black dots). This contrasts with the array tagging performance seen in human studies (gray dots), where only a small fraction of markers are needed to tag the bulk of the genome—a principle that array-based GWAS depends on. The open triangle depicts the actual performance of the Affymetrix-based Broad Institute *P. falciparum* SNP array [170].

Our data demonstrate that SNPs in *P. falciparum* have very little ability to tag neighboring SNPs due to the short LD in the African population from which they were sampled. While some portions of the genome exhibit significant LD, over 62% of the SNPs in the genome have no LD ($r^2 < 0.05$) between adjacent SNPs, and 87% of the SNPs have insufficient LD to tag their neighbor (Figure 3.1A) using the criterion derived from human GWAS ($r^2 < 0.8$) [42]. To measure tagging ability directly, we simulate genotyping arrays of various sizes by sampling random subsets of SNPs from our sequence data. We find that the simulated arrays are not able to tag a significant portion of unassayed markers, a result in stark contrast to the performance of human arrays (Figure 3.1B). The tagging performance of our own Affymetrix array (tagging only 22.6% of segregating SNPs in Senegal) is even lower than simulated arrays of similar size (Figure 3.1B), most likely due to population-based ascertainment biases [2] that were not modeled in our idealized approach. These findings lead us to conclude that array-based studies in *P. falciparum* will rarely be able to detect signals resulting from mutations not present on the array.

3.2.2 Sequence-based Genome-wide Association Studies

The goal of these studies is to identify genomic changes associated with changes in parasite response to antimalarial drugs, as measured in the set of 45 independent *P. falciparum* isolates. We assayed the cultured parasites for *in vitro* drug responses (measured by IC_{50}) to twelve standard antimalarials: amodiaquine, artemisinin, atovaquone, chloroquine, dihydroartemisnin, halofantrine, lumefantrine, mefloquine, piperaquine, primaquine, pyrimethamine, and quinine. These constitute the twelve phenotypes used in our association studies (Figure D.1). Not surprisingly, drugs with similar chemical structures (e.g. halofantrine, lumefantrine, and mefloquine) show a strong correlation in responses (Figure D.2), as has previously been observed [170, 185], and provide the opportunity for cross-validation of SNPs identified in association studies.

To test associations between SNP genotypes and drug response, we use efficient mixed-model association (EMMA). EMMA is a quantitative association ap-

proach well-suited for small sample sizes and partially inbred organisms, such as the malaria parasite [79]. It is a commonly used tool among mixed-model GWAS approaches [136] and has recently demonstrated effectiveness with *P. falciparum* drug studies [170]. After correcting for multiple testing (Bonferroni correction for 25,757 SNPs, $P < 2 \times 10^{-6}$), EMMA is able to detect a number of previously known markers of drug resistance, such as four non-synonymous SNPs in *pfcrt* [54, 184] associated with chloroquine response (N75E/K, K76T, Q271E, R371I), one *pfmdr1* SNP [49, 120] associated with halofantrine, lumefantrine, and mefloquine response (N86Y), and three *dhfr* SNPs [109] associated with pyrimethamine response (N51I, C59R, S108N). We note here that, although mitochondrial and apicoplast genomes were also sequenced, no significant associations were found and the known mitochondrial mutations associated with atovaquone resistance [48, 82] were fixed in all 45 individuals for the drug-sensitive alleles (*cytochrome b* 268Y, 133M, 280G). In all, EMMA detects 34 significant SNPs associated with parasite response to five drugs (Figure D.3). Most are in or near previously known associations [176], and five are novel associations with pyrimethamine response (Dataset 1).

While these sequence-based findings validate the previously known relationship between the *pfmdr1* gene and parasite responses to halofantrine, lumefantrine, and mefloquine, it is notable that this association is not detectable by our SNP array (Figure 3.2, Figure D.5), as the array lacks any markers in *pfmdr1* with a sufficiently high minor allele frequency. This exemplifies the type of association that can be missed by arrays due to limited LD. Additionally, the agreement between these three drugs at this locus provides validation of this result with respect to structurally related drugs.

3.2.3 Using Haplotype-based Selection Tests for Association

To test the hypothesis that drug resistance is largely driven by positive selection, we searched for long haplotypes associated with selection for drug resistance using the XP-EHH test [151]. This selection test has not previously been used as a GWAS tool, but it is well suited for this purpose when we presume that the phenotype

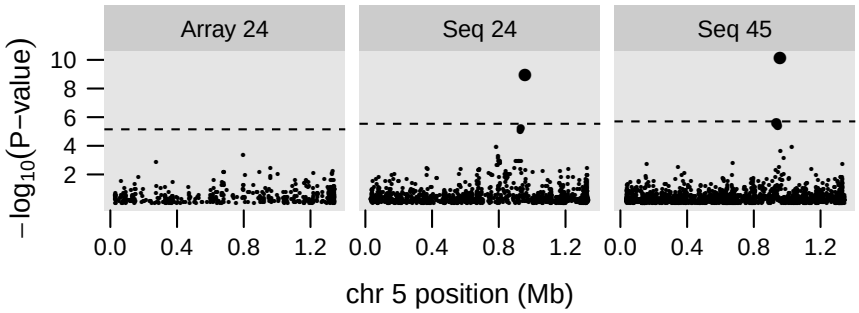


Figure 3.2: Mefloquine association signals around the known drug resistance locus *pfmdr1*. EMMA results are shown for all of chromosome 5 with P-values for each SNP on a $-\log_{10}$ scale against physical position. The array-based study (Array 24) does not detect any association at the known *pfmdr1* locus due to a lack of marker coverage within the gene and sufficient LD around the gene. The sequence-based study with the same 24 samples (Seq 24) detects the expected hit at 0.96Mb. Including all samples from the sequence-based study (Seq 45) increases the strength of this signal. The dashed line indicates the Bonferroni-corrected significance threshold ($P = 0.05$, genome-wide SNP counts are 7,068, 17,278, and 25,159 respectively).

we are studying is under positive selection. While this assumption is not valid for most human-based GWAS for non-communicable diseases, it is very likely to be the case when studying parasite genomes for resistance adaptations to widely used drugs, which represent a strong selective pressure. Used in this way, the XP-EHH test identifies areas in the genome where resistant parasites show much longer haplotypes than sensitive parasites, indicative of recent positive selection on the resistant population. In our data, the test detects a number of signals, including *pfcrt* and *dhfr*, as well as a number of other hits spanning a total of 32 genomic regions across eleven drugs (Figure 3.3, Figure D.4, Dataset 1). Seventeen of these regions are indicative of selection in the drug resistant population, whereas fifteen are consistent with selection in the drug sensitive population. With the exception of the regions containing *pfcrt* and *dhfr*, none of these loci were detected by EMMA alone.

While this approach does not detect the known *pfmdr1* locus, this is consistent with our expectations, due to the nature of the test. The N86Y mutation in *pfmdr1* confers increased susceptibility [49, 120] to many drugs when compared

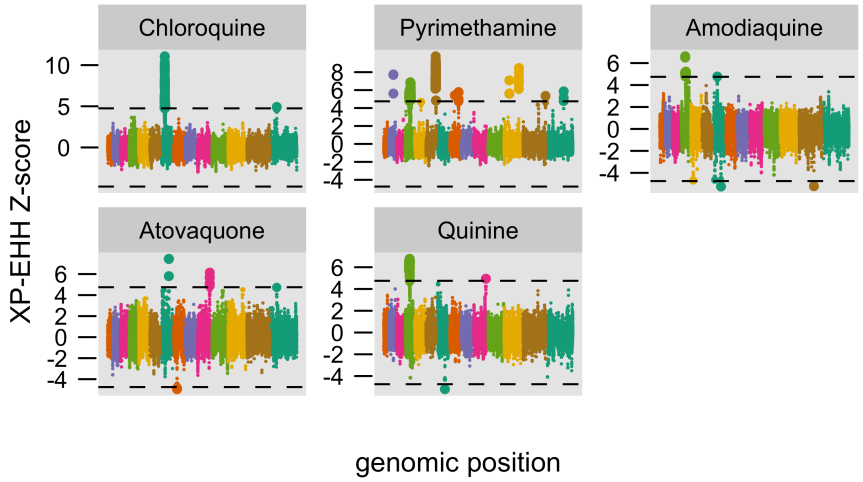


Figure 3.3: Significant signals of drug-associated selection across five antimalarial drugs. XP-EHH results are shown using a Manhattan-inspired plot, with SNP Z-scores plotted against genomic position, with each chromosome colored separately. Positive Z-scores suggest selection in drug resistant parasites, negative Z-scores suggest selection in sensitive parasites. The dashed lines indicate the two-sided Bonferroni significance thresholds ($P = 0.025$ and 0.975). Only drugs with significant hits are shown here, Z-score and quantile-quantile plots for all drugs are shown in Figure D.4.

to the wild-type allele. As such, this SNP would not be an expected candidate for positive natural selection on a novel variant—the type of selection XP-EHH is designed to detect. Moreover, the absence of a *pfmdr1* signal from the XP-EHH test is consistent with the lack of findings in this gene from previous genomic scans for positive selection based on the REHH, iHS, and XP-EHH tests in multiple populations [31, 107, 170].

In searching for long haplotypes, the XP-EHH test typically identifies a large number of significant SNPs in close proximity to each other. These regions often span many tens of kilobases and several annotated genes. This is expected because the process of positive natural selection increases the prevalence of both the selected variant as well as of nearby variants, generating local regions of extended haplotypes. Thus, while XP-EHH strongly implicates these 32 regions as areas of phenotype-associated positive selection, by itself it is usually unable to localize the source of this selection to a specific gene. We use *P*-values from EMMA to improve

signal localization by identifying the strongest signals of association within each region. This approach allows us to suggest a possible gene or mutation as a focus of phenotype-specific positive selection for each identified region (Dataset 1) and is reminiscent of earlier approaches that intersect selection and association results [31, 84].

A more comprehensive examination of the regions under drug-associated selection reveals discrete biological pathways and processes that may be particularly important as mediators of drug response in *P. falciparum* (Section D.3). The 59 genes in these 32 regions can be functionally classified as: surface molecules or transporters, genome maintenance or transcriptional regulation, metabolic enzymes including lipid metabolizers, and members of the ubiquitin proteasome system. Most surface molecule associated mutations and intergenic mutations are localized to intra-chromosomal clusters containing *var*, *rifin* and *stevor* genes; and a number of genes are found among molecules modulating ubiquitination, lipid metabolism, or folate metabolism. Members of these pathways are also represented in the large region of pyrimethamine-specific selection on chromosome 6, where it is difficult to localize the focus of selection. Collectively, these findings argue that certain biological processes in general, and genes in the ubiquitination and lipid metabolism pathways in particular, play important roles in modulating drug responses in *P. falciparum*.

3.3 Discussion

Complete genome sequencing provides many advantages over array-based genotyping for association studies. These include the ability to directly type the causal allele, the increased detection power from increased marker density, and the ability to overcome ascertainment biases that arise when studying different populations with a fixed marker set. In *P. falciparum*, the lack of tagging ability due to the near absence of long-range LD limits the utility of arrays for association studies. Furthermore, the small genome size of *P. falciparum* brings the cost of whole genome sequencing to approximate parity with traditional genotyping arrays, and recent advances in pathogen-specific DNA-enrichment and host-specific DNA-

depletion techniques for clinical samples makes the sequence-based GWAS approach more accessible and cost-effective than ever before [98, 171].

We introduce a selection-association approach based on the XP-EHH selection test. While this approach may not be appropriate for many association studies, it is sensible when the phenotype under study is under strong selection, which is likely the case for drug resistance in pathogens. As a haplotype-based test that takes advantage of multiple, adjacent SNPs, it has the advantage of being more sensitive than single-marker approaches like EMMA, given the same sample size [42]. In addition to detecting new signals of drug-associated selection, we also find that the directional nature of the test statistic, a Z-score, provides useful information about whether the selection is associated with drug sensitivity or resistance. Consequently, we also introduce an alternative visualization of the output: a Manhattan-like plot of Z-scores, instead of $-\log_{10} P$-values, to illustrate the directionality of the signals (Figure 3.3). In our data, we observed a tendency for many drugs (artemisnin, dihydroartemisnin, primaquine, halofantrine, lumefantrine, and mefloquine) to show highly significant signals of selection for drug sensitivity at *pfcrt*, the gene known to be responsible for chloroquine resistance (Figure D.4). While, in principle, this type of signal may result from selection towards drug sensitivity, in this particular case, it most likely results from the general pattern of anti-correlation between chloroquine and these six other drugs (Figure D.2). Additionally, the absence of a significant chloroquine sensitivity signal at *pfcrt* is consistent with reports that the return of chloroquine sensitive parasites in Africa did not result from a classic selective sweep [87]. In either case, the Manhattan-like Z-score plots allow us to note the presence of these drug sensitivity signals while keeping them visually separate from the drug resistance signals on which we wish to focus.

Our approaches identify a significant number of loci associated with changes in drug response (Dataset 1). The strongest of these contain previously known mediators of resistance, such as the mutations in *pfcrt*, *pfmdr1*, and *dhfr*. Curation of our remaining results using a variety of gene and protein prediction algorithms and literature searches [14] point to several cellular processes and pathways of potential interest, including the ubiquitin proteasome system, lipid metabolism, and folate

metabolism (Dataset 1). We argue that these findings point to biological processes used by the parasite to survive drug pressure or circumvent the action of anti-malarial compounds. Other genes of interest include three ABC transporters—a class of transporters known to modulate drug responses in other organisms [89]—and genes proposed to modulate chromatin [33, 37], DNA repair [25, 162], or RNA binding [99]—pathways that have been shown to potentially be altered in response to drug pressure.

A number of the signals of recent positive selection are unique to pyrimethamine-resistant parasites. While the known resistance locus, *dhfr*, is present among these, there are even stronger signals of pyrimethamine-associated selection on chromosome 6 and chromosome 12. The region on chromosome 6 contains two previously uncharacterized genes proposed to participate in folate metabolism (PFF1360w and PFF1490w), as well as six genes acting as either chaperones or in ubiquitination (PFF1365c, PFF1485w, PFF1445c; PFF1415c; and PFF1505w), and three molecules likely to modulate lipid metabolism (PFF1350c, PFF1375c-a/b, and PFF1420w). In the chromosome 12 region, the XP-EHH test produces significant *P*-values for eight SNPs over a 15kb region spanning five adjacent genes. The extended haplotypes surrounding these SNPs continue even further, spanning 28kb and fourteen genes in total (Figure 3.4A). These results present challenges for experimental validation, as the goal of association studies is to generate a small number of testable hypotheses about molecular mechanisms. Fortunately, the use of EMMA *P*-values in this region can assist in localizing the signal. We find that the strongest EMMA SNP coincides with the strongest XP-EHH SNP, which is a non-synonymous mutation in PFL2100w, a putative ubiquitin conjugating enzyme (E2) (Figure 3.4B). Additionally, a significant, pyrimethamine-specific selection signal on chromosome 8 is entirely contained within MAL8P1.23 (a putative HECT ubiquitin ligase E3) (Dataset 1), another gene in the ubiquitin-mediated pathway [133]. Given the role of this pathway in directing protein degradation and recycling, it is possible that alterations in these genes create changes in stress responses or protein turnover of key resistance modulators that allow the parasite to survive under drug pressure.

The evolution of drug resistance in the natural setting is likely to be a multistep

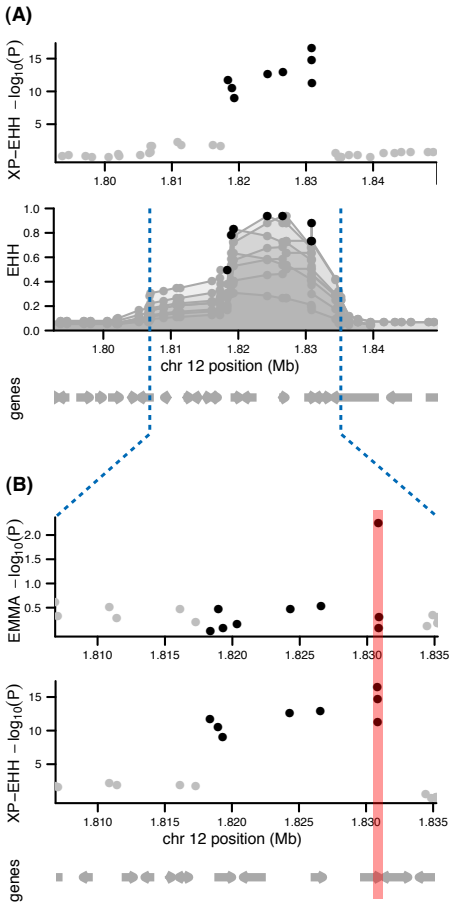


Figure 3.4: Localizing the pyrimethamine-associated selection signal on chromosome 12. **(A) Defining the region:** XP-EHH identifies eight genome-wide significant SNPs in close proximity on chromosome 12. Each of these eight SNPs represents the center of an area of extended haplotype homozygosity, as measured by the EHH statistic. Haplotype decay for resistant parasites is plotted for each of these eight SNPs, which defines a larger region from 1.807Mb to 1.835Mb in which the causal mutation may exist. This region spans 28kb and 14 genes. **(B) Localizing the signal:** focusing within this region, we utilize single-marker association signals from EMMA to localize the signal. The most significant EMMA SNP coincides with the most significant XP-EHH SNP and localizes to an E398D mutation in PFL2100w (ubiquitin conjugating enzyme E2).

process and our work potentially identifies key pathways involved in this process. Field-based evidence has demonstrated a reduced fitness for drug resistant parasites in the absence of drug pressure and laboratory-based work has demonstrated the relative fitness of different mutational changes in target enzymes. Our findings point to potential compensatory mutations in a pathway related to protein stability and turnover and it is tempting to speculate that such adaptations enable the "expression" of a resistant phenotype, such as has been observed in yeast [73]. Although molecular approaches are required to validate the role of this pathway in modulating drug response, these results demonstrate the potential for sequence-based GWAS approaches to identify pathways, in addition to individual genes, that may be responsible for the phenotype of interest.

Ultimately, all association results require experimental validation and follow-up work to explore possible mechanisms of action. Association studies, even in their ideal form, simply generate hypotheses based on correlations. However, improved methods for association studies can significantly reduce the necessary validation work by reducing false positive rates, increasing study detection power, and improving localization ability. This study successfully pilots the use of whole-genome sequence data for association studies in malaria and it demonstrates significant advantages in detection power over array-based studies. We strongly recommend that future association studies in low-LD, small-genome organisms adopt the sequence-based GWAS approach as well, given the relative costs. We additionally demonstrate the effectiveness of the XP-EHH selection test as an association test for phenotypes under positive selection. Finally, we combine data from both tests to localize long signals and reduce the number of hypotheses for follow-up validation. This combined approach identifies more candidate loci than with single-marker tests alone.

3.4 Methods

3.4.1 Sequencing

Parasites were obtained from patients with uncomplicated mild malaria in Senegal from 2001 to 2009 under ethical approval with informed consent for the study. Parasites were culture adapted by standard methods [165] and genomic DNA was extracted from 45 single-clone samples. Samples were determined to be monoclonal and genetically distinct by a 24 SNP molecular barcode [39]. Genomic DNA was sequenced using Illumina Hi-Seq machines. The first 12 parasites were sequenced with 76bp single-end reads and the remaining 33 were sequenced with paired-end reads ranging from 76bp to 101bp in length. The median sequence coverage depth was 144.8× after alignment (ranging from 32× to 400×).

Reads were aligned with BWA v0.5.9-r16 against the 3D7 reference assembly (PlasmoDB v7.1). A consensus sequence was called for each strain using the GATK Unified Genotyper v1.2.3-g61b89e2 [96] with the following parameters: `-A AlleleBalance -stand_emit_conf 0 --output_mode EMIT_ALL_SITES`. Bases were then removed if they exhibited poor quality (GQ less than 30 or QUAL less than 60) or if they called a heterozygous genotype. This left consensus calls for 56-91% of the genome (83% median) for each of 45 individuals. Of these sites, 225,623 positions are polymorphic among the 45 individuals. Of these SNPs, only 25,757 had genotypes in at least 36 individuals (80% call rate) and were non-singletons (i.e. minor allele count > 1 or minor allele frequency $> 4\%$). All analyses are based on this set of 25,757 SNPs. SNP data is available in dbSNP as batch Pf_0004 from submitter BROAD-GENOMEBIO. SNPs are being processed at PlasmoDB [14] for release later this year. SNP data and consensus calls for the whole genome are available as supplemental data files (Section D.2).

Principal component analysis was conducted using the program *SMARTPCA* [130] in the EIGENSOFT 3.0 package. We applied a local LD correction (nsnpldregress = 2) and found no significant eigenvectors in the population.

3.4.2 Tagging Analysis

Tagging analysis in Figure 3.1B was generated by using PLINK [142] to find tagging SNPs for each SNP that were within 10kb and at least $r^2 \geq 0.8$. We then simulated genotyping arrays by randomly sampling subsets of SNPs of varying subset sizes and calculating the fraction of total SNPs that are tagged by the subset. We first reduced the sequence data to 40 random individuals to simulate ascertainment bias against low allele-frequency markers, then randomly sampled markers that were still polymorphic among the smaller population size to simulate a genotyping array. We simulated 19 different array sizes, ranging from 5% of the sequenced SNPs (1,227) to 95% of the sequenced SNPs (22,087). 200 simulations per array size were run and the result was highly consistent: 95% confidence intervals were too small to visualize on the figure. Simulations for the human genome were based on 60 diploid individuals of European descent (CEU) from Hapmap release 23a. Each iteration chose 54 random individuals to simulate ascertainment bias, filtered SNPs to an 80% call rate and to non-singletons. Our Affymetrix array was able to tag 5,508 SNPs in our sequence data using the 4,894 SNPs on the array that overlapped with the 25,757 SNPs in our sequence data (open triangle in Figure 3.1B). Histograms in Figure 3.1A are binned into 20 evenly spaced bins of r^2 from 0 to 1. The plot is normalized such that the sum of all bars in each histogram is equal to 1 to show the relative proportions of SNPs in each bin. Simulation data is provided in supplemental data files (Section D.2).

3.4.3 Drug Assays

Drug assays were performed as described [132] with slight modifications for 384-well format (Section D.4). The range of drug concentrations are shown in Figure D.1, and the IC_{50} data, along with raw input data for all association tests, is provided in supplemental data files (Section D.2).

3.4.4 EMMA

Single marker association tests were run using EMMA [79]. Since not all drugs have complete phenotype data for all 45 individuals, SNPs are additionally filtered to those that met our previous call rate and minor allele criteria among the subset of samples for which drug data exists. This results in 23,000 to 25,180 SNPs for any given drug. $\log_{10}(\mathrm{IC}_{50})$ values were used for this quantitative test. Biological replicates of drug data were presented to EMMA as multiple individuals from the same genetic strain. This allows EMMA to use the additional data to discern heritable phenotypic variance from non-heritable variance [136], and mimics the use of clonally identical parasites in other studies [10, 11]. Significance was defined as SNPs that exceeded a Bonferroni-corrected threshold of $P < 0.05$ while also surviving 60% of jackknife simulations. EMMA results were jackknifed by performing 200 random subsets of 38 samples and requiring an FDR-corrected significance of $Q < 0.1$. SNPs that passed this threshold in 60% of jackknife simulations were considered to be robust against false positives due to small sample size effects.

3.4.5 XP-EHH

Selection-association tests were run using the cross population extended haplotype homozygosity test (XP-EHH) [151]. Each drug defined a partitioning of samples into two "subpopulations" ("sensitive" and "resistant") based on cutoffs shown in Figure D.1 and provided in supplemental data files (Section D.2). XP-EHH requires a recombination map as input, which we constructed with LDhat v2.1 [97] (Section D.4). XP-EHH also requires fully imputed genotypes. Imputation was performed using PHASE 2.1.1 [156], producing 29,605 non-singleton SNPs (Section D.4).

XP-EHH computes a significance value for each SNP in the genome, assuming that SNP comprises the haplotype "core" of selection. Because the test identifies long haplotypes, it results in a large number of genome-wide significant SNPs (defined by Bonferroni-corrected $P < 0.05$) in clustered stretches of the genome. We reduced the set of significant SNPs to a set of significant genomic regions by tak-

ing each significant core SNP, computing a window around each one where EHH decayed to 0.05, and merging overlapping windows. This resulted in a smaller list of significant regions for each drug (Dataset 1). Regions were further filtered by removing those which did not contain at least one core SNP that survived 50% of jackknife simulations. XP-EHH results were jackknifed by performing 200 random subsets of 38 samples and requiring a Bonferroni-corrected significance of $P < 0.1$.

3.4.6 Genotyping Arrays

A subset of 25 parasites was also hybridized to an Affymetrix array containing 74,656 markers [170]. SNPs were called using BRLMM-P from Affy Power Tools v1.10.2 and filtered according to the same methods as Van Tyne et al. [170], resulting in 15,075 validated SNPs, 8,778 of which were polymorphic among the 25 individuals from Senegal. SNP coordinates were converted from PlasmoDB v5.0 coordinates to v7.1 coordinates using whole genome nucmer alignments [85]. Concordance between array and sequencing data was measured for the set of markers in which genotype calls existed by both methods. For 24 samples, nearly perfect concordance between Affymetrix genotypes and sequence genotypes was observed for the 24 samples (averaging 99.2% concordance, with all 24 samples above 98.2% concordance). This level of concordance is similar to what is observed with technical replicate hybridizations of the same DNA sample [170]. One sample, SenP19.04.c, reported a 28.2% mismatch rate, suggestive of a sample identification error, and was removed from the analysis. EMMA analyses were run on the array data using the same filters and procedures as for sequence data described above, utilizing 4,514-4,653 SNPs per drug phenotype. Results are shown in Fig D.5. Array data for these 24 samples are in supplemental data files (Section D.2).

3.5 Acknowledgements

We thank the sample collection team in Senegal, including Younouss Diedhiou, Lamine Ndiaye, Amadou Moctar Mbaye, Baba Dieye, Moussa Dieng Sarr, Papa Diogoye Sene, and Ngayo Sy. We thank the technical staff at HSPH who maintained parasite cultures, including Kayla Barnes, Dave Rosen, Kate Fernandez, and Gilberto Ramirez. We thank members of the Sabeti lab for a careful review of our manuscript, including Kristian Andersen, Chris Edwards, Chris Matranga, Rachel Sealfon, Jesse Shapiro, Ilya Shlyakhter, Matt Stremlau, and Shervin Tabrizi.

We acknowledge contributions made to the community database, PlasmoDB.org, that facilitated biological curation of candidate genes presented in this work.

This study is supported by the Bill and Melinda Gates Foundation, National Institutes of Health (Grant: 1R01AI075080-01A1), Ellison Medical Foundation, ExxonMobil Foundation, NIH Fogarty, NIAID, and Broad SPARC. DJP is supported by an NSF Graduate Research Fellowship. PCS is supported by fellowships from the Burroughs Wellcome and Packard Foundations.

This chapter represents work not yet published. See page xv for details on author contributions.

4

Temporal Signatures of Selection

The final chapter of this dissertation explores an alternate approach at studying selection in the malaria parasite by examining temporal signatures of selection in-progress. Here, we take advantage of a fully sequenced set of 159 Senegalese samples spread out in time over a dozen generations. Recently developed Hidden Markov-based methods allow us to estimate the parameters for drift (N_e) and selection (s) in this type of time-series data. In particular, the estimation of s for every SNP in the genome allows us to conduct highly specific scans for markers with extremely rapid changes in allele frequency. These markers are candidates for very strong selection in an environment where malaria is currently subject to a strong eradication program. Finally, this new approach provides a complementary view into the selective environment of the parasite when combined with traditional tests for selection.

4.1 Introduction

The study of natural selection has been a significant focus of methods development in population genetics. These efforts have produced numerous statistical tests for identifying selection based on genetic data [149]. Most of these tests—such as long-haplotype tests or population differentiation tests—operate on data sampled from a present-day snapshot of the population and make inferences about historic selective events based on the telltale remnants in the genome that are produced by historic natural selection.

Many large parasite sample sets that are currently being produced are spread out in time over several years. As the sexual generation time of *P. falciparum* is thought to be on the order of three generations per year, these data sets can easily span the human-equivalent of several centuries in time. This breaks the contemporaneous sampling assumption made by most selection tests, but more importantly, it misses an opportunity to exploit time course data as evidence for selection. In this study, we seek to use this type of data as a direct observation of the effects of selection in action.

Studying selection using time series genetic data has gained recent attention. Based on previous methods for inferring effective population size (N_e) from time series data, Bollback et al. pioneered the use of a Hidden Markov Model (HMM) to jointly estimate both N_e and the selection coefficient (s), at a single locus [19]. Intuitively, the model treats the observed allele frequencies at each time point as noisy estimates of a hidden state, where the emission probabilities (measurement errors) and transition probabilities (changes in frequency over time due to drift and selection) of the HMM are well established probability distributions in population genetics.

Based on Bollback et al., a number of similar methods to estimate selection from time series data have emerged recently [70–72, 93]. Some of these are specific to asexual microbes, where the problem of clonal interference is a significant confounder. But these approaches have largely remained tests of a single marker or a small number of markers. Scaling this test to whole-genome data requires both speed optimizations and modified statistics. In this chapter, we adapt the original

Bollback et al. method into a genome-wide test for selection, create new statistics based on it, evaluate its behavior under simulated neutral conditions, and apply it to a set of 159 parasites from a population where the introduction of artemisinin drug pressure is relatively recent.

4.2 Results

We based our analyses on 139 sequenced samples collected from patients in Senegal from 2008 to 2011 (19 samples from 2008, 52 from 2009, 46 from 2010, and 22 from 2011). From these genomes, we identified a set of 27,255 variants (SNPs and small indels) that met a number of filtering criteria (Section 4.4.1). At each of these variants, we estimated the selection coefficient, s, using code based on the Bollback et al. algorithm (see Section 4.4.2).

We fixed the drift parameter, N_e, to 100, based on recent studies of drift in the same population [40]. This represents extremely strong drift and, as a consequence, suggests that we will only be able to detect the strongest signals of selection in this population. This should also result in a more conservative test, since the model will attribute more changes to the effects of strong drift.

The distribution of estimated selection coefficients across the genome is shown in Figure 4.1A. This statistic is unimodal and centered at zero, though it does not closely follow any standard distribution (Figure 4.1B). The distribution is asymmetric with a large left tail, suggesting thousands of candidate variants for strong negative selection. In order to compute significance of this statistic empirically, we modeled the behavior of s under the null hypothesis by producing forward-simulated time course data (drift only, no selection) and ran the HMM 27 million times (methods in Section 4.4). The resulting distribution is similarly left-tailed (Figure 4.1C). This suggests that the many variants under strong negative selection are artifacts of the HMM-based estimation (further analysis of this issue in Section E.3). Indeed, if you calculate empirical significance values based on the simulated null data, the resulting Z-scores are normally distributed with no significant outliers (Figure 4.1D). This suggests that our data set is not sufficiently powered to reject the null hypothesis.

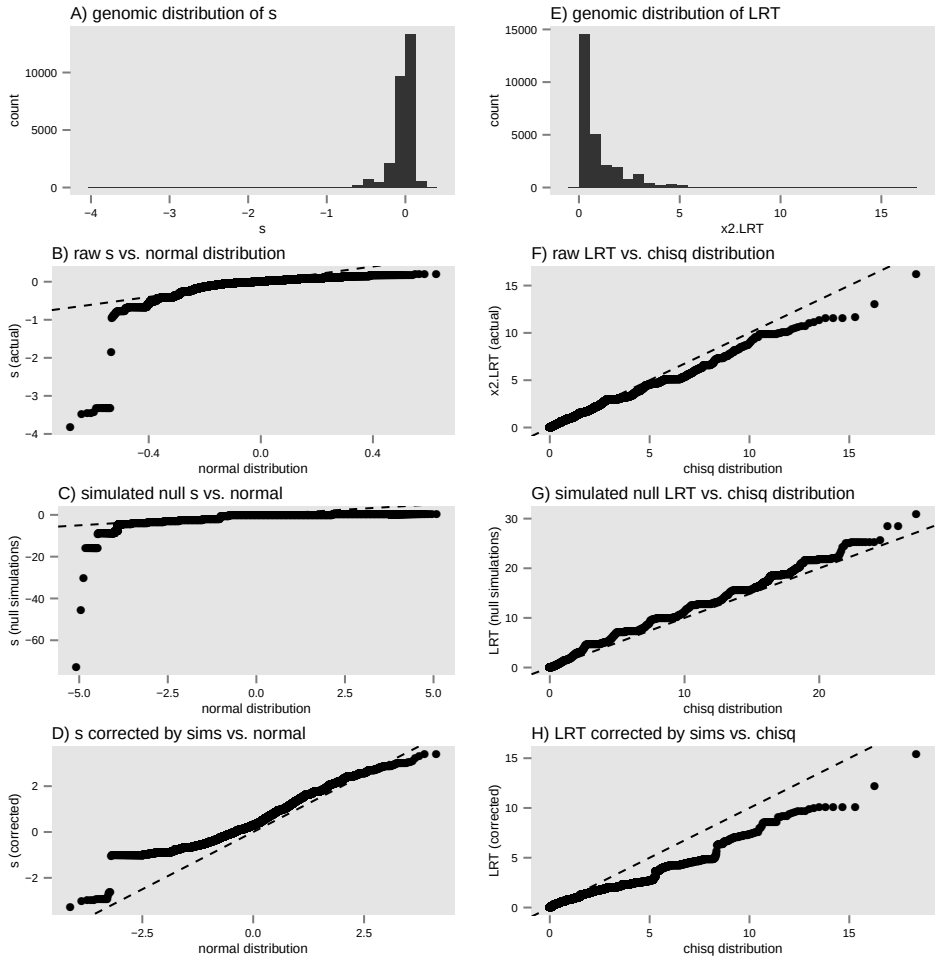


Figure 4.1: The genomic distributions of the two test statistics, s, at left, and the likelihood-ratio test (LRT), at right, do not allow us to reject the null hypothesis of no selection. The population size was fixed at $N_e = 100$. **(A)** Distribution of s (calculated by HMM) over all genomic markers. **(B)** QQ-plot showing the distribution of observed values of s against a normal distribution. It shows a large left-tail of markers reporting strong negative selection. **(C)** Distribution of s from null simulations. This displays a very similar distribution to the observed values in panel (B). **(D)** Corrected s based on the empirical distribution in (C). Observed values of s do not significantly depart from the simulated null distribution. **(E)** Distribution of the likelihood ratio test (LRT), which tests against the null hypothesis ($s = 0$), over all genomic markers. **(F)** QQ-plot showing the distribution of observed values of LRT against the expected χ^2 distribution. This matches the neutral expectation very closely. **(G)** Distribution of LRT from null simulations. This demonstrates that the neutral behavior of this test statistic is χ^2 distributed, as expected. **(H)** Corrected LRT based on the empirical distribution in (G). Observed values of the LRT do not significantly depart from the simulated null distribution.

We additionally evaluated an alternate test statistic based on the likelihood ratio test (LRT), a standard approach in statistical inference [182]. While estimating the maximum-likelihood value for s, the [19] HMM additionally emits the log-likelihood surface for all evaluated values of s. Using this, we can compute a test statistic that should be χ^2 distributed (with one degree of freedom) under the null hypothesis of no selection (Figure 4.1E):

$$\text{LRT} = 2\left[\log \mathcal{L}(s = \hat{s}_{MLE}) - \log \mathcal{L}(s = 0)\right]$$

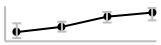
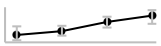
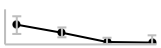
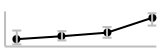
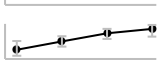
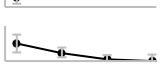
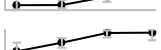
We find that the resulting distribution of LRT statistics across the genome matches the expected χ^2 null quite well (Figure 4.1F). As before, we also computed an empirical distribution of the LRT using 27 million iterations of our simulations of the null hypothesis and found an identical χ^2 distribution (Figure 4.1GH).

This test provides a better measure of statistical significance than s since its distribution is well characterized. However, our data set remains underpowered after correction for multiple-testing. It is likely that the overwhelming strength of genetic drift in this population will mask the signal of all but the strongest selective events, and that larger sample sizes and longer time spans of data would be required to confidently detect weaker selection.

However, in an absolute sense, most variants in the genome exhibit what is traditionally considered to be strong selection (24,046 out of 27,255 variants show $|N_e s| > 1$, that is, the product of effective population size and the selection coefficient are above one). Although we cannot confidently rule out neutrality for these markers, we can look at the tails of the distribution to understand the behavior of our method and the types of signals it is more sensitive to. Table 4.1 highlights the top candidates for strong selection.

Additionally, long-haplotype selection tests and GWAS tests were performed on the same data and are shown in Sections E.4 and E.5 but are not the focus of this chapter and are not elaborated on here.

Table 4.1: Top candidates for strong selection. This shows all variants with $P < 10^{-2.5}$ and at least one non-zero DAF value in 2010-2011 (rationale in Section E.3). P-values are computed by the likelihood ratio test (LRT). DAF estimates (black line) are drawn for each of four years (2008-2011) along with 95% confidence intervals for binomial sampling error. All thirteen hits lie in coding regions.

chr	pos	DAF by year	s	P	effect	gene	description
1	534,946		0.160	2.2×10^{-3}	non-syn	PF3D7_0113800	DBL containing protein
2	843,096		0.151	2.6×10^{-3}	syn	PF3D7_0220800	cytoadherence linked asexual protein 2 (CLAG2)
4	990,781		-0.213	3.0×10^{-3}	non-syn	PF3D7_0421700	unknown
5	249,925		0.160	1.5×10^{-3}	syn	PF3D7_0505700	unknown membrane protein
7	167,917		-0.380	1.1×10^{-3}	non-syn	PF3D7_0704000	unknown membrane protein
7	419,410		-0.239	1.4×10^{-3}	syn	PF3D7_0709300	Cg2
7	419,413		-0.241	1.2×10^{-3}	syn	PF3D7_0709300	Cg2
7	463,705		0.206	6.8×10^{-4}	non-syn	PF3D7_0710200	unknown
8	1,031,401		0.170	2.0×10^{-3}	non-syn	PF3D7_0823300	histone acetyltransferase GCN5 (GCN5)
10	512,935		-0.324	6.7×10^{-4}	non-syn	PF3D7_1013200	unknown
10	1,404,648		0.193	3.0×10^{-4}	non-syn	PF3D7_1035400	merozoite surface protein 3 (MSP3)
13	402,234		0.206	1.5×10^{-3}	codon insertion	PF3D7_1308700	unknown
14	1,852,794		0.171	1.5×10^{-3}	non-syn	PF3D7_1445100	histidine—tRNA ligase

4.3 Discussion

4.3.1 Candidates of Strong Selection

The candidates described in Table 4.1 are indicative of atypically strong changes in allele frequency over a relatively short time frame. Taken together, they describe the types of genetic regions that our approach is most likely to detect selection in. It is notable that all thirteen of these hits lie in coding regions of the genome (the genome is roughly 50% genic). Roughly half of the hits lie in proteins of unknown function—this is consistent with the proportion of unknown genes in the genome as whole.

It is also notable that five of these thirteen hits are surface proteins (DBL, CLAG2, MSP3, two unknown membrane proteins). These loci are typically thought to be under balancing selection due to their exposure to the host immune system (Section 1.4). Indeed, a number of these genes have been identified in tests for balancing selection previously [7]. However, balancing selection can behave like directional selection when observed over a relatively short time span. Balancing selection on a gene as a whole may also look like directional selection at individual SNPs within that gene. As such, though the HMM explicitly models directional selection, it is sensible for our method to detect signals of balancing selection as well. In our West African population, where many infections can be asymptomatic, one can speculate that balancing selection due to host immune pressure plays a stronger role than drug pressure in the short term: all parasites must encounter the host immune system during its life cycle, but not all will encounter drug treatment.

None of the major loci associated with chloroquine, pyrimethamine, or aminoquinolone resistance are observed in these results, but intriguingly, three of our hits occur within close proximity to *pfcrt*. Although these positions have previously been shown to reside within a long haplotype around this gene, the movements in our three markers appear to be entirely decoupled from the comparatively static allele frequencies in the central chloroquine resistance mutations of *pfcrt* (Figure E.2). This may simply represent the gradual breakdown of LD over

time and the reintroduction of favorable genetic diversity previously suppressed by the selective sweep 50 years ago.

The absence of well-characterized drug resistance loci in our findings suggests that this test is not sensitive to historic strong selective events in the way long-haplotype and other common selection tests are. Based on how selection is modeled, one would expect that even recent historic events would go undetected and that only selection currently in progress would produce a noticable signal. For this reason, this approach provides a very complementary set of insights to existing selection tests regarding the adaptive behavior of the parasite.

4.3.2 Future Directions

In this chapter, we have demonstrated the concept of scaling a temporal test for selection up to a genome-wide scale. We have established a number of useful genomic statistics based on this method originally developed by Bollback et al. in a single-marker setting. We have explored the behavior of this type of approach and the types of selective signals to which it is sensitive. And as an initial foray into a new type of selection test, we have also learned a number of ways in which the methods can be expanded upon in future work. These future directions are elaborated upon here.

The drift parameter, N_e, is currently set to a fixed value based on previous studies in this population [40]. Although we did manually examine the behavior of this statistic under several other values of this parameter (Figure E.4), this approach would benefit if it were combined with methods that can estimate N_e genome-wide. The existing HMM is able to provide a maximum-likelihood estimate of N_e for a single marker, but it is not able to jointly estimate the parameter across many markers. Additionally, recent evidence shows that effective population size may be changing over time as malaria control efforts take effect [40]. It would be ideal to be able to model a changing N_e over time, though this would require some significant changes to the underlying software.

Similarly, our model currently assumes a constant directional selection pressure over the entire time period sampled. When studying samples from a country

where the national drug treatment policies change over time (such as the introduction of a new standard of therapy in a certain year), this assumption does not hold. For these types of studies, it would not be hard to model a stepwise change in s at a given time point by splitting the data into two separate time intervals and running the HMM to estimate s separately for each time interval. A likelihood ratio test statistic can then be produced by testing against the null hypothesis of $s_{T1} = s_{T2}$. This would allow us to isolate regions of the genome where the selection coefficient significantly changes at a given point in time.

We have previously described how this test provides a complementary view of selection to existing long-haplotype-based selection tests. The natural next step is to compute composite statistics based on this and other tests to leverage the distinct information provided by each. This can be done using a thoroughly modeled approach as in Grossman et al. [63], or a more naive approach that simply combines significance values from each test [181, 186].

Very recent work shows that the complex life cycle of the malaria parasite (with multiple bottlenecks in the asexual and sexual stages) produces some important departures from the standard coalescent and Wright-Fisher models that our method is based on [27]. This is particularly true with regards to the strength of drift and selection—the very forces that this model tries to estimate. This complexity has often been overlooked in the past, but it is becoming clear that, particularly in studies of selection, it can no longer be ignored. It would be prudent to replace our forward simulations with the Chang et al. model while updating the transition probabilities in the Bollback et al. HMM with one that more closely matches those produced by the malaria life cycle.

Ideally, an expanded study would run forward simulations at a range of selection coefficients (here, we only simulated the null: $s = 0$), drift rates, sample sizes, number of time points, and total time spanned. This would produce a more formal calculation of the test's power and sensitivity to detect a certain level of selection and evaluate possible study designs before collecting the data.

Going forward, Dan Neafsey at the Broad Institute has proposed future work to utilize the methods developed in this chapter for a large scale study in Southeast

Asia—a population where clinical phenotypes appear to be in motion [47, 124]. The study would use the hybrid-selection approach developed at the Broad [98] to sequence 600 samples collected over 10 years from northwest Thailand in collaboration with François Nosten and Tim Anderson. The techniques developed here in this chapter can help inform the design of this study and provide the software framework on which to build that analysis.

The loci identified in Table 4.1 are biologically interesting and display some of the most significant movement of allele frequencies in our short time frame. Because this is based on genome-wide data from only a handful of time points, Diana Miao, a Harvard undergraduate, is presently validating these loci in the Wirth Lab at the Harvard School of Public Health. Using single-marker PCR assays, she is genotyping these top markers across a set of hundreds of unsequenced parasite samples from the same Senegalese population, across roughly a decade of time points. This will help determine whether these loci continue to show directional movement over a longer period of time, or whether they exhibit more drift-like behavior in the unsequenced years.

4.3.3 Conclusion

In the end, our time-series based approach is important for studying selection in malaria because of the very recent introduction of artemsinin-based selective pressures. From a disease control standpoint, it is also essential to detect the parasite's adaptations as close to real-time as possible, before resistance mutations become widespread. This necessitates using and combining as many distinct types of selection tests as possible to provide a clearer picture of the parasite's evolutionary responses. The approach piloted in this chapter provides a significant step in that direction.

This work, taken together with the chapters before it, make significant contributions to the field of malaria genomics. Each chapter of this dissertation demonstrates a new application of population genomic methods to the study of malaria, often adapting these methods in ways not originally foreseen. Each chapter also produces a number of candidate genes and mutations that lend insight to parasite

evolution, drug mechanisms, and point to possible genomic surveillance markers. This dissertation provides timely advances to both methods and knowledge for a field in need of innovative means to thwart the highly adaptable parasite responsible for one of the most prevalent human diseases.

4.4 Methods

4.4.1 Sample Collection and Sequence Analysis

190 parasite samples were collected from patients in Senegal over a ten year span from 2001 to 2011. Parasite DNA was isolated either by culture adaptation [165], hybrid-selection enrichment [98], or direct sequencing of patient DNA, depending on sample quality. Genomic DNA was sequenced at the Broad Institute using Illumina Hi-Seq machines using paired-end reads ranging from 76bp to 101bp in length.

Reads were aligned with BWA v0.6.2 against the 3D7 reference assembly (PlasmoDB v9.0). A consensus sequence was called for each sample individually using the GATK Unified Genotyper 2.4-9 [96] using the `EMIT_ALL_SITES` option. Both SNPs and indels were allowed. Bases were then removed if they exhibited poor quality (QUAL less than 60) or if they called a heterozygous genotype. After this filtering, all samples were merged into a single VCF file.

21 samples were then discarded that had genotype consensus calls for less than 50% of the genome (roughly 11.7Mb). After removing sites that were monomorphic in the remaining 159 samples, we had 631,032 variants. Of these, 62,599 were biallelic and non-singleton (at least two samples with a minor allele). Functional consequences of these variants were annotated using snpEff and PlasmoDB 9.0.

For this analysis, we further restricted our sample set to the 139 samples collected between 2008 and 2011, as these were the only four years with large enough sample sizes to reduce the errors in our estimate of annual allele frequencies. We reduced our marker set to those that had non-missing genotype calls for at least 12 samples in each of the four years, had a minimum minor allele frequency of 0.02

across all samples, had at least one year with a minor allele frequency of at least 0.05, and had a non-missing genotype call for the *P. reichenowi* genotype to infer ancestral/derived status. This resulted in 27,255 variants, of which 25,807 were SNPs and 1,448 of which were small indels.

In most selection studies (particularly with long-haplotype selection tests), SNP and indel data cannot be jointly analyzed due to the differences in mutation rates and, with some types of indels, recombination mechanisms. However, our approach measures selection against genetic drift (instead of recombination) which should affet both classes of variants equally. As such, we analyzed the data set without separating SNPs and indels while recognizing that our ability to ascertain indel variation with these methods (short reads, BWA, GATK) is restricted to only smaller and simpler indels.

The derived-vs-ancestral status of each variant was determined from genomic sequence from *P. reichenowi*, the closest known outgroup for *P. falciparum*. Raw reads from the "Dennis" isolate sequenced by the Wellcome Trust Sanger Institute were downloaded from the European Nucleotide Archive using project accession ERP000299 (see http://www.ebi.ac.uk/ena/data/view/ERP000299 and http://www.sanger.ac.uk/resources/downloads/protozoa/plasmodium-reichenowi.html). Reads were aligned to 3D7 with BWA and variants were genotyped with GATK's Unified Genotyper in the same manner as the *P. falciparum* samples described earlier. The resulting derived allele frequency distributions are shown in Figure E.1.

4.4.2 Estimating the Selection Coefficient

To estimate selection coefficients from our time series data, we started from the HMM implemented in Bollback et al. publicly available at http://www.simmap.com/bollback/software.html. This package, called sel2ns, models a single marker when provided with derived and ancestal allele counts for each time point. The user must specify a range of parameter space to explore (for N_e and s) and a resolution to explore it at (the number of intervals for each parameter) and the program will emit log-likelihoods for each combination of

parameter values in a linearly spaced grid within the specified boundaries. The output can then be examined to determine which parameter values result in the maximum likelihood.

This can result in many CPU-days of runtime when applied to a whole genome, but significant optimizations can be made to the likelihood maximization approach. As this software was originally intended for single locus studies, such optimizations were not a priority for the original authors. We wrapped the sel2ns software with our own script that uses well-established algorithms to optimize relatively well-behaved objective functions using an implementation of Brent's algorithm in the SciPy package (`scipy.optimize.minimize_scalar`). This approach efficiently finds the maximum likelihood estimate (MLE) of the parameters with as few executions of the HMM as possible.

We used a fixed value of N_e across the whole genome and, at each variant, estimated values for s and also computed a likelihood ratio test to produce a measure of significance against the null hypothesis of no selection. Effective population size is a concept used for multiple purposes in population genetics. Because N_e is used here to simulate the strength of drift over a short span of time, it makes sense to use a fixed value over the entire genome. Other interpretations of N_e that allow for varying values across the genome are used to describe how allele frequency spectra are affected by recent population history. However, since this is not how N_e is used by the HMM, it is sensible to fix a single value, since the entire genome will experience drift at the same rate within this period of time.

Forward simulations of the null model were produced with a simple implementation of the Wright-Fisher drift model using an initial allele frequency sampled randomly from the distribution of derived allele frequencies observed in the whole genome. Similar to the HMM, we introduced binomial sampling errors at each sampling time point, using annual sample sizes equal to those used in our actual data. The main difference between this model and the one described in Bollback et al. is the use of the Wright-Fisher model for drift here vs. the Kolmogorov backward equation used in sel2ns, but this difference should be negligible. These simulations are meant to model the behavior of the sel2ns output under the null hypothesis. The simulations can easily be adapted to incorporate directional selec-

tion by introducing a modification of the allele frequency at each Wright-Fisher generation prior to binomial sampling with the following additive model:

$$p_{\text{next}} = \frac{p(2 + s + ps)}{2 + 2ps}$$

Such a modification would allow for an examination of the sensitivity of sel2ns to detect selection at various strengths.

4.4.3 Orthogonal Tests for Selection

Additional analyses were also performed on this data set that are shown in Appendix E. These include traditional long-haplotype based tests for positive selection (Section E.4) and drug resistance GWAS for a subset of phenotyped parasites (Section E.5).

Long haplotype tests focused on sequence data from our Senegalese population combined with two other West African parasite populations sequenced by Manske et al. [95]: Burkina Faso and Mali. Raw reads from these samples were obtained from the European Nucleotide Archive using the accessions listed in Manske et al.. Alignment and variant calling was performed as described previously with our Senegalese and the *P. reichenowi* sample.

GWAS tests were performed exactly as in Chapter 3, with the exception that parasite drug response is now assesed using an *ex vivo* drug assay (manuscript in preparation) instead of the previous method relying on culture adaptation.

4.5 Acknowledgements

Daria Van Tyne (HSPH) has been involved with project concept and strategy. Hsiao-Han Chang (OEB), Clarissa Valim (HSPH), Hilary Finucane (MIT), Stephen Schaffner (Broad), and Dan Neafsey (Broad) have provided very valuable insight and guidance into the statistics and models. Sample management was handled by Sarah Volkman and colleagues in the Wirth Lab at HSPH. My thesis committee, comprised of Pardis Sabeti, Dan Hartl, John Wakeley, and Dyann

Wirth helped guide this project throughout.

This chapter describes thirteen recent manuscripts in which I have played an assistive role. Nine of these have reached publication, four more are in process.

A

Secondary Publications

During my time here, I also assisted with a number of publications led by other researchers. The following manuscripts are interior-author publications of mine from Fall 2010 to the present. Omitted from this listing is the Nature Reviews Genetics paper that forms the basis of Chapter 1.

A.1 Published manuscripts, Sept 2010 to July 2013

[40] Rachel F Daniels, Hsiao-Han Chang, Papa Diogoye Séne, Daniel J Park, Daniel E Neafsey, Stephen F Schaffner, Elizabeth J Hamilton, Amanda K Lukens, Daria Van Tyne, Souleymane Mboup, Pardis C Sabeti, Daouda Ndiaye, Dyann F Wirth, Daniel L Hartl, and Sarah K Volkman. Genetic Surveillance Detects Both Clonal and Epidemic Transmission of Malaria following Enhanced Intervention in Senegal. *PLoS ONE*, 8(4):e60780–e60780, April 2013. doi:

10.1371/journal.pone.0060780. URL http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0060780—This paper describes the use of a genetic barcode on a parasite sample set in Senegal spanning five years. It finds an increase in parasite clonality over time, consistent with a reduction in effective population size during a period of intense malaria control efforts. This effort was led by Harvard graduate students Rachel Daniels (BBS) and Hsiao-Han Chang (OEB). I played a minor role, providing some analysis for array-based validations and was involved in some discussions around data analysis.

[63] Sharon R Grossman, Kristian G Andersen, Ilya Shlyakhter, Shervin Tabrizi, Sarah Winnicki, Angela Yen, Daniel J Park, Dustin Griesemer, Elinor K Karlsson, Sunny H Wong, Moran Cabili, Richard A Adegbola, Rameshwar N K Bamezai, Adrian V S Hill, Fredrik O Vannberg, John L Rinn, Eric S Lander, Stephen F Schaffner, and Pardis C Sabeti. Identifying Recent Adaptations in Large-Scale Genomic Data. *Cell*, 152(4):703–713, February 2013. doi: 10.1016/j.cell.2013.01.035. URL http://dx.doi.org/10.1016/j.cell.2013.01.035—This paper describes a genome-wide scan of selection in human populations using composite statistics. I played a small role in this analysis, utilizing publicly available RNA-seq data to infer transcriptional activity of non-coding regions identified by this selection scan.

[28] Hsiao-Han Chang, Daniel J Park, Kevin J Galinsky, Stephen F Schaffner, Daouda Ndiaye, Omar Ndir, Soulyemane Mboup, Roger C Wiegand, Sarah K Volkman, Pardis C Sabeti, Dyann F Wirth, Daniel E Neafsey, and Daniel L Hartl. Genomic sequencing of Plasmodium falciparum malaria parasites from Senegal reveals the demographic history of the population. *Molecular Biology and Evolution*, 29:3427–3439, June 2012. doi: 10.1093/molbev/mss161. URL http://mbe.oxfordjournals.org/content/29/11/3427—This paper describes population genetic analyses and selection scans in 25 fully-sequenced Senegal parasites. They are a subset of the 45 parasites described in Chapter 3. This effort was led by Harvard OEB graduate student Hsiao-Han Chang. I provided analyses based on the iHS long-haplotype selection test and helped to analyze sequence data.

[6] Alfred Amambua-Ngwa, Daniel J Park, Sarah K Volkman, Kayla G Barnes, Amy Bei, Amanda K Lukens, Papa Sene, Daria Van Tyne, Daouda Ndiaye, Dyann F Wirth, David J Conway, Daniel E Neafsey, and Stephen F Schaffner. SNP genotyping identifies new signatures of selection in a deep sample of West African P. falciparum malaria parasites. *Molecular Biology and Evolution*, 29:3249–3253, June 2012. doi: 10.1093/molbev/mss151. URL http://mbe.oxfordjournals.org/content/29/11/3249—This paper describes population genetic analyses and selection scans in 75 parasites from Senegal and the Gambia genotyped on a high-density SNP array. This effort was led by Alfred Ngwa and David Conway (MRC Gambia) and Dan Neafsey and Steve Schaffner (Broad Institute). I processed and filtered array genotype data for this project and performed long haplotype selection tests (REHH and iHS).

[175] Sarah K Volkman, Daouda Ndiaye, Mahamadou Diakite, Ousmane A Koita, Davis Nwakanma, Rachel F Daniels, Daniel J Park, Daniel E Neafsey, Marc A T Muskavitch, Donald J Krogstad, Pardis C Sabeti, Daniel L Hartl, and Dyann F Wirth. Application of genomics to field investigations of malaria by the international centers of excellence for malaria research. *Acta Tropica*, 121(3):324–332, March 2012. doi: 10.1016/j.actatropica.2011.12.002. URL http://dx.doi.org/10.1016/j.actatropica.2011.12.002—This paper is a review paper malaria genomics as well as a description of the new ICEMR project (International Centers of Excellence for Malaria Research). This was led by Sarah Volkman (Harvard SPH). I played a minor role, and contributed small sections of text.

[101] Danny A Milner, Jimmy Vareta, Clarissa Valim, Jacqui Montgomery, Rachel F Daniels, Sarah K Volkman, Daniel E Neafsey, Daniel J Park, Stephen F Schaffner, Nira C Mahesh, Kayla G Barnes, David M Rosen, Amanda K Lukens, Daria Van Tyne, Roger C Wiegand, Pardis C Sabeti, Karl B Seydel, Simon J Glover, Steve Kamiza, Malcolm E Molyneux, Terrie E Taylor, and Dyann F Wirth. Human cerebral malaria and Plasmodium falciparum genotypes in Malawi. *Malaria Journal*, 11:35, March 2012. doi: 10.1186/1475-2875-11-35. URL http://www.malariajournal.com/content/11/1/35—This paper describes the application of a 24 SNP PCR barcode assay to explore the association between multiplicity of infection and severity of disease outcome in children. It was

led by Dan Milner (Harvard SPH). I played a minor role, mostly in the initial population genetic analyses that led towards the selection of the 24 SNPs.

[21] Kate M Broadbent, Daniel J Park, Ashley R Wolf, Daria Van Tyne, Jennifer S Sims, Ulf Ribacke, Sarah Volkman, Manoj Duraisingh, Dyann F Wirth, Pardis C Sabeti, and John L Rinn. A global transcriptional analysis of Plasmodium falciparum malaria reveals a novel family of telomere-associated lncRNAs. *Genome Biology*, 12(6):R56, June 2011. doi: 10.1186/gb-2011-12-6-r56. URL http://genomebiology.com/2011/12/6/R56—This paper describes a transcriptional scan and characterization of a family of long non-coding RNAs in the parasite *in vitro*. This was led by Harvard Systems Biology graduate student Kate Broadbent. I provided some statistical analyses and bridged the project from its initial rotation student to its final owner.

[115] Daniel E Neafsey, Mara K N Lawniczak, Daniel J Park, Seth N Redmond, M B Coulibaly, S F Traoré, N Sagnon, C Costantini, Charlie Johnson, Roger C Wiegand, Frank H Collins, Eric S Lander, Dyann F Wirth, Fotis C Kafatos, Nora J Besansky, George K Christophides, and Marc A T Muskavitch. Snp genotyping defines complex gene-flow boundaries among african malaria vector mosquitoes. *Science*, 330(6003):514–7, Oct 2010. doi: 10.1126/science.1193036. URL http://www.sciencemag.org/content/330/6003/514—This paper describes population genetic analyses of the vector mosquito, *Anopheles gambiae*. It was led by Dan Neafsey (Broad Institute), Mara Lawniczak (Imperial College London), and Marc Muskavitch (Boston College). I designed the genotyping array used for this study, and provided early-stage analysis of the array data.

[114] Daniel E Neafsey, Bridget M Barker, Thomas J Sharpton, Jason E Stajich, Daniel J Park, Emily Whiston, Chiung-Yu Hung, Cody McMahan, Jared White, Sean Sykes, David Heiman, Sarah Young, Qiandong Zeng, Amr Abouelleil, Lynne Aftuck, Daniel Bessette, Adam Brown, Michael Fitzgerald, Annie Lui, J Pendexter Macdonald, Margaret Priest, Marc J Orbach, John N Galgiani, Theo N Kirkland, Garry T Cole, Bruce W Birren, Matthew R Henn, John W Taylor, and Steven D Rounsley. Population genomic sequencing of Coccidioides fungi reveals recent hybridization and transposon control. *Genome Research*, 20(7):938–946, July 2010. doi: 10.1101/gr.103911.109. URL http://genome.cshlp.org/content/

20/7/938—This paper has nothing to do with malaria. It describes the population genetics and interspecific analyses of two related fungi: *Coccidioides immitis* and *Coccidioides posadasii*. This was led by Dan Neafsey (Broad Institute). I played a minor role, assisting with SNP discovery and gene annotation.

A.2 Manuscripts in preparation or review

Additionally, I have a few more manuscripts currently in preparation or review:

Ulf Ribacke, Mackenzie Bartlett, Saurabh D Patel, Niroshini Senaratne, Daniel J Park, Manoj T Duraisingh, Pardis C Sabeti, Sarah K Volkman, and Dyann F Wirth. Adaptive evolution of a ubiquitin ligase is linked to altered drug sensitivity in plasmodium falciparum. *Science Translational Medicine*. Manuscript under review.—I played a minor role, assisting with the interpretation and reanalysis of results from the REHH selection test in Van Tyne et al. [170] and incorporating derived allele analyses for mutations of interest.

Hsiao-Han Chang, Eli L Moss, Daniel J Park, Daouda Ndiaye, Soulyemane Mboup, Roger C Wiegand, Sarah K Volkman, Pardis C Sabeti, Dyann F Wirth, Daniel E Neafsey, and Daniel L Hartl. The malaria life cycle intensifies both natural selection and random genetic drift. *Proceedings of the National Academy of Sciences, USA*. Manuscript in preparation.—I played a minor role, providing helpful conversations and feedback on the population model used in this study and guiding the protocols for the analysis of raw sequence data. This constitutes the final chapter of Hsiao-Han's OEB dissertation, submitted to Harvard in May 2013.

Awa B Deme, Amy K Bei, Ousmane Sarr, Daniel E Neafsey, Stephen F Schaffner, Daniel J Park, Rachel F Daniels, Aida Sadikh Badiane, Papa El Hadji Omar Gueye, Ambroise Ahouidi, Daouda Ndiaye, Souleymane Mboup, Dyann F Wirth, and Sarah K Volkman. Analysis of the pfhrp2 genetic diversity in senegal and implications for rapid diagnostic test use. *Malaria Journal*. Manuscript in preparation.—I played a minor role, providing some advice on population genetic analyses and providing some summarized sequence data for the gene *pfhrp2*

from sequenced parasites in Senegal. This effort was led by Senegalese graduate student Awa Deme, who was a participant in a computational methods workshop that I taught in Dakar, Senegal in 2010.

Daria Van Tyne, Daniel J Park, and Dyann F Wirth. Understanding malaria drug resistance evolution in real-time. *Trends in Parasitology*. Manuscript in preparation.—I provided edits to this review paper in the areas of GWAS and selection studies.

This is a final term paper assignment for OEB242 (Coalescent Theory), taught in Fall 2010 by John Wakeley (Harvard University). See Section B.4 for details on author contributions.

B

Ascertainment Bias Corrections to Selection Studies in *Plasmodium falciparum*

This chapter explores and expands on a previous analysis of directional and balancing selection in the malaria parasite, *Plasmodium falciparum*. This previous analysis was a visualization of *P. falciparum* gene diversity vs. divergence in Chapter 2 [170]. Here, I explore ways that this analysis may have been affected by ascertainment bias and attempt to correct for it and calculate measures of significance utilizing coalescent approaches.

B.1 Introduction

Genomic studies of the *Plasmodium falciparum* parasite have advanced considerably in the past decade. From the initial sequencing and assembly of the genome to the characterizations of population diversity, and numerous studies since then, we now have a growing understanding of *P. falciparum*'s population history, global population structure, LD structure, and some of the strongest foci of selection pressure. Despite the rapidly decreasing cost of high throughput sequencing, a number of recent and continuing studies still rely on SNP genotyping chips to measure individual parasites for genomic variation [e.g. 107, 113].

However, it has been long known that a number of population genetic analyses from chip-derived data are sensitive to ascertainment bias [32, 117, 118]. This is due to the fact that the arrays only assay a pre-specified set of SNPs obtained in a smaller sample of individuals. This predisposes us to the detection of mostly common polymorphisms and biases the allele frequency spectra. In the case of *P. falciparum*, most current genotyping array designs are based on discovery data from less than 20 haploid strains originally described by a trio of 2007 papers [74, 106, 174]. Most studies in *P. falciparum* manage to avoid analyses that would be significantly affected by ascertainment bias, but some simply concede its possible influence. For example, Neafsey et al. and Van Tyne et al. use the terminology "SNP π" to emphasize its departure from the usual population genetic parameter, π, due to ascertainment bias.

In malaria genomics research, we ultimately wish to identify areas of the genome that may be under selection due to host immune pressure, drug pressure, or other factors. In this chapter, I wish to further explore a small analysis by Dan Neafsey (Figure 2.1B in this dissertation). I reproduce it here as Figure B.1A. This visualization plots the SNP π of a gene (within one subpopulation) against the gene's F_{ST} as a measure of within-population diversity vs. between-population divergence. The visualization suggests that data points with high diversity and low divergence may be candidates for genes under balancing selection. Conversely, genes with low diversity and high divergence may be candidates for population-specific positive selection. It is similar to Figure 4 from Ochola et al. [125], though Ochola

et al. use a interspecific metric of "divergence" (π/K and θ/K from a HKA test) whereas Van Tyne et al. use a intraspecific divergence metric (F_{ST}). Ochola et al. also does not suffer from ascertainment bias, as they PCR resequenced a set of candidate genes, whereas Van Tyne et al. utilizes the biased metric called "SNP π." In this chapter, I attempt to examine and correct for the effects of ascertainment bias in Figure B.1A and calculate measures of significance for the data.

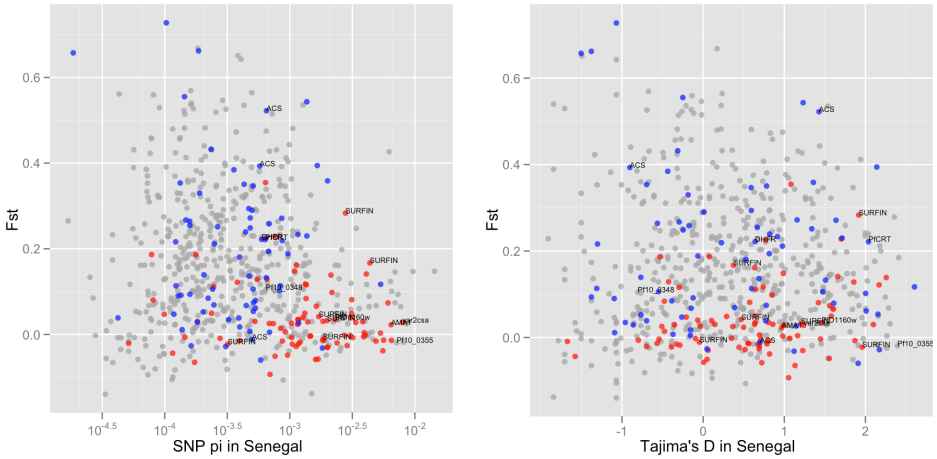


Figure B.1: Left: 650 genes in *P. falciparum* plotted by diversity within Senegal (SNP π) vs. divergence between Senegal and Thailand (F_{ST}). Genes on the lower right are suggestive of balancing selection and show an enrichment for known antigens, surface proteins, and highly polymorphic genes (red). Genes on the upper left are suggestive of directional positive selection and show an enrichment for known enzymes and transporters (blue). A handful of genes of biological interest are labeled. This figure was originally shown in Van Tyne et al. [170, Fig. 1B] (also Figure 2.1B in this dissertation) with a slightly different rendering. **Right:** Senegal Tajima's D vs. F_{ST}. Although Tajima's D should provide a better indication of balancing vs. positive selection, this graph fails to separate known antigens and enzymes as clearly as the one based on SNP π.

B.2 Results and Discussion

Even after correcting the values of SNP π in Figure B.1A for ascertainment bias, it would be helpful to give a measure of significance to genes at the extreme ends of the plot. The Tajima's D statistic can be thought of as an attempt to normalize the amount of polymorphism in a gene, Π, against the amount expected based on the

number of segregating sites, θ_W. This addresses some other concerns, such as the idea that a distibution of π might be affected by variations in mutation rate or gene age. Tajima's D should help discern candidates of balancing selection ($D > 0$) and positive selection ($D < 0$) from neutrality ($D = 0$). Unfortunately, this metric does a worse job of visually separating genes that, based on annotation, are thought to be candidates for balancing and positive selection (Figure B.1B). An alternate visualization of Tajima's D plots $\hat{\theta}_W$ against $\hat{\theta}_T$ (Figure B.2A), where Tajima's D is the normalized deviation from the unity line. Both figures also demonstrate an overall skew in the positive direction, which is characteristic of ascertainment bias.

Corrections for ascertainment bias were made based on Ramírez-Soriano and Nielsen [144] to values for $\hat{\theta}_T$ and $\hat{\theta}_W$ and Tajima's D (Section B.3). The corrected values are shown in Figure B.2B. These corrections succeed in centering Tajima's D around zero for most values of $\hat{\theta}$. Oddly, for low values of $\hat{\theta}_W$, Tajima's D is shifted quite negative. Although it has been long known that the *Plasmodium falciparum* population has experienced multiple bottlenecks and population expansions [78, 158] such effects should not be limited to genes with lower polymorphism, so the effect seen here is puzzling.

Ascertainment corrections to π and Tajima's D (Figure B.3) do not drastically change the overall picture from Figure B.1. Corrected π is a slightly better discriminator between balancing and positive selection candidate genes than SNP π, but the corrected Tajima's D is still unable to distinguish the two. Ultimately, we can see that while both Π (Figure B.2B) and π (Figure B.3A) can visually separate balancing and positive selection candidates, Tajima's D (Figure B.3B) cannot, even after corrections for ascertainment bias. This may imply that our previously defined categories (known highly variable genes/surface proteins vs. enzymes/transporters) may not be the best control groupings for this test. Even among the highly polymorphic genes where one might expect an enrichment for positive Tajima's D, many of these genes lie in large gene families (e.g. the vars, rifins, stevors, etc.), often in the subtelomeres, that result from many gene duplication events. Speculatively, this may mean that the "demographic history" of these gene copies vary widely and might confound a Tajima's D analysis of these families. Even so, since the resulting distribution of Tajima's D values shows rough concor-

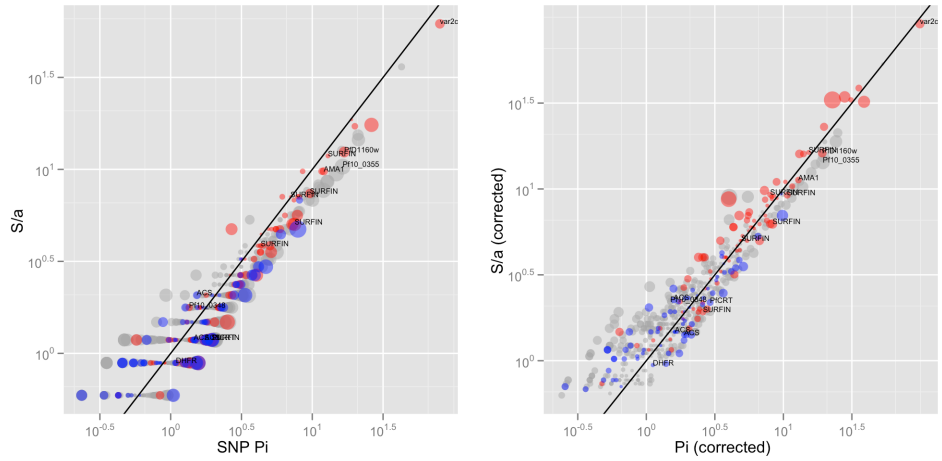


Figure B.2: A visualization of Tajima's D by plotting Watterson's $\hat{\theta}_W$ against Tajima's $\hat{\theta}_T$. **Left:** original data. **Right:** corrected for ascertainment bias. Genes to the right of the unity line have positive D and genes to the left are negative. Circle size corresponds to the absolute value of Tajima's D. As before, red genes correspond to candidates for balancing selection and blue genes correspond to candidates for positive selection. The SNP π plotted in Figure B.1A differs from the SNP Π shown here in that the former equals the latter divided by gene length.

dance with a normal distribution (Figure B.4C), we would not be too astray in interpreting these corrected values as Z-scores, using ± 2 as a significance threshold.

In the end, this chapter succeeds in correcting a number of parameters for ascertainment bias for *P. falciparum* genes in a population of Senegal-derived parasites. It uses these to correct and further explore and expand on a previous visualization of this data: a diversity-divergence plot that was used to identify candidates of balancing and positive selection. Although the corrected Tajima's D metric fails to separate a control set of genes thought to be under different types of selection, we are able to show that it provides a measure of significance. This chapter also shows a foundation of a naive coalescent model used to compute significance, which can be expanded upon in the future to account for confounding factors that are not currently addressed, such as demographic history. This model could also be used to simulate structured populations with limited gene flow. In addition to providing maximum likelihood estimates for migration rates between Senegal and Thailand,

it could also begin to explore the other axis of Figure B.1A not addressed in this chapter and possibly provide measures of significance for extreme F_{ST} values.

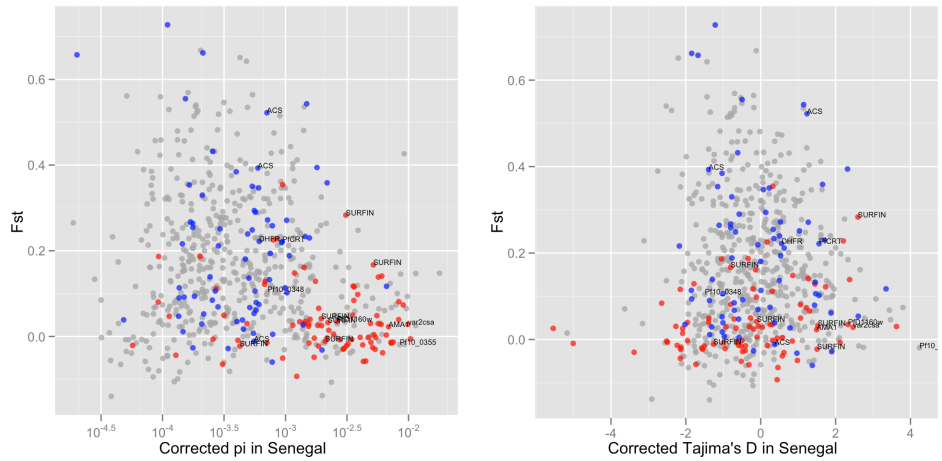


Figure B.3: Ascertainment corrections applied to the plots in Figure B.1.

B.3 Methods

Ascertainment bias corrections are derived from Ramírez-Soriano and Nielsen [144]. Although their downloadable Java program did not run on my data, all of the necessary equations for the corrections of $\hat{\theta}_W = S/a$ and $\hat{\theta}_T = \Pi$, are given in their main text (Equations 3, 5, and 7) and are implemented here.

Ascertainment bias corrections for Tajima's D was slightly more complicated, as the arithmetic for the denominator, the variance component of Tajima's D, was rather unwieldy [144, Eqn. 16]. As this author feared the possible programming errors in the absence of thorough testing, and because the denominator is based on assumptions of an underlying beta distribution, I opted instead to calculate $V(\Pi - S/a)$ from the results of 1000 coalescent simulations per gene. The numerator for Tajima's D is still based on ascertainment corrections from Ramírez-Soriano and Nielsen [144], so any remaining errors in my corrected Tajima's D values would be in magnitude, but not direction.

Simulations were performed using ms [68] using a simple, unstructured model. To simulate a distribution of locus lengths that matched a realistic distribution

of gene lengths, I run ms once for each gene with parameters scaled to the gene length. I estimated $\hat{\theta}/L$, where L is the length of the gene in bp, as the mean $S/(aL)$ and mean $\pi = \Pi/L$ values, which were both approximately 4×10^{-4}. A sample size of 17 individuals was used, to match the Senegalese sample size of our empirical data. A recombination parameter, $\hat{\rho}/L = 0.0057$, is taken from Chang [26]. However, unlike Chang, I do not attempt to model population growth. The final ms command line looks like this:

```
ms 17 1000*ngenes -t 0.0004*glen -r 0.0057*glen glen \
   | sample_stats > outputfile
```

where `glen` is the simulated gene length and `ngenes` is the number of genes that have the length `glen`. The results of this simulation is shown in Figure B.4A. The variance of $\Pi - S/a$ is calculated as a function of S/a and is shown to vary linearly with S/a (Figure B.4B). A linear fit is performed and used to extrapolate all of $V(\Pi - S/a)$ as a function of S/a.

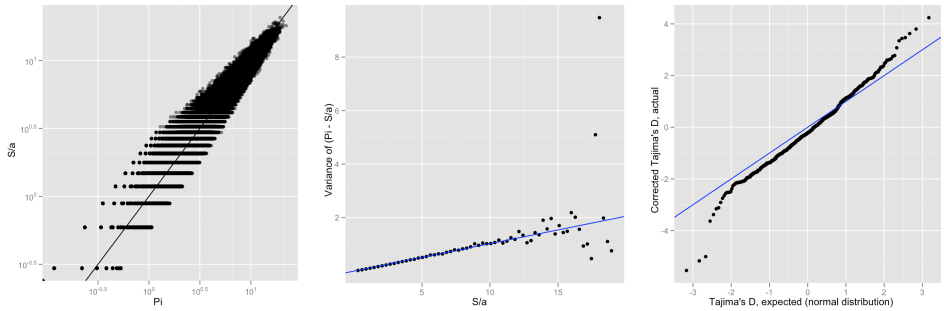


Figure B.4: Left: Watterson's $\hat{\theta}_W$ against Tajima's $\hat{\theta}_T$ for 650,000 simulated genes with the same gene length distribution as the empirical data. **Center:** simulated data shows a linear relationship between $V(\Pi - S/a)$ and S/a. A best fit line is shown in blue. This line is used to compute a simulated denominator of corrected Tajima's D for all genes. **Right:** quantile-quantile plot shows that the corrected Tajima's D values roughly conform to a normal distribution.

The square-root of this simulated variance is used as the denominator for the corrected Tajima's D, and the Ramírez-Soriano and Nielsen corrected values of $\hat{\theta}_{T,C} - \hat{\theta}_{W,C}$ are used as the numerator. The quantile-quantile plot in Figure B.4C shows the goodness of fit between the final, corrected Tajima's D values, and the

normal distribution ($\mu = 0, \sigma = 1$, which we use here as an approximator for Tajima's beta distribution) that might be expected under neutral, panmictic, constant population size conditions.

Figures were rendered using the ggplot2 package in R. Analysis was done in python, Java, and R (utilizing the plyr package).

B.4 Acknowledgements

I am grateful to Hsiao-Han Chang (Harvard OEB) and Dan Neafsey (Broad Institute) for helpful conversations, discussions and suggestions for this chapter. HHC also provided maximum-likelihood estimates of simulation parameters [26]. DN and Steve Schaffner (Broad Institute) originally produced Figure B.1A which inspired this modified analysis.

Sample collection was largely the work of collaborators at the Harvard School of Public Health (Boston, MA) and Chiekh Anta Diop University (Dakar, Senegal). Chip hybridizations were performed by Charlie Johnson (Broad Institute). I designed the chip in conjunction with Affymetrix, and the results were analyzed by me.

All data and analyses leading up to Figure B.1A are prior work represented in Chapter 2. All analyses beyond Figure B.1A are original to this chapter and this author. This work was undertaken in the context of a coalescent theory class (OEB252) taught by John Wakeley (Harvard OEB).

C

Supplemental Material for Chapter 2

C.1 Author Contributions to Supplemental Material

Most analyses and figures in this supplemental appendix were produced by me. In the Supplemental Methods, I performed all SNP Discovery, Array Development and Assessment, and the Genome-wide Association Study (GWAS). Copy Number Variation (CNV) Analysis of array data was performed by DEN. PCR Genotyping was peformed by numerous techs at HSPH.

Figure C.1 was produced by SFS and PCS. Figure C.2 and Table C.2 were produced by SFS. Figures C.4 and C.11 were produced by DEN. Figures C.5 and C.6 and Table C.3 were produced by EA. Figure C.12 and Tables C.6 and C.4 were produced by DVT. Figure C.3 and Tables C.1, C.5, and C.7 were compiled by SKV.

Both Supplemental Data Files were produced by me. I produced Figures C.7, C.8, C.9, C.10 and C.13. I also produced all Figures and Tables Supporting Supple-

mental Methods: Figures C.14, C.15, C.16, C.17, C.18, C.19, and C.20 and Table C.8.

C.2 Supplemental Methods

C.2.1 Parasites

Parasites (Table C.1) were obtained from the Malaria Research and Reagent Resource Repository (MR4, http://malaria.mr4.org/) or additional sources noted (Table C.1, Acknowledgements). The following parasite lines were obtained through MR4: parasite line 3D7 (MRA-151, deposited by D. Walliker); parasite line 7G8 (MRA-152, deposited by D. Walliker); parasite line HB3 (MRA-155, deposited by T.E. Wellems); parasite line Dd2 (MRA-156, deposited by T.E. Wellems); parasite line K1 (MRA-159, deposited by D.E. Kyle); parasite line V1/S (MRA-176, deposited by D.E. Kyle); parasite line RO-33 (MRA-200, deposited by D. Walliker, U. Certa and R. Reber-Liske); parasite line D10 (MRA-201, deposited by Y. Wu); parasite line TM90C2A (MRA-202, deposited by D.E. Kyle); parasite line TM90C6A (MRA-205, deposited by D.E. Kyle); parasite line TM91C235 (MRA-206, deposited by D.E. Kyle); parasite line WR87 (MRA-284, deposited by D.E. Kyle); parasite line D6 (MRA-285, deposited by D.E. Kyle); parasite line Malayan Camp (MRA-330, deposited by L.H. Miller and D. Baruch); parasite line Indochina I (MRA-347, deposited by W.E. Collins); parasite line Santa Lucia (MRA-362, deposited by W.E. Collins); parasite line FCC-2 (MRA-733, deposited by W. Trager); and parasite line T2-C6 (MRA-818, deposited by X. Su). Patient samples were obtained as part of ongoing studies in Senegal and Malawi described elsewhere in accordance with human subject guidelines. Additional parasites were the kind gift of: Alejandro Miguel Katzin (51, 10_54, 36_89, and 9_411); Christian Happi (APO41); Abdoulaye Djimde and Chris Plowe (PS189); Joseph Smith (A4); Karen Day (Muz51.1); Dennis Kyle and Sodsri Thaithong (TD203, TD257, TM327, TM345, GH2, and PR145); Sandra do Lago Moraes (JST); and Xin-zhuan Su (MR24). DNA from *P. reichenowi* was kindly provided by John Barnwell.

C.2.2 PCR Genotyping

Genomic regions (458850-459204) surrounding the *pfcrt* (MAL7P1.27) locus was amplified using the polymerase chain reaction using oligonucleotide primer sequences (CCTTGTCGACCTTAACAGATG, CTATTCCACCTACCAATATAAAAC) and the resulting DNA sample was sequenced using standard methods. In a similar manner the genomic region (754984-755584) surrounding the *dhfr* locus (PFD0830w) (oligonucleotide primer sequences: CAAGATTGATACATAAAGATAATAT, TTCTTGATAAACAACGGAACCTCCT); and the *pfmdr1* locus (PFE1150w) (oligonucleotide primer sequences: TGTTGAAAGATGGGTAAAGAGCAGAAAGAG, TACTTTCTTATTACATATGACACCACAAACA) were utilized [49].

C.2.3 SNP Discovery

The SNP discovery methodology was similar to those described in Volkman et al. [174]. 1× ABI shotgun sequence was obtained for nine geographically diverse parasite isolates that were previously sequenced to 0.25× coverage, bringing total coverage to 1.25× per isolate. These nine isolates include: 7G8, Santa Lucia (El Salvador), V1/S, D10, FCC-2/Hainan, D6, RO-33, Senegal V34.04 and K1. Three of the twelve previously sequenced isolates in Volkman et al. were excluded from additional sequencing, as they were previously found to be nearly genetically identical, suggesting possible contamination in culture [174]. Reads ends with low quality (PHRED < 10) bases were trimmed. Reads less than 100 bases, containing greater than 3% internal N's, or containing a mononucleotide repeat covering greater than 80% of the read were discarded. Reads were aligned to the PlasmoDB version 5 of the 3D7 genome using BLAT37 [81] requiring 95% identity, a minimum score of 100, less than 20% gaps, and coverage of at least half of the read. Only the highest scoring alignment for each read was kept and paired reads which aligned more than 10kb apart or in the wrong orientation were discarded. The Neighborhood Quality Standard (NQS) algorithm was used to distinguish real polymorphisms from sequence errors [4]. We required the SNP to have a minimum quality score of 25, and the five base neighborhood to have a minimum score

of 20. We allowed one mismatch and no indels in the neighborhood. We discarded SNPs when another read from the same sample met the NQS criteria at that position but did not have a sequence difference.

C.2.4 Array Development and Assessment

Based on all 111,536 discovered SNPs [74, 106, 174] in *P. falciparum*, and given design parameters and unique sequence constraints, we were able to design assays for 74,656 markers. Each of 74,656 SNPs is represented by a probe set of 12 to 84 probes, for a total of 4.4 million genotyping probes on the Affymetrix 49-format array. These were hybridized to 63 unique samples (totaling 81 arrays with replicates). Genotype calls were produced using the BRLMM-P algorithm [1], a variant of the RLMM algorithm [143], included in Affymetrix Power Tools version 1.8.5, and clustered over all 81 arrays. BRLMM-P was forced into a haploid calling mode by setting assigning all SNPs to the "Y chromosome" and setting all arrays to "male".

The array with sample TM93C1088 is eliminated immediately after clustering (arbitrarily, since the chip claiming to be TM90C6A and the chip claiming to be TM93C1088 are identical). We also remove samples CF04.010 and Senegal Th10.04, which were suspected to be multi-clonal based upon molecular barcode analysis [39]. A halofuginone-resistant version of Dd2, a human-DNA sample, and the *P. reichenowi* ancestral samples are also removed at this stage, leaving 57 unique samples (totaling 75 arrays with replicates) for analysis. We then calculate a call rate for each SNP and remove 7,778 SNPs that have below an 80% call rate, leaving 66,878 SNPs. Since technical replicates showed 99.9% repeatability between chips, we merged replicate data for each of the 57 samples, producing a no-call when the replicates indicated discordant genotypes.

Concordance against sequencing data was calculated in both major and minor alleles for 17 sequenced reference strains [113]. The following 17 samples were compared against sequencing data for concordance: 3D7, Dd2, FCC-2, Malayan Camp, D10, K1, V1/S, RO33, D6, Senegal P31.01, Senegal P51.02, Senegal V34.04, Senegal V35.04, 7G8, A4 (subclone of IT04 [147]), Santa Lucia,

and HB3. These are the 18 parasites presented in Fig 1 of Volkman et al. [174], removing the three found to be genetically identical, and adding the two strains 3D7 and A4. A total of 18,303 SNPs lacked call overlap between array genotypes and sequencing genotypes in minor alleles and were thus removed, since concordance in both alleles could not be fully calculated. Another 30,993 SNPs were removed due to imperfect concordance, and of these discordant SNPs, most (28,789) exhibited monomorphic behavior on the array, suggesting that much of the discordance may be attributed to either a faulty assay or false discovery. The remaining 17,582 perfectly concordant SNPs constituted the high confidence set of assays used in our analyses.

C.2.5 Copy Number Variation (CNV) Analysis

We examined the ability to detect copy number variants (CNV) using the array by first studying a known CNV using the hybridization intensity signal of the SNP genotyping probes on the array. Kidgell et al. [83] reported that the *pfmdr1* locus was present in 3-4 copies in the Dd2 strain relative to a collection of other strains. We compared Z-scores of the normalized hybridization intensity of perfect match probes for SNPs in the neighborhood of *pfmdr1* for Dd2 and six parasites estimated by Kidgell et al. to contain only 1 copy of the locus (3D7, 7G8, HB3, D10, D6, K1). For each SNP assay we utilized the average hybridization intensity of all perfect-match probes. Hybridization intensity values were background corrected and normalized to reduce inter-array variation artifacts. SNPs with a hybridization intensity standard deviation equal to or greater than half the magnitude of the average hybridization intensity across all arrays were excluded from analysis. Figure C.11 illustrates that probes for many of the SNPs assayed within the *pfmdr1* locus exhibit notably higher hybridization intensity values in Dd2 relative to the other parasites, with 13 assays exhibiting average intensities greater than 2 standard deviations higher than observed in the other parasites.

C.2.6 Genome-wide Association Study (GWAS)

We performed GWAS for drug resistance to thirteen antimalarials: amodiaquine (ADQ), artemether (ARTM), artesunate (ARTN), artemisinin (ARTS), atovaquone (ATV), chloroquine (CQ), dihydroartemisinin (DHA), halofuginone (HFG), halofantrine (HFN), lumefantrine (LUM), mefloquine (MFQ), piperaquine (PIP) and quinine (QN). 50 out of 59 samples had drug phenotype data. IC_{50} data are shown in Table C.4 and Figure C.13 for these 50 parasites against the 13 drugs.

The following drugs were obtained from Sigma Aldrich: artemisinin, dihydroartemisinin, chloroquine, mefloquine, and quinine. The following were obtained from AK Scientific: artemether, artesunate, halofantrine, lumefantrine, and piperaquine. The following were obtained from USP: amodiaquine and atovaquone. Each drug was tested in triplicate for each parasite. Additionally, some parasites were tested with multiple biological replicates: 3D7 (nine biological replicates per drug, each in triplicate), Dd2 (three replicates) and RO-33, D10, and 207-89 (two replicates).

SNPs were filtered down to a set that contained at least 5 strains with a minor allele as well as an 80% call rate under every phenotype condition. The final data set includes 7,437 SNPs. This gives us a genome-wide significance threshold of $\log_{10}(\text{P-value}) > 5.17$ by Bonferroni correction for multiple testing. For binary phenotype tests (Fisher's exact test, Fisher's permuted, CMH, and HLR), we used IC_{50} cutoffs shown in Table C.4. For tests requiring defined geographic clusters (CMH, HLR, Fisher's permuted), the three population clusters are defined by PCA, as in the LRH analysis, and the assignments are shown in Table C.1.

Pointwise P-values were computed using PLINK [142]. Quantile-quantile plots (qq-plots) were used to examine the resulting P-value distributions for inflating effects due to population structure (Figure C.7). Because most of the genome is assumed to fit the null hypothesis (most of the genome should not be in association with the phenotype), significant, early deviations from this expectation may result in a high false positive rate. The null expectation is plotted as the unity diagonal line in Figure C.7. Bonferroni significance is plotted as the dashed

line and Benjamini-Hochberg significance is marked with the dotted line. Since most Fisher's results show evidence of inflation, we do not report these results in Figure 2.2 or Table 2.1.

Permutations of Fisher's exact test can be used to compute empirical pointwise P-values based on a simulated null distribution. We used PLINK to perform this permutation while respecting the phenotype frequencies present in our three predefined population clusters. The resulting P-value distributions (Figure C.7) do not show inflation due to population structure, however no significant hits were found for any drug.

Similarly, the Cochran-Mantel-Haenszel (CMH) test can perform population-stratified analyses for association. We used PLINK to compute P-values (Figure C.7), and again, we see appropriate corrections for population structure, but no hits reach genome-wide (Bonferroni) significance.

The Efficient Mixed-Model Association (EMMA) test was specifically designed to handle quantitative trait associations to a data set with complex population structure using a linear mixed model [79]. It calculates a genotype similarity matrix instead of discrete categories and does not require a priori specification of population structure. The resulting P-value distributions demonstrate little remaining effect from population structure (Figure C.8) while retaining power to find a number of associations at genome-wide significance (Figure C.8, 2.2A, Table 2.1).

The Haplotype Likelihood Ratio (HLR) test is a multi-marker association test [90]. Unlike a standard, χ^2-based multi-marker test which looks for differences in haplotype frequencies in cases vs. controls, the HLR test specifically models the likelihood that a single haplotype rose to dominance in cases while all other haplotypes proportionally decreased. It produces a LOD score, which is the maximum likelihood estimate for the haplotype frequencies observed in cases $(O_1 \ldots O_k)$, given the distribution in controls $(f_1 \ldots f_k)$:

$$LOD_{ML} = \log_{10} \frac{P\left(O_1 \ldots O_k | ae_j + (1-a)(f_1 \ldots f_k)\right)}{P\left(O_1 \ldots O_k | f_1 \ldots f_k\right)}$$

where j is the haplotype on which the mutation arose, $1-a$ is the recombination

rate, and $e_j = 1$ when $i = j$ and $e_j = 0$ when $i \neq j$. The test produces maximum likelihood estimates for j and a.

So while a χ^2-based association finds any significant differences in haplotype frequencies, the HLR test models a specific scenario that is common in rapid selective events. The HLR test does not provide significantly more power than a single marker test in regions of high LD—in extreme cases, these regions may only have two haplotypes, and a multi-marker test will have the same power as a bi-allelic SNP test. But in regions where LD is low in controls and a single, long haplotype is prevalent in cases, the HLR test is highly sensitive. The HLR test is a one-sided test and we ran separate tests for both drug resistance (called "risk") and drug sensitivity ("protect"). Results for drug sensitivity are available in Figure C.10, but are not reported generally as we are more interested in selective events for drug resistance.

We used PLINK to produce sliding window haplotypes across the genome and calculate haplotype frequencies for input to the HLR test. We produced input for all two, four and six-marker windows. The resulting LOD scores did not map well to known distributions, such as the χ^2 1-degree of freedom distribution. We instead converted the pointwise LOD scores to empirical pointwise P-values by performing approximately 370,000 permutations of the null model for each test condition. This allows us to calculate empirical P-values up to a significance of about $\log_{10}(\text{P-value}) = 5.6$. Similar to the permuted Fisher's test, we preserved population-specific phenotype frequencies by only allowing permutations within each of our three defined populations. Resulting P-value distributions fit expectations well for the vast majority of test conditions (Figures C.9, C.10) and the test demonstrates power to detect a number of loci at genome-wide significance (Figure 2.2A, Table 2.1).

C.3 Supplemental Data Files

1. Drug data, *PF10_0355* copy number data, and top GWAS and LRH hits. Permanent URL (XLS, 60kB):
   doi://10.1371/journal.pgen.1001383.s001

2. Genotype data. Tab separated text file containing genotype data for 57 isolates across 17,582 SNPs. Additional information such as translation consequences (based on PlasmoDB v5.0 annotations) are also provided. Permanent URL (TXT, 3.3MB):
   doi://10.1371/journal.pgen.1001383.s002

C.4 Supplemental Figures

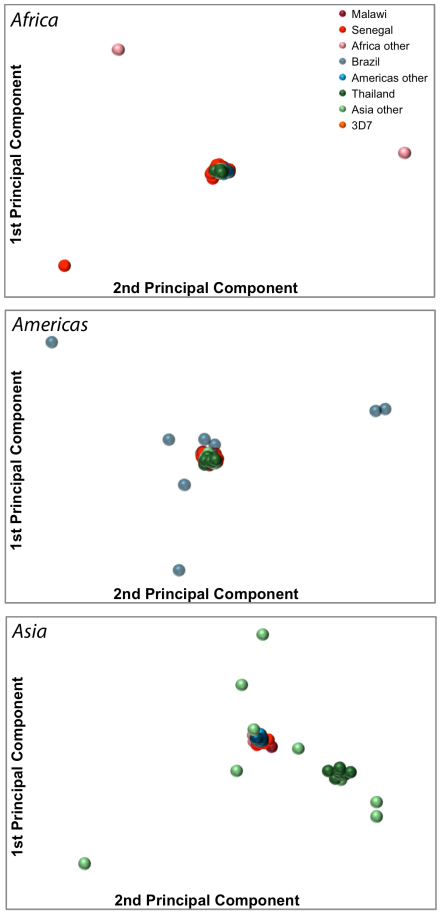


Figure C.1: Principal components analysis of population structure within **(A)** Africa, **(B)** the Americas, and **(C)** Asia. Plots of the first two principal components using Eigenstrat [130] using the Affymetrix array. Each solid circle represents an individual, and the color is assigned according to the reported origin.

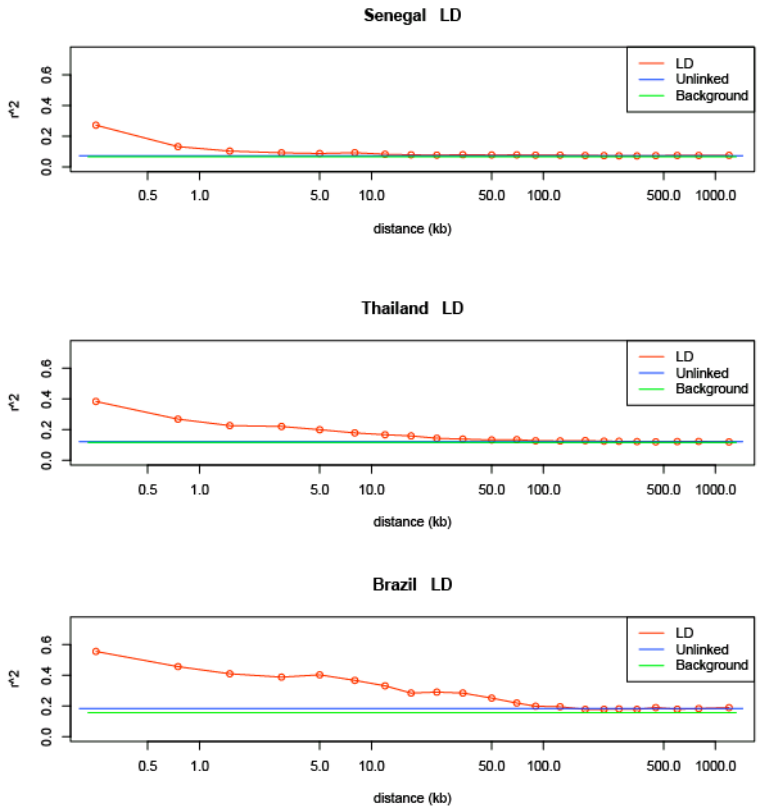


Figure C.2: Linkage disequilibrium (LD), measured by r^2, for each of the three population samples (Senegal, Thailand, Brazil). Plotted are r^2 for linked markers (red lines) and for unlinked markers (blue lines), as well as the level of background LD expected because of small sample size (green lines).

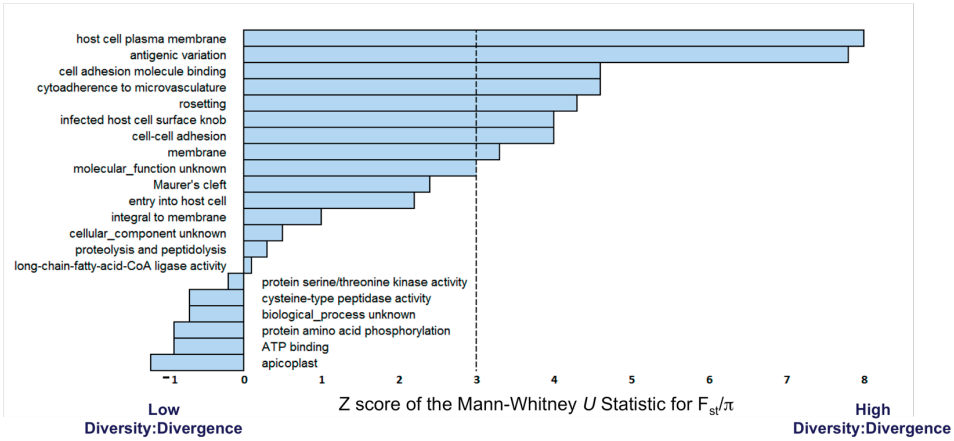


Figure C.3: Genes were classified by gene ontology (GO) functional categories and stratified by level of nucleotide diversity (π) as estimated by Z-scores. Select categories (highest five and lowest five categories along with categories in between that differ by incremental Z-scores) are shown. The majority of genes in GO categories for molecules found at the cell membrane have high levels of nucleotide diversity, while most of the genes classified into GO categories for conserved molecules lack nucleotide diversity.

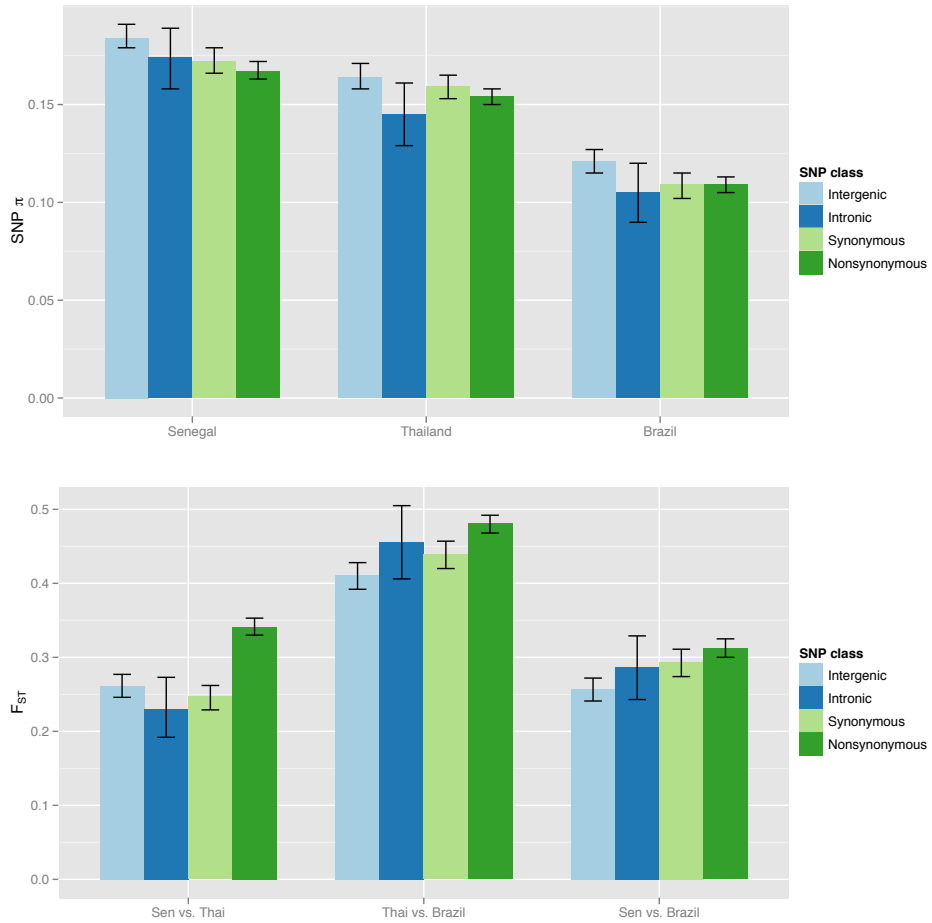


Figure C.4: SNP diversity and divergence by translation consequence. Diversity at assayed SNPs (SNP π) and Divergence between different populations as assayed by F_{ST}, for different classes of SNPs: intergenic (4,263 SNPs), intronic (584 SNPs), synonymous (3,957 SNPs), and nonsynonymous (8,778 SNPs). Intronic SNPs have the widest error bars due to their relative sparseness on the array. Non-synonymous SNPs are generally among the least diverse and most differentiated class of SNPs.

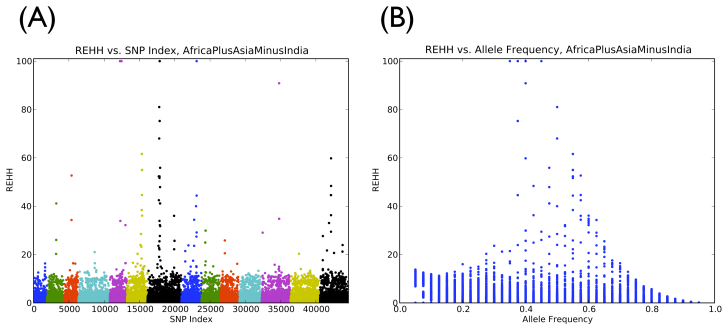


Figure C.5: Relative extended haplotype homozygosity (REHH) scores. Relative extended haplotype homozygosity (REHH) scores prior to any normalization, plotted for each core allele, **(A)** indexed by chromosome and position, and colored by chromosome, and **(B)** as a function of core allele frequency.

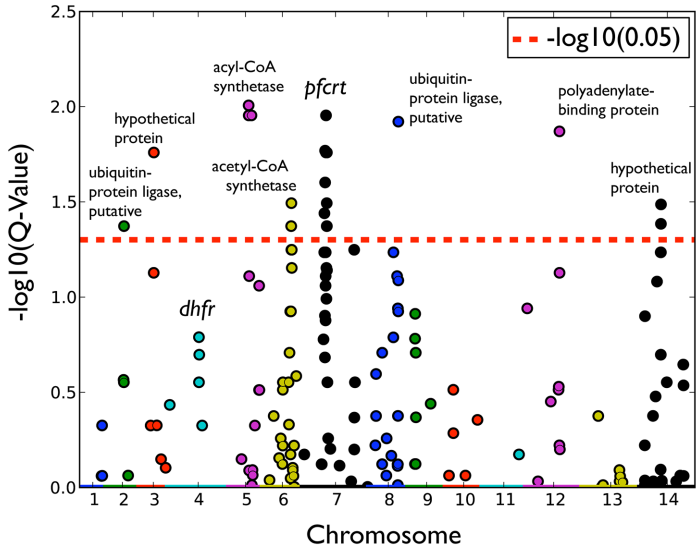


Figure C.6: Long-range haplotype (LRH) analysis yields genome-wide significant candidates for recent positive selection. For each core allele, we calculated relative extended haplotype homozygosity (REHH), and from the set of all REHH scores we calculated a corresponding distribution of Q-values. We plotted $-\log_{10}(Q)$, for all Q-values < 1, for each core allele, indexed by chromosome and position, and colored by chromosome. The red dotted line corresponds to the typical Q-value significance threshold of 0.05. Gene annotations from `http://plasmodb.org` for some significant scores are labeled. For comparison, the well-known sweeps around drug resistance loci *pfcrt* and *dhfr* are labeled. This data is also shown in tabular form in Table C.3.

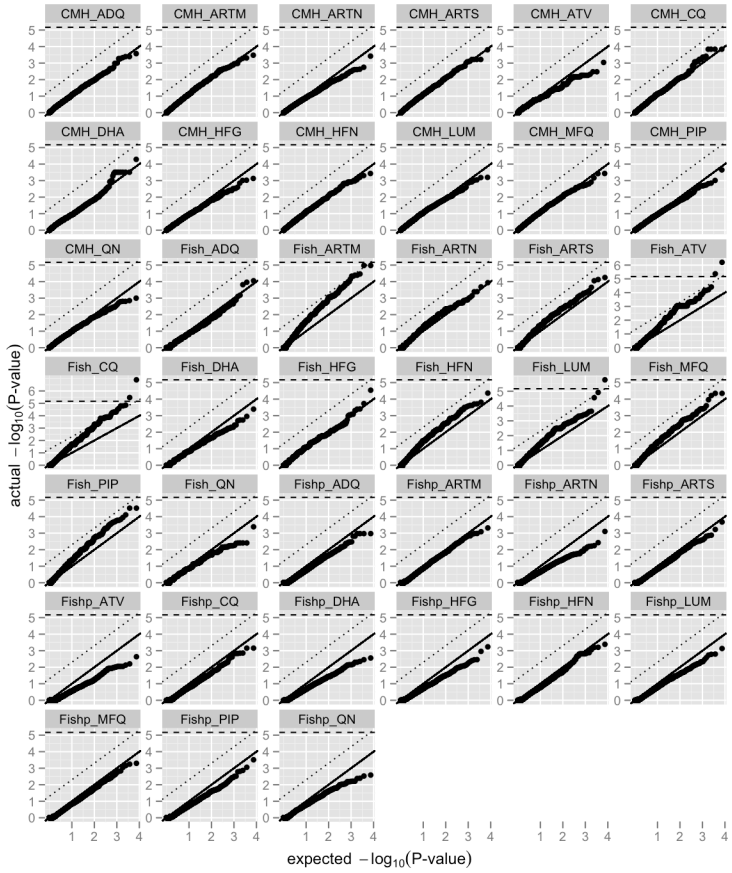


Figure C.7: GWAS P-value distributions for Fisher's exact test, permuted Fisher's exact test, and Cochran-Mantel-Haenszel (CMH) tests. Quantile-quantile plots (qq-plots) show $\log P$-values for every SNP on the y axis against the null expectation on the x axis. Fisher's exact test results generally show P-value inflation due to confounding effects from population structure for many drugs ("Fish"). As such, no results from this test are reported. To account for population structure, permutations of the null distribution were performed while preserving phenotypic associations to three predefined population clusters ("Fishp"). CMH also performs a stratified association test given predefined population clusters ("CMH"). The permuted Fisher's test and CMH test results show appropriate correction for population structure, but show no hits at genome-wide significance to report.

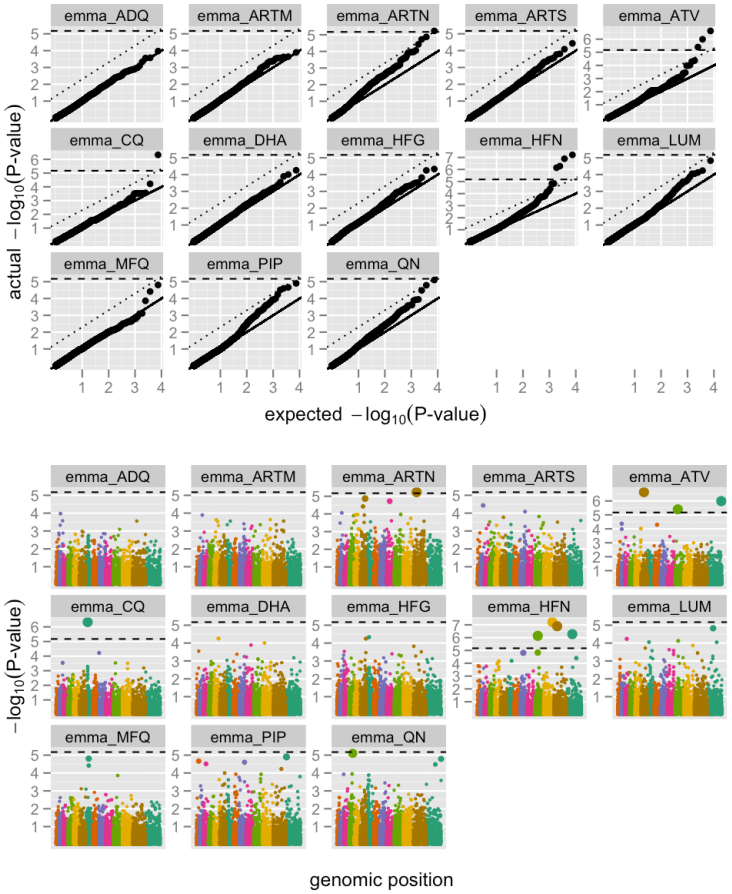


Figure C.8: GWAS results for the Efficient Mixed-Model Association (EMMA) test. QQ-plots show little to no confounding effect from population structure, with the possible exception of artesunate (ARTN). The significant ARTN result is not reported in Table 2.1 or Figure 2.2 for this reason. Manhattan plots depict the genomic location of significant hits, also reported in Table 2.1 and Figure 2.2.

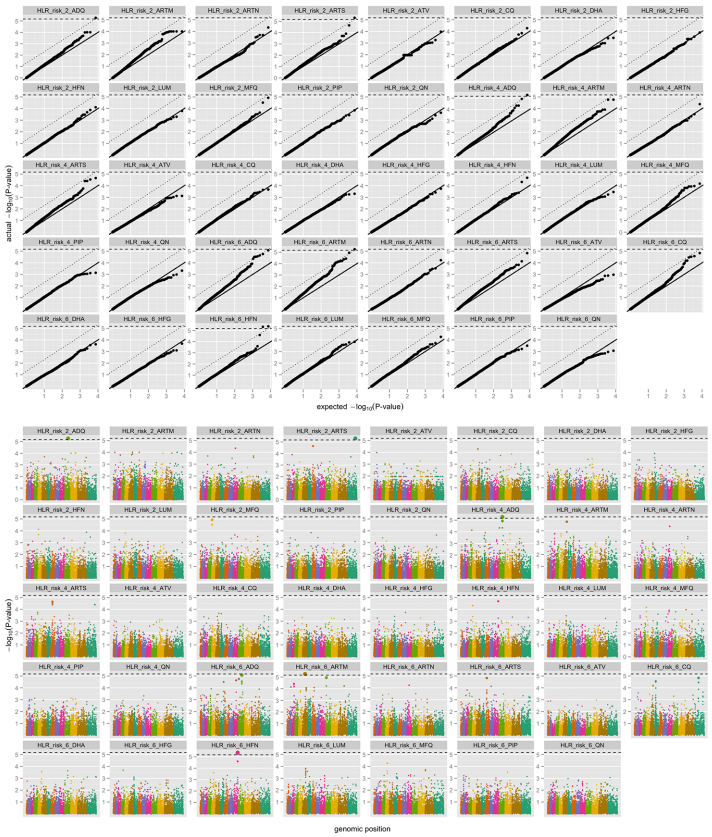


Figure C.9: GWAS P-value distributions for the Haplotype Likelihood Ratio (HLR) tests for association to drug resistance. Population-sensitive permutations of the null model were used to calculate P-values from LOD scores. Final distributions of P-values show little to no confounding effect from population structure for most tests. Exceptions include the 6-SNP artemether (HLR_risk_6_ARTM) test and the 4-SNP amodiaquine (HLR_risk_4_ADQ) test—these results are not reported in Table 2.1 or Figure 2.2. Manhattan plots for other tests that reached genome-wide significance are in Figure 2.2A.

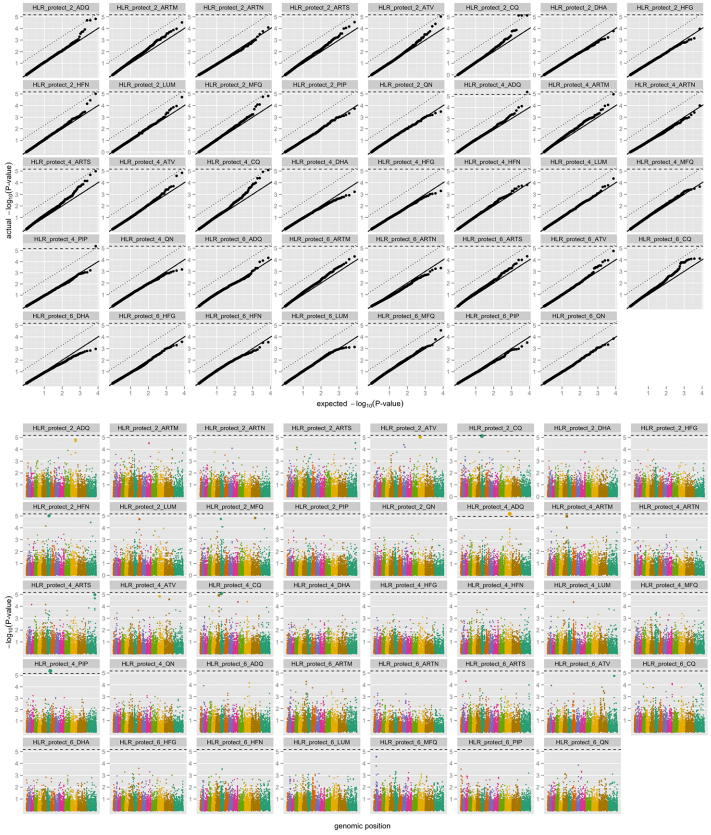


Figure C.10: GWAS P-value distributions for Haplotype Likelihood Ratio (HLR) tests for association to drug sensitivity. Population-sensitive permutations of the null model were used to calculate P-values from LOD scores. Final distributions of P-values show little to no confounding effect from population structure. Genome-wide significant hits include piperaquine (HLR_protect_4_PIP) on a haplotype that spans *PF07_0126*, *PF07_0127* and *MAL7P1_167* and amodiaquine (HLR_protect_4_ADQ) on a haplotype in *PFL1800w*. A chloroquine hit on *pfcrt* just misses genome-wide significance. These results are not reported in Table 2.1.

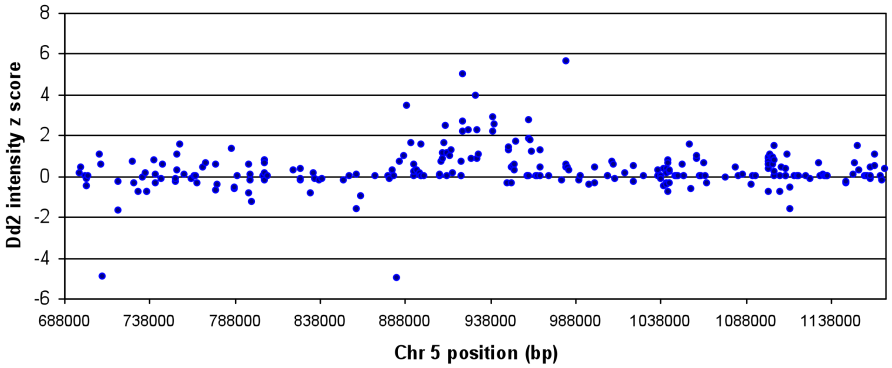
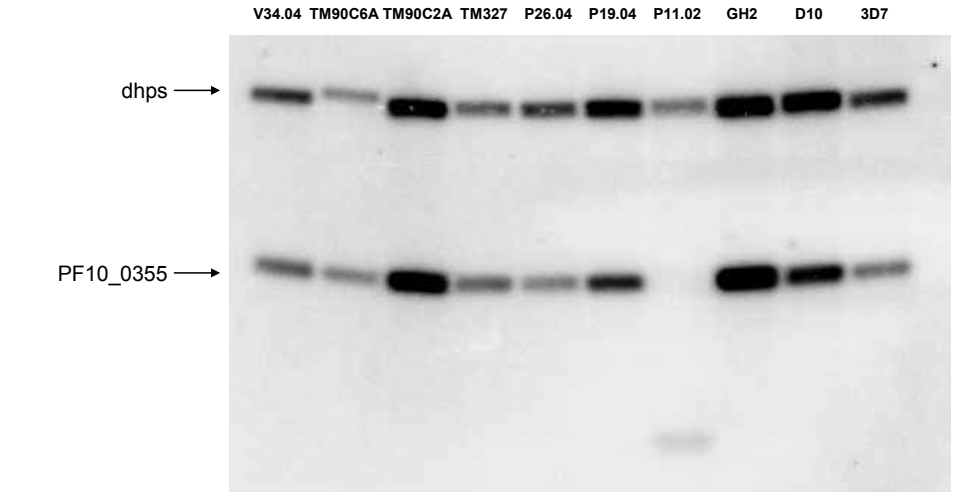


Figure C.11: Intensity Z-score for the Affymetrix array across chromosome 5. The results illustrate that probes for many of the SNPs assayed within the *pfmdr1* (888-988k) locus exhibit notably higher hybridization intensity values in Dd2 relative to the other parasites, with 13 assays exhibiting average intensities greater than 2 standard deviations higher than observed in other strains. This is consistent with the copy number variation reported in the *pfmdr1* locus, with 3-4 copies present in the Dd2 strain relative to a collection of other strains.



	V34.04	TM90C6A	TM90C2A	TM327	P26.04	P19.04	P11.02	GH2	D10	3D7
PF10_0355 copy number	1.54	**2.26**	**2.43**	**2.02**	1.21	1.18	0.53	**2.38**	1.06	1.00
HFN response	S	R	R	R	R	R	R	R	R	R

Figure C.12: *PF10_0355* copy number variation measured by Southern blotting. Select parasite isolates were digested with AflIII, EcoRV and XbaI and fragments were detected using probes to portions of the *PF10_0355* and *dhps* genes. Primers used for making probes were: *dhps* F: 5'-GTG ATT GTG TGG ATC AGA AGA TGA ATA ATC-3'; R: 5'-GGA TTA GGT ATA ACA AAA GGA CCA GAG G-3'; *PF10_0355* F: 5'-GGG GAA AGC ATA TAA TAA TAC TAT AGA TGC-3'; R: 5'-CTT GGA GGA ACA AGA ACC CCC TTA TTA TCA-3'; Radioactivity was measured using a phosphorimager plate and quantified using Quantity One software (version 4.6.5). Halofantrine (HFN) response is listed as sensitive (S) or resistant (R) for each strain.

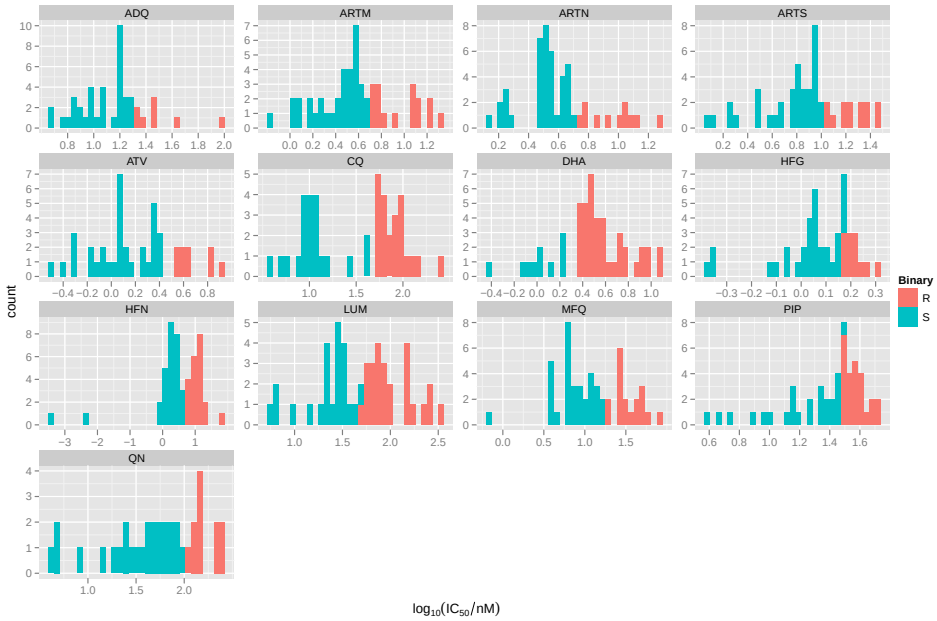


Figure C.13: Drug resistance phenotype classification for sweep and GWAS analyses. IC_{50} data were collected for thirteen antimalarial drugs against all genotyped parasite lines. Quantitative IC_{50}s were converted into binary "sensitive" and "resistant" phenotypes at the cutoffs shown (see also Table C.4). These binary phenotypes were only used for the Haplotype Likelihood Ratio (HLR) test. Drug abbreviations: amodiaquine (ADQ), artemether (ARTM), artesunate (ARTN), artemisinin (ARTS), atovaquone (ATV), chloroquine (CQ), dihydroartemisinin (DHA), halofugi- none (HFG), halofantrine (HFN), lumefantrine (LUM), mefloquine (MFQ), piperaquine (PIP) and quinine (QN).

C.5 Supplemental Tables

Table C.1: 63 parasites used in the study with the name (parasite), geographic origin (region, country), source, and molecular barcode [39], as well as which samples were included in SNP discovery (SEQ), population characterization (POP), long-range haplotype (LRH), and GWAS analyses. For GWAS, * indicates that the sample was used, but not included in any population cluster for stratified or permuted analyses. The human control sample and the ancestral *P. reichenowi* sample were not used in any analyses reported here.

Sample Information					*Used in Analysis*			
Parasite	***Region***	***Country***	***Source***	***Barcode***	***SEQ***	***POP***	***LRH***	***GWAS***
51	America	Brazil	Alejandro Miguel Katzin	CATTGCAGACTXCACCTTAGATTG		x		x
608	America	Brazil	Alejandro Miguel Katzin	TACCCGGGATTACAAACTAGACTT		x		x
10_54	America	Brazil	Alejandro Miguel Katzin	CACTGCAGACTXTACACTAACCTG		x		x
36_89	America	Brazil	Alejandro Miguel Katzin	TACTGCAGATCGCCCCTACGCCTG		x		x
365_89	America	Brazil	Alejandro Miguel Katzin	TGCTCCGGATTACAAACAAGACTT		x		
3D7	Europe	Netherlands	MRA-151	TACTCCGGTCCGCACCCACGATGG		x	x	*
7G8	America	Brazil	MRA-152	TACCCCAGACTXTCAATTAACCTG	x	x		x
9_411	America	Brazil	Alejandro Miguel Katzin	CACTCCAGACTGCAACTACGACTG		x		x
A4	America	Brazil	J. Smith	TATTCCGGTTCATACCCXAGATTG		x	x	
APO41	Africa	Nigeria	Christian Happi	CATTGGGGTCTACACCCAAGACTG		x	x	
CF04.008_12G	Africa	Malawi	Dan Milner	TACCCGGGACCGCCCACAAGATTG		x		x
CF04.008_1F	Africa	Malawi	Dan Milner	TATTCGGAACCGCACCCTAAATTG		x	x	x
CF04.009	Africa	Malawi	Dan Milner	TATTCCGAACCGTACCCTCGATTG		x	x	x
D10	Asia	PNG	MRA-201	CACTCCAGATTGCAACTTAGCTTG	x	x	x	x
D6	Africa	Sierra Leone	MRA-285	TACTGGAAACTGCAACCAAACTTG	x	x	x	x
Dd2	Asia	Indochina/Laos	MRA-156	CATCGCAATTCGCCCCTTAGACTG		x	x	x
FCC2	Asia	China	MRA-733	TACCCCAAATCGCACATTAACCTG	x	x	x	x
GH2	Asia	Thailand	S. Thaitong/D. Kyle	CACTGCGGTTTATCAATTAGCCTG		x	x	x
HB3	America	Honduras	MRA-155	TACTCCAGACTACACACTCACTTG		x		*
IGHCR14	Asia	India	Aditya Dash/Chetan Chitnis	TACCCCAGACCXCACACTAGACGG		x		*
Indochina_I	Asia	Indochina/Laos	MRA-347	TACTCGAGTCTACAACCACGATTG		x	x	x
JST	America	Brazil	Sandra do Lago Moraes	CACCGCGGTTTATAAACAAGATTG		x		x
K1	Asia	Thailand	MRA-159	TATTCGGATTTGTCCCTACGCCTG	x	x	x	x
M24	Africa	Kenya	X. Su	CATTGCGGTTTACCCATAAGCCTG		x	x	x
Malayan Camp	Asia	Malaysia	MRA-330	TATTCCGGATTGTCACTTAGACTG		x	x	*
Muz51.1	Asia	PNG	Karen Day	TACTCCAGATTATCACCTAGCCTG		x	x	x
PR145	Asia	Thailand	S. Thaitong/D. Kyle	CACTCCAGATCACAACCAAAACTG		x	x	x
PS189	Africa	Mali	C. Plowe/Djimde	CACTCCGGATTACAAACAAGCTTT		x	x	
RAJ116	Asia	India	Aditya Dash/Chetan Chitnis	CACTCCGAACTGCAACCACAACGG		x		*
RO33	Africa	Ghana	MRA-200	CACCCGGGATCGCAAACTAAACTT	x	x	x	x
Santa Lucia	America	El Salvador	MRA-362	CACCCGGGATTACAAACAAACCTT	x	x		*
SenP05.02	Africa	Senegal	S. Mboup	CACCCGGGATTACAAACAAGCTTT		x	x	x
SenP08.04	Africa	Senegal	S. Mboup	TACCCCGGATCGCAAACAAACTTT		x	x	x
SenP09.04	Africa	Senegal	S. Mboup	CACTCGGGTTTATACATXCAACGT		x	x	x
SenP11.02	Africa	Senegal	S. Mboup	CGCTCGAGATTACAAACTAGACTT		x	x	x
SenP19.04	Africa	Senegal	S. Mboup	TGTTCCGGTCTACAAATTAACCGT		x	x	x
SenP26.04	Africa	Senegal	S. Mboup	TATCCGAATTTATCAATACAACGT		x	x	x
SenP27.02	Africa	Senegal	S. Mboup	TACTCCGGTTTATACACTCAACGT		x	x	x
SenP31.01	Africa	Senegal	S. Mboup	TACTGCGGTCCGCAAACAAGATTG		x	x	x
SenP51.02	Africa	Senegal	S. Mboup	TATCCGGGACTGCAACTTCGACGG		x	x	x
SenP60.02	Africa	Senegal	S. Mboup	TACTCGAAACCGCAAACTAACCTT		x	x	x
SenT15.04	Africa	Senegal	S. Mboup	TGCTCCAAATCGTACCCAAGCTGT		x	x	x
SenT26.04	Africa	Senegal	S. Mboup	CGCTCGGATTTATCCCTACGCCGT		x	x	x
SenT28.04	Africa	Senegal	S. Mboup	CACCCCAGTTTATCAACAAACTGT		x	x	x
SenV34.04	Africa	Senegal	S. Mboup	TXCTCCAGATTATACCTAAACCTG	x	x	x	x
SenV35.04	Africa	Senegal	S. Mboup	CGCTCGAGTCCGTCCACACACTGT		x	x	x
SenV42.05	Africa	Senegal	S. Mboup	TGTTCCGATCTATCCACAAGACXT		x	x	x
T2_C6	Asia	Thailand	MRA-818	TACTCCGGATTATACACTAAACGT		x	x	
TD203	Asia	Thailand	S. Thaithong/D. Kyle	CGCCCCAGATCATAAATTAAACGT		x	x	x
TD257	Asia	Thailand	S. Thaithong/D. Kyle	TACTCGGGATCGTAACCACACTGT		x	x	x
TM327	Asia	Thailand	S. Thaithong/D. Kyle	TGTCCGAATCTATAAACACAACGT		x	x	x
TM345	Asia	Thailand	S. Thaithong/D. Kyle	CACCCCAGTCTATCACTACACCGT		x	x	
TM90C2A	Asia	Thailand	MRA-202	TATTCCGATTTATCAACTCACCGT		x	x	x
TM90C6A	Asia	Thailand	MRA-205	CACCCCAGTCTATCACTACACCGT		x	x	
TM91C235	Asia	Thailand	MRA-206	CGCTCCGGACTGCACCCAAGATTG		x	x	x
V1/S	Asia	Vietnam	MRA-176	TGCCCCAGATCACAACTAAGATTT	x	x	x	X
WR87	Asia	Vietnam	MRA-284	TACTGGAAATCACAACTAAGACTT		x	x	X
CF04.010	Africa	Malawi	Dan Milner	mixed				
Dd2_HFG_280	Asia	Indochina/Laos						
Human Control								
Preichenowi			John Barnwell					
SenT10.04	Africa	Senegal	S. Mboup	NACTNGGGACTATAACCAAACCTG				
TM93C1088	Asia	Thailand	MRA-207	CACCCCAGTCTATCACTACACCGT				

Table C.2: Analysis of the ability of SNPs on the array to act as a proxy for another. This ability is measured using the standard correlation metric r^2. In our data set, 28% of SNPs in the Brazilian sample (which has the most LD) had a nearby SNP on the array in strong LD ($r^2 > 0.5$) with it, while in the Senegal sample the proportion was only 16%. Most of the time, therefore, we will only be able to detect association with markers that have been directly typed. The exception is strong selective sweeps, which affect many markers within a region.

Fraction of SNPs	$r^2 > 0.3$	$r^2 > 0.5$	$r^2 > 0.8$
Senegal	26%	16%	10%
Thailand	34%	24%	18%
Brazil	33%	28%	24%

Table C.3: Long Range Haplotype (LRH) hits. All REHH hits with Q-value < 0.25.

chr	pos	core	hap len	qvalue	gene	description
2	608790	G	14428	0.2071	PFB0675w	hypothetical
2	617743	T	16930	0.2133	PFB0685c	acyl-CoA synthetase, PfACS9
2	623146	C	22333	0.0323	PFB0687c	RING zinc finger protein
3	466483	A	35321	0.0568	PFC0460w	hypothetical
3	466610	A	35448	0.0133	PFC0460w	hypothetical
4	755220	T	15162	0.1236	PFD0830w	dhfr
4	755243	C	15139	0.2133	PFD0830w	dhfr
4	764100	G	50270	0.1528	PFD0840w	hypothetical
5	1042120	A	18768	0.0075	PFE1250w	acetyl-CoA synthetase, PfACS10
5	1042527	A	18361	0.0085	PFE1250w	acetyl-CoA synthetase, PfACS10
5	1056621	T	14545	0.0591		
5	1159412	C	12844	0.0085	PFE1400c	beta adaptin protein
5	1159501	T	12755	0.0085	PFE1400c	beta adaptin protein
5	1333609	G	5132	0.2341	PFE1640w	PfEMP1, truncated
5	1333639	G	5102	0.2341	PFE1640w	PfEMP1, truncated
5	1333690	T	5051	0.0664	PFE1640w	PfEMP1, truncated
5	1333703	A	5038	0.2341	PFE1640w	PfEMP1, truncated
5	1333716	G	5025	0.2341	PFE1640w	PfEMP1, truncated
5	1333729	T	5012	0.2341	PFE1640w	PfEMP1, truncated
5	1333741	G	5000	0.2341	PFE1640w	PfEMP1, truncated
5	1333790	A	4951	0.2341	PFE1640w	PfEMP1, truncated
6	741192	A	55183	0.2335		
6	741293	C	55082	0.2133	PFF0855c	rifin
6	741366	A	55009	0.2133	PFF0855c	rifin
6	1025852	A	17872	0.2133	PFF1220w	hypothetical
6	1065237	G	33215	0.1492	PFF1280w	hypothetical
6	1098314	C	26112	0.0905	PFF1325c	c3h4-type ring finger protein
6	1114565	C	25158	0.0905	PFF1350c	acetyl-coenzyme a synthetase
6	1114929	G	23554	0.0323	PFF1350c	acetyl-coenzyme a synthetase
6	1115373	A	23998	0.0323	PFF1350c	acetyl-coenzyme a synthetase
6	1115454	C	23938	0.043	PFF1350c	acetyl-coenzyme a synthetase
6	1116047	G	24531	0.0244	PFF1350c	acetyl-coenzyme a synthetase
6	1116171	G	24655	0.043	PFF1350c	acetyl-coenzyme a synthetase
6	1116315	C	24799	0.043	PFF1350c	acetyl-coenzyme a synthetase
6	1117520	G	26004	0.043	PFF1350c	acetyl-coenzyme a synthetase
6	1124426	C	32910	0.0535	PFF1365c	HECT-domain (ubiquitin-transferase)
6	1283916	G	13004	0.1978		
7	428373	C	39799	0.1269	PF07_0027	DNA-directed RNA polymerase 2
7	449953	C	31672	0.0443		
7	459787	T	21838	0.0276	MAL7P1_27	pfcrt
7	460216	G	21409	0.0443	MAL7P1_27	pfcrt
7	461218	T	20407	0.0443	MAL7P1_27	pfcrt
7	465826	G	39654	0.0953	PF07_0035	cg1 protein
7	465826	G	16307	0.158	PF07_0035	cg1 protein
7	467846	G	41674	0.013	PF07_0036	Cg6 protein
7	467846	G	14287	0.019	PF07_0036	Cg6 protein
7	475935	T	46758	0.0664	PF07_0037	Cg2 protein
7	475935	T	26989	0.1009	PF07_0037	Cg2 protein
7	475948	A	46771	0.0664	PF07_0037	Cg2 protein
7	475948	A	26976	0.1009	PF07_0037	Cg2 protein
7	476288	G	47111	0.0591	PF07_0037	Cg2 protein
7	476305	C	47128	0.0443	PF07_0037	Cg2 protein
7	476305	C	26619	0.0443	PF07_0037	Cg2 protein
7	476305	C	47128	0.0443	PF07_0037	Cg2 protein
7	476305	C	26619	0.0443	PF07_0037	Cg2 protein
7	482133	T	10978	0.0133	PF07_0038	Cg7 protein
7	482133	T	20791	0.0085	PF07_0038	Cg7 protein
7	485744	G	14589	0.0133	MAL7P1_28	ribonucleases p/mrp protein subunit
7	485744	G	17180	0.0085	MAL7P1_28	ribonucleases p/mrp protein subunit
7	488164	A	15424	0.0777	MAL7P1_28	ribonucleases p/mrp protein subunit
7	488164	A	14760	0.0535	MAL7P1_28	ribonucleases p/mrp protein subunit
7	490748	C	18008	0.0323	PF07_0040	lysophospholipase-like protein
7	490748	C	14664	0.0244	PF07_0040	lysophospholipase-like protein
7	490877	T	18137	0.0323	PF07_0040	lysophospholipase-like protein
7	490877	T	14535	0.0133	PF07_0040	lysophospholipase-like protein
7	494285	A	21431	0.0551	MAL7P1_29	hypothetical
7	505396	G	17088	0.2133	MAL7P1_30	hypothetical
7	505412	G	17104	0.2133	MAL7P1_30	hypothetical
7	936167	A	15183	0.043	MAL7P1_105	hypothetical
7	940007	G	11940	0.2133	PF07_0085	ferrodoxin reductase-like protein
7	940111	T	12044	0.2133	PF07_0085	ferrodoxin reductase-like protein
7	940147	A	12080	0.2133	PF07_0085	ferrodoxin reductase-like protein
8	336524	T	6205	0.193	MAL8P1_135	hypothetical membrane protein
8	452794	A	778	0.1492	PF08_0105	rifin
8	862485	A	33846	0.1239	PF08_0054	heat shock 70 kDa protein
8	866334	C	29997	0.0443	MAL8P1_64	hypothetical
8	1104023	T	8994	0.0591		
8	1114567	A	8029	0.0873	MAL8P1_23	ubiquitin-protein ligase 1
8	1117372	G	10834	0.0905	MAL8P1_23	ubiquitin-protein ligase 1
8	1118090	T	11552	0.0622	MAL8P1_23	ubiquitin-protein ligase 1
8	1118190	T	11652	0.0091	MAL8P1_23	ubiquitin-protein ligase 1
9	272201	C	12632	0.1259	PFI0265c	RhopH3
9	282410	G	11553	0.0932	PFI0275w	hypothetical
9	284833	T	12260	0.1492	PFI0280c	autophagocytosis associated protein
9	284842	T	12269	0.1492	PFI0280c	autophagocytosis associated protein
9	284910	G	12337	0.1492	PFI0280c	autophagocytosis associated protein
10	324964	G	45666	0.2335	PF10_0078	histone deacetylase, putative
12	50106	T	288	0.0873	PFL0030c	PfEMP1
12	947550	A	63271	0.2335	PFL1130c	hypothetical
12	954384	C	56437	0.225	PFL1130c	hypothetical
12	990296	T	43303	0.0103	PFL1170w	polyadenylate-binding protein
12	1002740	T	55747	0.0568		
12	1002741	A	55748	0.0568		
14	279667	T	34730	0.0959	PF14_0074	hypothetical
14	960714	G	49486	0.0631	PF14_0228	hypothetical
14	1225984	A	8890	0.0443	PF14_0291	hypothetical
14	1226019	T	8925	0.0443	PF14_0291	hypothetical
14	1226103	C	9009	0.1528	PF14_0291	hypothetical
14	1226130	C	9036	0.0443	PF14_0291	hypothetical
14	1226242	A	9148	0.0314	PF14_0291	hypothetical
14	1226303	T	9209	0.0248	PF14_0291	hypothetical
14	1608531	A	54272	0.2133	PF14_0374	hypothetical
14	2812662	C	35145	0.1722	PF14_0653	hypothetical
14	2812679	T	35128	0.1722	PF14_0653	hypothetical
14	2838163	G	46456	0.2215	PF14_0660	hypothetical

Table C.4: IC_{50} drug resistance phenotype data (nM). ND: No data

sample	ADQ	ARTM	ARTN	ARTS	ATV	CQ	DHA	HFG	HFN	LUM	MFQ	PIP	QN
Resistance Threshold	20	5	5	10	3	50	2	1.5	5	50	20	30	100
10_54	16.48	1.551	3.079	2.002	3.335	63.57	0.7305	1.032	1.832	29.22	5.844	30.61	60.14
36_89	16.63	4.431	4.418	8.947	3.327	79.45	5.916	1.139	1.491	21.22	3.865	30.95	ND
3D7	6.8168	2.8094	3.6543	8.4817	2.6474	8.8972	2.5971	0.9846	6.6259	86.2639	20.8618	19.0066	24.5422
51	29.53	3.661	3.688	9.564	4.192	108.2	3.005	ND	2.509	20.59	7.559	41.53	125.9
608	15.16	3.089	3.07	4.351	6.978	93.25	2.784	1.155	1.198	10.09	5.83	10.47	104.9
7G8	18.37	1.464	3.1	2.988	4.161	56.18	2.789	0.7968	1.071	31.1	6.053	40.11	38.87
9_411	ND	1.158	1.66	5.99	ND	ND	0.8257	ND	1.525	5.687	0.6979	54.52	ND
CF04.008_1F	ND	3.718	3.035	6.777	0.8209	6.02	2.522	1.094	3.481	36.68	4.939	18.71	82.44
CF04.009	ND	4.1495	3.771	8.845	ND	ND	1.7316	ND	9.792	72.48	10.356	29.865	ND
D10	15.2461	3.5656	7.2633	8.7383	2.3062	11.6735	4.4617	1.5121	10.5989	92.63	25.4271	38.127	18.1534
D6	14.92	2.076	3.138	6.612	0.6697	3.611	1.333	0.4141	6.201	29.56	6.216	3.999	4.898
Dd2	10.0602	3.4412	3.5525	9.2125	1.2793	73.4253	2.3138	0.8717	4.7277	74.7528	14.9576	26.21	78.8966
FCC2	9.687	3.165	3.202	8.728	1.213	9.47	4.096	1.48	7.74	156.3	35.99	24.36	27.95
GA3	15.11	3.591	3.203	6.281	2.391	102.1	3.674	1.495	6.49	ND	27.22	43.25	31.96
GH2	23.3	6.718	5.983	16.78	2.366	98.03	5.694	1.707	16.03	346.7	17.82	48.54	149.1
HB3	8.703	1.169	3.019	3.117	1.181	9.78	0.9405	1.413	2.764	71.51	11.18	31.22	20.28
IGHCR14	4.879	0.6559	1.416	1.202	0.4847	4.967	0.382	1.451	1.725	53.57	11.51	9.564	4.99
Indochina_I	12	20.57	19.6	ND	ND	243.1	ND	0.8852	4.0E-04	6.178	3.902	28.99	153
JST	30.02	2.35	5.658	3.838	2.193	129.1	1.698	1.203	1.296	22.07	6.811	41.34	24.52
K1	16.07	1.902	3.124	3.176	2.442	86.42	3.509	1.52	1.907	30.84	13.57	35.78	77.81
M24	7.63	4.6	4.43	5.546	1.524	13.09	2.49	1.158	1.919	59.81	13.48	14.23	55.33
Malayan_Camp	7.944	2.887	1.811	1.775	1.171	7.279	2.319	1.126	1.283	ND	6.983	31.26	8.196
Muz51.1	19.64	3.29	3.18	7.449	1.939	60.44	2.271	1.13	1.696	19.5	8.501	28.23	39.73
PR145	11.11	15.79	13.66	29.7	6.59	51.24	12.02	1.106	16	140.8	53.43	31.58	149.1
RAJ116	19.23	1.075	1.805	1.367	1.315	68.59	1.035	1.546	0.0049	6.418	4.105	34.31	4.36
RO33	11.3737	2.6614	6.0825	6.9751	1.8615	11.0702	2.8762	1.6415	3.1416	69.66	8.094	34.0955	13.4901
Santa_Lucia	20.44	3.629	4.61	8.644	0.3903	11.76	3.567	1.171	0.7153	30.34	5.822	38.02	260.1
SenP05.02	16.24	5.695	3.178	11.25	2.493	ND	3.518	1.52	0.9853	30.43	6.798	29	ND
SenP08.04	17.33	4.873	4.15	11.12	1.19	14.85	3.015	0.4386	11.72	62.21	26.03	7.81	44.76
SenP09.04	4.517	5.332	4.177	15.7	0.7054	8.312	5.092	1.404	9.234	174	25.44	14.78	54.86
SenP11.02	92.05	12.75	11.15	19.99	1.881	25.83	8.258	0.7377	17.35	88.75	44.76	14.44	94.63
SenP19.04	7.157	11.66	10.98	24.04	ND	11.4	9.225	1.858	14.71	95.17	50.83	30.21	ND
SenP26.04	44.23	12.71	3.23	14.82	1.143	40.38	8.889	0.4341	20.6	57.79	84.86	5.288	ND
SenP27.02	9.269	3.232	3.635	8.222	ND	9.813	2.664	1.258	1.83	31.2	5.972	21.2	ND
SenP31.01	6.961	1.991	1.999	6.131	0.4997	8.854	2.966	1.214	3.177	78.89	17.03	22.56	ND
SenP51.02	29.62	5.041	4.748	8.875	0.3225	62.05	3.398	1.627	3.319	48.55	8.601	34.89	50.08
SenP60.02	15.8	4.084	4.978	9.62	0.9355	99.93	3.854	1.255	1.616	21.99	11.93	21.19	47.78
SenT15.04	ND	3.595	3.016	6.9315	ND	ND	3.3665	ND	5.262	51.32	3.948	49.53	ND
SenT26.04	16.05	3.149	3.116	4.827	1.19	79.98	2.991	1.43	2.905	30.92	7.877	24.72	121.1
SenT28.04	11.9	3.744	3.139	7.846	1.1	51.01	2.343	1.461	7.5	62.69	32.64	15.63	139.6
SenV34.04	21.39	5.1	4.645	6.781	0.5187	ND	3.024	1.627	2.37	30.34	4.018	28.48	ND
SenV35.04	5.567	3.078	3.68	4.706	0.8165	9.364	5.374	1.359	2.761	45.52	20.6	26.28	ND
SenV42.05	6.456	1.06	1.565	1.717	0.7363	9.594	1.731	1.016	4.476	13.6	10.04	4.485	ND
TD203	15.63	8.15	4.736	20.56	8.49	52.86	8.507	1.071	7.938	156.2	30.32	39.2	67.53
TD257	11.68	11.88	8.532	24.53	2.215	55.41	11.63	1.539	15.96	227.6	50.42	35.53	253.5
TM327	9.891	6.032	1.7	8.862	1.146	41.12	3.851	2.022	15.49	249.3	41.25	50.89	70.72
TM90C2A	18.85	16.17	10.31	28.32	3.704	58.94	5.376	1.086	12.51	264.5	28.04	30.04	209.1
TM91C235	ND	12.43	12.46	17.66	ND	ND	7.409	ND	70.65	154.5	57.96	12.69	ND
V1/S	17.98	1.625	3.459	6.172	4.02	155.5	1.035	1.538	1.739	26.4	12.84	38.58	224.3

Table C.5: Parasites used in the GWAS. Parasites used, indicating their nucleotide and amino acid sequence for various positions (indicated by number) in the *dhfr*, *pfcrt*, and *pfmdr1* gene loci.

	dhfr												***pfcrt***										***pfmdr1***									
	Nucleotide Sequence at AA Position						*Amino Acid Sequence*						*Nucleotide Sequence at AA Position*				*Amino Acid Sequence*						*Nucleotide Sequence at AA Position*					*Amino Acid Sequence*				
Parasite	16	51	59	108	164	506	16	51	59	108	164	506	72-76	326	356	371	72	74	75	76	326	356	371	86	184	1034	1042	1246	86	184	1034	1042	1246
51	C	T	T	A	A	T	A	I	C	N	I	Y	TCTGTAATGAATACA	aac	Tta	aGa	S	M	N	T	N	L	R	A	T	Tgt	Gat	Tat	N	F	C	D	Y
10_54	C	T	T	A	A	T	A	I	C	N	I	Y	TCTGTAATGAATACA	aac	Tta	aGa	S	M	N	T	N	L	R	A	T	Tgt	Gat	Tat	N	F	C	D	Y
36_89	C	T	T	A	A	T	A	I	C	N	I	Y	TCTGTAATGAATACA	aac	Tta	aGa	S	M	N	T	N	L	R	A	T	Tgt	Gat	Tat	N	F	C	D	Y
3D7	C	A	T	G	A	T	A	N	C	S	I	Y	TGTGTAATGAATAAA	aac	ata	aga	C	M	N	K	N	I	R	A	A	agt	aat	gat	N	Y	S	N	D
7G8	C	T	T	A	A	T	A	I	C	N	I	Y	AGTGTAATGAATACA	aac	Tta	aGa	S	M	N	T	N	L	R	A	T	Tgt	Gat	Tat	N	F	C	D	Y
9_411	C	T	T	G	A	T	A	I	C	N	I	Y	TCTGTAATGAATACA	aac	Tta	aGa	S	M	N	T	N	L	R	A	T	Tgt	Gat	Tat	N	F	C	D	Y
A4	C	T	C	A	A	T	A	I	C	T	I	Y	TGTGTAATTGAAACA	aac	Ata	aTa	C	I	E	T	N	I	I	T	A	Agt	Aat	Gat	Y	Y	S	N	D
APO41	C	T	C	A	A	T	A	I	R	N	I	Y	TGTGTAATTGAAACA	aac	Ata	aTa	C	I	E	T	N	I	I	A	T	Agt	Aat	Gat	N	F	S	N	D
CF04.008_12G	C	T	T	A	A	T	A	N	R	N	I	Y	TGTGTAATTGAAACA	aac	Ata	aTa	C	M	N	A	N	I	I	T	A	Agt	Aat	Gat	Y	Y	S	N	D
CF04.008_1F	C	T	C	A	A	T	A	I	R	N	I	Y	TGTGTAATGAATAAA	aac	Ata	aGa	C	M	N	A	N	I	R	T	A	Agt	Aat	Gat	Y	Y	S	N	D
CF04.009	C	T	C	A	A	T	A	I	R	N	I	Y	TGTGTAATGAATAAA	aac	Ata	aGa	C	M	N	A	N	I	R	A	T	Agt	Aat	Gat	N	F	S	N	D
D10	C	A	T	G	A	T	A	N	C	S	I	Y	TGTGTAATGAATAAA	aac	Ata	aGa	C	M	N	K	N	I	R	A	A	Agt	Aat	Gat	N	Y	S	N	D
D6	C	A	T	G	A	T	A	N	C	S	I	Y	TGTGTAATGAATAAA	aac	Ata	aGa	C	M	N	K	N	I	R	A	A	Agt	Aat	Gat	N	Y	S	N	D
Dd2	C	T	C	A	A	T	A	I	R	N	I	Y	TGTGTAATTGAAACA	aac	Ata	aTa	C	I	E	T	N	T	I	T	A	Agt	Aat	Gat	Y	Y	S	N	D
FCC2	C	A	T	A	A	T	A	N	C	N	I	Y	TGTGTAATGAATAAA	aac	Ata	aGa	C	M	N	K	N	I	R	A	A	Agt	Aat	Gat	N	Y	S	N	D
GH2	C	T	C	A	A	T	A	I	R	N	I	Y	TGTGTAATTGAAACA	aac	Ata	aTa	C	I	E	T	N	I	I	A	A	Agt	Aat	Gat	N	Y	S	N	D
HB3	C	A	T	G	A	T	A	N	C	S	I	Y	TGTGTAATGAATAAA	aac	Ata	aGa	C	M	N	K	N	I	R	A	T	Agt	Gat	Gat	N	F	S	D	D
Indochina_I	C	T	C	A	A	T	A	I	R	N	I	Y	TGTGTAATTGAAACA	aac	Ata	aTa	C	I	E	T	N	I	I	A	T	Tgt	Gat	Tat	N	F	C	D	Y
JST	C	T	T	A	A	T	A	N	C	N	I	Y	TCTGTAATGAATACA	aac	Tta	aGa	S	M	N	T	N	L	R	A	A	Tgt	Gat	Tat	N	Y	C	D	Y
K1	C	A	C	A	A	T	A	N	R	N	I	Y	TGTGTAATTGAAACA	aac	Ata	aTa	C	I	E	T	N	I	I	T	A	Agt	Aat	Gat	Y	Y	S	N	D
M24	C	A	T	G	A	T	A	N	C	S	I	Y	TGTGTAATGAATAAA	aac	Ata	aGa	C	M	N	K	N	I	R	A	T	Agt	Aat	Gat	N	F	S	N	D
Malayan Camp	C	A	T	A	A	T	A	N	C	N	I	Y	TGTGTAATGAATAAA	aac	Ata	aGa	C	M	N	K	N	I	R	A	T	Agt	Aat	Gat	N	Y	S	N	D
Muz51.1	C	A	T	G	A	T	A	N	C	S	I	Y	AGTGTAATGAATACA	aac	Tta	aGa	S	M	N	T	N	L	R	T	A	Agt	Aat	Gat	Y	Y	S	N	D
PR145	C	T	C	A	T	T	A	I	R	N	I	Y	GGTGTAATTGAAACA	aac	Ata	aTa	G	I	E	T	N	I	I	A	T	Agt	Aat	Gat	N	F	S	N	D
PS189	C	A	T	G	A	T	A	N	C	S	I	Y	TGTGTAATGAATAAA	aac	Ata	aGa	C	M	N	K	N	I	R	A	T	Agt	Aat	Gat	N	F	S	N	D
RO33	C	A	C	A	A	T	A	N	R	N	I	Y	TGTGTAATGAATAAA	aac	Ata	aGa	C	M	N	K	N	I	R	T	T	Agt	Aat	Gat	Y	F	S	N	D
Santa Lucia	G	A	T	G	A	T	A	N	C	S	I	Y	TGTGTAATGAATAAA	aac	Ata	aGa	C	M	N	T	N	I	R	A	T	Tgt	Gat	Gat	N	F	C	D	D
SenP05.02	C	T	C	A	A	T	A	N	C	N	I	Y	TGTGTAATTGAAACA	aac	Ata	aTa	C	I	E	T	N	I	I	T	T	Agt	Aat	Gat	Y	F	S	N	D
SenP08.04	C	T	C	A	A	T	A	I	R	N	I	Y	TGTGTAATGAATAAA	aac	Ata	aGa	C	M	N	K	N	I	R	A	T	Agt	Aat	Gat	N	F	S	N	D
SenP09.04	C	T	C	A	A	T	A	I	R	N	I	Y	TGTGTAATGAATAAA	aac	Ata	aGa	C	M	N	K	N	I	R	A	A	Agt	Aat	Gat	N	Y	S	N	D
SenP11.02	C	A	T	G	A	T	A	N	C	S	I	Y	TGTGTAATGAATAAA	aac	Ata	aGa	C	M	N	K	N	I	R	A	T	Agt	Aat	Gat	N	F	S	N	D
SenP19.04	C	T	C	A	A	T	A	N	R	N	I	Y	TGTGTAATGAATAAA	aac	Ata	aGa	C	M	N	K	N	I	R	A	A	Agt	Aat	Gat	N	Y	S	N	D
SenP26.04	C	T	C	A	A	T	A	N	R	N	I	Y	TGTGTAATTGAAACA	aac	Ata	aTa	C	I	E	T	N	I	I	T	T	Agt	Aat	Gat	Y	F	S	N	D
SenP27.02	C	T	C	A	A	T	A	I	R	N	I	Y	TGTGTAATGAATAAA	aac	Ata	aGa	C	M	N	K	N	I	R	T	T	Agt	Aat	Gat	Y	F	S	N	D
SenP31.01	C	A	T	G	A	C	A	N	C	S	I	Y	TGTGTAATGAATAAA	aac	Ata	aGa	C	M	N	K	N	I	R	A	T	Agt	Aat	Gat	N	F	S	N	D
SenP51.02	C	T	C	A	A	T	A	I	R	N	I	Y	TGTGTAATTGAAACA	aac	Ata	aTa	C	I	E	T	N	I	I	T	T	Agt	Aat	Gat	Y	F	S	N	D
SenP60.02	C	A	T	G	A	T	A	N	C	S	I	Y	TGTGTAATGAATAAA	aac	Ata	aGa	C	M	N	K	N	I	R	A	T	Agt	Aat	Gat	N	F	S	N	D
SenT15.04	C	T	C	A	A	T	A	I	R	N	I	Y	TGTGTAATTGAAACA	aac	Ata	aTa	C	I	E	T	N	I	I	A	T	Agt	Aat	Gat	N	F	S	N	D
SenT26.04	C	T	C	A	A	T	A	T	R	N	I	Y	TGTGTAATTGAAACA	aac	Ata	aTa	C	I	E	T	N	I	I	T	T	Agt	Aat	Gat	Y	F	S	N	D
SenT28.04	C	T	C	A	A	T	A		R	N	I	Y	TGTGTAATTGAAACA	aac	Ata	aTa	C	I	E	T	N	I	I	A	T	Agt	Aat	Gat	N	F	S	N	D
SenV34.04	C	A	T	G	A	T	A	N	C	S	I	Y	TGTGTAATTGAAACA	aac	Ata	aTa	C	I	E	T	N	I	I	T	T	Agt	Aat	Gat	Y	F	S	N	D
SenV35.04	C	A	T	G	A	T	A	N	R	S	I	Y	TGTGTAATGAATAAA	aac	Ata	aGa	C	M	N	K	N	I	R	A	T	Agt	Aat	Gat	N	F	S	N	D
SenV42.05	C	T	C	A	A	T	A	I	C	N	I	Y	TGTGTAATTGAAACA	aac	Ata	aTa	C	I	E	T	N	I	I	A	A	Agt	Aat	Gat	N	Y	S	N	D
T2_C6	C	A	T	G	A	T	A	N	R	S	I	Y	TGTGTAATGAATAAA	aac	Ata	aGa	C	M	N	K	N	I	R	A	A	Agt	Aat	Gat	N	Y	S	N	D
TD203	C	T	C	A	T	T	A	I	C	N	I	Y	TGTGTAATTGAAACA	aac	Ata	aTa	C	I	E	T	N	I	I	A	T	Agt	Aat	Gat	N	F	S	N	D
TD257	C	T	C	A	T	T	A	I	C	N	I	Y	TGTGTAATTGAAACA	aac	Ata	aTa	C	I	E	T	N	I	I	A	T	Agt	Aat	Gat	N	F	S	N	D
TM327	C	T	C	A	T	T	A	I	C	N	I	Y	TGTGTAATTGAAACA	aac	Ata	aTa	C	I	E	T	N	I	I	A	T	Agt	Aat	Gat	N	F	S	N	D
TM345	C	T	C	A	T	T	A	I	C	N	I	Y	TGTGTAATTGAAACA	aac	Ata	aTa	C	I	E	T	N	I	I	A	T	Agt	Aat	Gat	N	F	S	N	D
TM90C2A	C	T	C	A	T	T	A	I	C	N	I	Y	TGTGTAATTGAAACA	aac	Ata	aTa	C	I	E	T	N	I	I	A	T	Agt	Aat	Gat	N	F	S	N	D
TM90C6A	C	T	C	A	T	T	A	I	C	N	I	Y	TGTGTAATTGAAACA	aac	Ata	aTa	C	I	E	T	N	I	I	A	A	Agt	Aat	Gat	N	Y	S	N	D
TM91C235	C	T	C	A	T	T	A	I	C	N	I	Y	TGTGTAATTGAAACA	aac	Ata	aTa	C	I	E	T	N	I	I	A	T	Agt	Aat	Gat	N	F	S	N	D
V1/S	C	T	C	A	T	T	A	I	C	N	I	Y	TGTGTAATTGAAACA	aac	Ata	aTa	C	I	E	T	N	I	I	T	A	Agt	Aat	Gat	Y	Y	S	N	D

Table C.6: *PF10_0355* copy number summary for 38 parasites tested by qPCR using the Delta Delta Ct method. Copy number (CN) was compared to the reference locus *PF07_0076* and 3D7 was used as a reference strain. A cut-off of 1.4 was used to define *PF10_0355* copy number greater than 1; parasites with greater than 1 copy of *PF10_0355* are shaded. Parasites are ranked by Halofantrine (HFN) IC_{50}: HFN-sensitive parasites are indicated by an S and HFN-resistant parasites are indicated by an R.

Parasite	CN	HFN
Indochina_I	0.92	S
RAJ116	0.92	S
Santa_Lucia	1.03	S
SenP05.02	0.98	S
7G8	0.74	S
Malayan_Camp	0.94	S
JST	1.06	S
36_89	1.07	S
SenP60.02	1.19	S
Muz51.1	0.64	S
IGHCR14	1.63	S
V1/S	0.89	S
10_54	0.99	S
K1	0.87	S
M24	0.76	S
SenV34.04	4.92	S
51	0.77	S
SenV35.04	0.86	S
HB3	0.88	S
RO33	1.26	S
SenP31.01	0.77	S
SenP51.02	1.09	S
CF04.008_1F	0.95	S
SenV42.05	1.26	S
Dd2	1.07	S
SenT15.04	1.71	R
3D7	1	R
SenT28.04	0.73	R
FCC2	0.76	R
TD203	0.9	R
D10	1.06	R
SenP08.04	0.95	R
TM90C2A	1.43	R
SenP19.04	1.06	R
PR145	0.94	R
GH2	1.71	R
SenP11.02	7.14	R
SenP26.04	1.68	R

Table C.7: Annotation and GeneID Information for identified genes in Figure 2.1B.

GeneID	π	F_{ST}	Annotation	Category	Tag
MAL8P1_23	1.54E-04	0.646	ubiquitin-protein ligase 1, putative	enzymes, ACS and transporters	UBQ Ligase
PF13_0201	6.39E-03	0.216	thrombospondin-related anonymous protein, TRAP	other	TRAP
PFA0650w	4.37E-03	0.323	surface-associated interspersed gene pseudogene, (SURFIN) pseudogene	antigens, var, rifin, stevor, surfin	SURFIN
PF08_0105	6.10E-03	0.204	rifin	antigens, var, rifin, stevor, surfin	Rifin
PFB0960c	4.31E-03	0.036	P. falciparum Maurer's Cleft 2 transmembrane domain protein 2.1, PfMC-2TM_2.1	other	Mauer's Cleft
MAL7P1_27	6.36E-04	0.387	chloroquine resistance transporter	enzymes, ACS and transporters	PFCRT
PF10_0345	6.52E-03	0.240	merozoite surface protein 3	antigens, var, rifin, stevor, surfin	MSP3
PFI1475w	1.95E-03	0.221	merozoite surface protein 1, precursor	antigens, var, rifin, stevor, surfin	MSP1
PFB0972w	9.90E-03	0.077	hypothetical protein	other	*
PFL0030c	7.95E-03	0.050	erythrocyte membrane protein 1 (PfEMP1)	antigens, var, rifin, stevor, surfin	Var2CSA
PFD0830w	5.96E-04	0.459	bifunctional dihydrofolate reductase-thymidylate synthase	enzymes, ACS and transporters	DHFR
PF11_0344	6.46E-03	0.074	apical membrane antigen 1, AMA1	antigens, var, rifin, stevor, surfin	AMA1
PF10_0051	5.32E-03	0.215	ADP/ATP carrier protein, putative	enzymes, ACS and transporters	ADP/ATP Carrier
PFB0685c	5.75E-04	0.497	acyl-CoA synthetase, PfACS9	enzymes, ACS and transporters	ACS9
PFF1350c	2.00E-03	0.584	acetyl-coenzyme a synthetase	enzymes, ACS and transporters	ACS
PFE1250w	1.66E-03	0.602	acyl-CoA synthetase, PfACS10	enzymes, ACS and transporters	ACS10

C.6 Figures and Tables Supporting Supplemental Methods

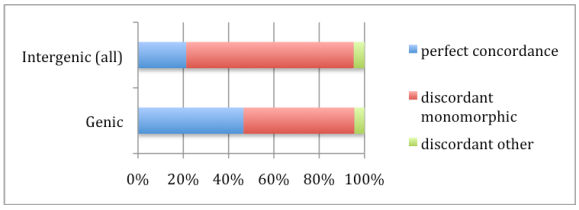


Figure C.14: SNPs located in genes (28,576) were more likely to pass concordance filtering than intergenic SNPs (19,999).

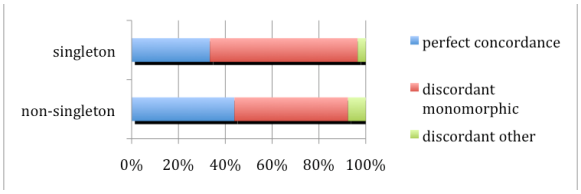


Figure C.15: SNPs that were discovered from only one sequenced strain (35,727 SNPs) show a higher rate of monomorphism on the array than those with higher minor allele counts (12,848 SNPs). Any amount of this discordance may be explained by false discovery from sequencing data.

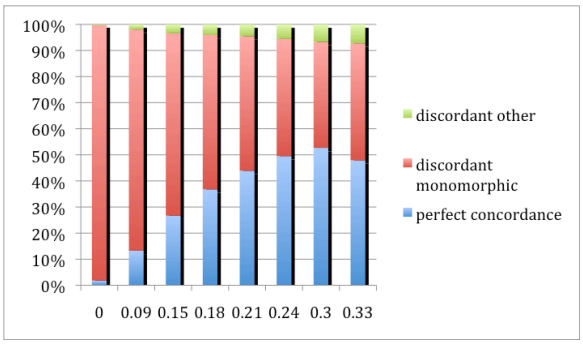


Figure C.16: To show the effect of GC composition, we took 16bp of flanking sequence on each side of the SNP to construct a 3D7-based 33mer and calculated the percent GC. The window boundaries for the graph below are chosen as the octiles of the GC distribution. SNP performance appears to worsen at GC levels below 20%, which accounts for roughly half of the SNPs.

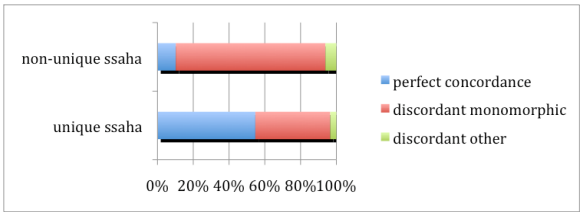


Figure C.17: The effect of unique sequences in flanking regions. Although the initial design of the array excluded probes that had exact matches elsewhere in the genome, many of the remaining SNPs are in neighborhoods that contain 1 or 2 base mismatch similarity to other parts of the genome. We took 16bp of flanking sequence on each side of the SNP to construct a 3D7-based 33mer and aligned it to 3D7 using SSAHA (word length 10, step length 1, max gap 2, max insert 1, min hits 24) [119]. 28,352 SNPs aligned uniquely to their location of origin. The 20,223 SNPs that aligned in multiple locations showed a much higher rate of discordance.

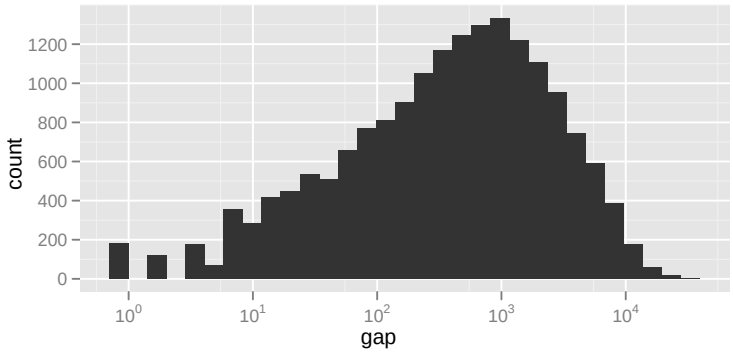


Figure C.18: A histogram of marker spacing. Most markers are spaced closely with a few large gaps in coverage. The mean spacing of concordant markers is 1316bp with a median spacing of 444bp.

Table C.8: Statistics on marker spacing (gaps) by chromosome. The table gives the number of gaps, as well as the median, mean gaps, 90% percentile (90% gap), and maximum gap length (max gap) by chromosome (1 – 14).

	number of gaps	median gap	mean gap	90% gap	max gap	sum of gaps
all	17568	444	1316	3639	39535	23118100
chr 1	644	150	957	2968	18798	616104
chr 2	855	346	1103	3090	16236	942754
chr 3	825	485	1282	3300	24076	1057642
chr 4	1462	182	817	2042	39535	1195109
chr 5	977	447	1344	3865	17583	1312630
chr 6	1088	422	1298	3801	24856	1412604
chr 7	1599	225	935	2446	19341	1495349
chr 8	1100	454	1284	3319	24320	1412004
chr 9	1087	557	1415	3820	27661	1538216
chr 10	1113	533	1520	4121	34097	1691290
chr 11	1357	636	1497	4115	21921	2031460
chr 12	1678	463	1351	3737	17010	2266590
chr 13	1831	669	1562	4333	19966	2860436
chr 14	1952	768	1683	4516	23942	3285912

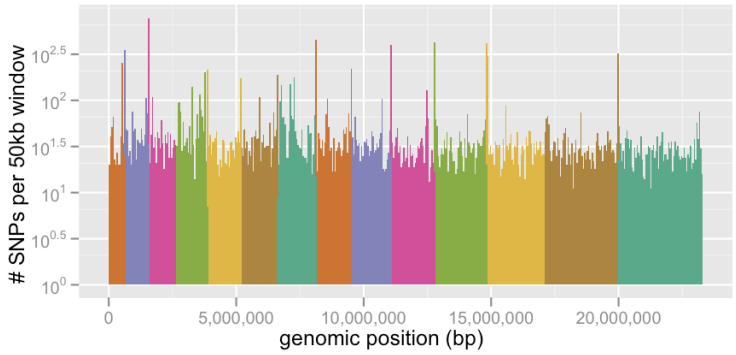


Figure C.19: Final concordant marker density across all fourteen chromosomes.

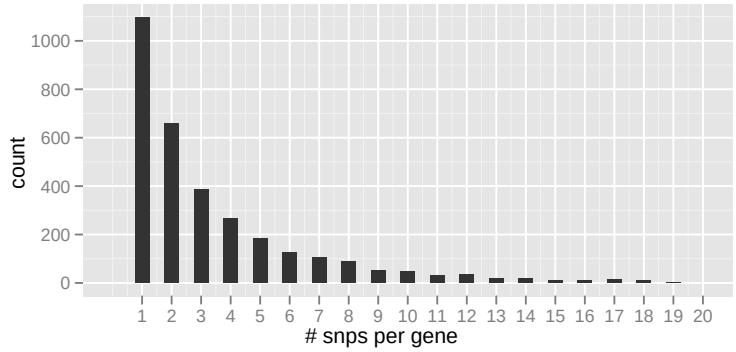


Figure C.20: Distribution of markers per gene: 59% of all *P. falciparum* genes contain at least one concordant marker. Below is a plot of the number of markers per gene. Most genes that have markers have only one or two (median marker count per gene is two). Mean marker count per gene is 4.1.

D

Supplemental Material for Chapter 3

D.1 Author Contributions to Supplemental Material

Most of the work in these Supplemental Materials (as in the main text of Chapter 3) are my own, though other authors made contributions in certain spots.

SKV and UR are jointly responsible for most of the text in Supplemental Results which delves into the functional characterization and categorization of hits. AKL performed the work described in Drug Assays.

I performed all XP-EHH and localization analyses described in Section D.4.2. I produced all Supplemental Data Files and Supplemental Figures. Dataset 1 also contains a number of code fragments that are also my work.

D.2 Supplemental Data Files

1. PLINK-formatted input data files describing genotype and phenotype data for sequence and array data. Recombination map and imputed genotype data for XP-EHH. Outputs from EMMA and XP-EHH. R code for GWAS figures. Permanent URL (ZIP, 57MB):
   ftp://ftp.broadinstitute.org/pub/malaria/pnas-park-2012-suppfile-1.zip

2. Consensus sequence calls for each of 45 strains and 23 million bases. VCF file is bgzip compressed and indexed by tabix and vcftools. Permanent URL (VCF.GZ, 2GB):
   ftp://ftp.broadinstitute.org/pub/malaria/pnas-park-2012-suppfile-2.vcf.gz

D.3 Supplemental Results

We identified 32 regions under selection using the XP-EHH test and identified top candidate mutations within each region associated with drug resistance using the EMMA test (Dataset 1). Of the 163 mutations, 48 (29%) were intergenic; 6 (4%) were intronic; 33 (20%) conferred synonymous changes; and 76 (47%) conferred non-synonymous changes. We evaluated the annotations for the 59 genes on this list using gene and protein prediction algorithms through PlasmoDB.org [14] and associated links, including predicted GO function, pathway, inter-pro domains or user comments, combined with any published literature for each gene. The great majority of the 59 genes (65%) can be collectively classified into the following categories: surface molecules or transporters (11/59 or 19%, of which 6/59 or 10% are transporters) including *pfcrt*; molecules involved in genome maintenance or transcriptional regulation (9/59 or 15%); metabolic enzymes (12/59 or 20%, of which 3/59 or 5% mediate lipid metabolism) including *dhfr*; and molecules involved in ubiquitination (6/59 or 10%). Remaining genes were determined as mediators of various other cellular functions including protein binding, invasion, and gamete fertilization (12/59 or 20%) or unclassified (9/59 or 15%).

We also analyzed all 35 genes within the chromosome 6 region (between position 1,117,269 and 1,390,662) found to be under selection in pyrimethamine-resistant parasites. This region contained a large stretch of intergenic mutations and it was difficult to localize the signal to any one gene. It contains a large number of metabolic genes (12/35 or 34%, of which 3/35 or 9% participate in lipid metabolism; 2/35 or 6% mediate folate metabolism); chaperones and genes involved in ubiquitination (5/35 or 14%); with additional genes classified as genome maintenance or transcription regulation (8/35 or 23%); surface molecules or transporters (3/35 or 9%); other biological functions including structural proteins (2/35 or 6%); and the remainder (5/35 or 14%) as unclassified.

Molecules implicated in the ubiquitination cascade were mainly associated with resistance to pyrimethamine and include a putative E2 conjugating enzyme (PFL2100w), which likely acts as a ubiquitin E2 variant (UEV) due to the lack of a catalytic cysteine and a HECT E3 ubiquitin ligase (MAL8P1.23). Within the chromosome 6 region there were several other molecules proposed to modulate ubiquitination including a HECT E3 (PFF1365c) and a Cullin-like E3 (PFF1445c). Two other molecules contain domains suggestive of a possible role in ubiquitination, including PF08_0080 that contains a PUB domain found in proteins linked to the ubiquitin proteasome system [3], and PFF1485w, which contains an ubiquitin interacting motif. Also in this region are two putative chaperones, including a protein containing a Dna J domain (PFF1415c) associated with heat shock molecules [38] and a TRP (PFF1505w) involved in RNA degradation [41] with a proposed chaperone function. Finally, there is a putative RING E3 (PFD0765w) in a region of selection associated with primaquine sensitivity.

Several genes putatively involved in lipid metabolism were identified in our regions of drug associated-selection, including an acyl-CoA synthetase, PfACS8 (PFB0695c) [18] and a putative phosphopantothenoylcysteine synthetase

HECT—homologous to the E6-AP carboxyl terminus

PUB—peptide:N-glycanase/UBA or UBX

TRP—tetratricopeptide repeat protein

(PFD0610w, under selection in quinine-resistant parasites) proposed to be involved in CoA biosynthesis. Other lipid-metabolism associated molecules in the chromosome 6 region include an acetyl-CoA synthetase (PFF1350c) [138]; an ethanolaminephosphotransferase (PFF1375c-a/b) [88, 172]; and a phosphatidylcholine-sterol acyltransferase precursor (PFF1420w). Finally, the PFD0350w gene, predicted to play a role in isoprenoid biosynthesis [77], is in a region under selection in artemisinin-sensitive parasites.

Folate pathway molecules in regions of selection specific to pyrimethamine-resistant parasites include the *dhfr* locus (PFD0830w); PF14_0487 (aminomethyl-transferase); as well as PFF1360w (6-pyruvolytetrahydropterin synthetase; and PFF1490w (methenyltetrahydrofolate activity) found within the chromosome 6 region. Folate metabolism has been shown to be a target of pyrimethamine resistance mechanisms, and specifically SNP changes in *dhfr* and *dhps*, as well as copy number variants in *gch1* [83] are associated with anti-folate resistance [110].

There are three ABC transporters among the gene lists including PF10_0049, MAL8P1.97, and PF08_0078, which are intriguing since these molecules have been shown to modulate drug responses in malaria (e.g. *pfmdr1*) and other organisms [89]. Finally, we believe that there are a large number of molecules from among the genome maintenance or transcriptional regulation classification may be candidates for drug modulation through changes in gene expression [103], chromatin or histone structure [33, 37], or RNA binding [99].

D.4 Supplemental Methods

D.4.1 Drug Assays

Drug assays were performed as described [132] with slight modifications for 384-well format. Synchronized ring-stage parasites were cultured in the presence of serial dilutions of test compounds in 40μL of RPMI supplemented with AlbuMAX II (Life Technologies 1021-045) at 1.0% hematocrit and an initial parasitemia of 1.0% in black clear-bottom plates (Greiner Bio-one 781090). Following a 72 hour incubation under standard culture conditions, SYBR Green I dye

(Invitrogen S7563) was added to a dilution of 1:5000 and plates were stored at room temperature until the flourescence signal was read on a Spectramax M5 plate reader (Molecular Devices, ex 480nm, em 530nm). Raw flourescence data were analyzed using the Prism v5.0 software package (GraphPad Software, Inc.). After background subraction and normalization, IC_{50} values were determined based on application of a nonlinear regression log(inhibitor)-response curve fit.

D.4.2 XP-EHH

Selection-association tests were run using the cross population extended haplotype homozygosity test (XP-EHH) [151]. Replicate IC_{50} data was geometrically averaged (equivalently, $\log_{10}(IC_{50})$ data was arithmetically averaged) and then converted to binary phenotypes ("sensitive" vs. "resistant") according to cutoffs shown in Figure D.1 and Dataset 1. For drugs with a bimodal distribution, binary cutoffs were chosen at positions that clearly separated the sensitive and resistant populations. For drugs with a more unimodal distribution, cutoffs were manually placed at a distribution minimum near the median IC_{50}, since the XP-EHH test, like many other tests, loses power when either of the two populations becomes too small (when the cutoff is too far from the median). Although these may not represent samples that are especially sensitive or resistant in the traditional sense, it is common in studies of quantitative phenotypes to simply compare the upper part of a distribution against the lower part for binary tests [15].

The recombination map was constructed with LDhat v2.1 [97] using a block penalty of 5.0, 10 million rjMCMC iterations, a missing data cutoff of 20%, minimum minor allele frequency of 8%, and otherwise default parameters. Since the XP-EHH test does not tolerate missing data, SNPs with data in at least 80% of individuals were imputed with PHASE 2.1.1 [156]. As PHASE requires "diploid" data, we dropped the sample with the lowest call rate (SenP60.02) to create an even number of haploid individuals, randomly paired together. 83,540 fully-imputed SNPs were polymorphic among the remaining 44 individuals. We then filtered out singleton SNPs and used only 29,605 SNPs that had at least two samples with a minor allele (minor allele frequency of 4%).

The XP-EHH test calculates haplotype decay separately for the resistant population and sensitive population using the extended haplotype homozygosity statistic (EHH). It then integrates these values with respect to genetic distance and computes a log ratio of these areas for the resistant population over the sensitive population. These log ratios, after normalization, are called XP-EHH scores or *Z*-scores, as they are found to correspond to a normal distribution, with the exception of the tails that diverge from the null expectation (Fig. D.4B). Other than its application to a phenotypically-divided population (instead of a geographically-divided population), the original algorithms were used as published [151] without any modifications. Significantly positive *Z*-scores are indicative of positive selection amongst resistant parasites. Negative scores are indicative of selection in sensitive parasites. We used a two-sided conversion of *Z*-scores to *P*-values, but generally focused our attention on positive *Z*-scores. It would be equally valid to do a one-sided, left- or right-tailed conversion for studies that are interested in specific selection scenarios.

We attempt to localize the signal in these regions by searching for the strongest EMMA signal in that window for that phenotype. We use this SNP to suggest a causal gene for the region (Dataset 1). We do not require significance from the EMMA test, as the region has already been identified as genome-wide significant by the XP-EHH test. We do not combine the results from these or any other statistics.

D.5 Supplemental Figures

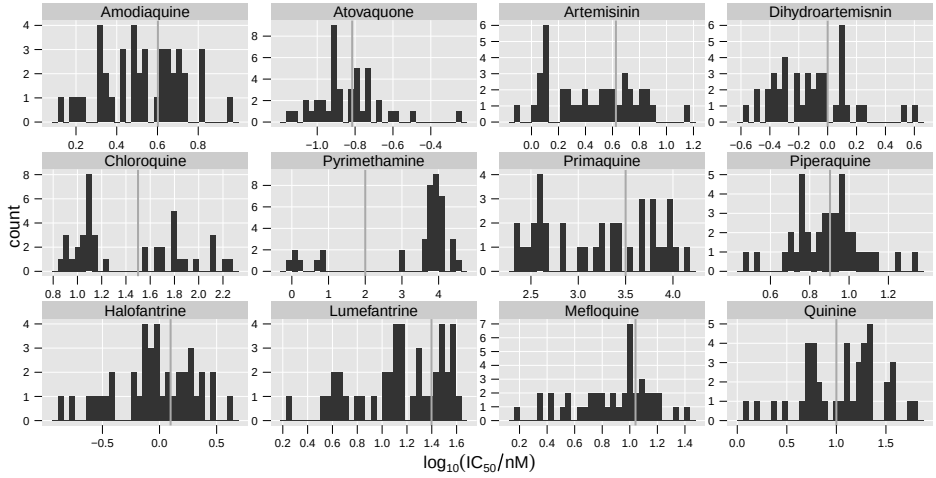


Figure D.1: Drug response distributions for the twelve drugs used in this study: histograms of $\log_{10}(IC_{50})$ across 45 strains. Dark gray lines indicate binary cutoff values used for the XP-EHH test.

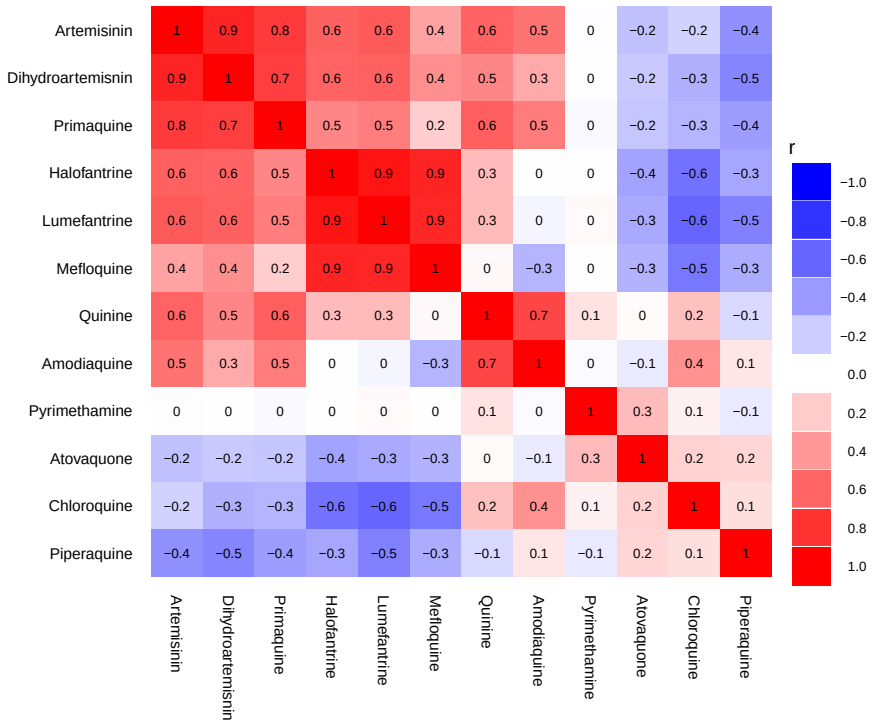


Figure D.2: Drug response correlation heat map for the twelve drugs used in this study. Pearson correlations are rendered for each pair of drugs, based on $\log_{10}(IC_{50})$ values for each strain.

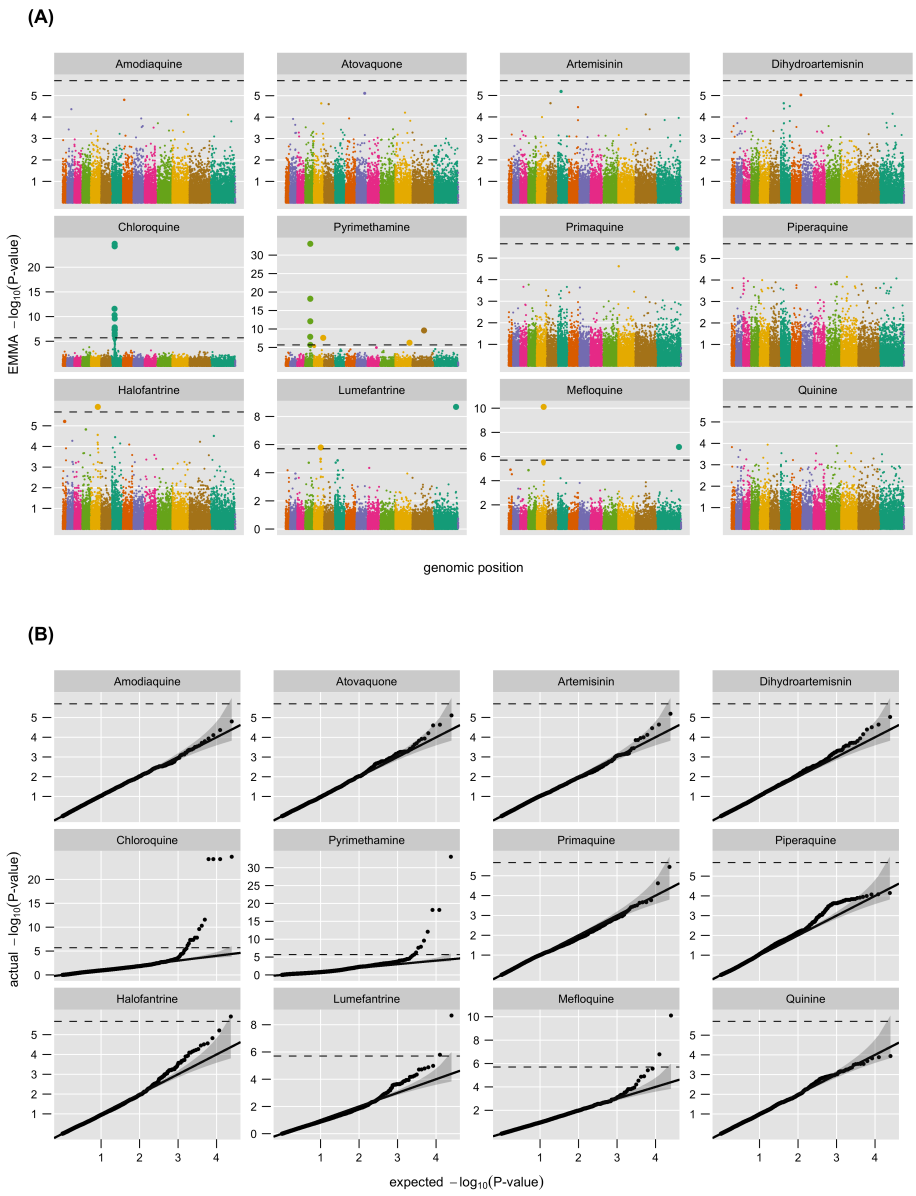


Figure D.3: EMMA GWAS plots (sequence data, 45 samples): **(A)** EMMA Manhattan plots of $-\log_{10}(P)$ against genomic position. **(B)** EMMA P-P plots against the expected uniform distribution. Dashed line indicates Bonferroni-corrected significance of 5%. The shaded area indicates a 95% confidence interval around the null.

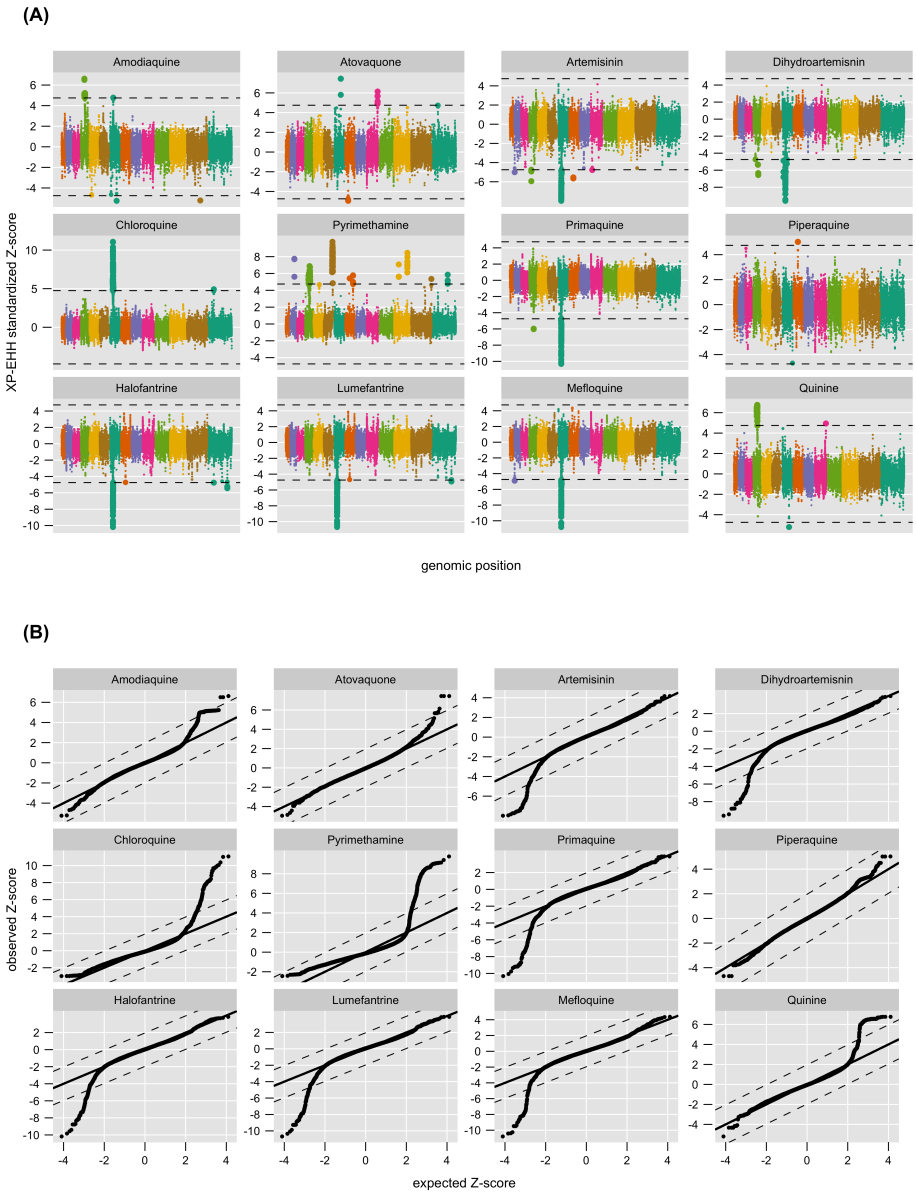


Figure D.4: XP-EHH GWAS plots (sequence data, 45 samples): **(A)** XP-EHH Manhattan-like plots of Z-scores against genomic position. **(B)** XP-EHH Q-Q plots of XP-EHH Z-scores against the expected normal distribution. Dashed lines indicate a 95% confidence interval around the null.

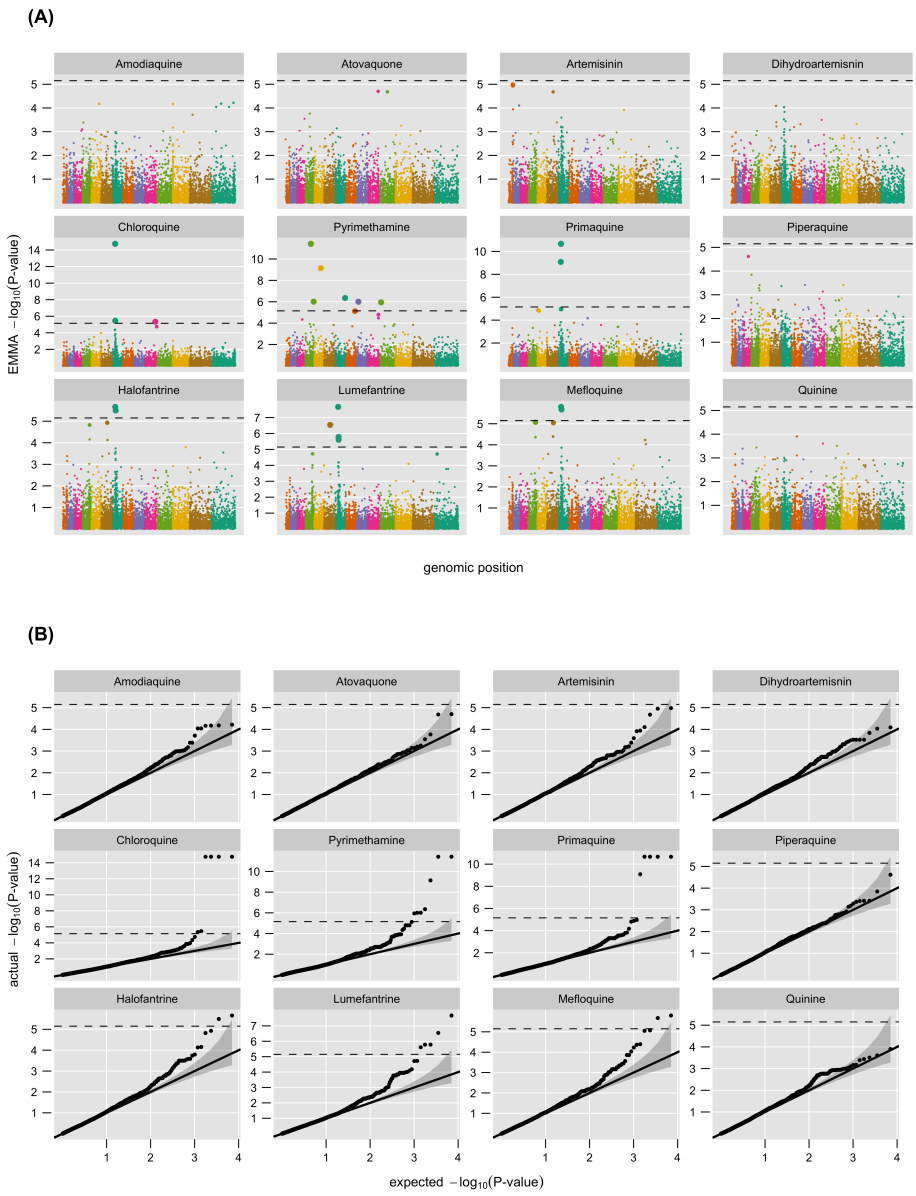


Figure D.5: EMMA GWAS plots (array data, 24 samples): **(A)** EMMA Manhattan plots of $-\log_{10}(P)$ against genomic position. **(B)** EMMA P-P plot against the expected uniform distribution. Dashed line indicates Bonferroni-corrected significance of 5%. The shaded area indicates a 95% confidence interval around the null.

E

Supplemental Material for Chapter 4

E.1 Author Contributions to Supplemental Material

I performed all analyses shown in this appendix and created all figures. *Ex vivo* drug data for GWAS was generated by DVT.

E.2 Supplemental Figures

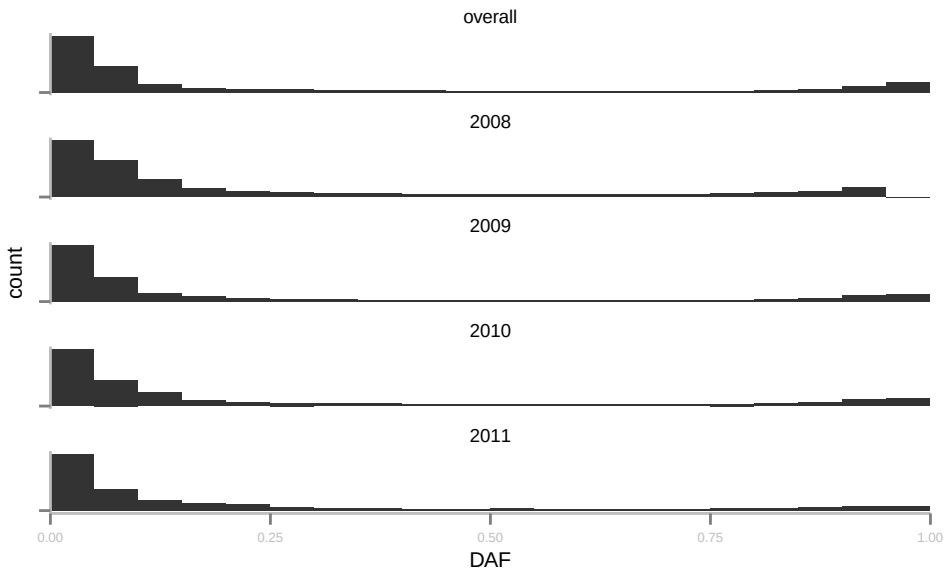


Figure E.1: Derived allele frequency (DAF) spectra over all samples and by year.

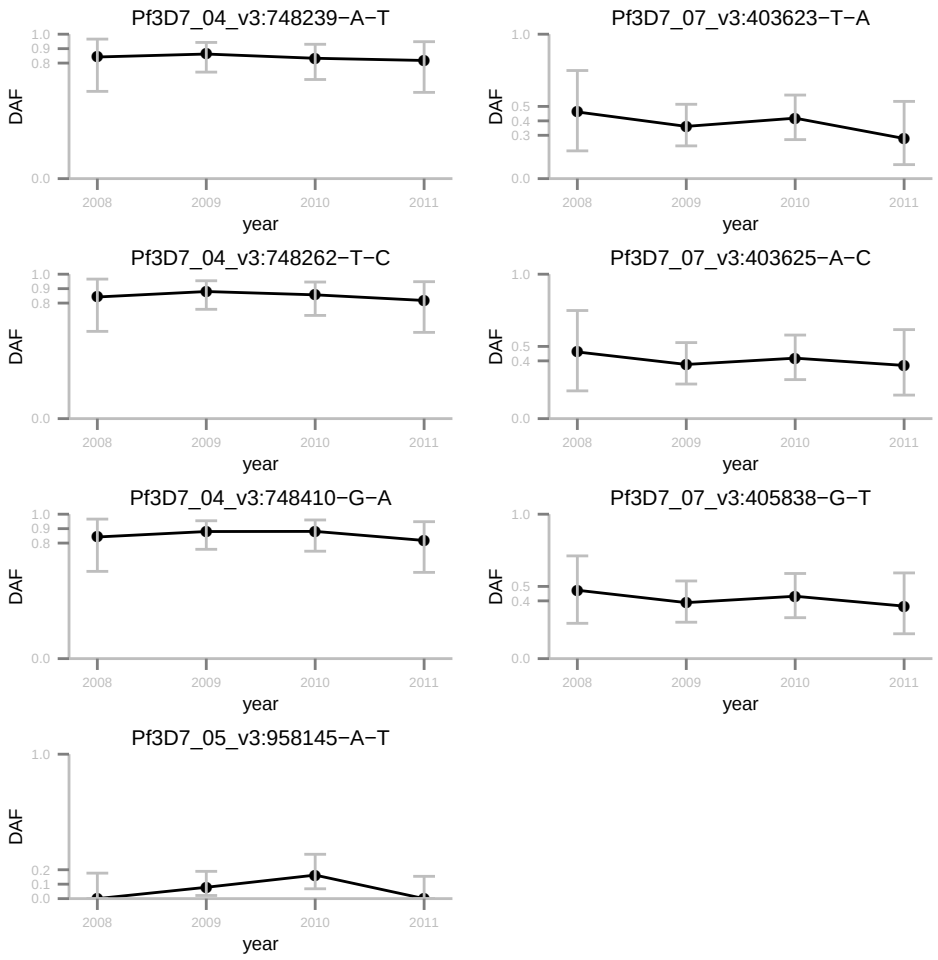


Figure E.2: DAF over time for major drug loci. This shows the estimated derived allele frequency at each year (black line) for seven known drug loci. The first three loci on chromosome 4 at left are *dhfr* mutations N51I, C59R, and S108N. The last locus at left on chromosome 5 is *pfmdr1* N86Y. The three loci at right on chromosome 7 are *pfcrt* mutations N75K, K76T, and R371I. 95% confidence intervals are drawn in gray for each estimate of the DAF based on binomial sampling error. None of these loci show significant movements over this time period.

E.3 Null models of the selection coefficient

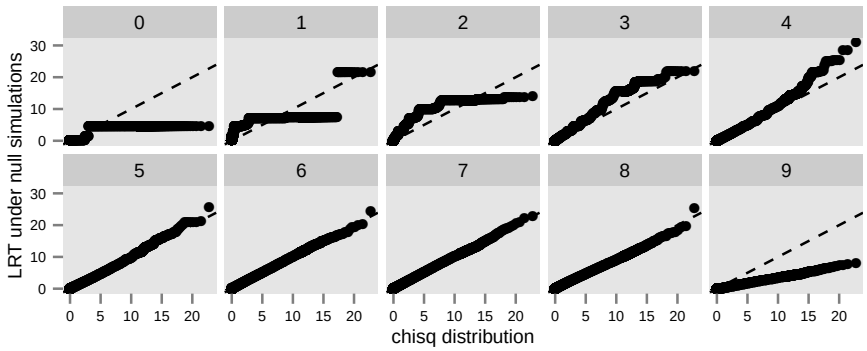


Figure E.3: HMM statistics are biased at extreme derived allele frequencies. This shows null simulations of the LRT statistic against the expected χ^2 distribution. The data is plotted separately for each of ten bins corresponding to deciles of the derived allele frequency distribution shown in Figure E.1. The first three deciles (corresponding to $DAF < 0.0365$) show significant departures from χ^2. Most of the intermediate frequencies fit well, but the final decile ($DAF > 0.927$) shows departures again. This suggests that the model has difficulty with extreme allele frequencies.

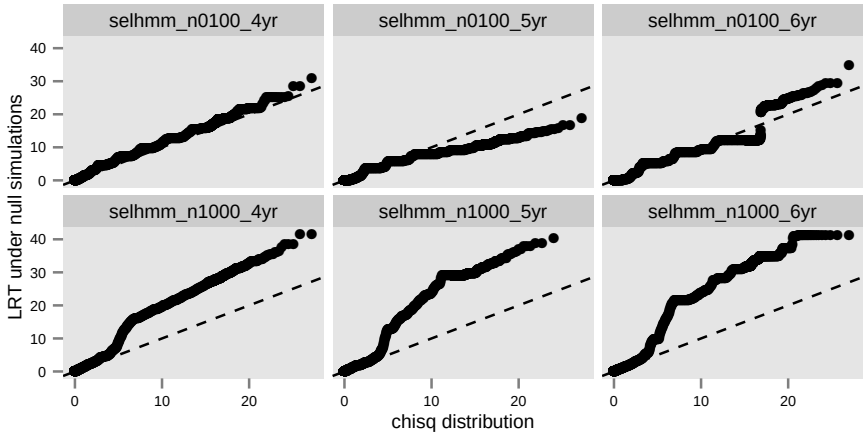


Figure E.4: HMM statistics at different drift strengths and time spans. This shows null simulations of the LRT statistic against the expected χ^2 distribution. Data is plotted separately for two values of N_e (100 on top, 1000 on bottom) and three thresholds for minimum number of samples per year, resulting in 4, 5, or 6 years of data. The four year data set comprises the years 2008-2011. The five year data set adds 2004. The six year data set adds 2002.

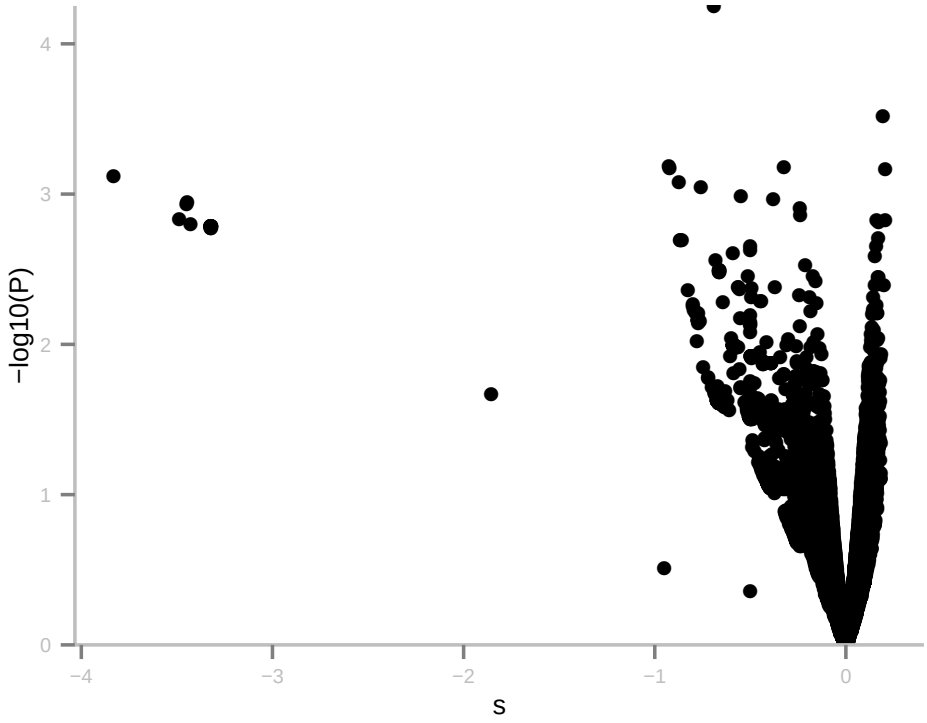


Figure E.5: Volcano plot. This visualization of statistical significance ($-\log_{10} P$) vs. effect size (s) is often used in RNA expression studies and allows one to visually prioritize selection strength, given a threshold for significance.

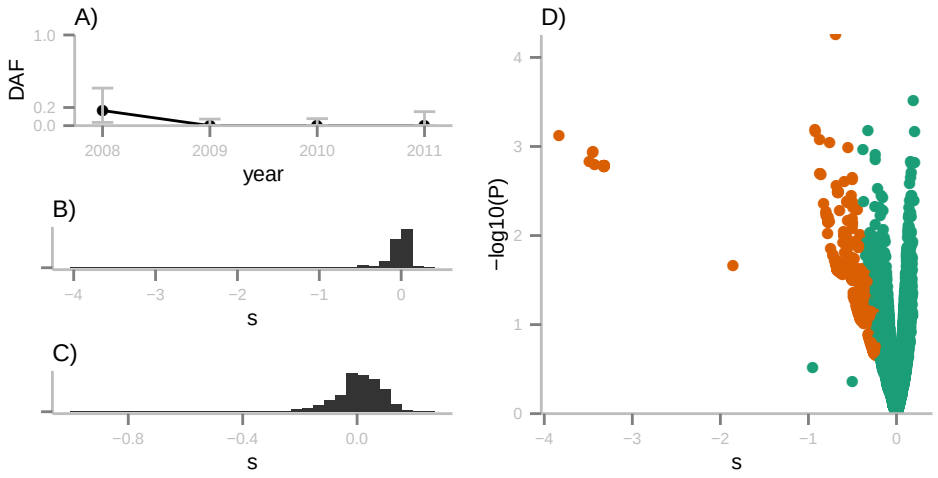


Figure E.6: Double-zero variants lead to inflated statistics. Many of the variants exhibiting strong negative selection are due to loci that have time series data similar to the example shown in (A). Small observations are seen in the first year, followed by complete absence in the final two years. This is interpreted by the HMM as evidence for very strong negative selection, but may simply be due to inaccurate estimates by the HMM at the edges of allele frequency space. The distribution of s including all variants (B) is reduced to a significantly smaller range when removing these "double zero" variants (C). A volcano plot colored by double zero status (D) illustrates that nearly all of the most extreme negative values are due to this artifact.

E.4 Long-haplotype tests

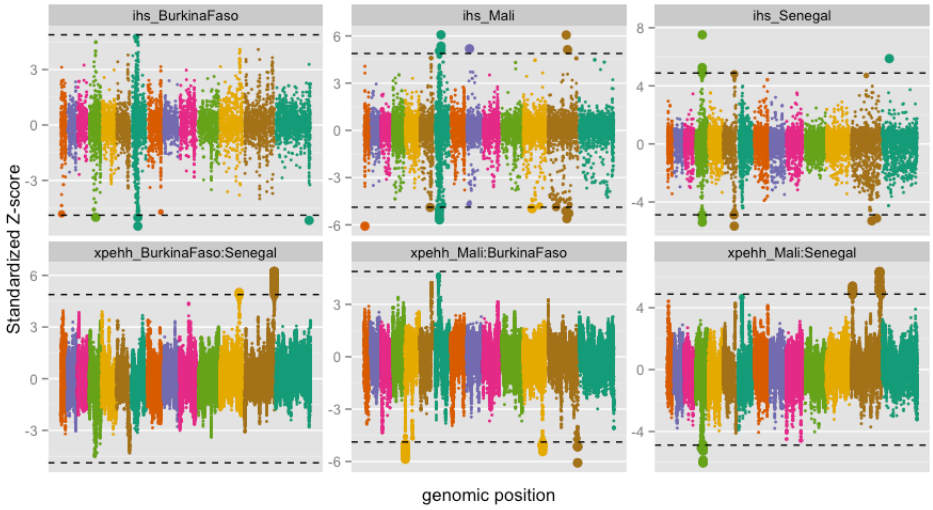


Figure E.7: Long-haplotype selection tests in three West African populations. The top three panels depict signals from the iHS long haplotype selection test [173] in Burkina Faso, Mali (both from Manske et al. [95]) and Senegal (described here). Positive values denote selection for the ancestral allele, negative values denote selection for the derived allele. The bottom three panels depict signals from the XP-EHH long haplotype differentiation test [151] between the three populations. Positive values denote selection in the first named sample, negative values denote selection in the second named sample.

E.5 Ex vivo GWAS tests

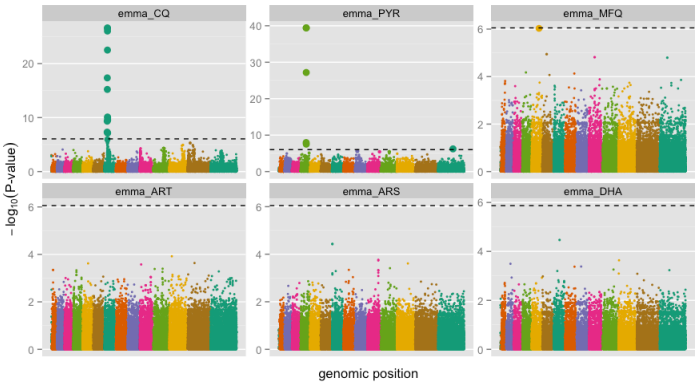


Figure E.8: EMMA results for GWAS on 125 Senegalese samples for resistance to chloroquine, pyrimethamine, mefloquine, artemisinin, artesunate, and dihydroartemisinin. Tests show strong signals at previously known loci at *pfcrt* (chloroquine), *dhfr* (pyrimethamine), and *pfmdr1* (mefloquine), with no significant signals from any of the artemisnin-related drugs.

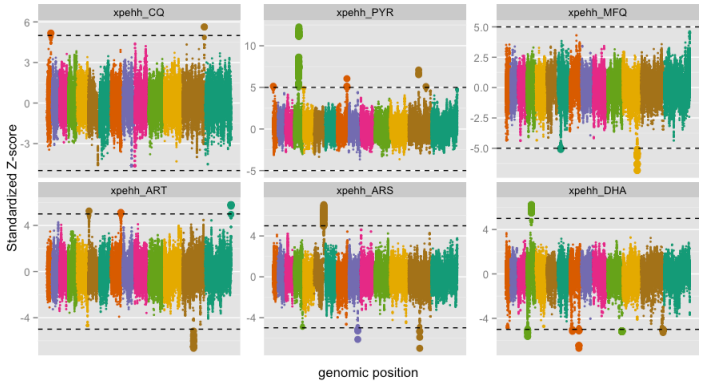


Figure E.9: XP-EHH GWAS results for drug resistance-associated positive selection. Drugs tested include chloroquine, pyrimethamine, mefloquine, artemisinin, artesunate, and dihydroartemisinin. Positive signals indicate positive selection in drug resistant parasites. Negative signals indicate positive selection in drug sensitive parasites.

References

[1] Affymetrix. BRLMM-P: a genotype calling method for the SNP 5.0 array. Technical report, Affymetrix, February 2007. URL http://www.affymetrix.com/support/technical/whitepapers/brlmmp_whitepaper.pdf.

[2] Anders Albrechtsen, Finn Cilius Nielsen, and Rasmus Nielsen. Ascertainment biases in snp chips affect measures of population divergence. *Molecular Biology and Evolution*, 27(11):2534–47, Nov 2010. doi: 10.1093/molbev/msq148. URL http://mbe.oxfordjournals.org/content/27/11/2534.long.

[3] Mark D Allen, Alexander Buchberger, and Mark Bycroft. The pub domain functions as a p97 binding module in human peptide n-glycanase. *Journal of Biological Chemistry*, 281(35):25502–8, Sep 2006. doi: 10.1074/jbc.M601173200. URL http://www.jbc.org/content/281/35/25502.long.

[4] David M Altshuler, V J Pollara, C R Cowles, W J Van Etten, Jennifer Baldwin, L Linton, and Eric S Lander. An SNP map of the human genome generated by reduced representation shotgun sequencing. *Nature*, 407 (6803):513–516, September 2000. doi: 10.1038/35035083. URL http://www.nature.com/nature/journal/v407/n6803/full/407513a0.html.

[5] David M Altshuler, Mark J Daly, and Eric S Lander. Genetic mapping in human disease. *Science*, 322(5903):881–8, Nov 2008. doi: 10.1126/science.1156409. URL http://www.sciencemag.org/cgi/content/full/322/5903/881.

[6] Alfred Amambua-Ngwa, Daniel J Park, Sarah K Volkman, Kayla G Barnes, Amy Bei, Amanda K Lukens, Papa Sene, Daria Van Tyne, Daouda Ndiaye, Dyann F Wirth, David J Conway, Daniel E Neafsey, and Stephen F Schaffner. SNP genotyping identifies new signatures of selection in a deep

sample of West African P. falciparum malaria parasites. *Molecular Biology and Evolution*, 29:3249–3253, June 2012. doi: 10.1093/molbev/mss151. URL http://mbe.oxfordjournals.org/content/29/11/3249.

[7] Alfred Amambua-Ngwa, Kevin K A Tetteh, Magnus Manske, Natalia Gomez-Escobar, Lindsay B Stewart, M Elizabeth Deerhake, Ian H Cheeseman, Christopher I Newbold, Anthony A Holder, Ellen Knuepfer, Omar Janha, Muminatou Jallow, Susana Campino, Bronwyn MacInnis, Dominic P Kwiatkowski, and David J Conway. Population genomic scan for candidate signatures of balancing selection to guide antigen characterization in malaria parasites. *PLoS Genetics*, 8(11): e1002992–e1002992, November 2012. doi: 10.1371/journal.pgen.1002992. URL http://www.plosgenetics.org/article/info%3Adoi%2F10.1371%2Fjournal.pgen.1002992.

[8] Tim J C Anderson, B Haubold, J T Williams, J G Estrada-Franco, L Richardson, R Mollinedo, M Bockarie, J Mokili, S Mharakurwa, N French, J Whitworth, I D Velez, A H Brockman, Francois Nosten, M U Ferreira, and K P Day. Microsatellite markers reveal a spectrum of population structures in the malaria parasite plasmodium falciparum. *Molecular Biology and Evolution*, 17(10):1467–82, Oct 2000. URL http://mbe.oxfordjournals.org/cgi/content/full/17/10/1467.

[9] Tim J C Anderson, Jigar J Patel, and Michael T Ferdig. Gene copy number and malaria biology. *Trends in Parasitology*, 25(7):336–43, Jul 2009. doi: 10.1016/j.pt.2009.04.005. URL http://www.sciencedirect.com/science/article/pii/S1471492209001184.

[10] Tim J C Anderson, Shalini Nair, Standwell Nkhoma, Jeff T Williams, Mallika Imwong, Poravuth Yi, Duong Socheat, Debashish Das, Kesinee Chotivanich, Nicholas P J Day, Nicholas J White, and Arjen M Dondorp. High heritability of malaria parasite clearance rate indicates a genetic basis for artemisinin resistance in western cambodia. *Journal of Infectious Diseases*, 201(9):1326–30, May 2010. doi: 10.1086/651562. URL http://www.journals.uchicago.edu/doi/abs/10.1086/651562.

[11] Tim J C Anderson, Jeff T Williams, Shalini Nair, Daniel Sudimack, Marion Barends, Anchalee Jaidee, Ric N Price, and Francois Nosten. Inferred relatedness and heritability in malaria parasites. *Proc. R. Soc. B*, 277 (1693):2531–40, Aug 2010. doi: 10.1098/rspb.2010.0196. URL http://rspb.royalsocietypublishing.org/content/277/1693/2531.long.

[12] Tim T Anderson, Standwell S Nkhoma, Andrea A Ecker, and David D Fidock. How can we identify parasite genes that underlie antimalarial drug resistance? *Pharmacogenomics*, 12(1):59–85, January 2011. doi: 10.2217/pgs.10.165. URL http://www.futuremedicine.com/doi/abs/10.2217/pgs.10.165.

[13] Sarah Auburn, Susana Campino, Olivo Miotto, Abdoulaye A Djimde, Issaka Zongo, Magnus Manske, Gareth Maslen, Valentina Mangano, Daniel Alcock, Bronwyn MacInnis, Kirk A Rockett, Taane G Clark, Ogobara K Doumbo, Jean Bosco Ouédraogo, and Dominic P Kwiatkowski. Characterization of within-host Plasmodium falciparum diversity using next-generation sequence data. *PLoS ONE*, 7(2):e32891, 2012. doi: 10.1371/journal.pone.0032891. URL http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0032891.

[14] Cristina Aurrecoechea, John Brestelli, Brian P Brunk, Jennifer Dommer, Steve Fischer, Bindu Gajria, Xin Gao, Alan Gingle, Greg Grant, Omar S Harb, Mark Heiges, Frank Innamorato, John Iodice, Jessica C Kissinger, Eileen Kraemer, Wei Li, John A Miller, Vishal Nayak, Cary Pennington, Deborah F Pinney, David S Roos, Chris Ross, Christian J Stoeckert, Charles Treatman, and Haiming Wang. Plasmodb: a functional genomic database for malaria parasites. *Nucleic Acids Res*, 37(Database issue): D539–43, Jan 2009. doi: 10.1093/nar/gkn814. URL http://nar.oxfordjournals.org/content/37/suppl_1/D539.long.

[15] Silviu-Alin Bacanu, Matthew R Nelson, and John C. Whittaker. Comparison of methods and sampling designs to test for association between rare variants and quantitative traits. *Genetic Epidemiology*, 35(4): 226–235, 2011. doi: 10.1002/gepi.20570. URL http://onlinelibrary.wiley.com/doi/10.1002/gepi.20570/abstract.

[16] Mary Lynn Baniecki, Dyann F Wirth, and Jon Clardy. High-throughput plasmodium falciparum growth assay for malaria drug discovery. *Antimicrobial Agents and Chemotherapy*, 51(2):716–23, Feb 2007. doi: 10.1128/AAC.01144-06. URL http://aac.asm.org/cgi/content/full/51/2/716?view=long&pmid=17116676.

[17] Sara E Beese, Takahiro Negishi, and David E Levin. Identification of positive regulators of the yeast fps1 glycerol channel. *PLoS Genetics*, 5 (11):e1000738–e1000738, November 2009. doi: 10.1371/journal.pgen.1000738. URL http://www.plosgenetics.org/article/info%3Adoi%2F10.1371%2Fjournal.pgen.1000738.

[18] Lara L Bethke, Martine Zilversmit, Kaare Nielsen, Johanna Daily, Sarah K Volkman, Daouda Ndiaye, Elena R Lozovsky, Daniel L Hartl, and Dyann F Wirth. Duplication, gene conversion, and genetic diversity in the species-specific acyl-coa synthetase gene family of plasmodium falciparum. *Molecular and Biochemical Parasitology*, 150(1):10–24, Nov 2006. doi: 10.1016/j.molbiopara.2006.06.004. URL http://www.sciencedirect.com/science/article/pii/S0166685106001770.

[19] Jonathan P Bollback, Thomas L York, and Rasmus Nielsen. Estimation of 2Nes from temporal allele frequency data. *Genetics*, 179(1):497–502, May 2008. doi: 10.1534/genetics.107.085019. URL http://www.genetics.org/content/179/1/497.

[20] Olivier Bouchaud, Patrick Imbert, Jean Touze, Alex NO Dodoo, Martin Danis, and Fabrice Legros. Fatal cardiotoxicity related to halofantrine: a review based on a worldwide safety data base. *Malaria Journal*, 8: 289–289, January 2009. doi: 10.1186/1475-2875-8-289. URL http://www.malariajournal.com/content/8/1/289.

[21] Kate M Broadbent, Daniel J Park, Ashley R Wolf, Daria Van Tyne, Jennifer S Sims, Ulf Ribacke, Sarah Volkman, Manoj Duraisingh, Dyann F Wirth, Pardis C Sabeti, and John L Rinn. A global transcriptional analysis of Plasmodium falciparum malaria reveals a novel family of telomere-associated lncRNAs. *Genome Biology*, 12(6):R56, June 2011. doi: 10.1186/gb-2011-12-6-r56. URL http://genomebiology.com/2011/12/6/R56.

[22] Jane M Carlton, Samuel V Angiuoli, Bernard B Suh, Taco W Kooij, Mihaela Pertea, Joana C Silva, Maria D Ermolaeva, Jonathan E Allen, Jeremy D Selengut, Hean L Koo, Jeremy D Peterson, Mihai Pop, Daniel S Kosack, Martin F Shumway, Shelby L Bidwell, Shamira J Shallom, Susan E van Aken, Steven B Riedmuller, Tamara V Feldblyum, Jennifer K Cho, John Quackenbush, Martha Sedegah, Azadeh Shoaibi, Leda M Cummings, Laurence Florens, John R Yates, J Dale Raine, Robert E Sinden, Michael A Harris, Deirdre A Cunningham, Peter R Preiser, Lawrence W Bergman, Akhil B Vaidya, Leo H van Lin, Chris J Janse, Andrew P Waters, Hamilton O Smith, Owen R White, Steven L Salzberg, J Craig Venter, Claire M Fraser, Stephen L Hoffman, Malcolm J Gardner, and Daniel J Carucci. Genome sequence and comparative analysis of the model rodent malaria parasite Plasmodium yoelii yoelii. *Nature*, 419 (6906):512–519, October 2002. doi: 10.1038/nature01099. URL

http://www.nature.com/nature/journal/v419/n6906/full/nature01099.html.

[23] Jane M Carlton, John H Adams, Joana C Silva, Shelby L Bidwell, Hernan Lorenzi, Elisabet Caler, Jonathan Crabtree, Samuel V Angiuoli, Emilio F Merino, Paolo Amedeo, Qin Cheng, Richard M R Coulson, Brendan S Crabb, Hernando A Del Portillo, Kobby Essien, Tamara V Feldblyum, Carmen Fernandez-Becerra, Paul R Gilson, Amy H Gueye, Xiang Guo, Simon Kang'a, Taco W A Kooij, Michael Korsinczky, Esmeralda V-S Meyer, Vish Nene, Ian Paulsen, Owen White, Stuart A Ralph, Qinghu Ren, Tobias J Sargeant, Steven L Salzberg, Christian J Stoeckert, Steven A Sullivan, Marcio M Yamamoto, Stephen L Hoffman, Jennifer R Wortman, Malcolm J Gardner, Mary R Galinski, John W Barnwell, and Claire M Fraser-Liggett. Comparative genomics of the neglected human malaria parasite Plasmodium vivax. *Nature*, 455(7214):757–763, October 2008. doi: 10.1038/nature07327. URL http://www.nature.com/nature/journal/v455/n7214/full/nature07327.html.

[24] Céline Karine Carret, Paul Horrocks, Bernard Konfortov, Elizabeth A Winzeler, Matloob Qureshi, Chris Newbold, and Alasdair Ivens. Microarray-based comparative genomic analyses of the human malaria parasite plasmodium falciparum using affymetrix arrays. *Molecular and Biochemical Parasitology*, 144(2):177–86, Dec 2005. doi: 10.1016/j.molbiopara.2005.08.010. URL http://dx.doi.org/10.1016/j.molbiopara.2005.08.010.

[25] Meryl A Castellini, Jeffrey S Buguliskis, Louis J Casta, Charles E Butz, Alan B Clark, Thomas A Kunkel, and Theodore F Taraschi. Malaria drug resistance is associated with defective dna mismatch repair. *Molecular and Biochemical Parasitology*, 177(2):143–7, Jun 2011. doi: 10.1016/j.molbiopara.2011.02.004. URL http://www.sciencedirect.com/science/article/pii/S0166685111000697.

[26] Hsiao-Han Chang. Whole-genome analysis of polymorphism in plasmodium falciparum. Graduate Thesis Proposal, Harvard OEB, 2010.

[27] Hsiao-Han Chang, Eli L Moss, Daniel J Park, Daouda Ndiaye, Soulyemane Mboup, Roger C Wiegand, Sarah K Volkman, Pardis C Sabeti, Dyann F Wirth, Daniel E Neafsey, and Daniel L Hartl. The malaria life cycle intensifies both natural selection and random genetic drift. *Proceedings of the National Academy of Sciences, USA*. Manuscript in preparation.

[28] Hsiao-Han Chang, Daniel J Park, Kevin J Galinsky, Stephen F Schaffner, Daouda Ndiaye, Omar Ndir, Soulyemane Mboup, Roger C Wiegand, Sarah K Volkman, Pardis C Sabeti, Dyann F Wirth, Daniel E Neafsey, and Daniel L Hartl. Genomic sequencing of Plasmodium falciparum malaria parasites from Senegal reveals the demographic history of the population. *Molecular Biology and Evolution*, 29:3427–3439, June 2012. doi: 10.1093/molbev/mss161. URL http://mbe.oxfordjournals.org/content/29/11/3427.

[29] Marina Chavchich, Lucia Gerena, Jennifer Peters, Nanhua Chen, Qin Cheng, and Dennis E Kyle. Role of pfmdr1 amplification and expression in induction of resistance to artemisinin derivatives in Plasmodium falciparum. *Antimicrobial Agents and Chemotherapy*, 54(6):2455–2464, June 2010. doi: 10.1128/AAC.00947-09. URL http://aac.asm.org/content/54/6/2455.

[30] Ian H Cheeseman, Natalia Gomez-Escobar, Celine K Carret, Alasdair Ivens, Lindsay B Stewart, Kevin K A Tetteh, and David J Conway. Gene copy number variation throughout the Plasmodium falciparum genome. *BMC Genomics*, 10:353, 2009. doi: 10.1186/1471-2164-10-353. URL http://www.biomedcentral.com/1471-2164/10/353.

[31] Ian H Cheeseman, Becky A Miller, Shalini Nair, Standwell Nkhoma, Asako Tan, John C Tan, Salma Al Saai, Aung Pyae Phyo, Carit Ler Moo, Khin Maung Lwin, Rose McGready, Elizabeth Ashley, Mallika Imwong, Kasia Stepniewska, Poravuth Yi, Arjen M Dondorp, Mayfong Mayxay, Paul N Newton, Nicholas J White, François Nosten, Michael T Ferdig, and Timothy J C Anderson. A major genome region underlying artemisinin resistance in malaria. *Science*, 336(6077):79–82, April 2012. doi: 10.1126/science.1215966. URL http://www.sciencemag.org/content/336/6077/79.full.html.

[32] Andrew G Clark, Melissa J Hubisz, Carlos D Bustamante, Scott H Williamson, and Rasmus Nielsen. Ascertainment bias in studies of human genome-wide polymorphism. *Genome Research*, 15(11):1496–502, Nov 2005. doi: 10.1101/gr.4107905. URL http://genome.cshlp.org/content/15/11/1496.long.

[33] Bradley I Coleman and Manoj T Duraisingh. Transcriptional control and gene silencing in plasmodium falciparum. *Cell Microbio*, 10(10):1935–46, Oct 2008. doi: 10.1111/j.1462-5822.2008.01203.x. URL http://www3.interscience.wiley.com/journal/120750891/abstract.

[34] D J Conway, C Roper, A M Oduola, D E Arnot, P G Kremsner, M P Grobusch, C F Curtis, and B M Greenwood. High recombination rate in natural populations of Plasmodium falciparum. *Proceedings of the National Academy of Sciences, USA*, 96(8):4506–4511, April 1999. doi: 10.1073/pnas.96.8.4506. URL http://www.pnas.org/content/96/8/4506.full.

[35] David J Conway, R L Machado, B Singh, P Dessert, Z S Mikes, M M Povoa, A M Oduola, and C Roper. Extreme geographical fixation of variation in the Plasmodium falciparum gamete surface protein gene Pfs48/45 compared with microsatellite loci. *Molecular and Biochemical Parasitology*, 115(2):145–156, July 2001. doi: 10.1016/S0166-6851(01)00278-X. URL http://www.sciencedirect.com/science/article/pii/S016668510100278X.

[36] B S Crabb, T Triglia, J G Waterkeyn, and Alan F Cowman. Stable transgene expression in Plasmodium falciparum. *Molecular and Biochemical Parasitology*, 90(1):131–144, December 1997. doi: 10.1016/S0166-6851(97)00143-6. URL http://www.sciencedirect.com/science/article/pii/S0166685197001436.

[37] Liwang Cui and Jun Miao. Chromatin-mediated epigenetic regulation in the malaria parasite plasmodium falciparum. *Eukaryotic Cell*, 9(8): 1138–49, Aug 2010. doi: 10.1128/EC.00036-10. URL http://ec.asm.org/content/9/8/1138.long.

[38] D M Cyr, T Langer, and M G Douglas. Dnaj-like proteins: molecular chaperones and specific regulators of hsp70. *Trends in Biochemical Sciences*, 19(4):176–81, Apr 1994. doi: 10.1016/0968-0004(94)90281-X. URL http://dx.doi.org/10.1016/0968-0004(94)90281-X.

[39] Rachel Daniels, Sarah K Volkman, Danny A Milner, Nira Mahesh, Daniel E Neafsey, Daniel J Park, David Rosen, Elaine Angelino, Pardis C Sabeti, Dyann F Wirth, and Roger C Wiegand. A general snp-based molecular barcode for plasmodium falciparum identification and tracking. *Malaria Journal*, 7:223, Oct 2008. doi: 10.1186/1475-2875-7-223. URL http://www.malariajournal.com/content/7/1/223.

[40] Rachel F Daniels, Hsiao-Han Chang, Papa Diogoye Séne, Daniel J Park, Daniel E Neafsey, Stephen F Schaffner, Elizabeth J Hamilton, Amanda K Lukens, Daria Van Tyne, Souleymane Mboup, Pardis C Sabeti, Daouda Ndiaye, Dyann F Wirth, Daniel L Hartl, and Sarah K Volkman. Genetic

Surveillance Detects Both Clonal and Epidemic Transmission of Malaria following Enhanced Intervention in Senegal. *PLoS ONE*, 8(4): e60780–e60780, April 2013. doi: 10.1371/journal.pone.0060780. URL http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0060780.

[41] A K Das, P W Cohen, and D Barford. The structure of the tetratricopeptide repeats of protein phosphatase 5: implications for tpr-mediated protein-protein interactions. *The EMBO Journal*, 17(5): 1192–9, Mar 1998. doi: 10.1093/emboj/17.5.1192. URL http://www.nature.com/emboj/journal/v17/n5/full/7590834a.html.

[42] Paul I W de Bakker, Roman Yelensky, Itsik Pe'er, Stacey B Gabriel, Mark J Daly, and David M Altshuler. Efficiency and power in genetic association studies. *Nature Genetics*, 37(11):1217–23, Nov 2005. doi: 10.1038/ng1669. URL http://www.nature.com/ng/journal/v37/n11/full/ng1669.html.

[43] Awa B Deme, Amy K Bei, Ousmane Sarr, Daniel E Neafsey, Stephen F Schaffner, Daniel J Park, Rachel F Daniels, Aida Sadikh Badiane, Papa El Hadji Omar Gueye, Ambroise Ahouidi, Daouda Ndiaye, Souleymane Mboup, Dyann F Wirth, and Sarah K Volkman. Analysis of the pfhrp2 genetic diversity in senegal and implications for rapid diagnostic test use. *Malaria Journal*. Manuscript in preparation.

[44] B Devlin and K Roeder. Genomic control for association studies. *Biometrics*, 55(4):997–1004, December 1999. doi: 10.1111/j.0006-341X.1999.00997.x. URL http://dx.doi.org/10.1111/j.0006-341X.1999.00997.x.

[45] Neekesh V Dharia, A Sidhu, M Cassera, Scott Westenberger, S Bopp, R Eastman, David Plouffe, S Batalov, Daniel J Park, Sarah K Volkman, Dyann F Wirth, Y Zhou, D Fidock, and Elizabeth A Winzeler. Use of high-density tiling microarrays to identify mutations globally and elucidate mechanisms of drug resistance in plasmodium falciparum. *Genome Biology*, 10(2):R21, Feb 2009. doi: 10.1186/gb-2009-10-2-r21. URL http://genomebiology.com/2009/10/2/R21.

[46] Neekesh V Dharia, David Plouffe, Selina E R Bopp, Gonzalo E González-Páez, Carmen Lucas, Carola Salas, Valeria Soberon, Badry Bursulaya, Tadeusz J Kochel, David J Bacon, and Elizabeth A Winzeler. Genome scanning of Amazonian Plasmodium falciparum shows subtelomeric

instability and clindamycin-resistant parasites. *Genome Research*, 20(11): 1534–1544, November 2010. doi: 10.1101/gr.105163.110. URL http://genome.cshlp.org/content/20/11/1534.long.

[47] Arjen M Dondorp, François Nosten, Poravuth Yi, Debashish Das, Aung Phae Phyo, Joel Tarning, Khin Maung Lwin, Frederic Ariey, Warunee Hanpithakpong, Sue J Lee, Pascal Ringwald, Kamolrat Silamut, Mallika Imwong, Kesinee Chotivanich, Pharath Lim, Trent Herdman, Sen Sam An, Shunmay Yeung, Pratap Singhasivanon, Nicholas P J Day, Niklas Lindegardh, Duong Socheat, and Nicholas J White. Artemisinin resistance in Plasmodium falciparum malaria. *New England Journal of Medicine*, 361(5):455–467, July 2009. doi: 10.1056/NEJMoa0808859. URL http://www.nejm.org/doi/full/10.1056/NEJMoa0808859.

[48] Carolyn K Dong, Sameer Urgaonkar, Joseph F Cortese, Francisco-Javier Gamo, Jose F Garcia-Bustos, Maria J Lafuente, Vishal Patel, Leila Ross, Bradley I Coleman, Emily R Derbyshire, Clary B Clish, Adelfa E Serrano, Mandy Cromwell, Robert H Barker, Jeffrey D Dvorin, Manoj T Duraisingh, Dyann F Wirth, Jon Clardy, and Ralph Mazitschek. Identification and validation of tetracyclic benzothiazepines as plasmodium falciparum cytochrome bc1 inhibitors. *Chemistry & Biology*, 18(12):1602–1610, December 2011. doi: 10.1016/j.chembiol.2011.09.016. URL http://dx.doi.org/10.1016/j.chembiol.2011.09.016.

[49] M T Duraisingh, P Jones, I Sambou, L von Seidlein, M Pinder, and D C Warhurst. The tyrosine-86 allele of the pfmdr1 gene of plasmodium falciparum is associated with increased sensitivity to the anti-malarials mefloquine and artemisinin. *Molecular and Biochemical Parasitology*, 108 (1):13–23, Apr 2000. URL http://www.sciencedirect.com/science/article/pii/S0166685100002012.

[50] Ananias A Escalante, Omar E Cornejo, Ascanio Rojas, Venkatachalam Udhayakumar, and Altaf A Lal. Assessing the effect of natural selection in malaria parasites. *Trends in Parasitology*, 20(8):388–395, August 2004. doi: 10.1016/j.pt.2004.06.002. URL http://dx.doi.org/10.1016/j.pt.2004.06.002.

[51] Michael T Ferdig, Roland A Cooper, Jianbing Mu, Bingbing Deng, Deirdre A Joy, Xin Zhuan Su, and Thomas E Wellems. Dissecting the loci of low-level quinine resistance in malaria parasites. *Molecular Microbiology*,

52(4):985–997, May 2004. doi: 10.1111/j.1365-2958.2004.04035.x. URL http://dx.doi.org/10.1111/j.1365-2958.2004.04035.x.

[52] Isabel D Ferreira, Virgílio E do Rosário, and Pedro V L Cravo. Real-time quantitative PCR with SYBR Green I detection for estimating copy numbers of nine drug resistance candidate genes in Plasmodium falciparum. *Malaria Journal*, 5:1, 2006. doi: 10.1186/1475-2875-5-1. URL http://www.malariajournal.com/content/5/1/1.

[53] D A Fidock and T E Wellems. Transformation with human dihydrofolate reductase renders malaria parasites insensitive to WR99210 but does not affect the intrinsic activity of proguanil. *Proceedings of the National Academy of Sciences, USA*, 94(20):10931–10936, September 1997. URL http://www.pnas.org/content/94/20/10931.full.

[54] D A Fidock, T Nomura, A K Talley, R A Cooper, S M Dzekunov, M T Ferdig, L M Ursos, A B Sidhu, B Naudé, Kirk W Deitsch, Xin-zhuan Su, John C Wootton, P D Roepe, and T E Wellems. Mutations in the p. falciparum digestive vacuole transmembrane protein pfcrt and evidence for their role in chloroquine resistance. *Molecular Cell*, 6(4):861–71, Oct 2000. URL http://linkinghub.elsevier.com/retrieve/pii/S1097-2765(05)00077-8.

[55] Quinton L Fivelman, Geoffrey A Butcher, Ipemida S Adagu, David C Warhurst, and Geoffrey Pasvol. Malarone treatment failure and in vitro confirmation of resistance of Plasmodium falciparum isolate from Lagos, Nigeria. *Malaria Journal*, 1:1–1, February 2002. doi: 10.1186/1475-2875-1-1. URL http://www.malariajournal.com/content/1/1/1.

[56] S J Foote, J K Thompson, Alan F Cowman, and D J Kemp. Amplification of the multidrug resistance gene in some chloroquine-resistant isolates of P. falciparum. *Cell*, 57(6):921–930, June 1989. doi: 10.1016/0092-8674(89)90330-9. URL http://www.sciencedirect.com/science/article/pii/0092867489903309.

[57] Malcolm J Gardner, Neil Hall, Eula Fung, Owen White, Matthew Berriman, Richard W Hyman, Jane M Carlton, Arnab Pain, Karen E Nelson, Sharen Bowman, Ian T Paulsen, Keith James, Jonathan A Eisen, Kim Rutherford, Steven L Salzberg, Alister Craig, Sue A Kyes, Man-Suen Chan, Vishvanath Nene, Shamira J Shallom, Bernard Suh, Jeremy Peterson, Sam Angiuoli, Mihaela Pertea, Jonathan Allen, Jeremy Selengut,

Daniel Haft, Michael W Mather, Akhil B Vaidya, David M A Martin, Alan H Fairlamb, Martin J Fraunholz, David S Roos, Stuart A Ralph, Geoffrey I McFadden, Leda M Cummings, G Mani Subramanian, Chris Mungall, J Craig Venter, Daniel J Carucci, Stephen L Hoffman, Chris Newbold, Ronald W Davis, Claire M Fraser, and Bart Barrell. Genome sequence of the human malaria parasite plasmodium falciparum. *Nature*, 419(6906):498–511, Oct 2002. doi: 10.1038/nature01097. URL http://www.nature.com/nature/journal/v419/n6906/full/nature01097.html.

[58] Markus Geisler, Marjolaine Girin, Sabine Brandt, Vincent Vincenzetti, Sonia Plaza, Nadine Paris, Yoshihiro Kobae, Masayoshi Maeshima, Karla Billion, Uner H Kolukisaoglu, Burkhard Schulz, and Enrico Martinoia. Arabidopsis immunophilin-like TWD1 functionally interacts with vacuolar ABC transporters. *Molecular Biology of the Cell*, 15(7): 3393–3405, July 2004. doi: 10.1091/mbc.E03-11-0831. URL http://www.molbiolcell.org/content/15/7/3393.

[59] Elodie Ghedin, Naomi A Sengamalay, Martin Shumway, Jennifer Zaborsky, Tamara Feldblyum, Vik Subbu, David J Spiro, Jeff Sitz, Hean Koo, Pavel Bolotov, Dmitry Dernovoy, Tatiana Tatusova, Yiming Bao, Kirsten St George, Jill Taylor, David J Lipman, Claire M Fraser, Jeffery K Taubenberger, and Steven L Salzberg. Large-scale sequencing of human influenza reveals the dynamic nature of viral genome evolution. *Nature*, 437(7062):1162–1166, October 2005. doi: 10.1038/nature04239. URL http://www.nature.com/nature/journal/v437/n7062/full/nature04239.html.

[60] GlaxoSmithKline. Halofantrine and fatal cardiac arrhythmia. Technical report, GlaxoSmithKline, 2005.

[61] Aric Gregson and Christopher V Plowe. Mechanisms of resistance of malaria parasites to antifolates. *Pharmacological Reviews*, 57(1):117–145, March 2005. doi: 10.1124/pr.57.1.4. URL http://pharmrev.aspetjournals.org/content/57/1/117.

[62] Sean M Griffing, Tonya Mixson-Hayden, Sankar Sridaran, Md Tauqeer Alam, Andrea M McCollum, César Cabezas, Wilmer Marquiño Quezada, John W Barnwell, Alexandre Macedo De Oliveira, Carmen Lucas, Nancy Arrospide, Ananias A Escalante, David J Bacon, and Venkatachalam Udhayakumar. South American Plasmodium falciparum after the malaria eradication era: clonal population expansion and survival of the fittest

hybrids. *PLoS ONE*, 6(9):e23486–e23486, January 2011. doi: 10.1371/journal.pone.0023486. URL http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0023486.

[63] Sharon R Grossman, Kristian G Andersen, Ilya Shlyakhter, Shervin Tabrizi, Sarah Winnicki, Angela Yen, Daniel J Park, Dustin Griesemer, Elinor K Karlsson, Sunny H Wong, Moran Cabili, Richard A Adegbola, Rameshwar N K Bamezai, Adrian V S Hill, Fredrik O Vannberg, John L Rinn, Eric S Lander, Stephen F Schaffner, and Pardis C Sabeti. Identifying Recent Adaptations in Large-Scale Genomic Data. *Cell*, 152(4):703–713, February 2013. doi: 10.1016/j.cell.2013.01.035. URL http://dx.doi.org/10.1016/j.cell.2013.01.035.

[64] Neil Hall, Marianna Karras, J Dale Raine, Jane M Carlton, Taco W A Kooij, Matthew Berriman, Laurence Florens, Christoph S Janssen, Arnab Pain, Georges K Christophides, Keith James, Kim Rutherford, Barbara Harris, David Harris, Carol Churcher, Michael A Quail, Doug Ormond, Jon Doggett, Holly E Trueman, Jacqui Mendoza, Shelby L Bidwell, Marie-Adele Rajandream, Daniel J Carucci, John R Yates, Fotis C Kafatos, Chris J Janse, Bart Barrell, C Michael R Turner, Andrew P Waters, and Robert E Sinden. A comprehensive survey of the Plasmodium life cycle by genomic, transcriptomic, and proteomic analyses. *Science*, 307(5706): 82–86, January 2005. doi: 10.1126/science.1103717. URL http://www.sciencemag.org/content/307/5706/82.

[65] Karen Hayton and Xin-Zhuan Su. Drug resistance and genetic mapping in Plasmodium falciparum. *Current Genetics*, 54(5):223–239, November 2008. doi: 10.1007/s00294-008-0214-x. URL http://www.springerlink.com/content/a455n61168nt1410/.

[66] William G Hill and A Robertson. The effect of linkage on limits to artificial selection. *Genetical Research*, 8:269–94, 1966.

[67] William G Hill and A Robertson. Linkage Disequlibrium in Finite Populations. *Theoretical and Applied Genetics*, 38:226–231, 1968.

[68] Richard R Hudson. Generating samples under a wright-fisher neutral model of genetic variation. *Bioinformatics*, 18(2):337–8, Feb 2002. URL http://bioinformatics.oxfordjournals.org/content/18/2/337.long.

[69] Richard R Hudson, M Slatkin, and W P Maddison. Estimation of levels of gene flow from dna sequence data. *Genetics*, 132(2):583–9, Oct 1992. URL http://www.genetics.org/cgi/reprint/132/2/583.

[70] Christopher J R Illingworth and Ville Mustonen. Distinguishing driver and passenger mutations in an evolutionary history categorized by interference. *Genetics*, 189(3):989–1000, November 2011. doi: 10.1534/genetics.111.133975. URL http://www.genetics.org/content/189/3/989.

[71] Christopher J R Illingworth and Ville Mustonen. Components of selection in the evolution of the influenza virus: linkage effects beat inherent selection. *PLoS Pathogens*, 8(12):e1003091–e1003091, December 2012. doi: 10.1371/journal.ppat.1003091. URL http://www.plospathogens.org/article/info%3Adoi%2F10.1371%2Fjournal.ppat.1003091.

[72] Christopher J R Illingworth, Leopold Parts, Stephan Schiffels, Gianni Liti, and Ville Mustonen. Quantifying selection acting on a complex trait using allele frequency time series data. *Molecular Biology and Evolution*, 29(4):1187–1197, April 2012. doi: 10.1093/molbev/msr289. URL http://mbe.oxfordjournals.org/cgi/doi/10.1093/molbev/msr289.

[73] Daniel F Jarosz and Susan Lindquist. Hsp90 and environmental stress transform the adaptive value of natural genetic variation. *Science*, 330(6012):1820–4, Dec 2010. doi: 10.1126/science.1195487. URL http://www.sciencemag.org/content/330/6012/1820.short.

[74] Daniel C Jeffares, Arnab Pain, Andrew Berry, Anthony V Cox, James Stalker, Catherine E Ingle, Alan Thomas, Michael A Quail, Kyle Siebenthall, Anne-Catrin Uhlemann, Sue A Kyes, Sanjeev Krishna, Chris Newbold, Emmanouil T Dermitzakis, and Matthew Berriman. Genome variation and evolution of the malaria parasite plasmodium falciparum. *Nature Genetics*, 39(1):120–5, Jan 2007. doi: 10.1038/ng1931. URL http://www.nature.com/ng/journal/v39/n1/abs/ng1931.html.

[75] Hongying Jiang, Ming Yi, Jianbing Mu, Louie Zhang, Al Ivens, Leszek J Klimczak, Yentram Huyen, Robert M Stephens, and Xin Zhuan Su. Detection of genome-wide polymorphisms in the at-rich plasmodium falciparum genome using a high-density microarray. *BMC Genomics*, 9:398, Aug 2008. doi: 10.1186/1471-2164-9-398. URL http://www.biomedcentral.com/1471-2164/9/398.

[76] Hongying Jiang, Na Li, Vivek Gopalan, Martine M Zilversmit, Sudhir Varma, Vijayaraj Nagarajan, Jian Li, Jianbing Mu, Karen Hayton, Bruce Henschen, Ming Yi, Robert Stephens, Gilean McVean, Philip Awadalla, Thomas E Wellems, and Xin Zhuan Su. High recombination rates and hotspots in a Plasmodium falciparum genetic cross. *Genome Biology*, 12 (4):R33, 2011. doi: 10.1186/gb-2011-12-4-r33. URL http://genomebiology.com/2011/12/4/R33.

[77] Fabiana Morandi Jordão, Emília Akemi Kimura, and Alejandro Miguel Katzin. Isoprenoid biosynthesis in the erythrocytic stages of plasmodium falciparum. *Mem Inst Oswaldo Cruz*, 106 Suppl 1:134–41, Aug 2011. URL http://www.ncbi.nlm.nih.gov/sites/entrez?Db=pubmed&Cmd=Retrieve&list_uids=21881768&dopt=abstractplus.

[78] Deirdre A Joy, Xiaorong Feng, Jianbing Mu, Tetsuya Furuya, Kesinee Chotivanich, Antoniana U Krettli, May Ho, Alex Wang, Nicholas J White, Edward Suh, Peter Beerli, and Xin Zhuan Su. Early origin and recent expansion of plasmodium falciparum. *Science*, 300(5617):318–21, Apr 2003. doi: 10.1126/science.1081449. URL http://www.sciencemag.org/content/300/5617/318.long.

[79] Hyun Min Kang, Noah A Zaitlen, Claire M Wade, Andrew Kirby, David Heckerman, Mark J Daly, and Eleazar Eskin. Efficient control of population structure in model organism association mapping. *Genetics*, 178(3):1709–23, Mar 2008. doi: 10.1534/genetics.107.080101. URL http://www.genetics.org/cgi/content/full/178/3/1709.

[80] Hyun Min Kang, Jae Hoon Sul, Susan K Service, Noah A Zaitlen, Sit-Yee Kong, Nelson B Freimer, Chiara Sabatti, and Eleazar Eskin. Variance component model to account for sample structure in genome-wide association studies. *Nature Genetics*, 42(4):348–354, March 2010. doi: 10.1038/ng.548. URL http://www.nature.com/doifinder/10.1038/ng.548.

[81] W James Kent. BLAT—the BLAST-like alignment tool. *Genome Research*, 12(4):656–664, April 2002. doi: 10.1101/gr.229202. URL http://genome.cshlp.org/content/12/4/656.long.

[82] Jacques J Kessl, Steven R Meshnick, and Bernard L Trumpower. Modeling the molecular basis of atovaquone resistance in parasites and pathogenic fungi. *Trends in Parasitology*, 23(10):494–501, October 2007. doi: 10.1016/j.pt.2007.08.004. URL http://www.sciencedirect.com/science/article/pii/S1471492207002176.

[83] Claire Kidgell, Sarah K Volkman, Johanna Daily, Justin O Borevitz, David Plouffe, Yingyao Zhou, Jeffrey R Johnson, Karine G Le Roch, Ousmane Sarr, Omar Ndir, Souleymane Mboup, Serge Batalov, Dyann F Wirth, and Elizabeth A Winzeler. A systematic map of genetic variation in plasmodium falciparum. *PLoS Pathogens*, 2(6):e57, Jun 2006. doi: 10.1371/journal.ppat.0020057. URL http://www.plospathogens.org/article/info%253Adoi%252F10.1371%252Fjournal.ppat.0020057.

[84] Sridhar Kudaravalli, Jean-Baptiste Veyrieras, Barbara E Stranger, Emmanouil T Dermitzakis, and Jonathan K Pritchard. Gene expression levels are a target of recent natural selection in the human genome. *Molecular Biology and Evolution*, 26(3):649–658, March 2009. doi: 10.1093/molbev/msn289. URL http://mbe.oxfordjournals.org/cgi/content/full/26/3/649.

[85] Stefan Kurtz, Adam Phillippy, Arthur L Delcher, Michael Smoot, Martin Shumway, Corina Antonescu, and Steven L Salzberg. Versatile and open software for comparing large genomes. *Genome Biology*, 5(2):R12, Jan 2004. doi: 10.1186/gb-2004-5-2-r12. URL http://genomebiology.com/content/5/2/R12.

[86] Eric S Lander, L M Linton, B Birren, C Nusbaum, M C Zody, J Baldwin, K Devon, K Dewar, M Doyle, W FitzHugh, R Funke, D Gage, K Harris, A Heaford, J Howland, L Kann, J Lehoczky, R LeVine, P McEwan, K McKernan, J Meldrim, J P Mesirov, C Miranda, W Morris, J Naylor, C Raymond, M Rosetti, R Santos, A Sheridan, C Sougnez, N Stange-Thomann, N Stojanovic, A Subramanian, D Wyman, J Rogers, J Sulston, R Ainscough, S Beck, D Bentley, J Burton, C Clee, N Carter, A Coulson, R Deadman, P Deloukas, A Dunham, I Dunham, R Durbin, L French, D Grafham, S Gregory, T Hubbard, S Humphray, A Hunt, M Jones, C Lloyd, A McMurray, L Matthews, S Mercer, S Milne, J C Mullikin, A Mungall, R Plumb, M Ross, R Shownkeen, S Sims, R H Waterston, R K Wilson, L W Hillier, J D McPherson, M A Marra, E R Mardis, L A Fulton, A T Chinwalla, K H Pepin, W R Gish, S L Chissoe, M C Wendl, K D Delehaunty, T L Miner, A Delehaunty, J B Kramer, L L Cook, R S Fulton, D L Johnson, P J Minx, S W Clifton, T Hawkins, E Branscomb, P Predki, P Richardson, S Wenning, T Slezak, N Doggett, J F Cheng, A Olsen, S Lucas, C Elkin, E Uberbacher, M Frazier, R A Gibbs, D M Muzny, S E Scherer, J B Bouck, E J Sodergren, K C Worley, C M Rives, J H Gorrell, M L Metzker, S L Naylor, R S Kucherlapati, D L Nelson, G M Weinstock, Y Sakaki, A Fujiyama, M Hattori, T Yada, A Toyoda, T Itoh, C Kawagoe,

H Watanabe, Y Totoki, T Taylor, J Weissenbach, R Heilig, W Saurin, F Artiguenave, P Brottier, T Bruls, E Pelletier, C Robert, P Wincker, D R Smith, L Doucette-Stamm, M Rubenfield, K Weinstock, H M Lee, J Dubois, A Rosenthal, M Platzer, G Nyakatura, S Taudien, A Rump, H Yang, J Yu, J Wang, G Huang, J Gu, L Hood, L Rowen, A Madan, S Qin, R W Davis, N A Federspiel, A P Abola, M J Proctor, R M Myers, J Schmutz, M Dickson, J Grimwood, D R Cox, M V Olson, R Kaul, C Raymond, N Shimizu, K Kawasaki, S Minoshima, G A Evans, M Athanasiou, R Schultz, B A Roe, F Chen, H Pan, J Ramser, H Lehrach, R Reinhardt, W R McCombie, M de la Bastide, N Dedhia, H Blöcker, K Hornischer, G Nordsiek, R Agarwala, L Aravind, J A Bailey, A Bateman, S Batzoglou, E Birney, P Bork, D G Brown, C B Burge, L Cerutti, H C Chen, D Church, M Clamp, R R Copley, T Doerks, S R Eddy, E E Eichler, T S Furey, J Galagan, J G Gilbert, C Harmon, Y Hayashizaki, D Haussler, H Hermjakob, K Hokamp, W Jang, L S Johnson, T A Jones, S Kasif, A Kaspryzk, S Kennedy, W J Kent, P Kitts, E V Koonin, I Korf, D Kulp, D Lancet, T M Lowe, A McLysaght, T Mikkelsen, J V Moran, N Mulder, V J Pollara, C P Ponting, G Schuler, J Schultz, G Slater, A F Smit, E Stupka, J Szustakowski, D Thierry-Mieg, J Thierry-Mieg, L Wagner, J Wallis, R Wheeler, A Williams, Y I Wolf, K H Wolfe, S P Yang, R F Yeh, F Collins, M S Guyer, J Peterson, A Felsenfeld, K A Wetterstrand, A Patrinos, M J Morgan, P de Jong, J J Catanese, K Osoegawa, H Shizuya, S Choi, Y J Chen, J Szustakowki, and International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921, February 2001. doi: 10.1038/35057062. URL http://www.nature.com/nature/journal/v409/n6822/full/409860a0.html.

[87] Miriam K Laufer, Shannon Takala-Harrison, Fraction K Dzinjalamala, O Colin Stine, Terrie E Taylor, and Christopher V Plowe. Return of chloroquine-susceptible falciparum malaria in malawi was a reexpansion of diverse susceptible parasites. *Journal of Infectious Diseases*, 202(5): 801–8, Sep 2010. doi: 10.1086/655659. URL http://jid.oxfordjournals.org/content/202/5/801.long.

[88] Karine G Le Roch, Jeffrey R Johnson, Hugues Ahiboh, Duk-Won D Chung, Jacques Prudhomme, David Plouffe, Kerstin Henson, Yingyao Zhou, William Witola, John R Yates, Choukri Ben Mamoun, Elizabeth A Winzeler, and Henri Vial. A systematic approach to understand the mechanism of action of the bisthiazolium compound t4 on the human malaria parasite, plasmodium falciparum. *BMC Genomics*, 9:513, Jan

2008. doi: 10.1186/1471-2164-9-513. URL http://www.biomedcentral.com/1471-2164/9/513.

[89] Philippe Leprohon, Danielle Légaré, and Marc Ouellette. Abc transporters involved in drug resistance in human parasites. *Essays in Biochemistry*, 50(1):121–44, Sep 2011. doi: 10.1042/bse0500121. URL http://dx.doi.org/10.1042/bse0500121.

[90] Kerstin Lindblad-Toh, Claire M Wade, Tarjei S Mikkelsen, Elinor K Karlsson, David B Jaffe, Michael Kamal, Michele Clamp, Jean L Chang, Edward J Kulbokas, Michael C Zody, Evan Mauceli, Xiaohui Xie, Matthew Breen, Robert K Wayne, Elaine A Ostrander, Chris P Ponting, Francis Galibert, Douglas R Smith, Pieter J DeJong, Ewen Kirkness, Pablo Alvarez, Tara M Biagi, William Brockman, Jonathan Butler, Chee-Wye Chin, April Cook, James Cuff, Mark J Daly, David DeCaprio, Sante Gnerre, Manfred Grabherr, Manolis Kellis, Michael Kleber, Carolyne Bardeleben, Leo Goodstadt, Andreas Heger, Christophe Hitte, Lisa Kim, Klaus-Peter Koepfli, Heidi G Parker, John P Pollinger, Stephen M J Searle, Nathan B Sutter, Rachael Thomas, Caleb Webber, Broad Institute Genome Sequencing Platform, and Eric S Lander. Genome sequence, comparative analysis and haplotype structure of the domestic dog. *Nature*, 438(7069):803–19, Dec 2005. doi: 10.1038/nature04338. URL http://www.nature.com/nature/journal/v438/n7069/full/nature04338.html.

[91] Weimin Liu, Yingying Li, Gerald H Learn, Rebecca S Rudicell, Joel D Robertson, Brandon F Keele, Jean-Bosco N Ndjango, Crickette M Sanz, David B Morgan, Sabrina Locatelli, Mary K Gonder, Philip J Kranzusch, Peter D Walsh, Eric Delaporte, Eitel Mpoudi-Ngole, Alexander V Georgiev, Martin N Muller, George M Shaw, Martine Peeters, Paul M Sharp, Julian C Rayner, and Beatrice H Hahn. Origin of the human malaria parasite Plasmodium falciparum in gorillas. *Nature*, 467(7314): 420–425, September 2010. doi: 10.1038/nature09442. URL http://www.nature.com/nature/journal/v467/n7314/full/nature09442.html.

[92] Kirk E Lohmueller, Celeste L Pearce, Malcolm Pike, Eric S Lander, and Joel N Hirschhorn. Meta-analysis of genetic association studies supports a contribution of common variants to susceptibility to common disease. *Nature Genetics*, 33(2):177–82, Feb 2003. doi: 10.1038/ng1071. URL http://www.nature.com/ng/journal/v33/n2/full/ng1071.html.

[93] Anna-Sapfo Malaspinas, Orestis Malaspinas, Steven N Evans, and Montgomery Slatkin. Estimating allele age and selection coefficient from time-serial data. *Genetics*, 192(2):599–607, October 2012. doi: 10.1534/genetics.112.140939. URL http://www.genetics.org/content/192/2/599.

[94] malERA Consultative Group on Drugs. A research agenda for malaria eradication: drugs. *PLoS Medicine*, 8(1):e1000402, Jan 2011. doi: 10.1371/journal.pmed.1000402. URL http://www.plosmedicine.org/article/info%253Adoi%252F10.1371%252Fjournal.pmed.1000402.

[95] Magnus Manske, Olivo Miotto, Susana Campino, Sarah Auburn, Jacob Almagro-Garcia, Gareth Maslen, Jack O'Brien, Abdoulaye Djimde, Ogobara Doumbo, Issaka Zongo, Jean Bosco Ouédraogo, Pascal Michon, Ivo Mueller, Peter Siba, Alexis Nzila, Steffen Borrmann, Steven M Kiara, Kevin Marsh, Hongying Jiang, Xin Zhuan Su, Chanaki Amaratunga, Rick M Fairhurst, Duong Socheat, François Nosten, Mallika Imwong, Nicholas J White, Mandy Sanders, Elisa Anastasi, Dan Alcock, Eleanor Drury, Samuel Oyola, Michael A Quail, Daniel J Turner, Valentin Ruano-Rubio, Dushyanth Jyothi, Lucas Amenga-Etego, Christina Hubbart, Anna Jeffreys, Kate Rowlands, Colin Sutherland, Cally Roper, Valentina Mangano, David Modiano, John C Tan, Michael T Ferdig, Alfred Amambua-Ngwa, David J Conway, Shannon Takala-Harrison, Christopher V Plowe, Julian C Rayner, Kirk A Rockett, Taane G Clark, Chris I Newbold, Matthew Berriman, Bronwyn MacInnis, and Dominic P Kwiatkowski. Analysis of Plasmodium falciparum diversity in natural infections by deep sequencing. *Nature*, June 2012. doi: 10.1038/nature11174. URL http://www.nature.com/doifinder/10.1038/nature11174.

[96] Aaron McKenna, Matthew Hanna, Eric Banks, Andrey Sivachenko, Kristian Cibulskis, Andrew Kernytsky, Kiran Garimella, David M Altshuler, Stacey Gabriel, Mark Daly, and Mark A Depristo. The genome analysis toolkit: a mapreduce framework for analyzing next-generation dna sequencing data. *Genome Research*, 20(9):1297–303, Sep 2010. doi: 10.1101/gr.107524.110. URL http://www.genome.org/cgi/doi/10.1101/gr.107524.110.

[97] Gil McVean, Philip Awadalla, and Paul Fearnhead. A coalescent-based method for detecting and estimating recombination from gene sequences. *Genetics*, 160(3):1231–41, Mar 2002. URL http://www.genetics.org/content/160/3/1231.long.

[98] Alexandre Melnikov, Kevin Galinsky, Peter Rogov, Timothy Fennell, Daria Van Tyne, Carsten Russ, Rachel Daniels, Kayla G Barnes, James Bochicchio, Daouda Ndiaye, Papa D Sene, Dyann F Wirth, Chad Nusbaum, Sarah K Volkman, Bruce W Birren, Andreas Gnirke, and Daniel E Neafsey. Hybrid selection for sequencing pathogen genomes from clinical samples. *Genome Biology*, 12(8):R73, Aug 2011. doi: 10.1186/gb-2011-12-8-r73. URL http://genomebiology.com/2011/12/8/R73/abstract.

[99] Xiangbing Meng, Danlin Zhu, Shujie Yang, Xinjun Wang, Zhi Xiong, Yuping Zhang, Pavla Brachova, and Kimberly K Leslie. Cytoplasmic metadherin (mtdh) provides survival advantage under conditions of stress by acting as rna-binding protein. *Journal of Biological Chemistry*, 287(7): 4485–91, Feb 2012. doi: 10.1074/jbc.C111.291518. URL http://www.jbc.org/content/287/7/4485.long.

[100] Karl E Miletti-González, Shiling Chen, Neelakandan Muthukumaran, Giuseppa N Saglimbeni, Xiaohua Wu, Jinming Yang, Kevin Apolito, Weichung J Shih, William N Hait, and Lorna Rodríguez-Rodríguez. The CD44 receptor interacts with P-glycoprotein to promote cell migration and invasion in cancer. *Cancer Research*, 65(15):6660–6667, August 2005. doi: 10.1158/0008-5472.CAN-04-3478. URL http://cancerres.aacrjournals.org/content/65/15/6660.

[101] Danny A Milner, Jimmy Vareta, Clarissa Valim, Jacqui Montgomery, Rachel F Daniels, Sarah K Volkman, Daniel E Neafsey, Daniel J Park, Stephen F Schaffner, Nira C Mahesh, Kayla G Barnes, David M Rosen, Amanda K Lukens, Daria Van Tyne, Roger C Wiegand, Pardis C Sabeti, Karl B Seydel, Simon J Glover, Steve Kamiza, Malcolm E Molyneux, Terrie E Taylor, and Dyann F Wirth. Human cerebral malaria and Plasmodium falciparum genotypes in Malawi. *Malaria Journal*, 11:35, March 2012. doi: 10.1186/1475-2875-11-35. URL http://www.malariajournal.com/content/11/1/35.

[102] Olivo Miotto, Jacob Almagro-Garcia, Magnus Manske, Bronwyn MacInnis, Susana Campino, Kirk A Rockett, Chanaki Amaratunga, Pharath Lim, Seila Suon, Sokunthea Sreng, Jennifer M Anderson, Socheat Duong, Chea Nguon, Char Meng Chuor, David Saunders, Youry Se, Chantap Lon, Mark M Fukuda, Lucas Amenga-Etego, Abraham V O Hodgson, Victor Asoala, Mallika Imwong, Shannon Takala-Harrison, François Nosten, Xin Zhuan Su, Pascal Ringwald, Frédéric Ariey, Christiane Dolecek, Tran Tinh Hien, Maciej F Boni, Cao Quang Thai,

Alfred Amambua-Ngwa, David J Conway, Abdoulaye A Djimde, Ogobara K Doumbo, Issaka Zongo, Jean Bosco Ouédraogo, Daniel Alcock, Eleanor Drury, Sarah Auburn, Oliver Koch, Mandy Sanders, Christina Hubbart, Gareth Maslen, Valentin Ruano-Rubio, Dushyanth Jyothi, Alistair Miles, John O'Brien, Chris Gamble, Samuel O Oyola, Julian C Rayner, Chris I Newbold, Matthew Berriman, Chris C A Spencer, Gilean McVean, Nicholas P Day, Nicholas J White, Delia Bethell, Arjen M Dondorp, Christopher V Plowe, Rick M Fairhurst, and Dominic P Kwiatkowski. Multiple populations of artemisinin-resistant Plasmodium falciparum in Cambodia. *Nature Genetics*, pages –, April 2013. doi: 10.1038/ng.2624. URL http://www.nature.com/ng/journal/vaop/ncurrent/abs/ng.2624.html.

[103] Sachel Mok, Mallika Imwong, Margaret J Mackinnon, Joan Sim, Ramya Ramadoss, Poravuth Yi, Mayfong Mayxay, Kesinee Chotivanich, Kek-Yee Liong, Bruce Russell, Duong Socheat, Paul N Newton, Nicholas P J Day, Nicholas J White, Peter R Preiser, François Nosten, Arjen M Dondorp, and Zbynek Bozdech. Artemisinin resistance in plasmodium falciparum is associated with an altered temporal pattern of transcription. *BMC Genomics*, 12:391, Jan 2011. doi: 10.1186/1471-2164-12-391. URL http://www.biomedcentral.com/1471-2164/12/391.

[104] E Mouzin, P M Thior, M B Diouf, and B Sambou. Focus on senegal. Technical Report 4, World Health Organization, Nov 2010. URL http://www.path.org/publications/detail.php?i=2072.

[105] Jianbing Mu, Philip Awadalla, Junhui Duan, Kate M McGee, Deirdre A Joy, Gilean A T McVean, and Xin Zhuan Su. Recombination hotspots and population structure in plasmodium falciparum. *PLoS Biology*, 3(10): e335, Oct 2005. doi: 10.1371/journal.pbio.0030335. URL http://www.plosbiology.org/article/info%253Adoi%252F10.1371%252Fjournal.pbio.0030335.

[106] Jianbing Mu, Philip Awadalla, Junhui Duan, Kate M McGee, Jon Keebler, Karl Seydel, Gilean A T McVean, and Xin Zhuan Su. Genome-wide variation and identification of vaccine targets in the plasmodium falciparum genome. *Nature Genetics*, 39(1):126–30, Jan 2007. doi: 10.1038/ng1924. URL http://www.nature.com/ng/journal/v39/n1/abs/ng1924.html.

[107] Jianbing Mu, Rachel A Myers, Hongying Jiang, Shengfa Liu, Stacy Ricklefs, Michael Waisberg, Kesinee Chotivanich, Polrat Wilairatana,

Srivicha Krudsood, Nicholas J White, Rachanee Udomsangpetch, Liwang Cui, May Ho, Fengzhen Ou, Haibo Li, Jianping Song, Guoqiao Li, Xinhua Wang, Suon Seila, Sreng Sokunthea, Duong Socheat, Daniel E Sturdevant, Stephen F Porcella, Rick M Fairhurst, Thomas E Wellems, Philip Awadalla, and Xin-zhuan Su. Plasmodium falciparum genome-wide scans for positive selection, recombination hot spots and resistance to antimalarial drugs. *Nature Genetics*, 42:268–271, Jan 2010. doi: 10.1038/ng.528. URL http://www.nature.com/ng/journal/vaop/ncurrent/abs/ng.528.html.

[108] Christopher J L Murray, Lisa C Rosenfeld, Stephen S Lim, Kathryn G Andrews, Kyle J Foreman, Diana Haring, Nancy Fullman, Mohsen Naghavi, Rafael Lozano, and Alan D Lopez. Global malaria mortality between 1980 and 2010: a systematic analysis. *The Lancet*, 379(9814): 413–31, Feb 2012. doi: 10.1016/S0140-6736(12)60034-8. URL http://www.sciencedirect.com/science/article/pii/S0140673612600348.

[109] Shalini Nair, Jeff T Williams, Alan Brockman, Lucy Paiphun, Mayfong Mayxay, Paul N Newton, Jean-Paul Guthmann, Frank M Smithuis, Tran Tinh Hien, Nicholas J White, François Nosten, and Tim J C Anderson. A selective sweep driven by pyrimethamine treatment in southeast asian malaria parasites. *Molecular Biology and Evolution*, 20(9): 1526–36, Sep 2003. doi: 10.1093/molbev/msg162. URL http://mbe.oxfordjournals.org/content/20/9/1526.long.

[110] Shalini Nair, Becky Miller, Marion Barends, Anchalee Jaidee, Jigar J Patel, Mayfong Mayxay, Paul Newton, Francois Nosten, Michael T Ferdig, and Tim J C Anderson. Adaptive copy number evolution in malaria parasites. *PLoS Genetics*, 4(10):e1000243, Oct 2008. doi: 10.1371/journal.pgen.1000243. URL http://www.plosgenetics.org/article/info%253Adoi%252F10.1371%252Fjournal.pgen.1000243.

[111] José A Nájera, Matiana González-Silva, and Pedro L Alonso. Some lessons for the future from the Global Malaria Eradication Programme (1955-1969). *PLoS Medicine*, 8(1):e1000412, 2011. doi: 10.1371/journal.pmed.1000412. URL http://www.plosmedicine.org/article/info%3Adoi%2F10.1371%2Fjournal.pmed.1000412.

[112] Daniel E Neafsey. Genome sequencing sheds light on emerging drug resistance in malaria parasites. *Nature Genetics*, 45(6):589–590, May 2013. doi: 10.1038/ng.2648. URL http://www.nature.com/doifinder/10.1038/ng.2648.

[113] Daniel E Neafsey, Stephen F Schaffner, Sarah K Volkman, Daniel J Park, Philip Montgomery, Danny A Milner, Amanda Lukens, David Rosen, Rachel Daniels, Nathan Houde, Joseph F Cortese, Erin Tyndall, Casey Gates, Nicole Stange-Thomann, Ousmane Sarr, Daouda Ndiaye, Omar Ndir, Souleymane Mboup, Marcelo U Ferreira, Sandra do Lago Moraes, Aditya P Dash, Chetan E Chitnis, Roger C Wiegand, Daniel L Hartl, Bruce W Birren, Eric S Lander, Pardis C Sabeti, and Dyann F Wirth. Genome-wide snp genotyping highlights the role of natural selection in plasmodium falciparum population divergence. *Genome Biology*, 9(12): R171, Dec 2008. doi: 10.1186/gb-2008-9-12-r171. URL http://genomebiology.com/2008/9/12/R171.

[114] Daniel E Neafsey, Bridget M Barker, Thomas J Sharpton, Jason E Stajich, Daniel J Park, Emily Whiston, Chiung-Yu Hung, Cody McMahan, Jared White, Sean Sykes, David Heiman, Sarah Young, Qiandong Zeng, Amr Abouelleil, Lynne Aftuck, Daniel Bessette, Adam Brown, Michael Fitzgerald, Annie Lui, J Pendexter Macdonald, Margaret Priest, Marc J Orbach, John N Galgiani, Theo N Kirkland, Garry T Cole, Bruce W Birren, Matthew R Henn, John W Taylor, and Steven D Rounsley. Population genomic sequencing of Coccidioides fungi reveals recent hybridization and transposon control. *Genome Research*, 20(7):938–946, July 2010. doi: 10.1101/gr.103911.109. URL http://genome.cshlp.org/content/20/7/938.

[115] Daniel E Neafsey, Mara K N Lawniczak, Daniel J Park, Seth N Redmond, M B Coulibaly, S F Traoré, N Sagnon, C Costantini, Charlie Johnson, Roger C Wiegand, Frank H Collins, Eric S Lander, Dyann F Wirth, Fotis C Kafatos, Nora J Besansky, George K Christophides, and Marc A T Muskavitch. Snp genotyping defines complex gene-flow boundaries among african malaria vector mosquitoes. *Science*, 330(6003):514–7, Oct 2010. doi: 10.1126/science.1193036. URL http://www.sciencemag.org/content/330/6003/514.

[116] Wang Nguitragool, Abdullah A B Bokhari, Ajay D Pillai, Kempaiah Rayavara, Paresh Sharma, Brad Turpin, L Aravind, and Sanjay A Desai. Malaria Parasite clag3 Genes Determine Channel-Mediated Nutrient Uptake by Infected Red Blood Cells. *Cell*, 145(5):665–677, May 2011. doi: 10.1016/j.cell.2011.05.002. URL http://dx.doi.org/10.1016/j.cell.2011.05.002.

[117] Rasmus Nielsen. Estimation of population parameters and recombination rates from single nucleotide polymorphisms. *Genetics*, 154(2):931–42,

Feb 2000. URL http://www.genetics.org/cgi/content/full/154/2/931.

[118] Rasmus Nielsen, Melissa J Hubisz, and Andrew G Clark. Reconstituting the frequency spectrum of ascertained single-nucleotide polymorphism data. *Genetics*, 168(4):2373–82, Dec 2004. doi: 10.1534/genetics.104.031039. URL http://www.genetics.org/cgi/content/full/168/4/2373.

[119] Z Ning, A J Cox, and James C Mullikin. SSAHA: a fast search method for large DNA databases. *Genome Research*, 11(10):1725–1729, October 2001. doi: 10.1101/gr.194201. URL http://genome.cshlp.org/content/11/10/1725.long.

[120] Standwell Nkhoma, Shalini Nair, Mavuto Mukaka, Malcolm E Molyneux, Stephen A Ward, and Timothy J C Anderson. Parasites bearing a single copy of the multi-drug resistance gene (pfmdr-1) with wild-type snps predominate amongst plasmodium falciparum isolates from malawi. *Acta tropica*, 111(1):78–81, Jul 2009. doi: 10.1016/j.actatropica.2009.01.011. URL http://www.sciencedirect.com/science/article/pii/S0001706X09000291.

[121] Standwell C Nkhoma, Kasia Stepniewska, Shalini Nair, Aung Pyae Phyo, Rose McGready, François Nosten, and Timothy J C Anderson. Genetic evaluation of the performance of malaria parasite clearance rate metrics. *Journal of Infectious Diseases*, 208(2):346–350, July 2013. doi: 10.1093/infdis/jit165. URL http://jid.oxfordjournals.org/content/208/2/346.

[122] Standwell C SC Nkhoma, Shalini S Nair, Ian H IH Cheeseman, Cherise C Rohr-Allegrini, Sittaporn S Singlam, François F Nosten, and Tim J C TJ Anderson. Close kinship within multiple-genotype malaria parasite infections. *Proceedings Biological Sciences*, 279(1738):2589–2598, July 2012. doi: 10.1098/rspb.2012.0113. URL http://rspb.royalsocietypublishing.org/content/279/1738/2589.

[123] Louis J Nkrumah, Paul M Riegelhaupt, Pedro Moura, David J Johnson, Jigar Patel, Karen Hayton, Michael T Ferdig, Thomas E Wellems, Myles H Akabas, and David A Fidock. Probing the multifactorial basis of Plasmodium falciparum quinine resistance: evidence for a strain-specific contribution of the sodium-proton exchanger PfNHE. *Molecular and Biochemical Parasitology*, 165(2):122–131, June 2009. doi:

10.1016/j.molbiopara.2009.01.011. URL http://dx.doi.org/10.1016/j.molbiopara.2009.01.011.

[124] Harald Noedl, Youry Se, Kurt Schaecher, Bryan L Smith, Duong Socheat, Mark M Fukuda, and Artemisinin Resistance in Cambodia 1 ARC1 Study Consortium. Evidence of artemisinin-resistant malaria in western Cambodia. *New England Journal of Medicine*, 359(24):2619–2620, December 2008. doi: 10.1056/NEJMc0805011. URL http://www.nejm.org/doi/full/10.1056/NEJMc0805011.

[125] Lynette Isabella Ochola, Kevin K A Tetteh, Lindsay B Stewart, Victor Riitho, Kevin Marsh, and David J Conway. Allele frequency-based and polymorphism-versus-divergence indices of balancing selection in a new filtered set of polymorphic genes in plasmodium falciparum. *Molecular Biology and Evolution*, 27(10):2344–2351, May 2010. doi: 10.1093/molbev/msq119. URL http://mbe.oxfordjournals.org/content/27/10/2344.

[126] World Health Organization. World malaria report. Technical report, World Health Organization, 2011. URL http://www.who.int/malaria/world_malaria_report_2011/.

[127] Samuel O Oyola, Thomas D Otto, Yong Gu, Gareth Maslen, Magnus Manske, Susana Campino, Daniel J Turner, Bronwyn Maclnnis, Dominic P Kwiatkowski, Harold P Swerdlow, and Michael A Quail. Optimizing illumina next-generation sequencing library preparation for extremely at-biased genomes. *BMC Genomics*, 13(1):1, Jan 2012. doi: 10.1186/1471-2164-13-1. URL http://www.biomedcentral.com/1471-2164/13/1/abstract.

[128] Samuel O Oyola, Yong Gu, Magnus Manske, Thomas D Otto, John O'Brien, Daniel Alcock, Bronwyn MacInnis, Matthew Berriman, Chris I Newbold, Dominic P Kwiatkowski, Harold P Swerdlow, and Michael A Quail. Efficient depletion of host DNA contamination in malaria clinical sequencing. *Journal of Clinical Microbiology*, 51(3):745–751, March 2013. doi: 10.1128/JCM.02507-12. URL http://jcm.asm.org/content/51/3/745.

[129] Daniel J Park, Amanda K Lukens, Daniel E Neafsey, Stephen F Schaffner, Hsiao-Han Chang, Clarissa Valim, Ulf Ribacke, Daria Van Tyne, Kevin Galinsky, Meghan Galligan, Justin S Becker, Daouda Ndiaye, Souleymane Mboup, Roger C Wiegand, Daniel L Hartl, Pardis C Sabeti, Dyann F

Wirth, and Sarah K Volkman. Sequence-based association and selection scans identify drug resistance loci in the Plasmodium falciparum malaria parasite. *Proceedings of the National Academy of Sciences, USA*, 109: 13052–13057, August 2012. doi: 10.1073/pnas.1210585109. URL http://www.pnas.org/cgi/doi/10.1073/pnas.1210585109.

[130] Nick Patterson, Alkes L Price, and David Reich. Population structure and eigenanalysis. *PLoS Genetics*, 2(12):e190, Dec 2006. doi: 10.1371/journal.pgen.0020190. URL http://www.plosgenetics.org/article/info%253Adoi%252F10.1371%252Fjournal.pgen.0020190.

[131] D Payne. Spread of chloroquine resistance in Plasmodium falciparum. *Parasitology Today*, 3(8):241–246, August 1987. doi: 10.1016/0169-4758(87)90147-5. URL http://dx.doi.org/10.1016/0169-4758(87)90147-5.

[132] David Plouffe, Achim Brinker, Case McNamara, Kerstin Henson, Nobutaka Kato, Kelli Kuhen, Advait Nagle, Francisco Adrián, Jason T Matzen, Paul Anderson, Tae-Gyu Nam, Nathanael S Gray, Arnab Chatterjee, Jeff Janes, S Frank Yan, Richard Trager, Jeremy S Caldwell, Peter G Schultz, Yingyao Zhou, and Elizabeth A Winzeler. In silico activity profiling reveals the mechanism of action of antimalarials discovered in a high-throughput screen. *Proceedings of the National Academy of Sciences, USA*, 105(26):9059–64, Jul 2008. doi: 10.1073/pnas.0802982105. URL http://www.pnas.org/content/105/26/9059.long.

[133] Nadia Ponts, Jianfeng Yang, Duk-Won Doug Chung, Jacques Prudhomme, Thomas Girke, Paul Horrocks, and Karine G Le Roch. Deciphering the ubiquitin-mediated pathway in apicomplexan parasites: a potential strategy to interfere with parasite virulence. *PLoS ONE*, 3(6): e2386, Jan 2008. doi: 10.1371/journal.pone.0002386. URL http://www.plosone.org/article/info%253Adoi%252F10.1371%252Fjournal.pone.0002386.

[134] Bruno Pradines, Philippe Hovette, Thierry Fusai, Henri Léonard Atanda, Eric Baret, Philippe Cheval, Joel Mosnier, Alain Callec, Julien Cren, Rémy Amalvict, Jean Pierre Gardair, and Christophe Rogier. Prevalence of in vitro resistance to eleven standard or new antimalarial drugs among Plasmodium falciparum isolates from Pointe-Noire, Republic of the Congo. *Journal of Clinical Microbiology*, 44(7):2404–2408, July 2006. doi: 10.1128/JCM.00623-06. URL http://jcm.asm.org/content/44/7/2404.

[135] Alkes L Price, Nick J Patterson, Robert M Plenge, Michael E Weinblatt, Nancy A Shadick, and David E Reich. Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*, 38(8):904–909, August 2006. doi: 10.1038/ng1847. URL http://www.nature.com/ng/journal/v38/n8/abs/ng1847.html.

[136] Alkes L Price, Noah A Zaitlen, David E Reich, and Nick Patterson. New approaches to population stratification in genome-wide association studies. *Nature Reviews Genetics*, 11(7):459–63, Jul 2010. doi: 10.1038/nrg2813. URL http://www.nature.com/nrg/journal/v11/n7/full/nrg2813.html.

[137] Ric N Price, Anne-Catrin Uhlemann, Alan Brockman, Rose McGready, Elizabeth Ashley, Lucy Phaipun, Rina Patel, Kenneth Laing, Sornchai Looareesuwan, Nicholas J White, François Nosten, and Sanjeev Krishna. Mefloquine resistance in Plasmodium falciparum and increased pfmdr1 gene copy number. *Lancet*, 364(9432):438–447, August 2004. doi: 10.1016/S0140-6736(04)16767-6. URL http://www.sciencedirect.com/science/article/pii/S0140673604167676.

[138] Sean T Prigge, Xin He, Lucia Gerena, Norman C Waters, and Kevin A Reynolds. The initiating steps of a type ii fatty acid synthase in plasmodium falciparum are catalyzed by pfacp, pfmcat, and pfkasiii. *Biochemistry*, 42(4):1160–9, Feb 2003. doi: 10.1021/bi026847k. URL http://pubs.acs.org/doi/abs/10.1021/bi026847k.

[139] Jonathan K Pritchard, Matthew Stephens, Noah A Rosenberg, and Peter Donnelly. Association Mapping in Structured Populations. *American Journal of Human Genetics*, 67(1):12–12, July 2000. doi: 10.1086/302959. URL http://www.cell.com/AJHG/retrieve/pii/S0002929707624422.

[140] Sara L Pulit, Benjamin F Voight, and Paul I W de Bakker. Multiethnic genetic association studies improve power for locus discovery. *PLoS ONE*, 5(9):e12600, 2010. doi: 10.1371/journal.pone.0012600. URL http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0012600.

[141] Tepanata Pumpaibool, Céline Arnathau, Patrick Durand, Naowarat Kanchanakhan, Napaporn Siripoon, Aree Suegorn, Chitr Sitthi-amorn, François Renaud, and Pongchai Harnyuttanakorn. Genetic diversity and population structure of Plasmodium falciparum in Thailand, a low transmission country. *Malaria Journal*, 8:155–155, January 2009. doi:

10.1186/1475-2875-8-155. URL http://www.malariajournal.com.ezp-prod1.hul.harvard.edu/content/8/1/155.

[142] Shaun Purcell, Benjamin M Neale, Kathe Todd-Brown, Lori Thomas, Manuel A R Ferreira, David Bender, Julian Maller, Pamela Sklar, Paul I W de Bakker, Mark J Daly, and Pak C Sham. Plink: a tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics*, 81(3):559–75, Sep 2007. doi: 10.1086/519795. URL http://linkinghub.elsevier.com/retrieve/pii/S0002-9297(07)61352-4.

[143] Nusrat Rabbee and Terence P Speed. A genotype calling algorithm for affymetrix SNP arrays. *Bioinformatics*, 22(1):7–12, 2006. doi: 10.1093/bioinformatics/bti741. URL http://bioinformatics.oxfordjournals.org/cgi/content/full/22/1/7.

[144] Anna Ramírez-Soriano and Rasmus Nielsen. Correcting estimators of theta and tajima's d for ascertainment biases caused by the single-nucleotide polymorphism discovery process. *Genetics*, 181(2):701–10, Feb 2009. doi: 10.1534/genetics.108.094060. URL http://www.genetics.org/cgi/content/full/181/2/701.

[145] Ulf Ribacke, Mackenzie Bartlett, Saurabh D Patel, Niroshini Senaratne, Daniel J Park, Manoj T Duraisingh, Pardis C Sabeti, Sarah K Volkman, and Dyann F Wirth. Adaptive evolution of a ubiquitin ligase is linked to altered drug sensitivity in plasmodium falciparum. *Science Translational Medicine*. Manuscript under review.

[146] Ulf Ribacke, Bobo W Mok, Valtteri Wirta, Johan Normark, Joakim Lundeberg, Fred Kironde, Thomas G Egwang, Peter Nilsson, and Mats Wahlgren. Genome wide gene amplifications and deletions in plasmodium falciparum. *Molecular and Biochemical Parasitology*, 155(1):33–44, Sep 2007. doi: 10.1016/j.molbiopara.2007.05.005. URL http://dx.doi.org/10.1016/j.molbiopara.2007.05.005.

[147] D J Roberts, A G Craig, A R Berendt, R Pinches, G Nash, K Marsh, and C I Newbold. Rapid switching to multiple antigenic and adhesive phenotypes in malaria. *Nature*, 357(6380):689–692, June 1992. doi: 10.1038/357689a0. URL http://www.nature.com/nature/journal/v357/n6380/abs/357689a0.html.

[148] Juliana Martha Sá, Olivia Twu, Karen Hayton, Sahily Reyes, Michael P Fay, Pascal Ringwald, and Thomas E Wellems. Geographic patterns of

Plasmodium falciparum drug resistance distinguished by differential responses to amodiaquine and chloroquine. *Proceedings of the National Academy of Sciences, USA*, 106(45):18883–18889, November 2009. doi: 10.1073/pnas.0911317106. URL http://www.pnas.org/cgi/doi/10.1073/pnas.0911317106.

[149] Pardis C Sabeti. Positive Natural Selection in the Human Lineage. *Science*, 312(5780):1614–1620, June 2006. doi: 10.1126/science.1124309. URL http://www.sciencemag.org/cgi/doi/10.1126/science.1124309.

[150] Pardis C Sabeti, David E Reich, John M Higgins, Haninah Z P Levine, Daniel J Richter, Stephen F Schaffner, Stacey B Gabriel, Jill V Platko, Nick J Patterson, Gavin J McDonald, Hans C Ackerman, Sarah J Campbell, David M Altshuler, Richard Cooper, Dominic Kwiatkowski, Ryk Ward, and Eric S Lander. Detecting recent positive selection in the human genome from haplotype structure. *Nature*, 419(6909):832–7, Oct 2002. doi: 10.1038/nature01140. URL http://www.nature.com/nature/journal/v419/n6909/full/nature01140.html.

[151] Pardis C Sabeti, Patrick Varilly, Ben Fry, Jason Lohmueller, Elizabeth Hostetter, Chris Cotsapas, Xiaohui Xie, Elizabeth H Byrne, Rachelle Gaudet, Stephen F Schaffner, Eric S Lander, and International HapMap Consortium. Genome-wide detection and characterization of positive selection in human populations. *Nature*, 449(7164):913–8, Oct 2007. doi: 10.1038/nature06250. URL http://www.nature.com/nature/journal/v449/n7164/full/nature06250.html.

[152] Cecilia P Sanchez, Sybille Mayer, Astutiati Nurhasanah, Wilfred D Stein, and Michael Lanzer. Genetic linkage analyses redefine the roles of PfCRT and PfMDR1 in drug accumulation and susceptibility in Plasmodium falciparum. *Molecular Microbiology*, 82(4):865–878, November 2011. doi: 10.1111/j.1365-2958.2011.07855.x. URL http://dx.doi.org/10.1111/j.1365-2958.2011.07855.x.

[153] Amar Bir Singh AB Sidhu, Anne-Catrin AC Uhlemann, Stephanie G SG Valderramos, Juan-Carlos JC Valderramos, Sanjeev S Krishna, and David A DA Fidock. 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, August 2006. doi: 10.1086/507115. URL http://jid.oxfordjournals.org/content/194/4/528.

[154] Subhash Singh, Soe Soe, Simon Weisman, John W Barnwell, Jean Louis Pérignon, and Pierre Druilhe. A conserved multi-gene family induces cross-reactive antibodies effective in defense against plasmodium falciparum. *PLoS ONE*, 4(4):e5410, Jan 2009. doi: 10.1371/journal.pone.0005410. URL http://www.plosone.org/article/info%253Adoi%252F10.1371%252Fjournal.pone.0005410.

[155] J M Smith and J Haigh. The hitch-hiking effect of a favourable gene. *Genetical Research*, 23(1):23–35, February 1974.

[156] Matthew Stephens and Peter Donnelly. A comparison of bayesian methods for haplotype reconstruction from population genotype data. *American Journal of Human Genetics*, 73(5):1162–9, Nov 2003. doi: 10.1086/379378. URL http://www.sciencedirect.com/science/article/pii/S0002929707619788.

[157] X Su, M T Ferdig, Y Huang, C Q Huynh, A Liu, J You, J C Wootton, and T E Wellems. A genetic map and recombination parameters of the human malaria parasite Plasmodium falciparum. *Science*, 286(5443):1351–1353, November 1999. doi: 10.1126/science.286.5443.1351. URL http://www.sciencemag.org/cgi/content/full/286/5443/1351.

[158] Xin Zhuan Su, Jianbing Mu, and Deirdre A Joy. The "malaria's eve" hypothesis and the debate concerning the origin of the human malaria parasite plasmodium falciparum. *Microbes and Infection*, 5(10):891–6, Aug 2003. doi: 10.1016/S1286-4579(03)00173-4. URL http://dx.doi.org/10.1016/S1286-4579(03)00173-4.

[159] Shannon L Takala, Drissa Coulibaly, Mahamadou A Thera, Alassane Dicko, David L Smith, Ando B Guindo, Abdoulaye K Kone, Karim Traoré, Amed Ouattara, Abdoulaye A Djimde, Paul S Sehdev, Kirsten E Lyke, Dapa A Diallo, Ogobara K Doumbo, and Christopher V Plowe. Dynamics of polymorphism in a malaria vaccine antigen at a vaccine-testing site in Mali. *PLoS Medicine*, 4(3):e93–e93, March 2007. doi: 10.1371/journal.pmed.0040093. URL http://www.plosmedicine.org/article/info:doi/10.1371/journal.pmed.0040093.

[160] Shannon Takala-Harrison, Taane G Clark, Christopher G Jacob, Michael P Cummings, Olivo Miotto, Arjen M Dondorp, Mark M Fukuda, François Nosten, Harald Noedl, Mallika Imwong, Delia Bethell, Youry Se, Chanthap Lon, Stuart D Tyner, David L Saunders, Duong Socheat,

Frédéric Ariey, Aung Pyae Phyo, Peter Starzengruber, Hans-Peter Fuehrer, Paul Swoboda, Kasia Stepniewska, Jennifer Flegg, Cesar Arze, Gustavo C Cerqueira, Joana C Silva, Stacy M Ricklefs, Stephen F Porcella, Robert M Stephens, Matthew Adams, Leo J Kenefic, Susana Campino, Sarah Auburn, Bronwyn MacInnis, Dominic P Kwiatkowski, Xin Zhuan Su, Nicholas J White, Pascal Ringwald, and Christopher V Plowe. Genetic loci associated with delayed clearance of Plasmodium falciparum following artemisinin treatment in Southeast Asia. *Proceedings of the National Academy of Sciences of the United States of America*, pages 240–245, January 2013. doi: 10.1073/pnas.1211205110. URL http://www.pnas.org/content/110/1/240.long.

[161] John C Tan, Jigar J Patel, Asako Tan, J Craig Blain, Tom J Albert, Neil F Lobo, and Michael T Ferdig. Optimizing comparative genomic hybridization probes for genotyping and snp detection in plasmodium falciparum. *Genomics*, 93(6):543–50, Jun 2009. doi: 10.1016/j.ygeno.2009.02.007. URL http://www.sciencedirect.com/science/article/pii/S088875430900055X.

[162] Mohammed Tarique, Akash Tripathi Satsangi, Moaz Ahmad, Shailja Singh, and Renu Tuteja. Plasmodium falciparum mlh is schizont stage specific endonuclease. *Molecular and Biochemical Parasitology*, 181(2): 153–61, Feb 2012. doi: 10.1016/j.molbiopara.2011.10.012. URL http://www.sciencedirect.com/science/article/pii/S0166685111002684.

[163] Mahamadou A Thera, Ogobara K Doumbo, Drissa Coulibaly, Matthew B Laurens, Amed Ouattara, Abdoulaye K Kone, Ando B Guindo, Karim Traoré, Idrissa Traore, Bourema Kouriba, Dapa A Diallo, Issa Diarra, Modibo Daou, Amagana Dolo, Youssouf Tolo, Mahamadou S Sissoko, Amadou Niangaly, Mady Sissoko, Shannon Takala-Harrison, Kirsten E Lyke, Yukun Wu, William C Blackwelder, Olivier Godeaux, Johan Vekemans, Marie-Claude Dubois, W Ripley Ballou, Joe Cohen, Darby Thompson, Tina Dube, Lorraine Soisson, Carter L Diggs, Brent House, David E Lanar, Sheetij Dutta, D Gray Heppner, and Christopher V Plowe. A field trial to assess a blood-stage malaria vaccine. *New England Journal of Medicine*, 365(11):1004–1013, September 2011. doi: 10.1056/NEJMoa1008115. URL http://www.nejm.org/doi/full/10.1056/NEJMoa1008115.

[164] Timothy Thornton and Mary Sara McPeek. ROADTRIPS: case-control association testing with partially or completely unknown population and

pedigree structure. *American Journal of Human Genetics*, 86(2):172–184, February 2010. doi: 10.1016/j.ajhg.2010.01.001. URL http://www.cell.com/AJHG/retrieve/pii/S0002929710000029.

[165] W Trager and J B Jensen. Human malaria parasites in continuous culture. *Science*, 193(4254):673–5, Aug 1976.

[166] Jean-François Trape, Adama Tall, Nafissatou Diagne, Ousmane Ndiath, Alioune B Ly, Joseph Faye, Fambaye Dieye-Ba, Clémentine Roucher, Charles Bouganali, Abdoulaye Badiane, Fatoumata Diene Sarr, Catherine Mazenot, Aïssatou Touré-Baldé, Didier Raoult, Pierre Druilhe, Odile Mercereau-Puijalon, Christophe Rogier, and Cheikh Sokhna. Malaria morbidity and pyrethroid resistance after the introduction of insecticide-treated bednets and artemisinin-based combination therapies: a longitudinal study. *The Lancet Infectious Diseases*, 11(12):925–932, December 2011. doi: 10.1016/S1473-3099(11)70194-3. URL http://www.sciencedirect.com/science/article/pii/S1473309911701943.

[167] Tony Triglia, Manoj T Duraisingh, Robert T Good, and Alan F Cowman. Reticulocyte-binding protein homologue 1 is required for sialic acid-dependent invasion into human erythrocytes by Plasmodium falciparum. *Molecular microbiology*, 55(1):162–174, January 2005. doi: 10.1111/j.1365-2958.2004.04388.x. URL http://onlinelibrary.wiley.com/doi/10.1111/j.1365-2958.2004.04388.x/abstract.

[168] A B Vaidya, O Muratova, F Guinet, D Keister, T E Wellems, and D C Kaslow. A genetic locus on Plasmodium falciparum chromosome 12 linked to a defect in mosquito-infectivity and male gametogenesis. *Molecular and Biochemical Parasitology*, 69(1):65–71, January 1995. doi: 10.1016/0166-6851(94)00199-W. URL http://dx.doi.org/10.1016/0166-6851(94)00199-W.

[169] Daria Van Tyne, Daniel J Park, and Dyann F Wirth. Understanding malaria drug resistance evolution in real-time. *Trends in Parasitology*. Manuscript in preparation.

[170] Daria Van Tyne, Daniel J Park, Stephen F Schaffner, Daniel E Neafsey, Elaine Angelino, Joseph F Cortese, Kayla G Barnes, David M Rosen, Amanda K Lukens, Rachel F Daniels, Danny A Milner, Charles A Johnson, Ilya Shlyakhter, Sharon R Grossman, Justin S Becker, Daniel Yamins, Elinor K Karlsson, Daouda Ndiaye, Ousmane Sarr, Souleymane Mboup, Christian Happi, Nicholas A Furlotte, Eleazar Eskin, Hyun Min

Kang, Daniel L Hartl, Bruce W Birren, Roger C Wiegand, Eric S Lander, Dyann F Wirth, Sarah K Volkman, and Pardis C Sabeti. Identification and functional validation of the novel antimalarial resistance locus *pf10_0355* in *plasmodium falciparum*. *PLoS Genetics*, 7(4):e1001383, Apr 2011. doi: 10.1371/journal.pgen.1001383. URL http://www.plosgenetics.org/article/info%3Adoi%2F10.1371%2Fjournal.pgen.1001383.

[171] Meera Venkatesan, Chanaki Amaratunga, Susana Campino, Sarah Auburn, Oliver Koch, Pharath Lim, Sambunny Uk, Duong Socheat, Dominic P Kwiatkowski, Rick M Fairhurst, and Christopher V Plowe. Using cf11 cellulose columns to inexpensively and effectively remove human dna from plasmodium falciparum-infected whole blood samples. *Malaria Journal*, 11:41, Jan 2012. doi: 10.1186/1475-2875-11-41. URL http://www.malariajournal.com/content/11/1/41.

[172] H J Vial, M J Thuet, and J R Philippot. Cholinephosphotransferase and ethanolaminephosphotransferase activities in plasmodium knowlesi-infected erythrocytes. their use as parasite-specific markers. *Biochimica et Biophysica Acta*, 795(2):372–83, Sep 1984. URL http://www.sciencedirect.com/science/article/pii/0005276084900882.

[173] Benjamin F Voight, Sridhar Kudaravalli, Xiaoquan Wen, and Jonathan K Pritchard. A Map of Recent Positive Selection in the Human Genome. *PLoS Biology*, 4(3):e72, March 2006. doi: 10.1371/journal.pbio.0040072. URL http://www.plosbiology.org/article/info%3Adoi%2F10.1371%2Fjournal.pbio.0040072.

[174] Sarah K Volkman, Pardis C Sabeti, David DeCaprio, Daniel E Neafsey, Stephen F Schaffner, Danny A Milner, Johanna P Daily, Ousmane Sarr, Daouda Ndiaye, Omar Ndir, Souleymane Mboup, Manoj T Duraisingh, Amanda Lukens, Alan Derr, Nicole Stange-Thomann, Skye Waggoner, Robert C Onofrio, Liuda Ziaugra, Evan Mauceli, Sante Gnerre, David B Jaffe, Joanne Zainoun, Roger C Wiegand, Bruce W Birren, Daniel L Hartl, James E Galagan, Eric S Lander, and Dyann F Wirth. A genome-wide map of diversity in plasmodium falciparum. *Nature Genetics*, 39(1):113–9, Jan 2007. doi: 10.1038/ng1930. URL http://www.nature.com/ng/journal/v39/n1/abs/ng1930.html.

[175] Sarah K Volkman, Daouda Ndiaye, Mahamadou Diakite, Ousmane A Koita, Davis Nwakanma, Rachel F Daniels, Daniel J Park, Daniel E Neafsey, Marc A T Muskavitch, Donald J Krogstad, Pardis C Sabeti, Daniel L Hartl, and Dyann F Wirth. Application of genomics to field

investigations of malaria by the international centers of excellence for malaria research. *Acta Tropica*, 121(3):324–332, March 2012. doi: 10.1016/j.actatropica.2011.12.002. URL http://dx.doi.org/10.1016/j.actatropica.2011.12.002.

[176] Sarah K Volkman, Daniel E Neafsey, Stephen F Schaffner, Daniel J Park, and Dyann F Wirth. Harnessing genomics and genome biology to understand malaria biology. *Nature Reviews Genetics*, 13(5):315–328, May 2012. doi: 10.1038/nrg3187. URL http://www.nature.com/nrg/journal/v13/n5/full/nrg3187.html.

[177] H K Webster, E F Boudreau, K Pavanand, K Yongvanitchit, and L W Pang. Antimalarial drug susceptibility testing of Plasmodium falciparum in Thailand using a microdilution radioisotope method. *American Journal of Tropical Medicine and Hygiene*, 34(2):228–235, March 1985. URL http://www.ajtmh.org/content/34/2/228.extract.

[178] Gareth D Weedall and David J Conway. Detecting signatures of balancing selection to identify targets of anti-parasite immunity. *Trends in Parasitology*, 26(7):363–369, July 2010. doi: 10.1016/j.pt.2010.04.002. URL http://dx.doi.org/10.1016/j.pt.2010.04.002.

[179] T E Wellems, A Walker-Jonah, and L J Panton. Genetic mapping of the chloroquine-resistance locus on Plasmodium falciparum chromosome 7. *Proceedings of the National Academy of Sciences, USA*, 88(8):3382–3386, April 1991. doi: 10.1073/pnas.88.8.3382. URL http://www.pnas.org/content/88/8/3382.abstract.

[180] Scott J Westenberger, Colleen M McClean, Rana Chattopadhyay, Neekesh V Dharia, Jane M Carlton, John W Barnwell, William E Collins, Stephen L Hoffman, Yingyao Zhou, Joseph M Vinetz, and Elizabeth A Winzeler. A systems-based analysis of Plasmodium vivax lifecycle transcription from human to mosquito. *PLoS Neglected Tropical Diseases*, 4 (4):e653–e653, January 2010. doi: 10.1371/journal.pntd.0000653. URL http://www.plosntds.org/article/info%3Adoi%2F10.1371%2Fjournal.pntd.0000653.

[181] M C Whitlock. Combining probability from independent tests: the weighted Z-method is superior to Fisher's approach. *Journal of Evolutionary Biology*, 18(5):1368–1373, September 2005. doi: 10.1111/j.1420-9101.2005.00917.x. URL http://onlinelibrary.wiley.com/doi/10.1111/j.1420-9101.2005.00917.x/abstract.

[182] Samuel S Wilks. The Large-Sample Distribution of the Likelihood Ratio for Testing Composite Hypotheses. *Annals of Mathematical Statistics*, 9 (1):60–62, March 1938. doi: 10.1214/aoms/1177732360. URL http://projecteuclid.org/euclid.aoms/1177732360.

[183] C M Wilson, A E Serrano, A Wasley, M P Bogenschutz, A H Shankar, and Dyann F Wirth. Amplification of a gene related to mammalian mdr genes in drug-resistant Plasmodium falciparum. *Science*, 244(4909):1184–1186, June 1989. URL http://www.sciencemag.org/content/244/4909/1184.long.

[184] John C Wootton, Xiaorong Feng, Michael T Ferdig, Roland A Cooper, Jianbing Mu, Dror I Baruch, Alan J Magill, and Xin-zhuan Su. Genetic diversity and chloroquine selective sweeps in plasmodium falciparum. *Nature*, 418(6895):320–3, Jul 2002. doi: 10.1038/nature00813. URL http://www.nature.com/nature/journal/v418/n6895/full/nature00813.html.

[185] Jing Yuan, Ken Chih-Chien Cheng, Ronald L Johnson, Ruili Huang, Sittiporn Pattaradilokrat, Anna Liu, Rajarshi Guha, David A Fidock, James Inglese, Thomas E Wellems, Christopher P Austin, and Xin-zhuan Su. Chemical genomic profiling for antimalarial therapies, response signatures, and molecular targets. *Science*, 333(6043):724–9, Aug 2011. doi: 10.1126/science.1205216. URL http://www.sciencemag.org/content/333/6043/724.short.

[186] D V Zaykin. Optimally weighted Z-test is a powerful method for combining probabilities in meta-analysis. *Journal of Evolutionary Biology*, 24(8):1836–1841, August 2011. doi: 10.1111/j.1420-9101.2011.02297.x. URL http://onlinelibrary.wiley.com/doi/10.1111/j.1420-9101.2011.02297.x/abstract.

Colophon

This thesis was typeset using LaTeX, originally developed by Leslie Lamport and based on Donald Knuth's TeX. The body text is set in Arno Pro, designed by Robert Slimbach in the style of book types from the Aldine Press in Venice, and issued by Adobe in 2007. The original version of this template, which can be used to format a PhD thesis with this look and feel, has been released under the permissive MIT (X11) license, and can be found online at github.com/suchow/. The author of this dissertation can be reached at dannypark@alum.mit.edu.