ACCESSING COMPLEX GENOMIC VARIATION IN *PLASMODIUM FALCIPARUM* NATURAL INFECTIONS

Jason Patrick Wendler

Genomic Medicine and Statistics Programme
Nuffield Department of Clinical Medicine
University of Oxford

A thesis submitted for the degree of *Doctor of Philosophy*
Wolfson College Trinity 2015

ABSTRACT

Accessing complex genomic variation in *Plasmodium falciparum* natural infections

Jason Patrick Wendler, Wolfson College, University of Oxford
Submitted for Doctor of Philosophy, Trinity term 2015

Genetic polymorphism in *Plasmodium falciparum* is a considerable obstacle to malaria intervention. Parasites have repeatedly evolved to overcome every front-line antimalarial deployed throughout history, and artemisinin resistant populations are expanding in Southeast Asia. Promising vaccine candidates routinely fail when challenged by the genetic diversity of natural parasite populations, and a recent trial using a blood-stage antigen showed immunity was allele specific. Modern sequencing technologies have revolutionized our understanding of parasite genomics and population genetics by providing access to single nucleotide variation, but characterizing more complex polymorphism remains a key challenge. Solving this problem is important because the selective pressures from drugs and host immunity often create complex polymorphism in the most clinically relevant genes that is missed using standard genotyping methods. In three sections, this thesis is a narrative about 1) encountering complex variation, 2) overcoming it with novel tools, and then 3) innovatively applying those tools to old and new questions. I first show examples of complex variation in a vaccine candidate (*EBA-175)* and a drug resistance gene (*pfcrt*) while reporting SNP based analyses of Kenyan and Tanzanian field isolates. While introducing this complex variation I also describe biological insights discovered in these populations. In Kenya I show evidence that chloroquine resistance selects for parasites that are primaquine sensitive, use a GWAS approach to discover new drug resistance loci, and catalogue variation in known resistance genes. In Tanzania I describe the population structure and allele frequencies of parasites from two geographic regions. In the second section of the thesis I develop methods for accessing complex variation and demonstrate their utility by producing *de novo* assemblies of *eba-175, pfcrt, ama1,* and *msp3.4* from thousands of sequenced samples. Finally, in the third section I apply these tools in depth to *eba-175*. I comprehensively characterize the SNP and structural variation in *eba-175* using an alignment of 1419 *de novo* assemblies. I use this resource to illustrate the profiles of positive selection across the gene, and corroborate these signals of balancing selection by showing the geographic distribution of the F/C indels and a lesser known 6bp indel positioned between the DBL domains. I then use the alignments to design Sequenom genotyping assays that facilitate a genome wide association study, testing for human associations with the *eba-175* indels in the infecting parasite. I close by reporting a potential association on human chromosome 14 with the 6bp indel in *eba-175.*

ACKNOWLEDGEMENTS

This DPhil is dedicated to the study participants who donate malaria infected blood for our work.

I would like to thank my supervisors, Dominic Kwiatkowski and Philip Bejon, for their guidance and patience. Witnessing Dominic direct his group has been an invaluable lesson in both science and leadership. I would also like to thank Kirk Rockett for his wisdom and direction throughout my PhD, but especially in the early days. I learned much about malaria and even more about Britain on our drives between Oxford and Sanger. My utmost gratitude goes to Vikki Cornelius for her incredible sincerity, perspective, common-sense, and ability to fix anything. We are so lucky to have Vikki in the Kwiatkowski group. Thank you to Jennifer Shelton and Antoine Claessens for taking the time to carefully review my thesis. There are many other impressive people in our large group who deserve to be named—Christina, Gavin, Quang, and others are mentioned for their contributions in chapter-specific acknowledgments. Thank you all.

I sincerely thank Patrick Duffy and Michal Fried for the excellent training opportunities over the years, which have been central to my achievements.

I am grateful to Jonathan Flint and Richard Mott for the opportunity to study in the Genomic Medicine and Statistics Programme, and to LMIV/NIAID/NIH and the Wellcome Trust for generous funding.

Finally, I thank my beautiful little family, Rachel and Patrick, for agreeing to sell everything we owned and move half a world away from home to support my PhD. Nothing is more important than family.

DECLARATION

This thesis represents my own original work. Areas of collaboration and contributions from others are acknowledged in each chapter.

In accordance with University guidelines, the main text of this thesis does not exceed 50,000 words.

Jason Patrick Wendler

PUBLICATIONS AND PRESENTATIONS OF THIS WORK

Publications

Wendler JP, Okombo J, et al. A genome wide association study of *Plasmodium falciparum* susceptibility to 22 antimalarial drugs in Kenya. PLoS One. 2014 May 8;9(5):e96486. doi: 10.1371/journal.pone.0096486.

Talks

EBA-175 variation. 2013. Genomic Epidemiology of Malaria meeting.

Malaria: epidemiology, genomics, & transcriptomics. 2012. Outreach for secondary students I hosted at the WTCHG. Also included a brief microscopy lab.

Posters

Assembling full-length gene sequences from field samples with mixed infections to comprehensively assess the variation in merozoite invasion proteins. 2014. Genomic Epidemiology of Malaria meeting.

Comprehensive assessment of variation in the invasion ligand, *eba-175*, by *de novo* assembly of 600 worldwide field isolates of *Plasmodium falciparum*. 2013. 6th MIM Pan-African Malaria Conference.

Parallel *de novo* assembly of full-length *eba-175* from parasites sequenced using short-reads. 2013. 9th Annual BioMalPar | EVIMalaR conference on the Biology and Pathology of the Malaria Parasite.

Candidate antigen discovery in severe malaria; leveraging the parasite genome in malaria vaccinology. 2012. 3rd prize at the Oxford MSDTC graduate research symposium.

From genomes to genes: what 2000 *Plasmodium falciparum* genomes can tell you about your favorite genes. 2012. Poster and mini-talk. 8th Annual BioMalPar | EVIMalaR conference on the Biology and Pathology of the Malaria Parasite.

From genomes to genes: Seeking feedback from the community as we develop MalariaGEN candidate gene reports. 2012. Poster and mini-talk. Genomic Epidemiology of Malaria meeting.

Association of host and parasite gene expression with outcomes of Tanzanian children with malaria. 2012. Molecular Approaches to Malaria.

Genomic variation in malaria parasites from Tanzanian mothers and children. 2011. Gordon Research Conference.

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS AND ACRONYMS

BWA	Burrows-Wheeler Aligner
CM	Cerebral Malaria
Country: BD	Bangladesh
Country: BF	Burkina Faso
Country: BJ	Benin
Country: BR	Brazil
Country: CM	Cameroon
Country: CO	Colombia
Country: EAF	East Africa
Country: GB	United Kingdom
Country: GH	Ghana
Country: GM	Gambia, The
Country: GN	Guinea
Country: GW	Guinea-Bissau
Country: HN	Honduras
Country: ID	Indonesia
Country: IN	India
Country: KE	Kenya
Country: KH	Cambodia
Country: KR	Korea, South
Country: LA	Laos
Country: LK	Sri Lanka
Country: ML	Mali
Country: MM	Myanmar (Burma)
Country: MW	Malawi
Country: MY	Malaysia
Country: MZ	Mozambique
Country: NG	Nigeria
Country: PE	Peru
Country: PG	PNG Papua New Guinea
Country: SD	Sudan
Country: SEA	South East Asia
Country: TH	Thailand
Country: TZ	Tanzania
Country: UG	Uganda
Country: US	United States
Country: VN	Vietnam
Country: WAF	West Africa
Country: ZW	Zimbabwe
DBL	Duffy Binding Ligand
DBP	Duffy Binding Protein
eba175, EBA-175	Erythrocyte Binding Antigen 175
EIR	Entomological Inoculation Rate

ENA	European Nucleotide Archive
FDR	False Discovery Rate
GIA	Growth Inhibitory Assay
GWAS	Genome Wide Association Study
GYP	Glycophorin
iCORN	Iterative Correction of Reference Nucleotides
IGV	Integrated Genomics Viewer (Broad Institute)
kDa	Kilodalton
LD	Linkage Disequilibrium
MAF	Minor Allele Frequency
MalariaGEN	Malaria Genomic Epidemiology Network
MOMS	Mother Offspring Malaria Study
MSA	Multiple Sequence Alignment
NCBI	National Center for Biotechnology Information
NRAF	Non-Reference Allele Frequency
OLC	Overlap Consensus Layout
PfCRT, *pfcrt,* CRT	*Plasmodium falciparum* chloroquine resistance transporter
PfEMP1	*P. falciparum* Erythrocyte Membrane Protein 1
PfRh	*P. falciparum* Reticulocyte Binding homolog
PGV	Plasmodium Genome Variation project
QQ	Quantile-Quantile
RD	Respiratory Distress
Rh	reticulocyte-binding protein homologue
SM	Severe Malaria
SMA	Severe Malarial Anemia
SNP	Single Nucleotide Polymorphism
UEP	Universal Extension Primer
VCF	Variant Call Format

LIST OF APPENDICES

1 BACKGROUND AND RATIONALE

Any story about malaria is really a story about evolution, and evolution is largely a story about genes changing through time. For millions of years the bloodline from our common ancestors has been infected by the ancestors of the very parasites that plague humanity today [1]. As hosts evolved from one species to the next, the parasite was there, likely causing fevers and certainly effecting the change of genes through time—both its own and those of its host. This thesis is about accessing parasite genes, particularly those that the host has pressured the parasite to change considerably. It would be unsatisfying to write about methods that will soon be irrelevant when technology advances, therefore this DPhil is also about applying these tools to investigate scientific questions, all of which boil down to parasite and host genes changing through time.

1.1 Malaria

Malaria is a complex disease caused by a single-celled, eukaryotic parasite that reflects this complexity at every level. The disease is complex because it manifests as different syndromes in different infections (discussed in 1.1.2). The genome is complex because it encodes proteins that enable the parasite to assume several morphological forms, and to survive within and outside multiple tissues of two different organisms—i.e., a mosquito vector and a vertebrate host [2]. The parasite actually carries three genomes. In addition to nuclear and mitochondrial genomes, it harbors a plastid containing 35Kb of DNA—a relic of an ancient heterotroph that acquired an algal symbiont [3,4]. There are five species of the genus *Plasmodium* known to naturally infect humans (*falciparum, vivax, ovale, malariae, and knowlesi*), and they have remote phylogenetic histories that suggest independent adaptations to this host [5,6]. Many dozens of *Plasmodium* species are known to cause malaria in a range of hosts, including birds, mammals and reptiles, and each parasite is narrowly committed to similar hosts [7,8,9]. *P. knowlesi* infects humans primarily as a

zoonosis from macaque monkeys in Southeast Asia, and these cases were thought to be extremely rare until relatively recently [10,11]. *P. vivax* and *P. ovale* can form hypnozoites (sleeping forms) in the liver that can hide for months in hepatocytes before causing relapse, an important consideration for radical cures of malaria and for elimination campaigns [12,13,14]. *P. vivax* and *P. falciparum* are responsible for nearly all of the human burden from malaria [15]. The most deadly of these species is *P. falciparum*, which is also the subject of the discussion going forward.

1.1.1 Lifecycle: stages and interventions

From an intervention standpoint, the complex lifecycle of *P. falciparum* can be divided into blood, liver (pre-erythrocytic), and transmission stages. Stopping parasites before they reach the blood will prevent infection (and thus transmission). Once in the blood, interventions target disease or transmission. Details of these general divisions and relevant interventions are outlined below.

1.1.1.1 Pre-erythrocytic host infection

P. falciparum is transmitted during the blood meal of a female *Anopheles* mosquito, usually *A. gambiae,* when she inoculates the host with a dozen or so motile sporozoites from her salivary glands [16]. The haploid sporozoites quickly enter a blood vessel and migrate to the liver, where they each become ensconced within a hepatocyte, after ironically traversing modified white blood cells called Kupffer cells [17,18].

Targeting the pre-erythrocytic stage is appealing because of the extreme bottleneck in population size (tens of sporozoites and hepatocytes versus tens of billions in the blood stage), however if even one parasite survives it could multiply exponentially in the blood stage and cause disease. The most advanced malaria vaccine to date, RTS,S, targets the circumsporozoite protein that coats the surface of this motile developmental stage [19]. Although early trials of this vaccine in malaria naïve adults provided sterilizing efficacy of 30-50%, 4-year efficacies against first and all clinical episodes of malaria in a phase III field trial were 30% and 17%, respectively [20,21,22]. Immunizing with sporozoites that have been attenuated by radiation or genetic modification, such that they arrest development before the blood stage, can confer sterilizing protection from reinfection [23,24]. Arrest at later stages of development in the liver elicits increasingly stronger immunity, and whole-organism vaccine strategies combined with chemoprophylaxis (thus, clearing infection immediately after the liver stage) are being investigated as well [25,26]. In the latter

approach parasites do enter the blood for a short time, however protection is mediated by pre-erythrocytic immunity [27].

Most licensed antimalarial drugs are ineffective against liver-stage parasites, however primaquine does have this "causal prophylactic" property against *P. falciparum* [28]. Primaquine is also the only drug indicated for radical cure of *P. vivax* and *P. ovale*—i.e., for clearing hypnozoites from hepatocytes [29].

1.1.1.2 The blood stage

It is during the blood stage that malaria symptoms and disease develop [30]. After a little over a week in the liver, the blood stage begins when a hepatocyte bursts, releasing thousands of merozoites into the bloodstream [31,32]. Each merozoite invades an individual RBC that it reorganizes both internally and externally to facilitate progression through ring, trophozoite, and schizont stages over a period of 48 hours. During this period of growth and mitotic division within its own parasitophorous vacuole, the parasite must uptake nutrients by creating a plasmodial surface anion channel (PSAC), and digests host hemoglobin for amino acids [33,34]. Other species undergo this intraerythrocytic development cycle (IDC) in shorter or longer periods, which is the basis for the disease names based on the recurrence of fever, and this timing has implications for disease [35]. Unlike some other human parasites, *P. falciparum* infects reticulocytes and mature erythrocytes, facilitating higher parasitemias and contributing to disease [36].

The parasite utilizes a suite of proteins and redundant pathways to attach, reorient, and actively enter the host red blood cell [37]. Early contact with erythrocytes is mediated through the merozoites surface protein (MSP) family. MSP-1 is an essential glycosyl-phosphatidyl-inositol (GPI) anchored vaccine candidate that forms a complex on the surface of the merozoite and is evidenced to interact with band 3 (a recent study indicates it may also interact with glycophorin A) [38,39,40]. The parasite is committed to invasion after reorienting its apical end to face the erythrocyte surface, exposing RBC receptors to ligands already released from microneme and rhoptry organelles (triggered by increased calcium levels) [41]. Unlike MSP-1, many of the erythrocyte binding-like (EBL) and reticulocyte binding-like homologs (PfRh) released from the micronemes and rhoptries, respectively, are redundant, thus multiple invasion phenotypes are available at this step [42]. PfRh5 is an exception. This conserved protein binds indispensably to basigin on the red cell surface during invasion, and is a promising vaccine candidate in phase I trials [43,44,45]. Invasion phenotypes have been traditionally defined by sensitivity or resistance to trypsin,

chymotrypsin, and neuraminidase [42,46]. For example, the micronemal protein EBA-175 binds to sialic acid residues on GYPA during invasion, and is sensitive to neuraminidase- and trypsin-treated erythrocytes, but resistant to those pre-treated with chymotrypsin [47]. Finally, apical membrane antigen-1 (AMA1) binds to a rhoptry neck protein complex (RON) of parasite origin that was embedded into the RBC membrane, and the merozoite is pulled into the red cell by an actin-myosin motor, forming a vacuole within which it will develop in the erythrocyte (i.e., the parasitophorous vacuole) [37,48]. All of these merozoite ligands are candidate vaccine targets [49].

In addition to targeting proteins on the merozoite surface, host immunity and thus vaccine interventions also target antigens exported by the parasite to the surface of the infected erythrocyte. The parasite establishes a protein trafficking complex and restructures the surface of the RBC with dense granules called knobs, which display members of the *P. falciparum* erythrocyte membrane protein-1 (PfEMP1) family [50,51,52]. These proteins play a fundamental role in antigenic variation and disease, as described in more detail in 1.1.2, making them primary targets for vaccine intervention. However, they are also among the most difficult genes to study due to their hyper-variability and paralogy (see 1.2.3).

The mitotic divisions during schizogony can result in extremely high parasite densities in non-immune hosts, increasing parasitemia approximately 10-fold with every cycle. Clinical symptoms manifest when merozoites are released from schizonts in a synchronized fashion. To progress to another asexual cycle, merozoites must escape from the parasitophorous vacuole and RBC via protease activity, and recent vaccine targets have been identified that arrest schizont egress and thereby significantly attenuate growth [53].

Rather than continuing on to invade a fresh erythrocyte, a subset of parasites commit midway through the IDC to becoming male (micro) and female (macro) sexual forms called gametocytes [54,55]. These forms are a prime target for transmission blocking interventions, including drugs and vaccines (see 1.1.1.3) [56].

Most drug treatments for malaria target the blood-stage of infection. Current front-line antimalarials are artemisinin derivatives deployed in combination with partner drugs as artemisinin combination therapies (ACT) [57]. The metabolic processes targeted by the partner drugs are most often related to heme digestion and the folate pathway, and parasites have a consistent history of evolving resistance to these activities [58]. Most famously (discussed in detail in chapter 3), mutations in the *P. falciparum* chloroquine

resistance transporter (PfCRT) counteract chloroquine's ability to interrupt heme detoxification in the digestive vacuole.

1.1.1.3 The mosquito stage

Like the sporozoites that initiate infection, gametocytes taken up by the mosquito are logs-fold fewer in number than asexual forms [59]. As such they are not under strong immune pressure to diversify, and thus make attractive targets for transmission blocking vaccines (TBV) [60]. Vaccines that clear infection at any stage would impact transmission and in this regard are grouped with TBVs as vaccines that interrupt malaria transmission (VIMT) [61]. Pfs25 and Pfs230 are two examples of TBV in phase I trials. These vaccines work by eliciting antibodies that are taken up by the mosquito to target mosquito-stage parasite proteins, thus protecting the mosquito from infection, and thereby blocking transmission [56]. Pfs25 begins expression in the mosquito mid-gut, and therefore isn't under immune pressure, whereas Pfs230 begins expression in blood stage gametocytes. As a result, a Pfs25 vaccine would not be boosted by natural immunity, but is a highly conserved target [62]. In contrast, as expected from an antigen under immune pressure, Pfs230 is much more polymorphic.

While VIMTs are a long term intervention strategy, there are other methods that target transmission stages. Bed-nets treated with long-lasting insecticides (LLITN) and indoor residual spraying (IRS) have already had a tremendous impact on transmission in many communities [63,64]. These interventions rely on effective insecticides, and pyrethroid resistant mosquitos are a major concern throughout Africa [65,66]. Separately, in addition to targeting liver stage parasites, primaquine has gametocytocidal activity. Although this drug can cause hemolysis in glucose-6-phosphate-dehydrogenase (G6PD) deficient individuals, use of low-dose primaquine to control *P. falciparum* transmission in certain areas has been recommended by the WHO, even if point of care enzyme testing is unavailable in areas with high G6PD deficiency prevalence [67].

Once inside the mosquito mid-gut, microgametocytes exflagellate and sexually recombine with macrogametocytes [68]. The diploid ookinete penetrates the lining of the mid-gut and migrates to the salivary glands where it undergoes meiosis to generate haploid sporozoites—bringing the lifecycle full circle.

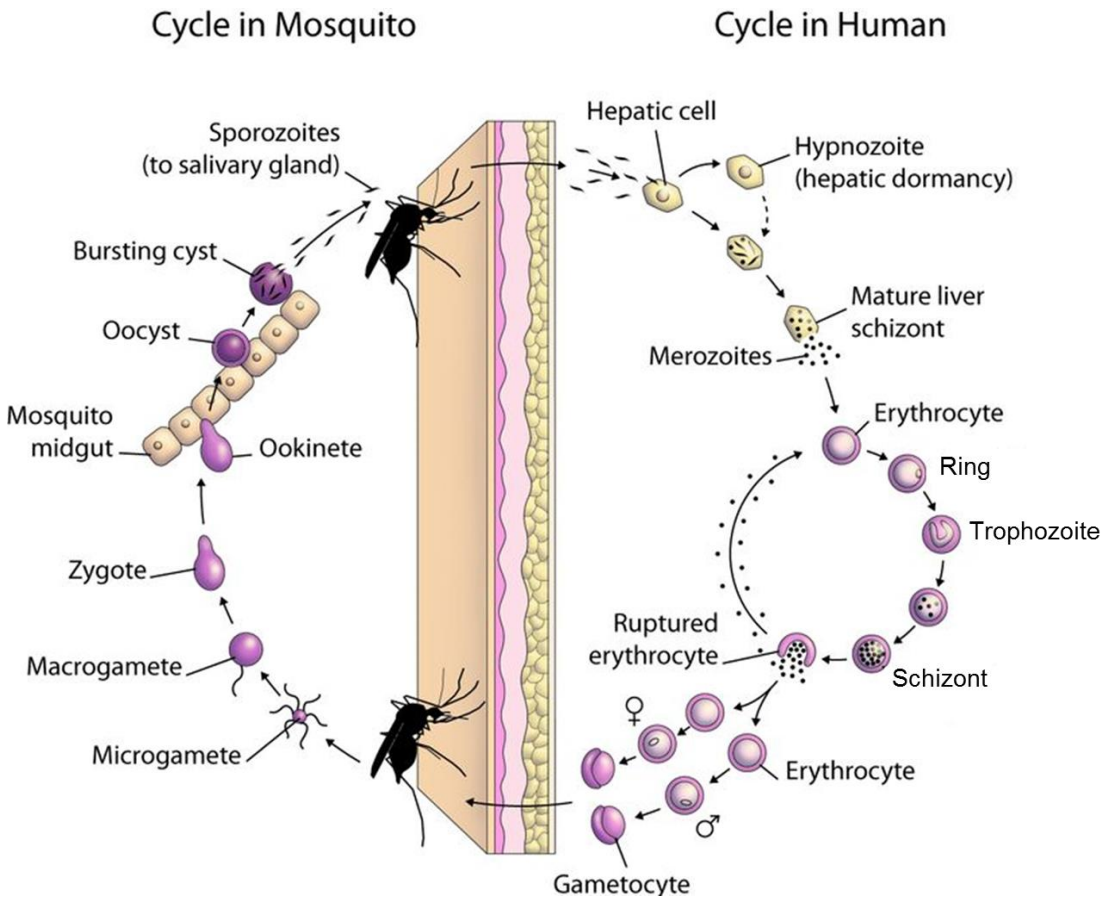


Figure 1-1. Lifecycle of the human malaria parasites. Modified under creative commons license from: Epidemiology of Infectious Diseases. Available at: http://ocw.jhsph.edu. Copyright © Johns Hopkins Bloomberg School of Public Health. Creative Commons BY-NC-SA.

1.1.2 Epidemiology and disease

Malaria will kill 600,000 children in Africa this year [69]. Hundreds of millions of clinical cases will occur in tropical zones across the globe, but the major burden occurs in African children under five and in first-time pregnant women [70]. This epidemiology is shaped by host immunity and transmission intensity [71]. One of the reasons for higher African transmission, and thus a more intense disease burden, is that *A. gambiae* is the most common vector there, which is more efficient and selective for humans compared to other mosquitos [72]. As illustrated in Figure 1-2, newborns in an area of stable malaria transmission are protected from disease for the first six months of life, partly by maternal antibodies [73]. Once this protection wanes they become susceptible to severe disease, which manifests as discrete syndromes (i.e., severe malarial anemia (SMA), cerebral malaria (CM), and respiratory distress (RD)) [74]. Only 1 in 100 symptomatic cases progress to severe disease. Severe anemia is more common in high transmission areas, and occurs more frequently in younger children than CM, which is more common in lower transmission

and seasonal malaria settings [75]. Immunity to severe disease is acquired rapidly, perhaps only after a single infection, and is not associated with parasitemia [76]. This is an important point because it signals the possibility that severe syndromes have a discrete pathogenesis, e.g., a specific receptor-ligand interaction, that could be targeted with a vaccine. Indeed, this appears to be exactly the case with another form of severe disease—placental malaria.

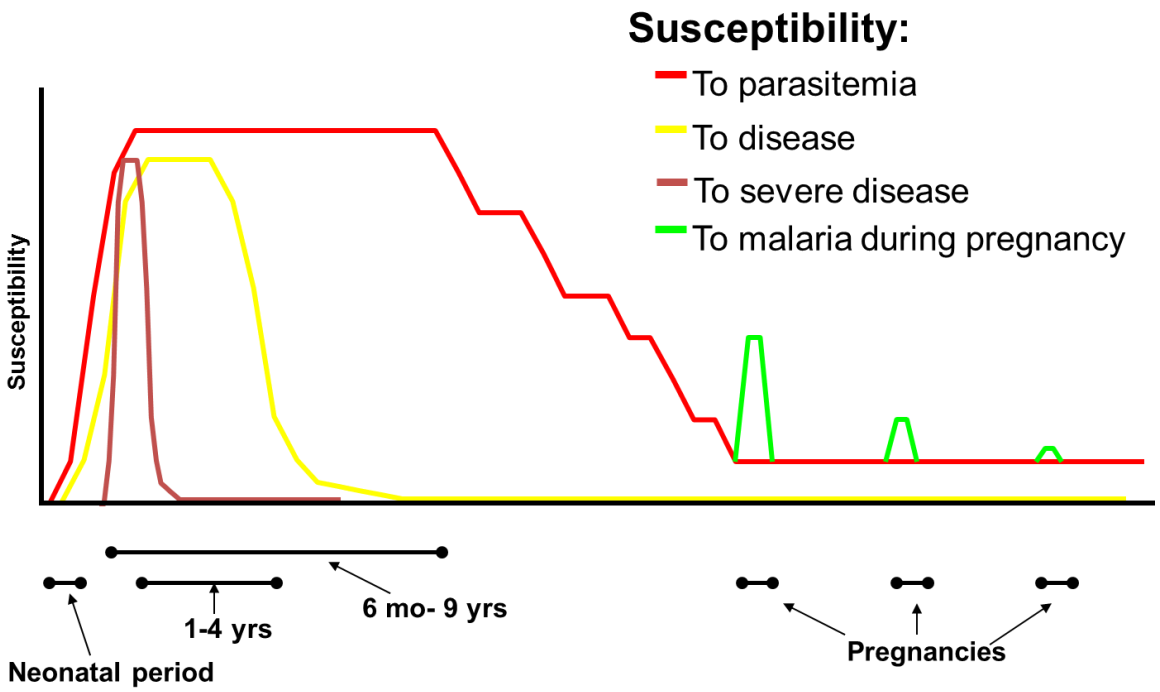


Figure 1-2. Epidemiology of *P. falciparum* malaria in an area of stable transmission. Figure kindly provided by Patrick Duffy.

Even after developing immunity that controls parasite density like their peers, women are again susceptible to disease and increased peripheral and placental parasitemias when they become pregnant for the first-time [77]. These infections can lead to severe anemia in the mother and abortion or low birthweight of her child, killing 100,000 infants each year and thousands of mothers [78,79]. As illustrated in Figure 1-2, mothers are most susceptible in their first two pregnancies, but acquire natural immunity over subsequent parities [80]. Infected erythrocytes adhere to the syncytiotrophoblast of the placenta via binding to the host receptor, chondroitin Sulfate A (CSA) [81]. This binding may be mediated by VAR2CSA, a PfEMP1 exported by the parasite to the red cell surface [82]. Sera from women who have acquired immunity to placental malaria over subsequent pregnancies effectively inhibit parasite binding to CSA *in vitro*. Significantly, the antibodies that block binding to CSA are pan-reactive—i.e., they block the binding activity of parasites from different geographical

regions [80]. This pan-reactive property is consistent with the hypothesis that the genes encoding the parasite ligand have regions under selective constraint that could be targeted by a vaccine with a single or limited number of components [83].

In lower transmission settings the epidemiology is different from than that of typical African communities with stable transmission because there is less immunity. In such settings disease occurs in all age groups, with greater risk of becoming severe, and presents as multi-organ disease, in contrast to the discrete syndromes described above [84].

1.2 Classes of genetic variation

The 23-megabase *P. falciparum* genome contains more than 5000 genes organized on 14 chromosomes [85]. A unique feature of this genome is the extremely high AT content (80.6% overall, and approximately 90% in introns and intergenic regions). Exacerbated by several forms of polymorphism, this extreme AT-bias creates difficulties for enzymes, sequencers, and bioinformatics tools, making *P. falciparum* a challenge to study in many respects.

1.2.1 SNPs and the "core genome"

As discussed below, and demonstrated in Section I, much of the genome is inaccessible using standard methods on short reads due to low complexity, hyper-polymorphism, and divergence. However, an extremely useful form of variation that is accessible in much of the exome is single nucleotide polymorphism. The Malaria Genomic Epidemiology Network (MalariaGEN) performed the first population-scale analysis of genome-wide, next generation sequence from malaria infected samples [86,87]. That work sought to define a panel of trustworthy SNPs to use for population genetic analyses. Extending standard methodologies, MalariaGEN ascertained SNPs by aligning short (~ 100bp) Illumina sequencing reads to the reference genome, and they determined that much of the genome should be excluded to avoid genotyping errors. For example, non-coding regions yielded much lower coverage depth than coding regions (likely due to low complexity, high AT content), as did known polymorphic gene families like *var, rifin, and stevor* (see also Figure 4-2). Thus to reduce the likelihood of genotyping errors due to low coverage and copy number variants, MalariaGEN focused on positions falling between the 15th and 85th percentiles of coverage in coding regions of the nuclear genome. After applying a variety of additional filters, MalariaGEN defined a catalog of 86,158 SNPs in this core genome of 227 samples as *typable* [88]. With the addition of more samples, each update to this catalog adds

mostly singleton SNPs—i.e., new variants only present in a single specimen. The current release identifies one million SNPs genotyped in 6000 samples.

1.2.2 Low complexity variation

Low complexity DNA, marked by an increase in Shannon entropy (i.e., low information content), is unusually abundant in both coding and non-coding regions of the *P. falciparum* genome [89]. Non-coding regions are not under selective constraint at the protein level, and thus these regions have more freedom to vary. Low complexity sequence reflects a more limited usage of nucleotides and/or amino acids in a contiguous stretch, therefore in non-coding regions (with 90% average AT content) these often take the form of homopolymer stretches and dinucleotide repeats of adenine and thymine. Enzyme slippage during replication of such regions is common, and the parasite is under no pressure to remove these mutations, hence introns are littered with size variation. An example of this is presented in detail in section 3.6.2, where the impact of this type of polymorphism on genotyping adjacent exonic SNPs is shown for the drug resistance gene, *pfcrt*.

In addition to non-coding DNA sequence, low complexity regions (LCR) occur in 90% of *P. falciparum* proteins [90]. In proteins these regions fall largely in hydrophilic coils, and appear to fall into three classes [91,92]. Interestingly, one of these categories is marked by polymorphism and some of the highest GC content in the genome, and may be associated with recombination. The two other LCR categories in coding regions are AT-rich, one of which is highly polymorphic with variable length asparagine repeats, and the other much more conserved and heterogeneous [92].

1.2.3 *var* genes

As introduced in the discussion about placental malaria (1.1.2), PfEMP1s are major surface antigens, exported by the intraerythrocytic parasite to the red cell surface, and thought to play a major role in sequestration via cytoadherence. In addition to binding CSA, PfEMP1s have been shown to adhere to several other host receptors, including: thrombospondin (TSP), CD36, ICAM-1, ELAM-1, VCAM-1, P-selectin, hyaluronic acid (HA), heparin sulphate, complement receptor-1 (CR1), PECAM-1, and endothelial protein C receptor (EPCR) [93,94,95]. The haploid genome of *P. falciparum* (3D7) contains 59 *var* genes that encode unique PfEMP1 antigens, yet they appear to be expressed one at a time [96]. This antigen switching may endow the parasite with new binding phenotypes, as well as with variant surface presentations for immune evasion. The PfEMP1 protein is highly variant in amino

acid sequence, however sufficient conservation exists to identify defined functional domains [97,98]. Two domains are particularly important in conferring different binding properties to PfEMP1: the Duffy binding like domain (DBL), and the cysteine-rich inter-domain region (CIDR). The six DBL types (α, β, γ, δ, ε, ζ) and the four CIDR types (α, β, γ, δ) vary in sequence, combination, and arrangement within each PfEMP1 protein [97,99]. Although PfEMP1 is thought to play an important role in cytoadherence, the rodent malaria parasite, *P. berghei*, also adheres to CD36 but lacks a PfEMP1 ortholog [100].

Multi-copy gene families present a challenge for next generation sequencing (NGS) technologies because identical sequence reads may derive from different paralogs, making mapping or assembly ambiguous. In addition to ~60 *vars*, the *P. falciparum* genome contains ~150 copies of the repetitive interspersed family (rifin) and 30-40 representatives from the sub-telomeric variable open reading frame (stevor) family. Like the *vars*, these primarily subtelomeric families are highly polymorphic and are present on the surface of the red cell [101,102].

1.2.4 Divergent and other complex variation

Another class of parasite genetic polymorphism, broadly referred to as complex variation here, includes large indels and highly divergent genomic regions, and is a primary focus of this thesis. Perhaps the earliest example of this type of variation in *P. falciparum* was described in the invasion ligand MSP1, for which monoclonal antibodies were shown to react to some forms of the protein but not others [103]. Sequence analysis later characterized the variation in *msp1* as falling into two distinct forms, described as dimorphic [104]. A hallmark of dimorphism is the existence of two genetic forms, or haplotypes, that do not recombine over longer-than-expected distances [105]. In many cases these forms are so dissimilar that calling them diverged may be misleading, as they may not even be homologous [106]. However, for simplicity these regions are referred to as divergent here, without implying homology. This type of variation is a problem for NGS technologies if the reads are processed by aligning to a reference genome. For example, Figure 1-3 shows the distinctive differences in coverage in the divergent regions of *msp1* when reads from several strains are aligned to 3D7. The coverage of the HB3 parasite across *msp1* (panel B) depicts large spans of the gene that are inaccessible using this standard approach. Further, the reads that do align have lots of mismatches (colored red) compared to the reference, which could lead to erroneous SNP calls.

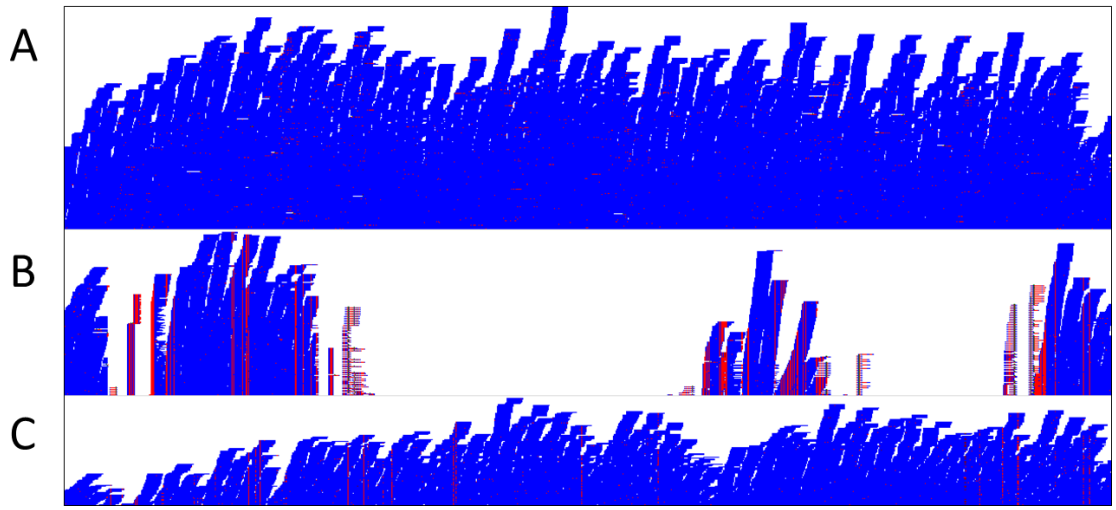


Figure 1-3. Read pileup onto *msp1*. Each panel shows Illumina reads (blue) from one of two lab lines or a field sample aligned to the 3D7 reference version of *msp1*. **A)** Reads came from sequencing a 3D7 parasite, thus they essentially align perfectly to the reference. **B)** The HB3 version of *msp1* diverges substantially from the reference, as illustrated by large areas of missing coverage and likely erroneous variation (red). Many other genes, especially invasion ligands, share this property, and thus we are missing this variation in SNPs aligned to the reference genome. **C)** A field sample from Tanzania appears to be less diverged from the reference than HB3, however, the dotted vertical red lines indicate more than one parasite genotype is present in this infection.

Genome-wide comparisons of sequenced parasites to the 3D7 reference genome indicate a number of extended, divergent regions. Similar patterns are seen in most other members of the MSP family, and dimorphism or extreme divergence has been described in these previously [104,106,107,108,109].

A similar pattern arises in the gene encoding another invasion ligand, *eba-175,* however in this case two mutually exclusive insertions at different loci define the dimorphism. These large insertions (423bp and 342bp, dubbed F and C, respectively, based on the original strains (FCR3 and Camp) in which they were defined) are separated by a conserved 279bp stretch, and there is no evidence of recombination between them [110]. In chapter 4 I show the impact of aligning C-type parasite reads to the F-type reference—i.e., lost and erroneous polymorphism. In chapters 5 and 6 I use novel tools to genotype and *de novo* assemble thousands of *eba-175* genes from clinical samples representing both dimorphic forms, and in chapter 7 I apply these tools for a more in-depth analysis of this gene.

1.3 Genetic variation is important in disease and intervention

As described below, parasite genetic diversity is an important source of failure of vaccine candidate antigens and drug interventions. Although the World Health Organization (WHO) abandoned a failed campaign for eradication a half-century ago, the malaria research community is calling once more for such an effort—this time armed with advances in parasite and vector control, and a grasp of past failures. No single intervention will accomplish the goal of malaria elimination and eradication, and of utmost importance in any multi-faceted approach will be real-time molecular surveillance of antigens and drug resistance genes at genome-wide resolution. Such efforts will rely on comprehensive assessments of the complex diversity in genes related to pathogenesis.

1.3.1 Invasion and cytoadherence phenotypes

Malaria infections result in disparate clinical outcomes and we do not fully understand why, nor do we understand how the parasite specifically contributes to disease [70]. In contrast to other malaria species that infect humans, a unique feature of *Plasmodium falciparum* is that it can adhere to endothelial surfaces in a manner that enables sequestration in deep vascular beds, and implicated in this phenotype is a highly antigenic and polymorphic gene family called *var* (see 1.2.3) [111,112]. Tissue sequestration enables the parasite to avoid splenic destruction [113]. Much focus has been placed on defining the host receptors to which IRBC bind, and on the parasite ligands that mediate this binding [93,95,114,115].

Genes involved in parasite invasion of host cells exhibit more moderate polymorphism, and in some cases circulate in populations as complex allelic forms (see 1.2.4). Beyond genetic variation, a further challenge to the host in the blood stage of infection is that parasites have several redundant pathways for invasion, each involving different genes harboring their own degrees of diversity [116]. Although this situation might elicit pessimism about the prospect of cataloguing important genetic variation of the parasite, the epidemiology of malaria provides some hope (see 1.1.2). In areas of stable malaria transmission, children acquire immunity to severe forms of disease quickly [76]. Similarly, women acquire protective immunity to the form of parasite that attacks the placenta after only one or two pregnancies, and this protection is effective against isolates from other continents [80]. These observations, coupled with evidence that allele-specific immunity to particular invasion proteins is acquired with increasing age, suggests the elusive malaria parasite is somehow constrained in its polymorphism, perhaps indicating a vaccine is a worthwhile

endeavor. This immuno-epidemiology also provides evidence that some forms of parasite diversity may be tractably characterized, and would inform such intervention strategies [117,118].

1.3.2 Drug resistance

P. falciparum has repeatedly demonstrated the ability to overcome front-line antimalarials by changing its genome. Most famously, resistance to chloroquine by genetic substitutions in digestive vacuole pumps has independently emerged at least four times since the 1950s [119,120]. Within six years of its introduction, parasites developed resistance to mefloquine in Southeast Asia [121]. Similarly, the combination of sulfadoxine and pyrimethamine was introduced to replace chloroquine in the late nineties in West Africa, and resistance-conferring mutations to folate pathway genes spread rapidly [122]. Finally, slow parasite clearance times in patients treated with artemisinin derivatives have been spreading from Western Cambodia for more than a decade, and recent genetic evidence has implicated mutations in the kelch propeller domain [123,124,125].

Drug resistance provides the most lucid example of the relevance of genomic variation to disease and death. Resistance to the above drugs is responsible for millions of deaths [126]. Chapter 3 goes into great detail about genetic variation in the *pfcrt* gene, which encodes the transporter involved in resistance to multiple quinolone drugs. The selective pressure on the parasite to adapt in this region has created a dense cluster of polymorphism that makes mapping reads to the reference genome problematic.

1.3.3 Vaccines

Disease eradication campaigns that have relied on vaccines have had success, while those built primarily on non-vaccine measures have failed [127,128]. Although substantial time and resources have been dedicated to vaccine development, promising preclinical antigens have repeatedly fallen short of expectations when tested in endemic regions. Vaccines based on merozoite surface proteins (AMA1 and MSP2) have shown evidence of allele-specific efficacy in field trials [117,118]. Further, a vaccine based on the MSP1 invasion ligand, which exhibits complex dimorphic variation (see 1.2.4), elicited antibodies that recognized MSP1 in Malian adults, but failed to protect Kenyan children [129,130]. Surface exposed parasite proteins are appealing vaccine targets, but are also encoded by the most polymorphic genes in the genome due to frequency-dependent selection by host immunity [131,132,133]. Consequently, it has been recommended that allele-specific efficacy be

included as a key endpoint in vaccine trials, and that molecular epidemiological studies be conducted early in the vaccine lifecycle [134]. The risk of selection of escape mutants is also a concern for transmission blocking vaccines, which must be administered to entire communities, thus creating a potential bottleneck that could substantially modify the parasite population [135]. Population-scale monitoring of parasites with the ability to assess complex variation would be valuable for vaccine design, testing, and deployment.

1.3.4 Surveillance and diagnostic tools

Accurate diagnosis of infections and large-scale surveillance of parasite populations is essential for clinical studies and elimination campaigns [136]. Avoiding antimalarial administration for non-malarial febrile illness is also important for controlling the spread of drug resistance and for proper treatment of other dangerous infections [137,138]. The gold-standard for diagnosis requires a trained microscopist to read a thick blood film, which is unrealistic in many endemic areas, and thus rapid diagnostic tests (RDT) have been widely adopted [139]. Ninety percent of RDTs target the histidine rich protein-2 (HRP2), and sequence polymorphism and spontaneous deletions of this gene have been attributed to false negative diagnoses [140,141]. Separately, as demonstrated for artemisinin resistance in Southeast Asia, real-time monitoring can forewarn about changing parasite populations, which is critical for intervention measures like seasonal malaria chemoprophylaxis, vector control, and future vaccines [142,143]. Genomic-wide monitoring at a population scale is now feasible, but these approaches focus on SNP variation, thus much information is lost for drug resistance and vaccine candidate loci due to complex polymorphism [86,144].

1.3.5 Impact of complex diversity on malaria research

The impact of complex variation on genomic studies of *P. falciparum* is profound, and reaches beyond those focused on parasite genetics. As mentioned in 1.2.4, most merozoite surface proteins exhibit indels or extended stretches of divergence [104,106,107,108,109]. As a rough gauge for how common complex variation is in the genome, the Cortex VCF described in section 5.3 was scanned for large stretches (>80bp) of divergence that occur in at least 12 distinct geographic regions, and this analysis yielded 148 unique genes. Approximately one-third of this list is represented by *var* and *rif genes*. This level of divergence has an impact on many types of experiments. For example, studies that attempted to describe *var* gene diversity in field isolates using degenerate primers were biased toward known sequences [145,146,147]. By extension, quantitative-PCR and

microarray results will be biased toward primer or probe sequences, yielding unreliable assessments of differential gene expression in divergent genes [148,149]. Similarly, gene-expression studies that align RNA-seq reads back to the 3D7 reference genome will miss complex variation or worse, will produce false positive results in diverse field isolates. A similar issue occurs in mass spectrometry experiments, which typically use a reference protein database to infer peptides from spectra [150]. Finally, as explored throughout this thesis, complex variation limits our ability to study important genes using modern sequencing techniques. Section 1.3 outlines why genetic variation is important in malaria disease and intervention, and provides several examples of the impact of complex polymorphism on vaccine and surveillance studies. Section 1.2.1 describes how large genome projects are relegated to studying a "core genome" because sequencing reads cannot be properly aligned to divergent regions in the genome. Hundreds of studies have used MalariaGEN SNPs and similar methods for genotyping, and all of these are limited to the same core genome.

The work in this thesis provides access to some of the genes outside of the core genome. As outlined in section 1.2, some classes of variation present issues that are beyond the reach and scope of this work—for example, the *var* gene family is too paralogous. However, as demonstrated in chapter 6, the large indels of *eba-175* and long divergent stretches of *msp3.4* can be reliably assembled. Further, even for genes that are difficult to assemble like *pfcrt*, particular exons can be accessed in lieu of the entire gene. This thesis focuses on a handful of genes, but others like *msp3.8, msp1, msp3,* and *msp6* have been assembled as well and will be contributed to the wider malaria community.

1.4 EBA-175

1.4.1 Introduction

Parasites express proteins that contain cysteine-rich Duffy binding-like (DBL) domains to facilitate adherence to host glycoproteins and endothelial receptors [151]. Different families of these ligands are either exported to the parasite surface membrane, or are trafficked to the surface of infected red blood cells [152,153]. As described in section 1.2.3, *var* genes encode DBL-containing erythrocyte membrane proteins that mediate parasite binding to various host receptors, enabling infected RBCs to avoid splenic clearance [154]. Other proteins with DBL domains are expressed directly on the merozoite surface and are important for invasion. The first and most extensively described DBL-containing parasite

ligand involved in red cell invasion is the 175 kilodalton erythrocyte binding antigen (EBA-175) [155]. Based on similarities in gene structure and sequence homology in the 5' and 3' cysteine-rich domains, EBA-175 is likely orthologous to the Duffy Binding Proteins of *P. vivax* and *P. knowlesi* [156]. Although these two species also infect humans, their most recent common ancestors with *P. falciparum* are millions of years older than the MRCA of *P. falciparum* with the chimpanzee parasite *P. reichenowi*, and this ancestry is reflected in *eba-175*. The DBPs of *P. vivax* and P. *knowlesi* contain a single DBL domain, whereas *eba-175* in *P. falciparum* and *P. reichenowi* have two DBL domains near the N-terminus. These tandem DBL domains are called F1/F2 in *P. falciparum* and R1/R2 in *P. reichenowi*, and together define Region II of *eba-175* (Figure 1-4). Phylogenetic studies suggest that the F1 and R1 domains are orthologous to the DBPs, and the F2/R2 domains are most closely related to the DBL domains found in the *var* multigene family [157].

The crystal structure of EBA-175 Region II was solved a decade ago, providing insights into this ligand's role in merozoite invasion [158]. This work showed that the RII domains from two EBA-175 molecules form a dimer in a handshake orientation—i.e., the F1 domain of one molecule interacts with the F2 domain of the other [159]. This complex forms two channels, containing binding sites for glycans presented by GYPA, particularly those encoded by exon 3 of this erythrocyte membrane protein [160]. As described in section 1.1.1.2, EBA-175 is released from the micronemes as the merozoite reorients its apical end to face the red cell surface, and is involved in forming a junction between the two cells by binding with GYPA. This interaction may occur when the two EBA-175 monomers assemble around GYPA (itself a dimer on the erythrocyte surface). Although substantial work has implicated RII in the interaction with GYPA, as discussed further in 7.6.2, a recent study shows that other EBA-175 regions may contribute to binding [161,162].

1.4.2 Structure

The *eba-175* gene contains 4 exons and is located on the sense strand of chromosome 7. This gene exists in two allelic forms, defined by insertion/deletions in region III (Figure 1-4). These forms are mutually exclusive within the same gene, and thus define a dimorphism dubbed F and C, based on the names of the original isolates in which each insert was detected (i.e., the FCR3 and Camp strains). The signature of the F indel is markedly visible as an artefact in MalariaGEN sequence coverage plots, and is also indicated by missingness in haplotype representations (Figure 5-1 and Supplementary figure 7-21). A smaller 6bp indel is present approximately between the two DBL domains, at the 3' end of

F1. Although not a DBL domain, a third cysteine rich region is located at the end of exon 1, just before the transmembrane domain in exon 2. The final 2 exons form a cytoplasmic tail.

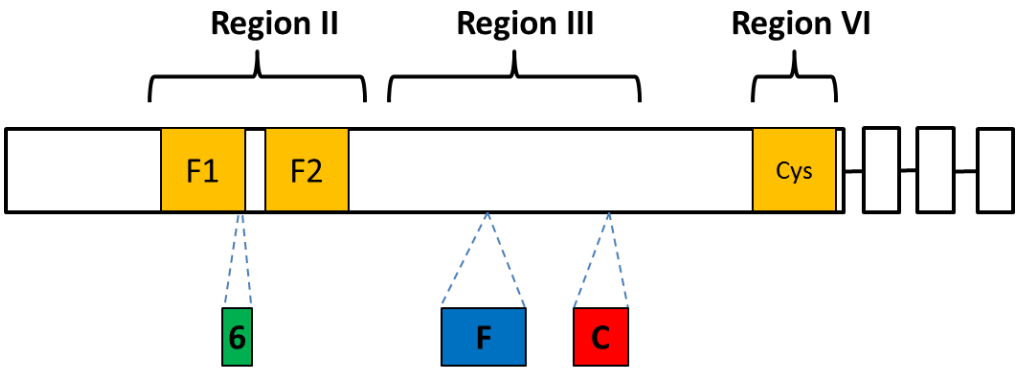


Figure 1-4. Schematic of *eba-175*. Exons are open boxes connected by short lines. Exon 1 contains several noteworthy features for this thesis. F1 and F2 are DBL domains. The approximate locations of the 3 structural variants (6bp, F and C indels) are shown as colored boxes below the gene model. Exact locations and sequences of these variants are given in Supplementary table 7-6 and in Supplementary table 7-7. The far right orange box illustrates the 6-cys domain.

1.4.3 Population genetics

Identifying genes or regions under balancing selection may reveal targets of natural host immunity [163]. This is relevant because humans develop natural protective immunity over time, and the targeted antigens would make rationale vaccine candidates. Identifying proteins under pressure to diversify is typically evaluated as a summary statistic, calculated across the entire gene or region using the available polymorphism. EBA-175 has previously been indicated as a target of balancing selection [164]. Naturally acquired antibodies to EBA-175 are present in immune individuals, but whether these are protective may depend on the domain targeted [165,166,167,168]. Further, antibodies targeting recombinant EBA-175 domains have been shown *in vitro* to inhibit invasion [169,170]. A hallmark of balancing selection is that polymorphism is maintained at a more intermediate frequency than expected (see 7.4.3). In virtually every study of the *eba-175* F and C dimorphic indels, both parasite forms are detected in every population [171,172,173,174]. The extent that immune-mediated selection and/or other mechanisms contribute to this phenomenon is unknown.

1.4.4 Interactions and association with disease

The dimorphic F and C indels of EBA-175 are visually striking in a multiple sequence alignment, and the thought that this pattern is a signature of host interaction has lured investigators for two decades (Supplementary figure 7-20). The host receptor GYPA is an

MNS blood-group antigen, and this blood-group is the first human genotype investigated for an association with the F and C dimorphism [175]. No association was detected in that study. Separate investigations have sought to connect the F/C dimorphism to disease outcomes. A study in Ghanaian children reported a possible association of the C-form with malaria fatality [171].

In contrast to the high level of interest and attention paid to the F and C indels of EBA-175, the smaller 6 base-pair indel in region II has been largely overlooked. Despite its presence in nearly every population around the globe (a strong indication of balancing selection), and its position between the two DBL binding domains, this indel is barely mentioned in the literature. A study investigating parasite genotypes with host outcomes did not detect an association of this indel with severe disease in Thailand [176]. To my knowledge, no investigations into the association of the 6bp indel with host genotypes have been published.

1.5 Antimalarial drugs and modes of action

In chapter 3 I describe a genome-wide study in which I tested for parasite SNP associations with 22 antimalarial drugs. These drugs fall into several broad categories that I describe below, highlighting the known mechanisms of action of those most relevant to the GWAS results.

1.5.1 Aminoquinolines

This class of antimalarials includes the 4-aminoquinolines: chloroquine, amodiaquine, pyronaradine, and piperaquine, as well as the quinolinemethanols: mefloquine, halofantrine, lumefantrine, and quinine. Chloroquine is perhaps the most recognized antimalarial because of its early success and disastrous decline due to drug resistance. Its (re)discovery was motivated by the lack of access to the quinine-yielding cinchona tree bark during World War II [177]. It was originally synthesized at Bayer before the war, but was not pursued at that time based on safety comparisons to its predecessor, atabrine [177,178].

The 4-aminoquinolines interfere with heme detoxification in the parasite food vacuole. During the intraerythrocytic stage, parasites must digest host haemoglobin for amino acids. This process results in reactive oxygen species and the toxic by-product Fe(II)-protoporphyrin IX (FP), which is part of the heme metabolite [34]. Chloroquine forms a complex with FP, interfering with its dimerization and crystallization into hemozoin [179]. As discussed in section 3.5, this leads to heme efflux from the digestive vacuole into the

cytosol, resulting in an oxidative challenge and membrane damage [180]. In the cytosol, glutathione can reduce free heme, but both CQ and AQ can interfere with this process [181]. Amodiaquine is rapidly metabolized into the pharmacologically active and more stable forms, mono- and bis-desethyl amodiaquine DEAQ, which have half-lives of 1-3 weeks [182,183,184]. Although the mode of action of AQ is thought to be similar to that of CQ, it is more effective than CQ against CQR parasites. In a study testing the activity of 108 4-aminoquinolines against *P. falciparum in vitro*, 55% were active against parasites that were both chloroquine sensitive and resistant [185]. Amodiaquine was shown *in vitro* to accumulate at a 2-3 fold greater concentration in *P. falciparum*, which may account for its increased efficacy.

The quinolinemethaonols also accumulate in the digestive vacuole but likely have a different mechanism of action. In mouse models, treatment with mefloquine and quinine does not lead to reduced hemozoin production (as with CQ), and mefloquine interacts weakly with free heme [186,187,188]. Further, the strong affinity of these quinolinemethanols to membrane phospholipids may account for their superiority in treating CQ resistant parasites [189].

As discussed in section 1.1.1, the 8-aminoquinoline, primaquine, is the only antimalarial clinically indicated for radical cure of *P.vivax* and *P. ovale* hypnozoites, and it has gametocytocidal activity against *P. falciparum*.

1.5.2 Artemisinin derivatives

Current front-line antimalarials are combination therapies of artemisinin derivatives with other drugs. This extract has been used to treat malaria fevers for thousands of years in China, and is derived from the herb *Artemisia annua* [190]. Three derivatives: dihydroartemisinin, artemether, and artesunate, are more potent than the primary extract, and are used in the artemisinin combination therapies (ACT) [190]. Although artemisinin derivatives rapidly clear blood-stage parasites, they have very short serum half-lives [191]. These half-lives, on the order of a few hours, are complemented in ACTs with longer-acting drugs, e.g., lumefantrine and piperaquine. Administering antimalarials in combination reduces the likelihood that resistant parasites will emerge from an infection, but as described in sections 1.3.2 and 4.1, artemisinin resistant parasites are spreading throughout Southeast Asia at an alarming rate. The mechanism of resistance has been localized to the PfKelch13 protein using *in vitro* and GWAS approaches (particularly substitution C580Y), and this resistance mechanism is consistent with a proposed mode of action for this

important antimalarial [124,143,192,193]. Recent work has implicated the *P. falciparum* phosphatidylinositol-3-kinase (PfPI3K) as the artemisinin target in the parasite [194]. Artemisinin is a strong inhibitor of PfPI3K, and the C580Y substitution is associated with decreased binding of Kelch13 to PfPI3K, resulting in increased levels of this kinase and its lipid product, phosphatidylinositol-3-phosphate (PI3P). Further, increased levels of PI3P induce artemisinin tolerance directly [194].

1.5.3 Antifolates

Folate metabolism is required for the synthesis of nucleic acids and for some amino acids, and is thus essential for parasite cell division and survival [195]. Through a series of enzymatic reactions, folate must be converted to tetrahydrofolate (THF), which is the cofactor involved in these essential biosynthesis pathways downstream [196]. One of the reactions along this pathway conjugates a heterocyclic pterin compound with para-aminobenzoic acid (pABA), and this step is catalysed by the enzyme dihydropteroate synthase (DHPS). Sulfa drugs (e.g., sulfadoxine) are structural analogs of pABA and interrupt this step via competitive inhibition of DHPS [197]. Downstream of this reaction, a double bond in one of the pterin rings is reduced by the enzyme dihydrofolate reductase (DHFR) to convert dihydrofolate (DHF) to THF. Interestingly, in *P. falciparum* and other protozoa the thymidylate synthase enzyme that converts THF back to DHF (a reaction important for thymine production) is encoded by a gene that is fused with the one encoding *dhfr* [198,199]. These genes are unfused in animals and fungi [200]. Drugs that directly target DHFR include pyrimethamine, methotrexate, and trimethoprim [196].

As shown in section 3.4.5, parasites have evolved a series of point mutations that counter the antifolate activity of drugs that target DHFR and DHPS [201,202]. These resistance-conferring mutations spread rapidly when the sulfadoxine-pyrimethamine combination, Fansidar, replaced chloroquine after its own diminished efficacy due to resistance in the 1990s [203]. The loss of efficacy of this antifolate combination is important because it is considered safe for use in all trimesters of pregnancy, and as discussed in section 1.1.2, malaria during pregnancy is a considerable cause of morbidity and mortality [204].

1.6 Overview

1.6.1 Section I

This section is comprised of chapters 3 and 4, which describe analyses based on SNP data. The first of these chapters is a GWAS that details parasite loci associated with drug resistance, as well as population genetic findings in this Kenyan sample-set. In chapter 4 I introduce a clinical field study I have been involved with at two locations in Tanzania. The sequenced parasites from these two Tanzanian sites are used for validation in the methods development chapters of Section II, but first I report some population genomic parameters on a subset of the dataset. A sub-theme in both of the Section I chapters is that using SNPs ascertained by aligning short reads to the reference genome makes some of the most interesting polymorphism inaccessible, and concrete examples of this are shown for two genes, *pfcrt* and *eba-175*. This 'missingness' motivates a deeper look into its origins, and I conclude that new methods are necessary to access complex variation.

1.6.2 Section II

This section is also comprised of 2 chapters (5 and 6), focused wholly on solving the problem of accessing the types of complex variation introduced in Section I. In chapter 5 I introduce software I developed called Malign—a tool for detecting known indels or divergent regions in short read data. After demonstrating the utility of Malign for parasite as well as for host samples, I conclude that more work should be done to develop methods for *de novo* assembly of full-length genes. Toward this end I initially develop an approach that patches Cortex variants into the 3D7 versions of genes, but show that this is subject to unacceptable systematic bias. Chapter 6 summarizes one of the major pieces of work in this thesis—an algorithm and software (called MalMOI) for *de novo* assembly of full-length genes from samples with a mixture of genomes. I demonstrate the power of MalMOI by assembling thousands of full and partial sequences of *msp3.4, ama1, pfcrt, and eba-175,* each exemplifying a different type of complex variation. Although I highlight a few applications of these sequence sets in Section II, in particular to two vaccine candidates (*msp3.4* and *ama1*), I reserve the in-earnest presentation of an application for Section III, where I go into depth about *eba-175.*

1.6.3 Section III

This section brings the thesis full-circle, revisiting a gene that appeared throughout the chapters, *eba-175*. Using the MalMOI *de novo* assemblies from Chapter 6 I provide a comprehensive catalog of variation in this vaccine candidate and define a universal IUPAC consensus (i.e., a single sequence representation for every parasite). I then use Malign to afford an accounting of the global distribution and allele frequencies of the three structural variants in *eba-175*, and go on to present some population genetic and molecular evolutionary analyses. Finally, I use the IUPAC consensus to design Sequenom assays for typing these structural variants in Kenyan and Gambian samples for which host genotypes were available. I conclude with a human GWAS testing for host interactions with the *eba-175* genotype of the infecting parasite, and report an interesting association for the 6bp indel that reaches genome-wide significance and yields signal in both populations.

2 MATERIALS AND METHODS

A substantial portion of this thesis (particularly Section II) describes the development of new genomics tools, and this work reads more coherently if those technical methods are printed within the relevant chapters. I use this space to describe methods and materials that are common to more than one chapter, and for those methods that require background that would distract from the spirit of the results sections.

2.1 General methods

2.1.1 Illumina sequencing

The "next generation" sequencing technology referred to in this thesis is short-read, paired-end, Illumina sequencing. This technology has had a massive impact on genetics and genomics over the past decade. Another generation of sequencers is on the horizon, promising to be cheaper and to generate reads that are 10s of kilobases (as opposed to 100bp Illumina reads) [205,206]. During my DPhil I was part of an early access program to one of these technologies via Oxford Nanopore.

Illumina sequencing works by fragmenting the genome into short ~300bp "inserts," which are sequenced from each end to ~100bp toward the center. These sequenced ends are called read-pairs, and it is tracked downstream which reads derive from the same insert. In other words, on average there is about 100bp of the insert that goes unsequenced. Thus when read pairs are aligned to the reference genome, those that map farther apart than expected can indicate an insert in the reference that is deleted in the given sample (see Figure 5-8 for an example in *eba-175*). The benefit of fragmenting the genome is the massive gain in efficiency from sequencing the inserts in parallel, but this comes at a cost downstream. Putting the short puzzle pieces back together to represent the newly

sequenced genome is complicated in repeat regions and for paralogous gene families. Reads are typically aligned to a reference genome, and polymorphism is characterized in regions where reliable mapping can be done. These variants are not phased (i.e., not grouped on a single haplotype) in parasite samples containing a mixture of genomes. Separately, in regions where the polymorphism is more complex than SNPs, read mapping can become unreliable or impossible, and this variation is inaccessible.

2.2 General Bioinformatic software

2.2.1 Candidate Gene Reports

The software package CGR.R was developed using the R programming language to produce visualizations of sequence data and other gene-specific features in one summary plot. The only required user input to this tool is the PlasmoDB gene accession, and several plots are output. Plots are structured with disparate datasets incorporated as separate tracks so multiple features can be positioned on the same gene coordinate.

2.2.1.1 Brief description of CGR.R plots

Haplotype plots: these are color-based matrix depictions of the genotypes of groups of samples. Columns represent SNPs and rows represent samples. Each cell of the matrix is colored blue if the allele for the given samples matches that of the 3D7 reference, and red otherwise. In samples with multiplicity of infection (MOI), the relative abundance of each parasite is represented with a corresponding mixture of blue and red in the same cell.

Coverage plots: the dominant feature of these plots is a central panel showing the read depth for each sample across the gene. Coverage plots also show SNP positions along the gene for various categories (e.g., synonymous/nonsynonymous) and sources (e.g., PlasmoDB) of SNPs. Other tracks include GC%, uniqueness, and flags highlighting high frequency SNPs.

Frequency plots: as in coverage plots, tracks showing SNP positions and categories are plotted. Additional tracks are plotted that illustrate within-country allele frequencies

Filter plots: in addition to SNP tracks, these plots highlight which of 12 MalariaGEN quality filters each SNP passed or failed.

2.2.1.2 Description of plot tracks

UQness: Uniqueness score. This is the shortest necessary oligo length spanning a particular position in the genome that is required in order to not match any other location in the genome. Values greater than 26 nucleotides are colored red. For computational efficiency, any position with a UQ greater than 46 is just designated 99—i.e., the dynamic range of UQ is 1-46.

GC%: GC content. Colors transition from red to green from 10% - 26% GC. Anything over 26% GC is solid green.

Gene: Physical layout of the intron and exon positions. The arrow points in the coding direction.

Domains: Interesting regions manually entered.

Variants (several tracks with different names)**:** These tracks indicate SNPs as defined by various sources or MalariaGEN filter thresholds. In each track, tall red bars represent non-synonymous SNPs, and short blue bars indicate those that are synonymous. Green bars may appear in the "No Filter" track, and represent intronic SNPs.

- **No Filter:** All 978,274 potential variants in the 26 April 2012 MalariaGEN VCF.
- **v2.0 SNPs:** 421,613 variants in the version 2.0 MalariaGEN release.
- **V1.0 SNPs:** The 86,158 variants described in the first MalariaGEN parasite SNP release [86].
- **PlasmoDB:** SNPs represented in PlasmoDB version 8.1.

Coverage: The coverage track contains a thin colored line for each sample (up to 1591 for MalariaGEN v2.0) representing the normalized sequence read coverage for each sample. Lines representing samples from the same country are the same color. The normalizing constant for each sample is the mean or median (indicated in the title) coverage of SNPs in all genes for the given sample. The single thick black line is the mean across all samples (median looks nearly identical). An optional thick blue line represents one standard deviation above the mean.

Freq Flags: For each SNP, a triangular symbol (flag) is plotted if the non-reference allele frequency reaches the following thresholds: red: >=20%, blue: >=10%, orange: >=5%.

2.2.1.3 Acknowledgments

Uniqueness values were provided by Magnus Manske. CGR.R grew out of project called JAGeneGen, on which I partnered with Antoine Claessens.

2.2.2 Pan-Conserved Stretch

The software package PCS.R was developed using the R language for identifying stretches of the reference genome with no evidence of polymorphism. This tool was primarily developed to facilitate the design of universal primers (i.e., primers that should work in any parasite). The software takes a PlasmoDB gene ID as input (other input options described below), and provides the user with a visualization of conserved domains at different allele frequency thresholds. At each minor allele frequency (MAF) a "masked" reference sequence is also output that replaces potentially variable nucleotides with the letter N. These sequences are convenient for input into the primer design software Primer3, which will avoid designing primers in masked regions [207]. Variation data is used from PlasmoDB version 8.1 and unfiltered SNPs in the MalariaGEN *P. falciparum* variation project version 2.0. Although the unfiltered SNPs contain many false positives, it provides for a more conservative estimate of non-variable regions.

2.2.2.1 Input options

- Gene ID: PlasmoDB identifier.
- Population: this is the geographic region of the samples to use for identifying SNPs. The default setting of "World" masks positions in which a SNP is found in any population. This can be set to an individual country or a region like East Africa (EAF) or Southeast Asia (SEA). The extremely high polymorphism in some genes makes primer design difficult in masked references created using the "World" setting, so a regional setting could be used in studies focusing on a narrower population.
- Stretch size: this is the minimum length of oligonucleotide with no intervening SNPs that will be indicated on the plot (default = 25).
- "NonSynonymousOnly": this is a true/false setting for whether to ignore synonymous SNPs. Setting to TRUE might be used to identify conserved peptides for antigen selection, for example.

2.2.2.2 Sample output and key

Population: input population used for identifying SNPs.

Coverage: median sequence coverage across all MalariaGEN version 2.0 samples. Odd coverage profiles may indicate regions to avoid for primer design, even if within a conserved stretch.

Frequency tracks: SNPs with a frequency, within the defined population, are featured as red vertical lines if their MAF exceed the cut-off threshold (0%, 0.5%, 10%, 20%). Ranges of nucleotides with no intervening SNPs that meet the stretch-length input minimum are indicated as colored horizontal lines. These panconserved stretches are colored gray, yellow, orange, red, and dark-red if they exceed 25, 50, 100, 200, and 300 nucleotides in length, respectively.

PlasmoDB: locations of SNPs in the PlasmoDB version 8.1 database.

CDS: gene model with introns spliced out.

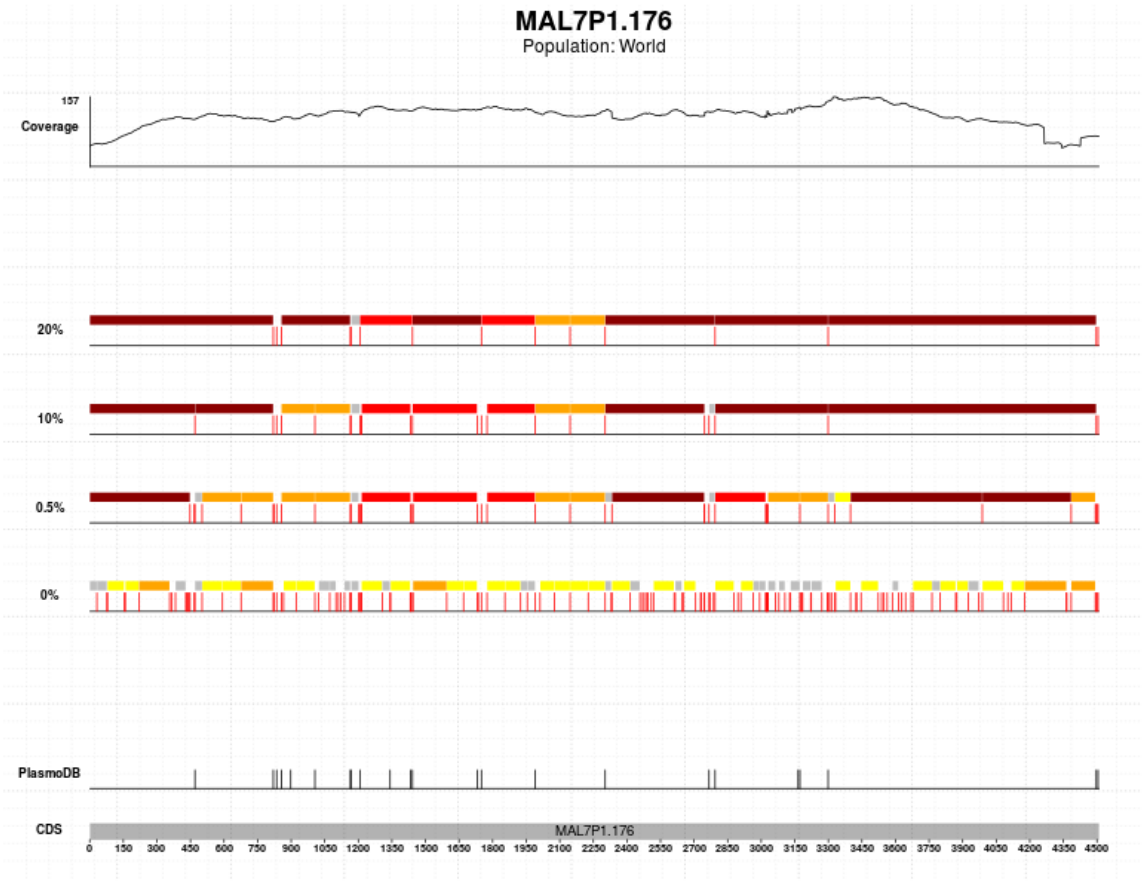


Figure 2-1. PCS.R plot for the gene *eba-175*.

2.2.3 fastqMixer

The software package fastqMixer.R was developed using the R language for creating artificial samples from *in silico* mixtures of reads from real samples. Complementing *in vitro* DNA mixtures, *in silico* mixtures are a powerful tool for studying MOI, and in particular the robustness and limits of the software created in Section II on mixed field samples. As opposed to *in vitro* approaches, *in silico* methods can be used to generate hundreds of thousands of control mixtures when targeting a specific gene. The program can mix any number of paired-end fastq files at user defined proportions.

2.2.3.1 Options

Option	*Description*
-h	Print help menu.
-name	Prefix for the output fastq files. The proportions (in the order entered) and "_1.fastq" or "_2.fastq" will be appended to the name.
-p	Comma separated fastq proportions. Must be same length and same order as the trailing fastq files.
-d	Depth (actually raw read number) in the output files. Will default to ((lowest covered sample * 1/prop)-1) if missing or too high.
-z	Want output fastq files to be gzipped? [T/F: Default=T]
-o	Out directory. [Default=calling directory]

Trailing the options, list each fastq pair in order, e.g., fqA_1 fqA_2 fqB_1 fqB_2, etc. There should be two fastq files for each proportion in the -p option. It is critical that the reads within each fastq pair are in the same order.

2.2.3.2 Usage

From the Linux command-line, an example call for fastq files covering a single gene:

$Rscript fastqMixer.R -z F –d 3000 -p 0.95,0.05 -name mygene_mixture -o my_directory fqA_1 fqA_2 fqB_1 fqB_2

2.3 Materials and methods for chapter 3

2.3.1 Ethics

Parasites were isolated from the peripheral blood of participants in two clinical trials on Artekin versus Coartem, conducted in Kilifi between 2005 and 2007. All studies obtained clearance from the Kenya Medical Research Institute (KEMRI) Ethical Review Committee under the protocol numbers SSC 945 and SSC 946.

2.3.2 Sequencing and genotyping

Extracted DNA was contributed to MalariaGEN for whole-genome sequencing and genotyping. Isolates were sequenced with an Illumina Genome Analyzer to a read depth of approximately 98x in genotyped loci, and reads of length 37-76 base pairs were aligned to the 3D7 reference genome as previously described [86]. Genotype calls for each sample were provided by MalariaGEN for more than 400,000 high-quality exonic SNPs in their version 2 catalog of genetic variation. Sequencing data for the parasites used in this study have been deposited in the European Nucleotide Archive, and are publicly available for download (http://www.ebi.ac.uk/ena/). SNP genomic coordinates and annotations are maintained by MalariaGEN, and the most updated tools for viewing this information can be found at http://www.malariagen.net/data. Accession numbers for data access in the European Nucleotide Archive and corresponding phenotype data are listed in Supplementary table 3-6.

2.3.3 Sample collection and processing

Infected blood pellets were cryopreserved using glycerolyte and later adapted to culture as described elsewhere [208]. Pellets were frozen for three months on average before culture adaptation and chemosensitivity testing, and were in continuous culture for approximately two months for these assays before DNA extraction and sequencing (Supplementary figure 3-18). DNA was extracted from adapted field isolates using the QIAamp DNA Blood Mini Kit (Qiagen, UK).

Of the thirty-five isolates used in the final analysis, thirteen were taken from patients admitted to Kilifi District Hospital with severe malaria, and twenty-two from participants in a study comparing Artekin to Coartem [209]. Of these latter twenty-two, twelve were collected at recruitment, and ten were collected 19-84 days later (mean = 48.7days), representing reinfections or recrudescence. Two of the ten follow-up samples are from

patients also represented at recruitment in this dataset. We classified both of these cases as reinfections because, based on the number of SNP identities, the recruitment and follow-up parasites were no more similar to one another than to those from other patients.

2.3.4 Chemosensitivity testing

Details of 50% growth inhibition (IC_{50}) determination for each parasite isolate have been previously described [210]. For a given assay, duplicate series of 200μl cultures containing 0.5% parasitemia and 1.5% hematocrit were established in 96-well microtiter plates and exposed to a gradient of drug concentrations. Drug sensitivity was approximated by standard incorporation of tritiated hypoxanthine, added after 24 hours of culture and measured by scintillation 18-20 hours later. The concentration at which IC_{50} was achieved was estimated using nonlinear regression. Chemosensitivity assays were performed two to four different times on each isolate, on separate days, and the median IC_{50} value was used as the phenotype in the final analysis. Median IC_{50} concentrations were determined for each of 22 drugs applied to 59 parasite isolates.

2.3.5 Analysis

All analyses were performed using R and Perl. For each SNP with greater than 9% MAF amongst these 35 samples (N = 6250), an independent hypothesis test was performed to assess whether $\log_{10}(IC_{50})$ levels differed between the reference (i.e., 3D7-like) and alternate allele groups. This was done separately for each drug. The MAF of 9% was chosen to ensure the minor allele group had at least 3 representative parasites. This value was arrived upon by using the phenotypic data for each drug to calculate the lowest p-value any SNP could possibly yield at various MAFs, given the data. Parasites with the highest and lowest IC_{50} values for a given drug were artificially assigned to different allele groups. With at least 3 parasites in the minor allele group (MAF = 3/35 = 9%), all drugs had the potential to reject at the 0.03 level of significance, and the median "best level" amongst the drugs at this MAF is 0.0001 (Supplementary figure 3-19). A MAF of 9% provided a reasonable balance between data loss, computational resources, and multiple hypothesis testing considerations.

The SNP-wise hypothesis tests assessed whether the dichotomous fixed effect of genotype (i.e., 3D7 vs. alternate alleles) was equal to zero in a linear model that also contained three surrogate variables to account for population structure. The surrogate variables were calculated from principal components analysis (PCA) performed on a matrix of 35 quality filtered samples and 12802 SNPs, in which each cell was the reference allele frequency. For

this PCA, SNPs with no missingness in any sample were included. The first three eigenvectors were projected onto the data, and these variables were modeled as direct, fixed-effects. Although mixed models accommodating within-isolate experimental replicates as random effects improved p-values, I chose to median-collapse repeated assays to avoid the possibility of pseudo-replication. Significant SNPs were also tested by Kruskal-Wallis, and residuals assessed for departure from normality by quantile-quantile (QQ) plots and the Shapiro-Wilk test. Spearman's rank was used for pairwise drug correlations and tests. To control for type I error inflation due to multiple hypothesis testing, the method proposed by Storey and Tibshirani for estimating the false discovery rate (FDR) was employed. Genome-wide significance was defined as q-value less than 0.05 after correcting for multiple comparisons [211]. This standard method estimates the expected null distribution of the thousands of p-values in the experiment and provides a significance cut-off that protects from false positives at some FDR. Rejecting SNPs with FDR < 0.05 is analogous to controlling for type I error with the traditional $\alpha = 0.05$. Further, we produced QQ plots and p-value histograms for each drug, and made sure there was no evidence of genome-wide inflation due to model misspecification (example provided for CQ in Supplementary figure 3-20).

2.3.5.1 SNP-specific raw data plots

The GWAS linear model employed for a given SNP (IC_{50} = Genotype + PC1 + PC2 + PC3) contains the parasite genotype, as well as several covariates. When plotting the "raw data," the visualization of interest is the genotype vs. phenotype, however the raw data doesn't accurately reflect the GWAS p-value, as the variation due to the covariates remains in the IC_{50} value. In order to show SNP-specific raw data plots that more accurately reflect GWAS result, I first fit a linear model without genotype. The residuals from this fit still contain the IC_{50} variation due to genotype, but with the variation due to the other covariates removed. These residual values were then used in the SNP-specific raw data plots (for example, Figure 3-3).

2.4 Materials and methods for chapter 4

2.4.1 Study site

The Laboratory of Malaria Immunology and Vaccinology at the US National Institute of Allergy and Infectious Disease (LMIV/NIAID/NIH) has established the Mother Offspring Malaria Study (MOMS) project, which aims to identify parasite ligands and soluble

mediators involved in malaria infections during early life. The project was first established by investigators at the Seattle Biomedical Research Institute in 2003, and those leaders subsequently moved to NIH and continued the study. Longitudinal and cross-sectional cohorts are recruited at hospitals in two separate regions of Tanzania: Muheza Designated District Hospital is located on the north-eastern coast of Tanzania, and the Morogoro Regional Hospital is 200km inland. Pregnant women are recruited at delivery, and followed for at least the next 5 years. Children under 5 are recruited during routine outpatient visits, and followed for five years. Infants under 12 months old are followed by bloodsmear every two weeks, and then monthly after their first birthday. Peripheral blood is taken at 3, 6, and 12 months, and then at 6 month intervals thereafter. Blood may also be drawn when patients are hospitalized on a non-routine visit. Clinical examinations are performed by project medical staff during each patient visit, and these data are stored in a secure relational database.

Tanzanian field studies were supported by the Division of Intramural Research, National Institute of Allergy and Infectious Diseases of the National Institutes of Health, the Bill & Melinda Gates Foundation (grant number 29202), the Foundation for the NIH through the Grand Challenges in Global Health Initiative (grant 1364), the National Institutes of Health Fogarty International Center (grant D43 TW005509), and the National Institutes of Health (grant R01 AI52059).

2.4.2 Ethics approval

The MOMS project has been approved by the International Clinical Studies Review Committee at US NIH, and ethical clearance was obtained from the Institutional Review Board of Seattle Biomedical Research Institute and the Tanzanian National Institute for Medical Research.

2.4.3 Parasite culture adaptation and processing

Cryopreserved parasites were thawed and short-term adapted to culture by technicians at the Seattle Biomedical Research Institute. Parasites were cultured until they expanded to a packed volume of 2-3ml of infected RBCs at greater than 3% parasitemia. The average time in culture was 2-6 weeks.

Frozen blood pellets were shipped on dry ice to Oxford where DNA was extracted from adapted field isolates using the QIAamp DNA Blood Mini Kit (Qiagen, UK).

2.4.4 Sequencing and genotyping

Extracted DNA was contributed to MalariaGEN for whole-genome sequencing and genotyping. Isolates were sequenced with an Illumina Genome Analyzer to a read depth of approximately 98x in genotyped loci, and reads of length 76-100 base pairs were aligned to the 3D7 reference genome as previously described [86]. Genotype calls for each sample were provided by MalariaGEN for more than 400,000 high-quality exonic SNPs in their current catalog of genetic variation. Sequencing data for the parasites used in this study have been deposited in the European Nucleotide Archive (ENA), and are publicly available for download (http://www.ebi.ac.uk/ena/). ENA accession numbers, geographic origins, sequencing outcomes, and sample qualities are listed in Table 2-1.

Table 2-1. Outcome and ENA number of Tanzanian sequence data.

MalariaGEN code	Origin	Status	QC	Sequencing reads
PE0001-C	—	Failed	—	—
PE0002-C	—	Failed	—	—
PE0002-CW	Tanzania: Muheza	Failed	99.90%	ERS010096
PE0003-C	—	Failed	—	—
PE0004-C	—	Failed	—	—
PE0005-C	—	Failed	—	—
PE0006-C	—	Failed	—	—
PE0007-C	—	Failed	—	—
PE0008-C	—	Failed	—	—
PE0009-C	Tanzania: Morogoro	Sequenced	99.80%	ERS010095
PE0009-CW	Tanzania: Morogoro	Sequenced	100%	ERS144056
PE0010-C	Tanzania: Muheza	Sequenced	99.90%	ERS010097
PE0011-C	Tanzania: Muheza	Sequenced	99.70%	ERS010098
PE0012-C	Tanzania: Muheza	Sequenced	99.90%	ERS010099
PE0013-C	Tanzania: Muheza	Sequenced	99.80%	ERS013053
PE0013-Cx	Tanzania: Muheza	Failed	—	ERS010100
PE0014-C	Tanzania: Muheza	Failed	—	ERS010101
PE0015-C	Tanzania: Muheza	Failed	—	ERS010102
PE0016-C	Tanzania: Muheza	Sequenced	100%	ERS013054
PE0017-C	Tanzania: Muheza	Sequenced	99.80%	ERS013055
PE0017-CW	Tanzania: Muheza	Sequenced	100%	ERS144059
PE0018-C	Tanzania: Muheza	Sequenced	99.80%	ERS013056
PE0018-CW	—	Failed	—	—
PE0019-C	Tanzania: Muheza	Sequenced	99.70%	ERS013057
PE0020-C	Tanzania: Muheza	Sequenced	99.90%	ERS013058

PE0021-C	Tanzania: Morogoro	Sequenced	99.80%	ERS013059
PE0021-CW	—	Failed	—	—
PE0022-C	Tanzania: Morogoro	Sequenced	99.90%	ERS013090
PE0022-CW	—	Failed	—	—
PE0023-C	Tanzania: Morogoro	Sequenced	100%	ERS013060
PE0024-C	Tanzania: Morogoro	Sequenced	99.90%	ERS013061
PE0025-C	Tanzania: Morogoro	Sequenced	100%	ERS013062
PE0027-C	Tanzania: Morogoro	Sequenced	100%	ERS144057
PE0028-C	Tanzania: Morogoro	Sequenced	100%	ERS144058
PE0030-C	Tanzania: Muheza	Sequenced	100%	ERS144060
PE0121-C	Tanzania: Morogoro	Sequenced	100%	ERS354264
PE0122-C	Tanzania: Morogoro	Sequenced	100%	ERS354267
PE0123-C	Tanzania: Muheza	Sequenced	100%	ERS354270
PE0124-C	Tanzania: Muheza	Sequenced	100%	ERS354273
PE0125-C	Tanzania: Morogoro	Sequenced	100%	ERS354276
PE0126-C	Tanzania: Morogoro	Sequenced	100%	ERS354279
PE0127-C	Tanzania: Muheza	Sequenced	100%	ERS354281
PE0128-C	Tanzania: Morogoro	Sequenced	100%	ERS354283
PE0129-C	Tanzania: Morogoro	Sequenced	100%	ERS354285
PE0130-C	Tanzania: Morogoro	Sequenced	99.90%	ERS354287
PE0131-C	Tanzania: Morogoro	Sequenced	100%	ERS354289
PE0132-C	Tanzania: Morogoro	Sequenced	100%	ERS354291
PE0133-C	Tanzania: Morogoro	Sequenced	100%	ERS354265
PE0134-C	Tanzania: Morogoro	Sequenced	100%	ERS354268
PE0135-C	Tanzania: Morogoro	Sequenced	100%	ERS354271
PE0136-C	Tanzania: Morogoro	Sequenced	100%	ERS354274
PE0137-C	Tanzania: Morogoro	Sequenced	100%	ERS354277
PE0138-C	Tanzania: Morogoro	Sequenced	100%	ERS354280
PE0139-C	Tanzania: Morogoro	Sequenced	100%	ERS354282
PE0140-C	Tanzania: Morogoro	Sequenced	100%	ERS354284
PE0141-C	Tanzania: Morogoro	Sequenced	100%	ERS354286
PE0142-C	Tanzania: Morogoro	Sequenced	100%	ERS354288
PE0143-C	Tanzania: Morogoro	Sequenced	100%	ERS354290
PE0144-C	Tanzania: Morogoro	Sequenced	100%	ERS354292
PE0145-C	Tanzania: Morogoro	Sequenced	100%	ERS354266
PE0146-C	Tanzania: Morogoro	Sequenced	100%	ERS354269
PE0147-C	Tanzania: Muheza	Sequenced	100%	ERS354272
PE0148-C	Tanzania: Morogoro	Sequenced	100%	ERS354275
PE0149-C	Tanzania: Morogoro	Sequenced	61.40%	ERS354278

2.4.5 Mixture modeling

To estimate parameters of the component curves in the distribution of gene coverages (Figure 4-2), the R package *normalmixEM* was used on a vector of mean coverage depths for 5324 parasite genes. The default settings were used with the number of expected distributions (k) set to 3. Figure 2-2 shows the resulting curve estimates alongside a curve generated from a random normal distribution using the means and standard deviations estimated by *normalmixEM*. The randomly generated curve is similar to the distribution of real coverage depths depicted in Figure 4-2.

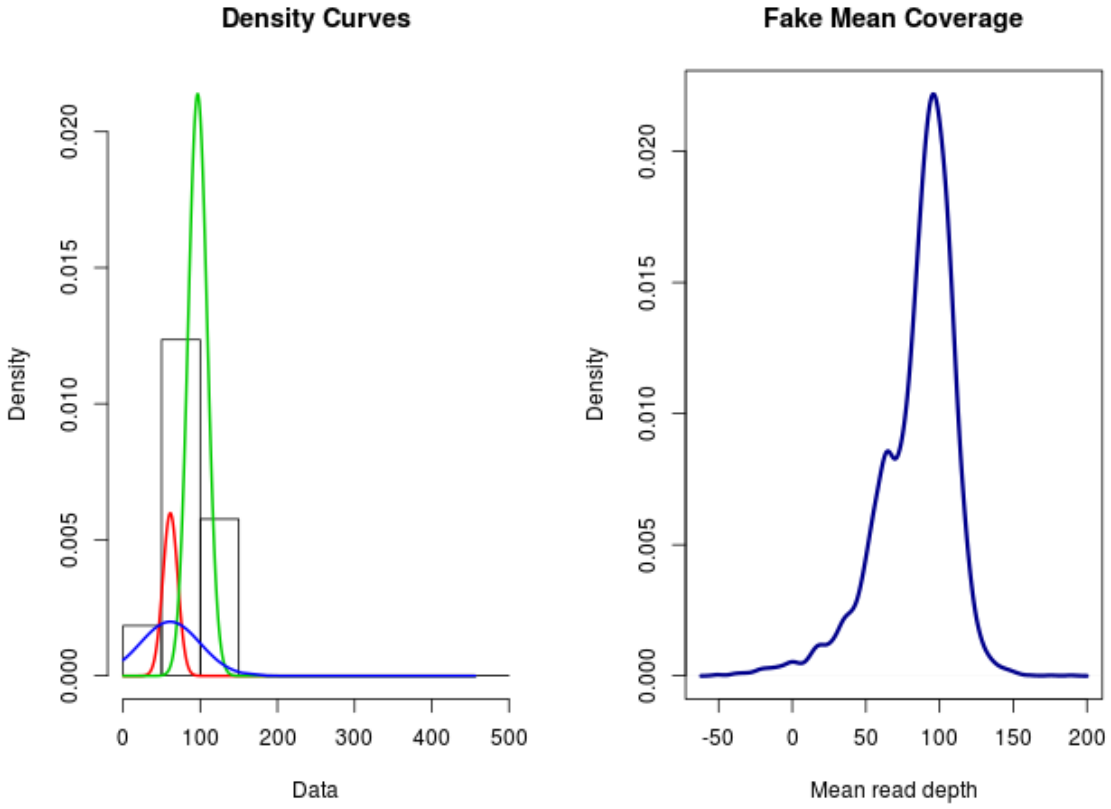


Figure 2-2. Mixture modeling of mean coverage depths of genes. Left) Density plot resulting from *normalmixEM* showing the 3 estimated distributions. **Right)** An artificial curve generated from 5324 generated values from random normal distributions with means and standard deviations set to those predicted on the left. This plot is a reality check comparison to the true distribution, and looks quite similar.

2.5 Materials and methods for chapter 5

2.5.1 Tanzanian field isolates

The source of the field samples used for these methods and validation are described in chapter 4.

2.5.2 PCR and qPCR typing of F and C indels in EBA-175

A multiplexed PCR reaction was designed containing primers that amplify a 190bp fragment from the F insert and a 157bp fragment from the C insert. For the F amplicon, the following primers were used: 5'TCAGGAACCAGCAAATAAGGA3', and 3'GTTTGGAATGGCGTCAGATT5'. For the C amplicon we designed the following primers: 5'GAAGTGCAACAGTGAGTGAATCT3', and 3'GCGCCTTTTTCATCTTCATC5'. 25µl reactions were performed using AccuPrime Pfx supermix (Life Technologies, cat. 12344-040) under the following conditions: initial denature, 95C for 2min; denaturation, 94C for 15s; annealing, 55C for 30s; extension, 65C for 1min for 35 cycles. 5µl of this product was loaded onto a 1% agarose gel. The same reaction was also used on a BioRad cfx96 qPCR machine, but with SYBR Green Real-Time PCR Master Mix (Life Technologies), and a post-run melt curve ramp from 55C to 95C in 0.1C increments was performed. In two cases (PE0030_C and PE0138_C) the melt-curve results looked strange and the gel results were used to determine genotype.

2.5.3 Running Malign

It is critical the names of the sequences in your fasta file be in the format below. The name "TPA_intron_h_rgn__1069__1379" has 3 elements separated by double underscores: i.e., "TPA_intron_h_rgn", "1069", and "1379". The last two elements in this nomenclature need to be the start and stop positions within the corresponding sequence that you want to test—i.e., your indel or a dimorphic region.

>TPA_intron_h_rgn__1069__1379
TCATGGAAGTGGTTCTTCCTGCTGAACCTGAAGATGTCCCCAGACTCTCTCTCAACATAGCAATCT
AGGGAG CCACTTCCGGTGGAGATGAAACCCC…

2.5.3.1 Installation

1. Install dependencies

   You need BWA, Samtools, R, and Java installed on your system and in your PATH. You should be able to enter "bwa aln", "Samtools view", "Rscript", and "Java" from

anywhere on your system and get an output of usage options if they are installed properly.

2. Clone repository (or download the zip file and enter the Malign directory).
   $ git clone https://github.com/jwendler/Malign.git
   $cd Malign
3. Make sure third party scripts are executable:
   $ chmod 755 pileup2baseindel_modified.pl

2.5.3.2 Examples

Example 1: Print available options.

$ Rscript Malign.R -h

Example 2: Use reads in fastq format tested against a fasta reference with two different indels. Notice the PDF provides two plots and the MalignResult.txt file has two columns for each indel.

$ Rscript Malign.R -r Exon1_3d7_Dd2.fa -sn HG02808_small_fq -F1 3d7_1.fastq.gz -F2 3d7_2.fastq.gz

Example 3: Use reads from a BAM file tested against a fasta reference containing one indel. Along with the reads from the -br region of the BAM, include unmapped reads (-u). This option is essential if your gene has large regions that are completely diverged, but not needed for a moderately conserved gene with even a large indel (<400bp). Notice in this output the -keep option prevented the fastq files created from the BAM reads from being deleted.

$ Rscript Malign.R -r TPA_h_rgn.fa -sn HG02808_small_keep -b HG02808_small.bam -br 8:42039279:42039968 -keep -u

2.5.3.3 Output

MalignResult.txt: contains a table of median coverages inside the defined window of each of your indels, and a TRUE/FALSE detection call as to whether this coverage reached the defined coverage threshold (option –ct). Following the sample name, each indel will have two columns giving median depth and the call. In example 2 above the columns would be: sample_name, indel1_depth, indel1_call, indel2_depth, indel2_call.

MalignPlots.pdf: plots for each sequence in the reference fasta input--two per page to make convenient for dimorphic genes.

allPileups.tab.gz: a zipped table containing the actual depths of each nucleotide at each position of your reference sequences.

[f1.fastq.gz, f2.fastq.gz]: optional outputs if the -keep option is used with a BAM input. These are fastq files created from the reads pulled out of your defined region of the BAM (-br option). Unmapped reads will also be included if the -u option is used.

2.5.3.4 Options

Option	Description
-h	This helpful list of options.
-b	Path to BAM file. Either need this OR F1/F2, not both.
-br	BAM region: e.g., Pf3D7_07_v3:1357455:1363529. If not provided with -b, only unmapped reads are used.
-u	Use unmapped reads from BAM (and -br region if provided).
-o	Out directory (also where files are created and deleted). [Default is MalignOut_sn]
-w	Window. Number of bp to ignore at the edges of the indel [(indel size < 3)=0, (indel size 3-10)=1, (indel size 11-100)=10%, (indel size > 100)=10].
-ct	Median coverage threshold needed to call indel as present [Default=1].
-sn	Sample name.
-r	Path to fasta file of references sequences. Each sequence name should give the coordinates of the insert: e.g., >id__start__stop
-F1	First fastq file (required if no -b BAM path provided).
-F2	Second fastq file (required if no -b BAM path provided).
-p	Path to pileup2baseindel_modified.pl. [can also set PILEUP_2_BASE_PATH in Malign.R]
-s	Path to SamToFastq.jar. [can also set SAM_2_FASTQ_PATH in Malign.R]
-keep	Keep fastq pulldowns if using -b option (zipped).

2.5.4 Meta-genes from sequence patching

A Perl program (CortexSeqPatcher.pl) was developed to patch Cortex discovered variants in the 3D7 reference version of a given gene. This program takes as input a Cortex VCF file, a reference sequence, and the 3D7 genomic coordinates for which the variants should be taken from the VCF. Variants considered are those with the VCF filter value of PASS. If more than one variant exists at the same position, only the first is taken forward. The 3D7 sequence is then fragmented around the variant coordinates and the Cortex allele calls are sliced in place of the 3D7 alleles. The nucleotide sequence and 3-frame translation is also part of the output.

2.5.5 Running Cortex

Fifty-nine samples representing parasites from 17 countries were genotyped using Cortex (Table 2-2). The script was run using kmer sizes of 31 and 61, with ploidy set to 1 (see full parameter list below for more detail). The combined VCF output from Cortex was parsed using Perl and read into R for analysis.

Table 2-2. Countries represented in Cortex run.

Country	BR	BF	GH	GM	GN	HN	KE	KH	LA	ML	MW	PE	PG	TH	TZ	UG	VN
count	1	1	6	4	4	2	4	4	5	4	4	3	3	4	4	2	4

2.5.5.1 Parameter settings

A Shell script was written to send Cortex call to the Sun Grid Engine with the SGE -l parameter set to `virtual_free=140G,h_vmem=140G,h_rt=9000:0:0`, and the cortex commands set as follows:

```
CORTEX_release_v1.0.5.14/scripts/calling/run_calls.pl --first_kmer 31
--last_kmer 61 --kmer_step 30 --fastaq_index index --auto_cleaning yes
--bc yes --pd yes --outdir OUTPUT  --outvcf mix62 --ploidy 1
--stampy_hash /home/jwendler/phd/data/ref/3d7
--stampy_bin /home/jwendler/installed_apps/stampy-1.0.13/stampy.py
--list_ref_fasta /home/jwendler/phd/data/reference/list_ref_chroms
--refbindir /home/jwendler/phd/data/ref/ --genome_size 24000000
--format FASTQ --mem_height 22 --qthresh 10 --mem_width 90
--vcftools_dir /home/jwendler/installed_apps/vcftools_0.1.9
--do_union yes --logfile log.txt --workflow independent
--ref CoordinatesAndInCalling --squeeze_mem
```

2.6 Materials and methods for chapter 6

2.6.1 Primer design

A modified version of the PCS.R software was used to produce a masked *eba-175* sequence for universal primer design. The modifications to PCS.R included adding remotely variable position as identified in Cortex genotyping of 59 worldwide isolates (section 2.5.5), GATK genotyping of three genetic crosses and progeny, as well as the MalariaGEN v2.0 SNPs.

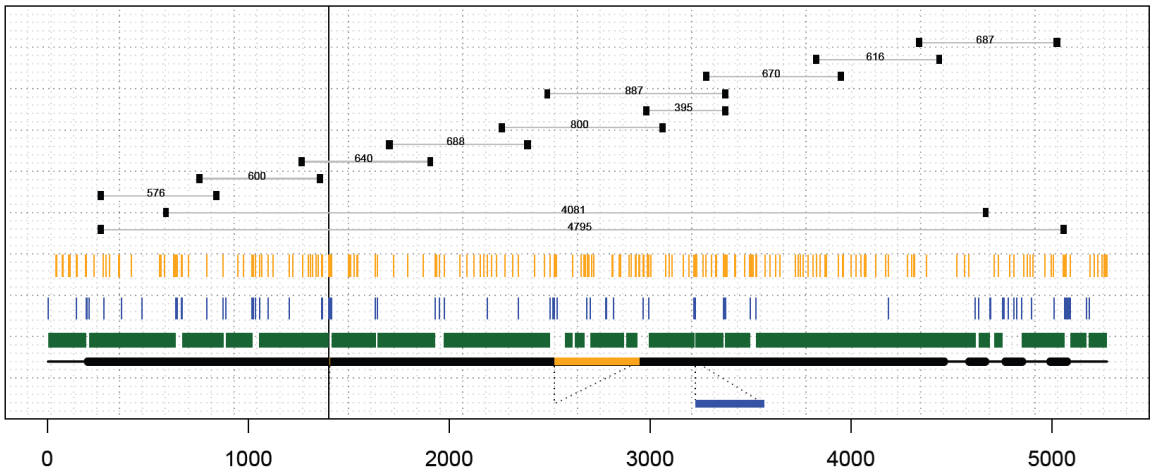


Figure 2-3. Modified PCS plot of *eba-175* for primer design. Bottom track shows the *eba*-175 gene model and the F and C indels respectively in orange and blue. Black vertical line at position 1200 indicates the position of the 6bp indel. Green track shows pan-conserved stretches (note, due to image resolution some gaps aren't visible). Variants with very loose filter criteria are shown as vertical bars for Cortex (blue) and GATK/MalariaGEN (orange). Tiling primer pairs are drawn plotted at the top, connected by thin lines with the predicted 3D7 amplicon length above (also shown in (Table 2-3).

Table 2-3. Primer list for *eba-175*.

Primer Name	start	length	Tm	GC%	Sequence	Product length
eba175_Full_F1	253	25	61.21	32	TTGCAAAAGCTAGGAATGAATATGA	4795
eba175_Full_R1	5047	24	61.47	37.5	TCCTCATGGTATTCAGAAAAATCG	4795
eba175_Full_F2	579	23	60.13	47.83	AACGCTGTACGTGTGTCTAGGAT	4081
eba175_Full_R2	4659	22	61.98	40.91	TTGGCTTGTGAAGCACCTAAAA	4081
eba175_Frag1_F	253	25	61.21	32	TTGCAAAAGCTAGGAATGAATATGA	576
eba175_Frag1_R	828	24	61.41	37.5	CAATGAAATGATCCTTCATGGTCT	576
eba175_Frag2_F	744	24	61	41.67	CCTGATCGTAGAATCCAATTATGC	600
eba175_Frag2_R	1343	26	61.79	34.62	CGATAACGTATGCCATTCAAATTTAC	600
eba175_Frag3_F	1255	23	60.02	39.13	ATGAAGCATGTGAGAAGGAATGT	640
eba175_Frag3_R	1894	24	60.65	29.17	TTGCTCAAATCATTCCAATAATCA	640
eba175_Frag4_F	1689	28	60.78	32.14	TGTCTTGGAAACATTGATAGAATATACG	688
eba175_Frag4_R	2376	28	60.46	28.57	CATCTTCGAAATTTAGGTTAGAACATTT	688
eba175_Frag5_F	2250	25	60.45	32	GCCAAACAATACCAAGAATATCAAA	800
eba175_Frag5_R	3049	26	60.96	38.46	GGTCTTGAATTTTCTGGTGATACACT	800
eba175_Frag6_F	2970	23	61.36	34.78	GGACCAAAAGGAAATGAACAAAA	395
eba175_Frag6_R	3364	23	60.26	43.48	GCTTTTCCTTCATCCAAGCTACT	395
eba175_Frag7_F	2478	20	62.19	50	TCGCAAGAAGCAGTTCCTGA	887
eba175_Frag7_R	3364	23	60.26	43.48	GCTTTTCCTTCATCCAAGCTACT	887
eba175_Frag8_F	3269	23	61.43	43.48	TGTTCAACAGTCTGGAGGAATTG	670
eba175_Frag8_R	3938	24	60.04	37.5	GTGTCTTTCGTTTTCCTCATTTCT	670
eba175_Frag9_F	3814	25	61.96	36	CAAAAATTCATAAGGCTGAAGAGGA	616
eba175_Frag9_R	4429	22	61.81	40.91	GGATCATCAAATTCCCTTTTCG	616
eba175_Frag10_F	4326	25	61.74	36	AAAACAAGAAATCTGTGTTGTGCAG	687
eba175_Frag10_R	5012	26	61.23	34.62	TTCCTGTACTTGTTGATTGACTGAAA	687

Primers were ordered from Metabion (Steinkirchen, Germany) and resuspended in nuclease-free water (Promega) to make a 100uM stock. Stock primers were diluted to 10uM to make working solutions.

2.6.1.1 Acknowledgments

The GATK VCF files from the three genetic crosses were kindly provided by Alistair Miles.

2.6.2 PCR amplification and Sanger sequencing

PCR reactions were performed on field and lab isolates to amplify the two longest products in the *eba-175* primer list (4081bp and 4795bp). In each reaction, 35ng of parasite DNA was combined with 800nM of the appropriate primer pair (Full_F1/R1 or Full_F2/R2) and 25µl of ProMega GoTaq master mix (M4021), and brought up to 50µl with nuclease-free water. The following thermocycler program was used:

Table 2-4. PCR program for *eba-175* long fragments.

Step	Temp (C)	Time
Initial denature	95	2 min
Denaturation	94	15 sec
Annealing	56 (~ Tm-5)	30 sec
Extension	65	5min (~ 1 min/ kb)
Cycle #	35x	
Final extension	72	10
Hold	4	infinity

5µl of each reaction product was run on a 1% agarose gel. The resulting bands were in the expected size range (Figure 2-4 A), but for validation before capillary sequencing with tiling primers, a restriction digestion was performed. It was determined that EcoRI would make a single cut in both the 4081bp and 4795bp products, yielding 8 potential digestion fragments for the F and C versions of *eba-175* (Table 2-5).

Table 2-5. EcoRI products for *eba-175* amplicons.

Allele	Primers	Fragment 1: size	Fragment 2: size
F	eba175_Full_F1, R1	1-2936: **2936**	2937-4795: **1859**
F	eba175_Full_F2, R2	1-2610: **2610**	2611-4081: **1471**
C	eba175_Full_F1, R1	1-2513: **2513**	2514-4691: **2178**
C	eba175_Full_F2, R2	1-2187: **2187**	2188-3979: **1792**

As shown in figure Figure 2-4 B, all digested amplicons yielded products of the expected size, so it was trusted that this product was indeed *eba-175* exon 1. This same protocol was used to amplify the largest fragment (4795bp) in a repeat of these samples, plus additional field isolates for capillary sequencing with 20 tiling primers each.

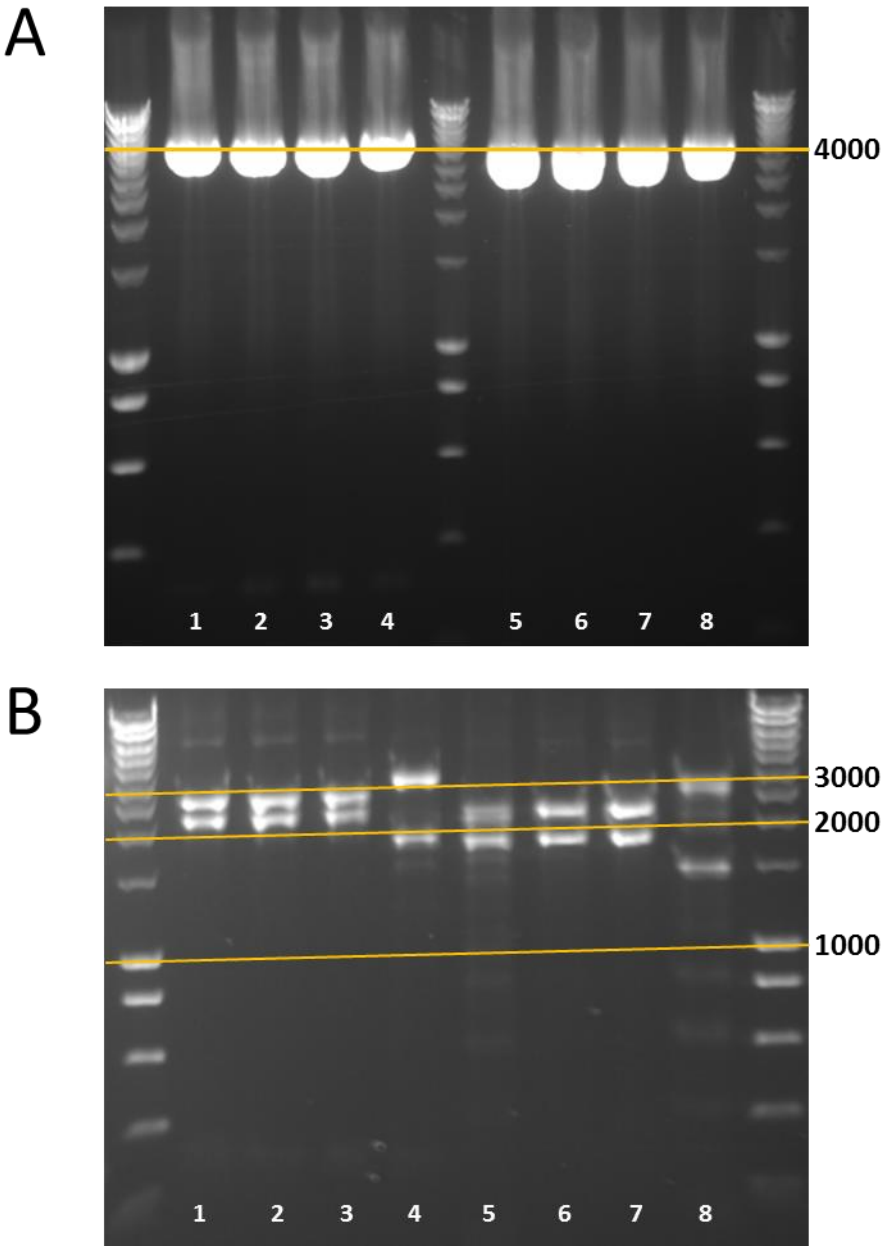


Figure 2-4. Amplification and EcoRI digestion of *eba-175* exon 1 amplicons. Bioline HyperLadder 1kb was run in non-numbered lanes. Lane contents and expected product sizes are listed in Table 2-6. **A)** initial amplifications from lab and field isolates. **B)** corresponding EcoRI digestions.

Table 2-6. Expected product sizes for Figure 2-4.

Lane	Sample	Genotype	Primers	Expected size
A1	PE0030	unknown	eba175_Full_F1/R1	4795 or 4714
A2	PG0190 (Dd2)	C	eba175_Full_F1/R1	4714
A3	PG0200 (HB3)	C	eba175_Full_F1/R1	4714
A4	PG0218 (3D7)	F	eba175_Full_F1/R1	4795
A5	PE0030	unknown	eba175_Full_F2/R2	4081 or 4000
A6	PG0190 (Dd2)	C	eba175_Full_F2/R2	4000
A7	PG0200 (HB3)	C	eba175_Full_F2/R2	4000
A8	PG0218 (3D7)	F	eba175_Full_F2/R2	4081
B1	PE0030	unknown	NA	unknown
B2	PG0190 (Dd2)	C	NA	2178 and 2513
B3	PG0200 (HB3)	C	NA	2178 and 2513
B4	PG0218 (3D7)	F	NA	1859 and 2936
B5	PE0030	unknown	NA	unknown
B6	PG0190 (Dd2)	C	NA	1792 and 2187
B7	PG0200 (HB3)	C	NA	1792 and 2187
B8	PG0218 (3D7)	F	NA	1471 and 2610

The above PCR amplification was repeated for all samples and two additional field samples (PE0027 and PE0028). Reactions were purified using the Qiagen QIAquick kit (cat. 28106) and eluted in 80μl buffer EB. Samples were diluted to 30ng/μl and sent for sequencing with the 20 tiling *eba-175* primers. Sequencing primers were diluted to 3.2 pmol/μl. Capillary sequencing reads were combined to one fasta file and assembled into one consensus sequence using Velvet (version 1.2.07) with kmer size of 81, expected coverage set to auto, and input type set to -long and -fasta. The longest resulting contig was taken as the consensus sequence.

2.6.3 *in silico* mixtures to assess MalMOI error rate

The software package fastqMixer (see section 2.2.3) was used to mix a dominant parasite line with one or more others at a range of proportions. 3D7 and Dd2 were used as the dominant parasites, and were mixed with between 1 and 4 other lab lines (7G8, GB4, HB3, 3D7, Dd2). Reads were randomly sampled from the dominant parasite sample such that it represented between 50-95% of the mixture in 2-5% intervals, depending on the number of parasites in the mixture. The rest of the proportion was split evenly between the non-dominant strains. At each interval, 100 mixtures were performed, each sampling reads with replacement. This was done separately for *msp3.4* and *ama1.*

2.7 Materials and methods for chapter 7

2.7.1 Multiple sequence alignments of DNA and protein from *de novo* assembled genes

The MalMOI pipeline was used to assemble individual *eba-175* exons, and where possible full length genes (including introns). For samples for which all 4 exons were assembled (but perhaps not introns), the exons were spliced together to make a coding sequence CDS. Individually assembled samples were concatenated into a single file, translated into protein sequences, and aligned using MAFFT with default settings. Assemblies with stop codons introduced within the coding sequence were excluded from the alignment (65 out of 2153). The protein alignment and DNA sequence was then used with the software RevTrans to produce an alignment at the nucleotide level. Alignments were checked visually in Jalview, particularly to ensure the indel boundaries were consistent in all sequences. I find that alignments produced at the protein level need fewer manual corrections, thus I use them as guides for the DNA alignments.

2.7.2 Genotyping indels at a population scale

2.7.2.1 Malign

The software package Malign, described in chapter 5, was used to detect the presence of the 6bp, F, and C indels in *eba-175*. Genotypes were determined for 2804 Illumina sequenced *P. falciparum* samples contributed by MalariaGEN partners from around the globe. For the 6bp indel, one nucleotide at each edge of the insert was ignored (i.e., the Malign window parameter w=1) when detecting coverage to both decrease noise at the indel boundary, and to avoid a SNP that lies at the 3' terminal position. A window size of 10bp was used for the larger F and C indels. As the absence of coverage within the indel region is used to infer a deletion, a conserved stretch of *eba-175* was used as a positive control, and 2526 samples were determined to have at least 5x coverage in this control region. There are 4 SNPs within 40bp on either side of the 6bp indel, so to reduce the possibility of false negative detection of this indel, two common haplotypes were used as reference sequences in Malign. Using a Malign coverage threshold of 2, there were 27 samples out of 2526 for which the 6bp indel calls were discordant when aligned to these two haplotypes. These samples were excluded from further consideration. Although all 3 indels are within exon 1, the smaller 6bp indel was genotyped separately so a smaller gene segment could be used as a reference to enhance visibility of the indel in the Malign pileups.

2.7.2.2 *de novo* assemblies

To determine the genotypes of the *eba-175* 6bp, F, and C indels based on the *de novo* assembled samples, the RevTrans exon 1 multiple sequence alignment was read into an R session using the seqinr package, and gap lengths were counted directly. Sequences were defined as having an F deletion if a 423bp gap was present between alignment positions 2325 and 2750, a C deletion if a 342bp gap was present between positions 3025 and 3375, and a 6bp deletion if a 6bp gap was present between nucleotide 1200 and 1207.

2.7.3 PCR for detecting F/C double insertions and deletions

Two sets of primers were combined from Table 2-3 to amplify a region of *eba-175* containing both the F and C indels: Set 1) eba175_Frag7_F with eba175_Frag7_R, and set 2) eba175_Frag5_F with eba175_Frag6_R. Table 2-7 lists the expected amplicon sizes for different F/C genotypes with different theoretical indel combinations and primer pairs.

Table 2-7. Expected amplicon sizes for double indel events.

F	C	Primer set	Size
+	+	1	1229
+	-	1	887
-	+	1	806
-	-	1	464
+	+	2	1457
+	-	2	1115
-	+	2	1034
-	-	2	692

2.7.4 Population genetic analyses

DNA Sequence Polymorphism (DNASP) version 5.10.01 was used to calculate Tajima's D and pairwise linkage disequilibrium(LD) based on r^2 [212]. SNPs with MAF<5% were excluded from consideration for LD. Tajima's D was calculated within 100bp windows, staggered by 25bp. This analysis was performed once excluding indels, and once with artificial SNPs

representing the indels. The R packages seqinr and ape were used for sequence manipulation and tree construction. Distance matrices were calculated based on percent identity, and the tree hierarchy was based on neighbor joining of pair-wise distances [213]. Geographic regions were classified as follows (country abbreviations are listed in the front-matter list of abbreviations and acronyms):

SEA: MM, PG, KH, VN, TH, BD, LA

SAM: BR, CO, PE, HN

EAF: TZ, MW, SD, KE, UG, MZ

WAF: GN, ML, GH, GM, BF, NG

Fixation index (Fst) estimates were performed on these geographic stratifications including indels as a fifth genetic state.

The MEGA6 software was used for codon-by-codon dN/dS calculations [214]. Gaps were eliminated from consideration in codon calculations. 1419 *de novo* assembled CDS sequences were used. Maximum likelihood estimations of dN and dS were performed using the HyPhy software [215].

2.7.5 Mapping variation to 3-dimensional protein structure

The *EBA-175* region II crystal structure solution (1ZRL) was downloaded from the NCBI Structure database and viewed in the Cn3D software [158,216]. Polymorphism identified using 1419 *de novo* assembled EBA-175 proteins was mapped to the Cn3D sequence and the image was exported.

2.7.6 Samples and ethics approvals

The samples used for the host-parasite interaction GWAS are a part of a larger consortial project (CP1) within the Malaria Genomic Epidemiology Network (MalariaGEN). Partners leading independent investigations in Kilifi, Kenya and Fajara, Gambia have contributed samples for genotyping from their ongoing projects. As stated in previous MalariaGEN publications using these data [217]: all research was reviewed and approvals granted by local Research Boards and Ethics committees in The Gambia: The Gambia Government/MRC Unit Joint Ethics Committee (SCC1029 and SCC670/630); Kenya: Research Ethics Committee from the KEMRI-Wellcome Research Programme, Kilifi, Kenya (SCC1192); and Oxford:

Oxford University Tropical Research Ethics committee (OXTREC), Oxford, United Kingdom (OXTREC 020-006).

2.7.7 GWAS

This GWA analysis benefitted tremendously from an existing MalariaGEN infrastructure established to investigate human associations with severe disease [217]. DNA was extracted either at the study site or by MalariaGEN and integrated into a standardized process of quality control and whole-genome amplification, as previously described [217,218]. Kenyan samples were genotyped on the Illumina HumanOmni2.5-4 microarray and those from Gambia on the HumanOmni2.5-8 chip. After rigorous quality controls, sample genotypes were phased within respective populations and imputed onto the 1000 genomes reference panel. Sample relatedness was determined by PCA using SHELLFISH.

Using the above infrastructure I separately ran SNPTEST v2.5 on Kenyan and Gambian imputed SNPs to test for associations with the parasite *eba*-175 indel genotyped from the same sample using Sequenom. The additive genetic model option in SNPTEST was used for all analyses reported here. Parasite genotype was dichotomized and modeled as the outcome in a logistic regression framework. For each human SNP, the host genotype was included in the model with the first five principal components to control for population structure. SNPs with MAF > 1% were considered.

2.7.8 Sequenom genotyping

Two Sequenom multiplexes typing 51 SNPs were applied to 2458 Kenyan and 2975 Gambian samples. Twenty-seven assays typed variants in *eba-175*, including three that directly typed the F, C, and 6bp indels. Five of the SNP assays typed the F and C indels with indicator SNPs. See section 7.3.1 for a description of direct typing versus indicator assays. To minimize errors due to low parasitemia, samples that didn't return a genotype for at least 90% of a panel of 18 SNP assays across *eba-175* were excluded (N = 935).

SECTION I: APPLICATIONS AND LIMITATIONS OF SINGLE NUCLEOTIDE POLYMORPHISM

This thesis is about accessing complex variation in malaria parasites isolated from field samples. The motivation for investigating complex variation derives from several artefacts encountered during SNP-based analyses. The two chapters of this section describe a GWAS and a population genetic study, both which utilize genome-wide parasite SNPs ascertained by aligning short reads to the 3D7 reference. While these chapters demonstrate the power of this type of variation along with resulting biological insights, they also present specific limitations, and conclude that different approaches are necessary to access particular regions of the genome. This aspect of Section I segues into methods development for accessing complex variation in Section II, with specific attention paid to problematic genes discussed in these two chapters.

3 A GWAS OF *P. FALCIPARUM* SUSCEPTIBILITY TO 22 ANTIMALARIAL DRUGS IN KENYA

3.1 Abstract

Drug resistance remains a chief concern for malaria control. In order to determine the genetic markers of drug resistant parasites, genome-wide association tests were performed on sequence-based genotypes from 35 Kenyan *P. falciparum* parasites with the activities of 22 antimalarial drugs.

Parasites isolated from children with acute febrile malaria were adapted to culture, and sensitivity was determined by *in vitro* growth in the presence of anti-malarial drugs. Parasites were genotyped using whole genome sequencing techniques. Associations between 6250 single nucleotide polymorphisms and resistance to individual anti-malarial agents were determined, with false discovery rate adjustment for multiple hypothesis testing. The expected associations in the *pfcrt* region with chloroquine (CQ) activity were found. Separately, novel loci were associated with amodiaquine, quinazoline, and quinine activities. Association signals for CQ and primaquine (PQ) overlap in and around *pfcrt*, and interestingly the phenotypes are inversely related for these two drugs. A catalog of the variation in *dhfr*, *dhps*, *mdr1*, *nhe*, and *crt* was created, including that of novel SNPs, and the presence of a *dhfr*-164L quadruple mutant in coastal Kenya is confirmed. Mutations implicated in sulfadoxine-pyrimethamine resistance are at or near fixation in this sample set.

Sequence-based GWA studies are powerful tools for phenotypic association tests. Using this approach on *P. falciparum* parasites from coastal Kenya, known and previously unreported genes associated with phenotypic resistance to anti-malarial drugs are identified. Further, high-resolution haplotype visualizations show a possible signature of an inverse selective relationship between CQ and PQ.

3.2 Introduction

Two million Kenyans will contract malaria this year, consuming at least a fifth of the country's hospitalization resources [219]. Prompt treatment with antimalarials can prevent mortality, but this efficacy is threatened by the parasite's ability to acquire drug resistance. This highlights the appeal of high-resolution genetic markers and data-sharing for early-warning surveillance [220]. Additionally, the elucidation of genetic loci underlying resistance is important for designing new formulations, and can reveal opposing selective pressures amongst drugs [221].

Targeted gene approaches have been successful tools for assessing allelic associations with drug resistance, but are obviously limited to known loci. Drug resistance loci in *P. falciparum* parasites have historically been discovered using genetic crosses for QTL analysis [222,223]. A number of recent studies targeting candidate parasite genes in coastal Kenya have described drug activity associations with familiar SNPs in *pfmdr1*, *pfcrt*, and *pfdhfr*, as well as structural associations with quinine (QN) tolerance in *pfnhe* [210,224,225]. Population-genetic approaches, such as sequence-based GWAS, provide the advantage of testing for phenotypic associations with novel SNPs while broadly surveying known polymorphisms [226]. In addition to important novel loci, null associations in adequately powered GWA studies are informative about which drugs might be effective in a given geographic region, and whether low frequency resistance variants are transmitting in the population. Parasites exposed *in vitro* to an array of chemical compounds cluster geographically in their responses, illustrating the tremendous impact drug exposure has had in shaping the parasite genome, and thus the relevance of population-based approaches for investigating drug resistance [227].

This work examines the association between SNPs ascertained from whole-genome sequencing of 35 Kenyan field isolates with the activities of 22 antimalarial drugs (Table 3-1). The cooperative efforts of the partnerships in MalariaGEN have created a panel of highly credible SNPs ascertained in the context of 1685 parasites (version 2 release), contributed from 17 countries, and this community resource is utilized here [87]. Given the difficulties tagging causal loci in genomes with short-range LD, it has been recently argued that sequence-based GWAS be fully adopted for such studies in *P. falciparum* [228].

Table 3-1. Drugs and abbreviations used in this study.

Drug	Abbreviation	Units
Amodiaquine	AMOD, AQ	nanomol
Atovaquone	ATV	nanomol
Chloroproguanil	CHLOPROG	nanomol
Chloroquine	CQ	nanomol
Cycloproguanil	CYCLOPG	nanomol
Desethylamodiaquine	DEAQ	nanomol
Dihydroartemisinin	DHA	nanomol
Halofantrine	HLF	nanomol
Isoquinine	ISOQIN	ng/ml
Lumafantrine	LUM	nanomol
Methotrexate	METHOT	nanomol
Methylene Blue	METHYLBL	nanomol
Mefloquine	MFL	nanomol
Piperaquine	PIQ	nanomol
Primaquine	PRIM, PQ	nanomol
Pyrimethamine	PYRIM	nanomol
Pyronaridine	PYRON	nanomol
Quinine	QIN, QN	nanomol
Quinazoline	QuiNazol	ng/ml
Trimethoprim	TRIMETHO	nanomol
Trimethotrexate	TRIMTX	nanomol
WR99210	WR99210	nanomol

3.3 Materials and Methods

Most of the methods for this chapter are found in section 2.3. The methods below define genotyping and various genetic models considered for the GWAS. This chapter reads more fluidly with this material contained herein.

3.3.1 Within-sample heterozygosity

For clarity in the upcoming discussion on model selection, an illustration of within-sample heterozygosity is first provided. More than one parasite genetic form is often present in the same infection or sample. In next-generation sequence data this multiplicity of infection manifests as heterozygosity in SNPs where the parasites have different alleles (i.e., different nucleotides). For example, Figure 3-1 shows a cartoon pileup of sample reads that might result from an infection containing two parasites. At the time of sample collection these two

parasites were present at a 6:4 ratio of the G and A alleles, respectively. For this hypothetical sample the genotype at the variable position 10 could be represented in many different ways—for example, as a majority call or as a ratio. Different genotyping methods will differentially impact phenotypic association tests, and this is discussed further in the next section.

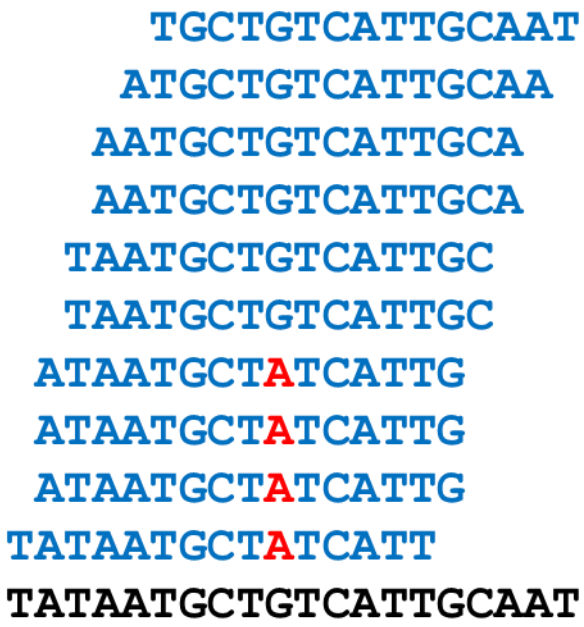


Figure 3-1. Illustration of within-sample heterozygosity. The black sequence along the bottom represents the reference genome, onto which the blue sequence reads from a sample with an MOI of 2 parasites are aligned, producing a "pileup." Position 10 of the reference is a SNP for which the two parasites have different alleles (G/A). At the time of the blood draw these two hypothetical parasites were present at a 6:4 G:A ratio.

3.3.2 Genetic model definitions

The following genetic models were considered for the association study. Here I provide concise definitions and example calculations for completeness (Figure 3-2). The final model selected was "z12." The process and rationale for selecting the z12 model is described in the results section 3.4.1.

To understand these definitions it is helpful to envision position 10 in Figure 3-1 (the heterozygous SNP). Genotyping (in this specific context of defining a genetic model) is the process of deciding whether position 10 should be classified as an A or G, something in between, or perhaps dropped. Although the pileup in Figure 3-1 represents a sample containing more than one parasite, genotyping will generate a single value representation for it. A different sample might have all G alleles, rendering the decision simple. If there

were 35 parasites in the GWAS, the end result will be a list of 35 genotypes (and missing values) to be matched to a list of 35 IC_{50} values, along with any other covariates (e.g., principal component for correcting population structure). For all of the models below, samples with read depths less than 5 at the SNP under consideration were dropped. The arcane model names match the variables used in the underlying R programs, thus for consistency were not changed to something more readable.

z12

This is the most conservative model and the one selected for use in the final GWAS. Any sample with evidence of within-sample heterozygosity is dropped (only for that particular SNP). Parasites matching the reference genome are assigned a 1, and otherwise a 0. The 0/1 genotypes are used as binary outcomes in generalized linear models (e.g., logistic regression).

f1

For f1 genotypes the within-sample reference allele frequency is calculated. As with the z12 model, a 1 represents a sample with 100% reference alleles in the reads at that SNP, and a 0 if all reads contain alternate alleles, however, samples with evidence of mixture are now retained. In our working example, position 10 would be assigned a genotype of 0.6. This value would be used as the outcome variable in a more traditional regression analysis.

Z12 soft

This model falls somewhere in between z12 and f1. A binary 0 or 1 is decided for each sample, but heterozygous positions are retained as long as the minor allele is present at very low abundance. A sample is classified as 1 as long as fewer than 5% of the reads represent the alternate allele. Similarly, a sample is classified as 0 as long as fewer than 5% of the reads represent the reference genome. The basis of this somewhat arbitrary 5% threshold is discussed below.

Dominant

This model was conceived under the hypothesis that even a small portion of resistant parasites in a mixed sample would have a dominant effect on the phenotyping assay. Under the assumption that non-reference alleles are dominant, samples are classified as 0 if the alternate allele frequency is at least 20%. One downside of this model is that genotyping under the assumption that the alternate allele is dominant yields different results than if the reference allele were dominant. This could be appropriate if 3D7 parasites are known to be

sensitive to the drug being tested, otherwise a separate GWAS would need to be performed under each assumption.

Read depth	Sample 1	Sample 2	Sample 3	Sample 4
Reference	10	6	96	2
Alternate	0	4	4	2

↓

Model	Sample 1	Sample 2	Sample 3	Sample 4
z12	1	NA	NA	NA
f1	1	0.6	0.96	NA
z12 soft	1	NA	1	NA
Dominant (alt)	1	0	1	NA

Figure 3-2. Example genotypes calculated under different genetic models. The top blue table depicts the reference and alternate allele read depths for a hypothetical SNP in 4 different samples. The bottom green table shows the resulting genotype under the different models. Dropped samples are designated NA.

3.4 RESULTS

3.4.1 Genetic model selection

3.4.1.1 The assumptions for the f1 model are violated

One possibility for the effect of MOI on drug resistance is that for an infection containing a mixture of resistant and susceptible parasites, the ratio of that mixture would directly relate to the drug sensitivity of the sample. For a given heterozygous SNP, this ratio would be reflected by its reference allele frequency, which could be used as a continuous variable in a regression analysis testing drug sensitivity. For example, if allele A in Figure 3-1 conferred drug resistance, under this hypothesis the sample would be 40% resistant. If the parasite relative abundances proportionally impacted the phenotyping assay, then a regression approach would provide greater statistical power than somehow categorizing a within-sample heterozygous SNP as one allele or the other. This concept is demonstrated by visualizing the raw data at a single locus in Figure 3-3. This SNP, located in a serine/threonine protein kinase, was selected because it exemplifies a dose-like response to CQ concentration—though almost certainly due to chance. In this example nearly 50% of

the data is lost using the z12 model (red dots in the central panel), as compared to the f1 model. Further, at this locus there appears to be a dose-like relationship, thus the log-fold drop in significance between the two models. If the underlying biology were truly driving this result, this SNP would demonstrate the power of using a model that incorporates MOI. However, most SNPs look more similar to Figure 3-4, and as discussed below, this contra-indicates the f1 model.

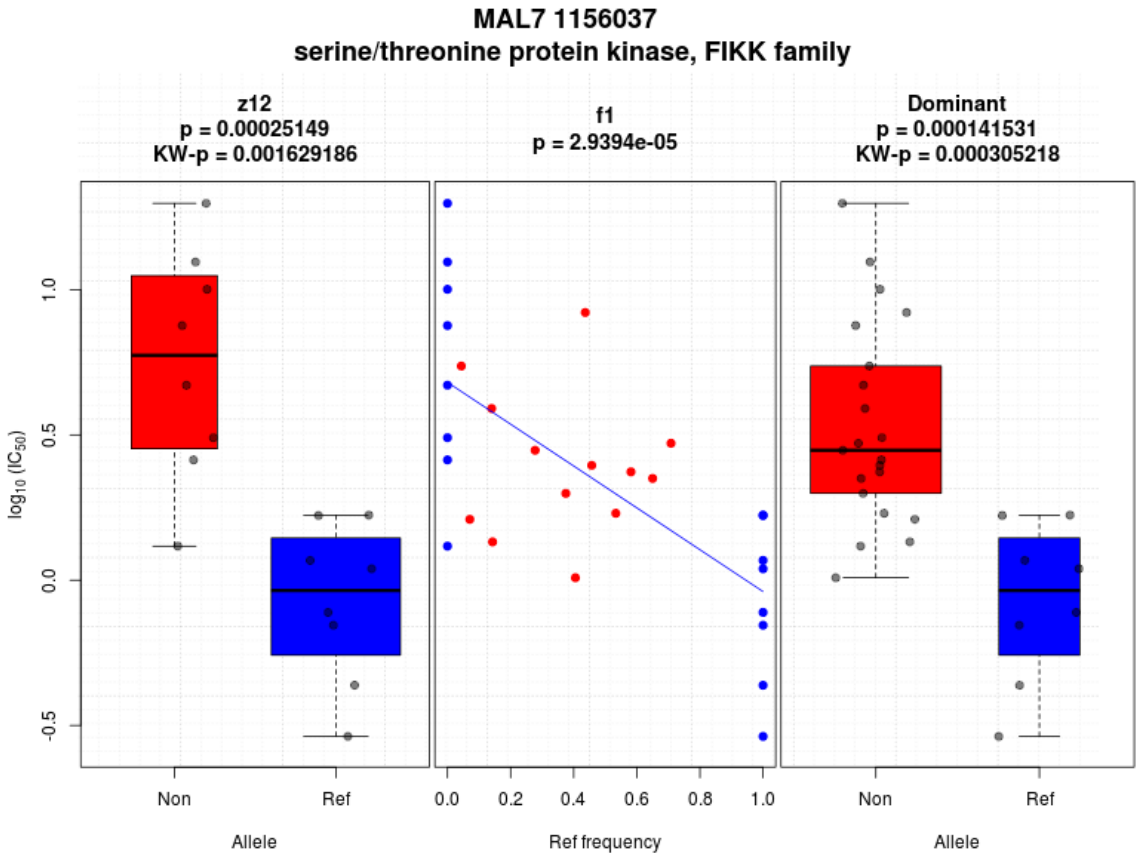


Figure 3-3. Association analysis for a heterozygous SNP under different models. This SNP was selected only to demonstrate an idealized example for the f1 model. Each panel shows the results and raw data of an association test calculated under the given genotyping model above. GWAS p-values are printed below the model name. Kruskal-Wallis (KW) nonparametric p-values are also given for dichotomous genotypes. Alleles are designated along the x-axes as Ref=reference, Non=non-reference, or the reference allele frequency in the case of the f1 model. For the f1 model, heterozygous samples at this locus are colored red, homozygous are colored blue, and the least squares regression line is drawn. Note in the z12 panel that the red dots are dropped. Boxplot box widths are proportional to samples size, and horizontal jitter added to dots for clarity.

A major confounder with the f1 hypothesis is that the relative abundance of parasites mixed within the same infection, or in culture, is a dynamic process, and the DNA used for sequencing is often collected at a different time from when phenotyping measurements are

taken. Indeed, in this study the chemosensitivity assays were performed up to two months before the parasites were harvested for sequencing (Supplementary figure 3-18). It is well known that some parasites adapt to culture more easily than others, and even stochastically the different populations within a sample may rise and fall [229]. Thus, the relative abundance of parasites in mixed samples would likely have changed over the 2 month phenotyping period. Although substantial within-sample heterozygosity remains in the 35 samples used in this study even after culture adaptation (Supplementary figure 3-22), it is for these reasons (i.e., the gap in time between genotyping and phenotyping, and the fact that most SNPs do not reflect the dynamic range seen in Figure 3-3) that I decided against modeling MOI as a continuous genotype with the f1 model.

3.4.1.2 The z12-soft model deflates the test statistic

In the samples used for this GWAS, most SNPs exhibiting within-sample heterozygosity have allele frequencies that are very close to 0 or 1, the implication being that in many samples the minor parasite is at very low abundance. Specifically, 93% of SNPs fall within 5% of either homozygous extreme (Supplementary figure 3-23). Given this, a rational approach to genotyping a heterozygous SNP might be to assign it the majority allele, as long as it represents at least some threshold percent of the reads, and to otherwise drop the sample for that position. A threshold of 100% would be the most conservative, discarding samples with any evidence of allele mixture (e.g., the z12 model). As the threshold is decreased (as with the z12-soft model) the sample size would increase for a given SNP, however the possibility of introducing noise would increase as well, perhaps resulting in a deflation of statistical power genome-wide. Indeed, based on the distribution of within-sample allele frequencies a GWAS was performed on SNPs genotyped using a threshold of 95%, and there is evidence indicating p-value deflation. If the null hypothesis were true for every SNP in a GWAS (i.e., that there is no difference in drug sensitivity), then the distribution of resulting p-values should be uniformly distributed between 0 and 1. A quantile-quantile plot is a visual depiction of this axiom, comparing the observed p-values to the expected values from a uniform(0,1) distribution. Significant deviations from this expectation indicate the statistical model may be misspecified or that a confounder, such as population structure, may be lurking. The genome-wide inflation factor (λ) is a separate metric for this, and should be close to 1. A λ value greater than 1 indicates more SNPs than expected are significant (i.e., p-values are inflated), and a value less than 1 suggests the association tests may be deflating significance. Using a 95% threshold for genotyping, a GWAS testing for chloroquine associations resulted in a λ of 0.95, whereas a threshold of 100% yielded a λ of

0.99. The deflated λ value in the first GWAS may be evidence that the gain in sample size from lowering the threshold also introduces noise, which more than offsets any potential gain in power.

3.4.1.3 The z12 model is conservative

In conclusion, genetic models that incorporate MOI (e.g., the f1, z12-soft, and dominant models) are theoretically promising for parasites direct from blood, but warrant further investigation. Little is understood about the dynamics of MOI as isolates adapt to culture. It was therefore decided to adopt a more conservative approach and discard heterozygous observations completely. Applying the z12 model to the 35 samples used in this GWAS yielded 6250 SNPs that represent the isolates without ambiguity.

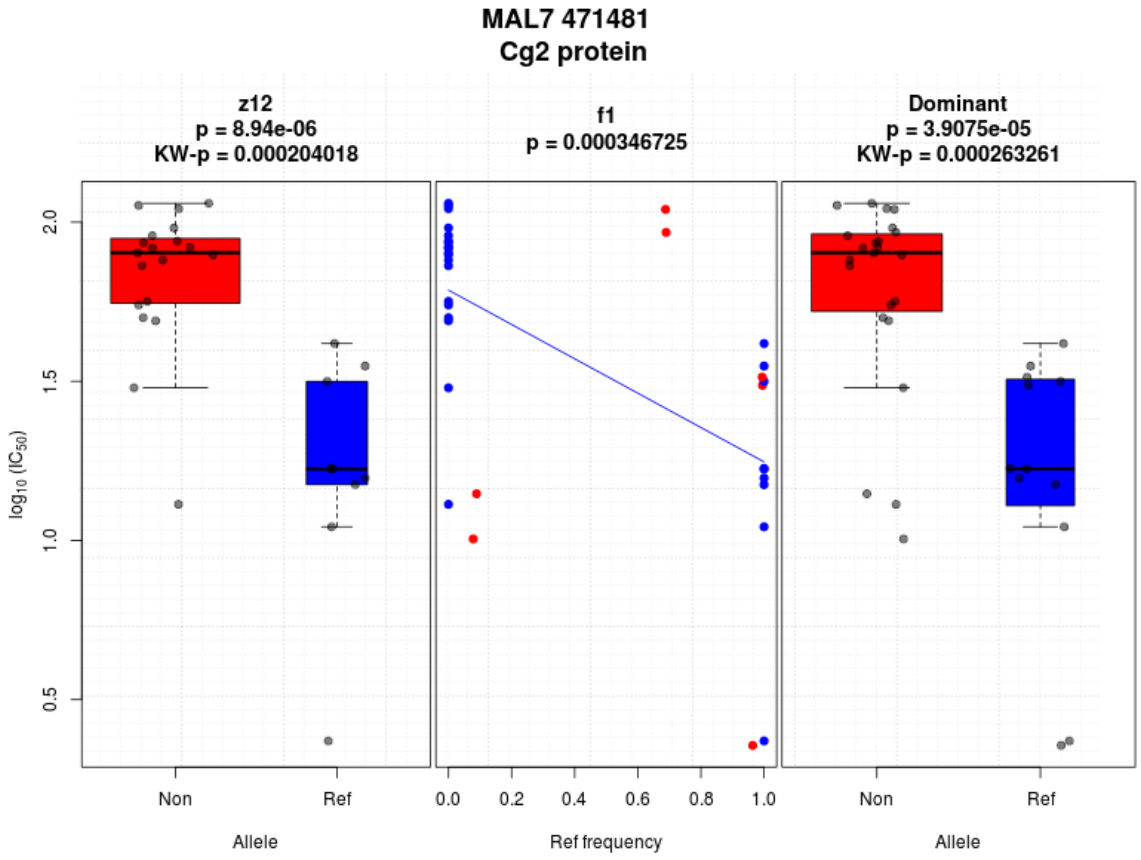


Figure 3-4. Association analysis for a heterozygous SNP under different models. This SNP was selected to demonstrate the effect on analysis if the f1 model assumption is wrong. Each panel shows the results and raw data of an association test calculated under the given genotyping model above. GWAS p-values are printed below the model name. Kruskal-Wallis (KW) nonparametric p-values are also given for dichotomous genotypes. Alleles are designated along the x-axes as Ref=reference, Non=non-reference, or the reference allele frequency in the case of the f1 model. For the f1 model, heterozygous samples at this locus are colored red, homozygous are colored blue, and the least squares regression line is drawn. Note in the z12 panel that the red dots are dropped. Boxplot box widths are proportional to samples size, and horizontal jitter added to dots for clarity.

3.4.2 Sample filtering

Whole-genome sequencing and genotyping was successful for 43 of 59 isolates. To assess for outliers and population structure, a principal components analysis was performed. Five samples were clear outliers from a dense cluster in the principal component describing the most variation (PC1), perhaps due to cross-contamination, and thus were excluded from further analyses (Supplementary figure 3-24).

Two samples were removed for having an excessively high proportion of missing SNPs (>60%, vs. less than 10% for most others), and an additional sample was excluded because it was identical at every position to another taken from the same patient one month prior (Supplementary figure 3-25).

3.4.3 GWAS

I tested 6250 SNPs for association with the activities of 22 drugs, and report 11 loci that meet genome-wide significance (Table 3-2). Two loci were significantly associated with CQ activity, and are within the genes *cg1* and *cg2*, adjacent to *pfcrt*. These two genes have frequently been associated with CQ resistance (CQR) in the literature, likely due to LD with *pfcrt* [230,231]. Two nonsynonymous SNPs in genes on chromosomes 2 and 6 were associated with QN sensitivity, and 5 SNPs with quinazoline activity on chromosomes 5, 9, 11, 13, and 14. Although the p-values for several of the QN and quinazoline hits are more significant than those for CQ, the Manhattan plot for CQ exhibits signal from a number of corroborating SNPs in proximity to *pfcrt* that do not reach genome-wide significance (Figure 3-5). This region is known to have uniquely long-range LD for *falciparum*, a remnant of the selective sweep of CQ resistance through the population [119]. Of note is that primaquine yields similar interesting signal in the *pfcrt* region, though no individual SNP meets genome-wide significance by association alone (Figure 3-6).

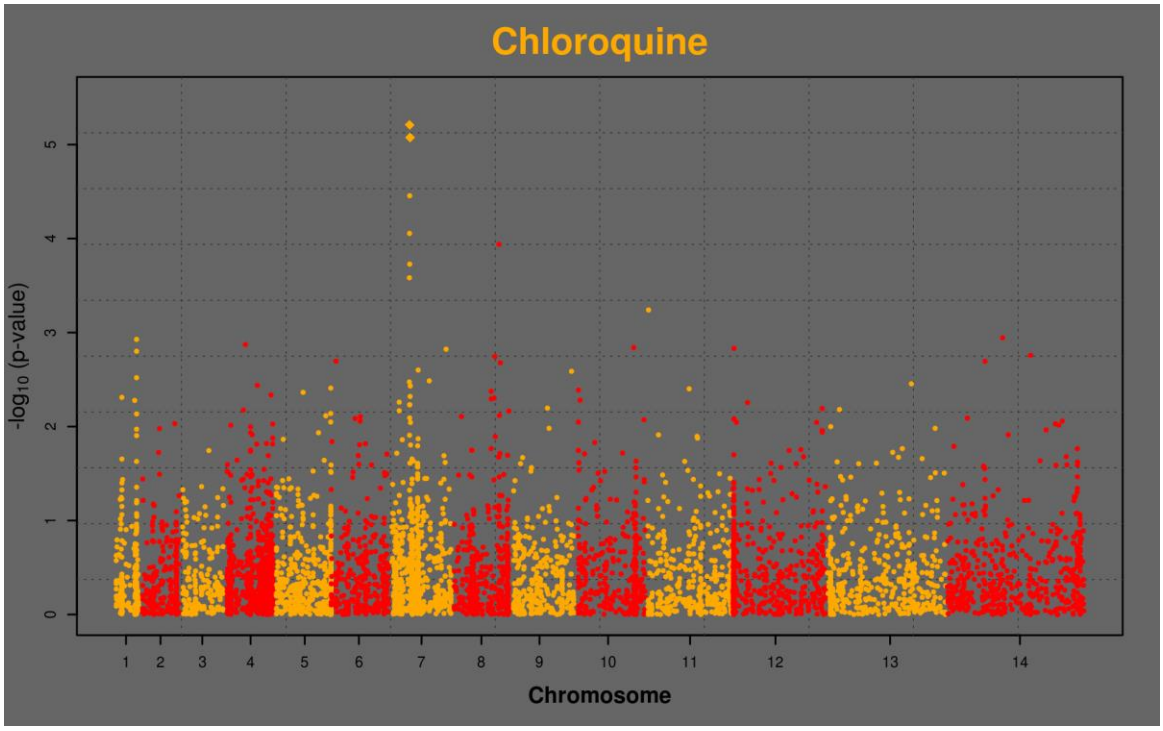


Figure 3-5. Manhattan plot of genome-wide associations with CQ activities from 35 parasite isolates. Horizontal axis is genome position, and vertical axis is $-\log_{10}$(p-value). Chromosomes alternate yellow and red, starting from chromosome 1 on the left. Yellow spire on chromosome 7 is in the region of *pfcrt*. SNPs reaching genome-wide significance are plotted as diamonds.

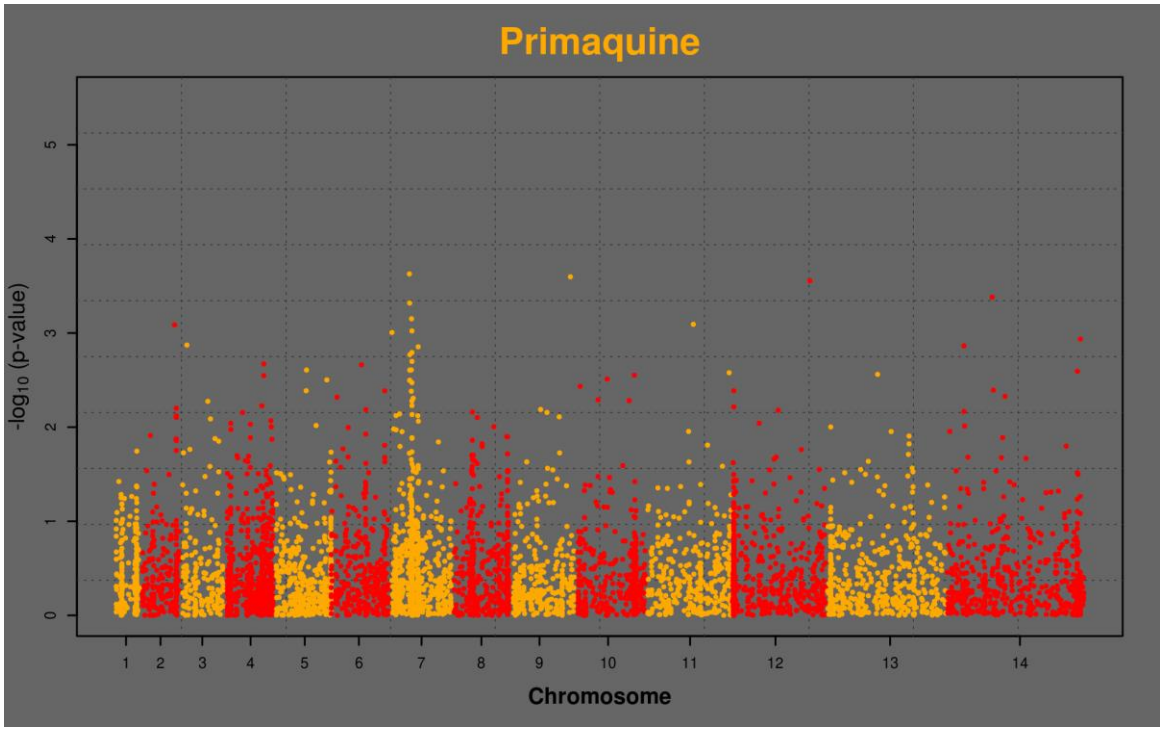


Figure 3-6. Manhattan plot of genome-wide associations with PQ activities from 35 parasite isolates. Horizontal axis is genome position, and vertical axis is $-\log_{10}$(p-value). Chromosomes

alternate yellow and red, starting from chromosome 1 on the left. Yellow spire on chromosome 7 is in the region of *pfcrt*.

Based on the quantile-quantile distribution of associations with the CQ phenotype, I used the first 3 principal components to correct a modestly deflated genome-wide inflation factor (λ=0.99), and applied this methodology to all drugs [232]. Amodiaquine activities were anticorrelated with the first two projected components (r=-0.33, r=-0.38), dampening signal from two adjacent loci in PF07_0068 that otherwise stood-out with genome-wide significance; therefore these are reported for thoroughness (Figure 3-7, Table 3-2). The ranks of these loci remain in the top 10 of AQ associated hits using either approach. All drugs were modeled as log normalized continuous variables and a subset of SNPs were tested for departures from normality (Figure 3-8).

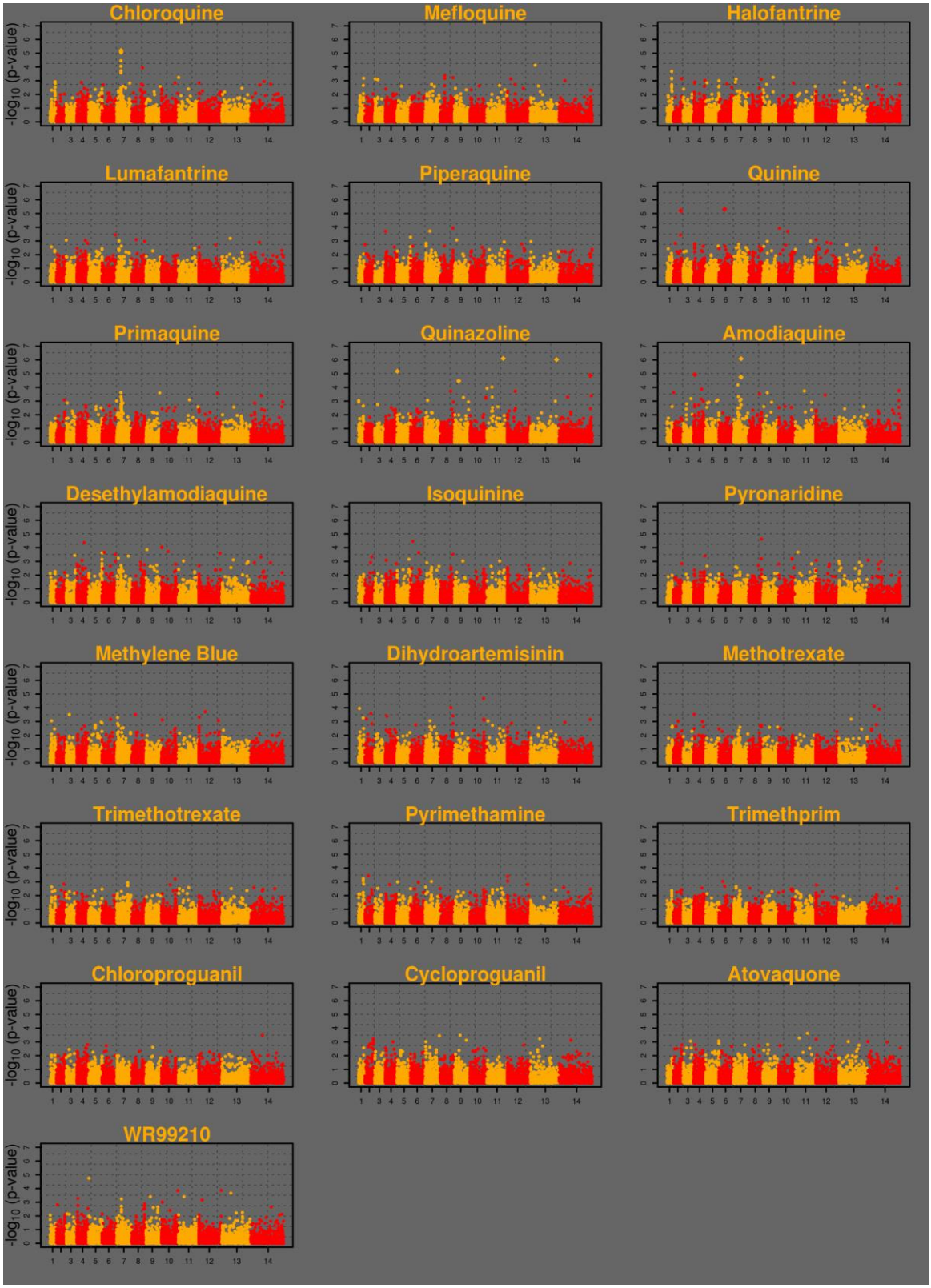


Figure 3-7. Manhattan plots for each of 22 drugs tested for association with 6250 SNPs in 35 parasite isolates. Chromosomes are numbered on the horizontal axes. Points alternate yellow and red based on chromosome. Vertical axes depicts negative $\log_{10}$(p-value) and all plots have the same y-axis scale (max of 7).

Table 3-2. Significant Kenyan GWAS SNPs. Loci achieving significance (q-value < 0.05) after correcting p-values for multiple hypothesis tests.

Drug	Gene	p-value	q-value	Gene Definition	AAC[1]
CQ	PF07_0035	6.15E-06	0.031	Cg1 protein	E161D
CQ	PF07_0037	8.4E-06	0.031	Cg2 protein	L1883V
QIN	PFB0870w	6.13E-06	0.023	conserved, unknown function	E1771K
QIN	PFF0670w	4.82E-06	0.023	transcription factor, putative	R1034C
QuiNazol	PF11_0420	7.70E-07	0.003	conserved, unknown function	R1208K
QuiNazol	PF13_0348	9.60E-07	0.003	rhoptry protein	.
QuiNazol	PF14_0726	1.34E-05	0.022	conserved, unknown function	T207P
QuiNazol	PFE0020c	6.7E-06	0.015	rifin	N226D
QuiNazol	PFI0495w	3.48E-05	0.046	conserved, unknown function	L268F
AQ[2]	PF07_0068	4.04E-06	0.012	cysteine desulfurase, putative	E339G
AQ[2]	PF07_0068	4.54E-06	0.012	cysteine desulfurase, putative	F361L

[1] Amino Acid Change. Synonymous substitutions indicated with a dot. Allele associated with drug tolerance in bold.

[2] Meets genome-wide significance without principal components in model (see Results).

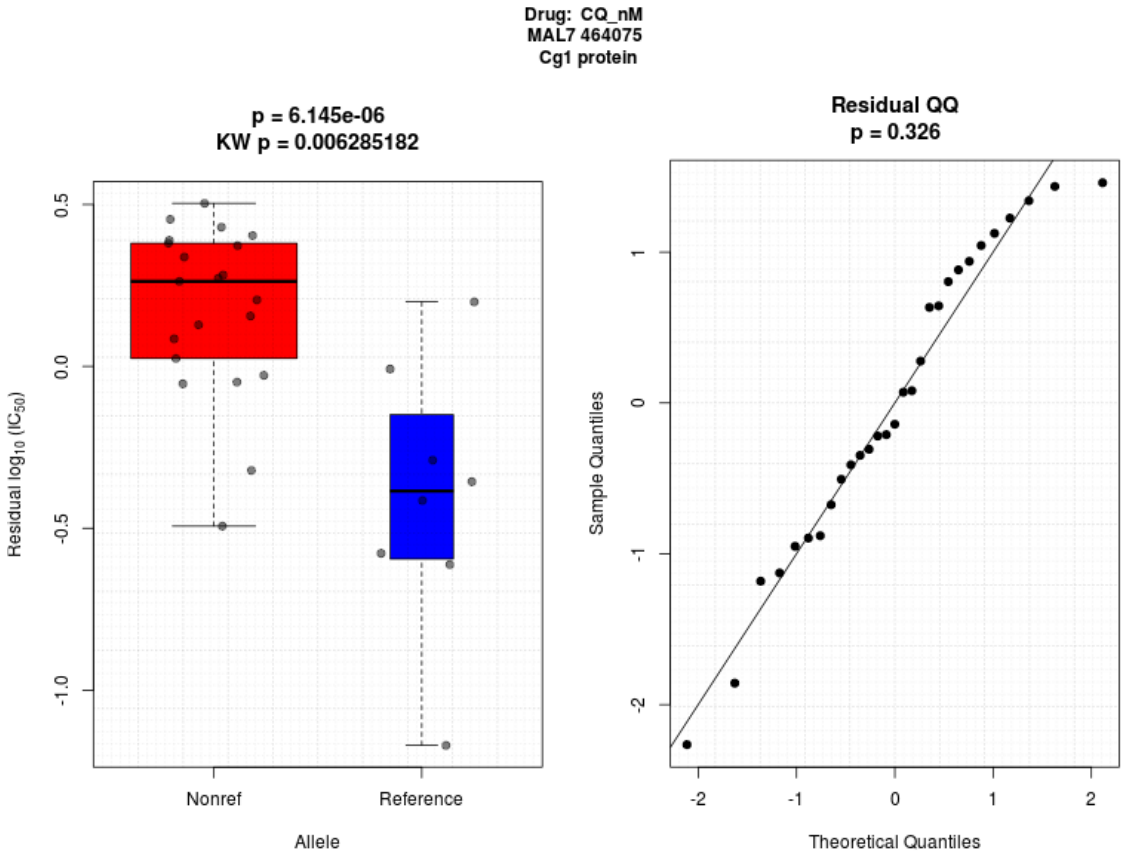


Figure 3-8. Raw data and normality test for a select SNP meeting genome wide significance. Left panel) Residual variation in $\log_{10}(IC_{50})$ values overlaid onto boxplots, stratified by reference and non-reference alleles. The residuals were generated by fitting a regression model in which genotype was held-out, allowing for better visualization of the genotype effect reflected in the GWAS p-value (p

= 6.145e-6, printed above the plot). The non-parametric Kruskal-Wallis (KW) significance is also given. **Right panel)** QQ plot comparing the residuals from the full model (i.e., genotype included) to the quantiles from a normal distribution. Departure from normality tested using the Shapiro-Wilk statistic, and the p-value given above the plot.

3.4.4 *pfcrt* haplotypes

Considering previously described *pfcrt* variants only, I observed two haplotypes representing 28 samples. For this particular analysis I excluded samples that were ambiguous due to missing genotype data or heterozygosity. Visualization of the haplotypes in this region highlights that this gene is difficult to assay with short reads, and explains why tagging SNPs of K76T yielded the strongest GWAS signal. At amino acid positions 72, 74-76, and 271, twenty isolates have residues CMNKQ, and 8 carry CIETE (Table 3-3). Non-synonymous variants at two other loci (positions 24 and 124) were also detected, which partitioned the 20 CMNKQ parasites into 3 haplotypes: 17 with DR at these positions, one with DQ, and two with amino acids YR.

Table 3-3. Amino acid haplotypes of variants in *PfCRT*. Column 'N' is the number of samples in this study represented by that haplotype

		Amino acid position					
24	**72**	**74**	**75**	**76**	**124**	**271**	**N**
D	C	I	E	T	R	E	8
D	C	M	N	K	Q	Q	1
D	C	M	N	K	R	Q	17
Y	C	M	N	K	R	Q	2

3.4.5 *pfdhfr*, *pfdhps*, and *pfmdr1*

Resistance to the antifolates pyrimethamine and sulfadoxine is attributed, respectively, to point mutations in *dhfr* and *dhps*, however no significant associations with loci in either gene were detected [233]. This was expected, as I did not test the activity of sulfadoxine, and the pyrimethamine resistance-conferring *dhfr* S108N mutation is at fixation in these samples (Figure 3-9, Table 3-4). Positions 51I and 59R in *dhfr* are nearly fixed as well, and one quadruple (I164L) mutant was detected in a mixed infection, corroborating previous reports of the emergence of this allele in Madagascar and Coastal Kenya [224,234].

Excluding mixed infections, I observed no occurrences in *dhps* of 437A-540E double mutants, but every parasite carried one or the other (Figure 3-10).

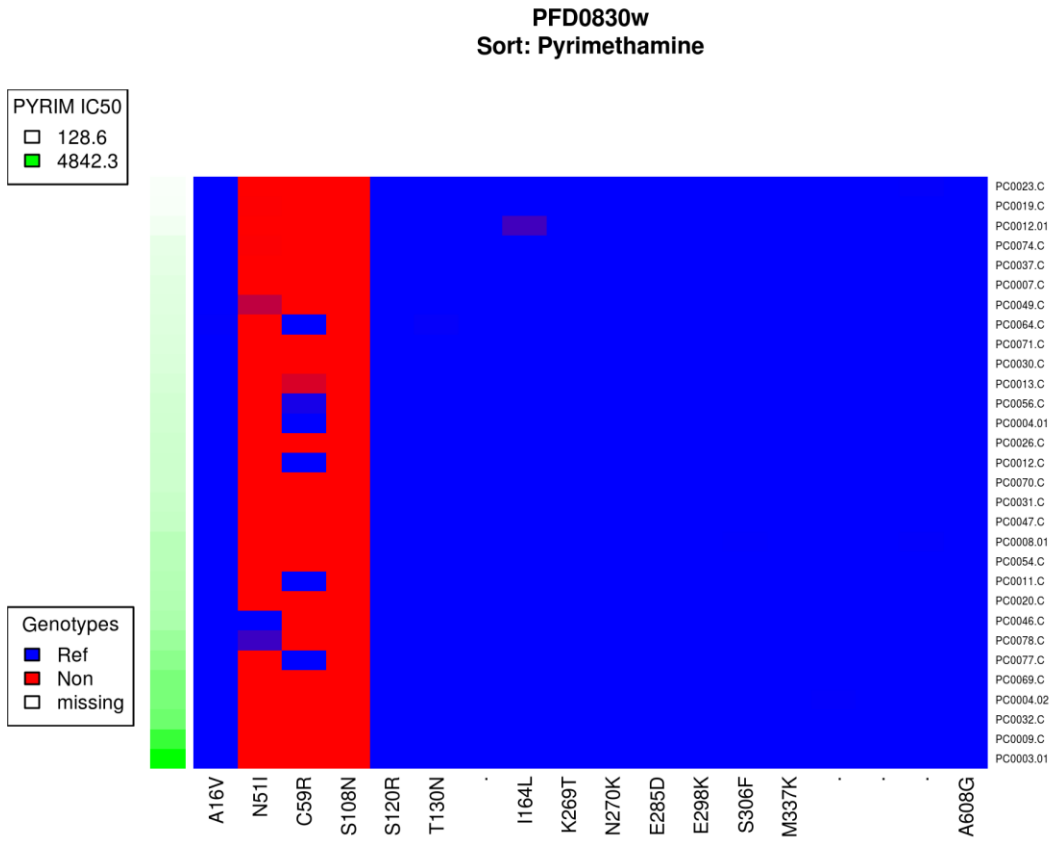


Figure 3-9. Haplotype plot for *pfdhfr* (PFD0830w). Each row represents a sample, and each column a potential SNP. Samples are sorted by pyrimethamine IC_{50}, indicated by the green bar on the far left. Blue cells indicate positions matching the reference genome, and red the alternate allele. Mixed infections are represented by blending of red and blue, proportional to the within-sample allele frequencies. White cells indicate missing data. Nonsynonymous SNPs are labeled with the amino acid substitution along the bottom, and with a dot if synonymous.

Table 3-4. Amino acid haplotypes of hallmark variants in *pfdhps* and *pfdhfr*. Column 'N' is the number of samples in this study represented by that haplotype.

DHPS			DHFR				
436	437	540	51	59	108	164	N
A	A	K	I	R	N	I	1
S	A	K	I	C	N	I	2
S	A	K	I	R	N	I	6
S	A	K	I	R	N	L	1
S	G	E	I	C	N	I	5
S	G	E	I	R	N	I	18
S	G	E	N	R	N	I	2

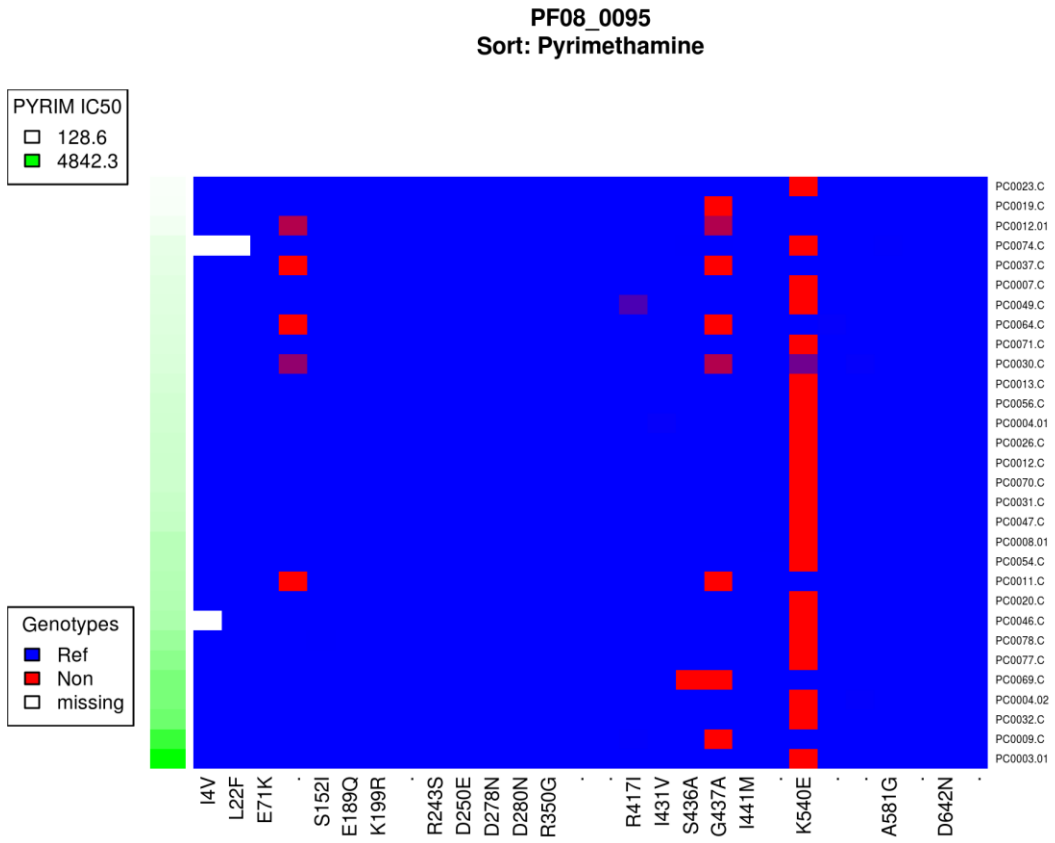


Figure 3-10. Haplotype plot for *pfdhps* (PF08_0095). Each row represents a sample, and each column a potential SNP. Samples are sorted by pyrimethamine IC_{50}, indicated by the green bar on the far left. Blue cells indicate positions matching the reference genome, and red the alternate allele. Mixed infections are represented by blending of red and blue, proportional to the within-sample allele frequencies. White cells indicate missing data. Nonsynonymous SNPs are labeled with the amino acid substitution along the bottom, and with a dot if synonymous.

Similarly, no signals of association in *pfmdr1* were detected. A previous study found an association of *pfmdr1* position 86 mutants with lumafantrine susceptibility in Coastal Kenya, however this SNP failed to meet quality thresholds, as did position 1246 [225]. Further, I observed little variation in this gene in SNPs that might otherwise have tagged position 86, or other commonly implicated loci (Figure 3-11). A larger sample size would be necessary to detect very low frequency variants in this gene.

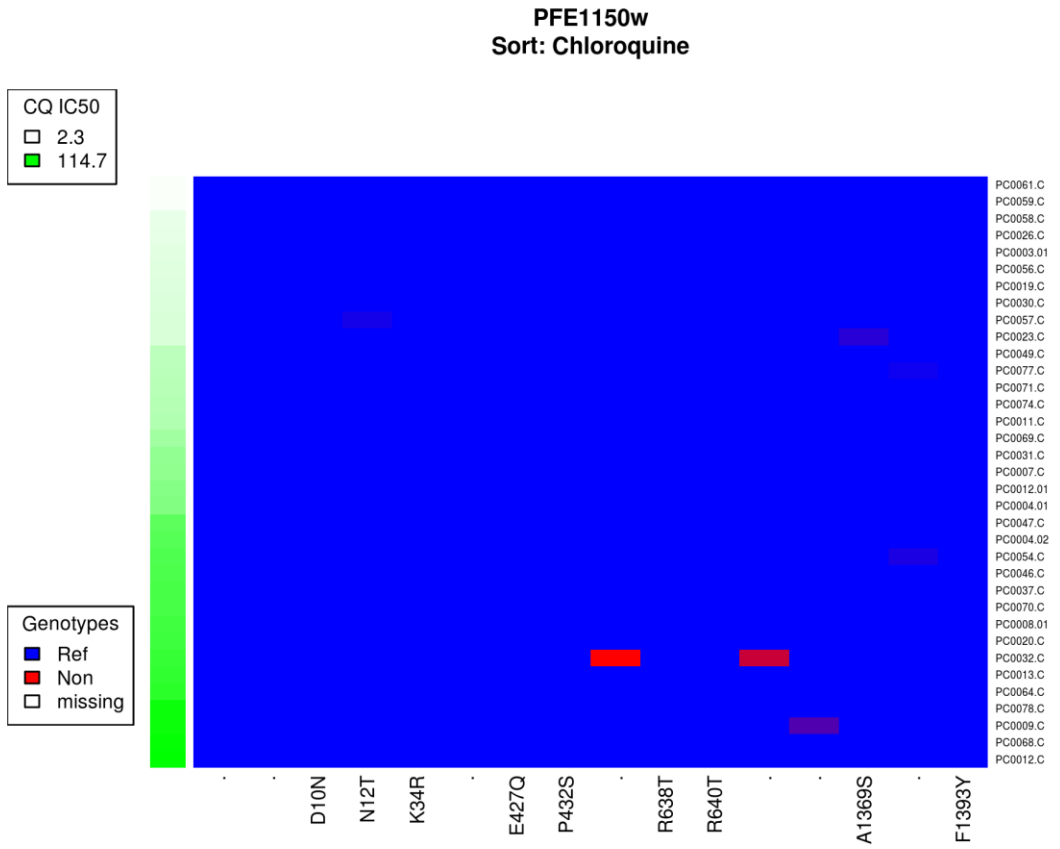


Figure 3-11. Haplotype plot for *pfmdr1* (PFE1150w). Each row represents a sample, and each column a potential SNP. Samples are sorted by chloroquine IC_{50}, indicated by the green bar on the far left. Blue cells indicate positions matching the reference genome, and red the alternate allele. Mixed infections are represented by blending of red and blue, proportional to the within-sample allele frequencies. White cells indicate missing data. Nonsynonymous SNPs are labeled with the amino acid substitution along the bottom, and with a dot if synonymous.

3.4.6 pfnhe

Previous reports have associated structural variants in the sodium/hydrogen exchanger gene (*pfnhe)* with quinine tolerance *in vitro* [210,235]. These structural variants in microsatellite ms4760 of *pfnhe* may be important markers for surveillance, and more work is needed to describe the natural variation in this gene [236]. While the analysis of structural variation is beyond the scope of this particular output, 15 nonsynonymous SNPs in *pfnhe* are reported (Table 3-5). N894K has been previously described and appears in 4 isolates. The most common variant was carried in 6 isolates (D209Y).

Table 3-5. Variants detected in *pfnhe*. Column 'N' is the number of samples in this study carrying that allele.

CHROM	POS	REF	ALT	AAC	N
MAL13	175711	T	C	K94R	1
MAL13	175367	C	A	D209Y	6
MAL13	175360	A	G	F211S	1
MAL13	175358	A	C	Y212D	1
MAL13	175357	T	G	Y212S	2
MAL13	175349	C	T	D215N	4
MAL13	175346	C	A	G216C	4
MAL13	173387	G	A	H869Y	2
MAL13	173347	T	C	K882R	1
MAL13	173310	A	T	N894K	4
MAL13	172529	C	T	D1155N	1
MAL13	171974	C	A	A1340S	1
MAL13	171818	T	C	N1392D	1
MAL13	171580	T	C	K1471R	2
MAL13	171448	C	A	R1515I	1

3.4.7 Drug correlations

Drugs with correlated activities may indicate related mechanisms of action, and perhaps more importantly, those with negative correlations might reveal synergistic partners for co-deployment or rotation strategies [227]. Several drugs, including lumafantrine, have been reported to select for parasites with inverse susceptibilities to CQ, and I find evidence of this here as well (Figure 3-12) [237]. CQ activity is significantly correlated with desethylamodiaquine (DEAQ, r=0.49, p=0.006) and anticorrelated with PQ (r=-0.48, p=0.008). Related to this, *pfcrt* haplotypes associated with CQ resistance sort inversely to PQ activity, and yield association signal in the same region (see discussion). Interestingly, the *dhfr*-targeting drug, WR99210, is negatively correlated with many of the other antifolates. Exceptions to this include trimethoprim, quinazoline, and pyrimethamine, which themselves form a tightly related cluster. Piperaquine activity is more highly correlated with the antifolates than with the aminoquinolines, with the exception of pyronaradine. Piperaquine and other bisquinolines have demonstrated effectiveness against CQ resistant parasites *in vitro*, and another study in Coastal Kenya found no association of *pfcrt* with activity for this drug [225,238].

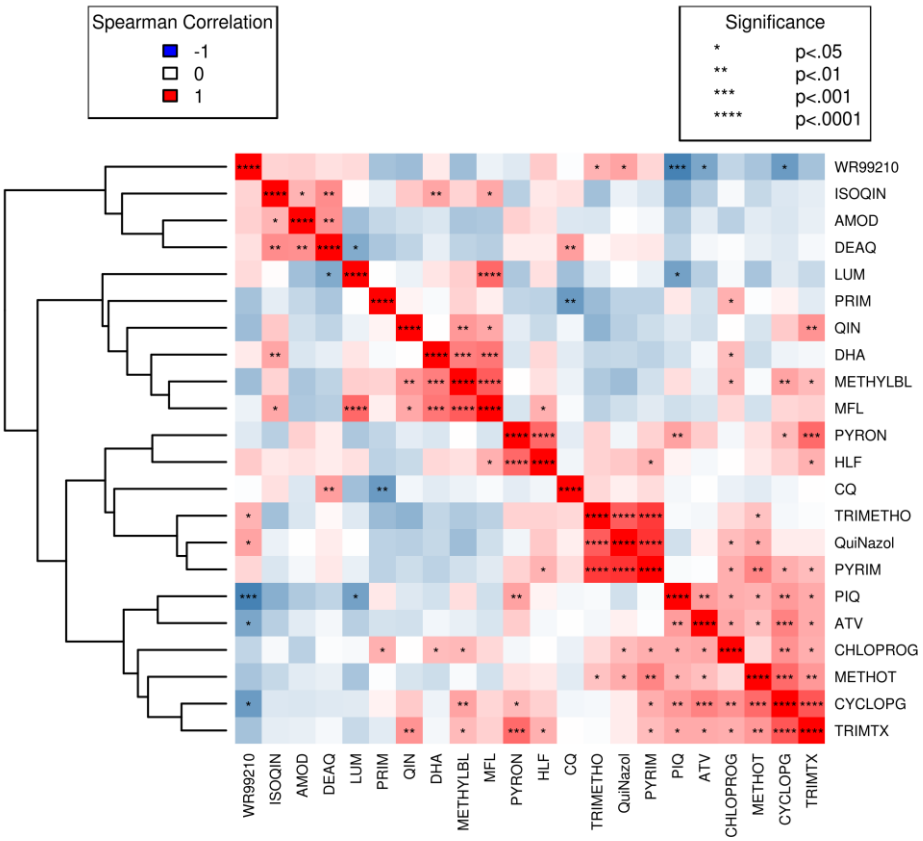


Figure 3-12. Cluster plot of drug correlations. Red to blue indicates the degree of positive to negative correlation. Significance levels of spearman rank tests are indicated with stars in each box (see legend).

3.5 Discussion

The expected signals of association with CQ activity were detected in the *pfcrt* region with these 35 samples, providing a positive control for the lower than optimal sample size for this study. At the time this work was published, most other parasite GWAS experiments were performed on genotypes generated using microarrays, and ranged in sample size from 57-331 parasites [193,226,239]. One other sequence-based study had been performed, and they enhanced their samples size of 45 parasites by including information about signatures of selection [228]. A more recent publication used more than 1600 samples across 15 Southeast Asian locations to confirm the Kelch gene's involvement in artemisinin resistance, and achieved significance levels as low as $4x10^{-26}$ [240].

CQ was a highly effective and cheap drug in Kenya for decades before the emergence of resistance in the mid-1980s [241]. National policy shifted from CQ to the antifolate, SP, in 1998, to which resistance also emerged in a short time [242]. Resistance to CQ remains

above 60% in Kenya, and prevalence of the important CQR conferring K76T mutation was measured at 63% in the Coastal region in 2006 [243,244]. A hallmark of selective sweeps, like that of CQR in Kenya, is uncharacteristically long haplotypes; i.e., segregating stretches of DNA carrying the resistance-conferring allele that have yet to be broken down by recombination. One of the significant CQ-associated SNPs detected within *pfcrt*, Q271E, is in complete LD with K76T for these samples—consistent with a report 4000 miles away in Senegal [245]. This level of LD might prove useful for imputation in similar populations of the important K76T variant, which is in a region that is relatively difficult to access with short-read sequencing. Indeed, outside of Papua New Guinea and South America, there is 99.8% agreement (1041/1043) between these two positions in homozygous MalariaGEN samples (Supplementary figure 3-26). Thirty-four percent of the Kenyan isolates used in this study carry the K76T substitution (46% if missing calls are inferred by Q271E).

I also report potentially novel associations for quinine, quinazoline, and amodiaquine. AQ tolerance is commonly associated with *pfcrt*, however this drug remains effective against some CQ resistant parasites—i.e., *pfcrt* alone does not encapsulate resistance [246,247]. The CIET haplotype observed in this study is not sufficient in isolation for conferring AQ resistance, and I do not detect significant signal for this drug in *pfcrt* [248,249,250]. I report two SNPs in a putative cysteine desulfurase gene (PF07_0068) that are significantly associated with AQ activity (Table 3-2). This gene is more than 300Kb from *pfcrt*, thus not likely tagging the CQR haplotype. 4-aminoquinolinines like CQ and AQ are thought to act by accumulating in the parasite digestive vacuole (DV) and preventing the crystallization of heme dimers into hemozoin [180]. The elevated concentration of toxic heme within the DV leads to increased efflux into the cytosol in a dose-dependent manner, resulting in an oxidative challenge to the parasite and membrane damage [180]. Free heme should be detoxified by glutathione in the cytosol, but both CQ and AQ directly compete with this activity [181]. One might speculate whether cysteine desulfurase affects this interaction, or is more broadly involved in parasite pathways related to alleviating increased oxidative stress, for example the thioredoxin or glutathione redox systems. In plants, cysteine desulfurase has been postulated to modify the catalytic properties of glutathione by changing cysteine content [251].

A decade after CQ withdrawal in Malawi, the proportion of circulating CQR parasites in the population has receded to nearly undetectable levels [252]. The velocity of this particular shift appears to be somewhat unique, nonetheless CQR in Kenya has also been on the decline since CQ withdrawal in 1999 [243]. The haplotypes and patterns of LD support that

this event in Malawi was due to an expansion of the existing CQ susceptible (CQS) parasite population, rather than a sweep or reversion, and our data are consistent with this model as well [253,254]. All parasites with the resistant *pfcrt*-76T allele are represented by a single haplotype across 7 positions, in contrast to the susceptible forms which are comprised of several haplotypes. This is consistent with the hypothesis that, relative to the homogenous CQR parasites originating from a selective sweep, a diverse pool of susceptible parasites has been maintained and serves as a reservoir of expansion in the absence of drug pressure. This stands-out visually when a second haplotype in *cg1*, found 2kb downstream, is juxtaposed with *pfcrt* (Figure 3-13). Although our inferences are limited due to small sample size, it would appear that CQS diversity was not completely extinguished under decades of drug pressure, indicated by the higher relative polymorphism in the parasites that are both most susceptible to CQ, and lack the 76T allele.

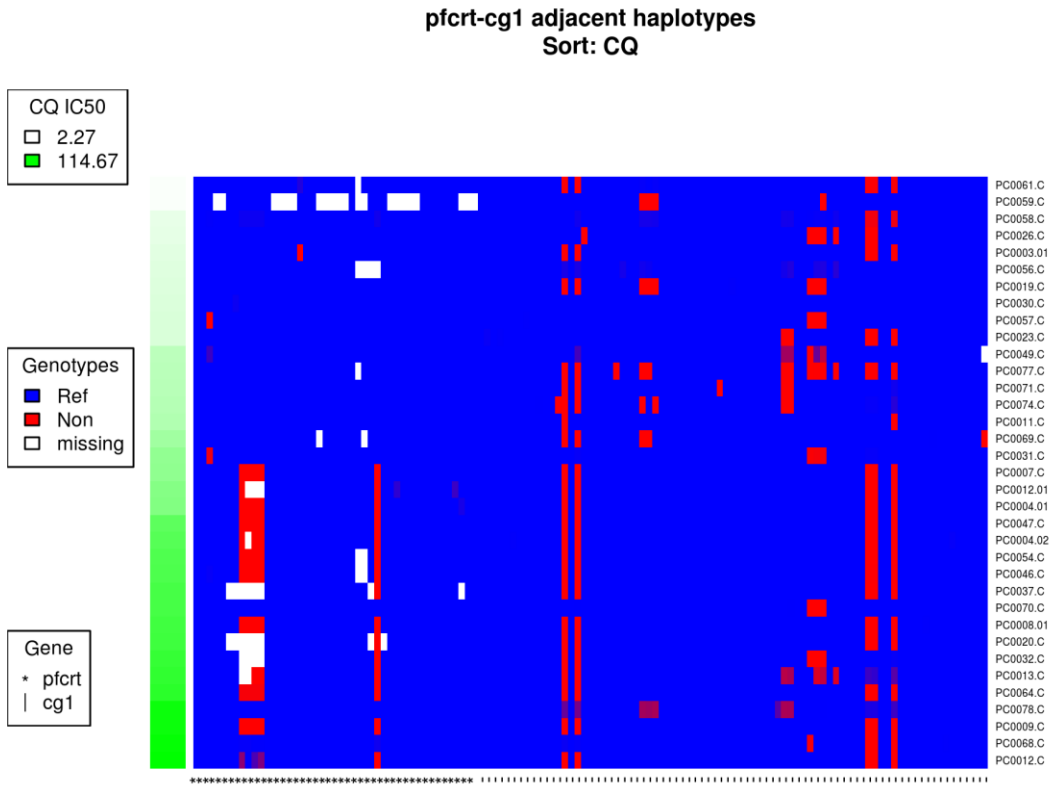


Figure 3-13. Haplotype plot for *pfcrt* (MAL7P1.27) and *cg1* (PF07_0035) combined. Each row represents a sample, and each column a potential SNP. Samples are sorted by chloroquine IC_{50}, indicated by the green bar on the far left. Blue cells indicate positions matching the reference genome, and red the alternate allele. Mixed infections are represented by blending of red and blue, proportional to the within-sample allele frequencies. White cells indicate missing data. SNPs in *pfcrt* are indicated with a "*" along the bottom, and those in *cg1* with the "|" symbol. More diversity is apparent in the top rows—i.e., those parasites that are most susceptible to CQ, and lack the 76T allele.

Like Verapamil (VP), PQ has been shown to reverse CQR in a dose-dependent manner, and there is evidence supporting direct inhibition of *pfcrt* as the underlying mechanism [255,256,257]. It is therefore intriguing to observe negatively correlated PQ and CQ activities, and correspondingly inverse *pfcrt* haplotype plots when sorted by drug activities (Figure 3-14, Supplementary table 3-7). Both PQ and CQ phenotypes yield convincing GWAS signal in the *pfcrt* region as well. Of the top 17 SNPs (by p-value) for these two drugs, 3 SNPs overlap identically in the CRT region, and another half-dozen are in the same vicinity, all of which have consistently inverse trends. It is tempting to speculate that in addition to PQ interacting with CRT mutants to reverse resistance directly, CQ might, separately, select for parasites that are more susceptible to PQ. If confirmed, the relevance of this would depend on whether the biochemical target of the high concentrations required for shizontocidal activity here is the same mechanism conventionally affected by lower concentrations in other stages. Primaquine's precise mechanism of action is unknown [258]. It is not possible to infer whether PQ, in reverse, would select for CQ sensitive parasites, as it is unlikely that our isolates were exposed to natural PQ pressure. Primaquine is primarily used for clearing *P. vivax* and *P. ovale* hypnozoites, and although it also has activity against gametocytes, this community benefit is counter-balanced by the risk of hemolysis to G6PD-deficient individuals [259]. Chloroquine is the first-line antimalarial given for *P. vivax* infections, and to prevent relapse it is recommended to be given in combination with primaquine [260]. Evidence of selective interactions as reported here would be salient in such drug policy decisions. A similar study in Senegal reported a highly significant signal of selection for PQ sensitivity in the *pfcrt* region, and those authors attribute this to PQ anticorrelation with CQ [228].

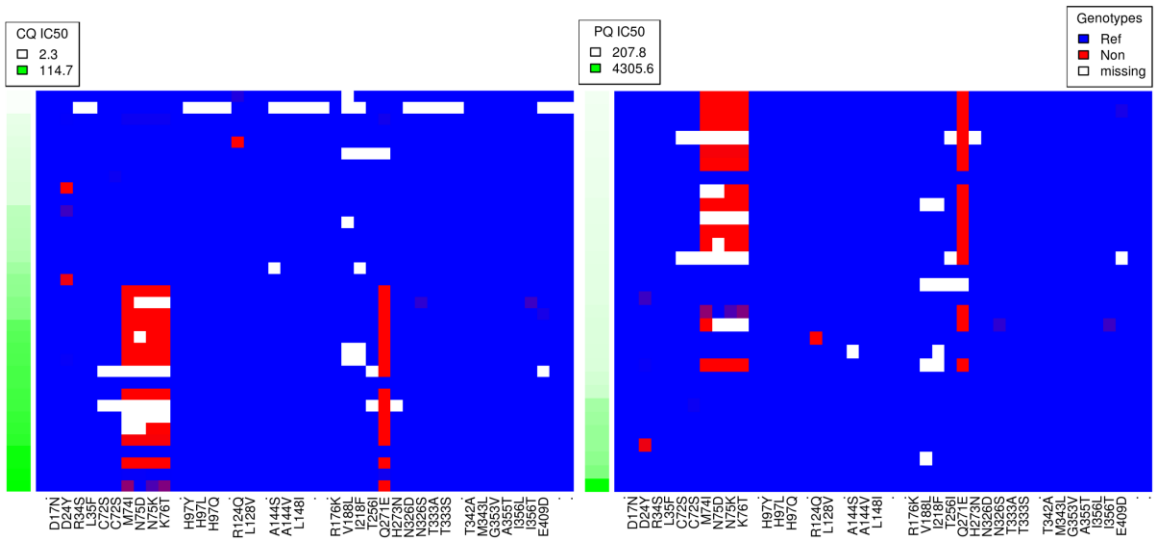


Figure 3-14. Haplotype plot for *pfcrt* (MAL7P1.27), sorted by CQ and PQ activities. Left panel is sorted top to bottom by increasing CQ IC_{50}, and the right panel is sorted by PQ IC_{50}. Each row represents a sample, and each column a potential SNP. Drug activity is shown as increasing green intensity in the far left column of each plot. Blue cells indicate positions matching the reference genome, and red the alternate allele. Mixed infections are represented by blending of red and blue, proportional to the within-sample allele frequencies. White cells indicate missing data. Nonsynonymous SNPs are labeled with the amino acid substitution along the bottom, and with a dot if synonymous.

With regard to selection, such relationships are not unprecedented—e.g., inverse pressures on *pfcrt* between CQ and LUM have been described in Tanzania and Kenya previously [225,237]. Lumafantrine is the partner drug in the ACT, Coartem, which has been the first-line treatment for uncomplicated malaria in Kenya since 2006. Although not as strong as with PQ, I similarly observe a modest "flip" in the ordering of haplotypes when CQ is compared to LUM (Figure 3-15). With only 35 parasites and a sample limited in time and geography, replicate studies and experiments are needed to confirm these observations.

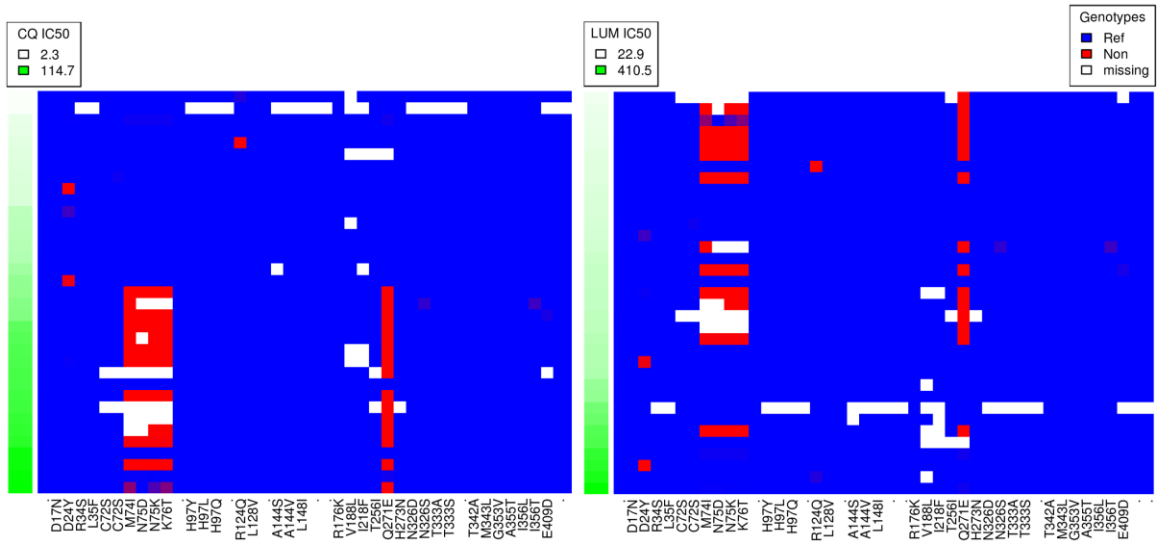


Figure 3-15. Haplotype plot for *pfcrt* (MAL7P1.27), sorted by CQ and LUM activities. Left panel is sorted top to bottom by CQ IC_{50}, and the right panel is sorted by LUM IC_{50}. Each row represents a sample, and each column a potential SNP. Drug activity is shown as increasing green intensity in the far left column of each plot. Blue cells indicate positions matching the reference genome, and red the alternate allele. Mixed infections are represented by blending of red and blue, proportional to the within-sample allele frequencies. White cells indicate missing data. Nonsynonymous SNPs are labeled with the amino acid substitution along the bottom, and with a dot if synonymous.

If adequately powered, null results from GWA studies of drug sensitivities are informative about which therapies might be most effectively deployed in the region of inference. Consistent with overlapping studies in the Kilifi region, no association of pyronaradine, methylene blue, piperaquine, or DHA activities with *pfcrt, pfmdr1,* or any other loci were detected [225,261]. The combination therapy of piperaquine and DHA (Artekin) might therefore be currently effective in this population, even with some degree of CQR prevalence. With the limited sample size here, interpretations of null associations must be heavily tempered; nonetheless, this study contributes precedent for planning future genome-wide association and surveillance studies.

In summary, the expected signals of association with chloroquine are confirmed, and I report novel loci related to the activities of AQ, QN, and quinazoline. The high resolution provided by sequence-based genotypes also revealed new polymorphisms in current candidates, and provided for haplotype visualizations that highlight relationships otherwise easily overlooked. Notably, these relationships are consistent with other reports, and if validated would be important for ethics and policy decisions involving PQ. Coastal Kenya has experienced a marked decline in transmission intensity over the past decade, and it is important to monitor the resulting dynamic immunoepidemiology in parallel with the

changing parasite population [71]. These developments, and the repeated emergence of drug resistance in Kenya, underscore the urgency for well-powered, sequence-based, genome-wide approaches to genetic association and surveillance of *Plasmodium falciparum*.

3.6 Limitations

3.6.1 Missingness in *pfcrt*

The *pfcrt* haplotype plots depict a notable pattern of missingness in the K76T region of exon 2, and this artefact has a systematic bias. Parasites that are more resistant to CQ, and thus have the non-reference genotype, are far more likely to display missingness (Figure 3-16). This can be mitigated to some extent by decreasing the read depth required for a positive genotype call. The left panel of Figure 3-15 is similar to Figure 3-16, except that the coverage filter was changed to 2 from the default minimum of 5. While this solution may be useful for qualitative visualizations, a more liberal filter setting will result in a higher error rate, and thus another approach to genotyping these complex regions is needed. Using the default coverage threshold, nearly a third of samples overall contain some degree of missingness in the K76T region, but within CQR samples the problem is far more acute. Using information from position 271 to infer the non-reference K76T haplotype, it appears that more than half of resistant parasites exhibit missingness. As mentioned previously, this explains why tagging SNPs yielded the GWAS signal for CQ sensitivity, rather than the known causal loci in *pfcrt*.

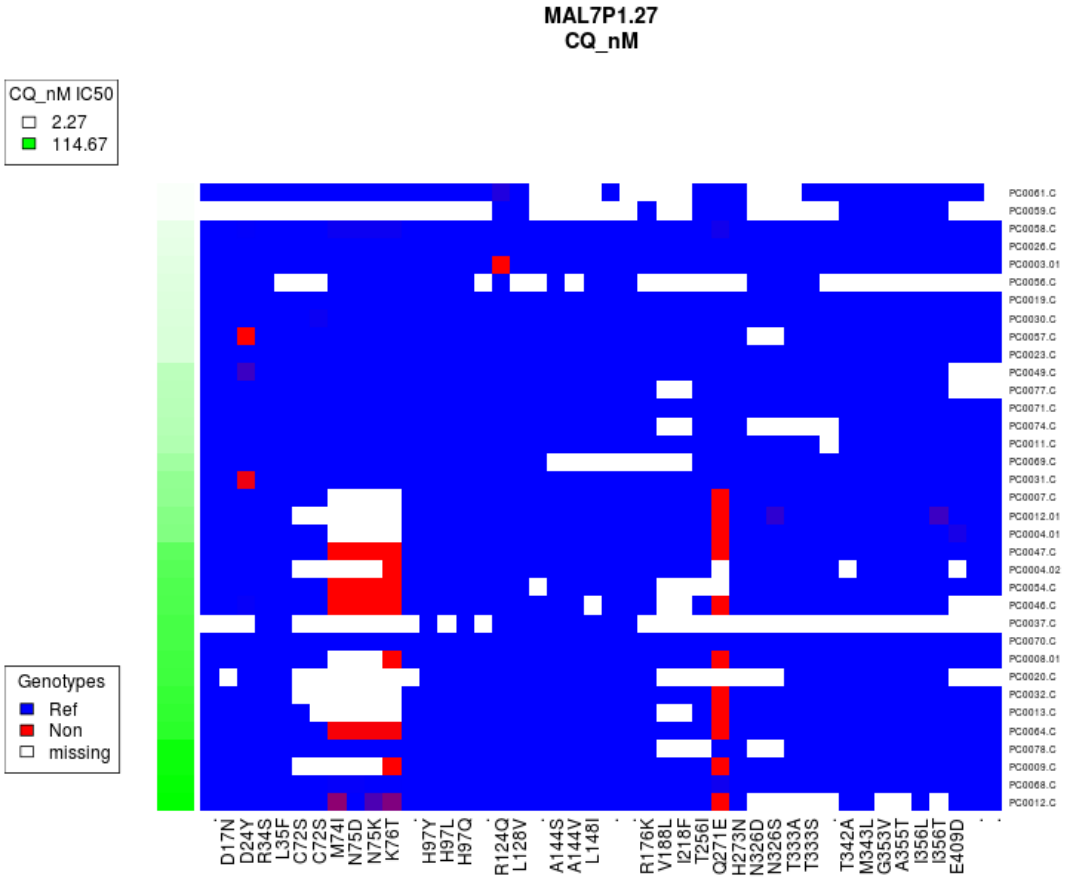


Figure 3-16. Haplotype plot of *pfcrt* with filters set to GWAS levels. Unlike other haplotype plots in which filters are set to minimize missingness, this plot reveals the amount of missingness in the association tests. Each row represents a sample, and each column a potential SNP. Samples are sorted by chloroquine IC_{50}, indicated by the green bar on the far left. Blue cells indicate positions matching the reference genome, and red the alternate allele. Mixed infections are represented by blending of red and blue, proportional to the within-sample allele frequencies. White cells indicate missing data. Nonsynonymous SNPs are labeled with the amino acid substitution along the bottom, and with a dot if synonymous.

3.6.2 Etiology

Several factors contribute to the difficulties with genotyping *pfcrt*. As shown in the coverage plot in Figure 3-17, many samples drop to zero coverage in intronic regions of the gene. The introns are littered with low complexity sequence. Homopolymer stretches (e.g., a long stretch of all the same nucleotide) and dinucleotide repeats are prone to errors in Illumina sequencing. These regions are also highly variable, likely due to polymerase slippage during replication and to the lack of selective constraint in introns. These factors, combined with the pockets of dense polymorphism in flanking exons that is associated with drug resistance, create a scenario in which sample reads contain too many mismatches to align properly to the reference gene. The BWA software as used by MalariaGEN will fail to map a read if there are more than 4 mismatches. Further, sequence reads that align to more than

one place in the genome are assigned a map quality of 0 and are discarded. As depicted in the bottom track of Figure 3-17, a large proportion of nearly every *pfcrt* intron contains non-unique sequence.

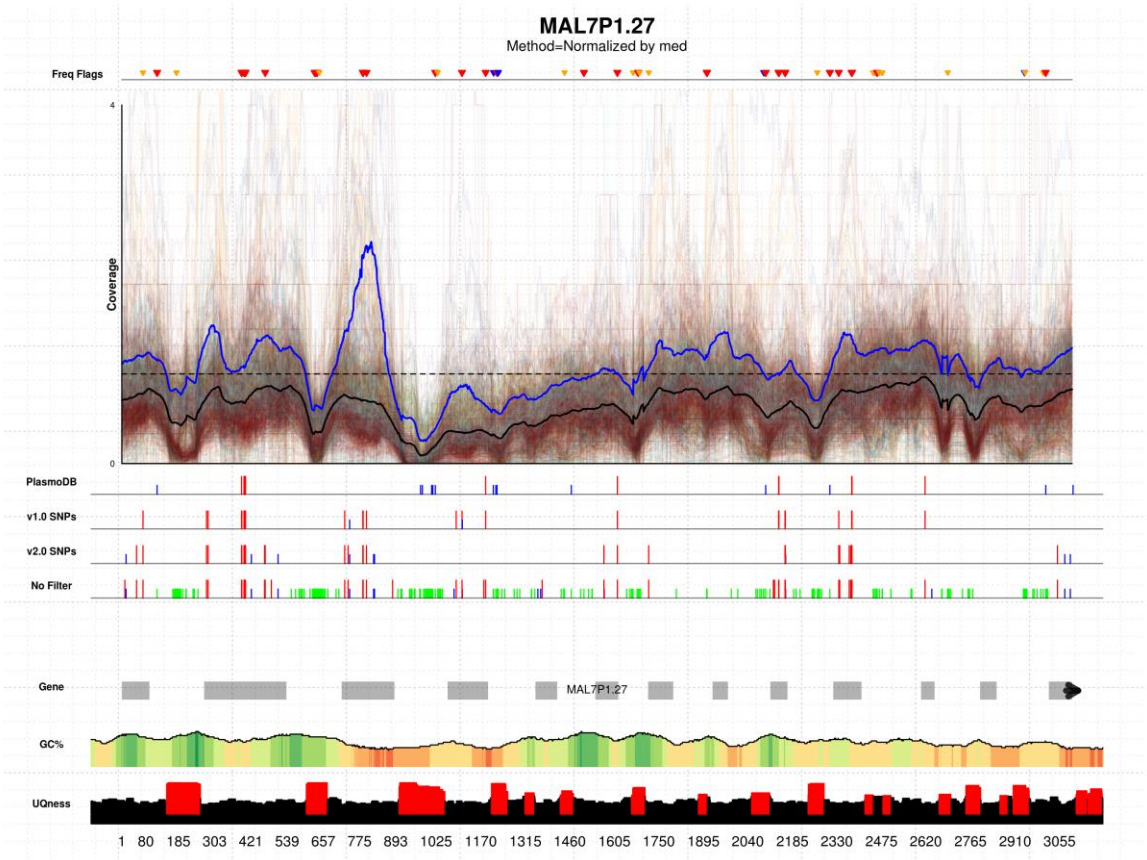


Figure 3-17. Candidate gene plot for *pfcrt* (Mal7P1.27). Complete details describing all tracks and data sources can be found in methods section 2.2.1. The coverage track contains 1591 lines representing the normalized sequence read coverage for each sample. Lines representing samples from the same country are the same color. The normalizing constant for each sample is the median (indicated in the title) coverage of SNPs in all genes for the given sample. The black line is the mean across all samples (median looks nearly identical). The "Gene" track indicates intron-exon boundaries, which correlate with uniqueness and coverage.

3.6.3 Significance and conclusions

At the very least, missingness will diminish statistical power, and in some applications it could lead to false conclusions. The primary impact of systematic missingness in *pfcrt* on these GWAS results is an increased type II (i.e., false negative) error rate. This is demonstrated in the chloroquine association study, where we expect the primary signal to emanate from the K76T region of exon 2. Yet, as seen in other studies, the strongest associations derive from tagging SNPs in adjacent genes. The biased nature of this missingness toward parasites with a particular phenotype could lead to incredibly

misleading results in other technologies. Differential expression in RNA-seq studies is determined by comparing normalized coverage depths between groups [262]. In this case, missingness from one group would be interpreted as lower transcript abundance, and could lead to type I error (false positives). This pathology would also exist in microarray and qPCR studies, where primers and probes could hybridize with different efficiencies, and even differential proteomic analyses that relied on a reference database for peptide identification. In the next chapter and in Section II I will present further examples of genes that are difficult to access with short sequence reads. One of the common denominators among these genes is divergence from the 3D7 reference.

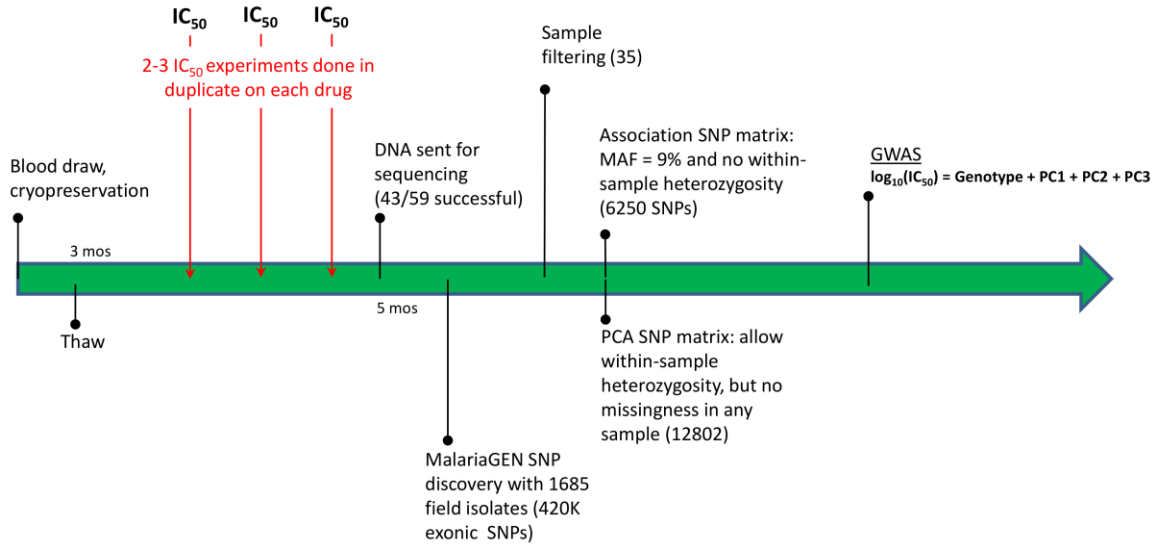
Alternative approaches are necessary to access the variation in certain genes. A possible solution to the *pfcrt* missingness problem would be to align the reads from each sample to a universal reference. In practice this might be achieved by aligning to multiple versions of the gene and then genotyping SNPs using the reference that provides the best coverage. An obvious limitation of this approach is that the "universal reference" is only as good as the existing database of genes. Another approach would be to assemble the gene *de novo*, and then to genotype directly. The advantage of this method is that it is reference-free, however it does have some drawbacks. Mapping sequence reads to a genome is highly efficient, and not using the information contained in the reference computationally expensive.

In Section II I describe a *de novo* assembly pipeline that also utilizes available reference sequence, and I apply it to *pfcrt*. As I describe in that section, this approach holds much promise for *pfcrt*, however it has even more value for some other divergent genes.

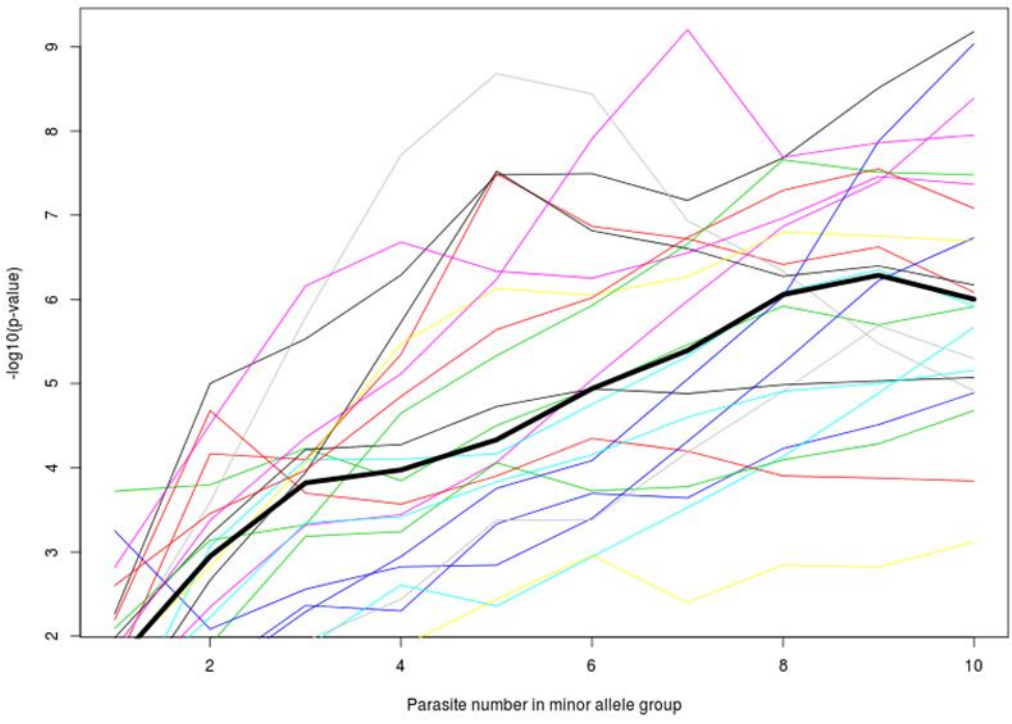
3.7 Acknowledgments and contributions

The samples used for this project were derived from the drug study cited in the methods. Members of Alexis Nzila's group in Kilifi, Kenya (particularly John Okombo) performed all culturing and chemosensitivity assays and organized parasite sequencing through a collaboration with the Kwiatkowski group in Oxford. I performed all analyses for this project and wrote the manuscript.

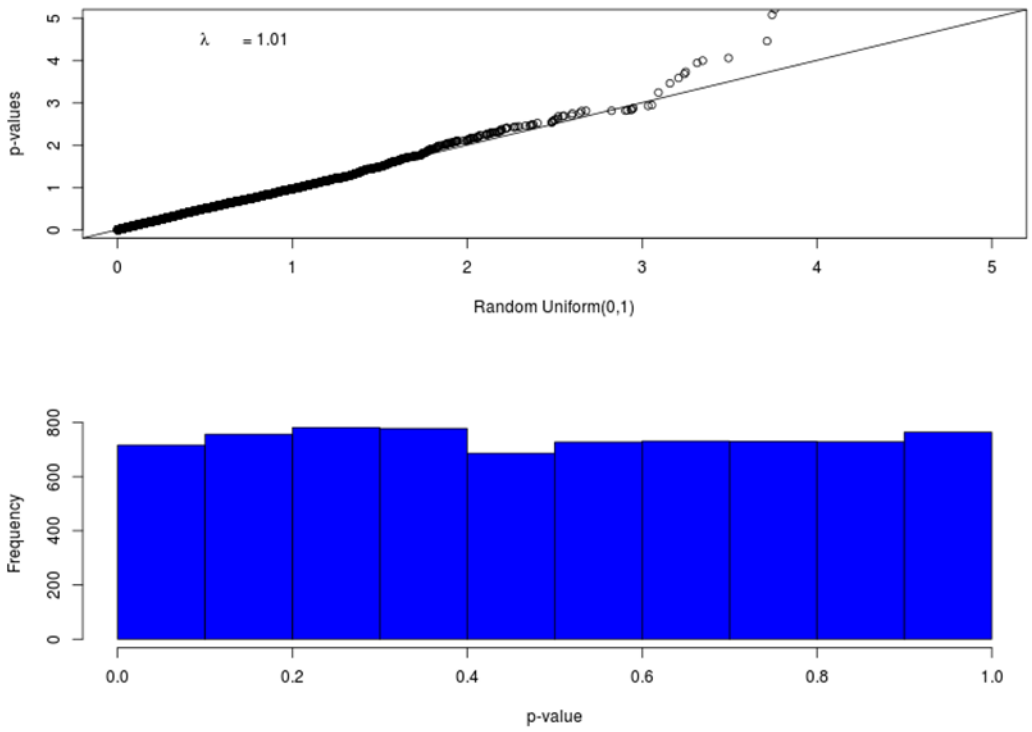
3.8 Supplementary material



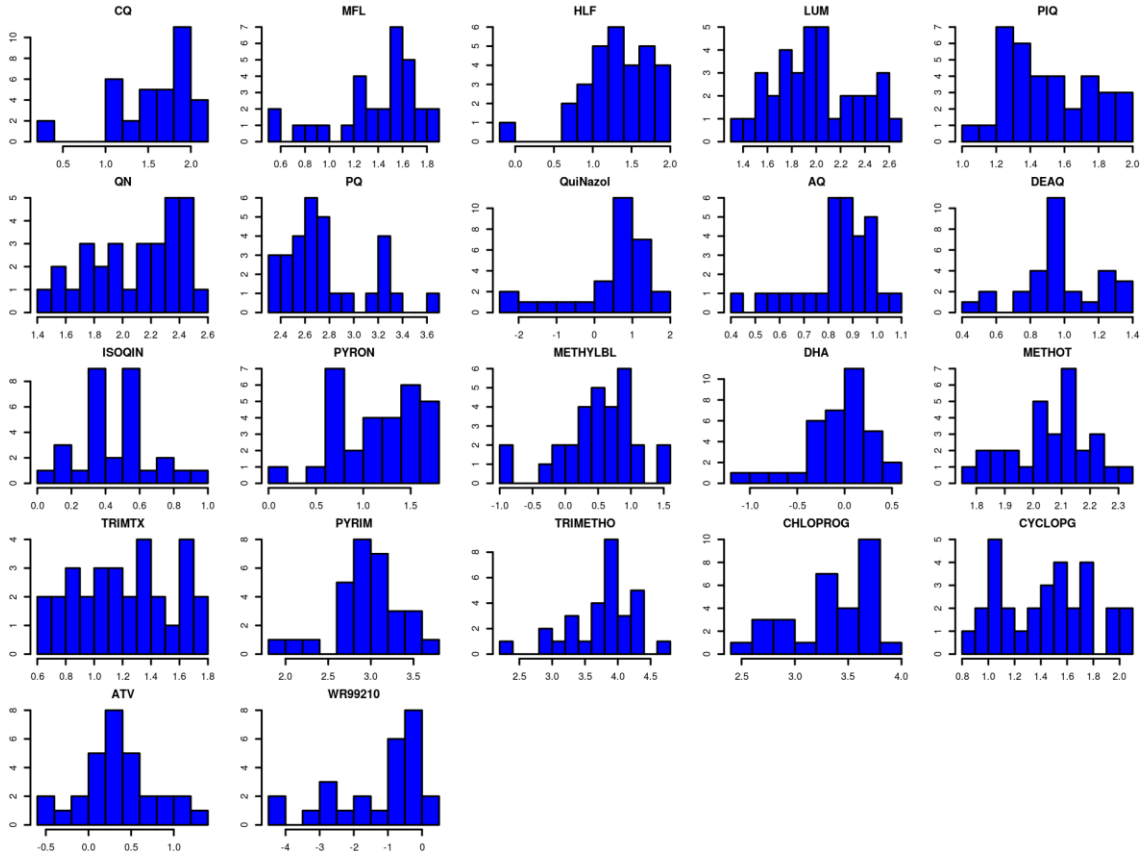
Supplementary figure 3-18. Overview schematic of the experimental and analytical workflow.



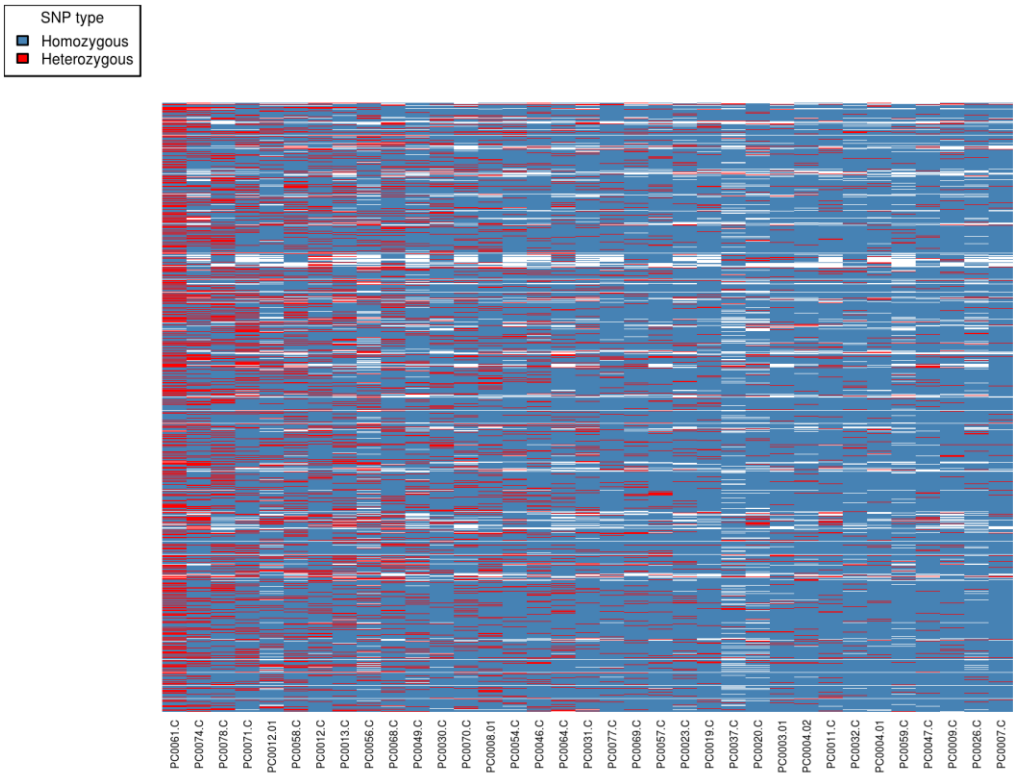
Supplementary figure 3-19. Highest possible significance level of each drug at various MAFs. P-values were calculated using a generalized linear model with dichotomous outcome, as described in the analysis section. For each drug, at each MAF, the highest and lowest IC_{50} values were artificially placed into opposite categories. Each of the 22 drugs is represented by a different colored line, and the thick black line is the median. The values for specific drugs were not of interest here, so drug labels are omitted for tidiness.



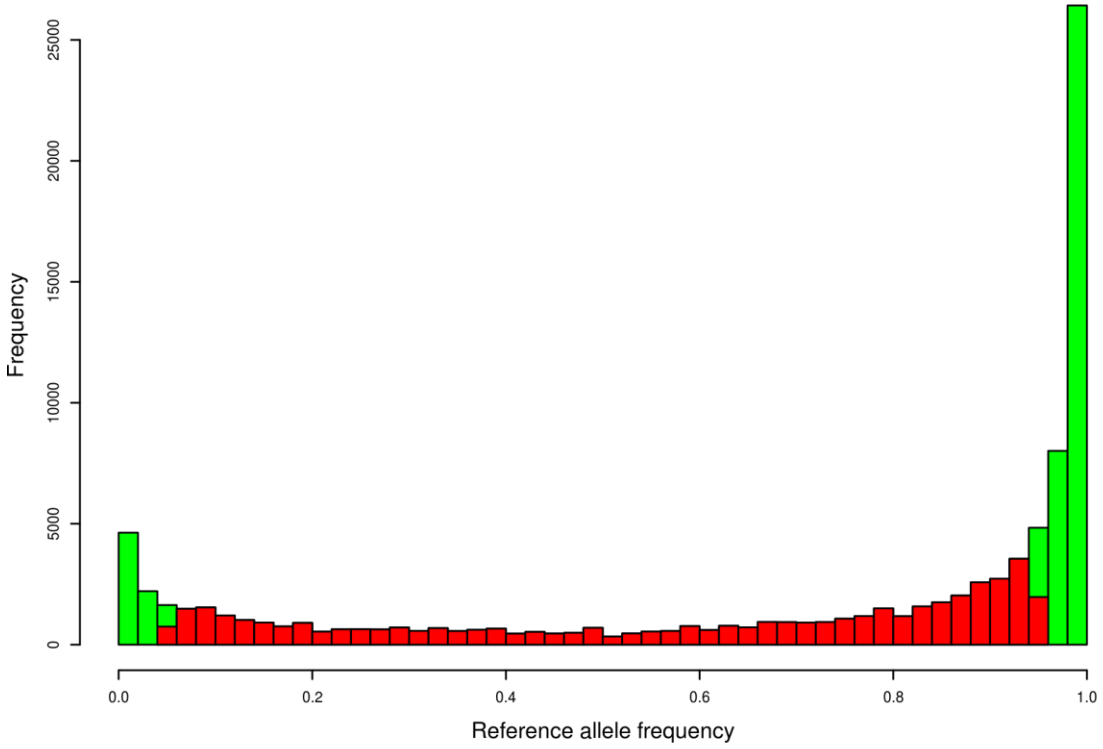
Supplementary figure 3-20. QQ plot and histogram of CQ GWAS p-values. Top) Sorted $-\log_{10}$(p-values) from the GWAS testing for associations with chloroquine are plotted on the y-axis versus randomly generated values from a uniform(0,1) distribution. The genome-wide inflation factor (λ) is shown in the top left of the QQ plot [263]. **Bottom)** Histogram of the same GWAS p-values. Under the null hypothesis p-values are expected to be distributed uniform(0,1). Based on these two plots and λ, there is no evidence for uncorrected population structure or model misspecification.



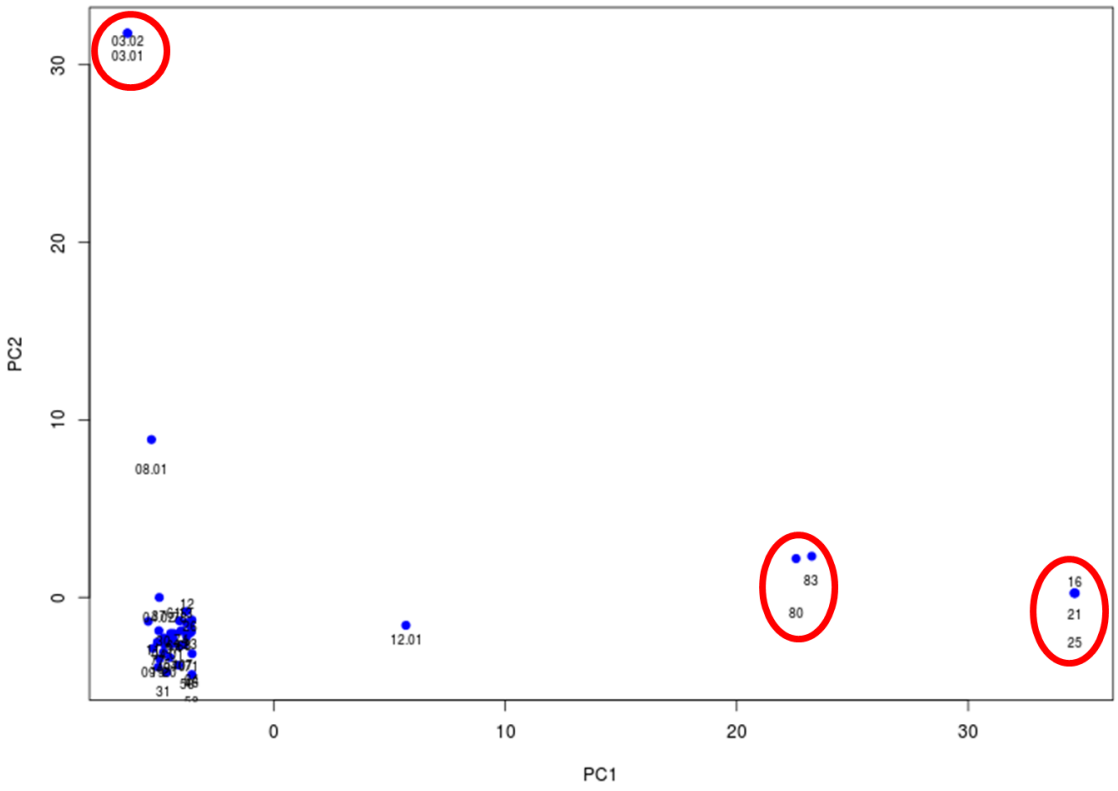
Supplementary figure 3-21. Histograms of $\log_{10}$(IC50) values for 22 drugs. Abbreviations and units are listed in Table 3-1.



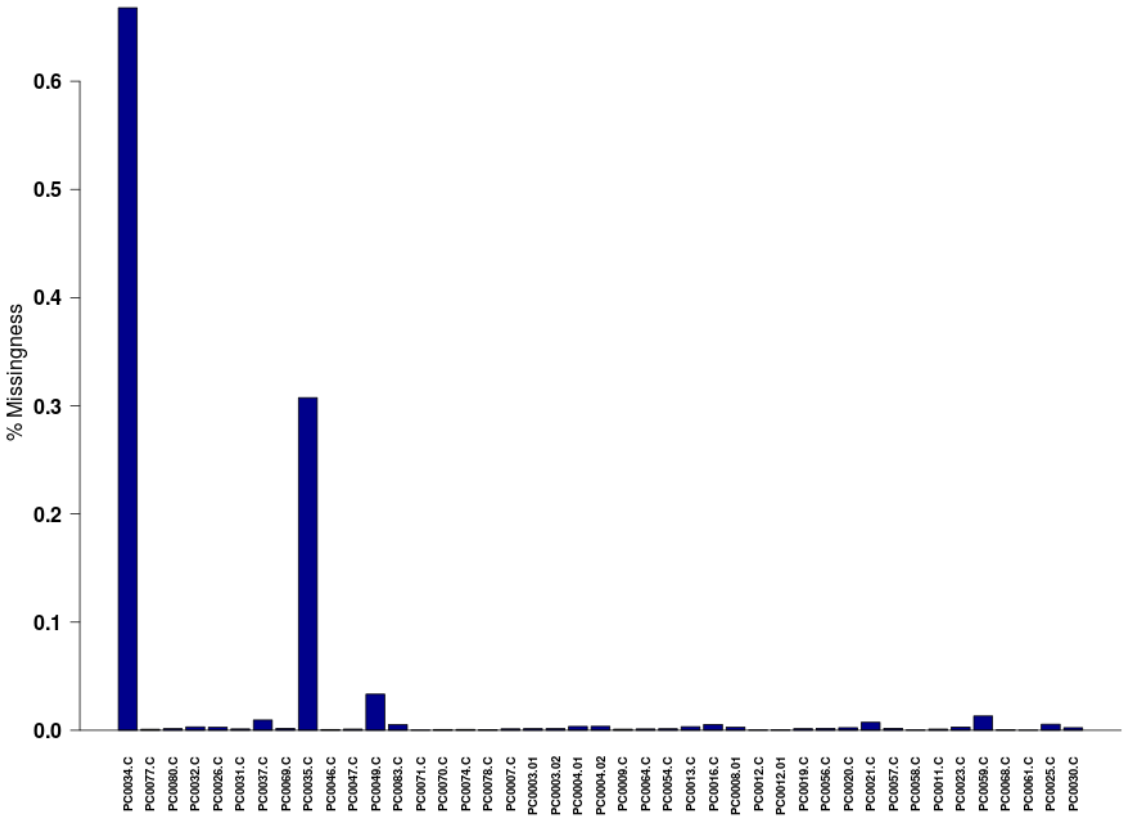
Supplementary figure 3-22. Heatmap depicting the level of heterozygosity in the sample set. SNPs (rows) are ordered by chromosome position. Samples (columns) are ordered by hierarchical clustering of Euclidean distances, based on the indicator variable 0=heterozygous, 1=homozygous.



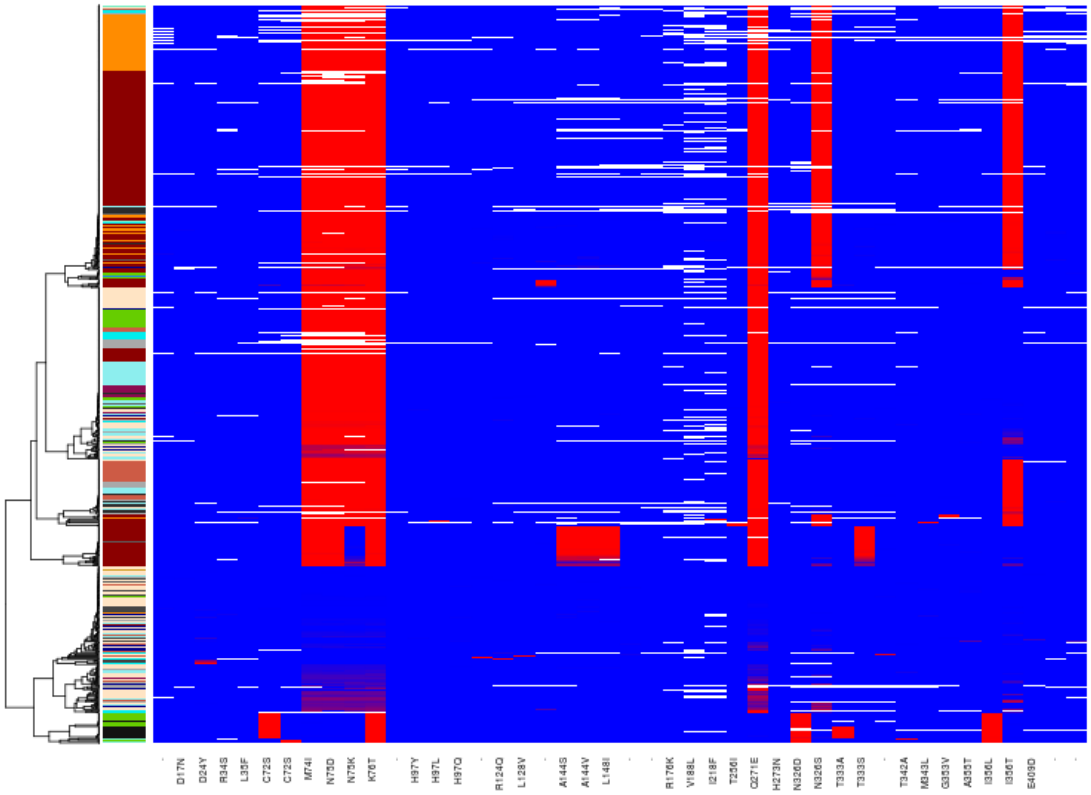
Supplementary figure 3-23. Histogram of within-sample allele frequencies. Red indicates the 7% of the data falling in the allele frequency range 0.05 to 0.95.



Supplementary figure 3-24. PCA plot of sequenced Kenyan samples. Each blue dot represents one sample. Each sample's SNPs were projected onto the first two PCA eigenvectors to determine its coordinates in the plot. Samples are labeled with jittered internal identifiers, so several identifiers may appear to label the same dot. Suspected outliers are circled in red. The 5 samples to the extreme right in PC1 (16, 21, 25, 80, and 83) are suspected contaminants from another country. PC2 appears to be driven by two identical samples (03.01 and 03.02).



Supplementary figure 3-25. SNP missingness in Kenyan sequenced samples. The percent missingness was calculated as the number of SNPs with coverage < 5 divided by the total possible for all samples.



Supplementary figure 3-26. *pfcrt* haplotype plot in an expanded MalariaGEN sample set. Each row represents a sample, and each column a MalariaGEN v2.0 SNP. The color of each cell indicates the allele frequency of the 3D7 allele, ranging from red (0% like 3D7) to blue (100% like 3D7). The gradient of colors between these extremes represents mixed infections. White cells indicate missing data. The far left column adjacent to the dendrogram is colored by country—samples are sorted by similarity of haplotype. Note the near perfect correlation between K76T and Q271E.

Supplementary table 3-6. Kenya GWAS data access. European Nucleotide Archive accession numbers and corresponding phenotypic data for the 35 samples used in this study. Drug abbreviations and concentration units are described in Table 3-1.

Accession	CQ	MFL	HLF	LUM	PIQ	QIN	PRIM	QuiNazol	AMOD	DEAQ	ISOQIN	PYRON	METHYLBL	DHA	METHOT	TRIMTX	PYRIM	TRIMETHO	CHLOPROG	CYCLOPG	ATV	WR99210
ERS016375	30.725	40.27175	92.505	169.28	36.28	122.345	1903.525	10.5665	7.79	17.43	3.7	34.85	7.8998	1.855	113.275	17.1174	2175.92	17394.313	4667.6	23	0.7	1.5329
ERS010455	90.705	42.05585	11.275	104.155	30.645	78.155	394.87	18.6265	5.475	8.515	2.18	4.325	5.7025	0.855	128.75	15.69085	2765.46	19404.3195	4602.09	38.035	2.24	0.28905
ERS010435	11.04	15.79805	10.6	93.375	30.985	33.73	458.2	6.3761	3.815	5.74	2.11	4.675	2.01295	1.04	147.355	5.71765	917.435	11735.093	830.595	10.995	1.31	0.31265
ERS010454	49.015	21.75495	7.965	123.575	26.92	27.825	1755.445	4.4796	2.615	3.915	1.43	3.085	3.01325	1.43	127.975	4.76005	1041.71	10088.879	4631.925	17.44	2.36	0.23785
ERS010464	83.05	3.22695	0.675	22.945	32.65	55.815	443.725	3.44875	6.23	9.385	2.175	1.51	0.9996	0.59	108.59	5.5181	479.82	4086.035	2308.06	7.08	1.7	0.128
ERS017458	41.575	40.19575	34.53	238.83	42.815	268.27	606.56	29.3145	7	7.245	3.75	5.81	9.26965	0.675	160.405	30.9052	2511.43	6268.2795	4067.38	57.545	0.775	0.0076
ERS016368	80.14	40.8371	71.9	97.635	24.795	198.595	623.14	23.2465	9.75	20.5	8.025	33.865	9.5151	1.93	88.32	22.7099	1562.23	10445.1625	7803.55	30.435	1.67	0.92305
ERS010641	73.015	9.01225	18.505	39.365	24.2	50.87	207.835	2.1535	9.675	16.675	3.495	10.845	1.5914	1.36	122.225	9.60425	1082.535	9515.328	973.545	11.895	0.29	0.40445
ERS016369	30.18	33.34135	27.305	69.345	20.925	202.775	484.19	4.363	7.665	7.425	2.885	11.585	2.327	1.425	101.455	11.27575	584.865	2963.924	2096.55	8.23	0.27	0.95345
ERS017460	31.61	36.7191	27.47	59.53	23.48	192	850.9	2.92	10.04	10.24	5.06	5.11	3.5172	0.98	171.98	13.4102	652.12	863.54	586.77	41.9	1.62	0.0025
ERS017459	83.37	33.914	16.86	123.165	17.475	207.235	326.705	3.6375	6.615	5.745	2.105	5.315	0.95855	0.47	74.035	7.15445	938.695	9971.096	544.465	14.03	0.435	0.66565
ERS017455	32.595	35.6631	10.52	60.26	18.21	262.77	1605.61	0.4965	8.49	8.46	3.465	18.98	4.5911	1.1	110.91	24.64135	446.02	1725.2375	1600.905	11.025	1.675	0.79115
ERS016376	109.72	53.5502	89.08	174.4	63.64	154.28	2154.58	20.534	9.31	13.43	6.22	23.35	15.3442	2.87	178.82	18.6319	1887.36	9864.529	4808.97	56.86	5.46	0.9129
ERS010401 ERS010402	50.125	17.11195	18.6	46.22	38.63	129.98	311.355	0.0142	9.99	8.59	2.25	29.98	8.66955	1.42	95.115	12.0291	550.895	2453.7405	1715.16	26.64	2.485	0.0213
ERS010403	12.99	5.3674	40.035	49.895	89.365	42.375	591.025	58.04855	8.405	8.64	1.295	45.7	0.1033	0.425	204.425	38.26965	4842.31	40546.7055	4568.83	93.04	19.825	0.01115
ERS010407 ERS010408	56.405	17.39265	89.455	90.81	56.885	32.795	240.125	3.78815	7.88	8.53	2.185	42.06	1.2626	0.2	78.51	12.90195	873.245	9687.469	1012.76	10.65	2.8	NA
ERS010409 ERS010410	76.03	7.8687	40.745	30.875	68.045	54.07	418.31	33.88525	8.555	9.785	1.37	40.18	0.14485	0.125	156.695	20.78705	2534.25	19865.7175	3278.8	38.805	8.345	NA
ERS010451	110.31	30.9385	36.195	50.185	19.8	181.54	242.125	27.01775	8.975	18.435	3.465	36.315	3.30645	1.07	165.715	49.75945	3796.325	18756.329	3218.945	107.84	4.69	0.43135
ERS010638	95.93	40.5743	41.47	115.96	17.13	67.68	282.595	3.2355	7.85	22.59	6.635	13.57	1.9529	1.2	58.37	9.6773	615.61	5825.9985	509.905	8.53	1.095	1.19815
ERS016370	78.965	63.86915	19.74	274.505	23.07	286.945	385.51	29.9905	6.505	6.53	4.18	4.24	3.3655	1.495	104.53	20.68515	1319.21	5734.073	3979.365	37.825	7.525	0.26375
ERS010411 ERS010412	92.955	52.9172	24.21	97.84	82.565	270.19	328.59	2.07865	6.375	7.91	2.5	26.005	25.80965	1.72	123.34	30.9761	777.255	16545.348	2389.96	37.8	1.99	0.00215
ERS010457	86.16	17.5761	14.91	36.76	34.91	145.77	414.345	3.94205	8.135	16.99	3.405	23.975	2.7722	0.84	139.385	11.49305	1314.605	9796.1825	922.155	13.57	2.595	0.06125
ERS010413 ERS010414	114.67	3.5479	5.33	33.43	56.44	238.76	542.54	7.088	7.86	21.67	3.4	16.49	1.3984	0.33	130.18	49.9524	932.9	7645.151	4301.14	52.7	3.9	0.0021
ERS010461	54.935	20.34035	5.45	75.09	28.87	91.615	576.885	0.00715	3.41	8.63	2.275	4.215	5.34225	0.475	63.535	4.0462	245.945	1440.326	1659.36	20.715	1.355	5e-05
ERS010452	14.98	31.4788	9.455	84.595	56.195	207.675	4305.555	0	7.905	8.51	3.485	9.17	6.578	1.2	83.595	6.9019	128.6	897.3005	4147.21	41.145	3.095	0
ERS016371	14.02	16.0402	10.37	284.74	20.92	81.96	467.18	5.126	11.52	8.1	2.11	12.8	0.5971	0.93	66.86	7.6363	840.1	9230.67	297.68	10.59	1.17	0.9154
ERS010462	87.195	36.1515	45.1	103.515	41.96	233.955	277.06	5.3539	6.425	11.145	2.885	45.255	26.1194	3.225	140.575	59.1883	1458.395	9990.156	4418.98	95.89	12.445	0.23345
ERS016372	16.765	NA	NA	346.385	12.145	NA	NA	NA	NA	NA	NA	NA	NA	1.59	NA	NA	NA	NA	NA	NA	NA	NA
ERS010644	10.11	NA	NA	332.435	16.15	NA	NA	NA	NA	NA	NA	NA	NA	0.45	NA	NA	NA	NA	NA	NA	NA	NA
ERS010416 ERS010415	35.33	34.9506	21.2	59.25	82.29	381.8	1748.64	0.0383	4.27	3.21	1.16	49.7	12.537	0.07	138.98	50.5258	1391.69	7271.692	1914.76	52.83	1.02	6e-04
ERS010453	16.84	71.5203	41.82	130.02	70.5	81.18	730.45	0.0036	4.94	3.05	1.74	35.09	9.7399	2.42	130.11	48.0824	99.85	215.963	2741.8	101.37	10.03	0
ERS016373	2.345	NA	NA	225.13	15.075	NA	NA	NA	NA	NA	NA	NA	NA	0.995	NA	NA	NA	NA	NA	NA	NA	NA
ERS010640	112.98	NA	NA	410.535	19.675	NA	NA	NA	NA	NA	NA	NA	NA	0.76	NA	NA	NA	NA	NA	NA	NA	NA
ERS016374	2.27	NA	NA	371.765	18.675	NA	NA	NA	NA	NA	NA	NA	NA	0.49	NA	NA	NA	NA	NA	NA	NA	NA
ERS010438	15.705	39.35435	9.535	67.115	56.685	270.42	1516.52	0.1584	6.805	8.8	3.46	8.095	4.5459	1.58	110.055	43.2382	687.22	2263.0595	5149.99	30.6	2.96	1e-04

Supplementary table 3-7. Pairwise drug correlations.

	WR99210	LUM	ISOQIN	AMOD	DEAQ	PRIM	QIN	DHA	MTHYLBL	MFL	PYRON	HLF	CQ	TRIMETH	QuiNazol	PYRIM	PIQ	ATV	CHLOPRO	METHOT	CYCLOPG	TRIMTX
WR99210	1	0.41	0.32	0.33	0.2	-0.08	-0.29	0.41	-0.17	0.21	0.1	0.4	0.03	0.28	0.34	0.14	-0.6	-0.32	-0.07	-0.28	-0.41	-0.27
LUM	0.41	1	0.19	-0.23	-0.37	0.18	0.1	0.26	0.34	0.67	-0.18	0.28	-0.23	0.04	0.17	0.08	-0.43	-0.13	0.14	-0.2	0.05	-0.07
ISOQIN	0.32	0.19	1	0.42	0.55	0.07	0.36	0.46	0.32	0.46	-0.13	0.25	0.28	-0.23	0.14	0.01	-0.34	-0.13	0.15	-0.08	0.03	0.06
AMOD	0.33	-0.23	0.42	1	0.53	-0.08	-0.02	0.02	-0.2	-0.18	0.33	0.27	0.03	0.01	0.21	0.13	-0.17	0.05	-0.13	0.05	0.02	0.08
DEAQ	0.2	-0.37	0.55	0.53	1	-0.19	-0.1	0.07	-0.1	-0.14	0.24	0.24	0.49	0.2	0.23	0.25	-0.1	0.13	0.16	0.08	0.04	0.14
PRIM	-0.08	0.18	0.07	-0.08	-0.19	1	0.21	0.15	0.32	0.23	-0.09	-0.11	-0.48	-0.26	-0.02	-0.11	0.25	-0.01	0.43	0.17	0.22	0.05
QIN	-0.29	0.1	0.36	-0.02	-0.1	0.21	1	0.18	0.51	0.46	0.06	-0.06	0.12	-0.32	-0.1	-0.05	0.04	-0.02	0.18	0	0.35	0.52
DHA	0.41	0.26	0.46	0.02	0.07	0.15	0.18	1	0.59	0.62	0.02	0.3	0.06	-0.08	-0.06	-0.11	-0.02	0.13	0.4	-0.04	0.11	0.13
MTHYLBL	-0.17	0.34	0.32	-0.2	-0.1	0.32	0.51	0.59	1	0.7	0.18	0.28	0.1	-0.17	-0.23	-0.02	0.29	0.08	0.41	0.15	0.47	0.4
MFL	0.21	0.67	0.46	-0.18	-0.14	0.23	0.46	0.62	0.7	1	0.01	0.43	0.15	-0.12	-0.04	0.04	-0.03	0.01	0.31	-0.01	0.31	0.33
PYRON	0.1	-0.18	-0.13	0.33	0.24	-0.09	0.06	0.02	0.18	0.01	1	0.67	0.06	0.33	0.1	0.33	0.47	0.34	0.13	0.28	0.4	0.65
HLF	0.4	0.28	0.25	0.27	0.24	-0.11	-0.06	0.3	0.28	0.43	0.67	1	0.12	0.32	0.31	0.42	0.21	0.14	0.18	0.26	0.33	0.43
CQ	0.03	-0.23	0.28	0.03	0.49	-0.48	0.12	0.06	0.1	0.15	0.06	0.12	1	0.3	0.18	0.29	0.13	0.18	0.09	0.06	0.12	0.18
TRIMETH	0.28	0.04	-0.23	0.01	0.2	-0.26	-0.32	-0.08	-0.17	-0.12	0.33	0.32	0.3	1	0.67	0.81	0.15	0.13	0.3	0.38	0.14	0.16
QuiNazol	0.34	0.17	0.14	0.21	0.23	-0.02	-0.1	-0.06	-0.23	-0.04	0.1	0.31	0.18	0.67	1	0.81	0.02	0.27	0.46	0.42	0.3	0.18
PYRIM	0.14	0.08	0.01	0.13	0.25	-0.11	-0.05	-0.11	-0.02	0.04	0.33	0.42	0.29	0.81	0.81	1	0.16	0.17	0.44	0.57	0.44	0.4
PIQ	-0.6	-0.43	-0.34	-0.17	-0.1	0.25	0.04	-0.02	0.29	-0.03	0.47	0.21	0.13	0.15	0.02	0.16	1	0.49	0.44	0.43	0.5	0.45
ATV	-0.32	-0.13	-0.13	0.05	0.13	-0.01	-0.02	0.13	0.08	0.01	0.34	0.14	0.18	0.13	0.27	0.17	0.49	1	0.46	0.39	0.58	0.45
CHLOPRO	-0.07	0.14	0.15	-0.13	0.16	0.43	0.18	0.4	0.41	0.31	0.13	0.18	0.09	0.3	0.46	0.44	0.44	0.46	1	0.3	0.53	0.44
METHOT	-0.28	-0.2	-0.08	0.05	0.08	0.17	0	-0.04	0.15	-0.01	0.28	0.26	0.06	0.38	0.42	0.57	0.43	0.39	0.3	1	0.66	0.51
CYCLOPG	-0.41	0.05	0.03	0.02	0.04	0.22	0.35	0.11	0.47	0.31	0.4	0.33	0.12	0.14	0.3	0.44	0.5	0.58	0.53	0.66	1	0.72
TRIMTX	-0.27	-0.07	0.06	0.08	0.14	0.05	0.52	0.13	0.4	0.33	0.65	0.43	0.18	0.16	0.18	0.4	0.45	0.45	0.44	0.51	0.72	1

4 GENOMIC DIVERSITY OF TANZANIAN FIELD ISOLATES

4.1 Abstract

This second chapter of Section I introduces a Tanzanian dataset that is used for validation of methods developed in Section II. Here I describe the MOMS project—a clinical field study through which these samples were collected. I discuss some issues related to accessibility of SNPs in the parasite genome, and then present results about population structure, within-sample mixture, and limitations of reference-based SNP variation in a candidate gene, *eba-175.*

4.2 Introduction

Rapid access to high resolution parasite genotypes will be an essential component of strategies for malaria control and elimination [142]. As discussed in the previous chapter, regardless of the therapy, the potential for parasites to become drug resistant is a consideration only of time. That holds for every antimalarial deployed to date, and now includes the only treatment standing in the way of widespread suffering and death—the front-line drug artemisinin [264]. Genome-wide SNP data can be used not only for drug association analyses, but for population genetic and parasite diversity studies that can provide real-time intelligence about looming threats. For example, delayed infection clearance times in individuals treated with artemisinin mono or combination therapies have been reported for years in western Cambodia, and investigations in Southeast Asia have demonstrated the power of population genomics to foreshadow such events [143,265]. Outlying clusters in population structure plots of Cambodian parasites, based on genome-wide SNPs, were highly correlated with slower *in vivo* parasite clearance times in a recent study, and these parasites showed classic evidence of founder effects that would be detected by organized genomic surveillance programs [143]. Until recently, ready access to genomic

data on the scale and resolution necessary to conduct such analyses has been limited, but this obstacle is being overcome by MalariaGEN partnerships that contribute field samples for aggregated investigations. In this chapter I present my own involvement in a MalariaGEN partnership through the MOMS project in Tanzania.

Through the MOMS partnership, 64 parasite samples have been submitted for Illumina sequencing, of which 50 were successfully sequenced (see methods section 2.4). The samples were processed in several batches, and at the time of this writing a population genetic analysis of all 50 samples is underway. Here I present results from the first batch of 16 samples.

4.3 Results

4.3.1 Genome-wide accessibility and limitations

Each sample yielded 1-2Gb of short read sequence data. An important problem with sequencing infected peripheral blood samples is that host DNA overwhelms that of the parasite—even in hyper-parasitemic samples. Typically samples are not sequenced unless human DNA comprises less than 80-90% of the extraction, otherwise coverage suffers or the cost must increase to attain sufficient sequence depth of the parasite. Fresh blood samples can be depleted of leukocytes by filtering through a cellulose column made with CF11 powder [266]. As described in the methods, these parasites were short-term adapted to culture after cryopreservation. The blood used for culture was leukodepleted by aspirating the buffy coat after several wash and centrifugation steps, thus host contamination was not an issue.

As illustrated with *pfcrt* in the previous chapter (Figure 3-17), low complexity regions like introns have lower-than-expected coverage depth when reads are aligned to the reference sequence. This property generalizes to the entire genome, and has consequences on SNP accessibility. For example, Figure 4-1 shows the read depth covering potential SNPs in both coding and non-coding positions, and highlights the stark contrast in the accessibility of these SNPs. Coverage in coding regions peaks at nearly 150x, whereas the peak in non-coding regions is closer to 12x. The lower coverage of non-coding regions in Tanzanian samples recapitulates similar findings in a broader sample set, and is partly due to the extreme AT bias in these regions (approximately 90%, vs. closer to 70% in coding regions) [86]. However, a closer look at coverage in only the coding SNPs (colored green in the

histogram) shows a second peak at the low end of the distribution, and further inspection at the gene level implicates polymorphism, as well as AT bias, as affecting coverage depth.

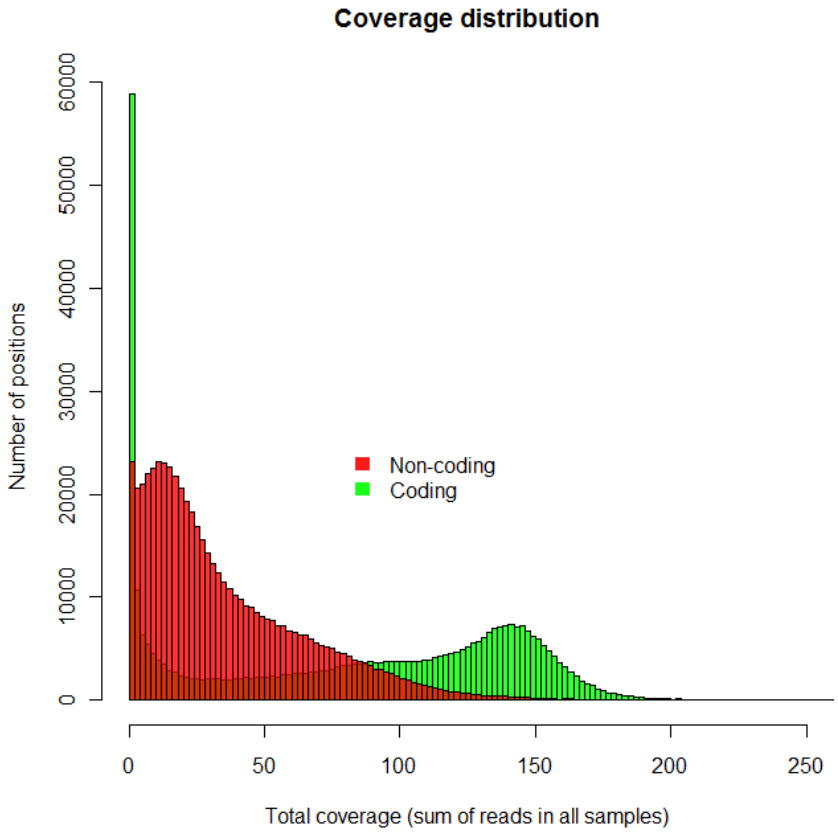


Figure 4-1. Distributions of SNP read depths in Tanzanian samples. MalariaGEN filters drop candidate SNPs that appear in non-coding regions. Before those filters were applied, the frequency distribution of the coverage depth of all candidate SNPs was calculated. Red indicates SNPs in non-coding regions and green are those within genes. The distribution of coding variants (green) appears to be bi-modal, and this is elaborated upon in the Figure 4-2.

Figure 4-2 provides a more detailed view of coverage at the gene level, and an ontological analysis of the genes comprising the lowest-covered peak (defined as those with mean coverage below 12.81 (see methods section 2.4.5 for details on mixture modeling)) enriches for gene families well-known to be some of the most highly polymorphic in *P. falciparum*—i.e., primarily *vars*, *rifs*, and *stevors* (N=165). The genes in this low-coverage peak are largely subtelomeric—a property of the aforementioned families (Figure 4-3). The gene coverage distribution appears to be bi-or tri-modal, and more than half of the genes in the higher coverage peaks are defined in PlasmoDB as conserved with unknown function. In other

words, more variable genes have lower coverage depth when reads are aligned to the 3D7 sequence.

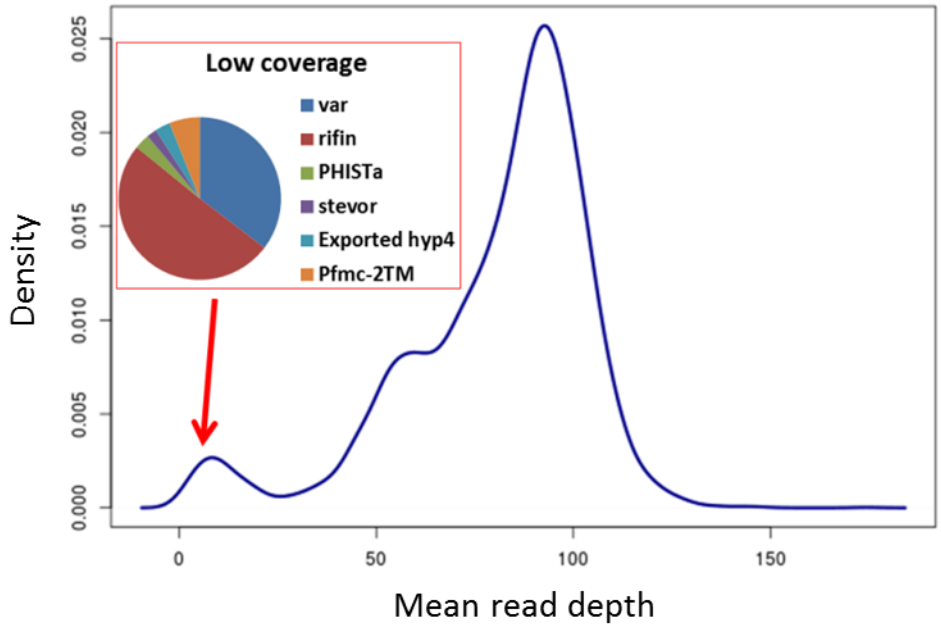


Figure 4-2. Coverage depth of genes in Tanzanian samples. Mean read depth of every parasite gene was calculated and the kernel density estimated for this distribution plot. This distribution appears to be bi- or tri-modal, and the left-most peak corresponds to the most polymorphic parasite genes. The inset pie-chart presents the ontological breakdown of the genes comprising the low-coverage hump.

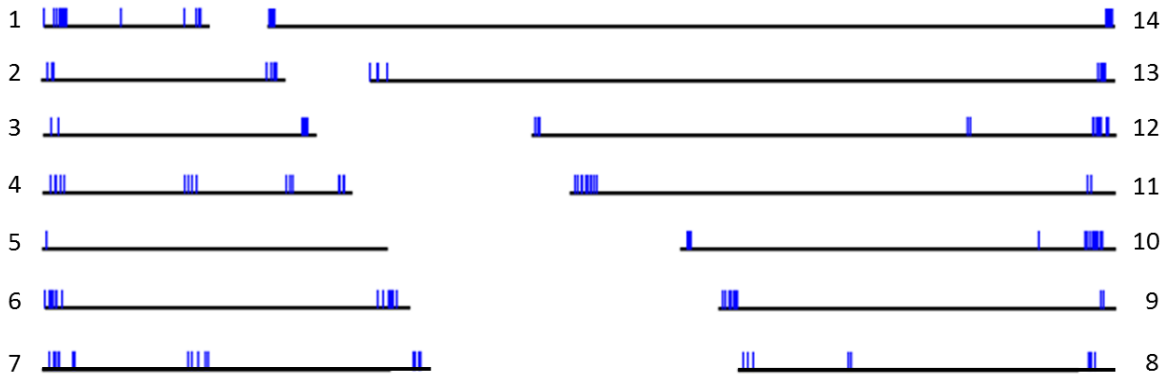


Figure 4-3. Chromosomal positions of the lowest coverage genes in Tanzanian samples. Genes with mean coverage depth lower than 12.81 (N=165), creating the low coverage peak in the Figure 4-2 density plot, are plotted as blue ticks according to chromosome position. Numbers along the outer edge of the plot indicate chromosome number. Plot was modified from MADIBA output [267].

In light of the precariousness of coverage in non-coding and polymorphic regions, coverage and coding filters are among those employed by MalariaGEN to define a list of high-quality "typable" SNPs. These SNPs represent the middle 70% of exonic positions in the distribution of coding regions, and are the variants used for population genetic analyses in this chapter. In the version 2.0 MalariaGEN release, SNPs were ascertained using 1685 samples from 20 countries. In an initial scan for variable positions in any sample, nearly 10% of the 23Mb genome become candidate polymorphisms (i.e., more than 2 million candidate SNPs before filtering). These candidates are filtered for exonic loci with at least 5x coverage in more than 80% of the samples, yielding 420241 typable SNPs.

4.3.2 Population structure

Using the MalariaGEN high quality SNPs, a neighbor-joining tree based on the pairwise genetic distances between 317 samples from 8 countries reflects some expected and perhaps unexpected population structure (Figure 4-4). These other countries were added for context, and the data have been previously published (without the Tanzanian samples) [86]. The largest distance in the tree is between the nodes separating African and Asian samples. As previously reported, there is far more population structure apparent in Asian samples than in those from Africa—particularly West Africa. This is not related to geographic distance, as the sites in West Africa are farther apart than those in Southeast Asia. As mentioned in the introduction, the source of some of the Cambodian structure is related to artemisinin tolerance in certain parasite sub-populations. The Tanzanian samples cluster with those from neighboring Kenya, and are part of an east-west population divide within Africa. In contrast to Asia, within East and West Africa parasites appear to undergo much more genetic mixing. One of the Kenyan parasites is a clear outlier—perhaps contaminated with an Asian sample.

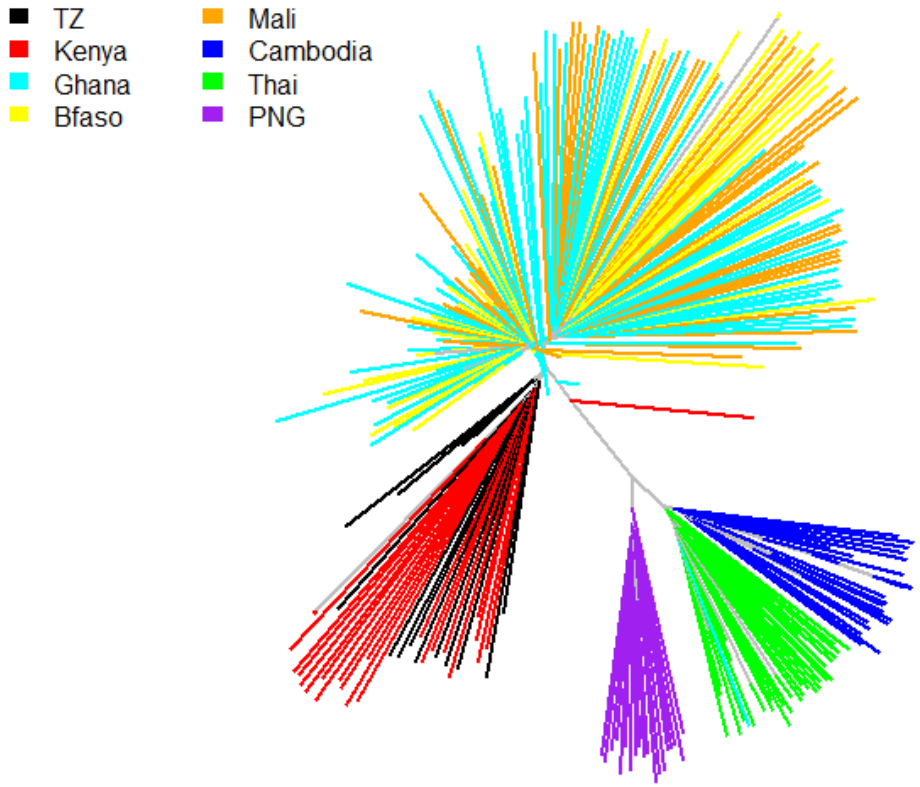


Figure 4-4. Neighbor-joining trees representing population structure. Pairwise distances are represented for 317 MalariaGEN isolates from 8 countries (see legend).

A similar tree using only Tanzanian samples fails to reveal structure between geographic regions, nor between parasites collected from mothers versus children (Figure 4-5). Although it may be that this sample of parasites in Tanzania lacks genetic structure as witnessed in West Africa, it is possible that genome-wide summaries are not the appropriate tool for detecting it at this resolution. The thickness of the lines in Figure 4-5 depicts an estimate of within-sample mixture (F_{ws}), discussed further in 4.3.4.

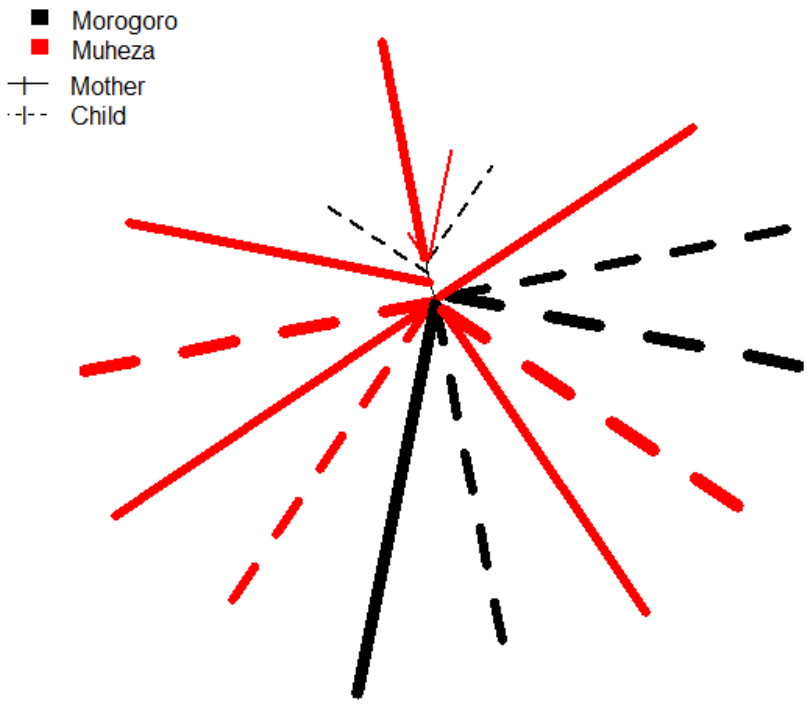


Figure 4-5. Neighbor-joining tree comparing Tanzanian samples. Pairwise distances are represented for 16 Tanzanian samples. Samples from Morogoro are labeled black, and Muheza red. Dashed lines indicate samples from children, and solid lines those from mothers. Line width is proportional to inbreeding coefficient (F_{ws}).

Previous studies have uncovered structure using principal components analysis that wasn't readily apparent in a neighbor-joining tree [143]. To determine if a similar situation was occurring here, a PCA calculation was performed on variable positions within these 16 samples, and for which no sample had a missing value. There are multiple approaches to creating the matrix to use for PCA in this context. In this case a binary matrix was generated, based on a score of 1 if the reference allele was represented in the majority of reads, and 0 otherwise. In other words, a matrix of dimension 8304 rows (SNPs) by 16 columns (samples) was generated, in which each cell was a 1 or 0, based on the majority allele within that sample (i.e., mixture was ignored). As illustrated with the first 4 principal components in Figure 4-6, using this approach there is still no evidence of sample stratification related to patient type or to geography.

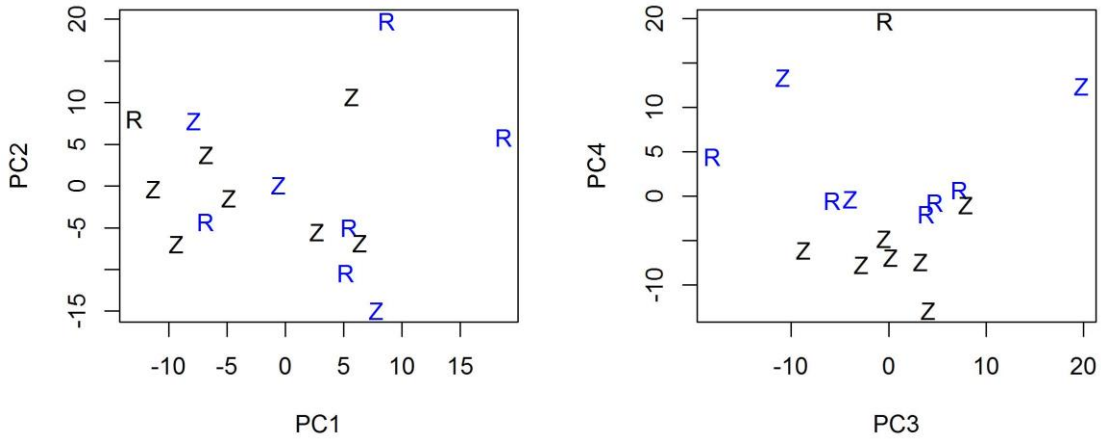


Figure 4-6. Principal components analysis of Tanzanian samples. PCA was calculated based on the majority call allele for each SNP in each sample. SNPs with missing values in any of 16 Tanzanian samples were excluded (N=8304 SNPs used). Left panel shows the first two principal components and the right panel shows components 3 vs 4. Samples from mothers are colored black and those from children are blue. Samples from Morogoro are labeled with and 'R' and those from Muheza are labeled with a 'Z'. No population structure is apparent between parasites infecting patient types or due to geography.

4.3.3 Allele frequencies

Of the SNPs ascertained by MalariaGEN across 20 populations, this set of Tanzanian samples shows variation at 14,261 sites, ignoring mixed infections. Examination of the allele frequency distribution is consistent with reports indicating a recent population expansion, marked by an excess of low frequency variants in Tanzanian parasites (Figure 4-7) [132]. This pattern is the same for both synonymous and non-synonymous variants.

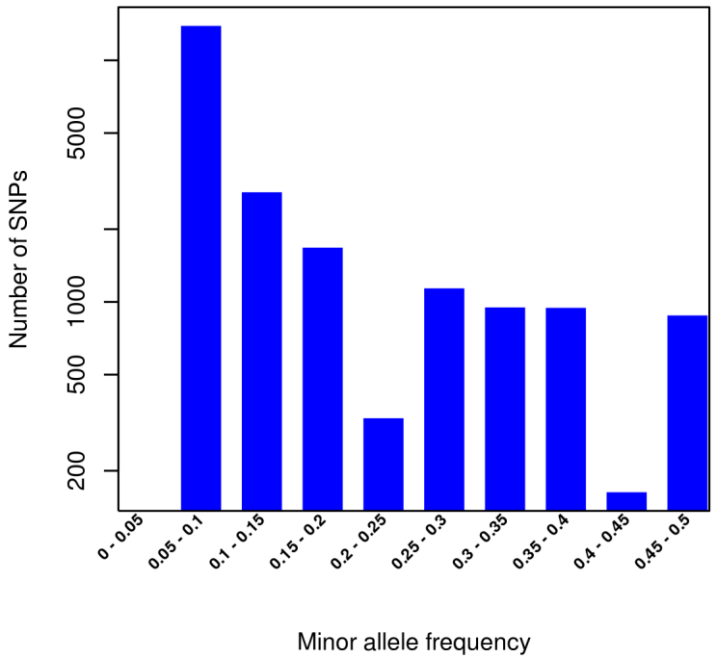


Figure 4-7. Allele frequency distribution in Tanzanian samples. Counts of SNPs in minor allele frequency (MAF) bins for 16 parasites sequenced from Tanzania. Y-axis is the actual count, and x-axis indicates the MAF bin. Note with only 16 samples the minimum detectable MAF is 0.0625.

Sixteen SNPs that occur in at least two isolates are private to Tanzania, compared to the rest of Africa (Table 4-1). The comparison dataset was a summary of 756 samples from East and West Africa that were available in the MalariaGEN v2.0 release. The majority of these private variants are found in known polymorphic genes, however several have unknown functions. Nothing in this list stands out as being potentially involved in drug resistance or selective pressures related to the vector. It should be restated that the most hyper-polymorphic and divergent regions of the genome are filtered out as candidate SNPs, and thus this list is likely a subset of all truly private alleles.

Table 4-1. Private SNPs in Tanzanian samples. SNPs present only in Tanzania, as compared to 756 samples from East and West Africa. SNPs must have appeared in at least 2 Tanzanian samples to make the list.

Chromosome	Position	Gene	Definition
MAL2	83423	PFB0085c	DnaJ protein, putative
MAL3	38123	PFC0005w	erythrocyte membrane protein 1, PfEMP1
MAL4	941349	PFD0995c	erythrocyte membrane protein 1, PfEMP1
MAL4	1105747	PFD1160w	surface-associated interspersed gene 4.2, (SURFIN4.2)
MAL4	1105853	PFD1160w	surface-associated interspersed gene 4.2, (SURFIN4.2)
MAL5	1341593	PFE1640w	erythrocyte membrane protein 1+(PfEMP1), truncated
MAL6	1337057	PFF1555w	rifin
MAL7	81465	MAL7P1.225-a	Plasmodium exported protein (PHISTa-like)
MAL8	563028	MAL8P1.105	conserved protein, unknown function
MAL10	980019	PF10_0224	dynein heavy chain, putative
MAL10	1629678	PF10_0402	rifin
MAL11	463900	PF11_0127	conserved Plasmodium protein, unknown function
MAL12	30373	PFL0015c	rifin
MAL12	51602	PFL0030c	erythrocyte membrane protein 1, PfEMP1
MAL12	315163	PFL0350c	conserved Plasmodium protein, unknown function
MAL13	2039445	MAL13P1.258	conserved Plasmodium protein, unknown function

4.3.4 Within-sample heterozygosity

Malaria infections in natural populations are frequently not clonal. This MOI can result when an already infected person is inoculated by a second mosquito, or from a single bite in which the vector itself harbored a mixture of parasites. Previous studies have reported a connection between transmission intensity and MOI, and this is somewhat intuitive—i.e., lots of biting by infected mosquitos fosters an environment in which parasites can intermix, however this is also dependent on the degree of population structure [86]. F_{ws} is a metric related to Wright's inbreeding coefficient and represents the chance that a sample will yield a meiotic outcross in the subsequent mosquito mid-gut [86]. This metric compares the heterozygosities of SNPs within an infection to those of the surrounding population, and is highly correlated with standard measures of MOI [268]. Although Muheza is an area of intense transmission, and Morogoro is more moderate, we do not find a pattern of MOI related to these geographic regions (Figure 4-8). This is perhaps unsurprising since these samples were cryopreserved and adapted to short-term culture before sequencing, which might be expected to decrease the within-sample diversity.

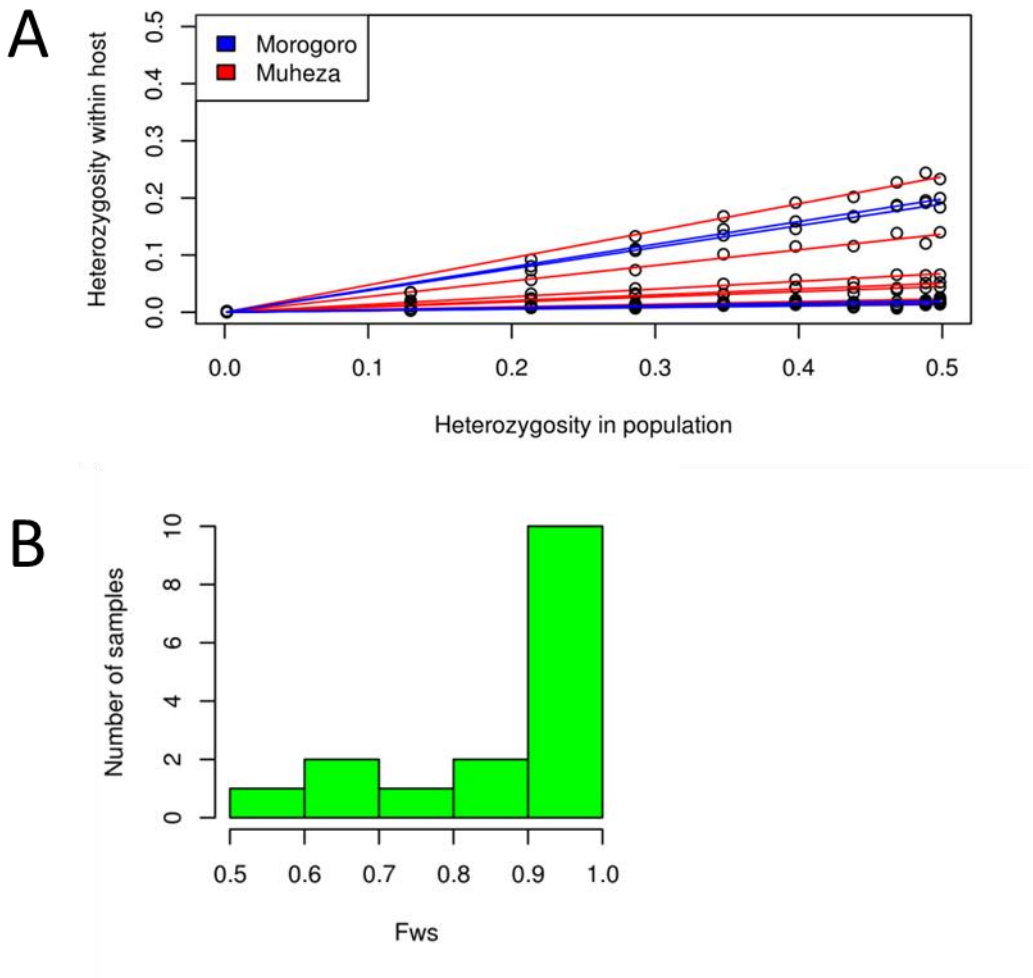


Figure 4-8. Representation of within-host diversity. A) Relationship of the heterozygosity in each sample with that present across all samples. Each line represents one isolate. SNPs were binned by allele frequency, and a regression line fitted for each sample using the mean within these bins compared to the population means. Inbreeding coefficient, F_{ws}, for each sample is estimated as 1 minus the slope of this line. **B)** Histogram of resulting F_{ws} values.

4.3.5 Missingness in *eba-175* and *msp3.4*

In the previous chapter *pfcrt* was used as an example to show how sequencing depth drops in low complexity regions, and, perhaps exacerbated by the dense polymorphism of the K76T region of exon 2, led to missingness in this important area. In that context it impacted GWAS results. Here I sought to show the etiology of another source of missingness using the gene *eba-175*, a vaccine candidate that contains two large indels. Exhaustive detail about *eba-175* is provided in Section III, but the motivation to look further into this complex variation came from the analysis of the Tanzanian samples, and thus I introduce it briefly here. Figure 4-9 shows a haplotype plot of the MalariaGEN SNP genotypes in *eba-175* for the

16 Tanzanian samples. Like for *pfcrt*, this plot depicts regions with extensive missingness (white cells). In contrast to *pfcrt*, this missingness is much more systematic in particular samples, and has more consistent boundaries.

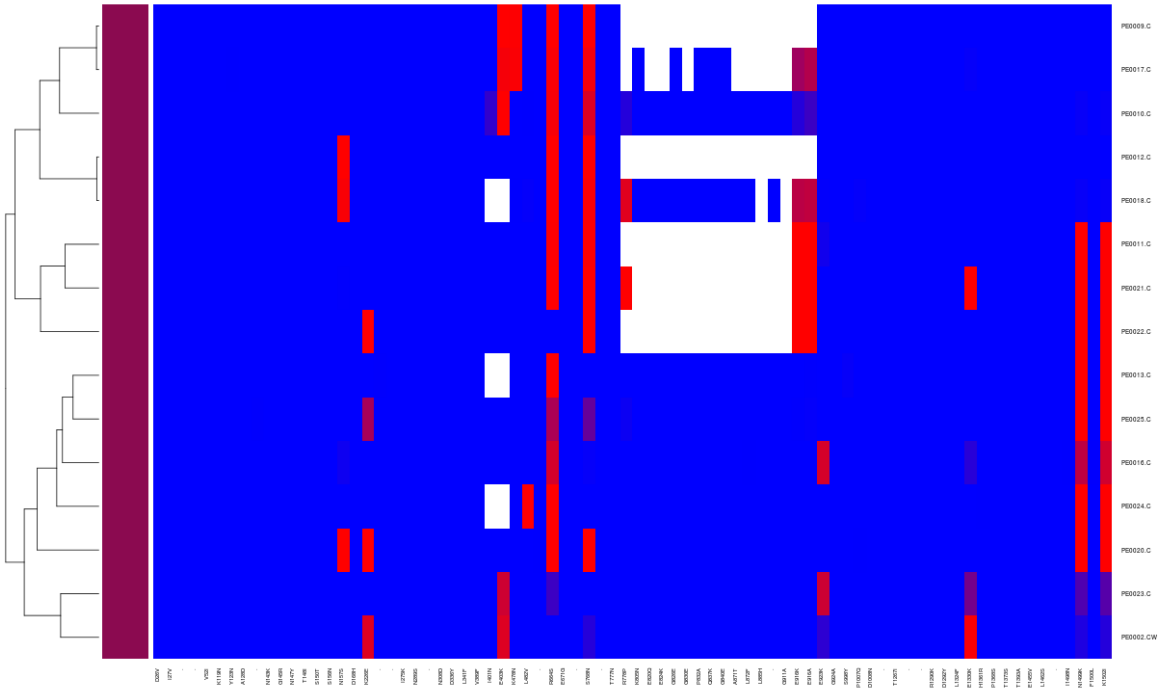


Figure 4-9. Haplotype plot for *eba-175* (MAL7P1.176). Each row reflects a sample, and each column a potential MalariaGEN v2.0 SNP. The color of each cell indicates the allele frequency of the reference allele, ranging from red (0% like 3D7) to blue (100% like 3D7). The gradient of colors between these extremes represents mixed infections. White cells indicate missing data. The far left column adjacent to the dendrogram is colored by country—in this case all maroon for Tanzania. The central block of missingness in half of the samples is in the region of the F-indel. SNPs that are filtered out in this region are left in so the missingness can be seen in the area of the indel.

The candidate gene report for *pfcrt* showed massive drops in coverage in intronic regions and around the dense polymorphism in exon 2 (Figure 3-17). A similar plot for *eba-175* shows a sharp drop to zero coverage in some samples but not others (Figure 4-10). This occurs directly over the location where the "F insert" track maps the location of that so-called indel in red. The reference parasite contains this insert, thus the drop is occurring for samples containing parasites with the deletion, as they lack reads covering this region. Recall that one of the MalariaGEN filters is to exclude SNPs with less than 5x coverage in more than 20% of samples. As described in explicit detail in Section III, this deletion occurs in more than 20% of parasites in nearly every population. Thus, while the unfiltered SNP

track shows candidates in this region, there is a void of v2.0 SNPs in the F-indel span. Many of these candidate SNPs are rightly filtered out, as read mapping in this area is noisy due to the indel, generating false positive variation.

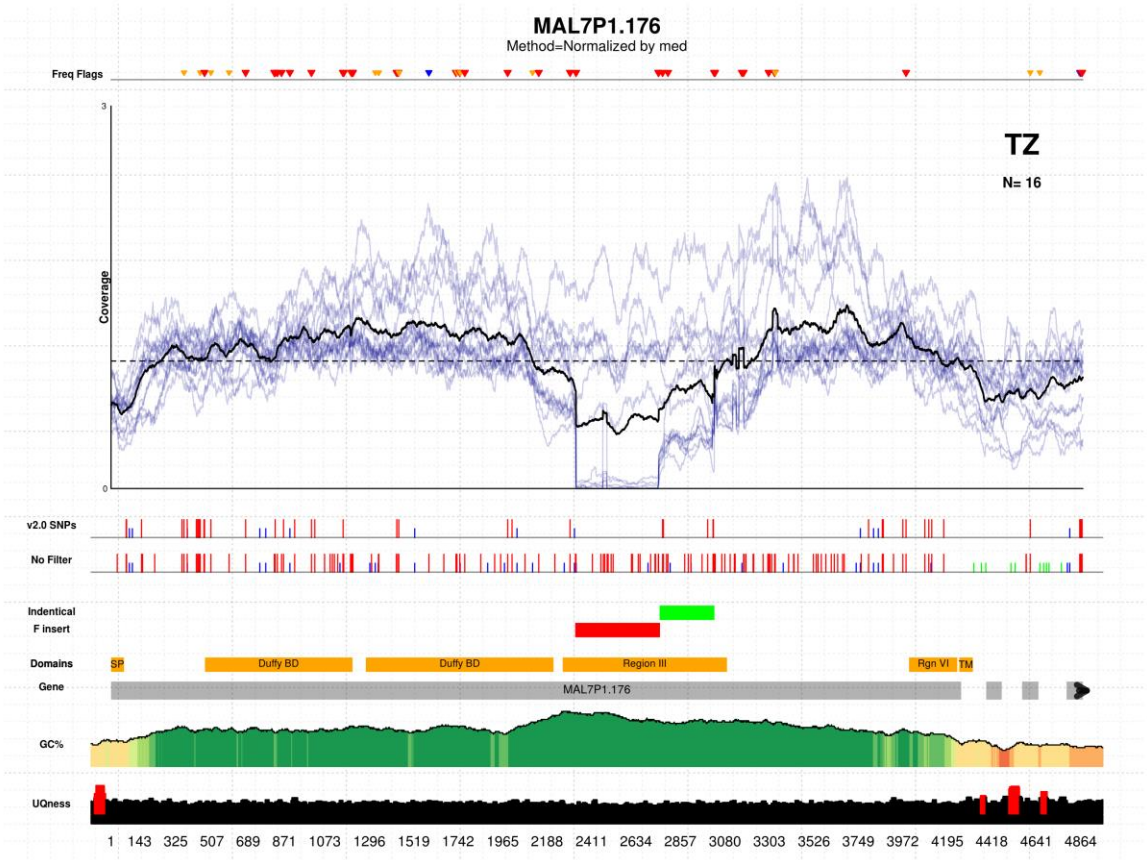


Figure 4-10. Candidate gene plot for *eba-175*. Complete details describing all tracks and data sources can be found in methods section 2.2.1. The coverage track contains 16 blue lines representing the normalized sequence read coverage for each Tanzanian sample. The normalizing constant for each sample is the median (indicated in the title) coverage of SNPs in all genes for the given sample. The black line is the mean across all samples. Two SNP tracks appear just below the coverage plot. Tall red SNPs are non-synonymous while blue lines are synonymous. The F-insert track highlights the location of that insert. The "identical" track indicates a stretch of near identity in all samples occurring between the F-indel and another that will be described in detail later.

A more recent vaccine candidate, *msp3.4*, displays missingness and coverage patterns that lie somewhere between those of *pfcrt* and *eba-175* (Supplementary figure 4-11 and Supplementary figure 4-12). As described in more detail in 6.7.2, *msp3.4* contains a 500bp stretch that in some parasites shares no apparent homology with the 3D7 version of the gene. Intriguingly, this complex structure appears in the region of the DBL domain (Supplementary figure 4-12), which in other gene families is a domain associated with adhesion to host receptors [115,269,270]. Like the indel in *eba-175*, this divergent region

appears in some parasites and not others, however the boundaries are not as clean as the indel. The ramification on MalariaGEN SNP filters is highlighted by comparing the v2.0 and "no filter" tracks in Supplementary figure 4-12. As depicted in Supplementary figure 6-21, these dimorphic forms are present at high frequency in every population, again tripping the MalariaGEN coverage filter if a position doesn't have a depth of at least 5x in more than 80% of samples.

4.4 Discussion and future work

The neighbor-joining tree based on pairwise SNP differences in the 16 Tanzanian samples did not suggest any obvious structure between parasites from Muheza and Morogoro, nor between those isolated from mothers versus infected children. This would perhaps be the most interesting question to address with this dataset. Although principal components analysis has been shown to provide better resolution for detecting population structure than neighbor-joining trees, PCA indicated a similar level of panmixis. It is possible that even though parasites are recombining with one another randomly across these populations, specific loci could be associated with, for example, patient type; however, that fine a scale of association may be beyond the limits of detection of the methods used above. A GWAS approach could be informative to answer this type of question, but should be performed using the larger samples size available with the more recently sequenced batches. A caveat to this position-wise approach using SNPs is that complex variation would be overlooked. As exemplified by *msp3.4* and *eba-175* above, this type of variation is found in some of the best candidates for host-parasite interaction studies.

4.5 Limitations

Consistent with the theme of Section I, SNPs ascertained by aligning short reads to the reference genome have tremendous utility, but also fail to represent important variation. Complex polymorphism, like indels and large dimorphic regions, is overlooked in SNP-centric analyses, and even candidate single nucleotide variation is lost within these regions due to coverage filters. Alternative methods are needed to access this complex variation to complement SNP-based population genetic and candidate gene studies.

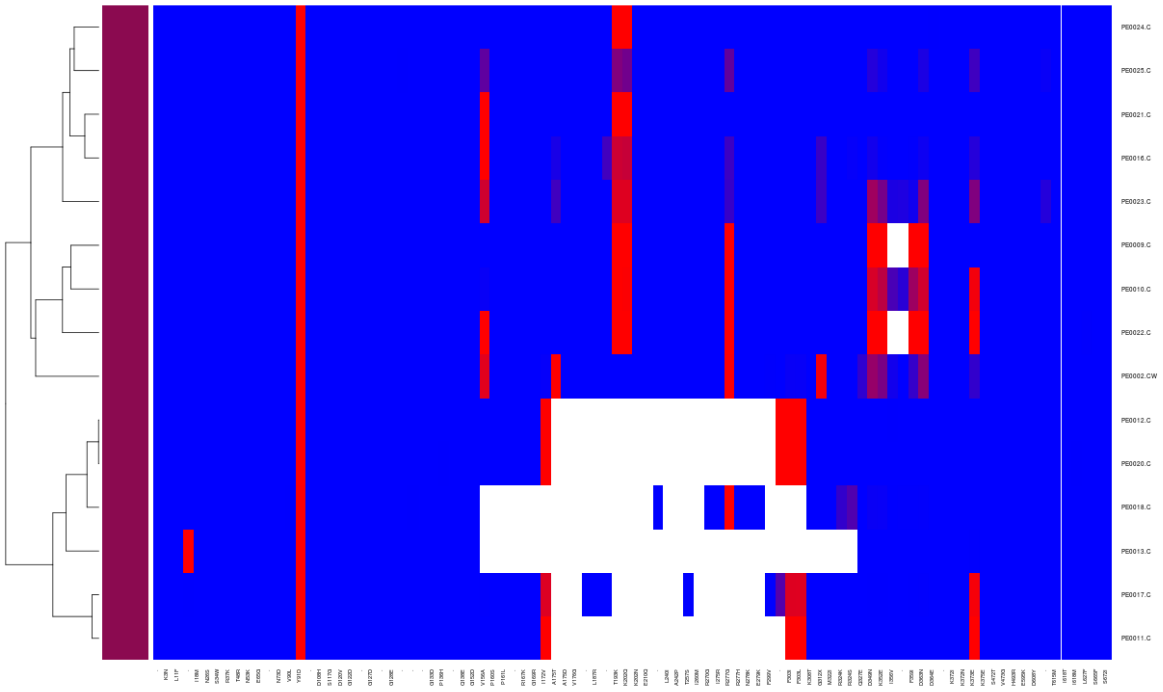
4.6 Acknowledgments and Contributions

My connection with the Tanzanian MOMS goes deeper than the samples I collected and sequenced for my DPhil. In 2001, while working as a research technician at Seattle

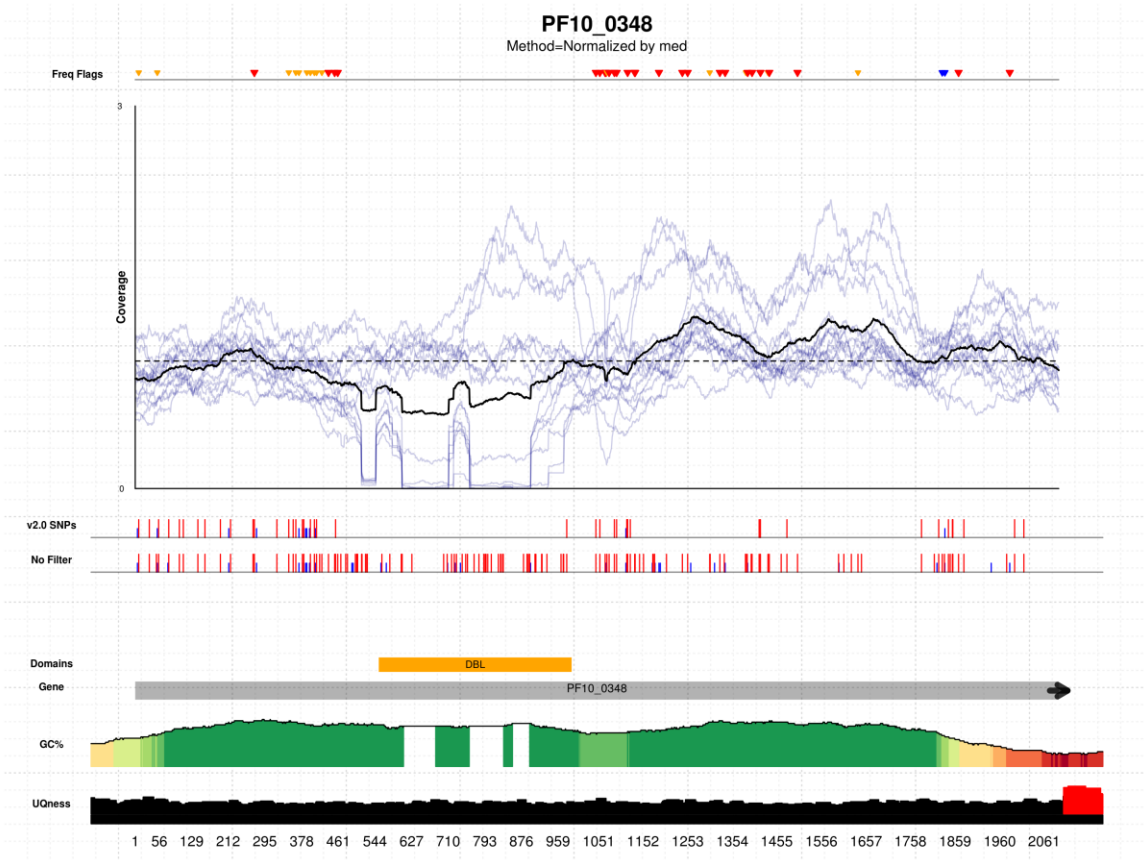
Biomedical Research Institute, I set up the study's computer network and clinical database when it launched in Muheza. Over the ensuing five years the database I designed grew to hold the information from thousands of mothers and hundreds of thousands clinical visits with their children. A few years later the MOMS project expanded 200km south to Morogoro, and I lived at the new site for a year under an independent NIH award as part of my MSc program. In Morogoro I extracted DNA and RNA from parasites, primarily for microarray studies comparing children with different clinical outcomes (see appendix).

The bulk of the culture adaptation work was performed by Katherine Williamson at Seattle Biomedical Research Institute.

4.7 Supplementary material



Supplementary figure 4-11. Haplotype plot for *msp3.4* (PF10_0348). Each row reflects a sample, and each column a potential MalariaGEN v2.0 SNP. The color of each cell indicates the allele frequency of the reference allele, ranging from red (0% like 3D7) to blue (100% like 3D7). The gradient of colors between these extremes represents mixed infections. White cells indicate missing data. The far left column adjacent to the dendrogram is colored by country—in this case all maroon for Tanzania. SNPs that are filtered out in the divergent region are left in so the missingness can be seen in this area.



Supplementary figure 4-12. Candidate gene plot for *msp3.4* (PF10_0348). Complete details describing all tracks and data sources can be found in methods section 2.2.1. The coverage track contains 16 blue lines representing the normalized sequence read coverage for each Tanzanian sample. The normalizing constant for each sample is the median (indicated in the title) coverage of SNPs in all genes for the given sample. The black line is the mean across all samples. Two SNP tracks appear just below the coverage plot. Tall red SNPs are non-synonymous while blue lines are synonymous. The domains track highlights the location of the Duffy binding-like (DBL) domain.

SECTION II: DEVELOPING TOOLS FOR ACCESSING COMPLEX VARIATION IN SHORT READ SEQUENCE DATA

The previous section presented two applications of single nucleotide polymorphism. I discussed how SNPs are ascertained by aligning short sequencing reads to the reference genome, and showed several examples of how this fails in genes containing complex variation. In this section I develop methods to access the sequence and genotypes in these complex regions. I start in chapter 5 by creating the software (Malign) that very directly answers the question, "if I know the complex variant I'm looking for, can I quickly detect if it's in my sequenced sample?" Malign does its job well, but I conclude that methods are still needed that return a full-length gene sequence without needing to *know* the potential variants. My first attempt at this, also in chapter 5, patches variants into the reference sequence, and I determine this has major drawbacks. To overcome these drawbacks, in chapter 6 I develop the *de novo* assembly algorithm, MalMOI. I apply this software to the difficult genes in Section I, and present the benefits and limitations of this approach.

5 ACCESSING COMPLEX SEQUENCE VARIATION

5.1 Overview

In two sections, this chapter describes two different approaches for accessing complex variation. The first approach (Malign) is alignment-based, and was designed for quick discovery of known variants in a BAM or fastq file (i.e., a sequenced sample). This program was designed with *eba-175* in mind. Thousands of samples have already been processed through the MalariaGEN sequencing pipeline, but the F/C indels in *eba-175* were not typed in any of them. Using a targeted approach, Malign genotyped the F/C indels in 3200 samples in a few hours on a Linux cluster. As covered in more detail below, Malign had several limitations that motivated exploring *de novo* assembly methods targeted at specific genes.

The second approach attempts to leverage the reference-free genotyping software called Cortex to overcome these limitations [271]. While the approach taken to *de novo* assemble full-length genes fell short, Cortex proves to be a powerful tool for accessing complex variation, and this section motivates the next chapter on *de novo* assembly of targeted genes from mixed infections.

5.2 Malign

5.2.1 Abstract

More than a petabase of open access sequence data is available to researchers in the Short Read Archive, but the utility of this data is limited for targeted meta-analyses of genes that diverge from reference genomes. Malign allows researchers to detect if regions of a custom reference, for example a novel dimorphic gene form or an indel, are represented in BAM or fastq inputs. The user is provided with a visualization of coverage depth and heterozygosity

across their region of interest, and with a table summarizing segments that reached their coverage threshold for detection. I demonstrate the value of this tool on an indel in a human gene associated with fibrinolysis-related diseases, and more in-depth on a dimorphic vaccine candidate in the malaria parasite, *Plasmodium falciparum*. I have validated this method using 1000 genomes data and PCR/qPCR results from malaria field samples introduced in chapter 4. Further, I exhaustively explored Malign's sensitivity and error rates using *in silico* mixtures of sequence reads, and went on to confirm these results by sequencing *in vitro* mixtures of parasite DNA to simulate multiplicity of infection (MOI). Malign is lightweight and useful even for low coverage samples, for example for pathogen detection in specimens dominated by host DNA.

5.2.2 Introduction

Large-scale sequencing pipelines typically align short reads to a reference genome, limiting the ascertainable variation in divergent genes. Investigators seeking to utilize these data are often focused much more narrowly—e.g., on the nuances of one or two divergent genes—and require *ad hoc* bioinformatics support to access this genetic variation. Excellent software exists for genome-wide *de novo* assembly and variant discovery [271], however there is great value in tools that scientists closer to the research problem can use to nimbly detect and visualize complex variation in their candidate genes. Malign is a straightforward tool that takes sequences of target genes as input in familiar fasta format, named with coordinates of interest, and provides the user with a coverage plot and detection call. Conveniently, sequence data can be input from fastq or BAM files, with the option of including unmapped reads in the latter case. Likely Malign users will be those who work in institutions with common genomics tools installed locally, but who may not have experience using these tools, or access to a bioinformatician who can build an improvised analysis pipeline for them. I present two examples here, applying Malign to both human and malaria parasite genomic data, with a special focus on a *Plasmodium falciparum* dimorphic vaccine candidate for validation.

5.2.2.1 Plasmodium falciparum example

Erythrocyte binding antigen 175 (EBA-175) is an important *Plasmodium falciparum* invasion ligand and candidate for a blood stage malaria vaccine [272,273]. This protein circulates in every population (see Figure 7-6) in two major dimorphic forms (F and C) that have been perennially investigated for associations with immunity, host genotype, and disease outcome [175,274]. As shown in Figure 5-1, the F/C indels are in two different

positions, and parasites appear to always be one form or the other. Population-scale investigations of the F/C variant would be informative for immuno-epidemiological studies and vaccine trials, and the ability to do so using modern sequencing technologies would expand the sample size available for such endeavors. A nested PCR reaction is most commonly used to classify infected blood samples as containing parasites with the F-insert, C-insert, or a mixture of both (in separate strains) [275]. Thousands of deep sequenced P. *falciparum* samples are available in the public domain, none of which have DNA readily available for traditional testing, and this variant is missed in standard pipelines [86,87].

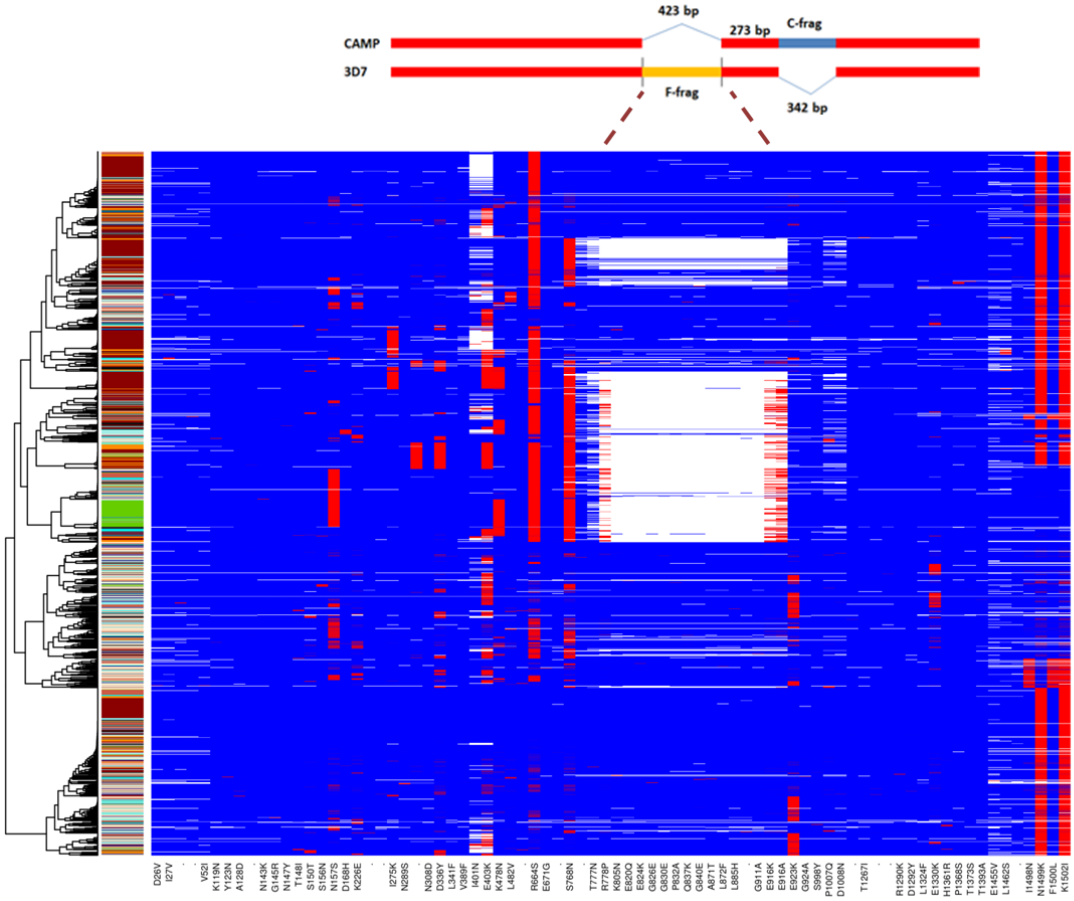


Figure 5-1. Global haplotypes and F/C indel schematic for *eba-175*. The top schematic depicts the approximate positions of the F and C indels in both types of parasite. The dotted lines show that the F region of 3D7 is where many field isolate have missing data when aligned to the reference. These parasites have the C-fragment, but lack the F-fragment. In the haplotype plot, each row represents a sample, and each column a MalariaGEN v2.0 SNP. The color of each cell indicates the allele frequency of the 3D7 allele, ranging from red (0% like 3D7) to blue (100% like 3D7). The gradient of colors between these extremes represents mixed infections. White cells indicate missing data. The far left column adjacent to the dendrogram is colored by country—samples are sorted by similarity of haplotype.

5.2.2.2 Human example

Tissue-type plasminogen activator (TPA) is a human serine protease important in tissue repair and development [276]. Intron 8 of TPA harbors an Alu sequence that is also an indel (I/D) variant associated with several disorders, including multiple sclerosis, bacterial osteomyelitis, and myocardial infarction [277,278,279]. Further, it is frequently used for studies of population ancestry [280]. To demonstrate the applicability of Malign in another organism, I applied this tool to several BAM files from the 1000 genomes project for which the donor TPA genotype was known. It would be unfeasible for many scientists to perform a similar investigation of a novel complex variant without substantial bioinformatics support.

5.2.3 Materials and methods

Other materials and methods are in section 2.5. Those listed below help describe some of the technical details in the results, so are included here for convenience.

5.2.3.1 Implementation

The two main inputs to Malign are the user's custom reference sequence and the paired-end short reads. The custom reference is simply a fasta file containing sequences representing the different versions (e.g., alleles) of a gene to be tested. The name of each sequence in the fasta file should contain 3 elements (ID, start position, and stop position) delimited by double underscores. The start and stop positions are coordinates within that sequence—perhaps an indel or a region that uniquely defines a dimorphic gene, for example. Malign takes fastq or BAM files as sequence input. BAM files are most ideal, as reads mapping to other parts of the genome can be ignored. Malign first extracts read pairs from BAMs for which either partner aligns to the original reference in a user defined region. Using this approach, read-pair insert sizes as low as 109bp fully recover the 342bp C-insert of *eba-175*, which is deleted from the 3D7 reference. However, if there is concern that large dimorphic regions will not be recovered, users can employ the -u option to also use unmapped reads from the BAM (or use the raw fastq). The software then infers the presence or absence of the indel or sequence of interest based on the coverage profile within the coordinates provided for each entry in the custom reference file. Development of this algorithm was guided using *in vitro* parasite mixtures to simulate multiplicity of infection. In addition to a table of genotype predictions, a visualization of the coverage profile of the sample's reads aligned to each gene form is produced, empowering the user to scrutinize parameter thresholds and to clearly see within-sample heterozygosity (Figure 5-2). Malign can be

launched with command line arguments, making it convenient for parallel processing, or run interactively in an R session.

5.2.3.2 *in silico* mixtures of parasites to simulate multiplicity of infection and estimate limits of detection

Mixtures were performed by two approaches. In the first approach, reads were randomly drawn from any of 10 fastq files from purely F-type parasites and 10 purely C-type parasite and combined at a particular F/C ratio and read depth. In other words, mixtures could contain reads from as many as 20 different parasites. This was repeated 200 times at each depth-ratio combination to estimate the false negative detection rate at those levels. Further, this was performed across a range of depths (approximately 11-117x) and MOIs (1%-10% for the minor allele) for a total of 112000 random mixtures. Effective coverage of the insert of interest (the one in minor abundance) was estimated from these parameters (approximately 0.1x-12x).

To ensure that including so many parasites in each mixture was not introducing a bias, and to depict FNR explicitly as a function of MOI, I also performed 2200 pairwise mixtures (the second approach mentioned above). Two laboratory lines (3D7/Dd2) and 18 Tanzanian field isolates were combined into 10 F-C pairs based on genotyping of the F and C indels by PCR and qPCR. For each sample pair, read pairs were randomly selected from original fastq files to produce a new, mixed set of fastq files at varying ratios. Twenty ratios were used, 10 in which the F-type parasite represented 1-10% (in steps of 1%) of the reads, and 10 in which the C-type parasite was represented in these proportions. At these 20 ratios, mixtures were randomly produced at each of 11 coverage depths—i.e., 220 mixtures were produced for each of the 10 F-C sample pairs (2200 mixtures in total). The two mixture approaches provided consistent results (Supplementary figure 5-13 and Supplementary figure 5-14).

5.2.3.3 *in vitro* mixtures of parasites to simulate multiplicity of infection and validate the *in silico* mixture approach

DNA was extracted from four laboratory parasite lines with known F and C genotypes, mixed at varying ratios to simulate MOI in natural infections, and submitted to the MalariaGEN pipeline for Illumina sequencing (Table 5-1) [87].

Table 5-1. DNA percentages for *in vitro* parasite mixtures.

		Parasite % (F/C type)				
sample	**ENA Accession**	**3D7 (F)**	**Dd2 (C)**	**HB3 (C)**	**7G8 (C)**	**Malign call**
PG0218_C	NA	100	0	0	0	F
PG0389_C	ERS319116	90	10	0	0	Mix
PG0390_C	ERS319119	80	20	0	0	Mix
PG0391_C	ERS319122	67	33	0	0	Mix
PG0392_C	ERS319125	33	67	0	0	Mix
PG0393_C	ERS319128	20	80	0	0	Mix
PG0394_C	ERS319130	10	90	0	0	Mix
PG0190_C	NA	0	100	0	0	C
PG0395_C	ERS319132	0	0	100	0	C
PG0396_C	ERS319134	0	0	99	1	C
PG0397_C	ERS319136	0	0	95	5	C
PG0398_C	ERS319138	0	0	90	10	C
PG0399_C	ERS319140	0	0	85	15	C
PG0400_C	ERS319142	0	0	80	20	C
PG0401_C	ERS319117	0	0	75	25	C
PG0402_C	ERS319120	0	0	70	30	C
PG0403_C	ERS319123	0	0	60	40	C
PG0404_C	ERS319126	0	0	50	50	C
PG0405_C	ERS319129	0	0	40	60	C
PG0406_C	ERS319131	0	0	30	70	C
PG0407_C	ERS319133	0	0	25	75	C
PG0408_C	ERS319135	0	0	20	80	C
PG0409_C	ERS319137	0	0	15	85	C
PG0410_C	ERS319139	0	0	10	90	C
PG0411_C	ERS319141	0	0	5	95	C
PG0412_C	ERS319143	0	0	1	99	C
PG0049_CW2	NA	0	0	0	100	C
PG0413_C	ERS319121	0	33.3	33.3	33.3	C
PG0414_C	ERS319124	0	25	25	50	C
PG0415_C	ERS319127	0	14.3	14.3	71.4	C

5.2.3.4 Artificial fastq generation

To get a sense of the impact that the size of the indel being tested has on Malign's error rate, artificial fastq files were generated from reference sequences in which I manually introduced deletions of various lengths. The artificial reads were generated with base quality scores from a real sample using the software ArtificialFastqGenerator [281]. For each of the 13 indel sizes listed in Figure 5-3, a high coverage fastq pair was generated based on a modified EBA-175 exon 1 sequence (i.e., I manually deleted part of this gene and used it to generate reads *in silico*). As these artificial samples were derived from a source with a defined faux deletion, I used them to estimate false positive rates (FPR). For consistency, I similarly generated artificial reads based on an unmodified EBA-175 sequence and subsampled these to estimate false negative rates (FNR) in the same indels created above. A Malign window size of 0 was used for the 1bp deletion, a window of 1 was used for

indels between 3 and 10bp, and a window of 10% of the indel size (up to 10bp, maximum) was used otherwise (Malign defaults).

5.2.4 Results

5.2.4.1 Sensitivity estimation: *in silico* parasite mixtures

To better understand Malign's limits of detection at various coverage depths, MOI proportions, and variant sizes, I performed 114200 *in silico* F/C mixtures of up to 20 parasite samples that were confirmed by PCR and qPCR to contain purely F or C genotypes, and 66000 sub-samplings of artificial reads generated from references with manual deletions. Testing on the F, C, and 13 artificial indels, these results indicate that false negative detection rates (FNR) approach zero at coverage depths of approximately 4x or greater for any indel size. For example, one might attain this FNR in a field sample with 150x coverage and in which the low abundance parasite represented 2.7% of the reads (Figure 5-2-D). Or similarly, an infected sample where host material overwhelmed the pathogen.

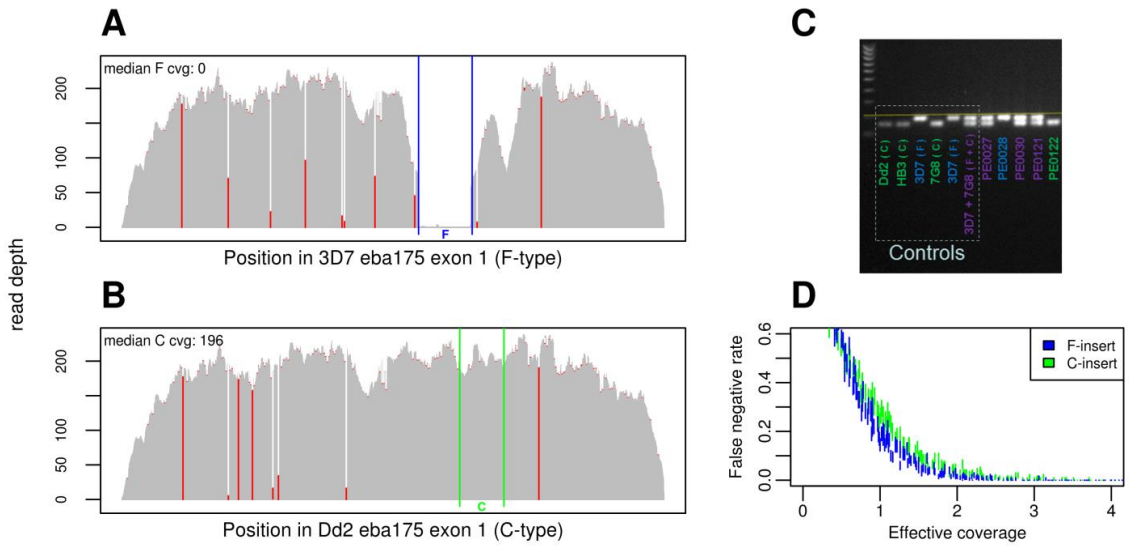


Figure 5-2. Malign sample output, validation, and error rate. A) Pileup of parasite sequence reads from a 6 year old Tanzanian child with moderately severe malaria (sample PE0122, for reference to the gel) onto *eba-175* from the 3D7 reference genome. Gray coverage bars indicate nucleotides that match the alignment sequence, and red bars are mismatches (highlighting MOI heterozygosity). Blue vertical lines mark the F-insert boundaries. Median coverage for the insert region is printed in the top left corner. B) Pileup of the reads from the same sample, but now aligned to the Dd2 reference, which contains the C-insertion. C) PCR gel from a multiplexed reaction with primers for products within the C (157bp band) and F (190bp band) inserts. Six controls are inside the dotted box. Sample PE0122 is in the far right lane and only indicates a C-band (as predicted by Malign). D) False negative detection rates of the F- and C-indels as a function of coverage, estimated from *in silico* mixtures.

5.2.4.2 Validation 1: *in vitro* parasite mixtures

To help ensure that *in silico* mixtures provide fidelity to natural MOI, I also mixed DNA of different parasites at various ratios and sequenced them at high coverage. Using sequence data from *in vitro* mixtures of laboratory line parasites with only the C allele (Dd2, 7G8, and HB3), with a window size of 10 and coverage threshold of 1, Malign yielded no false positive F-insert predictions in 23 samples. Further, no false negatives in six mixtures of F and C parasites were detected (Table 5-1). In the single lab line that was purely F-type, there was no false positive detection of the C-insert.

5.2.4.3 Validation 2: PCR/qPCR on Tanzanian field isolates

I went on to validate Malign in 40 sequenced Tanzanian field samples, on which I also performed PCR and qPCR melt-curve assays to genotype the F/C indels (Figure 5-2-C). Applying Malign with default settings to these samples, and considering PCR/qPCR the gold standard, there was one false positive and one false negative C-allele, and no errors for the F-allele. The melt-curve peak for the C-insert in the false negative sample is lower than average, and there is some evidence of very low coverage in the pileup, indicating this allele fell below the sensitivity of Malign with an effective coverage less than 3-4x (Supplementary figure 5-15). For the false positive result, there is clearly coverage in the pileup plot in the C region for this sample (Supplementary figure 5-16). The PCR should be repeated for this sample, perhaps with more cycles.

5.2.4.4 Estimating error rates in relation to indel size using artificially generated fastq files

These error rates are strikingly concordant with those estimated using *in silico* mixtures of real samples. As might be expected, FNRs are generally higher for larger variants, as the likelihood of coverage gaps is increased, however beyond coverages of 4-5x the estimated FNR drops to zero for any deletion size. It should be cautioned that different genes will have their own peculiarities. EBA-175 exon 1 is AT-rich (70%), and contains no repeats or long homopolymer stretches, thus may serve as a benchmark for similar genes.

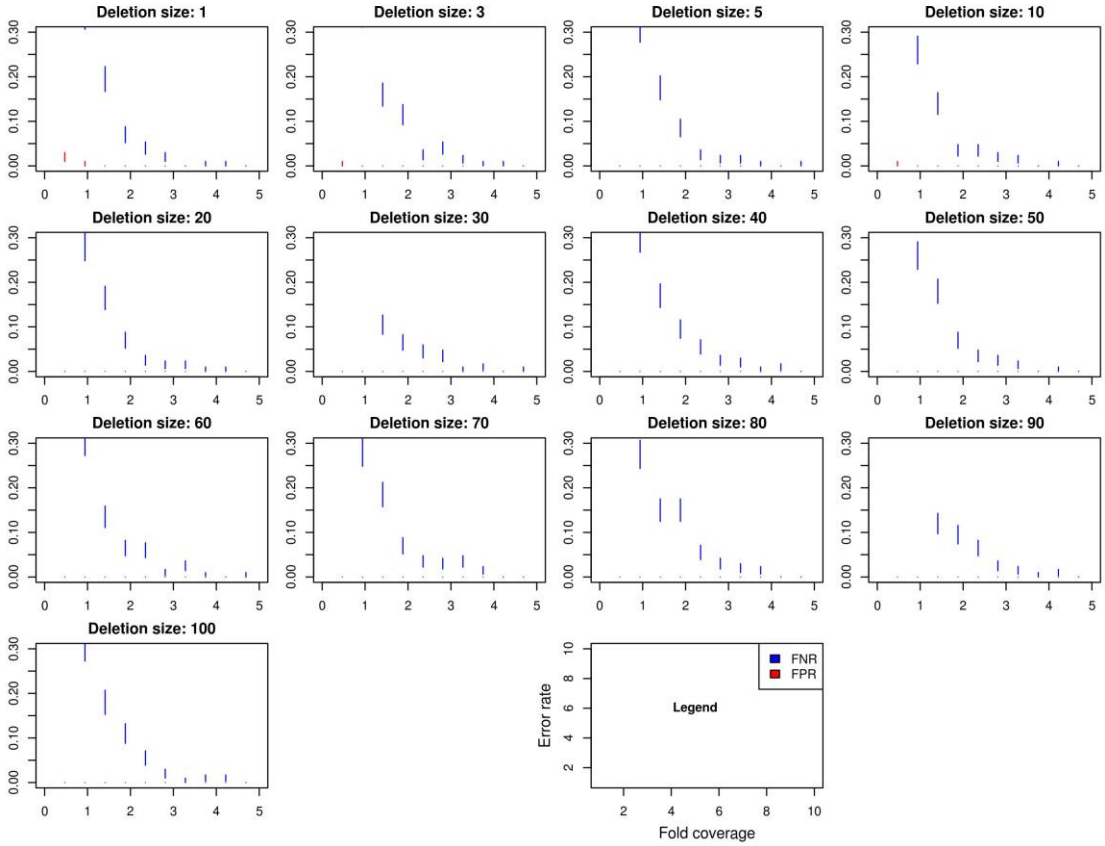


Figure 5-3. Malign error rates as a function of insert size and read depth. Each panel depicts the false positive (FPR) and false negative (FNR) rates for a specific indel size. For a given indel size (i.e., panel) and coverage depth, Malign was run on 400 subsamples taken from artificially generated fastq files. Half of the subsamples were generated from a 3D7 EBA-175 sequence in which a deletion was manually introduced (to test false positive detection). The other half were taken from fastq files artificially generated in a similar fashion, but were based on a normal 3D7 EBA-175 gene (to test false negative detection). Each short vertical line is the symmetric standard error around the proportion of false positives (red, only really apparent in panel 1) or false negatives (blue) in 200 tests.

5.2.4.5 Genotyping the Alu indel in intron h of the human tissue plasminogen activator (TPA)

I identified three individuals in the 1000 genomes phase 3 release with different inferred TPA Alu I/D genotypes: homozygous insertion (INS|INS), homozygous deletion (T|T), and a heterozygote (T|INS) [282]. The depth of coverage in these samples was 8-10x, and Malign accurately categorized the indel as present/absent in all cases using BAM files as input (Figure 5-4). The coverage profiles are consistent with the copy number of the insert as well—e.g., the heterozygote has approximately half the coverage depth in the region of the indel.

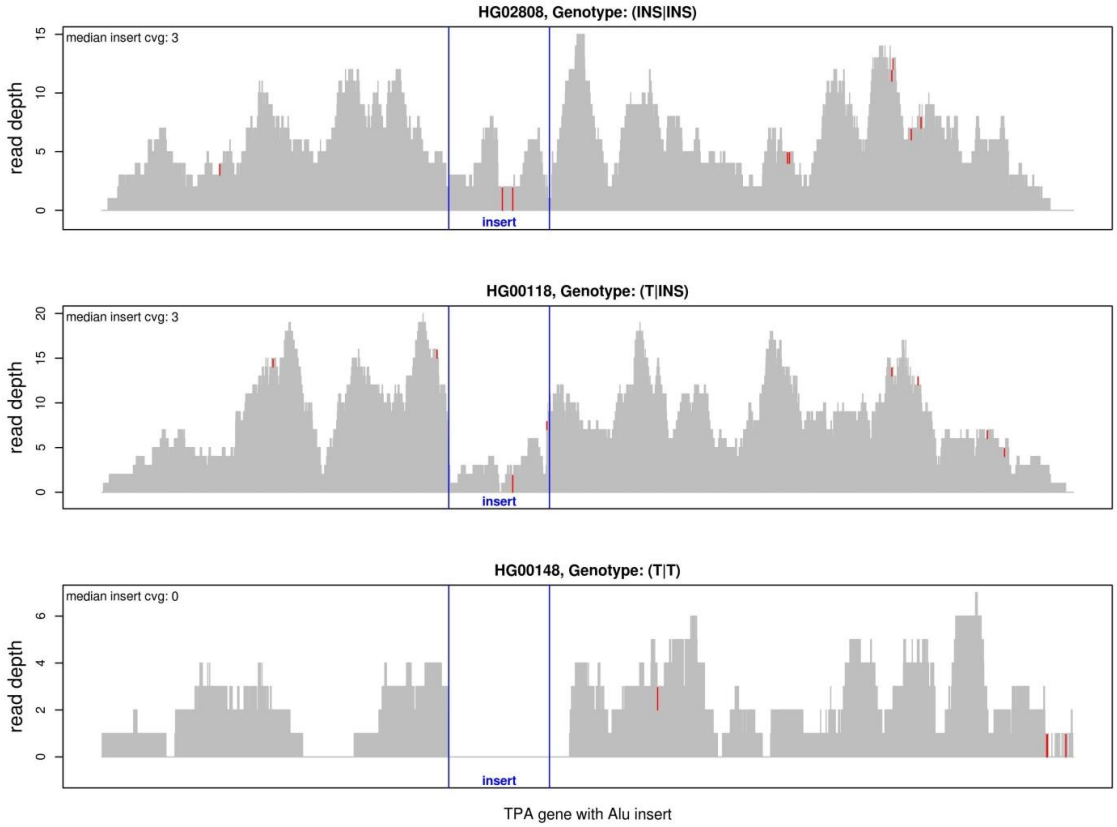


Figure 5-4. Malign output for the TPA Alu indel in three samples obtained from the 1000 Genomes (1kG) project. Sample ID and 1kG genotype calls are printed in the title of each plot: homozygous insertion (INS|INS), homozygous deletion (T|T), and heterozygote (T|INS). All samples were aligned to the same reference TPA sequence that included the Alu insert (blue vertical lines). All Malign calls match those from 1kG. As expected, the read depth within the Alu insert is approximately half of that for the rest of the region in the heterozygote (middle panel).

5.2.5 Discussion

Malign is a useful tool for scientists to quickly detect divergent genes or indels in paired-end, short read sequencing data. An important contribution of this report for malaria researchers is that I have validated this approach using *in vitro* and *in silico* parasite mixtures to simulate MOI, and with field samples genotyped by gold-standard methods. It is common for patients to be infected with multiple parasite strains, and a tool aimed at variant detection for associations or vaccine efficacy stratification would need to account for this. Further, I have demonstrated the broader utility of Malign by accurately detecting a clinically relevant indel in human samples. As Malign is designed for targeted analyses, run-times are orders of magnitude lower in time and memory requirements than whole-genome solutions. Analysis of the human TPA example runs in under an hour on a whole-genome

BAM file, and on the order of a few minutes using just chromosome 8. Similar *P. falciparum* analyses take seconds. This tool is straightforward to use and has high sensitivity with low error rates over a range of indel sizes. Malign will enable a broader group of investigators to perform *ad hoc* analyses of divergent variation in both raw and processed sequence data.

5.3 Cortex

5.3.1 Motivation

The previous section described a tool (Malign) that was designed to detect the presence or absence of a known indel variant. While this is a step forward in our ability to access complex variation, it is limited to variants for which a version of the gene is available to use as a reference for the alignment. There would be tremendous value in a tool that generated full-length genes by assembling sequence reads without relying on the reference. As described in more detail below, Cortex performs *de novo* assembly as part of its variant discovery algorithm, but these assemblies are usually shorter than full-length genes, and are not part of the standard output. Cortex is designed to yield a list of variants. In this chapter I describe a wrapper program that I developed in Perl for generating full-length "meta genes" by patching Cortex variants into the 3D7 reference version of the gene. While this approach perhaps has some utility, it also has significant drawbacks. Here I detail this methodology, some results and limitations, and conclude that an alternative approach for *de novo* assembly of *P. falciparum* genes should be developed. I should emphasize that Cortex is reliable and does its job very well. It was designed to discover and genotype variation, rather than to deliver long assemblies of target genes.

5.3.2 Introduction

In the *de novo* assembly world there are two broad classes of algorithms: overlap-layout-consensus (OLC) approaches and those that are graph-based. Cortex is a software package that utilizes de Bruijn graphs and is particularly well suited for complex variant discovery in samples containing a mixture of parasite genomes [271]. A de Bruijn graph is a mathematical structure that in this context is used to represent the overlap of all sequence reads (actually sub-reads called kmers) with all other reads in a sample [283]. Once such a graph is generated from sequence data, a particular path through the graph represents a contiguous stretch of reads—i.e., an assembly. Sometimes there is more than one path through the graph—for example, in a mixed sample the path would branch where two

parasites were different and then re-converge where the sequences were again the same. The reads from more than one sequenced sample can be loaded into the same graph and thought of as being different colors. Assembled regions in a de Bruijn graph that diverge are represented as bubbles (depicted in step 1 of Figure 5-5), which are typically collapsed and removed as errors in other software (for example, the Velvet assembler discussed later). The aim of Velvet is to find a unique path through the graph, whereas the point of Cortex is to focus explicitly on the bubbles. Cortex recognizes bubbles as potential variants, and will output the read counts covering each side of the bubble (each allele), even if present in the same sample. By way of analogy, a human sample that is heterozygous at a particular site would produce a bubble, and one would expect half of the reads to represent each side of the bubble—i.e., each allele. Although blood stage parasites are haploid, mixtures of strains will produce bubbles at variant positions, and rather than a 50/50 distribution of reads for each allele, the read counts will be proportional to the abundance of the different parasites in the infection at the time of the blood draw or tissue culture harvesting.

That Cortex can combine sequenced field samples into a graph along with the reference genome as separate colors, and that it can handle within-sample heterozygosity in the graph structure, made it a rational place to start when conceiving methods for *de novo* assembly of divergent genes. As mentioned previously, the assembled fragments produced by Cortex do not typically span entire genes and are not part of the standard output. One reason for the shorter fragments is that Cortex is run on a genome-scale, which makes it computationally infeasible to iterate over many parameter combinations (like kmer size) to optimize for different genomic regions. As a workaround, an approach was devised that patches Cortex variants discovered as bubbles between a field sample and the reference genome into the reference gene (Figure 5-5).

As the variants patched into the reference are dependent on the ascertainment power and accuracy of Cortex, these sequences are referred to as meta-genes. An immediate concern is that if a polymorphism is not detected by Cortex, rather than failing in some manner this method will simply default to the 3D7 sequence. Indeed, as described in the results section this concern was realized and led to the abandonment of the reference-patching approach early on.

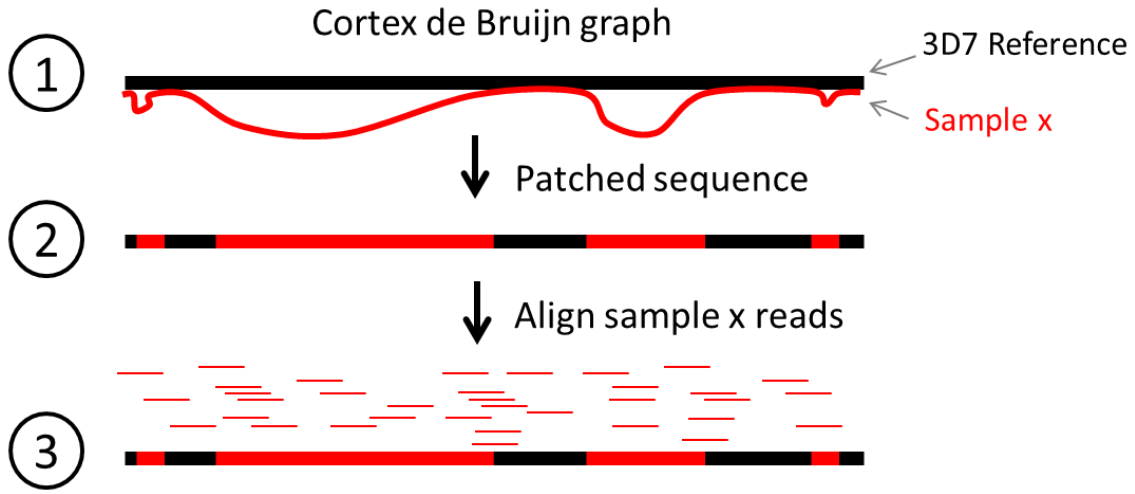


Figure 5-5. Schematic of the Cortex-reference patch algorithm. 1) Cortex de Bruijn graph is built with the sample of interest (red line) combined with the 3D7 reference (black line). This produces a list of variants with reference-based coordinates. The colored lines are adjacent where the two genomes are identical, and bubbles are formed where they diverge. The small bubbles might represent SNPs, and the large ones indels or other complex variants. **2)** The variants are substituted into the reference sequence to create a meta-gene. **3)** To assess how well the meta-gene represents reality, the original reads are aligned to it.

Although the reference-patch strategy adds the 3D7 sequence to the graph, it is nonetheless truly *de novo* with regard to variant discovery. The reference sequence is the one being patched into, so adding it to the graph is simply to yield the exact coordinates where the 3D7 sequence should be excised and replaced with alternative sequence. In fact, the appeal of using Cortex variants versus those from MalariaGEN is that larger structural variants could be ascertained because it doesn't rely on reference alignments.

5.3.3 Results

5.3.3.1 Cortex can detect structural variants

A demonstration of Cortex's ability to detect structural variation is shown in Figure 5-6, where a coverage artefact is clearly seen in one of the two samples aligned to the 3D7 version of *eba-175*. Cortex predicted a 6bp deletion in the sample displaying the artefact, and its precise location, but not in the sample with a normal coverage profile.

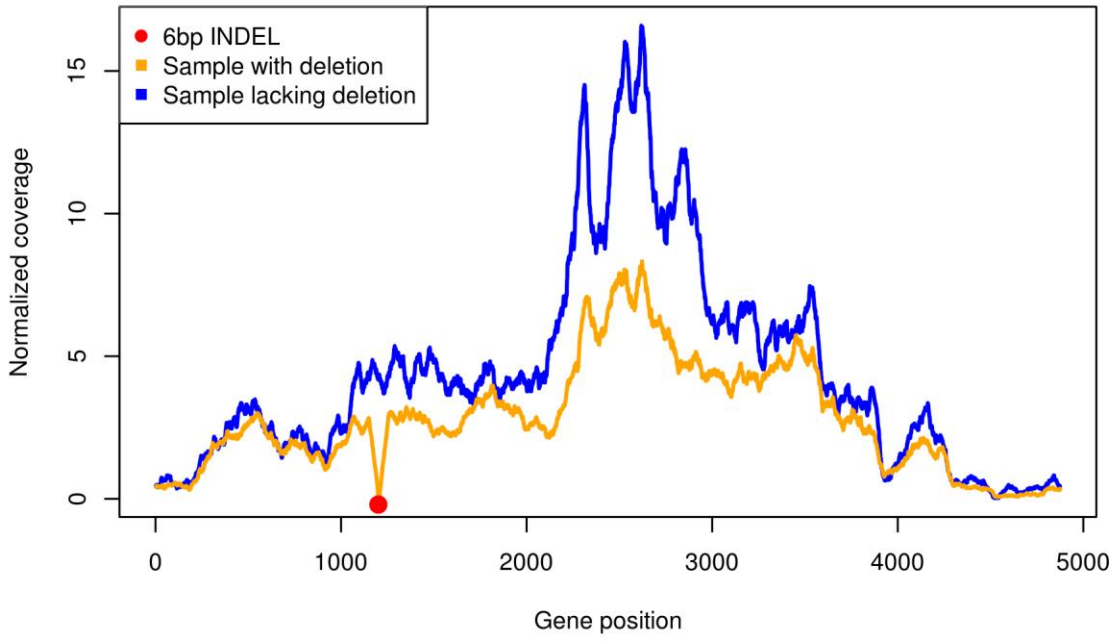


Figure 5-6. Cortex accurately discovers an indel in *eba-175*. The normalized read depths of two samples aligned to *eba-175* are plotted along the gene coordinates. The orange sample's coverage drops to 0 at position 1200. Cortex predicts a 6bp deletion as compared to the 3D7 sequence at position 1200 of the orange sample, but not in the blue sample.

5.3.3.2 Meta-assemblies

Meta-sequences were generated for 59 field and lab isolates using the Cortex variant patch approach for *eba-175*, *pfcrt*, and *var2csa*. As indicated in step three of Figure 5-5, to crudely assess how well the meta-genes represent the original sample, the reads are aligned to the patched meta-sequence. As shown in Figure 5-7 for *eba-175*, the reads aligned to the meta-gene do so without the large artefact as seen when aligned to the 3D7 version of the gene. Extensive detail about the structure of this gene is detailed in Section III, but briefly, *eba-175* has two large indels that are mutually exclusive (dubbed F and C, from the original description in FCR3 and Malayan Camp isolates). The 3D7 reference version of *eba-175* has an F insertion and C deletion, and vice versa for the HB3 lab-line. As shown in the Tanzanian parasites previously described, mapping paired-end reads of C-containing isolates like HB3 to the *eba-175* sequence of an F-parasite is problematic. The schematic in Figure 5-8 illustrates exactly what is happening to generate various errors and artefacts in such an alignment. The top panel shows real data in LookSeq—reads from a C-type parasite

(HB3) aligned to an F-type parasite (3D7), zoomed to the F-indel region of the reference genome [284]. A large gap appears because HB3 lacks reads representing the F-insert. Further, the read-pairs that span this gap have much longer than expected insert sizes. The bottom panel B shows several examples of read-pairs from a C-type parasite (e.g., a Camp strain), and how these reads would map to the reference genome. The read-pair in the center of the "Camp library" span the F-insert site and exhibit this artefact.

As seen in the top panel of Figure 5-7, when the HB3 reads are aligned to the 3D7 version of *eba-175* a large gap appears in the F-segment region of 3D7 (for which HB3 has no representative reads). Adjacent to this region most reads are highlighted as having mapping errors, which adds to the risk of assigning false positive SNPs. Nearly every read on both sides of the gap are flagged as having longer than expected insert sizes, as the mate-pairs are mapping with the additional length of the F-insert (423bp) apart. Also note that several SNPs appear in the top alignment, illustrated by the red, orange, and green vertical lines throughout the gene. The center panel (Figure 5-7-B) depicts HB3 reads aligned to the meta-gene produced by patching the Cortex HB3 variants into the 3D7 sequence. In contrast to the top panel, no SNPs or systematic artefacts appear in this alignment. Although a peppering of reads with mapping errors remain in the meta-gene alignment, qualitatively these are no more common than what is found in a control 3D7 sample aligned to the reference genome (Figure 5-7-C). In summary, at a high level it would appear that the meta-gene for *eba-175* provides a much better representation than the reference alignment.

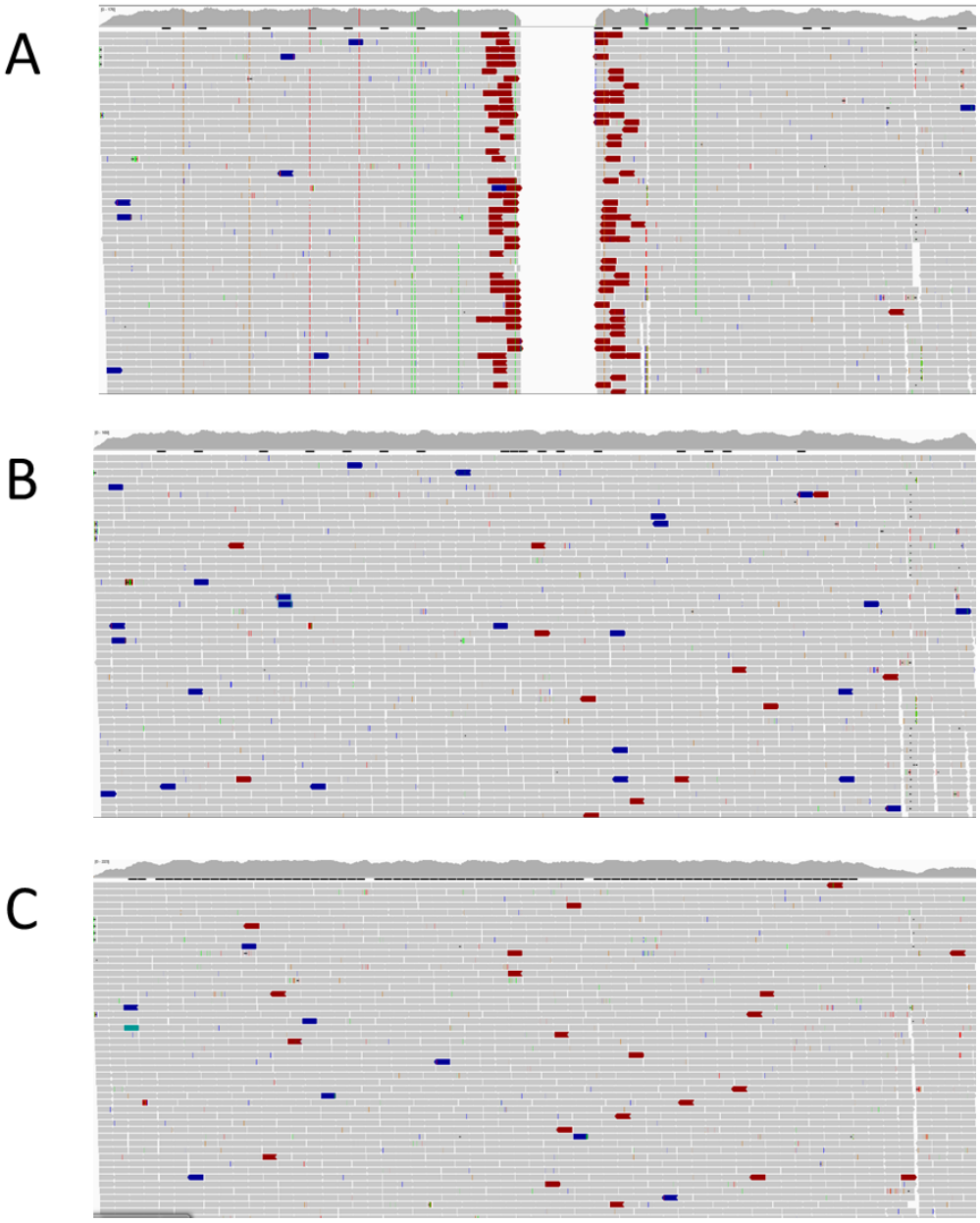


Figure 5-7. IGV pileup of HB3 reads onto 3D7 *eba-175* vs. the meta-gene. The gray, blue, and red horizontal bars peppered throughout the plot represent individual reads. Most reads are gray, indicating no alignment errors. Blue reads have shorter than expected distances to their mate (i.e., insert size), and longer than expected distances for red ones. The long, thin orange, red and green vertical lines indicate SNPs (colored by nucleotide). The top of each panel has a track showing the general trend in coverage. **A)** HB3 reads aligned to the 3D7 version of *eba-175.* The large gap in the center of this panel occurs where 3D7 has a 423bp F insert that HB3 lacks. Notice that the HB3 sample reads on either side of this artefact are red because they are mapping 423bp farther apart than expected. **B)** The same HB3 reads aligned to the meta-gene created by patching HB3 variants discovered by Cortex into the 3D7 reference. The SNPs and large indel artefact have been corrected. There is a general scattering of red and blue reads, which based on panel C appear to be in a normal range of what is expected by chance. **C)** Reads from a sequenced 3D7 sample aligned to the 3D7 version of *eba-175*. Note that even in this control we see a number of reads with longer and shorter than expected insert lengths.

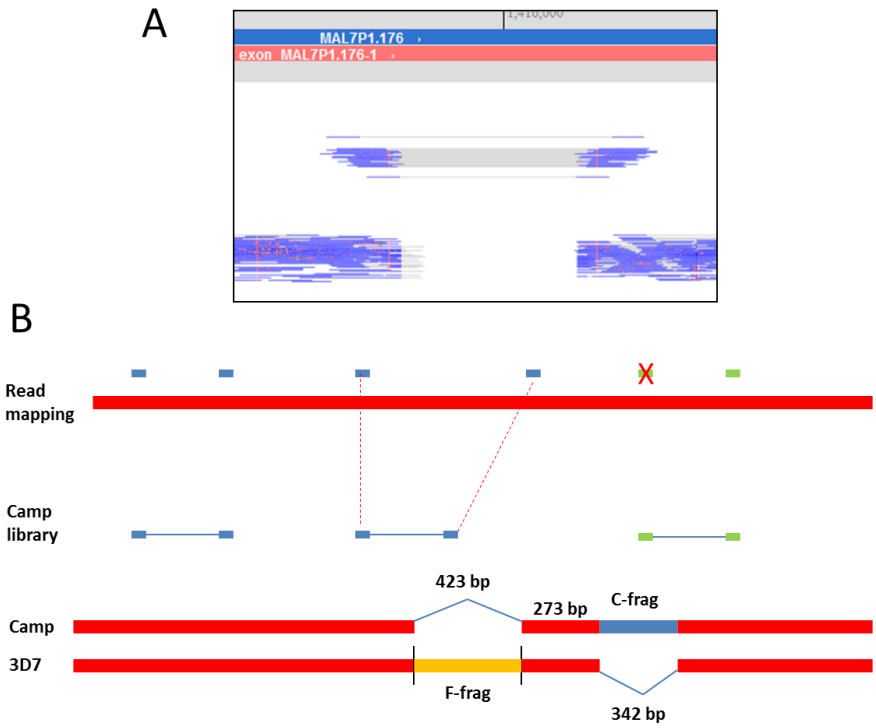


Figure 5-8. The impact of indels on read-pair mapping. A) LookSeq view of reads from a C-type parasite (HB3) mapping to 3D7 *eba-175* (F-type). Reads are colored blue and pairs are connected by a gray line. Pairs are sorted top to bottom by largest to smallest predicted insert size. The outlying reads above the gap are those for which the pairs were split by the F-insert. B) Schematic of reads from a C-type parasite mapping to the reference. The "Camp library" track shows 3 hypothetical read pairs. The read mapping track along the top shows where these reads would map to the 3D7 sequence. The leftmost pair map to the Camp and 3D7 gene in the same location. The middle pair spans the F-insert, and map with longer than expected insert size. One of the reads in the rightmost pair (green) maps within the C-insert (red 'X'), which is not present in 3D7, so the pair is lost.

A separate form of validation involves the use of genetic cross parents and progeny, and this approach was used to scrutinize the *pfcrt* meta-gene. Four experimental genetic crosses of *P. falciparum* have been performed, three of which have been published (strains 3D7xHB3, HB3xDd2, and 7G8xGB4) [285,286,287]. The parents and progeny from these crosses have been sequenced through a MalariaGEN partnership, and these data are used here for validation. The progeny from a genetic cross should match one of the two parents, or perhaps part of each in the off-chance that a recombination event occurred in the gene of interest. Substitutions in progeny that reflect neither parent are called Mendelian errors. Two Mendelian errors occur in separate meta-introns of the *pfcrt* parents and progeny of the HB3-Dd2 cross. No such errors occur in any of the 13 exons of *pfcrt*, and in exon 2, which contains the notorious chloroquine resistance conferring mutation, the expected

haplotypes were patched into the parental lines, and approximately half of the progeny reflected either parent (Figure 5-9).

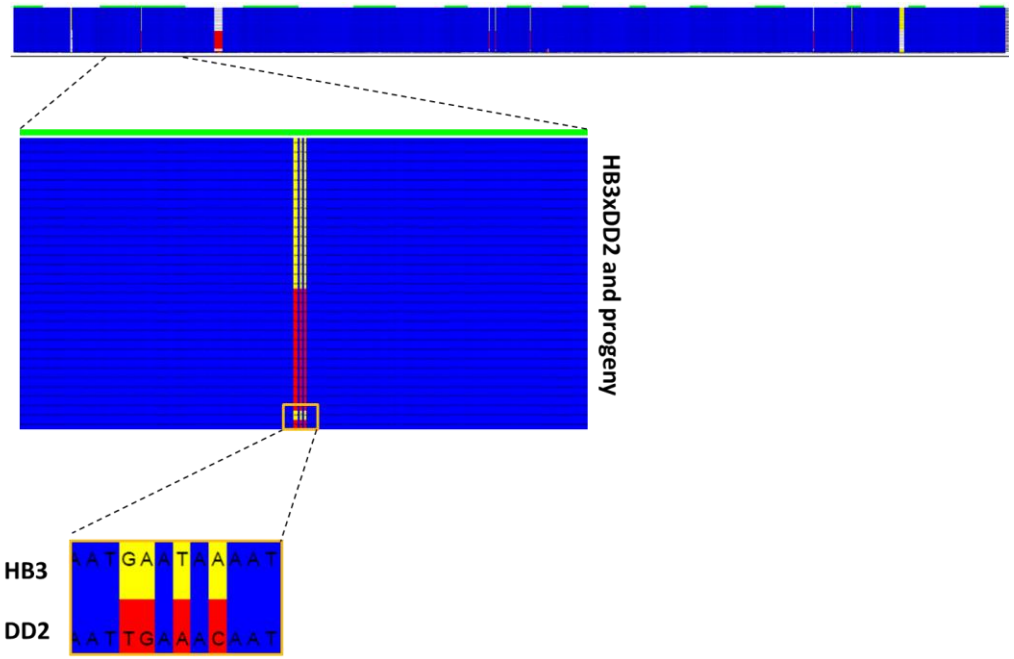


Figure 5-9. Comparing *pfcrt* meta-genes for HB3xDd2 parents and progeny. The top panel shows a bird's-eye view of an alignment of 31 meta-sequences produced by patching HB3 and Dd2 Cortex variants into the 3D7 version of *pfcrt*. Exons are indicated above the alignment with green bars. Blue positions are identical between HB3 and Dd2. Yellow positions are differences that match the HB3 parent, and red for those matching Dd2. The central panel zooms in on exon 2, which contains the canonical K76T chloroquine resistance locus. The HB3 and Dd2 parents are the bottom two sequences and match the expected genotypes, as shown in the bottom panel [288]. All progeny match one of the parents at this locus.

Taken together, the *eba-175* and *pfcrt* results bode fairly well for the reference-patch approach, however the lingering concern that missed variation would default to the 3D7 reference needed to be investigated further. The multiple sequence alignment in Figure 5-10 reveals this reality at position 1750. The 3D7 reference is aligned with the Dd2 patched meta-sequence, along with an assembly of overlapping Sanger sequenced amplicons and a Velvet assembly. At this position the 'A' allele did not pass the Cortex filters, and thus the 'C' allele remained in the meta-sequence. At this juncture there was really no point in pursuing the reference-patch approach further, as this issue would always haunt the meta-sequences, and few downstream analyses would pass major scrutiny.

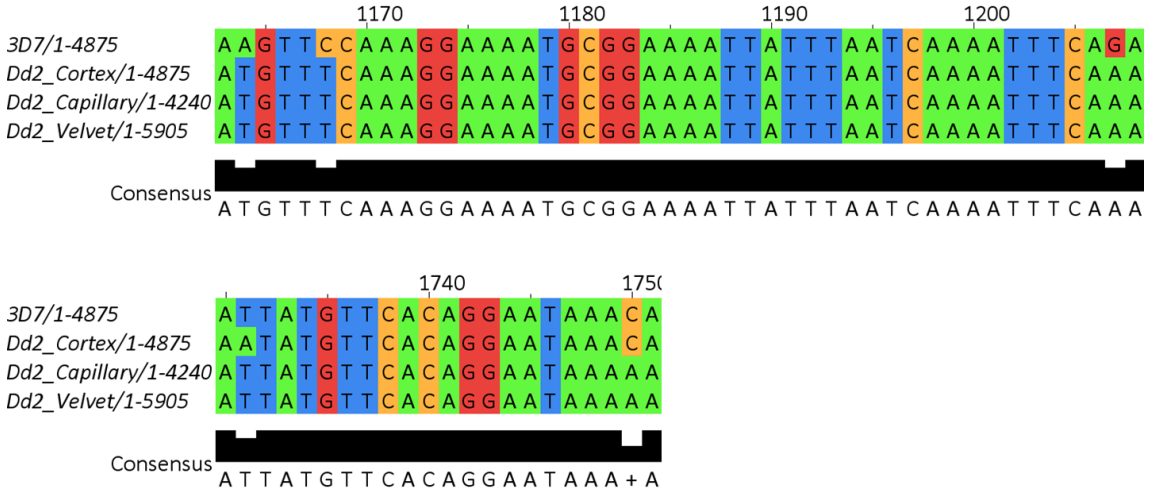


Figure 5-10. Meta-genes erroneously default to the 3D7 sequence. Two snippets of a multiple sequence alignment of *eba-175* sequences is shown in the top and bottom panels. The top sequence is the 3D7 reference from PlasmoDB. The bottom 3 sequences are of Dd2 parasites derived from Cortex variant patching, capillary sequencing of the same sample, and a *de novo* assembly using Velvet. In the top segment of the alignment all Dd2 sequences agree at the positions that vary from the reference. The bottom segment shows two SNPs for which the patching method has errors. At position 1731 Cortex had the wrong genotype. At position 1750 no variant passed in the Cortex VCF, and as expected in this scenario, an erroneous 3D7 allele appears in the meta-gene.

Fifty-nine *var2csa* meta-sequences were also generated by the reference-patching method. The hyper-variable and paralogous nature of *var* genes adds a level of difficulty for Cortex to build an accurate graph and filter true polymorphism from bubbles formed due to repeats and similarity to other genes. Using the most stringent variant filtering criteria, Cortex identifies 156 polymorphisms in 59 samples from 17 different countries, which is perhaps lower than expected in this 10,018bp gene. Given the concerns mentioned above of missed variants defaulting to the 3D7 sequence with the reference-patching method, meta-*var* genes would likely be particularly biased. Ironically this is a theme with *vars* that even affect the molecular biology. Criticisms of early *var* gene diversity studies that attempted to reverse transcribe *pfemp1* from field isolates were that the degenerate primers used would be biased toward known reference sequences [289]. For 100% of the 59 *var2csa* meta-sequences, the protein translation of exon 1 contained multiple stop codons, indicating either genotyping errors, or that some polymorphism is missing that would have kept the sequence in frame.

5.3.3.3 Other applications of Cortex to complex variation

Notwithstanding the issues described above with patching variants into the reference, Cortex remains an excellent resource for reference-free discovery of complex polymorphism

in *P. falciparum*. For example, Figure 5-11 revisits the *msp3.4* gene from Supplementary figure 4-11 that is highly divergent across large stretches, and is thus inaccessible using reference-based genotyping approaches. This apparent dimorphic complex variation is discussed in more detail in the next chapter, but first I show a brief application using Cortex. The haplotype plot shows long stretches of SNPs and small indels that appear to have identical within-sample, non-reference allele frequencies. This is because these variants actually comprise the same bubble in the de Bruijn graph. In other words, Cortex is yielding a phased haplotype. Unfortunately these haplotypes do not span the full length of the gene—i.e., full length composite assemblies of a mixture are not produced. Nonetheless, Cortex is identifying complex variation that is lost using reference-based methods, while also providing information about within-sample heterozygosity.

One might wonder if Figure 5-11 represents a scenario in which dimorphic classes of parasites are present in the same sample, or if in some parasites this gene is paralogous within the same genome. Cortex provides an opportunity to investigate this question. Defining the largest stretch of divergence as Class I (reference-like) and Class II, the within-sample ratio of these classes can be visualized using the Cortex output. As shown in Figure 5-12, all of the clonal lab-line samples fall exclusively into one class or the other. Further, the samples with evidence of MOI are mixed at a continuum of ratios. Although by no means conclusive, one might expect if the *msp3.4* dimorphism were paralogous that one of the lab-lines might contain both forms, and that the within-sample ratios of mixed samples would hover around 0.5.

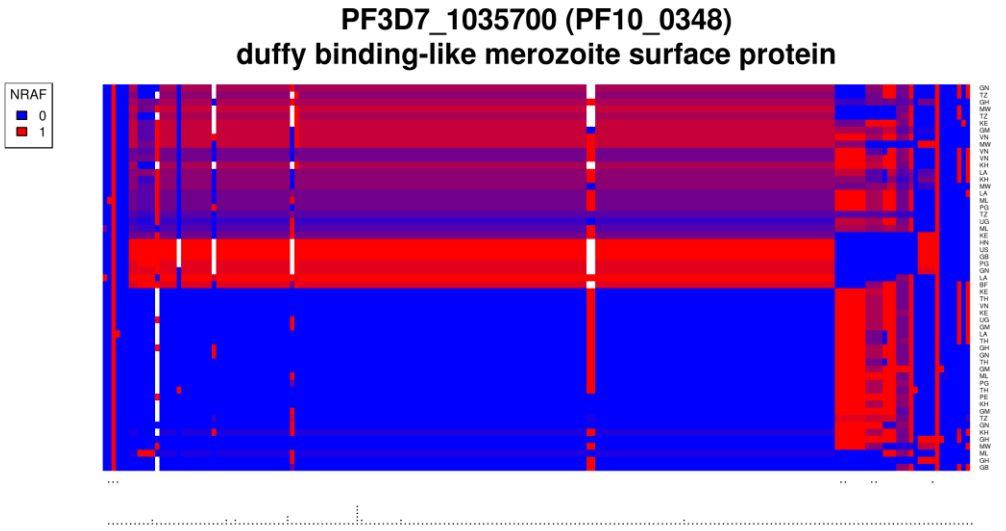


Figure 5-11. Haplotype heatmap of *msp3.4* based on Cortex variants. Each row represents a sample and each column is a variant detected by Cortex. Blue cells indicate positions matching the 3D7 reference genome, and red the alternate allele. Mixed infections are represented by blending of red and blue, proportional to the within-sample allele frequencies. Dots immediately below the colored plot show SNPs that are also in the MalariaGEN v2.0 database. The bottom row of dots indicates the size of the variant. Note the high number of variants missed by the MalariaGEN reference-based discovery. Also of interest are the long stretches of polymorphism that are in LD. Some regions appear to be dimorphic—i.e., parasites have one form or the other.

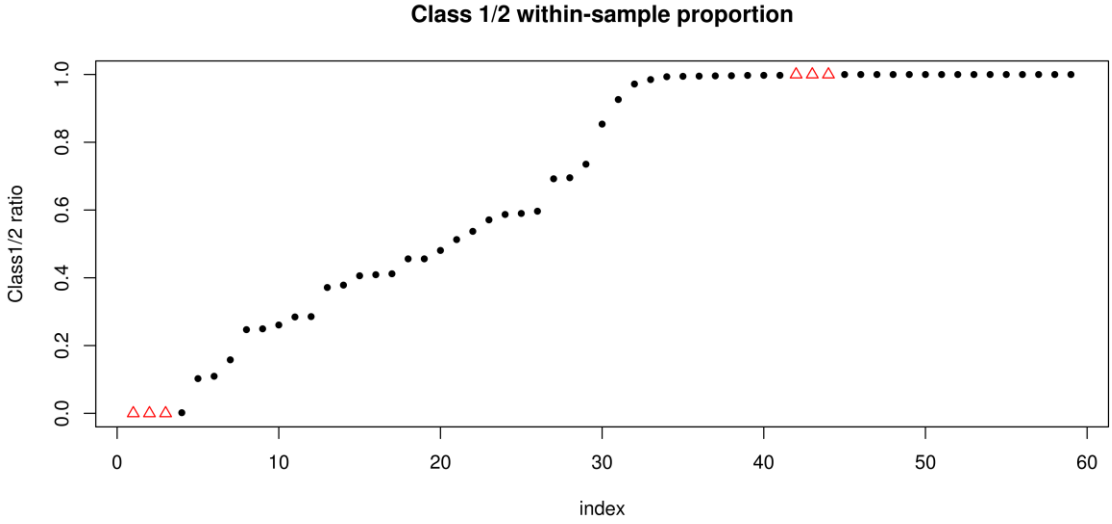


Figure 5-12. Within-sample ratio of *msp3.4* class I/II forms . Each dot or triangle represents a sample sorted by the ratio of Class I to Class II read depths within that sample. Clonal lab-lines are plotted as red triangles.

More work should be done to tune Cortex filters to the nuances of the *P. falciparum* genome and to define core genomic regions that can be reliably accessed. This could be done using genetic cross parents and progeny for genome-wide analyses.

5.3.4 Conclusions and discussion

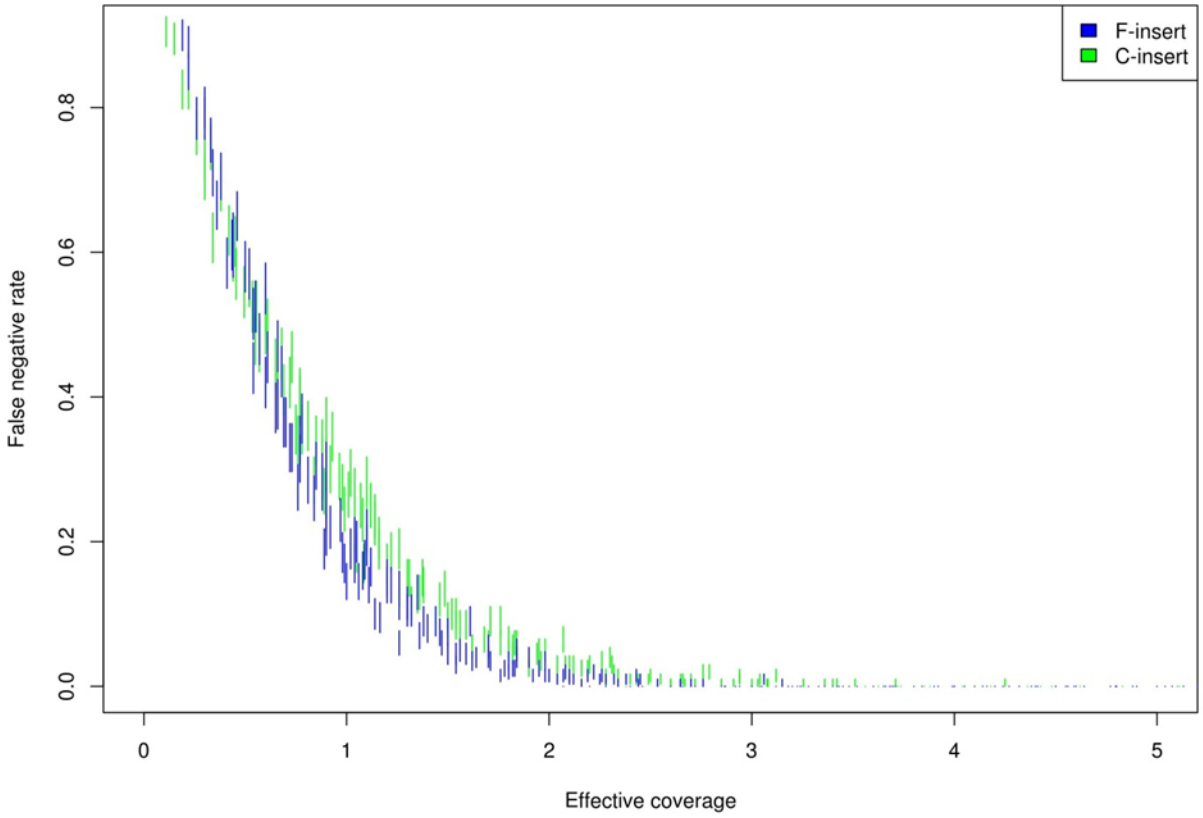
Based on the limitations described above, the approach of patching variants into the reference sequence is too problematic to develop further, and another approach should be taken to generate full-length gene sequences. Regardless of the variant discovery software used with this approach, the issue of defaulting to the reference sequence would always cause grave concern when interpreting downstream analyses and applications.

An early draw to Cortex was the theoretical possibility of phasing variants to deconvolute parasites in a mixed infection. Recall that within-sample heterozygosity is represented as bubbles in the de Bruijn graph, but since the different alleles come from the same sample both sides of the bubble would be the same color. If a gene has multiple bubbles, it is unclear which alleles among the many polymorphic sites should go together. This is a phasing problem. However, if the coverage depths of the alleles in each bubble corresponded to the abundance of the respective parasites in the sample, then this information could theoretically be used to phase the variants. In fact, it is precisely this approach that is taken by another graph-based algorithm called MetaVelvet to deconvolute different species in metagenomic samples [290]. The problem of phasing in mixed samples is one that I revisit multiple times going forward. In the next section I take a blunt force approach, simply dropping any sample with evidence of within-sample heterozygosity. While that method produces quite accurate results, it leads to unappealing level of data loss, and thus I follow on with a more precise approach.

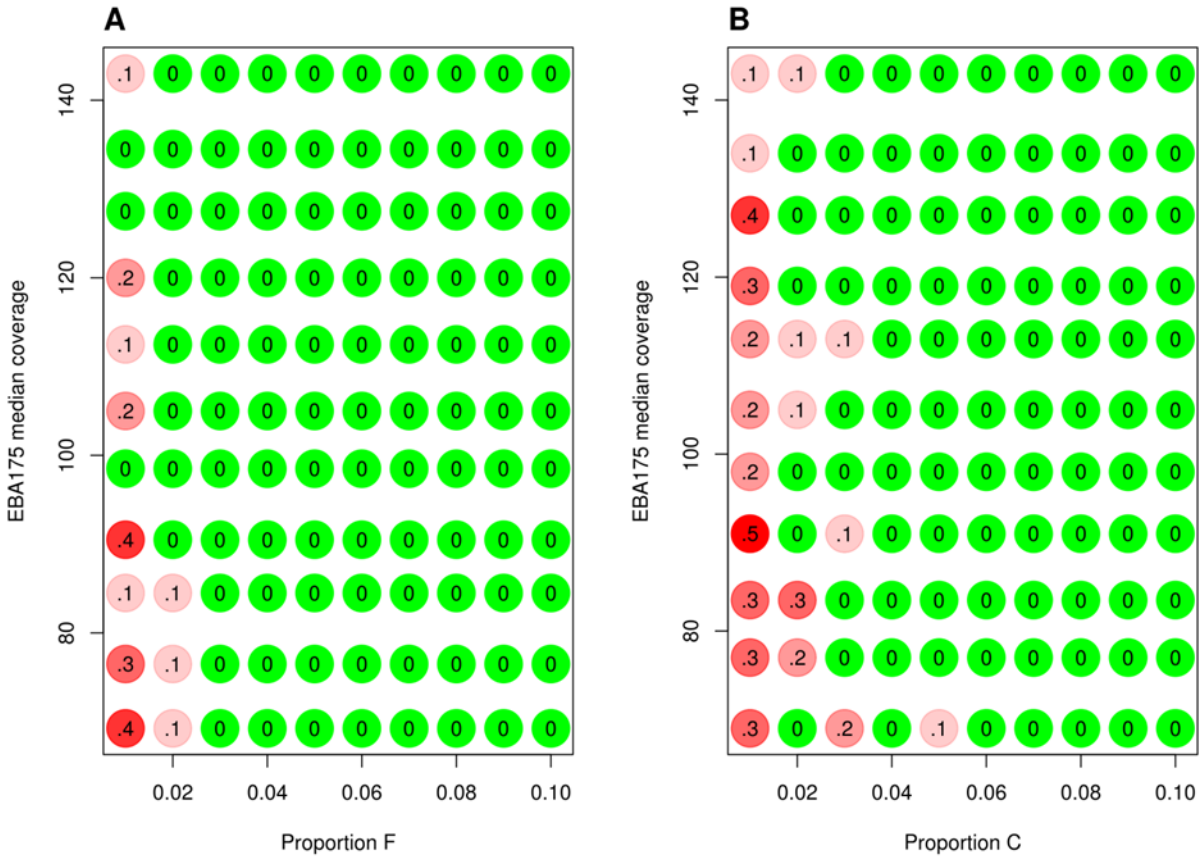
5.4 Acknowledgments

Zam Iqbal first classified the dimorphism in *msp3.4* as Class I/II, and I used his definition here.

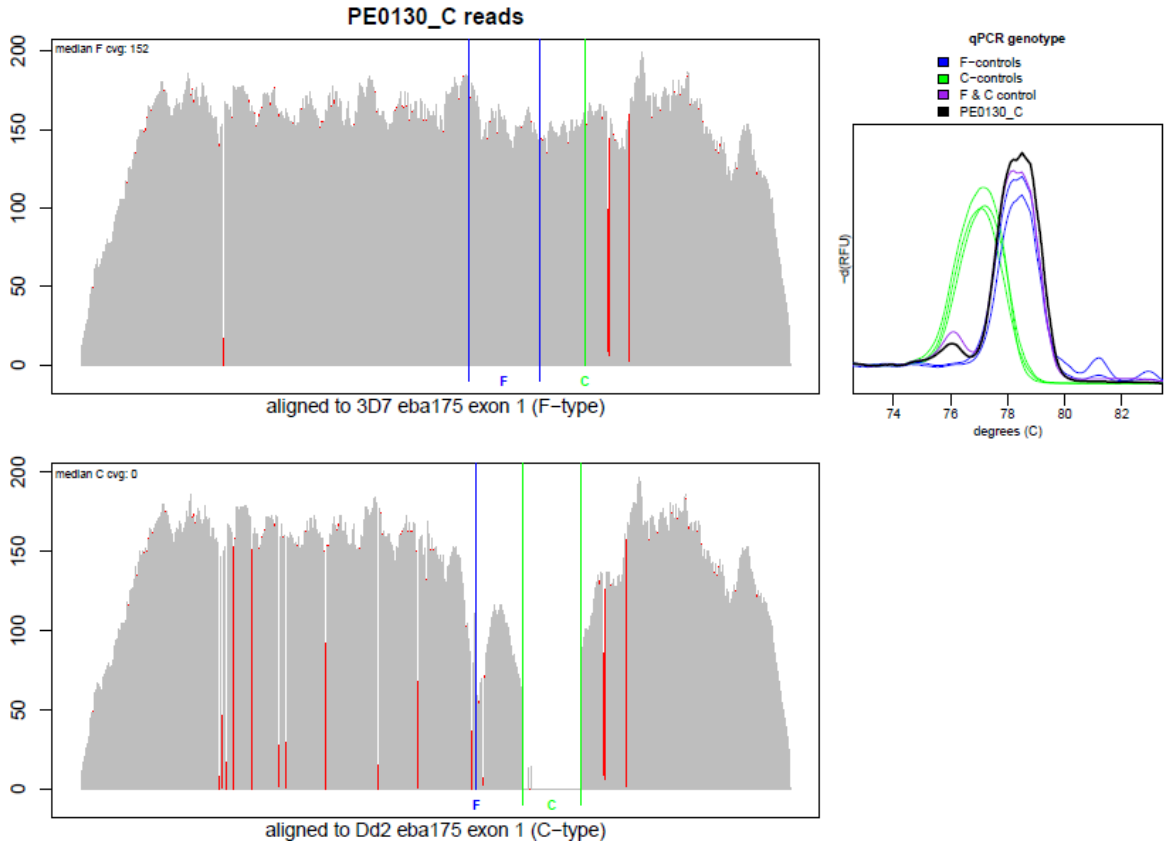
5.5 Supplementary material



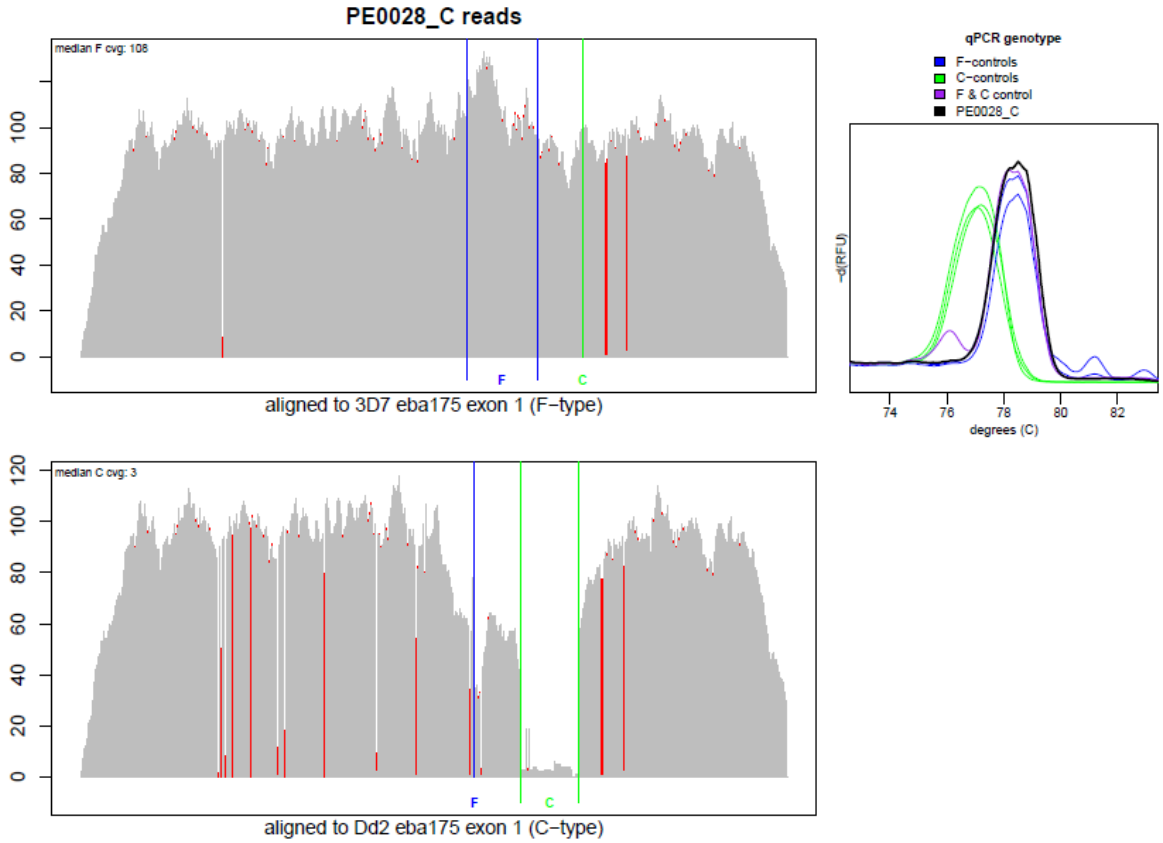
Supplementary figure 5-13. Malign error rates. This plot is an extension of Figure 5-2-D. False negative rates for the EBA-175 F and C indels as a function of effective coverage depth. Each short vertical line depicts the symmetric standard error around the proportion of false negatives in 200 random mixtures of reads drawn from 10 F and 10 C-type parasites. Blue lines were mixtures with the F-parasite in low abundance (testing the lower limits of detection of that indel), and green represents the C-indel. Effective coverage was calculated as the median coverage across the entire gene, multiplied but the proportion of the low abundance parasite in the mixture—i.e., it's an estimate of the coverage expected for the indel. The C-indel has a modestly higher FNR, however this drops to zero at around 4x effective coverage for both indels. For example, one might attain this FNR in a field sample with 150x coverage and the low abundance parasite in the infection representing 2.7% of the reads.



Supplementary figure 5-14. Malign error rates in pairwise mixtures. Malign false negative rates (FNR) as a function of overall EBA-175 sequence coverage depth and proportion of the F/C parasite in an *in silico* mixed infection. Each circle shows the FNR based on 10 *in silico* mixtures at the given coverage depth (y-axis) and proportion of the respective allele (x-axis). (A) F-type parasites are the minority in the mixture. (B) C-type parasites are the minority in the mixture. For example, the top-left value of 0.1 (pink circle) in panel A means 1 out of 10 samples didn't detect the F-insert with median coverage just above 140, and containing only 1% F-type parasites. Default Malign parameter values were used for the genotyping.



Supplementary figure 5-15. Malign coverage plots and PCR validation for a field sample yielding a false negative result. Coverage plots are described in Figure 5-2. Top-right panel: melt curve analysis showing the qPCR result of this sample (black curve) overlaid onto several controls: C-controls (green), F-controls (blue), and a control mixture (purple). The C-insert peak in the melt-curve analysis has lower intensity than average, and there is evidence of some coverage in the C-region of the Malign pileup (bottom panel), thus this sample likely contained a C-type parasite at very low abundance.



Supplementary figure 5-16. Malign coverage plots and PCR validation for a field sample yielding a false positive result. Coverage plots are described in Figure 5-2. Top-right panel: melt curve analysis showing the qPCR result of this sample (black curve) overlaid onto several controls: C-controls (green), F-controls (blue), and a control mixture (purple). Although the qPCR result shows no evidence of a C-type parasite present, there is clearly coverage across this region in the bottom panel.

6 MALMOI: ASSEMBLING GENES WITH COMPLEX VARIATION

The last chapter concluded that reference-free assembly of target genes by patching Cortex variants into the 3D7 version can result in systematic biases. This chapter picks up where chapter 5 left off, i.e., in pursuit of a method for *de novo* assembly of target genes, and addresses two specific aims:

1. Balancing the concepts of reference-free and targeted gene assembly. This is a bit of an uncertainty principal—that is, the more targeted an approach, the more that approach will rely on mapping to a reference. However, targeted assembly is essential due to computational constraints. Much iteration is required to find the right mix of parameters to assemble different genes in different samples. Further, the larger goal is to apply these methods to thousands of MalariaGEN samples.
2. Assembling phased contigs from samples containing a mixture of parasites.

Although this chapter is focused on methods development and the following chapter on their application, in the supplementary material I provide two vignettes demonstrating the utility of MalMOI.

6.1 Abstract

Malaria will kill half a million children this year, and a public health disaster looms as artemisinin resistant parasites expand in Southeast Asia. A comprehensive understanding of the genetic variation in *Plasmodium falciparum* genes encoding drug resistance proteins and candidate vaccine antigens is paramount for malaria surveillance and control. Thousands of parasites from dozens of populations have been sequenced in the past few years, but a void remains in the availability of full-length representations of important genes from these data. Methods that align sequencing reads to a reference genome fail to capture highly divergent polymorphism—a particular problem in genes under selective pressure

from drugs or immunity. Here I present MalMOI, an algorithm and software that utilizes the reference genome, public databases, and de novo methods to assemble phased genes from samples containing multiplicity of infection (MOI). I demonstrate the utility of this approach with the release of thousands of full-length gene and exon sequences representing parasites worldwide, and provide *in vitro* and *in silico* validation. This resource contains an unprecedented scale of haplotypic data that will facilitate vaccine antigen selection, assay design for parasite reconnaissance, and molecular evolutionary studies.

6.2 Introduction

Sequencing entire parasite genomes with short-read technologies has provided a major advance in our understanding of malaria parasite variation, however this approach falls short in several important respects [142]. Traditional pipelines that align short reads to a reference genome are only reliable in relatively static loci—e.g., a few SNPs or small indels in close proximity. Ironically the most important regions of the genome from an intervention perspective are also the most polymorphic due to selective pressure from drugs or host immunity. Exon 2 of the parasite's chloroquine resistance transporter (pfcrt) contains the well described K76T resistance locus, and is also a site with dense polymorphism. Southeast Asian samples frequently yield sequence reads that will have 4 or more SNPs within an individual oligonucleotide when aligned to the 3D7 reference genome, creating mapping issues for common bioinformatics tools [291]. Further, the 12 introns in this gene contain homopolymer stretches and di-nucleotide repeats that exacerbate the difficulty for these tools. The blood-stage vaccine candidate, AMA1, contains more than a dozen potential polymorphisms within a read-sized window (100bp). This antigen also elicits allele-specific immunity, highlighting the importance of not only characterizing its genetic variation, but of doing so in a way that conveys phased haplotypes. Resolving variation at a phased level would allow vaccine trials to measure efficacy stratified by haplotypes, and at the preclinical stage to group antigens for multicomponent vaccines.

Another situation in which reference-based methods fail is in genes that have diverged so extensively that large regions are not homologous in multiple sequence alignments. The vaccine candidate EBA-175 exhibits this property in two large indels (F and C segments) that are purportedly mutually exclusive. The 3D7 genome contains the F-form of this gene (i.e., a 423bp F-insert and 342bp C-deletion), and thus any polymorphism within the C-segment, and identification of the segment *per se*, is lost on reference-dependent methods. Further, alignments near the boundaries of these indels contain artifacts that can result in

erroneous SNP discovery and genotyping (see Figure 5-7). Similarly, MSP3.4 (PF3D7_1035700) is a novel vaccine candidate with one-quarter of the gene (~ 500bp) so diverged from the reference in some parasites that it cannot be aligned to 3D7. Importantly, this highly divergent variation exists in the domain implicated in RBC binding, and antibodies targeting this region inhibit invasion [292]. Characterization of the major divergent forms of this DBL domain would be crucial information for vaccine development and for population genetic analyses, yet this variation is invisible to the most widely used genotyping approaches.

A high proportion of malaria infections are comprised of a complex mixture of genetically distinct genomes, whether from multiple mosquito inoculations, and/or from a single bite containing a mixture [293]. These mixtures make phasing SNPs ascertained by aligning to the reference genome difficult, and make *de novo* assembly methods potentially unreliable because variable regions separated by a conserved stretch can be erroneously spliced together. Another limitation of *de novo* assemblers applied to whole-genome data is that they are difficult to scale to thousands of samples, especially if iteration is required to define optimal parameters. Each gene comes with its own difficulties, each of which that are overcome in assembly with different parameter settings or software tools.

Here I present a pipeline and software designed to assemble individual genes from the dominant parasite strain in patient samples infected with a mixture of malaria parasite genomes. This approach combines the strengths of reference-based and reference-free methods. I demonstrate the utility of this approach with several important vaccine candidates and a canonical drug resistance gene, each representing a particular class of difficulty. Erythrocyte binding antigen (EBA-175) contains two large indels (423bp and 342bp), separated by a fairly conserved region of 279bp. Apical membrane antigen 1 (AMA1) is expressed on the surface of the invading merozoite, and contains dense clusters of SNPs due to frequency dependent selection by host immunity. Merozoite surface protein 3.4 (MSP3.4) is dimorphic over a quarter of the gene, with each form sharing no homology. Finally, *pfcrt* contains 12 low complexity introns. Thousands of full length versions of these genes and partial cDNA sequences are assembled here, representing a massive increase in the publicly available sequence.

6.2.1 MalMOI algorithm overview

The MalMOI pipeline for assembly of specific genes has three major stages: read pull-down, assembly, and filtering (boxed respectively in Figure 6-1 in blue, yellow, and green). The

overall goals of these steps is to 1) reduce the sequencing reads to those only in the target region, 2) to assemble multiple contigs from these reads using various approaches, and 3) to filter the resulting candidates to remove those with evidence of mis-phased variants (e.g., SNPs or indels from different parasites in a mixed infection being spliced together). MalMOI has to balance two opposing forces. On one side is the benefit of using the reference genome and other sources like GenBank to target the assembly as narrowly as is possible. The opposing force to this is the need to remain purely *de novo* to avoid the reference-based issues that have been a sub-theme in every preceding chapter. The MalMOI algorithm lives somewhere between these two forces—perhaps a bit more on the *de novo* side of the spectrum, and the trade-offs and benefits of this are discussed below. The more technical details of the pipeline are provided in section 6.3.

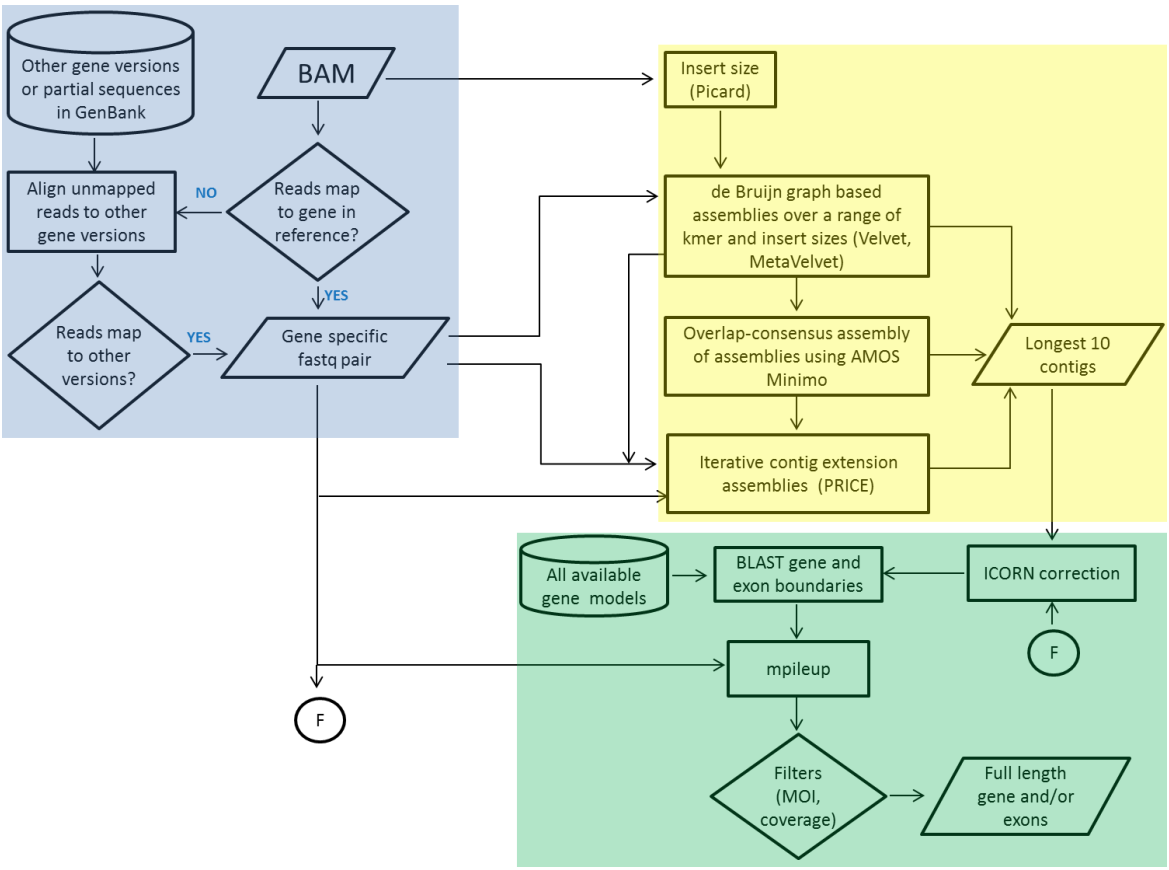


Figure 6-1. MalMOI assembly pipeline. The read pull-down stage is boxed in blue, the assembly stage in yellow, and the filtering stage in green. Starting with the original sample BAM file (top right of the blue box), an insert-size is estimated using all read pairs (to be used later during assembly). Separately, all reads in the BAM that map to the gene of interest (either to the reference or GenBank versions) are extracted. The extracted read-pairs are assembled using a multitude of approaches and the top contigs are put forward for iCORN correction. The corrected contigs are scanned for full-gene and exon boundary matches using BLAST comparisons to a database of available gene models. The original reads are mapped back to contigs for which genes and/or exons are identified, and MOI and

coverage filters are applied. If a full-length gene makes it to that stage, it is put forward as the winning assembly. Otherwise the contig with the most exons passing all filters is selected for output.

6.3 Materials and methods

6.3.1 Read pull-down

This step utilizes the reference genome and any publicly available versions of the gene of interest to drastically reduce the number of reads to be assembled. A common starting point for many researchers will be with a BAM file, produced by aligning the raw paired-end reads from fastq files to a reference genome. The BAM file contains both mapped reads with genomic coordinates, as well as those reads that did not align to the reference—thus fastq files can be reconstructed from BAM files. The final output of this step is two small paired-end fastq files containing reads for which either partner aligns to the target gene. These targeted reads come from two places: those that map to the region of interest in the reference alignment, and those of the original unmapped reads that map anywhere in a database of all publicly available versions of the gene. Selecting read-pairs for which either partner maps is enough to recover indels at least as large as 342bp, as we see in *eba-175*.

The mapped reads are extracted using a Linux Awk command, for example, the following would be used to pull-down reads on chromosome 7 from positions 402622 to 406917:

```
samtools view -h file.bam | awk ' ($3=="Pf3D7_07_v3" && $4>=402622 && $4<=406917) ||
(($7=="Pf3D7_07_v3" || ($3=="Pf3D7_07_v3" && $7=="=")) && $8>=402622 && $8<=406917) '
| sort > file.ordered.sam
```

Note that this yields a file without the original header information, which can cause problems downstream with some tools. The original header is retrieved with the -H option of 'samtools view' and concatenated to file.ordered.sam.

Separately, the unmapped reads are extracted from the original BAM using 'samtools view -h -f 12 -F 256' and converted to fastq format with SamToFastq.jar in Picard tools [294]. The resulting unmapped fastq files are subsequently aligned to a fasta file containing all available versions of the gene of interest using 'bwa aln -n0.01 -k4 -l32'. Any read-pairs for which at least one partner maps are extracted with Awk and appended to those that mapped to the reference originally.

The R script that performs the read pull-downs (GetReadsFromBAMs_batch_v3.R) is typically called in parallel on thousands of samples, so this is a convenient script within which to also retrieve insert size metrics from the original BAM using CollectInsertSizeMetrics.jar in Picard-tools.

6.3.2 Assembly

See 6.3.4 for a brief description of specific software described in this section. The targeted fastq files resulting from the read pull-downs are then *de novo* assembled using Velvet and MetaVelvet with a range of parameters [290,295]. Two key parameters affecting assemblies are kmer and insert sizes. Assemblies are therefore performed on a range of kmer sizes (21 to 91, by units of 10), and for each kmer, assemblies are done across a range of insert sizes around the median from Picard-tools (+/- 50 and +/- 100). The resulting contigs from these de Bruijn graph assemblies are fed into the overlap-layout-consensus algorithms AMOS-Minimo and PRICE [296,297]. The longest 10 contigs from any of the above methods are then corrected for small errors using ICORN, and taken forward for filtering [298]. As well as correcting small errors, the benefit of using iCORN is that it will 'flip' alleles to the majority base, which helps with phasing. In other words, if a SNP from a low abundance parasite gets erroneously spliced into a contig with a high abundance parasite, iCORN will flip to the majority allele.

6.3.3 Filtering

In the final step, the pulled-down reads are aligned back onto the top 10 contigs using Samtools mpileup and subjected to coverage and MOI filters [299]. The coverage filter is failed if any position in the pileup had coverage less than 10. The MOI filter is failed if any within-sample heterozygous position contains a minor allele that represents more than 20% of the coverage. Thus contigs that pass represent the strain in the infection that comprises at least 80% of the mixture. This threshold is somewhat arbitrary, but has been validated to be rather conservative (see results). As this parameter is likely different for different genes, future versions of MalMOI will output the level of MOI with the contig so filtering can be done more dynamically. Finally, a coverage plot is produced for each pileup and checked visually for artifacts.

6.3.4 Brief descriptions of key third party software

6.3.4.1 Velvet

The Velvet software builds a de Bruijn graph of portions of overlapping sequence reads (i.e., kmers) and constructs a contig by finding the appropriate path through this graph. Polymorphisms and sequencing errors can be visualized in the graph as 'bubbles,' as they represent divergent points where more than one path can be taken through the graph, flanked by conserved regions where the path converges. Velvet collapses these bubbles,

selecting the path through the bubble largely based on coverage. Kmer size and insert size are parameters supplied by the user, and have a substantial impact on the graph and its interpretation (Supplementary figure 6-18).

6.3.4.2 MetaVelvet

MetaVelvet attempts to deconvolute the bubbles in the Velvet de Bruijn graph to assemble individual contigs from related species in metagenomic samples. It is convenient to envision this as phasing individual haplotypes when applied to mixed malaria infected sample. The concept is that polymorphisms that are in *cis* will have sequencing coverage that correlates to the abundances of the comprising species in the sample. In my hands this algorithm does not work well to assemble more than one parasite in a mixture, however it does assemble single contigs that theoretically should have a better chance of being in phase than with Velvet alone. One of the benefits of the read-pulldown in my approach is that it is computationally feasible to use both programs with many parameter settings, increasing the chance that at least one will work.

6.3.4.3 AMOS Minimo

Minimo is an overlap-layout-consensus (OLC) based algorithm in the AMOS suite of assembly packages. OLC is an alternative paradigm to graph based methods, each with their own strengths and weaknesses. This approach straightforwardly looks for overlaps of contigs with a user defined length and percent identity. My pipeline uses this software on the contigs resulting from Velvet and MetaVelvet—producing an OLC assembly of graph-based assemblies. The key parameters values used in my pipeline can be changed, but are typically set at MIN_LEN=35 and MIN_IDENT=95.

6.3.4.4 PRICE

PRICE (paired-read iterative contig extension) is applied separately in my pipeline to the output contigs of Velvet, MetaVelvet, and Minimo. This software takes as input the assembled contigs from the other methods as well as the initial read-pairs. The reads are aligned to the contig ends, and using mate-pair and expected insert size information, the contigs are extended. This is done iteratively (20 cycles in my implementation), and as the contig grows it attempts also to collapse scaffolds using read-pairs that span gaps. PRICE was developed to assemble low abundance viral genomes from ultra-deep sequenced metagenomic samples. There are many parameters for this software, and the defaults used

in my implementation can be seen in the 'price' function of my script VelvetAssembly_batch.R.

6.3.4.5 iCORN

An essential component to the phasing of alleles in an assembled contig that derives from a sample with within-sample heterozygosity is iCORN (iterative Correction of Reference Nucleotides). This software iteratively aligns reads to a user specified reference (in this application the assembled contig), identifies loci where the coverage drops because the reads disagree with the reference, and modifies the contig to better reflect the reads [298]. Although iCORN was not designed with phasing in mind, it has the side-benefit of flipping reference alleles to the one most represented by the reads. For example, in our cartoon example of heterozygosity (Figure 3-1), the reads over position 10 contain 6 G alleles and 4 A alleles. If the reference contained an A in position 10, iCORN would flip it to a G and perform another iteration of alignment. In assemblies of mixed samples Velvet contigs frequently contain a mixture of the major and minor alleles at different positions, and iCORN helps to phase to those of the parasite in majority abundance in the sample. Further, as demonstrated in a controlled experiment with *pfcrt*, iCORN corrects 6 manually entered errors, ranging from SNPs, to small indels and scaffolding gaps (Supplementary figure 6-10).

6.4 Results

6.4.1 Full-length assemblies

Four genes were assembled from more than 3000 MalariaGEN samples to demonstrate MalMOI on genes with different variation properties. Two of the genes contained single exons (*ama1* and *msp3.4)*, while *pfcrt and eba-175* have multiple exons, and thus also have exon-targeted assemblies described below. The number of full-length genes and where applicable the number of full-length cDNA (i.e., exons only) representations that were assembled by MalMOI are shown in Table 6-1.

Table 6-1. Summary of MalMOI assemblies.

Gene	Samples >= 10x cvg	Full-length	Full-length cDNA
ama1	3009	739	NA
msp3.4	3033	1390	NA
eba-175	3199	1613	1893
pfcrt	2829	6	263

6.4.2 Exon-targeted assemblies (CDS output)

In the final step in the flowchart in Figure 6-1, both full length genes and individual exons are output by MalMOI. It is quite common in multi-exon genes for one or more of the exons to pass all MalMOI filters, while other exons fail. Rather than throw the baby out with the bathwater in this scenario, MalMOI returns these assembled regions as part of the coding segment (CDS) output. Only exons that were assembled on the same contig are returned—thus they represent phased haplotypes from one parasite. Specifically, after iCORN correction of the top 10 contigs, MalMOI first looks for a contig encompassing the entire gene for which every intron/exon position passes the coverage and MOI filters. If such an assembly is found, the corresponding exons comprise the CDS output. However, if no contig captures the full-length gene, the one containing the highest number of passing exons is selected, and these exons are returned as the CDS output. A downside of this approach is that a particular exon may be missed if it happens not to fall on the contig with the highest exon count. The upside is that the resulting exons can be reliably patched together, representing a full-length cDNA for which LD and haplotypes can be investigated. Of the genes investigated in this chapter, *pfcrt* and *eba-175* were the only two with multiple exons.

6.4.2.1 pfcrt

As described in 3.6.2, *pfcrt* contains 12 low complexity introns that are hard to access by aligning to the reference genome. It turns out that the homopolymer stretches and dinucleotide repeats comprising these introns are also extremely difficult for *de novo* assemblers, motivating the development of the CDS output in MalMOI. Indeed, in only 6 out of 2908 possible *pfcrt* assemblies did MalMOI yield a full-length gene (Table 6-2). However, 263 full-length cDNA sequences were assembled, and 1242 samples yielded a CDS with at least 10 exons.

Not all exons are equally likely to assemble. As portrayed in Figure 6-2, 87% of samples yielded an intact exon 8, whereas exon 1 could only be fully assembled in a quarter of samples. This property does not appear qualitatively to be associated with the degree of polymorphism, length or complexity of surrounding introns, or length the exons themselves. As mentioned in the introduction, each gene has different peculiarities that are most optimally assembled under particular conditions. Of the nearly 23,860 *pfcrt* exons that were assembled, 40% resulted from the PRICE step of the MalMOI pipeline, whereas this number for the *eba-175* exons is 2.3%.

Table 6-2. Exon-focused assembly counts. Number of parasites assembled with the given number of exons. Bottom category is 13 exons and all introns (i.e., full length assemblies).

Exons captured in one parasite	Number of parasites
1	98
2	113
3	136
4	94
5	142
6	167
7	214
8	395
9	307
10	658
11	173
12	142
13	**263**
13 + introns	**6**

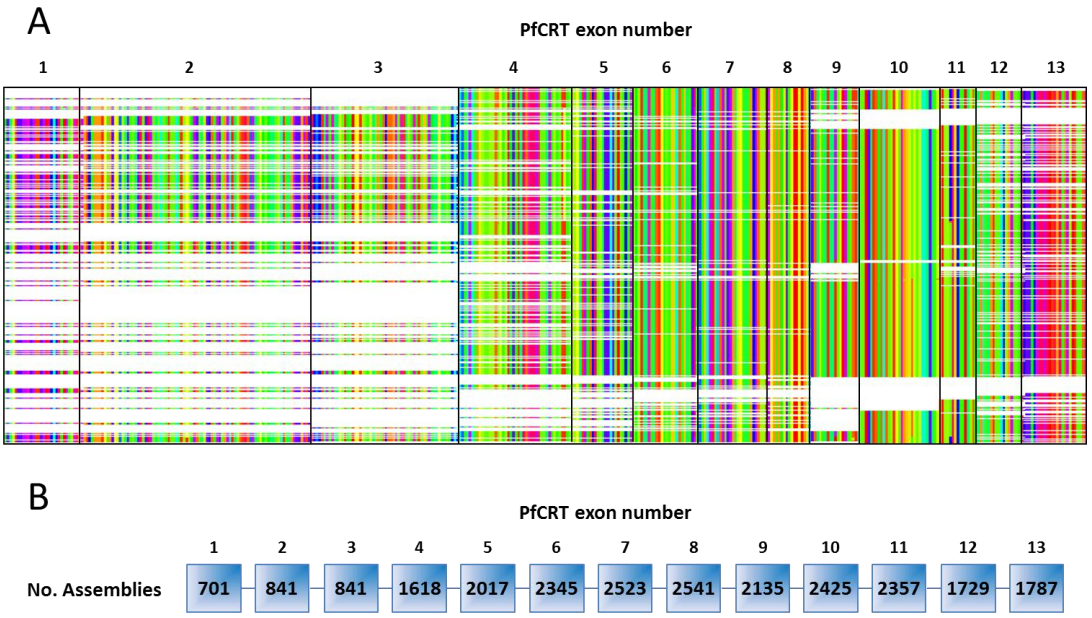


Figure 6-2. MalMOI gapped translated cDNA output and exon counts. A) Multiple sequence alignment of 1908 full or partial CDS sequences. Exons in a given sequence all derive from the same contig, thus from the same parasite. Exons are numbered along the top and are demarcated by black vertical lines. Amino acids are colored using the Taylor scheme in Jalview. **B)** Schematic of *pfcrt* gene model with the number of successful assemblies out of 3338 possible for each exon indicated within each box.

6.4.2.2 eba-175

Out of 3199 samples with *eba-175* coverage of at least 10x, 1613 yielded a full-length assembly (i.e., all exons and introns). The CDS output for this gene included 1893 sequences containing all 4 exons. In other words, approximately 15% of the potential full-length assemblies failed due to difficulties with the introns.

Figure 6-3. MalMOI assembled exon counts for *eba-175*. A) Schematic of *eba-175* gene model with the number of successful assemblies out of 3199 possible for each exon indicated within each box.

6.4.3 The impact of MOI on assemblies

The determination that it was necessary to apply MOI filters in the MalMOI pipeline came from comparisons of capillary sequenced genes to assemblies for several mixed and clonal samples. As illustrated in Figure 6-4 for using MalariaGEN SNPs, two field samples (PE0027 and PE0030) show substantial evidence of mixture in *eba-175*, whereas the other 4 samples appear to be clonal. Curiously, there is scant evidence of mixture in the supposedly clonal lab-line, HB3. PE0030, which shows the highest degree of mixture, yielded an assembly that did not match capillary sequence from the same sample (Supplementary figure 6-16). The other 5 sample produced accurate assemblies, including PE0027.

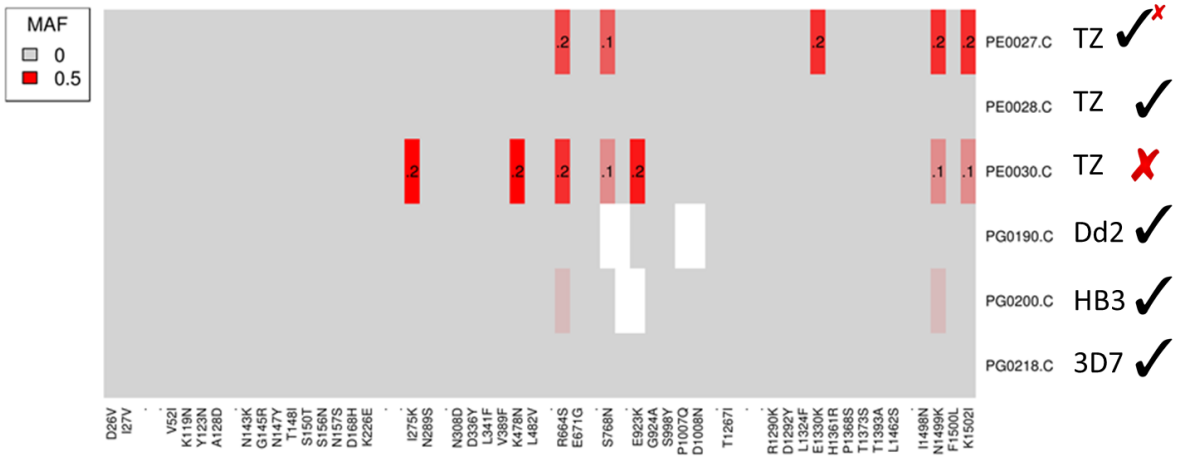


Figure 6-4. Within-sample heterozygosity in *eba-175* control samples. Each row represents one of six samples that were *de novo* assembled and capillary sequenced. Each column is one of 58 MalariaGEN v2.0 SNPs. Cells are colored on a scale between 0 and 0.5, based on within-sample minor allele frequency. To the right of the IDs, samples are labeled as being Tanzanian field isolates (TZ) or

clonal lab-lines (Dd2, HB3, 3D7). Samples are labeled with a check if their Velvet assembly agrees with the capillary sequence, otherwise with an 'X.' PE0027 is labeled with both, because as shown in Figure 6-5, the capillary sequence matched serendipitously.

On closer inspection, Figure 6-5 shows that the accurate PE0027 assembly may have been serendipitous, motivating the use of filters to ensure this sort of assembly is yielded intentionally, and excluded if SNPs are misphased.

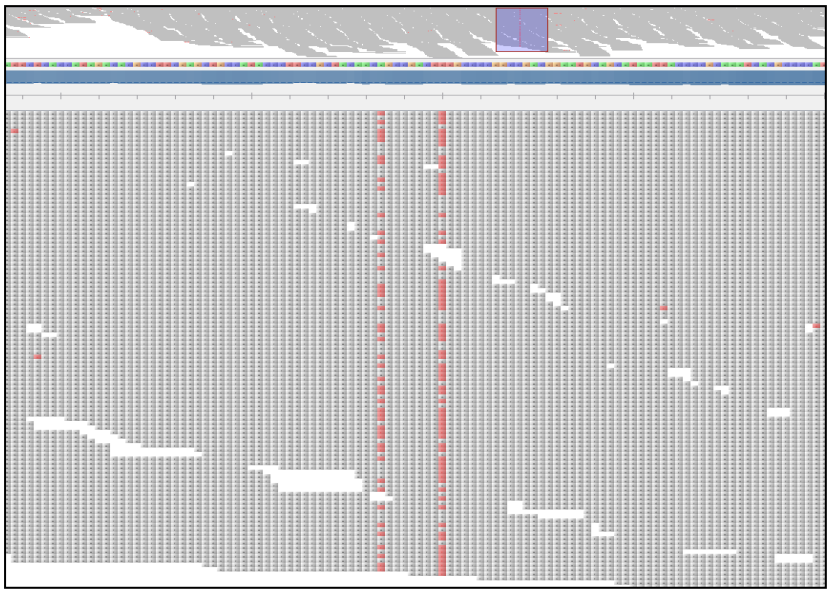


Figure 6-5. A closer look at the pileup for PE0027 onto its *eba-175* assembly. Tablet visualization of the pileup of PE0027 reads onto the *de novo* assembly of *eba-175*, zoomed to two adjacent heterozygous SNPs. An overview of the pileup is shown along the top, with the zoom region boxed in blue. In the lower plot, nucleotides are colored gray if they match the assembly and red if do not. As indicated in Figure 6-4, this sample appears to be an approximately 80/20 mix of two gene forms. The two SNPs here show that approximately 80-90% of the reads match the alternate alleles (red), thus although the capillary sequence matches the assembly, it is to some degree due to luck, and the filters applied by MalMOI are important for catching such events.

6.4.4 The trade-off of limiting to clonal samples

An early consideration for avoiding complications due to MOI was to focus only on samples that had no MalariaGEN SNP evidence of within-sample-heterozygosity in the target gene (see Figure 3-1 for an illustration of this concept). An obvious limitation of this is that it relies on MalariaGEN SNPs, which due to missingness in complex regions motivated this chapter to begin with. Another limitation is that it filters out a high percentage of samples (Figure 6-6).

A separate approach for defining clonal samples is to use only those with an inbreeding coefficient (Fws) close to 1 (see section 4.3.4). The drawback of this approach is that it is a genome-wide metric, which is lose-lose in this context. Not only would this not guarantee clonality in the target gene, it might exclude non-clonal samples that lack heterozygosity in the gene of interest.

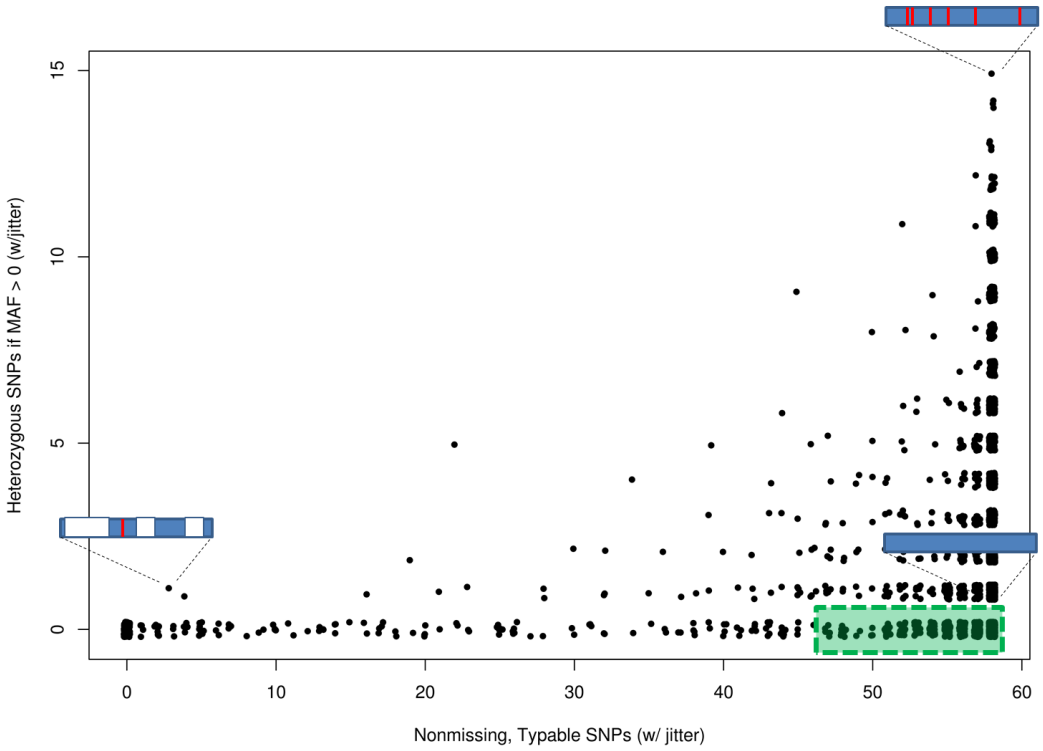


Figure 6-6. Filtering samples by missingness and MOI in *eba-175*. Each dot represents a sample with data based on 59 v2.0 MalariaGEN SNPs in *eba-175*. For each sample, the count of the number of missing SNPs is plotted against the number of the 58 SNPs with evidence of MOI. For a given SNP, evidence for MOI is defined as within-sample heterozygosity > 0. Three cartoon gene models are overlaid onto the plot as example genes in those regions of the plot. In these models, blue represents conserved regions, red indicates heterozygous SNPs, and white depicts missingness. The bottom left model has much missingness and few SNPs and the top right model has no missingness and many SNPs with evidence of mixture. Clonal isolates (for *eba-175)* are represented in the bottom right. The green box highlights SNPs with less than 20% missingness and no mixture.

In conclusion, limiting assemblies to clonal samples is not ideal because the methods for defining clonality are inexact, and the data loss is substantial. Based on inspection of sample reads aligned to the reference gene and to the assembled gene, an arbitrary filter of MOI<20% was chosen as the MalMOI default (Figure 6-7). In other words, when a sample's reads are aligned back onto its targeted assembly, any exon containing a SNP for which more than 20% of the reads have an alternate allele is filtered out. This is depicted in the

bottom panel of Figure 6-7, where the coverage and MOI filters are failed, but exons 2, 3, and 4 are labeled "good." Exon 1 has multiple SNPs that fail the MOI>20% filter.

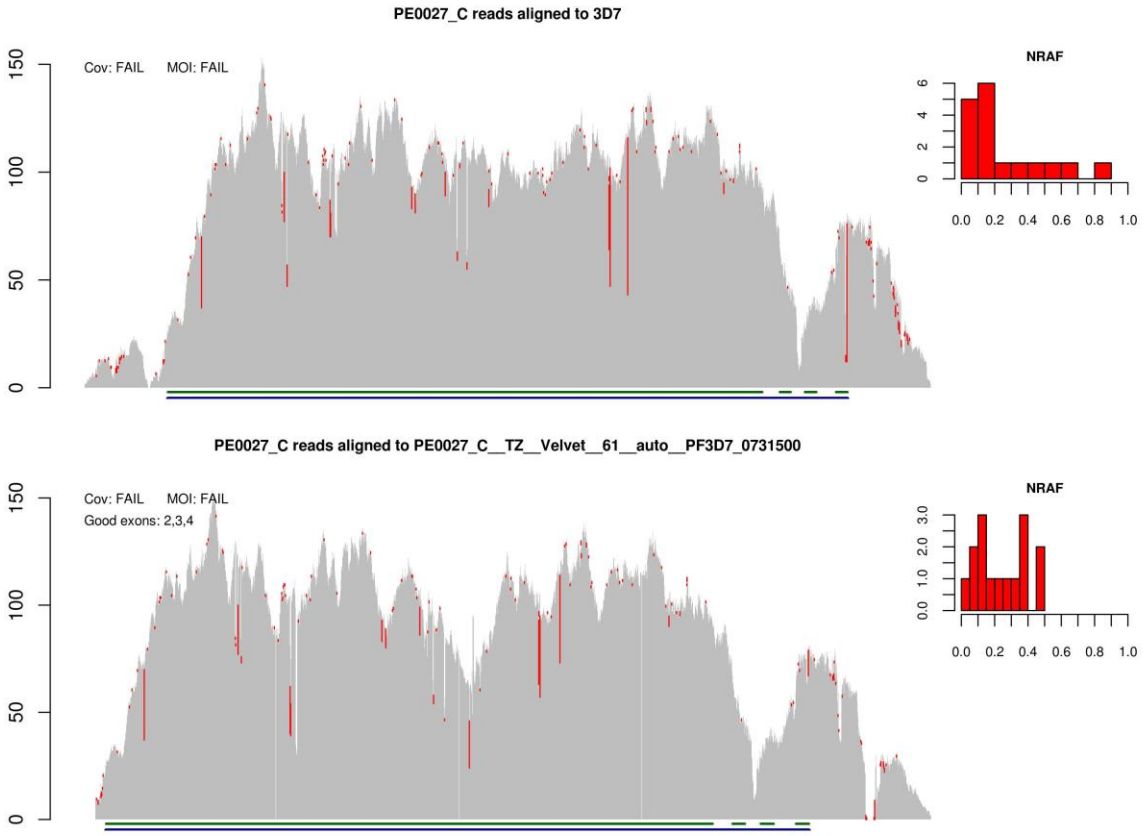


Figure 6-7. Assessing MalMOI assemblies with coverage plots. Each panel depicts the coverage depth of reads from PE0027 aligned to 3D7 *eba-175* (top), or the MalMOI assembly. Each pileup also has a histogram showing non-reference allele frequency (NRAF). These histograms show counts of heterozygous positions (with minimum depth of 3) at the given frequencies. The reference in each panel is the *eba-175* version to which reads are aligned. A bar at each position along the gene (models at bottom) show read depth. Gray bars indicate reads that match the reference, and red shows mismatching proportion. The text at the top left of each panel shows MalMOI filter results (Cov = coverage >=10, MOI<20%).

6.4.5 Validation

6.4.5.1 Capillary sequencing

As detailed in the methods section 2.6, conserved primers were designed tiling across *eba-175*. Large fragments (4081bp and 4795bp) of exon 1 were amplified and then these tiling oligonucleotides were used as forward and reverse capillary sequencing primers. The longest possible consensus sequences were derived from the overlapping capillary reads and compared to *de novo* assemblies from Illumina sequencing of the same DNA source.

Comparisons were made for three clonal laboratory lines (Dd2, HB3 and 3D7), as well as for three Tanzanian field isolates (Table 6-3). All three *eba-175* assemblies of the laboratory-line parasites perfectly matched the capillary consensus sequences. Further, one of the Tanzanian field samples showed no evidence of within-sample heterozygosity in any MalariaGEN *eba-175* SNP (PE0028, Figure 6-4), and this assembly was also a perfect match. The other two field samples showed evidence of mixture. Thus, all sequences were either assembled without error, or were properly filtered out due to MOI concerns.

Table 6-3. Summary of capillary sequencing vs. *eba-175* assemblies.

Parasite	Pairwise identity (bp)	Figure	Sample type	Outcome
Dd2	4240/4240	Supplementary figure **6-11**	Clonal lab-line	100% identity
PE0027	2394/2394	Supplementary figure **6-12**	Mixed field sample	100% identity (filtered out)
PE0028	4461/4461	Supplementary figure **6-13**	Clonal field sample	100% identity
HB3	4400/4400	Supplementary figure **6-14**	Clonal lab-line	100% identity
3D7	3294/3294	Supplementary figure **6-15**	Clonal lab-line	100% identity (trimmed 18bp)
PE0030	2518/2601	Supplementary figure **6-16**	Mixed field sample	Fail (filtered out)

6.4.5.2 Artificial *in silico* parasite mixtures

To get a more formal sense of the MalMOI error rate using larger numbers and known MOI proportions, artificial *in silico* mixtures were generated for the two single-exon genes; *ama1* and *msp3.4* (see methods section 2.6.3). As depicted in the Figure 6-8 and Figure 6-9, the proportion of genes assembled with errors (red bars) diminishes rapidly as one parasite becomes more dominant. This is seen in two respects. Within each panel (i.e., number of strains in each mixture) the errors decay as the proportion dominant strain increases. Separately, at a given proportion the error decays as more strains comprise a mixture. This is likely because the non-dominant strains are splitting the minor-proportion. These trends are similar for both genes. The mixture combinations for each gene generated approximately 100,000 assemblies, of which about 33,000 passed the MalMOI coverage and MOI filters (i.e., Coverage >= 10x and MOI < 20%). Based on these "passed" assemblies, the error rates for *ama1* and *msp3.4* are 0.0054 and 0.0018, respectively. In other words, if these genes are indicative of others, one might expect 1 error for every ~300 assembled

genes. This error does not account for post MalMOI filtering—for example dropping assemblies with stop codons introduced, which would lower the error.

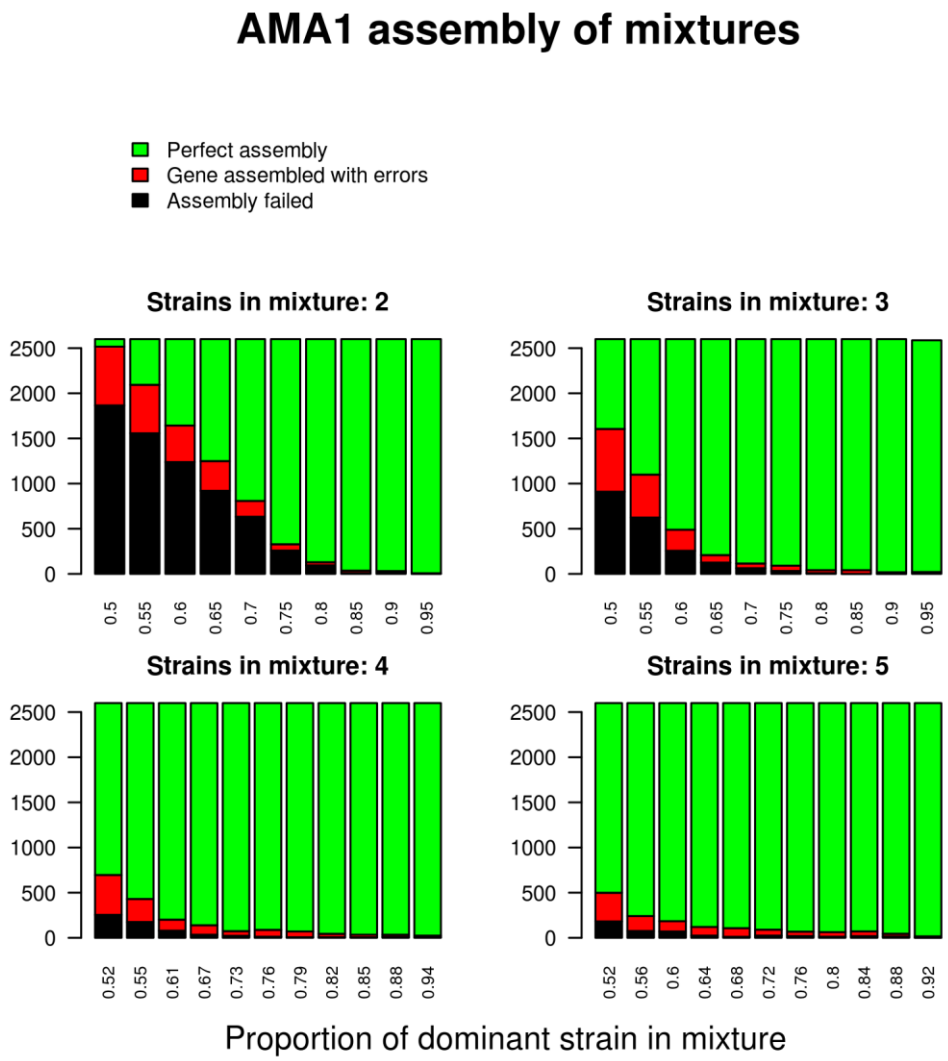


Figure 6-8. Error profile of *ama1* assemblies under various degrees of mixture. Artificial mixtures of different lab isolates was performed as described in the methods. Each of the 4 panels depicts a mixture of different numbers of parasites (labeled above each plot). The x-axis of each plot shows the proportion of the mixture comprising the strain being tested (3D7 or Dd2). At each proportion, 2500 random *in vitro* mixtures were created and assembled. Assemblies were compared to the test strain version of the gene, and the barplots shows the error profiles. The worst-case scenario is the red category, in which an incorrect assembly is returned. The black category is good in the sense that the MalMOI filters have done their job. The error rate for this gene is 0.0054.

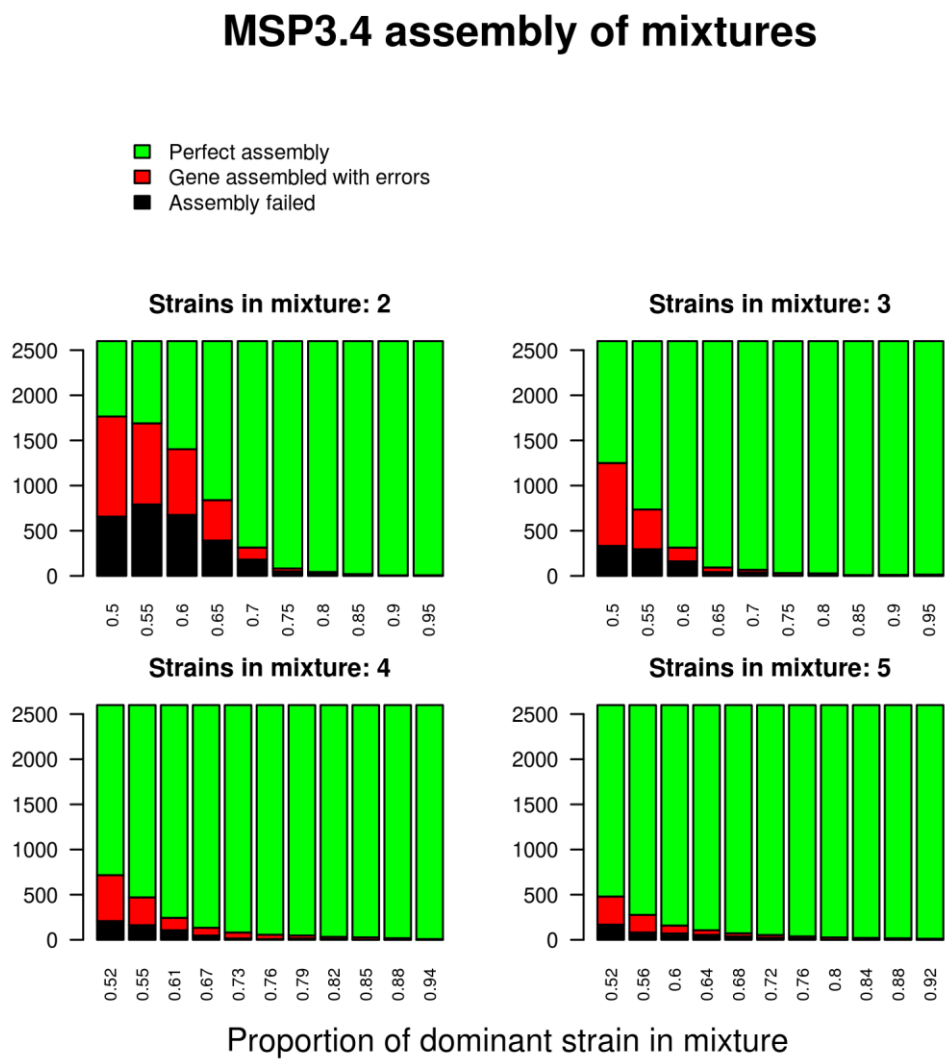


Figure 6-9. Error profile of *msp3.4* assemblies under various degrees of mixture. Artificial mixtures of different lab isolates was performed as described in the methods. Each of the 4 panels depicts a mixture of different numbers of parasites (labeled above each plot). The x-axis of each plot shows the proportion of the mixture comprising the strain being tested (3D7 or Dd2). At each proportion, 2500 random *in vitro* mixtures were created and assembled. Assemblies were compared to the test strain version of the gene, and the barplots shows the error profiles. The worst-case scenario is the red category, in which an incorrect assembly is returned. The black category is good in the sense that the MalMOI filters have done their job. The error rate for this gene is 0.0018.

6.5 Discussion

I have created a framework for assembling full-length genes representing the dominant strain from mixed *P. falciparum* infections. Applying this pipeline to *ama1*, *eba-175*, *pfcrt,* and *msp3.4* resulted in thousands full-length and partial CDS assemblies (Table 6-1). The accuracy of these assemblies was assessed by comparing several to capillary sequence data

and by assembly of *in silico* mixtures. The error rate is low (approximately 1 error for every 300 assemblies), and is likely lower in cases where most samples have a very dominant strain. Two key innovations of the MalMOI pipeline are the ability to target a given gene with many methods, and second, the ability to generate a phase contig from a mixed sample. A separate benefit it that as introns are difficult to assemble due to low complexity, MalMOI attempts to extract individual exons.

The idea of targeting one gene at a time is important, as it results in a reduction of computational resources of many logs compared to genome-wide assemblies. The benefit being that a multitude of methods and parameter combinations, each with different strengths in different situations, can be applied. A theme throughout this thesis is that different genes are hard to access for different reasons. For example, as detailed further in section 6.4.2.1, the iterative end extension applied by the PRICE step in the pipeline is much better at constructing full exons in *pfcrt* than Velvet. Without knowing in advance which settings will be most optimal for particular genes, it is useful to have the computational capacity to throw the kitchen sink at the problem. A typical run of MalMOI will process hundreds of assembly attempts, read mappings, and blast searches in a few minutes—a feat that would be virtually impossible genome-wide, even for just one sample. The targeted approach also requires less RAM (approximately 7GB max per run), and thus samples are typically processed in parallel on a Linux computing cluster. In fact, the genes presented here were assembled from more than 3000 MalariaGEN samples running in parallel 50-100 at a time, taking on the order of a few hours to complete.

Although the focus of this chapter was about developing the MalMOI algorithm (with an application to *eba-175* reserved for Section III), in the supplementary material I provide two brief examples of analyses that benefit from MalMOI output. For these examples I use *ama1* and *msp3.4*.

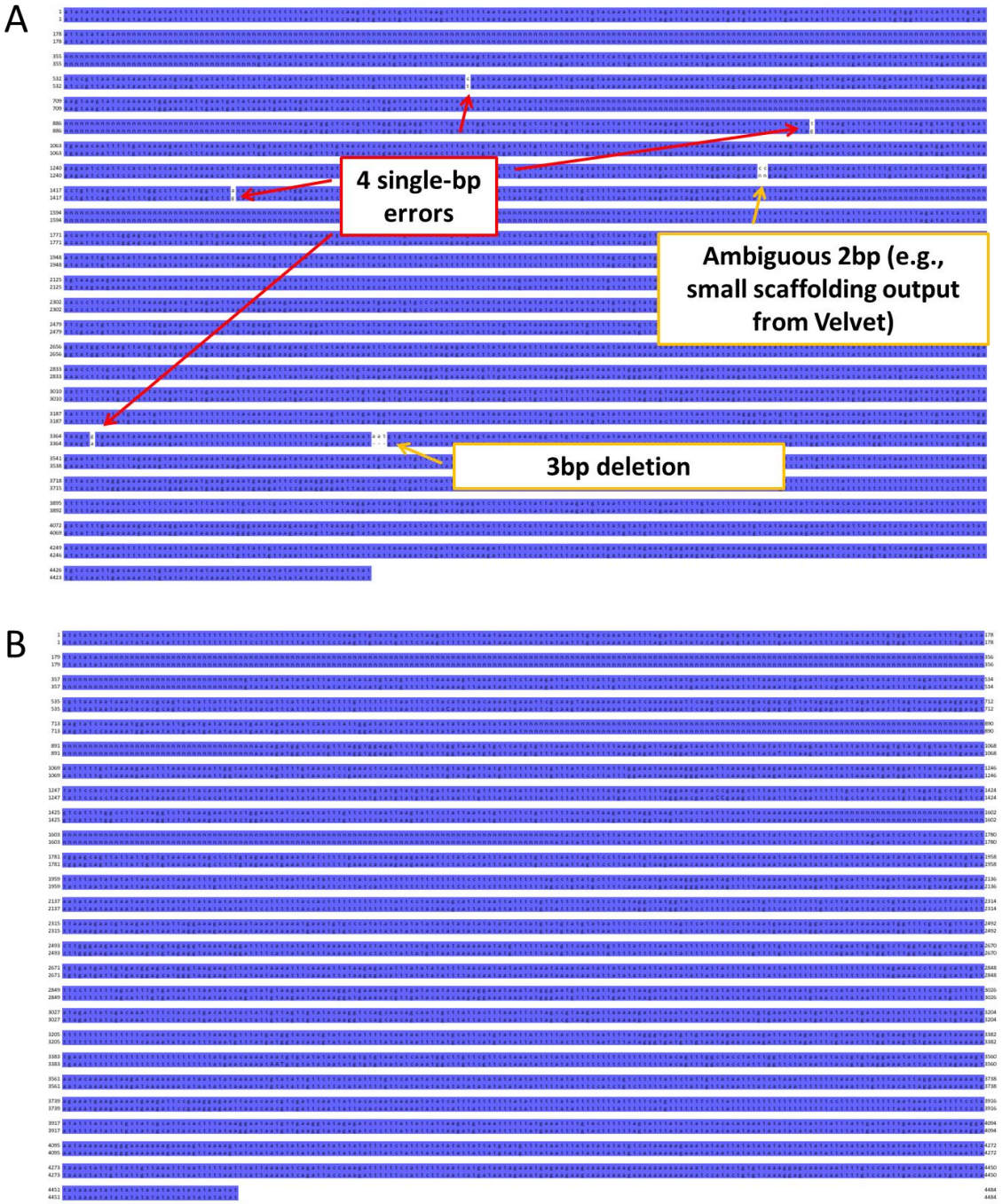
6.6 Limitations and future work

Although this method substantially expands the number of samples from which full-length products can be assembled, it is limited to samples that have at least one parasite at least 80% preponderant, and does not attempt to construct contigs from low abundance strains in an infection. The MetaVelvet step in the pipeline could theoretically assemble more than one version of a gene from the same infection, in fact this was the initial motivation for including it—however in practice this has not been observed. In this version of MalMOI, only the dominant parasite's gene would be retained.

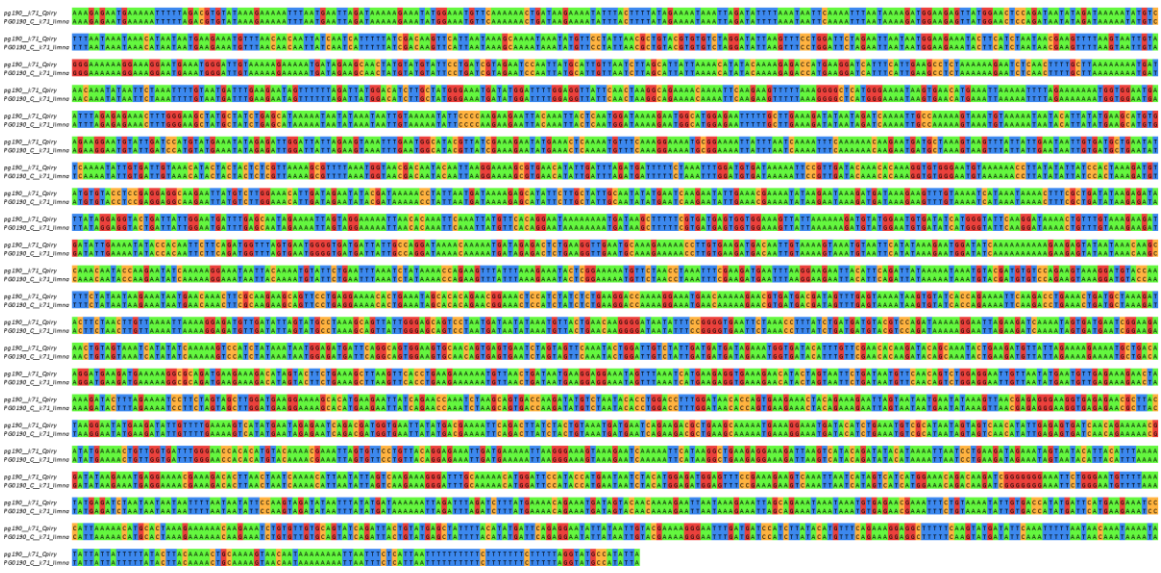
Highly divergent genes with minimal available sequence information will be problematic using an approach that relies on public databases. Although beyond the scope of this work, I did encounter this with a type 3 *var* gene, and can briefly recommend some approaches to overcome this issue. Rather than aligning the unmapped reads to the publicly available database as above, since in this situation there is no public database, simply append all unmapped reads to those that map to the reference. The downside of this is the substantial increase in time it takes to assemble many more reads, as well as the large number of contigs that result from the assembly, which must then be used as a BLAST database, searched with the target gene, to find the one of interest. One could use this approach on a subset of samples to build the database with different versions of the gene, and feed this into the original pipeline. This might also be done iteratively, expanding the database with each iteration in a spirit similar to psi-blast.

Future versions of MalMOI will not use a hard MOI cutoff, but rather will output all assemblies annotated with their MOI filter score. As different genes may have different error propensities, or some investigations may tolerate a higher or lower error rate, these should be tuned post-assembly to fit each situation.

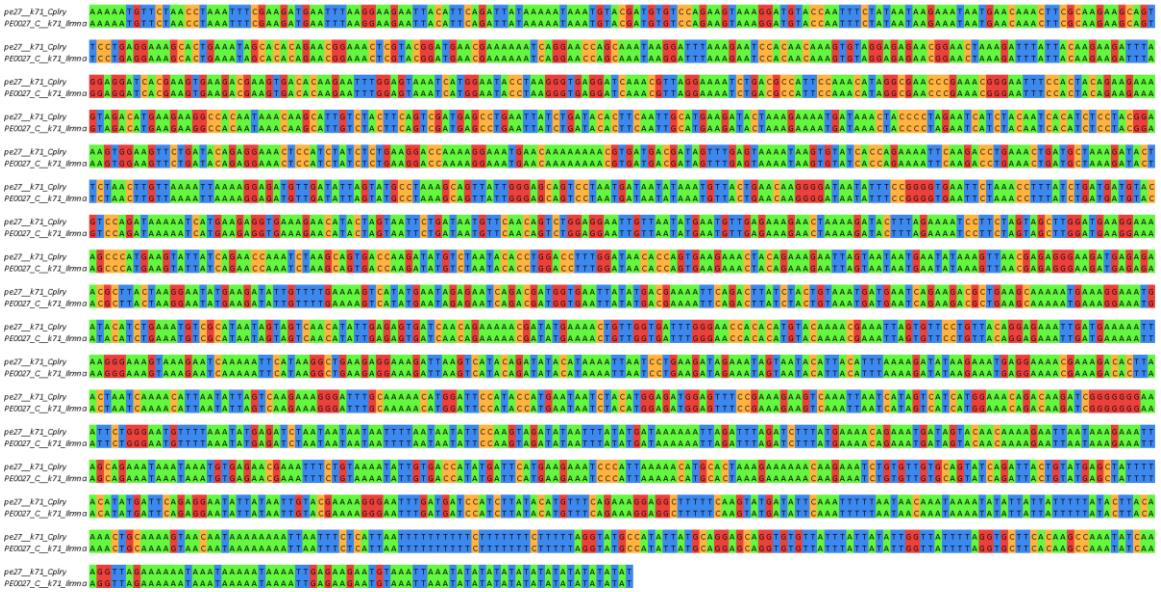
6.7 Supplementary material



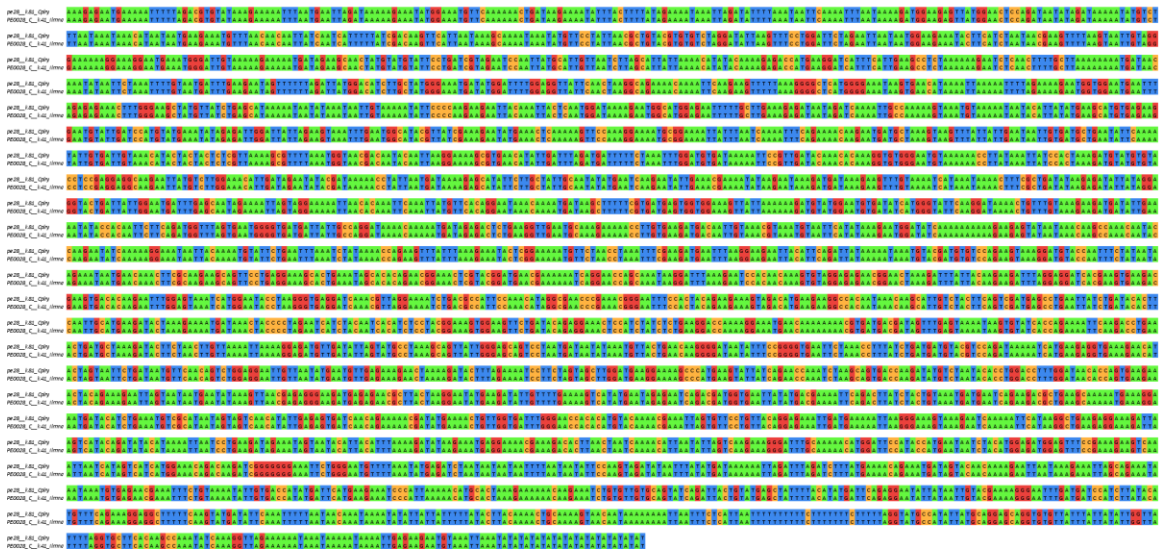
Supplementary figure 6-10. Controlled iCORN correction of *pfcrt*. A) Bird's-eye view of an alignment of two *pfcrt* sequences, one with 6 errors manually added. As labeled in the figure, 4 SNPs, a 3bp deletion, and a scaffolding error are added. The scaffolding error is an 'NN' sequence substituted into the reference. Velvet frequently does this to connect separate contigs that couldn't be completely assembled through, but that likely should be on the same scaffold because they are connected by read pairs. **B)** Same alignment after 6 iterations of iCORN. All errors have been corrected.



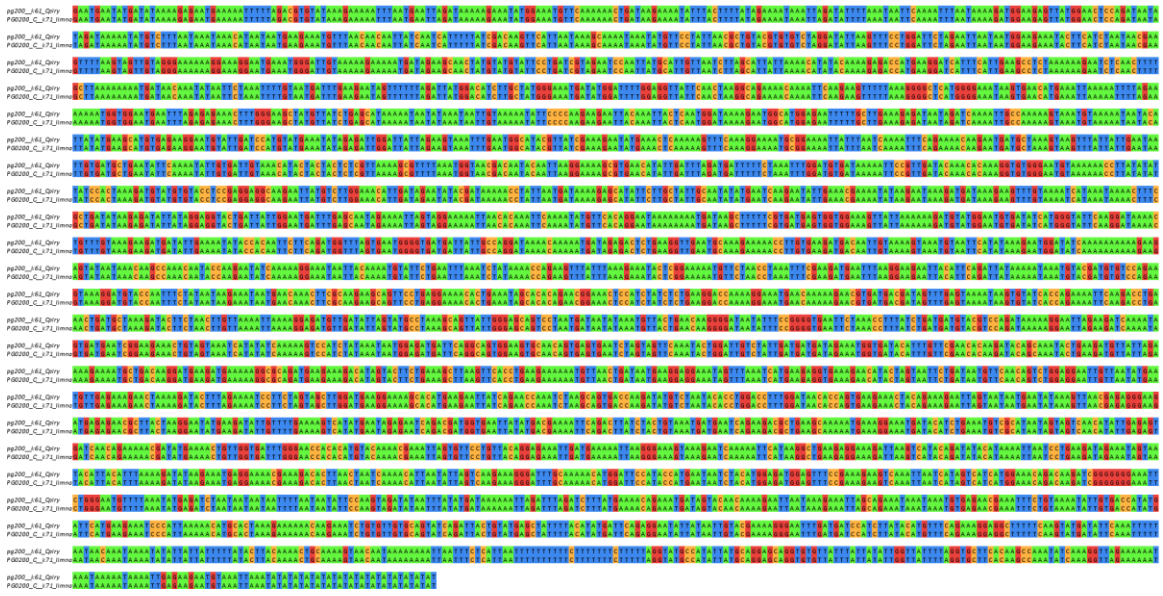
Supplementary figure 6-11. Capillary sequence validation of Dd2 assembly. Pairwise sequence alignment of a 4240bp fragment amplified and Sanger sequenced from *eba-175* with a *de novo* assembly from the same DNA source. Top sequence in the alignment is the capillary result. These sequences are 100% identical.



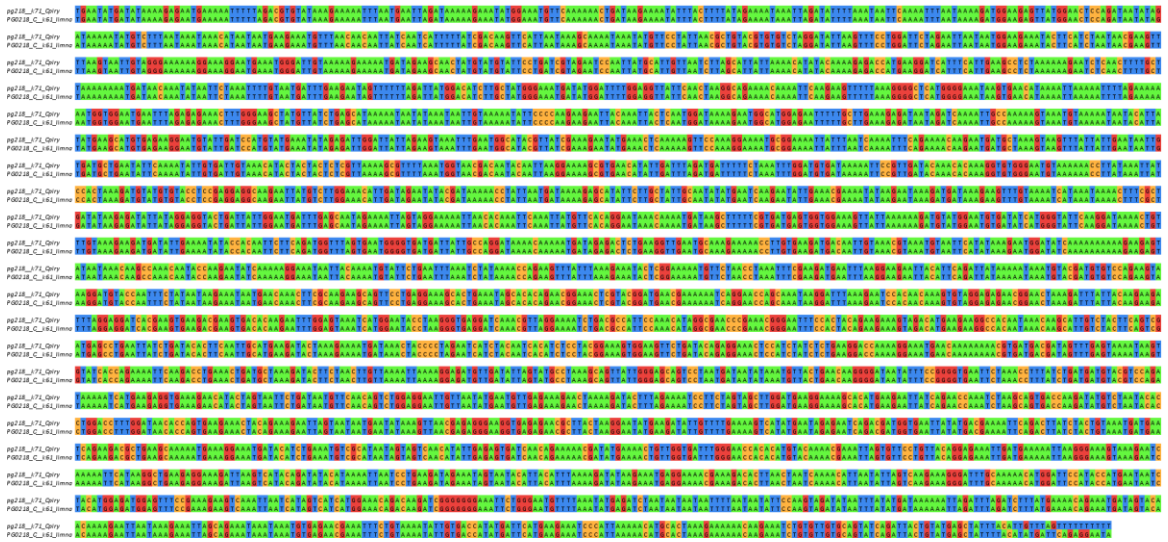
Supplementary figure 6-12. Capillary sequence validation of a field sample's assembly. Pairwise sequence alignment of a 2394bp fragment amplified and Sanger sequenced from *eba-175* with a *de novo* assembly from the same DNA source (PE0027). Top sequence in the alignment is the capillary result. These sequences are 100% identical.



Supplementary figure 6-13. Capillary sequence validation of a field sample's assembly. Pairwise sequence alignment of a 4461bp fragment amplified and Sanger sequenced from *eba-175* with a *de novo* assembly from the same DNA source (PE0028). Top sequence in the alignment is the capillary result. These sequences are 100% identical.



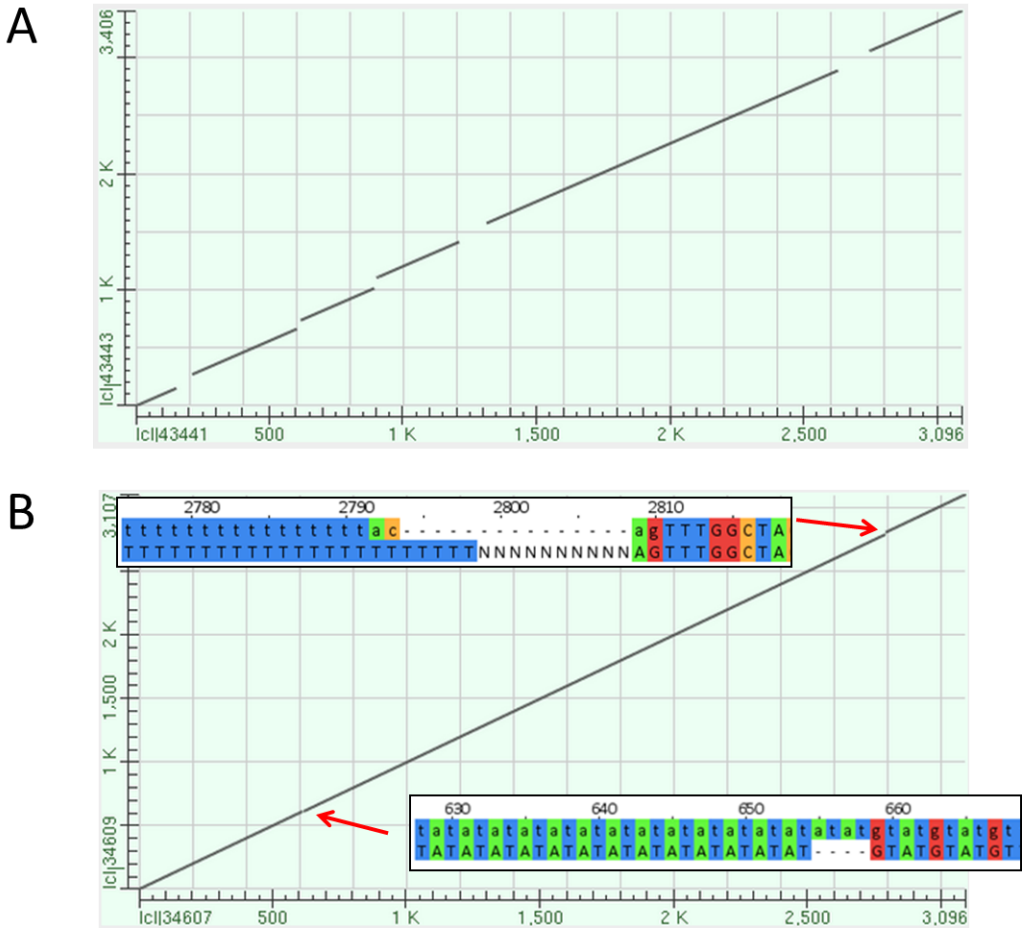
Supplementary figure 6-14. Capillary sequence validation of HB3 assembly. Pairwise sequence alignment of a 4400bp fragment amplified and Sanger sequenced from *eba-175* with a *de novo* assembly from the same DNA source. Top sequence in the alignment is the capillary result. These sequences are 100% identical.



Supplementary figure 6-15. Capillary sequence validation of 3D7 assembly. Pairwise sequence alignment of a 3312bp fragment amplified and Sanger sequenced from *eba-175* with a *de novo* assembly from the same DNA source. Top sequence in the alignment is the capillary result. These sequences are 100% identical for the first 3294bp. The final 18bp mismatch is attributed to read quality of the capillary data, not to an error in the assembly.

Supplementary figure 6-16. Capillary sequence validation of a field sample's assembly. Pairwise sequence alignment of a 2601bp fragment amplified and Sanger sequenced from *eba-175* with a *de novo* assembly from the same DNA source (PE0030). Top sequence in the alignment is the capillary result. 2518/2602 positions match in this alignment. The errors are attributed to a bad assembly due to MOI. This assembly was caught by the MalMOI filters.

Supplementary figure 6-17. A closer look at a singleton SNP based on assemblies of *eba-175*. A snippet of the *eba-175* multiple sequence alignment of *de novo* assemblies is shown in a location where one sample is different from all others ('g' vs. 'a' allele, circled in black). The inset panel shows the pileup of the sample reads onto the assembly (note the contig is in the reverse complement), and this singleton SNP is boxed in red. With a read depth of 33, all reads support the assembled genotype, suggesting that this is indeed a singleton SNP, rather than an assembly error.

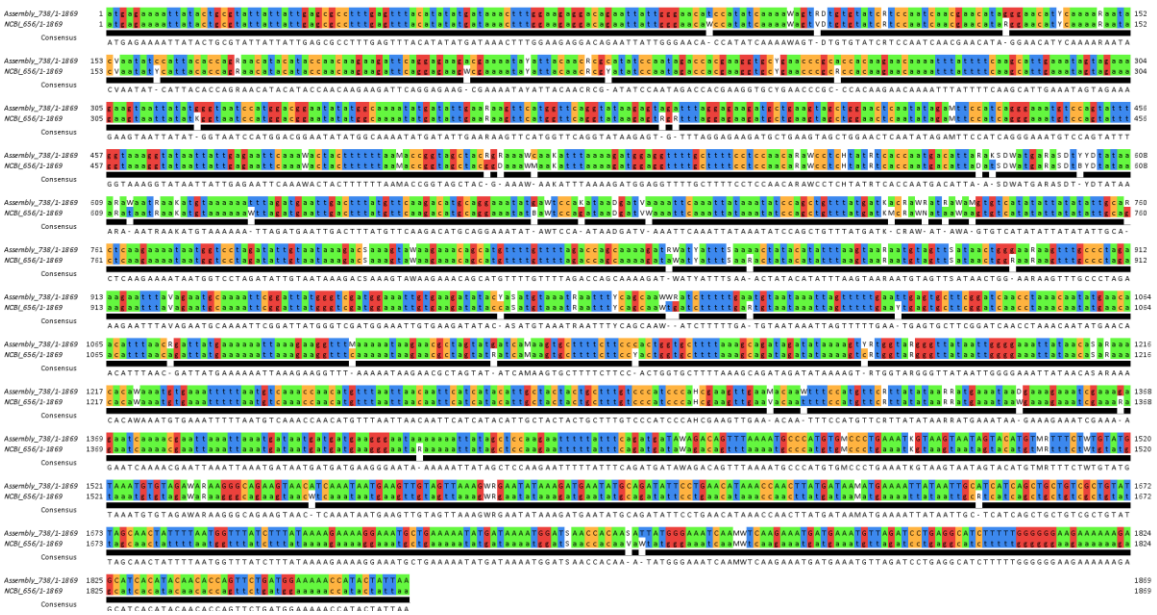


Supplementary figure 6-18. Insert-size matters in Velvet assembly setting. A) Dot-plot comparing the *pfcrt* reference sequence to a Velvet assembly of a 3D7 sample with the insert-size parameter set to auto. The gene is assembled in 6 contigs. **B)** Similar dot-plot but now comparing the reference to an assembly in which the insert-size was explicitly set as the median of that of all read-pairs in the BAM. Note that now only two gaps remain (red arrows), and are much smaller than above. The remaining gaps are within introns in areas of homopolymer and dinucleotide repeats (boxes). Boxes show the 3D7 sequence (top sequence in each box) aligned to the assembly.

6.7.1 Application: *ama1*

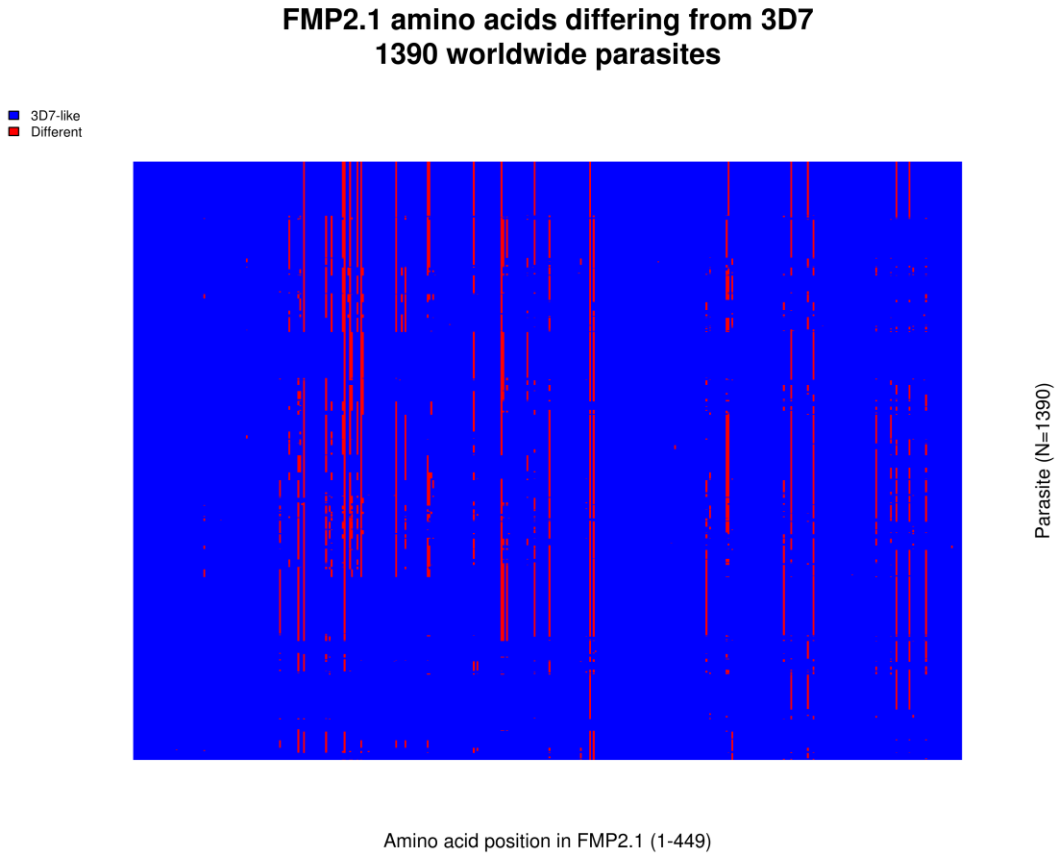
Although this chapter is focused on methods development, rather than the biology of the particular genes assembled, I highlight a few interesting discoveries in some genes that are not revisited later. Section III is dedicated entirely to applications of Section II tools to *eba-175*, so no elaboration is provided here for that gene. The blood-stage vaccine candidate *ama1* has been studied extensively, and already had a significant number of public sequences available. Comparing the 739 *de novo* assemblies to the 652 publicly available full-length *ama1* genes (downloaded from NCBI in July, 2014), at least 25 new variants are characterized. Outside of this, some positions are polymorphic in both sequence sets, but nonetheless a new variant may have been introduced. For example, at position 586 in the

alignment of the NCBI and *de novo* consensus sequences, an IUPAC R (A,G) in the MalariaGEN set is a D (A,G,T) in the NCBI set at that position—both are variable, but the T is new to the first set (Supplementary figure 6-19). There are 14 singleton SNPS in the new sequences. Fifteen SNPs in the NCBI database are not represented in the MalariaGEN assemblies.



Supplementary figure 6-19. IUPAC consensus sequence comparison of NCBI *ama1* to *de novo* assemblies. Each consensus sequences was generated from a multiple sequence alignment of either 652 *ama1* sequences or 739 *de novo* assemblies. The top sequence in each track of the alignment is the assembly consensus.

One of the applications of amassing a large number of full-length gene sequences is to the field of vaccinology. FMP2.1 is a vaccine antigen comprised of amino acids 83-531 of the 3D7 version of *ama1* [300]. Early growth inhibitory assay (GIA) evidence indicated that a vaccine against AMA1 might elicit allele-specific antibodies—a major concern in a gene with such a high degree of polymorphism. Indeed, a subsequent field trial measuring protection of Malian children from clinical disease showed that FMP2.1 was 64.3% efficacious when measured against the vaccine strain (3D7), versus 20% without stratifying by genotype [118]. Combining the NCBI and *de novo* assembled *ama1* sequences paints an unprecedented picture of parasite diversity in the AMA1 FMP2.1 region (Supplementary figure 6-20). Strikingly, this haplotype plot shows that the 3D7 form of this antigen has very low representation worldwide. This visualization and the underlying data are a valuable resource for selecting haplotypes to include in a multivalent AMA1 vaccine.



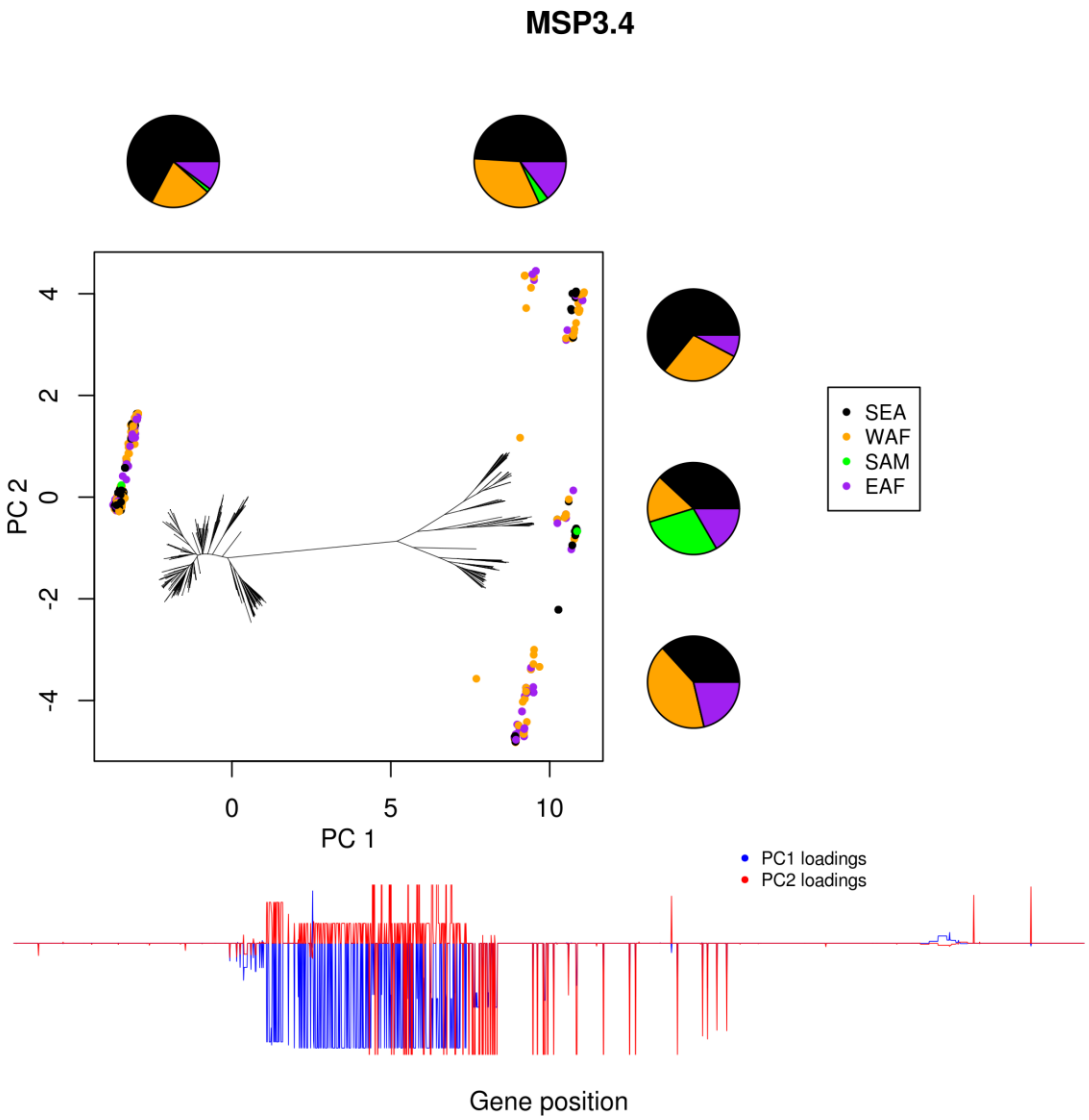
Supplementary figure 6-20. AMA1 FMP2.1 vaccine antigen haplotypes. Each row represents one of 1390 FMP2.1 protein sequences, either from NCBI or from the MalariaGEN *de novo* assemblies. Columns are amino acid positions. Cells are colored blue if that residue matches 3D7 and red otherwise.

6.7.2 Application: *msp3.4*

MSP3.4 (a.k.a: MSPDBL1, DBLMSP, PF3D7_1035700, and PF10_0348) is a much more recent blood-stage vaccine candidate. This protein, part of the MSP3 family, associates with the MSP1 complex on the surface of the merozoite, and along with the other DBL domain-containing merozoite surface protein (MSP3.8), it can bind directly to human erythrocytes [301]. Studies in Kenya and Papua New Guinea describe significant correlations between natural acquired immunity to MSP3.4 and protection from future clinical episodes of malaria [292,302]. The lead–up to identifying this protein as a vaccine candidate included genome-wide scans for genes evidenced to be under immune selective pressure [132]. Frequency-dependent balancing selection is an indicator of immune exposure, and thus a potential predictor for antigens involved in naturally acquired immunity [303].

In stark contrast to *ama1*, aside from reference sequences no full-length genes are publicly available for *msp3.4*. In this chapter I present 1390 full-length *de novo* assembled *msp3.4*

sequences from MalariaGEN samples contributed by partners across the globe. Genetic diversity studies of *msp3.4* fail to address its dimorphic properties. A multiple sequence alignment of these 1390 sequences reveals a strikingly divergent section of approximately 500bp from positions 465-985, interrupted by a perfectly conserved region in the single DBL domain, and just 5' of another completely conserved segment (see below). The dimorphic region is so diverged it is likely not homologous between the two forms, and thus should not be aligned for phylogenetic studies. This is important not only because of the bias it can introduce into population genetic analyses, but also could have implications for vaccine design. It would be useful to understand if these large dimorphic segments are under balancing selection, their geographic distributions, and which portions of the gene define the major forms. Supplementary figure 6-21 illustrates that the major driver of diversity in this gene is the dimorphic segment just 5' of center, encompassing the DBL domain. There are two remarkable segments of conservation within these seas of instability, NKGVLVPPRR (amino acids 183-192) and IPQYLRWFREWGTYVCSEYKNKFE, starting at position 329. That these peptides are 100% conserved in 1390 parasites taken from across the globe, and appear in the midst of regions that the parasite is under strong pressure to vary, is suggestive of selective constraint. Further, as these conserved sections are in or near the DBL domain that mediates erythrocyte binding and which is the target of inhibitory antibodies, suggests they may be attractive targets for a blood stage vaccine.



Supplementary figure 6-21. Global distribution and fine mapping of *msp3.4* dimorphic forms. Principal components analysis was performed on a multiple sequence alignment of 1390 *de novo* assembled *msp3.4* genes. The PCA was based on a binary matrix generated by pairwise comparisons of each sequence to an arbitrary sequence in the alignment. **Top)** Samples are projected onto the first two principal components and plotted as colored dots based on country of origin (see legend). A neighbor-joining tree based on the same multiple sequence alignment is overlaid onto the plot, and oriented to correspond to the PC clusters. The first PC accounts for the largest source of variation between samples. The cluster on the left of PC 1 is more condensed than the one on the right. The geographic distribution of these two clusters is illustrated in the pie charts above. Both dimorphic forms are present in all populations. PC 2 splits the tree into three apparent clusters, and the pie charts to the right show the geographic distributions of these groups. PC 2 may indicate some divergence due to geography, with the central cluster of sequences the only ones present in South America. **Bottom)** The values of the PC loadings for PCs 1 and 2 are plotted along the coordinates of *msp3.4* (nucleotide positions 1-2233). This indicates which regions of the gene are driving the stratifications depicted in the plot and tree. The dense blue lines show where in the gene the major dimorphism is located (which also stratifies the clusters in PC 1).

SECTION III: APPLYING TOOLS THAT DETECT COMPLEX VARIATION

In the previous section I developed two tools for accessing complex variation in a number of genes, including *eba-175*. In this section I apply both of these tools to *eba-175* to facilitate a population genetics study, and then to investigate host-parasite interactions. As I describe below, these analyses would not have been possible without novel methods.

7 EBA-175 POPULATION GENETICS AND HOST INTERACTIONS

7.1 Introduction

Three decades ago, perhaps to the month in which this thesis was submitted, Camus and Hadley would have been finalizing experiments and drafting a manuscript describing a 175 kilodalton parasite protein that correlates with merozoite invasion of erythrocytes in a sialic acid dependent manner [155]. The ensuing 30 years would see hundreds of reports attempting to characterize every imaginable aspect of the *eba-175* gene, and the 175kDa erythrocyte binding antigen it encodes. Motivated by the ubiquitous distribution yet unknown function of the F and C dimorphism in EBA-175, Binks, *et al* tested for associations of these indels with the M/N blood-group of its receptor, glycophorin A [175]. No associations were detected in that study, however the blotting technology used for genotyping the M/N alleles and their approach to classifying F/C mixed infections may have diminished their sensitivity. Further, this hypothesis-led experiment only tested for an interaction with one human locus. Here I take a GWAS approach to scan for host-parasite interactions. I affirm the conclusions of previous investigations that there is no evidence for a genetic association between EBA-175 and GYPA, and I expand the analysis to other host and parasite loci.

7.2 Aims

This chapter aims to revisit and expand on early investigations into the population genetics of EBA-175 and its potential interaction with host proteins using next-generation sequencing tools [175]. The following specific aims are explored:

1. Comprehensively describe the single nucleotide and structural polymorphism in *eba-175*, and create a universal reference sequence.

2. Identify specific regions of EBA-175 evidenced to be under balancing selective pressure.
3. Describe the global distribution of the F/C dimorphism and 6bp indel in *eba-175.*
4. Develop Sequenom assays to facilitate genome-wide human scans for host-parasite interactions.
5. Perform human genome-wide association studies, testing for relationships between human SNPs and parasite structural variants in Gambian and Kenyan populations.

7.3 Methods

The workflow below summarizes elements from both results sections (7.4 and 7.5) in this chapter, thus is placed here for convenience (rather than in the methods chapter). The other methods are outlined in section 2.7. The first part of the workflow describes how sequences were processed and filtered from chapter 6 to yield a trustworthy multiple sequence alignment for population genetic analyses. The MSA and polymorphism characterization from 7.4 feed into the second part of the workflow for Sequenom assay development, which in turn sets up the results for 7.5. Separately, this workflow has become a general tool that has been used for Sequenom assay design in other genes with complex variation.

7.3.1 Workflow

Sequenom assays use mass spectrometry to differentiate SNP alleles. First round primers amplify the genomic region of interest, and then a universal extension primer (UEP), designed to flank the SNP under scrutiny, extends by one nucleotide onto the polymorphism. Each SNP requires a separate assay, 30-40 of which can be multiplexed in the well of a 384-well plate. The degree to which assays can be combined depends on whether the spectral peaks yielded by the various single-extension amplicons can be resolved. In other words, the peaks of the different alleles on the mass spectrum must be far enough apart to differentiate, and the more SNPs in the multiplex, the more peaks on the spectrum. Many assays will fail to design using the Sequenom proprietary software, MassARRAY Designer, even before the multiplexing step. This can occur for many reasons—for example, surrounding polymorphism may prevent reliable primer design. Due to these constraints on multiplexing and individual assay design, it is beneficial to provide as many SNP options as possible to the software. Although the point of these assays is to genotype polymorphic sites, assays designed to type monomorphic positions can be informative if

they target sequence unique to an indel or divergent region. For example, although position 2401 of the *eba-175* IUPAC consensus is monomorphic, it types sequence within the F-insert that is exclusive to F-type parasites (Supplementary table 7-6). A null result in such an "indicator assay" could indicate an F-deletion, or could just be a failed assay. Therefore, it is important to combine indicator assays with one or more positive controls. I developed indicator assays and a pipeline for their design (Figure 7-1) to enable Sequenom typing of complex variants for which a direct assay cannot be found.

The early steps in this pipeline also generated the alignments used for the *eba-175* population genetic analyses. This pipeline was devised to work in general for designing Sequenom assays in genes containing complex variation, however I describe it using *eba-175* as an example for coherence. The 1893 MalMOI assemblies of *eba-175* that contained all 4 exons (i.e., the CDS assemblies from 6.4.2.2) were translated to protein, and those containing stop codons were dropped (N = 59). Genetic cross progeny and expansions of lab-lines that were sequenced for various experiments were removed (N = 415). The remaining 1419 protein sequences were aligned using MAFFT. Using the RevTrans software, this protein alignment was used to guide a DNA alignment of the CDS assemblies [304]. A neighbor-joining tree based on pairwise percent identity and, separately, a PCA were calculated based on the RevTrans alignment with gaps included in the distance metrics. Based on this tree and the first principal component, the sequences were split into two major groups, which as expected were defined by the F and C indels. This pipeline will be used on other genes with less well understood complex variation, so this result provides some level of validation to the approach. IUPAC consensus sequences were then generated based on the RevTrans alignment and the alignments of the major groups, and these were used to design Sequenom assays. Assays were designed for polymorphic and monomorphic positions (i.e., for every position in the gene with indicator assays based on fake SNPs where necessary). Nucmer was used to compare the primers designed for the two major groups to the opposite group's original alignment to designate group-specific assays. Positions for which assays could be successfully designed were then taken forward for multiplexing.

Although indicator assays are often necessary for typing complex variation, in some cases a direct assay can be engineered. For example, the eba-175 IUPAC sequence flanking the 6bp indel can be represented as follows:

TATATGAAGCATAAACAATTATTTAAAA[-/**T**TTCAR]**A**AAAM

The insertion and deletion are in brackets between flanking sequence. If a primer just upstream of the indel extended by one nucleotide, the assay would behave just like a T/A SNP (the first nucleotide in either the indel or in the 3' flank in the case of a deletion). Notice a primer could not be designed typing from the 3' direction because the IUPAC M indicates it would prime over another SNP. This strategy worked for the F and C indels as well, and the results from these direct assays are used for the host-parasite interaction GWAS.

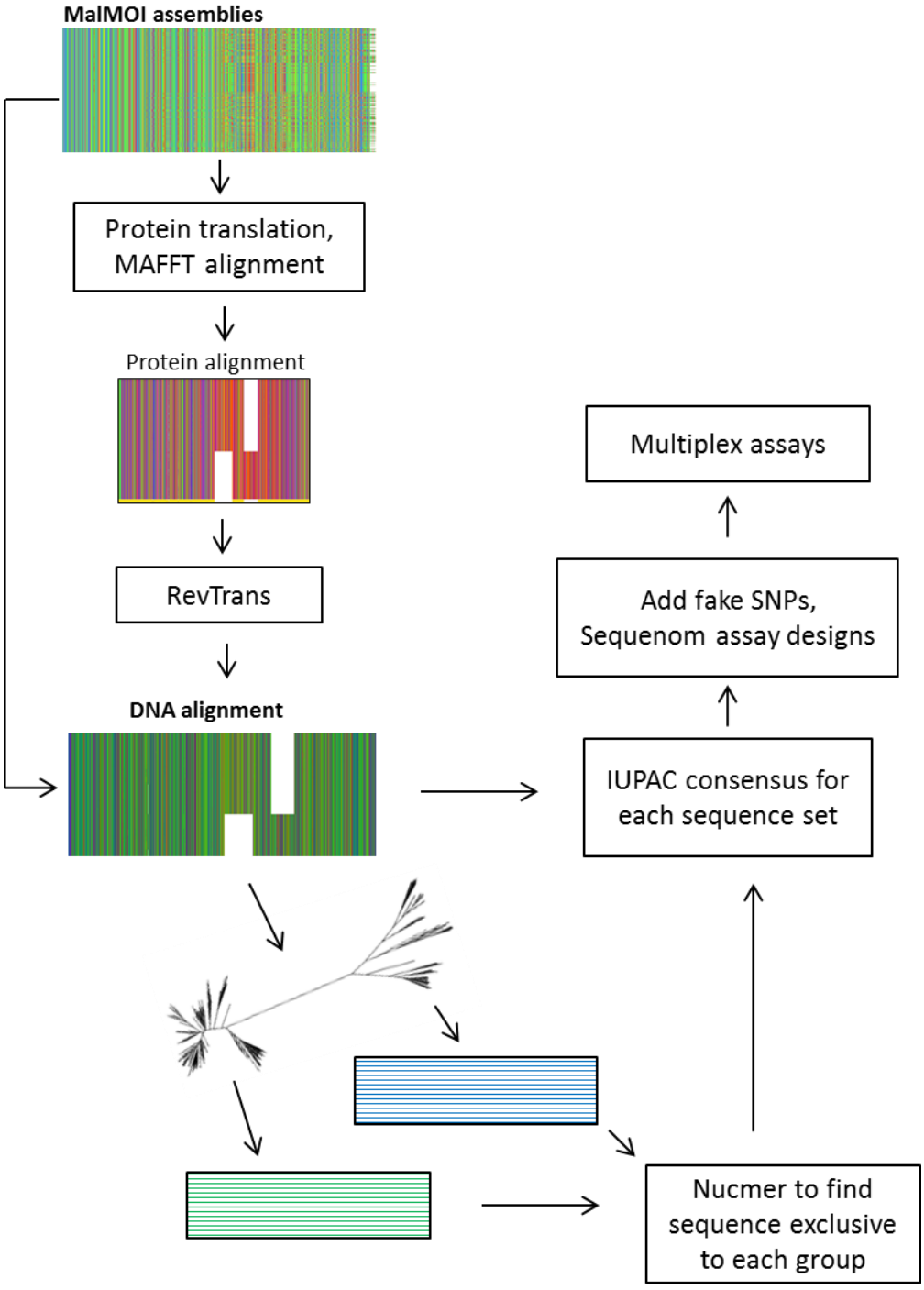


Figure 7-1. Workflow for MSA and Sequenom assay design. Details of the pipeline are described in the main text. Briefly, a multiple sequence alignment of the MalMOI assemblies is generated with guidance from the protein translations using RevTrans. Major complex variant forms are defined using neighbor-joining trees and PCA. Group-specific assays are designed, as well as assays using all sequences. The latter type of assay is preferred if one can be found that genotypes the complex variant, otherwise indicator assays are combined from each group. Finally, assays that successfully "design" are tested for multiplexing potential.

7.4 Results: population genetics

7.4.1 Comprehensive variation in *eba-175*

In addition to the three the structural variants highlighted in Figure 1-4, the coding regions of *eba-175* contain 121 single nucleotide polymorphisms in this dataset (Supplementary table 7-8). This includes 12 SNPs within the indels (i.e., positions in the F or C insert that vary between parasites of that respective class). The F-insert contains 9 low frequency SNPs (all below 1% MAF). The C-insert contains 2 SNPs, also with MAFs below 1%, and a nonsynonymous SNP present at 3.5% in these samples. This variant has not been previously described. It results in the substitution of a guanine to a cytosine at nucleotide position 3127 of the universal IUPAC reference (changing the amino acid from a glycine to an arginine), and is found in East and West Africa, as well as Southeast Asia. Including those located within the indels, only 14 SNPs do not result in a change to the protein sequence (i.e., they are synonymous substitutions).

Three positions (593, 812, and 1750) are tri-allelic SNPs. Position 1750 is particularly interesting, as all 3 variants occur at high frequency across Africa and Asia (Table 7-1). The 'G' allele doesn't appear in South America, though this could be a limitation of sample size.

Two complex codons were discovered in this multiple sequence alignment—i.e., codons containing more than one polymorphism in the triplet. One of these codons also contains the high-frequency tri-allelic SNP (position 1750). Along with the low frequency SNP, 1752, this complex codon can exist as AAA, AAG, GAA, GAG, CAA, and CAG. The last position in this codon is synonymous. Neither the AAG nor GAG haplotypes were detected in this dataset. The second complex codon occurs at positons 3640-3642, in which the first two of the triplet are bi-allelic SNPs. The four possibilities (TAT, TGT, GAT, and GGT) encode different amino acids (Y, C, D, and G), however no parasites were found to have a cysteine at this position.

Table 7-1. Global distribution of an *eba-175* tri-allelic SNP. Position 1750 of the IUPAC consensus is a tri-allelic SNP. All three alleles occur at high frequency in Africa and Asia.

	Nucleotide		
Country	A	C	G
BD	2	8	6
BF	1	2	3
CO	3	6	0
GH	27	125	16
GM	25	51	8
GN	21	39	6
HN	1	0	0
KE	6	15	2
KH	100	157	116
LA	45	18	15
ML	4	14	1
MM	46	13	18
MW	33	65	25
NG	0	2	0
PE	5	3	0
PG	3	5	15
TH	65	48	36
TZ	13	15	5
UG	3	4	1
VN	55	52	20

7.4.2 Linkage disequilibrium

Linkage disequilibrium (LD) refers to a situation in which the alleles in different polymorphisms on the same chromosome are not independent. In the case of comparing two bi-allelic SNPs (as is done pairwise for the *eba-175* variants below), the estimation of LD is reminiscent of a 2x2 contingency table analysis. The two most common metrics for reporting LD are D' and r^2, both of which are derived from a comparison of the observed frequency at which the alleles co-occur to the estimated probability if they were independent. LD metrics are biased if the MAFs greatly differ between the SNPs. D' attempts to control for this by normalizing by the maximum value possible, given the allele frequencies. While more susceptible to this allele frequency bias, r^2 is more intuitive, as it is the square of the familiar correlation coefficient between two binary vectors [305]. Low frequency SNPs are often excluded in LD analyses to avoid spuriously significant correlations, and to minimize bias using r^2, and that approach is taken below.

Genome-wide linkage disequilibrium (LD) decays rapidly with the distance between the variants in most *P. falciparum* populations. For example, average values of the standard r^2 metric rarely exceed 0.4 after 1kb in any population, and more typically decays to 0.1-0.2 at this distance [86,226]. Higher LD occurs in Southeast Asian and South American parasites compared to those from Africa—particularly West Africa, where there is little population structure. Transmission intensity also impacts LD. For example, lower average LD in Malawian parasites than in those from Cambodia and Thailand is attributed to much higher inoculation rates and MOI, consistent with higher transmission [306].

As depicted in Figure 7-2, LD decays rapidly with distance as expected (note these are individual pairwise SNP estimates, not mean values, and most points are near 0). Two associations are worth mentioning from this plot. The higher than expected LD between position 2303 and the C-insert is really marking a SNP that tags the F-insert with a high degree of association. Position 2303 is only 22bp upstream of the F indel, and has an r^2 = 0.77 with this structural variant. Although not a perfect tag for the F/C dimorphism, in 94% of samples a guanine at position 2303 marks an F-insert, and an adenine marks an F-deletion. This is a global calculation, thus the relationship may be stronger or break-down in specific populations, but there is clearly some level of recombination between these sites. As expected, the F and C indels are essentially in perfect LD (r^2 = 0.99). The few discrepancies may either be assembly errors or rare occurrences of double insertion and double deletion events (see supplementary section 7.8.4).

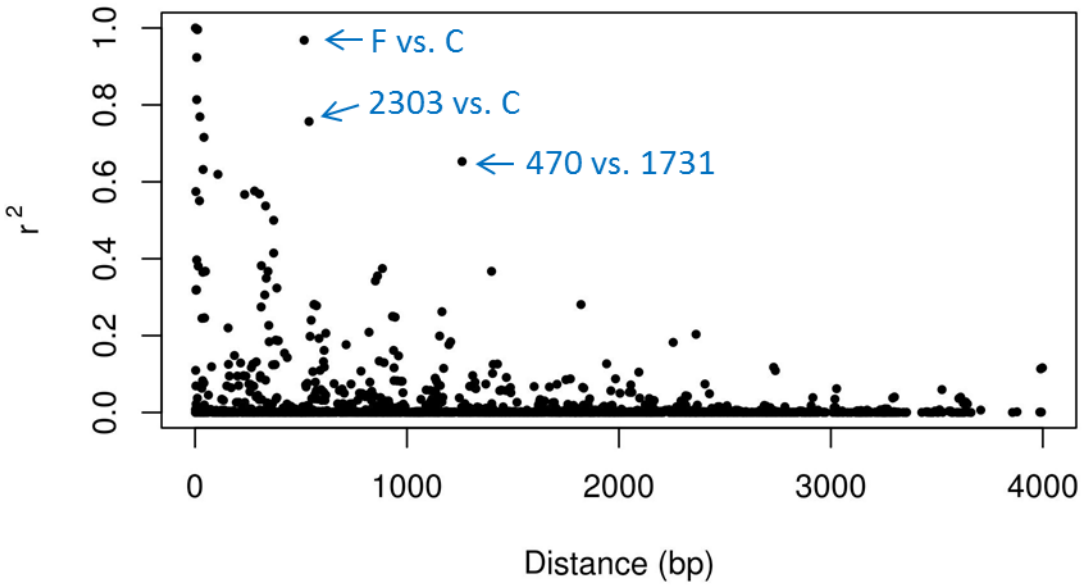


Figure 7-2. LD in *eba-175*. Pairwise r^2 values are plotted as a function of distance between all SNP and indel pairs with MAF>5%.

7.4.3 Signatures of selection

A signature of host immune selection on parasite proteins is that more intermediate-frequency polymorphism is maintained than is expected under neutral drift [163]. This form of positive selection balances allele frequencies because those antigens circulating at high prevalence in a population are negatively selected by acquired immunity, thus reducing the imbalance with low frequency polymorphism. The most widely cited statistic for testing for the presence of balancing selection is Tajima's D [307]. The numerator of D is essentially the difference between the number of SNPs observed compared to the number expected under neutrality, and this is divided by its standard deviation. Thus, D represents the number of standard deviations the observed minus expected SNP difference is above or below 0. Though not a level of significance, D > 2 is strongly suggestive of balancing selection and/or decreasing population size, while negative values indicate purifying selection or a population expansion. To identify regions of *eba-175* potentially under balancing selective pressure, Tajima's D was calculated for 100bp windows every 25bp across the gene (Figure 7-3, bottom panel). A maximum value of 3.63 occurs at the 3' end of the first DBL domain. This region contains 5 SNPs with MAFs ranging from 26-47%

surrounding the 6bp indel, which occurs at intermediate frequency in every population (Figure 7-7 and Supplementary table 7-8).

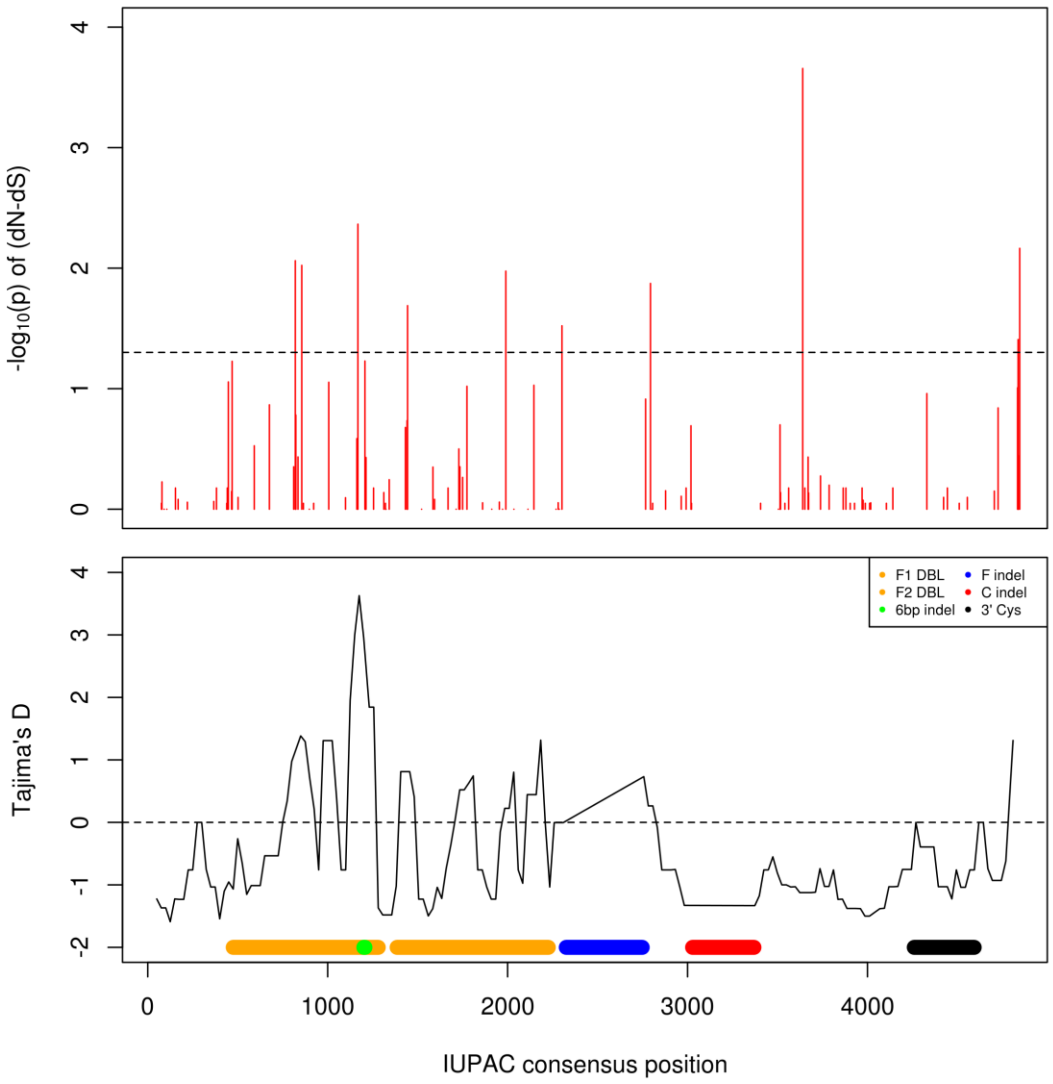


Figure 7-3. Signatures of selection in *eba-175*. Gene positions correspond to the IUPAC consensus sequence in both panels. **Top)** Per-codon -$\log_{10}$(p) values from a HyPhy test of dN > dS. Dashed line plotted at p = 0.05. **Bottom)** Tajima's D based on 100bp windows calculated every 25bp.

This calculation of Tajima's D excluded gapped regions of the alignment, which can be seen in Figure 7-3 as straight lines connecting windows on either side of the labeled indels. A separate analysis was performed in which an artificial SNP was introduced to represent the indel alleles to ensure the results were consistent. In that analysis, including the faux SNPs

raises the highest peak near the 6bp indel to 4.03, and results in slight increases near the F and C indels, however the overall interpretation across the gene remains the same.

Nucleotide substitutions that change the encoded amino acid can affect protein function and thereby parasite fitness. Comparing the rates of synonymous (dS) and nonsynonymous (dN) substitutions can be informative about the selective pressures on a protein. Changes to functionally constrained proteins will often be detrimental, and fewer nonsynonymous substitutions than expected (dS > dN) is an indicator of purifying selection. In contrast, dN greater than dS is suggestive of positive selection—e.g., the parasite is under pressure to maintain protein diversity or adapt in some direction. The manifestation of chloroquine resistance at *pfcrt* position K76T, as discussed in chapter 3, is an example of positive directional selection [308]. Diversifying selection can manifest through several mechanisms, one of which is pressure on immune epitopes [309]. To detect evidence of these forms of selection in *eba-175*, dN and dS were investigated codon-by-codon, as well as overall. The average dN/dS ratio across the entire CDS was 4.93. Codon-specific calculations are shown in the top panel of Figure 7-3. Statistically significant positions testing for an excess of nonsynonymous vs synonymous substitutions occur throughout the gene. Interestingly, while one of the highest peaks overlaps with that of Tajima's D (approximately positions 1150-1250), the most significant codon starts at position 3640 ($p = 0.0002$), where Tajima's D is negative. As elaborated upon in the discussion, this codon may be at an important intersection of selective pressures—i.e., immune pressure to diversify, but in a region under functional constraint.

7.4.4 Neighbor-joining trees

As a first step in investigating the relationship between *eba-175* polymorphism and the geographic distribution of parasites, neighbor-joining trees were constructed based on percent identity between sequence pairs. Figure 7-4-A shows a striking stratification of parasites, splitting the sample-set almost in half. The distance matrix used for this tree included the gapped positions in the alignment, indicating that the F, C, and perhaps the 6bp indels may be driving this tree structure.

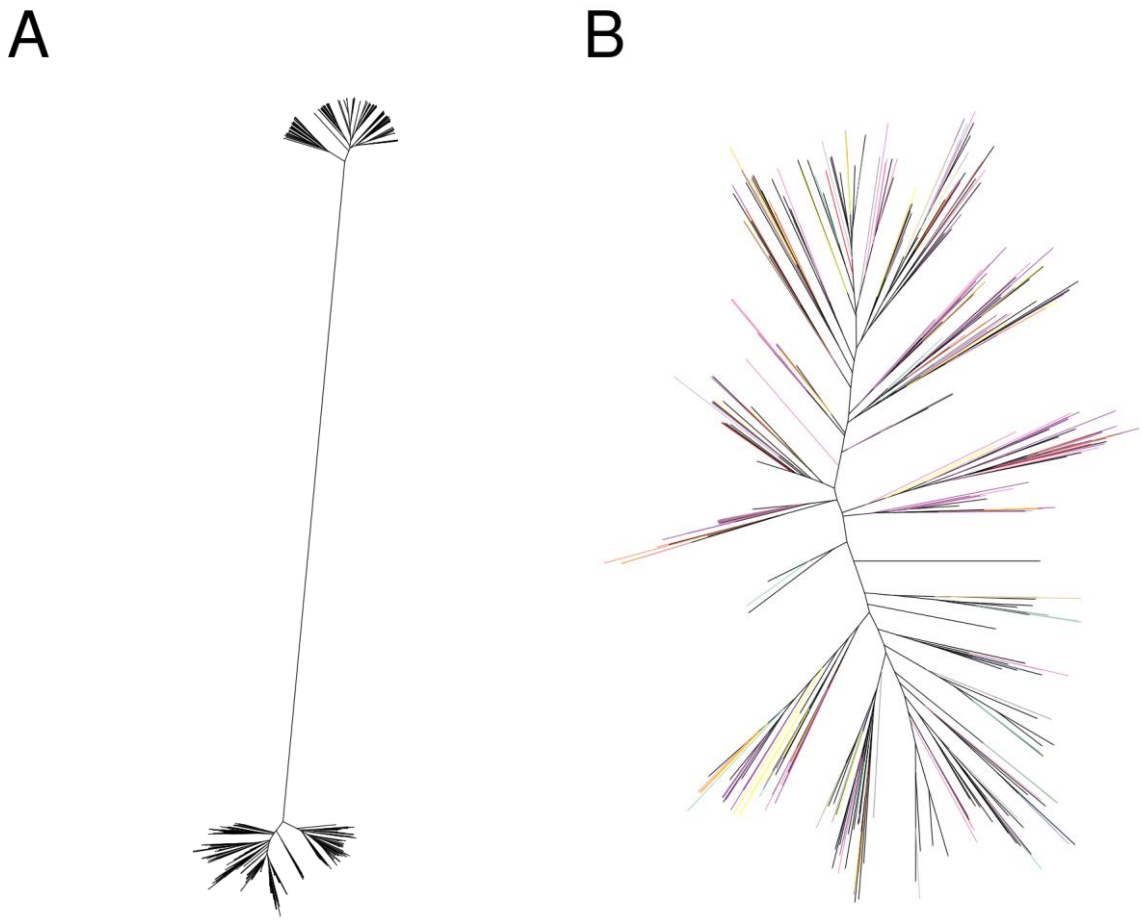


Figure 7-4. Neighbor-joining trees of 1419 *eba-175* CDS assemblies. Distance matrices were based on pairwise percent identities **A**) including all gaps in the scores and **B)** excluding gaps. In the tree excluding gaps, lines are colored by country of origin to show there is no clustering by country.

To assess this question with finer resolution, a PCA was calculated based on the same alignment, also retaining gaps. In addition to the traditional approach of visualizing each parasite's sequence projected onto the eigenvectors (e.g., a PC1 vs PC2 plot), the loadings of the eigenvectors were also visualized along the IUPAC consensus coordinates to determine which parts of the gene are driving the tree separations (Figure 7-5). As predicted above, the largest source of variation (represented by PC1) is overwhelmingly driven by the F and C indels. This is portrayed in the bottom panel of Figure 7-5 by the PC1 loadings (blue), which are large in relative magnitude spanning the F and C indel regions. As confirmation, position 2303 (shown in Figure 7-2 to tag the F indel 22bp upstream) also has a large loading in PC1—but is slightly smaller than the F-indel loadings, as predicted by its 94% tagging identity with that indel. All 4 geographic regions are present on both extremes of PC1 as illustrated in the two pie charts above the plot. This holds even when stratifying by

individual country (see Figure 7-6 discussed later). Consistent with balancing selection, the PC2 loadings are distributed across the gene and like PC1 do not stratify parasites by population. In other words, parasites from vastly different geographic regions share *eba-175* haplotypes to a similar extent as those from the same region due to immune selection and recombination. As described in the workflow Figure 7-1 this type of analysis is helpful for characterizing divergent regions that define parasite forms for Sequenom assay design.

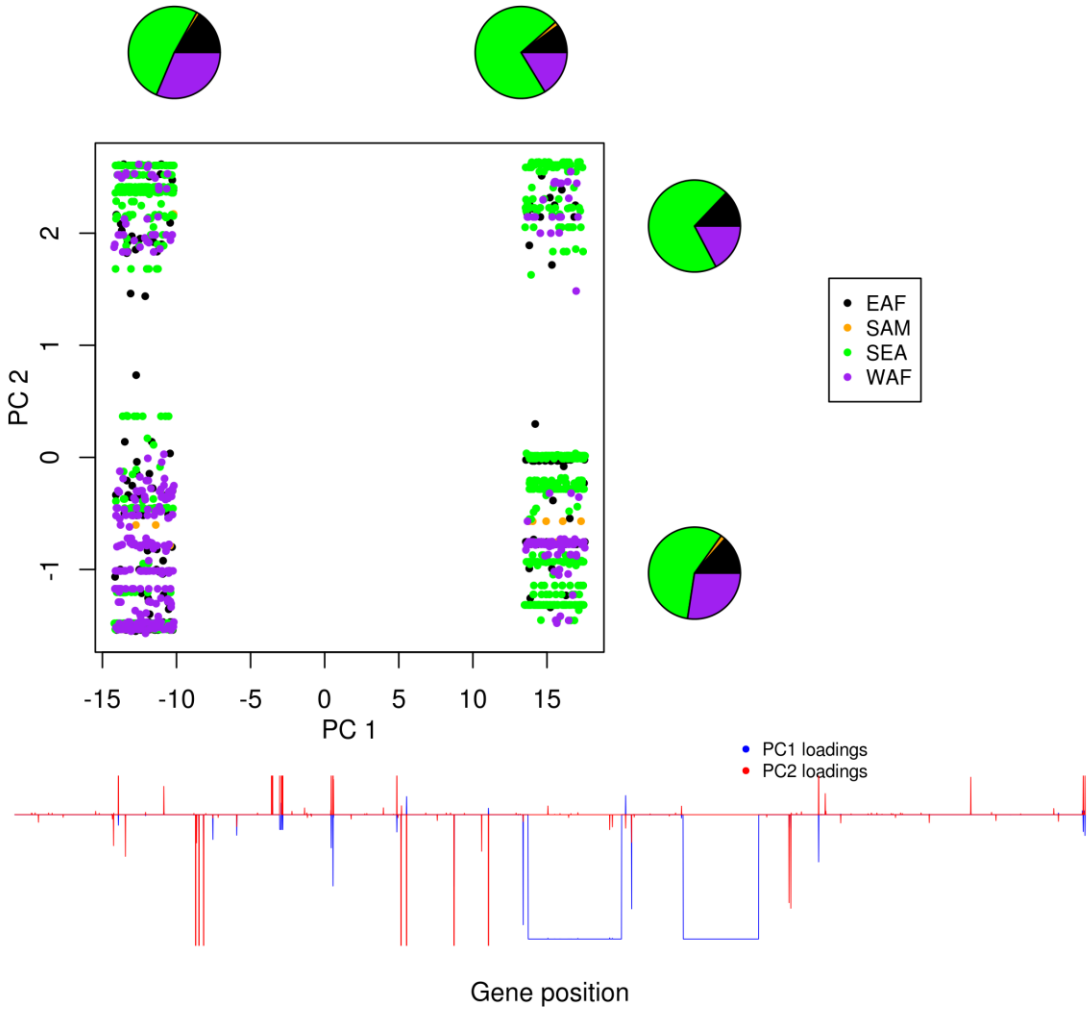


Figure 7-5. Global distribution and fine mapping of *eba-175* polymorphism. Principal components analysis was performed on a multiple sequence alignment of 1419 *de novo* assembled *eba-175* genes. The PCA was based on a binary matrix generated by pairwise comparisons of each sequence to an arbitrary sequence in the alignment, with gaps retained. **Top)** Samples are projected onto the first two principal components and plotted as colored dots based on country of origin. The first PC (PC 1) accounts for the largest source of variation between samples, and accounts for the major structure in Figure 7-4-A. The geographic distribution of these two clusters is illustrated in the pie charts above. Both dimorphic forms are present in all populations. **Bottom)** The values of the PC

loadings for PCs 1 and 2 are plotted along the coordinates of the IUPAC consensus. This indicates which regions of the gene are driving the stratifications depicted in the plot and tree. The long blue lines in the F and C regions show where in the gene the major dimorphism is located (which also stratifies the clusters in PC 1).

To further assess inter-population differences in allele frequencies, Fst values were calculated for all pairwise geographic stratifications (Table 7-2). The only comparison appreciably different from 0 was between SEA and WAF. PCA plots and loadings were evaluated for the first 10 principal components, but none indicated a pattern of differentiation between any of the populations.

Table 7-2. Fixation indices (Fst) between populations, averaged across SNPs.

Pop 1	Pop 2	Fst
EAF	**SAM**	0.025
EAF	**SEA**	0.067
EAF	**WAF**	0.001
SAM	**SEA**	-0.022
SAM	**WAF**	0.058
SEA	**WAF**	0.104

7.4.5 Genotyping *eba-175* indels on a population scale

MalMOI and Malign were used to genotype the F, C, and 6bp indels. Malign detects the presence of inserts even in mixed samples, whereas MalMOI only generates a single assembly from a highly abundant parasite in a mixture. Thus, MalMOI could produce a biased representation of population frequencies if the most abundant parasite form in a sample isn't random. On the other hand, this also provides a motivation to investigate a very interesting question—i.e., whether there is a pattern to which parasite (F or C, for example) is most dominant in mixed infections. This is investigated further with Malign output in section 7.5.3.

The most striking thing to note about the global distribution of the F and C indels is that both forms are maintained at high frequency in every population (Figure 7-6). This is a strong indication of balancing selection, and consistent with the distributions in Figure 7-5. This is also true for the 6bp indel—i.e., every one of the 18 countries depicted contains both

alleles (Figure 7-7). The F and C indels are dimorphic—i.e., parasites seem to be of one form or the other, and thus Malign can detect samples with both forms. In contrast, if Malign detects a 6bp insert in a sample there is no way of knowing whether a deleted genotype was present as well. A conserved region of the gene was used as a positive control for genotyping deletions (to distinguish from no coverage at all).

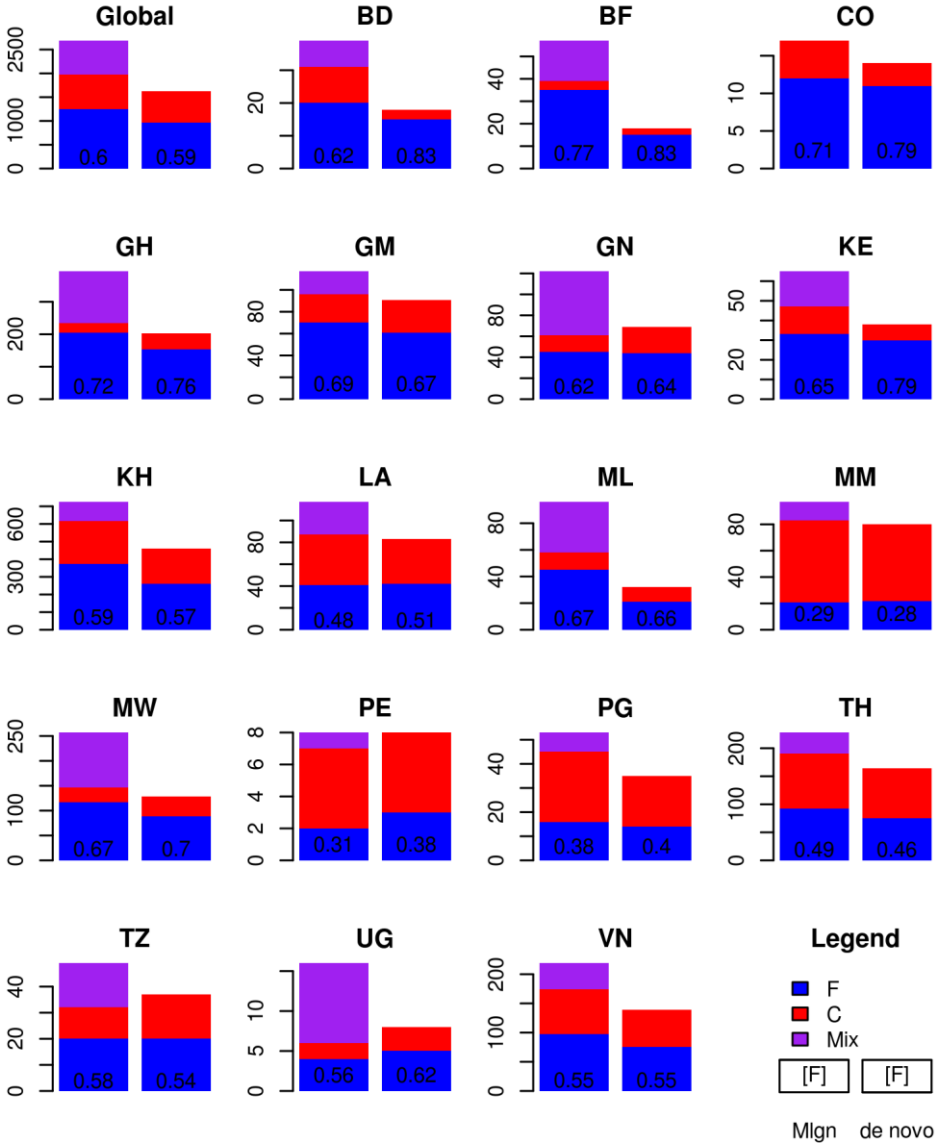


Figure 7-6. Global distribution of the *eba-175* F and C indels. For each country (and for all combined in the global panel), the summary of F (blue), C (red), and mixed (purple) genotypes are depicted as estimated by Malign (left bar) and MalMOI (right bar). The number printed at the bottom of each bar is the F-insert frequency. The y-axis shows counts. Country abbreviations are defined in the list of abbreviations and acronyms on page xxii.

In most populations the F-allele is more abundant than the C-allele, with notable exceptions in Myanmar (MM), Peru (PE), and Papua New Guinea (PN). Malign and MalMOI give generally similar trends, especially after accounting for F/C mixtures. This is consistent with a separate comparison of MalMOI frequencies for a SNP compared to MalariaGEN read counts (Supplementary figure 7-22). In every case MalMOI estimates a higher frequency of 6bp deletions than malign, which is expected because Malign does not detect deletion in mixed samples.

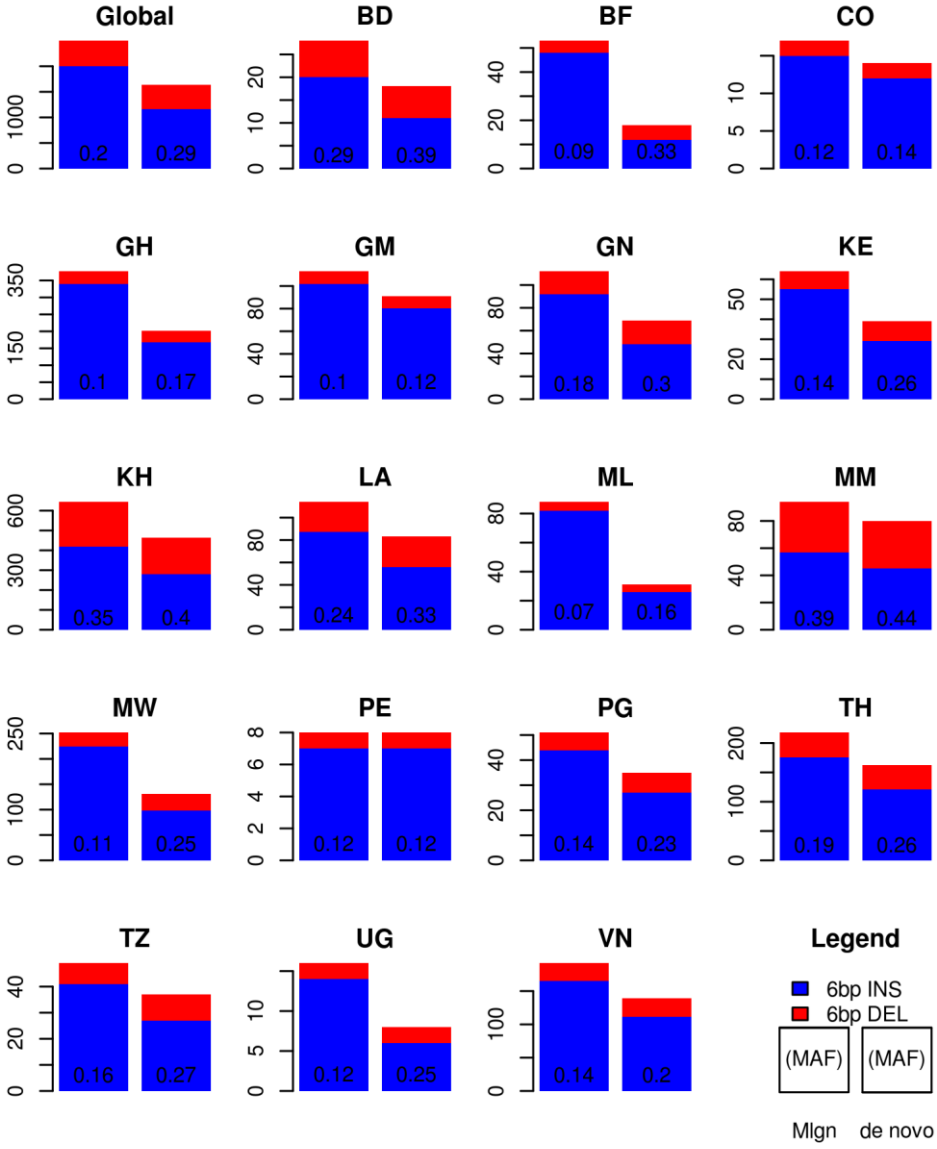


Figure 7-7. Global distribution of the *eba-175* 6bp indel. For each country (and for all combined in the global panel), the summary of the 6bp insertion (blue) and deletion (red) genotypes are depicted as estimated by Malign (left bar) and MalMOI (right bar). The number printed at the bottom of each bar is the minor allele frequency, which turns out to always be the deletion. Country abbreviations are defined in the list of abbreviations and acronyms on page xxii.

The 6bp insertion circulates at higher frequency in every population measured. This is supported by Sequenom data from Kenya and The Gambia using the direct assay (Table 7-3), although the Sequenom MAF is much higher for The Gambia using this approach.

Table 7-3. Sequenom genotypes for 6bp indel in Kenya and Gambia.

Population	Insertion	Deletion	Mix	MAF
Kenya	1161	264	371	0.29
The Gambia	1638	359	263	0.24

7.4.6 Mapping variation to protein structure

A decade ago, Tolia *et al* solved the crystal structure of EBA-175 Region II, providing detailed visualization of a dimer formed with the RII of a second EBA-175 in its interaction with glycan [158]. To assess the tertiary position of nonsynonymous polymorphism in this homodimer, SNPs and the 6bp indel described in 7.4.1 were mapped to this 3D structure (Figure 7-8). The 6bp indel is located at the end of the 10th alpha helix (yellow arrows), perhaps partially in the coil between helices 10 and 11. This position is near the end of the first DBL domain (F1) on an external part of the dimer. The highest Tajima's D peak in Figure 7-3 corresponds to an outer coil between helices 9 and 10, just upstream of the 6bp indel. The high-frequency tri-allelic SNP (IUPAC consensus position 1750) falls on helix T within the binding pocket of the dimer (white arrows). Interestingly, this SNP is described as being flanked directly on either side by residues involved in glycan binding [158].

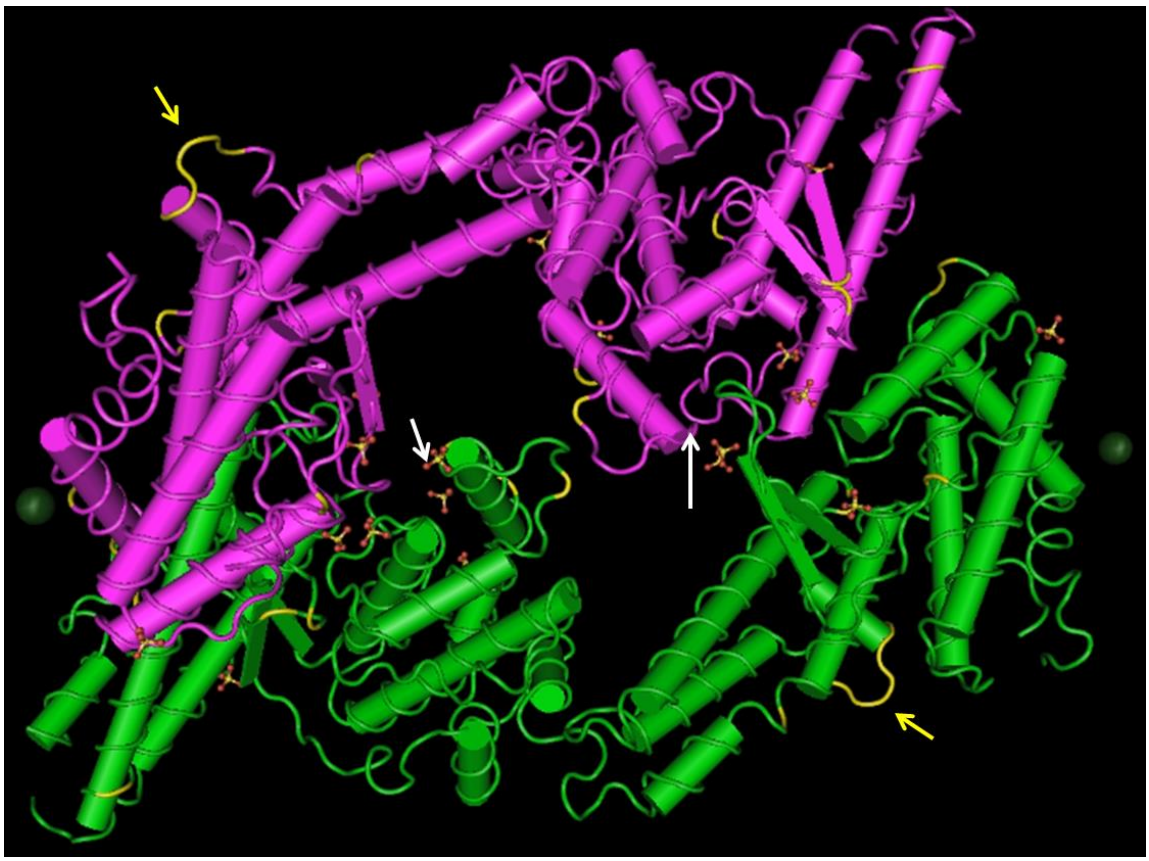


Figure 7-8. Polymorphism mapped to the EBA-175 region II crystal structure. The structure depicts RII from two EBA-175 proteins in a homodimer. Monomers are colored green and purple. Nonsynonymous variation defined in Supplementary table 7-8 is highlighted in yellow. Yellow arrows point to the 6bp indel positions. The indel consists of 2 amino acids (IS) at the C-terminal side of the 10th alpha helix (hJ). The indel comes off of the helix and begins the externally exposed coil. The two amino acids following IS are also highlighted, thus the insert itself does not extend as far into the coil as what is highlighted. White arrows (center) indicate the high-frequency tri-allelic SNP (IUPAC consensus position 1750).

7.5 Results: host-parasite interaction

In this section of the results I show the benefit of having a multiple sequence alignment of hundreds of *de novo* assemblies representing diverse MalariaGEN populations for host-parasite interaction studies. I developed a pipeline for Sequenom assay design that leverages the MSA, targeting divergent regions both directly and with assays that can impute complex genotypes (see 7.3.1). Following assay design I perform a human GWAS, scanning for human associations to the complex genotypes of the infecting parasite.

7.5.1 Human genome-wide associations with the 6bp indel

Sequenom assays for the 6bp indel yielded a genotype for 2226 Kenyan and 2764 Gambian samples that contained both host and parasite DNA. Analogous to a traditional case-control

association analysis, parasite genotype was analyzed as a binary outcome variable—i.e., 1 = 6bp insert and 0 = 6bp deletion, and therefore mixed samples were excluded (N = 811). Of the remaining samples, 1011 from Kenya (224 deletions and 787 insertions) and 1883 from Gambia (328 deletions and 1555 insertions) also had genome-wide SNP information available for the host (see methods section 2.7.7).

Association tests were performed on each population separately. Imputation of Illumina Omni2.5M data onto the 1000G reference panel resulted in tests for association of the 6bp indel with over 10 million human SNPs. Five SNPs in the Kenyan cohort reached genome-wide significance of $p < 5x10^{-8}$ (Figure 7-9, Table 7-4) [310]. All five of these SNPs fall within an 8kb intergenic region on chromosome 14. While no SNPs in the Gambian cohort reached this level of significance, an interesting spire in the $p < 5x10^{-7}$ range occurs on chromosome 7 (Figure 7-10). This cluster of Gambian SNPs spans a 25kb region that is 50kb downstream of a glucuronidase pseudogene (GUSBP10). No signal is apparent in this region in the Kenyan cohort (Supplementary table 7-9).

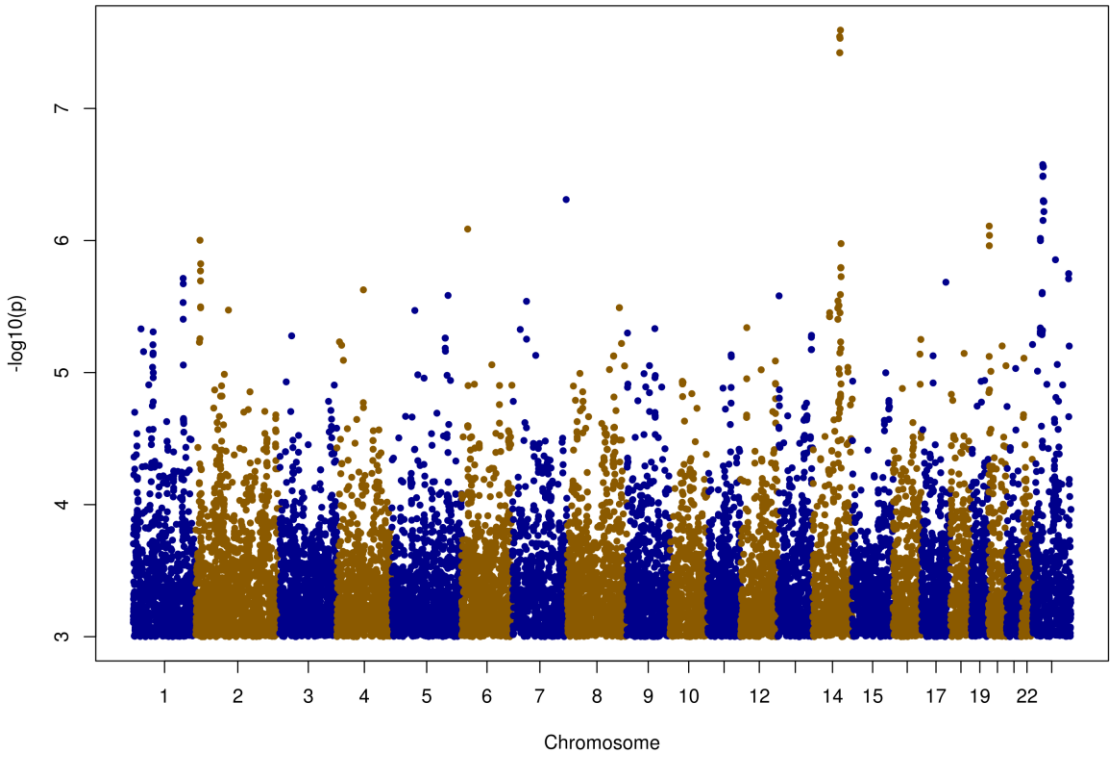


Figure 7-9. Manhattan plot of the *eba-175* 6bp indel associations in Kenyan samples.
Chromosome X is to the far right in dark blue.

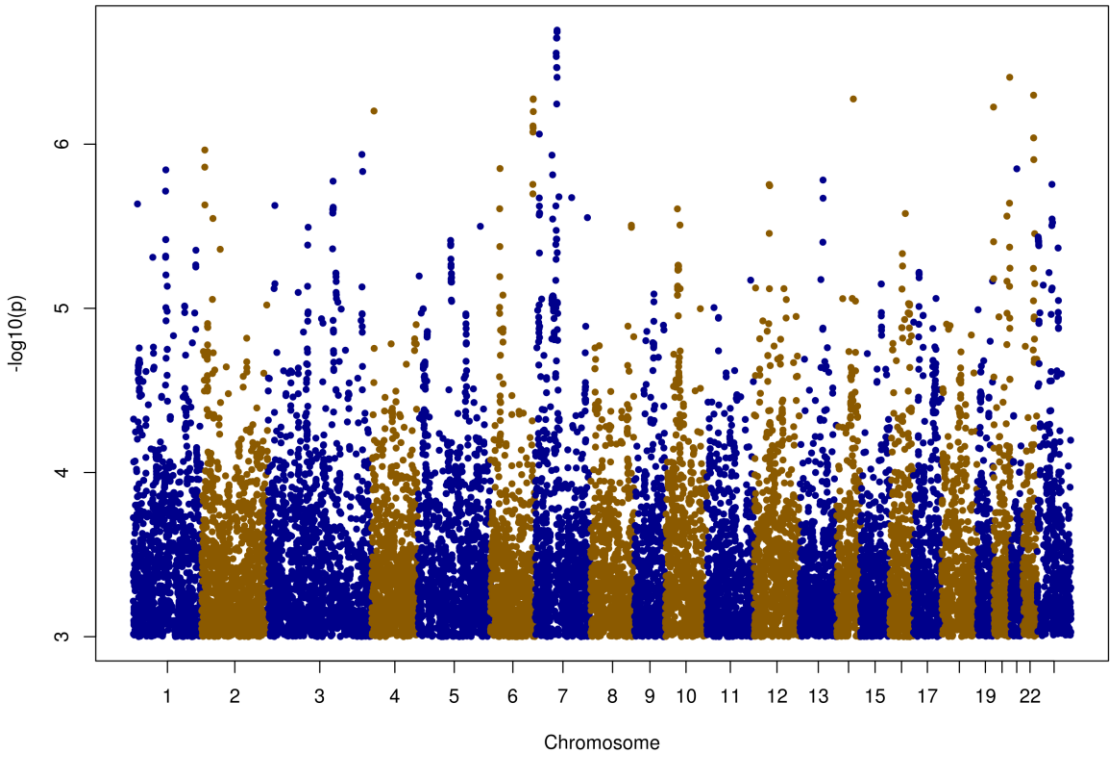


Figure 7-10. Manhattan plot of the *eba-175* 6bp indel associations in Gambian samples. Chromosome X is to the far right in dark blue.

To avoid following false positive leads I focused on the region of chromosome 14 yielding genome-wide significance, and produced a chromosome-level Manhattan plot combining both cohorts (Figure 7-11). This plot reveals that the Gambian population has some evidence of association in this region as well, with one SNP reaching a significance of $p = 5.03 \times 10^{-7}$ (Table 7-4). The five significant Kenyan SNPs span an 8kb intergenic region in chromosome 14q22, directly between estrogen-related receptor beta (ESRRB) and vasohibin 1 (VASH1); approximately 200kb from each gene (Figure 7-12).

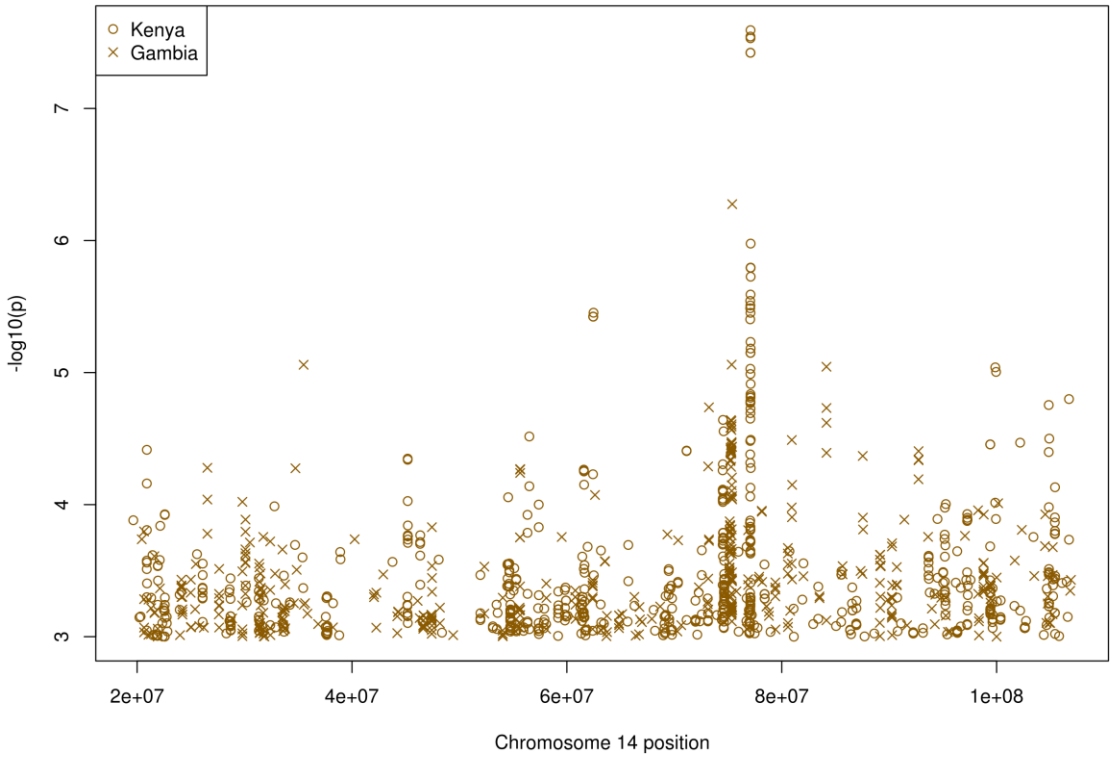


Figure 7-11. Manhattan plot combining Kenyan and Gambian results for chromosome 14. Circles represent Kenyan results and "x" characters depict those from Gambian.

Table 7-4. Chromosome 14 top GWAS candidates. Combined results for the Kenyan and Gambian populations

Position	Country	rsid	MAF	p-value
75392298	Gambia	rs11159110	0.118696	5.30E-07
77087221	Kenya	rs12590729	0.11746	2.85E-08
77087672	Kenya	rs11159218	0.121374	3.78E-08
77087730	Kenya	rs11159219	0.121374	3.79E-08
77090397	Kenya	rs17104696	0.117438	2.94E-08
77095148	Kenya	rs4506820	0.117705	2.55E-08

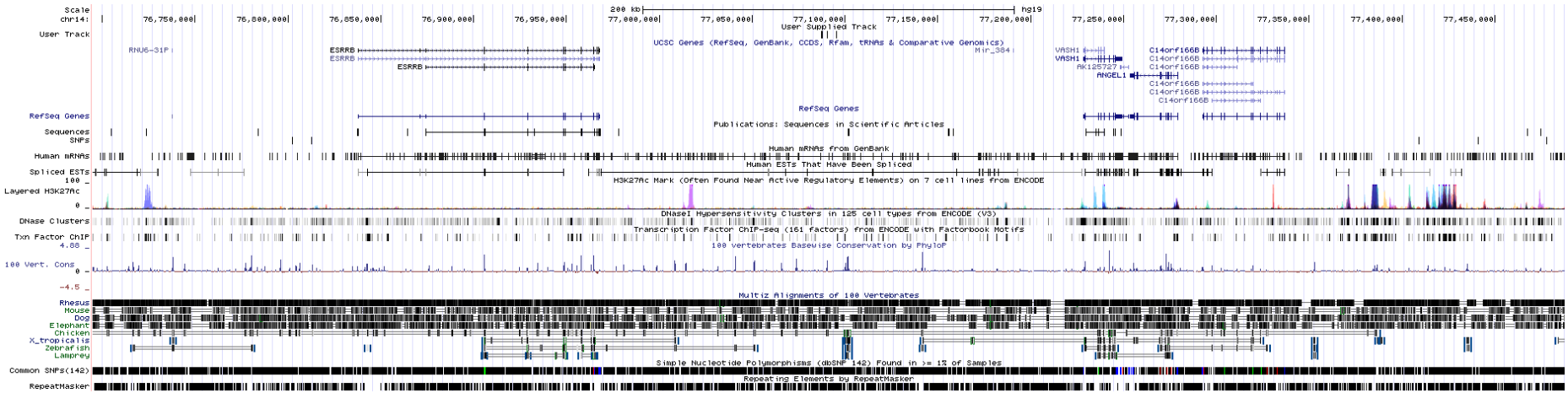


Figure 7-12. UCSC Genome browser view of the significant Kenyan GWAS region [311,312]. The user track shows the location of the five significant SNPs (top track, center of the plot). The next track down shows the closest Refseq defined genes (ESRRB and VASH1).

The 1.7Mb region spanned by the significant SNPs from both cohorts contains 18 genes defined by Refseq (Table 7-5).

Table 7-5. Genes near significant SNPs in Kenya and Gambia. SNPs were tested for association with the *eba-175* 6bp indel. The genes in the chromosome 14q22 region spanned by the most significant SNPs are listed below.

Gene name
placental growth factor (PGF)
eukaryotic translation initiation factor 2B, subunit 2 beta, 39kDa (EIF2B2)
mutL homolog 3 (E. coli) (MLH3)
acylphosphatase 1, erythrocyte (common) type (ACYP1)
zinc finger, C2HC-type containing 1C (ZC2HC1C)
zinc finger, C2HC-type containing 1C (ZC2HC1C)
NIMA-related kinase 9 (NEK9)
transmembrane emp24-like trafficking protein 10 (yeast) (TMED10)
FBJ murine osteosarcoma viral oncogene homolog (FOS)
Jun dimerization protein 2 (JDP2)
basic leucine zipper transcription factor, ATF-like (BATF)
feline leukemia virus subgroup C cellular receptor family, member 2 (FLVCR2)
tubulin tyrosine ligase-like family, member 5 (TTLL5)
transforming growth factor, beta 3 (TGFB3)
intraflagellar transport 43 homolog (Chlamydomonas) (IFT43)
G patch domain containing 2-like (GPATCH2L)
estrogen-related receptor beta (ESRRB)
vasohibin 1 (VASH1)

7.5.2 Human genome-wide associations with the F/C dimorphism

As the F and C indels are at different loci, two different direct Sequenom assays were used to genotype them. However, these indels are dimorphic, and thus one F/C state was derived from a combination of these assays. I took the most conservative approach and only retained samples for which the two assays were concordant. Sequenom assays for the F indel yielded a genotype for 1613 Kenyan and 1882 Gambian samples. After filtering for parasitemia (see methods section 2.7.7) and concordance, 2006 samples remained (C: 439, F: 1567). After matching with those samples for which human genotypes were also available, 375 from Kenya (C: 100, F: 275) and 1301 from Gambia (C: 275, F: 1026) remained for the final analysis.

No human SNPs met genome-wide significance in association with the F/C genotype of the infecting parasite in either cohort (Figure 7-13, Figure 7-14). Further, no signal of association was found in the GYPA region of chromosome 4.

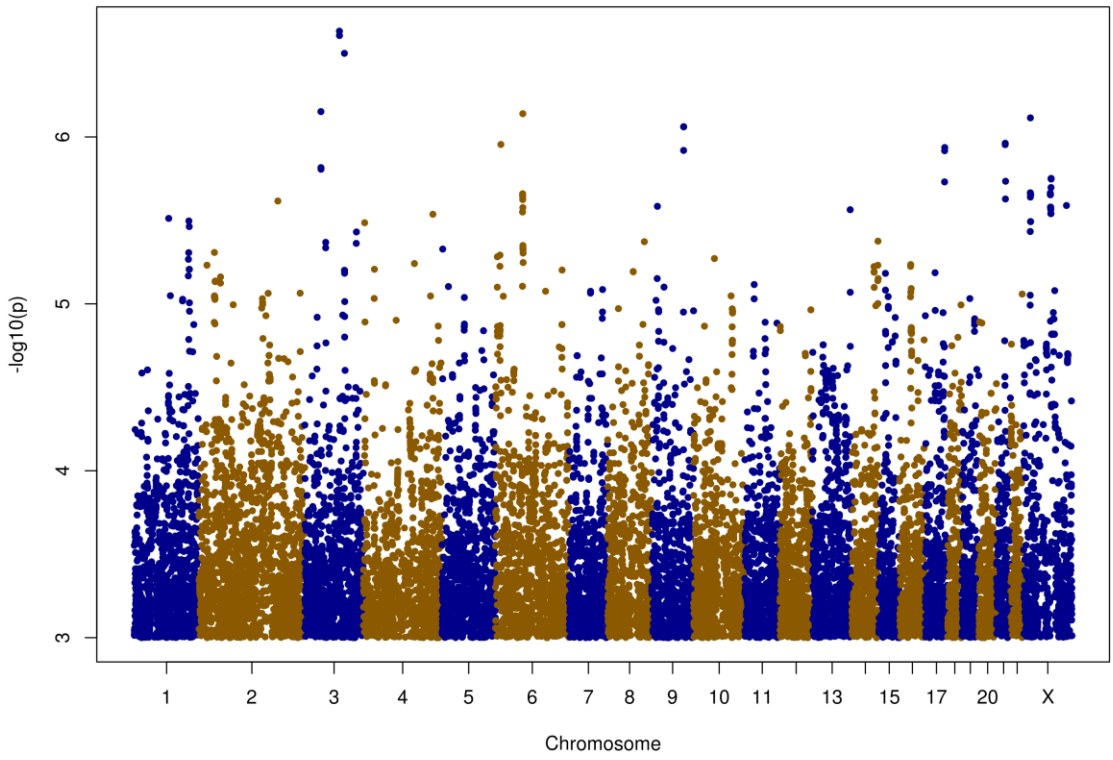


Figure 7-13. Manhattan plot of the *eba-175* F/C indel associations in Kenyan samples.
Chromosome X is to the far right in dark blue.

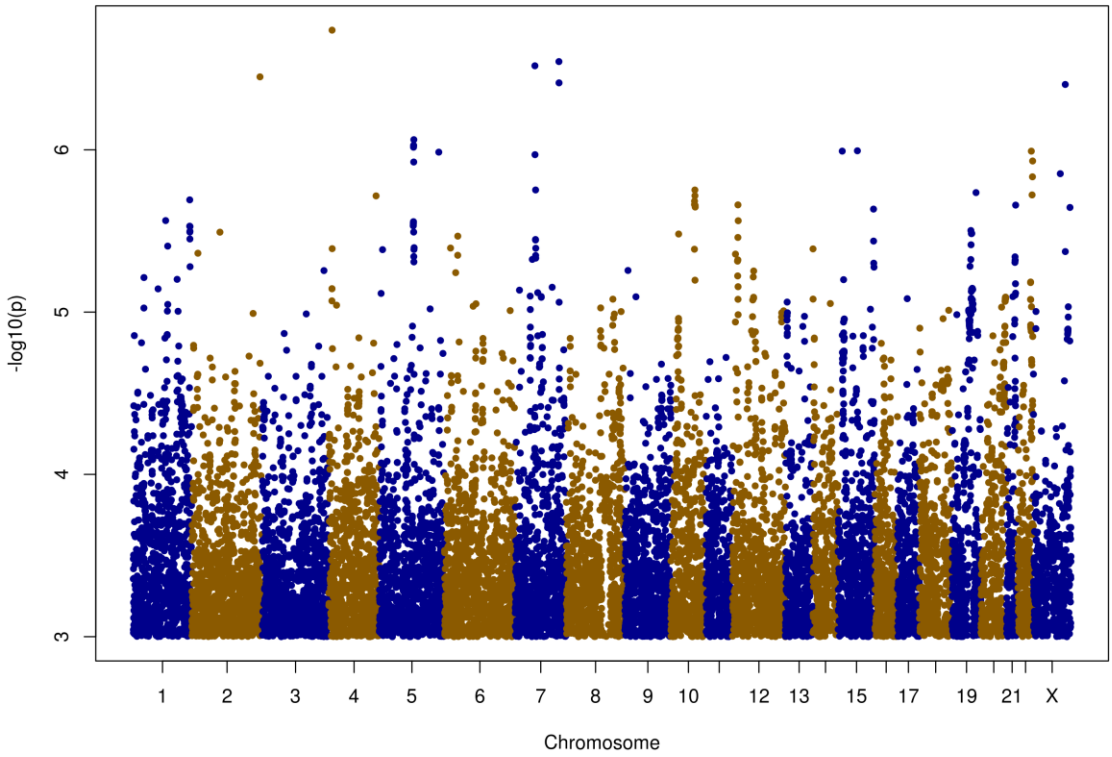


Figure 7-14. Manhattan plot of the *eba-175* F/C indel associations in Gambian samples. Chromosome X is to the far right in dark blue.

7.5.3 F and C relative abundance within mixed infections

To investigate whether F and C-type parasites might have differential fitness when competing in the same infection, Malign read depths were used to compare the relative abundance of F and C parasites in mixed samples. Although in Ghana (GH) it appears there is a statistically significant difference between the F and C abundances when found in the same infection (Figure 7-15), generally the results are unconvincing—especially after correcting for multiple hypothesis testing. Without broader context one might speculate that if an infection contains a mixture of both F and C type parasites, one form is overwhelmingly dominant at any given snapshot. This is depicted visually in Figure 7-15 by lines (representing mixed infections) with steep slopes. This would be particularly exciting if longitudinally the same parasite form was consistently dominant in a given host—possibly indicating that specific host genotypes better control specific parasite forms. However, this observation was recapitulated when tested on several other polymorphisms, and thus some (or all) of this effect may simply result if one parasite at a time tends to cycle in dominance in a mixed infection.

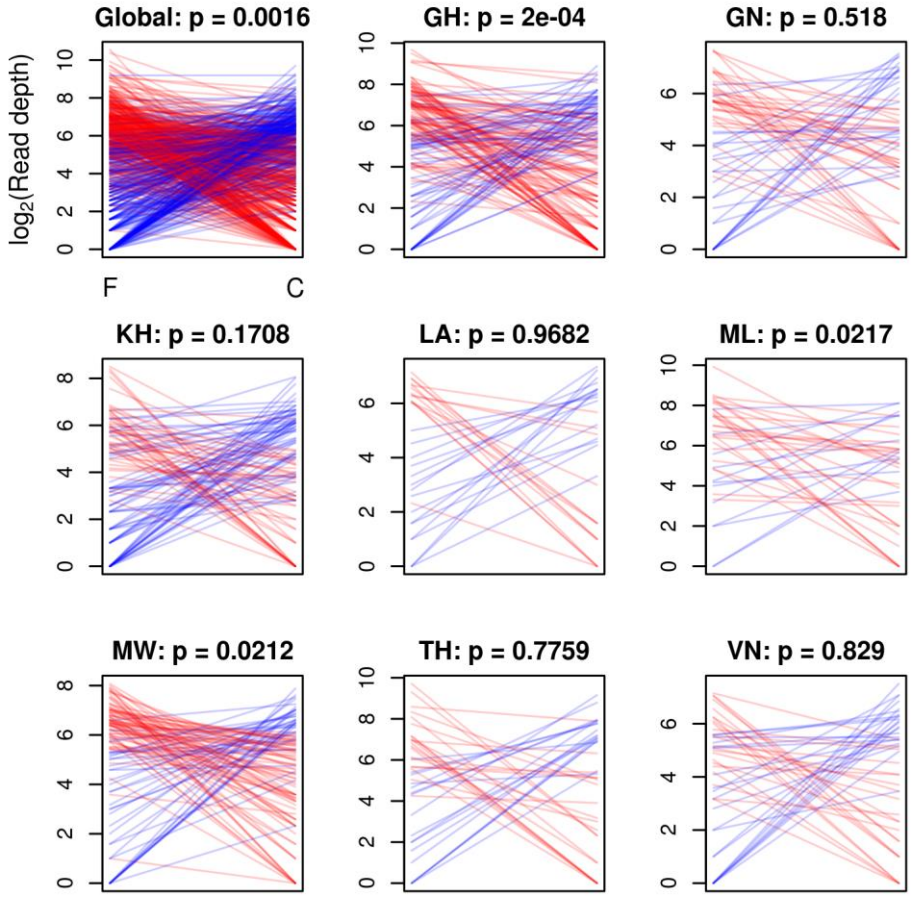


Figure 7-15. Comparing F vs C parasite abundance within mixed infections. Each blue or red line represents a single infection in which both an F and C parasite were detected by Malign. The Malign read depths for the F and C indels are then connected from left to right by each line. Lines are colored red if F is more abundant, and blue if C is more abundant. Each panel represents a country that had at least 30 parasites with an F/C mixture. Depths are $\log_2$ transformed at the p-value represents a paired t-test of the difference of these transformed depths within a sample.

7.6 Discussion

7.6.1 Population genetics and molecular evolution

One of the most striking population characteristics about *eba-175* is that all alleles of the F, C, and 6bp indels occur in every population surveyed (Figure 7-6 and Figure 7-7). In most populations the F-form is at higher frequency than the C-form of *eba-175.* Consistent with a previous study near the Brazilian Amazon that reported a higher frequency of the C-insert over a 15 year period, Figure 7-6 shows a higher frequency of the C-insert in neighboring Peru [172]. The fact that the C-form was at higher frequency over time in this study is interesting, as it supports the idea that particular alleles are better suited to specific populations. A separate study from 2010 reports a similar dominance of the C-form in

Brazil [313]. The previous study also suggests that inter-population divergence (Fst) is generally low for the *eba-175* dimorphism, which is supported by other reports and indicative of frequency-dependent selection [175]. Although balancing selection may indeed maintain both alleles in all populations, it doesn't explain why some populations appear consistently to have a dominant form. A study in Laos in 2003 found that F and C-type parasites were equally abundant in the south, but that the F-fragment was more abundant in the north [314]. The MalariaGEN parasites reported here were collected in central and southern Lao, and consistent with this previous study from more than 10 years ago, the F and C forms are equal in frequency.

The 6bp insert is far more abundant in every population than is the deletion (Figure 7-7). This indel has received far less attention than the dimorphism, but the sequences used in a 2006 study would appear to support this trend [315].

A previous study comparing the *P. falciparum eba-175* sequence to an ortholog suggests that the *P. reichenowi* gene is the F-form [316]. Aligning the *P. reichenowi* sequence to the translated d*e novo* assemblies indeed supports that *P. reichenowi* contains the F-insert (Supplementary figure 7-18), however, there is also evidence that part of the C-insert is present in this species as well (Supplementary figure 7-19). In fact, with the exception of a 36bp deletion in the *P. reichenowi* C-fragment, the level of amino acid conservation is similar to that of the F-fragment. This is interesting because it suggests that the F/C dimorphism may have a history of ancient balancing selection.

7.6.2 Signatures of selection

Individuals living in areas of stable malaria transmission naturally acquire protective immunity to disease as well as to parasitemia [30,76]. Although many parasite proteins are antigenic, not all elicit protective immune responses. Identifying genes with signatures of balancing selection is predictive of which antigens comprise the natural immune response, providing leads on those that may be protective. Parasite proteins that evoke acquired protective immunity in natural infections are rational candidates for vaccines [131]. However, it should be noted that evidence of balancing selection is by no means a prerequisite for a promising blood-stage vaccine target [317,318]. Interestingly, GYPA (the most abundant glycoprotein on the erythrocyte surface and target receptor for EBA-175 during invasion) also exhibits strong signatures of diversifying selection. However, this may be because sialylated glycoproteins on the red cell surface serve as decoys to viruses and bacteria that would otherwise invade nucleated cells [319].

Previous studies have identified *eba-175* as being under strong positive selective pressure, particularly the region containing the DBL domains (Region II) [164,239]. In fact, genome-wide scans identify *eba-175* as having one of the strongest signatures of balancing selection in the exome, and this information contributes to its appeal as a blood-stage vaccine antigen [132,303]. Consistent with these previous findings, Figure 7-3 (bottom panel) suggests elevated frequency-dependent selection in Region II, with a spike at the 3' end of the first DBL domain (F1). Region II has been implicated in the direct interaction with GYPA on the erythrocyte surface, and thus antibodies targeting this region could inhibit invasion, consistent with the Tajima's D profile across the gene [158,320]. It was more recently shown that full-length recombinant *eba-175* binds GYPA with a 10-fold stronger affinity than RII alone, and antibodies targeting outside of RII contribute to inhibition of invasion [161]. In light of this, the two position-wise profiles in Figure 7-3 paint an interesting picture. While RII indicates balancing selection and an excess of nonsynonymous changes, the 2kb region following the F-indel has an excess of low frequency polymorphism and, aside from position 3640, no significant positive selection. The codon starting at position 3640 is the most significant triplet in the entire gene when comparing dN to dS, thus it would be interesting to know if the non-RII antibodies that inhibit invasion target this vicinity.

7.6.3 Host-parasite interaction

The role of EBA-175 in erythrocyte invasion by interaction with the human GYPA receptor has long been known, and this has motivated previous investigations into whether the F/C dimorphism is associated with the MN blood-group, which is defined by polymorphism in GYPA [273,321]. No association was previously detected in an analysis of host and parasite genotypes in Gabon, Nigeria, and South Africa [175]. The human GWAS results here are consistent with this conclusion, and further, detect no convincing signals of association of the F/C dimorphism with any other loci. As mentioned in section 8.4.2, the glycophorin region of chromosome 4 is highly complex due to SNPs, unequal crossing over, and gene conversion, thus accurately genotyping these genes is difficult, which may diminish power to detect associations.

Separately, the GWAS testing the 6bp indel for host associations yielded a cluster of SNPs in the Kenyan cohort that meet genome-wide significance, and the association of this region shows some evidence in the Gambian population as well (Figure 7-11). Considering both cohorts, this region spans 1.7Mb on chromosome 14q22, which includes 18 genes. The

cluster of 5 SNPs meeting genome-wide significance in the Kenyan population lie equidistant from two genes: ESRRB and VASH1. VASH1 is a 7 exon gene expressed by epithelial cells to inhibit angiogenesis [322]. ESRRB encodes a gene similar to the estrogen receptor, and in mice is involved in placental development [323]. It would be premature to speculate deeply about these genes, or on those 18 in the 1.7Mb region, without further validation in larger cohorts or different populations. However, as discussed in 8.4.2 patient outcome should be considered as a potential lurking variable in this type of analysis. If there truly is an interaction between a host gene and a parasite invasion ligand, it may have arisen as a protective adaptation. In such a scenario there would be a third variable in the equation—i.e., host outcome, which could shed more light on the functional connection between EBA-175, the host gene, and pathogenesis.

7.6.4 Contributions of Malign and MalMOI to these investigations

Aside from the structural variants in *eba-175*, this gene is fairly "well behaved" with regard to aligning reads from field samples to the 3D7 reference. Indeed, MalariaGEN classifies 58 SNPs in *eba-175* in the v2.0 release. This release overlaps most closely with the samples assembled for this analysis, but in an upcoming release there will be 180 SNPs defined from a sample set of nearly 5000 parasites (MalariaGEN, personal communication). Clearly as more samples are analyzed, more low frequency SNPs will be ascertained, but as long as this is done by aligning to the 3D7 reference, discovery will be limited to positions represented in that genome. A key value added here is the addition of structural variation, SNPs within these structural variants, and tri-allelic polymorphism. An even larger gain would be realized with highly divergent genes like *msp3.4*. Further, mapping variation to a universal IUPAC reference is a convenient way to represent this information. This universal reference made possible the Sequenom assay designs that facilitated the host-parasite interaction GWASs.

7.7 Acknowledgements

As referenced in the methods, the infrastructure for performing the human genome-wide scans in this chapter, from sample collection and genotyping, to imputation, quality-control, and software development, represents thousands of hours of work from an army Kwiatkowski group members and collaborators. Dr. Kirk Rockett enabled me to navigate the available Kenyan and Gambian sample data, and he and his lab team, particularly Christina Hubbart, mentored the design of the Sequenom multiplexes and performed the

genotyping. Dr. Si Quang Le graciously introduced me to the group's human GWAS pipeline that he and Dr. Gavin Band built, and Drs. Band and Rockett were advisors for much of the work in the chapter.

7.8 Supplementary material

7.8.1 EBA-175 IUPAC consensus

Supplementary table 7-6. Universal IUPAC consensus for *eba-175*. A consensus sequence was derived from the multiple sequence alignment of 1419 spliced *eba-175* exons. The appropriate IUPAC symbol is used to represent all alleles in polymorphic positions. Gapped positions are represented by those sequences with nucleotides in the alignment (e.g., only F-type parasites comprise the consensus in the region of the F-insert). The 3 indels (6bp, F, and C) are colored green, blue, and red, corresponding to Figure 1-4.

ATGAAATGTA ATATTAGTAT ATATTTTTTT GCTTCCTTCT TTGTGTTATA TTTTGCAAAA	60
GCTAGGAATG AATATGWTRT AAAAGAGAAT GARAAATTTT TAGACGTRTA TAAAGAAAAA	120
TTTAATGAAT TAGATAAAAA GAAATATGGA AATRTTCAAA AAACTGATAW GAAAATATTT	180
ACTTTTATAG AAAATAAATT AGATATTTTA AATAATTCAR AATTTAATAA AAGATGGAAG	240
AGTTATGGAA CTCCAGATAA TATAGATAAA AATATGTCTT TAATAAATAA ACATAATAAT	300
GAAGAAATGT TTAACAACAA TTATCAATCA TTTTTATCGA CAAGTTCATT AATAAAGCAA	360
AATAAAWATG TTCCTATTAA CGMTGTACGT GTGTCTAGGA TATTAAGTTT CCTGGATTCT	420
AGAATTAATA ATGGAAGAWA TAYTTCAWMT AATAACGAAG TTTTAARTAR TTGTAGGGAA	480
AAAAGGAAAG GAATGAAATG GSATTGTAAA AAGAAAAATG ATAGAAGCAA CTATGTATGT	540
ATTCCTGATC GTAGAATCCA ATTATGCATT GTTAATCTTA GCATTATTAA AAVATATACA	600
AAAGAGACCA TGAAGGATCA TTTCATTGAA GCCTCTAAAA AAGAATCTCA ACTTTTGCTT	660
AAAAAAAATG ATAACRAATA TAATTCTAAA TTTTGTAATG ATTTGAAGAA TAGTTTTTTA	720
GATTATGGAC ATCTTGCTAT GGGAAATGAT ATGGATTTTG GAGGTTATTC AACTAAGGCA	780
GAAAACAAAA TTCAAGAAGT TTTTAAAGGG GHTCATGGGR AAAWAAGTGA ACATRAAATT	840
AAAAATTTTA GAAAARAATG GTGGARTGAA TTTAGAGAGA AACTTTGGGA AGCTATGYTA	900
TCTGAGCATA AAAATAATAT ARATAATTGT AAAAATATTC CCCAAGAAGA ATTACAAATT	960
ACTCAATGGA TAAAAGAATG GCATGGAGAA TTTTTGCTTG AAAGAKATAA TAGATCAAAA	1020
TTGCCAAAAA GTAAATGTAA AAATAATACA TTATATGAAG CATGTGAGAA GGAATGTATT	1080
GATCCATGTA TGAAATATAR AGATTGGATT ATTAGAAGTA AATTTGAATG GCATACGTTA	1140
TCGAAAGAAT ATGAAACTCA AAAWGTTYCA AAGGAAAATG CGGAAAATTA TTTAATCAAA	1200
ATTTCARAAA AMAWGAATGA TGCTAAAGTA AGTTTATTAT TGAATAATTG TGATGMTGAA	1260
TATTCAAAAT ATTGTGATTG TAAACATACT ACTACTCTCG TTAAAAGCGT TTYAAATGGT	1320
AAMGAYAATA CAATTAAGGA ARAKCGTGAA CATATTGATT TAGATGATTT TTCTAAATTT	1380
GGATGTGATA AAAATTCCGT TGATACAAAC ACAAAGGTGT GGGAATGTAA ARAMCCTTAT	1440
AWAKTATCCA CTAAAGATGT ATGTGTACCT CCGAGGAGGC AAGAATTATG TCTTGGAAAC	1500
ATTGATAGAA TATACGATAA AAAYCTATTA ATGATAAAAG AGCATATTCT TGCTATTGCA	1560
ATATATGAAT CAAGAATATT GAAACKAAAA TATARGAATA AAGATGATAA AGAAGTTTGT	1620
AAAATCATAA ATAAAACTTT CGCTGATATA AGAGATATTA TAGGAGGTAM TGATTATTGG	1680
AATGATTTGA GCAATAGAAA ATTAGTAGGA AAAATYAACA CAAATTCAAA WTATGYTCAC	1740
AGGAATAAAV ARAATGATAA GCTTTTTCGT GATGMGTGGT GGAAAGTTAT TAAAAAAGAT	1800
GTATGGAATG TGATATCATG GGTATTCAAG GATAAAACTG TTTGTAAAGA AGATGATATT	1860
GMAAATATAC CACAATTCTT CAGATGGTTT AGTGAATGGG GTGATGATTA TTGYCAGGAT	1920
AAAACAAAAA TGATAGAGAC TCTGAAGGTT GAATRCAAAG AAAAACCTTG TGARGATGAC	1980
AATTGTAAAM GTAAATGTAA TTCATATAAA GAATGGATAT CAAAAAAAAA AGAAGARTAT	2040
AATAAACAAG CCAAACAATA CCAAGAATAT CAAAAAGGAA ATAATTACAA AATGTATTCT	2100
GAATTTAAAT CTATMAAACC AGAAGTTTAT TTAAAGAAAT ACTCGRAAAA ATGTTCTAAC	2160
CTAAATTTCG AAGATGAATT TAAGGAAGAA TTACATTCAG ATTATAAAAA TAAATGTACG	2220
ATGTGTCCAG AAGTAAAGGA TGTACCAATT TCTATAATAA GAAATAATGA RCAAACTTCG	2280
MAAGAAGCMG TTCCTGAGGA AARCACTGAA ATAGCACACA GAACGGAAAC TCGTACGGAT	2340
GAACGAAAAA ATCAGGAACC AGCAAATAAG GATTTAAAGA ATCCACAACA AAGTGTAGGA	2400

GAGAACGGAA CTAAWRATTT ATTACAAGAA GATTTAGGAG GATCAMGAAG TGAAGACGAA	2460
GTGACACAAR AATTTGGAGT AAATCATGGA ATACCTAAGG GTGAGGATMA AACGTTAGRA	2520
AAATCTGACG CCATTCCAAA CATAGGCGAA YCCGAAACGG GAATTTCCAC TACAGAAGAA	2580
AGTAGACATG AAGAAGGCCA CAATAAACAA GCATTGTCTA CTTCAGTCGA TGAGCCTGAA	2640
TTATCTGATA CACTTCAATT GCATGAAGAT ACTAAAGAAA ATGATAAACT ACCCMTAGAA	2700
TCATCTRCAA TCACATCTCC TACGGAAAGT GGAAGTTCTG ATACAGAGGA AACTCCATCT	2760
ATCTCTRAAG GACCAAAAGG AAATGAACAA AAARAACGTG ATGAYKATAG TTTGAGTAAA	2820
ATAAGTGTAT CACCAGAAAA TTCAAGACCT GAAACTGATG CTAAAGATAC TTCTAACWTG	2880
TTAAAATTAA AAGGAGATGT TGATATTAGT ATGCCTAAAG CAGTTATTGG GAGCAGTCCT	2940
AATGATAATA TAAATGTTAC TGAASAAGGG GATAATATTT CCGGGGTGAA TTMTAAACCT	3000
TTATCTGATG ATGTACGTCM ARATAAAAAG GAATTAGAAG ATCAAAATAG TGATGAATYG	3060
GAAGAAACTG TAGTAAATCA TATATCAAAA AGTCCATCTA TAAATAATGG ADATGATTCA	3120
GGCAGTSGAA GTGCAACAGT GAGTGAATCT AGTAGTTCAA ATACTGGATT GTCTATTGAT	3180
GATGATAGAA ATGGTGATAC ATTTGTTCGA ACACAAGATA CAGCAAATAC TGAAGATGTT	3240
ATTAGAAAAG AAAATGCTGA CAAGGATGAA GATGAAAAAG GCGCAGATGA AGAAAGACAT	3300
AGTACTTCTG AAAGCTTAAG TTCACCTGAA GAAAAAATGT TAACTGATAA TGAAGGAGGA	3360
AATAGTTTAA ATCATGAAGA GGTGAAAGAA CATACTAGTA ATTCTRATAA TGTTCAACAG	3420
TCTGGAGGAA TTGTTAATAT GAATGTTGAG AAAGAACTAA AAGATACTTT AGAAAATCCT	3480
TCTAGTAGCT TGGATGAAGG AAAAGCMCAT GAAGWAWTAT CAGAACCAAA TCTAAGCAGT	3540
RACCAAGATA TGTCTAATAC ACSTGGACCT TTGGATAACA CCAGTGAAGA AACTACAGAA	3600
AGAATTAGTA ATAATGAATA TAAAGTTAAC GAGAGGGAAK RTGAGAGAAC GCKTACTAAG	3660
GAATATGAAG RTAYTGTTTT GAAAAGTCAT ATGAATAGAG AATCAGACGA TGGTGAATTA	3720
TATGACGAAA ATTCAGACTT WTCTACTGTA AATGATGAAT CAGAAGACGC TGAAGCAAAA	3780
ATGAAAGRAA ATGATACATC TGAAATGTCG CATAATAGTA GTCAACATAT TGAGAGTGAT	3840
CAACAGAAAA ACGATATGAA AACTKTTGGT GATTTGGGAR CCACACATGT ACAAAACGAA	3900
ATTARTGTTC CTGTTACAGG AGAAATTRAT GAAAAATTAA GGGAAAGTAA AGAATCAAAA	3960
ATTCATAAGR CTGAAGWGGA AAGATTAART CATACAGATA TACATAAAAT TWATCCTGAW	4020
GATAGAAATA GTAATACATT ACATTTAAAA GATATAAGAA ATGAGGAAAA CGAAAGACAC	4080
TTAACTAATC AAAACATTAA TATTRGTCAA GAAAGGGATT TGCAAAAACA TGGATTCCAT	4140
AYCATGAATA ATCTACATGG AGATGGAGTT TCCGAAAGAA GTCAAATTAA TCATAGTCAT	4200
CATGGAAACA GACAAGATCG GGGGGGAAAT TCTGGGAATG TTTTAAATAT GAGATCTAAT	4260
AATAATAATT TTAATAATAT TCCAAGTAGA TATAATTTAT ATGATAAAAA ATTAGATTTA	4320
GATCTTTATR AAAACAGAAA TGATAGTACA ACAAAAGAAT TAATAAAGAA ATTAGCAGAA	4380
ATAAATAAAT GTGAGAACGA AATTTCTGTA AAATATTGTG ACCRTATGAT TCATGAAGAA	4440
ATCYCATTAA AAACATGCAC TAAAGAAAAA ACAAGAAATC TGTGTTGTGC AGTATCAGAT	4500
TACTGTATGM GCTATTTTAC ATATGATTCA GAGGAATATT ATAATTGTAC GAAAARGGAA	4560
TTTGATGATC CATCTTATAC ATGTTTCAGA AAGGAGGCTT TTTCAAGTAT GCCATATTAT	4620
GCAGGAGCAG GTGTGTTATT TATTATATTG GTTATTTTAG GTGCTTCACA AGCCAAATAT	4680
CAAAGTTCTG AAGGAGTTAT GAATGWGAAT AATGAGAATA ATTTTTYATT TGAAGTTACT	4740
GATAATTTAG ATAAATTATC CAATATGTTC AATCAACAAG TACAGGAAAC TAATATCAAC	4800
GATTTTTCTG AATACCATGA GGATATAAAT GATAWTAAWT TWAAGAWA	4848

7.8.2 Mapping major features of *eba-175*

Supplementary table 7-7. Mapping major features to the *eba-175* IUPAC consensus. Positions refer to the universal IUPAC consensus obtained by aligning 1419 MalMOI assembled field isolates (Supplementary table 7-6).

Feature	Position in IUPAC consensus
Exon 1	1-4606
Exon 2	4607-4685
Exon 3	4686-4768
Exon 4	4769-4848
F1 DBL domain	472-1279
F2 DBL domain	1384-2224
6bp indel	1201-1206
F insert	2326-2748

C insert	3028-3369
3' cysteine-rich domain	4258-4591

7.8.3 Full list of SNPs in *eba-175*

Supplementary table 7-8. Polymorphism in *eba-175*. Position refers to the universal IUPAC reference. Allele frequencies are for the 2nd allele listed for bi-allelic SNPs (the minor allele), otherwise frequencies are listed in the corresponding allele order. SNPs within the F and C-inserts are noted as F or C.

Position	Note	MAF	Allele (major, minor)	AAC
77	Singleton	0.0007	A,T	D,V
79		0.0014	A,G	I,V
93		0.0014	A,G	E,E
108		0.0021	G,A	V,V
154	Singleton	0.0007	G,A	V,I
170	Singleton	0.0007	A,T	K,M
220	Singleton	0.0007	A,G	K,E
367		0.0014	T,A	Y,N
383	Singleton	0.0007	C,A	A,D
439	Singleton	0.0007	A,T	N,Y
443		0.0014	C,T	T,I
448		0.0162	T,A	S,T
449	Singleton	0.0007	C,A	S,Y
467		0.0035	G,A	S,N
470		0.1001	A,G	N,S
502		0.012	G,C	D,H
593	tri-allelic	0.994,0.004,0.0007	C,A,G	T,K,R
676		0.0247	A,G	K,E
812	tri-allelic	0.997,0.002,0.0007	C,A,T	A,D,V
820		0.456	G,A	E,K
824		0.0761	T,A	I,K
835		0.3122	G,A	E,K
856		0.4341	A,G	K,E
866	Singleton	0.0007	G,A	S,N
898		0.0909	T,C	L,L
922	Singleton	0.0007	A,G	N,D
1006		0.1325	G,T	D,Y
1100	Singleton	0.0007	G,A	R,K
1164		0.4538	A,T	K,N
1168		0.4087	T,C	S,P
1207		0.4729	A,G	K,E
1212		0.2622	C,A	N,K
1214		0.2622	A,T	K,M

1256		0.0014	C,A	A,D
1313		0.0028	T,C	L,S
1323	Singleton	0.0007	C,A	N,K
1326	Singleton	0.0007	C,T	D,D
1342		0.0021	G,A	E,K
1344		0.0021	G,T	K,N
1432		0.0028	A,G	K,E
1434		0.1572	A,C	K,N
1442		0.3199	A,T	K,I
1444		0.0162	T,G	L,V
1524		0.0014	C,T	N,N
1586	Singleton	0.0007	G,T	R,L
1595	Singleton	0.0007	A,G	K,R
1670		0.0035	C,A	T,N
1716	Singleton	0.0007	T,C	I,I
1731		0.0881	T,A	N,K
1736		0.0021	T,C	V,A
1750	tri-allelic	0.46, 0.33, 0.21	C,A,G	Q,K,E
1752		0.0042	A,G	Q/K/E,Q/K/E
1775		0.1832	A,C	E,A
1862	Singleton	0.0007	A,C	E,A
1914		0.0014	C,T	C,C
1955	Singleton	0.0007	G,A	C,Y
1974	Singleton	0.0007	A,G	E,E
1990		0.3693	A,C	S,R
2037	Singleton	0.0007	G,A	E,E
2115		0.0106	A,C	I,I
2146		0.2755	G,A	E,K
2271	Singleton	0.0007	A,G	E,E
2281		0.0014	C,A	Q,K
2289	Singleton	0.0007	A,C	A,A
2303		0.4954	G,A	S,N
2415	F	0.0063	A,T	K,N
2416	F	0.0013	G,A	D,N
2446	F	0.0013	C,A	R,R
2470	F	0.0013	G,A	E,K
2509	F	0.0013	C,A	Q,K
2519	F	0.0013	G,A	G,E
2551	F	0.0051	C,T	P,S
2695	F	0.0088	C,A	L,I
2707	F	0.0063	A,G	T,A
2767		0.0853	G,A	E,K
2794		0.4242	G,A	E,K
2805	Singleton	0.0007	C,T	D,D

2806	Singleton	0.0007	G,T	D,Y
2878	Singleton	0.0007	T,A	L,M
2965		0.0014	C,G	E,Q
2993	Singleton	0.0007	C,A	S,Y
3020		0.012	C,A	P,Q
3022	Singleton	0.0007	G,A	D,N
3059	C	0.0032	C,T	S,L
3112	C	0.0016	G,T	D,Y
3127	C	0.0352	G,C	G,R
3406	Singleton	0.0007	G,A	D,N
3507		0.0634	A,C	A,A
3515		0.0684	A,T	E,V
3517	Singleton	0.0007	T,A	L,I
3541	Singleton	0.0007	G,A	D,N
3563		0.0014	C,G	P,R
3640	Singleton	0.0007	G,T	D,Y
3641		0.3199	A,G	D,G
3653		0.0014	T,G	L,R
3671		0.0204	A,G	D,G
3674		0.0014	T,C	I,T
3741		0.0028	A,T	L,F
3788	Singleton	0.0007	G,A	G,E
3865	Singleton	0.0007	G,T	V,F
3880	Singleton	0.0007	A,G	T,A
3905		0.0014	G,A	S,N
3928	Singleton	0.0007	G,A	D,N
3970	Singleton	0.0007	G,A	A,T
3977	Singleton	0.0007	A,T	E,V
3989	Singleton	0.0007	G,A	S,N
4012	Singleton	0.0007	A,T	N,Y
4020	Singleton	0.0007	A,T	E,D
4105	Singleton	0.0007	A,G	S,G
4142		0.0021	C,T	T,I
4330		0.0402	G,A	E,K
4424		0.0014	A,G	H,R
4444		0.0014	C,T	P,S
4510	Singleton	0.0007	A,C	S,R
4556	Singleton	0.0007	G,A	R,K
4706		0.0028	A,T	E,V
4727		0.0155	T,C	L,S
4835		0.0381	T,A	I,N
4839		0.265	A,T	K,N
4842		0.0381	T,A	F,L
4847		0.2269	T,A	I,K

7.8.4 F/C indel double insertions and deletions

The most frequently used method for genotyping F and C parasites is based on a nested PCR reaction using primers that span both indels. This reaction yields a 714bp amplicon from C parasites, and a 795bp amplicon from F-type parasites. Using the Nested PCR method, Soulama, *et al* found two alternate fragments (360 and 400bp), which represented more than a quarter of genotypes in northern Ghana [324]. This may be consistent with a deletion, but similar amplicons weren't reported in four other studies conducted in similar regions representing 1132 participants [171,172,173,174]. The MalMOI assemblies yielded 4 products with both the F and C-inserts, and 7 with both deleted. Of these 11 double insertion/deletion samples, legacy DNA was available from MalariaGEN for one of each putative genotype (i.e., one +/+ (PFL0063) and one -/- (PFL0005)). To assess whether these rare genotypes truly exist or are assembly errors, two sets of primers that flank both indels were used to amplify products in these legacy samples, along with F and C controls (Figure 7-16). These reactions and expected products are described in section 2.7.3.

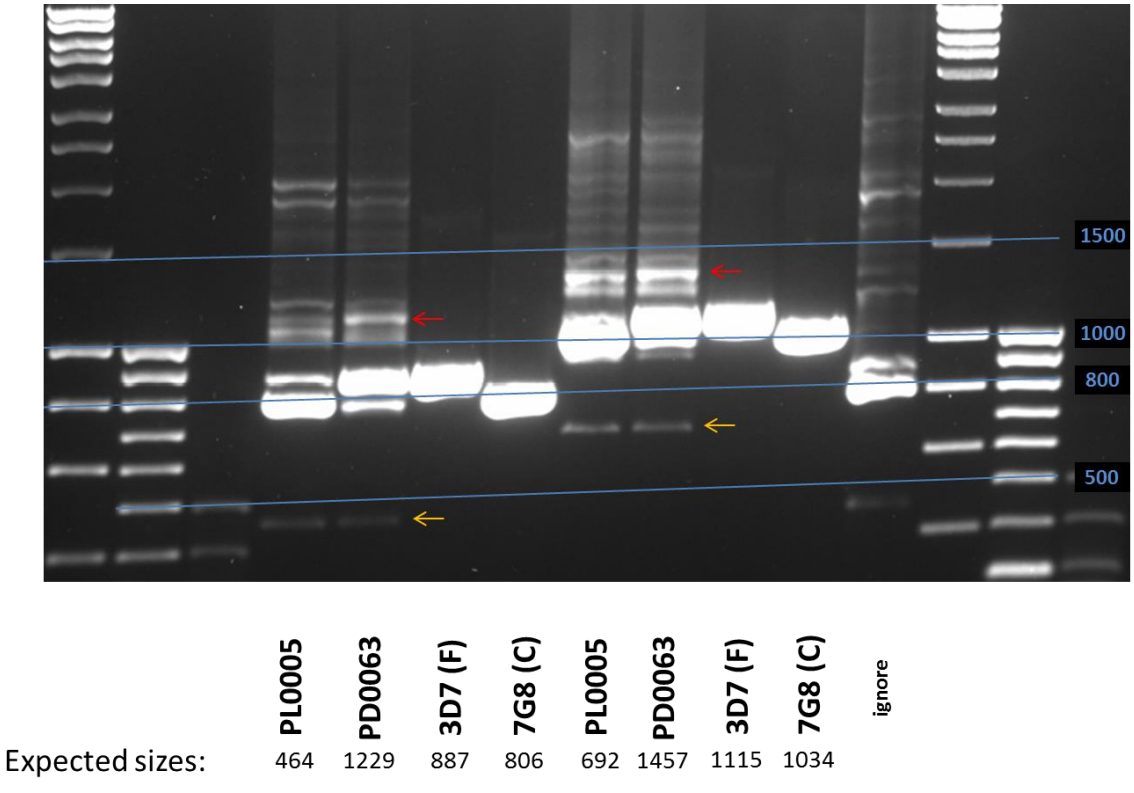


Figure 7-16. PCR products investigating the existence of *eba-175* double F/C insertions and deletions. PL0005 yielded a -/- *de novo* assembly and PD0063 yielded a +/+ product from MalMOI. The first 4 lanes and the second 4 lanes are the same samples PCR amplified with different primers, both primer sets encompassing the F and C indels. Orange arrows indicate the expected -/- product

sizes and red arrows indicate the expected sizes for +/+ PCR products. Both field samples show bands for F, C, +/+, and -/-. The higher weight bands are messy, adding to the suspicion of chimeras in mixed infections.

Although this reaction yields bands exactly where expected for double insertions and deletions, both legacy samples show evidence of all 4 genotypes: F, C, +/+. and -/-. To test whether this was an artifact from PCR chimeras I performed a second reaction with the F and C controls, both individually and artificially mixed, and indeed the same +/+ and -/- bands appeared (Figure 7-17). Although many field samples contain mixtures, the fact that both of the legacy reactions generated all 4 genotypes is suspicious. No conclusions can be drawn from these results to support or disclaim the existence of double insertions and deletions.

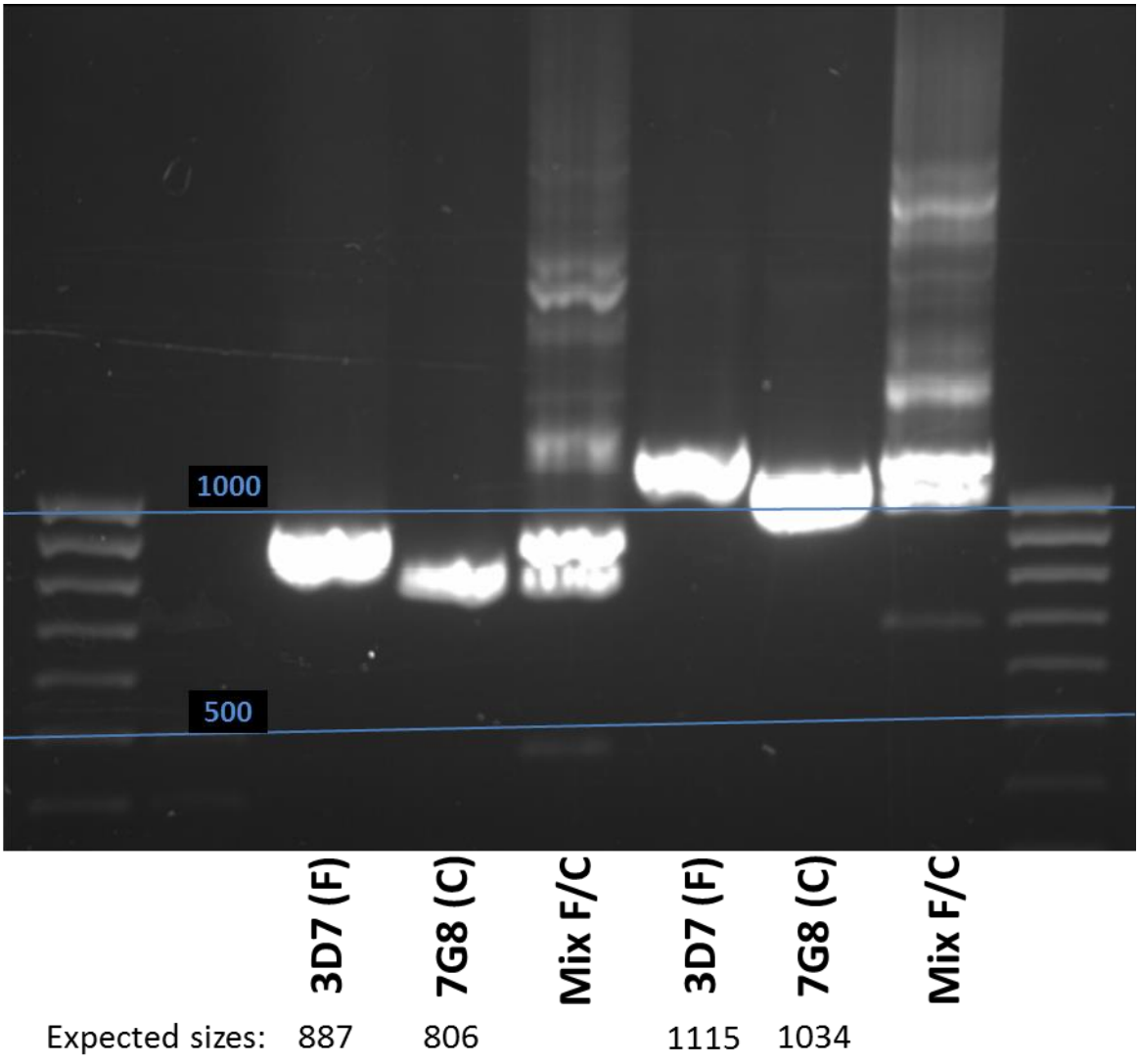
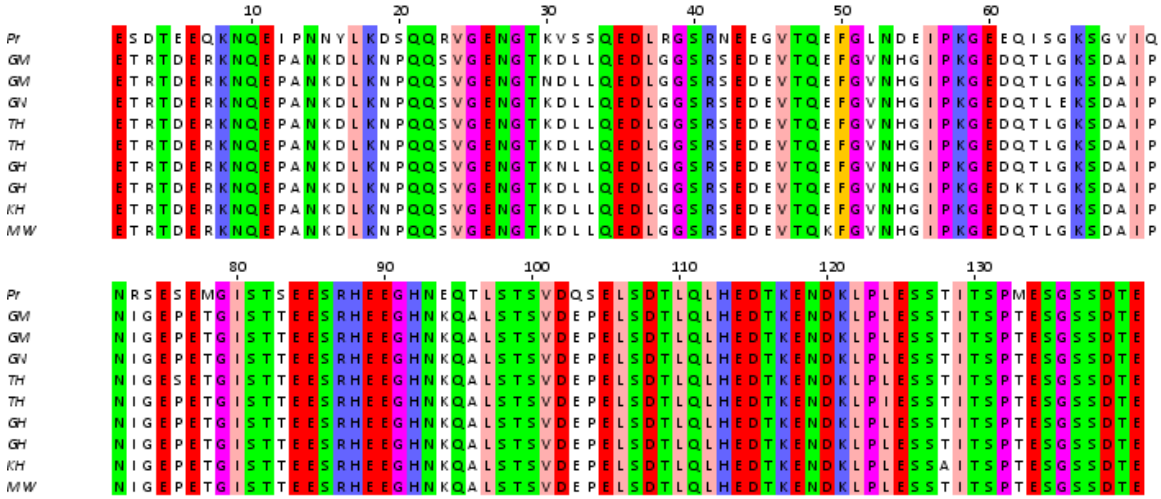


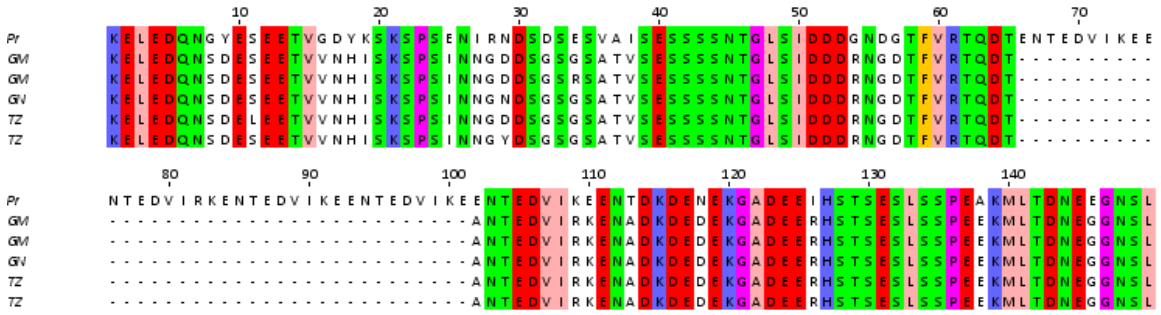
Figure 7-17. PCR products testing for PCR chimeras in F/C mixtures. Clonal F (3D7) and C (7G8) lab-line parasites were used as templates for the primers testing for double insertions and deletions. Reactions with artificial mixtures were also performed to investigate the possibility of PCR chimeras.

For both primer sets, chimeric products matching the expected +/+ and -/- sizes are artifactually produced.

7.8.5 Comparison of the F/C between *P. falciparum* and *P. reichenowi*



Supplementary figure 7-18. *P. falciparum* and *P. reichenowi* protein alignment of the EBA-175 F segment. The top sequence is that of *P. reichenowi*, and the other 9 are a random selection of *P. falciparum sequences* from a variety of countries.



Supplementary figure 7-19. *P. falciparum* and *P. reichenowi* protein alignment of the EBA-175 C segment. The top sequence is that of *P. reichenowi*, and the other 5 are a random selection of *P. falciparum sequences* from a variety of countries. *P. reichenowi* has a 36bp deletion compared to its ortholog.

7.8.6 Supplementary GWAS hits

Supplementary table 7-9. Host-parasite interaction GWAS SNPs with $p < 1e^{-6}$.

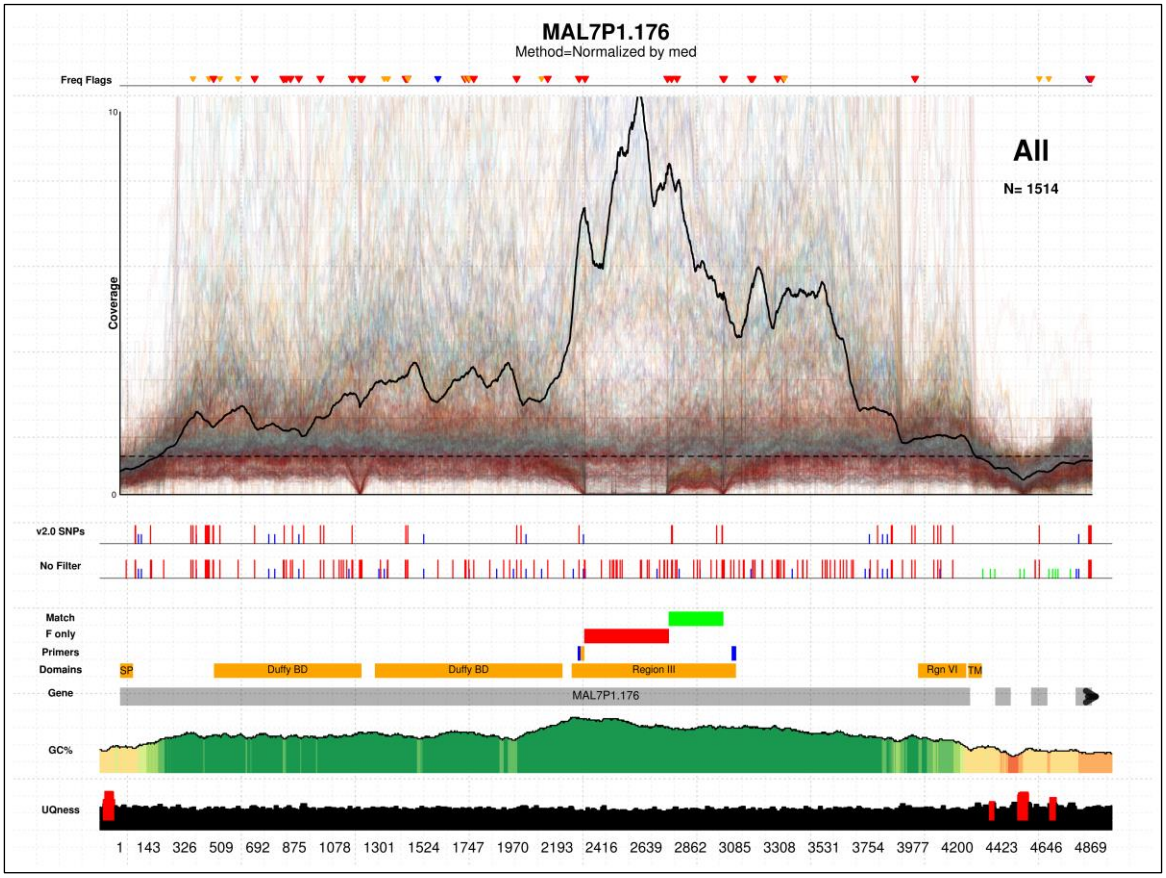
Indel	Cohort	rsid	Chr	Position	MAF	p-value
6bp	Gambia	rs12498304	chr04	8520610	0.198576	6.28E-07

6bp	Gambia	rs142818843	chr06	162198444	0.016431	8.44E-07
6bp	Gambia	rs146209842	chr06	162204110	0.01623	7.74E-07
6bp	Gambia	rs140822816	chr06	162211749	0.0162	7.96E-07
6bp	Gambia	rs148671237	chr06	162216379	0.016719	5.34E-07
6bp	Gambia	rs142046836	chr06	162220961	0.01678	5.30E-07
6bp	Gambia	rs140948727	chr06	162230251	0.016733	6.33E-07
6bp	Gambia	rs6955390	chr07	4009511	0.230578	8.68E-07
6bp	Gambia	rs113026531	chr07	57300653	0.072594	2.79E-07
6bp	Gambia	kgp13741002	chr07	57302621	0.068242	2.92E-07
6bp	Gambia	rs112526474	chr07	57309703	0.071045	2.25E-07
6bp	Gambia	rs112763844	chr07	57311729	0.071044	2.26E-07
6bp	Gambia	rs113498505	chr07	57312149	0.067977	3.41E-07
6bp	Gambia	rs3903054	chr07	57312977	0.069126	5.69E-07
6bp	Gambia	rs113908032	chr07	57316529	0.071041	2.25E-07
6bp	Gambia	rs113279125	chr07	57317604	0.067977	3.42E-07
6bp	Gambia	kgp13361519	chr07	57323749	0.071163	2.07E-07
6bp	Gambia	rs113606080	chr07	57323937	0.067712	3.91E-07
6bp	Gambia	rs112037547	chr07	57324573	0.071267	2.02E-07
6bp	Gambia	rs112599751	chr07	57325749	0.071164	2.07E-07
6bp	Gambia	rs11159110	chr14	75392298	0.118696	5.30E-07
6bp	Gambia	rs6051936	chr20	406559	0.347769	5.94E-07
6bp	Gambia	rs2208299	chr20	59635953	0.260489	3.91E-07
6bp	Gambia	kgp10422564	chr22	45676678	0.148433	5.03E-07
6bp	Gambia	rs2742625	chr22	45677125	0.147876	9.16E-07
6bp	Kenya	rs183253676	chr02	10697620	0.02413	9.96E-07
6bp	Kenya	chr6:28561574:I	chr06	28561574	0.020171	8.19E-07
6bp	Kenya	rs55757328	chr07	155746142	0.107833	4.89E-07
6bp	Kenya	rs12590729	chr14	77087221	0.11746	2.85E-08
6bp	Kenya	rs11159218	chr14	77087672	0.121374	3.78E-08
6bp	Kenya	rs11159219	chr14	77087730	0.121374	3.79E-08
6bp	Kenya	rs17104696	chr14	77090397	0.117438	2.94E-08
6bp	Kenya	rs4506820	chr14	77095148	0.117705	2.55E-08
6bp	Kenya	rs145593950	chr20	356815	0.036594	7.77E-07
6bp	Kenya	rs115859734	chr20	365436	0.03688	9.15E-07
F/C	Gambia	rs112108926	chr02	238999088	0.337294	3.55E-07
F/C	Gambia	rs56046910	chr04	14062119	0.038758	1.83E-07
F/C	Gambia	rs192098257	chr05	76606067	0.018551	9.64E-07
F/C	Gambia	rs182465929	chr05	76614814	0.018586	9.43E-07
F/C	Gambia	rs149576060	chr05	76615978	0.018593	9.38E-07
F/C	Gambia	rs147053952	chr05	76630183	0.018506	8.66E-07
F/C	Gambia	rs117029919	chr07	63672420	0.032184	3.04E-07
F/C	Gambia	rs58461899	chr07	132411363	0.02802	2.86E-07
F/C	Gambia	rs73438842	chr07	132413126	0.025464	3.87E-07
F/C	Kenya	rs76343721	chr03	61181922	0.012483	7.05E-07

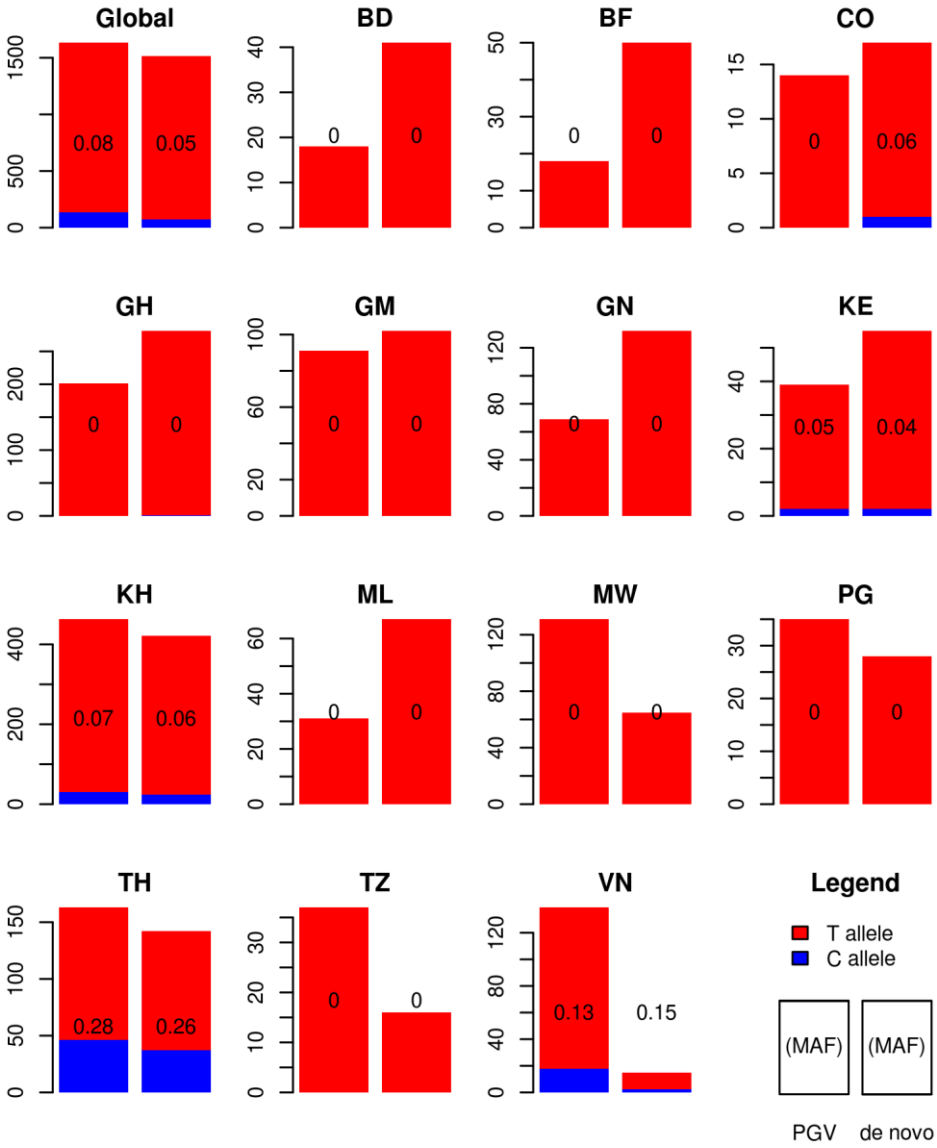
F/C	Kenya	rs76010254	chr03	115947943	0.068391	2.32E-07
F/C	Kenya	kgp18041158	chr03	115949266	0.068	2.47E-07
F/C	Kenya	rs111494462	chr03	133717410	0.021117	3.15E-07
F/C	Kenya	rs192697267	chr06	48553219	0.020444	7.25E-07
F/C	Kenya	rs80262290	chr09	115672948	0.281685	8.70E-07
F/C	Kenya	rs79462584	chr09	115672982	0.281697	8.68E-07

7.8.7 Other supplementary figures

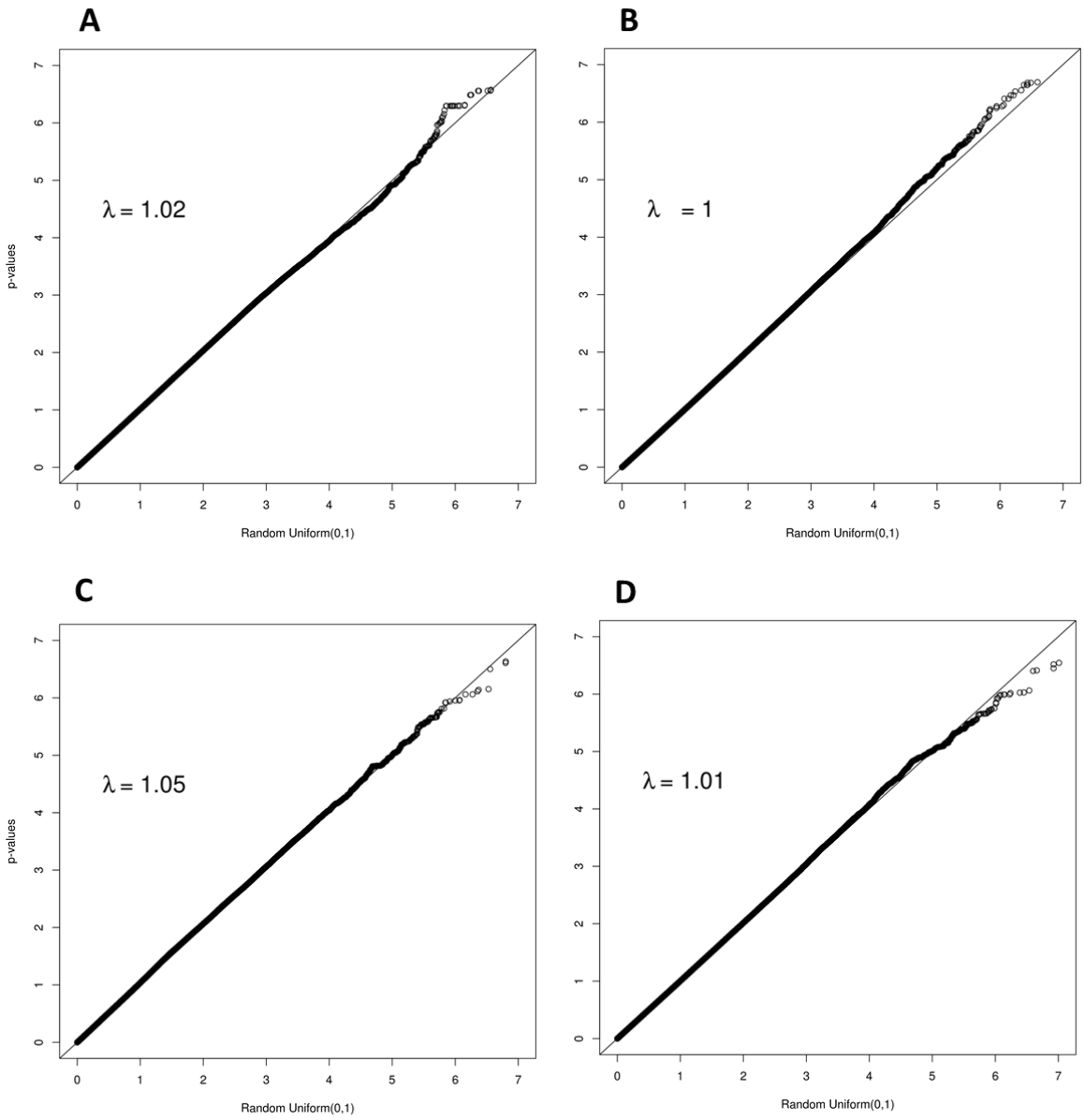
Supplementary figure 7-20. Multiple sequence alignment of EBA-175 translated assemblies. Parasite assemblies from 1419 MalMOI processed samples were translated, aligned using MAFFT, and sorted by length—thus there are 1419 rows in this plot. Each vertical line is an amino acid position (differences in color along a vertical line indicate nonsynonymous polymorphism). Amino acids are colored using the Zappo scheme [325]. The two striking patterns of missingness are the F and C indels. The F-indel is larger, and 5'. Notice the dimorphic property—i.e., parasites have either the F or C-insert, never both and never neither. Several cases of double insertions and deletions did occur, but it is unclear whether these were natural events or assembly errors, so they were excluded (see 7.8.4).



Supplementary figure 7-21. Candidate gene plot for *eba-175*. Complete details describing all tracks and data sources can be found in methods section 2.2.1. The coverage track contains 1514 lines representing the normalized sequence read coverage for each sample. Lines are colored by country. The normalizing constant for each sample is the median (indicated in the title) coverage of SNPs in all genes for the given sample. The black line is the mean across all samples. Two SNP tracks appear just below the coverage plot. Tall red SNPs are non-synonymous while blue lines are synonymous. The F-only track highlights the location of the F insert. The "match" track indicates a stretch of near identity in all samples occurring between the F and C indels.



Supplementary figure 7-22. Allele frequencies for *eba-175* SNP L300L by two methods. Each panel displays a barplot comparing the allele frequencies of the synonymous SNP L300L as calculated by two different methods (PGV and *de novo*). The left bar in each panel is calculated from MalariaGEN *Plasmodium* Genome Variation project read counts as previously described. The right bar is calculated directly from MalMOI *de novo* assembled genes, where only a dominantly abundant parasite is assembled from an infection. The first plot combines all samples broken down by country in the following panels. Country abbreviations are listed in the Abbreviations and Acronyms section. The figure legend is shown in the bottom right panel. The methods appear to be highly concordant—particularly in regions with appreciable sample sizes.



Supplementary figure 7-23. QQ plots for host-parasite interaction GWA studies. A) Kenya 6bp study. **B)** Gambia 6bp study. **C)** Kenya F/C study. **D)** Gambia F/C study. Genome-wide inflation factor (λ) printed in each box.

8 GENERAL DISCUSSION

8.1 Section I

This section presented two themes. The first theme was about the biological insights gained from analysing genome-wide SNP data. The second theme motivates the ensuing chapters by presenting artefacts in the SNP data that arise in the presence of complex variation. On the latter theme, the first example shows a systematic artefact of missing SNPs in the K76T encoding region of *pfcrt* in samples from Kenya. Importantly, this missingness is associated with a particular genotype that confers chloroquine tolerance, providing an augury for confounding that can occur when analysing complex variation with SNP data. Through haplotype and candidate gene plots I connect these artefacts with low complexity introns that flank dense polymorphism in exon 2. Separately, while analysing SNPs for a population genetic analysis of Tanzanian samples I show a different kind of artefact for the invasion ligand, *eba-175*. In this case, two large indels produce a different pattern of missingness in SNP data from that seen in *pfcrt*. The reference version of this gene contains the F-insert but not the C-insert, thus samples with the opposite configuration that are aligned to 3D7 produce a well-defined span of missingness in the F-region. Further, as the reference is missing the C-insert, no assessment of that region is made for any sample. Using the Tanzanian SNP data I provide a third example of artefacts derived from complex variation in the gene *msp3.4*. The signature of missingness in this gene is large like that of *eba-175*, however with less well-defined boundaries. These examples motivate the development of alternative approaches for accessing the complex variation in these and similar genes.

8.2 Section II

8.2.1 Malign and first attempts at assembly

Malign was designed to take a brute-force approach to detecting complex variation. A key value of this software is that it can be used to quickly assay samples that have already been processed in a standard sequencing pipeline that didn't genotype the variant of interest (i.e., BAM files), and can also be used on freshly sequenced samples (i.e., fastq files). Malign provides accurate detection over a range of indel sizes, but has some limitations. The utility of this tool was demonstrated on 3000 MalariaGEN samples that had already been processed, in which the F, C, and 6bp indels of *eba-175* were quickly genotyped (section 7.4.5). One limitation of Malign is that the variant being assayed must be known, as sample reads are aligned to a user supplied version of the gene in fasta format. A second limitation is that variants identified in mixed samples are not necessarily phased. These limitations motivated the first attempt at *de novo* assembly, which involved patching Cortex variants into the 3D7 version of target genes. Although Cortex proves to be a useful tool, patching into the reference created systematic biases, and this approach was abandoned.

8.2.2 MalMOI

The second chapter in Section II describes an algorithm for assembling full-length genes and exons. Important insights in this work are the gains in computational efficiency that result from targeting the assemblies, and the ability to assemble phased genes from mixed samples. Although only written as a single chapter, this work represents a high proportion of the effort for this DPhil. MalMOI has several limitations that will be addressed in future versions (see 8.4.1).

8.3 Section III

The thesis concludes by revisiting a gene that was discussed in every section, *eba-175.* Leveraging hundreds of full-length exon sequences, the coding variation of this important vaccine candidate is characterized with unprecedented resolution. I create an IUPAC consensus sequence from a multiple sequence alignment of 1419 assemblies, which serves as a universal reference for mapping and referring to polymorphism. This consensus also fed directly into Sequenom assay design software. The MSA provided a foundation for designing a pipeline for identifying complex variation that is unique to each allele class, and

then for designing Sequenom assays to type this polymorphism—either directly, or with sets of indicator assays.

Using these assays, the F, C, and 6bp indels were genotyped in Gambian and Kenyan cohorts for which genome-wide human SNP data was also available. I then used this unique dataset to innovate a host-parasite interaction GWAS, in which the host SNPs were tested for association with the indel class of the infecting parasite. The most striking result from this analysis is a region on chromosome 14 that shows evidence of association with the 6bp indel in both populations, but most significantly in Kenya. It is also noteworthy that like a previous study, no association was detected between the F/C dimorphism and the glycophorin region of chromosome 4 [175]. However, the GYP region is complex in its own right, so the corresponding SNPs may not sufficiently tag these genes. Separate work is being undertaken within the Kwiatkowski group to access this human complex variation. The host-parasite interaction scans reported here are the very tip of the iceberg of what remains to be explored, and this is discussed in more detail below (8.4.2).

8.4 Future work

8.4.1 MalMOI

Currently MalMOI attempts to output either a full-length gene or full-length exons. In many contexts this may be an overly demanding expectation, as long contigs may be discarded because they fall slightly short, and these representations could be useful. Upcoming versions of MalMOI will return large contigs (size defined by the user) that pass filters. Currently, if a single position in the post-assembly pileup exhibits heterozygosity in which more than 20% of the alleles at that site disagree with the assembled contig, then the assembly is labeled a failure. As indicated in the *in silico* validation analysis of MalMOI, different genes and degrees of mixture may require different MOI cutoffs. Further certain investigations may have more or less tolerance for error. Thus, all assemblies will be output with the coverage and MOI filter results appended to their names, providing an easy format for changing the filters post-assembly.

Further validation should also be performed using *in silico* mixtures on more genes, including those with introns. In fact, it may be that every gene assembled should undergo this type of analysis to determine the best filter settings. In addition to controlling error, this could increase the sample size for certain genes, or in areas where transmission (and thus mixture) is low.

8.4.2 Host-parasite interaction scans

In my mind the most exciting part of this thesis is the host-parasite interaction work. It took an immense amount of effort to build the tools and infrastructure to enable these investigations, so they necessarily occurred at the very final stages of the DPhil. In the short term, complex variation in the other parasite genes included in the Sequenom multiplex will be tested for host genetic associations. These genes include *msp3.4* and *msp3.8*, for which Sequenom indicator assays were designed (see 7.3.1). These genes will first be tested using a GWAS with an additive model, as was done for the *eba-175* association scans. As this work is undertaken it will be necessary to assess the accuracy of the indicator assays for typing complex variation. The *eba-175* dataset will be useful for this analysis, as indirect and direct assays were performed on the F and C indels in the same multiplex. Recall the primary concern with the indicator assays is that a failed assay is meant to indicate the lack of a particular complex variant. This will clearly be confounded with assays that fail for other reasons (e.g., low parasitemia), which is why they are designed as a set of three—one for each complex form (in the case of a dimorphism), and one assay that should always return a genotype. This system should be evaluated for potential biases. For example, it's possible that differential performance of the indicator assays could lead to mixed samples being systematically classified as one particular form.

In the medium term, a variety of genetic models should be investigated for each association test. For the GWASs in this work, mixed parasite genotypes (e.g., F and C) were excluded from analysis. The rationale behind this was to mimic a binary case-control outcome variable, however this comes at the expense of sample size. Having three possible "outcomes" (F, C, F/C) might provide more statistical power, not only by increasing sample size, but also by more realistically modeling the relationship with the possible host SNP genotypes. This could be done using multivariate models, and the current beta version of SNPTEST provides this option. A second approach and extension of this could be to keep patient disease outcome as the binary response variable, and include host and parasite genotypes in the model with an interaction term. The interaction term would require more degrees of freedom, however mixed samples would be included. This analysis would test for the possibility that patient outcomes change if particular parasite genotypes infect particular host genotypes—a fundamental question in malaria pathogenesis.

In the longer term, more genes should be assembled to facilitate the characterization of complex variation for Sequenom design. On the fast approaching horizon, whole-genome sequencing will be performed on both the host and parasite from the same samples, and

investigation will be on the scale of whole-genome by whole-genome interaction scans. In the meantime, targeting parasite genes is a rational approach. One method for selecting candidate genes with complex variation is to use Cortex output to select genes with large variants that are present in many populations.

8.5 Final remarks

As I mentioned at the start of this thesis, any story about malaria is really a story about evolution, and evolution is largely a story about genes changing through time. This story about malaria covered a lot of ground, but each sub-theme was about genes that have been pressured to change, and about accessing the variation in those genes. In Section I we saw parasite genes that changed on a short time scale due to drug pressure, but also hints of genes undergoing more complex changes. After developing methods to access that complex variation we saw some striking patterns—uncharacteristically long haplotypes present in every population that haven't been broken down by recombination. Why do these complex variants exist, and how are they maintained in every population? To address this question I performed a unique type of GWAS, testing for host-parasite interactions. While much more work remains in this area, initial results indicate a possible association of a parasite indel with host genotypes. The pressure to evolve is applied by both host and parasite to one another, and these interactions result in both genomes changing, sometimes orchestrated, through time.

APPENDIX A: A MICROARRAY STUDY OF SEVER MALARIA IN TANZANIAN CHILDREN

Part of my DPhil funding came through the US National Institutes of Health Graduate Partnership program. Through this program I worked with the Laboratory of Malaria Immunology and Vaccinology, directed by the same investigators that lead the MOMS project in Tanzania (see chapter 4). One of my activities for LMIV was to analyze a microarray dataset that I generated during a one-year fellowship in Morogoro, Tanzania, prior to starting my DPhil. This analysis represents a substantial amount of effort expended during my DPhil. Although my project went in another direction after this analysis, I feel an obligation to report these findings, so I briefly describe this project in the appendix.

Background and design

I analyzed data from 236 4-color microarray hybridizations of RNA from the peripheral blood of 71 Tanzanian children in the MOMS project. This activity presented several analytical challenges requiring novel methods to overcome. These challenges include host RNA contamination, variable parasite developmental stage, repeated measures, batch effects, and 4-color normalization.

After extensive quality filtering, expression values for 1766 parasite genes and 355 human genes were analyzed for 66 children presenting with one of 4 disease classifications (table A.1).

Table A.1. Breakdown of 66 children by disease category. Repeated measure counts are represented in the technical replicates column.

Category	Biological Replicates	Technical Replicates
Asymptomatic	13	42
Mild	30	108
cLinical	10	33
Severe	13	39

Stage correction

To correct for parasite stage, a subset of 300 probes matching those used in a study that characterized the transcriptome of the intraerythrocytic developmental cycle (IDC) was correlated for each child to each of the 48 hourly time-points in that study [326]. This was done for three parasite isolates, yielding for each child's hybridization a sinusoidal curve of correlations across the IDC (figure A.1). The parasite hour post invasion (HPI) was estimated by the maximum correlation above, and the median of replicate assays was assigned as the HPI for a given child's sample. Samples yielding non-sinusoidal curves may represent poor quality RNA or high gametocytemia, and were therefore excluded from final analyses. The estimated HPI was used as a covariate in the mixed linear model used to test for differential expression (see below).

Interestingly, HPI appears to be associated with hemoglobin level (figure A.2). Children in this population with parasites older than 14 HPI have significantly lower hemoglobin levels.

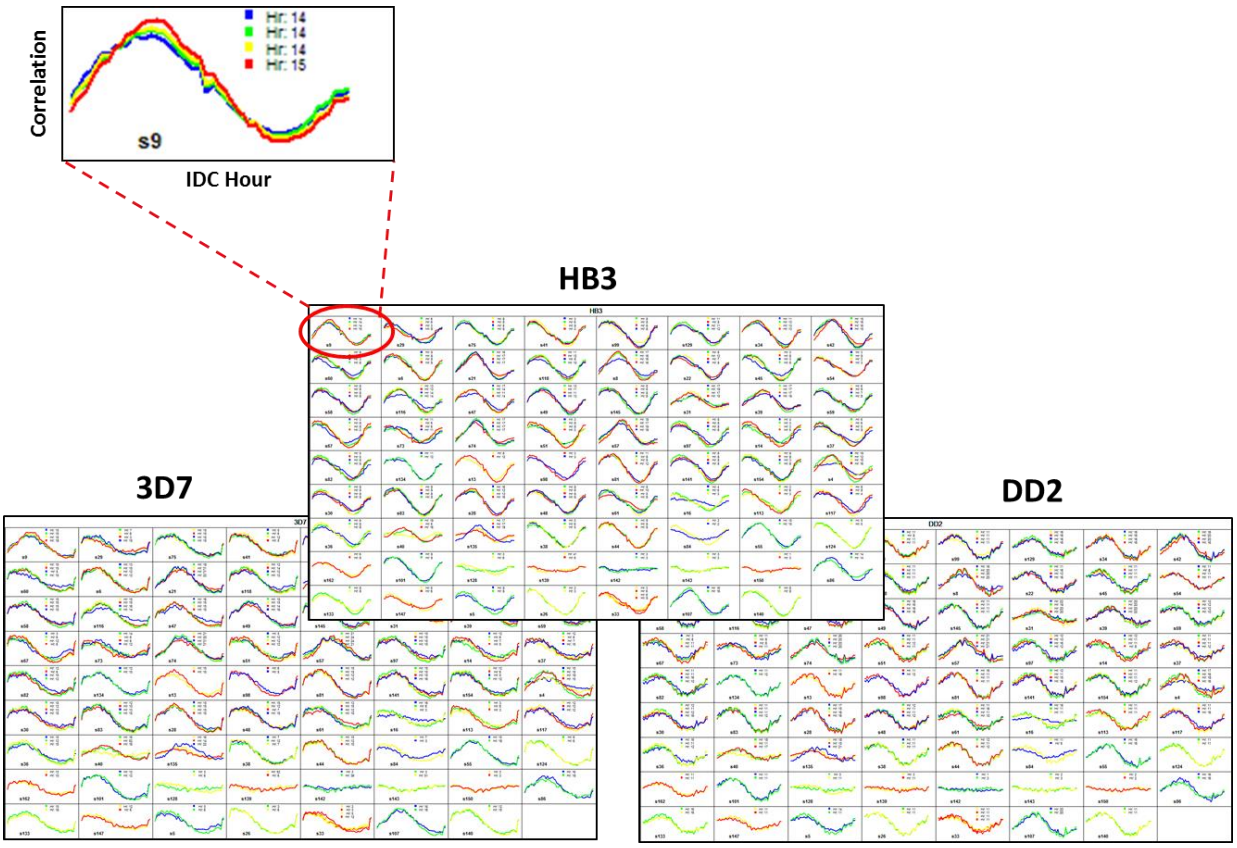


Figure A.1. Estimation of hour post invasion. Top blown-up image is a plot of correlations of gene expression values from child s9's parasites with each 48-hour time-point from HB3. X-axis is hour post invasion of HB3, and y-axis is the correlation. Curve colors represent dyes used for technical replicates of this sample. Similar analyses were performed for each child and 3 parasites (HB3, DD2, 3D7).

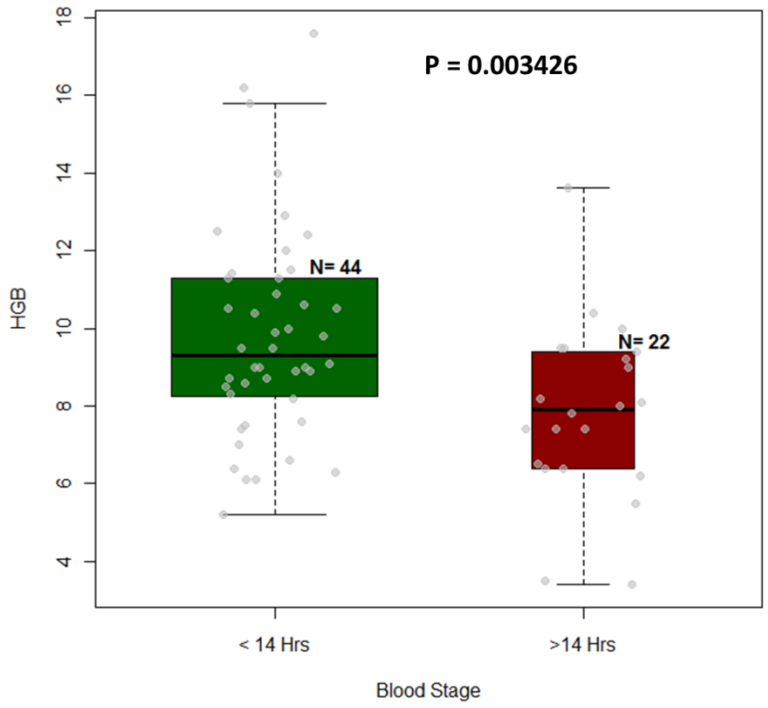


Figure A.2. Parasite stage is associated with hemoglobin level. Parasites were categorized as younger or older than 14 hours post invasion, and a 2-sided t-test used to detect a difference amongst these groups in hemoglobin levels of children from which the parasites were taken. Children with more mature parasites have significantly lower hemoglobin levels (p-value =0.003).

Expression heterogeneity

A ubiquitous problem in high-throughput datasets is subsets of observations with correlated values due to batch effects, or other unknown sources of variation [211,327]. To confront this issue here, a method similar to surrogate variable analysis and PCA correction of population structure in GWA studies was developed. Briefly, a linear model was fit for each microarray probe, testing for differential expression using all known covariates (i.e., HPI, Disease category, microarray, dye, parasitemia, and host contamination). Principal components (PC) were determined using the residuals from this model, and the first 5 PCs were taken back and included as covariates in the original model. To determine if this approach was yielding anything sensible, it was performed separately without HPI in the original model, and then the PCs were compared to HPI to asses if parasite stage would have been corrected were it completely unknown. Indeed, this test revealed that the first two PCs would have been strong surrogates for HPI (figure A.3). In the final analysis, the estimate of

HPI is included in the linear model, and thus the first 5 PCs are meant to control for unknown sources of spurious variance.

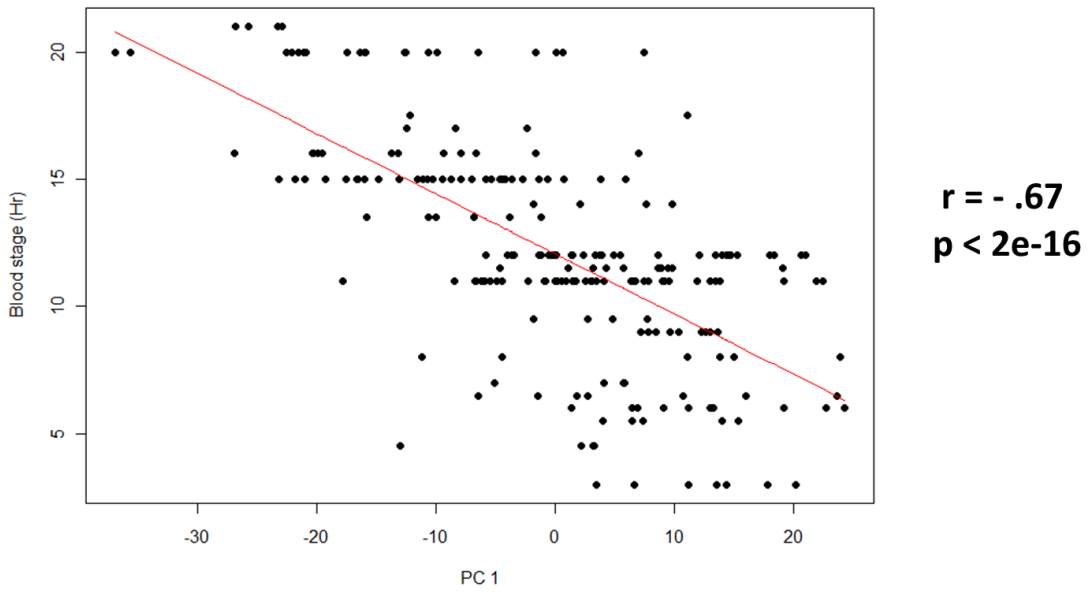


Figure A.3. Test of PCA method for controlling unknown effects. X-axis indicates the first principal component from a PCA on residuals from a model in which HPI was held out, and the y-axis is the estimated HPI. This PC strongly predicts HPI, and would thus have effectively corrected for this covariate were it unknown (p-value < 2e-16). PC2 was also significantly correlated with HPI (not shown).

Differential expression

To detect differential expression between disease categories, a linear mixed model was fit for each probe with microarray intensity as the response, fixed effects of parasitemia, HPI, disease category, dye, PC1-PC5, and random effects for child and microarray. Many genes were represented on the microarray by multiple probes, and these were used to validate the analytical approach. As expected if the method were yielding signal, probes with more significant p-values are highly enriched for the same genes (figure A.4). As an example to highlight one of the significant genes with multiple significant probes, *gbph2* is associated with the surface of the merozoite, and if validated may thus make an interesting candidate antigen (figure A.5). This gene is one that will be validated by qPCR, but is highly variable and non-unique in some regions, and thus qualifies as a candidate for characterizing conserved regions for primer design.

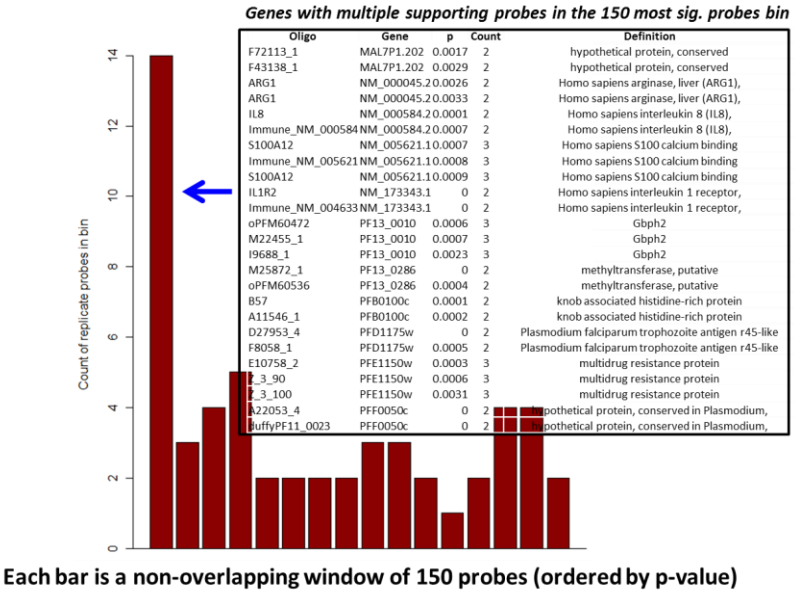


Figure A.4. Significant probes are enriched for the same genes. P-values were ordered and split into groups of 150. Each bin in the histogram represents an ordered section from this list—i.e., the left most bin represents the smallest 150 p-values. The y-axis is the count of probes in the 150 for that bin that map to the same gene. The probability of seeing a result this extreme in the left-most bin was determined by permutation to be less than 1 in 10 million. Genes in the left-most bin with multiple probes are listed in the box.

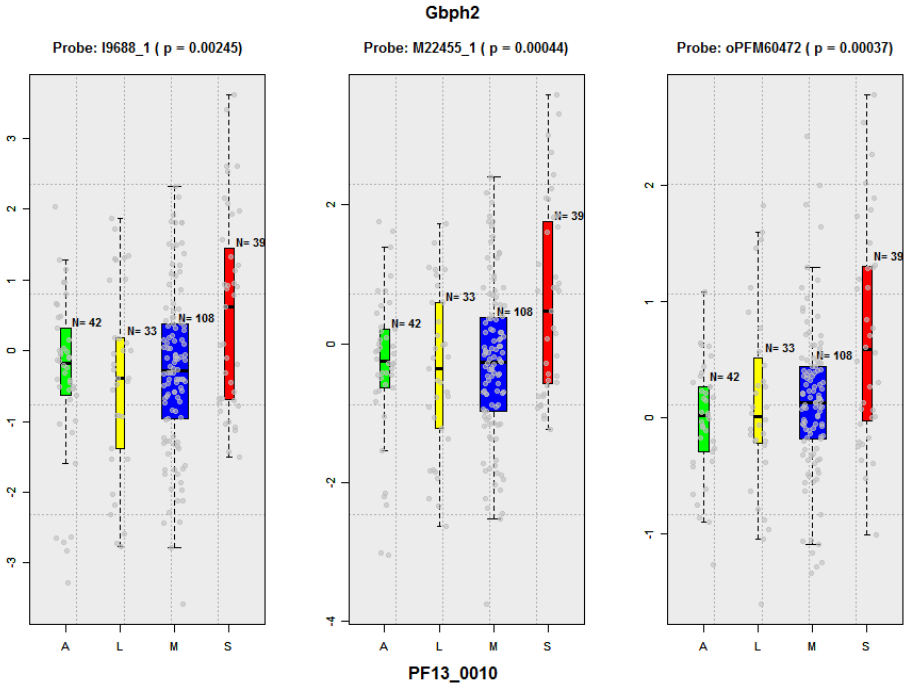


Figure A.5. Boxplots of the raw data for each of the 3 probes for *gbph2*. The p-value above each plot represents the test of whether all 4 disease categories are the same. X-axis labeled with disease: A=asymptomatic, L=clinical (a.k.a., moderately severe), M=mild, S=severe. Y-axis indicates normalized microarray intensity.

Acknowledgment

The Tanzanian microarray dataset was generated and analysed separately for my MSc final project. No part of that analysis is presented here. All of the work described here resulted from a complete reanalysis of this dataset, and was conducted during my DPhil.

CITATIONS

1. Silva JC, Egan A, Arze C, Spouge JL, Harris DG (2015) A new method for estimating species age supports the co-existence of malaria parasites and their mammalian hosts. Molecular biology and evolution.

2. Bannister L, Mitchell G (2003) The ins, outs and roundabouts of malaria. Trends in parasitology 19: 209-213.

3. McFadden GI, Reith ME, Munholland J, Lang-Unnasch N (1996) Plastid in human parasites. Nature 381: 482.

4. Fast NM, Kissinger JC, Roos DS, Keeling PJ (2001) Nuclear-encoded, plastid-targeted genes suggest a single common origin for apicomplexan and dinoflagellate plastids. Molecular biology and evolution 18: 418-426.

5. Prugnolle F, Durand P, Ollomo B, Duval L, Ariey F, et al. (2011) A fresh look at the origin of Plasmodium falciparum, the most malignant malaria agent. PLoS pathogens 7: e1001283.

6. Escalante AA, Ayala FJ (1994) Phylogeny of the malarial genus Plasmodium, derived from rRNA gene sequences. Proceedings of the National Academy of Sciences of the United States of America 91: 11373-11377.

7. Hall N (2012) Genomic insights into the other malaria. Nature genetics 44: 962-963.

8. Bensch S, Hellgren O, Perez-Tris J (2009) MalAvi: a public database of malaria parasites and related haemosporidians in avian hosts based on mitochondrial cytochrome b lineages. Molecular ecology resources 9: 1353-1358.

9. Escalante AA, Freeland DE, Collins WE, Lal AA (1998) The evolution of primate malaria parasites based on the gene encoding cytochrome b from the linear mitochondrial

genome. Proceedings of the National Academy of Sciences of the United States of America 95: 8124-8129.

10. Chin W, Contacos PG, Coatney GR, Kimball HR (1965) A Naturally Acquited Quotidian-Type Malaria in Man Transferable to Monkeys. Science 149: 865.

11. Singh B, Daneshvar C (2013) Human infections and detection of Plasmodium knowlesi. Clinical microbiology reviews 26: 165-184.

12. Cogswell FB (1992) The hypnozoite and relapse in primate malaria. Clinical microbiology reviews 5: 26-35.

13. White NJ (2011) Determinants of relapse periodicity in Plasmodium vivax malaria. Malaria journal 10: 297.

14. Price RN, Tjitra E, Guerra CA, Yeung S, White NJ, et al. (2007) Vivax malaria: neglected and not benign. The American journal of tropical medicine and hygiene 77: 79-87.

15. Mendis K, Sina BJ, Marchesini P, Carter R (2001) The neglected burden of Plasmodium vivax malaria. The American journal of tropical medicine and hygiene 64: 97-106.

16. Rosenberg R, Wirtz RA, Schneider I, Burge R (1990) An estimation of the number of malaria sporozoites ejected by a feeding mosquito. Transactions of the Royal Society of Tropical Medicine and Hygiene 84: 209-212.

17. Pradel G, Frevert U (2001) Malaria sporozoites actively enter and pass through rat Kupffer cells prior to hepatocyte invasion. Hepatology 33: 1154-1165.

18. Naito M, Hasegawa G, Takahashi K (1997) Development, differentiation, and maturation of Kupffer cells. Microscopy research and technique 39: 350-364.

19. Foquet L, Hermsen CC, van Gemert GJ, Van Braeckel E, Weening KE, et al. (2014) Vaccine-induced monoclonal antibodies targeting circumsporozoite protein prevent Plasmodium falciparum infection. The Journal of clinical investigation 124: 140-144.

20. Kester KE, Cummings JF, Ofori-Anyinam O, Ockenhouse CF, Krzych U, et al. (2009) Randomized, double-blind, phase 2a trial of falciparum malaria vaccines RTS,S/AS01B and RTS,S/AS02A in malaria-naive adults: safety, efficacy, and immunologic associates of protection. The Journal of infectious diseases 200: 337-346.

21. Olotu A, Fegan G, Wambua J, Nyangweso G, Awuondo KO, et al. (2013) Four-year efficacy of RTS,S/AS01E and its interaction with malaria exposure. The New England journal of medicine 368: 1111-1120.

22. Kester KE, McKinney DA, Tornieporth N, Ockenhouse CF, Heppner DG, et al. (2001) Efficacy of recombinant circumsporozoite protein vaccine regimens against experimental Plasmodium falciparum malaria. The Journal of infectious diseases 183: 640-647.

23. Hoffman SL, Goh LM, Luke TC, Schneider I, Le TP, et al. (2002) Protection of humans against malaria by immunization with radiation-attenuated Plasmodium falciparum sporozoites. The Journal of infectious diseases 185: 1155-1164.

24. Spring M, Murphy J, Nielsen R, Dowler M, Bennett JW, et al. (2013) First-in-human evaluation of genetically attenuated Plasmodium falciparum sporozoites administered by bite of Anopheles mosquitoes to adult volunteers. Vaccine 31: 4975-4983.

25. Epstein JE, Richie TL (2013) The whole parasite, pre-erythrocytic stage approach to malaria vaccine development: a review. Current opinion in infectious diseases 26: 420-428.

26. Roestenberg M, McCall M, Hopman J, Wiersma J, Luty AJ, et al. (2009) Protection against a malaria challenge by sporozoite inoculation. The New England journal of medicine 361: 468-477.

27. Bijker EM, Bastiaens GJ, Teirlinck AC, van Gemert GJ, Graumans W, et al. (2013) Protection against malaria after immunization by chloroquine prophylaxis and sporozoites is mediated by preerythrocytic immunity. Proceedings of the National Academy of Sciences of the United States of America 110: 7862-7867.

28. Baird JK, Fryauff DJ, Hoffman SL (2003) Primaquine for prevention of malaria in travelers. Clinical infectious diseases : an official publication of the Infectious Diseases Society of America 37: 1659-1667.

29. Fernando D, Rodrigo C, Rajapakse S (2011) Primaquine in vivax malaria: an update and review on management issues. Malaria journal 10: 351.

30. Langhorne J, Ndungu FM, Sponaas AM, Marsh K (2008) Immunity to malaria: more questions than answers. Nature immunology 9: 725-732.

31. Hermsen CC, de Vlas SJ, van Gemert GJ, Telgt DS, Verhage DF, et al. (2004) Testing vaccines in human experimental malaria: statistical analysis of parasitemia measured by a quantitative real-time polymerase chain reaction. The American journal of tropical medicine and hygiene 71: 196-201.

32. Murphy JR, Baqar S, Davis JR, Herrington DA, Clyde DF (1989) Evidence for a 6.5-day minimum exoerythrocytic cycle for Plasmodium falciparum in humans and confirmation that immunization with a synthetic peptide representative of a region of the circumsporozoite protein retards infection. Journal of clinical microbiology 27: 1434-1437.

33. Nguitragool W, Bokhari AA, Pillai AD, Rayavara K, Sharma P, et al. (2011) Malaria parasite clag3 genes determine channel-mediated nutrient uptake by infected red blood cells. Cell 145: 665-677.

34. Francis SE, Sullivan DJ, Jr., Goldberg DE (1997) Hemoglobin metabolism in the malaria parasite Plasmodium falciparum. Annual review of microbiology 51: 97-123.

35. Hawking F, Worms MJ, Gammage K (1968) 24- and 48-hour cycles of malaria parasites in the blood; their purpose, production and control. Transactions of the Royal Society of Tropical Medicine and Hygiene 62: 731-765.

36. Triglia T, Thompson J, Caruana SR, Delorenzi M, Speed T, et al. (2001) Identification of proteins from Plasmodium falciparum that are homologous to reticulocyte binding proteins in Plasmodium vivax. Infection and immunity 69: 1084-1092.

37. Cowman AF, Crabb BS (2006) Invasion of red blood cells by malaria parasites. Cell 124: 755-766.

38. Blackman MJ, Heidrich HG, Donachie S, McBride JS, Holder AA (1990) A single fragment of a malaria merozoite surface protein remains on the parasite during red cell invasion and is the target of invasion-inhibiting antibodies. The Journal of experimental medicine 172: 379-382.

39. Baldwin MR, Li X, Hanada T, Liu SC, Chishti AH (2015) Merozoite surface protein 1 recognition of host glycophorin A mediates malaria parasite invasion of red blood cells. Blood 125: 2704-2711.

40. Goel VK, Li X, Chen H, Liu SC, Chishti AH, et al. (2003) Band 3 is a host receptor binding merozoite surface protein 1 during the Plasmodium falciparum invasion of

erythrocytes. Proceedings of the National Academy of Sciences of the United States of America 100: 5164-5169.

41. Singh S, Alam MM, Pal-Bhowmick I, Brzostowski JA, Chitnis CE (2010) Distinct external signals trigger sequential release of apical organelles during erythrocyte invasion by malaria parasites. PLoS pathogens 6: e1000746.

42. Duraisingh MT, Triglia T, Ralph SA, Rayner JC, Barnwell JW, et al. (2003) Phenotypic variation of Plasmodium falciparum merozoite proteins directs receptor targeting for invasion of human erythrocytes. The EMBO journal 22: 1047-1057.

43. Baum J, Chen L, Healer J, Lopaticki S, Boyle M, et al. (2009) Reticulocyte-binding protein homologue 5 - an essential adhesin involved in invasion of human erythrocytes by Plasmodium falciparum. International journal for parasitology 39: 371-380.

44. Crosnier C, Bustamante LY, Bartholdson SJ, Bei AK, Theron M, et al. (2011) Basigin is a receptor essential for erythrocyte invasion by Plasmodium falciparum. Nature 480: 534-537.

45. Douglas AD, Baldeviano GC, Lucas CM, Lugo-Roman LA, Crosnier C, et al. (2015) A PfRH5-based vaccine is efficacious against Heterologous strain blood-stage Plasmodium falciparum infection in aotus monkeys. Cell host & microbe 17: 130-139.

46. Binks RH, Conway DJ (1999) The major allelic dimorphisms in four Plasmodium falciparum merozoite proteins are not associated with alternative pathways of erythrocyte invasion. Molecular and biochemical parasitology 103: 123-127.

47. Baum J, Maier AG, Good RT, Simpson KM, Cowman AF (2005) Invasion by P. falciparum merozoites suggests a hierarchy of molecular interactions. PLoS pathogens 1: e37.

48. Riglar DT, Richard D, Wilson DW, Boyle MJ, Dekiwadia C, et al. (2011) Super-resolution dissection of coordinated events during malaria parasite invasion of the human erythrocyte. Cell host & microbe 9: 9-20.

49. Schwartz L, Brown GV, Genton B, Moorthy VS (2012) A review of malaria vaccine clinical projects based on the WHO rainbow table. Malaria journal 11: 11.

50. Kilejian A (1979) Characterization of a protein correlated with the production of knob-like protrusions on membranes of erythrocytes infected with Plasmodium

falciparum. Proceedings of the National Academy of Sciences of the United States of America 76: 4650-4653.

51. Haldar K, Mohandas N, Samuel BU, Harrison T, Hiller NL, et al. (2002) Protein and lipid trafficking induced in erythrocytes infected by malaria parasites. Cellular microbiology 4: 383-395.

52. Wickert H, Wissing F, Andrews KT, Stich A, Krohne G, et al. (2003) Evidence for trafficking of PfEMP1 to the surface of P. falciparum-infected erythrocytes via a complex membrane network. European journal of cell biology 82: 271-284.

53. Raj DK, Nixon CP, Nixon CE, Dvorin JD, DiPetrillo CG, et al. (2014) Antibodies to PfSEA-1 block parasite egress from RBCs and protect against malaria infection. Science 344: 871-877.

54. Bruce MC, Alano P, Duthie S, Carter R (1990) Commitment of the malaria parasite Plasmodium falciparum to sexual and asexual development. Parasitology 100 Pt 2: 191-200.

55. Dyer M, Day KP (2000) Commitment to gametocytogenesis in Plasmodium falciparum. Parasitology today 16: 102-107.

56. Kaslow DC (1997) Transmission-blocking vaccines: uses and current status of development. International journal for parasitology 27: 183-189.

57. Nosten F, White NJ (2007) Artemisinin-based combination treatment of falciparum malaria. The American journal of tropical medicine and hygiene 77: 181-192.

58. White NJ (2004) Antimalarial drug resistance. The Journal of clinical investigation 113: 1084-1092.

59. Ghosh A, Edwards MJ, Jacobs-Lorena M (2000) The journey of the malaria parasite in the mosquito: hopes for the new century. Parasitology today 16: 196-201.

60. Carter R, Graves PM, Quakyi IA, Good MF (1989) Restricted or absent immune responses in human populations to Plasmodium falciparum gamete antigens that are targets of malaria transmission-blocking antibodies. The Journal of experimental medicine 169: 135-147.

61. malERA (2011) A research agenda for malaria eradication: vaccines. PLoS medicine 8: e1000398.

62. Kaslow DC (1990) Immunogenicity of Plasmodium falciparum sexual stage antigens: implications for the design of a transmission blocking vaccine. Immunology letters 25: 83-86.

63. Nevill CG, Some ES, Mung'ala VO, Mutemi W, New L, et al. (1996) Insecticide-treated bednets reduce mortality and severe morbidity from malaria among children on the Kenyan coast. Tropical medicine & international health : TM & IH 1: 139-146.

64. Pluess B, Tanser FC, Lengeler C, Sharp BL (2010) Indoor residual spraying for preventing malaria. The Cochrane database of systematic reviews: CD006657.

65. Martinez-Torres D, Chandre F, Williamson MS, Darriet F, Berge JB, et al. (1998) Molecular characterization of pyrethroid knockdown resistance (kdr) in the major malaria vector Anopheles gambiae s.s. Insect molecular biology 7: 179-184.

66. Ranson H, N'Guessan R, Lines J, Moiroux N, Nkuni Z, et al. (2011) Pyrethroid resistance in African anopheline mosquitoes: what are the implications for malaria control? Trends in parasitology 27: 91-98.

67. WHO (2012) WHO Evidence Review Group: The Safety and Effectiveness of Single Dose Primaquine

as a P. falciparum gametocytocide. Malaria Policy Advisory Committee Meeting.

68. Kuehn A, Pradel G (2010) The coming-out of malaria gametocytes. Journal of biomedicine & biotechnology 2010: 976827.

69. WHO (2011) World Malaria Report 2011. World Health Organization.

70. Miller LH, Baruch DI, Marsh K, Doumbo OK (2002) The pathogenic basis of malaria. Nature 415: 673-679.

71. O'Meara WP, Bejon P, Mwangi TW, Okiro EA, Peshu N, et al. (2008) Effect of a fall in malaria transmission on morbidity and mortality in Kilifi, Kenya. Lancet 372: 1555-1562.

72. Sinka ME, Bangs MJ, Manguin S, Coetzee M, Mbogo CM, et al. (2010) The dominant Anopheles vectors of human malaria in Africa, Europe and the Middle East: occurrence data, distribution maps and bionomic precis. Parasites & vectors 3: 117.

73. Kitua AY, Smith T, Alonso PL, Masanja H, Urassa H, et al. (1996) Plasmodium falciparum malaria in the first year of life in an area of intense and perennial transmission. Tropical medicine & international health : TM & IH 1: 475-484.

74. Miller LH, Ackerman HC, Su XZ, Wellems TE (2013) Malaria biology and disease pathogenesis: insights for new treatments. Nature medicine 19: 156-167.

75. Greenwood BM (1997) The epidemiology of malaria. Annals of tropical medicine and parasitology 91: 763-769.

76. Gupta S, Snow RW, Donnelly CA, Marsh K, Newbold C (1999) Immunity to non-cerebral severe malaria is acquired after one or two infections. Nature medicine 5: 340-343.

77. Fowkes FJ, McGready R, Cross NJ, Hommel M, Simpson JA, et al. (2012) New insights into acquisition, boosting, and longevity of immunity to malaria in pregnant women. The Journal of infectious diseases 206: 1612-1621.

78. Desai M, ter Kuile FO, Nosten F, McGready R, Asamoa K, et al. (2007) Epidemiology and burden of malaria in pregnancy. The Lancet Infectious diseases 7: 93-104.

79. Steketee RW, Nahlen BL, Parise ME, Menendez C (2001) The burden of malaria in pregnancy in malaria-endemic areas. The American journal of tropical medicine and hygiene 64: 28-35.

80. Fried M, Nosten F, Brockman A, Brabin BJ, Duffy PE (1998) Maternal antibodies block malaria. Nature 395: 851-852.

81. Fried M, Duffy PE (1996) Adherence of Plasmodium falciparum to chondroitin sulfate A in the human placenta. Science 272: 1502-1504.

82. Salanti A, Dahlback M, Turner L, Nielsen MA, Barfod L, et al. (2004) Evidence for the involvement of VAR2CSA in pregnancy-associated malaria. The Journal of experimental medicine 200: 1197-1203.

83. Springer AL, Smith LM, Mackay DQ, Nelson SO, Smith JD (2004) Functional interdependence of the DBLbeta domain and c2 region for binding of the Plasmodium falciparum variant antigen to ICAM-1. Molecular and biochemical parasitology 137: 55-64.

84. Snow RW, Marsh K (2002) The consequences of reducing transmission of Plasmodium falciparum in Africa. Advances in parasitology 52: 235-264.

85. Gardner MJ, Hall N, Fung E, White O, Berriman M, et al. (2002) Genome sequence of the human malaria parasite Plasmodium falciparum. Nature 419: 498-511.

86. Manske M, Miotto O, Campino S, Auburn S, Almagro-Garcia J, et al. (2012) Analysis of Plasmodium falciparum diversity in natural infections by deep sequencing. Nature 487: 375-379.

87. Malaria_Genomic_Epidemiology_Network (2008) A global network for investigating the genomic epidemiology of malaria. Nature 456: 732-737.

88. Manske HM, Kwiatkowski DP (2009) SNP-o-matic. Bioinformatics 25: 2434-2435.

89. DePristo MA, Zilversmit MM, Hartl DL (2006) On the abundance, amino acid composition, and evolutionary dynamics of low-complexity regions in proteins. Gene 378: 19-30.

90. Pizzi E, Frontali C (2001) Low-complexity regions in Plasmodium falciparum proteins. Genome research 11: 218-229.

91. Aravind L, Iyer LM, Wellems TE, Miller LH (2003) Plasmodium biology: genomic gleanings. Cell 115: 771-785.

92. Zilversmit MM, Volkman SK, DePristo MA, Wirth DF, Awadalla P, et al. (2010) Low-complexity regions in Plasmodium falciparum: missing links in the evolution of an extreme genome. Molecular biology and evolution 27: 2198-2209.

93. Kirchgatter K, Del Portillo HA (2005) Clinical and molecular aspects of severe malaria. Anais da Academia Brasileira de Ciencias 77: 455-475.

94. Kraemer SM, Smith JD (2006) A family affair: var genes, PfEMP1 binding, and malaria disease. Current opinion in microbiology 9: 374-380.

95. Turner L, Lavstsen T, Berger SS, Wang CW, Petersen JE, et al. (2013) Severe malaria is associated with parasite binding to endothelial protein C receptor. Nature 498: 502-505.

96. Scherf A, Hernandez-Rivas R, Buffet P, Bottius E, Benatar C, et al. (1998) Antigenic variation in malaria: in situ switching, relaxed and mutually exclusive transcription of var genes during intra-erythrocytic development in Plasmodium falciparum. The EMBO journal 17: 5418-5426.

97. Smith JD, Subramanian G, Gamain B, Baruch DI, Miller LH (2000) Classification of adhesive domains in the Plasmodium falciparum erythrocyte membrane protein 1 family. Molecular and biochemical parasitology 110: 293-310.

98. Salanti A, Jensen AT, Zornig HD, Staalsoe T, Joergensen L, et al. (2002) A sub-family of common and highly conserved Plasmodium falciparum var genes. Molecular and biochemical parasitology 122: 111-115.

99. Bull PC, Berriman M, Kyes S, Quail MA, Hall N, et al. (2005) Plasmodium falciparum variant surface antigen expression patterns during malaria. PLoS pathogens 1: e26.

100. Franke-Fayard B, Janse CJ, Cunha-Rodrigues M, Ramesar J, Buscher P, et al. (2005) Murine malaria parasite sequestration: CD36 is the major receptor, but cerebral pathology is unlinked to sequestration. Proceedings of the National Academy of Sciences of the United States of America 102: 11468-11473.

101. Kyes SA, Rowe JA, Kriek N, Newbold CI (1999) Rifins: a second family of clonally variant proteins expressed on the surface of red cells infected with Plasmodium falciparum. Proceedings of the National Academy of Sciences of the United States of America 96: 9333-9338.

102. Cheng Q, Cloonan N, Fischer K, Thompson J, Waine G, et al. (1998) stevor and rif are Plasmodium falciparum multicopy gene families which potentially encode variant antigens. Molecular and biochemical parasitology 97: 161-176.

103. McBride JS, Newbold CI, Anand R (1985) Polymorphism of a high molecular weight schizont antigen of the human malaria parasite Plasmodium falciparum. The Journal of experimental medicine 161: 160-180.

104. Tanabe K, Mackay M, Goman M, Scaife JG (1987) Allelic dimorphism in a surface antigen gene of the malaria parasite Plasmodium falciparum. Journal of molecular biology 195: 273-287.

105. Roy SW, Ferreira MU (2015) A new model for the origins of allelic dimorphism in Plasmodium falciparum. Parasitology international 64: 229-237.

106. Roy SW, Ferreira MU, Hartl DL (2008) Evolution of allelic dimorphism in malarial surface antigens. Heredity 100: 103-110.

107. Pearce JA, Triglia T, Hodder AN, Jackson DC, Cowman AF, et al. (2004) Plasmodium falciparum merozoite surface protein 6 is a dimorphic antigen. Infection and immunity 72: 2321-2328.

108. Snewin VA, Herrera M, Sanchez G, Scherf A, Langsley G, et al. (1991) Polymorphism of the alleles of the merozoite surface antigens MSA1 and MSA2 in Plasmodium

falciparum wild isolates from Colombia. Molecular and biochemical parasitology 49: 265-275.

109. McColl DJ, Anders RF (1997) Conservation of structural motifs and antigenic diversity in the Plasmodium falciparum merozoite surface protein-3 (MSP-3). Molecular and biochemical parasitology 90: 21-31.

110. Ware LA, Kain KC, Lee Sim BK, Haynes JD, Baird JK, et al. (1993) Two alleles of the 175-kilodalton Plasmodium falciparum erythrocyte binding antigen. Molecular and biochemical parasitology 60: 105-109.

111. Deitsch KW, Chitnis CE (2012) Molecular basis of severe malaria. Proceedings of the National Academy of Sciences of the United States of America 109: 10130-10131.

112. Greenwood BM, Fidock DA, Kyle DE, Kappe SH, Alonso PL, et al. (2008) Malaria: progress, perils, and prospects for eradication. The Journal of clinical investigation 118: 1266-1276.

113. Duffy PE, Fried M (2003) Plasmodium falciparum adhesion in the placenta. Current opinion in microbiology 6: 371-376.

114. Newbold CI, Craig AG, Kyes S, Berendt AR, Snow RW, et al. (1997) PfEMP1, polymorphism and pathogenesis. Annals of tropical medicine and parasitology 91: 551-557.

115. Smith JD, Craig AG, Kriek N, Hudson-Taylor D, Kyes S, et al. (2000) Identification of a Plasmodium falciparum intercellular adhesion molecule-1 binding domain: a parasite adhesion trait implicated in cerebral malaria. Proceedings of the National Academy of Sciences of the United States of America 97: 1766-1771.

116. Lopaticki S, Maier AG, Thompson J, Wilson DW, Tham WH, et al. (2011) Reticulocyte and erythrocyte binding-like proteins function cooperatively in invasion of human erythrocytes by malaria parasites. Infection and immunity 79: 1107-1117.

117. Fluck C, Smith T, Beck HP, Irion A, Betuela I, et al. (2004) Strain-specific humoral response to a polymorphic malaria vaccine. Infection and immunity 72: 6300-6305.

118. Thera MA, Doumbo OK, Coulibaly D, Laurens MB, Ouattara A, et al. (2011) A field trial to assess a blood-stage malaria vaccine. The New England journal of medicine 365: 1004-1013.

119. Wootton JC, Feng X, Ferdig MT, Cooper RA, Mu J, et al. (2002) Genetic diversity and chloroquine selective sweeps in Plasmodium falciparum. Nature 418: 320-323.

120. Payne D (1987) Spread of chloroquine resistance in Plasmodium falciparum. Parasitology today 3: 241-246.

121. Price RN, Uhlemann AC, Brockman A, McGready R, Ashley E, et al. (2004) Mefloquine resistance in Plasmodium falciparum and increased pfmdr1 gene copy number. Lancet 364: 438-447.

122. Pearce RJ, Pota H, Evehe MS, Ba el H, Mombo-Ngoma G, et al. (2009) Multiple origins and regional dispersal of resistant dhps in African Plasmodium falciparum malaria. PLoS medicine 6: e1000055.

123. Dondorp AM, Yeung S, White L, Nguon C, Day NP, et al. (2010) Artemisinin resistance: current status and scenarios for containment. Nature reviews Microbiology 8: 272-280.

124. Ariey F, Witkowski B, Amaratunga C, Beghain J, Langlois AC, et al. (2014) A molecular marker of artemisinin-resistant Plasmodium falciparum malaria. Nature 505: 50-55.

125. White NJ (2014) Malaria: a molecular marker of artemisinin resistance. Lancet 383: 1439-1440.

126. White N (1999) Antimalarial drug resistance and mortality in falciparum malaria. Tropical medicine & international health : TM & IH 4: 469-470.

127. Henderson DA (1999) Lessons from the eradication campaigns. Vaccine 17 Suppl 3: S53-55.

128. Thera MA, Plowe CV (2012) Vaccines for malaria: how close are we? Annual review of medicine 63: 345-357.

129. Thera MA, Doumbo OK, Coulibaly D, Diallo DA, Sagara I, et al. (2006) Safety and allele-specific immunogenicity of a malaria vaccine in Malian adults: results of a phase I randomized trial. PLoS clinical trials 1: e34.

130. Ogutu BR, Apollo OJ, McKinney D, Okoth W, Siangla J, et al. (2009) Blood stage malaria vaccine eliciting high antigen-specific antibody concentrations confers no protection to young children in Western Kenya. PloS one 4: e4708.

131. Conway DJ (1997) Natural selection on polymorphic malaria antigens and the search for a vaccine. Parasitology today 13: 26-29.

132. Amambua-Ngwa A, Tetteh KK, Manske M, Gomez-Escobar N, Stewart LB, et al. (2012) Population genomic scan for candidate signatures of balancing selection to guide antigen characterization in malaria parasites. PLoS genetics 8: e1002992.

133. Takala SL, Coulibaly D, Thera MA, Batchelor AH, Cummings MP, et al. (2009) Extreme polymorphism in a vaccine antigen and risk of clinical malaria: implications for vaccine development. Science translational medicine 1: 2ra5.

134. Takala SL, Plowe CV (2009) Genetic diversity and malaria vaccine design, testing and efficacy: preventing and overcoming 'vaccine resistant malaria'. Parasite immunology 31: 560-573.

135. Nunes JK, Woods C, Carter T, Raphael T, Morin MJ, et al. (2014) Development of a transmission-blocking malaria vaccine: progress, challenges, and the path forward. Vaccine 32: 5531-5539.

136. Feachem RP, AA.; Targett, GA.; Eds (2009) Shrinking the Malaria Map A Prospectus on Malaria Elimination. The Global Health Group, Global Health Sciences, University of California.

137. Wongsrichanalai C, Barcus MJ, Muth S, Sutamihardja A, Wernsdorfer WH (2007) A review of malaria diagnostic tools: microscopy and rapid diagnostic test (RDT). The American journal of tropical medicine and hygiene 77: 119-127.

138. Reyburn H, Mbatia R, Drakeley C, Carneiro I, Mwakasungula E, et al. (2004) Overdiagnosis of malaria in patients with severe febrile illness in Tanzania: a prospective study. BMJ 329: 1212.

139. Bejon P, Andrews L, Hunt-Cooke A, Sanderson F, Gilbert SC, et al. (2006) Thick blood film examination for Plasmodium falciparum malaria has reduced sensitivity and underestimates parasite density. Malaria journal 5: 104.

140. Mouatcho JC, Goldring JP (2013) Malaria rapid diagnostic tests: challenges and prospects. Journal of medical microbiology 62: 1491-1505.

141. Koita OA, Doumbo OK, Ouattara A, Tall LK, Konare A, et al. (2012) False-negative rapid diagnostic tests for malaria and deletion of the histidine-rich repeat region of the hrp2 gene. The American journal of tropical medicine and hygiene 86: 194-198.

142. Kwiatkowski D (2015) Malaria genomics: tracking a diverse and evolving parasite population. International health 7: 82-84.

143. Miotto O, Almagro-Garcia J, Manske M, Macinnis B, Campino S, et al. (2013) Multiple populations of artemisinin-resistant Plasmodium falciparum in Cambodia. Nature genetics 45: 648-655.

144. Wendler JP, Okombo J, Amato R, Miotto O, Kiara SM, et al. (2014) A genome wide association study of Plasmodium falciparum susceptibility to 22 antimalarial drugs in Kenya. PloS one 9: e96486.

145. Chen Q, Fernandez V, Sundstrom A, Schlichtherle M, Datta S, et al. (1998) Developmental selection of var gene expression in Plasmodium falciparum. Nature 394: 392-395.

146. Barry AE, Leliwa-Sytek A, Tavul L, Imrie H, Migot-Nabias F, et al. (2007) Population genomics of the immune evasion (var) genes of Plasmodium falciparum. PLoS pathogens 3: e34.

147. Fried M, Duffy PE (2002) Two DBLgamma subtypes are commonly expressed by placental isolates of Plasmodium falciparum. Molecular and biochemical parasitology 122: 201-210.

148. Benovoy D, Kwan T, Majewski J (2008) Effect of polymorphisms within probe-target sequences on olignonucleotide microarray experiments. Nucleic acids research 36: 4417-4423.

149. Bozdech Z, Zhu J, Joachimiak MP, Cohen FE, Pulliam B, et al. (2003) Expression profiling of the schizont and trophozoite stages of Plasmodium falciparum with a long-oligonucleotide microarray. Genome biology 4: R9.

150. Fried M, Wendler JP, Mutabingwa TK, Duffy PE (2004) Mass spectrometric analysis of Plasmodium falciparum erythrocyte membrane protein-1 variants expressed by placental malaria parasites. Proteomics 4: 1086-1093.

151. von Itzstein M, Plebanski M, Cooke BM, Coppel RL (2008) Hot, sweet and sticky: the glycobiology of Plasmodium falciparum. Trends in parasitology 24: 210-218.

152. Tham WH, Healer J, Cowman AF (2012) Erythrocyte and reticulocyte binding-like proteins of Plasmodium falciparum. Trends in parasitology 28: 23-30.

153. Baruch DI, Pasloske BL, Singh HB, Bi X, Ma XC, et al. (1995) Cloning the P. falciparum gene encoding PfEMP1, a malarial variant antigen and adherence receptor on the surface of parasitized human erythrocytes. Cell 82: 77-87.

154. Smith JD, Chitnis CE, Craig AG, Roberts DJ, Hudson-Taylor DE, et al. (1995) Switches in expression of Plasmodium falciparum var genes correlate with changes in antigenic and cytoadherent phenotypes of infected erythrocytes. Cell 82: 101-110.

155. Camus D, Hadley TJ (1985) A Plasmodium falciparum antigen that binds to host erythrocytes and merozoites. Science 230: 553-556.

156. Adams JH, Sim BK, Dolan SA, Fang X, Kaslow DC, et al. (1992) A family of erythrocyte binding proteins of malaria parasites. Proceedings of the National Academy of Sciences of the United States of America 89: 7085-7089.

157. Michon P, Stevens JR, Kaneko O, Adams JH (2002) Evolutionary relationships of conserved cysteine-rich motifs in adhesive molecules of malaria parasites. Molecular biology and evolution 19: 1128-1142.

158. Tolia NH, Enemark EJ, Sim BK, Joshua-Tor L (2005) Structural basis for the EBA-175 erythrocyte invasion pathway of the malaria parasite Plasmodium falciparum. Cell 122: 183-193.

159. Chattopadhyay D, Rayner J, McHenry AM, Adams JH (2006) The structure of the Plasmodium falciparum EBA175 ligand domain and the molecular basis of host specificity. Trends in parasitology 22: 143-145.

160. Salinas ND, Paing MM, Tolia NH (2014) Critical glycosylated residues in exon three of erythrocyte glycophorin a engage Plasmodium falciparum EBA-175 and define receptor specificity. mBio 5: e01606-01614.

161. Wanaguru M, Crosnier C, Johnson S, Rayner JC, Wright GJ (2013) Biochemical analysis of the Plasmodium falciparum erythrocyte-binding antigen-175 (EBA175)-glycophorin-A interaction: implications for vaccine design. The Journal of biological chemistry 288: 32106-32117.

162. Healer J, Thompson JK, Riglar DT, Wilson DW, Chiu YH, et al. (2013) Vaccination with conserved regions of erythrocyte-binding antigens induces neutralizing antibodies against multiple strains of Plasmodium falciparum. PloS one 8: e72504.

163. Weedall GD, Conway DJ (2010) Detecting signatures of balancing selection to identify targets of anti-parasite immunity. Trends in parasitology 26: 363-369.

164. Baum J, Thomas AW, Conway DJ (2003) Evidence for diversifying selection on erythrocyte-binding antigens of Plasmodium falciparum and P. vivax. Genetics 163: 1327-1336.

165. Richards JS, Stanisic DI, Fowkes FJ, Tavul L, Dabod E, et al. (2010) Association between naturally acquired antibodies to erythrocyte-binding antigens of Plasmodium falciparum and protection from malaria and high-density parasitemia. Clinical infectious diseases : an official publication of the Infectious Diseases Society of America 51: e50-60.

166. Osier FH, Fegan G, Polley SD, Murungi L, Verra F, et al. (2008) Breadth and magnitude of antibody responses to multiple Plasmodium falciparum merozoite antigens are associated with protection from clinical malaria. Infection and immunity 76: 2240-2248.

167. Okenu DM, Riley EM, Bickle QD, Agomo PU, Barbosa A, et al. (2000) Analysis of human antibodies to erythrocyte binding antigen 175 of Plasmodium falciparum. Infection and immunity 68: 5559-5566.

168. John CC, Moormann AM, Pregibon DC, Sumba PO, McHugh MM, et al. (2005) Correlation of high levels of antibodies to multiple pre-erythrocytic Plasmodium falciparum antigens and protection from infection. The American journal of tropical medicine and hygiene 73: 222-228.

169. Jiang L, Gaur D, Mu J, Zhou H, Long CA, et al. (2011) Evidence for erythrocyte-binding antigen 175 as a component of a ligand-blocking blood-stage malaria vaccine. Proceedings of the National Academy of Sciences of the United States of America 108: 7553-7558.

170. Pandey KC, Singh S, Pattnaik P, Pillai CR, Pillai U, et al. (2002) Bacterially expressed and refolded receptor binding domain of Plasmodium falciparum EBA-175 elicits invasion inhibitory antibodies. Molecular and biochemical parasitology 123: 23-33.

171. Cramer JP, Mockenhaupt FP, Mohl I, Dittrich S, Dietz E, et al. (2004) Allelic dimorphism of the erythrocyte binding antigen-175 (eba-175) gene of Plasmodium falciparum and severe malaria: Significant association of the C-segment with fatal outcome in Ghanaian children. Malaria journal 3: 11.

172. Perce-da-Silva DS, Banic DM, Lima-Junior JC, Santos F, Daniel-Ribeiro CT, et al. (2011) Evaluation of allelic forms of the erythrocyte binding antigen 175 (EBA-175) in

Plasmodium falciparum field isolates from Brazilian endemic area. Malaria journal 10: 146.

173. Soulama I, Bougouma EC, Diarra A, Nebie I, Sirima SB (2010) Low-high season variation in Plasmodium falciparum erythrocyte binding antigen 175 (eba-175) allelic forms in malaria endemic area of Burkina Faso. Tropical medicine & international health : TM & IH 15: 51-59.

174. Toure FS, Bisseye C, Mavoungou E (2006) Imbalanced distribution of Plasmodium falciparum EBA-175 genotypes related to clinical status in children from Bakoumba, Gabon. Clinical medicine & research 4: 7-11.

175. Binks RH, Baum J, Oduola AM, Arnot DE, Babiker HA, et al. (2001) Population genetic analysis of the Plasmodium falciparum erythrocyte binding antigen-175 (eba-175) gene. Molecular and biochemical parasitology 114: 63-70.

176. Chokejindachai W, Conway DJ (2009) Case-control approach to identify Plasmodium falciparum polymorphisms associated with severe malaria. PloS one 4: e5454.

177. Coatney GR (1963) Pitfalls in a discovery: the chronicle of chloroquine. The American journal of tropical medicine and hygiene 12: 121-128.

178. Krafts K, Hempelmann E, Skorska-Stania A (2012) From methylene blue to chloroquine: a brief review of the development of an antimalarial therapy. Parasitology research 111: 1-6.

179. Bray PG, Mungthin M, Ridley RG, Ward SA (1998) Access to hematin: the basis of chloroquine resistance. Molecular pharmacology 54: 170-179.

180. Combrinck JM, Mabotha TE, Ncokazi KK, Ambele MA, Taylor D, et al. (2013) Insights into the role of heme in the mechanism of action of antimalarials. ACS chemical biology 8: 133-137.

181. Ginsburg H, Famin O, Zhang J, Krugliak M (1998) Inhibition of glutathione-dependent degradation of heme by chloroquine and amodiaquine as a possible basis for their antimalarial mode of action. Biochemical pharmacology 56: 1305-1313.

182. Laurent F, Saivin S, Chretien P, Magnaval JF, Peyron F, et al. (1993) Pharmacokinetic and pharmacodynamic study of amodiaquine and its two metabolites after a single oral dose in human volunteers. Arzneimittel-Forschung 43: 612-616.

183. Krishna S, White NJ (1996) Pharmacokinetics of quinine, chloroquine and amodiaquine. Clinical implications. Clinical pharmacokinetics 30: 263-299.

184. Stepniewska K, Taylor W, Sirima SB, Ouedraogo EB, Ouedraogo A, et al. (2009) Population pharmacokinetics of artesunate and amodiaquine in African children. Malaria journal 8: 200.

185. Hocart SJ, Liu H, Deng H, De D, Krogstad FM, et al. (2011) 4-aminoquinolines active against chloroquine-resistant Plasmodium falciparum: basis of antiparasite activity and quantitative structure-activity relationship analyses. Antimicrobial agents and chemotherapy 55: 2233-2244.

186. Chou AC, Fitch CD (1993) Control of heme polymerase by chloroquine and other quinoline derivatives. Biochemical and biophysical research communications 195: 422-427.

187. Chou AC, Chevli R, Fitch CD (1980) Ferriprotoporphyrin IX fulfills the criteria for identification as the chloroquine receptor of malaria parasites. Biochemistry 19: 1543-1549.

188. Foley M, Tilley L (1998) Quinoline antimalarials: mechanisms of action and resistance and prospects for new agents. Pharmacology & therapeutics 79: 55-87.

189. Chevli R, Fitch CD (1982) The antimalarial drug mefloquine binds to membrane phospholipids. Antimicrobial agents and chemotherapy 21: 581-586.

190. Klayman DL (1985) Qinghaosu (artemisinin): an antimalarial drug from China. Science 228: 1049-1055.

191. Benakis A, Paris M, Loutan L, Plessas CT, Plessas ST (1997) Pharmacokinetics of artemisinin and artesunate after oral administration in healthy volunteers. The American journal of tropical medicine and hygiene 56: 17-23.

192. Cheeseman IH, Miller BA, Nair S, Nkhoma S, Tan A, et al. (2012) A major genome region underlying artemisinin resistance in malaria. Science 336: 79-82.

193. Takala-Harrison S, Clark TG, Jacob CG, Cummings MP, Miotto O, et al. (2013) Genetic loci associated with delayed clearance of Plasmodium falciparum following artemisinin treatment in Southeast Asia. Proceedings of the National Academy of Sciences of the United States of America 110: 240-245.

194. Mbengue A, Bhattacharjee S, Pandharkar T, Liu H, Estiu G, et al. (2015) A molecular mechanism of artemisinin resistance in Plasmodium falciparum malaria. Nature 520: 683-687.

195. Ferone R (1977) Folate metabolism in malaria. Bulletin of the World Health Organization 55: 291-298.

196. Hyde JE (2005) Exploring the folate pathway in Plasmodium falciparum. Acta tropica 94: 191-206.

197. Triglia T, Cowman AF (1999) The mechanism of resistance to sulfa drugs in Plasmodium falciparum. Drug resistance updates : reviews and commentaries in antimicrobial and anticancer chemotherapy 2: 15-19.

198. Ferone R, Roland S (1980) Dihydrofolate reductase: thymidylate synthase, a bifunctional polypeptide from Crithidia fasciculata. Proceedings of the National Academy of Sciences of the United States of America 77: 5802-5806.

199. Garrett CE, Coderre JA, Meek TD, Garvey EP, Claman DM, et al. (1984) A bifunctional thymidylate synthetase-dihydrofolate reductase in protozoa. Molecular and biochemical parasitology 11: 257-265.

200. Stechmann A, Cavalier-Smith T (2003) The root of the eukaryote tree pinpointed. Current biology : CB 13: R665-666.

201. Peterson DS, Walliker D, Wellems TE (1988) Evidence that a point mutation in dihydrofolate reductase-thymidylate synthase confers resistance to pyrimethamine in falciparum malaria. Proceedings of the National Academy of Sciences of the United States of America 85: 9114-9118.

202. Cowman AF, Morry MJ, Biggs BA, Cross GA, Foote SJ (1988) Amino acid changes linked to pyrimethamine resistance in the dihydrofolate reductase-thymidylate synthase gene of Plasmodium falciparum. Proceedings of the National Academy of Sciences of the United States of America 85: 9109-9113.

203. Sibley CH, Hyde JE, Sims PF, Plowe CV, Kublin JG, et al. (2001) Pyrimethamine-sulfadoxine resistance in Plasmodium falciparum: what next? Trends in parasitology 17: 582-588.

204. White NJ (1996) The treatment of malaria. The New England journal of medicine 335: 800-806.

205. Branton D, Deamer DW, Marziali A, Bayley H, Benner SA, et al. (2008) The potential and challenges of nanopore sequencing. Nature biotechnology 26: 1146-1153.

206. Clarke J, Wu HC, Jayasinghe L, Patel A, Reid S, et al. (2009) Continuous base identification for single-molecule nanopore DNA sequencing. Nature nanotechnology 4: 265-270.

207. Koressaar T, Remm M (2007) Enhancements and modifications of primer design program Primer3. Bioinformatics 23: 1289-1291.

208. Sasi P, Abdulrahaman A, Mwai L, Muriithi S, Straimer J, et al. (2009) In vivo and in vitro efficacy of amodiaquine against Plasmodium falciparum in an area of continued use of 4-aminoquinolines in East Africa. The Journal of infectious diseases 199: 1575-1582.

209. Borrmann S, Sasi P, Mwai L, Bashraheil M, Abdallah A, et al. (2011) Declining responsiveness of Plasmodium falciparum infections to artemisinin-based combination treatments on the Kenyan coast. PloS one 6: e26005.

210. Okombo J, Kiara SM, Rono J, Mwai L, Pole L, et al. (2010) In vitro activities of quinine and other antimalarials and pfnhe polymorphisms in Plasmodium isolates from Kenya. Antimicrobial agents and chemotherapy 54: 3302-3307.

211. Storey JD, Tibshirani R (2003) Statistical significance for genomewide studies. Proceedings of the National Academy of Sciences of the United States of America 100: 9440-9445.

212. Rozas J, Sanchez-DelBarrio JC, Messeguer X, Rozas R (2003) DnaSP, DNA polymorphism analyses by the coalescent and other methods. Bioinformatics 19: 2496-2497.

213. Saitou N, Nei M (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. Molecular biology and evolution 4: 406-425.

214. Tamura K, Stecher G, Peterson D, Filipski A, Kumar S (2013) MEGA6: Molecular Evolutionary Genetics Analysis version 6.0. Molecular biology and evolution 30: 2725-2729.

215. Pond SL, Frost SD, Muse SV (2005) HyPhy: hypothesis testing using phylogenies. Bioinformatics 21: 676-679.

216. Hogue CW (1997) Cn3D: a new generation of three-dimensional molecular structure viewer. Trends in biochemical sciences 22: 314-316.

217. Band G, Le QS, Jostins L, Pirinen M, Kivinen K, et al. (2013) Imputation-based meta-analysis of severe malaria in three African populations. PLoS genetics 9: e1003509.

218. MalariaGEN (2014) Reappraisal of known malaria resistance loci in a large multicenter study. Nature genetics 46: 1197-1204.

219. DOMC (2009) National Malaria Strategy 2009–2017. Kenya Division of Malaria Control Ministry of Public Health and Sanitation.

220. Sibley CH, Guerin PJ, Ringwald P (2010) Monitoring antimalarial resistance: launching a cooperative effort. Trends in parasitology 26: 221-224.

221. Ecker A, Lehane AM, Clain J, Fidock DA (2012) PfCRT and its role in antimalarial drug resistance. Trends in parasitology 28: 504-514.

222. Wellems TE, Walker-Jonah A, Panton LJ (1991) Genetic mapping of the chloroquine-resistance locus on Plasmodium falciparum chromosome 7. Proceedings of the National Academy of Sciences of the United States of America 88: 3382-3386.

223. Fidock DA, Nomura T, Talley AK, Cooper RA, Dzekunov SM, et al. (2000) Mutations in the P. falciparum digestive vacuole transmembrane protein PfCRT and evidence for their role in chloroquine resistance. Molecular cell 6: 861-871.

224. Kiara SM, Okombo J, Masseno V, Mwai L, Ochola I, et al. (2009) In vitro activity of antifolate and polymorphism in dihydrofolate reductase of Plasmodium falciparum isolates from the Kenyan coast: emergence of parasites with Ile-164-Leu mutation. Antimicrobial agents and chemotherapy 53: 3793-3798.

225. Mwai L, Kiara SM, Abdirahman A, Pole L, Rippert A, et al. (2009) In vitro activities of piperaquine, lumefantrine, and dihydroartemisinin in Kenyan Plasmodium falciparum isolates and polymorphisms in pfcrt and pfmdr1. Antimicrobial agents and chemotherapy 53: 5069-5073.

226. Van Tyne D, Park DJ, Schaffner SF, Neafsey DE, Angelino E, et al. (2011) Identification and functional validation of the novel antimalarial resistance locus PF10_0355 in Plasmodium falciparum. PLoS genetics 7: e1001383.

227. Yuan J, Cheng KC, Johnson RL, Huang R, Pattaradilokrat S, et al. (2011) Chemical genomic profiling for antimalarial therapies, response signatures, and molecular targets. Science 333: 724-729.

228. Park DJ, Lukens AK, Neafsey DE, Schaffner SF, Chang HH, et al. (2012) Sequence-based association and selection scans identify drug resistance loci in the Plasmodium falciparum malaria parasite. Proceedings of the National Academy of Sciences of the United States of America 109: 13052-13057.

229. Chin W, Collins WE (1980) Comparative studies of three strains of Plasmodium falciparum isolated by the culture method of Trager and Jensen. The American journal of tropical medicine and hygiene 29: 1143-1146.

230. Su X, Kirkman LA, Fujioka H, Wellems TE (1997) Complex polymorphisms in an approximately 330 kDa protein are linked to chloroquine-resistant P. falciparum in Southeast Asia and Africa. Cell 91: 593-603.

231. Fidock DA, Nomura T, Cooper RA, Su X, Talley AK, et al. (2000) Allelic modifications of the cg2 and cg1 genes do not alter the chloroquine response of drug-resistant Plasmodium falciparum. Molecular and biochemical parasitology 110: 1-10.

232. Zheng G, Freidlin B, Gastwirth JL (2006) Robust genomic control for association studies. American journal of human genetics 78: 350-356.

233. Duah NO, Quashie NB, Abuaku BK, Sebeny PJ, Kronmann KC, et al. (2012) Surveillance of Molecular Markers of Plasmodium falciparum Resistance to Sulphadoxine-Pyrimethamine 5 Years after the Change of Malaria Treatment Policy in Ghana. The American journal of tropical medicine and hygiene.

234. Andriantsoanirina V, Ratsimbasoa A, Bouchier C, Jahevitra M, Rabearimanana S, et al. (2009) Plasmodium falciparum drug resistance in Madagascar: facing the spread of unusual pfdhfr and pfmdr-1 haplotypes and the decrease of dihydroartemisinin susceptibility. Antimicrobial agents and chemotherapy 53: 4588-4597.

235. Ferdig MT, Cooper RA, Mu J, Deng B, Joy DA, et al. (2004) Dissecting the loci of low-level quinine resistance in malaria parasites. Molecular microbiology 52: 985-997.

236. Henry M, Briolant S, Zettor A, Pelleau S, Baragatti M, et al. (2009) Plasmodium falciparum Na+/H+ exchanger 1 transporter is involved in reduced susceptibility to quinine. Antimicrobial agents and chemotherapy 53: 1926-1930.

237. Sisowath C, Petersen I, Veiga MI, Martensson A, Premji Z, et al. (2009) In vivo selection of Plasmodium falciparum parasites carrying the chloroquine-susceptible pfcrt K76 allele after treatment with artemether-lumefantrine in Africa. The Journal of infectious diseases 199: 750-757.

238. Basco LK, Ringwald P (2003) In vitro activities of piperaquine and other 4-aminoquinolines against clinical isolates of Plasmodium falciparum in Cameroon. Antimicrobial agents and chemotherapy 47: 1391-1394.

239. Mu J, Awadalla P, Duan J, McGee KM, Keebler J, et al. (2007) Genome-wide variation and identification of vaccine targets in the Plasmodium falciparum genome. Nature genetics 39: 126-130.

240. Miotto O, Amato R, Ashley EA, MacInnis B, Almagro-Garcia J, et al. (2015) Genetic architecture of artemisinin-resistant Plasmodium falciparum. Nature genetics 47: 226-234.

241. Watkins WM, Sixsmith DG, Spencer HC, Boriga DA, Kariuki DM, et al. (1984) Effectiveness of amodiaquine as treatment for chloroquine-resistant Plasmodium falciparum infections in Kenya. Lancet 1: 357-359.

242. Shretta R, Omumbo J, Rapuoda B, Snow RW (2000) Using evidence to change antimalarial drug policy in Kenya. Tropical medicine & international health : TM & IH 5: 755-764.

243. Mwai L, Ochong E, Abdirahman A, Kiara SM, Ward S, et al. (2009) Chloroquine resistance before and after its withdrawal in Kenya. Malaria journal 8: 106.

244. WHO (2007) Antimalarial Medicines in Kenya. A baseline study undertaken prior to nationwide distribution of artmether-lumefantrine (AL) in Kenya. World Health Organization.

245. Daily JP, Roberts C, Thomas SM, Ndir O, Dieng T, et al. (2003) Prevalence of Plasmodium falciparum pfcrt polymorphisms and in vitro chloroquine sensitivity in Senegal. Parasitology 126: 401-405.

246. Holmgren G, Gil JP, Ferreira PM, Veiga MI, Obonyo CO, et al. (2006) Amodiaquine resistant Plasmodium falciparum malaria in vivo is associated with selection of pfcrt 76T and pfmdr1 86Y. Infection, genetics and evolution : journal of molecular epidemiology and evolutionary genetics in infectious diseases 6: 309-314.

247. Ochong EO, van den Broek IV, Keus K, Nzila A (2003) Short report: association between chloroquine and amodiaquine resistance and allelic variation in the Plasmodium falciparum multiple drug resistance 1 gene and the chloroquine resistance transporter gene in isolates from the upper Nile in southern Sudan. The American journal of tropical medicine and hygiene 69: 184-187.

248. Beshir K, Sutherland CJ, Merinopoulos I, Durrani N, Leslie T, et al. (2010) Amodiaquine resistance in Plasmodium falciparum malaria in Afghanistan is associated with the pfcrt SVMNT allele at codons 72 to 76. Antimicrobial agents and chemotherapy 54: 3714-3716.

249. Nsobya SL, Dokomajilar C, Joloba M, Dorsey G, Rosenthal PJ (2007) Resistance-mediating Plasmodium falciparum pfcrt and pfmdr1 alleles after treatment with artesunate-amodiaquine in Uganda. Antimicrobial agents and chemotherapy 51: 3023-3025.

250. Sa JM, Twu O, Hayton K, Reyes S, Fay MP, et al. (2009) Geographic patterns of Plasmodium falciparum drug resistance distinguished by differential responses to amodiaquine and chloroquine. Proceedings of the National Academy of Sciences of the United States of America 106: 18883-18889.

251. Heis MD, Ditmer EM, de Oliveira LA, Frazzon AP, Margis R, et al. (2011) Differential expression of cysteine desulfurases in soybean. BMC plant biology 11: 166.

252. Kublin JG, Cortese JF, Njunju EM, Mukadam RA, Wirima JJ, et al. (2003) Reemergence of chloroquine-sensitive Plasmodium falciparum malaria after cessation of chloroquine use in Malawi. The Journal of infectious diseases 187: 1870-1875.

253. Mita T, Kaneko A, Lum JK, Zungu IL, Tsukahara T, et al. (2004) Expansion of wild type allele rather than back mutation in pfcrt explains the recent recovery of chloroquine sensitivity of Plasmodium falciparum in Malawi. Molecular and biochemical parasitology 135: 159-163.

254. Laufer MK, Takala-Harrison S, Dzinjalamala FK, Stine OC, Taylor TE, et al. (2010) Return of chloroquine-susceptible falciparum malaria in Malawi was a reexpansion of diverse susceptible parasites. The Journal of infectious diseases 202: 801-808.

255. Bray PG, Deed S, Fox E, Kalkanidis M, Mungthin M, et al. (2005) Primaquine synergises the activity of chloroquine against chloroquine-resistant P. falciparum. Biochemical pharmacology 70: 1158-1166.

256. Martin SK, Oduola AM, Milhous WK (1987) Reversal of chloroquine resistance in Plasmodium falciparum by verapamil. Science 235: 899-901.

257. Patel JJ, Thacker D, Tan JC, Pleeter P, Checkley L, et al. (2010) Chloroquine susceptibility and reversibility in a Plasmodium falciparum genetic cross. Molecular microbiology 78: 770-787.

258. Vale N, Moreira R, Gomes P (2009) Primaquine revisited six decades after its discovery. European journal of medicinal chemistry 44: 937-953.

259. Baird JK, Surjadjaja C (2011) Consideration of ethics in primaquine therapy against malaria transmission. Trends in parasitology 27: 11-16.

260. Graf PC, Durand S, Alvarez Antonio C, Montalvan C, Galves Montoya M, et al. (2012) Failure of Supervised Chloroquine and Primaquine Regimen for the Treatment of Plasmodium vivax in the Peruvian Amazon. Malaria research and treatment 2012: 936067.

261. Okombo J, Kiara SM, Mwai L, Pole L, Ohuma E, et al. (2012) Baseline in vitro activities of the antimalarials pyronaridine and methylene blue against Plasmodium falciparum isolates from Kenya. Antimicrobial agents and chemotherapy 56: 1105-1107.

262. Otto TD, Wilinski D, Assefa S, Keane TM, Sarry LR, et al. (2010) New insights into the blood-stage transcriptome of Plasmodium falciparum using RNA-Seq. Molecular microbiology 76: 12-24.

263. Clayton DG, Walker NM, Smyth DJ, Pask R, Cooper JD, et al. (2005) Population structure, differential bias and genomic control in a large-scale, case-control association study. Nature genetics 37: 1243-1246.

264. Sibley CH (2015) Infectious diseases. Understanding artemisinin resistance. Science 347: 373-374.

265. Dondorp AM, Nosten F, Yi P, Das D, Phyo AP, et al. (2009) Artemisinin resistance in Plasmodium falciparum malaria. The New England journal of medicine 361: 455-467.

266. Venkatesan M, Amaratunga C, Campino S, Auburn S, Koch O, et al. (2012) Using CF11 cellulose columns to inexpensively and effectively remove human DNA from Plasmodium falciparum-infected whole blood samples. Malaria journal 11: 41.

267. Law PJ, Claudel-Renard C, Joubert F, Louw AI, Berger DK (2008) MADIBA: a web server toolkit for biological interpretation of Plasmodium and plant gene clusters. BMC genomics 9: 105.

268. Auburn S, Campino S, Miotto O, Djimde AA, Zongo I, et al. (2012) Characterization of within-host Plasmodium falciparum diversity using next-generation sequence data. PloS one 7: e32891.

269. Buffet PA, Gamain B, Scheidig C, Baruch D, Smith JD, et al. (1999) Plasmodium falciparum domain mediating adhesion to chondroitin sulfate A: a receptor for human placental infection. Proceedings of the National Academy of Sciences of the United States of America 96: 12743-12748.

270. Rowe JA, Moulds JM, Newbold CI, Miller LH (1997) P. falciparum rosetting mediated by a parasite-variant erythrocyte membrane protein and complement-receptor 1. Nature 388: 292-295.

271. Iqbal Z, Caccamo M, Turner I, Flicek P, McVean G (2012) De novo assembly and genotyping of variants using colored de Bruijn graphs. Nature genetics 44: 226-232.

272. El Sahly HM, Patel SM, Atmar RL, Lanford TA, Dube T, et al. (2010) Safety and immunogenicity of a recombinant nonglycosylated erythrocyte binding antigen 175 Region II malaria vaccine in healthy adults living in an area where malaria is not endemic. Clinical and vaccine immunology : CVI 17: 1552-1559.

273. Pasvol G, Wainscoat JS, Weatherall DJ (1982) Erythrocytes deficiency in glycophorin resist invasion by the malarial parasite Plasmodium falciparum. Nature 297: 64-66.

274. Fowkes FJ, Richards JS, Simpson JA, Beeson JG (2010) The relationship between anti-merozoite antibodies and incidence of Plasmodium falciparum malaria: A systematic review and meta-analysis. PLoS medicine 7: e1000218.

275. Toure FS, Mavoungou E, Ndong JM, Tshipamba P, Deloron P (2001) Erythrocyte binding antigen (EBA-175) of Plasmodium falciparum: improved genotype determination by nested polymerase chain reaction. Tropical medicine & international health : TM & IH 6: 767-769.

276. Ludwig M, Wohn KD, Schleuning WD, Olek K (1992) Allelic dimorphism in the human tissue-type plasminogen activator (TPA) gene as a result of an Alu insertion/deletion event. Human genetics 88: 388-392.

277. Valle-Garay E, Montes AH, Corte JR, Meana A, Fierer J, et al. (2013) tPA Alu (I/D) polymorphism associates with bacterial osteomyelitis. The Journal of infectious diseases 208: 218-223.

278. van der Bom JG, de Knijff P, Haverkate F, Bots ML, Meijer P, et al. (1997) Tissue plasminogen activator and risk of myocardial infarction. The Rotterdam Study. Circulation 95: 2623-2627.

279. Zivkovic M, Starcevic Cizmarevic N, Lovrecic L, Klupka-Saric I, Stankovic A, et al. (2014) The role of TPA I/D and PAI-1 4G/5G polymorphisms in multiple sclerosis. Disease markers 2014: 362708.

280. Branco CC, Palla R, Lino S, Pacheco PR, Cabral R, et al. (2006) Assessment of Azorean ancestry by Alu insertion polymorphisms. American journal of human biology : the official journal of the Human Biology Council 18: 223-226.

281. Frampton M, Houlston R (2012) Generation of artificial FASTQ files to evaluate the performance of next-generation sequencing pipelines. PloS one 7: e49110.

282. Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, et al. (2012) An integrated map of genetic variation from 1,092 human genomes. Nature 491: 56-65.

283. Pevzner PA, Tang H, Waterman MS (2001) An Eulerian path approach to DNA fragment assembly. Proceedings of the National Academy of Sciences of the United States of America 98: 9748-9753.

284. Manske HM, Kwiatkowski DP (2009) LookSeq: a browser-based viewer for deep sequencing data. Genome research 19: 2125-2132.

285. Walliker D, Quakyi IA, Wellems TE, McCutchan TF, Szarfman A, et al. (1987) Genetic analysis of the human malaria parasite Plasmodium falciparum. Science 236: 1661-1666.

286. Wellems TE, Panton LJ, Gluzman IY, do Rosario VE, Gwadz RW, et al. (1990) Chloroquine resistance not linked to mdr-like genes in a Plasmodium falciparum cross. Nature 345: 253-255.

287. Hayton K, Gaur D, Liu A, Takahashi J, Henschen B, et al. (2008) Erythrocyte binding protein PfRH5 polymorphisms determine species-specific pathways of Plasmodium falciparum invasion. Cell host & microbe 4: 40-51.

288. Sidhu AB, Verdier-Pinard D, Fidock DA (2002) Chloroquine resistance in Plasmodium falciparum malaria parasites conferred by pfcrt mutations. Science 298: 210-213.

289. Duffy MF, Caragounis A, Noviyanti R, Kyriacou HM, Choong EK, et al. (2006) Transcribed var genes associated with placental malaria in Malawian women. Infection and immunity 74: 4875-4883.

290. Namiki T, Hachiya T, Tanaka H, Sakakibara Y (2012) MetaVelvet: an extension of Velvet assembler to de novo metagenome assembly from short sequence reads. Nucleic acids research 40: e155.

291. Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25: 1754-1760.

292. Chiu CY, Hodder AN, Lin CS, Hill DL, Li Wai Suen CS, et al. (2015) Antibodies to the Plasmodium falciparum Proteins MSPDBL1 and MSPDBL2 Opsonize Merozoites, Inhibit Parasite Growth, and Predict Protection From Clinical Malaria. The Journal of infectious diseases.

293. Pinkevych M, Petravic J, Bereczky S, Rooth I, Farnert A, et al. (2015) Understanding the relationship between Plasmodium falciparum growth rate and multiplicity of infection. The Journal of infectious diseases 211: 1121-1127.

294. Picard http://broadinstitute.github.io/picard.

295. Zerbino DR (2010) Using the Velvet de novo assembler for short-read sequencing technologies. Current protocols in bioinformatics / editoral board, Andreas D Baxevanis [et al] Chapter 11: Unit 11 15.

296. Ruby JG, Bellare P, Derisi JL (2013) PRICE: software for the targeted assembly of components of (Meta) genomic sequence data. G3 3: 865-880.

297. Treangen TJ, Sommer DD, Angly FE, Koren S, Pop M (2011) Next generation sequence assembly with AMOS. Current protocols in bioinformatics / editoral board, Andreas D Baxevanis [et al] Chapter 11: Unit 11 18.

298. Otto TD, Sanders M, Berriman M, Newbold C (2010) Iterative Correction of Reference Nucleotides (iCORN) using second generation sequencing technology. Bioinformatics 26: 1704-1707.

299. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, et al. (2009) The Sequence Alignment/Map format and SAMtools. Bioinformatics 25: 2078-2079.

300. Polhemus ME, Magill AJ, Cummings JF, Kester KE, Ockenhouse CF, et al. (2007) Phase I dose escalation safety and immunogenicity trial of Plasmodium falciparum apical membrane protein (AMA-1) FMP2.1, adjuvanted with AS02A, in malaria-naive adults at the Walter Reed Army Institute of Research. Vaccine 25: 4203-4212.

301. Lin CS, Uboldi AD, Marapana D, Czabotar PE, Epp C, et al. (2014) The merozoite surface protein 1 complex is a platform for binding to human erythrocytes by Plasmodium falciparum. The Journal of biological chemistry 289: 25655-25669.

302. Tetteh KK, Osier FH, Salanti A, Kamuyu G, Drought L, et al. (2013) Analysis of antibodies to newly described Plasmodium falciparum merozoite antigens supports MSPDBL2 as a predicted target of naturally acquired immunity. Infection and immunity 81: 3835-3842.

303. Conway DJ (2015) Paths to a malaria vaccine illuminated by parasite genomics. Trends in genetics : TIG 31: 97-107.

304. Wernersson R, Pedersen AG (2003) RevTrans: Multiple alignment of coding DNA from aligned amino acid sequences. Nucleic acids research 31: 3537-3539.

305. VanLiere JM, Rosenberg NA (2008) Mathematical properties of the r2 measure of linkage disequilibrium. Theoretical population biology 74: 130-137.

306. Ocholla H, Preston MD, Mipando M, Jensen AT, Campino S, et al. (2014) Whole-genome scans provide evidence of adaptive evolution in Malawian Plasmodium falciparum isolates. The Journal of infectious diseases 210: 1991-2000.

307. Tajima F (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123: 585-595.

308. Nwakanma DC, Duffy CW, Amambua-Ngwa A, Oriero EC, Bojang KA, et al. (2014) Changes in malaria parasite drug resistance in an endemic population over a 25-year period with resulting genomic evidence of selection. The Journal of infectious diseases 209: 1126-1135.

309. Lee AJ, Das SR, Wang W, Fitzgerald T, Pickett BE, et al. (2015) Diversifying selection analysis predicts antigenic evolution of 2009 pandemic H1N1 influenza A virus in humans. Journal of virology.

310. Panagiotou OA, Ioannidis JP (2012) What should the genome-wide significance threshold be? Empirical replication of borderline genetic associations. International journal of epidemiology 41: 273-286.

311. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, et al. (2002) The human genome browser at UCSC. Genome research 12: 996-1006.

312. Karolchik D, Hinrichs AS, Furey TS, Roskin KM, Sugnet CW, et al. (2004) The UCSC Table Browser data retrieval tool. Nucleic acids research 32: D493-496.

313. Ahouidi AD, Bei AK, Neafsey DE, Sarr O, Volkman S, et al. (2010) Population genetic analysis of large sequence polymorphisms in Plasmodium falciparum blood-stage antigens. Infection, genetics and evolution : journal of molecular epidemiology and evolutionary genetics in infectious diseases 10: 200-206.

314. Dittrich S, Schwobel B, Jordan S, Vanisaveth V, Rattanaxay P, et al. (2003) Distribution of the two forms of Plasmodium falciparum erythrocyte binding antigen-175 (eba-175) gene in Lao PDR. Malaria journal 2: 23.

315. Verra F, Chokejindachai W, Weedall GD, Polley SD, Mwangi TW, et al. (2006) Contrasting signatures of selection on the Plasmodium falciparum erythrocyte binding antigen gene family. Molecular and biochemical parasitology 149: 182-190.

316. Ozwara H, Kocken CH, Conway DJ, Mwenda JM, Thomas AW (2001) Comparative analysis of Plasmodium reichenowi and P. falciparum erythrocyte-binding proteins reveals selection to maintain polymorphism in the erythrocyte-binding region of EBA-175. Molecular and biochemical parasitology 116: 81-84.

317. Douglas AD, Williams AR, Illingworth JJ, Kamuyu G, Biswas S, et al. (2011) The blood-stage malaria antigen PfRH5 is susceptible to vaccine-inducible cross-strain neutralizing antibody. Nature Communications 2.

318. Tran TM, Ongoiba A, Coursen J, Crosnier C, Diouf A, et al. (2014) Naturally acquired antibodies specific for Plasmodium falciparum reticulocyte-binding protein homologue 5 inhibit parasite growth and predict protection from malaria. The Journal of infectious diseases 209: 789-798.

319. Baum J, Ward RH, Conway DJ (2002) Natural selection on the erythrocyte surface. Molecular biology and evolution 19: 223-229.

320. Sim BK, Chitnis CE, Wasniowska K, Hadley TJ, Miller LH (1994) Receptor and ligand domains for invasion of erythrocytes by Plasmodium falciparum. Science 264: 1941-1944.

321. Reid ME (2009) MNS blood group system: a review. Immunohematology / American Red Cross 25: 95-101.

322. Sato Y (2012) The vasohibin family: Novel regulators of angiogenesis. Vascular pharmacology 56: 262-266.

323. Luo J, Sladek R, Bader JA, Matthyssen A, Rossant J, et al. (1997) Placental abnormalities in mouse embryos lacking the orphan nuclear receptor ERR-beta. Nature 388: 778-782.

324. Soulama I, Bigoga JD, Ndiaye M, Bougouma EC, Quagraine J, et al. (2011) Genetic diversity of polymorphic vaccine candidate antigens (apical membrane antigen-1, merozoite surface protein-3, and erythrocyte binding antigen-175) in Plasmodium falciparum isolates from western and central Africa. The American journal of tropical medicine and hygiene 84: 276-284.

325. Procter JB, Thompson J, Letunic I, Creevey C, Jossinet F, et al. (2010) Visualization of multiple alignments, phylogenies and gene family evolution. Nature methods 7: S16-25.

326. Llinas M, Bozdech Z, Wong ED, Adai AT, DeRisi JL (2006) Comparative whole genome transcriptome analysis of three Plasmodium falciparum strains. Nucleic acids research 34: 1166-1173.

327. Leek JT, Storey JD (2007) Capturing heterogeneity in gene expression studies by surrogate variable analysis. PLoS genetics 3: 1724-1735.