A Tale of Brothers, Sisters, Aunts and Uncles: Using Genomics and Modeling to Uncover the Nature of P. Falciparum Polygenomic Infections and Cotransmission

A tale of brothers, sisters, aunts and uncles: using genomics and modeling to uncover the nature of P. falciparum polygenomic infections and cotransmission

A dissertation presented

by

Wesley Wong

to

The Department of Immunology and Infectious Diseases

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

in the subject of

Biological Sciences in Public Health

Harvard University

Cambridge, Massachusetts

December 2017

Dissertation Advisor: Professor Dyann F. Wirth Wesley Wong

A tale of brothers, sisters, aunts and uncles: using genomics and modeling to uncover the nature of P. falciparum polygenomic infections and cotransmission

Abstract

A curious feature of malaria epidemiology is the presence of polygenomic (multiple strain) infections in natural parasite populations. Polygenomic infections are an important aspect of malaria transmission and a necessary prerequisite for outcrossing. From a public health perspective, the genomic composition of polygenomic infections can be used to better understand malaria transmission and to monitor changes in transmission intensity. From an evolutionary perspective, polygenomic infections allow genetic exchange between coinfecting strains and alter parasite population genetics.

In this thesis, I use a mix of computational biology tools, ranging from bioinformatics and sequencing analysis to mathematical modeling, to understand the genomic composition of polygenomic infections and the consequences of coinfection in the context of malaria population genetics and public health.

First, I analyzed the genetic relatedness of coinfecting strains in polygenomic infections collected from Thiès, Senegal. I show that the relatedness of coinfecting strains in polygenomic infections are incompatible with the expectations of pure superinfection, which suggests that cotransmission is common in natural populations.

Second, I used a mathematical model to quantify the expected relatedness of cotransmitted strains. I demonstrate that there are only 9 different ways that cotransmitted parasites can be related to one another. I show that the relatedness of polygenomic infections depends on the conditions of the initial infection and that different transmission lineages have different expectations of polygenomic relatedness.

Third, I analyzed the sequencing quality of lab-generated mock infections to determine whether selective whole genome amplification could be used to accurately sequence polygenomic infections. I found that selective whole genome amplification could be used to characterize the genomic composition of polygenomic infections, even when there is a significant amount of contaminating host DNA present.

Finally, I interrogate how coinfection and transmission topology affects malaria population genetics and evolution by performing evolutionary invasion analyses. This work borrows heavily from theoretical evolutionary population genetics and is designed to show how modeling can be used to highlight importance features of malaria transmission.

The use of population genomics for understanding parasite transmission and evolution hinges on our ability to integrate population genetics into existing epidemiological frameworks. The integration of these fields will require advances in both data generation and theory development. This research contributes to our understanding of malaria population genomics and the importance of coinfection and sexual recombination in the context of transmission.

Acknowledgements

The work reported in this dissertation was performed under the supervision of Professor Dyann F. Wirth. When I first started my PhD, I had decided that I was a terrible molecular biologist and that I wanted to work on population genomics. This was a bold decision at the time, for I had neither the population genetic background nor the computational skills needed to carry out such work. I am forever indebted to Dyann for being willing to take a chance with me and guide me through the PhD process. She has been an excellent source of inspiration and support throughout my PhD. Equal credit must also be given to Professor Daniel L. Hartl at the Organismal and Evolutionary Department at Harvard University. He has been a constant source of mentoring and provided me with the best critical feedback a PhD student could hope for.

I would also like to thank Professor Daniel E. Neafsey and his team at the Broad Institute. He and his team provided an excellent learning environment where I could immerse myself in the complexities of computational genomics. I would also like to thank Sarah K. Volkman, who has also been a great mentor and a constant source of inspiration. Finally, I must thank Edward Wenger at the Institute for Disease modeling. Without him, the modeling sections of this thesis would be sorely lacking.

Due to the sheer number of mentors that I have had, I owe my success to all members of the Wirth Lab, Hartl Lab and Neafsey lab. Members of each lab had something different to offer and have contributed to a richer PhD experience. Selina Bopp has been especially helpful and interesting to work with.

I am also indebted to the many friends that have kept me sane throughout my PhD. These include members of my PhD cohort and people that I met in Boston. Without people such as Stanley Wang, Emma (Ye) Tang, Erika Ilagan, Ava Lee, Michelle Bau, Anny (I-Ni) Hsieh, Jasin Wong, Jason Lee, Zhouwei Zhang, Mia (Chun-jou) Tsai, I doubt that I could have finished my PhD with my mental health intact.

Last, but certainly not least, I am especially grateful to my parents and sister for all that they have provided me. Without them, I would certainly not have made it this far. Their unconditional support has been a constant morale boost during the most difficult periods of my life.

Finally, I would like to thank whoever is currently reading this thesis. This thesis chronicles my personal development as I transitioned away from molecular biology and into computational biology and mathematical modeling. I hope that you will find it an interesting read, for it was certainly an interesting experience.

Contents

Chapter 1: Introduction

1.1 Malaria as a public health threat

Despite the historic availability of cheap, effective drugs and intense public health interventions, malaria still remains a global public health concern. Malaria is a mosquito-borne disease caused by single-celled, eukaryote parasites from the genus *Plasmodium*. These parasites have a complicated life cycle that alternates between a human host and a mosquito vector. Of the human malaria-causing species, *Plasmodium falciparum* is widely regarded as the deadliest.

As of 2015, the WHO reported 212 million new cases of malaria world wide and estimated a total of 429,000 malaria deaths [1]. The African continent bears the brunt of the global malaria burden, accounting for 90% of the cases and deaths reported in 2015. Despite these grim numbers, renewed interest in malaria elimination and eradication has resulted in drastic decreases in transmission. At the turn of the century, the Roll Back Malaria initiative was established with the goal of halving malaria deaths by 2010 and heralded the first major effort against malaria in four decades [2]. In 2007, the Bill and Melinda Gates Foundation challenged the malaria community to work towards complete malaria eradication. Today, these efforts are beginning to bear fruit as global malaria incidence rates decline. From 2000 to 2015, *P. falciparum* infection prevalence and incidence in Africa have fallen by 40% and have severely altered the transmission landscape [3,4]. In 2015, the Malaria Atlas Project estimated that 90% of the African population lived in either meso- or hypo-endemic regions, as compared to the 66% in 2000 [3]. Despite this success, maintaining the effectiveness of these interventions in the face of changing transmission dynamics will be significant challenges moving forward. The failure of previous public health interventions and the emergence of drug resistance are stark reminders of the challenges ahead.

1.2 Malaria population genetics: a look backwards

Malaria has had a surprisingly long and storied relationship with population genetics. In 1949, J. B. S. Haldane published a review where he summarized the (then) current knowledge regarding natural selection and infectious disease [5,6]. He hypothesized that pathogens were a driving force for diversification and that highly polymorphic traits, such as red blood cell polymorphisms, were an adaptive response to infectious diseases. Haldane is often credited as being the first to propose a link between thalassemia heterozygosity and malaria resistance, but offered no definitive proof at the time. In 1954, A.C. Allison provided conclusive evidence supporting Haldane's hypothesis when he discovered that sickle cell heterozygotes were protected against malaria [7]. Today, a wide variety of blood polymorphisms are known to confer some degree of malaria protection [8,9]. Malaria is often used as the textbook example of balancing selection and credited as one of the most significant drivers of human evolution [8].

Since the 1950s, malaria population genetics has accelerated with the availability of genetic data. Thousands of parasite genomes have been sequenced since the original *P. falciparum* reference genome assembly in 2002

[10]. Population genomic studies have revealed much about the demographic histories of natural parasite populations [11–13]. African parasite populations are highly admixed and have the highest levels of genetic diversity while Southeast Asia and South America are more structured and have lower levels of genetic diversity [12,13]. Haplotype analyses of the PfCRT mutations, which confers chloroquine resistance, show that resistance first arose independently in Southeast Asia and South America, after which it invaded the African continent [14]. This pattern of emergence and invasion has been repeated with many other drug resistance mutations [15–17], including the recently discovered kelch13 artmimisinin resistance mutations [18]. The increasingly large collections of parasite genomes spanning multiple populations and years are a goldmine for future population genomic studies. These sequences provide us with historical records of how parasite populations evolve over time. At the time of this writing, 2,512 whole genome sequences of field isolates collected from global parasite populations are publicly available through the Pf3k Project, an international collaboration whose goal is to provide a high-resolution view of natural variation in *P. falciparum* (https://www.malariagen.net/projects/pf3k).

At a more practical level, population genomic analyses are also beginning to be used to monitor transmission and evaluate public health interventions. Genomics has the potential to establish directionality in parasite movement, identify source-sink populations, and identify drug resistance mutations before they threaten the efficacy of current malaria therapies [18–22]. As transmission intensities decline and traditional metrics of transmission intensity become more difficult to collect [20,23], genetic metrics could provide an easier method of monitoring transmission intensity [24–26]. These metrics include (COI, number of strains per infection), the frequency of polygenomic infections, and the incidence of parasite clonality. Whether genomics will be useful for future public health interventions will depend on how easily genomic data can be integrated into existing epidemiological techniques and frameworks.

Today, the challenge lies in interpretation and it is here where mathematical models can be useful. However, neither traditional population genetic models nor epidemiology models are sufficient for modeling malaria transmission dynamics in relation to population genetics. In fact, it is with great irony that fields historically awash in theory are now insufficient to explain all the patterns observed in genomic data. Epidemiology models excel in simulating complex transmission structures, but are generally strain agnostic and do not incorporate differences in strain biology. Conversely, traditional population genetic models fail to account for the complexities of the malaria life cycle. Future models will need to incorporate techniques from both fields to accurately predict changes in malaria population genetics and characterize its relationship with changing transmission conditions.

1.3 The problem of the sexual recombination and coinfection

Unlike most pathogens, malaria must sexually reproduce in a mosquito vector during each transmission event. When a mosquito feeds on an infected human host, she ingests haploid gametocytes which differentiate into male and female gametes. Male and female gametes fuse in the mosquito midgut to create a diploid zygote that develops into a motile ookinete that traverses the midgut wall and creates a sack-like structure known as the oocyst. Within the oocyst, the parasite undergoes meiotic division and mitotic amplification, resulting in the generation of thousands of haploid sporozoites. These sporozoites travel to the mosquito salivary glands and are deposited in a new human host during the next mosquito blood meal (**Figure 1.1**).

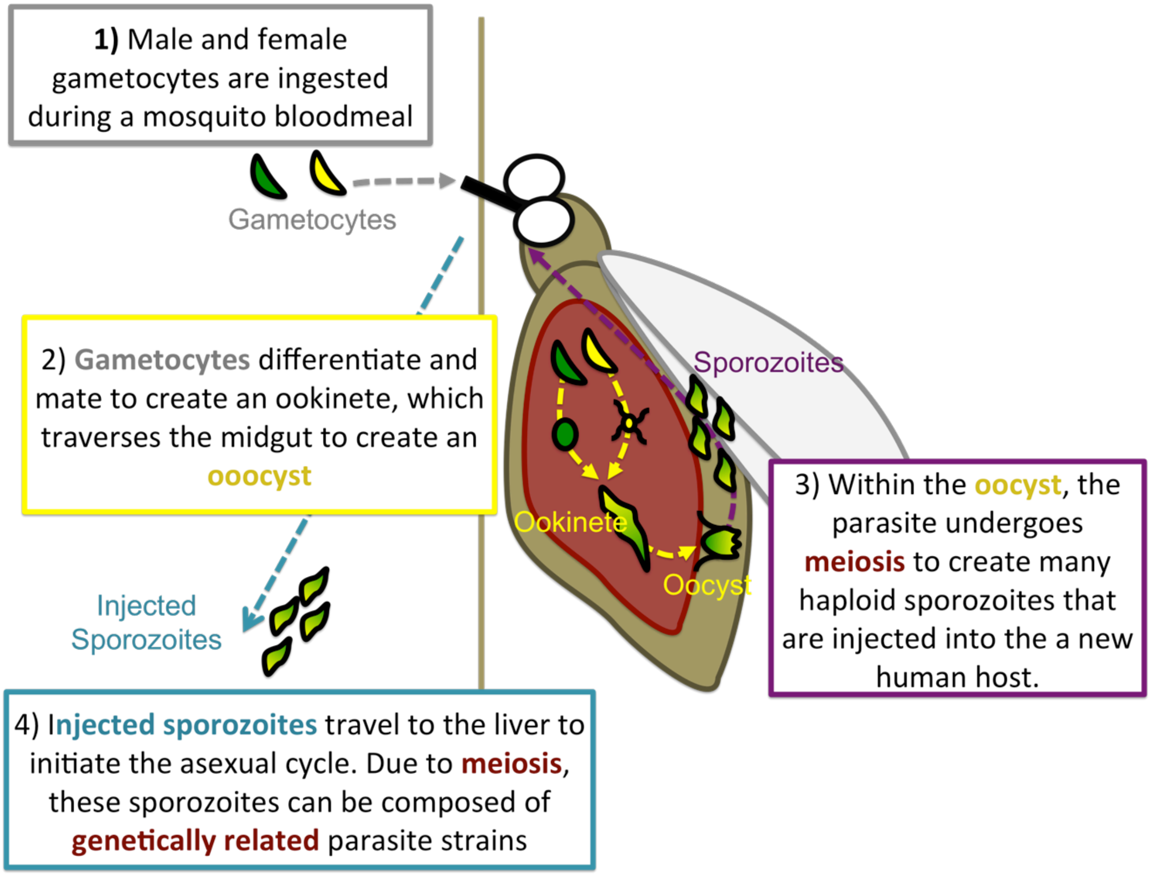

1) Male and female gametocytes are ingested during a mosquito bloodmeal
Gametocytes
2) Gametocytes differentiate and mate to create an ookinete, which traverses the midgut to create an ooocyst
Sporozoites
Ookinete
Oocyst
3) Within the oocyst, the parasite undergoes meiosis to create many haploid sporozoites that are injected into the a new human host.
Injected Sporozoites
4) Injected sporozoites travel to the liver to initiate the asexual cycle. Due to meiosis, these sporozoites can be composed of genetically related parasite strains

Figure 1.1 Sexual Phase of the malaria life cycle

The sexual phase of the malaria life cycle has surprisingly deep consequences for parasite genetics. Transmission provides the parasite an opportunity for genetic exchange through sexual recombination, which occurs during meiosis and facilitates chromosomal crossover. This process allows genomic regions to be swapped and increases genetic variation in the population (**Figure 1.2**). Recombination breaks existing genetic associations and reduces linkage disequilibrium, the non-random association of alleles at different sites in the population. For a purely outcrossing population (mating between unrelated individuals), the rate with which recombination reduces linkage disequilibrium depends on the genetic distance between sites. Although recombination occurs during every transmission event, not all events will result in observable genetic exchange. Effective recombination only occurs between two genetically distinct genomic sequences; selfing (mating between genetically identical individuals) does not reduce linkage disequilibrium because the exchanged genetic sequences are identical. Recombination is a powerful force for diversification and theorized to help diploid organisms purge deleterious mutations. For malaria, recombination and sexual reproduction allows cotransmitted parasite (those from the same mosquito vector) to be genetic siblings and share genetic relatedness.

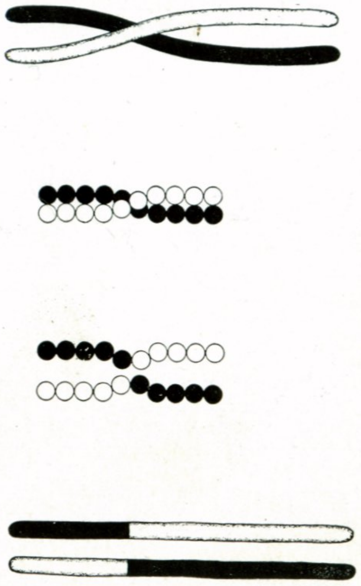


FIG. 64. Scheme to illustrate a method of crossing over of the chromosomes.

Figure 1.2 Thomas Hunt Morgan’s 1916 illustration of crossing over

Why is recombination so problematic? From a population genetic standpoint, recombination allows different regions of the genome to have different evolutionary histories. This limits the usefulness of coalescent-based phylodynamic analyses of transmission and population history reconstruction [27]. Phylodynamics methods use genomic sequences to reconstruct transmission trees and are particularly popular for tracing the origin and evolution of viral outbreaks [28–31]. However, these methods struggle with incorporating recombination, choosing either to ignore it or to partition the genome into blocks that allow recombination between but not within blocks [32,33]. Advances in coalescent methods have improved our ability to accommodate recombination and identify recombination hotspots, but recombination remains a significant computational and theoretical challenge [34]. Coalescent or phylogenetic approaches for interrogating malaria population genetics are not widely used.

The problem of recombination is even more complicated in malaria because its effective recombination rate depends on epidemiological conditions such as transmission intensity. In malaria, outcrossing and effective recombination can only occur when a mosquito feeds on a polygenomic (multiple-strain) infection. Mosquitoes feeding on monogenomic (single-strain) infections force the parasite to self, resulting in no changes in linkage disequilibrium. Parasite populations in high transmission settings have higher outcrossing rates and less linkage disequilibrium than those in low transmission settings [12,35]. This is because individuals in high transmission areas are more likely to be coinfected with multiple strains introduced by superinfection, the repeated infection of the individuals from independent transmission events [26]. Superinfection in high transmission areas forms the rationale for using complexity of infection as a genetic proxy for transmission intensity. Coinfection and sexual reproduction are major aspects of malaria biology and important for our understanding of malaria transmission and population genetics.

Coinfection is also a major problem and can have one of two evolutionary relevant consequences. As described above, coinfection allows genetic exchange between coinfecting pathogens. It also introduces a new level of competition at the intra-host level, and accurately modeling these dynamics in the context of host immunity remains a significant challenge. Coinfection has been extensively modeled in virulence evolution studies, but the problems associated with it are reflected in the dizzying array of frameworks used to accommodate (or fail to accommodate) it [36,37] (**Figure 1.3**). Polygenomic infections are generally assumed to be comprised of unrelated strains that are randomly sampled from the greater population. Whether such an assumption is applicable to malaria, which undergoes sexual reproduction during every transmission event, is a major focus of this thesis.

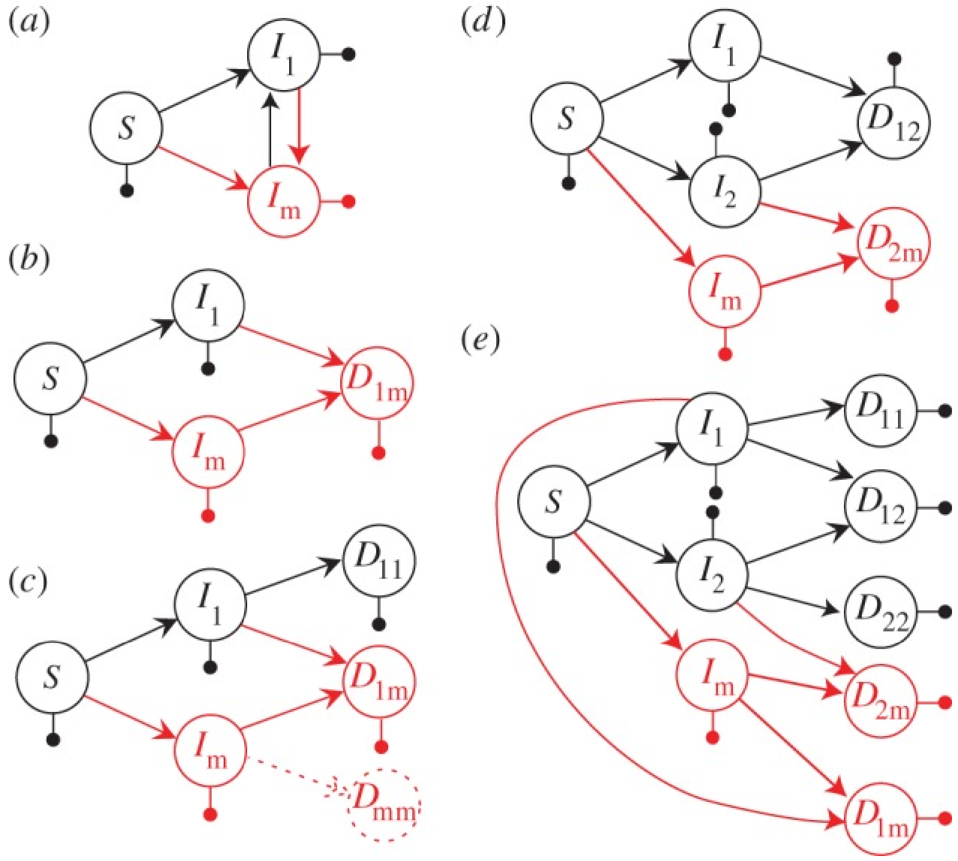


Figure 1.3 Epidemiological models with coinfection

Each model starts off with a susceptible individual (S) that transitions to an infection class with a single parasite type (I). These infections can be superinfected and transition to an infection class with multiple parasite types (D). Numbered subscripts are used to differentiate parasite species. “m” is used to

(*Figure 1.3, continued*) denote mutant strains of the same parasite species. a) A superinfection model that does not allow coinfection. Superinfection immediately supplants the resident strain. b) A coinfection model that does not allow repeat infection of concurrent strains. c) A coinfection model that allows repeat infection of concurrent strains. d) A coinfection model with different species e) a coinfection model with parasites from the same species in a dimorphic population. For d), numbered subscripts differentiate strains from the same species but different populations. Figure taken from [37].

1.4 Thesis structure

In this thesis, I focus on the nature of coinfection and polygenomic infections in *P. falciparum* malaria. Throughout this thesis, I use “coinfection” to refer to the simultaneous infection of two or more *P. falciparum* strains, not the simultaneous infection of two or more species. To emphasize this distinction, I use the term “polygenomic infection” to refer to multiple strain infections. In malaria, coinfection alters pathogen dynamics and has one of two evolutionary relevant consequences: 1) it provides an opportunity for effective recombination and 2) introduces a new level of competition at the within-host level. This thesis focuses on the first consequence and its implications for population genomics and future genetic epidemiology models.

The chapters in this thesis are self-sufficient and presented in the order with which they were performed. In chapter 2, I quantified the genetic relatedness of coinfecting strains in polygenomic infections collected from Thiès, Senegal. I showed that the coinfecting strains in polygenomic infections are highly related and incompatible with the expectations of pure superinfection. These results suggest that cotransmission is common in natural populations.

In chapter 3, I used a mathematical model to quantify the expected relatedness of cotransmitted strains. I demonstrate that there are only 9 different ways that cotransmitted parasites can be related to one another. I show that the relatedness of polygenomic infections depends on the conditions of the initial infection and that different transmission lineages have different expectations of polygenomic relatedness.

In chapter 4, I analyzed the sequencing quality of lab-generated mock infections to determine whether selective whole genome amplification could be used to accurately sequence polygenomic infections. I found that selective whole genome amplification could be used to characterize the genomic composition of polygenomic infections, even when there is a significant amount of contaminating host DNA present.

In chapter 5, I interrogate how coinfection and transmission topology affects malaria population genetics and evolution by performing evolutionary invasion analyses. This work borrows heavily from theoretical evolutionary population genetics and is designed to show how modeling can be used to highlight importance features of malaria transmission.

The use of population genomics for understanding parasite transmission and evolution hinges on our ability to integrate population genetics into existing epidemiological frameworks. The integration of these fields will require advances in data analysis, sequencing generation, and theory development. This thesis touches on all three aspects in order to understand the genetic consequences of coinfection and cotransmission.

Chapter 2: Genetic relatedness analysis reveals the cotransmission of genetically related Plasmodium falciparum parasites in Thies, Senegal[1]

Wesley Wong[a], Allison D. Griggs [b], Rachel F. Daniels[a,b], Stephen F. Schaffner[b], Daouda Ndiaye[c], Amy K. Bei[a, c], Awa B. Deme[c], Bronwyn MacInnis[b], Sarah K. Volkman[a,b,d], Daniel L. Hartl[a,e], Daniel E. Neafsey[b], and Dyann F. Wirth[a,b*]

*Corresponding author

Affiliations:

[a] Department of Immunology and Infectious Diseases, Harvard T. H. Chan School of Public Health, Boston, Massachusetts, 02115, U.S.A.

[b] Broad Institute, Cambridge, Massachusetts, 02142, U.S.A.

[c] Faculty of Medicine and Pharmacy, Cheikh Anta Diop University, Dakar, Senegal.

[d] School of Nursing and Health Sciences, Simmons College, Boston, Massachusetts, 02115, U.S.A.

[e] Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, Massachusetts, 02138, U.S.A.

[1] This chapter was originally published as
Wong W, Griggs AD, Daniels RF, et al. Genetic relatedness analysis reveals the cotransmission of genetically related *Plasmodium falciparum* parasites in Thiès, Senegal. *Genome Medicine*. 2017;9:5. doi:10.1186/s13073-017-0398-0.

2.1 Abstract

As public health interventions drive parasite populations to elimination, genetic epidemiology models that incorporate population genomics can be powerful tools for evaluating the effectiveness of continued intervention. However, current genetic epidemiology models may not accurately simulate the population genetic profile of parasite populations, particularly with regard to polygenomic (multi-strain) infections. Current epidemiology models simulate polygenomic infections via superinfection (multiple mosquito bites) despite growing evidence that cotransmission (single mosquito bite) may contribute to polygenomic infections. Here, we quantified the relatedness of strains within 31 polygenomic infections collected from patients in Thiès, Senegal using a Hidden Markov model to measure the proportion of the genome that is inferred to be identical by descent. We found that polygenomic infections can be composed of highly related parasites and that superinfection models drastically underestimate the relatedness of strains within polygenomic infections. Our findings suggest that cotransmission is a major contributor to polygenomic infections in Thiès, Senegal. The incorporation of cotransmission into existing genetic epidemiology models may enhance our ability to characterize and predict changes in population structure associated with reduced transmission intensities and the emergence of important phenotypes like drug resistance that threaten to undermine malaria elimination activities.

2.2 Background

The recent push for malaria eradication highlights a growing need to accurately monitor changes in malaria transmission and assess the impact of interventions. Population genomic analyses and genetic epidemiology models can be powerful tools for monitoring declining transmission rates and evaluating the efficacy of public health interventions. Metrics of population genetic structure have been used to characterize parasite populations in low transmission regions [24,25,38,39] and, in combination with epidemiological modeling, to monitor changes in transmission rate [40].

Previous studies have largely relied on the sequences obtained from monogenomic (single-strain) infections, which may not provide an accurate representation of the genetic structure within the population. Polygenomic (multiple-genome) infections exhibit reduced genetic diversity relative to the total genetic diversity of all strains in the local population [13] and are known to be composed of genetically similar parasite strains [41–44], regardless of the genetic markers used. Understanding how polygenomic infections are formed, and incorporating the consequences of these infections on transmission patterns into genetic epidemiology models would help improve monitoring and evaluating systems within malaria elimination programs.

Historically, the formation of polygenomic infections has been assumed to be a function of the entomological inoculation rate (EIR), or the number of infectious bites per human per day [45] because multiple mosquito bites greatly enhance the probability of independent infections within a single human host from multiple mosquitoes (superinfection). Current epidemiology models largely operate under the assumptions of superinfection [36,46,47], which has been supported by the increased incidence of polygenomic infections in high transmission areas [48]. In high transmission areas, patients are exposed to numerous infectious bites, thus raising the chance of superinfection and the creation of new polygenomic infections. Under superinfection, strains within polygenomic infections are randomly and independently sampled from the local population.

The assumption of superinfection in epidemiology models is at odds with the observed similarity of strains within polygenomic infections [41–44] because superinfection cannot easily account for the high degree of similarity between strains within polygenomic infections. Relatedness among genomes in polygenomic infections is commonly attributed to cotransmission, or the simultaneous transfer of multiple, distinct parasite genomes from a single mosquito bite. Because the parasite undergoes sexual reproduction within the mosquito vector, cotransmitted parasites are expected to be genetically related to one another [42]. After a single cotransmission event, cotransmitted infections may be composed of F_1 hybrids as well as unrecombined parental genomes. Subsequent cotransmission events (serial cotransmission) may result in high degrees of relatedness within polygenomic infections. Serial cotransmission chains constrain parasites to mating with their relatives, resulting in a steady increase in the average relatedness between cotransmitted strains. Extremely high degrees of genetic relatedness have been proposed to be signatures of serial cotransmission events that could be used to identify infections due to serial cotransmission [42] .

Determining whether current epidemiological models can realistically simulate the relatedness within polygenomic infections is of key public health interest when these models use population genomics to monitor declining transmission rates. Here, we quantified the genetic relatedness of genomes within individual polygenomic infections using a Hidden Markov Model (HMM) to measure the proportion of the genome that is inferred to be identical by descent (IBD). Our HMM allows us to distinguish regions of the genome that are more likely to be identical due to random chance and population structure from regions of the genome that are more likely to be identical due to shared inheritance. These IBD estimates were compared to the relatedness expected with superinfection, which was simulated as the random sampling of parasites from Thiès, Senegal, which was represented by 146 monogenomic infections previously collected from the region.

Our polygenomic infections comprised of 31 infections collected from patients in Thiès, Senegal in the years 2011–2013. Thiès lies 70km away from the capital city of Dakar, a hypoendemic region with an EIR < 5 [49]. In 2005, Senegal implemented a redesigned National Malaria Control Programme (NMCP) aimed at improving insecticide-treated mosquito net coverage, indoor residual spraying coverage, preventative treatment coverage for pregnant women and children under five, and antimalarial treatment coverage. Since then, there has been a significant decrease in the number of confirmed cases, going from 1,555,000 cases in 2006 to 174,000 cases in 2009 [50]. As of 2009, the prevalence in Thiès was ~3%[50] and has since fallen further.

Our findings indicate that cotransmission is common in Thiès, Senegal, and that genetic epidemiology models can be made to more accurately reflect relatedness within polygenomic infections by incorporating cotransmission. These findings have important implications for the application and use of genetic tools to understand malaria transmission dynamics, to assess the impact of malaria elimination interventions, and to study the consequences of these interventions on potentially undermining traits such as drug resistance emergence.

2.3 Methods

2.3.1 Sample and Sequence Collection

All patient samples were collected at clinics located in three different areas of Senegal: Thiès, Pikine, and Velingara. These samples were collected between approximately September and December each year, which roughly corresponds to the period just following the rainy season in Senegal. Participants reporting acute fevers and suspected of being infected with malaria (e.g., mild uncomplicated malaria infection) with no reported history of antimalarial therapy were considered for inclusion in our study. Participants were diagnosed for malaria based on microscopy and rapid diagnostic tests. Samples were anonymous and coded as to Country (Senegal or Sen); collection village (T =

Thies, P= Pikine, V = Velingara); sample number collected from the clinic (001 to 999); and identified by year (e.g., 2011 or 11) to create a sample number of 'SenT009.11', which was collected from Thies, Senegal in the year 2011 and represents the ninth sample (009) collected that year.

We sequenced 190 *P. falciparum* genomes from patient-derived material collected from Senegal, of which 176 were collected from Thiès, 4 from Velingara, and 10 from Pikine. These samples were initially identified as monogenomic using a 24-SNP molecular barcode [51]. Barcodes were genotyped using an high-resolution melting (HRM)-based assay [39,51]. The parasite strains were culture adapted at the Harvard T.H. Chan School of Public Health and sequenced at the Broad Institute using Illumina Hi-Seq (Illumina, Inc., San Diego, CA) machines.

We also sequenced a set of 111 samples collected exclusively from Thiès, Senegal during the years 2011-2013. Unlike our previously mentioned samples, genomic DNA was extracted directly from patient samples to avoid strain ascertainment bias and the potential loss of low frequency strains. Genomic DNA was extracted using a QiAmp DNA Blood Mini kit (Qiagen, Valencia, CA) according to manufacturer's specifications. These samples were sequenced at the Broad institute using Illumina Hi-Seq machines.

Sequencing reads were aligned using the Burrows-Wheeler Aligner (version 0.5.9-r16) [52] against the 3D7 reference assembly (PlasmoDBv7.1)[53] to create BAM files. Variant calls and consensus sequences for each sample was determined using GATK Unified Genotyper [54]. A full list of the individual parameter and quality-score thresholds can be found in the supplementary information of [40].

2.3.2 Defining our monogenomic infection dataset

To determine the expected relatedness of superinfection, we needed to identify a set of monogenomic infections to represent the parasites present in Thiès, Senegal. To do this, we relied on a set of 190 samples that were previously sequenced and identified as monogenomic using a 24-SNP barcode. For this study, we decided to use stricter criteria to identify monogenomic samples. Within each of the 190 sequences classified as monogenomic by barcode, all sites with a non-unanimous read pileup were first identified, resulting in 1.1 million variant positions. These positions were then filtered to have a read depth of at least 10 across 90% of the samples, to be strictly biallelic, and to be found in at least 2 of the 190 samples. A preliminary set of 440,000 SNPs passed these criteria, which were then used to reclassify each of the 190 putatively monogenomic samples. Monogenomic samples were reclassified by calculating the proportion of the 440,000 sites with a unanimous read support within each of the 190 samples. Those samples where the proportion was 80% or higher were considered monogenomic, which identifed 146 monogenomic samples, all of which originated from Thiès. The read pileups of these samples over the preliminary set of 440,000 SNPs have less than 0.0005% non-unanimous reads **(Supplemental Figure S2.1)**.

Because our set of 440,000 SNPs was derived using information from all 190 samples, which could represent a mix of monogenomic and potentially cryptic polygenomic samples, we chose a more stringent set of SNPs based solely on the information drawn from monogenomic samples. 56 of the 146 monogenomic samples were randomly chosen to further filter our set of 440,000 preliminary SNPs. Sites where the read-pileup across all 56 samples was less than or equal to 0.01%, or that lacked reads in more than 1 of the 56 samples, were also removed. After applying these filters, we identified a set of 3132 SNP positions that were used to analyze the genetic relatedness within polygenomic infections.

2.3.3 Defining our final polygenomic infection dataset

These 3132 SNPs were then used to identify polygenomic infections from the set of 111 samples collected from Senegal during the years 2011-2013. Samples where less than 30% of the 3132 SNPs had at least one read were excluded from our analysis, leaving us with 31 polygenomic infections. For each of the remaining samples, we removed sites that were supported by a single read. All samples in which at least 95% of the remaining sites were completely unanimous were classified as monogenomic, while any sample with a proportion less than 95% was classified as polygenomic.

2.3.4 Estimating relatedness using a Hidden Markov Model

For each sample, we calculated relatedness between sample pairs by first identifying regions of the genome that are inferred to be IBD based on the likelihood of observing identity due to random chance using a Hidden Markov Model [40]. The model has two hidden states: IBD, inherited from the same ancestor, or Different-By-Descent (DBD), inherited from different ancestors. Sequence pairs are reduced to a series of discordant and concordant calls, depending on the observations made at each SNP site. Sites where both sequences have the same allele are considered concordant while sites where each sequence has a different allele are considered discordant. It then calculates the probability of observing concordant or discordant genotypes under the assumption of IBD or DBD by using the population allele frequencies at that site, the error rate, and the probability of transitioning from one hidden state to the other. The probability of transitioning from IBD to DBD between two SNPs is proportional to the physical distance between them and influenced by the overall recombination rate. The HMM then uses a Viterbi algorithm to identify the most probable path of hidden states. An overall estimate of relatedness for each comparison was obtained by summing the total proportion of the optimum path that is in IBD.

Delete-a-group jackknife analysis was performed to obtain jackknife estimates of the mean and jackknife estimates of the standard error of the mean. Groups were defined by dividing the genome into 10 mutually exclusive groups by scanning across the genome and placing the *i*th SNP into the *i*th group. After all 10 groups have at least one SNP, the process is repeated, placing the *i*+10th

SNP into the *i*th group, and continuing until the end of the genome. This effectively randomizes the SNPs in each group and ensures that the number of SNPs and distribution of SNP locations within each group is evenly distributed.

2.3.5 Generating artificial mixed genome samples

Genomic DNA mixtures were generated by mixing DNA obtained from five distinct culture-adapted parasite strains (SenT148.09, SenT111.09, SenT165.09, SenT033.09, and SenT015.09) in proportions described in **Supplemental Table S2.1.** Genomic DNA was extracted from adapted parasite cultures using a QiAmp DNA Blood Mini kit (Qiagen, Valencia, CA) according to manufacturer specifications. DNA concentrations were determined by a NanoDrop Spectrophotometer (Thermo Fisher Scientific) and a barcode-based quantification assay[51]. Each mixture had a total DNA concentration of 5ng/ul.

2.3.6 Constructing pseudohaplotypes

Pseudohaplotypes were constructed by examining the read-pileups at each of the available 3132 SNPs for each polygenomic infection. Sites were categorized into heterozygous sites, a site where at least one read had an alternate allele, and homozygous sites, a site where all the reads had the same allele. Pseudohaplotypes were constructed by randomly assigning the allelic states of each site to one of two constructed haplotypes. For homozygous sites, both haplotypes received the same allelic state. For heterozygous sites, one haplotype received the major allelic state (the allele with the greater read support) while the other haplotype received the minor allelic state (the allele with the lower read support). These pseudohaplotypes preserve the physical order and distance between each of the available 3132 trusted SNPs and the order of concordant and discordant calls, but do not establish true linkage-phase.

2.3.7 Generating subsets to test the limitations of the HMM

Subsets were generated by randomly choosing without replacement from the 3132 SNPs. The largest of these subsets contained 90% of the 3132 SNPs while the smallest contained 10% of them. Each subset was repeated 40 times to obtain estimates of the mean and standard deviation.

2.3.8 Calculating concordance

For each pairwise comparison, concordance was calculated as the number of sites with the same allelic identity divided by the number of sites examined. Due to the presence of missing data, the number of sites examined fluctuated. If a site was missing in one or both of the strains being compared, then the site was excluded from the analysis. In addition, sites where only the major allele was present were also excluded.

2.3.9 Simulating expected relatedness under superinfection

Superinfection was simulated as a random sampling of parasites collected throughout Thiès, Senegal. We assumed that the parasite population in this region was completely mixed, with no heterogeneity in population structure or transmission intensity. The expected relatedness under the superinfection hypothesis was calculated by quantifying the relatedness between our set of 146 monogenomic infections.

To make the data from our simulation more comparable with the data obtained from our polygenomic infections, we generated a series of bootstrap resampled distributions of the mean relatedness. Simple random sampling bootstrap distributions were generated by randomly sampling 40,000 sets of 31 monogenomic pairs and calculating the average relatedness among these sample pairs. To create weighted bootstrap distributions, we extracted the barcode sequence from each of the monogenomic infection whole genome sequences and identified it with one of the barcode sequences within our 24-SNP barcode dataset. The identities of at least 22 of the 24 barcode positions needed to be identical to be considered the same sequence. The observed frequency of each 24-SNP barcode was used to infer the population frequency of the parasite strain within each monogenomic infection. A weighted bootstrap distribution of mean relatedness was created by calculating the randomly sampling 40,000 sets of 31 monogenomic infection pairs, where each pair was weighted according to the probability of drawing that particular sample pair.

P-values for each bootstrap distribution was calculated by counting the number of times our sample mean was greater than or equal to the observed mean relatedness in our 31 polygenomic infections (relatedness = 0.38).

2.3.10 Identifying monogenomic infections that were related to polygenomic infections

For each polygenomic infection, we used the HMM to compare the observed within-polygenomic infection IBD segments with the corresponding genomic regions in each of the 146 monogenomic samples. Related monogenomic infections were identified as those that contributed a significant fraction of the polygenomic infection's IBD segments.

2.4 Results

2.4.1 Relatedness within polygenomic infections

To quantify the relatedness of strains within each infection, we identified a set of 3132 SNPs that had passed a set of read-mapping filters designed to remove variant positions liable to yield erroneous heterozygous signals due to read mapping and/or base calling errors. These trusted SNPs form a sensitive panel for detecting heterozygous positions within polygenomic samples, and can be used to mark IBD segment boundaries **(Figure 2.1)**. The majority of our SNPs fall within coding regions (77% coding, 23% noncoding). The proportion of reads supporting the major allele at each of these sites reflect the expected ratio of individual strains in sets of mixtures created from genomic DNA to control for both genome diversity and relative proportions **(Supplemental Figure S2.2**).

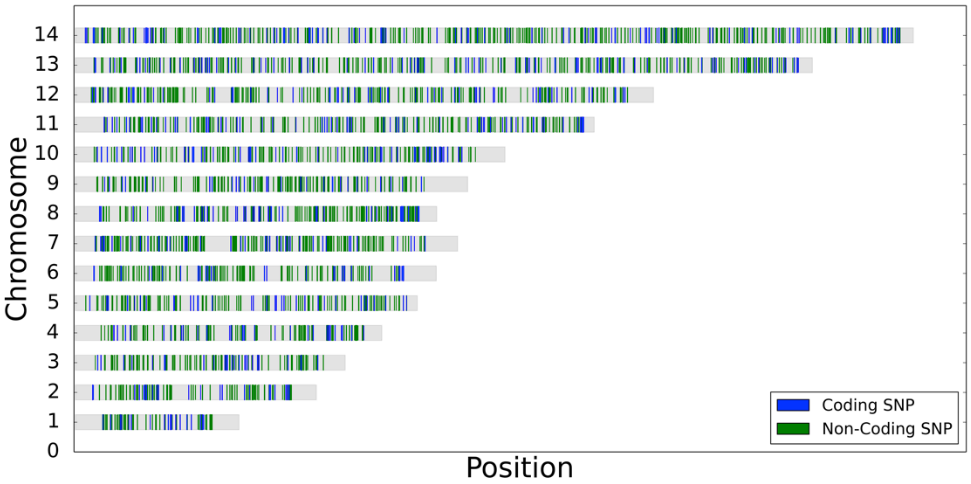


Figure 2.1. Trusted SNP set marker map.

A representation of the *P. falciparum* genome and the location of each of the 3132 trusted SNPs. Grey bars represent individual chromosomes. Blue lines indicate the location of coding SNPs and green lines represent the location of non-coding SNPs.

We sequenced 111 polygenomic infections collected from patients in Senegal arriving at clinic for treatment for mild uncomplicated malaria infection during the years 2011–2013. Each sample had an average of 58 million reads, but because genomic DNA was extracted directly from patient material and not depleted of host material before sequencing, only 1% of them aligned to the *P. falciparum* genome. As a result, some of the polygenomic infections lacked coverage at all the trusted SNP locations. Samples where > 30% of the trusted

SNP sites lacked sequencing reads were excluded from our analysis, leaving us with a total of 31 polygenomic infections. For each of the remaining polygenomic infections, we excluded sites with < 1 read from our analysis. After excluding these sites, we found that the range of usable sites per sample spanned from 300 to 3132 SNPs. Samples collected from 2011 had the highest mean number of usable sites (3113 sites) while samples collected in 2012 and 2013 had a lower mean number of usable sites (865 and 1172 sites, respectively) (**Supplemental Figure S2.3**). At sites where there were at least two reads, we found that the average read depth in our samples was 7.68; read depth in samples collected from 2011 was higher (12.74) and those collected from 2012 and 2013 had a lower read depth (3.08 and 3.62, respectively).

To quantify the relatedness, or proportion of the genome that is identical by descent (IBD), within each polygenomic infection, we used a Hidden Markov Model (HMM) that was previously used to quantify the relatedness of genomes present in monogenomic infections collected in Senegal [40]. Because our HMM examines sequence pairs as a series of discordant and concordant calls, we constructed two pseudohaplotypes that preserve the order and position of discordant and concordant calls to represent the genetic similarity of genomes within each infection. We use the term pseudohaplotype because the inferred haplotype does not necessarily establish the true linkage-phase of haplotypes within polygenomic infections. These pseudohaplotypes are actually conservative representations of genetic similarity because they underestimate the true similarity between genomes when the polygenomic infection is composed of more than two strains. During the sampling timeframe and setting in Thiès, Senegal, the average complexity of infection (COI) in polygenomic infections is two[55], and the pseudohaplotypes reflect the genetic similarity of the genomes.

We first ran tests to determine if the variation in number of assayable SNPs would affect our estimates of relatedness. We calculated the relatedness between 27 monogenomic sample pairs using different numbers of SNPs taken from the complete set of 3132 SNPs. We found that the HMM is robust to differences in SNP number and that estimates of relatedness based on as few as 313 SNPs will consistently provide the same estimate as those based either on 3132 SNPs or an ever larger set of 14,972 SNPs with a minor allele frequency of ≥ 0.05 among the samples from Senegal **(Supplemental Figure S2.4 & S2.5**).

We found that the estimated genetic relatedness within the 31 polygenomic infections are evenly distributed, ranging from completely unrelated (relatedness = 0.0) to highly related (relatedness = 0.90) **(Figure 2.2, Supplemental Figure S2.6**). Across all years, we found that the average relatedness within a polygenomic infection was 0.38. To examine the distribution of IBD block sizes within each infection, we mapped each IBD block to its corresponding location in the *P. falciparum* genome (**Figure 2.3**). There was a trend in genetic relatedness and IBD block size. Across all samples, the average IBD block size within the 31 polygenomic infections was 0.92 Mbp. After dividing infections into highly related infections, which was defined as having a relatedness of ≥ 0.30 (a value exceeding that expected of half- siblings, 0.25, but allowing for some uncertainty in the accuracy of our HMM) and less related infections (relatedness < 0.30), we found that the average IBD block size among highly related infections was significantly longer (*p*-value = 2.70 x 10^{-8}, Mann-Whitney U). IBD blocks among highly related parasites (average IBD block size = 1.05 Mbp) were on average 0.73 Mbp longer than the block sizes across less related parasites (average IBD blocksize = 0.32 Mbp) **(Figure 2.4)**.

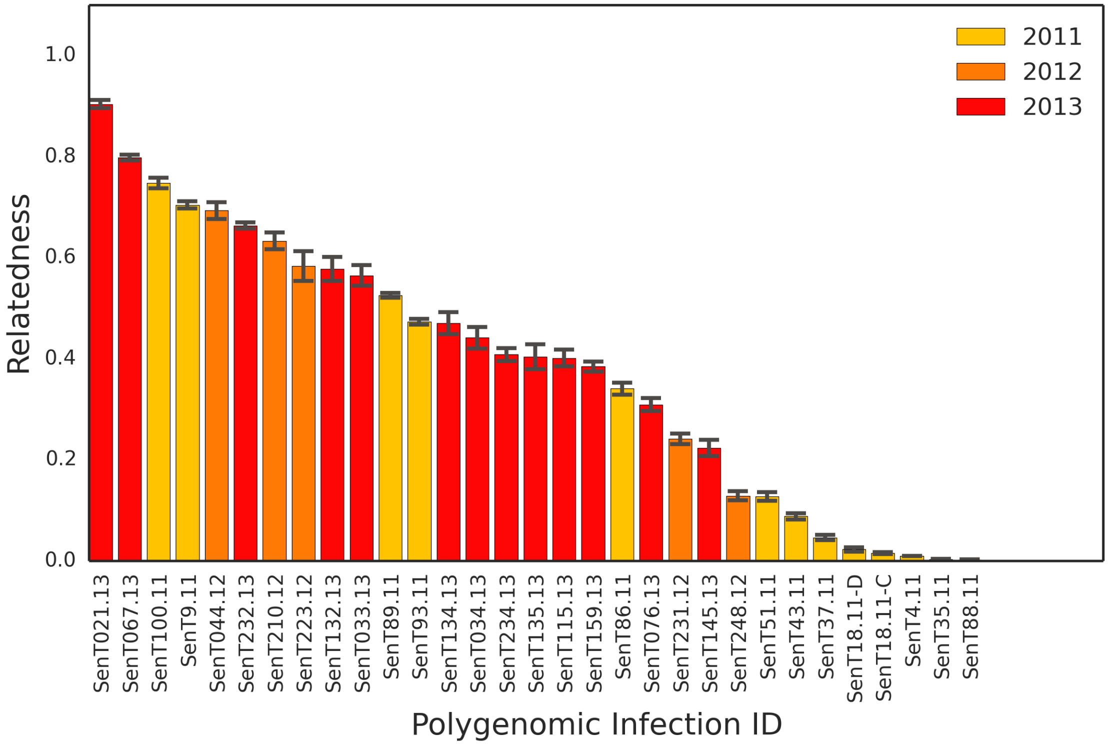


Figure 2.2 Relatedness within polygenomic infections.

Barplots of jackknife estimates of the mean relatedness within 31 polygenomic infections collected from Senegal from 2011-2013. Error bars represent one jackknife estimate of the standard error of the mean. Relatedness is defined as the proportion genome shared IBD between the strains comprising each polygenomic infection. While there is no clustering of relatedness by year, samples collected in 2011 are less related (average relatedness = 0.24) than

(Figure 2.2, continued) samples collected in 2012 and 2013 (average relatedness = 0.46 and 0.50, respectively) (*p*-value = 0.048, 1-way ANOVA). Samples collected from 2012 and (*Figure 2.2, continued*) 2013 had lower coverage than those in 2011, which may contribute to their higher relatedness values.

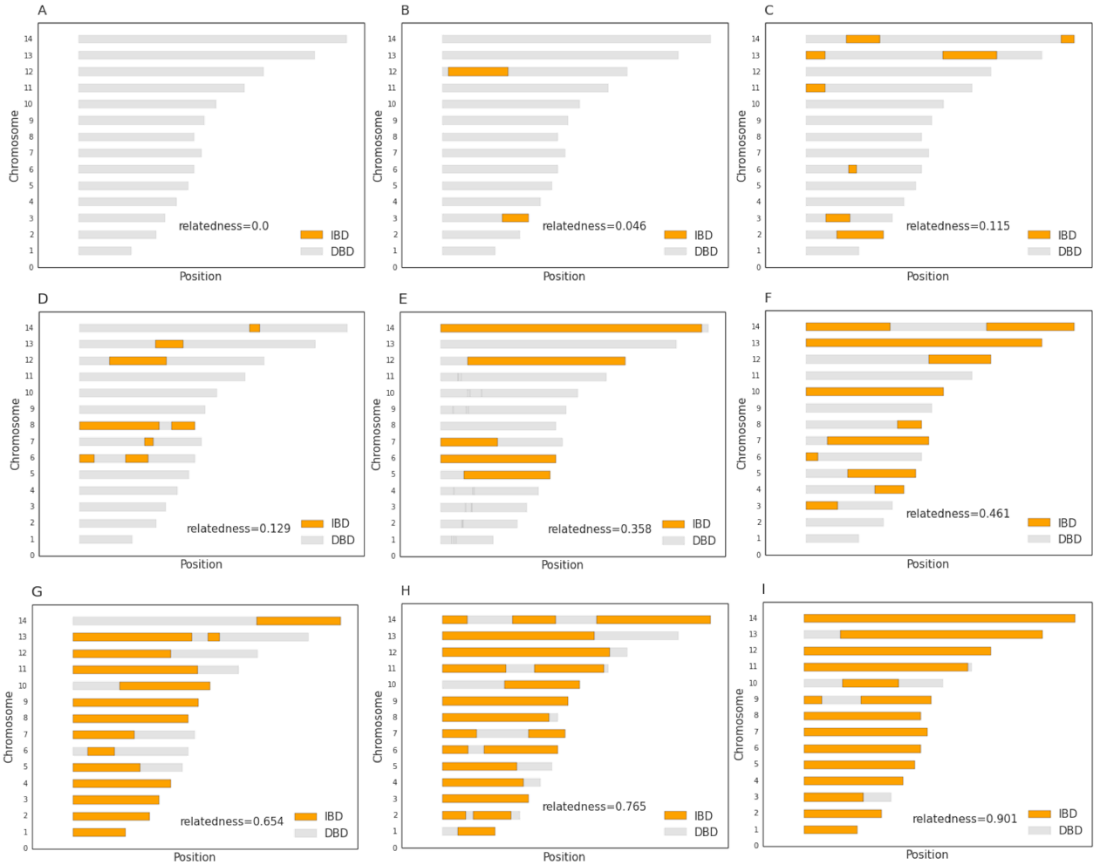


Figure 2.3 Polygenomic infection IBD maps.

Representative IBD maps of nine different polygenomic infections. Grey bars represent sections of the genome that are not IBD among the strains present within the polygenomic infections. Orange sections represent regions of the genome that are IBD. A=SenT88.11, B=SenT37.11, C=SenT51.11, D=SenT248.12, E=SenT223.12, F=SenT093.11, G=SenT232.13, H=SenT100.11, I=SenT021.13

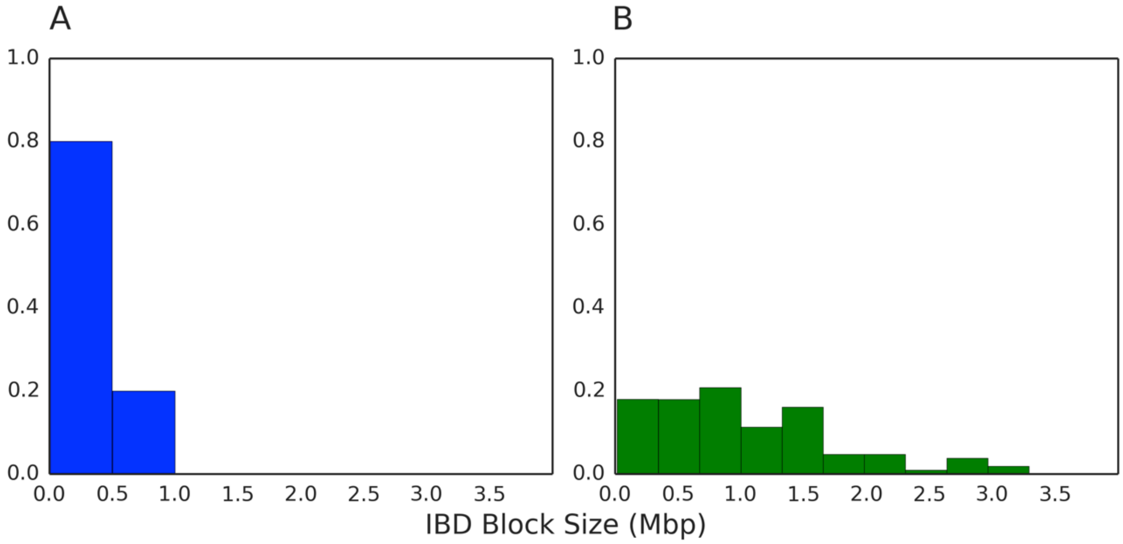


Figure 2.4 IBD block distributions within polygenomic infections

Distribution of IBD block sizes in megabase pairs (Mbp). IBD blocks were defined as contiguous segments of the genome that are IBD and are longer in highly related polygenomic infections (*p*-value = 2.70 x 10^{-8}, Mann-Whitney U). **A)** Distribution of IBD block size in less related polygenomic infections (relatedness < 0.30). Average block size is 0.31 Mbp with a standard deviation of 0.21 Mbp. **B)** Distribution of IBD block sizes in highly related polygenomic infections (relatedness > 0.30). Average block 1.04 Mbp with a standard deviation of 0.73 Mbp.

We also found that some of these polygenomic infections were related to parasite strains independently sampled from within the local population. We used the within-polygenomic IBD segment boundaries to generate IBD maps between the strains within polygenomic infections to the strains from monogenomic infections **(Figure 2.5)**. IBD segments create localized regions of the genome where the phase is known, allowing us to compare the strains from polygenomic infections to strains from the local population. For each of the polygenomic samples, we determined whether there were monogenomic samples sharing IBD segments with those within polygenomic infections and identified monogenomic samples that shared a large fraction of IBD with the within-polygenomic IBD segments.

For one polygenomic infection collected in 2011, SenT009.11, we identified two related strains, both of which were collected in the previous year (2010) among monogenomic infections. In the case of SenT009.11, the monogenomic samples SenT076.10 and SenT104.10 collectively shared IBD with 71% of the within-polygenomic IBD segments, contributing 33% and 36% of shared IBD, respectively. In this case, SenT076.10 and SenT104.10 each contributed to approximately half of the identifiable within-polygenomic IBD segments, with little overlap in the ancestral IBD segments. We also found that the relatedness between SenT076.10 and SenT104.10 was negligible (relatedness = 0.01) (**Supplemental Figure S2.7**), which could suggest that SenT009.11 is the result of a natural genetic cross between SenT076.10 and SenT104.10.

For five other polygenomic infections, we could identify one strain that was highly related to an independent monogenomic infection. The proportion of shared IBD blocks between each polygenomic infection and related monogenomic infection varied but was on average 0.51. One polygenomic infection shared an unusually large proportion of its IBD segments shared with its related monogenomic infection, where 93% of its IBD segments were with SenT044.11.

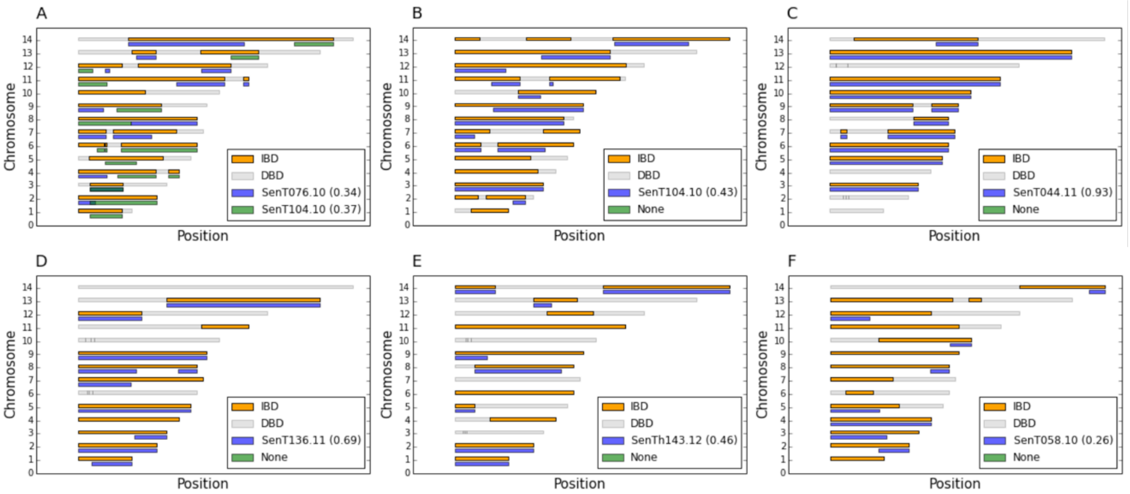


Figure 2.5 IBD maps within polygenomic infections and between monogenomic infections

Each subplot represents an individual polygenomic infection. A= SenT009.11, B=SenT100.11, C=SenT044.12, D=SenT210.12, E=SenT232.13, F=SenT232.13. Orange/grey color scheme represents the IBD map of the polygenomic infection, with orange representing regions of the genome that are IBD and grey representing regions of the genome not IBD. Blue/green color schemes represent regions of the genome that are IBD between the strains

(*Figure 2.5, continued*) found within the polygenomic infections and a related monogenomic strain. Blue bars indicate that region of the genome is IBD with one of the monogenomic strains while green bars indicate that region of the genome is IBD with the other monogenomic strain. Values in parenthesis indicate the proportion of the within-polygenomic infection IBD block that is explained by a particular monogenomic infection.

2.4.2 Expected relatedness with superinfection

Under the superinfection hypothesis, polygenomic infections are composed of parasite strains sampled from the local population. Here, we simulated the formation of polygenomic infections through superinfection by sampling from a set of 146 monogenomic infections previously collected from Senegal around the same time and place as our 31 polygenomic samples. These samples exhibit negligible population structure [56]. A polygenomic infection was simulated by drawing two random sets of SNPs from the full set of 3132, where each set of SNPs represents one of a pair of genomes in a superinfection. We assumed pairs of genomes because the average number of unique strains in our sample of polygenomic infections is two [55].

Our first sampling scheme did not correct for either differences in sample size or any potential bias in the monogenomic samples. We created a naive simulation of superinfection by quantifying the relatedness between all possible 146-choose-2 monogenomic sample pairs. We found that the distribution of relatedness is positively skewed, with 99% of the comparisons having a relatedness of 0. Under this naive simulation, the average relatedness of simulated polygenomic infections under superinfection is only 0.007 (**Supplemental Figure S2.8**).

Because the distribution of relatedness within real polygenomic infections was based on only 31 samples, we wanted to generate a simulation that took into account sampling variation. To do this, we generated simple random sampling bootstrap distributions of the mean relatedness between sample pairs **(Figure 2.6, blue)**. We calculated the mean relatedness of 31 randomly chosen sample pairs and repeated this process 40,000 times. We found that the mean relatedness of this distribution was extremely low (0.02). In addition, to correct for any potential strain bias in the set of 146 monogenomic samples, we also generated a weighted bootstrap distribution where monogenomic sample pairs were weighed according to the frequency of the corresponding 24-SNP barcode for each strain **(Figure 2.6, green)**. The 24-SNP barcode consists of 24 putatively neutral, unlinked sites that were used to profile parasite diversity in Senegal[24]. After correcting for potential ascertainment bias that would lead to an underestimate of true relatedness among monogenomic samples in the population, we found that the expected relatedness under superinfection was still very low (0.048.)

However analyzed, the simulated superinfections severely underestimate the level of relatedness within polygenomic infections (*p*-value in the naïve simulations = 1.1×10^{-21}, Mann-Whitney U). Attempts to correct for sample size and strain bias failed to recapitulate the level of relatedness actually observed within polygenomic infections. In both bootstrap simulations, the relatedness within simulated superinfections is significantly lower than the relatedness observed within polygenomic infections, with *p*-values ≤ 2.5×10^{-5} for both (*p*-value calculated using resampling techniques).

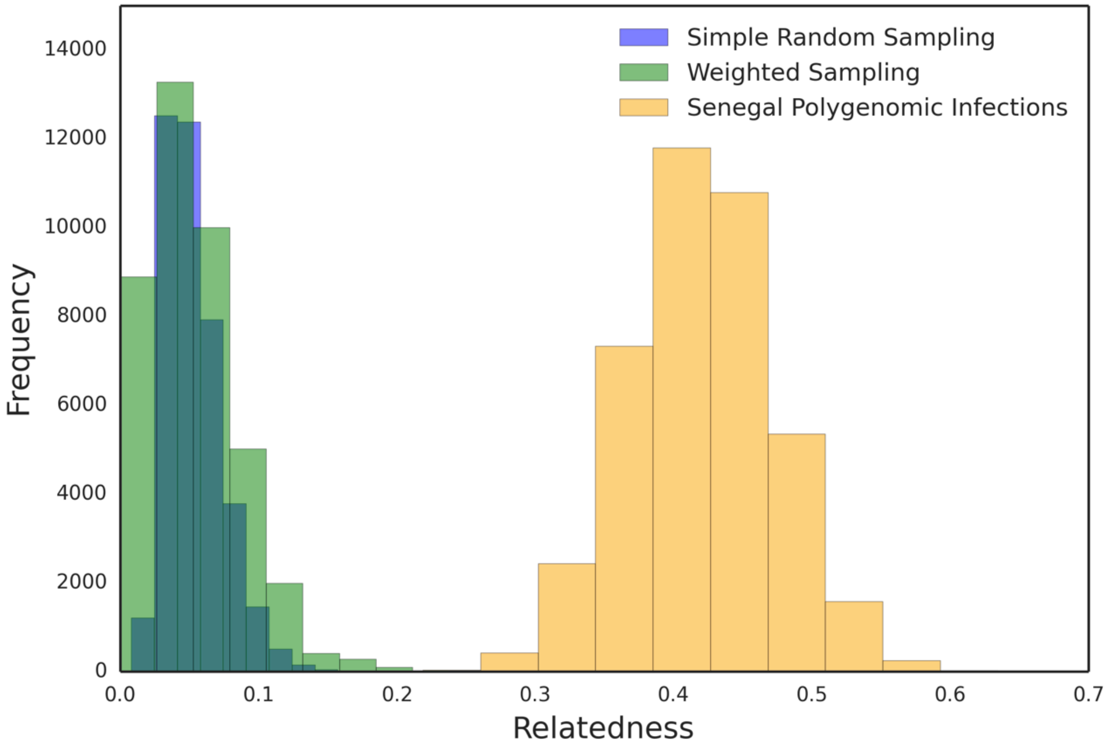


Figure 2.6 Expected relatedness under superinfection

Bootstrap distributions for the expected relatedness under superinfection were generated by randomly sampling with replacement 31 monogenomic pairs. For each set of 31 monogenomic pairs, we calculated the average relatedness and repeated this process 40,000 times to generate bootstrapped distributions of the mean relatedness between monogenomic infection pairs. Superinfection was simulated with either a simple random sampling scheme (blue), in which all

(*Figure 2.6, continued*) sample pairs were equally likely, or a weighted sampling scheme (green), which uses the barcode frequencies of the corresponding monogenomic samples to weigh each sample pair. Bootstrap resampled distributions of expected relatedness in polygenomic infections are shown in orange. *P*-values for both sampling schemes ≤ 2.5 x 10^{-5}.

2.5 Discussion

Understanding the genomic composition of polygenomic infections is crucial for the assessment of transmission based on the genetic profile of malaria infections and for generating epidemiological models relating population genomics to transmission intensity. In this study, we investigated whether polygenomic infections simulated under superinfection conditions would accurately recapitulate the genetic relatedness observed in 31 natural polygenomic infections collected from patients in Thiès, Senegal. We first developed a strategy that offers a simple, cost-effective way of quantifying the relatedness within polygenomic infections without serial dilution or flow sorting single cells. Previous studies have characterized the relatedness within polygenomic infections by isolating individual parasite haplotypes through culture adaptation, serial dilution or flow sorting [41–43]. Our pipeline uses standard Illumina sequencing reads to interpret the relatedness within polygenomic infections from direct patient samples without needing to establish linkage phase, which greatly increases the number of polygenomic infections one can examine. This approach trades the resolution of previous approaches in exchange for reduced sample preparation requirements and does not require that cells be preserved intact. Our methodology is more applicable to a broader range of samples, which may be useful for understanding the relatedness of polygenomic infections in different transmission settings.

However, alternative sequencing approaches should be considered when analyzing polygenomic infections with a COI > 2. While our approach works well when COI is 2, it underestimates the relatedness of polygenomic infections with COI > 2, since the constructed pseudohaplotypes will combine the differences across all strains in the infection. Polygenomic infections identified as being composed of apparently unrelated parasites by our method may in fact be composed of 3 or more strains of varying degrees of relatedness. Thus, the genomic haplotypes of more complex polygenomic infections should be established prior to using our HMM. Haplotypes can be established using sequencing technologies that generate longer reads, but haplotype reconstruction can be computationally challenging, especially in situations where the relative frequency of strains are not the same (reviewed in [57]). Single-cell sequencing, which was previously used to calculate the relatedness of strains in polygenomic infections for both *P. falciparum* and *P. vivax* [43], has the advantage of avoiding complex haplotype reconstruction algorithms but is extremely labor intensive. Although our HMM will be useful for quantifying the relatedness of more complex infections, quantifying the relatedness of more complex polygenomic infections will require more sophisticated sequencing technologies or haplotype reconstruction algorithms.

Our study also contributes to a growing body of evidence indicating that cotransmission is common in natural parasite populations. Studies in low transmission areas, such as the Peruvian Amazon [44] and Thai-Burma border [25,42,43], have reported highly related parasite strains within polygenomic infections. Highly related polygenomic infections are also observed in high transmission areas [41,42], despite the fact that patients are exposed to a large numbers of infectious mosquito bites. Here, we simulated superinfection as the random sampling of parasites from those found in Thiès, Senegal and found that a pure superinfection model fails to explain the observed relatedness within natural polygenomic infections.

When constructing our superinfection simulations, we assumed that the parasite population in Thiès, Senegal was completely mixed, with no hidden population structure. This is an oversimplification, since malaria transmission becomes clustered around transmission foci at low transmission settings [58]. To date, there is no genetic evidence of population structure in this region [56], but this could be because the sample collection was insufficient to capture the effects of localized transmission foci or other spatial heterogeneity effects. Spatial clustering can result in localized inbreeding events that raise the relatedness of parasites in the surrounding region and thus increase the relatedness of true superinfections. We believe it is unlikely that the relatedness in our polygenomic infections is due solely to the sampling of infections from transmission clusters, since the majority of parasites in Senegal are unrelated to one another [40] and because patients reporting to clinic do not necessarily live in the same areas of

Thiès. However, since patient data regarding residence and travel history were not made available, we cannot exclude this possibility. We recognize that the relatedness of superinfection events could be influenced by the inhibition of future strains due to the host immune response, but we suspect these are effects are small and previous studies have observed similar findings in children with little or no premunition [42].

The wide range of polygenomic relatedness values in Senegal suggests that our polygenomic infections may represent a mix of both superinfection and cotransmission events. Some polygenomic infections include apparently unrelated parasite genomes, but it is unclear whether these result from superinfection or the cotransmission of unrecombined parasite genomes. With self-fertilization in the mosquito, it is theoretically possible for two unrelated genomes to be cotransmitted by a single mosquito host. This problem could be exacerbated if there is a preference for self-fertilization or selection occurring within the mosquito vector and human host. These complications make it difficult to estimate the rate of cotransmission based solely on the frequency of highly related genomes in polygenomic infections. Nonetheless, our data suggest that cotransmission is frequent in Thiès, Senegal and may be a dominant mechanism by which polygenomic infections persist in low transmission settings.

Previously, Nkhoma et al [42] suggested that extreme degrees of genetic relatedness within polygenomic infections could be the result of repeated cotransmission events, or serial cotransmission chains. Analyses of experimental crosses indicate that the mean relatedness between F_1 progeny is approximately normally distributed with a mean of 0.52 and a standard deviation of 0.08 [59]. In our data (Figure 2), 6.5% of polygenomic infections exhibit genomic relatedness exceeding 0.76, which is three standard deviations above the mean in experimental crosses, and also suggests serial cotransmission. The relatively low frequency of such closely related genomes might suggest that serial cotransmission over multiple generations is rare in this population. However, because polygenomic infections were identified based on the proportion of sites with non-unanimous reads, some of the infections classified as monogenomic may actually be polygenomic infections with extremely related parasite strains. This issue could be resolved by analyzing samples with higher read depth coverage. Because we were concerned about the loss of low frequency strains, our samples were directly sequenced from patient samples. This meant that the majority of generated reads aligned to the human genome. The genomes of parasites within some of these samples were only represented by 300 SNPs, which complicates the detection of sites with non-unanimous reads in highly related samples. Future studies could use selective whole genome amplification or hybrid selection to generate higher quality samples but will need to consider the potential for strain amplification bias.

A major implication of this work is that genetic epidemiology models can be improved by accounting for the genetic relatedness within polygenomic infections. The rates of superinfection and cotransmission may change depending on the transmission setting. In high transmission settings, genetic epidemiology models that simulate polygenomic infections as the result of superinfection may be sufficient, since superinfection is expected to be more common than cotransmission [48]. However, this assumption may be suspect, due to the observation of highly related haplotypes in polygenomic infections from high transmission settings [42], and cotransmission could still play a major role in these areas. In mid-low transmission settings, genetic epidemiology models should be adjusted to take into account the genetic relatedness of polygenomic infection owing to cotransmission, since superinfection will underestimate the genetic relatedness of polygenomic infections. Future studies are needed to quantify the relative rates of cotransmission and superinfection, but cotransmission can be incorporated into existing models by simulating the sampling of parental genotypes and sexual reproductive processes within the mosquito vector to determine the relatedness of the subsequent polygenomic infection. The explicit modeling of cotransmission connects the relatedness of polygenomic infections to the genetic composition of local parasite population, allowing it to be affected by changes in transmission intensity and is applicable across any epidemiological setting.

The incorporation of related strains within polygenomic infection is important for understanding the genetic composition of parasite populations, particularly those in low transmission settings, since it can lead to differences in modeled expectations. Theoretical models of superinfection suggest that superinfection can greatly increase selection efficiency within the host [60] and can affect the fitness of drug resistant parasites [61]. However, the presence of related strains within infections can alter these effects. For example, one study found that simulated infections composed of unrelated parasite strains can have different infection lengths compared to those of related strains [62]. Models that incorporate cotransmission should provide more accurate predictions, which will be helpful in malaria elimination activities to monitor transmission, assess the impact of interventions, and improve our understanding of the underlying biology and consequences on important traits such as drug resistance that threaten to undermine our elimination efforts.

Finally, the high prevalence of highly related polygenomic infections suggests that current methods for estimating COI can be improved. We previously published a method for estimating the COI of polygenomic infections based off a set of biallelic SNP markers [55]. Our method, known as COIL, assumes that polygenomic infections are composed of unrelated parasite strains, which we now know is not always the case in natural populations. Recognition that polygenomic infections can be composed of related parasite strains suggests that estimated COI levels could be reported as continuous rather than discrete values in settings where co-transmission is prevalent.

2.6 Conclusions

To conclude, we find that models that simulate polygenomic infections through superinfection do not produce the high degree of relatedness observed within a set of 31 natural polygenomic infections collected from patients in Thiès, Senegal. The relatedness within these polygenomic infections suggests that cotransmission plays a major role in the persistence of polygenomic infections.

Our data support the hypothesis that the cotransmission of genetically related parasite strains is common, and that this aspect of transmission should be incorporated into existing genetic epidemiology models. These findings have important implications for our understanding of malaria transmission, and potentially how important phenotypes like drug resistance that threaten to undermine malaria elimination activities may be promoted. As public health interventions drive parasite populations toward elimination, these models will play a critical role in understanding the changes in population structure associated with declining transmission rates and influencing the future of public health policy.

2.7 Addendum

2.7.1 List of abbreviations

EIR: Entomological inoculation rate**,** SNP: Single Nucleotide Polymorphism, DNA: Deoxyribonucleic acid, IBD: Identical by descent, DBD: Different by Descent, COI: Complexity of Infection, HMM: Hidden Markov Model, NMCP: National Malaria Control Programme

2.7.2 Declarations

Ethics approval and consent to participate

All human samples were collected after recruitment and with written consent from either the subject or a parent/guardian. This protocol was reviewed and approved by the ethical committees of the Senegal Ministry of Health (Senegal) and the

Harvard T.H. Chan School of Public Health (16330-110, 2008) for Senegalese subjects. This study conforms to the principles established in the Declaration of Helsinki.

2.7.3 Authors' contributions

WW performed the analysis and was a major contributor in writing the manuscript. ADG identified the trusted SNP set. RFD performed the lab mixtures, extracted parasite DNA from samples, and provided barcode data. SFS wrote the source code for the HMM. DN, AKB, ABD collected the samples from Senegal. SKV, BM, DEN, DLH, and DFW helped supervise the project. All authors reviewed and approved this manuscript.

2.7.4 Acknowledgements

We thank Hsiao-Han Chang, Angela Early, Seth Redmond, and Aimee Taylor for their helpful advice and discussions. We would like to especially acknowledge Katelyn Renke, who helped with the genomic DNA extractions and library preparations. We also acknowledge the contributions of Ngayo Sy, Younouss Diedhiou, Lamine Ndiaye and Amadou Moctar Mbaye for their help in collecting samples from Thiès, Senegal.

Chapter 3: Modeling the genetic relatedness of *Plasmodium falciparum* parasites following meiotic recombination and cotransmission[2]

Wesley Wong[1], Edward A. Wenger[2], Daniel L. Hartl[1,3], Dyann F. Wirth[1,4*]

[1] Department of Immunology and Infectious Diseases, Harvard T. H. Chan School of Public Health, Boston, Massachusetts, United States of America

[2] Institute for Disease Modeling, Bellevue, Washington, United States of America

[3] Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, Massachusetts, United States of America

[4] Broad Institute, Cambridge, Massachusetts, United States of America

* Corresponding author

E-mail: dfwirth@hsph.harvard.edu (DFW)

[2] This chapter has been submitted to PloS Computational Biology and accepted under the title "*Modeling the genetic relatedness of Plasmodium falciparum parasites following meiotic recombination and cotransmission*"

3.1 Abstract

Unlike in most pathogens, multiple-strain (polygenomic) infections of *P. falciparum* are frequently composed of genetic siblings. These genetic siblings are the result of sexual reproduction and can coinfect the same host when cotransmitted by the same mosquito. The degree with which coinfecting strains are related varies among infections and populations. Because sexual recombination occurs within the mosquito, the relatedness of cotransmitted strains could depend on transmission dynamics, but little is actually known of the factors that influence the relatedness of cotransmitted strains. Part of the uncertainty stems from an incomplete understanding of how within-host and within-vector dynamics affect cotransmission. Cotransmission is difficult to examine experimentally but can be explored using a computational model. We developed a malaria transmission model that simulates sexual reproduction in order to understand what determines the relatedness of cotransmitted strains. This study highlights how the relatedness of cotransmitted strains depends on both within-host and within-vector dynamics including the complexity of infection. We also used our transmission model to analyze the genetic relatedness of polygenomic infections following a series of multiple transmission events and examined the effects of superinfection. Understanding the factors that influence the relatedness of cotransmitted strains could lead to a better understanding of the population-genetic correlates of transmission and therefore be important for public health.

3.2 Background

Unlike most bacterial and viral pathogens, the malaria parasite *P. falciparum*, while predominantly haploid, must sexually reproduce in a mosquito vector before infecting a new human host. Sexual recombination has a significant impact on the population genomics of the parasite, and its effects depend on epidemiological conditions such as transmission intensity [35,38,63]. One outcome of sexual recombination is that parasites transmitted by a mosquito vector can be genetically related, which can be measured as the proportion of the genome that is identical-by-descent (IBD). IBD segments are region of the genome that originate from a recent common parental strain. A number of studies have used IBD to study transmission [40–42,64,65], survey antimalarial resistance [21], and detect signals of selection [66].

The effects of sexual recombination are also apparent in polygenomic (multi-strain) infections. Polygenomic infections can be formed through a series of infectious mosquito bites (superinfection) or through the transmission of multiple strains from the a single mosquito bite (cotransmission) [42,61,64]. Coinfecting strains resulting from superinfection are assumed to be unrelated while those resulting from cotransmission are assumed to be genetically related [41,42,64]. While superinfection is believed to be common in high transmission settings, owing to high entomological inoculation rates and complexity of infections (COI, the number of strains per infection) [12,48], the frequency with which cotransmission occurs is less clear. Studies of genetic relatedness in symptomatic polygenomic infections reporting to clinics in mid-to-low transmission settings show that cotransmission is prevalent in these regions [41,42,64,65], but little is known of the frequencies of cotransmission and superinfection across transmission settings. Genetic relatedness studies reveal a large amount of variation in the relatedness of polygenomic infections. The fact that sexual recombination occurs within the mosquito suggests that the relatedness in these polygenomic infections is associated with transmission. High relatedness in polygenomic infections could be indicative of serial cotransmission chains [43], but it is unclear what other factors may influence the relatedness of polygenomic infections.

Part of the uncertainty stems from an incomplete understanding of the cotransmission process. When a female Anopheline mosquito bites an individual infected with malaria, she ingests male and female gametocytes. The ingestion of these gametocytes activates them to form gametes that fuse to create a diploid zygote. Gametes can fuse with other gametes of the same genotype, resulting in self-fertilization (selfing), or can fuse with gametes from other genotypes resulting in outcrossing. The zygote undergoes meiosis and develops into a motile ookinete that traverses the midgut epithelial layer and forms an oocyst. Within the oocyst, the parasite undergoes many rounds of mitosis to create thousands of haploid sporozoites. These sporozoites travel to the mosquito salivary glands and are stored until deposited by the mosquito into the human host during a blood meal. Only those sporozoites that invade the liver will survive to continue the malaria life cycle. How then could variation in within-host and within-vector transmission dynamics, such as the number of oocysts formed and the number of sporozoites infecting the liver, affect the relatedness of cotransmitted strains, and how could these variables in turn affect the relatedness of polygenomic infections in natural populations?

To address the complexity of this transmission cascade and better understand the process of cotransmission, we devised a classification framework based on parasite pedigrees and kinships to develop an understanding of how the various sampling and mating events within the mosquito vector affects the relatedness of transmitted sporozoites. We then created a transmission model to quantify the relatedness of cotransmitted strains under a variety of within-host and within-vector dynamics and used this model to examine the relatedness of polygenomic infections in transmission chains. Our study reveals new insights into the cotransmission process, which we believe will be useful for the interpretation of population genomic signals obtained from more complicated population-level models or from natural populations.

3.3 Results

3.3.1 Simulating sexual recombination

To simulate sexual recombination, we developed a *P. falciparum*-specific meiosis model based on the whole genome sequences of 69 genetically distinct progeny derived from 3 previously generated *P. falciparum* crosses involving different laboratory-adapted strains (3D7, HB3, Dd2, 7G8, and GB4) [67–71]. The whole genome sequences generated from these crosses are one of best sources of data for designing a *P. falciparum*-specific meiosis model because the genotypes of the parental strains are known. Furthermore, we can be confident of the number of sexual reproduction cycles separating progeny and parental strains. While previous IBD analyses of parasites from natural parasite populations have identified putative F_1 progeny [42,44,64,72], having complete knowledge of parental ancestry simplifies the identification of IBD segments and allows us to better identify recombination events throughout the genome. We calculated the number of crossover events and inter-crossover distances (**Supplemental Figure S3.1 & Supplemental Table S3.1**) using a hidden Markov model (HMM) [40,73] to identify IBD segments shared between progeny and parental strains (Methods). We then used this data to test the fit of two different meiosis models, one with and one without obligate chiasma formation. Both were based off the gamma model of crossover formation, which has been used to characterize recombination events in a wide variety of taxa, including *H. sapiens, D. melanogaster*, and *S. cerevisiae* [74–77]. The gamma model is an improvement over simpler Poisson-based crossover models because it allows us to explore a wide range of crossover interferences.

Regardless of whether obligate chiasma formation was modeled, the number of crossover events and intercrossover distances in our simulated meiotic events resembled those of the laboratory-crossed progeny (**Figure 3.1A, 1B**). However, both meiosis models underestimated the frequency of short intercrossover distances (< 50 cM) (**Figure 3.1A**), which we suspect is because our HMM overestimated the frequency of short intercrossover distances in the laboratory-cross data (**Supplemental Figure S3.2**). We found that the obligate chiasma model generated crossover events that were more consistent with that of the laboratory-crossed progeny, but overestimated the number of chromosomes with two crossover events. Using a pseudo-likelihood function (Methods), we determined that an obligate chiasma model fit the data better than a non-obligate chiasma model (**Figure 3.1C**). However, we could not estimate the level of crossover interference. Because crossover interference is observed in a wide-variety of organisms spanning multiple taxa [74], we chose to use an obligate chiasma meiosis model with a weak level of interference (gamma distribution with shape = 2, scale = 0.38) for all of our transmission simulations.

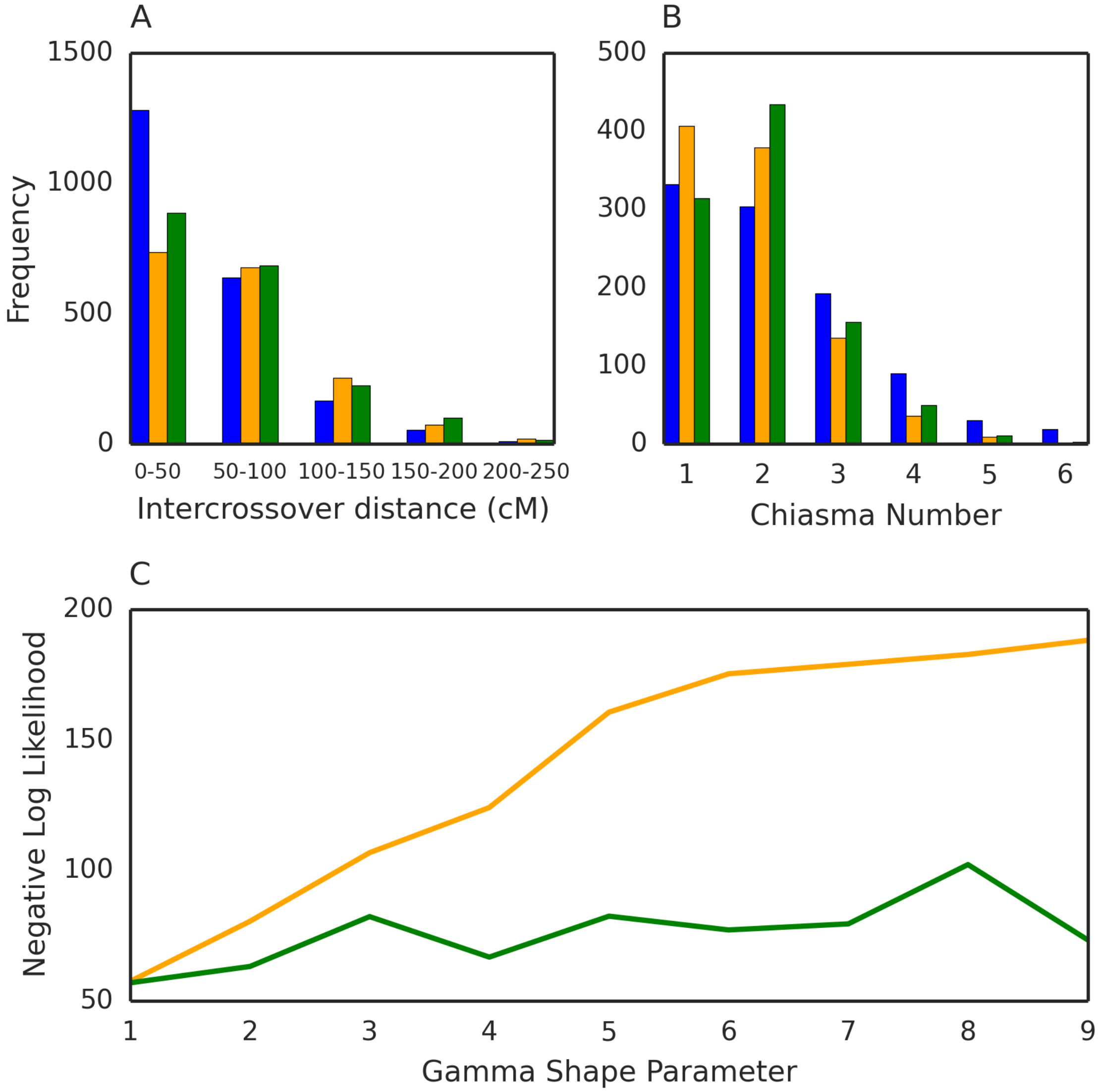


Figure 3.1 Meiosis simulations.

Blue indicates data obtained from the progeny of strains crossed in the laboratory. Orange indicates simulated data using the non-obligate chiasma meiosis model while green indicates simulated data from the obligate chiasma meiosis model. **A**) Barplots of the intercrossover distances of all 14 chromosomes of the *P. falciparum* genome. **B**) Barplots of the number of chiasma scattered throughout the genome. For **A** and **B**, simulated data were

(*Figure 3.1, continued*) generated using a shape parameter of 2. **C**) Line plots of the negative pseudolikelihood values of the non-obligate and obligate meiosis models at different levels of crossover interference. Along the x-axis are different levels of crossover interference (determined by the value of the shape parameter). A shape parameter of 1 indicates no crossover intereference. Lower negative pseudo-likelihood values indicate a better fit to the data obtained from the progeny of experimentally lab-crossed strains.

3.3.2 Role of pedigree and kinship in determining genetic relatedness.

To develop an intuition of how the relatedness of cotransmitted strains is influenced by within-host and within-vector transmission dynamics, we developed a framework for understanding how these aspects could affect the relatedness of cotransmitted strains. We reasoned that changes in oocyst counts and COI could alter the relatedness of cotransmitted strains by influencing how gametocytes are sampled and mate within the mosquito vector. This framework is based on one of nine possible pedigrees describing sporozoite pairs. These pedigrees are defined by 1) whether sporozoites are sampled from multiple oocysts and 2) whether these oocysts are the result of selfing or outcrossing. If sporozoites are sampled from multiple oocysts, we consider a third criterion: the number of parental strains shared between oocyst pairs.

We used our meiosis simulations to quantify the relatedness of sporozoites described by each possible pedigree. Based on the parental ancestries described by our nine pedigrees and the estimate of relatedness provided by our meiosis simulation, we also grouped parasites using kinship definitions. These kinship definitions are analogous to those used in diploid organisms and have been used in other IBD analyses [42]. However, we found that sporozoites sampled from a single, outcrossed oocyst (pedigree 3) could not be described by existing kinship categories. Because they originate from the same meiotic event, we describe their kinship as "meiotic siblings." Although the average relatedness of meiotic siblings is 0.5, our meiosis simulation revealed that the distribution is bimodal, with one mode at 1.0 (the expected relatedness of genetically identical meiotic siblings) and one mode at 0.33 (the expected relatedness of genetically distinct meiotic siblings) (S3 Fig). Our pedigree/kinship framework and meiosis simulation results are summarized in **Figure 3.2.**

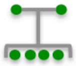
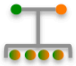
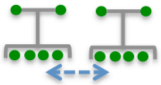
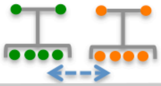
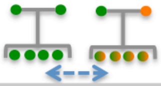
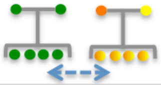
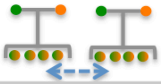
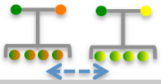
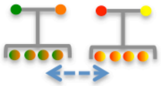
Pedigree ID	Oocyst pedigree	Parents shared between oocysts	Expected relatedness	Kinship
1			1.0	Clones
2			0.5 (1.0 / 0.33)	Meiotic Siblings
3		2	1.0	Clones
4		0	0.0	Unrelated
5		1	0.5	Parent-offspring
6		0	0.0	Unrelated
7		2	0.5	Full Siblings
8		1	0.25	Half-siblings
9		0	0.0	Unrelated

Figure 3.2 Pedigrees between parasites and their expected relatedness.

The 9 possible pedigrees describing parasite pairs. Pedigrees represent the genetic ancestry of parasites and the oocysts they are sampled from. Circles at the top of each pedigree represent the gametes that fuse and undergo meiosis while circles at the bottom represent the sporozoites that are generated following meiosis and expansion in the oocyst. Different colors represent different genomes. Sporozoites with mixed colors indicate they are the result of outcrossing. Blue arrows between pedigrees indicate that sporozoites are sampled from different oocysts. Parasites can be sampled from the same oocyst

(*Figure 3.2, continued*) (pedigrees 1 and 2) or from multiple oocysts (pedigrees 3-9). For pedigree 2, the distribution of expected relatedness was bimodal, and we provide the average across the entire distribution (top) as well as the two modes (bottom). Pedigree 6 and 8 are only accessible when COI ≥ 3 and pedigree 9 is only accessible when COI ≥ 4.

3.3.3 Transmission Model Description

We then designed a transmission model that partitions transmission into three steps: 1) The host-vector sampling of gametocytes from an initial host infection 2) the sequence of events starting from gamete fusion and meiosis to the development of the oocyst within the mosquito vector, and 3) the vector-host injection of sporozoites and subsequent invasion of the liver to determine the genetic composition of the next human host (**Figure 3.3**). We initiate our model by simulating a mosquito blood-feeding event on a polygenomic infection comprised of unrelated strains and parameterized by 1) COI, 2) oocyst count, and 3) the infected hepatocyte count. The number of unique strains present in the initial infection is determined by COI. In our model, we consider oocyst formation as the final outcome of gamete fusion and subsequent meiosis. Based on the oocyst count, our model samples gamete pairs, which fuse and undergo meiosis to create an oocyst consisting of four unique meiotic products. Competition within the oocyst is not modeled and we assume that each meiotic product is present at equal proportion in the oocyst. After all oocysts are created, the model samples sporozoites according to the infected hepatocyte count to determine the genetic composition of the subsequent host infection. If the resulting infection harbors multiple strains, we calculated the relatedness of cotransmitted strains as the average pairwise relatedness between each of the unique genotypes present in the final host infection.

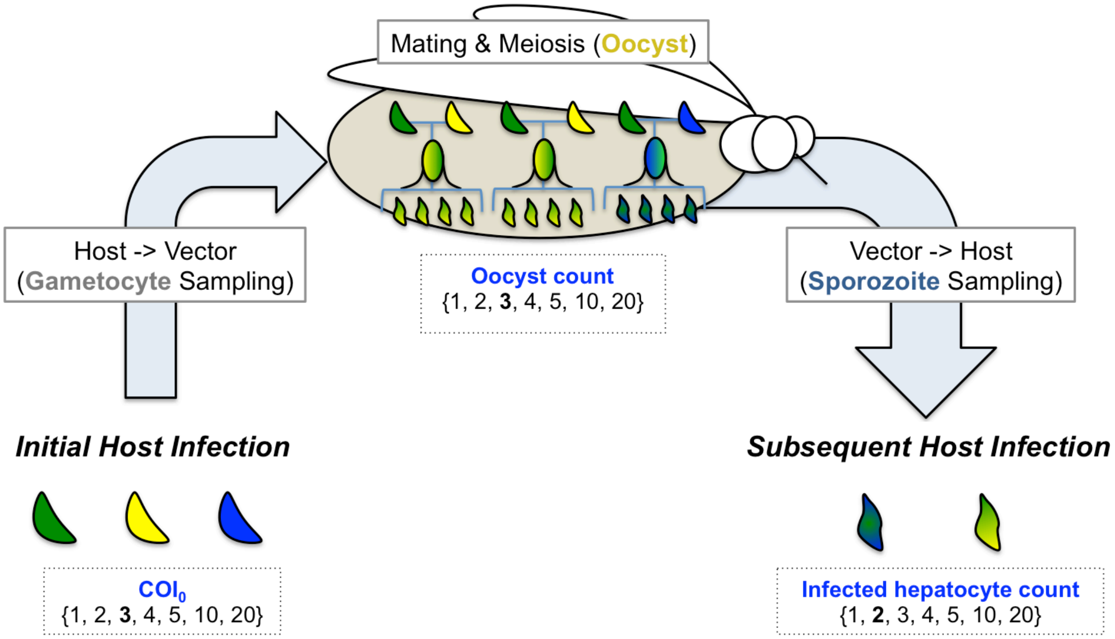


Figure 3.3 Model of parasite transmission.

Model of transmission where genetically distinct parasites are distinguished by color. Parameter values are drawn from the set {1, 2, 3, 4, 5, 10, 20}, which represents the range of values observed in real life simulations. The bold number indicates the value in the example shown in the figure. Gametocytes are sampled from an initial polygenomic infection comprised of unrelated parasite strains. Sampled gametocytes are used to create oocysts within the mosquito midgut (middle). Each oocyst summarizes the entire sequence of events starting from

(*Figure 3.3, continued*) gamete fusion to the end of meiosis. Oocyst are represented using a stylized pedigrees tree, where the crescents at the top represent parental strains undergoing meiosis, the oval in the center indicates whether the mating event is the result of selfing (solid color) or outcrossing (color-gradient), and the sporozoites at the bottom represent the four meiotic products generated through meiosis. Those with multiple colors indicate that genomes have undergone effective recombination and are genetically distinct from the parental strains and to each other. Sporozoites are sampled from the total pool of meiotic products to determine the genetic composition of the subsequent host infection. The number of sporozoites sampled is determined by the infected hepatocyte count used in the simulation.

The values for the infected hepatocyte count are pre-specified and drawn from the set {1, 2, 3, 4, 5, 10, 20}. Simulations with COI =1 were excluded because they always result in selfing and the transmission of genetic clones. Simulations with an infected hepatocyte count = 1 were also excluded, as they cannot result in cotransmission. Small values are overrepresented to reflect the right-skewed distributions of oocyst counts observed in mosquito feeding assays and infected hepatocyte counts estimated from a malaria-challenge study [78–80]. These values also include the COI observed in naturally occurring polygenomic infections from mid-to-low endemic settings (COI ranging from 2-6 in polygenomic infections).

3.3.4 Single oocyst simulations guarantee the transmission of meiotic siblings or genetic clones

From our pedigree/kinship framework, we knew that sporozoites sampled from a single oocyst would be either genetic clones or meiotic siblings. Our transmission simulation confirmed this prediction and found that the expected relatedness of cotransmitted strains in single-oocyst transmission simulations was always 0.33 (**Figure 3.4**), which is the expected relatedness of genetically distinct meiotic siblings. In single oocyst transmission simulations, cotransmission can only be achieved by the transmission of two or more genetically distinct meiotic siblings. The distinction between genetically distinct and genetically identical meiotic siblings is relevant in the context of cotransmission, as the transmission of clonal meiotic siblings cannot result in cotransmission. Changes to the infected hepatocyte count do not affect the expected relatedness values, but higher infected hepatocyte counts caused the distribution to be more concentrated around the mean.

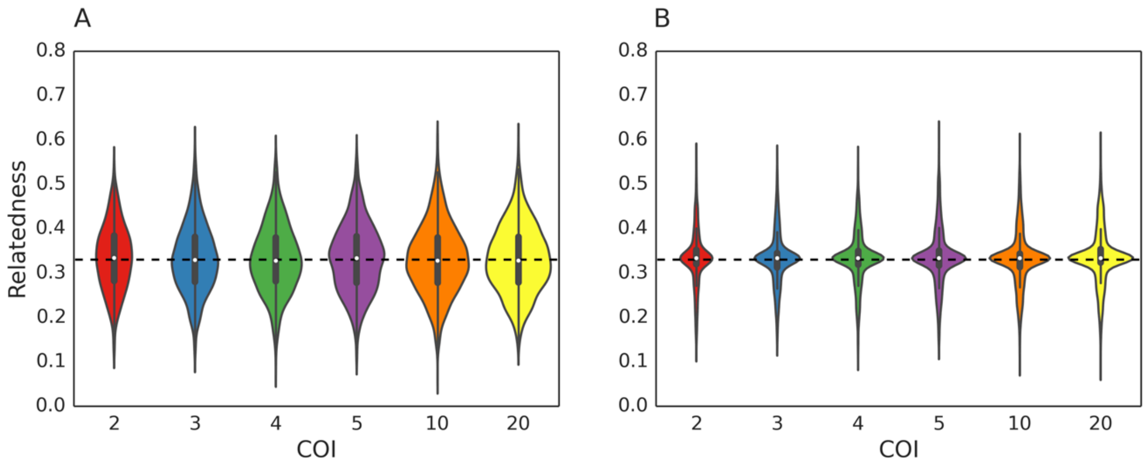


Figure 3.4 Relatedness of cotransmitted strains when oocyst count is 1.

Violin plots of the relatedness of cotransmitted strains in single oocyst simulations. Only the results for simulations with infected hepatocyte count of 2 (**A**) or 5 (**B**) are shown. The expected relatedness (in terms of both median and mean) is always 0.33. A box plot is drawn in the center of each violin plot, where the white dot represents the median of the distribution, the thicker line represent the interquartile range, and the thinner line represents the whiskers of the box plot, up to 1.5 times the interquartile range. The horizontal dotted line represents the value of 0.33.

3.3.5 The expected relatedness of cotransmitted strains in multiple oocyst simulations depends on COI

In multiple oocyst transmission simulations, the relatedness of cotransmitted strains is not as easy to predict, since multiple kinships can be transmitted. Based on our pedigree/kinship framework, we hypothesized that COI modulates the expected relatedness of cotransmitted strains by limiting the transmission of half-siblings and unrelated strains; the transmission of half-siblings and unrelated strains described by pedigrees 6 are only possible when COI ≥ 3. The transmission of unrelated strains described by pedigree 9 only applies when COI ≥ 4.

Our transmission simulation simulations confirmed these predictions and revealed a simple relationship between COI, oocyst count, and the relatedness of cotransmitted strains (**Figure 3.5, 3.6**): the relatedness of cotransmitted strains declines with increasing COI. All COI = 2 simulations have an expected relatedness > 0.33, with a larger increase in high oocyst count simulations. The increase in relatedness is a reflection of the increased transmission of full-siblings and parent-offspring strains. When COI = 3, increasing oocyst counts no longer increased the expected relatedness of cotransmitted strains due to the additional transmission of half-siblings. Once COI > 4, increasing oocyst counts decreased the expected relatedness of cotransmitted strains. This was due to the increased transmission of unrelated strains, particularly those described by pedigree 9 (outcrossed oocysts that do not share any parental strains) (Fig 6C-D, c-d). When COI = 20, the majority of transmitted parasites are either meiotic siblings or unrelated strains described by pedigree 9.

We found that different infected hepatocyte counts altered the distribution of relatedness (**Supplemental Figure S3.4 Fig & Supplemental Figure S3.5**) but had no effect on the trends established by either COI or oocyst count. Again, simulations with a COI = 2 consistently had the highest expected relatedness values while simulations with higher COIs had lower expected relatedness values, regardless of the infected hepatocyte count.

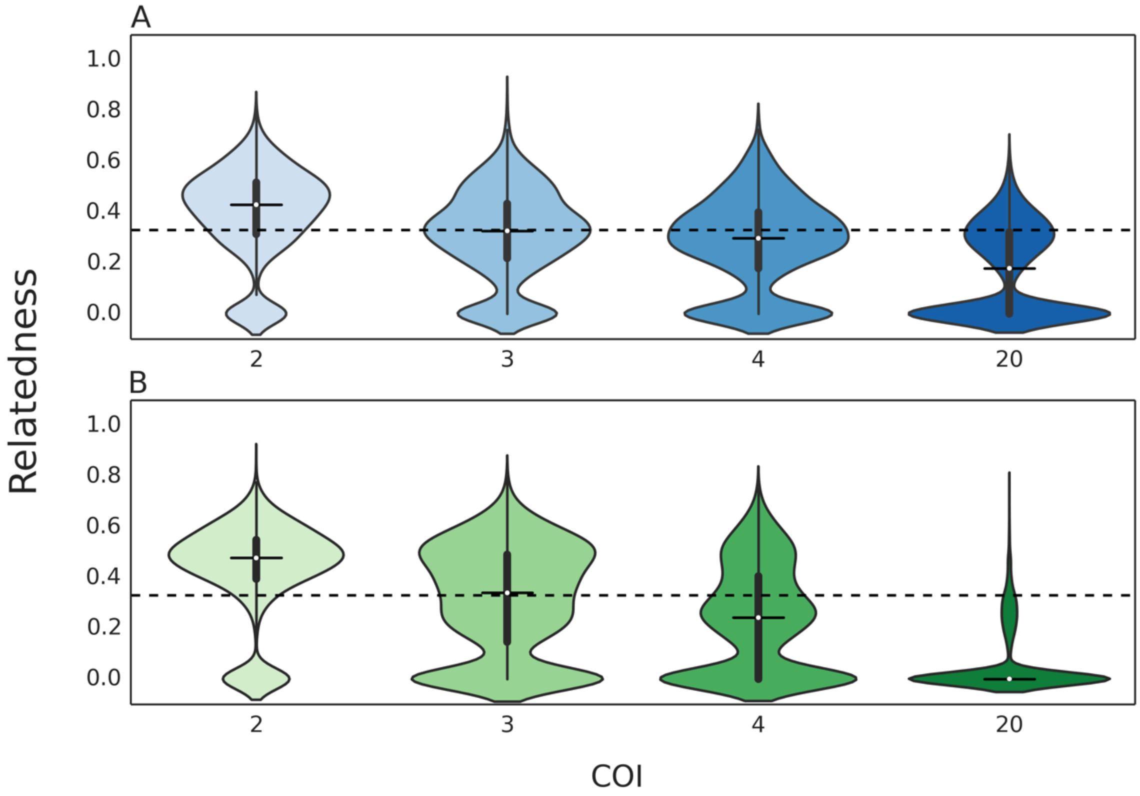


Figure 3.5 Relatedness of cotransmitted strains in multiple oocyst simulations.

Violin plots of the relatedness of cotransmitted strains in multiple oocyst simulations. Only the results for simulations with oocyst counts of 2 (A) and 20 (B) and infected hepatocyte counts of 2 are shown. The expected relatedness of cotransmitted strains declines with increasing COI. A box plot is drawn in the center of each violin plot, where the white dot represents the median of the distribution, the thicker line represent the interquartile range, and the thinner line represents the whiskers of the box plot, up to 1.5 times the interquartile range. The horizontal dotted line represents the value of 0.33.

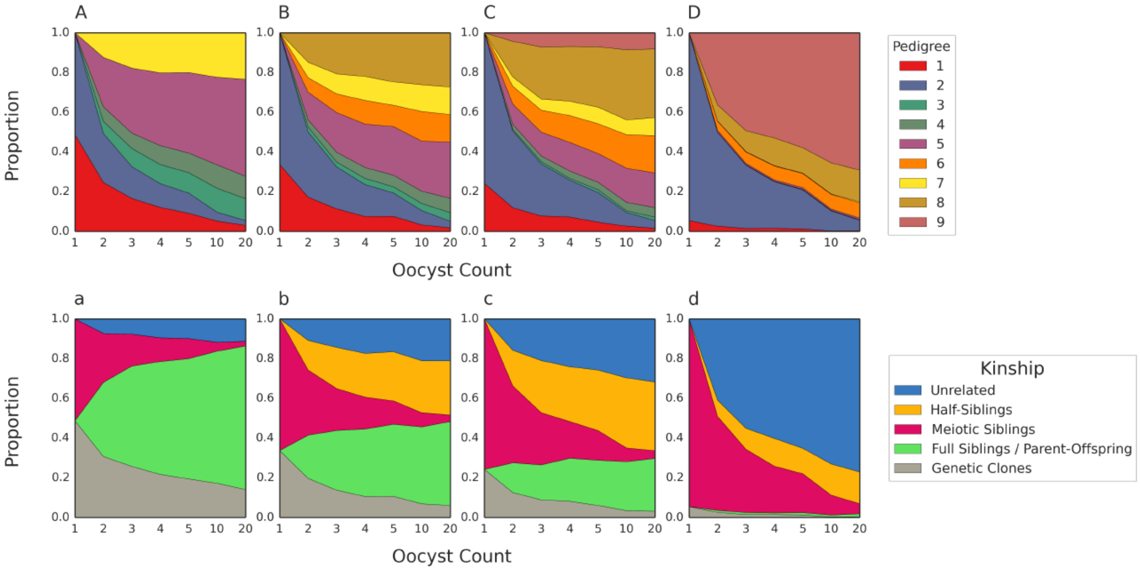


Figure 3.6 Pedigree and kinship frequencies from multiple oocyst simulations.

Stacked line charts of the frequencies of different pedigrees (A-D) and kinships (a-d) plotted against oocyst count. Each subplot represents a scenario with a different COI (A/a = 2, B/b = 3, C/c = 4, D/d = 20). Only the results from simulations where infected hepatocyte count = 10 are shown. Genetic clones are defined as those emerging from oocysts characterized by pedigree 1 and 3; genetically identical meiotic siblings are still classified as meiotic siblings in this graph

3.3.6 Investigating the effect of non-uniform gametocyte sampling probabilities

Thus far, our simulations have assumed that the strains making up polygenomic infection are present and sampled in equal proportions. However, strain proportions in natural polygenomic infections can be highly skewed. Furthermore, different strains can have different transmissibility relating to factors such as gametocyte production. To investigate how skewed gametocyte sampling probabilities could affect the relatedness of cotransmitted strains, we devised a weighted sampling scheme defined by the ratio of the most frequent to the least frequent strain in the infection (**Methods**).

Predictably, skewing the gametocyte strain ratios increased the rate of selfing and the transmission of genetic clones (**Supplemental Figure S3.6**). Skewed ratios of up to 10:1 increased relatedness of cotransmitted strains by a small amount. Ratios ranging from 1:1 to 10:1 increased the expected relatedness of cotransmitted strains by 0.01 – 0.10. This increase depended on both COI and the magnitude by which strains proportions differed. The relatedness of cotransmitted strains from high COI infections was more robust to differences in strain proportions; a 10:1 ratio in a COI = 20 infection increased relatedness by only 0.02 while a 10:1 ratio in a COI = 3 infection increased relatedness by 0.03-0.06.

3.3.7 Generating a combined model of cotransmission and examining serial cotransmission chains

The genetic composition of natural polygenomic infections can result from multiple transmission events and influenced by population-level transmission dynamics. However, developing a model that take into account all possible population-level transmission dynamics is beyond the scope of this paper. Instead, we used our model to quantify the relatedness of polygenomic infections in three different multiple transmission simulations, which we refer to as transmission lineages. Each transmission lineage is designed to resemble transmission chains that occur in natural populations and initiated by simulating a mosquito blood-feeding event on a polygenomic infection comprised of unrelated strains. The first transmission lineage does not allow superinfection; all subsequent transmission events in the chain must infect uninfected hosts. The second and third transmission lineages allow superinfection and are differentiated by the nature of the resident strain in the soon-to-be superinfected host. For the second transmission lineage, the resident strain is identical to one of the parental strains in the initial polygenomic infection (resembling natural backcrossing events). For the third transmission lineage, the resident strain is not related to any of the parental strains in the initial polygenomic infection but is the same in all transmission events. In the last transmission lineage, the resident strain is not related to any of the parental strains in the initial polygenomic infection and is different in all transmission events.

For our transmission lineage simulations, we modified our cotransmission model so that oocyst and infected hepatocyte counts are determined by randomly sampling from distributions reflecting those of found in previous studies [80,81]. Subsequent transmission events sample parasites from the infection generated by the previous transmission event. Allowing oocyst and infected hepatocyte counts to be chosen from these distributions did not affect the previously observed relationship between COI and the relatedness of cotransmitted strains (**Supplemental Figure S3.7**). The relatedness of cotransmitted strains following single cotransmission events from infections COI = 2 had an expected relatedness greater than 0.33 while those with a COI > 3 had an expected relatedness less than 0.33.

As expected of serial cotransmission chains, we found that the relatedness of polygenomic infections increases with each transmission event (**Figure 3.7**). Transmission lineages with superinfection had lower relatedness values and smaller proportions of serial transmission simulations that converged to the transmission of single strains. The reduction in relatedness was greatest in those where the resident strain was unrelated to the parental strains of the original infection (Fig 7, purple). Changing the resident strain after each transmission event prevented the relatedness of polygenomic relatedness from increasing beyond 0.10 even after five transmission events. We also saw that the COI of the initial infection could have a lasting effect on the relatedness of polygenomic infections. Transmission lineages initiated with low COI polygenomic infections had higher relatedness values than those initiated with high COI polygenomic infections. This effect was weaker in superinfection lineages with unrelated resident strains. While skewed gametocyte-sampling ratios had a modest effect on the relatedness of polygenomic infection, it drastically increased the rate with which transmission lineages converged to the transmission of single strains for all transmission lineages except the one where unrelated resident strains were changed after each transmission event.

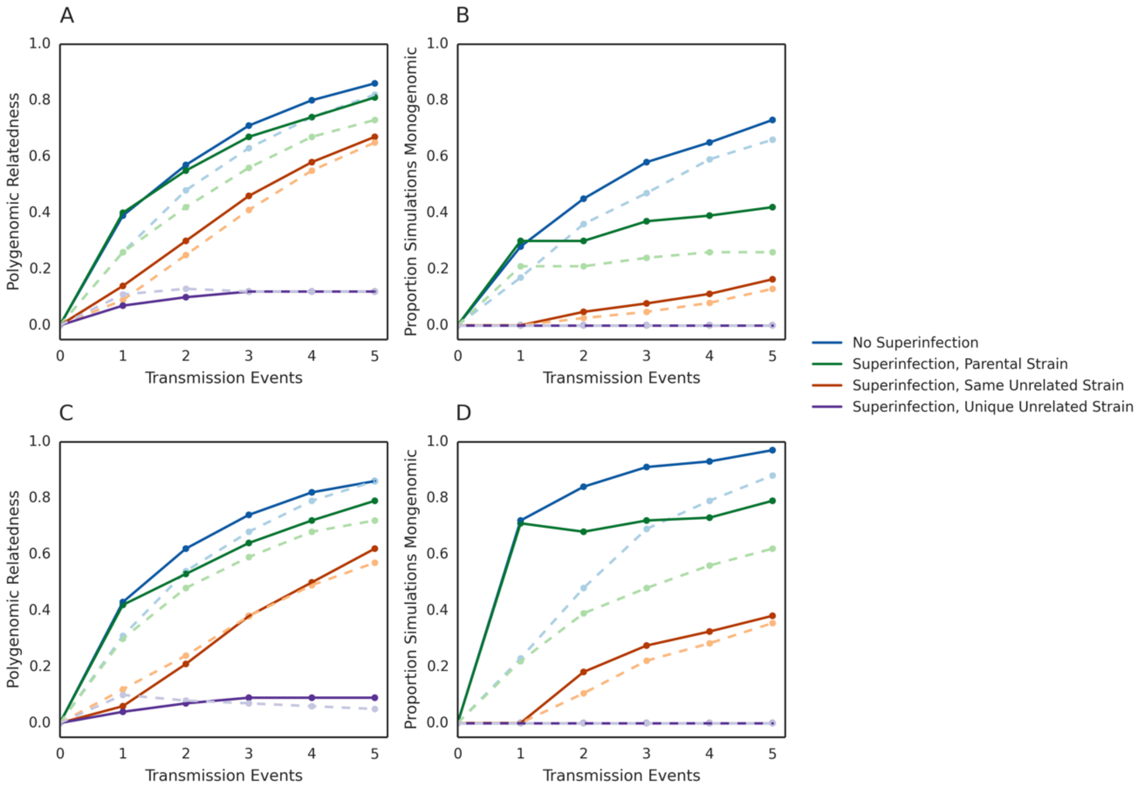


Figure 3.7 Relatedness of polygenomic infections after multiple transmission events

Line plots of the average relatedness of polygenomic infections **(A,C)** and proportion of simulations that have converged to the transmission of a single strain after multiple transmission events across 500 simulations **(B,D)**. Only

(*Figure 3.7, continued*) results from simulations where gametocyte-sampling probabilities are equal (**A,B**) and where gametocyte sampling probabilities are skewed by a 10:1 ratio between the most frequent and least frequent strain **(C,D**) are shown. Blue = No superinfection. Green = Superinfection where the resident strain is the same as one of the parental strains in the initial infection. Red = Superinfection where the resident strain is unrelated to the parental strains in the initial infection. Purple = Superinfection where the resident strain is unrelated and different in each transmission event. Solid dark lines indicate results where the initial polygenomic infection had a COI =2 and light dotted lines indicate results where the initial polygenomic infection had a COI = 5.

3.4 Discussion

Parasite strains in polygenomic infections are often genetically related, but it is unclear why there is so much variation between infections or whether the relatedness of polygenomic infections can be used to understand parasite transmission. In order to help bridge the gaps in our understanding, we developed a pedigree/kinship framework for understanding how COI and oocyst counts affect the relatedness of cotransmitted strains. We then tested the predictions of this framework using a parasite transmission model to quantify changes in the relatedness of cotransmitted strains. We demonstrated that multiple oocyst simulations in low COI conditions favor the transmission of full-siblings / parent-offspring strains and limit the transmission of half-siblings and unrelated strains, causing an increase in the expected relatedness of cotransmitted strains. Multiple oocyst simulations in high COI conditions decrease the relatedness of cotransmitted strains by favoring the transmission of half-siblings and unrelated strains. Alterations to the number of sporozoites that invade the liver have little effect on relatedness, conditioned on the fact that multiple sporozoites invade.

We also examined how non-uniform gametocyte-sampling probabilities could affect the relatedness of cotransmitted strains. Previous studies have established that intra-host parasite dynamics depend on patient age [82,83] disease severity (reviewed in [84]), and eco-epidemiological factors such as seasonal transmission [85,86]. These dynamics are strongly influenced by host immunity [87] and can fluctuate over the course of a single infection [62,83,88–90]. Furthermore, gametocyte sampling is not completely random [91] and not reliant on peripheral blood gametocyte densities at low parasitemias [85,92]. Our results show that the relatedness of cotransmitted strains is robust to variations in intra-host strain proportions and gametocyte-sampling probabilities. Even infections where the ratio of the most frequent to least frequent strain is 10:1 do not result in drastic changes to that observed from infections with even strain proportions. This suggests that the relatedness of cotransmitted strains is consistent across differences in patient-age, disease severity, and host immunity.

Our results are in agreement with the frequent assumption that cotransmission events are comprised of genetically related parasite strains [41,42,64,65]. A large fraction of simulated cotransmission events result in the transmission of genetically distinct meiotic siblings, as evidenced by the peaks at

0.33 for all simulations where oocyst counts and hepatocyte counts were randomly sampled. However, we also found that the transmission of unrelated strains is a major aspect of cotransmission. The cotransmission of unrelated strains was present in all multiple oocyst simulations and increased in frequency with COI. Polygenomic infections comprised of unrelated strains are typically assumed to be the result of superinfection, but these findings suggest that some are the result of cotransmission. Current estimates of the prevalence of cotransmission are underestimates, since they rely on the subset of cotransmission events resulting in polygenomic infections comprised of genetically related strains [64].

Our results reveal an inverse relationship between the relatedness of cotransmitted strains and COI. COI is correlated with high entomological inoculation rates [25,26] and a known genetic correlate of transmission intensity [25,26]. COI is higher in high transmission areas than in low transmission areas due to increased superinfection rates. The association between the relatedness of cotransmitted strains and COI suggests that polygenomic infections in low transmission areas are comprised of more related strains than those in high transmission areas. We previously found that the average relatedness of 32 symptomatic polygenomic patients collected from a clinic in a low transmission region of Senegal (mean COI of two) was 0.38 [64]. This value exceeds the expected relatedness of meiotic siblings and may reflect an increase in the transmission of full-siblings / parent-offspring parasites but could also result from factors such as population structure. Previous studies of genetic relatedness have focused on areas of mid-to-low transmission setting [41,42,64,65] and a comparison of genetic relatedness of polygenomic infections across transmission settings have yet to be performed. High relatedness from low COI infections could have implications for the spread of drug resistance traits in low transmission settings, as the increased relatedness could increase the chance that multi-locus drug resistant genes are passed on together to the next generation.

It remains to be seen whether the relationship between relatedness and COI can be reflected in polygenomic infections collected from natural parasite populations. If the inverse relationship between COI and relatedness holds, then the relatedness of coinfecting strains could be a potential population genetic correlate of transmission intensity. Population genetic correlates of transmission are valuable in the context of malaria control and can be used to supplement or supplant traditional epidemiological measures, which can be difficult to collect in low transmission areas [20,25]. With regards to polygenomic infections, only the frequency and COI of polygenomic infections are known to correlate with transmission intensity [25,40,55]. Other population genetic metrics, such as parasite clonality [25], currently rely on data obtained from monogenomic infections, which are limited in high transmission areas where polygenomic infections are frequent. By providing an additional source of information, genetic relatedness could increase the granularity by which we use genetic signals to monitor changes in transmission. However, spatial-temporal transmission, such as the seasonality or the existence of transmission hotspots, and host immunity can influence population genetic structure [87]. Neither of these are taken into consideration in this study, and it is unclear how these might affect polygenomic relatedness. Population-level models and epidemiological sampling will be needed to understand the effects of cotransmission and establish whether the relatedness of polygenomic infections correlates with transmission intensity.

An alternative method of dissecting population-level dynamics is to focus on the characterization of transmission lineage. Transmission lineages consist of chained transmission events and are a simplification of the transmission processes within populations. Our transmission lineages were designed to examine the effect of multiple transmission events and to examine how the co-occurrence of superinfection affects the relatedness of polygenomic infections. They show that superinfection depresses the relatedness of polygenomic infections, but also show how sensitive these lineages are to the conditions of the host infection. Strikingly, they show that cotransmission fails to increase the relatedness of polygenomic infections if each host in the transmission chain harbors a different, genetically unrelated parasite strain. They also reveal the fragility of serial cotransmission chains. In the absence of superinfection, serial cotransmission chains quickly converge to the transmission of single strains. High COI in the initial infection delays this process but a large fraction of serial cotransmission chains still converge within five transmission events. Because these transmission lineages are analogous to the introduction of a polygenomic infection to a new population, polygenomic relatedness could be useful for studying transmission in import scenarios.

In conclusion, our study uses a model of parasite transmission to provide mechanistic insight into the process of cotransmission to help understand the factors that influence the relatedness of cotransmitted strains. Understanding the effects of sexual recombination and transmission on malaria population genomics is of key public health interest in an era where parasite populations are experiencing rapid declines in transmission intensity. We believe mechanistic models such as the one used in this study reveal new insights that can be applied to the results obtained from more complicated conditions. Our model highlights the importance of COI in influencing the relatedness of cotransmitted strains, but future models and epidemiology studies are needed uncover how transmission intensity and cotransmission affects the genetic composition of strains in polygenomic infections in natural populations.

3.5 Methods and Models

3.5.1 Simulating meiosis and recombination

We simulated meiosis under two different frameworks: one with and one without obligate chiasma formation. Both frameworks sample from a constrained gamma distribution where the average distance between randomly sampled distances is 50 centimorgans to determine the location of chiasma along a bivalent [75,93]. For each placed chiasma, our meiosis model chose one sister chromatid from each homolog to undergo recombination. Sister chromatids were independently chosen for each recombination event. Once all recombination events were complete, the model independently segregated and randomly combined sister chromatids from other bivalents to create haploid parasite genomes.

For the non-obligate chiasma framework, our meiosis model placed the first chiasma 10^5 base pairs before the beginning of each chromosome. It then drew a distance, *d*, from a gamma distribution with shape = *v* and scale = 1/(2*v*) [93] to determine the location of the next chiasma. New chiasma were placed *d* units after the previous chiasma and a new distance was drawn for each chiasma. Chiasma locations were filtered to include only those that fell within the boundaries of the chromosome under consideration.

For the obligate chiasma framework, the position of the first chiasma was determined by drawing from a uniform distribution that spans the length of each chromosome. Subsequent chiasma were placed by drawing distances from a constrained gamma distribution (described in the next paragraph) and placing the next chiasma *d* units before it. This was repeated until the start of the chromosome was reached. Afterwards, the process was repeated in the other direction until the end of the chromosome was reached.

Due to the forced placement of chiasma, we could not use the formulas used in the non-obligate chiasma framework to generate appropriately constrained gamma distributions. We used an approximate Bayesian computation (ABC) Markov chain Monte Carlo (MCMC) to solve the appropriate scale parameter and shape parameters. Shape parameters varied from 1 - 9 and scale parameters were sampled from a uniform distribution with a range of 0 - 5.

For each set of scale and shape parameters, we counted the number of chiasma on a bivalant 100 centiMorgans (cM) in length and repeated this process 1000 times to estimate the average and standard deviation. We evaluated the fit of each proposed parameter using the following distance metric:

$$D' = \frac{(2-u)^2}{0.05^2 + \delta^2}$$

where u and δ are the simulated mean and standard deviations of the number of chiasma, 2 represents the desired number of chiasma per 100 centiMorgans, and 0.05 represents a small error term. We then constructed an estimate of the pseudo-likelihood as:

$$L = \frac{1}{e^{D'}}$$

The proposed scale parameter was accepted if the proposed pseudo-likelihood was greater than the pseudo-likelihood of the previously proposed parameter. If the new pseudo-likelihood was smaller, then the probability of rejection was decided by the ratio of the current pseudo-likelihood over the previous pseudo-likelihood. This process was repeated 2,500 times to form a MCMC chain. After our MCMC chain was completed, we calculated the mean of the accepted scale parameters from the last 1500 steps to serve as our estimate of the scale for each shape parameter.

3.5.2 Model selection: Non-obligate vs obligate chiasma model

We calculated the average number of crossover events and intercrossover distances for each chromosome in the genome using SNP data from 69 genetically distinct progeny generated from 3 different laboratory crosses [67,69–71]. These data were previously generated by the Pf3k project (www.malargen.net/pf3k) [67–71]. VCF files were downloaded and filtered based on the available INFO strings. We removed non-Mendelian sites, sites that did not pass the quality filters used, and sites that were invariant between the parental strains used in the cross. Samples from each laboratory cross were represented by an average of 1028 SNPs. From this filtered dataset, we performed pairwise calculations of percent similarity to identify and remove duplicate strains. Duplicate strains were defined as those having greater than 90% SNP similarity.

For each chromosome, we used a modified version of an IBD Hidden Markov Model (HMM) [40,73] to quantify the average number of crossover events and the average intercrossover distance for each chromosome. Our previously published HMM relied on population SNP frequencies to infer IBD, which is problematic when using cultured strains with vague demographic histories. For each laboratory cross, we used SNP data to infer IBD between progeny and parental strains using the following emission probabilities:

$$P(Concordance \mid IBD) = (1-\varepsilon)^2 + (\varepsilon)^2$$

$$P(Concordance \mid non-IBD) = 2\varepsilon(1-\varepsilon)$$

$$P(Discordance \mid IBD) = 2\varepsilon(1-\varepsilon)$$

$$P(Discordance \mid non-IBD) = 1-2\varepsilon(1-\varepsilon)$$

where ε refers to the rate of sequencing error, concordance refers to having the same SNP identity, discordance refers to having different SNP identities, and IBD refers to identical-by-descent.

The resulting IBD maps closely mirror the parental inheritance boundaries specified in [68], but sometimes identifies very short IBD fragments that are unlikely to be real (S5 Fig). Crossover events were identified as the points in the chromosome where the IBD map switches from IBD to non-IBD and intercrossover distance was calculated as the distance (in cM) between each of the identified crossover points. Intercrossover distances were converted to centiMorgans using the estimates reported in [67,68] (15 kb/cM). If no crossovers were observed, then the intercrossover distance was defined as the length of the entire chromosome.

We then used the average number of crossovers and intercrossover distances to determine whether a non-obligate or obligate chiasma model of meiosis would fit the data best. Each simulation was run 20 times to get an average and standard deviation of the number of crossover events and crossover distances per chromosome. We then devised a distance metric defined as:

$$D_j = \sum_{i}^{14} \frac{(u_{i,sim} - u_{i,observed})^2}{\delta^2{}_{i,sim} + \delta^2{}_{i,observed}}$$

where *u* is the mean, □is the standard deviation, *j* is the feature (number of crossover events or interarrival distance), *i* is the chromosome number, *sim* indicates the simulation result, and *observed* indicates the value observed in the 69 progeny strains. We defined a pseudo-likelihood as

$$L = \prod_{j}^{2} \frac{1}{e^{D_j}}$$

and used it to determine the model that fit the data best.

3.5.3 Model design: Modeling transmission

To quantify the average relatedness of cotransmitted strains, we developed an agent-based mosquito transmission model that simulates the sampling processes that occur as parasites enter and exit the mosquito vector and parameterized by COI, oocyst count, and infected hepatocyte count. The values for oocyst count and infected hepatocyte count were drawn from the set {1, 2, 3, 4, 5, 10, 20} while the values for COI were drawn from the set {2, 3, 4, 5, 10, 20}. Each set of parameters was run 2000 times. Each simulation was initiated by creating an initial infection comprised of unrelated parasite strains; the number of strains within the initial infection was determined by COI.

To model differences in intra-host strain proportions and differences in sampling probabilities, we assumed that strain proportions followed an exponential equation of the form:

$$f(x) = Ae^{Bx}$$

where x is a discrete variable representing each strain in the infection. We used an exponential equation to magnify the difference in frequency between the most frequent strain and the other strains present in the infection.

For an infection with COI = *n*, *x* ranges from 0 to *n -1*. We fit this equation to two points, (0, *f*(0)) and (*n - 1*, *f*(*n - 1*)), based on the ratio of the most frequent to the least frequent strain in the infection. These ratios ranged from 1:1 to 10:1, reflecting the observed strain proportions in a set of polygenomic infections collected from Thiès, Senegal (**Supplemental Figure 3.6**). *f*(0) is the ratio of the most frequent to the least frequent strain. *f*(*n - 1*) is the ratio of the final strain to the least frequent strain and always equal to one. The ratios of all other strains present in the infection was determined by *f*(1), *f*(2), ...*f*(*n*-1). We then drew from a Dirichlet distribution with a concentration parameter = {*f*(0), *f*(1), *f*(2), … *f*(*n* - 1)} 1000 times to calculate the expected frequency of each strain in the infection.

Based on the specified oocyst count, our model sampled gametocyte pairs by their intra-host strain proportions to create oocysts, allowing for multiple samplings of the same strain. Each sample pair underwent meiosis to create four meiotic products. The progeny from all the meiotic events were combined without the removal of repeat strains to represent the sporozoites within the mosquito vector. Our model assumed mating success and oocyst formation could be simulated as the random sampling of gametocytes from the human host. It is unclear whether the parasite has a preference for self-fertilization or outcrossing. Evidence for non-random mating is based on the observation of highly inbred oocysts within the mosquito midgut [94], but it is unknown to what extent self-fertilization occurs more frequently than expected by chance.

We then sampled sporozoites to represent the strains in the infected hepatocytes. Multiply-infected hepatocytes were not allowed. At this point, our model performed pairwise comparisons between all the parasites in the infected hepatocytes, regardless of whether or not the pair consisted of genetically distinct parasites, to determine the frequency of the different pedigrees specified in Fig 2. The expected relatedness of cotransmitted strains was calculated as the average pairwise relatedness between genetically distinct strains. This average is not weighted by the frequency of strains within the infected hepatocytes. Because cotransmission must result in the creation of polygenomic infections, we excluded infections where the infected hepatocytes consisted of a single strain. When an infected hepatocytes consisted of two or more genetically distinct strains, the relatedness of cotransmitted strains was calculated as the relatedness between the two strains; when an infection was comprised of 20 genetically distinct strains, the relatedness of cotransmitted strains is calculated as the average pairwise relatedness from all 20-choose-2 comparisons.

Source code is available on GitHub, under the project name Cotransmission (https://github.com/weswong/Cotransmission). The code is written using Python 2.7.0 and is platform independent.

3.5.4 Quantifying relatedness in simulated genomes

We defined relatedness as the proportion of the genome that is identical-by-descent (IBD) owing to inheritance from the same common ancestor. Because the genetic ancestry of all input strains was known and assumed to be genetically unrelated, IBD segments were identified as segments of the genome that originated from the same parental input strain.

3.5.5 Calculating the expected relatedness of the nine pedigrees

To calculate the expected relatedness of parasites described by our 9 pedigrees, we generated simulations with the appropriate number of oocysts (1 or 2), the appropriate pedigreess for each oocyst, and the appropriate method of sampling parasite pairs (within or between oocysts) for each pedigree and quantified the relatedness of a single randomly drawn parasite pair. This process was repeated 800 times to generate distributions of relatedness and to get an estimate of the mean.

3.6 Addendum

3.6.1 Author’s Contributions

EW helped create the initial transmission simulation design. WW performed all simulations and analysis. DLH and DFW supervised the project

3.6.2 Acknowledgements

We thank Hsiao-Han Chang, Angela Early, Seth Redmond, and Daniel E. Neafsey for helpful discussions and early manuscript critiques. The computations in this paper were run on the Odyssey cluster supported by the FAS Division of Science, Research Computing Group at Harvard University.

Chapter 4: Selective whole genome amplification of lab-generated mock infections

Chapter 4: Selective whole genome amplification of lab-generated mock infections[3]

Wesley Wong [a], Selina Bopp [a], Rachel F. Daniels [a,b], Amanda Lukens [a], Caroline Keroack [a], Sarah K. Volkman [a,b,c], Daniel E. Neafsey [a,b], Dyann F. Wirth [a]

Affiliations:

[a] Department of Immunology and Infectious Diseases, Harvard T. H. Chan School of Public Health, Boston, Massachusetts, 02115, U.S.A.

[b] Broad Institute, Cambridge, Massachusetts, 02142, U.S.A.

[c] School of Nursing and Health Sciences, Simmons College, Boston, Massachusetts, 02115, U.S.A

[3] The contents of this chapter have not yet been prepared for publication

4.1 Abstract

Selective whole genome amplification (SWGA) is a method of preferentially amplifying species-specific DNA from a sample with DNA from two or more species. SWGA could be particularly useful for sequencing malaria parasites, because genomes are extracted from patient blood containing a mixture of both human and parasite DNA. Determining whether SWGA has strain-specific amplification biases is important for the characterization of intra-host strain dynamics of polygenomic infections, of which little is currently known. We generated samples using four lab-cultured strains to determine the limitations of selective whole genome amplification. In the presence of human DNA, SWGA outperforms standard whole genome amplification and direct sequencing. We found no significant differences in genome coverage or average read depth associated with either the source material used or the number of strains present in the sample. These results demonstrate the potential of SWGA for whole genome sequencing analysis and demonstrate that dried filter paper material can provide equivalent genomic information as that derived from blood pellets. We also show that there is little evidence for strain-bias with SWGA and validate its use for characterizing the genomic composition of polygenomic infections.

4.2 Introduction

In *Plasmodium falciparum, a*dvances in whole genome sequencing technologies have enabled large-scale population genomic analyses that have revealed crucial insights into parasite evolution and drug resistance [22,56,18,21,19]. In the context of public health, genomic analyses are uniquely positioned to identify molecular markers of drug resistance and played a major role in identifying the genetic basis of artemisinin and piperaquine resistance [22,95,96] In recent years, genomics has also been used to track changes in transmission by monitoring changes in population genetic correlates of transmission such as parasite clonality [25,40].

Obtaining sufficient amounts of genomic DNA for whole genome sequencing can be challenging due to the overwhelming presence of contaminating host DNA and to the limited amount of DNA present in the sample. To date, the most common method of sample retrieval for whole genome sequencing is from venous blood collected from infected patients. This process is logistically demanding and requires highly trained personnel to draw and filter large volumes of blood (~5-10ml per sample). Once drawn, samples needs to be maintained at 4-8°C until genomic DNA (gDNA) extraction [97]. These samples are then run through columns to remove leukocytes. Even after leukocyte depletion, the amount of human DNA in the sample can be high, constituting anywhere from 10-50% of the DNA in the sample [97,13]. gDNA can also be extracted from dried filter paper blood spots generated during routine diagnostic surveys. Post gDNA extraction methods of depleting host DNA, such as hybrid selection, can also be used to improve parasite DNA yields [98]. Hybrid selection uses biotinylated RNA baits that bind to the *P. falciparum* genome and later pulled down using streptavidin-coated magnetic beads for whole genome sequencing. However, hybrid selection is technically demanding and requires the use and construction of costly RNA baits. Neither of these methods address the issue of low input DNA and requires a separate genomic DNA amplification step.

Recently, selective whole genome amplification (SWGA) has been used to generate high quality sequences from mixed-species DNA samples [99–101]. Unlike leukocyte depletion or hybrid selection, SWGA simultaneously addresses the issue of low input DNA and the presence of human DNA. SWGA uses primers designed to target species-specific DNA motifs present in the genome of interest but absent in the genomes of other species present in the sample. SWGA also uses a unique strand-displacing phi29 DNA polymerase. Instead of stopping when it encounters a stretch it double-stranded DNA, the phi29 polymerase displaces the complimentary strand and continues with DNA extension [99,102]. This allows simultaneous amplification of DNA in regions where primer binding is frequent.

For *P. falciparum*, SWGA has been used to generate whole genome sequences from both venous blood and dried filter paper blood spots collected from the field [100,101,103]. SWGA has also been used to successfully generate whole genome sequences from *Plasmodium vivax*, another human malaria species [104]. These studies show that SWGA is a cost-effective way of obtaining *P. falciparum* whole genome sequences from low DNA samples with significant amounts of contaminating human DNA and can be used to reliably recover whole genome sequences from both venous blood and dried filter paper blood spots.

However, it is unclear whether SWGA can accurately recover genomic information from polygenomic (multiple-strain) samples. Polygenomic infections are common in natural populations and the frequency of polygenomic infections and the complexity of infection (number of strains per infection) are known population genetic correlate of transmission intensity [25,26,105]. The intra-host strain dynamics of polygenomic infections can yield much needed insight to the nature of within-host strain composition and the effects of immune-mediated selection. Characterizing the genomic composition polygenomic infections can also reveal the dynamics of superinfection and cotransmission [64,41,65,42]. Using genomics to characterize intra-host strain dynamics would require deep sequencing coverage and need to be free of any strain-specific amplification bias. The primers used in previous *P. falciparum* SWGA studies were based off the DNA motifs in the 3D7 reference strain [103,106]. It is unclear whether this primer design favors the amplification of 3D7 over that of other *P. falciparum* strains.

The goal of this study was to confirm the results of previous SWGA studies and to determine whether SWGA could be used to reliably characterize intra-host strain dynamics. We generated a series of mock infections and dried filter paper blood spots that explore the range of parasitemias observed in real clinical infections and test how well SWGA performs compares against direct sequencing and standard whole genome amplification (WGA), which does not use the phi29 DNA polymerase or the primers designed to preferentially bind to *P. falciparum* specific DNA motifs. Using SWGA to characterize polygenomic infections could lead to better insights into how intra-host competition, host immunity, and transmission affect parasite evolution and population genetics.

4.3 Results

4.3.1 Mock Infection and sample creation

To test the efficacy of SWGA, we created a series of mock infections with varying parasitemias using four lab-adapted parasite strains: 3D7, Dd2, SenT120.11 and SenT185.10 (**Methods**). 3D7 and Dd2 are standard lab-adapted strains that have been in culture for over 30 years [107,71,108]. SenT120.11 and SenT185.10 were culture-adapted from samples collected from patients with *P. falciparum* malaria reporting to a clinic in Thiès, Senegal in 2011 and 2012 respectively. These mock infections contained red blood cells (RBC) infected with live parasites and synchronized so that all parasites were at the ring-stage. The parasitemias for these mixtures ranged from 3% to 0.003% to reflect the parasitemias observed in natural infections. For the 3% parasitemia infections, we created mock infections with and without the presence of contaminating human DNA. For the remaining parasitemias, we only created mock infections with human DNA. In terms of strain composition, we created seven genetically distinct mock infections. Four were monogenomic (one for each strain) and three were polygenomic. Our polygenomic mixtures consisted of a 1:1 mixture of 3D7 and Dd2, a 7:3 mixture of SenT120.11 and SenT185.10, and a 9:1 mixture of SenT120.11 and SenT185.10. In summary, a total of 35 mock infections were generated, seven with and 28 without human DNA (**Supplementary Table S4.1**). We generated 16 monogenomic and 19 polygenomic mock infections.

Once our mock infections were created, we extracted gDNA from the RBC pellet of spun down mock infections. We also generated a set of dried filter paper blood spots from each of the 28 mock infections with human DNA. These filter paper samples mimic the dried bloodspots collected from clinics in the field. We did not create any filter paper samples from the seven mock infections without human DNA, since such mixtures exist only in laboratory settings and the use of dried bloodspots there is less common. Once dried, we extracted genomic DNA (gDNA) from the filter paper samples. Extracted gDNA was amplified using either standard whole genome amplification (WGA) or SWGA prior to whole genome sequencing. gDNA was also directly sequenced without any pre-amplification step for a subset the RBC pellet samples. Using gDNA extracted from our 3D7 and Dd2 monogenomic cultures, we also created one polygenomic gDNA mixture mixed at a 1:1 ratio. For polygenomic infections, we performed two technical replicates of direct sequencing, WGA, and SWGA using the same gDNA. The steps leading to sample creation and gDNA extraction are described in **Figure 4.1**.

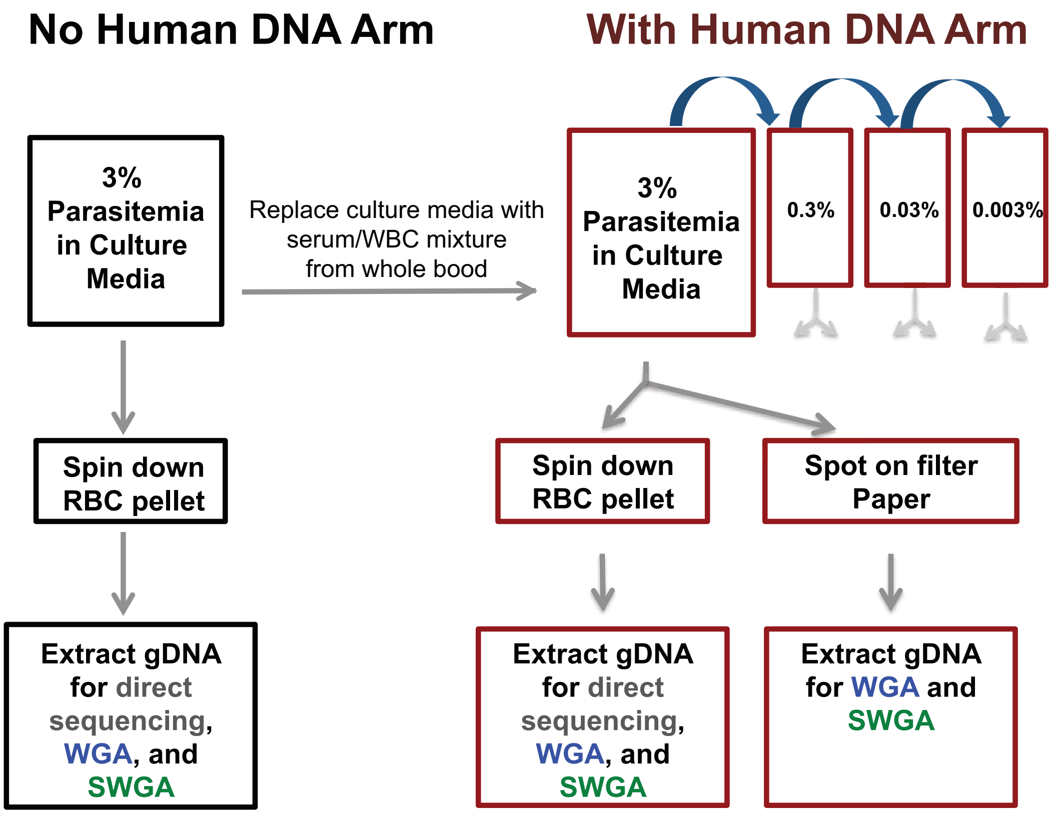

No Human DNA Arm
With Human DNA Arm
3% Parasitemia in Culture Media
Replace culture media with serum/WBC mixture from whole bood
3% Parasitemia in Culture Media
0.3%
0.03%
0.003%
Spin down RBC pellet
Spin down RBC pellet
Spot on filter Paper
Extract gDNA for direct sequencing, WGA, and SWGA
Extract gDNA for direct sequencing, WGA, and SWGA
Extract gDNA for WGA and SWGA

Figure 4.1 Mock infection and sample creation design

Mock infections were split into two arms: one with and without human DNA. We first created a 3% parasitemia mock infection with 40% hematocrit using cultured parasite material. This initial mock infection was used as the source material for all subsequent mock infections with the same strain composition. For the “no human DNA arm,” a portion of this mixture was spun down to create an RBC pellet. gDNA extracted from the RBC pellet was used for direct sequencing or amplified using either WGA or SWGA prior to sequencing. For the “with human DNA arm”, we replaced the culture media with a serum/WBC mixture obtained from spinning down uninfected non-leukocyte-depleted whole blood. We diluted this mixture using whole blood to create our lower parasitemia samples. For samples with human DNA, we created a set of dried filter paper blood spots and

(*Figure 4.1, continued*) a set of spun down RBC pellets. gDNA was extracted from both the dried filter paper blood spots and the RBC pellets. Not all mock infections were sequenced for this study. Only the 3D7 monogenomic infections were sequenced at parasitemias below 0.3%.

In summary, a total of 96 samples were submitted for whole genome sequencing (**Supplemental Table S4.1**). These samples differed based on the strain composition, parasitemia, and presence/absence of human blood in the initial mock infection. They also differed based on how gDNA was extracted (from pellet or filter paper), and the type (or lack thereof) of pre-amplification prior to whole genome sequencing. Although mock infections and dried filter bloodspots were created at all parasitemias, not all were whole genome sequenced; only the 3D7 monogenomic mock infections were sequenced at parasitemias below 0.3%.

4.3.2 Comparing sequencing quality

To examine the sequencing quality of SWGA, we quantified the total number of sequencing reads generated, the proportion of reads aligning to the 3D7 *P. falciparum* reference genome, mean coverage, proportion of genome with at least 5x coverage, and the proportion of reads that are PCR duplicates. We first examined whether SWGA performed worse than either direct sequencing or WGA in the absence of human DNA. All samples without human DNA had a parasitemia of 3% and had gDNA extracted from RBC pellet. We found no statistically significant difference between direct sequencing, WGA, and SWGA for the five metrics mentioned above (1-way ANOVA: p-value > 0. 064 for all five metrics) (**Figure 4.2A**). Regardless of pre-amplification strategy, well over 90% of the reads generated aligned to the 3D7 *P. falciparum* reference genome, with a mean coverage > 8. Despite this, only about half of the genome had at least 5x coverage. Genome-wide coverage varied greatly throughout the genome (**Figure 4.2B)**.

We next expanded our analysis to consider samples with human DNA. We found that SWGA consistently outperformed both direct sequencing and WGA (**Figure 4.3**). We also saw that SWGA resulted in a smaller number of total reads than either direct sequencing or WGA. However, a greater proportion of them aligned to the 3D7 *P. falciparum* genome. Even with SWGA, lower parasitemia samples had worse sequencing quality. When parasitemia was 0.03%, ~50% of the reads from SWGA samples aligned to the 3D7 *P. falciparum* genome. This percentage was higher than that of any of the direct sequencing or WGA samples, which remained stable at 8-10%. We found no statistically significant difference in sequencing quality associated with extracting DNA from either RBC pellet or filter paper (1-way ANOVA: p-value > 0.20 for all sequencing quality metrics).

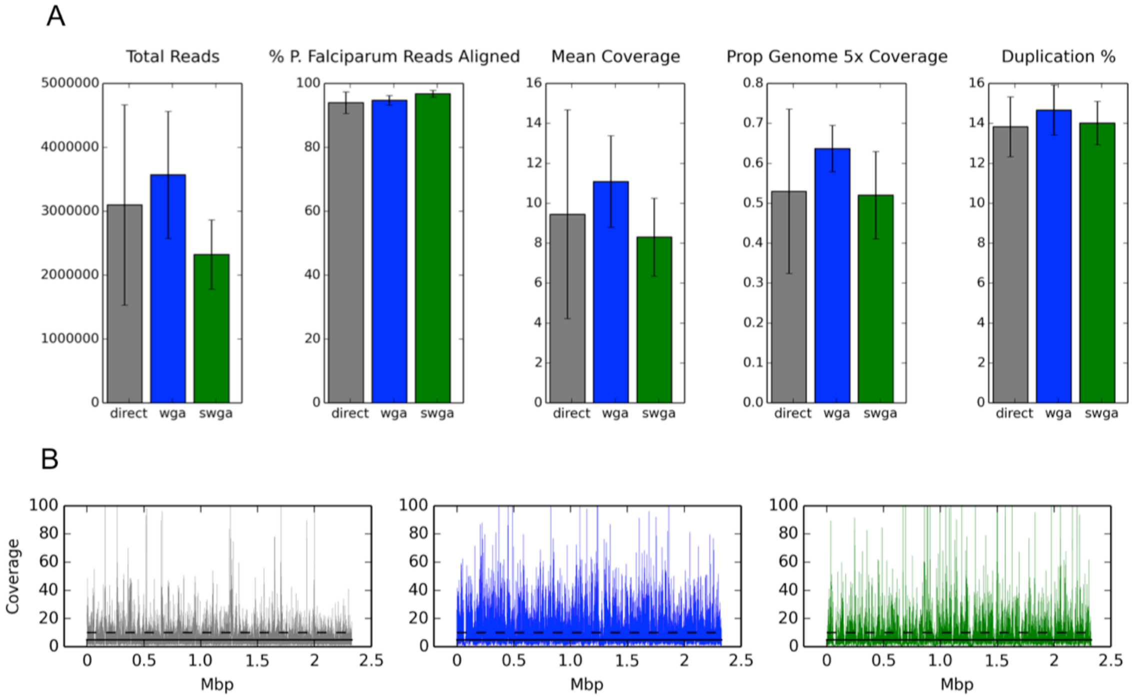


Figure 4.2 Sequencing quality of 3D7 monogenomic samples (no human DNA)

A) Barpots of the total read count, percent of reads aligning to the 3D7 *P. falciparum* genome, mean coverage, proportion genome with >5x coverage, and PCR duplication rates for all samples with 3% parasitemia and no human DNA. Colors indicate whether gDNA was directly sequenced (grey) or amplified using either WGA (blue) or SWGA (green) and error bars represent 1 standard deviation. There are no statistically significant differences in any of these 5 metrics associated with direct sequencing, WGA, or SWGA. (1-way ANOVA p-values for total reads = 0.152, percent reads aligning to the 3D7 *P. falciparum* genome = 0.064, mean coverage = 0.406, proportion genome with >5x coverage = 0.331, and PCR duplication rates= 0.507.) There were also no differences between these 5 metrics associated with the extraction of gDNA from either RBC

(*Figure 4.2, continued*) pellet or filter paper. (Student's T test p-values for total reads = 0.374, percent reads aligning to the 3D7 *P. falciparum* genome = 0.197, mean coverage = 0.206, proportion genome with >5x coverage = 0.202, and PCR duplication rate = 0.15.) **B)** Coverage plot across the entire genome.

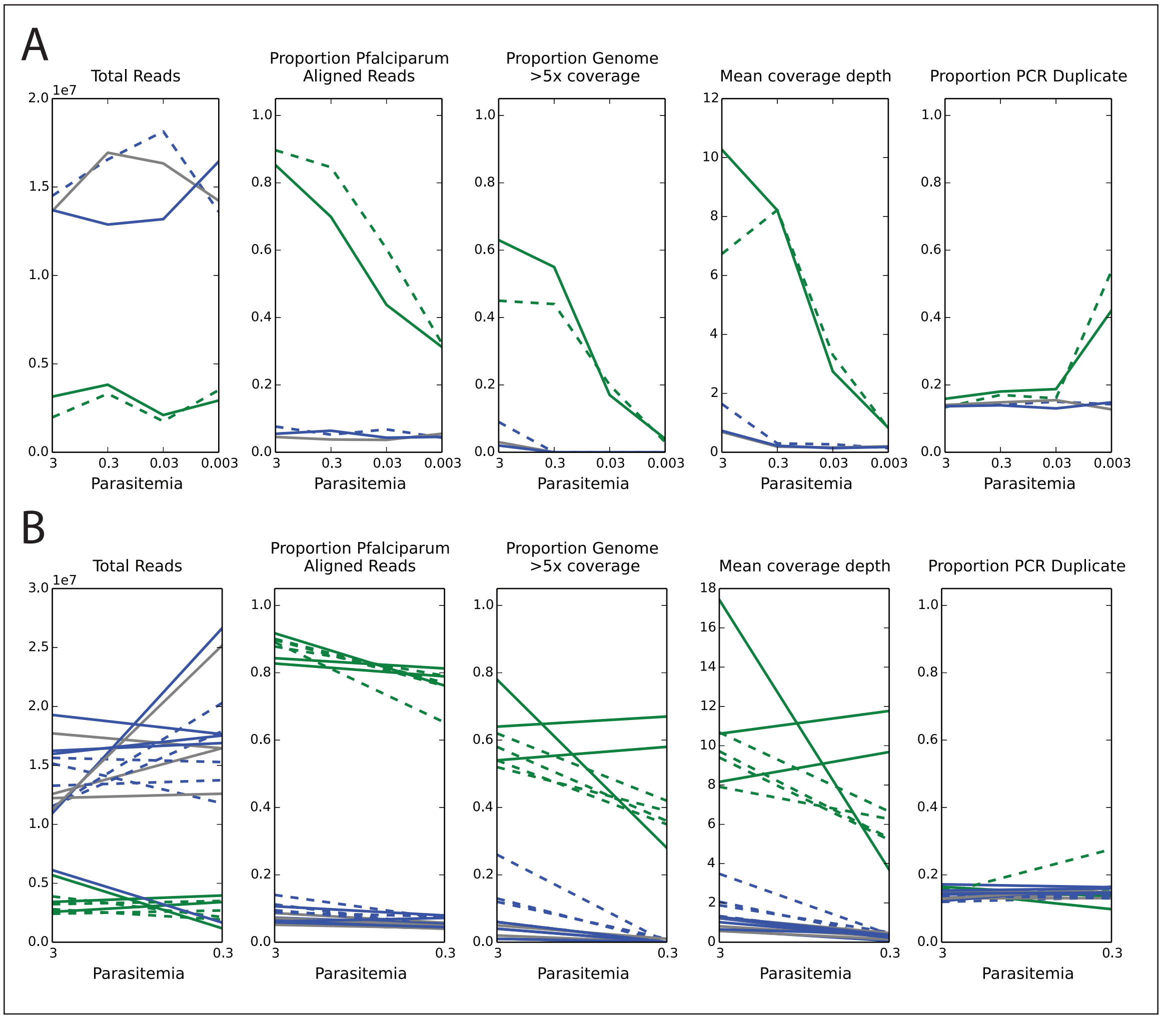


Figure 4.3 Sequencing quality of samples with human DNA

Line plots of the total read count, percent of reads aligning to the 3D7 *P. falciparum* genome, mean coverage, proportion genome with >5x coverage, and PCR duplication rates for all samples with human DNA plotted against

(Figure 4.3, continued) parasitemia. Colors indicate whether gDNA was directly sequenced (grey) or amplified using either WGA (blue) or SWGA (green). Solid lines indicate that gDNA was extracted from RBC pellet while dotted lines indicate that gDNA was extracted from dried filter paper blood spots. Only results for the 3D7 monogenomic samples (**A**) and results from the polygenomic samples (**B**) are shown. In the presence of human DNA, there was a statistically significant difference in sequence quality between direct sequencing, WGA, and SWGA samples (1-way ANOVA: p-values <1.22e-10 for all metrics, after dividing samples by parasitemia and pre-amplification strategy. Statistics were not performed on the 3D7 monogenomic samples with parasitemias ≤ 0.03% because there were only 2 WGA and SWGA samples.). There was no statistically significant difference in sequence quality associated with extracting gDNA from either RBC pellet or filter paper material (Student's T-test: p-values > 0.201 for all metrics). Samples amplified using SWGA have fewer total reads than either direct sequencing or WGA, but have a higher proportions of reads aligning to *P. falciparum* 3D7 reference genome, higher mean coverage depths, and a higher proportion of the genome with >5x coverage across all parasitemias.

4.3.3 Base Call analysis

To determine whether SWGA could be used for future population genetic analyses, we called the allelic identity of every nucleotide position in the genome using a pipeline based on the GATK best practices. Based on the SWGA samples with parasitemia of 3%, we identified a core genome where greater than 70% of samples had a non-missing base call. This core genome covers 88.5% of the entire genome and excludes subtelomeric regions and highly repetitive regions such as rifins, stevor, and var genes (**Figure 4.4, Supplemental Table S2**). The presence of human DNA or other parasite strains did not affect the boundaries of the core genome. Sequencing error rates within the core genome for SWGA were small and comparable to those from direct or post-WGA sequences (**Table 4.1**).

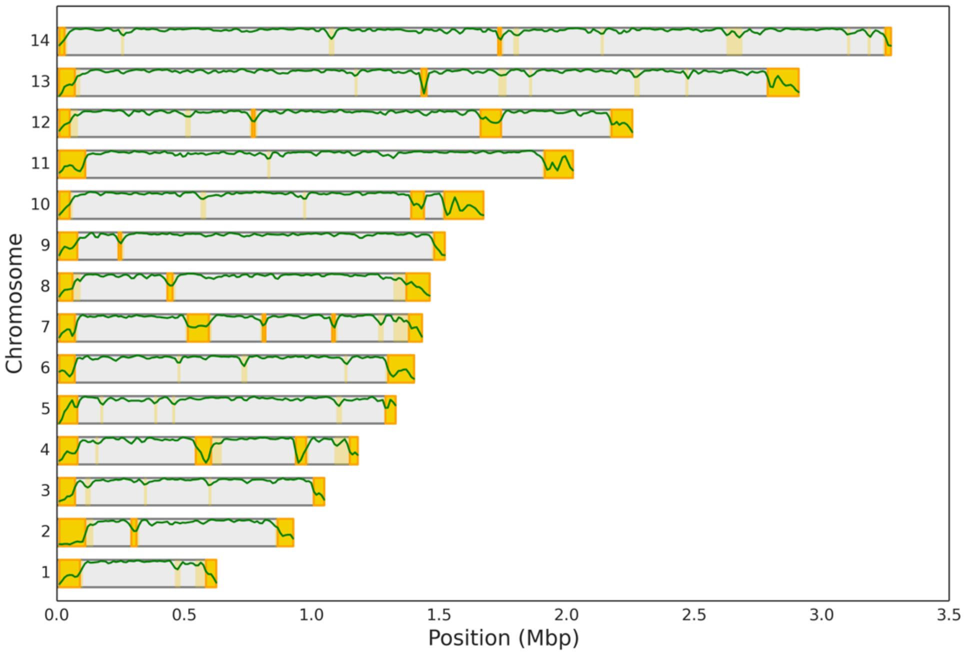


Figure 4.4 Map of the core genome of using SWGA

The core genome is defined as the region of the genome where >70% of the 3% parasitemia SWGA samples (no human DNA) have non-missing data averaged across a 1000kb window. Each of the grey blocks represents one of the 14 chromosomes in the *P. falciparum* genome. The green line plot indicates the average proportion of samples with a non-missing call averaged across a 1000kb sliding window. Dark yellow regions indicate regions where <70% of the samples have non-missing data. Light yellow regions indicate regions where <80% of the samples have non-missing data. Mitochodrial and apicoplast sequences were not examined.

Table 4.1 Sequencing error rates for all 3D7 monogenomic samples

We compared the base calls of all sites for each sample to the known 3D7 reference genome. All sites with calls different from the 3D7 reference genome were considered sequencing error. This included all erroneous SNPs as well as insertion and deletions. Sites with missing calls were not included in these calculations.

Strain	Parasitemia	Pre-Amp Method	Source Material	Human DNA	Total Positions Called	Miscalled Positions	Error Rate
3D7	3	direct	pellet	no	18503517	1089	5.89E-05
3D7	3	swga	pellet	no	18953705	1154	6.09E-05
3D7	3	wga	pellet	no	19554073	1484	7.59E-05
3D7	3	direct	pellet	yes	5328600	391	7.34E-05
3D7	3	swga	pellet	yes	19584034	1213	6.19E-05
3D7	3	wga	pellet	yes	6301208	251	3.98E-05
3D7	3	wga	filter paper	yes	9498244	566	5.96E-05
3D7	3	swga	filter paper	yes	18451951	1142	6.19E-05
3D7	0.3	direct	pellet	yes	1869074	362	0.0001937
3D7	0.3	swga	pellet	yes	19212601	1259	6.55E-05
3D7	0.3	wga	pellet	yes	1747160	274	0.0001568
3D7	0.3	swga	filter paper	yes	17939837	1297	7.23E-05
3D7	0.3	wga	filter paper	yes	2449314	568	0.0002319
3D7	0.03	direct	pellet	yes	1671329	697	0.000417
3D7	0.03	swga	pellet	yes	14181148	801	5.65E-05
3D7	0.03	wga	pellet	yes	1388040	265	0.0001909
3D7	0.03	swga	filter paper	yes	14362807	747	5.20E-05
3D7	0.03	wga	filter paper	yes	1583158	566	0.0003575
3D7	0.003	direct	pellet	yes	1670135	689	0.0004125
3D7	0.003	wga	pellet	yes	1481282	265	0.0001789
3D7	0.003	swga	pellet	yes	7151129	556	7.77E-05
3D7	0.003	wga	filter paper	yes	1299172	278	0.000214
3D7	0.003	swga	filter paper	yes	7637867	673	8.81E-05

When compared to direct sequencing and WGA, we found that SWGA has similar base call rates within the core genome for samples without human DNA (**Figure 4.5**). In the presence of human DNA, SWGA performed better than either direct sequencing or WGA. We observed no statistically significant difference in the average proportion of called sites between samples with 3% parasitemia and samples with 0.3%, regardless of the presence or absence of human DNA (1-way ANOVA, p-value = 0.64). More than 80% of the core genome was callable in our post-SWGA sequences. This percentage is similar to the percentage observed in our gDNA mixture. Once parasitemia dipped below 0.3%, there was a statistically significant decline in the proportion of callable sites compared to that of post-SWGA sequences from 3% parasitemia samples with human DNA (Student's T-test, p-value = 0.02 for the 0.03% sample and p-value = 0.0048 for the 0.003% sample). Across all pre-amplification strategies, we also noticed higher call rates in coding regions than in non-coding regions.

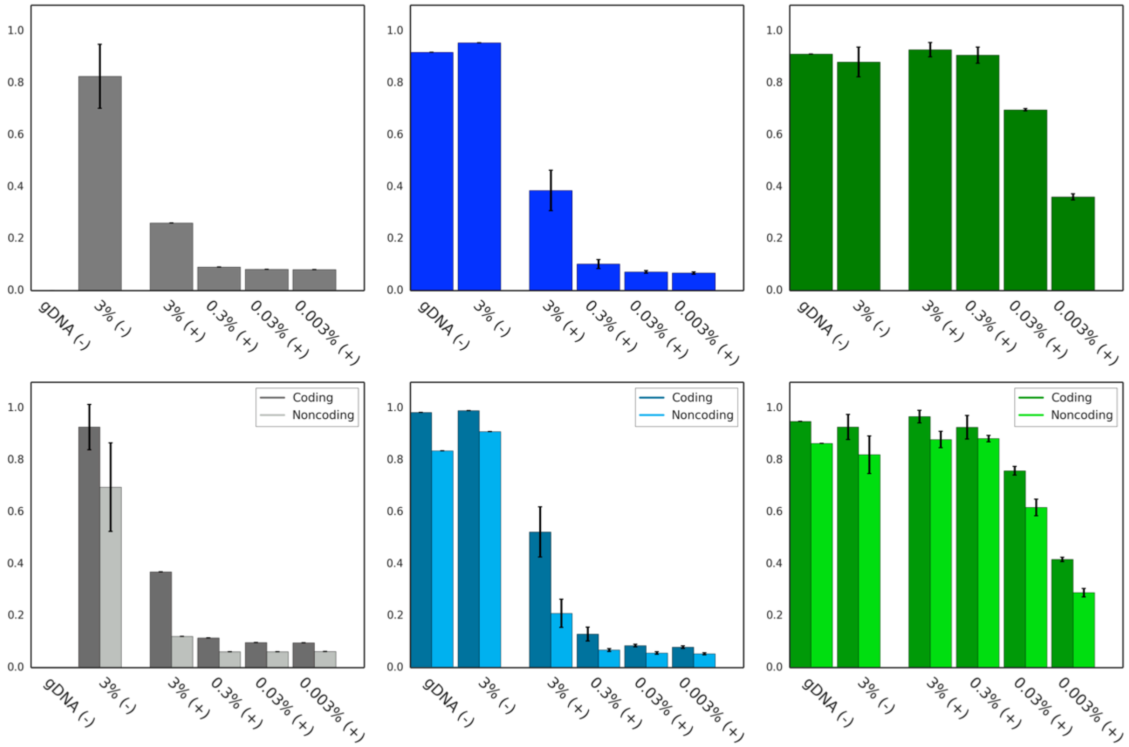


Figure 4.5 Barplots of the proportion of callable sites per sample

Colors indicate whether gDNA was directly sequenced (grey) or amplified using either WGA (blue) or SWGA (green). Minus signs indicate categories lacking human DNA while plus signs indicate categories containing human DNA. The "gDNA (-)" category refers to the polygenomic gDNA mixture obtained by mixing gDNA extracted from our 3D7 and Dd2 monogenomic samples (no human DNA, 3% parasitemia). There are no standard deviation bars for any of the direct sequencing categories with human DNA, or for WGA 3% (-) and gDNA (-) because these categories are represented by a single sample.

4.3.4 Polygenomic infections

To determine whether SWGA could be used to interrogate intra-host strain dynamics, we went back to our sequencing read data and characterized the read pileups of sites with at least 5x coverage. Because greater than 99% of sites with 5x coverage were unanimous in our polygenomic samples, we focused our analysis to variant sites. We identified these variant sites by comparing the 3D7, Dd2, SenT120.11, and SenT185.10 whole genome sequences in the Pf3K database (https://www.malariagen.net/projects/pf3k). SenT129.11and SenT185.10 were first sequenced and published in [40]. We identified 9,157 variant sites between 3D7 and Dd2 and 2,167 variant sites between SenT120.11 and SenT185.10. For samples without human DNA, our 3D7/Dd2 mixtures had an average of 5658 variant sites with > 5x while our Senegal mixtures had an average of 1664 variant sites with >5x coverage. For samples with human DNA, sequences amplified using WGA or directly sequenced had fewer than 500 variant sites with >5x coverage for our 3D7/Dd2 and Senegal polygenomic mixtures. Conversely, sequences amplified using SWGA had an average of 4197 variant sites with >5x coverage in our 3D7 mixture. For our Senegal mixtures, we found an average of 1201 variant sites and 1323 variant sites with >5x coverage for our 7:3 and 9:1 mixtures, respectively. For each of these mixtures, we found that the allele balances corresponded to the proportions with which each strain was present (**Figure 4.6)**.

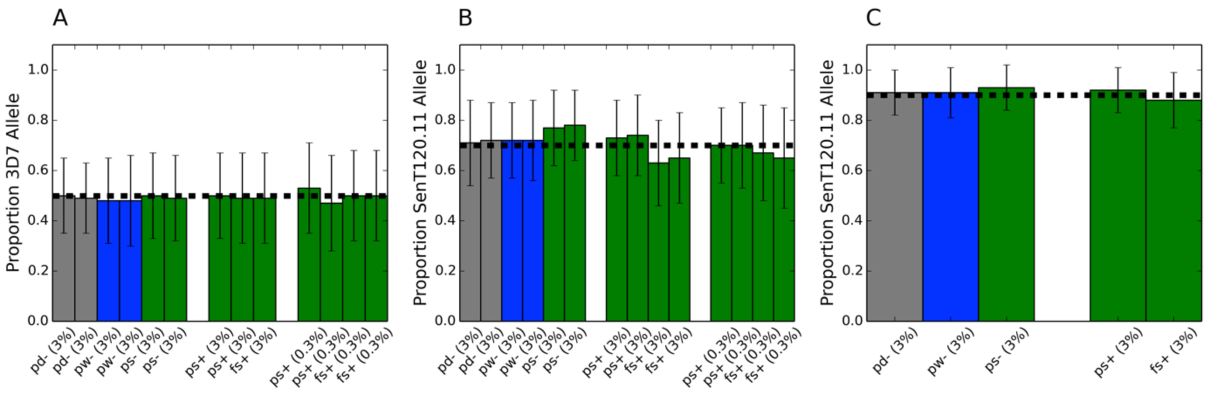


Figure 4.6 Allele proportions at variant sites

Barplots of the read pileup allele proportions at known variant sites for **A)** the 1:1 3D7/Dd2 mixture, **B)** the 7:3 SenT120.11/SenT185.10 mixture, and **C)** the 9:1 SenT120.11/SenT185.10 mixture. Only SWGA results are shown for samples with human DNA, as sequences amplified using WGA or directly sequenced had fewer then 500 sites with >5x coverage. Colors indicate whether gDNA was directly sequenced (grey) or amplified using either WGA (blue) or SWGA (green). The dotted black line represents the expected read depths for each of the polygenomic mixtures. Each bar represents an individual sample. The first letter indicates the source from which gDNA was extracted from (p = RBC pellet, f = filter paper). The second letter indicates whether gDNA was directly sequenced (d) or amplified using either WGA (w) or SWGA (s). +/- indicates the presence or absence of human blood. The percentage in parenthesis indicates the parasitemia of the mock infection. Technical replicates were not sequenced for our 9:1 Senegal mock infections.

Finally, we determined whether SWGA could be used to quantify the relatedness of co-infecting strains in polygenomic infections. Although 3D7 and Dd2 have been lab-culture for many years, they have different sampling and population genetic histories. 3D7 was first derived from a parasite line first isolated in the Netherlands but population genetic analyses show that it is genetically similar to African parasites [109]. Dd2 was derived from a parasite line isolated in Southeast Asia [107]. Based on these histories, we expected 3D7 and Dd2 to be unrelated and for their genomes to have no shared IBD regions. Conversely, SenT120.11 and SenT185.10 were isolated more recently and from the same population. Based on a panel of 3,132 SNPs that were filtered to have the least potential for sequencing error using WGA, we previously identified these strains as genetically related [64]. For this study, we wanted to test whether SWGA could be used to accurately quantify the relatedness between different strains.

We focused our analysis to a set of 11,450 SNPs that have a minor allele frequency of at least 5% in Thiès, Senegal. As expected, our HMM identified 3D7 and Dd2 as unrelated when comparing gDNA extracted from our monogenomic cultures (**Figure 4.7**). For each polygenomic sample, we constructed pseudohaplotypes (**Methods**) and used an IBD Hidden Markov Model (HMM) to quantify the relatedness between the two strains. Our pseudohaplotypes do not establish true genomic phase but maintain the order and positions of concordant and discordant sites between sequence pairs. Our HMM correctly identified our 3D7/Dd2 mixtures as unrelated when SWGA was used. When WGA or direct sequencing was used, our HMM correctly identified the mixtures as unrelated when parasitemia was 3% (with and without human DNA). However, it incorrectly identified them as genetically related when parasitemia was 0.3%.

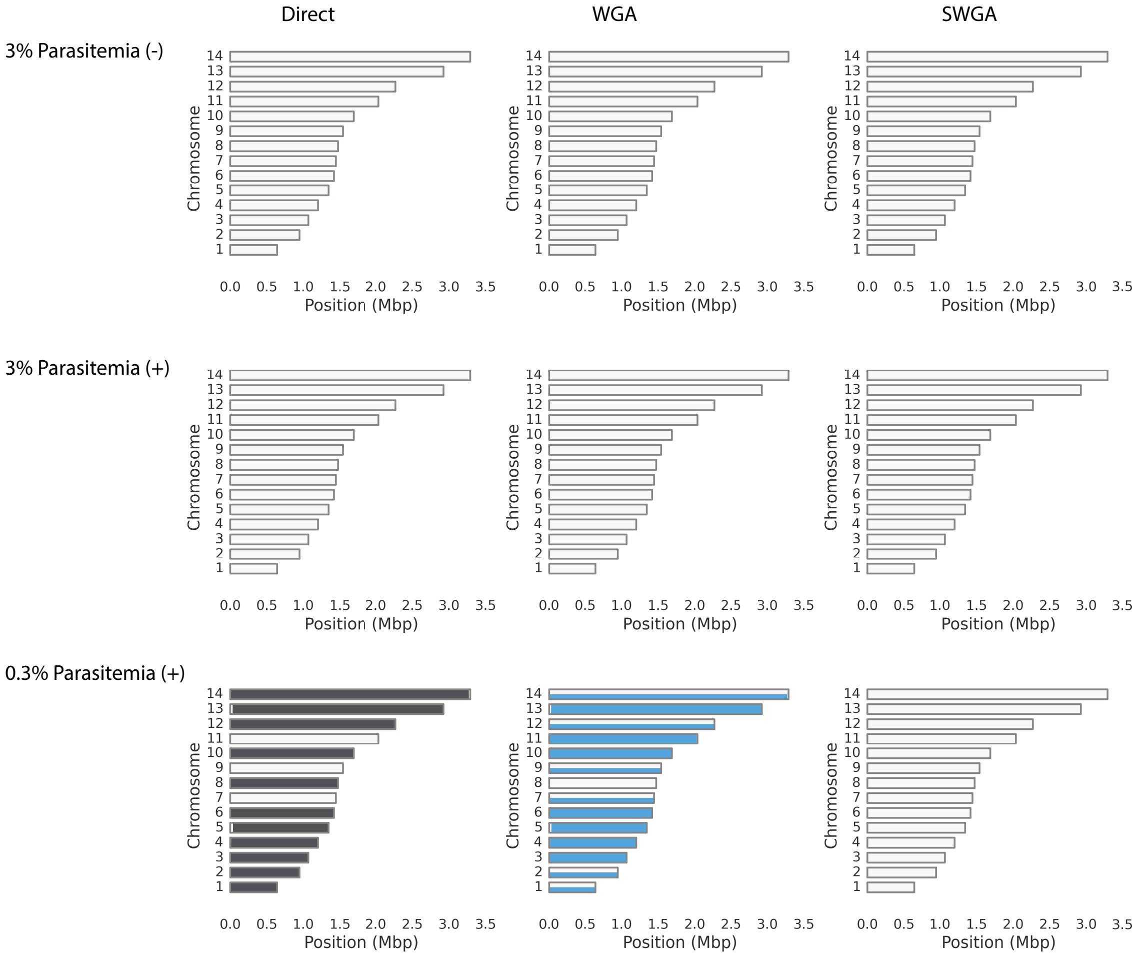


Figure 4.7 IBD maps of the relatedness of strains present in our 3D7/Dd2 mock infections.

3D7 and Dd2 are known to be unrelated to one another. Data is organized by row: **A)** 3% parasitemia without human DNA **B)** 3% parasitemia with human DNA, and **C)** 0.3% parasitemia with human DNA. Data is also organized by

(*Figure 4.7, continued*) column. From left to right, the first column shows the relatedness results from direct sequencing (dark grey), the second column shows the results from WGA (blue), and third column shows the results from SWGA (green). Each bar represents a one of the 14 chromosome in the *P. falciparum* genome and is shaded proportionally based on the number of sample pairs that are IBD at that position. Light grey shading indicates that the region is not IBD while other colors indicate that region is IBD. All three pre-amplification strategies correctly identify the sample as being comprised of unrelated strains except in the 0.3% parasitemia mock infection with human DNA. Only SWGA correctly identifies this mixture as being comprised of unrelated strains.

For our Senegal samples, we first characterized IBD between gDNA extracted from our SenT120.11 and SenT185.10 monogenomic samples. These monogenomic infections had a parasitemia of 3% and had no contaminating DNA. These sequences were directly sequenced with no pre-amplification step. The results from this comparison were treated as our gold standard and used to quantify the sensitivity and specificity of direct sequencing, WGA, and SWGA in our lab-generated polygenomic mixtures (**Figure 4.8**). In the absence of human DNA, direct sequencing, WGA, and SWGA correctly identified IBD in our 3% parasitemia polygenomic samples with high sensitivity and specificity. However, in the presence of human DNA, direct sequencing and WGA resulted in a loss in sensitivity at 3% parasitemia and a loss in specificity at 0.03% parasitemia.

Regardless of parasitemia or the presence of human DNA, SWGA had consistently high sensitivity and specificity.

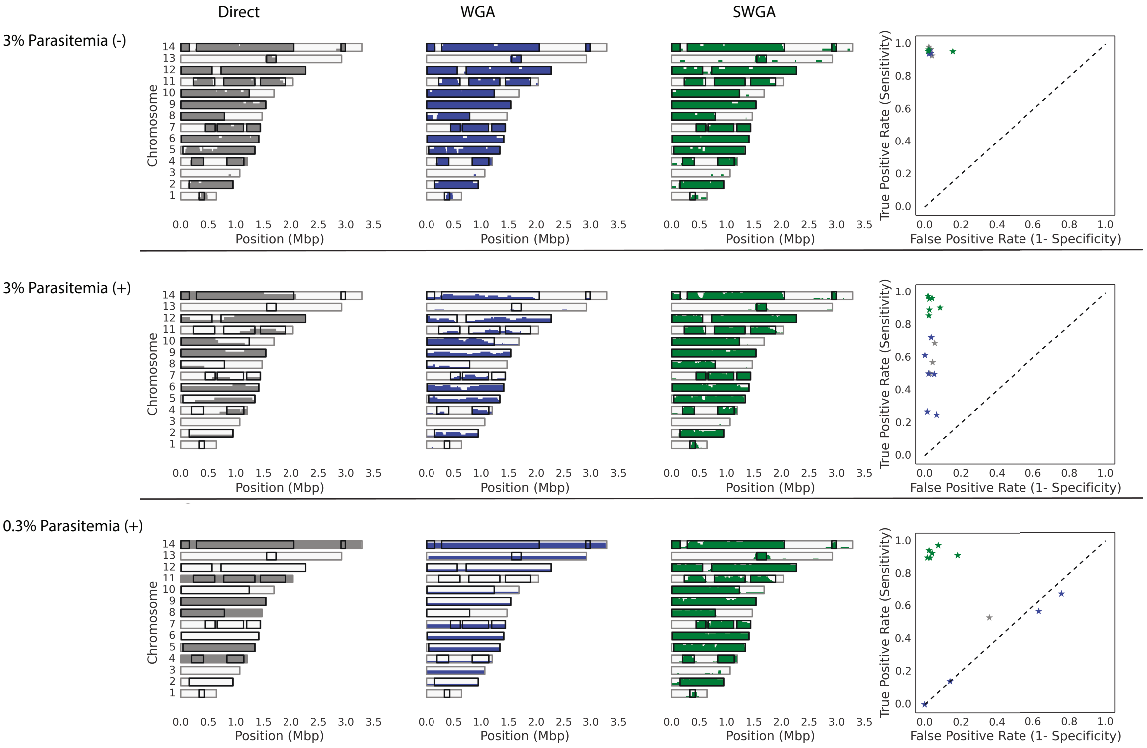


Figure 4.8 IBD maps of the relatedness of strains present in our SenT120.11 and SenT185.10 mock infections

The expected positions of IBD blocks obtained from comparing SenT120.11 and SenT185.10 sequences obtained from our 3% monogenomic mock infections (without human DNA) are outlined in black boxes. Data is organized by row: **A)** 3% parasitemia without human DNA **B)** 3% parasitemia with human DNA, and **C)** 0.3% parasitemia with human DNA. Data is also organized by column. From left to right, the first column shows the relatedness results from direct sequencing (dark grey), the second column shows the results from WGA (blue), and third

(*Figure 4.8, continued*) column shows the results from SWGA (green). Each bar represents one of the 14 chromosome in the *P. falciparum* genome and is shaded proportionally based on the number of sample pairs that are IBD at that position. Light grey shading indicates that the region is not IBD while other colors indicate that region is IBD. The fourth column is an ROC plot of the true positive and false positive rates for each sample. Each dot represents a different sample and the color indicates whether it was directly sequenced (grey) or amplified using either WGA (blue) or SWGA (green).

4.4 Discussion

Here, we generated a series of mock infections to test the limitations of SWGA and to determine whether SWGA could be used to interrogate intra-host parasite variation in polygenomic samples. Our study compared the quality of whole genome sequences that were either directly sequenced or pre-amplified using either WGA or SWGA prior to sequencing. We found that SWGA performs equivalently to WGA and direct sequencing for samples lacking human DNA. For samples with human DNA, SWGA performs significantly better, even though it generates a smaller number of total reads. The sequences generated by SWGA are of high quality and can be used to call the positions of > 88% of the *P. falciparum* genome with low error rate. The areas of the genome that cannot be called are either sub-telomeric or highly repetitive, which are known to make sequencing difficult. Our results also show that dried filter paper blood spots can be used a source of genomic DNA. We observed no difference in sequencing quality associated with extracting gDNA from either RBC pellet or from dried filter paper material. These results confirm the results of a previous study [106] and show that dried filter paper blood spots are a potential source of parasite genomic DNA. Filter paper blood spots have less stringent storage requirements than blood draws and more easily maintained. Whole genome sequencing analyses would benefit from using SWGA to amplify gDNA from dried filter paper blood spots and expand the areas where population genomic analyses are amenable without the need for complex techniques such as hybrid-selection or leukocyte depletion.

Despite this, our results show that sequencing quality declines in low parasitemia samples. Lower parasitemia samples consistently had lower coverage and average read depths than those of higher parasitemia samples, and we found that samples with lower than 0.3% parasitemias resulted in subpar whole genome sequences. Compared to our 3% parasitemia samples, we observed a small decline in sequencing quality based on the percent reads aligning to the genome, the mean coverage, and the proportion of the genome with >5x coverage. However, these declines were not enough to affect base calling, and there was little difference in the number of callable bases or in the sequencing error rate. The SWGA results from our 0.3% parasitemia samples closely matched the results obtained from our pure gDNA mixtures. These results suggest that SWGA can be used to generate high quality whole genome sequences for all infections with at least 0.3% parasitemia. This concentration falls within the range of typical of symptomatic, clinical malaria cases. However, we observed a sharp decline in sequencing quality for samples with parasitemias at or below 0.03%. Previous studies have also reported a declining relationship between sequencing quality and parasitemia, but identified 0.003% (~150 parasites/ul) as the cut-off for which sequencing becomes unreliable. Our threshold is higher because mock infections were synchronized to early ring-stages and have an average of one genome per RBC. The threshold in other studies is lower due to the presence of late-stage parasite stages in their samples, which increases the average number of genomes per RBC.

Regardless of the cut-off, these results suggest that a single round of SWGA will be insufficient to recover enough genomic material from subpatent infections for whole genome sequencing. Diagnostic tests based on light microscopy can detect parasites at 10-100 parasites per ul while PCR-based assays can detect parasites at concentrations of 0.05-10 parasites/ul [110]. Despite their low densities, submicroscopic infections are still infectious and can contribute to a significant proportion of mosquito infections in the population [111]. This can be problematic in low transmission areas, where it has been hypothesized that infections are on average older and more likely to have parasite densities undetectable using light microscopy [112]. It may also be difficult to sequence parasites from asymptomatic or chronic infections, which oftentimes have subpatent parasitemias [110]. One way to improve the quality of post-SWGA sequences from low parasitemia samples may be to perform multiple amplification reactions. However, we advise that each amplification step use different primer sets due to the high PCR duplication rates observed in our

0.0003% parasitemia samples. Alternatively, increasing the amount of input DNA used for SWGA or increasing the number of lanes used for sequencing could improve yields as well.

Finally, our study found little evidence of strain-amplification bias associated SWGA, suggesting that it can be used to characterize the intra-host strain dynamics within polygenomic infections. Polygenomic infections are common in malaria endemic areas, but can be challenging to study. The use of genomics for understanding intra-host strain dynamics could hinge on methods such as SWGA to amplify gDNA from whole genome sequencing. Intra-host dynamics are an important aspect of malaria biology crucial to our understanding of both disease progression and parasite evolution. These dynamics are a reflection of both within-host parasite competition and host immune selection, but our understanding is still incomplete. It is known that strain proportions vary throughout the course of natural infections, but little is known of what drives these fluctuations or whether host factors such as age or immunity affect them. Our study also found that whole genome sequences generated using SWGA could be used to accurately quantify the genetic relatedness of strains within polygenomic infection with high sensitivity and specificity. The relatedness of coinfecting strains within polygenomic infections can be used to characterize cotransmission and superinfection events in natural parasite populations and help us better understand parasite transmission.

In conclusion, our study shows that SWGA can be used to amplify gDNA for whole genome sequencing. These sequences are of high quality, and unaffected by the presence or absence of human DNA or the type of material used to extract gDNA. SWGA is a cheap, cost-effective method of generating whole genome sequences from unprocessed blood samples and can be used to study the intra-host variation within polygenomic infections. SWGA has the potential to increase the number and types of samples amenable to population genomic analysis, thereby improving our understanding of malaria evolution and the use of genomics for public health interventions.

4.5 Methods

4.5.1 Sample collection and culture maintenance

SenT120.11 and SenT185.10 were previously collected from individuals after recruitment and written consent of either the subject or a parent/guardian [40]. This protocol was reviewed and approved by the ethical committees of the Senegal Ministry of Health (Senegal) and the Harvard School of Public Health (16330, 2008) for Senegalese subjects. SenT120.11 and SenT185.10 were obtained through passive case detection from patients over 12 years of age reporting to clinic with acute fever and with no reported history of antimalarial use for suspected malaria. SenT185.10 was collected sometime between September and December of 2010 while SenT120.11 was collected sometime between September and December of 2011. SenT185.10 and SenT120.11 were culture adapted by thawing cryopreserved material containing infected RBCs that had been mixed with glycerolyte. These samples were confirmed to be monogenomic using a 24-SNP barcode [51].

All parasite cultures were maintained in fresh, leukocyte-depleted human blood (O+) and Hepes buffered RPMI containing 10% O+ human serum (heat inactivated and pooled). Cultures were placed in modular incubators and gassed with 1% O2/5% CO2/balance N2 gas and incubated with rotation (50rpm) in a 37°C incubator.

4.5.2 Mock infection creation

3D7, Dd2, SenT120.11, and SenT185.10 monogenomic cultures were synchronized to 0-6 hour rings and grown until parasitemia exceeded 3%. Thick smears were checked prior to mock infection creation to check parasitemia and to ensure that parasites were synchronized. Parasitemias were also verified using a SYBR green FACS assay described in [113]. Briefly, parasites were stained in 10x SYBR Green I in 1x BS for 30 min in the dark at 37°C. Cells were washed and resuspended in 5x the initial volume of PBS used in the assay. FACS data acquisition was performed on a MACSQuant VYB (Milteni Biotec) with a 488nm laser and a 525 nm filter and analyzed with FlowJo 2. RBCs were gated on the forward light scatter and side scatter. Infected RBCS were detected in channel B1. At least 100,000 events were analyzed for each sample.

Once all the monogenomic cultures had a parasitemia of at least 3%, cultures were spun down to separate the RBC pellet. Culture media was removed and packed, uninfected RBCs were added to normalize the parasitemias of all the cultures to 3%. Polygenomic infections were created by mixing the infected RBC pellets from the corresponding monogenomic cultures at the appropriate ratios. A portion of the 3% parasitemia RBC pellet was used to generate mock infections without human by adding hepes buffered RPMI media to the infected RBC pellet until hematocrit was 40% hematocrit. Lower parasitemia samples were not created using this mixture. The remainder of the 3% parasitemia RBC pellet was combined with serum/WBC mixture obtained by mixing the serum and buffy coat from non-leukocyte depleted whole blood to create a 40% hematocrit mock inection. This serum/buffy coat mixture was obtained from separated whole blood. To create lower parasitemia mock infections, we diluted this mixture using non-leukocyte depleted whole blood obtained from the same patient. We did not correct for any differences in hematocrit between the whole patient blood (which was estimated to be ~40% hematocrit by sample volume) and our 3% parasitemia 40% hematocrit mock infections.

Once each mock infection was created, we spotted 500ul of each mock infection onto Whattman filter papers to create filter paper blood spots. Dried filter paper blood spots were not created for the 3% parasitemia mock infection diluted in hepes buffered RPMI media. Dried filter paper blood spots were allowed to dry overnight before storage in boxes filled with dessicant. The remainder of each mock infection was frozen in equal volumes of glycerolyte solution at -80°C until gDNA extraction.

4.5.3 DNA extraction and pre-amplification

gDNA was extracted from filter papers using a Promega DNA IQ Casework Pro kit for Maxwell 16 (Promega Corp., Madison, WI, USA). gDNA was extracted from culture-adapted material using the QIAamp DNA Blood Minikit. Both kits are commercially available and extractions were done according to manufacture specification. gDNA was amplified with WGA using a commercially available WGA kit. gDNA was amplified using SWGA following a protocol obtained from the Wellcome Trust Sanger Institiute and used in [106]. Briefly, 40ng of extracted gDNA is added to a 30ul reaction containing the phi29 polymerase and SWGA primers (**Supplemental Table S4.3)**. The reaction is run in a PCR machine that cycles between 35°C for 5 min, 34°C for 10 min, 33°C for 15 min, 32°C for 20 min, 30°C for 16 hours, and 65°C for 15 min (heat-inactivation of phi29 polymerase).

4.5.4 Sequencing and sequencing analysis

DNA was submitted to the Broad Institute for next generation Illumina short-read sequencing. A total of 1ug of DNA obtained after amplifying gDNA with WGA and SWGA were used to contruct libraries. 1ug of unamplified gDNA was submitted for direct library construction and sequencing. Illumina libraries were constructed using commericially available Nextera XT Sample Prep kits (Illumina, San Diego, CA, USA) and sequenced on an Illumina Hiseq 2000 (Illumina, San Diego, CA, USA). Reads were aligned to the *P. falciparum* 3D7 reference assembly (PlasmoDb v 7.1) using a combination of Burrows-Wheeler

Aligner (version 0.5.9-r16)[40] and PicardTools (v2.14.0). Base calls were determined using the GATK Unified Genotyper based on a pipeline that follows GATK best practices for variant calling [41]. Variant quality score recalibration (VQSR) was not run on these samples because there were an insufficient number of variant sites between 3D7, Dd2, SenT120.11, and SenT185.10. Downstream sequencing analyses were performed using SAMtools (v1.3), VCFtools (v0.1.14), BCFtools (v1.5), and a set of custom python scripts.

4.5.5 Quantifying relatedness

The relatedness between two strains was calculated as the proportion of the genome inherited from the same ancestor, or identical by descent. Relatedness was calculated from 11,450 SNPs that have a minor allele frequency in of at least 5% in Senegal. These SNPs were used in hidden Markov model described in [26] to identify IBD blocks and quantity genetic relatedness. The eps parameter (sequencing error rate) in the HMM was set to 0.005. Specificity and sensitivity were calculated by quantifying the number of sites that were correctly/incorrectly identified as being IBD and the number of sites that were correctly/incorrectly identified as not being IBD. Sites that were expected to be IBD were based on the IBD between the whole genome sequences of 3D7 vs Dd2 and SenT185.10 and SenT120.11 obtained form our 3% parasitemia monogenomic samples. These samples did not have any human DNA and were directly sequenced without pre-amplification.

4.6 Addendum

4.6.1 Author's Contributions

CK, SKV, AL, and SB participated in the culturing and culture-adaptation of parasite strains. SB created all mock infection samples, extracted gDNA from RBC pellet samples, and performed all amplification reactions prior to library construction. SKV extracted gDNA from filter paper samples and helped with sample preparation. Project design and all genomic analyses were performed by WW. SKV, DEN and DFW guided and supervised the project.

4.6.2 Acknowledgements

The computations in this paper were run on the Odyssey cluster supported by the FAS Division of Science, Research Computing Group at Harvard University

Chapter 5: Modeling the effects of coinfection and transmission topology on malaria population genetics[4]

Wesley Wong[1], Edward A. Wenger[2], Daniel L. Hartl[1,3], Dyann F. Wirth[1,4*]

Affiliations

[1] Department of Immunology and Infectious Diseases, Harvard T. H. Chan School of Public Health, Boston, Massachusetts, United States of America

[2] Institute for Disease Modeling, Bellevue, Washington, United States of America

[3] Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, Massachusetts, United States of America

[4] Broad Institute, Cambridge, Massachusetts, United States of America

* Corresponding author

E-mail: dfwirth@hsph.harvard.edu (DFW)

[4] This chapter has not yet been prepared for publication

5.1 Abstract

Renewed interest in malaria eradication has emphasized the need to accurately monitor changes in malaria transmission. Genetic epidemiology models provide a framework for predicting how these signals react to changes in transmission, but lack defined transmission topologies and assume that infected individuals are equally likely to transmit to any other member of the population. Previous models of infectious disease evolutionary dynamics are insufficient for malaria because they ignore coinfection or focus on asexually reproducing organisms. Here, we performed evolutionary invasion analyses with and without coinfection on networks representing highly clustered populations or populations containing super-spreaders. Our results show that clustered transmission makes the parasite population more resilient to the invasion by a newly arrived mutation but that superspreading has no effect. These results highlight the importance of incorporating different transmission structures in genetic epidemiology models and contribute to our understanding of the emergence of new beneficial mutations such as those involved with drug resistance.

5.2 Introduction

Renewed interest in malaria eradication and improvements in sequencing technology have increased the number of studies applying population genomics to understand trends genetic diversity, tracking the emergence and spread of drug resistance, and for monitoring changes in transmission intensity [19,20,25,114]. These studies have led to the discovery of new population genetic correlates of transmission intensity and are of significant public health value. This is particularly true in low transmission regions where standard epidemiological measures are difficult to collect [115]. The application of population genetics for understanding malaria transmission and epidemiology promises to be a fruitful endeavor and is critical for understanding how public health interventions affect parasite populations. Genetic epidemiology models integrate population genetics concepts into traditional epidemiology models, which are strain agnostic. We previously used a genetic epidemiology model to show that parasite clonality, polygenomic (multiple-strain) infection frequency, and complexity of infection (number of strains/ infection) track with declining transmission intensities [40].

However, genetic epidemiology models typically assume transmission is well-mixed. This is not the case in natural populations, where transmission differs by geographic region and is clustered by household [58,116–118]. Mosquito biting exposure is also heterogeneous, with a small minority of individuals exposed to a large number of bites [118,119]. Previous epidemiology models have stressed the importance of contact structure and transmission topology in promoting or suppressing the spread of epidemics [120,121]. Differences in transmission structures also affect parasite population genetics and evolution [64], but it is unclear how it could affect malaria population genetics. One way of understanding how transmission structure affects population genetics and evolutionary dynamics is by performing evolutionary invasion analyses [122]. Evolutionary invasion analyses are used to understand the long-term consequences of mutations with an “invasion fitness” which quantifies the growth rate of a novel rare variant. In the context of infectious diseases, the invasion fitness is usually defined as R_0, the number of new cases per infection in an infinitely large population [36]. A recent study by Leventhal et al. showed that superspreading impedes the fixation of newly derived mutations [123].

However, current evolutionary invasion models are not suitable for understanding malaria evolutionary dynamics. Evolutionary invasion models do not allow coinfection (the infection of multiple strains in the same individual) or focus on asexually reproducing organisms. This limits their usefulness in organisms like malaria, which undergoes sexual reproduction during every transmission event. In malaria, cotransmission (infection of two or more strains from a single mosquito bite) is common, as evidenced by the genetic relatedness of polygenomic (multiple strain) infections in natural parasite populations [42,64]. How then, would evolutionary invasion models behave in the presence of coinfection and cotransmission?

Here, we performed evolutionary invasion analyses on populations whose transmission topologies are represented by one of four idealized networks.

Unlike previous studies, our model allows for both coinfection and cotransmission. These four idealized networks (complete, random regular, Barabasi scale free, and a ring lattice) represent different aspects of transmission in natural populations. Complete networks are equivalent to the well-mixed populations used in most epidemiology and population genetic models. Random regular networks limit the number of transmission routes per individual; each infection can only infect a limited number of individuals. Barabasi scale free networks have exposure heterogeneity and can be used to model superspreading. Finally, ring lattices represent the most extreme form of clustering, and do not allow transmission between distal nodes. While these idealized networks do not recapitaulate the transmission topologies of actual populations, they allow us to separate quantifiable network metrics such as the degree heterogeneity (the variation in edges per node) and the clustering coefficient (the extent to which nodes in a graph cluster together). Because all networks can be characterized by these metrics, understanding the individual effects of each will be beneficial for understanding malaria evolutionary dynamics in more complicated transmission topologies.

How infections spread through a population is just as important as the rate with which it spreads. In fact, changes in transmission intensity can result in changes in transmission transmission topology. Evidence of hotspots and clustered transmission are more frequent in low transmission areas [58]. Networks are a useful way of simulating transmission topologies because they can be summarized by metrics such as the degree heterogeneity and clustering coefficient. Understanding the consequences of structured transmission on parasite population genetics is of key public health importance, especially as population genomic metrics such as identity-by-descent are being used understand patterns in parasite movement and migration [124].

5.3 Results

5.3.1 Model Description

To model the effect of transmission topology on malaria population genetics, we performed evolutionary invasion analyses on four different idealized networks (**Figure 5.1**). Evolutionary invasion analyses were initiated by simulating single-strain epidemics on networks with 200 individuals. Throughout this study, we refer to the initial epidemic strain as the resident strain. During transmission, our model samples gametocytes from infected individuals. Parasite sexual reproduce to create recombinant sporozoites that are injected into the recipient host. Gametocyte sampling probabilities are based on the intra-host parasite densities within infected individuals. Once recombinant sporozoites were created, we randomly chose an individual connected by an edge to the source infection to receive recombinant sporozoites.

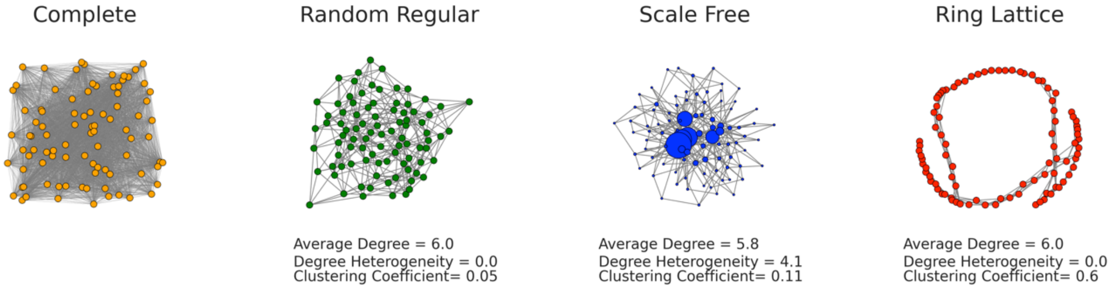

Complete
Random Regular
Scale Free
Ring Lattice
Average Degree = 6.0
Degree Heterogeneity = 0.0
Clustering Coefficient= 0.05
Average Degree = 5.8
Degree Heterogeneity = 4.1
Clustering Coefficient= 0.11
Average Degree = 6.0
Degree Heterogeneity = 0.0
Clustering Coefficient= 0.6

Figure 5.1 Networks

Representations of the four networks used in this study. Each of the four networks can be described by the mean degree (average number of edges per node), the degree heterogeneity (the standard deviation in the number of edges per node), and its clustering coefficient (a measure of how clustered interconnected nodes are). Complete networks (Orange) are equivalent to completely mixed populations where transmission random. Nodes in random regular (green), scale free, (blue) and ring lattice networks (red) have a limited number of edges. Edges in random regular and scale free are randomly drawn while edges in ring lattices are preferentially drawn to neighboring nodes. Scale free networks are characterized by high degree heterogeneity and used to represent superspreading. The degree heterogeneity of Barabasi scale free networks follows a Power distribution. Ring lattice networks have a high clustering coefficient and are used to represent highly clustered transmission, such as in transmission hotspots. In this figure, node sizes are weighted by the number of connecting edges.

Our model allows both superinfection and cotransmissions but does not allow infected individuals to be reinfected by concurrent strains. Susceptibility is strain-specific and there is no resistant or recovered class in our model. Individuals can be reinfected with previous strains once it has been cleared. Biologically, this means that host immunity is strain specific but not long-lasting. Strains within each infection have a maximum duration time that is drawn from a distribution reflecting the duration times of untreated malaria infections [83,125]. For polygenomic infections, our model assumes that asexual parasite and gametocyte densities are independently determined for each coinfecting strain. We also assume there is no competition between coinfecting strains.

5.3.2 Prevalence in single-strain epidemics

We first examined the steady-state prevalences of single-strain epidemics run on different networks (**Figure 5.2A**). We define prevalence as the proportion of infected individuals within the population. Under this definition, polygenomic infections are treated the same as monogenomic infections. As expected, different transmission topologies supported different prevalence values. Random regular, scale free, and ring lattice networks had smaller prevalences than the complete network. Of these, we saw that scale free networks had the smallest steady-state prevalence values. These networks also differed in how uninfected and infected individuals were connected to one another. One property of epidemics in ring lattice networks is that uninfected individuals tend to be clustered; uninfected individuals in these networks are more likely to be connected with other uninfected individuals. Our simulation results agree with this assertion. A large proportion of uninfected individuals in ring lattice are only connected to other uninfected individuals (**Supplemental Figure S1**). This proportion is much larger than that of our other idealized networks.

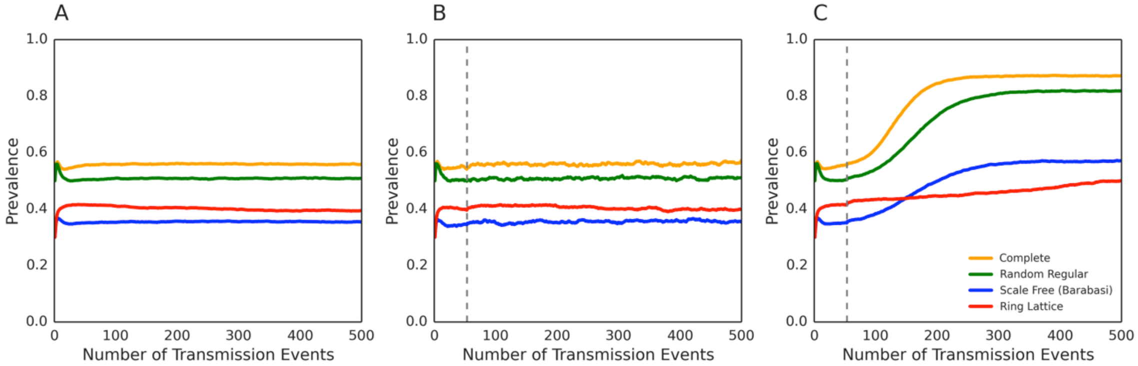


Figure 5.2 Prevalence curves

Line plots of the prevalence in our simulations, conditioned on successful invasion. **A**) Prevalence in simulations without invasion. **B)** Prevalence in our invasion simulations where coinfection is not allowed (traditional model) and **C**) in our invasion simulations where coinfection is allowed (malaria coinfection model). The grey dotted line in **B** and **C** marks the point in time when importation occurred. Orange lines are the prevalence values of epidemics run on complete networks, green on random regular networks, blue on scale free networks (Barabasi) and red on ring lattices.

5.3.3 Coinfection increases post importation steady-state prevalence and importation success

To simulate importation, we randomly infected an uninfected individual with a second strain after the resident strain epidemic reached steady-state prevalence. Traditional evolutionary invasion analyses do not allow either coinfection or cotransmission. To determine how evolutionary invasion models behave once these aspects of malaria biology are allowed, we considered two frameworks: one with and one without coinfection. Both simulations allow sexual recombination, but only simulations with coinfection allow outcrossing and effective recombination. Here, we refer to simulations without coinfection as our "traditional" model and simulations with coinfection and cotransmission as our malaria coinfection model.

We found that coinfection and cotransmission has a substantial impact on both steady-state prevalences and on importation success probabilities. In our traditional model, the importation of a second strain had no effect on steady-state prevalence and was the same as that observed in our single-strain epidemics (**Figure 5.2B**). In our malaria coinfection model, prevalence rose after importation and continued to rise until a new steady-state equillibrium was established (**Figure 5.2C**). Of the networks examined, we found that ring lattices had the smallest increase in prevalence following importation.

We also found that probability of importation success differed between our traditional model and our malaria coinfection model (**Figure 5.3**). We define importation success by the presence or absence of non-resident strains at the end of each simulation. If non-resident strains were present, we considered the simulation to be an example of importation success. Overall, we found that importation success was low in the traditional model, with success rates below 15%. Of the networks examined, ring lattices in our traditional model were the most susceptible to invasion by an imported strain. Allowing for coinfection made importation success much more likely. We found that more than 30% of simulations in our malaria coinfection model had successful importation. Relative to the complete network, both the random regular and scale free networks had reduced importation success probabilities. Surprisingly, our malaria coinfection model showed that the ring lattice had the lowest importation success probabilities, suggesting that it was the least susceptible to invasion by an incoming strain.

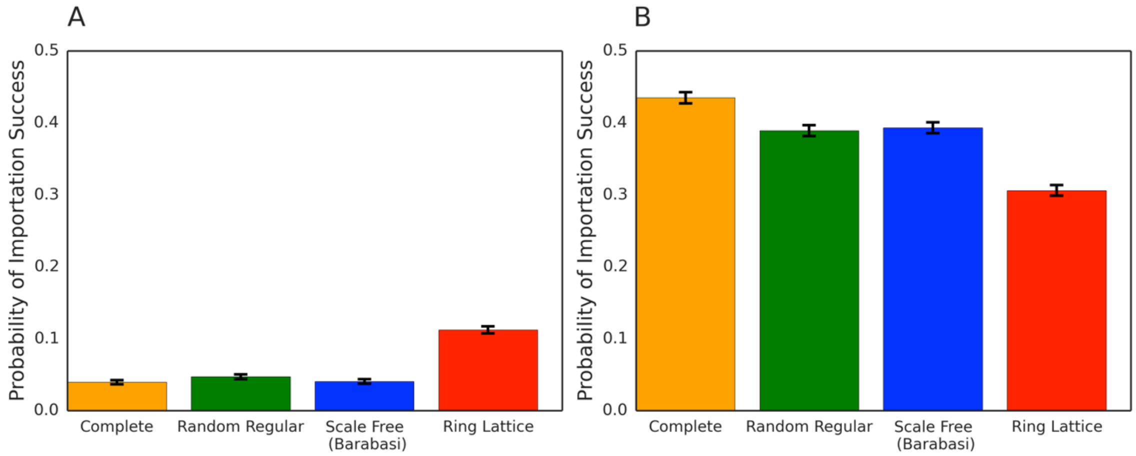


Figure 5.3 Importation success probabilities

Bar graphs of the importatation success probabilities from 4000 simulations run using our **A**) traditional model (no coinfection) and **B)** malaria coinfection model.

(Figure 5.3, continued) Error bars indicate one standard error from the mean. Importation success was defined by the presence or absence of non-resident strains in the population by the end of the simulation. Simulations with non-resident strains were considered succesful importation events. There was no statistically significant difference in importation success probabilities between the random regular and scale free networks in either the traditional model or the malaria coinfection model (Two proportion z-test: traditional model p-value = 0.78, malaria coifection model p-value = 0.71). All other comparisons were statistically significant at an alpha level of 0.004 after Bonferroni correction.

5.3.4 Incorporating more refined transmission topologies decreases the rate with which the resident strain is replaced

We then examined how coinfection and transmission topologies affected the resident strain replacement rate. Here, we focused our analyses to the subset of simulations with successful importation. Again, we found substantial differences between our traditional model and our coinfection model (**Figure 5.4 and Figure 5.5**). The most obvious difference concerns the generation of recombinant strains. Because coinfection is not possible, parasites in our traditional model can never outcross or generate new recombinant strains. There are only ever two strains present in our traditional model: the resident strain and the imported strain. In contrast, there can be a limitless number of strains in our coinfection model because each outcrossing event results in the formation of new recombinant strains. In our malaria coinfection model, we divided the parasite population into three categories consisting of 1) resident strains, 2) imported strains, and 3) recombinant strains.

None of the simulations in our traditional model resulted in the complete replacement of the resident strain (**Figure 5.4**). However, they showed that scale free networks had the lowest resident strain frequencies while ring lattices had the highest resident strain frequencies. There was no difference in resident strain frequencies between the random regular and complete networks. We found a completely different association between resident strain frequencies and transmission topology in our malaria coinfection model (**Figure 5.5**). Here, the resident strain was completely replaced by recombinant strains in the complete, random regular and scale free networks. Complete networks had the fastest replacement rates, but we observed no difference between the random regular and scale free networks. Only epidemics run on ring lattices failed to replace the resident strain. Epidemics run on ring lattices had the highest resident strain frequences and the lowest recombinant strain frequencies. Across all networks, imported strain fractions remained small. There was a small increase in the imported strain fraction immediately after importation, but this declined as the epidemic proceeded.

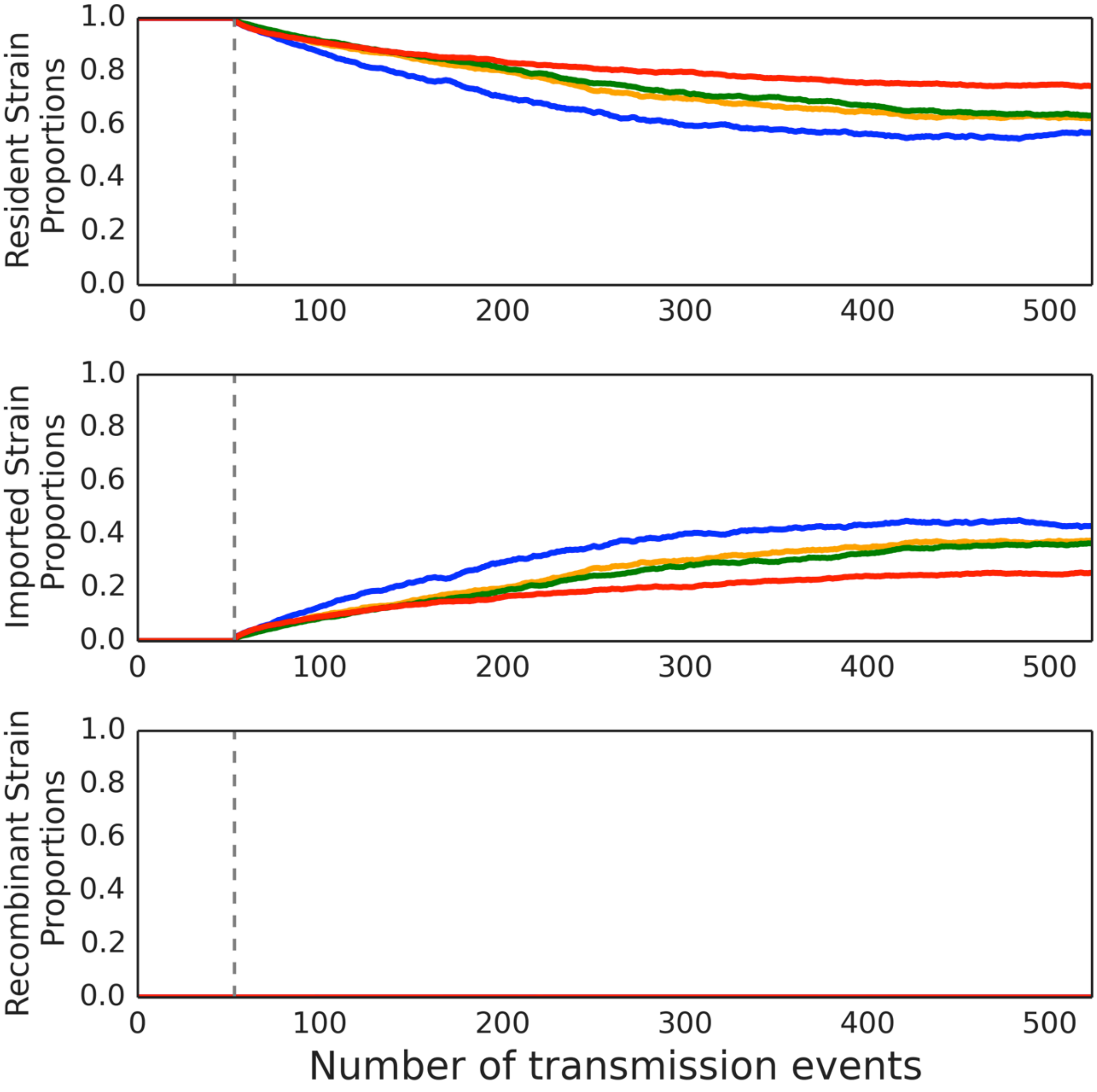


Figure 5.4 Strain proportion plots in our traditional model

Resident strain, imported strain, and recombinant strain proportions in our traditional model. Each line represents the average proportions in simulations with successful importation. Orange lines are the results obtained from pidemics run on complete networks, green on random regular networks, blue on scale free networks (Barabasi) and red on ring lattics. Scale free networks had the lowest resident strain proportions. There was little difference in strain proportions

(Figure 5.4, continued) between the random regular and the complete networks. No recombinant strains were present in our traditional model.

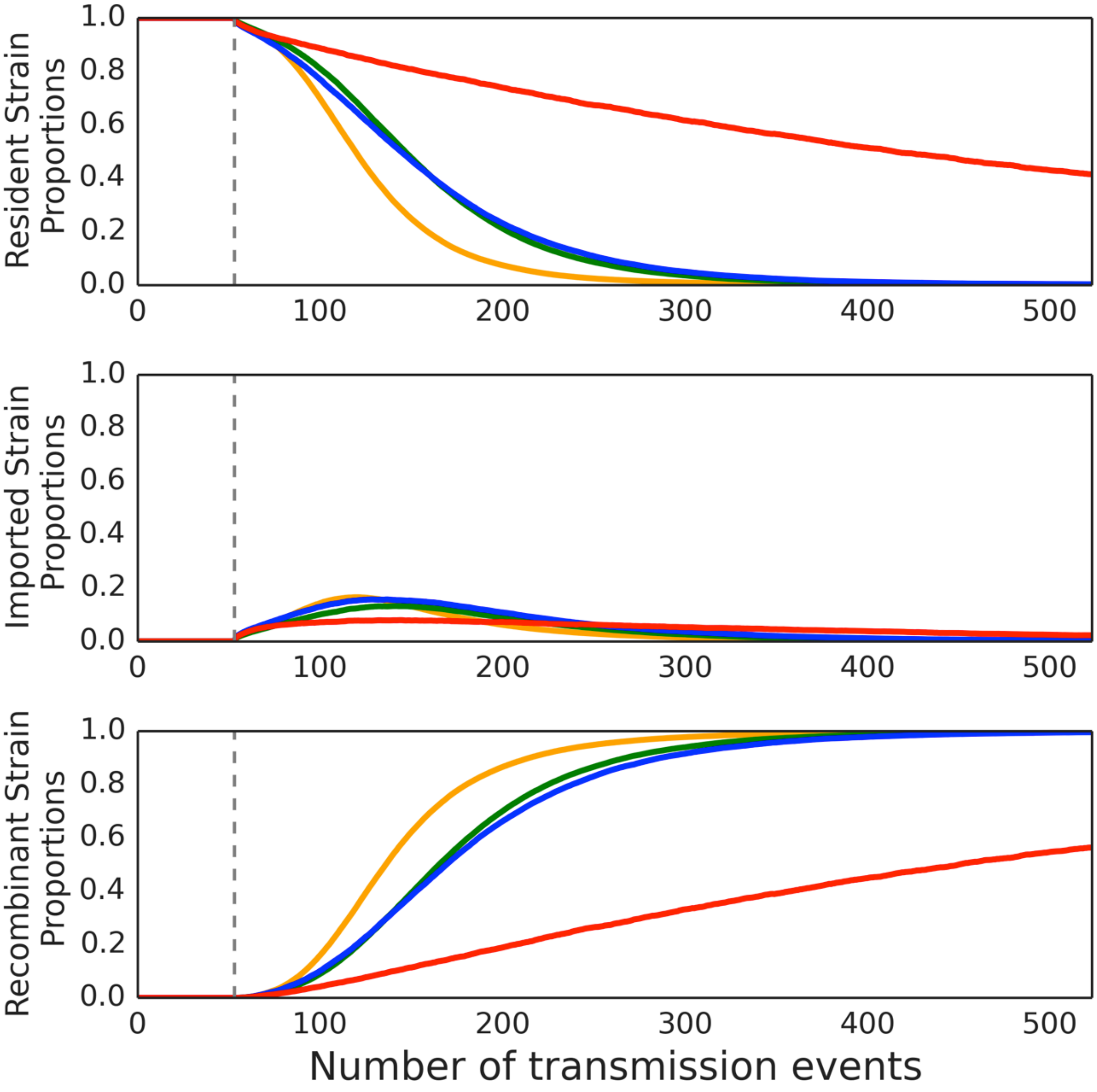


Figure 5.5 Strain proportion plots in our malaria coinfection model

Resident strain, imported strain, and recombinant strain proportions in our malaria coinfection model. Each line represents the average proportions in simulations with successful importation. Orange lines are the results obtained

(*Figure 5.5, continued*) from epidemics run on complete networks, green on random regular networks, blue on scale free networks (Barabasi) and red on ring lattics. Ring lattice networks had the highest resident strain proportions and the lowest recombinant strain proportions. There was little difference in strain proportions between the random regular and the scale free networks.

We reasoned that the high proportion of resident strains in our ring lattice networks was because clustered transmission made coinfection difficult. To address this, we examined whether the replacement of the resident strain was associated with increased outcrossing opportunity by measuring the proportion of polygenomic infections in the population. Simulations were not performed using our traditional model because polygenomic infection formation is impossible. We found that complete networks supported the highest frequency of polygenomic infections while ring lattices had the lowest frequencies (**Supplemental Figure S5.2**).

5.3.5 Limited transmission routes and clustered transmission increase allele extinction probabilities and suppress fixation probabilities

Next, we examined how network topologies influenced the emergence and spread of single point mutations in our malaria coinfection model. Here, we examined the fate of an allele carried by the imported strain by looking at the combined allele frequency spectrum across all simulations. Strains carrying this allele have a transmission advantage that makes them to more likely to be sampled by the mosquito vector than strains without the allele. This transmission advantage is relative and dependent on the frequency of the allele in the population. The imported allele confers the greatest advantage when it is rare and offers no transmission advantage when it is fixed.

For a neutral imported allele, we found that random regular, scale free, and ring lattice networks had increased extinction probabilities compared to the complete network (**Figure 5.6**). Ring lattices had the highest extinction probabilities and there was no statistically significant difference in extinction probabilities between our random regular and scale free networks (Chi-Square: p-value = 0.73). As might be expected, the imported neutral allele rarely reached fixation. For an imported beneficial allele with a 5x transmission advantage, we found that random regular, scale free, and ring lattices had reduced extinction probabilities and suppressed fixation probabilities. Complete networks had the highest fixation probabilities while ring lattices had the lowest. There was no statistically significant difference in either the fixation or extinction probabilities between our random regular and scale free networks (Chi-square: p-values = 0.17).

Finally, we examined the linkage disequillirium (LD) of nearby sites to examine the effect these topologies have on the effective recombination rate **(Figure 5.7A**). LD is the non-random association between genetic markers and the rate at which it is broken down provides a sense of how frequent outcrossing occurs. We found that the random regular and scale free networks resulted in stronger linkage disequilibrium patterns than the complete network but that ring lattices maintained the higher levels of linkage disequilibrium, even on the most distally located sites. Doubling the host population size had no effect on LD decay in simulation run with complete, random regular and scale free networks but increased LD decay in simulations run on ring lattices (**Figure 5.7B**).

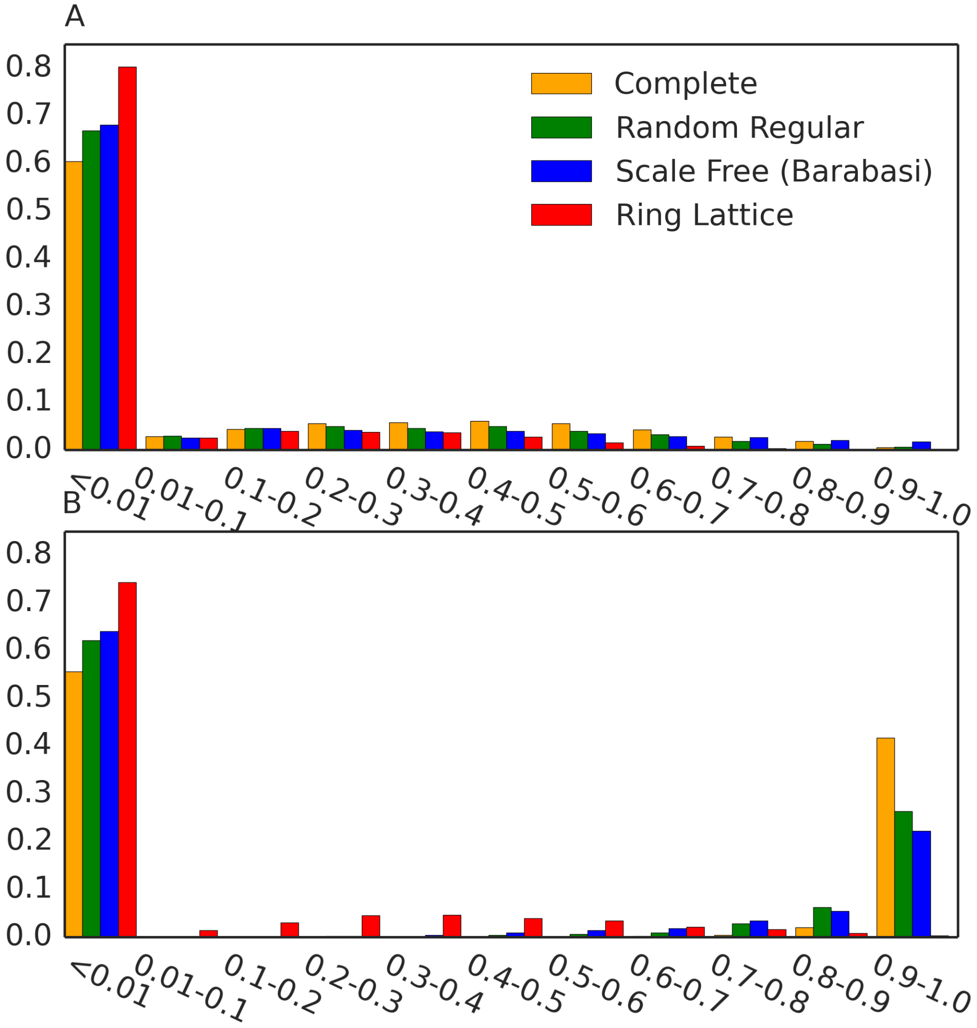


Figure 5.6 Allele frequency spectrum of the imported allele in our malaria coinfection model

The allele frequency spectrum of **A)** a neutral imported allele or **B)** an imported allele that confers a transmission advantage across all 4000 simulations. In **B**), strains with the imported allele are 5x more likely to be sampled by the mosquito vector than strains without the imported allele. The allele frequency spectrum is not conditioned on successful invasion. Simulations where the allele frequency of

(*Figure 5.6, continued)* the imported allele was less than 0.01 were considered extinct while those with allele frequencies between 0.9 and 1.0 were considered fixed or near fixation.

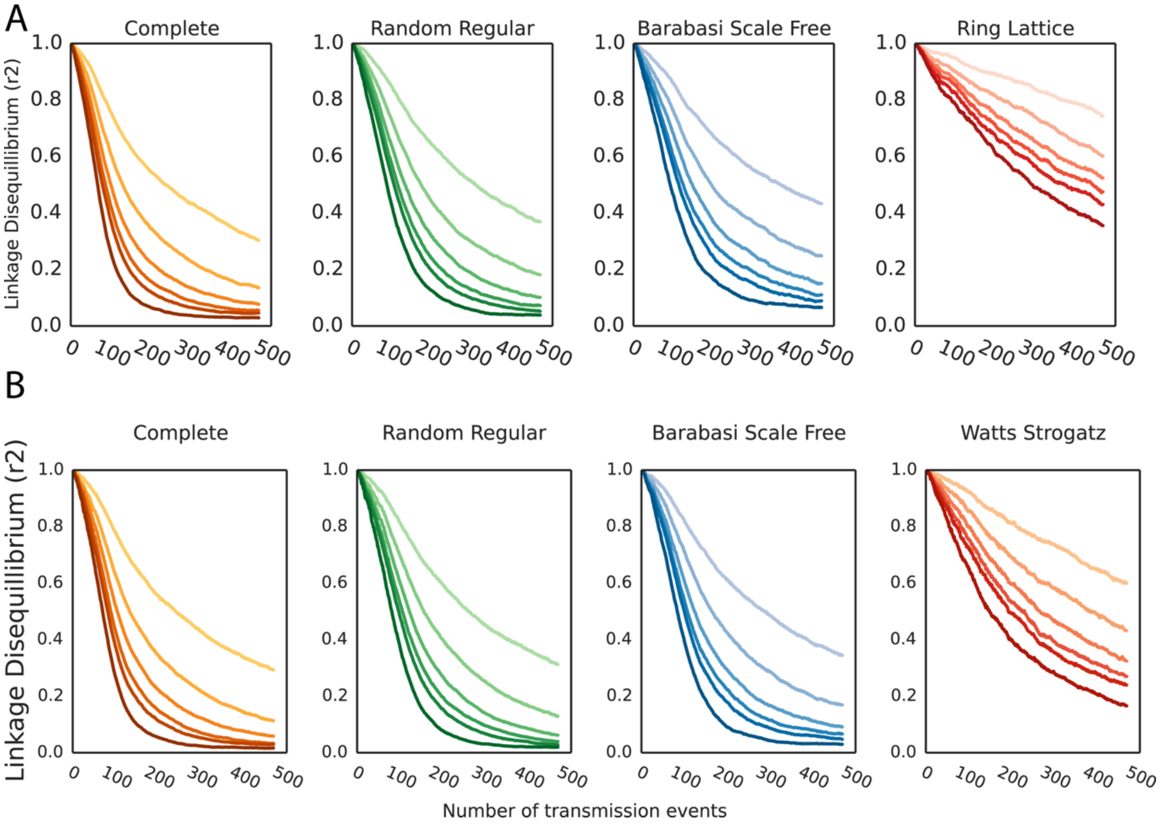


Figure 5.7 Linkage disequilibrium of alleles proximal to the imported allele

From lightest to darkest, linkage disequilibrium (LD) plots of alleles located 10cM, 20cM, 30cM, 40cM, 50cM from the site of the imported allele. The darkest line shows the LD at a site located on a different chromosome. We assumed that 1cM was equal to 1.5 Mbp [67]. LD was calculated using r^2, which is better at detecting LD between low frequency alleles than D prime.

5.4 Discussion

Here, we examined how different transmission topologies could affect malaria parasite population genetics using evolutionary invasion models. Because coinfection and cotransmission is an important part of malaria transmission, we also examined whether the coinfection and sexual recombination would affect the predictions made from evolutionary invasion models. Our results highlight the importance of transmission topology and show how populations with the same transmission intensity can have different epidemiological and population genetic outcomes. Models that assume transmission is completely mixed have higher prevalence values, invasion success probabilities, and allele fixation probabilities than models that incorporate more complicated transmission dynamics. While superspreading was previously shown to have an inhibitory effect in traditional evolutionary invasion models [123], we found it had little effect compared to a network with the same level of limited transmission once coinfection and cotransmission was allowed.

Our study shows that clustered transmission impedes the spread of new alleles in the population and promotes inbreeding of parasites in locally spaced infections. Clustered transmission limits mating to neighboring nodes, resulting in the maintenance of higher levels of LD. While this makes intuitive sense, it also runs counter to the historical patterns of malaria drug resistance emergence and spread. Despite having the greatest levels of parasite diversity [12,13], drug resistance mutations of global importance rarely originate from the African continent [14,18,126,127]. Southeast Asian and South American parasite populations are more clustered, structured, and have smaller effective population sizes than African parasite populations [12,13,22]. Such conditions enhance the effects of genetic drift, which is theorized to increase the number of slightly deleterious mutations in the population and allow new mutational pathways to be accessed [128–130]. Migration between populations and other metapopulation (multiple, loosely connected communities) structures could make it easier for mutations to fixate than in our model, which consists of a single, closed population. Alterations in population size could also play a role, but we found there no differences in results from run with 200 or 400 individuals. It is likely that our understanding of how structure influences drug resistance will need to incorporate differential treatment coverage and efficacy to truly understand the emergence of drug resistance in these regions.

Another major conclusion of this study is that traditional evolutionary invasion models should not be used to understand malaria population genetics and evolution. Traditional evolutionary invasion models are characterized by the competition for uninfected hosts. In these models, infected individuals cannot be reinfected until the current infection expires. Using the assumptions of traditional evolutionary invasion models, we found that clustered transmission makes a network more prone to importation. However, imported strains are maintained at low frequencies. This is because epidemics in ring lattices form small clusters of uninfected individuals. This make invasion easier by providing an environment where the imported strain does not compete with the resident strain for uninfected hosts [120]. However, these clusters also make it more difficult for imported strains to reach high frequencies because other uninfected pockets in the population are difficult to reach. Conditioned on successful invasion, we found that scale free networks supported the highest frequency of imported strains. This outcome is consistent with the expectation that epidemics on scale free networks spread more rapidly than on other networks [131]. This is due to the presence of central hubs that serve as rapid access points to other individuals in the population [123,131,132].

Why do these predictions not hold when coinfection and cotransmission are allowed? Part of the reason is because uninfected host availability no longer limits transmission. While the dynamics of traditional evolutionary invasion models can be framed as the competition for uninfected hosts, the dynamics in our malaria coinfection model are better framed as the competition for susceptible hosts. Because we assume infections cannot be reinfected by concurrent strains, susceptibility is strain-specific. Unlike in our traditional evolutionary invasion model, clustered transmission networks are now the least susceptible to invasion. This is because of two factors. First, isolated pockets of uninfected individuals do no affect transmission dynamics because all individuals immediately following importation are considered susceptible to the imported strain. Second, infections spread slowly in clustered transmission networks, as evidenced by the higher resident strain frequencies, lower polygenomic infection frequencies, and maintenance of high LD in our ring lattice networks. From a population genetic perspective, the slow dissemination of infections on clustered transmission networks makes it much more difficult for an allele to reach fixation, even if the mutation has a high selective coefficient. It also enhances the effect of random genetic drift, as neighboring individuals are more likely to have the same genetic composition and serve as dead ends for transmission.

Our coinfection framework precludes reinfection with concurrent strains and is similar to the framework used in other coinfection models [133,134]. Previous studies have argued that this framework is inaccurate for evolutionary invasion analyses because it gives imported strains a frequency dependent transmission advantage during the initial stages of invasion [37]. This scenario may not be inappropriate for a disease like malaria, which has strong strain-specific immunity [87,135–137]. Strain-specific immunity limits the spread of pre-existing diversity but is less effective at limiting the spread of novel strains. This provides a biological mechanism for strain-dependent frequency in the initial stages of invasion. Our model also does not does not take into account lasting immunity. Again, this may not be inappropriate for malaria, where strain-specific host immunity wanes quickly over time. The efficacy of the RTSS-vaccine against the 3D7 reference strain in recent clinical trials is remarkably short wanes over the course of a few weeks [138]. However, host immunity is also “strain-transcending” [87,139] and cross-reactive immune responses make hosts less susceptible to infection by new strains than simulated here. Amplicon sequencing data from the RTSS-vaccine trials suggest that cross-reactive immunity is stronger if targeted against genetically-similar parasites [138]. Although our model simulates sexual reproduction, we assumed that there were no cross-reactive immune responses between strains, even for highly related strains.

Thus, the models used in our study represent two opposite extremes. The traditional model represents complete strain-transcending immunity while our malaria coinfection model represents complete “strain-specific” immunity. Real transmission dynamics likely lie somewhere between these two extremes and future models will need to address how host immunity affects the strain dynamics of related parasite strains. One approach is to simulate parasites with different var antigens and weigh immunity according to the similarity of var types between different strains [140,141]. This approach is similar to the ones used in [62,142]. However, var antigens are susceptible to mitotic recombination [143], and it is unclear how strain-specific immunity could act on parasites with genome-wide relatedness. Data from immunogenic antigens, such as the *csp1* and *sera*2 antigens used in the RTSS vaccine, could also be used to model cross-reactivity and are not believed to undergo mitotic recombination. Future work will also need to examine how differences in infectiousness, such as that between acute and chronic infections, and asymptomatic infections could affect the results reported in this study

The use of population genetics for monitoring and assessing public health interventions relies on our ability to accurately predict how changing transmission affects parasite populations. Mechanistic mathematical models can reveal gaps in our knowledge and reveal critical features that need to be taken into account. While more complicated models are needed to accurately simulate and predict changes in natural populations, our framework enables us to develop a stronger intuition of the relative importance of different transmission structures. Such intuition is critical for interpreting data from the field and for the application of genetics for monitoring transmission and parasite evolution.

5.5 Methods

5.5.1 Model description

Our model is a stochastic, agent-based simulation that simulates the transmission of different parasite strains. Simulations were run on randomly generated networks with 200 individuals and a total of 4000 simulations were run on each network type. Resident strain epidemics were initiated with initial frequencies of at least 30% to prevent stochastic loss. After 53 transmission cycles (~30 years), an uninfected individual was randomly infected with an imported strain. Each strain is represented by a genome consisting of 66,110 sites. These sites were based off the sites with a minor allele frequency of > 0.01 in Thiès, Senegal. Only the genomic positions of these sites were used in this study and we did not incorporate allele frequencies in our simulations. We assumed that the resident and imported strain were completely unrelated to one another and had no allelic variants in common.

5.5.2 Modeling Transmission

Transmission happened in discrete steps that occurred once every 21 days. Each simulation lasts 500 transmission cycles. We determined the number of transmission per infection by drawing from a Poisson distribution with a rate parameter of 0.21 transmissions/transmission cycle. Variations in transmission rate were not explored and we assumed that the infectiousness of each infection did not vary with either parasitemia or infection age.

To simulate transmission, we used the framework described in (Plos CompBio paper, thesis chapter 2). During transmission, mosquitos sample gametocytes according to the gametcoyte parasite densities of the infected individual and the transmission advantage of each strain in the infection. Each strain's transmission advantage was determined by the presence or absence of the imported allele in its genome. By definition, the transmission coefficient of the resident strain allele was always equal to one. The probability of sampling gametocyte strains was determined by multiplying the strain's transmission advantage, *s*, with its gametocyte parasite density. These were then normalized so that the total sampling probabilities of each coinfecting strain was one. Sampled gametocyte pairs form gametes that fuse and undergo meiosis to create sporozoites that are then sampled from to determine the genomic composition of the next host.

Our model allows for both superinfection and cotransmission, but reinfection of concurrent strains is not allowed. Superinfection and cotransmission rates are not specified by our model, and instead depend on how recipient hosts are chosen. Superinfection occurs if the model chooses a previoulsy infected individual to be the recipient host. Recipient hosts are based on the connections in the network used to represent the transmission topology of the population. Random networks were generated for each simulation using the random graph generators in NetworkX, a python package for the creation, manipulation, and study of complex networks. Scale free networks were generated using the Barabasi-Albert preferential attachment model. Ring lattices were generated using the Watts Strogatz algorithm for generating small world networks. Despite using the Watts Strogatz algorithm, we did not create small world networks and the rewiring probability was set to 0.

5.5.3 Modeling Intra-host dynamics

For each infected individual, our model simulates asexual and gametocyte parasite densities which change throughout infection time. These parasite dynamics are based on the parasite densities of children under the age of five in the EMOD/DTK malaria transmission model [144] (**Supplemental Figure S5.3**). Briefly, our model simulates asexual parasite densities by tracking the parasite densities of up to 10 antigenic variants per strain. These antigenic variants are strain-specific and characterized by a normal distribution whose peak height depends on the time the strain has been present in the infection. We assume an incubation period of 7 days and that gametocytes appear 10 days after infection. The time separating antigen wave peaks are loosely based on the rate of var antigen switching in natural infections and drawn from a uniform distribution ranging from 8-30 days. While we allowed antigen waves to overlap, we assumed that antigens went extinct 15 days after peak parasitemia for that antigenic wave was reached.

Gametocyte densities are proportional to the asexual parasite densities and assumed to follow an exponentially modified Gaussian probability density function (sigma = 1.2, lambda = 0.5) over time. Each antigen has a different gametocyte production rate. Gametocyte production rates are drawn from a bounded lognormal variate distribution (mu = -3, sigma = 2) with a maximum value of 0.1 and a minimum value of 3e-4 [145]. Strains durations within each infection was determined by drawing from a lognormal variate distribution (mu =5.13, sigma = 0.8). This distribution reflects the infection duration times from the cross-sectional surveys of village populations used in the Garki project and in villages in Ghana and Tanzania [23,125]. Our model assumes that the asexual and gametocyte parasite densities of coinfecting strains are independent of one another and models each separately. However, a maximum of 10 coinfecting strains were allowed in each polygenomic infection. Once a strain expired, the individual was immediately susceptible to reinfection by the same strain. No lasting immunity was modeled.

5.6 Addendum

5.6.1 Author's contributions

EW provided the initial model design, which was heavily modified by WW. WW was responsible for project design and all data analyses. DLH and DFW supervised the project.

5.6.2 Acknowledgements

We thank Hsiao-Han Chang and Angela Early for helpful discussions. The computations in this paper were run on the Odyssey cluster supported by the FAS Division of Science, Research Computing Group at Harvard University.

Chapter 6: Conclusion

Using population genomics to understand parasite transmission and evolution hinges on our ability to integrate population genetics into existing epidemiological frameworks. The integration of these fields will require advances in data analysis and theory development: how do we layer data from multiple sources into a single conceptual framework? Importantly, these advances need to act in concert and should not be developed in isolation. Theory is needed to correctly interpret population genomic signals and predict how changing transmission conditions will affect parasite evolution. Likewise, because all theoretical models rely on simplifying assumptions, data analysis can reveal when assumptions are appropriate and when assumptions are erroneous. Ideally, data analysis and theory development should form a feedback loop because neither can advance without the other. To quote Dr. Sean Caroll, a theoretical cosmologist at Caltech, "*Theory without data is blind. Data without theory is lame."*

As with all major scientific works, the scope of this thesis has expanded since its initial conception. I first described the genetic relatedness of coinfecting strains in polygenomic infections collected from Thiès, Senegal. At the time, our understanding of cotransmission and superinfection was fairly simplistic: coinfecting strains in superinfections were unrelated while those in cotransmissions were related. Based on the relatedness of these polygenomic infections, we concluded that cotransmission is common in natural parasite populations. Our study showed that cotransmission is highly prevalent in natural populations and that it is unlikely that all polygenomic infections are the result of superinfection. This study provided evidence against the assumption used in mathematical models that polygenomic infections are the result of superinfection.

However, this study also raised new issues. How reliable is genetic relatedness for identifying cotransmission events? Are, as a previous study suggested [1], polygenomic infections with highly related strains evidence of serial cotransmission? When considering transmission and parasite mating, we quickly realized that cotransmitted parasites are not always related; of the nine ways pedigrees characterizing cotransmitted parasites, three of them involve the cotransmission of unrelated parasites. Far from the simple assertion that "cotransmission = related," cotransmission is a complicated process and cotransmitted parasites can have a wide range in relatedness values. Our simulations showed that the reliability of using genetic relatedness for identifying cotransmissions depends on the greater population genetic and epidemiological context. Our simulations predict that cotransmission events in highly diverse populations are much less likely to transmit related strains than those in lower diversity populations.

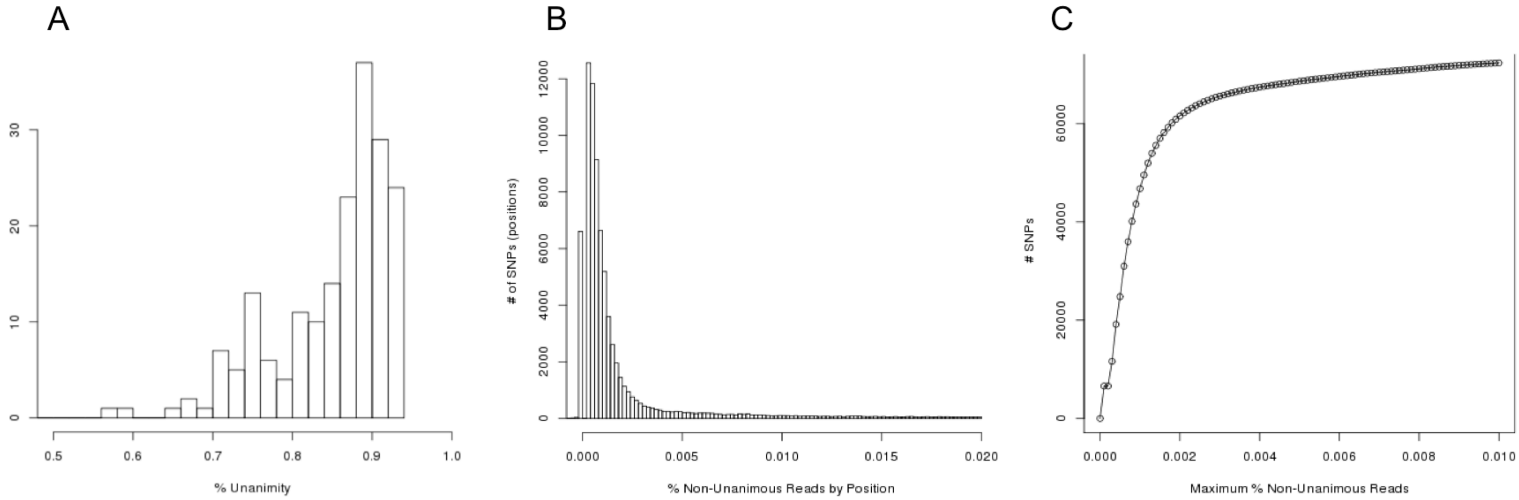
Looking back, it was fortunate that my first study analyzed the relatedness of polygenomic infections from a mid-low transmission setting. If the initial study had analyzed polygenomic infections collected from highly diverse populations in high transmission setting, we may have erroneously concluded that cotransmission is not a significant aspect of parasite transmission. Our simulations shows how theory can help us understand the data collected in the field and generate new lines of investigation. One hypothesis I am interested in pursuing is whether the relatedness of polygenomic infections differs between high and low transmission settings. To date, there has been no systematic study comparing the relatedness of polygenomic infections collected from low and high transmission settings. Optimistically, the relatedness of polygenomic infections may be a new population genetic correlate of transmission and thus of interest for future for future genetic epidemiology studies.

Although I previously stated that today's challenges lie in data interpretation, there are also major challenges in data collection and data generation that need to be addressed. In our initial study, because we were concerned about strain ascertainment bias, we sequenced patient samples directly without leukocyte depletion or hybrid selection. As a result, the sequences used in our initial study had poor sequencing depth and coverage. This prompted our investigation of whether selective whole genome amplification could be used to accurately characterize the strain dynamics of polygenomic infections. Selective whole genome amplification could increase the number of high quality whole genome sequences from poorly sampled populations where dried filter paper blood spots are easier to obtain than venous blood draws.

Finally, my last study provides an example of why it is critical that population genetic models be integrated with existing epidemiological frameworks of understanding transmission. Contact structure and transmission topology are important epidemiology concepts that are acknowledged but not well explored in population genomic studies. Current genetic epidemiology models do not handle sexual reproduction and coinfection very well, and thus unsatisfactory for making predictions regarding malaria population genetics. Using an evolutionary invasion framework, I show how transmission topology can affect malaria evolutionary dynamics.

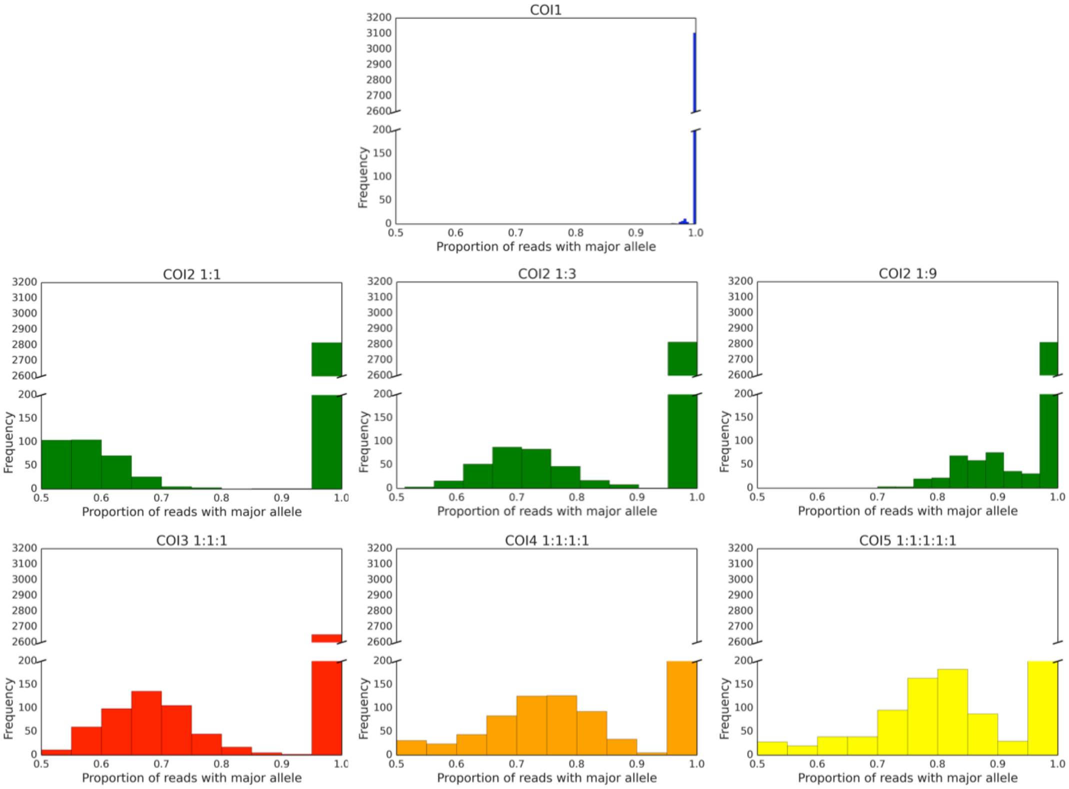
In this thesis, I focused on the consequences of malaria coinfection and how it provides the opportunity for sexual recombination and cotransmission. However, the story of how coinfection and cotransmission affects malaria genetics is far from complete. Intra-host strain dynamics due to host immunity and inter-strain competition are only lightly touched upon in this thesis and will require additional genomic analyses and modeling. Even with our limited scope, we show that coinfection and cotransmission are major aspects of malaria transmission that have an impact on malaria population genomics. This research contributes to our understanding of malaria population genomics and the importance of coinfection and sexual recombination in the context of transmission.

Appendix A: Chapter 2 Supplemental



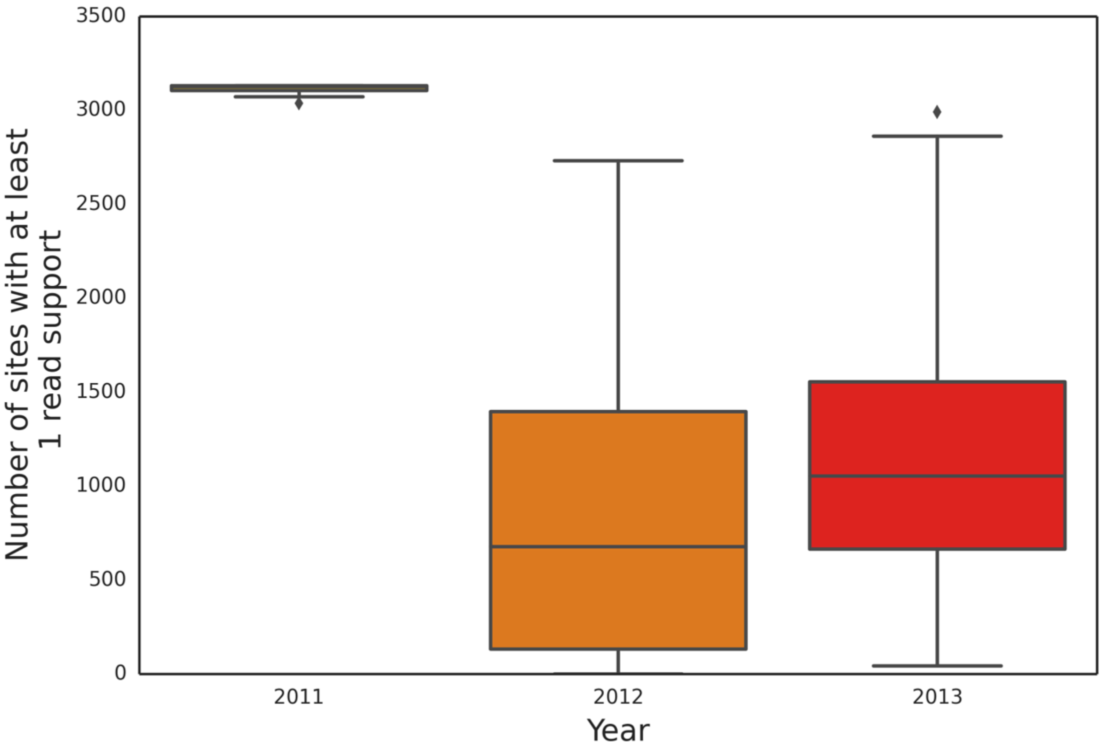
Supplemental Figure S2.1 Developing a trusted SNP set.

A) Histogram of allele balance across our set of preliminary trusted SNPs by sample. The preliminary SNP set consisted of 440,000 SNPs identified as having a non-unanimous pileup across 190 Senegal samples. An initial cut off of 80% was used to identify a set of putatively monogenomic infections. **B)** The percent non-unanimous reads across all the 440,000 SNPs in the 56 randomly chosen putative monogenomic infections. **C)** Cumulative density plot of the maximum percent of non-unanimous reads at each of the 440,000 sites. The read pileup over most sites has less than 0.2% of their reads with an alternative allele.



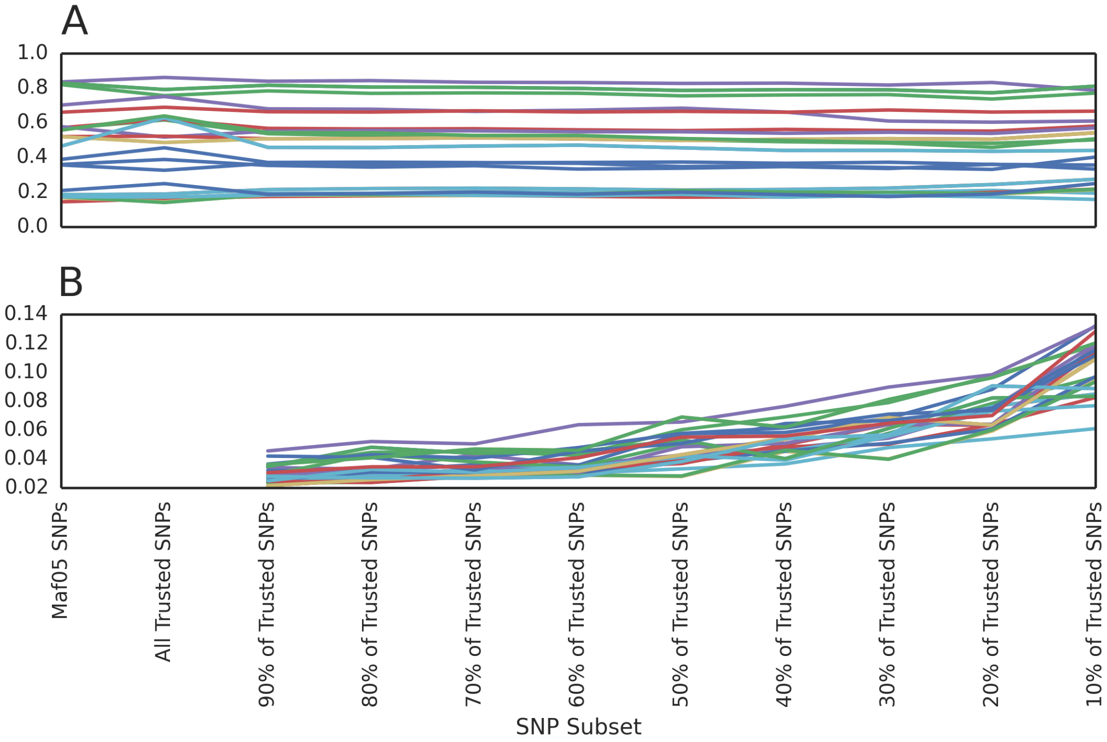
Supplemental Figure S2.2 Histogram of proportions of reads supporting the major read

Histograms showing the proportions of reads supporting the major read over each of the 3132 trusted SNP sites for a variety of lab-generated mixtures. Sites where the proportion is equal to one indicates a unanimous read pileup. Sites where the proportion is 0.6 indicates that 60% of the reads show support for the major variant present in the read pileup while 40% if the reads show support for the minor variant. Blue histogram represents the lab mixture where COI = 1, green COI = 2, red COI = 3, orange COI = 4, and yellow COI = 5.



Supplemental Figure S2.3 Trusted SNP set coverage for all 111 samples collected from 2011-2013.

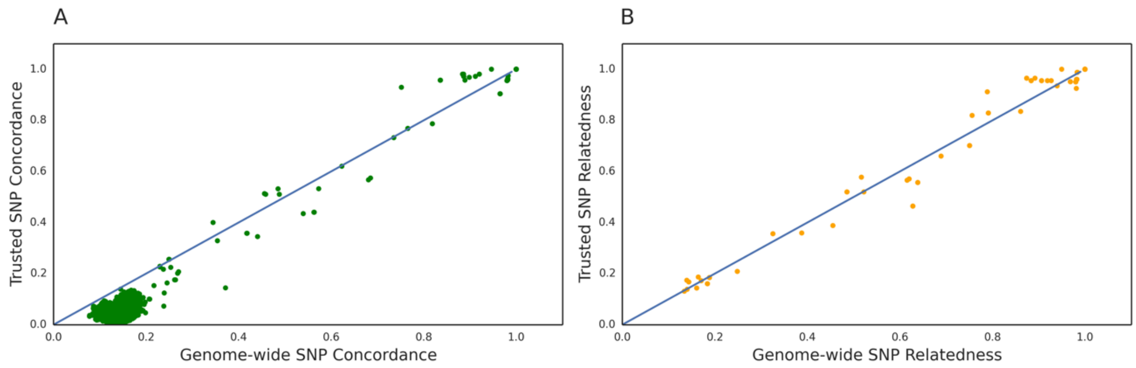
Boxplots showing the number of trusted SNP sites with at least 1 read support. Each dot represents a sample collected from 2011-2013. Samples collected from 2011 had the most number of trusted SNP sites represented, while samples collected from 2012-2013 had a broad range in the number of represented sites.



Supplemental Figure S2.4 Sensitivity of HMM to different SNP subsets.

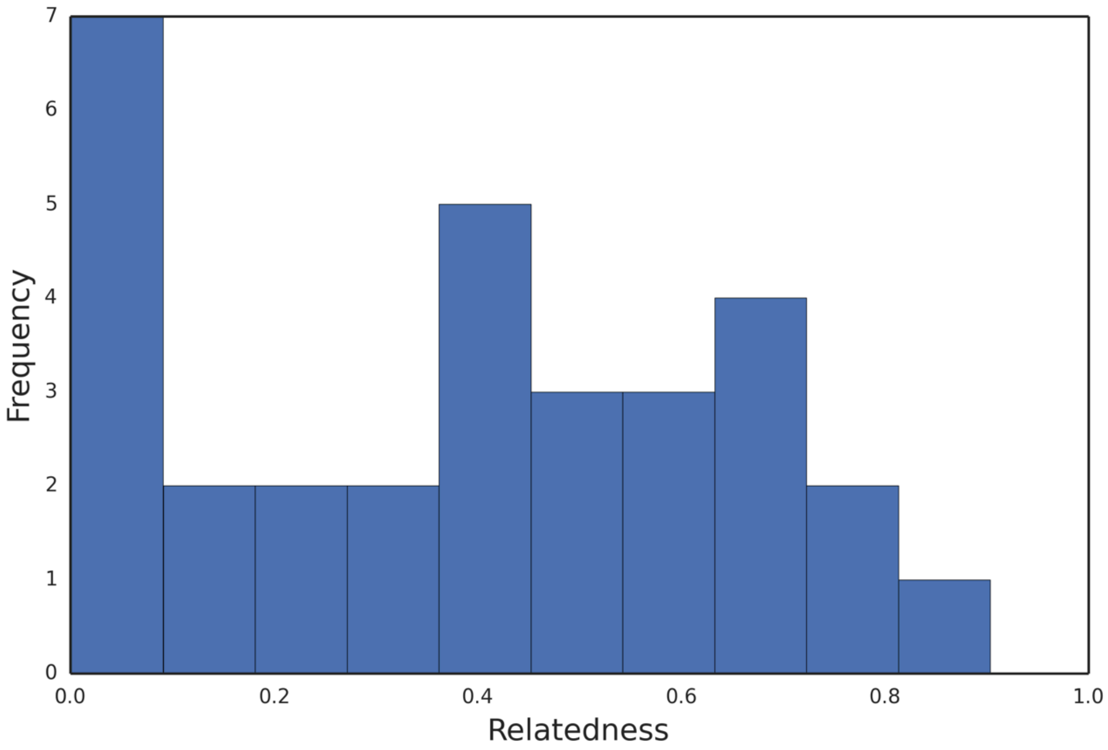
A) Mean relatedness estimates across different SNP subsets. For the genome-wide SNP set and the trusted SNP set, the mean is equal to the observed relatedness. For all other SNP sets, the mean represents the average relatedness across 40 randomly generated subsets. Each line represents a particular pairwise comparison

B) Standard deviation of relatedness estimates across different SNP subsets. For the genome-wide SNP set and the full trusted SNP set, we cannot calculate a standard deviation and are thus not plotted. For all other SNP sets, we calculated the standard deviation across 40 randomly generated subsets. Each line represents a particular pairwise comparison.



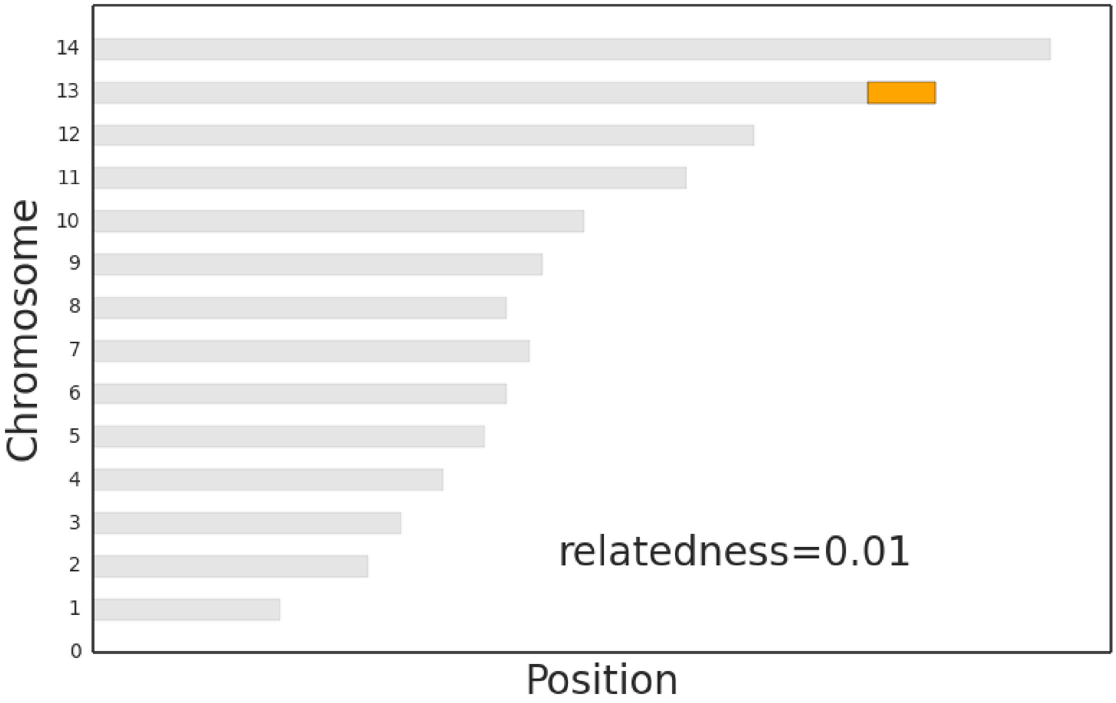
Supplemental Figure S2.5 3132 trusted SNP set vs genome-wide SNP set

Scatterplots representing the concordance (**A**) and relatedness (**B**) calculating using a set of 3132 trusted SNPs or the a set of 14,197 genome-wide SNPs. Concordance is the percent similarity at minor allele sites while relatedness is the proportion of the genome that is IBD. Individual points represent individual sample pairs. The blue diagonal line represents the 1-1 expectation.



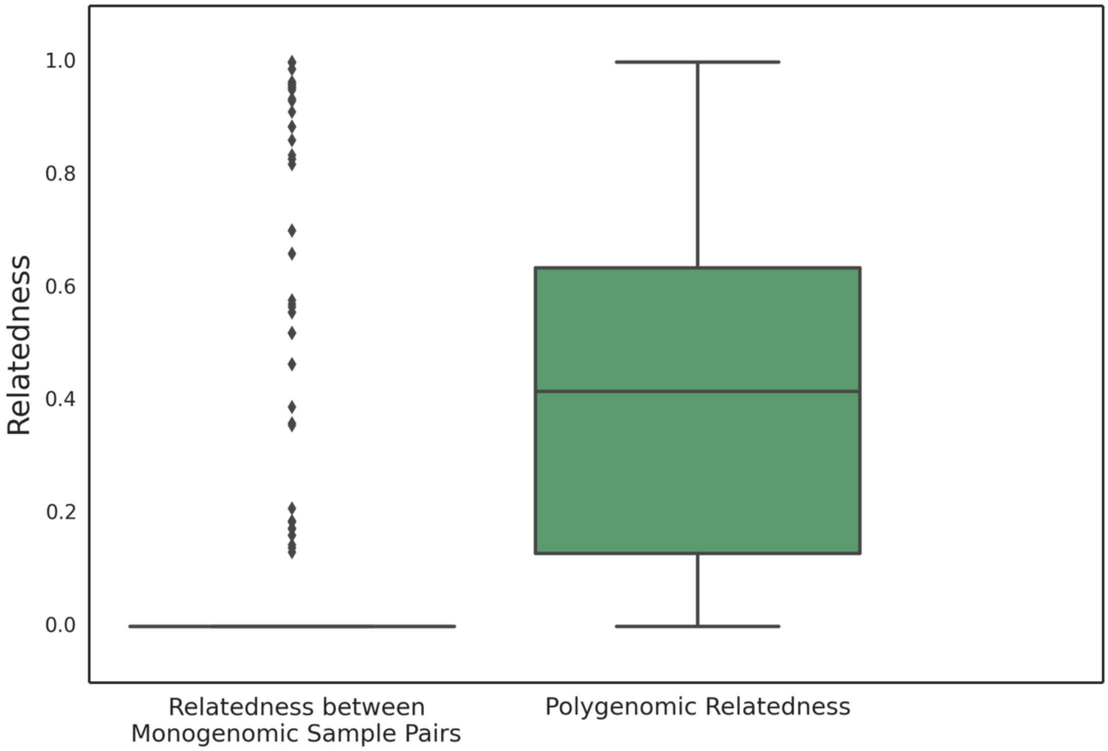
Supplemental Figure S2.6 Histogram of relatedness within polygenomic infections

Histogram of jackknife estimates of the mean relatedness within 31 polygenomic infections collected from patients in Senegal from 2011-2013.



Supplemental Figure S2.7 IBD map of the monogenomic strains related to SenT009.11

Orange represents the section of the genome that is IBD while grey represents the section of the genome that is not IBD.



Supplemental Figure S2.8 Naive resampling relatedness distributions

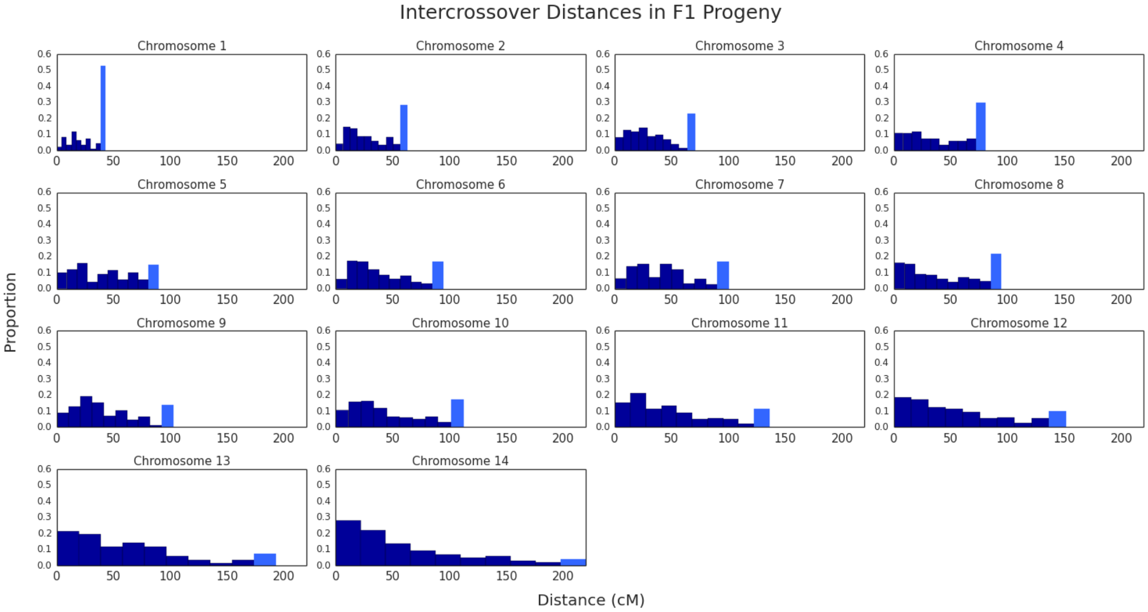
Boxplots of the relatedness between monogenomic sample pairs and the observed polygenomic infections. The distribution of relatedness among monogenomic infections is highly skewed, with 99% of the data having a relatedness of 0. The mean relatedness of this distribution is 0.07. The distribution of relatedness within polygenomic infections is much less skewed, with a mean relatedness of 0.4.

Supplemental Table S2.1

Table of the strain composition and ratios of the lab-generated strain mixtures. A = SenT148.2009, B = SenT111.2009, C = SenT165.2009, D = SenT033.2009, and E = SenT015.2009. Sample concentrations were determined by nanodrop (Thermo).The mixtures of genomic DNA were made at final concentrations of 5ng/ul.

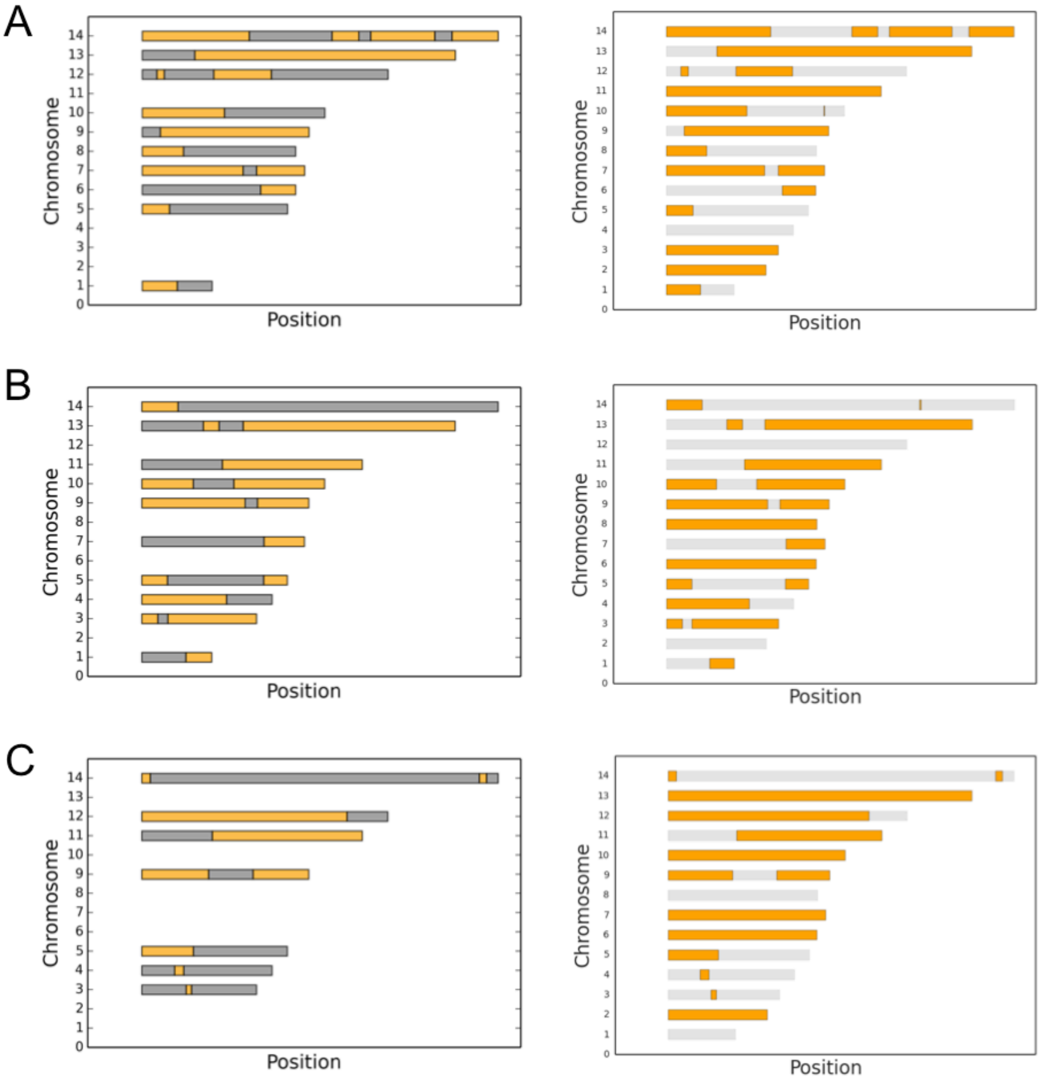
ID	Combination	Mixture
COI5	ABCDE	20:20:20:20:20
COI4	ABCD	25:25:25:25
COI3a	CDE	33:33:33
COI3b	CDE	40:20:40
COI3C	CDE	45:10:45
COI3d	CDE	49:02:49
COI2a	AB	50:50
COI2b	AB	75:25
COI2c	AB	90:10
COI2d	AB	95:05
COI2e	BC	50:50
COI2f	BC	75:25
COI2g	BC	90:10
COI2h	BC	95:05
COI1A	A	100
COI1B	B	100
COI1C	C	100
COI1D	D	100
COI1E	E	100

Appendix B: Chapter 3 Supplemental



Supplemental Figure S3.1 Observed intercrossover distances from progeny of lab crossed strains.

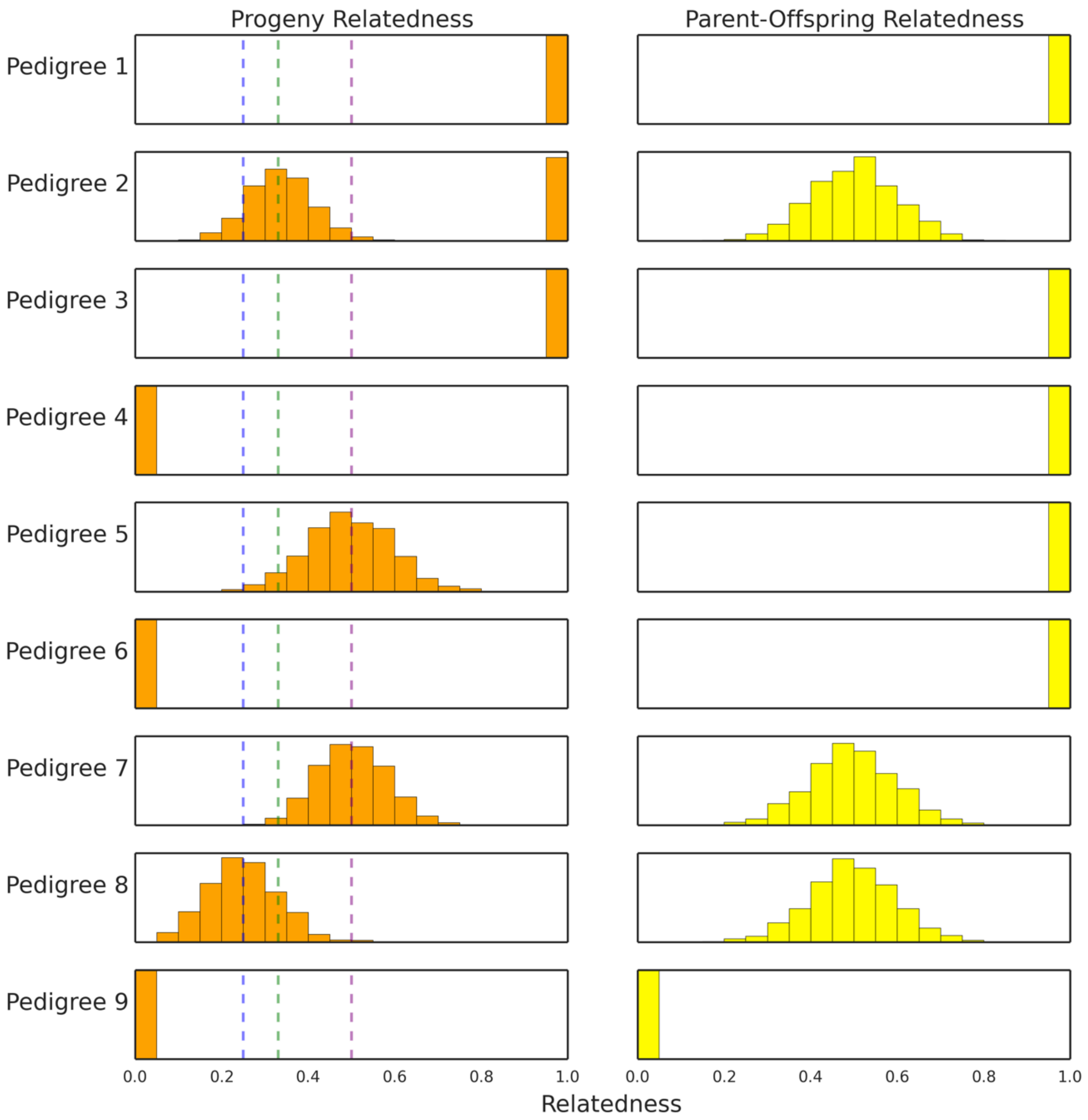
Histograms of the distribution of intercrossover distances (cM) for each chromosome in the *P. falciparum* genome. Dark blue indicate distances whose boundaries fall within each chromosome and light blue represents distances that span the entire chromosome.



Supplemental Figure S3.2 IBD Map Comparison.

Comparison of the parental inheritance boundaries defined by [68] (left) and our HMM (right) for A) 3D7_ERR019061 (parental) vs C12_ERR019063 (progeny) B) 7G8_ERR027099 (parental) vs AUD_ERR029406 (progeny) and C) DD2_ERR012840 (parental) vs 3BA6_ERR126027 (progeny). For the maps based on the boundaries defined by [68], only the results from chromosomes with evidence of recombination are shown. Orange coloration indicates a section

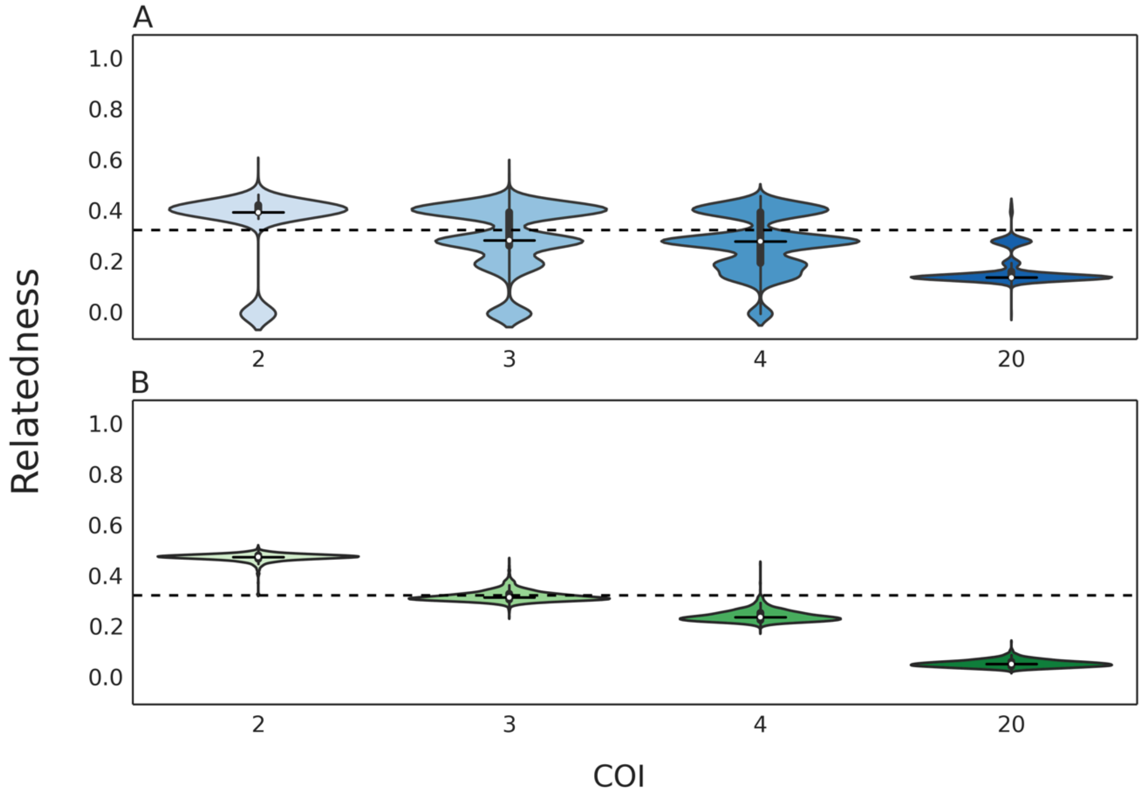
(*Supplemental Figure S3.2, continued*) of the genome inherited from 3D7_ERR019061, 7G8_ERR027099, or DD2_ERR012840 while grey sections indicate a section inherited by the other parent in the cross. Our HMM occasionally identified short IBD segments not present in the data from [68].



Supplemental Figure S3.3 Relatedness distributions for each of the 9 pedigrees.

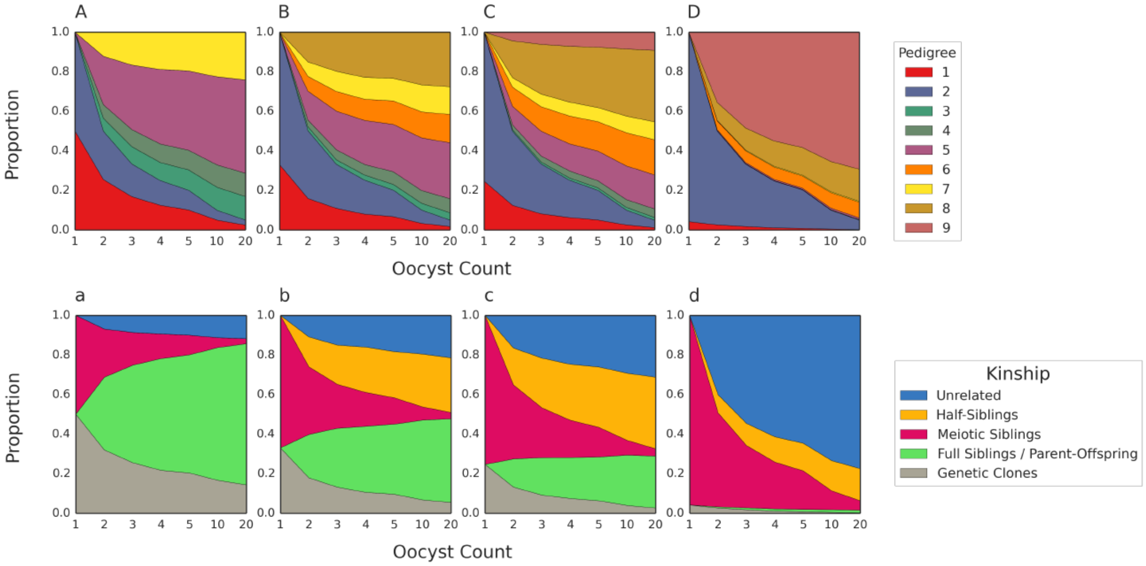
Histograms of the expected relatedness for each pedigree. Orange: Relatedness of between progeny strains. Yellow: relatedness of progeny strains vs one of the parental strains. The blue dotted line represents the expected relatedness of half-siblings (0.25), the green dotted line represents the expected relatedness of

(*Supplemental Figure S3.3, continued*) unique meiotic siblings (0.33), and the purple dotted line represents the expected relatedness of full-siblings / parent-offspring strains (0.5).



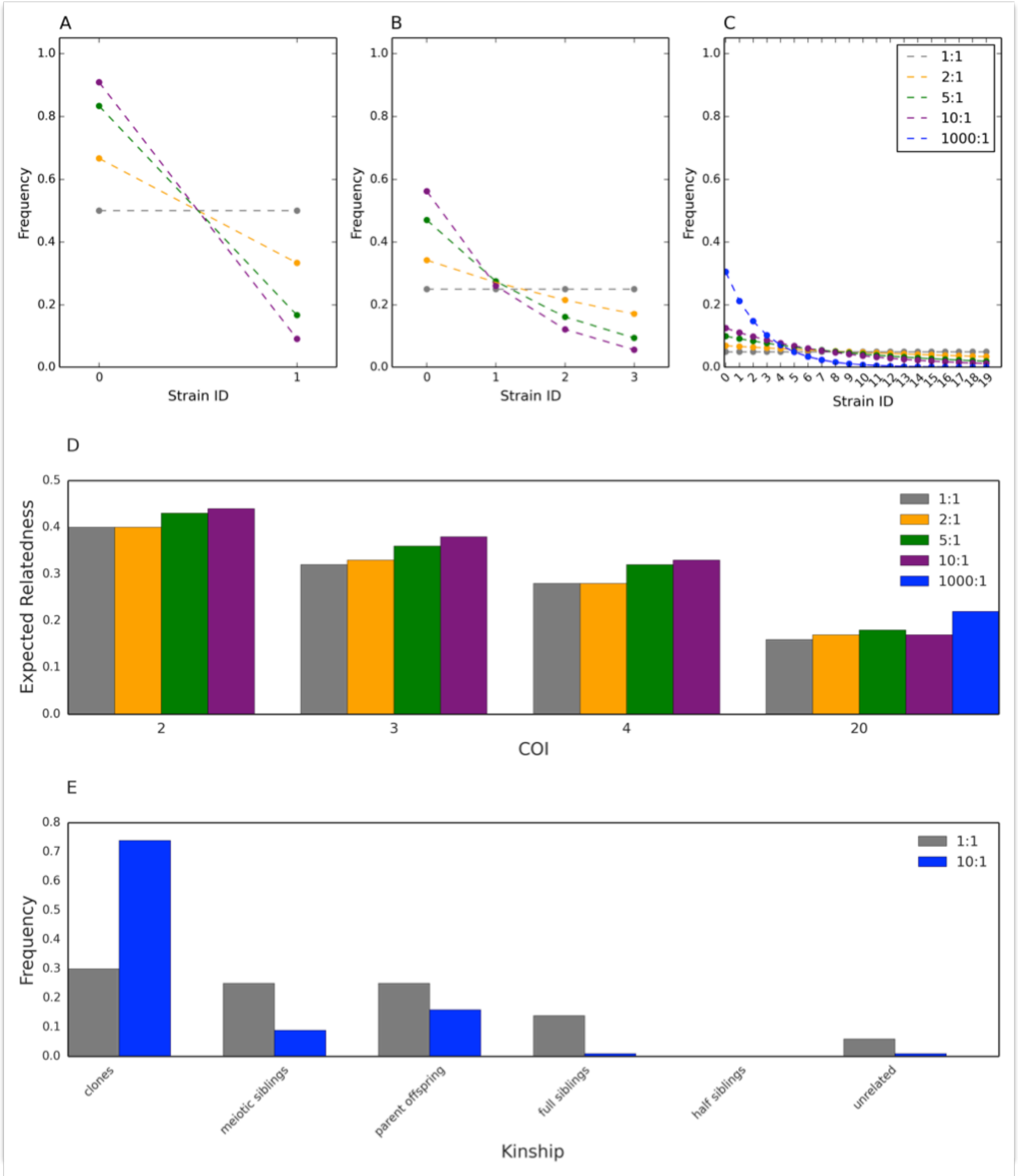
Supplemental Figure S3.4 Relatedness of cotransmitted strains in multiple oocyst simulations with high infected hepatocyte counts.

Violin plots of the relatedness of cotransmitted strains in simulations where the infected hepatocyte count was 20 and the oocyst count was 2 (**A**) or 20 (**B**). A box plot is drawn in the center of each violin plot, where the white dot represents the median of the distribution, the thicker line represent the interquartile range, and the thinner line represents the whiskers of the box plot, up to 1.5 times the interquartile range. The horizontal dotted line represents the value of 0.33.



Supplemental Figure S3.5 Pedigree and kinship frequencies from multiple oocyst simulations with high infected hepatocyte counts.

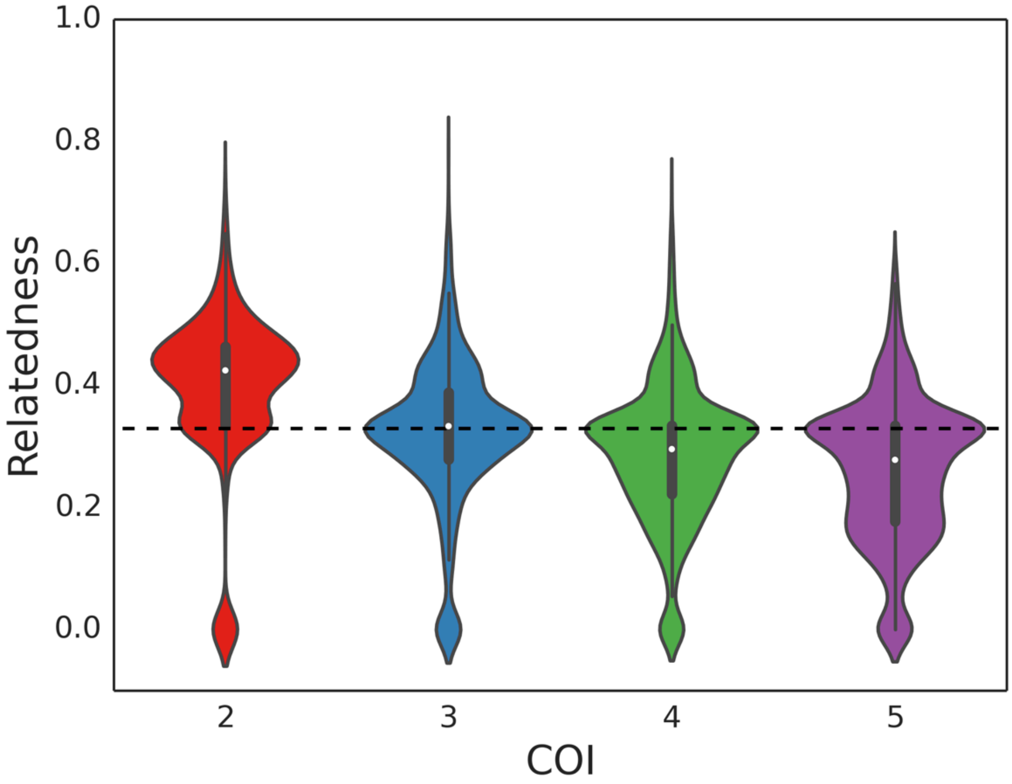
Stacked line charts of the frequencies of different pedigrees (A-D) and kinships (a-d) plotted against oocyst count. Each subplot represents a scenario with a different COI (A/a = 2, B/b = 3, C/c = 4, D/d = 20). Results from simulations where infected hepatocyte count = 20 are shown. Genetic clones are defined as those emerging from oocysts characterized by pedigree 1 and 3; genetically identical meiotic siblings are still classified as meiotic siblings in this graph



Supplemental Figure S3.6 Non-uniform gametocyte sampling probabilities

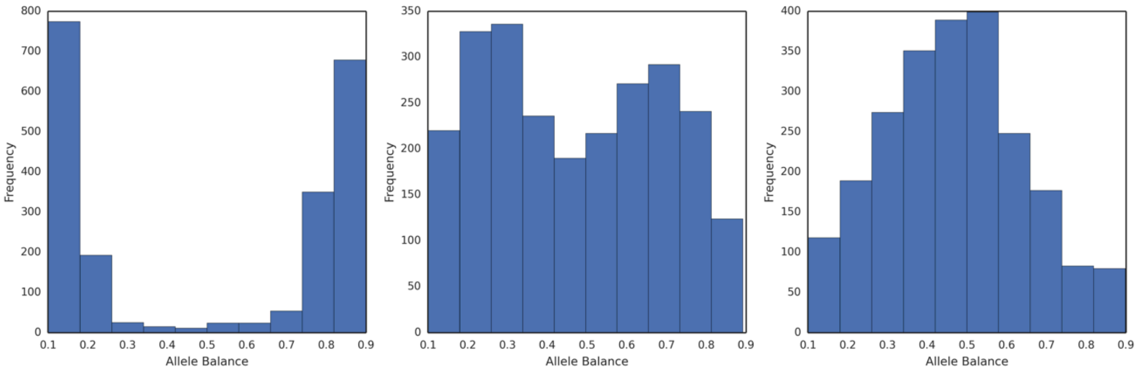
Strain frequencies in COI 2 (A), 4 (B) and 20 (C) infections. We examined strain proportions ranging from 1:1 to 10:1 for all COI infections. We also examined a 1000:1 ratio for COI = 20 infections. Ratios exceeding 10:1 were not examined in the COI =2 and COI =4 infections because the minor strains become so

(*Supplemental Figure S3.6, continued*) infrequent that the infections could be considered lower COI infections. **B**) Expected relatedness of cotransmitted strains after a single cotransmission event. Only results using oocyst and infected hepatocyte counts of 2 are shown. **C**) Kinships among transmitted parasites from infections with different strain proportions. Only results using oocyst and infected counts of 2 from a COI = 2 infection are shown.



Supplemental Figure S3.7 Relatedness of cotransmitted strains under variable oocyst and infected hepatocyte conditions.

Violin plots of the relatedness of cotransmitted strains where oocyst and infected hepatocyte counts were drawn from distributions resembling those in real transmission events. A box plot is drawn in the center of each violin plot, where the white dot represents the median of the distribution, the thicker line represent the interquartile range, and the thinner line represents the whiskers of the box plot, up to 1.5 times the interquartile range. The horizontal dotted line represents the value of 0.33.



Supplemental Figure S3.8 3D7 reference allele proportions in polygenomic infections collected from Thiès, Senegal

Representative 3D7 reference allele proportions in the pileups of all sites with a non-uniform read pileup from three COI = 2 polygenomic infections collected from Thiès, Senegal. These samples were previously sequenced and used in [40]. These reference allele proportions reveal a wide range in strain proportions, ranging from 1:1 to 9:1.

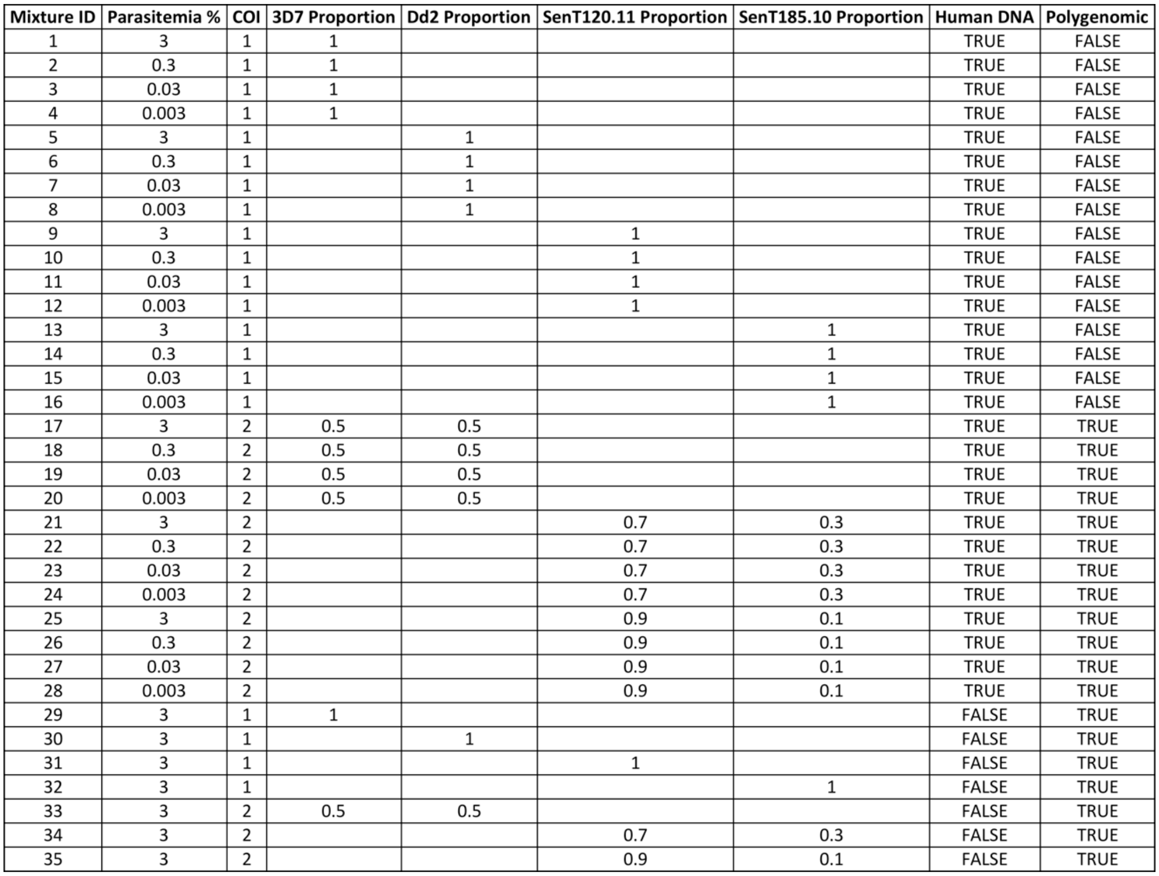
Supplemental Table S3.1 Observed chiasma events from progeny of lab crossed strains.

Chromosome	Average Number of Chiasma Events	Standard Deviation
1	0.38	0.68
2	0.78	0.99
3	0.96	1.01
4	0.75	0.84
5	1.01	0.83
6	1.14	1.03
7	1.06	1.1
8	1.06	1.2
9	1.26	1.04
10	1.19	1.13
11	1.55	1.41
12	1.57	1.31
13	1.97	1.34
14	2.45	1.66

Appendix C: Chapter 4 Supplemental

Supplemental Table S4.1

Sheet 1: List of all the mock infections generated for this study.



Mixture ID	Parasitemia %	COI	3D7 Proportion	Dd2 Proportion	SenT120.11 Proportion	SenT185.10 Proportion	Human DNA	Polygenomic
1	3	1	1				TRUE	FALSE
2	0.3	1	1				TRUE	FALSE
3	0.03	1	1				TRUE	FALSE
4	0.003	1	1				TRUE	FALSE
5	3	1		1			TRUE	FALSE
6	0.3	1		1			TRUE	FALSE
7	0.03	1		1			TRUE	FALSE
8	0.003	1		1			TRUE	FALSE
9	3	1			1		TRUE	FALSE
10	0.3	1			1		TRUE	FALSE
11	0.03	1			1		TRUE	FALSE
12	0.003	1			1		TRUE	FALSE
13	3	1				1	TRUE	FALSE
14	0.3	1				1	TRUE	FALSE
15	0.03	1				1	TRUE	FALSE
16	0.003	1				1	TRUE	FALSE
17	3	2	0.5	0.5			TRUE	TRUE
18	0.3	2	0.5	0.5			TRUE	TRUE
19	0.03	2	0.5	0.5			TRUE	TRUE
20	0.003	2	0.5	0.5			TRUE	TRUE
21	3	2			0.7	0.3	TRUE	TRUE
22	0.3	2			0.7	0.3	TRUE	TRUE
23	0.03	2			0.7	0.3	TRUE	TRUE
24	0.003	2			0.7	0.3	TRUE	TRUE
25	3	2			0.9	0.1	TRUE	TRUE
26	0.3	2			0.9	0.1	TRUE	TRUE
27	0.03	2			0.9	0.1	TRUE	TRUE
28	0.003	2			0.9	0.1	TRUE	TRUE
29	3	1	1				FALSE	TRUE
30	3	1		1			FALSE	TRUE
31	3	1			1		FALSE	TRUE
32	3	1				1	FALSE	TRUE
33	3	2	0.5	0.5			FALSE	TRUE
34	3	2			0.7	0.3	FALSE	TRUE
35	3	2			0.9	0.1	FALSE	TRUE

(Supplemental Table S4.1, continued)

Sheet 2: List of the 96 samples submitted for sequencing.

Sample	parasitemia	source_material	pre_amp	human	poly	3D7 Proportion	Dd2 Proportion	SenT120.11 Proportion	SenT185.10 Proportion
3D7_3_pd-	3	pellet	direct	FALSE	FALSE	1	0	0	0
3D7_3_pw-	3	pellet	wga	FALSE	FALSE	1	0	0	0
3D7_3_ps-	3	pellet	swga	FALSE	FALSE	1	0	0	0
Dd2_3_pd-	3	pellet	direct	FALSE	FALSE	1	0	0	0
Dd2_3_ps-	3	pellet	swga	FALSE	FALSE	1	0	0	0
S1_3_pd-	3	pellet	direct	FALSE	FALSE	1	0	0	0
S1_3_ps-	3	pellet	swga	FALSE	FALSE	1	0	0	0
S2_3_pd-	3	pellet	direct	FALSE	FALSE	1	0	0	0
S2_3_ps-	3	pellet	swga	FALSE	FALSE	1	0	0	0
3D7_3_pd+	3	pellet	direct	TRUE	FALSE	1	0	0	0
3D7_0.3_pd+	0.3	pellet	direct	TRUE	FALSE	1	0	0	0
3D7_0.03_pd+	0.03	pellet	direct	TRUE	FALSE	1	0	0	0
3D7_0.003_pd+	0.003	pellet	direct	TRUE	FALSE	1	0	0	0
3D7_3_pw+	3	pellet	wga	TRUE	FALSE	1	0	0	0
3D7_0.3_pw+	0.3	pellet	wga	TRUE	FALSE	1	0	0	0
3D7_0.03_pw+	0.03	pellet	wga	TRUE	FALSE	1	0	0	0
3D7_0.003_pw+	0.003	pellet	wga	TRUE	FALSE	1	0	0	0
3D7_3_ps+	3	pellet	swga	TRUE	FALSE	1	0	0	0
3D7_0.3_ps+	0.3	pellet	swga	TRUE	FALSE	1	0	0	0
3D7_0.03_ps+	0.03	pellet	swga	TRUE	FALSE	1	0	0	0
3D7_0.003_ps+	0.003	pellet	swga	TRUE	FALSE	1	0	0	0
3D7_3_fw+	3	filter	wga	TRUE	FALSE	1	0	0	0
3D7_0.3_fw+	0.3	filter	wga	TRUE	FALSE	1	0	0	0
3D7_0.03_fw+	0.03	filter	wga	TRUE	FALSE	1	0	0	0
3D7_0.003_fw+	0.003	filter	wga	TRUE	FALSE	1	0	0	0
3D7_3_fs+	3	filter	swga	TRUE	FALSE	1	0	0	0
3D7_0.3_fs+	0.3	filter	swga	TRUE	FALSE	1	0	0	0
3D7_0.03_fs+	0.03	filter	swga	TRUE	FALSE	1	0	0	0
3D7_0.003_fs+	0.003	filter	swga	TRUE	FALSE	1	0	0	0
3D7.1_Dd2.1_gDNA_w	-	gDNA	wga	FALSE	TRUE	0.5	0.5	0	0
3D7.1_Dd2.1_gDNA_s	-	gDNA	swga	FALSE	TRUE	0.5	0.5	0	0
3D7.1_Dd2.1_3_pd-_1	3	pellet	direct	FALSE	TRUE	0.5	0.5	0	0
3D7.1_Dd2.1_3_pd-_2	3	pellet	direct	FALSE	TRUE	0.5	0.5	0	0
3D7.1_Dd2.1_3_pw-_1	3	pellet	wga	FALSE	TRUE	0.5	0.5	0	0
3D7.1_Dd2.1_3_pw-_2	3	pellet	wga	FALSE	TRUE	0.5	0.5	0	0
3D7.1_Dd2.1_3_ps-_1	3	pellet	swga	FALSE	TRUE	0.5	0.5	0	0
3D7.1_Dd2.1_3_ps-_2	3	pellet	swga	FALSE	TRUE	0.5	0.5	0	0
3D7.1_Dd2.1_3_pd+_1	3	pellet	direct	TRUE	TRUE	0.5	0.5	0	0
3D7.1_Dd2.1_3_pd+_2	3	pellet	direct	TRUE	TRUE	0.5	0.5	0	0
3D7.1_Dd2.1_3_pw+_1	3	pellet	wga	TRUE	TRUE	0.5	0.5	0	0
3D7.1_Dd2.1_3_pw+_2	3	pellet	wga	TRUE	TRUE	0.5	0.5	0	0
3D7.1_Dd2.1_3_ps+_1	3	pellet	swga	TRUE	TRUE	0.5	0.5	0	0
3D7.1_Dd2.1_3_ps+_2	3	pellet	swga	TRUE	TRUE	0.5	0.5	0	0
3D7.1_Dd2.1_3_fw+_1	3	filter	wga	TRUE	TRUE	0.5	0.5	0	0
3D7.1_Dd2.1_3_fw+_2	3	filter	wga	TRUE	TRUE	0.5	0.5	0	0
3D7.1_Dd2.1_3_fs+_1	3	filter	swga	TRUE	TRUE	0.5	0.5	0	0
3D7.1_Dd2.1_3_fs+_1	3	filter	swga	TRUE	TRUE	0.5	0.5	0	0
3D7.1_Dd2.1_0.3_pd+_1	0.3	pellet	direct	TRUE	TRUE	0.5	0.5	0	0
3D7.1_Dd2.1_0.3_pd+_2	0.3	pellet	direct	TRUE	TRUE	0.5	0.5	0	0
3D7.1_Dd2.1_0.3_pw+_1	0.3	pellet	wga	TRUE	TRUE	0.5	0.5	0	0
3D7.1_Dd2.1_0.3_pw+_2	0.3	pellet	wga	TRUE	TRUE	0.5	0.5	0	0
3D7.1_Dd2.1_0.3_ps+_1	0.3	pellet	swga	TRUE	TRUE	0.5	0.5	0	0
3D7.1_Dd2.1_0.3_ps+_2	0.3	pellet	swga	TRUE	TRUE	0.5	0.5	0	0
3D7.1_Dd2.1_0.3_fs+_1	0.3	filter	swga	TRUE	TRUE	0.5	0.5	0	0
3D7.1_Dd2.1_0.3_fs+_2	0.3	filter	swga	TRUE	TRUE	0.5	0.5	0	0
3D7.1_Dd2.1_0.3_fw+_1	0.3	filter	wga	TRUE	TRUE	0.5	0.5	0	0
3D7.1_Dd2.1_0.3_fw+_2	0.3	filter	wga	TRUE	TRUE	0.5	0.5	0	0

(Supplemental Table S4.1, sheet 2, continued)

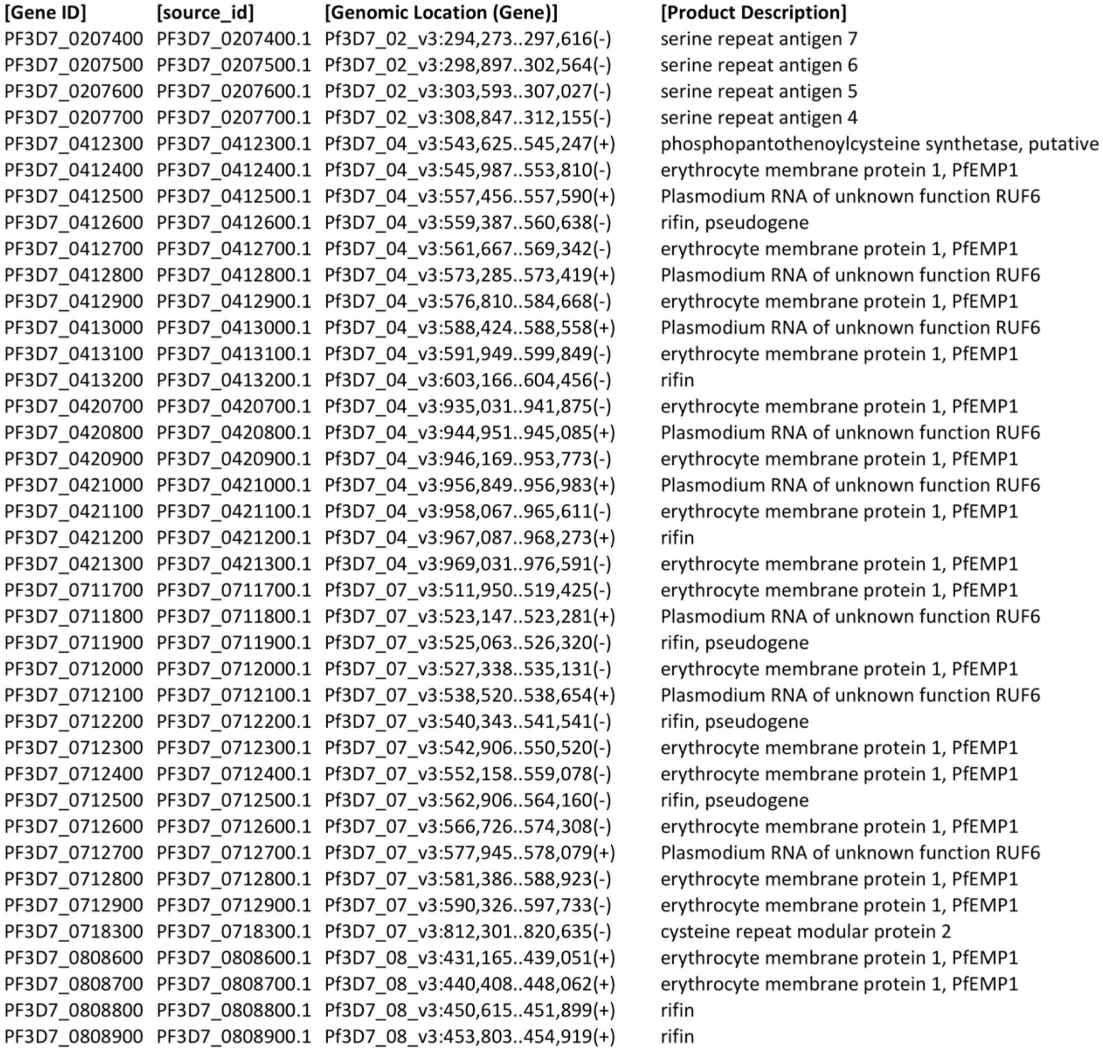
S1.1_S2.1_3_pd-_1	**3**	**pellet**	direct	**FALSE**	**TRUE**	0	0	0.7	0.3
S1.1_S2.1_3_pd-_2	**3**	**pellet**	direct	**FALSE**	**TRUE**	0	0	0.7	0.3
S1.1_S2.1_3_pw-_1	**3**	**pellet**	**wga**	**FALSE**	**TRUE**	0	0	0.7	0.3
S1.1_S2.1_3_pw-_2	**3**	**pellet**	**wga**	**FALSE**	**TRUE**	0	0	0.7	0.3
S1.1_S2.1_3_ps-_1	**3**	**pellet**	**swga**	**FALSE**	**TRUE**	0	0	0.7	0.3
S1.1_S2.1_3_ps-_2	**3**	**pellet**	**swga**	**FALSE**	**TRUE**	0	0	0.7	0.3
S1.1_S2.1_3_pd+_1	**3**	**pellet**	direct	**TRUE**	**TRUE**	0	0	0.7	0.3
S1.1_S2.1_3_pd+_2	**3**	**pellet**	direct	**TRUE**	**TRUE**	0	0	0.7	0.3
S1.1_S2.1_3_pw+_1	**3**	**pellet**	**wga**	**TRUE**	**TRUE**	0	0	0.7	0.3
S1.1_S2.1_3_pw+_2	**3**	**pellet**	**wga**	**TRUE**	**TRUE**	0	0	0.7	0.3
S1.1_S2.1_3_ps+_1	**3**	**pellet**	**swga**	**TRUE**	**TRUE**	0	0	0.7	0.3
S1.1_S2.1_3_ps+_2	**3**	**pellet**	**swga**	**TRUE**	**TRUE**	0	0	0.7	0.3
S1.1_S2.1_3_fw+_1	**3**	**filter**	**wga**	**TRUE**	**TRUE**	0	0	0.7	0.3
S1.1_S2.1_3_fw+_2	**3**	**filter**	**wga**	**TRUE**	**TRUE**	0	0	0.7	0.3
S1.1_S2.1_3_fs+_1	**3**	**filter**	**swga**	**TRUE**	**TRUE**	0	0	0.7	0.3
S1.1_S2.1_3_fs+_2	**3**	**filter**	**swga**	**TRUE**	**TRUE**	0	0	0.7	0.3
S1.1_S2.1_0.3_pd+_1	**0.3**	**pellet**	direct	**TRUE**	**TRUE**	0	0	0.7	0.3
S1.1_S2.1_0.3_pd+_2	**0.3**	**pellet**	direct	**TRUE**	**TRUE**	0	0	0.7	0.3
S1.1_S2.1_0.3_pw+_1	**0.3**	**pellet**	**wga**	**TRUE**	**TRUE**	0	0	0.7	0.3
S1.1_S2.1_0.3_pw+_2	**0.3**	**pellet**	**wga**	**TRUE**	**TRUE**	0	0	0.7	0.3
S1.1_S2.1_0.3_ps+_1	**0.3**	**pellet**	**swga**	**TRUE**	**TRUE**	0	0	0.7	0.3
S1.1_S2.1_0.3_ps+_2	**0.3**	**pellet**	**swga**	**TRUE**	**TRUE**	0	0	0.7	0.3
S1.1_S2.1_0.3_fw+_1	**0.3**	**filter**	**wga**	**TRUE**	**TRUE**	0	0	0.7	0.3
S1.1_S2.1_0.3_fw+_2	**0.3**	**filter**	**wga**	**TRUE**	**TRUE**	0	0	0.7	0.3
S1.1_S2.1_0.3_fs+_1	**0.3**	**filter**	**swga**	**TRUE**	**TRUE**	0	0	0.7	0.3
S1.1_S2.1_0.3_fs+_2	**0.3**	**filter**	**swga**	**TRUE**	**TRUE**	0	0	0.7	0.3
S1.4_S2.1_3_pd-_1	**3**	**pellet**	direct	**FALSE**	**TRUE**	0	0	0.9	0.1
S1.4_S2.1_3_pw-_1	**3**	**pellet**	**wga**	**FALSE**	**TRUE**	0	0	0.9	0.1
S1.4_S2.1_3_ps-_1	**3**	**pellet**	**swga**	**FALSE**	**TRUE**	0	0	0.9	0.1
S1.4_S2.1_3_pd+_1	**3**	**pellet**	direct	**TRUE**	**TRUE**	0	0	0.9	0.1
S1.4_S2.1_3_pw+_1	**3**	**pellet**	**wga**	**TRUE**	**TRUE**	0	0	0.9	0.1
S1.4_S2.1_3_ps+_1	**3**	**pellet**	**swga**	**TRUE**	**TRUE**	0	0	0.9	0.1
S1.4_S2.1_3_fw+_1	**3**	**filter**	**wga**	**TRUE**	**TRUE**	0	0	0.9	0.1
S1.4_S2.1_3_fs+_1	**3**	**filter**	**swga**	**TRUE**	**TRUE**	0	0	0.9	0.1
S1.4_S2.1_0.3_pd+_1	**0.3**	**pellet**	direct	**TRUE**	**TRUE**	0	0	0.9	0.1
S1.4_S2.1_0.3_pw+_1	**0.3**	**pellet**	**wga**	**TRUE**	**TRUE**	0	0	0.9	0.1
S1.4_S2.1_0.3_ps+_1	**0.3**	**pellet**	**swga**	**TRUE**	**TRUE**	0	0	0.9	0.1
S1.4_S2.1_0.3_fw+_1	**0.3**	**filter**	**wga**	**TRUE**	**TRUE**	0	0	0.9	0.1
S1.4_S2.1_0.3_fs+_1	**0.3**	**filter**	**swga**	**TRUE**	**TRUE**	0	0	0.9	0.1

Supplemental Table S4.2

Sheet 1: The genomic coordinates of all non-core regions.

chrom	start positioi	end position
1	0	91322
1	585164	640851
2	0	111632
2	292812	312923
2	866659	947102
3	0	71327
3	1008376	1067971
4	0	81778
4	545057	605567
4	938035	978324
4	1149701	1200490
5	0	80710
5	1288639	1343557
6	0	71005
6	1299642	1418242
7	0	71058
7	514555	596497
7	808203	818237
7	1079869	1089911
7	1381731	1445207
8	0	61363
8	434037	454527
8	1371148	1472805
9	0	81128
9	242026	252072
9	1480810	1541735
10	0	51172
10	1390455	1440735
10	1521264	1687656
11	0	111804
11	1913933	2038340
12	0	51182
12	765849	776062
12	1662030	1743224
12	2176552	2271494
13	0	70880
13	1430633	1450662
13	2788938	2925236
14	0	30583
14	1732075	1742119
14	3251598	3291936

(Supplemental Table S4.2, continued) Sheet 2: All genes that fall within the non-core regions of the genome.



[Gene ID]	[source_id]	[Genomic Location (Gene)]	[Product Description]
PF3D7_0207400	PF3D7_0207400.1	Pf3D7_02_v3:294,273..297,616(-)	serine repeat antigen 7
PF3D7_0207500	PF3D7_0207500.1	Pf3D7_02_v3:298,897..302,564(-)	serine repeat antigen 6
PF3D7_0207600	PF3D7_0207600.1	Pf3D7_02_v3:303,593..307,027(-)	serine repeat antigen 5
PF3D7_0207700	PF3D7_0207700.1	Pf3D7_02_v3:308,847..312,155(-)	serine repeat antigen 4
PF3D7_0412300	PF3D7_0412300.1	Pf3D7_04_v3:543,625..545,247(+)	phosphopantothenoylcysteine synthetase, putative
PF3D7_0412400	PF3D7_0412400.1	Pf3D7_04_v3:545,987..553,810(-)	erythrocyte membrane protein 1, PfEMP1
PF3D7_0412500	PF3D7_0412500.1	Pf3D7_04_v3:557,456..557,590(+)	Plasmodium RNA of unknown function RUF6
PF3D7_0412600	PF3D7_0412600.1	Pf3D7_04_v3:559,387..560,638(-)	rifin, pseudogene
PF3D7_0412700	PF3D7_0412700.1	Pf3D7_04_v3:561,667..569,342(-)	erythrocyte membrane protein 1, PfEMP1
PF3D7_0412800	PF3D7_0412800.1	Pf3D7_04_v3:573,285..573,419(+)	Plasmodium RNA of unknown function RUF6
PF3D7_0412900	PF3D7_0412900.1	Pf3D7_04_v3:576,810..584,668(-)	erythrocyte membrane protein 1, PfEMP1
PF3D7_0413000	PF3D7_0413000.1	Pf3D7_04_v3:588,424..588,558(+)	Plasmodium RNA of unknown function RUF6
PF3D7_0413100	PF3D7_0413100.1	Pf3D7_04_v3:591,949..599,849(-)	erythrocyte membrane protein 1, PfEMP1
PF3D7_0413200	PF3D7_0413200.1	Pf3D7_04_v3:603,166..604,456(-)	rifin
PF3D7_0420700	PF3D7_0420700.1	Pf3D7_04_v3:935,031..941,875(-)	erythrocyte membrane protein 1, PfEMP1
PF3D7_0420800	PF3D7_0420800.1	Pf3D7_04_v3:944,951..945,085(+)	Plasmodium RNA of unknown function RUF6
PF3D7_0420900	PF3D7_0420900.1	Pf3D7_04_v3:946,169..953,773(-)	erythrocyte membrane protein 1, PfEMP1
PF3D7_0421000	PF3D7_0421000.1	Pf3D7_04_v3:956,849..956,983(+)	Plasmodium RNA of unknown function RUF6
PF3D7_0421100	PF3D7_0421100.1	Pf3D7_04_v3:958,067..965,611(-)	erythrocyte membrane protein 1, PfEMP1
PF3D7_0421200	PF3D7_0421200.1	Pf3D7_04_v3:967,087..968,273(+)	rifin
PF3D7_0421300	PF3D7_0421300.1	Pf3D7_04_v3:969,031..976,591(-)	erythrocyte membrane protein 1, PfEMP1
PF3D7_0711700	PF3D7_0711700.1	Pf3D7_07_v3:511,950..519,425(-)	erythrocyte membrane protein 1, PfEMP1
PF3D7_0711800	PF3D7_0711800.1	Pf3D7_07_v3:523,147..523,281(+)	Plasmodium RNA of unknown function RUF6
PF3D7_0711900	PF3D7_0711900.1	Pf3D7_07_v3:525,063..526,320(-)	rifin, pseudogene
PF3D7_0712000	PF3D7_0712000.1	Pf3D7_07_v3:527,338..535,131(-)	erythrocyte membrane protein 1, PfEMP1
PF3D7_0712100	PF3D7_0712100.1	Pf3D7_07_v3:538,520..538,654(+)	Plasmodium RNA of unknown function RUF6
PF3D7_0712200	PF3D7_0712200.1	Pf3D7_07_v3:540,343..541,541(-)	rifin, pseudogene
PF3D7_0712300	PF3D7_0712300.1	Pf3D7_07_v3:542,906..550,520(-)	erythrocyte membrane protein 1, PfEMP1
PF3D7_0712400	PF3D7_0712400.1	Pf3D7_07_v3:552,158..559,078(-)	erythrocyte membrane protein 1, PfEMP1
PF3D7_0712500	PF3D7_0712500.1	Pf3D7_07_v3:562,906..564,160(-)	rifin, pseudogene
PF3D7_0712600	PF3D7_0712600.1	Pf3D7_07_v3:566,726..574,308(-)	erythrocyte membrane protein 1, PfEMP1
PF3D7_0712700	PF3D7_0712700.1	Pf3D7_07_v3:577,945..578,079(+)	Plasmodium RNA of unknown function RUF6
PF3D7_0712800	PF3D7_0712800.1	Pf3D7_07_v3:581,386..588,923(-)	erythrocyte membrane protein 1, PfEMP1
PF3D7_0712900	PF3D7_0712900.1	Pf3D7_07_v3:590,326..597,733(-)	erythrocyte membrane protein 1, PfEMP1
PF3D7_0718300	PF3D7_0718300.1	Pf3D7_07_v3:812,301..820,635(-)	cysteine repeat modular protein 2
PF3D7_0808600	PF3D7_0808600.1	Pf3D7_08_v3:431,165..439,051(+)	erythrocyte membrane protein 1, PfEMP1
PF3D7_0808700	PF3D7_0808700.1	Pf3D7_08_v3:440,408..448,062(+)	erythrocyte membrane protein 1, PfEMP1
PF3D7_0808800	PF3D7_0808800.1	Pf3D7_08_v3:450,615..451,899(+)	rifin
PF3D7_0808900	PF3D7_0808900.1	Pf3D7_08_v3:453,803..454,919(+)	rifin

(Supplemental Table S4.2, sheet 2, continued)

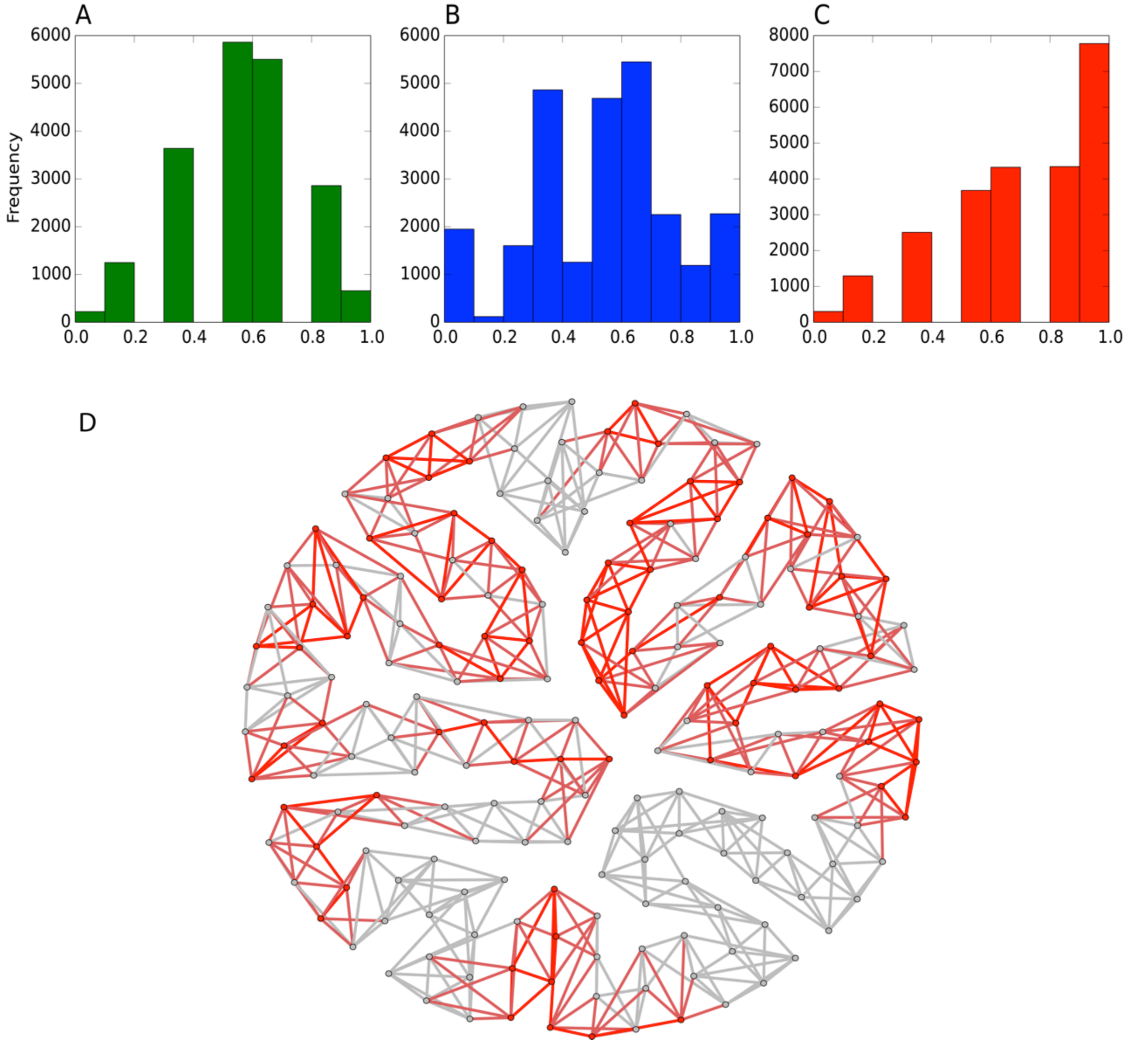
PF3D7_0905100	PF3D7_0905100.1	Pf3D7_09_v3:236,359..242,969(-)	nucleoporin NUP100/NSP100, putative
PF3D7_0905200	PF3D7_0905200.1	Pf3D7_09_v3:245,261..248,860(-)	mitochondrial carrier protein, putative
PF3D7_0905300	PF3D7_0905300.1	Pf3D7_09_v3:251,353..269,709(-)	dynein heavy chain, putative
PF3D7_1035100	PF3D7_1035100.1	Pf3D7_10_v3:1,391,445..1,393,130(+)	probable protein, unknown function
PF3D7_1035200	PF3D7_1035200.1	Pf3D7_10_v3:1,394,839..1,396,596(+)	S-antigen
PF3D7_1035300	PF3D7_1035300.1	Pf3D7_10_v3:1,399,195..1,402,896(+)	glutamate-rich protein
PF3D7_1035400	PF3D7_1035400.1	Pf3D7_10_v3:1,404,195..1,405,259(+)	merozoite surface protein 3
PF3D7_1035500	PF3D7_1035500.1	Pf3D7_10_v3:1,407,276..1,408,391(+)	merozoite surface protein 6
PF3D7_1035600	PF3D7_1035600.1	Pf3D7_10_v3:1,409,281..1,410,555(+)	merozoite surface protein
PF3D7_1035700	PF3D7_1035700.1	Pf3D7_10_v3:1,413,200..1,415,293(+)	duffy binding-like merozoite surface protein
PF3D7_1035800	PF3D7_1035800.1	Pf3D7_10_v3:1,420,533..1,422,671(+)	probable protein, unknown function
PF3D7_1035900	PF3D7_1035900.1	Pf3D7_10_v3:1,423,983..1,425,683(+)	probable protein, unknown function
PF3D7_1036000	PF3D7_1036000.1	Pf3D7_10_v3:1,427,127..1,428,344(+)	merozoite surface protein
PF3D7_1036300	PF3D7_1036300.1	Pf3D7_10_v3:1,432,498..1,434,786(+)	duffy binding-like merozoite surface protein 2
PF3D7_1036400	PF3D7_1036400.1	Pf3D7_10_v3:1,436,316..1,439,804(+)	liver stage antigen 1
PF3D7_1219300	PF3D7_1219300.1	Pf3D7_12_v3:766,654..774,197(-)	erythrocyte membrane protein 1, PfEMP1
PF3D7_1239800	PF3D7_1239800.1	Pf3D7_12_v3:1,660,451..1,677,754(+)	conserved Plasmodium protein, unknown function
PF3D7_1239900	PF3D7_1239900.1	Pf3D7_12_v3:1,678,438..1,681,536(-)	vacuolar protein sorting-associated protein 16, putative
PF3D7_1240000	PF3D7_1240000.1	Pf3D7_12_v3:1,682,835..1,684,460(+)	3-hydroxyisobutyryl-coenzyme A hydrolase, putative
PF3D7_1240100	PF3D7_1240100.1	Pf3D7_12_v3:1,685,094..1,685,411(-)	early transcribed membrane protein 12
PF3D7_1240200	PF3D7_1240200.1	Pf3D7_12_v3:1,688,606..1,691,162(-)	erythrocyte membrane protein 1 (PfEMP1), pseudogene
PF3D7_1240300	PF3D7_1240300.1	Pf3D7_12_v3:1,694,150..1,703,087(+)	erythrocyte membrane protein 1, PfEMP1
PF3D7_1240400	PF3D7_1240400.1	Pf3D7_12_v3:1,704,512..1,712,490(+)	erythrocyte membrane protein 1, PfEMP1
PF3D7_1240500	PF3D7_1240500.1	Pf3D7_12_v3:1,715,847..1,715,981(-)	Plasmodium RNA of unknown function RUF6
PF3D7_1240600	PF3D7_1240600.1	Pf3D7_12_v3:1,719,574..1,727,456(+)	erythrocyte membrane protein 1, PfEMP1
PF3D7_1240700	PF3D7_1240700.1	Pf3D7_12_v3:1,728,774..1,730,007(+)	rifin, pseudogene
PF3D7_1240800	PF3D7_1240800.1	Pf3D7_12_v3:1,731,821..1,731,955(-)	Plasmodium RNA of unknown function RUF6
PF3D7_1240900	PF3D7_1240900.1	Pf3D7_12_v3:1,735,543..1,743,408(+)	erythrocyte membrane protein 1, PfEMP1
PF3D7_1335300	PF3D7_1335300.1	Pf3D7_13_v3:1,428,874..1,438,852(-)	reticulocyte binding protein 2 homologue b
PF3D7_1335400	PF3D7_1335400.1	Pf3D7_13_v3:1,440,779..1,450,385(+)	reticulocyte binding protein 2 homologue a
PF3D7_1442600	PF3D7_1442600.1	Pf3D7_14_v3:1,730,087..1,740,916(+)	TRAP-like protein
PF3D7_1442700	PF3D7_1442700.1	Pf3D7_14_v3:1,742,113..1,749,790(+)	conserved Plasmodium protein, unknown function

Supplemental Table S4.3

List of primers used in the SWGA reaction

sWGA primers for *Plasmodium falciparum*			
Primer name	Primer sequence*	Primer quantity to order	Primer formulation
Pf1	ATATATATAT*A	250 nmole	STD
Pf2	TATATATATAT*T	250 nmole	STD
Pf3	TATATATATA*A	250 nmole	STD
Pf4	TAATATATA*T	250 nmole	STD
Pf5	TATATATATT*T	250 nmole	STD
Pf6	ATTATTATTA*T	250 nmole	STD
Pf7	TAATAATAAT*A	250 nmole	STD
Pf8	AAAAAAAAAAA*A	250 nmole	STD
Pf9	AATAATAATA*A	250 nmole	STD
Pf10	TATTATATA*T	250 nmole	STD

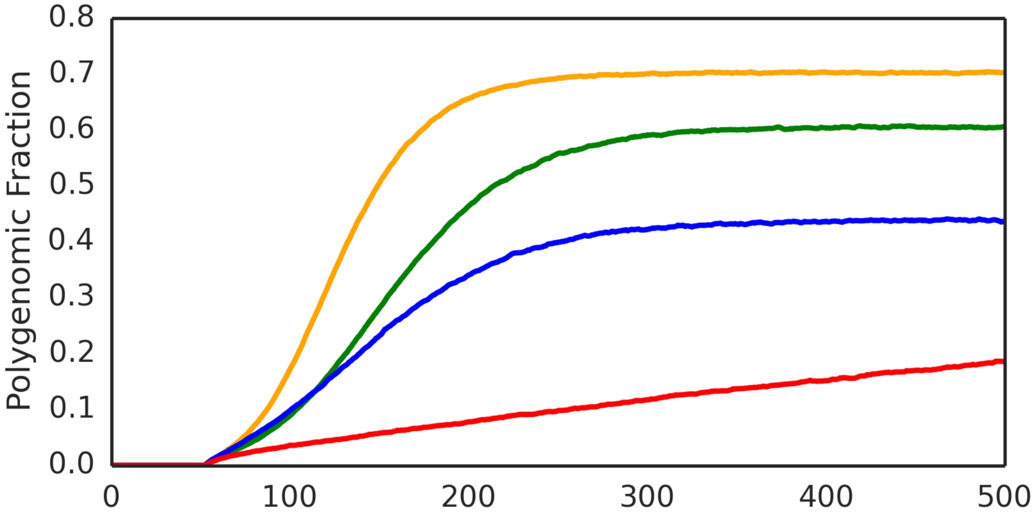
Appendix D: Chapter 5 Supplemental



Supplemental Figure S5.1 Uninfected individuals on ring lattice networks are clustered

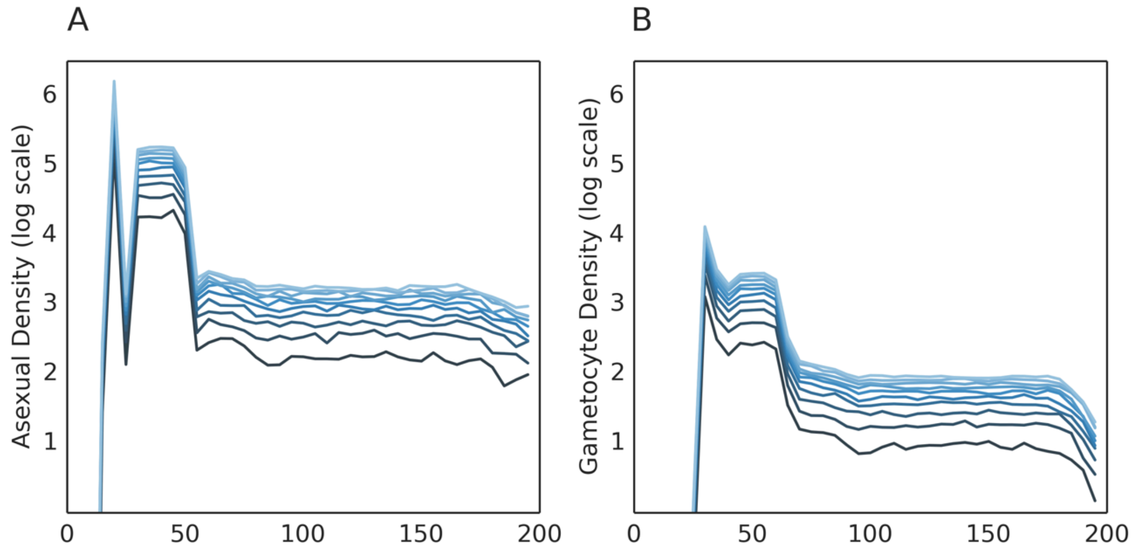
Distributions of the average proportion of edges connected to uninfected individuals per uninfected node in **A)** random regular network, **B)** Barabasi scale free network and **C**) ring lattice network from 200 simulations. The distribution is skewed in the ring lattice networks and most uninfected nodes are only connected to other uninfected nodes. **D**) An epidemic run on a ring lattice

(*Supplemental Figure 5.1, continued*) network immediately prior to the importation of a second strain. Nodes infected with the resident strain are colored red while uninfected nodes are colored grey. Edges that are drawn between one or more infected individuals are also in red. Edges connecting uninfected nodes are in grey. Uninfected individuals tend to cluster in epidemics run on ring lattice networks.



Supplemental Figure S5.2 Polygenomic infection fraction our malaria coinfection model

The fraction of infected individuals that are polygenomic in epidemics run on complete (orange), random regular (green), scale free (blue) and ring lattices (red). Complete and random regular networks have the highest proportion of polygenomic fractions because transmission is evenly distributed and random. Ring lattices had the lowest polygenomic fractions.



Supplemental Figure S5.3 Polygenomic infection fraction our malaria coinfection model

Average asexual parasite densities and gametocyte densities of infections coinfected with up to 10 strains. Densities are independently simulated for all coinfecting strains. This allows us to simulate time-dependent parasite strain proportions, but has the side effect of increasing parasite densities in polygenomic infections. Asexual and gametocyte parasite densities are used to guide mosquito sampling probabilities and have no bearing on infection duration or infectivity.

References

1. WHO. World Malaria Report 2016. 2016.

2. Nabarro DN, Tayler EM. The &quot;Roll Back Malaria&quot; Campaign. Science (80-). 1998;280: 2067 LP-2068. Available: http://science.sciencemag.org/content/280/5372/2067.abstract

3. Bhatt S, Weiss DJ, Cameron E, Bisanzio D, Mappin B, Dalrymple U, et al. The effect of malaria control on Plasmodium falciparum in Africa between 2000 and 2015. Nature. 2015;526: 207–211. doi:10.1038/nature15535

4. Noor AM, Kinyoki DK, Mundia CW, Kabaria CW, Mutua JW, Alegana VA, et al. The changing risk of Plasmodium falciparum malaria infection in Africa: 2000–10: a spatial and temporal analysis of transmission intensity. Lancet. Lancet Publishing Group; 2014;383: 1739–1747. doi:10.1016/S0140-6736(13)62566-0

5. Lederberg J. J. B. S. Haldane (1949) on infectious disease and evolution. Genetics. 1999;153: 1–3. Available: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1460735/

6. J.B.S. H. No Title. La Ric Sci Suppl A. 1949;19: 68–76.

7. Allison AC. Protection Afforded by Sickle-cell Trait Against Subtertian Malarial Infection. Br Med J. 1954;1: 290–294. Available: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2093356/

8. Kwiatkowski DP. How Malaria Has Affected the Human Genome and What Human Genetics Can Teach Us about Malaria. Am J Hum Genet. The

American Society of Human Genetics; 2005;77: 171–192. Available: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1224522/

9. Ruwende C, Khoo SC, Snow RW, Yates SN, Kwiatkowski D, Gupta S, et al. Natural selection of hemi- and heterozygotes for G6PD deficiency in Africa by resistance to severe malaria. Nature. 1995;376. doi:10.1038/376246a0

10. Gardner MJ, Hall N, Fung E, White O, Berriman M, Hyman RW, et al. Genome sequence of the human malaria parasite Plasmodium falciparum. Nature. Macmillian Magazines Ltd.; 2002;419: 498. Available: http://dx.doi.org/10.1038/nature01097

11. Chang H-. H, Park DJ, Galinsky KJ, Schaffner SF, Ndiaye D, Ndir O. Genomic sequencing of Plasmodium falciparum malaria parasites from Senegal reveals the demographic history of the population. Mol Biol Evol. 2012;29. doi:10.1093/molbev/mss161

12. Anderson TJ, Haubold B, Williams JT, Estrada-Franco JG, Richardson L, Mollinedo R. Microsatellite markers reveal a spectrum of population structures in the malaria parasite Plasmodium falciparum. Mol Biol Evol. 2000;17. doi:10.1093/oxfordjournals.molbev.a026247

13. Manske M, Miotto O, Campino S, Auburn S, Almagro-Garcia J, Maslen G, et al. Analysis of Plasmodium falciparum diversity in natural infections by deep sequencing. Nature. Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved.; 2012;487: 375–379. Available:

http://dx.doi.org/10.1038/nature11174

14. Wootton JC, Feng X, Ferdig MT. Genetic diversity and chloroquine selective sweeps in Plasmodium falciparum. 2002;418: 18–21.

15. Alam MT, de Souza DK, Vinayak S, Griffing SM, Poe AC, Duah NO, et al. Selective Sweeps and Genetic Lineages of Plasmodium falciparum Drug - Resistant Alleles in Ghana. J Infect Dis . 2011;203: 220–227. doi:10.1093/infdis/jiq038

16. McCollum AM, Mueller K, Villegas L, Udhayakumar V, Escalante AA. Common Origin and Fixation of Plasmodium falciparum dhfr and dhps Mutations Associated with Sulfadoxine-Pyrimethamine Resistance in a Low-Transmission Area in South America. Antimicrob Agents Chemother. 2007;51: 2085–2091. doi:10.1128/AAC.01228-06

17. Yalcindag E, Elguero E, Arnathau C, Durand P, Akiana J, Anderson TJ, et al. Multiple independent introductions of Plasmodium falciparum in South America. Proc Natl Acad Sci . 2012;109: 511–516. Available: http://www.pnas.org/content/109/2/511.abstract

18. Ariey F, Witkowski B, Amaratunga C, Beghain J, Langlois A-C, Khim N, et al. A molecular marker of artemisinin-resistant Plasmodium falciparum malaria. Nature. Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved.; 2014;505: 50–55. Available: http://dx.doi.org/10.1038/nature12876

19. Neafsey DE, Volkman SK. Malaria Genomics in the Era of Eradication.

Cold Spring Harb Perspect Med . 2017;7. doi:10.1101/cshperspect.a025544

20. Volkman SK, Neafsey DE, Schaffner SF, Park DJ, Wirth DF. Harnessing genomics and genome biology to understand malaria biology. Nat Rev Genet. Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved.; 2012;13: 315–328. Available: http://dx.doi.org/10.1038/nrg3187

21. Cerqueira GC, Cheeseman IH, Schaffner SF, Nair S, McDew-White M, Phyo AP, et al. Longitudinal genomic surveillance of Plasmodium falciparum malaria parasites reveals complex genomic architecture of emerging artemisinin resistance. Genome Biol. London: BioMed Central; 2017;18: 78. doi:10.1186/s13059-017-1204-4

22. Miotto O, Amato R, Ashley EA, Macinnis B, Dhorda M, Imwong M, et al. Genetic architecture of artemisinin-resistant Plasmodium Falciparum. 2015;47: 226–234. doi:10.1038/ng.3189.Genetic

23. Smith T, Maire N, Dietz K, Killeen GF, Vounatsou P, Molineaux L, et al. Relationship between the entomologic inoculation rate and the force of infection for Plasmodium falciparum malaria. Am J Trop Med Hyg. 2006;75.

24. Daniels R, Chang H-H, Séne PD, Park DC, Neafsey DE, Schaffner SF, et al. Genetic Surveillance Detects Both Clonal and Epidemic Transmission of Malaria following Enhanced Intervention in Senegal. PLoS One. Public Library of Science; 2013;8: e60780. Available:

http://dx.doi.org/10.1371%2Fjournal.pone.0060780

25. Nkhoma SC, Nair S, Al-Saai S, Ashley E, McGready R, Phyo AP, et al. Population genetic correlates of declining transmission in a human pathogen. Mol Ecol. Blackwell Publishing Ltd; 2013;22: 273–285. doi:10.1111/mec.12099

26. Arnot D. Unstable malaria in Sudan: the influence of the dry season: Clone multiplicity of Plasmodium falciparum infections in individuals exposed to variable levels of disease transmission. Trans R Soc Trop Med Hyg . 1998;92: 580–585. doi:10.1016/S0035-9203(98)90773-8

27. Yang Z, Rannala B. Molecular phylogenetics: principles and practice. Nat Rev Genet. Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved.; 2012;13: 303–314. Available: http://dx.doi.org/10.1038/nrg3186

28. Didelot X, Gardy J, Colijn C. Bayesian Inference of Infectious Disease Transmission from Whole-Genome Sequence Data. Mol Biol Evol. 2014;31: 1869–1879. Available: http://dx.doi.org/10.1093/molbev/msu121

29. Holmes EC, Dudas G, Rambaut A, Andersen KG. The evolution of Ebola virus: Insights from the 2013–2016 epidemic. Nature. Macmillan Publishers Limited, part of Springer Nature. All rights reserved.; 2016;538: 193–200. Available: http://dx.doi.org/10.1038/nature19790

30. Mena I, Nelson MI, Quezada-Monroy F, Dutta J, Cortes-Fernández R, Lara-Puente JH, et al. Origins of the 2009 H1N1 influenza pandemic in

swine in Mexico. Neher RA, editor. Elife. eLife Sciences Publications, Ltd; 2016;5: e16777. doi:10.7554/eLife.16777

31. Pybus OG, Rambaut A. Evolutionary analysis of the dynamics of viral infectious disease. Nat Rev Genet. Nature Publishing Group; 2009;10: 540. Available: http://dx.doi.org/10.1038/nrg2583

32. Lemey P, Pybus OG, Rambaut A, Drummond AJ, Robertson DL, Roques P, et al. The molecular population genetics of HIV-1 group O. Genetics. United States; 2004;167: 1059–1068. doi:10.1534/genetics.104.026666

33. Frost SDW, Pybus OG, Gog JR, Viboud C, Bonhoeffer S, Bedford T. Eight challenges in phylodynamic inference. Epidemics. 2015;10: 88–92. doi:https://doi.org/10.1016/j.epidem.2014.09.001

34. Yang T, Deng H-W, Niu T. Critical assessment of coalescent simulators in modeling recombination hotspots in genomic sequences. BMC Bioinformatics. 2014;15: 3. doi:10.1186/1471-2105-15-3

35. Conway DJ, Roper C, Oduola AMJ, Arnot DE, Kremsner PG, Grobusch MP, et al. High recombination rate in natural populations of Plasmodium falciparum. Proc Natl Acad Sci . 1999;96: 4506–4511. doi:10.1073/pnas.96.8.4506

36. Alizon S, de Roode JC, Michalakis Y. Multiple infections and the evolution of virulence. Ecol Lett. 2013;16: 556–567. doi:10.1111/ele.12076

37. Alizon S. Co-infection and super-infection models in evolutionary epidemiology. Interface Focus. The Royal Society; 2013;3: 20130031.

doi:10.1098/rsfs.2013.0031

38. Anderson TJ, Haubold B, Williams JT, Estrada-Franco JG, Richardson L, Mollinedo R, et al. Microsatellite markers reveal a spectrum of population structures in the malaria parasite Plasmodium falciparum. Mol Biol Evol. 2000;17: 1467–1482. doi:10.1093/oxfordjournals.molbev.a026247

39. Obaldia N 3rd, Baro NK, Calzada JE, Santamaria AM, Daniels R, Wong W, et al. Clonal Outbreak of Plasmodium falciparum Infection in Eastern Panama. J Infect Dis. United States; 2015;211: 1087–1096. doi:10.1093/infdis/jiu575

40. Daniels RF, Schaffner SF, Wenger EA, Proctor JL, Chang H-H, Wong W, et al. Modeling malaria genomics reveals transmission decline and rebound in Senegal. Proc Natl Acad Sci . 2015;112: 7067–7072. Available: http://www.pnas.org/content/112/22/7067.abstract

41. Conway DJ, Greenwood BM, McBride JS. The epidemiology of multiple-clone Plasmodium falciparum infections in Gambian patients. Parasitology. 1991;103: 1–5.

42. Nkhoma SC, Nair S, Cheeseman IH, Rohr-Allegrini C, Singlam S, Nosten F, et al. Close kinship within multiple-genotype malaria parasite infections. Proc R Soc London B Biol Sci. 2012; Available: http://rspb.royalsocietypublishing.org/content/early/2012/02/28/rspb.2012.0113.abstract

43. Nair S, Nkhoma SC, Serre D, Zimmerman PA, Gorena K, Daniel BJ, et al.

Single-cell genomics for dissection of complex malaria infections. Genome Res. Cold Spring Harbor Laboratory Press; 2014;24: 1028–1038. doi:10.1101/gr.168286.113

44. Sutton PL, Neyra V, Hernandez JN, Branch OH. Plasmodium falciparum and Plasmodium vivax Infections in the Peruvian Amazon: Propagation of Complex, Multiple Allele-Type Infections without Super-Infection. Am J Trop Med Hyg. 2009;81: 950–960. doi:10.4269/ajtmh.2009.09-0132

45. Ross R. Some Quantitative Studies in Epidemiology. Nature. 1911; 466–467.

46. Lion S. Multiple infections, kin selection and the evolutionary epidemiology of parasite traits. J Evol Biol. Switzerland; 2013;26: 2107–2122. doi:10.1111/jeb.12207

47. van Baalen M, Sabelis MW. The Dynamics of Multiple Infection and the Evolution of Virulence. Am Nat. [University of Chicago Press, American Society of Naturalists]; 1995;146: 881–910. Available: http://www.jstor.org.ezp-prod1.hul.harvard.edu/stable/2463102

48. Smith DL, Battle KE, Hay SI, Barker CM, Scott TW, McKenzie FE. Ross, Macdonald, and a Theory for the Dynamics and Control of Mosquito-Transmitted Pathogens. Chitnis CE, editor. PLoS Pathog. San Francisco, USA: Public Library of Science; 2012;8: e1002588. doi:10.1371/journal.ppat.1002588

49. Trape JF, Lefebvre-Zante E, Legros F, Ndiaye G, Bouganali H, Druilhe P,

et al. Vector density gradients and the epidemiology of urban malaria in Dakar, Senegal. Am J Trop Med Hyg. 1992;47.

50. Mouzin E, Thior P, Diouf M, Sambou B. Focus on Senegal Roll Back Malaria: Progress and Impact Series.

51. Daniels R, Volkman S, Milner D, Mahesh N, Neafsey D, Park D, et al. A general SNP-based molecular barcode for Plasmodium falciparum identification and tracking. Malar J. 2008;7: 223. Available: http://www.malariajournal.com/content/7/1/223

52. Li H, Durbin R. Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinforma . 2009;25: 1754–1760. doi:10.1093/bioinformatics/btp324

53. Aurrecoechea C, Brestelli J, Brunk BP, Dommer J, Fischer S, Gajria B, et al. PlasmoDB: a functional genomic database for malaria parasites. Nucleic Acids Res . 2009;37: D539–D543. doi:10.1093/nar/gkn814

54. DePristo MA, Banks E, Poplin R, Garimella K V, Maguire JR, Hartl C, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet. Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved.; 2011;43: 491–498. Available: http://dx.doi.org/10.1038/ng.806

55. Galinsky K, Valim C, Salmier A, de Thoisy B, Musset L, Legrand E, et al. COIL: a methodology for evaluating malarial complexity of infection using likelihood from single nucleotide polymorphism data. Malar J. BioMed

Central Ltd; 2015;14: 4. doi:10.1186/1475-2875-14-4

56. Chang H-H, Park DJ, Galinsky KJ, Schaffner SF, Ndiaye D, Ndir O, et al. Genomic Sequencing of Plasmodium falciparum Malaria Parasites from Senegal Reveals the Demographic History of the Population. Mol Biol Evol. 2012;29: 3427–3439. doi:10.1093/molbev/mss161

57. Browning SR, Browning BL. Haplotype phasing: existing methods and new developments. Nat Rev Genet. Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved.; 2011;12: 703–714. Available: http://dx.doi.org/10.1038/nrg3054

58. Bousema T, Griffin JT, Sauerwein RW, Smith DL, Churcher TS, Takken W, et al. Hitting Hotspots: Spatial Targeting of Malaria for Control and Elimination. PLoS Med. Public Library of Science; 2012;9: e1001165. Available: http://dx.doi.org/10.1371%2Fjournal.pmed.1001165

59. Vaughan AM, Pinapati RS, Cheeseman IH, Camargo N, Fishbaugher M, Checkley LA, et al. Plasmodium falciparum genetic crosses in a humanized mouse model. Nat Methods. 2015;12: 631–633. doi:10.1038/nmeth.3432

60. Chang H-H, Childs LM, Buckee CO. Variation in infection length and superinfection enhance selection efficiency in the human malaria parasite. Sci Rep. The Author(s); 2016;6: 26370. Available: http://dx.doi.org/10.1038/srep26370

61. Klein EY, Smith DL, Laxminarayan R, Levin S. Superinfection and the evolution of resistance to antimalarial drugs. Proc R Soc B Biol Sci. The

Royal Society; 2012;279: 3834–3842. doi:10.1098/rspb.2012.1064

62. Childs LM, Buckee CO. Dissecting the determinants of malaria chronicity: why within-host models struggle to reproduce infection dynamics. J R Soc Interface. The Royal Society; 2015;12: 20141379. doi:10.1098/rsif.2014.1379

63. Dye C, Williams BG. Multigenic drug resistance among inbred malaria parasites. Proc R Soc London Ser B Biol Sci. 1997;264: 61 LP-67. Available: http://rspb.royalsocietypublishing.org/content/264/1378/61.abstract

64. Wong W, Griggs AD, Daniels RF, Schaffner SF, Ndiaye D, Bei AK, et al. Genetic relatedness analysis reveals the cotransmission of genetically related Plasmodium falciparum parasites in Thiès, Senegal. Genome Med. 2017;9: 5. doi:10.1186/s13073-017-0398-0

65. Branch OH, Takala S, Kariuki S, Nahlen BL, Kolczak M, Hawley W, et al. Plasmodium falciparum genotypes, low complexity of infection, and resistance to subsequent malaria in participants in the asembo bay cohort project. Infect Immun. 2001;69. doi:10.1128/IAI.69.12.7783-7792.2001

66. Henden L, Lee S, Mueller I, Barry A, Bahlo M. Detecting Selection Signals In Plasmodium falciparum Using Identity-By-Descent Analysis. bioRxiv. 2016; Available: http://biorxiv.org/content/early/2016/11/16/088039.abstract

67. Su X, Ferdig MT, Huang Y, Huynh CQ, Liu A, You J, et al. A Genetic Map and Recombination Parameters of the Human Malaria Parasite

Plasmodium falciparum. Science (80-). 1999;286: 1351–1353. Available: http://science.sciencemag.org/content/286/5443/1351.abstract

68. Miles A, Iqbal Z, Vauterin P, Pearson R, Campino S, Theron M, et al. Genome variation and meiotic recombination in Plasmodium falciparum: insights from deep sequencing of genetic crosses. bioRxiv. 2015; Available: http://biorxiv.org/content/early/2015/12/23/024182.abstract

69. Wellems TE, Panton LJ, Gluzman IY, do Rosario VE, Gwadz RW, Walker-Jonah A, et al. Chloroquine resistance not linked to mdr-like genes in a Plasmodium falciparum cross. Nature. ENGLAND; 1990;345: 253–255. doi:10.1038/345253a0

70. Hayton K, Gaur D, Liu A, Takahashi J, Henschen B, Singh S, et al. Erythrocyte binding protein PfRH5 polymorphisms determine species-specific pathways of Plasmodium falciparum invasion. Cell Host Microbe. United States; 2008;4: 40–51. doi:10.1016/j.chom.2008.06.001

71. Walliker D, Quakyi IA, Wellems TE, McCutchan TF, Szarfman A, London WT, et al. Genetic analysis of the human malaria parasite Plasmodium falciparum. Science. UNITED STATES; 1987;236: 1661–1666.

72. Conway DJ, McBride JS. Population genetics of Plasmodium falciparum within a malaria hyperendemic area. Parasitology. 1991;103. doi:10.1017/S0031182000059229

73. Schaffner SF, Taylor AR, Wong W, Wirth DF, Neafsey DE. hmmIBD: software to infer pairwise identity by descent between haploid genotypes.

bioRxiv. 2017; Available: http://biorxiv.org/content/early/2017/09/12/188078.abstract

74. Berchowitz LE, Copenhaver GP. Genetic Interference: Don't Stand So Close to Me. Curr Genomics. Bentham Science Publishers Ltd.; 2010;11: 91–102. doi:10.2174/138920210790886835

75. Falque M, Mercier R, Mézard C, de Vienne D, Martin OC. Patterns of Recombination and MLH1 Foci Density Along Mouse Chromosomes: Modeling Effects of Interference and Obligate Chiasma. Genetics. 2007;176: 1453–1467. Available: http://www.genetics.org/content/176/3/1453.abstract

76. McPeek MS, Speed TP. Modeling Interference in Genetic Recombination. Genetics. 1995;139: 1031–1044. Available: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1206354/

77. Krishnaprasad GN, Anand MT, Lin G, Tekkedil MM, Steinmetz LM, Nishant KT. Variation in Crossover Frequencies Perturb Crossover Assurance Without Affecting Meiotic Chromosome Segregation in &lt;em&gt;Saccharomyces cerevisiae&lt;/em&gt; Genetics. 2015;199: 399 LP-412. Available: http://www.genetics.org/content/199/2/399.abstract

78. Stone WJR, Eldering M, van Gemert G-J, Lanke KHW, Grignard L, van de Vegte-Bolmer MG, et al. The relevance and applicability of oocyst prevalence as a read-out for mosquito feeding assays. Sci Rep. The Author(s); 2013;3: 3418. Available: http://dx.doi.org/10.1038/srep03418

79. Gnémé A, Guelbéogo WM, Riehle MM, Sanou A, Traoré A, Zongo S, et al. Equivalent susceptibility of Anopheles gambiae M and S molecular forms and Anopheles arabiensis to Plasmodium falciparum infection in Burkina Faso. Malar J. 2013;12: 1–12. doi:10.1186/1475-2875-12-204

80. Bejon P, Andrews L, Andersen RF, Dunachie S, Webster D, Walther M, et al. Calculation of Liver-to-Blood Inocula, Parasite Growth Rates, and Preerythrocytic Vaccine Efficacy, from Serial Quantitative Polymerase Chain Reaction Studies of Volunteers Challenged with Malaria Sporozoites. J Infect Dis . 2005;191: 619–626. doi:10.1086/427243

81. Lin Ouédraogo A, Gonçalves BP, Gnémé A, Wenger EA, Guelbeogo MW, Ouédraogo A, et al. Dynamics of the Human Infectious Reservoir for Malaria Determined by Mosquito Feeding Assays and Ultrasensitive Malaria Diagnosis in Burkina Faso. J Infect Dis. 2016;213: 90–99. Available: http://dx.doi.org/10.1093/infdis/jiv370

82. Molineaux L, Gramiccia G. The Garki Project: Research on the Epidemiology and Control of Malaria in the Sudan Savanna of West Africa. Geneva: World Health Organization; 1980.

83. Felger I, Maire M, Bretscher MT, Falk N, Tiaden A, Sama W, et al. The Dynamics of Natural Plasmodium falciparum Infections. Gosling RD, editor. PLoS One. San Francisco, USA: Public Library of Science; 2012;7: e45542. doi:10.1371/journal.pone.0045542

84. Lindblade KA, Steinhardt L, Samuels A, Kachur SP, Slutsker L. The silent

threat: asymptomatic parasitemia and malaria transmission. Expert Rev Anti Infect Ther. Taylor & Francis; 2013;11: 623–639. doi:10.1586/eri.13.45

85. Ouedraogo AL, de Vlas SJ, Nebie I, Ilboudo-Sanogo E, Bousema JT, Ouattara AS, et al. Seasonal patterns of Plasmodium falciparum gametocyte prevalence and density in a rural population of Burkina Faso. Acta Trop. 2008;105. doi:10.1016/j.actatropica.2007.09.003

86. McKenzie FE, Killeen GF, Beier JC, Bossert WH. Seasonality, parasite diversity, and local extinctions in Plasmodium falciparum malaria. Ecology. 2001;82. doi:10.1890/0012-9658(2001)082[2673:SPDALE]2.0.CO;2

87. Artzy-Randrup Y, Rorick MM, Day K, Chen D, Dobson AP, Pascual M. Population structuring of multi-copy, antigen-encoding genes in Plasmodium falciparum. Bergstrom CT, editor. Elife. eLife Sciences Publications, Ltd; 2012;1: e00093. doi:10.7554/eLife.00093

88. Bruce MC, Galinski MR, Barnwell JW, Donnelly CA, Walmsley M, Alpers MP, et al. Genetic diversity and dynamics of plasmodium falciparum and P. vivax populations in multiply infected children with asymptomatic malaria infections in Papua New Guinea. Parasitology. England; 2000;121 (Pt 3): 257–272.

89. Nwakanma D, Kheir A, Sowa M, Dunyo S, Jawara M, Pinder M, et al. High gametocyte complexity and mosquito infectivity of Plasmodium falciparum in the Gambia. Int J Parasitol. 2008;38: 219–227. doi:https://doi.org/10.1016/j.ijpara.2007.07.003

90. Owusu-Agyei S, Smith T, Beck H-P, Amenga-Etego L, Felger I. Molecular epidemiology of Plasmodium falciparum infections among asymptomatic inhabitants of a holoendemic malarious area in northern Ghana. Trop Med Int Heal. Blackwell Science Ltd; 2002;7: 421–428. doi:10.1046/j.1365-3156.2002.00881.x

91. Pichon G, Awono-Ambene HP, Robert V. High heterogeneity in the number of Plasmodium falciparum gametocytes in the bloodmeal of mosquitoes fed on the same host. Parasitology. ENGLAND; 2000;121 (Pt 2: 115–120.

92. Bousema T, Dinglasan RR, Morlais I, Gouagna LC, van Warmerdam T, Awono-Ambene PH, et al. Mosquito feeding assays to determine the infectiousness of naturally infected Plasmodium falciparum gametocyte carriers. PLoS One. 2012;7. doi:10.1371/journal.pone.0042821

93. Broman KW, Weber JL. Characterization of human crossover interference. Am J Hum Genet. 2000;66: 1911–1926. Available: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1378063/

94. Mzilahowa T, McCall PJ, Hastings IM. “Sexual” Population Structure and Genetics of the Malaria Agent P. falciparum. PLoS One. Public Library of Science; 2007;2: e613. Available: http://dx.plos.org/10.1371/journal.pone.0000613

95. Witkowski B, Duru V, Khim N, Ross LS, Saintpierre B, Beghain J, et al. A surrogate marker of piperaquine-resistant <em>Plasmodium falciparum</em> malaria: a phenotype&#x2013;genotype association

study. Lancet Infect Dis. Elsevier; 2017;17: 174–183. doi:10.1016/S1473-3099(16)30415-7

96. Amato R, Lim P, Miotto O, Amaratunga C, Dek D, Pearson RD, et al. Genetic markers associated with dihydroartemisinin&#x2013;piperaquine failure in <em>Plasmodium falciparum</em> malaria in Cambodia: a genotype&#x2013;phenotype association study. Lancet Infect Dis. Elsevier; 2017;17: 164–173. doi:10.1016/S1473-3099(16)30409-1

97. Auburn S, Campino S, Clark TG, Djimde AA, Zongo I, Pinches R. An effective method to purify Plasmodium falciparum DNA directly from clinical blood samples for whole genome high-throughput sequencing. PLoS One. 2011;6. doi:10.1371/journal.pone.0022213

98. Melnikov A, Galinsky K, Rogov P, Fennell T, Tyne D, Russ C. Hybrid selection for sequencing pathogen genomes from clinical samples. Genome Biol. 2011;12. doi:10.1186/gb-2011-12-8-r73

99. Leichty AR, Brisson D. Selective whole genome amplification for resequencing target microbial species from complex natural samples. Genetics. 2014;198. doi:10.1534/genetics.114.165498

100. Guggisberg AM, Sundararaman SA, Lanaspa M, Moraleda C, González R, Mayor A. Whole genome sequencing to evaluate the resistance landscape following antimalarial treatment failure with fosmidomycin-clindamycin. J Infect Dis. 2016;214. doi:10.1093/infdis/jiw304

101. Oyola SO, Gu Y, Manske M, Otto TD, O'Brien J, Alcock D. Efficient

depletion of host DNA contamination in malaria clinical sequencing. J Clin Microbiol. 2013;51. doi:10.1128/JCM.02507-12

102. Dean FB, Nelson JR, Giesler TL, Lasken RS. Rapid Amplification of Plasmid and Phage DNA Using Phi29 DNA Polymerase and Multiply-Primed Rolling Circle Amplification. Genome Res. Cold Spring Harbor Laboratory Press; 2001;11: 1095–1099. doi:10.1101/gr.180501

103. Sundararaman SA, Plenderleith LJ, Liu W, Loy DE, Learn GH, Li Y, et al. Genomes of cryptic chimpanzee Plasmodium species reveal key evolutionary events leading to human malaria. Nat Commun. The Author(s); 2016;7: 11078. Available: http://dx.doi.org/10.1038/ncomms11078

104. Cowman AF, Morry MJ, Biggs BA, Cross GA, Foote SJ. Amino acid changes linked to pyrimethamine resistance in the dihydrofolate reductase-thymidylate synthase gene of Plasmodium falciparum. Proc Natl Acad Sci USA. 1988;85. doi:10.1073/pnas.85.23.9109

105. Galinsky K, Valim C, Salmier A, Thoisy B, Musset L, Legrand E. COIL: a methodology for evaluating malarial complexity of infection using likelihood from single nucleotide polymorphism data. Malar J. 2015;14. doi:10.1186/1475-2875-14-4

106. Oyola SO, Ariani C V, Hamilton WL, Kekre M, Amenga-Etego LN, Ghansah A, et al. Whole genome sequencing of Plasmodium falciparum from dried blood spots using selective whole genome amplification. Malar

J. 2016;15: 597. doi:10.1186/s12936-016-1641-7

107. Guinet F, Dvorak JA, Fujioka H, Keister DB, Muratova O, Kaslow DC, et al. A developmental defect in Plasmodium falciparum male gametogenesis. J Cell Biol. The Rockefeller University Press; 1996;135: 269–278. Available: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2121010/

108. Ponnudurai T, Leeuwenberg AD, Meuwissen JH. Chloroquine sensitivity of isolates of Plasmodium falciparum adapted to in vitro culture. Trop Geogr Med. Netherlands; 1981;33: 50–54.

109. Preston MD, Campino S, Assefa SA, Echeverry DF, Ocholla H, Amambua-Ngwa A, et al. A barcode of organellar genome polymorphisms identifies the geographic origin of Plasmodium falciparum strains. The Author(s); 2014;5: 4052. Available: http://dx.doi.org/10.1038/ncomms5052

110. Bousema T, Okell L, Felger I, Drakeley C. Asymptomatic malaria infections: detectability, transmissibility and public health relevance. Nat Rev Microbiol. Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved.; 2014;12: 833. Available: http://dx.doi.org/10.1038/nrmicro3364

111. Lin JT, Saunders DL, Meshnick SR. The role of submicroscopic parasitemia in malaria transmission: what is the evidence? Trends Parasitol. Elsevier; 2017;30: 183–190. doi:10.1016/j.pt.2014.02.004

112. Okell LC, Bousema T, Griffin JT, Ouédraogo AL, Ghani AC, Drakeley CJ. Factors determining the occurrence of submicroscopic malaria infections

and their relevance for control. Nat Commun. The Author(s); 2012;3: 1237. Available: http://dx.doi.org/10.1038/ncomms2241

113. Johnson JD, Dennull RA, Gerena L, Lopez-Sanchez M, Roncal NE, Waters NC. Assessment and Continued Validation of the Malaria SYBR Green I-Based Fluorescence Assay for Use in Malaria Drug Screening . Antimicrob Agents Chemother. American Society for Microbiology; 2007;51: 1926–1933. doi:10.1128/AAC.01607-06

114. Auburn S, Barry AE. Dissecting malaria biology and epidemiology using population genetics and genomics. Int J Parasitol. 2017;47: 77–85. doi:https://doi.org/10.1016/j.ijpara.2016.08.006

115. Tusting LS, Bousema T, Smith DL, Drakeley C. Measuring changes in Plasmodium falciparum transmission: Precision, accuracy and costs of metrics. Adv Parasitol. 2014;84: 151–208. doi:10.1016/B978-0-12-800099-1.00003-X

116. Kerkhof K, Sluydts V, Heng S, Kim S, Pareyn M, Willen L, et al. Geographical patterns of malaria transmission based on serological markers for falciparum and vivax malaria in Ratanakiri, Cambodia. Malar J. 2016;15: 510. doi:10.1186/s12936-016-1558-1

117. Gething PW, Patil AP, Smith DL, Guerra CA, Elyazar IRF, Johnston GL, et al. A new world malaria map: Plasmodium falciparum endemicity in 2010. Malar J. 2011;10: 378. doi:10.1186/1475-2875-10-378

118. Perkins TA, Scott TW, Le Menach A, Smith DL. Heterogeneity, Mixing, and

the Spatial Scales of Mosquito-Borne Pathogen Transmission. PLOS Comput Biol. Public Library of Science; 2013;9: e1003327. Available: https://doi.org/10.1371/journal.pcbi.1003327

119. Amino R, Thiberge S, Martin B, Celli S, Shorte S, Frischknecht F, et al. Quantitative imaging of Plasmodium transmission from mosquito to mammal. Nat Med. Nature Publishing Group; 2006;12: 220–224. Available: http://dx.doi.org/10.1038/nm1350

120. Keeling MJ. The effects of local spatial structure on epidemiological invasions. Proc R Soc B Biol Sci. 1999;266: 859–867. Available: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1689913/

121. Arifin SMN, Madey GR, Collins FH. Agent-Based Modeling and Malaria. Spatial Agent-Based Simulation Modeling in Public Health. John Wiley & Sons, Inc; 2016. pp. 17–38. doi:10.1002/9781118964385.ch3

122. Hurford A, Cownden D, Day T. Next-generation tools for evolutionary invasion analyses. J R Soc Interface. The Royal Society; 2010;7: 561–571. doi:10.1098/rsif.2009.0448

123. Leventhal GE, Hill AL, Nowak MA, Bonhoeffer S. Evolution and emergence of infectious diseases in theoretical and real-world networks. Nat Commun. The Author(s); 2015;6: 6101. Available: http://dx.doi.org/10.1038/ncomms7101

124. Taylor AR, Schaffner SF, Cerqueira GC, Nkhoma SC, Anderson TJC, Sriprawat K, et al. Quantifying connectivity between local Plasmodium

falciparum malaria parasite populations using identity by descent. PLOS Genet. Public Library of Science; 2017;13: e1007065. Available: https://doi.org/10.1371/journal.pgen.1007065

125. Maire N, Smith T, Ross A, Owusu-Agyei S, Dietz K, Molineaux L. A model for natural immunity to asexual blood stages of Plasmodium falciparum malaria in endemic areas. Am J Trop Med Hyg. 2006;75.

126. Fidock DA, Nomura T, Talley AK, Cooper RA, Dzekunov SM, Ferdig MT. Mutations in the P. falciparum digestive vacuole transmembrane protein PfCRT and evidence for their role in chloroquine resistance. Mol Cell. 2000;6. doi:10.1016/S1097-2765(05)00077-8

127. Maïga O, Djimdé AA, Hubert V, Renard E, Aubouy A, Kironde F, et al. A Shared Asian Origin of the Triple-Mutant *dhfr* Allele in *Plasmodium falciparum* from Sites across Africa. J Infect Dis. 2007;196: 165–172. doi:10.1086/518512

128. Ohta T. the Nearly Neutral Theory of Molecular. Annu Rev Ecol Syst. 2992;23: 263–286.

129. Hughes AL. Near-Neutrality: the Leading Edge of the Neutral Theory of Molecular Evolution. Ann N Y Acad Sci. 2008;1133: 162–179. doi:10.1196/annals.1438.001.Near-Neutrality

130. Ariey F, Duchemin J-B, Robert V. Metapopulation concepts applied to falciparum malaria and their impacts on the emergence and spread of chloroquine resistance. Infect Genet Evol. Netherlands; 2003;2: 185–192.

131. Pastor-Satorras R, Vespignani A. Epidemic Spreading in Scale-Free Networks. Phys Rev Lett. American Physical Society; 2001;86: 3200–3203. Available: https://link.aps.org/doi/10.1103/PhysRevLett.86.3200

132. May RM, Lloyd AL. Infection dynamics on scale-free networks. Phys Rev E Stat Nonlin Soft Matter Phys. United States; 2001;64: 66112. doi:10.1103/PhysRevE.64.066112

133. Adler FR, Brunet RC. The dynamics of simultaneous infections with altered susceptibilities. Theor Popul Biol. 1991;40: 369–410. doi:https://doi.org/10.1016/0040-5809(91)90061-J

134. Mosquera J, Adler FR. Evolution of Virulence: a Unified Framework for Coinfection and Superinfection. J Theor Biol. 1998;195: 293–313. doi:https://doi.org/10.1006/jtbi.1998.0793

135. Cheesman S, Raza A, Carter R. Mixed Strain Infections and Strain-Specific Protective Immunity in the Rodent Malaria Parasite Plasmodium chabaudi chabaudi in Mice. Infect Immun . 2006;74: 2996–3001. doi:10.1128/IAI.74.5.2996-3001.2006

136. Doolan DL, Dobaño C, Baird JK. Acquired Immunity to Malaria. Clin Microbiol Rev. American Society for Microbiology (ASM); 2009;22: 13–36. doi:10.1128/CMR.00025-08

137. Wu J, Tian L, Yu X, Pattaradilokrat S, Li J, Wang M, et al. Strain-specific innate immune signaling pathways determine malaria parasitemia dynamics and host mortality. Proc Natl Acad Sci U S A. National Academy

of Sciences; 2014;111: E511–E520. doi:10.1073/pnas.1316467111

138. Neafsey DE, Juraska M, Bedford T, Benkeser D, Valim C, Griggs A, et al. Genetic Diversity and Protective Efficacy of the RTS,S/AS01 Malaria Vaccine. N Engl J Med. Massachusetts Medical Society; 2015;373: 2025–2037. doi:10.1056/NEJMoa1505819

139. Ghumra A, Semblat J-P, Ataide R, Kifude C, Adams Y, Claessens A, et al. Induction of Strain-Transcending Antibodies Against Group A PfEMP1 Surface Antigens from Virulent Malaria Parasites. PLoS Pathog. Public Library of Science; 2012;8: e1002665. Available: http://dx.doi.org/10.1371%2Fjournal.ppat.1002665

140. Day KP, Artzy-Randrup Y, Tiedje KE, Rougeron V, Chen DS, Rask TS, et al. Evidence of strain structure in Plasmodium falciparum var gene repertoires in children from Gabon, West Africa. Proc Natl Acad Sci . 2017;114: E4103–E4111. doi:10.1073/pnas.1613018114

141. Eckhoff P. A malaria transmission-directed model of mosquito life cycle and ecology. Malar J. 2011;10. doi:10.1186/1475-2875-10-303

142. Eckhoff PA. Malaria parasite diversity and transmission intensity affect development of parasitological immunity in a mathematical model. Malar J. 2012;11: 419. doi:10.1186/1475-2875-11-419

143. Claessens A, Hamilton WL, Kekre M, Otto TD, Faizullabhoy A, Rayner JC, et al. Generation of Antigenic Diversity in Plasmodium falciparum by Structured Rearrangement of Var Genes During Mitosis. PLOS Genet.

Public Library of Science; 2014;10: e1004812. Available: https://doi.org/10.1371/journal.pgen.1004812

144. Eckhoff P. P. falciparum Infection Durations and Infectiousness Are Shaped by Antigenic Variation and Innate and Adaptive Host Immunity in a Mathematical Model. PLoS One. Public Library of Science; 2012;7: e44950. Available: https://doi.org/10.1371/journal.pone.0044950

145. Eichner M, Diebner HH, Molineaux L, Collins WE, Jeffery GM, Dietz K. Genesis, sequestration and survival of Plasmodium falciparum gametocytes: parameter estimates from fitting a model to malariatherapy data. Trans R Soc Trop Med Hyg. 2001;95. doi:10.1016/S0035-9203(01)90016-1