Genomic variation and evolution of the human malaria parasite Plasmodium falciparum

Genomic variation and evolution of the human malaria parasite *Plasmodium falciparum*

A dissertation presented

By

Hsiao-Han Chang

To

The Department of Organismic and Evolutionary Biology

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

In the subject of

Biology

Harvard University

Cambridge, Massachusetts

April 2013

Advisors: Profs. Daniel L. Hartl and John Wakeley Hsiao-Han Chang

Genomic variation and evolution of the human malaria parasite *Plasmodium falciparum*

Abstract

Malaria is a deadly disease that causes nearly one million deaths each year. Understanding the demographic history of the malaria parasite *Plasmodium falciparum* and the genetic basis of its adaptations to antimalarial treatments and the human immune system is important for developing methods to control and eradicate malaria. To study the long-term demographic history and recent effective size of the population in order to identify genes under selection more efficiently and predict the effectiveness of selection, in Chapter 2 we sequenced the complete genomes of 25 cultured *P. falciparum* isolates from Senegal. In addition, in Chapter 3 we estimated temporal allele frequencies in 24 loci among 528 strains from the same population across six years. Based on genetic diversity of the genome sequences, we estimate the long-term effective population size to be approximately 100,000, and a major population expansion of the parasite population approximately 20,000–40,000 years ago. Based on temporal changes in allele frequencies, however, the recent effective size is estimated to be less than 100 from 2007–2011. The discrepancy may reflect recent aggressive efforts to control malaria in Senegal or migration between populations.

Moreover, using inferred demographic history as a null model for coalescent simulation, we identified candidate genes under selection, including genes identified before, such as *pfcrt* and *PfAMA1*, as well as new candidate genes in Chapter 2. Interestingly, we also found selection against G/C-to-A/T changes that offsets the large mutational bias toward A/T, and two unusual patterns: similar synonymous and nonsynonymous allele-frequency spectra, and 18% of genes having a nonsynonymous-to-synonymous polymorphism ratio greater than 1.

In Chapter 4, we studied how the malaria life cycle affects the evolution of malaria parasites and whether the unique life cycle could explain the observed unusual patterns of polymorphism using computer simulations of the life cycle. The results show that both random genetic drift and efficiency of purifying selection are intensified under the malaria life cycle model when comparing with the commonly used Wright-Fisher model and suggest that the malaria life cycle could possibly lead to the unusual patterns.

Table of Contents

Acknowledgments

Throughout my doctoral studies I received invaluable advice and support from many people, without which the research presented in this thesis would not have been possible. First, I would like to thank my advisor Dan Hartl for giving me the opportunity to work on exciting projects in population genetics of malaria parasites. Dan gave me the freedom to explore the questions I am interested in and helped me whenever I felt lost. Dan has also given me countless comments and advice leading to the improvement of my writing and presentation style. Dan is always very supportive, for any kinds of things I asked him. I cannot thank him enough for such things.

I would like to thank my co-advisor John Wakeley for his constant support. It was great to learn coalescent theory from John, and this is one of the best courses I have ever taken. I enjoyed the discussions with John about all different kinds of questions in population genetics. He has given me many helpful suggestions on all the projects I have been working on, and also been patient with my questions. I will miss the time of deriving equations and deep thinking with him very, very much.

I also owe special thanks to a third mentor, Dan Neafsey. Dan has been a patient mentor. He introduced me to the field of malaria parasites and spent much time teaching me the basics and gave me the guidance I needed. Dan provided valuable advice and comments that contributed to the successful completion of the work reported in this thesis. I also thank Dan for his encouragement and availability every time I wanted to discuss my projects or my career with him.

I am grateful for all the support I received from Hopi Hoekstra. As a committee member, Hopi gave me helpful suggestions and trustful support regarding my abilities in completing this PhD.

I also have to thank Chris Marx for giving me advice. It was great to work with him on microbe evolution in my first year. He has provided insightful discussions about the research. I hope that I could be as lively, enthusiastic, and energetic as Chris.

Next, I would like to thank my previous advisors, Chau-Ti Ting, Hwei-yu Chang, Chuck Langley, and Chuan-Chin Chiao. I would not have had the chance to pursue my PhD at Harvard without the help and training from them. I thank Chau-Ti for introducing me to the field of evolution, motivating me, and encouraging me greatly to extend my education. I thank Hwei-yu for her constant support and encouragement. I thank Chuck for sparking and developing my interests in population genetics. I thank Chuan-Chin for helping me begin my graduate career and valuable suggestions.

My thesis has benefitted greatly from the fruitful collaborations that I have had with researchers at the Broad Institute and Harvard School of Public Health.

Rachel Daniels has been an excellent collaborator. We worked on multiple projects together. I would like to thank her for teaching me all the relevant background in malaria parasites that I did not know. Her encouragement and kind words always made me feel warm and happy inside. Danny Park is extremely smart, nice and helpful. I feel very lucky to have had them as my colleagues. It also was my honor to work with Steve Schaffner. Steve is one of the smartest scientists I have ever met while he is so humble and extremely friendly and patient. I would like to thank him for his scientific advice and knowledge and many insightful discussions and suggestions. I also owe my gratitude to Sarah Volkman and Dyann Wirth. Their different perspectives on the important issues in my research have been invaluable.

I am also indebted to the whole Hartl lab, past and present, for valuable contributions to my research and for advice on presentations during my graduate work. These include Pan-Pan Jiang, Russ Corbett-Detig, Jenny Pham, Kerry Geiler-Samerotte, Rebekah Rogers, Mauricio Carneiro, Marna Costanzo, Tim Sackton, Jun Zhou, Luciana Araripe, Horácio Montenegro, Bernardo Lemos, So Nakagawa, Marjorie Liénard, Julien Ayroles, Lene Martinsen, Kalsang Namgyal, Elena Lozovsky, Rubing Liang, and Mimi Velazquez. Pan-Pan and Jenny are the best officemates ever. I could not imagine how my graduate life would have been without Pan-Pan. She has been an extremely friendly and caring colleague and friend, and amazingly, she could know what I am going to say before I say it from my face and gesture. Jenny has been an excellent audience for listening to all my complaints while preparing the thesis and hunting for postdoctoral jobs. I thank Jenny for her patience and encouragement. I also owe special thanks to Russ and Tim. Russ was the one who was there for me to help me quickly proofread and give critical suggestions. Tim let me know how useful R could be, and this became one of the most important tools for my thesis work.

I would like to thank members at the Wakeley lab for giving me suggestions for my research and presentation. Pleuni Pennings has been an amazing, extremely helpful, interested, and understanding colleague and friend. I will always remember the time we discussed our projects in the café. Pleuni has provided critical advice towards many aspects of my projects and my career.

I am also grateful for all the support I received from my friends in the United States and Taiwan throughout my time here, especially Doory Kim, Yu-Ping Poh and Grace Lee. The weekly lunch meeting with Doory and Yu-Ping has always been the best way to refresh myself during my PhD life. Grace always shared with me how she overcame the frustrating moments during her PhD years and still loves science so much.

I am forever grateful to my parents Shih-Mou Chang and Li-Yu Yang for their support and encouragement throughout my graduate and early education. I would not have even arrived here at Harvard University if it were not due to the support and love from them. It is not very clear to them what I study, but they collected all the news from newspapers related to biology, evolution, and malaria for me, and tried to understand me more. I cannot thank them enough for their unlimited love.

Finally, I would like to extend my deepest gratitude to Cheng-Sheng Lee for coming to Boston with me, providing partnership, moral support, fun and understanding throughout these years. I am extremely thankful to have him in my life.

Dedicated to my parents for their unlimited love, encouragement and support

Chapter 1

Introduction

Mutation, selection, and random genetic drift are three major components of evolution. Mutation provides genetic diversity upon which the forces that lead to evolution can act, and the changes in frequencies of mutations depend on random genetic drift and selection. The predictions of genetic diversity or allele frequency changes are usually based on simple idealized models. Differences between actual data and predictions of these models provide a variety of approaches for drawing inferences about evolution through analysis of genetic data. This thesis describes inferences about the evolution of malaria parasites made from comparing actual data with predictions of simple models, and attempts to understand how the malaria life cycle affects the evolution of malaria parasites using computer simulations of the life cycle.

An idealized model that is commonly used in population genetics is the Wright–Fisher model. The key simplifying assumptions are a finite and constant population size (N), nonoverlapping generations, and random sampling with replacement from alleles in the current generation to produce those in the next generation. In its simplest form, the Wright–Fisher model also assumes no new mutations by dealing only with those that already exist in the population. Random genetic drift is defined as a change in allele frequency due entirely to stochastic sampling from generation to generation in a finite population. The effective size of a population corresponds to the size of the ideal population that would have the same magnitude of random genetic drift as an actual population.

The effective population size is an important characteristic of a population that determines the effects of genetic drift interacting with selection and other processes. The effective size is often more relevant to evolutionary processes such as adaptation than is the census population size. The relative importance of selection and random genetic drift depends on the absolute value of the product of effective population size N_e and selection coefficient s. If $|N_e s|$ » 1, selection dominates over genetic drift and beneficial mutations are more likely to become fixed in the population, while deleterious mutations are efficiently eliminated. If $|N_e s|$ « 1, selection is swamped by drift and the frequency of non-neutral mutations fluctuate as if they were neutral. If $|N_e s| \approx 1$, selection and genetic drift both play important roles in determining the evolutionary fate of a new mutant allele. These principles not only imply that the efficacy of selection increases with N_e but they also imply that more genes are affected by selection in populations with a larger effective population size. Mutations with the same selection coefficient therefore can have different fates in populations with different effective sizes.

Various factors affect effective population size, such as fluctuations in population size, inbreeding, variance in offspring number, and the ratio of females to males. If the population size changes through time, the effective population size is approximately the harmonic mean of the census sizes through time. Effective population size can be directly estimated by genetic diversity if the mutation rate of the organism is known. Under neutrality, genetic diversity is proportional to the product of the effective population size and the mutation rate (π = $4N_e\mu$ and π = $2N_e\mu$ for diploid and haploid systems, respectively). Alternatively, if temporal polymorphism data are available, effective population sizes can be estimated by variance in allele frequency changes. If there are large changes in population size, these two methods could give very different estimates. The first N_e represents the effective population size over longer time periods, whereas the second N_e is more related to current population size. If we want to predict the effectiveness of selection in the near future, the current N_e is more suitable, but if inference of past evolutionary processes is the goal, then the long-term N_e is appropriate.

The changes in population size could be estimated from fitting a demographic model to allele-frequency spectrum of the real data. The allele-frequency spectrum can be represented as a histogram of the distribution of allele frequencies, which depicts the probability distribution of allele frequencies of the frequencies of all the alleles under consideration (such as all alleles of a given gene, or of all synonymous nucleotide sites). In the Wright-Fisher model with constant population size, the proportion of mutations with frequency *i* is proportional to $1/i$. When population size increases, there are more new mutations, and the spectrum will be more skewed toward low frequency alleles compared to the neutral model with constant population size; and when population size decreases, there are fewer new mutations, and the spectrum will be more skewed toward high frequency alleles. Therefore, the allele-frequency spectrum can be used for inferring demographic changes.

In Chapters 2 and 3, we estimated effective population sizes of the population of the human malaria parasite, *Plasmodium falciparum*, in Senegal by using either genetic diversity or temporal changes in allele frequencies. The average genetic diversity of synonymous sites suggests that long-term effective population size is on the order of 100,000. In addition, we fitted a two-epoch model to the *P. falciparum* synonymous allele frequency spectrum and estimated a major 60-fold population expansion of the parasite population from 43,000 to 2.7 million in Senegal approximately 20,000–40,000 years ago. In stark contrast, the estimate of effective population size from temporal changes in allele frequency between 2007 and 2011 is less than 100.

These seemingly contradictory results could be explained by *either of* two *possible* scenarios. First, if the population size of malaria parasites had undergone dramatic reduction very recently, it is possible that the influence of the population reduction on the allele-frequency spectrum would be smaller than that of the earlier population expansion, and therefore we could not yet detect the recent population reduction from the allele-frequency spectrum. Second, if there is migration between Senegal with a small population size and a nearby population with a much larger population size, then migrants could increase the coalescent time and the overall genetic diversity, leading to higher estimate of effective population size while the effect of migrants on temporal changes in allele frequencies is nevertheless small and hard to detect (Figure 1.1 shows a simple example).

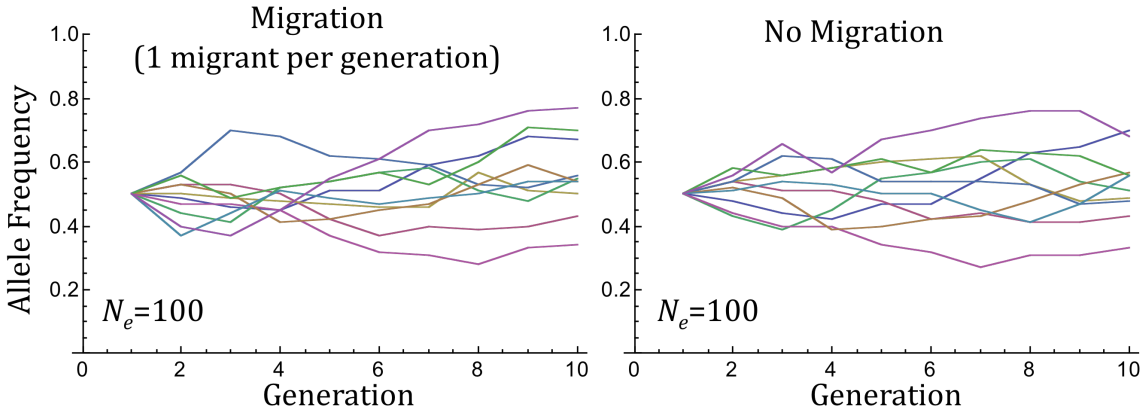


Figure 1.1. Migrants have little effect on allele frequency changes. The left figure shows the pattern of allele frequency changes in a population with size 100 and a migrant from an outside population in each generation, and the right figure shows temporal changes in allele frequencies when there is no migrant. The initial allele frequency is 0.5 and 10 replicates are simulated for 10 generations. There is only one type of allele present in the outside population, and both figures show the frequency of this allele type.

The allele-frequency spectrum can also be used for detecting genes under selection (Figure 1.2). Purifying selection removes deleterious alleles from the population and leads to a spectrum that is skewed toward lower frequencies compared to the neutral expectation. Positive selection increases the frequency of beneficial alleles and the fraction of high-frequency alleles. Because demographic history can alter the null distribution under the neutral hypothesis, if we detect demographic changes, it is better to estimate the null distribution of allele frequency spectrum by coalescent simulation rather than to assume a constant population size under the neutral hypothesis.

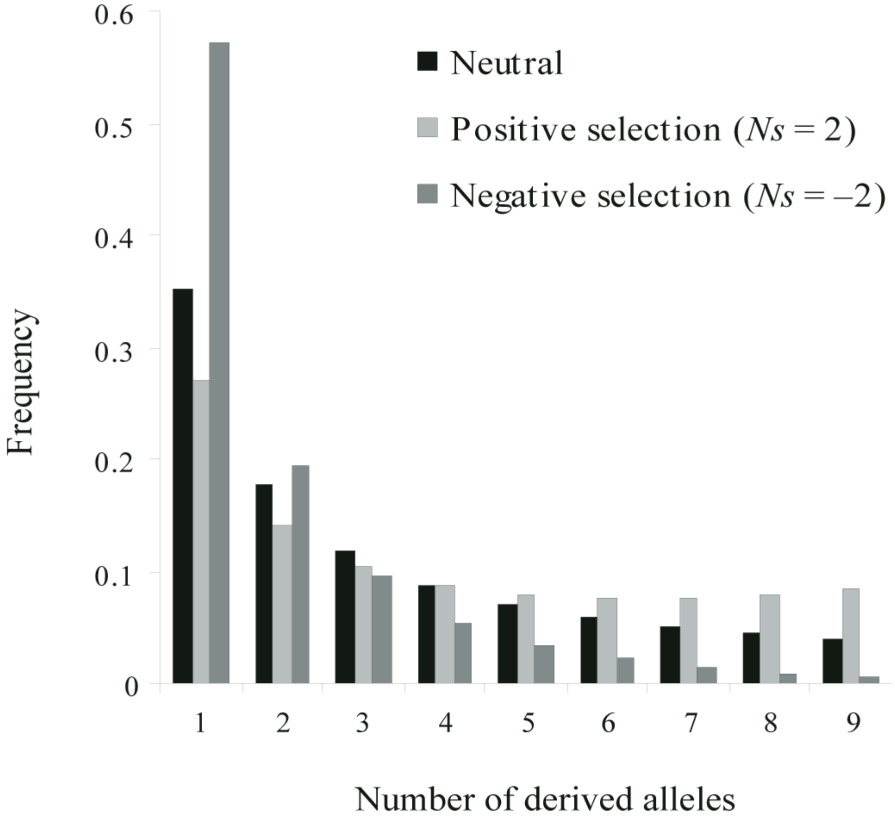


Figure 1.2. Allele frequency spectra under neutrality, positive selection, and negative selection. Equation 11 from Sawyer and Hartl (1992) was used to generate the spectra under positive and negative selection. The example here has sample size equal to 10.

In Chapter 2, besides estimating demographic parameters, we also looked for signals of selection by comparing the observed allele-frequency spectra of genes with the null allele-frequency spectrum under the neutral model after taking into account the estimated demographic history. In addition, we describe two unusual patterns of genetic diversity in *P. falciparum* genome that have not to our knowledge been reported in any other organisms.

We discussed the possible explanations for these unusual patterns of genetic diversity in Chapter 2. Because malaria parasites have a complex life cycle that includes two types of hosts, in Chapter 4 we used computer simulation to study whether the malaria life cycle could lead to the observed unusual patterns. Chapter 4 shows important qualitative differences in population genetic behavior between the Wright-Fisher model and the malaria life cycle model, and explains how these differences could possibly lead to the unusual patterns observed. Because of the intrinsic differences between the Wright-Fisher model and the malaria life cycle model, knowing the effective population size itself is not sufficient information for understanding the interaction between random genetic drift and selection, and further approaches are needed.

This thesis describes the evolutionary forces shaping the genome variation of malaria parasites in Senegal, and shows an important case where a complex life cycle can lead to a totally different population genetic behavior than expected from the classical Wright-Fisher model. Other parasite species that are transmitted among hosts with population-size expansion within hosts may also evolve in a way that is qualitatively different from the Wright-Fisher model, and this study cautions of misinterpretation when analyzing data based on standard population genetic methods in organisms with unconventional life histories. ¶

¶ Part of this chapter has been published in Chapter 1 of "Evolution of Virulence in Eukaryotic Microbes" (ISBN: 978-1-1180-3818-5) by Wiley-Blackwell on 2012 (Pp. 3-16.)

Chapter 2

Genomic sequencing of *Plasmodium falciparum* malaria parasites from Senegal reveals the demographic history of the population[§]

Hsiao-Han Chang, Daniel J. Park, Kevin J. Galinsky, Stephen F. Schaffner, Daouda Ndiaye, Omar Ndir, Soulyemane Mboup, Roger C. Wiegand, Sarah K. Volkman, Pardis C. Sabeti, Dyann F. Wirth, Daniel E. Neafsey, and Daniel L. Hartl

[§] This chapter has been published by Oxford University Press on behalf of the Society for Molecular Biology and Evolution on June 24, 2012 (Volume 29, Issue 11, Pp. 3427-3439).

2.1 Introduction

Malaria caused by *Plasmodium falciparum* is one of the major fatal diseases in the world that infects over 200 million people and causes nearly one million deaths annually (WHO World Malaria Report 2011). The main strategies for controlling the disease have been chemotherapy and mosquito control; however, the emergence of drug-resistant parasites and insecticide-resistant mosquitoes allowed a resurgence of malaria (reviewed in Hartl 2004). Understanding the demographic history and evolutionary forces shaping sequence variation across the genome has practical implications for developing methods of disease control.

Studying the demographic history of malaria parasites is useful for multiple reasons. First, inferences about the timing and magnitude of population size changes can provide clues to the causes of those changes. In addition, demographic history and natural selection are the two major forces affecting genome variation, and therefore attaining a thorough understanding of selection from genome variation requires disentangling its effects from those of demography. Genes involved in evading the natural defenses of the human immune system or offering resistance to antimalarial drugs are under strong selective pressure. To identify selective sweeps and drug resistant alleles, the influence of demographic history or population substructure on genome variation must be considered. For example, in order to evaluate evidence for selection, it is important to have a null distribution of a population genetic statistic, such as Tajima's *D* (Tajima 1989a), under the proper demographic model.

Previous studies have given evidence for worldwide population structure and identified some regions with signatures of recent selective sweeps in drug-resistant parasites (Anderson et al. 2000; Conway et al. 2000; Joy et al. 2003; Mu et al. 2005; Volkman et al. 2007). However, previous studies of demographic changes of *P. falciparum* are not completely consistent (Hartl et al. 2002; Su, Mu, and Joy 2003; Hartl 2004), and this question was strongly contested 10 years ago. The original studies could not adequately account for demographic changes because they were based on too few loci, had problems of ascertainment bias, or included too few parasite isolates from each sampled location. Some evidence suggests that the current worldwide population derives from a small number of parasites in the recent past, inferred from the paucity of genetic diversity in synonymous and non-coding regions (Rich et al. 1998; Conway et al. 2000; Volkman et al. 2001); however, regions with high polymorphism suggest that the current worldwide populations are descended from multiple ancient lineages (Verra and Hughes 2000; Hughes and Verra 2001; Polley and Conway 2001; Volkman et al. 2002). Among the clearest studies is that of Joy et al. (2003), who studied the 6-kb mitochondrial genome of 96 worldwide parasites and found evidence supporting a demographic model in which some ancient lineages emerged out of Africa 50,000 to 100,000 years ago and a major population expansion occurred in Africa about 10,000 years ago followed by extensive migration to other regions. This model helped explain the conflicting patterns of genetic diversity in the previous studies. Because of recombination breaking up linkage, different regions in the genome have different demographic histories. Regions having their most recent common ancestor (MRCA) before migration out of Africa can have high diversity while regions having their MRCA after migration out of Africa have lower genetic diversity. The estimated times are point estimates, however, and lack of recombination in mitochondrial DNA leaves many details about demographic history in doubt. Moreover, owing again to the lack of recombination, the estimate from mitochondrial sequences could be biased by selection.

In order to make stronger inferences about demographic history, identify genes under selection, and understand the genome-wide patterns of genetic diversity in *P. falciparum*, we fully sequenced 25 isolates from a local population in Senegal. Because *P. falciparum* likely originated in Africa (Conway et al. 2000), studying an African population may yield more information about its population history. Studying a local population in depth has two advantages over worldwide samples. First, polymorphisms within a local population are better at revealing signatures of recent selection than worldwide polymorphisms, and second, the existence of population structure in worldwide samples can result in patterns of linkage disequilibrium and the allele-frequency spectrum that artifactually resemble those expected from selection.

In this study, we investigated the genome-wide variation patterns of *P. falciparum*. We used principal component analysis and Bayesian clustering methods to test whether there is any significant population substructure in Senegal; we estimated demographic history using the allele-frequency spectrum; and we identified genes and gene categories under various forces of selection based on deviations from the null allele-frequency spectrum obtained by coalescent simulations with the inferred demographic parameters. Finally, we find evidence that the unusually high A/T composition of the *P. falciparum* genome (81%; Gardner et al. 2002) is maintained as a dynamic equilibrium between mutation and selection pressures operating on nucleotide composition.

2.2 Materials and Methods

Dataset and data processing

Twenty-five culture-adapted isolates of *P. falciparum* from different sites in Senegal were sequenced and investigated in this study. Ten of them are from Pikine (P05.02, P08.04, P09.04, P11.02, P19.04, P26.04, P27.02, P31.01, P51.02, P60.02), four from Velingara (V34.04, V35.04, V42.05, V92.05), and eleven from Thiès (T074.08, T10.04, T105.07, T113.09, T130.09, T15.04, T230.08, T231.08, T232.08, T26.04, T28.04). The first letter in the isolate names indicates the site of collection (P for Pikine; V for Velingara; and, T for Thiès), and the last two digits in the isolate names indicate the year of collection (for example, P05.02 was isolated in 2002).

Sequencing reads (101 bp, paired-end) were generated using Illumina HiSeq machines and were aligned to the *P. falciparum* 3D7 reference available from PlasmoDB version 7.1 (Gardner et al. 2002) using the Picard pipeline. The Picard pipeline includes BWA aligner (Li and Durbin 2010) and the Samtools data processing tool (Li et al. 2009). Genotypes were called from the reads using the Unified Genotyper (DePristo et al. 2011), which is part of the GATK package, for each isolate separately. Ambiguous calls as well as genotypes with a PHRED-style quality score of less than 30 were discarded. Repeat-rich sequences near the telomeres of each chromosome arm were excluded from the analyses (using the same bounds as in Volkman et al. 2007, Supplementary Methods). *PfEMP1* (*var*) genes were excluded from the analyses since the reads from these genes are difficult to align to the reference genome correctly due to their extremely high variability and the fact that they undergo ectopic recombination. The average depth (number of reads that map to the same location) is 46 per site. The mean of the number of isolates sequenced per aligned base is 20. More than half (51.68%) of sites have nucleotide information of all 25 isolates. SNPs have been submitted to dbSNP (submitter handle BROAD-GENOMEBIO; batch id Pf_0004), and will be released with dbSNP build B136, mid-2012.

Genetic diversity and linkage disequilibrium

We measured genetic diversity using π (the average number of pairwise differences per site among different isolates) and Watterson's theta θ_W (number of segregating sites normalized by $\sum_{i=1}^{n-1}\frac{1}{i}$, where n is the number of aligned parasite sequences) (Watterson 1975). We used the yn00 program in the PAML package (Yang and Nielsen 2000) to calculate the synonymous substitution rate (d_S). Linkage disequilibrium, measured by r^2 (Hill and Robertson 1968), was calculated for pairs of SNPs with different physical separation. Confidence intervals were obtained by bootstrapping 10,000 times over genes or chromosomes. The r^2 measure is the square of the correlation coefficient between two SNPs. Since the distance between SNPs affects the level of linkage disequilibrium, sliding LD was calculated only for SNP pairs that were 1–3 kb from each other, and the average distance within each window was also calculated. The background level of LD was estimated by calculating r^2 between pairs of SNPs from different chromosomes.

Nonsynonymous and synonymous polymorphism (π_N and π_S) were calculated by dividing the average number of nonsynonymous or synonymous pairwise differences by the number of nonsynonymous or synonymous sites, respectively. The number of nonsynonymous or synonymous sites was calculated using the method described in Table 1 of Ina (1995), and the mutation matrix estimated by the divergence of intergenic regions between *P. reichenowi* and *P. falciparum* (Table 3).

Population substructure

We investigated population structure in Senegal using two kinds of analyses: principal component analysis and a Bayesian model-based clustering method. Principal component analysis was conducted using the program *SMARTPCA* (Patterson et al. 2006) in the software EIGENSOFT 3.0. We applied a local LD correction (*nsnpldregress* = 2) and calculated the top ten eigenvectors or principal components from all the SNPs of the Senegal population. The Bayesian model-based clustering method was performed using the program *STRUCTURE* (version 2.2.3) (Pritchard, Stephens, and Donnelly 2000; Falush, Stephens, and Pritchard 2003). Each run used 20,000 iterations after a burn-in of 10,000 iterations and a model allowing for admixture and correlated allele frequencies. We conducted a series of independent runs with different numbers of clusters. Analyses with or without a prior based on geographical location were both applied. All other parameters were set to the default parameters. To ensure the consistency of results, we performed 5 independent replicates for each condition.

Demographic modeling and inference

We used the deviations in the allele-frequency spectrum from that expected under a neutral, panmictic, constant-size model to assess demographic history. The 'folded' allele-frequency spectrum is the frequency spectrum of minor alleles, which is useful when outgroup sequence data is sparse, as is the case with *P. reichenowi*. Demographic history parameters were inferred using the program $\partial a \partial i$ version 1.5.2 (Gutenkunst et al. 2009), which infers demographic parameters by fitting different demographic models to the allele-frequency spectrum. We used the folded allele-frequency spectrum of synonymous SNPs for demographic inference. Two demographic models (a two-epoch model and an exponential growth model) were used. The parameters in these two models are the time in the past at which the change in size began (T) and the ratio of current to ancient population size (N_A/N_0). Ten independent runs were performed on each model, and the parameter set with the highest log-likelihood was selected as the point estimate of the parameters. For each model, 100 conventional bootstraps were performed to obtain the 95% confidence intervals. Optimized θ from the output of ∂a∂i is equal to $2N_A\mu L$ (for haploid), where μ is mutation rate and L is the effective sequence length. With L and μ, N_A can be calculated and the unit of time (T) can be converted to years. Coalescent simulations under the best-fit demographic models were performed using the *ms* program of Hudson (2002).

Identifying genes under selection

An allele-frequency spectrum-based test, Tajima's *D* (Tajima 1989a), a long range haplotype test (Voight et al. 2006), and the ratio of nonsynonymous and synonymous polymorphism (π_N/π_S) were applied to identify genes or gene categories under selection. The null distribution of Tajima's *D* was obtained by coalescent simulations using the *ms* program and inferred demographic parameters from *∂a∂i*. Since the null distribution is sensitive to mutation rate per gene, the null distributions of different genes were simulated separately with different mutation rates that are proportional to their gene lengths. We investigated the broader biological basis of evolution by determining whether certain Gene Ontology (GO) terms were overrepresented among the genes found to be significant in Tajima's *D* test. The associations between Plasmodb gene IDs and GO terms were downloaded from the FTP site of the Wellcome Trust *Sanger* Institute (ftp://ftp.sanger.ac.uk/pub/pathogens/Plasmodium/falciparum/3D7/3D7.archive/gene_association_file/gene_association.GeneDB_Pfalciparum.20100519). *P*-values were determined according to a hypergeometric distribution. A Mann-Whitney U-test was used to test whether a GO term had significant higher or lower π_N/π_S than the rest of genes. We considered the problem of multiple testing by estimating *q*-values using the *qvalue* package (Storey and Tibshirani 2003) in the R-environment.

The significance of π_N/π_S is determined by the null distribution obtained from sampling 10,000 times two values of π_S from the empirical distribution of π_S and calculating the ratio of them. This is a very conservative test because the variation in the empirical distribution of π_S could be caused by variation in coalescent time between genes due to recombination, whereas π_N and π_S for the same gene should be based on similar coalescent histories.

The iHS statistic was computed according to the methods of Voight et al. (2006). Recombination maps were generated with LDhat 2.1 (McVean, Awadalla, and Fearnhead 2002) using a block penalty of 5.0, 10 million rjMCMC iterations, a missing data cutoff of 20%, minimum minor allele frequency of 8%, and all other parameters set to default values. Since the iHS test does not tolerate missing data, SNPs with data in at least 80% of individuals were imputed with PHASE 2.1.1 (Stephens and Donnelly 2003). As PHASE requires "diploid" data, we dropped the sample with the lowest call rate (SenP60.02) to create an even number of haploid individuals, randomly paired. 17,572 fully-imputed SNPs had at least two minor alleles among the remaining 24 individuals. The unstandardized iHS was defined as ln (iHH_A/iHH_D), where iHH_A is the integrated EHH in both directions for haplotypes with the major allele at the core SNP, and iHH_D is the integrated EHH for haplotypes with the minor allele. These unstandardized iHS scores were then normalized within allele frequency bins. Since the *P. reichenowi* sequence is sparse and the ancestral allele is not available for most SNPs in our data set, we ignored the sign of the normalized iHS score and instead present the scores as *P*-values based on a two-tailed conversion from a normal distribution.

Nucleotide transition matrix and equilibrium base compositions

To obtain empirical nucleotide transition matrices, we used *P. reichenowi* to infer the ancestral state for observed *P. falciparum* polymorphisms. The *P. reichenowi* genomic sequences (Jeffares et al. 2007) were aligned to *P. falciparum* 3D7 genomic sequence using the NUCmer program from the MUMmer 3.22 package (mummer.sourceforge.net) (Kurtz et al. 2004). Reciprocal best hits were selected and trimmed by 10 bases from each end to avoid errors stemming from poor alignment at the ends of segments. Additionally, subtelomeric regions (using the same bounds as in Volkman et al. 2007, Supplementary Methods) and *PfEMP1* genes were also filtered out from the alignment set. Since the *P. reichenowi* sequence is sparse, only 21.6% of the *P. falciparum* genome has informative *P. reichenowi* alleles. We excluded nucleotide positions where there is a lack of a *P. reichenowi* allele, and for each nucleotide position, we examined all the alleles for the 25 Senegal *P. falciparum* isolates. If any of these alleles matched the *P. reichenowi* allele, we used the *P. reichenowi* allele as the ancestral allele and inferred the nucleotide change, and otherwise we excluded that position. Then we summed up all the changes to obtain empirical nucleotide transition matrices with the counts, and normalized them by dividing units of each row by the sum of each row (so each row sums up to 1).

From the empirical nucleotide transition matrix, we obtained the equilibrium base compositions by solving for the left eigenvector of the empirical nucleotide transition matrix. We did not correct for multiple hits, but the low level of divergence between *P. falciparum* and *P. reichenowi* makes this correction negligible.

2.3 Results

The genomic sequencing was carried out on a sample of parasites isolated from a relatively small geographical region in Senegal. After verifying that the sample appeared to consist of isolates from a single, genetically homogeneous population, we set out to infer the population's demographic history, estimate key population genetic parameters, identify the selective forces (balancing, positive, negative) likely impacting individual genes, and obtain genome-wide estimates of mutation bias.

Polymorphism and linkage disequilibrium

Among the 25 isolates from Senegal we found 78,596 polymorphic sites (SNPs). The average polymorphism (pairwise mismatches, π, and normalized segregating sites, θ_W) on different chromosomes is shown in Table 2.1 and Table 2.2. Average pairwise synonymous polymorphism (π_S) [0.000601, 95% CI = (0.000544, 0.000663)] is higher than average pairwise nonsynonymous polymorphism (π_N) [0.000317, 95% CI = (0.000298, 0.000337)], indicating that many nonsynonymous changes are harmful and are removed from the population by purifying selection. Moreover, average polymorphism in intergenic regions ($\pi_{Intergenic}$) [0.000478, 95% CI = (0.000415, 0.000544)] and polymorphism in introns (π_{Intron}) [0.000420, 95% CI = (0.000378, 0.000464)] are not significantly different, and there are fewer polymorphisms, on average, in both these regions than the average for

Table 2.1. Average pairwise polymorphism (π)

Chromosome	Genic	Synonymous	Nonsynonymous	Intergenic	Intron
1	0.000635	0.000675	0.000349	0.000356	0.000487
2	0.000459	0.000645	0.000359	0.000574	0.000378
3	0.000372	0.000610	0.000286	0.000336	0.000365
4	0.000743	0.000941	0.000487	0.000951	0.000602
5	0.000407	0.000607	0.000313	0.000409	0.000447
6	0.000379	0.000579	0.000297	0.000496	0.000521
7	0.000572	0.000883	0.000381	0.000695	0.000561
8	0.000430	0.000666	0.000299	0.000569	0.000371
9	0.000380	0.000561	0.000306	0.000427	0.000334
10	0.000416	0.000621	0.000319	0.000438	0.000481
11	0.000393	0.000585	0.000308	0.000384	0.000491
12	0.000353	0.000494	0.000293	0.000354	0.000313
13	0.000379	0.000527	0.000291	0.000455	0.000414
14	0.000371	0.000519	0.000305	0.000467	0.000347
Average	0.000420	0.000601	0.000317	0.000478	0.000420

Table 2.2. Average polymorphism (θ_W)

Chromosome	Genic	Synonymous	Nonsynonymous	Intergenic	Intron
1	0.001060	0.001425	0.000676	0.000693	0.001152
2	0.000879	0.001181	0.000706	0.001097	0.000765
3	0.000782	0.001256	0.000608	0.000855	0.000997
4	0.001135	0.001438	0.000782	0.001455	0.001089
5	0.000801	0.001154	0.000636	0.000903	0.000993
6	0.000730	0.001081	0.000587	0.000889	0.000979
7	0.000993	0.001534	0.000740	0.001103	0.001195
8	0.000870	0.001404	0.000629	0.001199	0.000872
9	0.000813	0.001181	0.000664	0.000991	0.000858
10	0.000831	0.001214	0.000653	0.000934	0.001183
11	0.000839	0.001282	0.000679	0.000957	0.001127
12	0.000768	0.001133	0.000624	0.000857	0.000837
13	0.000846	0.001216	0.000681	0.001045	0.000968
14	0.000832	0.001209	0.000683	0.001050	0.000906
Average	0.000848	0.001241	0.000666	0.001007	0.000981

synonymous polymorphism. The finding of apparent selective constraints in intergenic regions and introns suggests that functionally important nucleotide sites in these regions affecting processes such as gene expression and RNA processing are sufficiently numerous that uniform selective neutrality across these regions should not be assumed. The value of θ_W for synonymous sites, nonsynonymous sites, and intergenic regions show similar patterns (Table 2.2).

Furthermore, synonymous polymorphism varies between chromosomes (Figure 2.1A). The maximum and minimum average chromosomal π_S are 5×10^{-4} and 9×10^{-4}, respectively. Three possible explanations are as follows: first, regions of extremely high diversity cause the variation between chromosomes; second, variation in the rate of recombination among chromosomes causes variation in polymorphism reduction due to selective sweeps or background selection at adjacent sites; or, third, mutation rates might vary substantially from one chromosome to the next. Highly polymorphic *PfEMP1* genes were excluded from analysis due to low confidence in appropriate read mapping and differences in gene family composition among isolates. To rule out the possibility that this variation is caused by regions of extremely high diversity, such as antigenic genes, we performed non-parametric test (Mann–Whitney *U* test) to examine the difference of π_S among chromosomes, and the results are consistent.

To understand whether the variation of synonymous polymorphism can be explained by differences in recombination rates, we calculated the correlation between chromosomal recombination rates estimated by Jiang et al. (2011) and

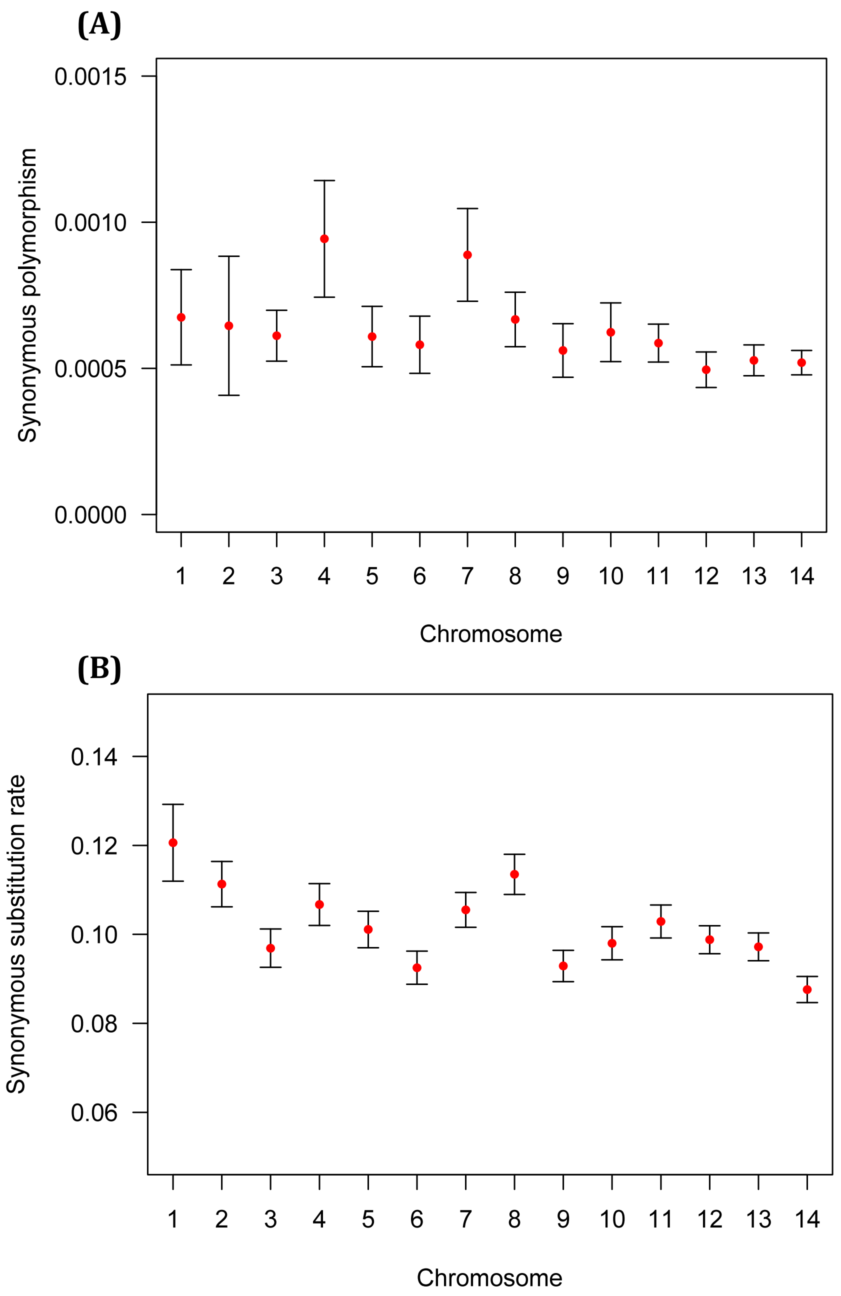


Figure 2.1. Synonymous variation in different chromosomes. (A) Synonymous polymorphism (π) in different chromosomes. Synonymous polymorphism varies significantly between chromosomes. The red dots are averages; the black vertical lines represent 95% confidence interval obtained by bootstrapping 10,000 times over genes. **(B)** Synonymous substitution rate (d_S) in different chromosomes. The d_S value also varies significantly between chromosomes.

synonymous polymorphism. The correlation is significant (Spearman's rank correlation ρ = 0.54, two-sided *P*-value = 0.048), suggesting that the variation in recombination rates can explain some of the variation in polymorphism among chromosomes. Divergence between species is less affected by recombination rates than polymorphism within species, and is a better indicator of mutation rate than polymorphism. Synonymous substitution rates (d_S) between *P. falciparum* and the chimpanzee parasite *P. reichenowi* also differ among chromosomes (Figure 2.1B), and d_S is positively correlated with π_S (Spearman's rank correlation ρ = 0.75, two-sided *P*-value = 0.003), suggesting that the variation of π_S can at least in part be explained by differences in mutation rate among chromosomes. We emphasize that recombination rates and synonymous substitution rates are not highly correlated (Spearman's rank correlation test, two-sided *P*-value = 0.10), and therefore we are detecting two different correlations. Furthermore, the difference in π_S and d_S among chromosomes cannot be explained by the variation of gene density between chromosomes (Spearman's rank correlation test, two-sided *P*-value = 0.37 and 0,74 for π_S and d_S, respectively).

We then looked at linkage disequilibrium (LD) (r^2) and polymorphism (π and θ_W) across all chromosomes using a sliding windows approach (Figure 2.2) and determined that LD decays rapidly in the Senegal population (Figure 2.3). The average r^2 decreases to the baseline level at a distance of only 1 kb, which is consistent with previous studies (Neafsey et al. 2008; Van Tyne et al. 2011), however the present study is at a finer scale. Low linkage disequilibrium suggests

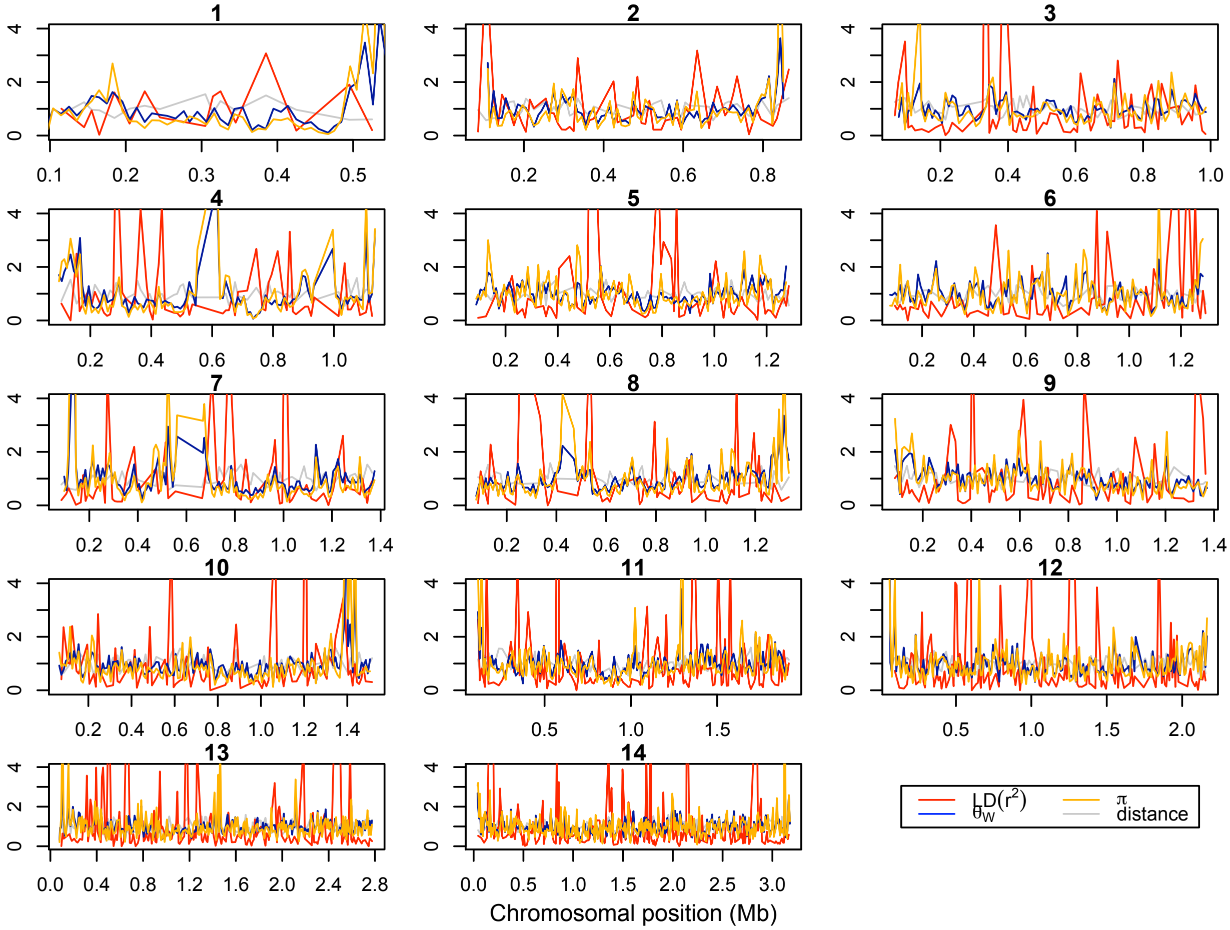


Figure 2.2. Linkage disequlibrium and polymorphism. Values on the *y* axis are expressed as multiples of the overall mean. The sliding window size is 10 kb. Only polymorphic sites that are separated by 1 to 3 kb were used in the calculation of r^2. The gray line shows the average of distances between all the pairs of polymorphic sites used for calculating r^2 in every window.

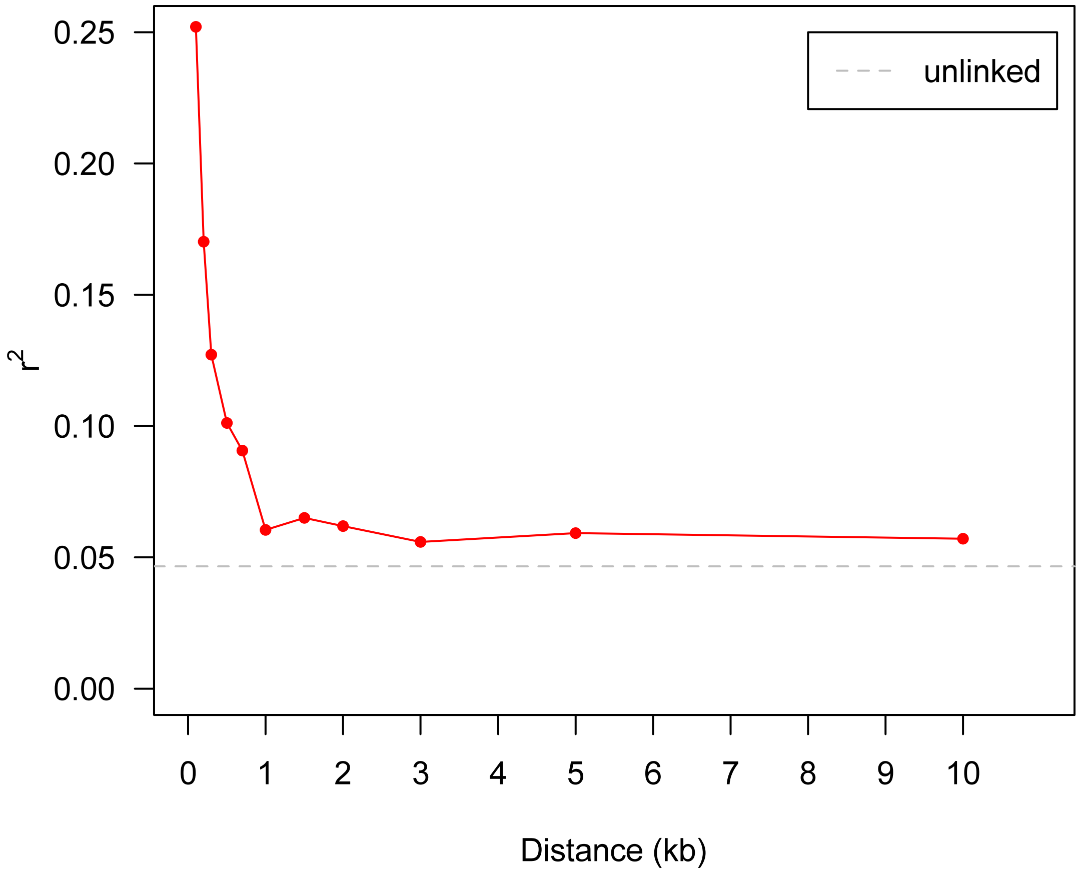


Figure 2.3. The decay of linkage disequilibrium with distance. Pairs of sites separated by an interval of 0.975 × distance to 1.025 × distance were used in calculating r^2. Unlinked r^2 (dashed grey line) was computed using random pairs of SNPs on different chromosomes. The average value of linkage disequilibrium decreases to the unlinked LD level after 1 kb.

high levels of recombination in Senegal. Since most mosquitoes bite only once and meiosis and recombination happen in the mosquito gut, outcrossing (and effective recombination) only happens when patients are infected with multiple strains. Therefore, high levels of recombination in the Senegal population suggest relatively high transmission intensity as compared to regions with higher levels of LD, such as Brazil and Thailand (Neafsey et al. 2008). Moreover, high levels of recombination suggest the potential of fine-scale genetic mapping from genomic sequences. High recombination rates also mean that the signatures of positive selection persist for a shorter period of time. Hence any significant long-range haplotypes indicate a very recent episode of positive selection.

The abundance of dramatic LD spikes in Figure 2.2 could be due to both selection and variation in recombination rate. For example, the dramatic LD spikes between 0.75 and 0.85 Mb on chromosome 5 (Figure 2.2) are likely to be caused by a selective sweep because the EHH test of selection also captures this region (Table S2.2). A high variation of recombination rate has also been found in other organisms, such as human and *Drosophila* (McVean et al. 2004; Kulathinal et al. 2008; Sella et al. 2009). Sella et al. (2009) showed that level of synonymous polymorphism is positively correlated with estimated recombination rates.

Evident absence of population substructure

Since population substructure could cause false-positive results when examining the data for signals of selection, we next investigated whether there is any population substructure among the samples from three different cities in Senegal that are less than 250 miles away from each other. Van Tyne et al. (2011) found no population substructure within the Senegal when studying worldwide strains, but that study employed a SNP genotyping array rather than sequencing, and therefore may have been less sensitive than the present approach. We also find no significant principal component for this population using *SMARTPCA*. The likely cause is the lack of population substructure, however we cannot formally exclude the possibility that unsupervised principal component analysis failed to detect variation among samples from the three locations.

The second analysis for population structure made use of a Bayesian model-based clustering approach implemented in the software *STRUCTURE*. Similar to PCA, this analysis suggests that there is no obvious population substructure in Senegal. The number of clusters (K) with $K = 1$ fits the data much better than $K = 2$ or $K = 3$, irrespective of whether the sampling location is used as a prior (log-likelihoods of $K = 1$, $K = 2$, and $K = 3$ without the sampling location as a prior are –508380.7, –527733.4, and –529739.8, respectively; log-likelihoods of $K = 1$, $K = 2$, and $K = 3$ with the sampling location as a prior are –508313.1, –518471.7, and -668800.1, respectively). The large differences in log-likelihoods for $K = 1$, $K = 2$, and $K = 3$ mean that, if we assume that the prior probabilities of $K = 1$, $K = 2$, and $K = 3$ are equal, the posterior probability of $K = 1$ is almost 100% (specifically, $\frac{e^{-508380.7}}{e^{-508380.7}+e^{-527733.4}+e^{-529739.8}} \approx 1$). In addition, inconsistent individual membership coefficients among five replicate runs when $K > 1$ support the inference that there is no significant population substructure detected by genetic diversity in Senegal. Thus, if we nevertheless set $K = 3$, isolates from the three different cities are intermixed among the three clusters, again consistent with the observations that there is a lack of significant substructure among parasites in this study population from Senegal. The estimated ancestry proportion of each isolate when $K = 3$ is shown in Figure 2.4A and Table 2.3. Since isolates that are used in this study are samples from seven years between 2001 and 2009, we also set $K = 7$ (Figure 2.4B), with the result that isolates from different years have similar patterns of ancestry proportion indicating no significant differences among years.

Demographic history

A proper null model for identifying genes under selection must incorporate the demographic history of the sampled population. To know whether the Senegal population underwent demographic changes in the past, we compared the folded allele-frequency spectrum of biallelic polymorphic sites in the data with the expected allele-frequency spectrum under the standard Wright-Fisher model with constant population size. In a folded allele-frequency spectrum, the ancestral states

Table 2.3. Estimated subpopulation admixture proportions when *K* = 3

Sample origin	Isolate names	Cluster 1	Cluster 2	Cluster 3
Pikine	SenP05.02	0.032	0.022	0.946
Pikine	SenP08.04	0.001	0.001	0.998
Pikine	SenP09.04	0.012	0.542	0.447
Pikine	SenP11.02	0.005	0.103	0.892
Pikine	SenP19.04	0.008	0.015	0.977
Pikine	SenP26.04	0.003	0.002	0.995
Pikine	SenP27.02	0.042	0.323	0.635
Pikine	SenP31.01	0.110	0.147	0.743
Pikine	SenP51.02	0.081	0.004	0.914
Pikine	SenP60.02	0.009	0.034	0.957
Thiès	SenT074.08	0.005	0.860	0.134
Thiès	SenT10.04	0.613	0.103	0.284
Thiès	SenT105.07	0.016	0.982	0.002
Thiès	SenT113.09	0.255	0.007	0.738
Thiès	SenT130.09	0.042	0.725	0.233
Thiès	SenT15.04	0.980	0.007	0.013
Thiès	SenT230.08	0.572	0.070	0.358
Thiès	SenT231.08	0.011	0.976	0.013
Thiès	SenT232.08	0.872	0.126	0.003
Thiès	SenT26.04	0.004	0.002	0.994
Thiès	SenT28.04	0.075	0.055	0.870
Velingara	SenV34.04	0.004	0.031	0.965
Velingara	SenV35.04	0.005	0.987	0.008
Velingara	SenV42.05	0.952	0.015	0.033
Velingara	SenV92.05	0.954	0.022	0.024

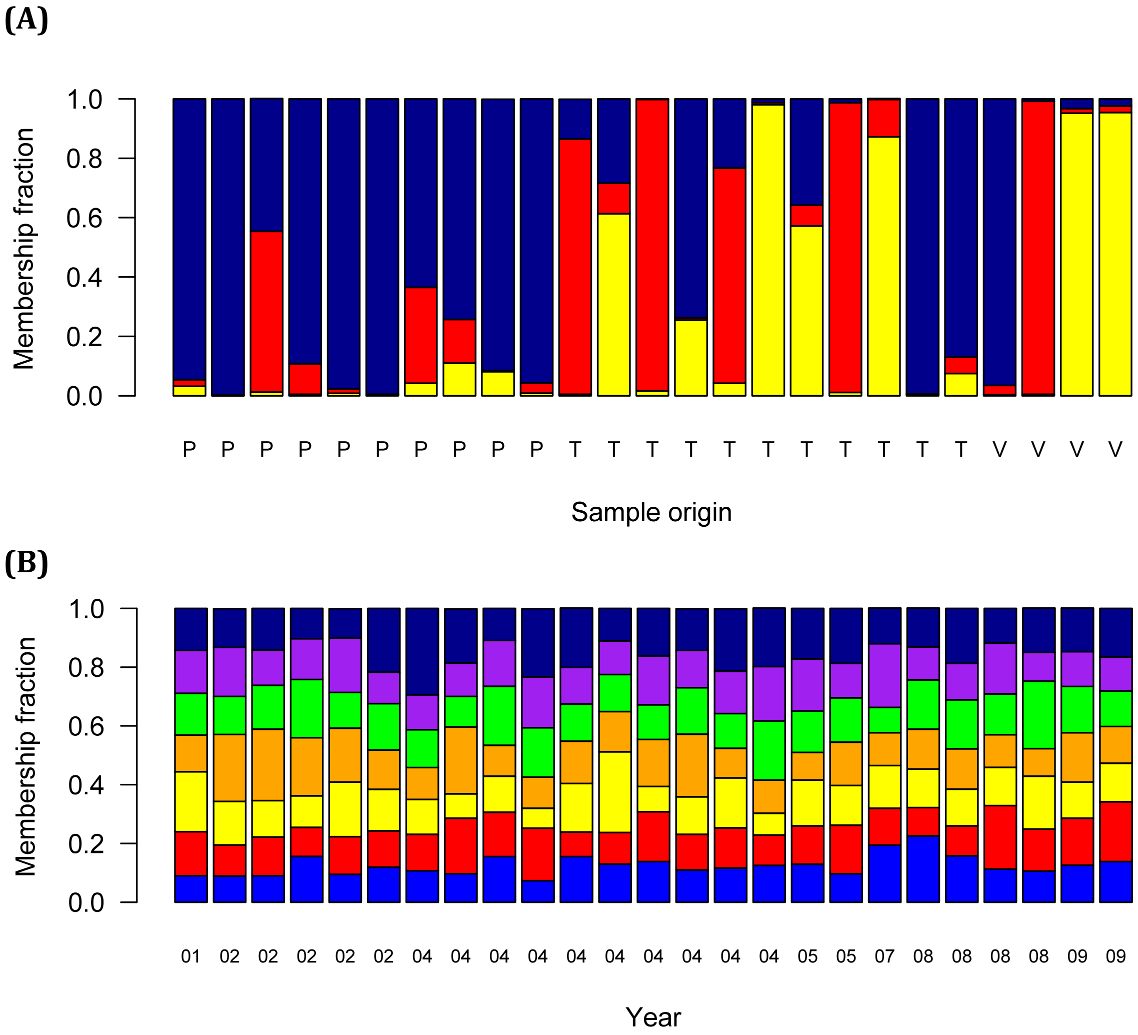


Figure 2.4. Subpopulation admixture proportion. (A) Estimated subpopulation admixture proportion of isolates from Pikine (P), Thiès (T), and Velingara (V). Each isolate is represented by a vertical bar divided into colored segments representing the isolate's estimated membership fraction in each of 3 clusters. Results from the replicate run with lowest log-likelihoods are shown here. Isolates from three different cities share the same major colors, indicating no obvious geographical substructure. **(B)** Estimated subpopulation admixture proportion of isolates from different years. Each isolate is represented by a vertical bar divided into colored segments representing the isolate's estimated membership fraction in each of 7 clusters. Results from the replicate run with lowest log-likelihoods are shown. Isolates from different years have similar patterns, indicating no obvious temporal substructure.

of alleles are assumed to be unknown, and hence alleles with frequency *x* are pooled with those of frequency 1 – *x*, so the allele frequencies in the folded frequency spectrum range from 0 to 0.5. In the Senegal data, the folded allele-frequency spectra of all polymorphic sites, genic sites, and intergenic sites are all much more skewed than expected under neutrality with a constant population size (Figure 2.5). This result suggests that there was a population expansion in the recent past, since the number of low-frequency alleles in the population increases when a population grows in size.

The nonsynonymous allele-frequency spectrum is only slightly more skewed than the synonymous allele-frequency spectrum. If nonsynonymous changes are disfavored by selection (as suggested by lower nonsynonymous pairwise polymorphism than nonsynonymous polymorphism), then the nonsynonymous allele-frequency spectrum should be more skewed toward low-frequency variants than the synonymous allele-frequency spectrum. The possible explanations for the lack of a larger difference are discussed below in the context of the ratio of nonsynonymous to synonymous polymorphism.

To further estimate demographic parameters, we fit the synonymous allele-frequency spectrum of observed data to two kinds of demographic models—a two-epoch model and an exponential growth model (Figure 2.6)—by using likelihood-based software *∂a∂i* (Gutenkunst et al. 2009). We used the synonymous allele-frequency spectrum because it is less likely to be affected by selection. Both models include two demographic parameters: the time in the past when change in size took

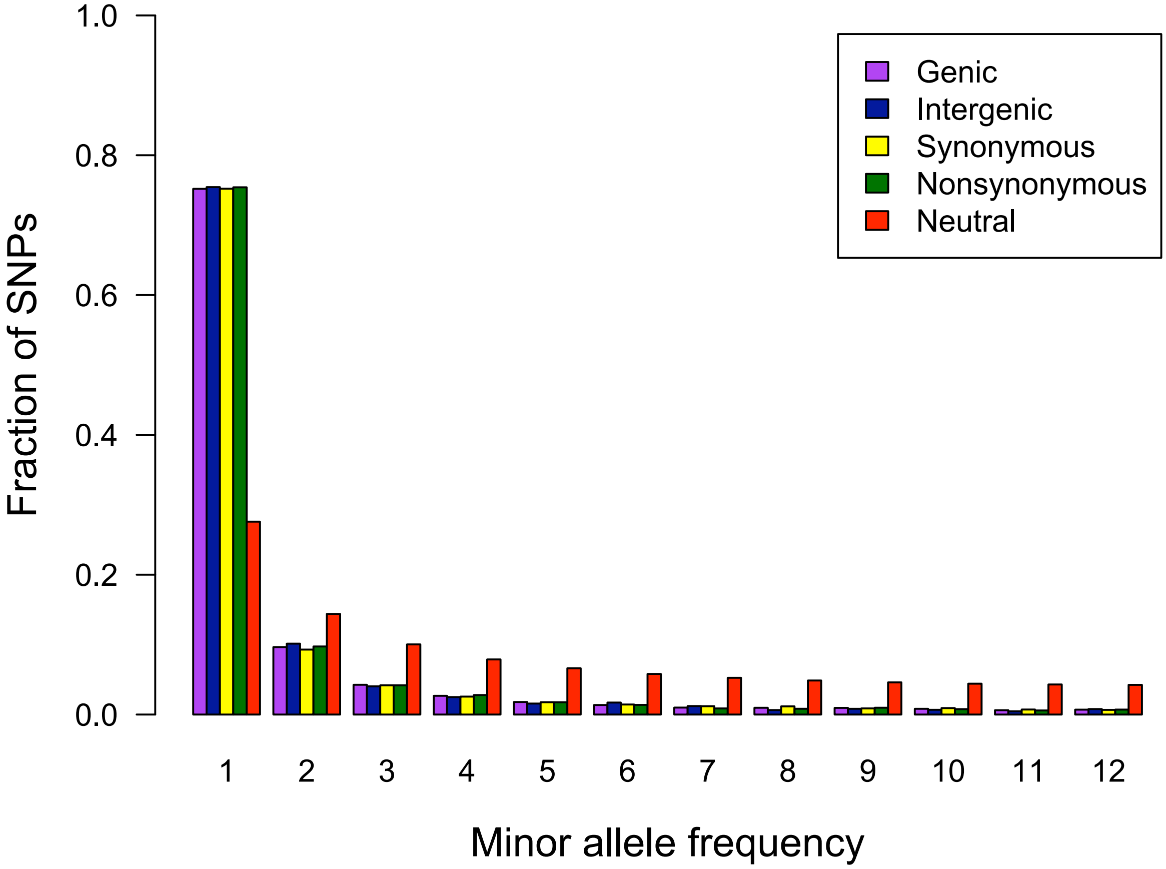


Figure 2.5. Folded allele-frequency spectrum. Allele-frequency spectra of different classes of nucleotide sites all show excess of rare alleles. The neutral allele-frequency spectrum was obtained assuming a constant population size. The excess of rare alleles in empirical spectrum indicates a population expansion in the past.

place (T) and the ratio of ancient to current population size (N_A/N_0). The only difference between these two models is that population size changes instantaneously in the two-epoch model whereas population size grows exponentially in the exponential growth model ($N_0 = N_A e^{\alpha T}$, where α is the growth rate).

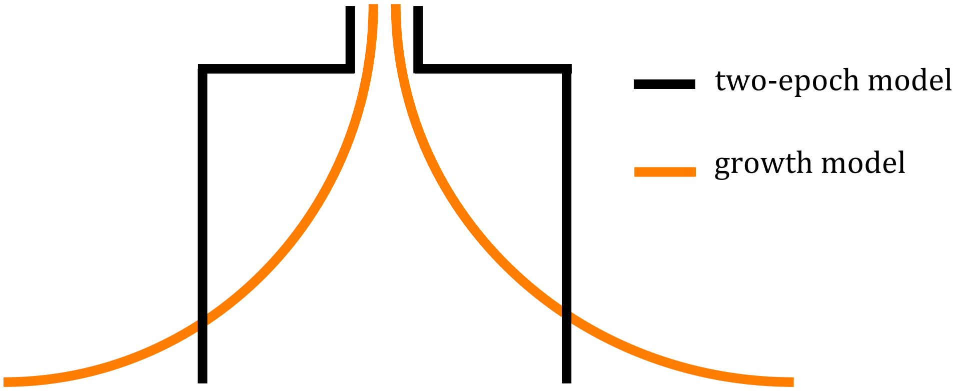


Figure 2.6. Demographic models. In the exponential model, population size starts increasing earlier than in the two-epoch model, and the current population size is larger.

To solve the equations for population size and population expansion time in the two models, an estimate of mutation rate is required. For this estimate, we calculated substitution rates (d_S) between *P. falciparum* and *P. reichenowi*. Since Liu et al. (2010) showed that *P. falciparum* likely originated from parasites of gorilla origin, we assume that *P. falciparum* and *P. reichenowi* diverged at the same time as the ancestors of chimpanzees and gorillas, and used and the divergence time between chimpanzees and gorillas, 7.3 million years (Chen and Li 2001), as the credible upper limit of divergence time between *P. falciparum* and *P. reichenowi*. The resulting mutation rate was estimated to be 6.82 × 10^{-9} per site per year. We then applied this estimate to the two models and obtained the estimates of population sizes and expansion time (Table 2.4). The estimated two-epoch model starts with a ancestral population size of 43,000, followed by an increase of 62.03 fold at a time 24,000 years ago. Under the exponential growth model, the ancestral population size is estimated as 40,000, and population size increases exponentially by 395.93-fold beginning 30,000 years ago. A three-epoch model was also used, but it did not show significantly lower likelihood, and the estimates from replicated runs are different.

Table 2.4. Inferred demographic parameters under two-epoch model and exponential growth model.

Model	**Parameter**	**Estimate**	**95% CI**
Two-epoch	N_A (× 10^5)	0.43	0.34 – 0.55
	N_0 (× 10^5)	26.97	22.36 – 32.16
	T (× 10^4 years)	2.41	2.26 – 2.55
	N_0 / N_A	62.03	42.84 – 82.28
Exponential growth	N_A (× 10^5)	0.40	0.20 – 0.52
	N_0 (× 10^5)	107.79	1.20 – 180.39
	T (× 10^4)	3.07	2.63 – 4.02
	N_0 / N_A	395.93	69.77 – 671.29

Identification of genes under selection

To identify genes under selection, Tajima's *D* was calculated for genes that are polymorphic. Positive Tajima's *D* is considered as an indicator of balancing selection or partial sweep, and negative Tajima's *D* could be caused by either negative selection or selective sweep. Since the null distribution of Tajima's *D* under the neutral hypothesis is sensitive to the demographic history and gene length (mutation rate per gene), the significance level was estimated using the null distribution from coalescent simulations that consider both gene length and inferred demographic parameters. Among 4,281 genes examined, 29 genes have extremely high or low Tajima's *D* scores and *q*-values (estimated false discovery rates, Storey and Tibshirani 2003) of less than 0.30 (Table 2.5). Among these genes, 26 have positive Tajima's *D*, including *PfAMA1* (PF11_0344), which has been found to be under balancing selection in many previous studies, and genes related to host-parasite interaction, such as acyl-CoA synthetase (*PFB0695c*) and tryptophan-rich antigen (*PF10_0026*). Only 3 of the genes have significant negative Tajima's *D* values, and these genes are of unknown function. Gene ontology enrichment analysis was used for summarizing the genes with extreme Tajima's *D* values to obtain an overall picture of genes that are under selection. GO terms that enriched for significant positive Tajima's *D*, such as pathogenesis (GO:0009405), COPI vesicle coat (GO:0030126), and RNA helicase activity (GO:0003724), are listed in Table 2.6. However, no GO term was found to be enriched for significant negative Tajima's *D* values.

Table 2.5. The 29 genes with Tajima's *D* having *q*-values < 0.30.

Gene ID	Gene Function	*P*-value	*q*-value
Negative Tajima's *D*			
PFE1365w	conserved Plasmodium protein, unknown function	4×10^{-4}	0.259
MAL8P1.141	conserved Plasmodium protein, unknown function	4×10^{-4}	0.259
PF13_0265	conserved Plasmodium protein, unknown function	0.003	0.259
Positive Tajima's *D*			
PFA0330w	MORN repeat protein, putative	6×10^{-4}	0.259
PFB0695c	serine/threonine protein kinase, FIKK family	6×10^{-4}	0.259
PFC0910w	Cg1 protein	8×10^{-4}	0.259
PFD0830w	pseudouridylate synthase, putative	8×10^{-4}	0.259
PFD0865c	aminophospholipid-transporting P-ATPase	8×10^{-4}	0.259
PFD0975w	acyl-CoA synthetase, PfACS8	0.001	0.259
PFD0980w	AAA family ATPase, CDC48 subfamily	0.001	0.259
PFE0560c	nucleoside transporter 2	0.001	0.259
PFF0475w	tryptophan-rich antigen 3	0.002	0.259
PFF1350c	endoplasmic reticulum-resident calcium binding protein	0.002	0.259
PFF1485w	cdc2-related protein kinase 1	0.002	0.259
PF07_0029	heat shock protein 90	0.002	0.259
PF07_0035	small ribosomal subunit assembling AARP2 protein	0.002	0.259
PF07_0047	conserved Plasmodium protein, unknown function	0.002	0.259
MAL8P1.32	holo-(acyl-carrier protein) synthase, putative	0.002	0.259
PFI0120c	acetyl-CoA synthetase	0.002	0.259
PFI0685w	DNA polymerase epsilon subunit b, putative	0.002	0.259
PF10_0026	bifunctional dihydrofolate reductase-thymidylate synthase	0.002	0.259
PF10_0246	surface-associated interspersed gene 13.1 (SURFIN13.1)	0.002	0.259
PF11_0098	conserved Plasmodium protein, unknown function	0.003	0.259
PF11_0344	RIO-like serine/threonine kinase, putative	0.003	0.259
PFL0950c	hypothetical protein	0.003	0.259
PFL1425w	conserved Plasmodium protein, unknown function	0.003	0.259
PFL1655c	conserved Plasmodium protein, unknown function	0.003	0.259
PF13_0075	T-complex protein 1, gamma subunit, putative	0.003	0.259
PF13_0015	apical membrane antigen 1	0.004	0.282

Table 2.6. GO terms enriched for positive Tajima's *D* and extreme π_N/π_S

Gene Ontology ID	Names	*P*-value[1]
Positive Tajima's *D*		
GO:0009405	Pathogenesis	0.006
GO:0030126	COPI vesicle coat	0.011
GO:0003724	RNA helicase activity	0.011
GO:0006890	Retrograde vesicle-mediated transport, Golgi to ER	0.011
GO:0006310	DNA recombination	0.017
GO:0005832	Chaperonin-containing T-complex	0.020
GO:0006260	DNA replication	0.023
GO:0004467	Long-chain-fatty-acid-CoA ligase activity	0.033
GO:0009982	Pseudouridine synthase activity	0.033
GO:0001522	Pseudouridine synthesis	0.033
GO:0007035	Vacuolar acidification	0.033
GO:0008283	Cell proliferation	0.033
GO:0030260	Entry into host cell	0.041
GO:0006281	DNA repair	0.042
GO:0005773	Vacuole	0.047
High π_N/π_S		
GO:0016020	Membrane	1.0×10^{-4}
GO:0016255	Attachment of GPI anchor to protein	1.6×10^{-4}
GO:0003779	Actin binding	2.3×10^{-4}
GO:0003774	Motor activity	2.7×10^{-4}
GO:0016459	Myosin complex	4.7×10^{-4}
Low π_N/π_S		
GO:0003735	Structural constituent of ribosome	6.2×10^{-10}
GO:0005840	Ribosome	1.3×10^{-7}
GO:0006412	Translation	9.1×10^{-6}
GO:0005842	Cytosolic large ribosomal subunit	1.5×10^{-5}
GO:0005843	Cytosolic small ribosomal subunit	3.7×10^{-5}

[1] *q*-values are ≤ 0.15 for each entry of Tajima's *D* and *q*-values are ≤ 0.05 for each entry of π_N/π_S.

The ratio of nonsynonymous and synonymous polymorphism (π_N/π_S) was also used to find genes that are under potential selection. In contrast to Tajima's *D*, π_N/π_S is less sensitive to demographic history since synonymous sites and nonsynonymous sites are both affected by demographic history. The top 10 highest π_N/π_S genes are exported protein (*PFL0070c*), acyl-CoA synthetase PfACS8 (*PFB0695c*), acyl-CoA synthetase PfACS7 (*PFL0035c*), sporozoite invasion-associated protein 1 (*PFD0425w*), cysteine repeat modular protein 1 (*PFI0550w*), cysteine repeat modular protein 3 (*PFL0410w*), protein kinase (*PFB0520w*), oocyst capsule protein (*PFC0905c*), cysteine-rich surface protein (*PF13_0338*), and subtilisin-like protease 2 (*PF11_0381*). Among these, acyl-CoA synthetase was reported to have high genetic diversity and long-range haplotype before (Bethke et al. 2006; Van Tyne et al. 2011). We also tested whether any GO term has significantly higher or lower π_N/π_S than others. GO categories with significant higher or lower π_N/π_S (*q*-values < 0.05) are listed in Table 2.6.

Since π_N/π_S is not expected to be higher under positive selection before the favorable allele is fixed (π_S is expected to increase with π_N in this situation), it is helpful for distinguishing two potential selective forces suggested by positive Tajima's *D*, balancing selection and partial sweep. Figure 2.8 and Figure 2.7 compares Tajima's *D* and π_N/π_S. Genes with high Tajima's *D* and high π_N/π_S, such as acyl-CoA synthetase (*PFL0035c* and *PFB0695c*), are more likely to be under balancing selection. Genes with π_N/π_S greater than 3 have significantly higher

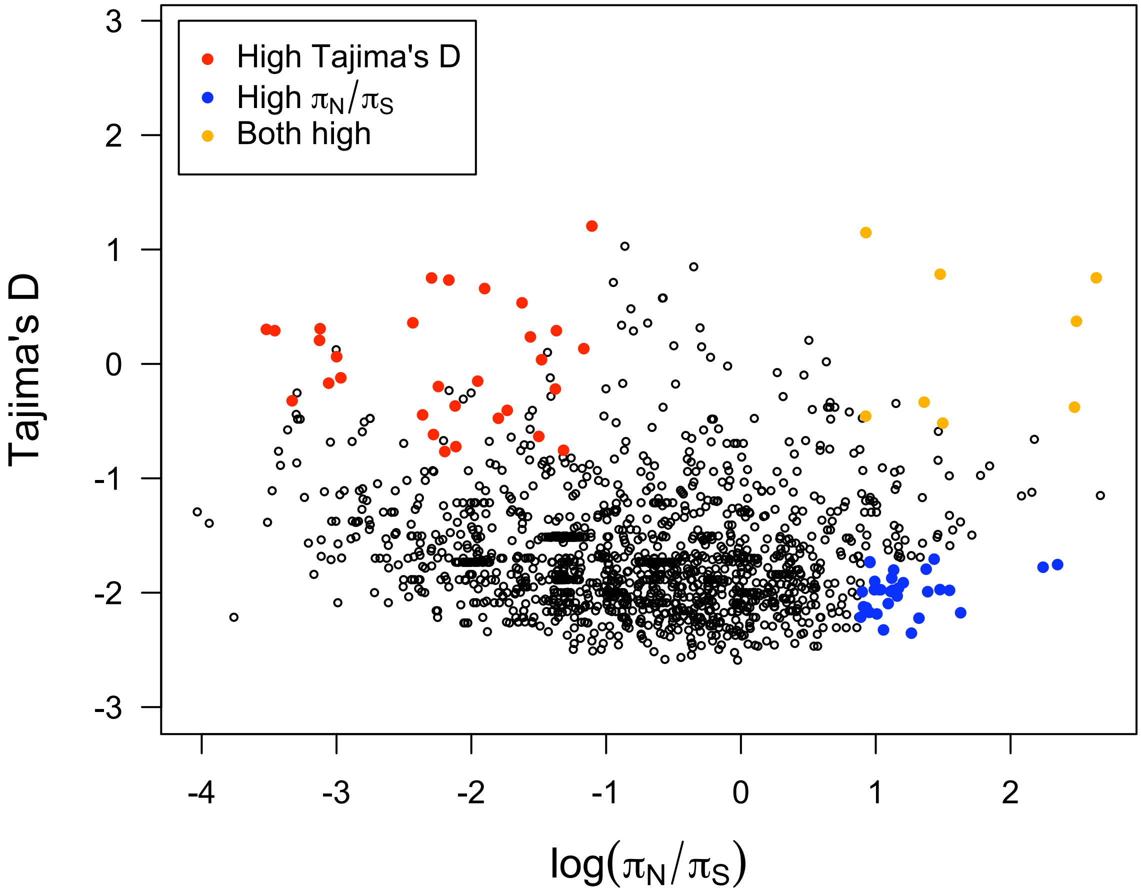


Figure 2.7. Tajima's *D* versus log(π_N/π_S). Tajima's *D* and π_N/π_S do not always have the same pattern. By considering both together, genes under selection can more readily be identified. Orange dots show genes with both high Tajima's *D* (*P*-value < 0.05) and π_N/π_S (top 5%), red dots show genes with only high Tajima's *D*, and blue dots represent genes with only high π_N/π_S.

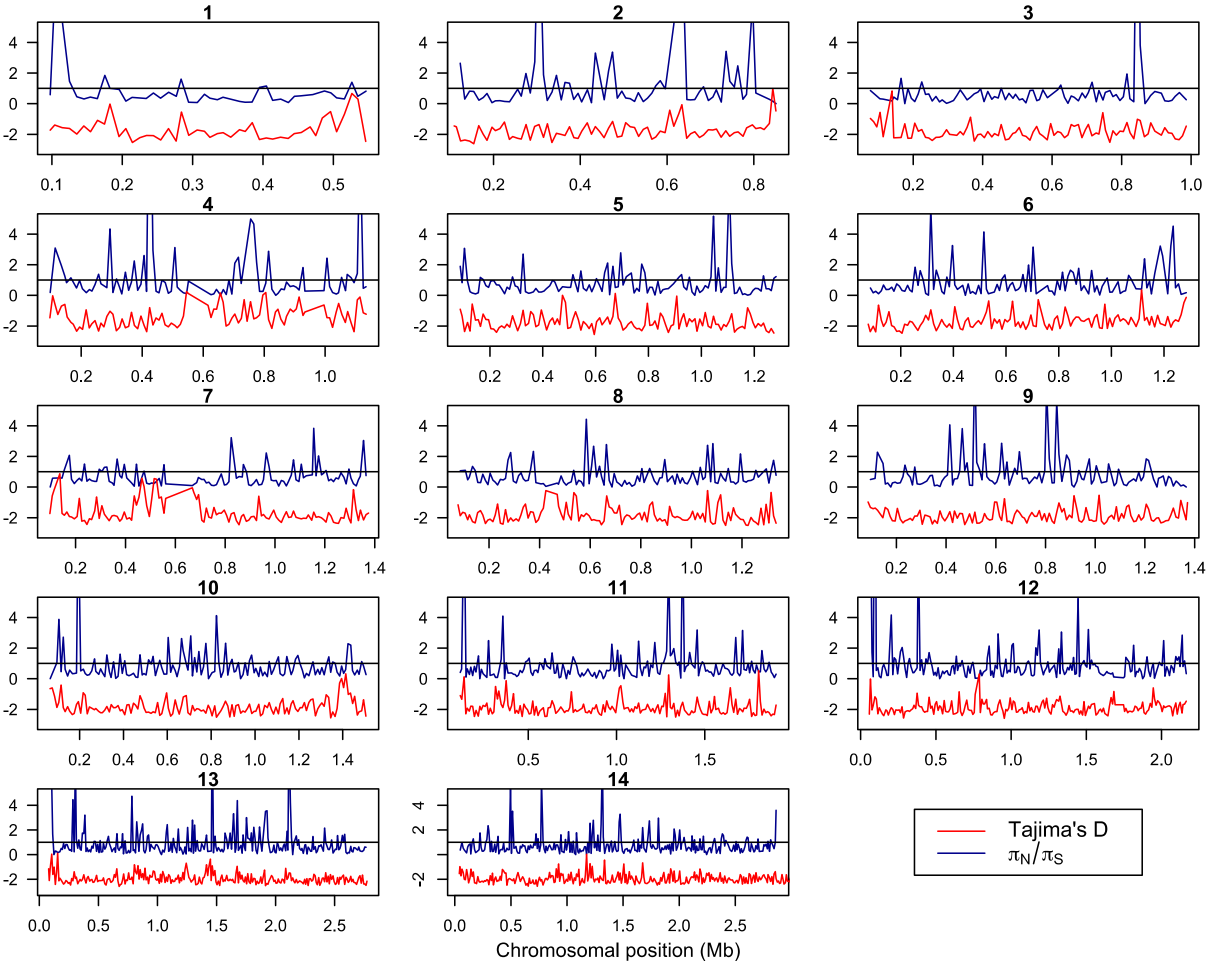


Figure 2.8. Tajima's *D* and π_N/π_S in a sliding window. The window size is 10 kb. In some regions Tajima's *D* and π_N/π_S are correlated, but in some regions they show the opposite pattern. Tajima's *D* and π_N/π_S are both high when genes are under diversifying selection, and only Tajima's *D* would be high if genes undergo partial sweep.

Tajima's *D* than genes with π_N/π_S lower than 3 (Mann-Whitney U-test, *P*-value = 0.002), suggesting that genes with higher π_N/π_S tend to be under balancing selection. Genes with high Tajima's *D* but not high π_N/π_S, such as single-strand binding protein (*PFE0435c*), might indicate partial sweeps caused by positive selection, recent balancing selection (before the selected allele reaches the equilibrium frequency), or balancing selection on noncoding sites like UTRs or introns. Genes having high π_N/π_S but not high Tajima's *D*, such as cysteine repeat modular protein 1(*PFI0550w*), might be under balancing selection, and Tajima's *D* is not significant due to the occurrence of rare alleles.

Interestingly, 18% of genes have π_N/π_S greater than 1 (Figure 2.9 and Table S2.1). Such a high value is very unusual and to our knowledge has not been reported in any other organism. This finding, together with the observed similar synonymous and nonsynonymous spectra, may be due to one or more of the following. First, one must consider sequencing artifacts or faulty annotation. Sequencing error could increase the number of singletons, and if the number of false singletons is much larger than the number of true ones, the difference between synonymous and nonsynonymous spectra would be diminished. Sequencing error could also increase π_N/π_S. To test this possibility, we increased the quality score threshold from 30 to 50, and found that the nonsynonymous and synonymous frequency spectra are still very similar (Figure 2.10), with 17% of genes having π_N/π_S greater than 1. In regard to annotation, incorrect annotation could cause high π_N/π_S and similar synonymous

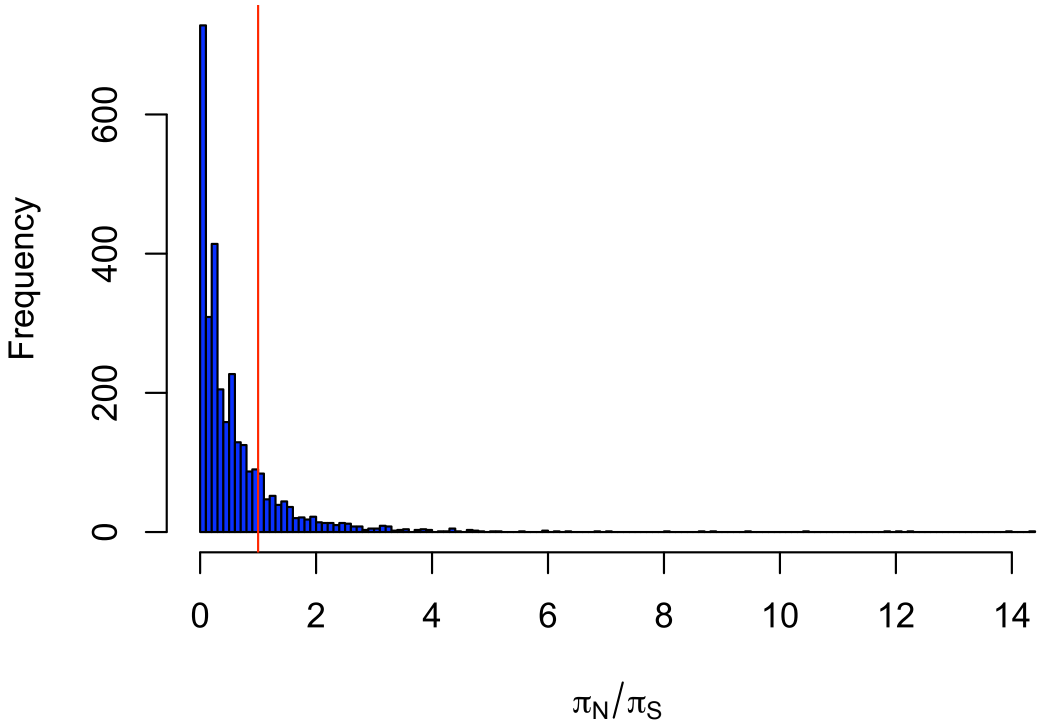


Figure 2.9. The distribution of π_N/π_S across all genes with π_S greater than 0.

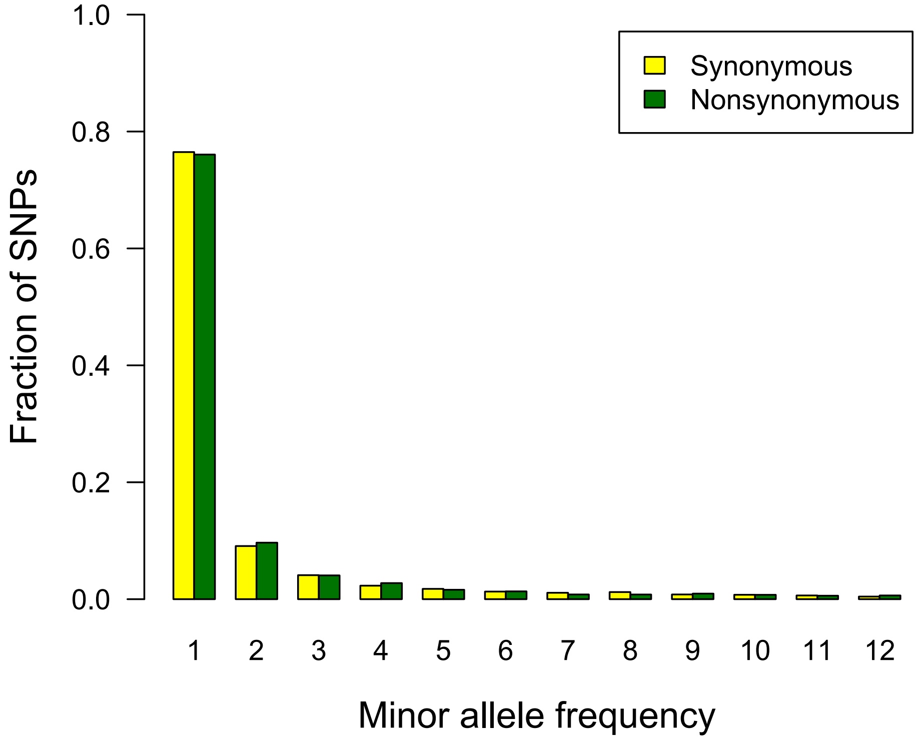


Figure 2.10. Nonsynonymous and synonymous frequency spectra when the quality score threshold is 50. Nonsynonymous and synonymous frequency spectra are still very similar after we increased the quality score threshold from 30 to 50.

and nonsynonymous spectra. However, this is unlikely to be a major factor because 456 out of 556 genes (82% of genes) with π_N/π_S greater than 1 were found to have mass spectrometry-based evidence of expression (PlasmoDB, http://plasmodb.org/plasmo/).

A second possibility is relaxed selection. Relaxed constraint or inefficient selection could contribute to the patterns. The possible relaxed constraint or inefficient selection could be caused by population expansion, recent intervention to reduce transmission, and clonal interference within the human host. Deleterious mutations have more chance to stay in population under population expansion. Recent intervention by drug treatments or bed net distribution may have reduced the effective population size very recently and hence reduced the efficacy of selection. This effect would not be captured by the allele frequency spectrum because it is so recent, and also because the effect is smaller than that of population expansion. Clonal interference within the human host could also increase the probability of deleterious mutations remaining in the population if they are linked to beneficial mutations. The finding of significantly low polymorphism in nonsynonymous sites compared with synonymous sites indicates that some genes at least are under effective purifying selection.

Another factor might be purifying selection. Purifying selection on synonymous sites could increase π_N/π_S and reduce the difference between synonymous and nonsynonymous spectra. Finally, the abundance of genes under long-term balancing selection or diversifying selection due to strong selective pressure from the host immune system could explain the pattern. It is possible that the effect on the nonsynonymous frequency spectrum of genes that are under balancing or diversifying selection offsets the effect of purifying selection, and genes under long-term multi-allelic balancing selection or diversifying selection tend to have higher π_N/π_S ratio. We observed 26 genes with significantly positive Tajima's *D*, and 17 genes have significant π_N/π_S ratio. However, it requires very high number of genes under multi-allelic balancing selection or diversifying selection to explain the whole pattern, and this high amount has never been seen in other organisms.

LD-based tests were also used for finding regions affected by putative selective sweeps. A quantile-quantile plot of integrated haplotype (iHS) scores is shown in Figure 2.12. We performed an iHS test for positive selection on 17,572 imputed SNPs (Figure 2.11). Contiguous regions of positive selection were identified by taking each genome-wide significant core SNP from the iHS test, and extending out in each direction until the extended haplotype of the major allele at a site (EHH_A) and the extended haplotype of the minor allele at a site (EHH_D) both decayed below 0.05. Each core SNP thereby defined a window of putative positive selection, and overlapping windows were merged to define the regions described in Table S2.2. The linkage-based tests suggest several regions of recent positive selection, including areas near the known drug resistance loci *pfcrt* (*MAL7P1.27*) and *pfmdr1* (*PFE1150w*). Additional regions include areas on chromosomes 2, 4, 5, 6, 7, 8, and 12. Since partial selective sweep also causes positive Tajima's *D*, Tajima's *D* values are also listed in Table S2.2 for comparison.

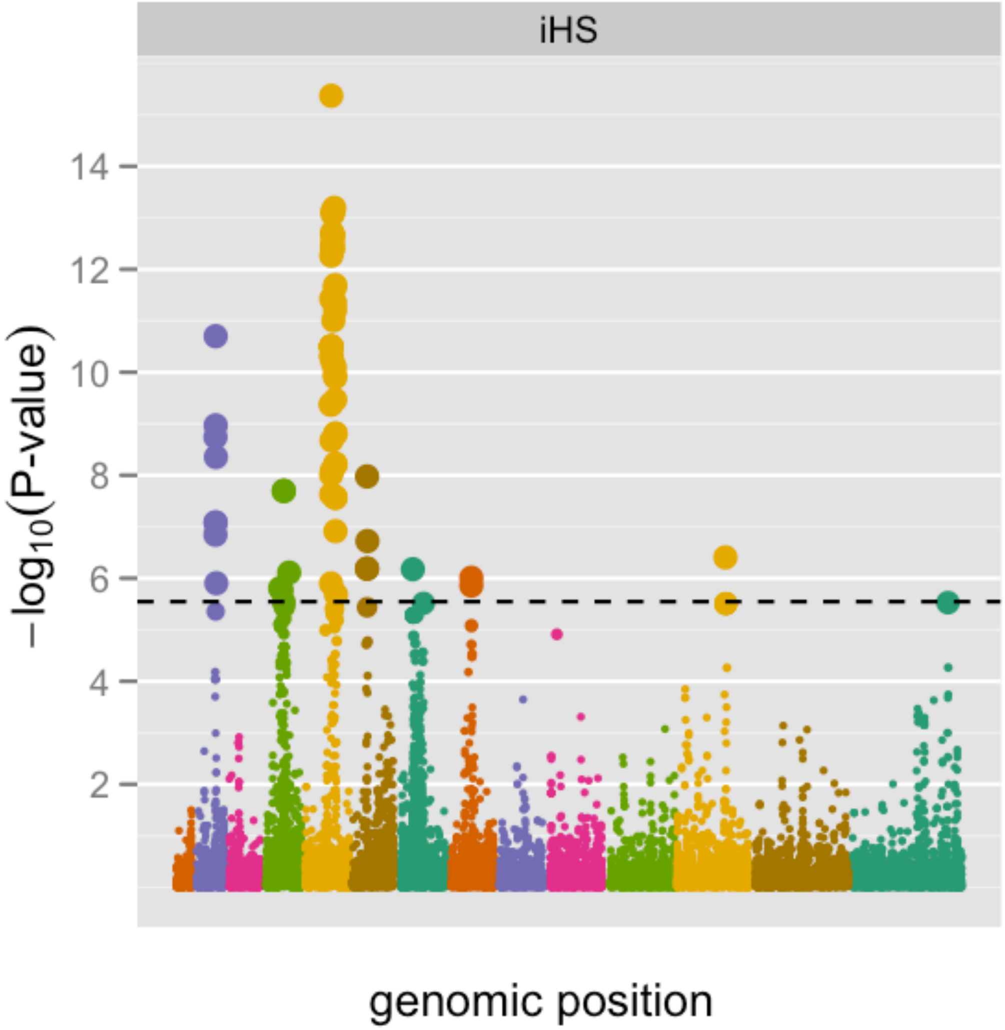


Figure 2.11. Manhattan plot of iHS *P*-values on a negative $\log_{10}$ scale. The dashed line indicates a Bonferroni threshold for significance. Dots are colored by chromosome, and their sizes are scaled according to *P*-value. Several regions of recent positive selection are suggested by the iHS statistic, including areas near the know drug resistance loci *pfcrt* and *pfmdr1*.

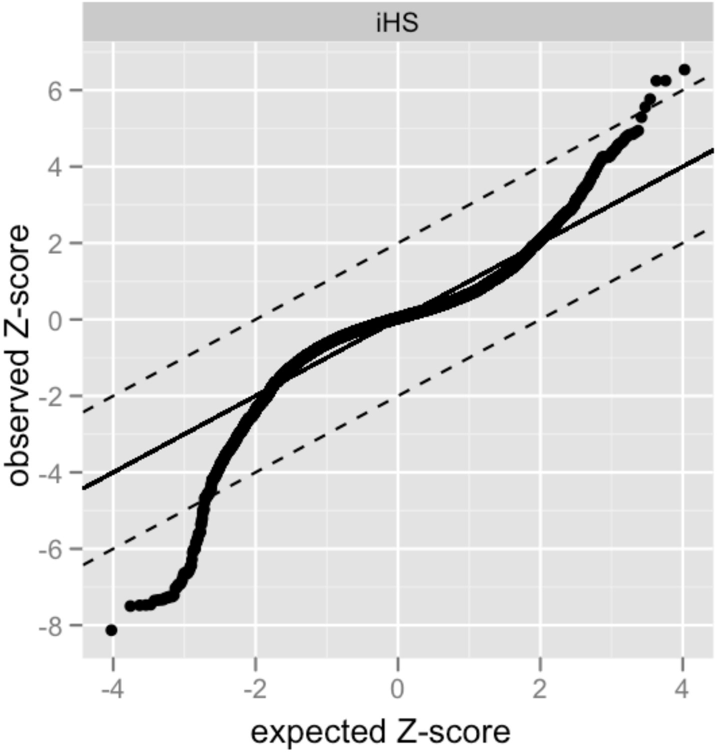


Figure 2.12. Quantile-quantile plot of iHS scores against the expected normal distribution.

Selection on base composition

The genomes of *P. falciparum* and *P. reichenowi* have an A/T base composition of 81% (Pollack et al. 1982; Gardner et al. 2002; Jeffares et al. 2007). This is much higher than what is observed in the genomes of other primate malarial parasites like *P. vivax* and *P. knowlesi*, which have an A/T content of about 60% (Williamson et al. 1985; Carlton 2003; Pain et al. 2008). To study the dynamics of changes in base composition, we used *P. reichenowi* to infer the ancestral state for observed *P. falciparum* polymorphisms. We first generated empirical nucleotide transition matrices and calculated the equilibrium states (Table 2.7 and Table 2.8). We found that the observed mean A/T composition at fourfold degenerate sites in coding regions as well as in intergenic regions are close to the predicted equilibrium A/T composition based on the empirical nucleotide transition matrices. This suggests that the nucleotide composition in *P. falciparum* may be close to equilibrium.

Table 2.7. Empirical nucleotide transition matrices

		A	T	C	G
4-fold degenerate sites	**A**	0.9905	0.0033	0.0020	0.0043
	T	0.0033	0.9903	0.0043	0.0021
	C	0.0078	0.0154	0.9717	0.0051
	G	0.0141	0.0062	0.0043	0.9754
Intergenic regions	**A**	0.9947	0.0028	0.0007	0.0019
	T	0.0028	0.9945	0.0020	0.0007
	C	0.0042	0.0084	0.9853	0.0020
	G	0.0085	0.0042	0.0021	0.9852

Table 2.8. The observed and predicted equilibrium nucleotide composition

	4-fold degenerate sites		Intergenic regions	
	Predicted	**Observed**	**Predicted**	**Observed**
A	0.40	0.41	0.42	0.43
T	0.38	0.41	0.41	0.42
C	0.10	0.09	0.09	0.08
G	0.12	0.09	0.08	0.07
A/T	0.78	0.82	0.83	0.85

Within-species polymorphism provides suitable material to study to what extent the high A/T percentage in the genome is due to mutation pressure offset by selection. We generated separate derived allele-frequency (DAF) spectra for mutations that would serve to decrease the A/T composition ("A/T to G/C") and mutations that would increase the A/T composition "G/C to A/T") in silent coding and intergenic regions (Figure 2.13). The G/C to A/T spectrum has more rare derived alleles and fewer high frequency derived alleles than the A/T to G/C spectrum in both intergenic and silent coding regions, suggesting purifying selection against A/T nucleotides and/or positive selection favoring C/G nucleotides. That is, on average, mutations that would result in a higher A/T composition of the genome are acted against by selection, and/or mutations that would result in lower G/C composition are selected for, and therefore the high equilibrium AT composition in *P. falciparum* likely reflects a dynamic equilibrium balanced between A/T biased mutation pressure and selection to augment the relative fixation probability of G/C mutations. While an incorrect ascertainment of ancestral state due to ancestral polymorphism could be responsible for a small fraction of the high frequency derived alleles observed, we do not expect a difference in the accuracy of ancestral state inferences between the two classes of mutations; therefore the observation of a relative difference in the frequencies of mutations that would further increase vs. decrease A/T composition should be robust.

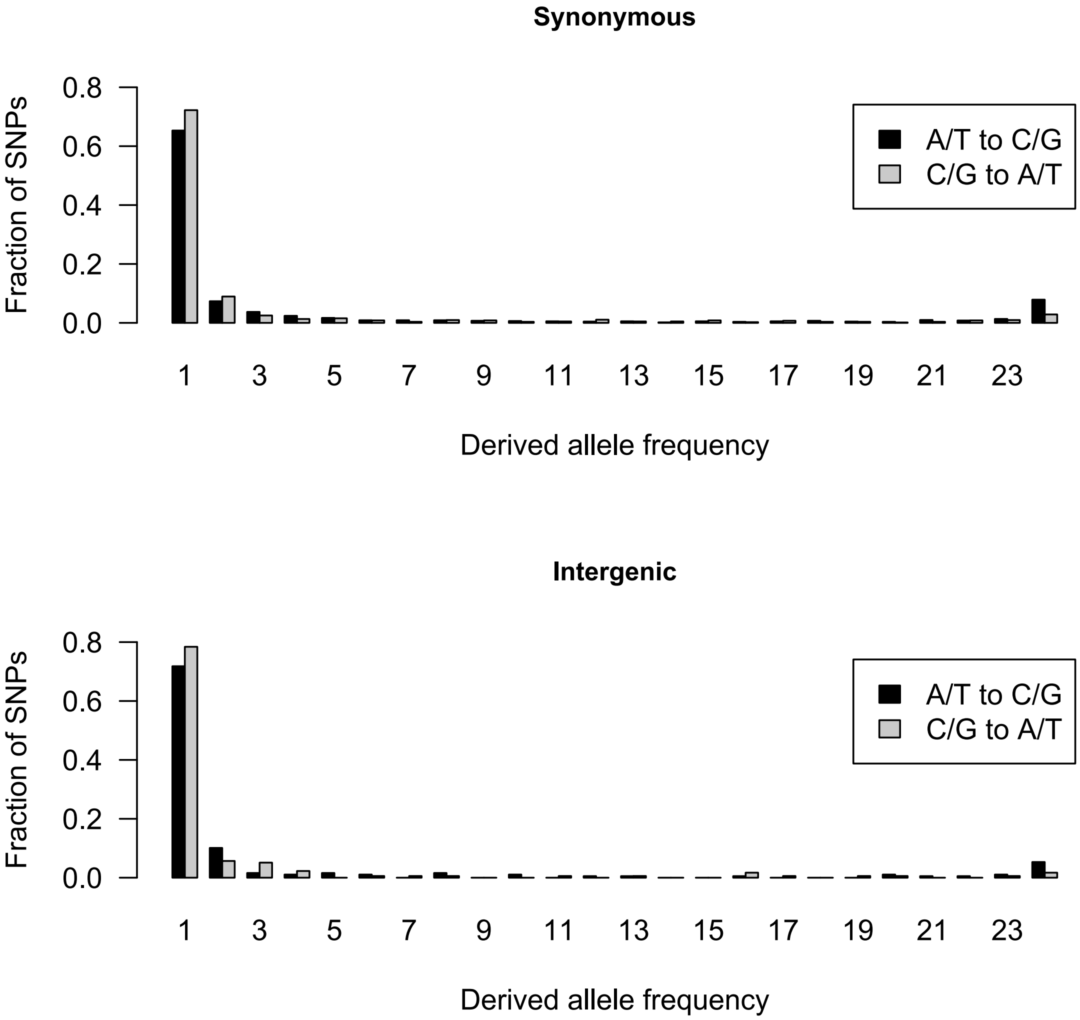


Figure 2.13. Derived site-frequency spectrum. The unfolded site-frequency spectrum was generated by using *P. reichenowi* as an outgroup. The G/C to A/T spectrum is more skewed toward low frequencies than the A/T to G/C spectrum, suggesting positive selection favoring C/G nucleotides or purifying selection against A/T nucleotides.

2.4 Discussion

In this study we carried out whole-genome sequencing of 25 isolates of *P. falciparum* from Senegal, and used population genetic methods to investigate the evolutionary forces shaping genetic diversity in this local African population. Fully sequenced population genomic data provide an important resource for understanding genome-wide patterns of diversity as well as the evolution of genes of particular interest.

Previous studies of the evolutionary history of *P. falciparum* are not completely consistent (Hartl et al. 2002; Su, Mu, and Joy 2003; Hartl 2004), which may in part reflect variation in the evolutionary history and genetic diversity across different regions or genes in the genome due to both selection and recombination. With the complete genome sequences of a deep sample from a single African population, we have been able to use genome-wide synonymous changes that are known to be minimally affected by selective forces (especially when linkage disequilibrium is low) to better estimate demographic parameters than in previous studies which relied on only a few loci. Moreover, the absence of any obvious population substructure, which can also affect the allele-frequency spectrum (Tajima 1989b), suggests that our estimates for Senegal are not biased by substructure. Because the effect on the synonymous frequency spectrum of selection on AT content (Figure 2.13) may bias the estimates of demographic parameters, we also fitted the same model to the synonymous frequency spectrum with only A/T to T/A and C/G to G/C changes, and the estimates of demographic parameters were consistent. However, it should be noted that selection on linked sites could affect synonymous sites and bias our estimates of demographic parameters, even if average linkage disequilibrium over the genome is low.

We estimate a major 60-fold population expansion approximately 20,000–40,000 years ago by fitting demographic models to the synonymous allele-frequency spectrum. The effective population size prior to the population expansion was about 20,000–55,000. Although the effective population size after the expansion is sensitive to the demographic model (varying from 2.2 million to 18 million), it is consistently very large. Our estimate of population expansion time in Africa is close to the 95% confidence interval (2,000–14,500) of the estimate in Joy et al. (2003) based on mitochondrial DNA of pan-African strains and different methods of analysis. Moreover, it was recently suggested that *P. falciparum* first infected human ancestors 365,000 years ago (Baron, Higgins, and Dzik 2011). This indicates that this event was well before the population expansion we identified and our estimates are unlikely biased by it. However, it should be noted that our estimates of population sizes and population expansion time are highly dependent on the estimate of mutation rate, and therefore on the choice of divergence time between *P. falciparum* and *P. reichenowi*. While we have curated our data as carefully as possible, any remaining sequencing error, reference-sequence bias, and alignment error caused by nearby insertions or deletions, would all potentially bias the allele-frequency spectrum and therefore our estimate of demographic parameters. Sequencing errors and reference-sequence bias both can skew the site-frequency spectrum to rare alleles, resulting in an overestimate of the changes in population sizes or the population expansion time.

Our estimate of population expansion time of 20,000 to 40,000 years overlaps with the Upper Paleolithic era (10,000 to 40,000 years ago) and the Mousterian Pluvial (30,000 to 50,000 years ago), and is immediately after human migration out of Africa (40,000 to 130,000 years ago). During the Mousterian Pluvial, northern Africa was a land of lakes, swamps, and rivers, and this may have increased the spread of malaria parasites by mosquitoes and facilitated epidemic transmission. As such, we believe our estimate of the date of expansion offers a good fit to archaeological and climatological explanations.

The recent availability of large-scale genomic datasets such as this one affords an opportunity to refine our understanding of the general evolutionary patterns in *P. falciparum*. First, we found that both synonymous polymorphism and synonymous substitution rates differ among chromosomes, suggesting that the mutation rates are not the same for all chromosomes. Since population genetic statistics, such as Tajima's *D* per gene, are sensitive to mutation rate, it is desirable to correct for the variation in mutation rate when conducting this test, provided that reliable estimates of mutation rate for each chromosome are available. Second, lower intergenic and intronic polymorphism as compared to synonymous polymorphism suggests weak selection constraints might be common in intergenic regions and introns. Divergence between *P. falciparum* and *P. reichenowi* is also lower in intergenic regions and introns (Neafsey, Hartl, and Berriman 2005). Third, low linkage disequilibrium indicates the potential of fine-scale mapping of selection or association signals and fewer problems when applying analytic tools that assume SNPs are unassociated with each other, such as *∂a∂i* and *STRUCTURE*. Fourth, we reported two unusual findings, the similar synonymous and nonsynonymous allele-frequency spectra, and the large number of genes with high nonsynonymous-to-synonymous polymorphism, and discussed some possible explanations.

Fifth, we found that the nucleotide composition of *P. falciparum*, which has the highest A/T content of any eukaryotic genomes described so far, is at or near equilibrium. Moreover, we showed that the unfolded site-frequency spectra of G/C to A/T polymorphic sites in both synonymous sites and intergenic regions have more rare alleles than that of A/T to C/G polymorphic sites, which suggests that selection acts against A/T at least at certain sites in the *P. falciparum* genome. The empirical nucleotide transition matrices (Table 2.7) also support this trend toward high A/T composition, since A, T, C and G all tend to change to A or T, and A and T also change more often. Since the empirical nucleotide transition matrix is influenced by both mutation and selection and we know that selection acts, on average, against C/G to A/T changes, it can be inferred that mutations are toward A/T. This result differs from that of Escalante, Lal and Ayala (1998), who found a trend toward higher A/T but no clear evidence of A/T mutational bias in studies of 10 highly polymorphic genes including six encoding surface antigens; signals of strong selection on five of the genes may have obscured a signal of A/T biased mutation. Mutational bias toward A/T has been found in other organisms (for example, Hershberg and Petrov 2010), and selective advantages of biased mutation rates have been discussed (Rocha and Danchin 2002; Dalpke et al. 2006). It was suggested that G and C were less favored by natural selection in obligatory pathogens and symbionts because GTP and CTP nucleotides cost more energy and there is less availability of GTP and CTP in the cell (Rocha and Danchin 2002). Also, as toll-like receptor 9 specifically recognizes non-methylated CpG dinucleotides (Dalpke et al. 2006), and there is lack of evidence of DNA methylation in *P. falciparum* (Choi, Keyes, and Horrocks 2006), reducing G and C in the genome would be one mechanism to reduce the innate immune response and therefore be favored by selection. However, when mutations are strongly biased to A/T, it may affect amino acid composition. Singer and Hickey (2000) found that A/T codon bias in *P. falciparum* was so severe that it was affecting amino acid composition. Although mutational bias can change amino acid composition in some genes, genes under strong selective constraint are less influenced by mutational bias. It was shown in *P. falciparum* that amino acids encoded by GC-rich codons are significantly more frequent in highly expressed genes (Chanda, Pan, and Dutta 2005), perhaps because highly expressed genes are more conserved, and the ancestral state presumably exhibited less biased AT content. Our findings that suggest a mutation-selection balance in the A/T content of the genome (with mutation favoring higher A/T and selection against A/T) add a new layer of understanding to this otherwise puzzling aspect of *P. falciparum*'s genome composition.

The absence of population substructure in Senegal means that signatures of selection can be identified with greater confidence. Here we used the Tajima's *D* test, the ratio of nonsynonymous to synonymous polymorphism, and the iHS test, and successfully detected signatures of selection in some genes identified previously, such as *pfcrt* and *pfmdr1,* as well as some new candidate genes, such as two acyl-CoA synthetases (*PFL0035c* and *PFB0695c*). Additionally, we identified the gene categories that are more likely to be under negative selection (for example, ribosome and translation) and balancing/diversifying selection (for example, membrane and attachment of GPI anchor to protein). The identification of these GO categories not only provides a broader view about types of selection in various biological processes but also helps with functional characterization of genes whose function is unknown.

Our study provides the one of first population genomic analyses of a deeply sampled local population of *P. falciparum*. The estimation of demographic parameters by genome-wide SNPs offers, for the first time, a proper null distribution for identifying genes under various selective forces. Genes identified here could be validated by follow-up functional assays, and the results have practical implications for finding functional variants of medical relevance and developing methods of disease control. If whole genome sequences of closely related species, such as *P. reichenowi* and other species from chimpanzee and gorilla, are available in the future, our dataset can be used for investigating evolutionary questions with more confidence, such as controlling for variation in mutation rates and detecting selection using the ratio of divergence to polymorphism.

Chapter 3

Genetic surveillance detects both clonal and epidemic transmission of malaria following enhanced intervention in Senegal[†]

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

[†] This chapter has been published in PLoS ONE 8(4): e60780.
[*] Authors with equal contributions.

3.1 Introduction

The *Plasmodium falciparum* malaria parasite causes nearly 700,000 deaths annually, primarily in sub-Saharan Africa (WHO World Malaria Report 2011), where disease prevalence and transmission intensity are highest. Because parasite populations are large in Africa, they are more genetically diverse there than elsewhere. They also exhibit less correlation between allelic states at different loci (i.e. less linkage disequilibrium, or LD), reflecting both the large population and also higher disease transmission rates, which facilitate sexual outcrossing (Anderson et al. 2000; Mzilahowa, Mccall, and Hastings 2007; Volkman et al. 2007; Neafsey et al. 2008; Van Tyne et al. 2011).

We sought to use changes in parasite population diversity to detect longitudinal changes in disease transmission, and thereby to develop useful metrics for monitoring antimalarial interventions. As a tool to track parasite diversity, we employed a previously developed 'molecular barcode', composed of assays for 24 single nucleotide polymorphisms (SNPs) across *the P. falciparum* genome (Daniels et al. 2008). We applied the barcode to samples from Senegal. Since 2005, Senegal has dramatically increased deployment of intervention strategies, including ITNs for prevention, RDTs for detection, and ACTs for treatment, resulting in an overall decline in a number of malaria indicators (Malaria RB 2010), and making it a good site for detecting changes in parasite diversity.

3.2 Materials and Methods

Study site

We obtained *P. falciparum*-positive clinical samples from patients evaluated at the SLAP clinic in Thiès, Senegal under ethical approval for human subjects and informed consent conditions. Full written consent was obtained in a protocol approved by Harvard School of Public Health, Office of Human Research Administration (P16330-110, Wirth PI) and the Ministry of Health, Senegal.

The site, located 75km southeast of the country capital of Dakar, is characterized by perennial hypo-endemic transmission with the greatest number of malaria cases by primarily *Anopheles gambiae s.l* and *A. funestus* vectors occurring approximately from September to December, at the end of the rainy season. Samples are collected passively; with patients over the age of 12 months admitted to this study with self-reported acute fevers within 24 hours of visiting the clinic and no recent anti-malarial use. Patients are screened by slide smears and rapid diagnostic test (RDT) to diagnose *P. falciparum* infection (Ndiaye et al. 2010; Ndiath et al. 2011).

DNA extraction and quantification

Whole blood spots from 2006-2011 were preserved on Whatman FTA filter paper (Whatman catalog #WB120205). We extracted genomic DNA from 4-6mm punches from the FTA cards using the manufacturer protocol for Promega Maxwell DNA IQ Casework Sample kit (Promega catalog #AS1210). After extraction, we quantified and generated a molecular barcode for each sample as described previously (Daniels et al. 2008). Extracted samples were excluded from analysis if the concentration (and corresponding parasitemia of the patient) were too low for successful amplification. The sample size in each year is shown in Table 3.1.

Table 3.1. The number of mixed and single infections in each year

Year	Single infection	Mixed infection	Total
2006	90	41	131
2007	54	26	80
2008	95	13	108
2009	77	15	92
2010	100	25	125
2011	112	26	138

Sequencing **csp**

We sequenced across the T-epitope region of the *P. falciparum csp* gene. Primer sequences were: 5’- AAATGACCCAAACCGAAATG-3’ forward and 5’-TTAAGGAACAAGAAGGATAATACCA-3’ reverse. We used 1ul of each sample as a template in 25ul PCR reactions using iProof master mix (Bio-Rad cat# 172-5310) (initial denaturation 98°C 30s, followed by 35 cycles of 98°C denaturation (30s), 55°C annealing (30s), 72°C extension (30s), and a final extension of 72°C for 5min) and sent post-PCR processed samples (exoSAP-IT, usb catalog #78201) for sequencing (Genewiz, Inc., South Plainfield, NJ).

Affymetrix array analysis

Using an Affymetrix array containing 74,656 markers (Van Tyne et al. 2011), we hybridized parasites with identical barcodes and parasites within the same collection but with different barcodes as well as technical replicates of control strains. We called SNPs using BRLMM-P from Affy Power Tools v1.10.2.Haploid genotypes were forced_by designating all SNPs as "Y chromosome" and all individuals as "male". We counted the number of differing SNP genotypes for pairs of arrays, with pairings sorted into three categories: 1) technical replicates (same parasite sample hybridized to two arrays); 2) identical barcodes (distinct patient samples with identical barcodes); and, 3) unrelated parasites (distinct barcodes).

Human Genotyping

We used a set of SNPs selected by The Broad Institute for human typing on their analysis platforms to distinguish patient samples from one another. From an original set of 23 assays, we selected 18 as robust under conditions with low template concentrations. We ran these pre-developed TaqMan-MGB probes (Life Technologies, Inc.) on an Applied Biosystems 7900HT qrt-PCR system (LifeTechnologies, Inc.) using the standard amplification and analysis protocols (see Table 3.2 for SNP identity and human typing results).

Table 3.2. Human Genotyping data (TaqMan probes from Broad Institute set)

Sample Name	rs1009806	rs898500	rs1000797	rs1000026	rs1000005	rs10242744	rs1036689	rs1000203	rs1037439	rs242076	rs1012315	rs1015939	rs1000158	rs1571256	rs1025412	rs1000192	rs10775365	rs1000053	rs1000121	rs10513695
SenT005.11	TT	AG	GG	CC	CG	AG	GG	TT	AG	AG	AG	CC	GG	AA	GG	GG	GG	CC	CT	GT
SenT014.11	CT	AG	-	CC	CG	AA	GG	TT	AG	AG	AA	-	GG	AA	GG	GG	AG	CC	CT	GT
SenT018.11	TT	AA	-	CC	GG	GG	GG	TT	AA	AA	AG	-	AG	AA	GG	GG	GG	CC	CC	GG
SenT019.11	TT	AA	GG	CC	GG	AG	AG	TT	AA	AA	AA	CC	GG	AA	AA	AG	GG	CT	CT	GT
SenT021.11	TT	AA	GG	CT	GG	AA	AG	TT	GG	AA	AG	-	AG	AA	AG	GG	-	CC	CT	TT
SenT026.11	TT	AG	GG	CC	CG	GG	GG	TT	GG	AG	GG	-	GG	AA	AG	GG	GG	CT	TT	GT
SenT054.11	TT	AG	GG	CT	CG	AG	AG	TT	AA	AG	AA	-	GG	AA	GG	GG	GG	CC	CT	GT
SenT064.11	TT	AG	GG	CC	CG	GG	GG	TT	AA	-	AG	CC	X	AA	AG	GG	GG	CC	CT	TT
SenT078.11	TT	AA	GG	CT	GG	AG	AG	TT	AG	AG	GG	CC	GG	AA	AG	AG	GG	CC	TT	TT
SenT095.11	TT	AG	-	CC	GG	AG	GG	TT	AA	AA	AG	CC	GG	AA	AG	GG	GG	CC	CT	TT
SenT098.11	TT	GG	GG	CT	GG	AA	GG	TT	AG	AG	GG	-	AG	AA	AG	GG	GG	CT	TT	GT
SenT101.11	CT	AG	-	CT	CG	AG	GG	TT	AA	AG	AA	CC	AG	AA	AA	GG	GG	CC	TT	TT
SenT103.11	TT	AG	-	CC	GG	GG	AG	TT	AA	AG	AG	CC	GG	AA	GG	GG	GG	CC	TT	TT
SenT104.11	TT	AG	-	CT	GG	AG	GG	TT	AG	AG	AG	-	GG	AA	GG	GG	GG	CC	CT	GT
SenT105.11	TT	AG	-	CC	GG	AG	AG	TT	GG	AA	GG	CC	AG	AG	GG	GG	GG	CC	TT	TT
SenT106.11	CT	AG	GG	CT	CC	AA	GG	TT	AG	AG	GG	CC	GG	AA	AG	AG	GG	CC	CT	TT
SenT110.11	TT	AG	GG	CT	CG	AA	GG	TT	AA	AG	AG	CC	AG	AA	GG	GG	GG	CC	TT	TT
SenT111.11	CT	AG	GG	CC	CG	GG	AG	TT	AA	AG	AG	TT	GG	AA	AA	AG	GG	CC	CC	TT
SenT117.11	TT	AG	GG	CT	GG	AA	AG	TT	GG	AG	GG	CC	AG	AA	AA	AG	AA	CC	-	TT
SenT122.11	TT	GG	GG	CT	GG	AG	GG	TT	AA	AG	AA	CC	AG	AA	AA	GG	GG	CC	CC	TT
SenT123.11	CT	GG	-	CT	GG	AA	GG	TT	AG	AA	AG	CC	AG	AA	AA	GG	GG	CT	TT	TT
SenT127.11	CT	AG	GG	CC	CG	AA	GG	CT	AA	AG	AG	-	AG	AA	AG	-	AG	CC	CC	TT
SenT132.11	TT	AG	GG	CT	GG	AG	-	TT	AA	AG	AG	-	GG	AA	AG	-	GG	CC	CC	GT
SenT134.11	CT	GG	GG	CC	GG	AG	GG	TT	GG	AG	AA	CT	GG	AA	GG	GG	GG	TT	CT	GT
SenT135.11	TT	AG	GG	CC	CG	AG	GG	TT	AA	AG	GG	CC	-	AA	AA	GG	GG	CC	CT	TT
SenT136.11	TT	AG	-	CT	CG	GG	AG	TT	AG	AA	AG	CC	GG	AA	AA	AG	AG	CT	TT	GG
SenT145.11	TT	AA	GG	CT	GG	AA	AG	TT	AA	AG	AA	CC	AG	AA	AG	AG	GG	CC	CT	TT
SenT146.11	CT	AA	-	CC	GG	GG	GG	TT	AA	AG	AG	CC	GG	AA	GG	GG	GG	CC	CT	TT

In addition, we sent several samples for STR genotyping on an ABI 3130 Genetic Analyzer to detect the STR alleles amplified using the ABI AmpFlSTR Profiler Plus Kits (Life Technologies catalog # 4303326) at the Histocompatibility and Tissue Typing Laboratory, Brigham and Women's Hospital, Boston, MA. See Table 3.3 for results of this genotyping.

Table 3.3. STR genotyping on an ABI 3130 Genetic Analyzer

STR Locus	Th153.09	Th093.09	Th138.09	Th142.09	Th108.09	Th109.09
D3S1358	16,18	15,16	17,18	15,–*	15,16	15,16
vWA	15,18	19,20	13,16	15,20	15,16	15,16
FGA	24,26	21, –	22,26	24, –	22,23	21,24
Amelogenin	X,Y	X,Y	X,X	X,Y	X,Y	X,Y
D8S1179	11,14	12, –	15,16	14,15	15, –	13,16
D21S11	28, –	33,34	28,29	29, –	28,30	28,30
D18S51	11,15	12,19	16,17	17,20	17,18	17,18
D5S818	10,12	11,12	11,13	10,13	10,12	10,12
D13S317	12,13	12, –	12,13	11,12	10,12	11,12
D7S820	10,11	11, –	10, –	8,10	9,13	9, –

* "– " represents missing data.

Data Analysis

We excluded from analysis those samples with missing data on more than four SNP positions. We determined that samples with more than one site showing both fluorescent signals in genotyping (indicating that more than one allele were present) were "mixed infections" with more than one genome present in the patient sample. For simplicity, the results we show in the paper are all based on samples with single genome. We also considered mixed infection in the analyses, and the results do not change qualitatively.

We calculated the standardized index of association (I_A^S) by the program LIAN, version 3.5 (Haubold and Hudson 2000). The number of re-samplings was set to be 10,000. We assumed there are two generations per year and estimated variance effective population size through temporal changes in allele frequencies by both the moment method (Waples 1989) and likelihood approximation implemented in program CoNe (Anderson 2005). We calculated the ratio of parasites persisting between years in each year through dividing the number of barcodes that are shared with other years by the total number of barcodes in a particular year.

3.3 Results

Identification of repeated barcodes

We sampled patients annually from 2006–2011, from the Service de Lutte Anti-Parasitaire (SLAP) clinic in Thiès, Senegal under ethical approval, and genotyped the samples using the barcode. We first compared molecular barcodes within and between years. We confined this analysis to infections caused by a single parasite strain to reduce ambiguity from heterozygosity. The most prominent signal in our longitudinal collection of molecular barcode data was a steady increase in the number of identical barcodes observed in distinct patient samples (Figure 3.1). Whereas 10% of samples shared barcodes during the 2006 transmission season, more than 50% were within identical-barcode clusters in 2010 and 2011. Repeated instances of the same barcode were not limited to clusters of 2 or 3; in 2008 one barcode was observed in 22 distinct patient samples, and in 2011 nearly a quarter of the sampled infections exhibited another shared barcode. Overall, the proportion of unique parasite types decreased significantly over the study period (Figure 3.1B; P = 0.006, ANOVA). We investigated whether parasite samples exhibiting identical SNP barcodes are also genetically identical at other sites in the genome by hybridizing multiple clusters of samples with shared barcodes to a whole-genome SNP array that interrogates 17,000 polymorphic positions (Van Tyne et al. 2011). Parasite samples sharing barcodes exhibited array-based genotype profiles as similar to each other as technical replicate hybridizations of a single laboratory reference strain (Figure 3.2), suggesting that samples sharing barcodes are nearly

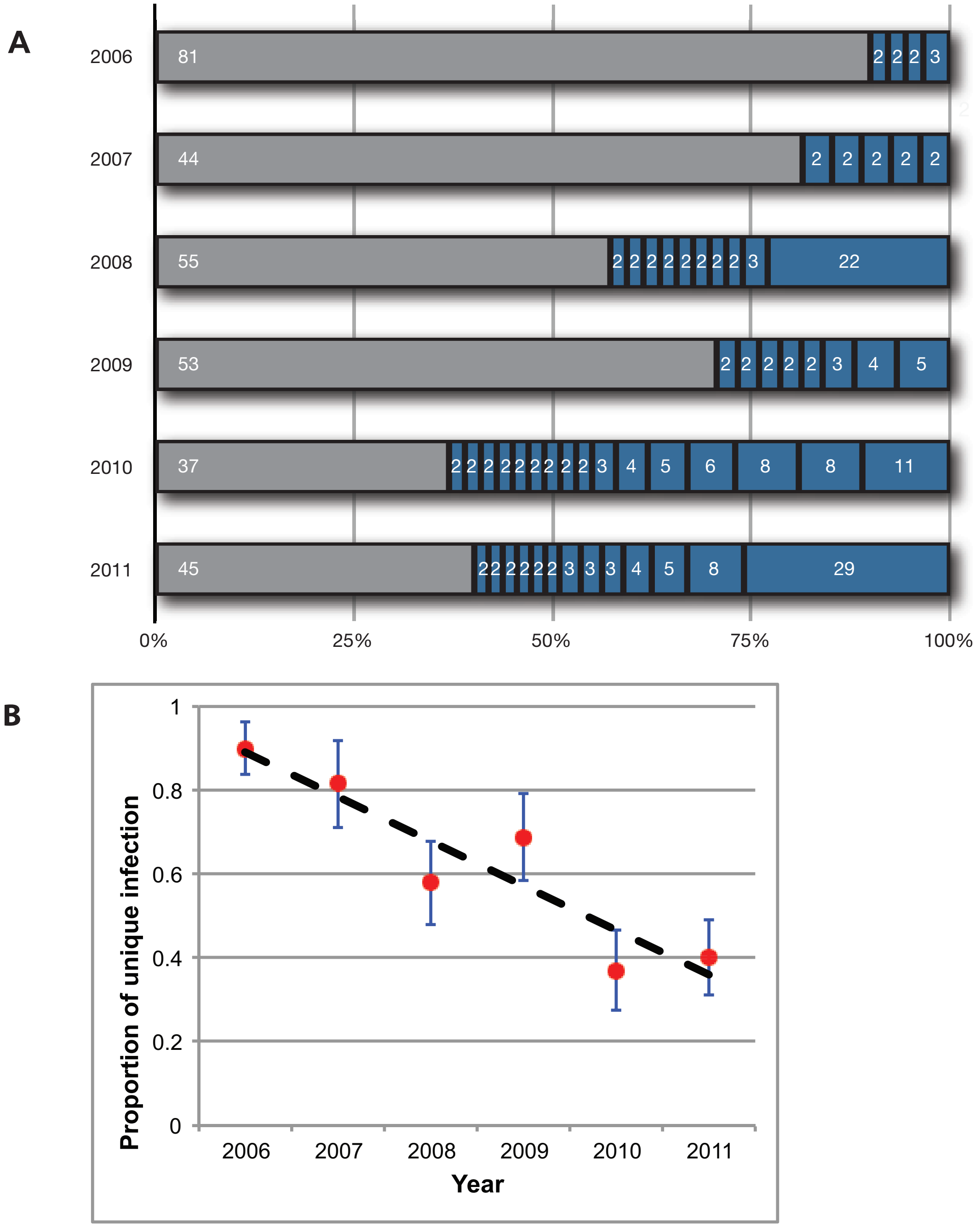
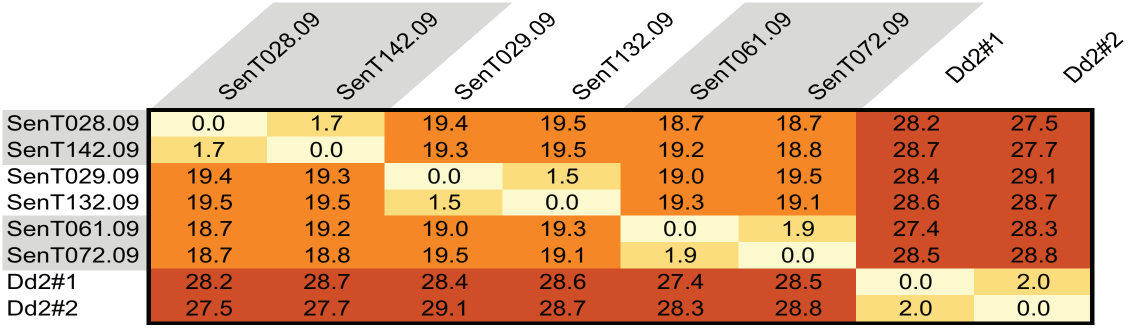


Figure 3.1. (A) Decreasing prevalence of unique parasite barcode profiles. For every collection season, the number of samples with unique barcodes (grey) and the number of samples in each shared-barcode cluster (blue) are shown. **(B) Ratio of shared vs. unique barcode profiles.** The proportion of samples residing outside of shared-barcode clusters is shown per year. The error bars show 95% confidence interval of mean (±1.96 SE).



	SenT028.09	SenT142.09	SenT029.09	SenT132.09	SenT061.09	SenT072.09	Dd2#1	Dd2#2
SenT028.09	0.0	1.7	19.4	19.5	18.7	18.7	28.2	27.5
SenT142.09	1.7	0.0	19.3	19.5	19.2	18.8	28.7	27.7
SenT029.09	19.4	19.3	0.0	1.5	19.0	19.5	28.4	29.1
SenT132.09	19.5	19.5	1.5	0.0	19.3	19.1	28.6	28.7
SenT061.09	18.7	19.2	19.0	19.3	0.0	1.9	27.4	28.3
SenT072.09	18.7	18.8	19.5	19.1	1.9	0.0	28.5	28.8
Dd2#1	28.2	28.7	28.4	28.6	27.4	28.5	0.0	2.0
Dd2#2	27.5	27.7	29.1	28.7	28.3	28.8	2.0	0.0

Figure 3.2. The percent differences between hybridized biological replicates and samples with identical barcodes. Array analysis shows that the percentage of SNP differences between samples with identical barcodes is similar to those seen in biological replicates, suggesting that samples with identical barcodes are nearly genetically identical.

genetically identical and likely derived from the same ancestor.

Clonal propagation vs. epidemic expansion

The increasing occurrence of repeated barcodes (i.e. nearly genetically identical samples) in later years could be attributed to either “clonal propagation” or “epidemic expansion”, or both. Clonal propagation is intrinsically linked to low parasite transmission, owing to the life history of *Anopheles* mosquito vectors. Female *Anopheles* mosquitoes ingest haploid *P. falciparum* gametocytes during a blood meal from a human host. The gametocytes differentiate into gametes in the mosquito midgut, where they unite to form a diploid zygote, which in turn undergoes meiosis to restore haploidy prior to inoculation of the next human host. Genetic outcrossing during the parasite’s sexual stage occurs only when a mosquito bites a host infected simultaneously by multiple parasite strains and gametocytes from multiple genetically distinct strains circulate in the blood of a host; bites of singly-infected hosts result in the union of nearly genetically identical gametes in the mosquito midgut, and consequently result in self-fertilization and clonal parasite transmission. To test this possibility, we compared the proportion of multiple infections over time and found that the proportion of mixed infections was significantly greater in 2006-2007 compared to subsequent years (Figure 3.3 and Table 3.1). While the patient parasitemia reported for those years varied between years, there was no trend in decreased parasitemia or sampling bias that could

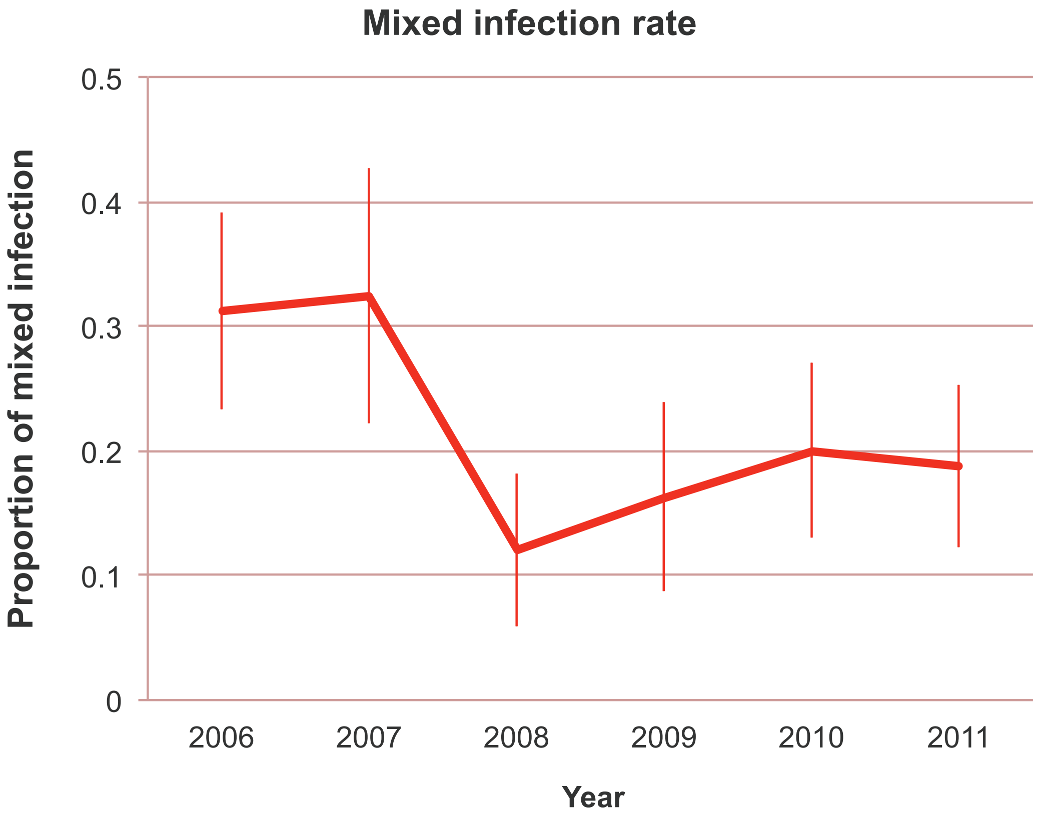


Figure 3.3. Mixedness over time. Proportion of mixed infections decreased between 2007 and 2008. The error bars show 95% confidence interval of mean (±1.96 SE).

contribute to the trend (Figure 3.4). This pattern of decreasing proportion of multiple infections is consistent with the decrease in the proportion of unique barcodes in Figure 3.1B, suggesting that "clonal propagation" due to decreased outcrossing is also consistent with the appearance and increase of repeated barcodes.

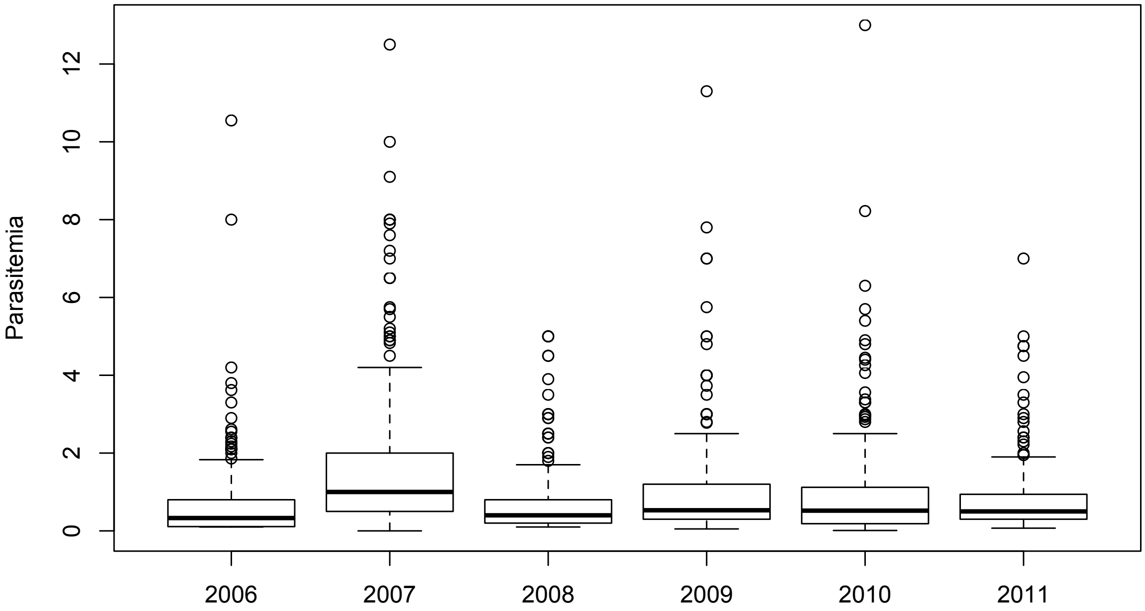


Figure 3.4. Parasitemia variation by year. Kruskal-Wallis rank-sum test indicates variance between years (p-value < 2.2e-16); however, there is no decreasing trend over time.

"Epidemic expansion" means that particular clones expand in the population, perhaps due to advantageous haplotypes, or a founder effect at the beginning of each transmission season, or both. Factors promoting variance in reproductive success, such as enhanced production of gametocytes, evasion of the host immune response, or enhanced transmission by selected or alternative mosquito vectors could select and enrich for favored parasite lineages in the population. Epidemic expansion is supported by the observation of two exceptionally prevalent barcodes in 2008 and 2011 (shown in Figure 3.5). To further test the possibility of epidemic expansion in our population, we used the framework described in Maynard Smith et al. (Smith, Smith, and O'Rourke 1993) and Anderson et al. (Anderson et al. 2000). We compared multilocus linkage disequilibrium (LD) using the standardized index of association (I_A^S) (Haubold and Hudson 2000), when including and excluding samples with the same barcode. The result shows significant LD from 2008 to 2011 when all samples are included, and no significant LD when only considering unique barcodes (Table 3.4), suggesting that the significant LD from 2008 to 2011 is caused by repeated barcodes; that is, some epidemic clones. There is no significant LD in 2006 and 2007 whether we included or omitted repeated barcodes. The lack of significant LD in 2006 and 2007, and the restoration of linkage equilibrium from 2008 to 2011 after excluding repeated barcodes suggest that the background population is still under linkage equilibrium and the decrease in the population recombination rate due to lowered transmission is very recent. Taken together with our analyses of the proportion of mixed infections, a likely explanation for these

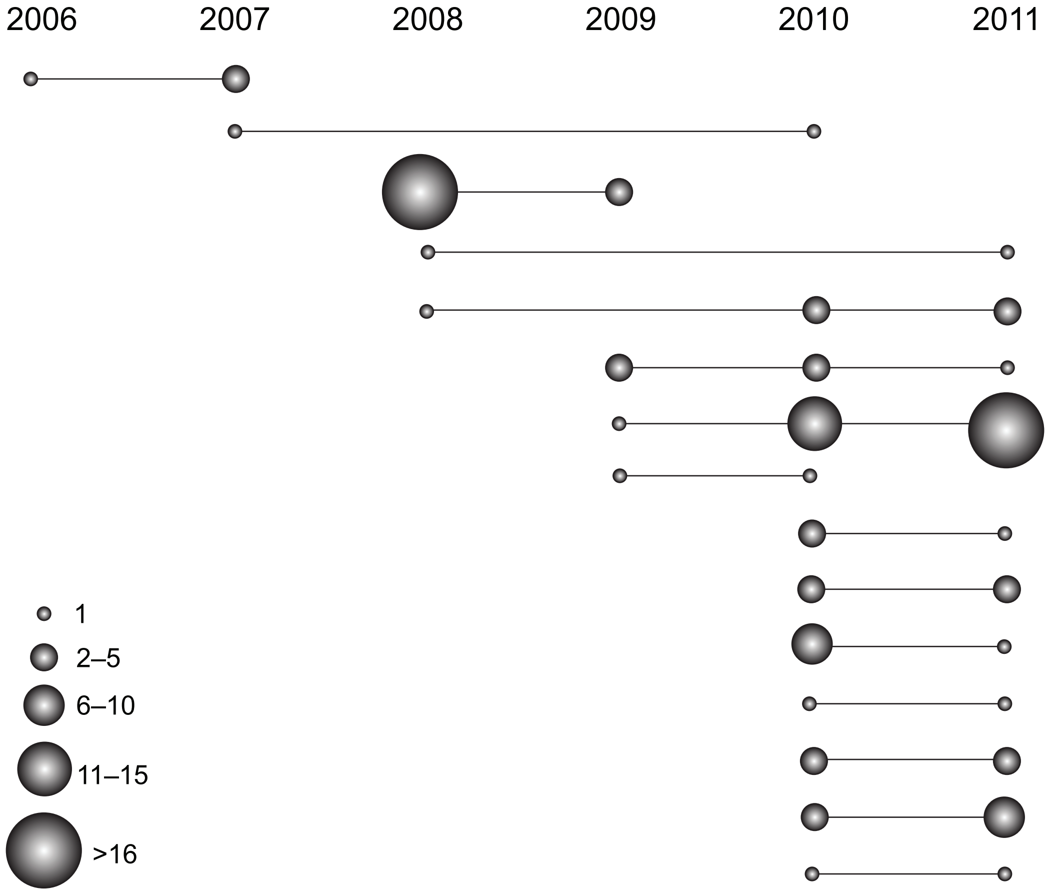


Figure 3.5. Clonal transmission of parasites across transmission seasons. Size and distribution of same parasite types across collection years.

Table 3.4. Multilocus linkage disequilibrium.

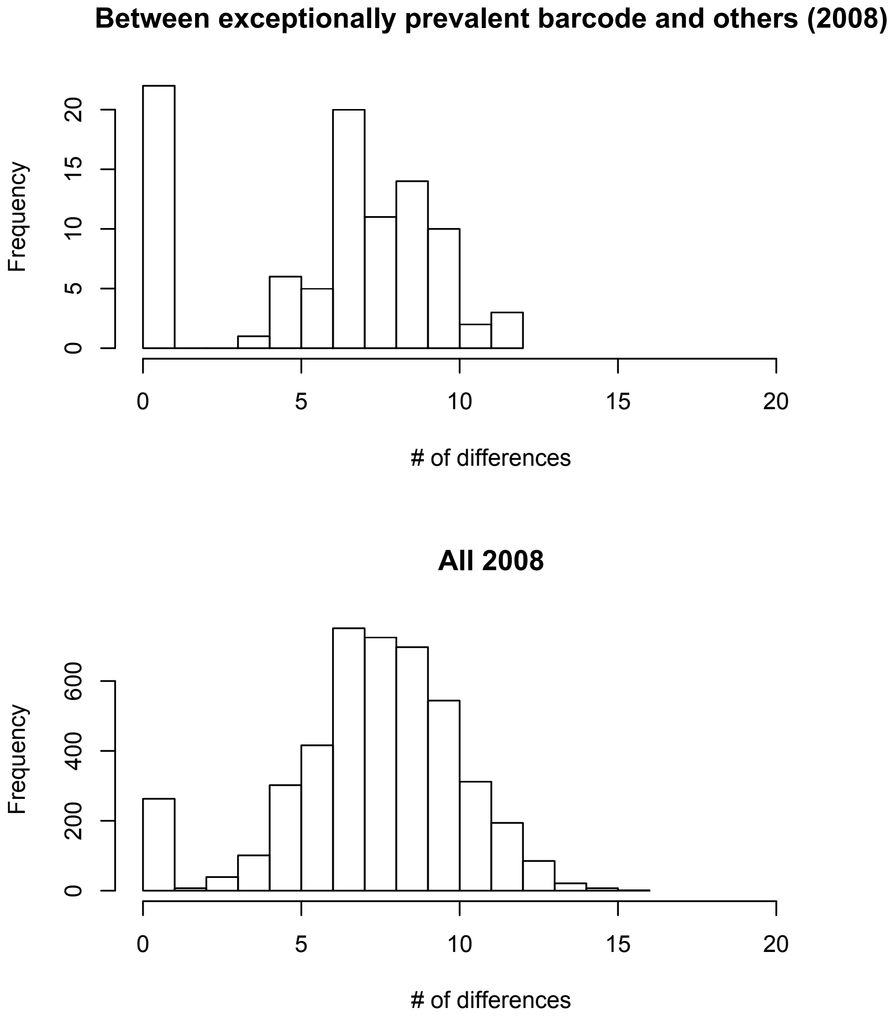
		2006	2007	2008	2009	2010	2011
All samples	I_A^S	-0.0049	0.0049	0.0306	0.0072	0.0293	0.1111
	P-value	0.989	0.162	$<10^{-4}$	0.021	$<10^{-4}$	$<10^{-4}$
Unique barcodes only	I_A^S	-0.0059	0.002	0.0012	0.0038	0.0042	0.0041
	P-value	0.997	0.332	0.329	0.151	0.145	0.128

observations is a reduction of outcrossing in 2007-2008 followed by an expansion of individual parasite genotypes.

Moreover, we examined whether parasite samples with shared barcodes were collected in proximal dates. The difference in collection dates among samples with identical barcodes is significantly smaller than that among samples with different barcodes (Wilcoxon rank sum test, *P* = 0.005), suggesting temporal expansion of particular clones in the population. However, because there was no temporal trend of increasing prevalence of a single parasite type (Figure 3.5) we do not believe that this was a selection event caused by emergence of drug resistance. It could possibly be a selection caused by emergence of resistance to host immune response, but the advantage disappears over time due to the corresponding changes in host, or non-selective forces. Alternatively, it might be possible that the parasite clones that appear to expand in the community were derived from an imported line novel to the area and thus the local population has little "strain-specific" immunity. We compared the pairwise differences between two exceptionally prevalent barcodes in 2008 and 2011 and the rest of strains with the pairwise differences among all strains from the same year, and found that the differences between two prevalent repeated barcodes and the rest of strains are not significantly higher than the differences among all strains from the same year (Figure 3.6). This result indicates that we do not observe the evidence of imported lines from the current data. Additional sequence information of polymorphic sites will be helpful to distinguish migrants from local population.

Effective population size

Reduced transmission can lead to lower parasite effective population size (N_e). To test whether the deployment of intervention strategies in recent years reduces malaria transmission, we examined the parasite effective population size (N_e). Population genetic theory predicts that a decreasing population should undergo increased genetic drift, manifested as increasingly variable allele frequencies across generations. The relevant measure of effective size in this context is the variance effective population size; this we estimated by measuring the fluctuation in allele frequencies across transmission seasons of the SNPs comprising the molecular barcode. We observed large fluctuations in allele frequencies over time (Figure 3.7). The variance N_e was calculated by all polymorphic SNPs using a likelihood approximation (Methods) and observed an extremely small variance N_e over time (Table 3.5). The estimated variance effective size in 2011 is only 10, a strikingly low value that reflects large fluctuations in allele frequencies. In



Between exceptionally prevalent barcode and others (2008)

All 2008

Figure 3.6. Pairwise differences between two exceptionally prevalent barcodes and other barcodes. The number of pairwise differences between the exceptionally prevalent barcode and other barcodes is not significantly higher than the number of pairwise differences among all the barcodes in the same year.

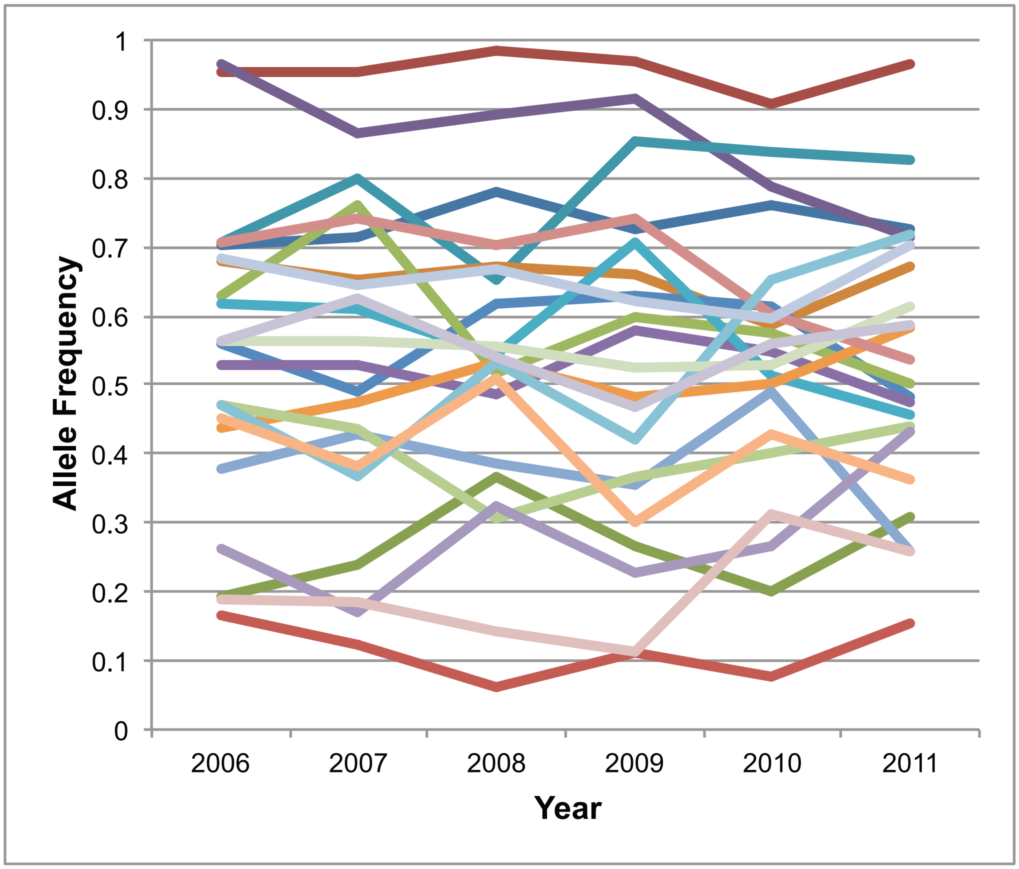


Figure 3.7. Changes in allele frequencies between transmission seasons. Each colored line shows the year-to-year variability in allele frequency for each non-fixed SNP. Each colored line shows the year-to-year variability in allele frequency for each non-fixed SNP. Allele frequencies fluctuate substantially across years, suggesting high random genetic drift and low effective population size.

Table 3.5. Variance effective population size.

	Moment		Likelihood	
	Mean	95% CI	Mean	95% CI
All samples				
2006-2007	ND*	(70, ND)	ND	(226, ND)
2007-2008	18	(7, 46)	19	(9, 49)
2008-2009	24	(9, 67)	29	(12, 90)
2009-2010	16	(7, 36)	18	(9, 42)
2010-2011	9	(4, 16)	10	(6, 18)
Ignoring duplicate barcodes				
2006-2007	ND	(116, ND)	ND	(132, ND)
2007-2008	82	(16, ND)	85	(19, ND)
2008-2009	138	(21, ND)	197	(27, ND)
2009-2010	59	(14, ND)	77	(18, ND)
2010-2011	240	(22, ND)	214	(24, ND)

* ND represents “Not Determinable”.

order to exclude the possibility that some particular parasite types are so successful in the population that lower the estimate of effective population size, we also calculated N_e by counting each repeated barcode once (Table 3.5). The estimates of N_e are still very small (less than 250) although some of the confidence intervals could not be determined. This extremely small effective population size predicts low effectiveness of selection efficiency and low rate of adaptation in Senegal.

Persistence across years

We also investigated the barcode dataset for evidence of clonal parasite persistence across years. Malaria transmission in Senegal is sharply seasonal, coinciding with annual rainfall patterns. Some parasite clones did indeed appear in more than one transmission season (Figure 3.5). These included clonal parasite types that persisted into the subsequent year and some that persisted longer, sometimes reappearing two or three seasons after initial detection. The increasing ratio of parasites persisting between years from 2006 to 2011 was statistically significant (P = 0.008, ANOVA) (Figure 3.8). Notably, we found an increase in the frequency of identical-barcode parasites persisting between 2010 and 2011: of the 15 identical barcodes that persisted for at least one year, ten were found during that pair of years. Because parasite samples sharing the same barcode are likely to be identical by descent, the persistence of identical barcodes across years suggests multiple sequential transmission cycles among singly-infected hosts, and indicates clonal propagation.

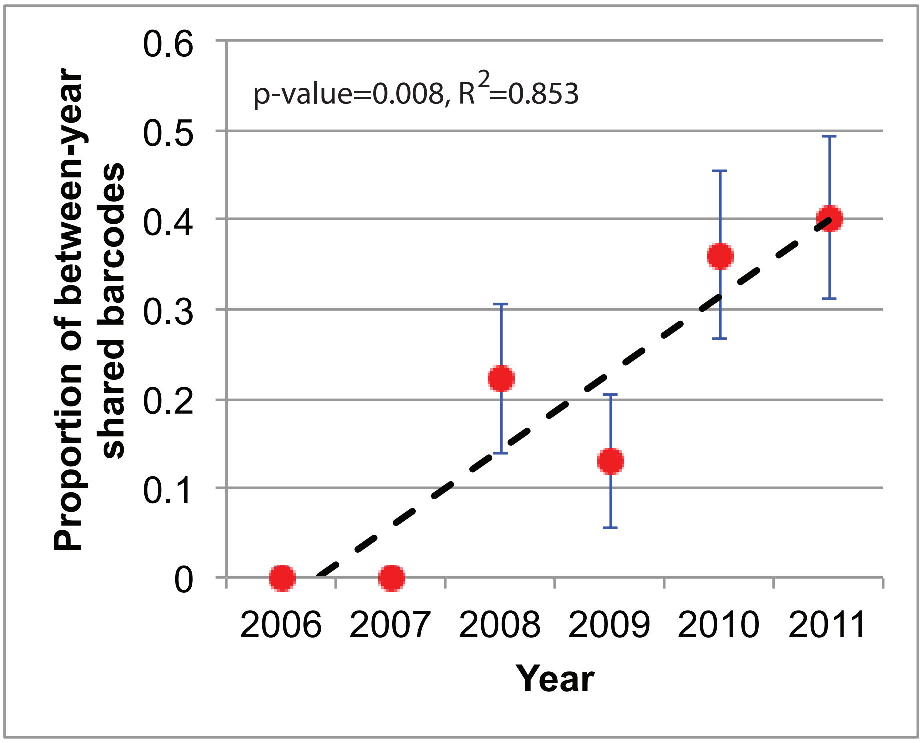


Figure 3.8. Proportion of between-year shared barcodes. Proportion of between-year shared barcodes increased significantly. The error bars show 95% confidence interval of mean (±1.96 SE).

To explore the patterns of repeated barcodes, and to rule out sampling biases in our study design, we examined the spatial and temporal relationships between samples exhibiting identical barcodes. We insured that clonal parasites were derived from independent natural infections by assaying 18 SNPs in the human host genetic material. We found no evidence of serial sampling of the same host among samples exhibiting the same barcode (Table 3.2 and Table 3.3). Examination of patient data confirmed that barcodes observed more than once were not clustered by household, ruling out a simple hypothesis of transmission among family members. Further analysis of the parasites within these samples by sequencing of the highly-polymorphic T-epitope region of the *csp* gene provided further evidence of highly related parasites (Table 3.6). We found that samples with identical barcodes are distributed across the entire transmission season and clinical catchment area, indicating a lack of temporal or spatial clustering. Our data therefore suggest a regional-level change in transmission dynamics from 2006 to 2011, rather than localized shifts.

Moreover, we compared ages of hosts before and after we observed the significant increase in the frequency of repeated barcodes. There is no significant difference in host ages between 2006-2007 and 2008-2011 (t test, *P*=0.094), suggesting that the patterns of identical barcodes are unlikely to be confounded by host ages.

Table 3.6. ***csp*** **sequences.**

3D7	A T T T A A A C A A A A T A C A A A A T T N C T C T T T C A A C T G A A T G G T C C C C A T G T A G T G T A A C T T
Th110.08	A T T T A A A G A A A A T A C A A A A T T - C T C T T T C A A C T G A A T G G T C C C C A T G T A G T G T A A C T T
Th250.08	A T T T A A A G A C A A T A C A A A A T T - C T C T T T C A A C T G A A T G G T C C C C A T G T A G T G T A A C T T
Th045.08	A T T T G A A G - A A A T A C T T A A T T - C T C T T T C A A C T G A A C G G T C C C C A T G T A G T G T A A C T T

Th216.08	A T T T A A A G G A A A T A C A A A A T T - C T C T T T C A A C T G A A T G G T C C C C A T G T A G T G T A A C T T
Th109.09	A T T T A A A G G A A A T A C A A A A T T - C T C T T T C A A C T G A A T G G T C C C C A T G T A G T G T A A C T T
Th088.08	A T T T A A A G G A A A T A C A A A A T T - C T C T T T C A A C T G A A T G G T C C C C A T G T A G T G T A A C T T
Th068.08	A T T T A A A G G A A A T A C A A A A T T - C T C T T T C A A C T G A A T G G T C C C C A T G T A G T G T A A C T T
Th042.08	A T T T A A A G G A A A T A C A A A A T T - C T C T T T C A A C T G A A T G G T C C C C A T G T A G T G T A A C T T
Th228.08	A T T T A A A G G A A A T A C A A A A T T - C T C T T T C A A C T G A A T G G T C C C C A T G T A G T G T A A C T T
Th096.08	A T T T A A A G G A A A T A C A A A A T T - C T C T T T C A A C T G A A T G G T C C C C A T G T A G T G T A A C T T
Th188.08	A T T T A A A G G A A A T A C A A A A T T - C T C T T T C A A C T G A A T G G T C C C C A T G T A G T G T A A C T T
Th173.08	A T T T A A A G G A A A T A C A A A A T T - C T C T T T C A A C T G A A T G G T C C C C A T G T A G T G T A A C T T
Th120.08	A T T T A A A G G A A A T A C A A A A T T - C T C T T T C A A C T G A A T G G T C C C C A T G T A G T G T A A C T T
Th101.08	A T T T A A A G G A A A T A C A A A A T T - C T C T T T C A A C T G A A T G G T C C C C A T G T A G T G T A A C T T
Th108.09	A T T T A A A G G A A A T A C A A A A T T - C T C T T T C A A C T G A A T G G T C C C C A T G T A G T G T A A C T T
Th112.08	A T T T A A A G G A A A T A C A A A A T T - C T C T T T C A A C T G A A T G G T C C C C A T G T A G T G T A A C T T
Th105.08	A T T T A A A G G A A A T A C A A A A T T - C T C T T T C A A C T G A A T G G T C C C C A T G T A G T G T A A C T T
Th214.08	A T T T A A A G G A A A T A C A A A A T T - C T C T T T C A A C T G A A T G G T C C C C A T G T A G T G T A A C T T
Th073.08	A T T T A A A G G A A A T A C A A A A T T - C T C T T T C A A C T G A A T G G T C C C C A T G T A G T G T A A C T T
Th044.08	A T T T A A A G G A A A T A C A A A A T T - C T C T T T C A A C T G A A T G G T C C C C A T G T A G T G T A A C T T
Th004.08	A T T T A A A G G A A A T A C A A A A T T - C T C T T T C A A C T G A A T G G T C C C C A T G T A G T G T A A C T T
Th014.08	A T T T A A A G G A A A T A C A A A A T T - C T C T T T C A A C T G A A T G G T C C C C A T G T A G T G T A A C T T
Th220.08	A T T T A A A G G A A A T A C A A A A T T - C T C T T T C A A C T G A A T G G T C C C C A T G T A G T G T A A C T T

Th096.06	A T T T A A A G A C A A T A A A A A A T T - C T C T T T C A A C T G A A T G G T C C C C A T G T A G T G T A A C T T
Th134.06	A T T T A A A G A C A A T A A A A A A T T - C T C T T T C A A C T G A A T G G T C C C C A T G T A G T G T A A C T T

Th127.08	A T T T A A A G A G A A T A C A A A A T T - C T C T T T C A A C T G A A T G G T C C C C A T G T A G T G T A A C T T
Th246.08	A T T T A A A G A G A A T A C A A A A T T - C T C T T T C A A C T G A A T G G T C C C C A T G T A G T G T A A C T T

3.4 Conclusions and Discussion

With the restructuring of the National Malaria Control Programme (NMCP) in 2005, Senegal implemented an organized approach to malaria control and elimination. From 2006 to 2010, the NMCP increased access to insecticide-treated bednets (ITNs) and residual insecticide spraying, with the number of reported bednets per home increasing more than 35% from 2008 to 2010. Combined with no-charge access to ACTs from 2007, the country reported a 41% drop in the number of malaria cases between 2008 and 2009 (Malaria RB 2010). The findings of increasing repeated barcodes, persistence, and proportion of single infections across transmission seasons demonstrate the usefulness of genetic tools for monitoring the effectiveness of intervention strategies against infectious disease. This type of evidence could inform control efforts as a real-time gauge of the progress towards control, elimination, or eradication. Our ability to differentiate between clonal and epidemic population structures and to track these changes within the population could lend a more refined view of the subtle effects and varying degrees of effectiveness in control programs.

While our study reported the first evidence of clonal propagation and epidemic expansion in Africa, other groups have also used genetic tools to study parasite dynamics in geographically distinct regions, and reported clonal lineages and persistence over time (Branch et al. 2011; Griffing et al. 2011; Nkhoma et al. 2012). Roper et al. showed the persistence of parasites over the dry season in Sudan and Echeverry et al. showed similar in Colombia (Roper et al. 1996; Echeverry et al.

2002). Both Branch et al and Griffing et al point to distinct genetic types within South America and Peru in particular, and attribute population patterns to periodic epidemics in regions with relatively low transmission levels (Branch et al. 2011; Griffing et al. 2011). Similarly, Nkhoma et al showed the decreases in the proportion of unique parasite genotypes and the proportion of multiple infections along with large reduction in transmission over time. However, they found no evidence of reduction in N_e during the same period of time, which was possibly caused by migrations between nearby populations, or the lack of power in analysis of temporal data when the true N_e is not small enough (Nkhoma et al. 2012). Moreover, Mobegi et al. 2012 showed that the background of non-clonal population structure has been widespread elsewhere surrounding our study area in West Africa, indicating that there has been dramatic changes in the population structure of this site in contrast to the surrounding regional parasite population structure (Mobegi et al. 2012). These studies, including our study, indicate the power of using genetic tools to study parasite population structure, and highlight the need for further detailed study of parasite population dynamics in more extensive geographical regions to understand the interactions and migrations between different parasite populations.

Further applications of this approach might be to differentiate between parasite recrudescence or re-emergence in selected populations to allow facile decision-making in the face of a very changeable parasite where resistance emerges quickly (Daniels et al. 2008). With additional evidence provided by other types epidemiological studies to more directly link these parameters to parasite population genetics, changes in the profile of parasites with different molecular barcodes might be used as an indicator of parasite transmission. The finding is also one beneficial outcome of a genomic diversity project undertaken by the malaria community five years ago (Jeffares et al. 2007; Mu et al. 2007; Volkman et al. 2007). The decreasing cost and increasing translation of sequencing and genotyping tools into clinical environments will make genetic data invaluable for rapidly understanding diverse aspects of infectious disease epidemiology, particularly when such information is combined with population genetic inferences and knowledge of pathogen biology.

Chapter 4

The malaria life cycle intensifies both natural selection and random genetic drift

Hsiao-Han Chang and Daniel L. Hartl

4.1 Introduction

Malaria, caused by the parasite, *Plasmodium falciparum*, is one of the major causes of death in the world. To aid in developing vaccines or drug treatments for malaria, researchers have studied the *P. falciparum* genome and identified genes that are essential to malaria parasites as well as genes that are related to drug-resistance phenotypes using population genetic tools (Mu et al. 2010; Van Tyne et al. 2011; Amambua-Ngwa et al. 2012; Chang et al. 2012; Manske et al. 2012; Park et al. 2012). Researchers have also focused on particular genes related to drug resistance and characterized the evolutionary pathways of emerging drug resistance using *Escherichia coli* and *Saccharomyces cerevisiae* as model systems (Lozovsky et al. 2009; Brown et al. 2010; Costanzo et al. 2011; Toprak et al. 2012).

Malaria parasites have a unique and complex life cycle with two types of host organisms — humans and female *Anopheles* mosquitoes. Malaria parasites are transmitted from mosquito to humans through the bite of an infected mosquito. In the human host, the parasite reproduces asexually multiple times, and the within-human population size increases from $10–10^2$ at the time of infection to $10^8–10^{13}$ within a few weeks. When another female mosquito feeds on the blood of the infected human, $10–10^3$ malaria gametocytes are transmitted back to mosquito host, and these immature gametes undergo maturation, fuse to form zygotes, undergo sexual recombination and meiosis, and the resulting haploid cells reproduce asexually and form sporozooites that migrate to the salivary glands to complete the life cycle.

Much of population genetics is based on the concept of a Wright-Fisher (WF) population (Fisher 1930; Wright 1931). In the WF model, the population size is constant, generations are non-overlapping, and each new generation is formed by sampling parents with replacement from the current generation. The major differences between the malaria life cycle and the WF model are that each malaria life cycle includes two transmissions, multiple generations of asexual reproduction, and population expansions and bottlenecks. Before population genetic inferences can be conducted through methods based on WF assumptions, it is necessary to determine whether the malaria life cycle is sufficiently well described by the WF model. If the life cycle impacts population genetic behaviors, then inferences based on conventional interpretations of the WF model may need to be adjusted.

In a previous study, we observed two unusual patterns in the *P. falciparum* genome that had not been reported in any other organism (Chang et al. 2012). First, we observed synonymous and nonsynonymous allele-frequency spectra that were more similar than expected given that nonsynonymous sites likely experience stronger selection. Second, 18% of genes showed a ratio of nonsynonymous to synonymous polymorphism (π_N/π_S) greater than 1. Because nonsynonymous mutations result in a change in amino acid, they are likely to have a deleterious effect and exist in low frequencies in the population (if not yet being removed from the population). As expected, nonsynonymous allele-frequency spectrum was found to be more skewed toward low frequency alleles than synonymous allele-frequency spectrum. Examples include humans (Torgerson et al. 2009; Fujimoto et al. 2010; Li et al. 2010), *Oryctolagus cuniculus* (Carneiro et al. 2012), *Drosophila melanogaster*

(Lee and Reinhardt 2012), and *Capsella grandiflora* (Slotte et al. 2010). Moreover, in *D. melanogaster* (Langley et al. 2012), only less than 2% of genes have π_N/π_S greater than 1.

Potential explanations for these unusual patterns including sequencing error and annotation error could be ruled out, and dramatically relaxed selection or 18% of genes under diversifying selection (especially because some of the GO categories that enriched for significantly high π_N/π_S are less likely to be under diversifying selection) seems unlikely. Because of the uniqueness and complexities of the malaria life cycle, we wondered whether the malaria life cycle itself could explain part of these unusual patterns. More recent work in *P. vivax*, a close relative with similar life history to *P. falciparum*, also revealed large numbers of genes with π_N/π_S greater than 1 (Neafsey et al. 2012), supporting the idea that factors common to *Plasmodium* species but different from most other species may cause allele-frequency patterns that deviate from WF expectations.

Although the behavior of the WF model is relatively robust to deviations from many underlying assumptions, there are examples in which the WF model is known to perform poorly. For instance, Der et al. (2012) showed that the effect of selection is increased relative to the WF model when large family sizes in the distribution of offspring number are allowed. Their results highlight the importance of studying the effect of non-standard reproductive mechanisms on basic evolutionary outcomes. Although there has been research on the evolution of drug resistance under the malaria life cycle, in both mathematical models and computational simulations (Hastings 1997; Hastings 2006; Antao and Hastings 2011), it has not been ascertained whether the underlying processes of random genetic drift, natural selection, and their interactions yield outcomes under the malaria life cycle that are congruent with those of the WF model.

Here, we used computational simulation to examine how the malaria life cycle influences random genetic drift, natural selection, and their interactions. First, we compared quantities from generation to generation, including number of mutations after one generation and probability of loss, between a malaria model and the WF model. Second, we considered longer time-scale properties, including time to fixation or loss, segregation time, and probability of fixation or loss. Third, we simulated the allele-frequency spectrum under a neutral model with the malaria life cycle. The flexibility of the simulation framework enables us to investigate various combinations of selection coefficients.

4.2 Methods

Simulation

To simulate the evolution under the malaria life cycle, we used the following forward-time simulation framework (Figure 4.1):

(i) Assume there are N human hosts and $a \times N$ mosquito hosts, with D parasites (sporozoites) transmitted from the mosquito host to the human host. The initial condition is the number of mutations in the initial parasite pool within each human host.

(ii) Within a human host, the probability of a parasite that carries a particular allele surviving from one asexual generation to the next is $P \times (1 + s_h)$, where P is the probability that a parasite survives a given round of replication and s_h is selection coefficient of the allele within the human host. Whether or not a parasite dies is determined by the outcomes of a Bernoulli trial. If parasites remain alive, they reproduce to create two daughter cells. The maximum population size within a single human host is N_{eH}. After the population size reaches the maximum, it stays at the same population size for e additional WF generations.

(iii) The number of mosquitoes that obtain parasites (gametocytes) from each human host is based on multinomial sampling. If the mutation has a transmission advantage, the human host with the mutation has a $(1 + t_m)$-fold higher probability of transmitting gametocytes to mosquitoes. D parasites (gametocytes) are transmitted from the human host to the mosquito host during each bite.

(iv) Within a mosquito host, the probability of surviving from one generation to the next is $P \times (1 + s_m)$, where s_m is the selection coefficient of the allele within the mosquito host. As in step (ii) in the human host, whether or not a parasite dies is determined by the outcomes of a Bernoulli trial, and if parasites remain alive, they reproduce to create two daughter cells. The maximum population size within the mosquito host is N_{eM}. The population size increases until it reaches maximum size.

(v) The number of humans that acquire parasites (sporozoites) from each mosquito host is based on multinomial sampling. If the mutation has a transmission advantage, the mosquito host with the mutation has a [$(1+ t_h)$-fold] higher probability to transmit sporozoites to human individuals.

(vi) Repeat steps (ii) to steps (v) until the mutation becomes lost or fixed in the entire population. Steps (ii) to (v) correspond to a “generation” in the malaria life cycle model.

We repeated the simulation 500,000 times for each initial condition. Table 4.1 lists all the relevant parameters and their default values. Unless stated otherwise, the default values were used in the simulations. Simulations were performed using custom code written in C. This is available from the authors by request.

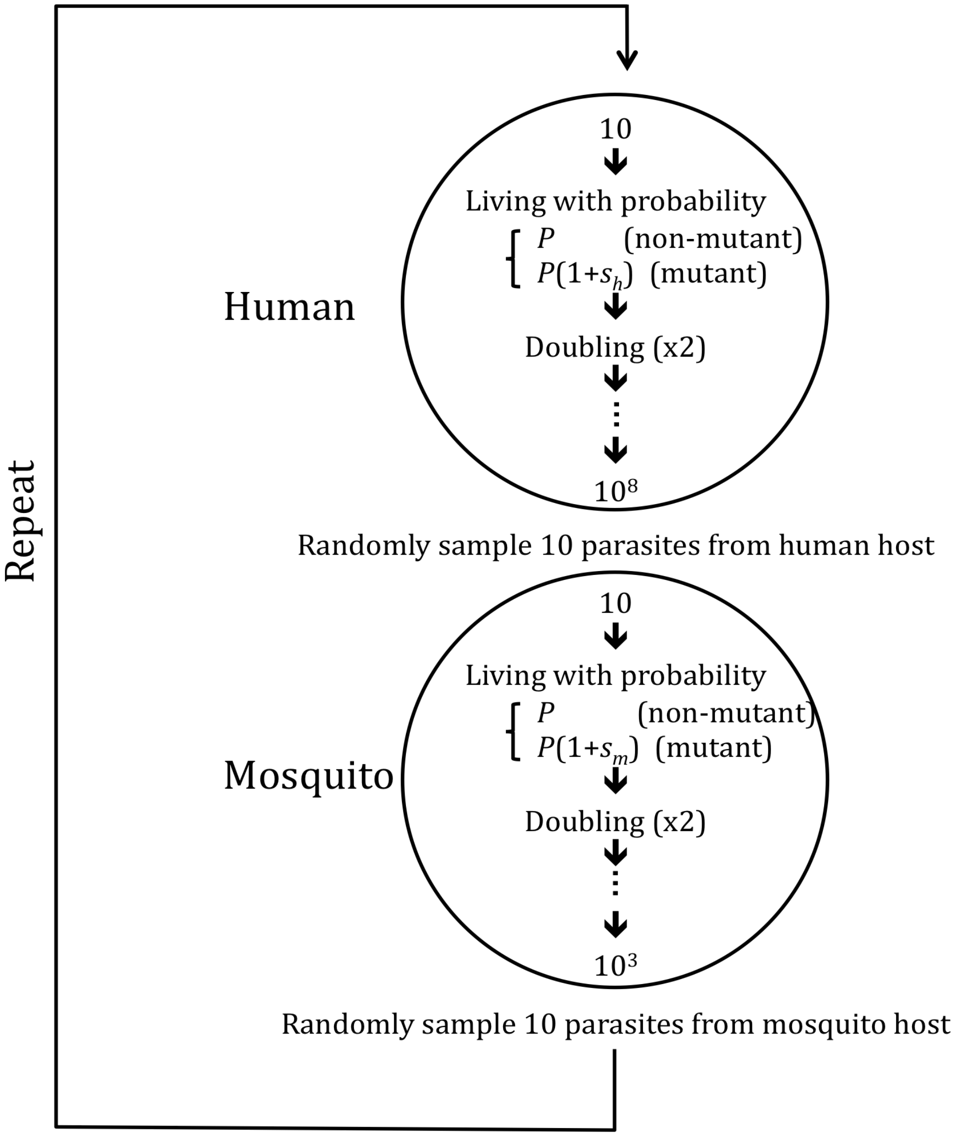


Figure 4.1. Simulation diagram.

Table 4.1. List of relevant parameters.

Parameters	Description	Default values
e	The number of additional generations after population stops expanding within human host	10
N	The number of human hosts	1000
a	The ratio of the number of mosquito hosts to the number of human hosts	1
D	The number of parasites transmitted from one host to another	10
P	The probability of living from asexual generation to generation during population expansion within hosts	0.9
s_h	Selection coefficient within human hosts	0
s_m	Selection coefficient within mosquito hosts	0
n	Population size in Wright-Fisher model	10000
N_{eH}	The maximum population size of malaria parasites within one human host	10^8
N_{eM}	The maximum population size of malaria parasites within one mosquito host	1000
t_m	Transmission coefficient from human to mosquito	1
t_h	Transmission coefficient from mosquito to human	1

The results of the WF models used for comparison were also obtained by simulations. We used 10,000 as the population size in the WF model because the default total parasite population size is 10,000 (10 transmitted parasites per host × 1000 hosts). The results of the WF models with population sizes 10^8 and 10^{11} are also shown in Figure 4.2. They are not qualitatively different from those obtained assume a population size of 10^4 and hence do not significantly alter the results of comparing the WF model with the malaria model.

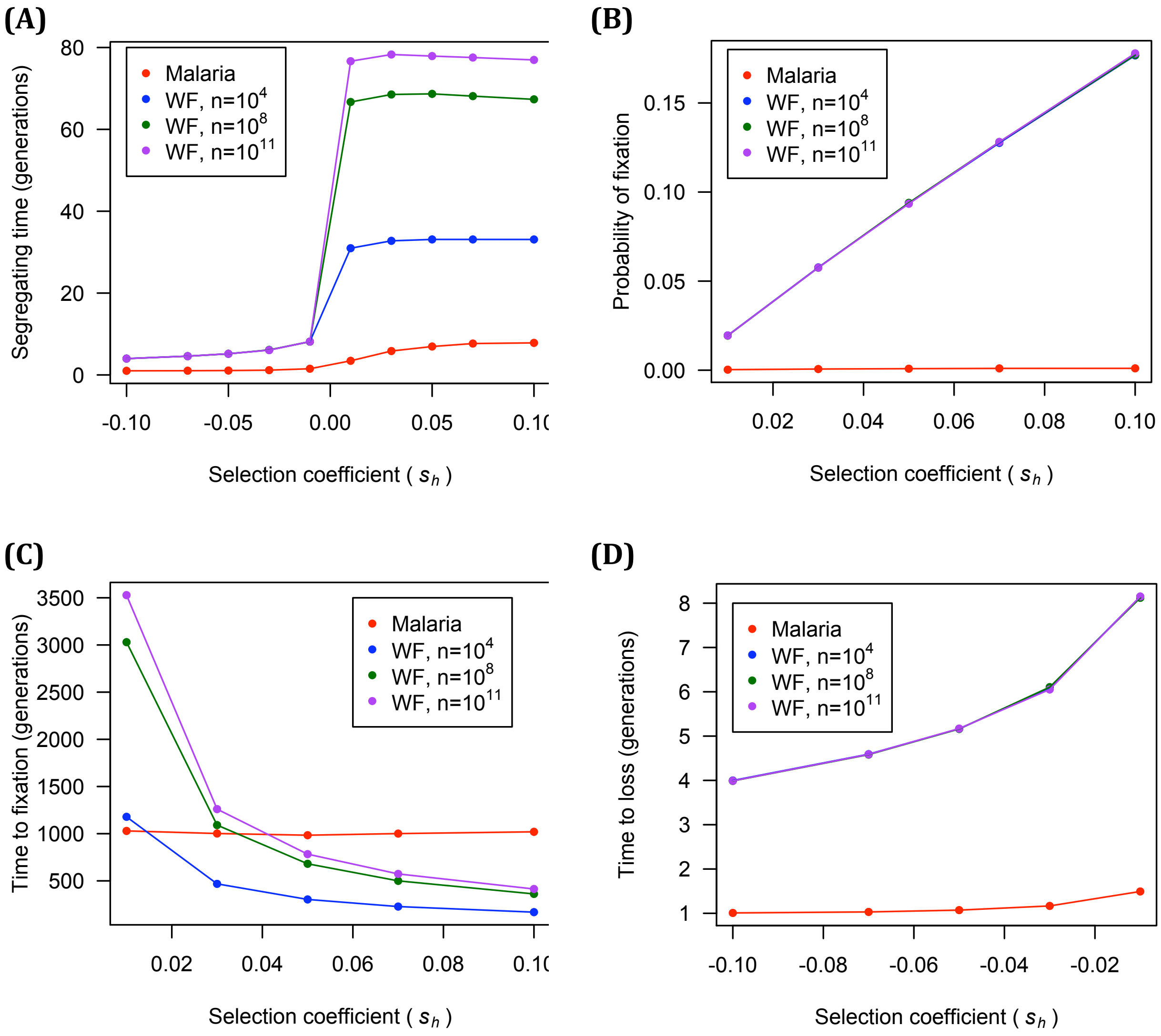


Figure 4.2. The Wright-Fisher model with various population sizes. The results from the Wright-Fisher model with different population sizes do not differ qualitatively and all of them are different from malaria model.

Allele frequency spectrum

To obtain the null allele frequency spectrum under the malaria life cycle, we simulated mutations and kept track of them until they became fixed or lost in the population. Then we sampled mutations weighted by the time that they remained segregating in the population. Mutations that remain in the population longer are more likely to be sampled. After a mutation was chosen, we randomly selected one time point during the interval that the mutation was segregating in the population, and at that time point sampled 25 parasites from 25 different human hosts, and recorded the allele frequency. To minimize the computational time for simulating new mutations, we first calculated the relative probabilities of different initial conditions, and combined the allele frequency spectra according to their weighted average . Initial conditions with probability less than 1/1000 of the most probable conditions contribute little to the null distribution and therefore were ignored in the analysis.

4.3 Results

From Generation to Generation

We first examined random genetic drift under the malaria life cycle and the WF model by comparing the probability of loss after one generation. Under the malaria life cycle, the probability of loss of a new neutral allele is as high as 74%, whereas it is approximately $e^{-1} \approx 37\%$ under the WF model. The discrepancy indicates that random genetic drift has much stronger effects under the malaria life cycle.

If we consider a non-neutral allele, the probability of loss under the malaria life cycle is greater than that in the WF model (Figure 4.3A), no matter whether the mutation is beneficial or deleterious. Interestingly, the average number of copies of a mutant allele one generation after its occurrence is also more extreme under the malaria life cycle (Figure 4.3B). On average, after one generation, beneficial alleles have more copies than in the WF model and deleterious alleles fewer copies, suggesting that selection works more efficiently under the malaria life cycle. This result is in contrast to that expected from existing population genetic theories. It is commonly thought that, when random genetic drift intensifies, the selection efficiency is reduced in the WF model. The results in Figure 4.3 show that random genetic drift and selection efficiency can both increase at the same time in the malaria life cycle.

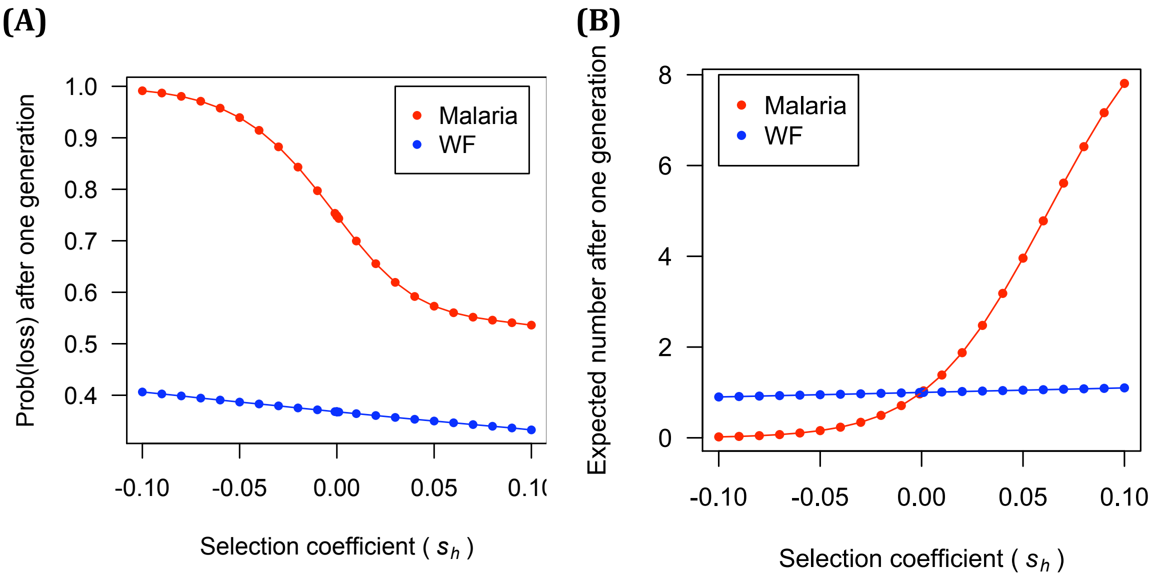


Figure 4.3. Comparison of probability of loss and average number of mutations after one generation between the malaria life cycle and the WF model. (A) Probability of loss is greater under the malaria life cycle. **(B)** Average number of mutations after one generation is more extreme under the malaria life cycle except when the mutation is neutral.

We tested whether the difference between the malaria life cycle and the WF model is sensitive to other parameters in the simulation by varying the values of other parameters, including *a*, *e*, *P* and *D*, in the simulation. The results are consistent and differ quantitatively but not qualitatively (Figure 4.4).

Longer time scale

We then considered properties on a longer time scale, including the segregation time (the average time until a mutation becomes fixed or lost in the population), the time to fixation, time to loss, and the fixation probability. The

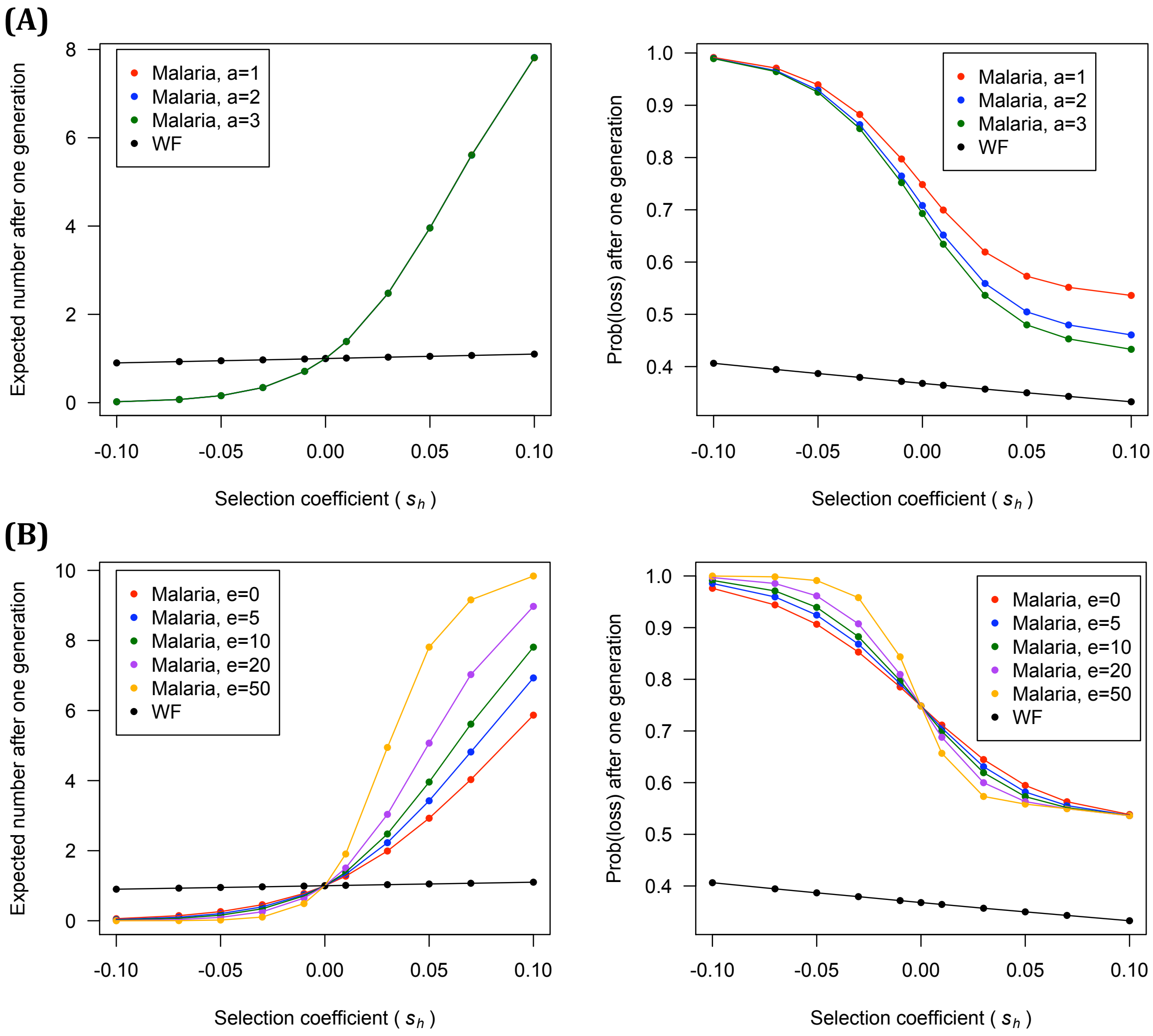


Figure 4.4. Malaria model when other variables are different from default settings.

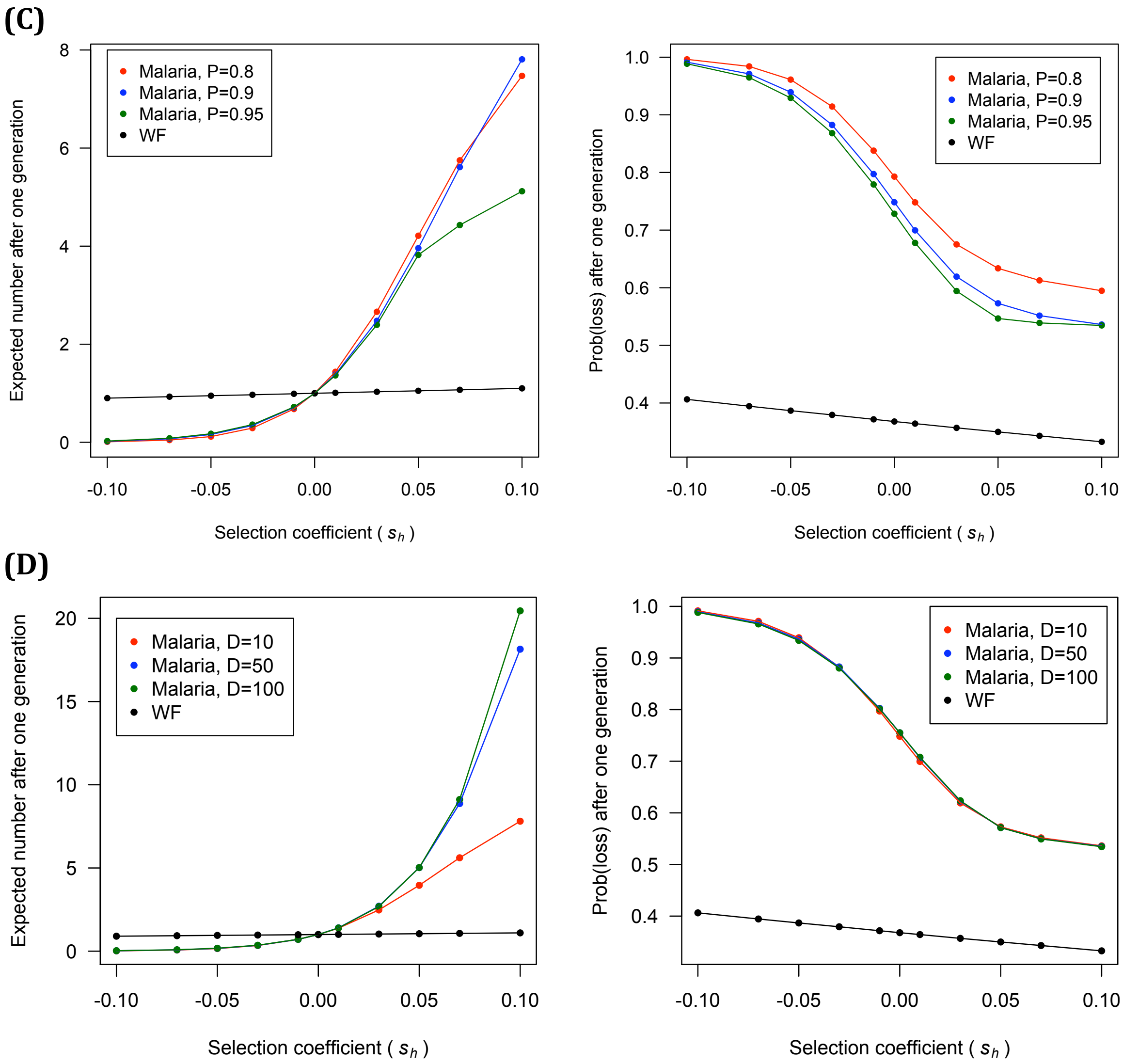


Figure 4.4 (Continued).

results indicate that the segregation time under the malaria life is shorter than in the WF model (Figure 4.5A). The mutations segregate on average for less than 8 generations under the malaria life cycle, even when the selection coefficient is as high as 0.1 because of the high genetic drift under malaria life cycle. The shortening of the segregation time indicates that a large proportion of segregating sites in the genome of malaria parasites are expected to be recently derived.

Time to fixation for beneficial mutations under the malaria life cycle is shorter than that in the WF model when the selection coefficient is smaller than a threshold value, and longer for larger selection coefficients (Figure 4.5B). When the selection coefficient is small, the time to fixation of beneficial mutations is shorter under the malaria life cycle because, after a mutation becomes fixed in one host, it has a greater increment in allele frequency in each generation because of the transmission of multiple parasites between human and mosquito. However, when the selection coefficient exceeds a threshold, then selection in the WF model is so efficient that fixation takes less time than in the parasite in spite of the transmission of multiple parasites. Nevertheless, the probability of fixation of beneficial alleles under the malaria life cycle is always smaller than that in the WF model (Figure 4.5C) owing to the enhanced random genetic drift and the stochastic nature of parasite transmissions among hosts.

Time to loss for deleterious mutations under the malaria life cycle is also shorter than that in the WF model (Figure 4.5D), suggesting that purifying selection is more efficient under the malaria life cycle and deleterious mutations are removed

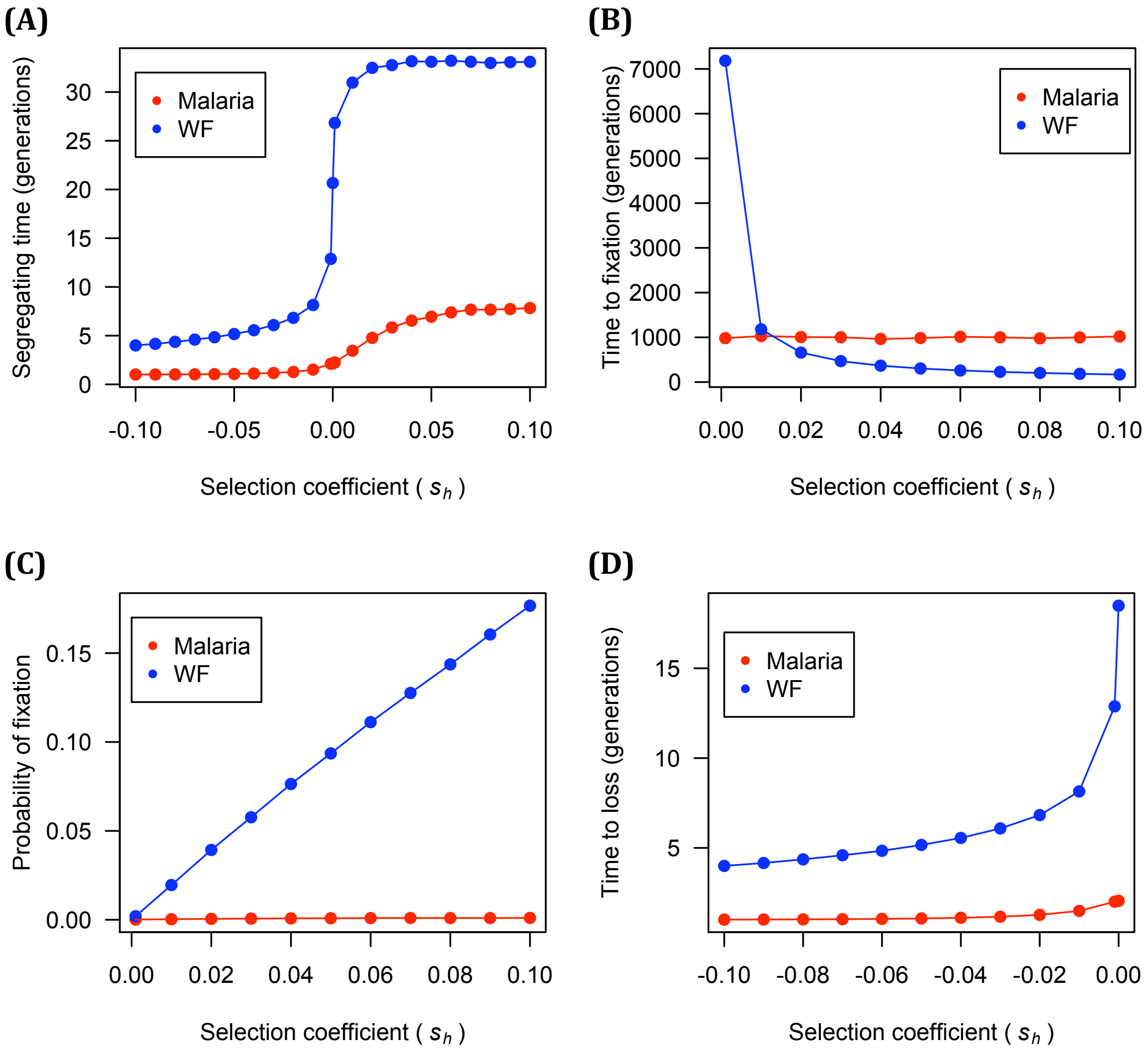


Figure 4.5. Comparison of longer time scale properties between the malaria life cycle and the WF model. (A) Segregation time is shorter under the malaria life cycle. **(B)** Time to fixation for beneficial alleles is shorter under the malaria life cycle when the selection coefficient is smaller than threshold value (*s* = 0.01 under the default settings), and is greater than in the WF model if the selection coefficient exceeds the threshold. **(C)** Probability of fixation of beneficial alleles under the malaria life cycle is smaller than in the WF model, due to greater effects of random genetic drift and stochastic transmission among hosts in the malaria life cycle. **(D)** Time to loss of deleterious alleles is shorter under the malaria life cycle, suggesting highly efficient purifying selection in the malaria parasite.

from the population very quickly, hence segregating mutations in the malaria parasite are less likely to be deleterious than mutations observed in other organisms with similar effective population sizes that evolve in accord with the WF model.

We examined whether these results are sensitive to values of parameters other than the selection coefficient by varying the values of other parameters, including a, e, P and D, in the simulation. The results are consistent and differ only quantitatively but not qualitatively (Figure 4.6).

Non-neutral Transmission

It has been shown that genetic factors influence the rate of conversion of gametocytes into male or female gametes (reviewed in Talman et al. 2004). Because gametocyte differentiation is critical for forming zygotes in the mosquito host and successful transmission, transmission among hosts could be affected by mutations in the parasite genome. We therefore performed the simulations in which transmission probabilities could be altered by mutations.

When mutation only affects the transmission probability (t_m-only model), beneficial mutations have even shorter segregation times than the s_{h}-only model, and deleterious mutations have slightly longer segregation time than the s_{h}-only model (Figure 4.7A). Fixation times for beneficial mutations in the t_m-only model and the "$t_{\mathrm{m}} = s_{\mathrm{h}}$" model are both shorter than in the s_{h}-only model and the WF model

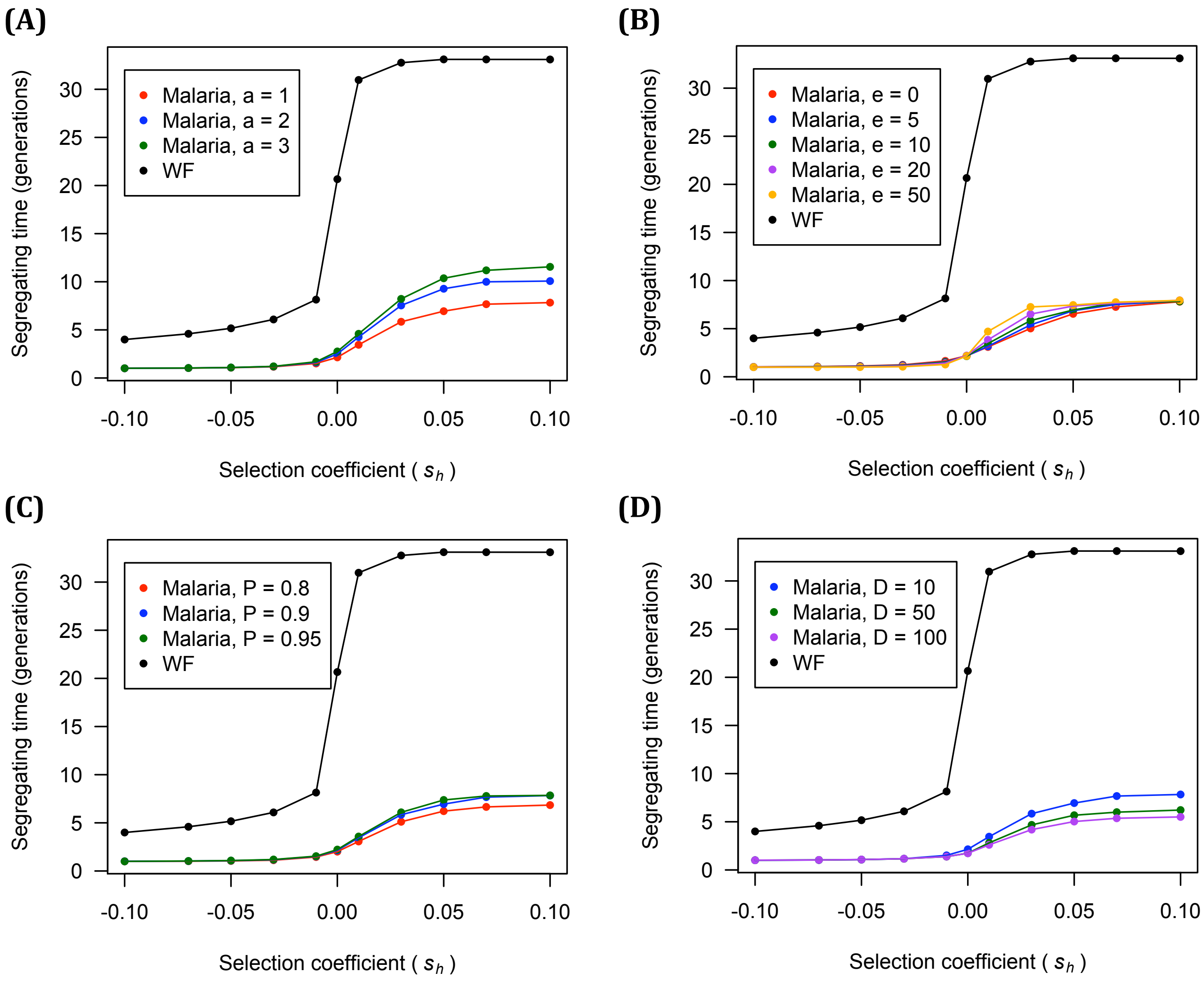


Figure 4.6. Longer time scale properties of malaria model when other variables are different from default settings.

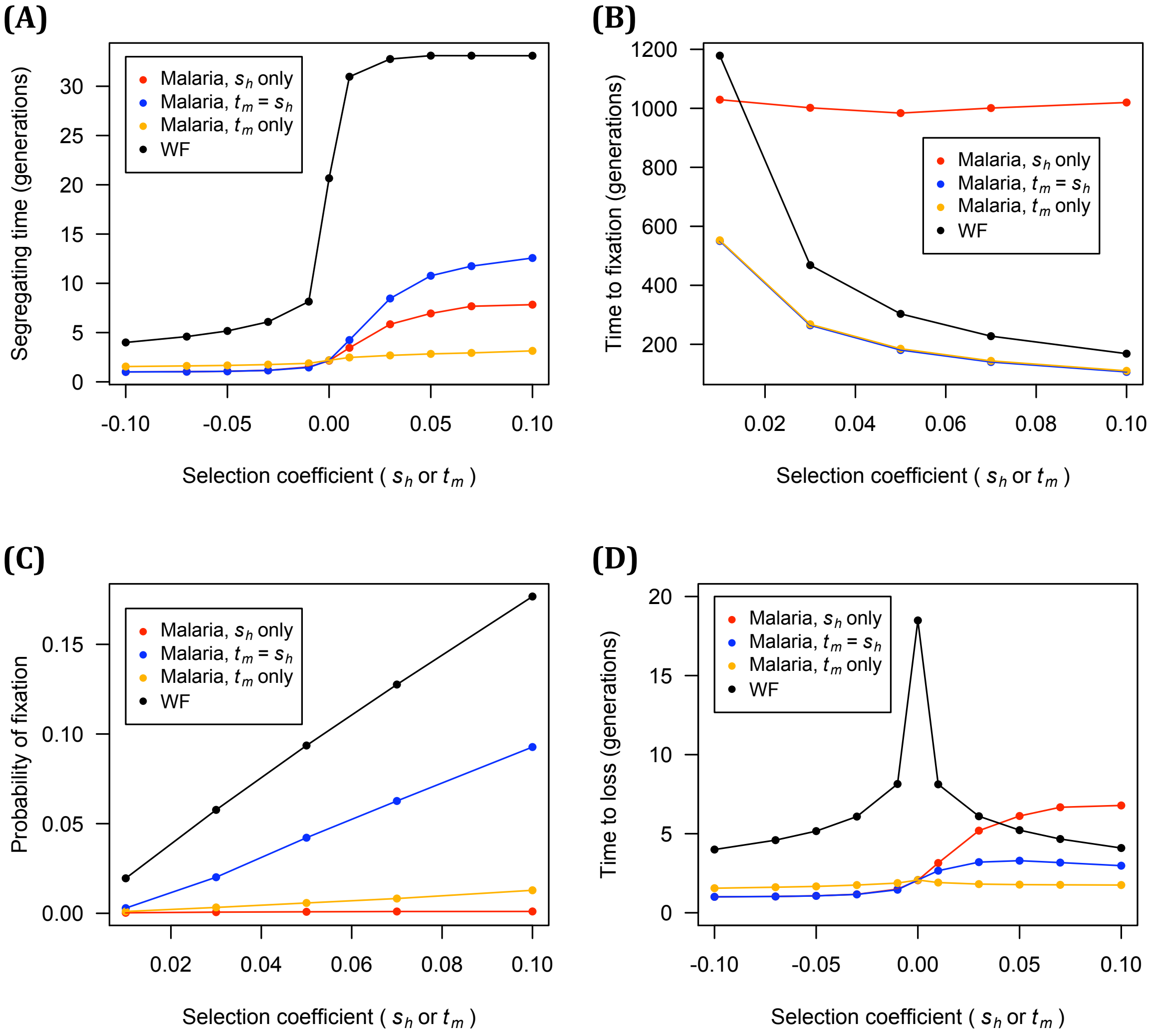


Figure 4.7. Simulations for non-neutral transmissions. In the "s_h-only" model (red line), only s_h varies and transmission probabilities are all the same. In the "t_m-only" model (yellow line), only t_m varies and s_h are all zero. In the "$t_m = s_h$" model (blue line), t_m and s_h are the same.

(Figure 4.7B). The fixation probability for beneficial mutations in the t_m-only model is larger than in the s_h-only model, but still smaller than in the WF model (Figure 4.7C). Among the three malaria models, the $t_\mathrm{m} = s_\mathrm{h}$ model has the greatest efficiency for positive selection because it has higher probability of fixation and shorter time to fixation for beneficial alleles. All three malaria models show patterns that are qualitatively different from the WF model.

Allele Frequency Spectrum

We simulated the null allele frequency spectrum when the sample size is 25 (matching the sample size in Chang et al. 2012). The result shows that the malaria life cycle skews the allele-frequency spectrum to the lower frequency alleles (Figure 4.8). This result implies that, even if the overall parasite population size has not changed recently, the allele-frequency spectrum resembles that of a WF population with an increasing population size. This effect is partly due to the population expansion within hosts in each generation, and it makes estimation of parasite demographic history more difficult than in other organisms. The results also imply that intrinsic differences in evolutionary processes caused by the complex malaria life cycle alter the null distributions of frequency-spectrum–based tests of selection. Hence it is important to consider the complexities of the malaria life cycle when analyzing genomic data to infer demographic history and to identify genes under selection.

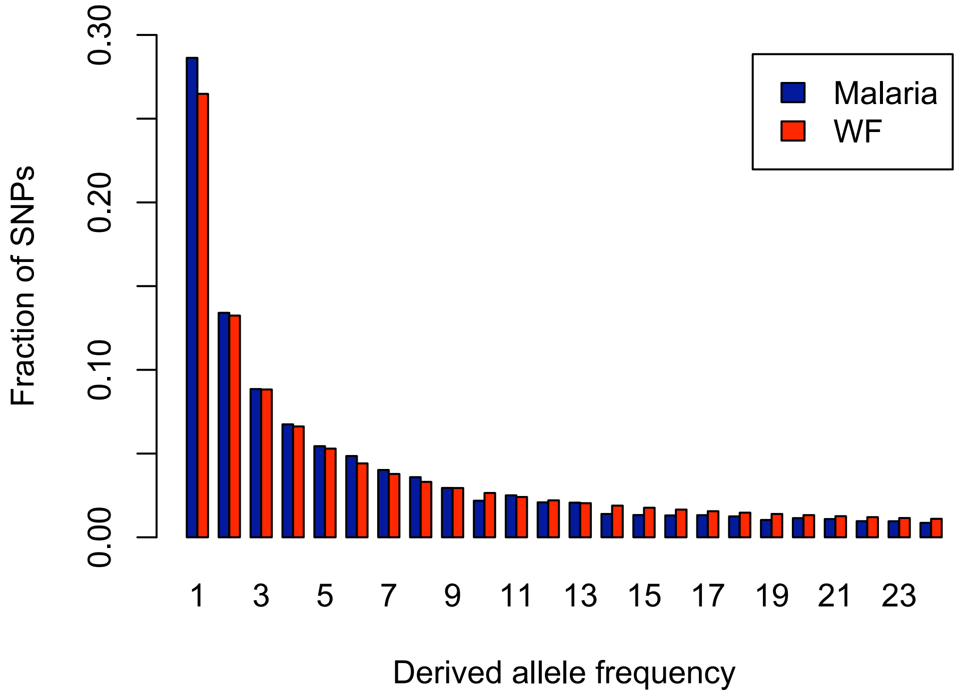


Figure 4.8. Allele-frequency spectrum of neutral alleles under the malaria life cycle compared with WF. The allele-frequency spectrum under the malaria life cycle is more skewed toward lower frequency alleles than in the WF model.

4.4 Discussion

Chang et al. (2012) found two unusual patterns of polymorphism in the *P. falciparum* genome. Here we investigated whether these patterns might reflect the complexities of the malaria life cycle. Specifically, we showed that mutations segregate in the population for a very short time (Figure 4.5A), which suggests that most of segregating mutations are either very new or very nearly neutral. We also showed that, for deleterious alleles, the probability of loss is greater (Figure 4.3B) and the time to loss shorter (Figure 4.5D) in the malaria life cycle than in the classical WF model. These results suggest that purifying selection works more efficiently in the malaria life cycle. Because purifying selection works so efficiently in the parasite life cycle, and the probability of loss is so high, most of the segregating mutations are either very new or very nearly neutral, and the expected difference between the synonymous and nonsynonymous allele-frequency spectra is reduced.

Moreover, because of the high efficiency of purifying selection in the parasite, sites with relatively small selection coefficients could nevertheless have their ultimate fate determined by selection whereas the same selection coefficients would be regarded as nearly neutral in the WF model. It is possible that polymorphisms at synonymous sites, which are commonly thought to be effectively neutral or under weak selection, experience more efficient selection in the malaria parasite, which would result in a higher than expected ratio of nonsynonymous to synonymous polymorphism (because of reduced polymorphism at synonymous sites).

Nonsynonymous polymorphism is less affected by enhanced efficiency of purifying selection due to malaria life cycle than synonymous polymorphism because selection coefficients of nonsynonymous mutations are in average more negative than synonymous mutations and more nonsynonymous mutations could be removed efficiently without the increase in selection efficiency. Indeed, selection for base composition at synonymous sites and in intergenic regions is quite effective, as shown by the significant differences between the C/G to A/T and the A/T to C/G site-frequency spectra in the genomes of malaria parasites from Senegal (Chang et al. 2012). In addition, not only *P. falciparum*, but another *Plasmodium* species with a similar life cycle, *P. vivax*, also has a large number of genes with π_N/π_S greater than 1 (Neafsey et al. 2012). Therefore, we concluded that malaria life cycle could at least partially explain the unusual patterns that were observed in the *P. falciparum* genome.

Random genetic drift and natural selection are two major forces that shape genetic variation. In standard population genetic models, when population size increases, the influence of random genetic drift decreases and that of natural selection increases, and conversely. Our main finding reported in this paper is that, in the malaria life cycle, both random genetic drift and natural selection are intensified simultaneously. Because of the unique parasite features of multiple asexual generations, population expansion within hosts, and stochastic transmission in each iteration of the life cycle, natural selection and random genetic drift can both increase at the same time. The lack of any tradeoff between random drift and selection contrasts with classical theoretical population genetics, and it demonstrates the importance of taking the parasite life cycle into account when interpreting genomic sequence data. Many microbial populations and parasite species have population growth and population bottlenecks during their life cycles and these species may be affected by similar dynamics. Our results also suggest that other parasite species that are transmitted among hosts with population-size expansion within hosts may evolve in a way that is qualitatively different from the Wright-Fisher model. Caution is therefore in order when interpreting data based on standard population genetic methods in organisms with unconventional life histories.

Whether the efficiency of positive selection is higher or lower with the malaria life cycle depends on the selection coefficient. When the selection coefficient is small, the time to fixation for beneficial alleles is shorter than in the WF model. When a mutation increases both the selection coefficient and the transmission probability, or when a mutation increases only the transmission probability, the time to fixation in the malaria life cycle is less than in the WF model across the whole range of simulated selection coefficients. However, it should be noted that the probability of fixation with malaria life cycle is less than in the WF model owing to the enhanced random genetic drift.

Comparing with WF models with multiple generations

It could be argued that the simulated malaria life cycle shows different allele-frequency patterns from the WF model simply because we treat multiple asexual generations in one life cycle as one "generation." We therefore compared the probability of loss, and the average number of mutations, after multiple WF generations with the model of the malaria life cycle (Figure 4.9). The comparison makes it clear that, even if we increase the generation numbers in the WF model, the malaria life cycle gives qualitatively different results from the WF model. The change in probability of loss as a function of change in selection coefficient is consistently greater in the malaria life cycle.

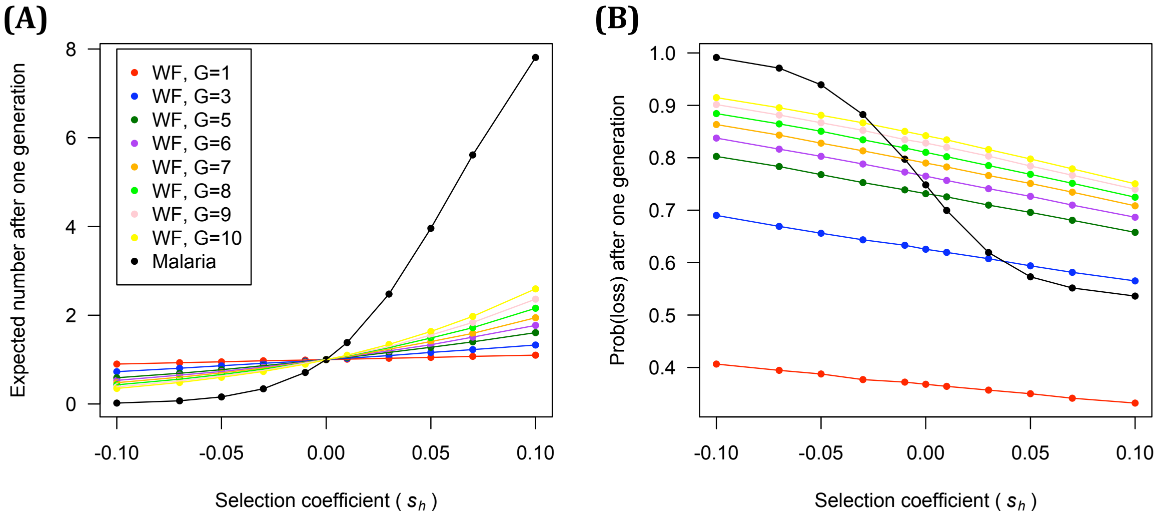


Figure 4.9. Comparison between the WF model and malaria models with differing numbers of generations (*G*). The WF model, even with multiple generations, is consistently different from the malaria model.

An effective s for malaria?

Effective population size of a real population is defined as the population size of a standard WF population that has the same level of random genetic drift as the population of interest. It is useful when the population under consideration deviates from the WF model but we want to make predictions or inferences from models or tools based on the WF model. Because the efficiency of selection is intensified under the malaria life cycle, we examined whether, analogous to effective population size, an effective selection coefficient could be defined. The effective selection coefficient corresponds to that in an ideal WF model that has the same population dynamics as in the malaria model. We therefore identified selection coefficients in the WF model that have the same probability of loss and average number of mutations as in the malaria (Table 4.2), however we found that it is not possible to fit these two properties at the same time. Hence an effective selection coefficient that holds for all aspects of the population dynamics does not exist.

Table 4.2. Fitted selection coefficients for expected number are different from fitted s for probability of loss.

Malaria *s*	Expected number	Fitted WF *s*	Prob of loss	Fitted WF *s*
-0.10	0.02045	-0.97955	0.99136	-0.99136
-0.09	0.03108	-0.96892	0.98697	-0.98697
-0.08	0.04730	-0.95270	0.98050	-0.98050
-0.07	0.07122	-0.92878	0.97105	-0.97105
-0.06	0.10666	-0.89334	0.95761	-0.95761
-0.05	0.15872	-0.84128	0.93926	-0.93800
-0.04	0.23408	-0.76592	0.91447	-0.91000
-0.03	0.34171	-0.65829	0.88244	-0.87500
-0.02	0.49482	-0.50518	0.84308	-0.83000
-0.01	0.71010	-0.28990	0.79722	-0.77000

Comparing with neutral condition

Besides comparing the absolute values of time to fixation or probabilities of fixation in the two models, we also examined the ratio of these quantities in the neutral case in the malaria model versus the WF model (Figure 4.10). Figure 4.10 shows that both the relative time to fixation of beneficial alleles and the relative time to loss of deleterious alleles are longer in malaria model. This indicates that equal transmission probability among hosts reduces the fold-difference between the neutral and selective cases. However, for deleterious mutations, the absolute time to loss is probably more relevant than the relative values because most of deleterious mutations are lost within the first host in which the mutation happens, and therefore neutral transmission among hosts does not play an important role. Random genetic drift (and the probability of loss) is so high that even neutral mutations are quickly lost, and the fold-difference between deleterious mutations and neutral mutations is smaller than in the WF model. Hence, this result also supports the conclusion of greater efficiency of purifying selection under the malaria life cycle.

In summary, this study used simulation to investigate the effect of malaria life cycle on population genetic behaviors. The results suggest that both genetic drift and efficiency of purifying selection are intensified by malaria life cycle, and because these two properties are typically not enhanced at the same time, this demonstrates the intrinsic differences between the WF model and the malaria life cycle model. Furthermore, allele-frequency spectrum under malaria life cycle is shown to be more skewed toward low frequency alleles even if the total number of hosts has not changed recently. Our study suggests that life cycle should be considered explicitly in order to study the evolution of malaria parasites or other organisms with similar life cycle through patterns of genetic diversity.

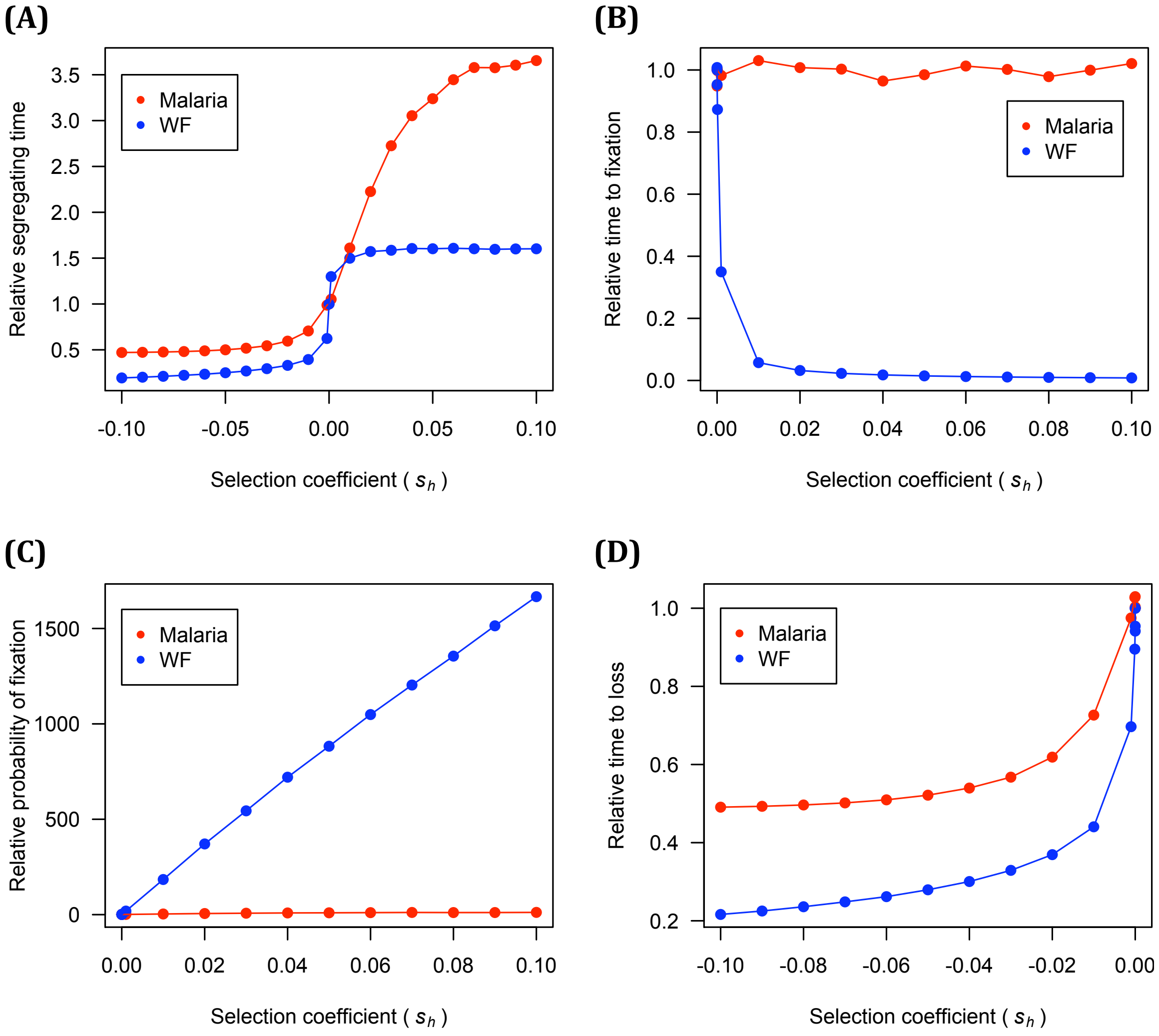


Figure 4.10. Ratio of quantities of studied case to that of the neutral case.

Supplementary Materials

Table S2.1. Genes with $\pi_N/\ \pi_S$ greater than 1

Gene ID	$\pi_N/\ \pi_S$	Gene ID	$\pi_N/\ \pi_S$	Gene ID	$\pi_N/\ \pi_S$	Gene ID	$\pi_N/\ \pi_S$	Gene ID	$\pi_N/\ \pi_S$
PFL0070c	14.39	PF14_0684	4.31	PF14_0740	3.19	PF10_0146	2.70	PFD0085c	2.42
PFB0695c	13.96	PFL2420w	4.28	PF13_0248	3.18	PF13_0107	2.69	PFD1200c	2.41
PFA0205w	12.20	PFB0340c	4.19	PFD0215c	3.16	PF14_0495	2.68	MAL13P1.121	2.40
PFL0035c	12.05	PFL2190c	4.00	PF11_0206	3.15	PFF0935c	2.68	PFF0460w	2.40
PFD0425w	11.87	PFB0755w	3.99	PF08_0034	3.14	PFF0370w	2.63	PFA0120c	2.39
PFI0550w	10.47	PFL1190c	3.95	PFB0485c	3.14	PF14_0216	2.61	PF10_0176	2.38
PFL0410w	9.40	PFA0110w	3.89	PFI1170c	3.12	PFB0835c	2.60	PF14_0238	2.37
PFB0520w	8.82	PF14_0706	3.85	PF13_0146	3.12	MAL13P1.107	2.60	PF10_0196	2.36
PFC0905c	8.65	PFA0310c	3.84	PFL0840c	3.10	PF14_0075	2.60	PF14_0389	2.35
PF13_0338	8.01	PFI0495w	3.82	PFF1440w	3.08	PF13_0064	2.59	PFL1455w	2.31
PFD0080c	7.00	PFL0440c	3.79	MAL13P1.88	3.06	MAL7P1.129	2.54	PFF1010c	2.31
PFF0595c	6.88	PF13_0079	3.78	PF07_0126	3.04	PFE0775c	2.54	PF11_0273-b	2.30
PF11_0381	6.33	PFL0290w	3.75	PF14_0315	3.02	PF14_0408	2.53	PF10_0064	2.30
PFE1325w	6.16	PFD0545w	3.59	MAL13P1.243	3.01	MAL8P1.32	2.53	PF11_0273-a	2.30
PF13_0078	5.94	PF14_0710	3.58	PF14_0511	2.99	PF14_0537	2.52	PF10_0231	2.29
MAL13P1.60	5.91	PF07_0010	3.54	PFF0820w	2.98	PF14_0455	2.52	PF10_0139	2.28
PFA0125c	5.55	PF14_0343	3.54	PFD0765w	2.98	PF11_0302	2.52	PFE0740c	2.27
PFI0970c	5.10	PFL1865w	3.45	PFI1085w	2.96	PFD0400w	2.52	MAL8P1.68	2.26
PFI1010w	5.10	PF13_0126	3.44	MAL13P1.212	2.94	PF08_0027	2.50	PF11_0189	2.26
PF10_0161	4.81	PF13_0221	3.43	PF14_0250	2.88	PFB0640c	2.50	PFL1205c	2.25
PFA0380w	4.75	PFF1280w	3.38	PF11_0131	2.85	PFL0920c	2.48	PF11_0045	2.24
PF07_0019	4.71	PF10_0186	3.33	PF14_0294	2.82	PF10_0214	2.47	PFL1720w	2.23
PF13_0104	4.70	PF07_0105	3.27	PF11_0347	2.79	PFF1255w	2.46	PF11_0328	2.23
PF13_0103	4.70	PFD0260c	3.26	PF10_0025	2.79	PFE1015c	2.46	PF14_0609	2.23
PFB0345c	4.69	PFL0555c	3.24	PFI0415c	2.78	PF08_0091	2.45	PFI0840w	2.23
PF11_0452	4.47	PF07_0030	3.23	PFL1810w	2.76	PFA0495c	2.43	PFB0380c	2.21
PFD0830w	4.38	PFL1750c	3.22	PF14_0404	2.76	PFB0335c	2.43	PF08_0089	2.19
PF10_0129	4.38	PF14_0557	3.22	PF10_0171	2.74	PFL1690w	2.42	PFL2365w	2.19
PF11_0351	4.32	PFI1485c	3.21	MAL13P1.203	2.74	PF14_0342	2.42	PFF0490w	2.18
PF13_0120	4.31	PFL0390c	3.21	PF13_0233	2.73	PFC0355c	2.42	PF14_0305	2.18

Table S2.1 (Continued)

Gene ID	π_N/ π_S	Gene ID	π_N/ π_S	Gene ID	π_N/ π_S	Gene ID	π_N/ π_S	Gene ID	π_N/ π_S
PFB0325c	2.17	PF14_0649	1.95	PFL0620c	1.80	PFF1460c	1.67	PFL2135c	1.55
MAL13P1.301	2.16	PF14_0615	1.95	PF14_0692	1.80	PF14_0657	1.66	PF11_0316	1.55
PFI0710c	2.15	PF10_0281	1.94	PF13_0067	1.80	PF13_0034	1.66	MAL7P1.37	1.55
PF08_0130	2.15	PFE0355c	1.94	PFI1475w	1.80	PF14_0198	1.65	PF11_0094	1.54
MAL8P1.24	2.14	PF11_0267	1.94	PF13_0020	1.80	PFI0335w	1.65	PF11_0388	1.54
PF10_0179	2.13	PFE0340c	1.94	MAL13P1.323	1.78	PF13_0076	1.64	PFL1395c	1.53
PF14_0478	2.11	PFE1240w	1.93	MAL13P1.316	1.77	PFL2430c	1.62	PFI1345c	1.53
PF14_0489	2.11	PFI1210w	1.93	PFE0715w	1.77	PF14_0291	1.62	PFE1515w	1.52
PF14_0598	2.10	MAL13P1.155	1.93	PF11_0354	1.77	PFD1140w	1.62	PFC0145c	1.52
PFF0795w	2.10	PF07_0111	1.93	MAL8P1.154	1.76	PF08_0078	1.62	PFL0610w	1.52
MAL13P1.151	2.09	PF11_0174	1.93	PF14_0179	1.75	PFF0325c	1.61	PF10_0260	1.52
PF13_0336	2.08	PF14_0298	1.93	PFD0835c	1.75	PF11_0373	1.61	PF10_0168-a	1.52
PFE0725c	2.08	PF07_0067	1.91	PF13_0044	1.74	MAL13P1.39	1.60	PFL2460w	1.52
MAL13P1.52	2.08	PFC0710w-a	1.90	PF14_0073	1.74	MAL13P1.308	1.60	PF14_0683	1.52
PFE0485w	2.07	PFE0085c	1.90	PF11_0191	1.73	PFA0330w	1.60	PF14_0347	1.52
PFC0415c	2.06	PF11_0374	1.90	PF10_0151	1.73	PFE0205w	1.59	PF11_0077	1.51
PF13_0239	2.05	PFI0975c	1.89	PF10_0052	1.72	MAL13P1.122	1.59	MAL13P1.322	1.51
PF13_0162	2.05	PF10_0164	1.88	PFL0755c	1.72	PF14_0707	1.59	PF14_0400	1.51
PF10_0286	2.03	PFF0770c	1.88	PFB0880w	1.72	PFC0210c	1.58	PFL2125c	1.51
PF13_0292	2.01	MAL13P1.114	1.87	PF11_0342	1.72	PFL1315w	1.58	PFE0305w	1.50
MAL13P1.32	2.01	PF11_0162	1.87	PF14_0591	1.71	PF14_0538	1.58	PFD1155w	1.50
PFF0280w	2.01	PF14_0533	1.86	PFE0127c	1.70	PF11_0535	1.57	PF11_0060	1.49
PFF0975c	2.01	MAL7P1.144	1.85	PFC0830w	1.70	PFE0360c	1.57	MAL13P1.240	1.49
PF14_0739a	1.99	PF14_0637	1.84	PF13_0295	1.70	PFF0870w	1.57	PF10_0323	1.49
PF10_0045	1.99	PF14_0509	1.82	PFI1045w	1.70	PF11_0054	1.56	MAL13P1.336	1.49
PFL0135w	1.97	PF14_0662	1.82	MAL13P1.18	1.69	PF14_0062	1.56	MAL7P1.113	1.48
PF07_0125	1.97	PF10_0145	1.82	PF14_0431	1.69	PFI1587c	1.56	PFF0800w	1.48
PF08_0013	1.97	PFL0815w	1.81	PF07_0022	1.69	PF10_0285	1.56	PFI1445w	1.48
PF11_0107	1.96	PF11_0203	1.81	PF13_0183	1.68	MAL8P1.60	1.55	PFL1085w	1.48
MAL13P1.313	1.95	MAL13P1.226	1.80	PFD0403w	1.68	PF10_0078	1.55	PFC0850c	1.47

Table S2.1 (Continued)

Gene ID	$\pi_N/\ \pi_S$	Gene ID	$\pi_N/\ \pi_S$	Gene ID	$\pi_N/\ \pi_S$	Gene ID	$\pi_N/\ \pi_S$	Gene ID	$\pi_N/\ \pi_S$
PF14_0116	1.47	PFL2050w	1.42	PFF1500c	1.33	PFI0915w	1.26	MAL13P1.116	1.21
PFE0520c	1.47	PFL1990c	1.40	PFL2120w	1.33	PFE0830c	1.26	PF10_0199	1.21
PF10_0313a	1.47	PF10_0221	1.40	PFL1710c	1.33	PFL2165w	1.26	PFI0510c	1.21
PF14_0077	1.47	PF14_0344	1.40	PFL0765w	1.32	PF10_0142	1.26	MAL13P1.201	1.20
PF11_0419	1.47	PFC0770c	1.40	PF13_0087	1.32	PF10_0284	1.26	PF14_0245	1.20
PF14_0034	1.46	PF08_0120	1.40	PF11_0408	1.32	PF14_0162	1.26	PF13_0314	1.20
PF13_0219	1.46	PFI0480w	1.40	PF14_0372	1.32	PF08_0058	1.26	PF14_0506	1.19
PFI1635w	1.46	MAL7P1.208	1.39	MAL13P1.140	1.31	PFL1670c	1.26	PF14_0664	1.19
MAL13P1.342	1.46	PFL1350w	1.39	PFB0650w	1.31	PFF0920c-a	1.26	PF11_0528	1.19
PF08_0046	1.45	PF14_0402	1.39	PF13_0197	1.31	PFF0965c	1.25	MAL7P1.22	1.19
PFL0745c	1.45	PFB0400w	1.39	PFL1790w	1.31	PFL1400c	1.25	PFI0250c	1.18
MAL8P1.203	1.45	PF13_0168	1.39	PFL2505c	1.31	PF08_0084	1.25	PFC0320w	1.18
PFE1270c	1.45	PF10_0363	1.38	PFL0685w	1.30	PFC0065c	1.25	PFL0530c	1.18
PFF1370w	1.45	PF13_0321	1.38	PFC1035w	1.30	PFD0872w	1.24	PFE1245w	1.18
PFI1180w	1.45	PF11_0168a	1.38	PF11_0056	1.29	PF14_0613	1.24	PFB0350c	1.18
PF10_0319	1.44	MAL8P1.80	1.37	PF10_0292	1.29	PFL0350c	1.23	PF14_0304	1.18
PF11_0371	1.44	PF10_0322	1.37	PF14_0546	1.29	PFD0940w	1.23	PF13_0137	1.18
PFL1785c	1.44	PFL2085w	1.37	PFE0930w	1.29	MAL7P1.146	1.23	MAL7P1.32	1.17
PF14_0647	1.44	PF11_0353	1.37	PFC0470w	1.29	PF14_0680	1.23	MAL13P1.395	1.17
PFE1185w	1.43	PF11_0364	1.37	PF11_0227	1.28	MAL13P1.262	1.23	PFB0920w	1.17
PF13_0220	1.43	MAL7P1.89	1.37	MAL8P1.42	1.28	PF14_0171	1.23	PFI0265c	1.16
PFL0310c	1.43	PF11_0527	1.36	MAL8P1.11	1.28	PF10_0057a	1.22	PF11_0175	1.16
PF07_0021	1.43	PFE0765w	1.35	MAL8P1.29	1.28	PF11_0279	1.22	PFF0683c	1.15
PF14_0722	1.43	PFL1245w	1.35	PFB0205c	1.28	PFL1335w	1.22	PFI1315c	1.15
PFL1320w	1.43	PF10_0234	1.34	MAL7P1.145	1.28	MAL7P1.6	1.22	PF13_0097	1.15
PF14_0416	1.43	PFL0510c	1.34	PF11_0092	1.27	PF13_0237	1.22	PFL0270c	1.15
PFL0110c	1.43	PF14_0644	1.34	PFD0470c	1.27	PFF0410w	1.22	MAL8P1.30	1.14
PFF0295c	1.43	PFB0405w	1.34	MAL7P1.112	1.26	PFC0130c	1.21	PF14_0654	1.14
PF13_0041	1.42	PF08_0050	1.33	PF10_0132	1.26	PF14_0708	1.21	PF14_0419	1.14
PF11_0349	1.42	PF14_0620	1.33	MAL8P1.45	1.26	PF14_0561	1.21	PF13_0196	1.14

Table S2.1 (Continued)

Gene ID	π_N/ π_S	Gene ID	π_N/ π_S	Gene ID	π_N/ π_S	Gene ID	π_N/ π_S	Gene ID	π_N/ π_S
PF11_0456	1.13	PF14_0435	1.10	MAL7P1.92	1.06	PFI0805w	1.04	MAL8P1.123	1.02
PFL1430c	1.13	PF13_0286	1.10	PF10_0095	1.06	PFI0485c	1.03	PF07_0037	1.01
PFC0940c	1.13	PF14_0394	1.10	PFL0895c	1.05	PF10_0185	1.03	PF14_0132	1.01
PFE1120w	1.13	PFF1100c	1.09	MAL13P1.234	1.05	PFL1045w	1.03	PF07_0024	1.01
PFB0300c	1.13	PF14_0558	1.09	PFF0720w	1.05	PF11_0158	1.03	PF10_0162	1.01
MAL13P1.390	1.13	PFL1915w	1.09	MAL13P1.195	1.05	PFA0150c	1.03	PF14_0031	1.01
PFC0590c	1.13	PF14_0512	1.09	PFB0375w	1.05	PF10_0333	1.03	PF10_0153a-a	1.01
MAL7P1.204	1.12	PFD1150c	1.08	PF14_0660	1.05	PFE0620c	1.03	PF10_0047	1.01
PF14_0720	1.12	PF10_0254	1.08	PFI0625c	1.05	MAL7P1.14	1.03	PFE0215w	1.01
MAL8P1.205	1.12	PF13_0052	1.08	PF10_0352	1.05	PFL0495c	1.03	PF10_0183	1.00
PFE1365w	1.12	PF10_0044	1.07	PF14_0632	1.05	MAL8P1.204	1.02	PF10_0065	1.00
MAL7P1.147	1.11	PF13_0027	1.07	PFE0090w	1.05	PFF0715c	1.02	PFA0170c	1.00
PFE1555c	1.11	PFD0285c	1.07	PF13_0125	1.05	PFD0345c	1.02	PFL1390w	1.00
PF11_0480	1.11	PFD0840w	1.07	PF10_0032	1.05	PF13_0214	1.02	PF10_0200	1.00
PFL1330c	1.11	PF07_0120	1.07	PF10_0182	1.05	PF14_0468	1.02	PFI0825w	1.00
PFA0415c	1.11	PFI0240c	1.07	PF14_0292	1.05	PFE0100w	1.02	PF14_0117	1.00
PF11_0392	1.11	PF10_0331	1.07	PF11_0226	1.05	MAL7P1.125	1.02	PF13_0165	1.00
PF08_0096	1.10	PFC0950c	1.07	PF14_0403	1.05	PFI1615w	1.02	PF14_0471	1.00
PFB0280w	1.10	PF10_0131	1.07	PF11_0311	1.05	PF14_0703	1.02	MAL13P1.150	1.00
PF14_0119	1.10	PF10_0371	1.07	MAL13P1.385	1.04	PF13_0143	1.02	PF13_0173	1.00
PF10_0177a	1.10	PF14_0059	1.07	PFL1715w	1.04	PFE0245c	1.02		
PF10_0225	1.10	PF10_0291	1.07	PF14_0031b	1.04	PF11_0333	1.02		
MAL7P1.138	1.10	PF14_0738	1.06	PFL1155w	1.04	PF11_0177	1.02		

Table S2.2. Significant regions of the iHS test

Chr.	start	end	gene ID	Tajima's *D*	*P* value	Chr.	start	end	gene ID	Tajima's *D*	*P* value
2	554110	646016	PFB0615c	-1.5946	0.401	4	474738	697839	PFD0620c	NA	NA
2	554110	646016	PFB0620w	NA	NA	4	474738	697839	PFD0625c	NA	NA
2	554110	646016	PFB0625w	-1.0186	0.270	4	474738	697839	PFD0630c	NA	NA
2	554110	646016	PFB0630c	-2.5098	0.002	4	474738	697839	PFD0635c	NA	NA
2	554110	646016	PFB0635w	-0.6983	0.137	4	474738	697839	PFD0640c	0.1005	0.070
2	554110	646016	PFB0640c	-1.2969	0.214	4	474738	697839	PFD0645w	NA	NA
2	554110	646016	PFB0645c	NA	NA	4	474738	697839	PFD0655w	NA	NA
2	554110	646016	PFB0650w	-1.8693	0.734	4	474738	697839	PFD0660w	-0.2052	0.142
2	554110	646016	PFB0655c	-1.5141	0.662	4	474738	697839	PFD0665c	-2.1061	0.111
2	554110	646016	PFB0660w	-1.1575	0.692	4	474738	697839	PFD0670c	NA	NA
2	554110	646016	PFB0665w	-1.6859	0.537	4	474738	697839	PFD0675w	-0.1761	0.153
2	554110	646016	PFB0670c	-1.5141	0.914	4	474738	697839	PFD0680c	-1.0665	0.209
2	554110	646016	PFB0675w	0.0470	0.005	4	474738	697839	PFD0685c	-0.6633	0.037
2	554110	646016	PFB0680w	-1.8859	0.867	4	474738	697839	PFD0690c	-1.1691	0.177
2	554110	646016	PFB0685c	-1.1119	0.210	4	474738	697839	PFD0692c	1.1008	0.019
2	554110	646016	PFB0687c	NA	NA	4	474738	697839	PFD0669c	-1.2137	0.624
2	554110	646016	PFB0690w	NA	NA	4	474738	697839	PFD0695w	-1.3417	0.787
2	554110	646016	PFB0695c	0.7525	0.001	4	474738	697839	PFD0700c	0.2673	0.030
2	554110	646016	PFB0700c	-1.9994	0.697	4	474738	697839	PFD0705c	2.2053	0.000
2	554110	646016	PFB0705w	-1.9495	0.700	4	474738	697839	PFD0710w	-1.8859	0.753
2	554110	646016	PFB0710c	-1.5141	0.988	4	474738	697839	PFD0715c	-1.1575	0.892
2	554110	646016	PFB0715w	-1.7333	0.668	4	474738	697839	PFD0720w	-1.2001	0.764
2	554110	646016	PFD0505c	-0.8316	0.103	4	474738	697839	PFD0725c	-1.2748	0.622
4	474738	697839	PFD0515w	-1.5141	0.917	4	474738	697839	PFD0730w	-1.2137	0.658
4	474738	697839	PFD0520c	-1.1081	0.623	4	474738	697839	PFD0735c	-1.5355	0.328
4	474738	697839	PFD0525w	-1.1575	0.501	4	474738	697839	PFD0740w	-1.6411	0.577
4	474738	697839	PFD0530c	-1.8859	0.917	4	474738	697839	PFD0745c	NA	NA
4	474738	697839	PFD0535w	-0.9707	0.113	4	474738	697839	PFD0750w	NA	NA
4	474738	697839	PFD0540c	-1.9994	0.676	4	474738	697839	PFD0755c	-1.1575	0.850
4	474738	697839	PFD0545w	-1.9417	0.970	4	474738	697839	PFD0795w	-1.2593	0.202
4	474738	697839	PFD0550c	-1.1575	0.901	4	474738	697839	PFD0800c	NA	NA
4	474738	697839	PFD0555c	-1.6486	0.692	4	474738	697839	PFD0805w	-0.2808	0.193
4	474738	697839	PFD0560w	-2.1581	0.299	4	474738	697839	PFD0807c	NA	NA
4	474738	697839	PFD0565c	-1.7422	0.776	4	474738	697839	PF04TR001	NA	NA
4	474738	697839	PFD0580c	NA	NA	4	474738	697839	PFD0810w	NA	NA
4	474738	697839	PFD0585c	-2.2097	0.460	4	474738	697839	PFD0815c	-1.6859	0.496
4	474738	697839	PFD0590c	-0.6910	0.027	4	474738	697839	PFD0820w	NA	NA
4	474738	697839	PFD0595w	1.2467	0.001	4	474738	697839	PFD0825c	-1.5141	0.791
4	474738	697839	PFD0600c	-1.8859	0.410	4	474738	697839	PFD0830w	0.7839	0.002
4	474738	697839	PFD0605c	1.4302	0.007	4	474738	697839	PFD0835c	-1.1338	0.287
4	474738	697839	PFD0610w	0.1119	0.058	4	474738	697839	PF04TR002	NA	NA
4	474738	697839	PFD0615c	NA	NA	4	474738	697839	PFD0840w	-1.4503	0.164

Table S2.2 (Continued)

Chr.	start	end	gene ID	Tajima's *D*	*P* value	Chr.	start	end	gene ID	Tajima's *D*	*P* value
4	474738	697839	PFE0805w	-0.7053	0.025	5	673558	878001	PFE1005w	NA	NA
4	727478	773803	PFE0810c	NA	NA	5	673558	878001	PFE1010w	-1.7333	0.962
4	727478	773803	PFE0815w	0.4816	0.006	5	673558	878001	PFE1015c	-1.9922	0.642
4	727478	773803	PFE0820c	-0.6403	0.243	5	673558	878001	PFE1020w	NA	NA
4	727478	773803	PFE0825w	-1.1575	0.414	5	673558	878001	PFE1025c	-1.1575	0.684
4	727478	773803	PFE0830c	-2.0729	0.736	5	673558	878001	PFE1030c	NA	NA
4	727478	773803	PFE0835w	-1.7333	0.688	5	673558	878001	PFE1035c	-0.8789	0.546
4	727478	773803	PFE0840c	-2.2407	0.454	5	673558	878001	PFE1040c	-1.0710	0.215
4	727478	773803	PF05TR001	NA	NA	5	673558	878001	PFE1045c	-1.2179	0.142
4	727478	773803	PFE0845c	-1.1575	0.708	5	673558	878001	PFE1050w	-0.6983	0.159
4	727478	773803	PFE0850c	NA	NA	5	673558	878001	PFE1055c	-1.5141	0.751
4	727478	773803	PFE0855c	-1.8859	0.332	5	673558	878001	PFE1060c	-1.8923	0.962
4	727478	773803	PFE0860c	-1.8859	0.779	5	673558	878001	PFE1065w	-0.6983	0.346
4	727478	773803	PFE0865c	NA	NA	5	673558	878001	PFE1070c	-2.0875	0.621
4	727478	773803	PF05TR002	NA	NA	5	673558	878001	PFE1075c	-1.8859	0.688
5	673558	878001	PFE0870w	-1.5141	0.471	5	673558	878001	PFE1120w	-1.7421	0.344
5	673558	878001	PFE0875c	-0.6949	0.079	5	673558	878001	PFE1125w	-1.1575	1.000
5	673558	878001	PFE0880c	-1.7333	0.385	5	673558	878001	PFE1130w	-1.1575	0.407
5	673558	878001	PFE0885w	-1.5041	0.581	5	673558	878001	PFE1135w	NA	NA
5	673558	878001	PFE0890c	-0.9134	0.466	5	673558	878001	PFE1140c	NA	NA
5	673558	878001	PFE0895c	NA	NA	5	673558	878001	PFE1145w	-1.9472	0.998
5	673558	878001	PFE0900w	-1.8859	0.287	5	673558	878001	PF05TR004	NA	NA
5	673558	878001	PFE0905w	-1.8742	0.951	5	673558	878001	PFE1150w	-0.8487	0.058
5	673558	878001	PFE0910w	NA	NA	5	673558	878001	PFE1155c	-1.1575	0.377
5	673558	878001	PFE0915c	NA	NA	5	673558	878001	PFE1160w	-1.7805	0.908
5	673558	878001	PFE0920c	NA	NA	5	673558	878001	PFF0525w	-1.5141	0.083
5	673558	878001	PFE0925c	-2.0875	0.603	5	673558	878001	PFF0530w	-1.5041	0.590
5	673558	878001	PFE0930w	-2.0875	0.638	5	673558	878001	PFF0535c	-0.6949	0.046
5	673558	878001	PFE0935c	-2.0062	0.849	5	673558	878001	PFF0540c	-1.1575	0.711
5	673558	878001	PFE0940c	-1.5141	0.937	5	673558	878001	PFF0545c	NA	NA
5	673558	878001	PFE0950c	NA	NA	5	673558	878001	PFF0550w	-1.1575	0.763
5	673558	878001	PFE0955w	-1.1575	0.657	5	673558	878001	PFF0555w	-1.1575	0.586
5	673558	878001	PFE0960w	NA	NA	5	922308	966309	PFF0560c	-1.5141	0.552
5	673558	878001	PFE0965c	NA	NA	5	922308	966309	PFF0565c	-0.9414	0.354
5	673558	878001	PFE0970w	-0.5075	0.069	5	922308	966309	PFF0570c	-1.8407	0.836
5	673558	878001	PFE0975c	0.3154	0.066	5	922308	966309	PFF0573c	NA	NA
5	673558	878001	PFE0980c	-1.9655	0.918	5	922308	966309	PFF0575c	-1.9409	0.833
5	673558	878001	PFE0985w	-1.6106	0.752	5	922308	966309	PFF0580w	NA	NA
5	673558	878001	PFE0990w	-1.1575	0.599	5	922308	966309	PFF0585c	-1.1575	0.364
5	673558	878001	PFE0995c	-1.5141	0.605	5	922308	966309	PFF0590c	-1.6821	0.906
5	673558	878001	PF05TR003	NA	NA	5	922308	966309	PFF0595c	-1.3304	0.195
5	673558	878001	PFE1000c	-0.5765	0.143	5	922308	966309	PFF0600w	-1.5141	0.491

Table S2.2 (Continued)

Chr.	start	end	gene ID	Tajima's *D*	*P* value	Chr.	start	end	gene ID	Tajima's *D*	*P* value
5	922308	966309	PFF0602w	NA	NA	7	414963	533798	PF08_0083	-1.7333	0.877
5	922308	966309	PFF0605c	NA	NA	7	414963	533798	PF08_0082	-1.1575	0.334
5	922308	966309	PFF0610c	-1.7333	0.780	7	414963	533798	MAL8P1.96	-1.5141	0.938
6	451552	542102	PFF0615c	-0.9407	0.345	7	414963	533798	MAL8P1.95	-1.1575	0.637
6	451552	542102	PFF0620c	0.5338	0.026	7	414963	533798	PF08_0081	-0.6949	0.131
6	451552	542102	PFF0625w	-0.1975	0.030	7	414963	533798	PF08_0080	-0.3554	0.151
6	451552	542102	PFF0630c-a	NA	NA	7	414963	533798	PF08_0079	-1.5118	0.891
6	451552	542102	PFF0630c-b	NA	NA	7	414963	533798	PF08_0078	-0.7671	0.059
6	451552	542102	PFF0630c-c	NA	NA	7	414963	533798	PF08_0077	-0.6705	0.223
6	451552	542102	PFF0635w	NA	NA	7	414963	533798	PF08_0076	-0.6949	0.596
6	451552	542102	PFF0640w	NA	NA	7	414963	533798	PFL1645w	-2.2124	0.584
6	451552	542102	PFF0645c	-1.9646	0.936	7	414963	533798	PFL1650w	-1.8591	0.772
6	451552	542102	PF07_0026	-1.5141	0.538	7	414963	533798	PFL1655c	0.7212	0.002
6	451552	542102	MAL7P1.23	-1.1575	0.223	7	414963	533798	PFL1660c	-1.5141	0.727
6	451552	542102	MAL7P1.24	-1.1575	0.717	7	414963	533798	PFL1665c	NA	NA
6	451552	542102	MAL7P1.25	-2.0358	0.634	7	414963	533798	PFL1670c	-2.0875	0.563
6	451552	542102	PF07_0027	NA	NA	7	414963	533798	PFL1675c	-1.5747	0.496
6	451552	542102	PF07_0028	-1.0683	0.202	7	414963	533798	PFL1680w	-2.0875	0.679
6	451552	542102	MAL7P1.26	-1.1575	0.849	7	414963	533798	PFL1685w	-1.1575	0.923
6	451552	542102	PF07_0029	0.5391	0.002	7	414963	533798	PFL1690w	-1.9163	0.572
6	451552	542102	PF07_0030	-1.4367	0.457	7	414963	533798	PFL1695c	-0.6983	0.624
6	451552	542102	PF07_0031	NA	NA	7	414963	533798	PFL1700c	-1.6988	0.677
6	451552	542102	PF07_0032	-1.0664	0.544	7	414963	533798	PFL1705w	-2.3144	0.246
6	451552	542102	PF07_0033	-0.9336	0.132	8	642093	682038	PFL1710c	-1.8407	0.944
6	451552	542102	PF07_0034	-1.5141	0.976	8	642093	682038	PFL1715w	-1.2748	0.411
6	451552	542102	MAL7P1.27	1.3621	0.002	8	642093	682038	PFL1720w	-0.4204	0.100
6	451552	542102	PF07_0035	0.5766	0.001	8	642093	682038	PFL1725w	-1.5141	0.731
6	451552	542102	PF07_0036	0.1587	0.071	8	642093	682038	PFL1730c	-2.3409	0.231
6	451552	542102	PF07_0037	0.1024	0.001	8	642093	682038	PFL1735c	-1.5141	0.576
6	451552	542102	PF07_0038	-0.9571	0.105	8	642093	682038	PFL1740w	-1.9994	0.270
6	451552	542102	PF07_0039	-1.1575	0.706	8	642093	682038	PFL1745c	NA	NA
6	451552	542102	MAL7P1.28	-1.1343	0.161	8	642093	682038	PFL1750c	-1.9605	0.939
6	451552	542102	PF07_0040	-0.2537	0.096	8	642093	682038	PFL1755w	-1.5118	0.810
6	451552	542102	MAL7P1.29a	-0.6983	0.985	8	642093	682038	PFL1760w	-1.5141	0.809
6	451552	542102	MAL7P1.29	-2.0169	0.943	8	642093	682038	PFL1765c	-1.1575	0.699
6	451552	542102	MAL7P1.30	-0.0020	0.001	8	642093	682038	PFL1770c	-1.1575	0.846
7	414963	533798	PF07_0041	0.2362	0.029	12	1E+06	1544106	PFL1775c	-1.7333	0.877
7	414963	533798	PF07_0042	0.5110	0.000	12	1E+06	1544106	PFL1780w	-1.1575	0.450
7	414963	533798	MAL7P1.31	-1.1575	0.924	12	1E+06	1544106	PFL1785c	-0.9414	0.373
7	414963	533798	MAL7P1.32	-0.2677	0.140	12	1E+06	1544106	PFL1790w	-0.9414	0.186
7	414963	533798	MAL7P1.33	-0.6004	0.314	12	1E+06	1544106	PFL1795c	-1.6520	0.462
7	414963	533798	MAL8P1.97	-2.4224	0.095						

Bibliography

Amambua-Ngwa, A., K. K. Tetteh, M. Manske, N. Gomez-Escobar, L. B. Stewart, M. E. Deerhake, I. H. Cheeseman, C. I. Newbold, A. A. Holder, E. Knuepfer, O. Janha, M. Jallow, S. Campino, B. Macinnis, D. P. Kwiatkowski, and D. J. Conway. 2012. Population genomic scan for candidate signatures of balancing selection to guide antigen characterization in malaria parasites. PLoS Genet **8**:e1002992.

Anderson, E. C. 2005. An efficient Monte Carlo method for estimating Ne from temporally spaced samples using a coalescent-based likelihood. Genetics **170**:955-967.

Anderson, T. J., B. Haubold, J. T. Williams, J. G. Estrada-Franco, L. Richardson, R. Mollinedo, M. Bockarie, J. Mokili, S. Mharakurwa, N. French, J. Whitworth, I. D. Velez, A. H. Brockman, F. Nosten, M. U. Ferreira, and K. P. Day. 2000. Microsatellite markers reveal a spectrum of population structures in the malaria parasite *Plasmodium falciparum*. Mol Biol Evol **17**:1467-1482.

Antao, T., and I. M. Hastings. 2011. ogaraK: a population genetics simulator for malaria. Bioinformatics **27**:1335-1336.

Baron, J. M., J. M. Higgins, and W. H. Dzik. 2011. A Revised Timeline for the Origin of *Plasmodium falciparum* as a Human Pathogen. J Mol Evol.

Bethke, L. L., M. Zilversmit, K. Nielsen, J. Daily, S. K. Volkman, D. Ndiaye, E. R. Lozovsky, D. L. Hartl, and D. F. Wirth. 2006. Duplication, gene conversion, and genetic diversity in the species-specific acyl-CoA synthetase gene family of *Plasmodium falciparum*. Mol Biochem Parasitol **150**:10-24.

Branch, O. H., P. L. Sutton, C. Barnes, J. C. Castro, J. Hussin, P. Awadalla, and G. Hijar. 2011. *Plasmodium falciparum* genetic diversity maintained and amplified over 5 years of a low transmission endemic in the Peruvian Amazon. Mol Biol Evol **28**:1973-1986.

Brown, K. M., M. S. Costanzo, W. Xu, S. Roy, E. R. Lozovsky, and D. L. Hartl. 2010. Compensatory mutations restore fitness during the evolution of dihydrofolate reductase. Mol Biol Evol **27**:2682-2690.

Carlton, J. 2003. The *Plasmodium vivax* genome sequencing project. Trends Parasitol **19**:227-231.

Carneiro, M., F. W. Albert, J. Melo-Ferreira, N. Galtier, P. Gayral, J. A. Blanco-Aguiar, R. Villafuerte, M. W. Nachman, and N. Ferrand. 2012. Evidence for widespread positive and purifying selection across the European rabbit (*Oryctolagus cuniculus*) genome. Mol Biol Evol **29**:1837-1849.

Chanda, I., A. Pan, and C. Dutta. 2005. Proteome composition in Plasmodium falciparum: higher usage of GC-rich nonsynonymous codons in highly expressed genes. J Mol Evol **61**:513-523.

Chang, H. H., D. J. Park, K. J. Galinsky, S. F. Schaffner, D. Ndiaye, O. Ndir, S. Mboup, R. C. Wiegand, S. K. Volkman, P. C. Sabeti, D. F. Wirth, D. E. Neafsey, and D. L. Hartl. 2012. Genomic sequencing of *Plasmodium falciparum* malaria parasites from senegal reveals the demographic history of the population. Mol Biol Evol **29**:3427-3439.

Chen, F. C., and W. H. Li. 2001. Genomic divergences between humans and other hominoids and the effective population size of the common ancestor of humans and chimpanzees. Am J Hum Genet **68**:444-456.

Choi, S. W., M. K. Keyes, and P. Horrocks. 2006. LC/ESI-MS demonstrates the absence of 5-methyl-2'-deoxycytosine in *Plasmodium falciparum* genomic DNA. Mol Biochem Parasitol **150**:350-352.

Conway, D. J., C. Fanello, J. M. Lloyd, B. M. Al-Joubori, A. H. Baloch, S. D. Somanath, C. Roper, A. M. Oduola, B. Mulder, M. M. Povoa, B. Singh, and A. W. Thomas. 2000. Origin of *Plasmodium falciparum* malaria is traced by mitochondrial DNA. Mol Biochem Parasitol **111**:163-171.

Costanzo, M. S., K. M. Brown, and D. L. Hartl. 2011. Fitness trade-offs in the evolution of dihydrofolate reductase and drug resistance in *Plasmodium falciparum*. PLoS One **6**:e19636.

Dalpke, A., J. Frank, M. Peter, and K. Heeg. 2006. Activation of toll-like receptor 9 by DNA from different bacterial species. Infect Immun **74**:940-946.

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

DePristo, M. A., E. Banks, R. Poplin, K. V. Garimella, J. R. Maguire, C. Hartl, A. A. Philippakis, G. del Angel, M. A. Rivas, M. Hanna, A. McKenna, T. J. Fennell, A. M. Kernytsky, A. Y. Sivachenko, K. Cibulskis, S. B. Gabriel, D. Altshuler, and M. J. Daly. 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet **43**:491-498.

Der, R., C. Epstein, and J. B. Plotkin. 2012. Dynamics of neutral and selected alleles when the offspring distribution is skewed. Genetics **191**:1331-1344.

Echeverry, D. F., S. Nair, L. Osorio, S. Menon, C. Murillo, and T. J. Anderson. 2013. Long term persistence of clonal malaria parasite *Plasmodium falciparum* lineages in the Colombian Pacific region. BMC Genet **14**:2.

Escalante, A. A., A. A. Lal, and F. J. Ayala. 1998. Genetic polymorphism and natural selection in the malaria parasite *Plasmodium falciparum*. Genetics **149**:189-202.

Falush, D., M. Stephens, and J. K. Pritchard. 2003. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. Genetics **164**:1567-1587.

Fisher, R. A. 1930. The genetical theory of natural selection. Clarendon Press, Oxford.

Fujimoto, A., H. Nakagawa, N. Hosono, K. Nakano, T. Abe, K. A. Boroevich, M. Nagasaki, R. Yamaguchi, T. Shibuya, M. Kubo, S. Miyano, Y. Nakamura, and T. Tsunoda. 2010. Whole-genome sequencing and comprehensive variant analysis of a Japanese individual using massively parallel sequencing. Nat Genet **42**:931-936.

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

Griffing, S. M., T. Mixson-Hayden, S. Sridaran, M. T. Alam, A. M. McCollum, C. Cabezas, W. Marquino Quezada, J. W. Barnwell, A. M. De Oliveira, C. Lucas, N. Arrospide, A. A. Escalante, D. J. Bacon, and V. Udhayakumar. 2011. South American P*lasmodium falciparum* after the malaria eradication era: clonal population expansion and survival of the fittest hybrids. PLoS One **6**:e23486.

Gutenkunst, R. N., R. D. Hernandez, S. H. Williamson, and C. D. Bustamante. 2009. Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. PLoS Genet **5**:e1000695.

Hartl, D. L. 2004. The origin of malaria: mixed messages from genetic diversity. Nat Rev Microbiol **2**:15-22.

Hartl, D. L., S. K. Volkman, K. M. Nielsen, A. E. Barry, K. P. Day, D. F. Wirth, and E. A. Winzeler. 2002. The paradoxical population genetics of *Plasmodium falciparum*. Trends Parasitol **18**:266-272.

Hastings, I. M. 2006. Complex dynamics and stability of resistance to antimalarial drugs. Parasitology **132**:615-624.

Hastings, I. M. 1997. A model for the origins and spread of drug-resistant malaria. Parasitology **115 (Pt 2)**:133-141.

Haubold, B., and R. R. Hudson. 2000. LIAN 3.0: detecting linkage disequilibrium in multilocus data. Linkage Analysis. Bioinformatics **16**:847-848.

Hershberg, R., and D. A. Petrov. 2010. Evidence that mutation is universally biased towards AT in bacteria. PLoS Genet **6**.

Hill, W. G., and A. Robertson. 1968. Linkage disequilibrium in finite populations Theor. Appl. Genet **38**:226-231.

Hudson, R. R. 2002. Generating samples under a Wright-Fisher neutral model of genetic variation. Bioinformatics **18**:337-338.

Hughes, A. L., and F. Verra. 2001. Very large long-term effective population size in the virulent human malaria parasite *Plasmodium falciparum*. Proc Biol Sci **268**:1855-1860.

Ina, Y. 1995. New methods for estimating the numbers of synonymous and nonsynonymous substitutions. J Mol Evol **40**:190-226.

Jeffares, D. C., A. Pain, A. Berry, A. V. Cox, J. Stalker, C. E. Ingle, A. Thomas, M. A. Quail, K. Siebenthall, A. C. Uhlemann, S. Kyes, S. Krishna, C. Newbold, E. T. Dermitzakis, and M. Berriman. 2007. Genome variation and evolution of the malaria parasite *Plasmodium falciparum*. Nat Genet **39**:120-125.

Jiang, H., N. Li, V. Gopalan, M. M. Zilversmit, S. Varma, V. Nagarajan, J. Li, J. Mu, K. Hayton, B. Henschen, M. Yi, R. Stephens, G. McVean, P. Awadalla, T. E. Wellems, and X. Z. Su. 2011. High recombination rates and hotspots in a *Plasmodium falciparum* genetic cross. Genome Biol **12**:R33.

Joy, D. A., X. Feng, J. Mu, T. Furuya, K. Chotivanich, A. U. Krettli, M. Ho, A. Wang, N. J. White, E. Suh, P. Beerli, and X. Z. Su. 2003. Early origin and recent expansion of *Plasmodium falciparum*. Science **300**:318-321.

Kulathinal, R. J., S. M. Bennett, C. L. Fitzpatrick, and M. A. Noor. 2008. Fine-scale mapping of recombination rate in *Drosophila* refines its correlation to diversity and divergence. Proc Natl Acad Sci U S A **105**:10051-10056.

Kurtz, S., A. Phillippy, A. L. Delcher, M. Smoot, M. Shumway, C. Antonescu, and S. L. Salzberg. 2004. Versatile and open software for comparing large genomes. Genome Biol **5**:R12.

Langley, C. H., K. Stevens, C. Cardeno, Y. C. Lee, D. R. Schrider, J. E. Pool, S. A. Langley, C. Suarez, R. B. Corbett-Detig, B. Kolaczkowski, S. Fang, P. M. Nista, A. K. Holloway, A. D. Kern, C. N. Dewey, Y. S. Song, M. W. Hahn, and D. J. Begun. 2012. Genomic variation in natural populations of *Drosophila melanogaster*. Genetics **192**:533-598.

Lee, Y. C., and J. A. Reinhardt. 2012. Widespread polymorphism in the positions of stop codons in *Drosophila melanogaster*. Genome Biol Evol **4**:533-549.

Li, H., and R. Durbin. 2010. Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics **26**:589-595.

Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, and R. Durbin. 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics **25**:2078-2079.

Li, Y., N. Vinckenbosch, G. Tian, E. Huerta-Sanchez, T. Jiang, H. Jiang, A. Albrechtsen, G. Andersen, H. Cao, T. Korneliussen, N. Grarup, Y. Guo, I. Hellman, X. Jin, Q. Li, J. Liu, X. Liu, T. Sparso, M. Tang, H. Wu, R. Wu, C. Yu, H. Zheng, A. Astrup, L. Bolund, J. Holmkvist, T. Jorgensen, K. Kristiansen, O. Schmitz, T. W. Schwartz, X. Zhang, R. Li, H. Yang, J. Wang, T. Hansen, O. Pedersen, and R. Nielsen. 2010. Resequencing of 200 human exomes identifies an excess of low-frequency non-synonymous coding variants. Nat Genet **42**:969-972.

Liu, W., Y. Li, G. H. Learn, R. S. Rudicell, J. D. Robertson, B. F. Keele, J. B. Ndjango, C. M. Sanz, D. B. Morgan, S. Locatelli, M. K. Gonder, P. J. Kranzusch, P. D. Walsh, E. Delaporte, E. Mpoudi-Ngole, A. V. Georgiev, M. N. Muller, G. M. Shaw, M. Peeters, P. M. Sharp, J. C. Rayner, and B. H. Hahn. 2010. Origin of the human malaria parasite *Plasmodium falciparum* in gorillas. Nature **467**:420-425.

Lozovsky, E. R., T. Chookajorn, K. M. Brown, M. Imwong, P. J. Shaw, S. Kamchonwongpaisan, D. E. Neafsey, D. M. Weinreich, and D. L. Hartl. 2009. Stepwise acquisition of pyrimethamine resistance in the malaria parasite. Proc Natl Acad Sci U S A **106**:12025-12030.

Malaria RB. 2010. Focus on Senegal—progress and impact series. Geneva: World Health Organization.

Manske, M., O. Miotto, S. Campino, S. Auburn, J. Almagro-Garcia, G. Maslen, J. O'Brien, A. Djimde, O. Doumbo, I. Zongo, J. B. Ouedraogo, P. Michon, I. Mueller, P. Siba, A. Nzila, S. Borrmann, S. M. Kiara, K. Marsh, H. Jiang, X. Z. Su, C. Amaratunga, R. Fairhurst, D. Socheat, F. Nosten, M. Imwong, N. J. White, M. Sanders, E. Anastasi, D. Alcock, E. Drury, S. Oyola, M. A. Quail, D. J. Turner, V. Ruano-Rubio, D. Jyothi, L. Amenga-Etego, C. Hubbart, A. Jeffreys, K. Rowlands, C. Sutherland, C. Roper, V. Mangano, D. Modiano, J. C. Tan, M. T. Ferdig, A. Amambua-Ngwa, D. J. Conway, S. Takala-Harrison, C. V. Plowe, J. C. Rayner, K.

A. Rockett, T. G. Clark, C. I. Newbold, M. Berriman, B. MacInnis, and D. P. Kwiatkowski. 2012. Analysis of *Plasmodium falciparum* diversity in natural infections by deep sequencing. Nature **487**:375-379.

McVean, G., P. Awadalla, and P. Fearnhead. 2002. A coalescent-based method for detecting and estimating recombination from gene sequences. Genetics **160**:1231-1241.

McVean, G. A., S. R. Myers, S. Hunt, P. Deloukas, D. R. Bentley, and P. Donnelly. 2004. The fine-scale structure of recombination rate variation in the human genome. Science **304**:581-584.

Mobegi, V. A., K. M. Loua, A. D. Ahouidi, J. Satoguina, D. C. Nwakanma, A. Amambua-Ngwa, and D. J. Conway. 2012. Population genetic structure of *Plasmodium falciparum* across a region of diverse endemicity in West Africa. Malar J **11**:223.

Mu, J., P. Awadalla, J. Duan, K. M. McGee, D. A. Joy, G. A. McVean, and X. Z. Su. 2005. Recombination hotspots and population structure in *Plasmodium falciparum*. PLoS Biol **3**:e335.

Mu, J., P. Awadalla, J. Duan, K. M. McGee, J. Keebler, K. Seydel, G. A. McVean, and X. Z. Su. 2007. Genome-wide variation and identification of vaccine targets in the *Plasmodium falciparum* genome. Nat Genet **39**:126-130.

Mu, J., R. A. Myers, H. Jiang, S. Liu, S. Ricklefs, M. Waisberg, K. Chotivanich, P. Wilairatana, S. Krudsood, N. J. White, R. Udomsangpetch, L. Cui, M. Ho, F. Ou, H. Li, J. Song, G. Li, X. Wang, S. Seila, S. Sokunthea, D. Socheat, D. E. Sturdevant, S. F. Porcella, R. M. Fairhurst, T. E. Wellems, P. Awadalla, and X. Z. Su. 2010. *Plasmodium falciparum* genome-wide scans for positive selection, recombination hot spots and resistance to antimalarial drugs. Nat Genet **42**:268-271.

Mzilahowa, T., P. J. McCall, and I. M. Hastings. 2007. "Sexual" population structure and genetics of the malaria agent *P. falciparum*. PLoS One **2**:e613.

Ndiath, M. O., C. Mazenot, A. Gaye, L. Konate, C. Bouganali, O. Faye, C. Sokhna, and J. F. Trape. 2011. Methods to collect *Anopheles* mosquitoes and evaluate malaria transmission: a comparative study in two villages in Senegal. Malar J **10**:270.

Ndiaye, D., V. Patel, A. Demas, M. LeRoux, O. Ndir, S. Mboup, J. Clardy, V. Lakshmanan, J. P. Daily, and D. F. Wirth. 2010. A non-radioactive DAPI-based high-throughput in vitro assay to assess *Plasmodium falciparum* responsiveness to antimalarials--increased sensitivity of *P. falciparum* to chloroquine in Senegal. Am J Trop Med Hyg **82**:228-230.

Neafsey, D. E., K. Galinsky, R. H. Jiang, L. Young, S. M. Sykes, S. Saif, S. Gujja, J. M. Goldberg, S. Young, Q. Zeng, S. B. Chapman, A. P. Dash, A. R. Anvikar, P. L. Sutton, B. W. Birren, A. A. Escalante, J. W. Barnwell, and J. M. Carlton. 2012. The malaria parasite *Plasmodium vivax* exhibits greater genetic diversity than Plasmodium falciparum. Nat Genet **44**:1046-1050.

Neafsey, D. E., D. L. Hartl, and M. Berriman. 2005. Evolution of noncoding and silent coding sites in the *Plasmodium falciparum* and *Plasmodium reichenowi* genomes. Mol Biol Evol **22**:1621-1626.

Neafsey, D. E., S. F. Schaffner, S. K. Volkman, D. Park, P. Montgomery, D. A. Milner, Jr., A. Lukens, D. Rosen, R. Daniels, N. Houde, J. F. Cortese, E. Tyndall, C. Gates, N. Stange-Thomann, O. Sarr, D. Ndiaye, O. Ndir, S. Mboup, M. U. Ferreira, L. Moraes Sdo, A. P. Dash, C. E. Chitnis, R. C. Wiegand, D. L. Hartl, B. W. Birren, E. S. Lander, P. C. Sabeti, and D. F. Wirth. 2008. Genome-wide SNP genotyping highlights the role of natural selection in *Plasmodium falciparum* population divergence. Genome Biol **9**:R171.

Nkhoma, S. C., S. Nair, S. Al-Saai, E. Ashley, R. McGready, A. P. Phyo, F. Nosten, and T. J. Anderson. 2012. Population genetic correlates of declining transmission in a human pathogen. Mol Ecol **22**:273-285.

Pain, A., U. Bohme, A. E. Berry, K. Mungall, R. D. Finn, A. P. Jackson, T. Mourier, J. Mistry, E. M. Pasini, M. A. Aslett, S. Balasubrammaniam, K. Borgwardt, K. Brooks, C. Carret, T. J. Carver, I. Cherevach, T. Chillingworth, T. G. Clark, M. R. Galinski, N. Hall, D. Harper, D. Harris, H. Hauser, A. Ivens, C. S. Janssen, T. Keane, N. Larke, S. Lapp, M. Marti, S. Moule, I. M. Meyer, D. Ormond, N. Peters, M. Sanders, S. Sanders, T. J. Sargeant, M. Simmonds, F. Smith, R. Squares, S. Thurston, A. R. Tivey, D. Walker, B. White, E. Zuiderwijk, C. Churcher, M. A. Quail, A. F. Cowman, C. M. Turner, M. A. Rajandream, C. H. Kocken, A. W. Thomas, C. I. Newbold, B. G. Barrell, and M. Berriman. 2008. The genome of the simian and human malaria parasite *Plasmodium knowlesi*. Nature **455**:799-803.

Park, D. J., A. K. Lukens, D. E. Neafsey, S. F. Schaffner, H. H. Chang, C. Valim, U. Ribacke, D. Van Tyne, K. Galinsky, M. Galligan, J. S. Becker, D. Ndiaye, S. Mboup, R. C. Wiegand, D. L. Hartl, P. C. Sabeti, D. F. Wirth, and S. K. Volkman. 2012. Sequence-based association and selection scans identify drug resistance loci in the *Plasmodium falciparum* malaria parasite. Proc Natl Acad Sci U S A **109**:13052-13057.

Patterson, N., A. L. Price, and D. Reich. 2006. Population structure and eigenanalysis. PLoS Genet **2**:e190.

Pollack, Y., A. L. Katzen, D. T. Spira, and J. Golenser. 1982. The genome of *Plasmodium falciparum*. I: DNA base composition. Nucleic Acids Res **10**:539-546.

Polley, S. D., and D. J. Conway. 2001. Strong diversifying selection on domains of the *Plasmodium falciparum* apical membrane antigen 1 gene. Genetics **158**:1505-1512.

Pritchard, J. K., M. Stephens, and P. Donnelly. 2000. Inference of population structure using multilocus genotype data. Genetics **155**:945-959.

Rich, S. M., M. C. Licht, R. R. Hudson, and F. J. Ayala. 1998. Malaria's Eve: evidence of a recent population bottleneck throughout the world populations of *Plasmodium falciparum*. Proc Natl Acad Sci U S A **95**:4425-4430.

Rocha, E. P., and A. Danchin. 2002. Base composition bias might result from competition for metabolic resources. Trends Genet **18**:291-294.

Roper, C., I. M. Elhassan, L. Hviid, H. Giha, W. Richardson, H. Babiker, G. M. Satti, T. G. Theander, and D. E. Arnot. 1996. Detection of very low level *Plasmodium falciparum* infections using the nested polymerase chain reaction and a reassessment of the epidemiology of unstable malaria in Sudan. Am J Trop Med Hyg **54**:325-331.

Sella, G., D. A. Petrov, M. Przeworski, and P. Andolfatto. 2009. Pervasive natural selection in the *Drosophila* genome? PLoS Genet **5**:e1000495.

Singer, G. A., and D. A. Hickey. 2000. Nucleotide bias causes a genomewide bias in the amino acid composition of proteins. Mol Biol Evol **17**:1581-1588.

Slotte, T., J. P. Foxe, K. M. Hazzouri, and S. I. Wright. 2010. Genome-wide evidence for efficient positive and purifying selection in *Capsella grandiflora*, a plant species with a large effective population size. Mol Biol Evol **27**:1813-1821.

Smith, J. M., N. H. Smith, M. O'Rourke, and B. G. Spratt. 1993. How clonal are bacteria? Proc Natl Acad Sci U S A **90**:4384-4388.

Stephens, M., and P. Donnelly. 2003. A comparison of bayesian methods for haplotype reconstruction from population genotype data. Am J Hum Genet **73**:1162-1169.

Storey, J. D., and R. Tibshirani. 2003. Statistical significance for genomewide studies. Proc Natl Acad Sci U S A **100**:9440-9445.

Su, X. Z., J. Mu, and D. A. Joy. 2003. The "Malaria's Eve" hypothesis and the debate concerning the origin of the human malaria parasite *Plasmodium falciparum*. Microbes Infect **5**:891-896.

Tajima, F. 1989a. DNA polymorphism in a subdivided population: the expected number of segregating sites in the two-subpopulation model. Genetics **123**:229-240.

Tajima, F. 1989b. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics **123**:585-595.

Talman, A. M., O. Domarle, F. E. McKenzie, F. Ariey, and V. Robert. 2004. Gametocytogenesis: the puberty of *Plasmodium falciparum*. Malar J **3**:24.

Toprak, E., A. Veres, J. B. Michel, R. Chait, D. L. Hartl, and R. Kishony. 2012. Evolutionary paths to antibiotic resistance under dynamically sustained drug selection. Nat Genet **44**:101-105.

Torgerson, D. G., A. R. Boyko, R. D. Hernandez, A. Indap, X. Hu, T. J. White, J. J. Sninsky, M. Cargill, M. D. Adams, C. D. Bustamante, and A. G. Clark. 2009. Evolutionary processes acting on candidate cis-regulatory regions in humans inferred from patterns of polymorphism and divergence. PLoS Genet **5**:e1000592.

Van Tyne, D., D. J. Park, S. F. Schaffner, D. E. Neafsey, E. Angelino, J. F. Cortese, K. G. Barnes, D. M. Rosen, A. K. Lukens, R. F. Daniels, D. A. Milner, Jr., C. A. Johnson, I. Shlyakhter, S. R. Grossman, J. S. Becker, D. Yamins, E. K. Karlsson, D. Ndiaye, O. Sarr, S. Mboup, C. Happi, N. A. Furlotte, E. Eskin, H. M. Kang, D. L. Hartl, B. W. Birren, R. C. Wiegand, E. S. Lander, D. F. Wirth, S. K. Volkman, and P. C. Sabeti. 2011. Identification and functional validation of the novel antimalarial resistance locus PF10_0355 in *Plasmodium falciparum*. PLoS Genet **7**:e1001383.

Verra, F., and A. L. Hughes. 2000. Evidence for ancient balanced polymorphism at the Apical Membrane Antigen-1 (AMA-1) locus of *Plasmodium falciparum*. Mol Biochem Parasitol **105**:149-153.

Voight, B. F., S. Kudaravalli, X. Wen, and J. K. Pritchard. 2006. A map of recent positive selection in the human genome. PLoS Biol **4**:e72.

Volkman, S. K., A. E. Barry, E. J. Lyons, K. M. Nielsen, S. M. Thomas, M. Choi, S. S. Thakore, K. P. Day, D. F. Wirth, and D. L. Hartl. 2001. Recent origin of *Plasmodium falciparum* from a single progenitor. Science **293**:482-484.

Volkman, S. K., D. L. Hartl, D. F. Wirth, K. M. Nielsen, M. Choi, S. Batalov, Y. Zhou, D. Plouffe, K. G. Le Roch, R. Abagyan, and E. A. Winzeler. 2002. Excess polymorphisms in genes for membrane proteins in *Plasmodium falciparum*. Science **298**:216-218.

Volkman, S. K., P. C. Sabeti, D. DeCaprio, D. E. Neafsey, S. F. Schaffner, D. A. Milner, Jr., J. P. Daily, O. Sarr, D. Ndiaye, O. Ndir, S. Mboup, M. T. Duraisingh, A. Lukens, A. Derr, N. Stange-Thomann, S. Waggoner, R. Onofrio, L. Ziaugra, E. Mauceli, S. Gnerre, D. B. Jaffe, J. Zainoun, R. C. Wiegand, B. W. Birren, D. L. Hartl, J. E. Galagan, E. S. Lander, and D. F. Wirth. 2007. A genome-wide map of diversity in *Plasmodium falciparum*. Nat Genet **39**:113-119.

Waples, R. S. 1989. A generalized approach for estimating effective population size from temporal changes in allele frequency. Genetics **121**:379-391.

Watterson, G. A. 1975. On the number of segregating sites in genetical models without recombination. Theor Popul Biol **7**:256-276.

World Health Organization. 2011. World malaria report 2011. Geneva: World Health Organization 246.

Williamson, D. H., R. J. Wilson, P. A. Bates, S. McCready, F. Perler, and B. U. Qiang. 1985. Nuclear and mitochondrial DNA of the primate malarial parasite *Plasmodium knowlesi*. Mol Biochem Parasitol **14**:199-209.

Wright, S. 1931. Evolution in Mendelian Populations. Genetics **16**:97-159.

Yang, Z., and R. Nielsen. 2000. Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. Mol Biol Evol **17**:32-43.