MCGEE, KATE MARIE. Evolutionary Forces Shaping Haplotype and Nucleotide Diversity in Humans and Malaria. (Under the direction of Philip Awadalla.)

Cheaper and more rapid DNA sequencing has led to the accumulation of large amounts of genetic data and has fueled the development of new methods to analyze this data. Using population genetics theory and computational methods we can explore the evolutionary forces that shape genetic variation within and among populations of humans and malaria parasites. Demographic events such as population size change influence current patterns of genetic variation. Accounting for the demographic history of a population is critical in the interpretation of population genetic analyses, particularly in detecting of regions under selection and in making inferences about linkage disequilibrium. Characterizing how recombination rates evolve is critical for the efficient design of association studies and, in turn, the understanding of the genetics behind complex phenotypes. In malaria parasites, recombination is a key element in the creation of a wide array of antigens, which help invade host cells. We examine patterns of genetic variation in humans and malaria and explore how demographic history and recombination rates affect these patterns.
Evolutionary Factors Shaping Haplotype and Nucleotide Diversity in Humans and Malaria

by

Kate Marie McGee

A dissertation submitted to the Graduate Faculty of North Carolina State University in partial fulfillment of the requirements for the Degree of Doctor of Philosophy

Bioinformatics

Raleigh, North Carolina

2007

APPROVED BY:

_________________________       _________________________
Jeffrey Thorne                                      Eric Stone

__________________________        _________________________
Philip Awadalla                           Trudy Mackay
Co-chair of Advisory Committee            Co-chair of Advisory Committee
DEDICATION

I dedicate this to my parents, whose unwavering support and encouragement helped me survive this journey.
BIOGRAPHY

Kate Marie McGee was born on October 8, 1981 in Towson, Maryland to Scott and Deborah McGee; she is the youngest of two daughters. Kate attended James Madison University in Harrisonburg, VA from 1999 to 2003 where she earned her Bachelor’s degree in Integrated Science and Technology. During her undergraduate career she completed a summer internship at the University of Maryland (Marlene and Stuart) Greenebaum Cancer Center in Baltimore. It was during her stay there that she was motivated to continue her education. She accepted the offer to pursue her Ph.D. in Bioinformatics at North Carolina State University in 2003 and hopes to pursue research in the discovery of genes underlying common human diseases.
ACKNOWLEDGEMENTS

I would like to thank my advisor, Philip Awadalla, for all of his help and guidance over the last several years and for putting up with massive amounts of emails on a daily basis. I also want to thank my advisory committee, who made sure I understood everything I was doing and who gave helpful advice over the course of my research. In addition, I am grateful to Rachel Myers and Jon Keebler for sitting through lab meetings and listening to me babble about my latest problems. Finally, I want to thank all of the other students, faculty and staff at the BRC who were willing to listen to my research presentations and provide feedback and suggestions.
# TABLE OF CONTENTS

LIST OF TABLES ........................................................................................................... viii

LIST OF FIGURES ........................................................................................................ ix

1 POPULATION GENETICS REVIEW: MAJOR CONCEPTS AND RECENT
   ACHIEVEMENTS ................................................................................................... 1

   Introduction ........................................................................................................ 2

   Coalescent Theory ................................................................................................ 2

   The Effects of Linkage Disequilibrium and Recombination on
   Patterns of Genetic Variation ............................................................................... 5

   Overview ............................................................................................................. 5

   Estimating Recombination Rates from Genetic Variation Data ..................... 6

   Observations in Humans and Malaria .............................................................. 8

   Natural Selection ................................................................................................ 10

   Host-pathogen interaction: An arms race between humans
   and malaria ........................................................................................................... 10

   Research Chapters ............................................................................................. 12

   References ......................................................................................................... 14

   Figures ............................................................................................................... 18

2 GENETIC VARIATION ACROSS THE Plasmodium falciparum Genome .......... 20

   Abstract .............................................................................................................. 21

   Introduction ........................................................................................................ 22

   Results and Discussion ...................................................................................... 23

   Conclusions ........................................................................................................ 27

   Methods ............................................................................................................. 28

   References ......................................................................................................... 32

   Tables ............................................................................................................... 34

   Figures ............................................................................................................... 35
### 3 Inferring the Demographic History of *Plasmodium falciparum* Using Polymorphism and Divergence Data

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abstract</td>
<td>51</td>
</tr>
<tr>
<td>Introduction</td>
<td>52</td>
</tr>
<tr>
<td>Results</td>
<td>55</td>
</tr>
<tr>
<td>Discussion</td>
<td>60</td>
</tr>
<tr>
<td>Methods</td>
<td>61</td>
</tr>
<tr>
<td>References</td>
<td>66</td>
</tr>
<tr>
<td>Tables</td>
<td>69</td>
</tr>
<tr>
<td>Figures</td>
<td>70</td>
</tr>
</tbody>
</table>

### 4 The Evolution of Population Recombination Landscapes: Implications for Demographic History and Association Studies

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abstract</td>
<td>76</td>
</tr>
<tr>
<td>Introduction</td>
<td>78</td>
</tr>
<tr>
<td>Results/Discussion</td>
<td>83</td>
</tr>
<tr>
<td>Conclusions</td>
<td>92</td>
</tr>
<tr>
<td>Methods</td>
<td>93</td>
</tr>
<tr>
<td>References</td>
<td>99</td>
</tr>
<tr>
<td>Tables</td>
<td>104</td>
</tr>
<tr>
<td>Figures</td>
<td>106</td>
</tr>
</tbody>
</table>

### 5 Detecting Selection and Exploring Associations in Functionally-Related Genes

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abstract</td>
<td>117</td>
</tr>
<tr>
<td>Introduction</td>
<td>118</td>
</tr>
<tr>
<td>Results and Discussion</td>
<td>120</td>
</tr>
<tr>
<td>Conclusions</td>
<td>125</td>
</tr>
<tr>
<td>Methods</td>
<td>126</td>
</tr>
<tr>
<td>References</td>
<td>131</td>
</tr>
</tbody>
</table>
## LIST OF TABLES

| Table 2-1 | Summary of sequenced genes and diversity among five isolates of *P. falciparum* | 34 |
| Table 2-2 | Summary of polymorphisms and substitutions in genes with significant McDonald-Kreitman $p$-values | 35 |
| Table 3-1 | Summary statistics across Chromosome 3 in 98 parasite isolates | 69 |
| Table 4-1 | Summary of correlation in recombination rates among two simulated populations when all polymorphisms are used to estimate $\rho$ across region and only sites that are segregating in both populations are used to estimate $\rho$ | 104 |
| Table 4-2 | Proportion of shared polymorphisms and correlation in recombination landscapes | 105 |
| Table 5-1 | Summary of population differentiation | 134 |
| Table 5-2 | Genes with significant Fay and Wu’s $H$ | 135 |
| Table 5-3 | Summary of LD results for the Clotting and PAR pathways | 137 |
| Table 5-4 | Summary of MHC-1 pathway LD results for all HapMap Populations | 137 |
LIST OF FIGURES

Figure 1-1 Example coalescent trees under neutral and population expansion models ................................................................. 4

Figure 1-2 The basic principle behind the coalescent ................................................. 5

Figure 1-3 Effect of a selective sweep on genetic variation ........................................ 9

Figure 2-1 Physical maps showing the distribution of polymorphisms across chromosomes ........................................................................................................................................ 35

Figure 2-2 Correlation between effective number of codons and GC3s ................. 36

Figure 2-3 Highly polymorphic genes grouped by GO functional terms ............. 36

Figure 2-4 Distribution of Tajima’s D across each chromosome for genes with at least one polymorphism ................................................................. 37

Figure 2-5 Distribution of \( \alpha \) across each chromosome for genes with at least one synonymous polymorphism ................................................................. 41

Figure 2-6 Genes influence by natural selection (\( \alpha=1 \)) grouped by GO process terms ................................................................................................................................. 44

Figure 2-7 Recombination matrices for each chromosome showing the minimum number of recombination events between each pair of sites .......... 45

Figure 3-1 Sampling locations in Africa, Asia, Central and South America, and Papua New Guinea ................................................................. 70

Figure 3-2 Inferred population structure of global parasite populations ............ 70

Figure 3-3 Power analyses ....................................................................................... 71

Figure 3-4 Likelihood surface for population models ........................................... 72

Figure 3-5 Schematic of global models explored .................................................. 73

Figure 3-6 Likelihood surface for global models ................................................... 74

Figure 4-1 Persistence of ancestral recombination events over time ................ 106
Figure 4-2  Distribution of correlation in recombination landscapes between pairs of populations that diverge at different time intervals .............................. 106

Figure 4-3  Distribution of correlation in population recombination rates across values of $\theta$ ........................................................................................................ 107

Figure 4-4  Effects of sample size and total recombination map length on correlations of map length among populations .................................................... 108

Figure 4-5  Estimates of population recombination rate variation across three 3Mb regions using the RJMCMC with a jump penalty of 20 for HapMap individuals from Europe (red), Africa (blue), and Asia (green) ................................................................. 109

Figure 4-6  Power analysis ........................................................................................ 110

Figure 4-7  Probability of the split time ($T_{split}$) given the observed correlation in recombination landscapes .............................................................. 111

Figure 4-8  Distribution of correlation coefficients across 86 genes sequenced in the Seattle SNPs Project ................................................................. 112

Figure 4-9  Comparison of empirical and simulated correlation in recombination landscapes among two populations ............................................... 113

Figure 4-10 Distribution of recombination events along the first 4.5Mb of the extended MHC region ............................................................................ 114

Figure 4-11 The distribution of detectable recombination events on Chromosome 3 of $P. falciparum$ ........................................................................ 115

Figure 5-1  Distribution of $c_j$ values across populations .................................... 138

Figure 5-2  Distribution of genes that show a significant departure from neutrality according to the McDonald-Kreitman test ................................. 138

Figure 5-3  Correlation of $H/S$ among populations ............................................ 139

Figure 5-4  Schematic of the (a) Clotting and (b) PAR pathways .......................... 140

Figure 5-5  Schematic of the MHC-1 pathway .................................................... 141
CHAPTER 1

POPULATION GENETICS REVIEW: MAJOR CONCEPTS AND RECENT ACHIEVEMENTS
INTRODUCTION

Population genetics is the study of changes in allele frequencies in a population over time, the forces responsible for those changes, and their effects on evolution and adaptation\(^1, 2\). The major goals of population genetics are to enhance our knowledge of biological processes, to investigate the evolutionary history of species, and to elucidate the link between DNA sequence and physical traits (i.e. phenotype). The advent of technologies that allow cheaper and more rapid DNA sequencing has led to the accumulation of large amounts of genetic data, which fueled the development of new methods to analyze this data. From identifying genes underlying diseases susceptibility and interpreting forensic DNA evidence to designing breeding programs to preserve endangered species, population genetics methods have wide reaching applications. The following sections detail concepts and achievements that are relevant to the subsequent research chapters.

COALESCENT THEORY\(^{[3-8]}\)

Coalescent theory describes a probabilistic model of the genealogical process that can be used to infer details about the history of a sample. It is a retrospective approach that starts with a sample and traces backwards in time to infer events in the history of the sample. The coalescent process can be used to generate a distribution of genealogies (or trees) made up of coalescence events. A coalescence event occurs when two sequences come from the same parent in the previous generation. All coalescence events are independent and it is assumed that no more than one event occurs in each generation. A genealogy is simulated until all
lineages coalesce into a single lineage, known as the most recent common ancestor (MRCA). When time is considered on a continuous scale, the time to coalescence is exponentially distributed and the probability that two sequences coalesce in the previous generation is $1/2N$, where $N$ is the population size. The probability that two sequences had a common ancestor $t+1$ generations in the past is

$$\left(\frac{1}{2N}\right)^*\left(1 - \frac{1}{2N}\right).$$

This can be generalized to a sample of $n$ sequences such that the probability that any two sequences coalesced $t+1$ generations ago is

$$\left(\frac{n(n-1)}{4N}\right)^*\left(1 - \frac{n(n-1)}{4N}\right).$$

As the sample size approaches $2N$, the expected time to the MRCA of all alleles in a population is $4N$. This is the “coalescent effective size” and times are measured in terms of $4N$ generations.

The basic coalescent makes all of the assumptions of the Wright-Fisher model of evolution (an ideal population): constant population size, no population structure, no migration, no recombination, and no selection (i.e. all mutations are neutral and have no effect on fitness). Most of these assumptions are violated by real populations and the coalescent has been extended to allow more realistic models including (but not limited to) population size change and recombination. A change in population size affects the probability distribution of coalescence times. For a model of exponential growth where $N_t =$
\( N_0 e^{-\alpha t} \), genealogies have longer terminal branches and more, shorter branches near the root (Figure 1-1) because it takes longer for sequences to coalesce when the population is growing. Incorporating recombination into the coalescent model is complex, especially for an intermediate rate of recombination between loci because the segments will have correlated genealogies. Hudson first described a coalescent model with recombination using a set of exponentially distributed waiting times for one of two events: coalescence or recombination\(^9\). Considering only two sequences, the time back until an event occurs is \( \left( \frac{2N}{1+\rho} \right) \), where \( \rho = 4Nr \), and the probability that this event is a coalescent event is \( \left( \frac{1}{1+\rho} \right) \).

If instead the first event is a recombination, occurring with probability \( \left( \frac{\rho}{1+\rho} \right) \), one of the lineages will split into two producing a total of three lineages. By doing this, recombination increases the size of a genealogy.

Using coalescent theory, it is only necessary to keep track of individuals that are ancestral to the sample rather than the entire population (Figure 1-2) and is therefore computationally more efficient than traditional, forward approaches. It has many uses including detecting deviations from the neutral model, simulating sequence data, and performing likelihood analyses.

---

\(^{1}\) \( N_t \): population size at time \( t \), \( N_0 \): current population size, \( \alpha \): growth rate, \( t \): time growth began in \( 4N_0 \) generations
THE EFFECTS OF LINKAGE DISEQUILIBRIUM AND RECOMBINATION ON PATTERNS OF GENETIC VARIATION\textsuperscript{[7, 10, 11]}

Overview

Linkage disequilibrium (LD) is the non-random association of alleles, that is two or more alleles are more likely to be inherited together than expected by chance. More specifically, genotype frequencies are different than expected under Hardy-Weinberg equilibrium expectations, which predicts that under random mating, the frequency of a genotype is equal to the product of each of the allele frequencies. LD is affected by many forces including genetic drift, population admixture, natural selection, and recombination. Chance associations will occur in small populations due to random changes in allele frequencies or if a new mutation arises and has not yet increased in frequency in the population\textsuperscript{[7]}. If a specific combination of alleles is beneficial, natural selection may maintain this haplotype in high frequency in a population, thereby increasing levels of allelic association. LD can also arise if two populations with differing allele frequencies mix. For example if the major alleles at two loci are $A$ and $B$ in one sub-population and $a$ and $b$ in another, these two loci would be in complete LD. While drift, sub-structure and selection create and maintain levels of LD, recombination breaks down these associations by, essentially, shuffling alleles.

Recombination is the exchange of genetic material between homologous chromosomes and is an important factor in DNA sequence evolution. It generates genetic variation by creating novel combinations of alleles and it breaks down allelic associations (LD), effectively generating independent instances of the genealogical process. By breaking
down LD, recombination may also increase the efficiency of natural selection. Interference is the effect of allelic associations or LD on the efficacy of selection at those loci. Hill and Robertson hypothesized that these associations would impede the action of selection by increasing the probably of an unfavorable allele becoming fixed in the population and decreasing the probably of fixation of a favorable allele\[12\]. By breaking up these associations, recombination can bring multiple beneficial (or deleterious) mutations together so they can spread through (or be removed from) the population, thereby allowing selection to be more effective.

**Estimating Recombination Rates Using Genetic Variation Data**

Direct measurement of recombination rates (i.e. sperm typing) is laborious and costly and pedigree studies have little power to determine rates at a fine scale\[13-15\]. Large-scale surveys of genetic variation spurred the development of statistical methods to estimate recombination rates. The simplest method of identifying recombination events is to look at all pairs of segregating sites and identify pairs where all four possible haplotypes are present. This method, known as the four gamete test (FGT), estimates the minimum number of events that have occurred in the history of a sample \(R_m\). \(R_m\) is a conservative estimate because it assumes that all overlapping regions where a recombination is observed originate from a single recombination event\[16\]. Myers and Griffiths extended this idea by developing a new technique that combines local recombination bounds to generate a bound on a larger region\[17\]. They derived two estimators using this technique: \(R_h\) estimates the number of events based on the difference in the number of haplotypes observed and the number of
segregating sites, while \( R_s \) is based on approximating the history of the data by minimizing the number of recombination events\(^{17}\).

While we can learn something about recombination from counting methods, they generally underestimate the amount of recombination that has occurred in the history of the sample\(^{18,19}\). To obtain better estimates of the recombination landscape from population genetic data likelihood approaches, which estimate the probability of the observed data under a given model, have been developed. Full likelihood methods are too computationally intensive to implement on large datasets and, as a result, composite (or approximate) likelihood methods are often used to estimate recombination rates for large regions. Hudson first described such a method assuming an infinite-sites model and bi-allelic markers\(^{20}\). Data are examined two loci at a time and all possible two-locus genealogies are generated using the observed mutation rate and a range of population recombination rate \( (\rho=4N_{e}r) \) values. The likelihood of the sample configuration for each pair of sites is the average over all genealogies and the overall likelihood of a specific value of \( \rho \) is obtained by multiplying the likelihood at each pair of sites. In this way, a single likelihood value is estimated for each value of \( \rho \). The point on the likelihood surface where the likelihood is the greatest (maximum likelihood) is the estimate of \( \rho \) for the sampled region. This method has been extended to allow for recurrent mutation\(^{21}\) and multi-allelic sites\(^{22}\), and to incorporate variable recombination rates\(^{23}\).

Li and Stephens developed a model that directly relates the distribution of sampled haplotypes to the underlying recombination rate\(^{24}\). An approximation is used to obtain the probability distribution of the observed haplotypes given the recombination rate. Simply, if
we observed \( k \) haplotypes, what is the conditional distribution of the type of the next sampled chromosome? This unseen type is a mosaic of imperfectly copied segments of the previously observed haplotypes. The probability of observing a particular type is obtained using a Markov model to sum over all possible mosaic patterns\(^{[24]}\). This method can be used to obtain a point estimate of the recombination rate when it is assumed to be constant and to estimate variation in rates across a region.

**Observations in Humans and Malaria**

Similar to the pattern observed in yeast\(^{[25]}\), the landscape of recombination in humans and malaria parasites is heterogeneous and marked by regions with recombination rates much higher than the surrounding regions (hotspots). While differences in the magnitude of recombination rates have been observed among populations, the location of major hotspots appears to be conserved\(^{[23, 26, 27]}\). Very little is known about the mechanisms influencing the location and pattern of hotspots. Several studies in humans have shown that variation in alleles at nearby polymorphic loci seemingly affect hotspot activity at MS32, DNA2 and NID1, three well-characterized hotspots in humans\(^{[28-30]}\). Recently, Myers, *et al.*\(^{[31]}\) matched hotspots with coldspots of similar size and SNP density and explored differences between the two. They found that terminal repeats of two retrotransposons and CT-rich repeats are overrepresented in hotspots\(^{[31]}\).

By producing new combinations of alleles, recombination is a critical process in the creation of antigenic variation. Extensive sequence diversity creates an arsenal of antigens that help attack host cells and evade host immunity. The *var* gene family encodes the *P.*
*falciparum* erythrocyte membrane binding protein-1 (PfEMP1) family, which is transported to the surface of the infected red blood cells where it can bind to host receptors\(^{32}\). While more than 50 *var* genes are typically found in the genome of this parasite, only a single gene is expressed at any given time\(^{33}\). Members of the *var* gene family have been detected on all chromosomes except 14, and are located mostly among the conserved, repetitive sequences in the sub-telomeric regions of chromosomes\(^{34}\). Taylor, *et al* examined conserved motifs in *var* genes in the 3D7 strain of *P. falciparum* and found that, on average, variability is as great within isolates as it is among isolates\(^{35}\). In addition, Freitas-Junior found that parasites from diverse locations have minimal overlap in their set of *var* genes\(^{36}\) and Kraemer, *et al* observed that only 7 of 31 domain architectures are found in all three isolates examined (3D7, HB3, IT4) and 19 were unique to a single isolate\(^{37}\), supporting diversity among isolates from different locations. Genes located on the ends of heterologous chromosomes are more similar than those found in central regions of homologous chromosomes suggesting that recombination occurs between chromosome ends\(^{34}\). In addition, crosses of two different strains result in progeny with non-parental *var* types more often than expected based on estimated recombination frequencies\(^{35, 36}\). Evidence for intra- and inter-genic recombination suggests that the location of these genes in regions of enhanced recombination on the chromosome ends contributes to the diversity among genes, which is necessary for efficient host invasion.
NATURAL SELECTION

Natural selection is a major force driving sequence evolution. Genes that have been targets of selection contribute to variation in fitness among individuals, therefore identifying regions of the genome that show signatures of selection will help elucidate genotype-phenotype associations and broaden our understanding of genome evolution. Different types of selection leave distinctive signatures on genetic variation. Positive selection (i.e. selective sweep) depletes the pool of genetic variation and increases levels of linkage disequilibrium by rapidly increasing the frequency of a favorable allele in a population, taking the background haplotype with it (Figure 1-3). The generation of genome-wide polymorphism data has allowed the identification of genes that show a signature of a recent selective sweep (e.g. [38-41]). Contrarily, balancing selection acts to maintain variation (multiple alleles) at a locus rather than increasing the frequency of a single allele. In humans, this is most evident in genes involved in the immune response including HLA genes[42] and G6PD[43].

Host-pathogen interaction: An arms race between humans and malaria

Selection plays a major role in the ability of pathogens to infect their host and the efficiency of host immunity to recognize and destroy invading pathogens. This type of arms race can be seen clearly at two human loci that are thought to be under selection from malaria. The HBB locus, which produces beta hemoglobin, is the classic example of balanced polymorphism in the human genome. Three different single nucleotide polymorphisms in the HBB gene that confer protection against malaria have arisen independently and are present at different frequencies among populations[44]. The HbS, or
sickle cell variant is common in Africa and confers a 10-fold reduction in risk of severe malaria in heterozygotes\[^{45}\], but results in sickle cell anemia in homozygotes. The HbE variant is rare in Africa, but common in Southeast Asia. One study found that individuals in reference group were nearly 7 times more likely to suffer from severe malaria than patients with the HbE variant\[^{46}\]. The selective advantage for the third variant, HbC, is greatest for homozygotes (> 90%) and *in vitro* studies suggest that a reduction in PfEMP1 expression is the mechanism for the protective effect against severe disease\[^{47, 48}\].

Glucose-6-phosphate dehydrogenase (G6PD) protects cells against oxidative stress, and deficiencies in this enzyme result in several blood disorders. Correlation in the distribution of these deficiencies with endemicity of malaria suggests that natural selection by malaria has caused them to persist and increase in frequency. It was reported that G6PD deficiencies are correlated with protection against severe malaria in Nigerian children\[^{49}\]. The common African form (G6PD A-) results in an 88% reduction in enzyme activity, but has been shown to reduce the risk of severe malaria by 46% in females who are heterozygous for the allele and by 58% in hemizygous males\[^{50}\]. It is thought that red blood cells deficient in G6PD reduce the replication of parasites, thereby reducing parasite load and conferring resistance. Parasites have a memory of previously inhabited cells, and when they are grown in G6PD-deficient cells for several cycles it appears that *P. falciparum* compensates for this deficiency by making the enzyme itself\[^{51}\]. These two examples illustrate how both host and pathogen accumulate traits in response to environmental stressors to aid in survival.
RESEARCH CHAPTERS

Chapter 2 – Genetic Variation Across the *Plasmodium falciparum* Genome

Sequencing of the *Plasmodium falciparum* genome brings the promise of new drugs and possibly a vaccine to treat and prevent the most lethal form of malaria. In this chapter, we discuss the basic patterns of variation observed across the genomes of 5 isolates of *P. falciparum* as well as characterization of genes by GO process classification and results from tests of selection.

Chapter 3 - Inferring the Demographic History of *Plasmodium falciparum* Using Polymorphism and Divergence Data

The demographic history of a species is important when trying to detect signatures of natural selection or to characterize levels of linkage disequilibrium. In this chapter, we use a set of summary statistics to attempt to infer the evolutionary history of four worldwide populations of *Plasmodium falciparum*.

Chapter 4 - The Evolution of Population Recombination Landscapes: Implications for Demographic History and Association Studies

Variation in the landscape of recombination has been observed both within and among species. Specifically, little similarity has been seen in recombination rates for chimpanzees and humans. This chapter investigates the change in recombination rates over time by simulating sequence data under a range of parameter values for two populations with a common ancestor. We also explore what these simulations may be able to tell us about how long ago real populations have diverged based on their recombination patterns.
Chapter 5 – Detecting Selection and Exploring Associations in Functionally-Related Genes

New, less expensive technologies have allowed the accumulation of polymorphism data, and with it, the widespread use of population genetics methods to analyze this data. Chapter 3 revisits analysis to detect population structure in human populations and departures from neutrality at multiple classes of sites and examines the idea of conserved associations between functionally related genes. We also look at phenotypic differences between populations, as measured by differences in expression levels.
REFERENCES


Figure 1-1. Example coalescent trees under (a) neutral and (b) population expansion models.

Figure 1-2. The basic principle behind the coalescent.
Figure 1-3. Effect of a selective sweep on genetic variation. Each line is a chromosome and each circle represents a mutation. The mutation in red is the new, beneficial mutation. The haplotype on which the beneficial mutation arose spreads through the population, virtually eliminating variation until new mutations arise (blue).
CHAPTER 2

GENETIC VARIATION ACROSS THE *Plasmodium falciparum* GENOME

This chapter is an extension of reference [1].
ABSTRACT

Malaria causes the death of over two million people every year, mainly in sub-Saharan Africa. A major goal in sequencing the genome of \textit{P. falciparum} is to discover new drug targets; however more than half of the genes have unknown function making this very difficult. Known malaria antigens are highly polymorphic and under selective pressure, therefore identifying genes with signatures of selection may lead to the discovery of new target genes. We surveyed more than 3500 genes for polymorphisms and were able to characterize several genes with previously unknown function as antigens using immunoassays. In addition, we identified a number of genes showing signatures of selection.
INTRODUCTION

Malaria is a parasitic infection of red blood cells that can result in a range of symptoms from anemia and fever to coma and possibly death. Four species of Plasmodium parasites infect humans: P. malariae, P. ovale, P. vivax and P. falciparum, the agent that causes the most lethal form of the disease. Malaria is endemic in the Americas, parts of Asia and Africa. There are 300-500 million cases reported worldwide every year resulting in more than 2 million deaths, most of which occur in sub-Saharan Africa[2]. As such, large efforts are being made to develop vaccines and treatments to prevent and combat this disease. Each stage of the Plasmodium life cycle exhibits different biochemical properties and therefore drugs that may be effective at one stage of the cycle may have no effect at another[3]. Use of anti-malarial drugs in conjunction with wide-spread pesticide use to control mosquito populations has proved unsuccessful in the eradication of the disease[3]. Parasites have developed resistance to anti-malarial drugs, such as chloroquine, that were once widely used and successful in saving millions of lives rendering them practically useless in treating falciparum malaria[3], and mosquitoes have built up immunity to pesticides. In order to significantly reduce the incidence of this disease worldwide and to prevent the further development of parasite resistance to new therapeutic agents, we must continue to investigate the patterns of and mechanisms behind parasite evolution.

Uncovering patterns of genotypic and phenotypic variation among populations and across life stages is critical in understanding the mechanisms of evolution. A genome-wide map of polymorphic markers will greatly enhance our ability to identify genes underlying
important phenotypic traits such as host invasion and drug resistance. More than 3,500 predicted genes across the genome were sequenced and nearly 4,000 well-validated single-nucleotide polymorphisms (SNPs) were extracted. We found that genes with known antigenic function were among the most highly polymorphic and we were able to characterize several genes with previously unknown function using immuno-assays. We also identified many genes that have signatures of selection and found that some genes exhibit a codon usage bias.

RESULTS AND DISCUSSION

Patterns of variation

After eliminating polymorphisms that fell in regions of poor alignment or repetitive elements (e.g. microsatellites), 3918 SNPs remained giving us a density of approximately 1 SNP per 5.9kb. Figure 2-1 shows the distribution of SNPs and microsatellites across each chromosome. The genome-wide average population mutation rate ($\theta=4N_e\mu$) based on the number of segregating sites ($\theta_H$) is $5.05 \times 10^{-4}$ and based on the average number of pairwise differences ($\theta_S$) is $4.83 \times 10^{-4}$. Most genes have less than three SNPs and SNP density is inversely related to chromosome size. (Per-chromosome estimates are shown in Table 2-1). The majority of polymorphisms observed are nonsynonymous suggesting beneficial amino acid substitutions have arisen, potentially to adapt to changes in host environments or as a response to eradication attempts.

We estimated the effective number of codons ($N_c$) for each gene, a measure that quantifies how far codon usage departs from equal usage of synonymous codons$^{[4]}$. $N_c$ falls
between 20, when only one codon is effectively used for each codon, and 61, when all synonymous codons are equally likely\(^4\). We found that \(N_c\) is normally distributed across genes with a mean of approximately 40 and a standard deviation of 3.7. There are 14 genes with an \(N_c\) less than 30, suggesting that some genes may be experiencing negative selection for non-optimal codons (i.e. codon bias). In humans, there is a strong correlation between \(N_c\) and the fraction of codons, of those with a guanine or cytosine at the third position, that are synonymous at the third codon position (GC3s)\(^4\). As shown in Figure 2-2, this is also true in \textit{P. falciparum} \((r^2=0.52)\).

Grouping genes with similar biological functions can not only tell us how specific classes of genes are evolving but may also elucidate the functions of previously uncharacterized genes. Gene ontology (GO) process classifications\(^5\) revealed that the most polymorphic genes encode antigens, cell adhesion molecules, and proteins involved in drug susceptibility and cell communication (\textbf{Figure 2-3}). More than 100 highly polymorphic genes were expressed and detected using pooled human immune sera. 7 of the 65 expressed proteins recognized by the immune sera were previously unknown antigens\(^1\).

\textit{Signatures of selection}

Tajima’s \(D\) statistic\(^6\) is designed to detect departures from the neutral model of evolution using two estimates of the population mutation rate. Values significantly different from zero indicate a violation of one or more assumptions of the neutral model. We calculated Tajima’s \(D\) for all genes with at least one polymorphism. Both evolutionary and selective forces can produce non-zero \(D\) values. Evolutionary forces such as population size
change will affect all loci equally, while selection has a locus-specific effect. In addition, due to the small sample size it is inappropriate to assign significance to the statistic; therefore we applied the genome-wide average ± 1 standard deviation (0.251 ± 2.779) as a threshold for identifying outlier genes that may be under selection. The distribution across each chromosome is shown in Figure 2-4. Of the 1917 genes with at least one polymorphism, 101 genes fall outside the threshold values. Over 95% of these genes have positive $D$ values, suggesting that they may be under balancing selection to maintain genetic variation. According to their GO process classification, many of the genes are involved in cell growth and/or maintenance and metabolism, and three are involved in transport. The remaining three outlier genes (PFE0335w, PF11_0024, and PF14_0086), all of which are hypothetical proteins, have negative Tajima’s $D$ values, a signal that they may have experienced a selective sweep. When the threshold is increased to the genome average ± two standard deviations, 66 percent of the previously identified genes remain outliers, including 2 of the 3 genes with negative Tajima’s $D$ values.

The neutral theory of evolution assumes that most genetic differences between species are selectively neutral, or that they have no effect on the fitness of an organism, and changes in allele frequencies are due solely to random genetic drift\(^7\). As such, the ratio of synonymous to nonsynonymous polymorphisms within species should be proportional to the ratio of synonymous to nonsynonymous differences between species. The McDonald-Kreitman (MK) test\(^8\) assesses whether observed data violate this assumption. We found six genes with a significant departure from the neutral expectation ($p \leq 0.05$), all of which have a deficiency of synonymous polymorphism (5 of the 6 contain no synonymous polymorphism),
suggesting there may be a codon usage bias (Table 2-2). The estimated $N_c$ for each of these genes is less than or equal to 40, supporting codon bias as a reasonable explanation for a deficiency of synonymous polymorphisms. Using an extension of the MK test developed by Smith and Eyre-Walker[9] we estimate that at least 34% of substitutions are driven by positive selection on each chromosome, with the exception of chromosome 1. The distribution of $\alpha$, the proportion of adaptive amino acid substitutions, across all genes with at least one synonymous polymorphism is shown in Figure 2-5 and is highly biased toward extreme values (1 and < 0). $\alpha$ is negative when the ratio of nonsynonymous to synonymous polymorphisms within species is large or the ratio of the number of nonsynonymous to synonymous substitutions between species is small. The most likely cause of negative values is the lack of synonymous polymorphisms found in the *P. falciparum* genome. Values of 1 indicate that all substitutions are driven by positive selection indicating that these genes have undergone adaptive evolution since the divergence of *P. falciparum* from *P. reichenowi*. The GO process classifications most highly represented in this group of genes are cell growth and/or maintenance, metabolism, biosynthesis and transport (Figure 2-6).

We estimated the per-generation genome-wide amino acid mutation rate ($M$) and the genome-wide deleterious mutation rate ($U$) using the method described in [10] and found that the two rates are nearly equal ($M=0.055$ and $U=0.049$), and are much lower than estimates previously reported for mammals which range from 0.59 to 6.3 and 0.49 to 3.0 for $M$ and $U$, respectively[10]. These values indicate that selection eliminates nearly 90-percent of nonsynonymous mutations, suggesting that the large number of amino acid mutations detected in the genome are either neutral or confer a selective advantage.
Pattern of recombination

*Plasmodium* species have a haploid phase in the vertebrate host and a diploid phase in the *Anopheles* mosquito, where recombination can occur. In regions with high endemicity a single mosquito host may contain multiple parasite genotypes. During zygote formation a male and female gamete with different genotypes can fuse, resulting in a recombinant genotype\[^{[11]}\]. Previously published results for chromosome 3 show that the population recombination rates ($\rho=4N_{e}r$) for *P. falciparum* are extremely high\[^{[12]}\]. Not surprising, strains from Africa, where infection rates are highest, exhibit the highest amount of recombination. We further investigated the recombination landscape across the entire genome by estimating the minimum number of recombination events in our sample as described in [13]. We found that recombination events tend to cluster in the sub-telomeric regions of the chromosomes (Figure 2-7).

CONCLUSIONS

Hundreds of millions of cases of malaria are reported and more than 1 million people die of the disease every year. Malaria parasites have developed resistance to once widely used and effective drug treatments. A dense map of genetic variation across the genome of *P. falciparum* is critical in the fight against malaria. This study provides thousands of well-validated markers that can be utilized in the identification of genes responsible for disease severity and drug resistance. We have shown that surveying genome-wide genetic variation for signatures of selection can aid in the identification of previously uncharacterized
antigens. This map can also be utilized to obtain fine-scale recombination rates, which is critical for association studies.

METHODS

Data

Whole genome sequence data for 5 isolates of *P. falciparum* was obtained from colleagues at the National Institutes of Health, National Institute for Allergies and Infectious Diseases (NIH/NAID). The isolates come from Africa (3D7), Asia (Dd2), Central and South America (HB3, 7G8) and Papua New Guinea (D10).

Sequence data for homologous gene regions in *P. reichenowi*, the parasite that infects primates, was extracted from PlasmoDB (Sept. 2005) using the BLAST search engine provided. Data for a total of 959 genes across the genome was obtained. In addition, data was also obtained from the supplementary table in [14].

Sequence alignment and SNP identification

Raw sequence data (chromatograms) were run through *Phred*\(^{[15, 16]}\) to obtain base calls and quality scores for each contig. *Phrap*\(^{[17]}\) was used to assemble consensus sequences for each of the four isolates and Mace aligned the consensus sequences. The genomic sequence for 3D7 was obtained for each gene from PlasmoDB (www.plasmodb.org) and used as a template. A Java program was written (by Jon Keebler) to reverse complement certain sequences to facilitate a final alignment of all five sequences using ClustalW\(^{[18]}\).
A set of Perl scripts was written to identify SNPs. Searching started 25 bp from the first position in the alignment where all sequences have a nucleotide (no gaps). A polymorphic site was identified as a true SNP if the 30 bases upstream and the 30 bases downstream have a quality score of at least 20.

*Data quality control*

Each alignment was manually inspected to further eliminate “junk” SNPs that fall in regions of low quality sequence or poor alignment. This occurred most frequently at the beginning and end of the sequence reads or in repeat regions (i.e. microsatellites). The automated SNPs were also compared to those manually extracted by our colleagues at the NIH/NIAID and a final set of 3,918 well-validated SNPs was assembled.

*Summary Statistics*

The population mutation rate ($\theta=4N_e\mu$) based on the number of segregating sites ($\theta_W$), average pairwise nucleotide diversity ($\theta_a$), and Tajima’s $D$ were all calculated using a Perl script. Base usage and codon bias indices were obtained using CodonW\cite{19}

*Divergence*

The McDonald-Kreitman (MK) test was performed via the Fisher’s Exact Test module of the statistical software package R to test for deviations in the ratio of synonymous to nonsynonymous polymorphism within *P. falciparum* isolates ($P_s$ and $P_n$, respectively) and substitutions between *P. falciparum* and *P. reichenowi* ($D_s$ and $D_n$, respectively). $P_s$, $P_n$, $D_s$, $D_n$
and $D_n$ represent raw, observed counts. Using an extension of the MK test developed by Smith and Eyre-Walker\textsuperscript{[9]} we estimated the proportion of adaptive amino acid substitutions ($\alpha$) for genes with at least one synonymous polymorphism (Eq. 1) and across each chromosome (Eq. 2) as follows:

\begin{align*}
\alpha &= 1 - \frac{D_n P_s}{D_n P_s} \\
\bar{\alpha} &= 1 - \frac{D_s}{D_n} \left( \frac{P_n}{P_s + 1} \right)
\end{align*}

We applied the method described in [10] to calculate the genome-wide amino acid mutation rate ($M$) and the genome-wide deleterious mutation rate ($U$). We estimated the rate of synonymous substitutions per nonsynonymous site ($dS$) and nonsynonymous substitutions per nonsynonymous site ($dN$) by counting the number of observed synonymous and nonsynonymous changes between species and estimated the total number of synonymous and nonsynonymous sites by examining the degeneracy of each codon. The genome-wide mutation rates were then calculated as follows: $M = Z \left( \overline{dS} * \overline{D_n} \right)$ and $U = M - Z dN / 3$, where averages are across 959 genes and $Z$ is a constant that converts the per site estimate to a per genome per generation estimate. We assume there are 5300 genes with an average length of 4000bp, 4 generations per year, and 1.2 million years of divergence between $P. reichenowi$ and $P. falciparum$.

**Recombination**

We estimated the minimum number of recombination events in the history of our sample using the program RecMin, which calculates the statistics ($R_h$ and $R_v$) described in [13]. $R_h$ estimates the number of events based on the difference in the number of haplotypes
observed and the number of segregating sites, while $R_s$ is based on approximating the history of the data by minimizing the number of recombination events. The analyses described above are based on estimates of $R_s$. 

REFERENCES


Table 2-1. Summary of sequenced genes and diversity among five isolates of *P. falciparum*.

<table>
<thead>
<tr>
<th>Chr.</th>
<th>Bp sequenced&lt;sup&gt;a&lt;/sup&gt;</th>
<th>Percentage sequenced</th>
<th>Genes sequenced&lt;sup&gt;b&lt;/sup&gt;</th>
<th>Genes with SNPs</th>
<th>Number of SNPs</th>
<th>Genes with &gt;4 SNPs</th>
<th>sSNPs</th>
<th>nsSNPs</th>
<th>ncSNPs</th>
<th>MS</th>
<th>θ (10^-4)</th>
<th>π (10^-4)</th>
<th>Rec. events</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>81,256</td>
<td>12.6</td>
<td>73 (155)</td>
<td>44</td>
<td>120</td>
<td>3</td>
<td>35</td>
<td>81</td>
<td>4</td>
<td>62</td>
<td>6.21</td>
<td>6.18</td>
<td>18</td>
</tr>
<tr>
<td>2</td>
<td>136,980</td>
<td>14.6</td>
<td>116 (224)</td>
<td>70</td>
<td>134</td>
<td>3</td>
<td>48</td>
<td>82</td>
<td>4</td>
<td>70</td>
<td>6.71</td>
<td>6.52</td>
<td>12</td>
</tr>
<tr>
<td>3</td>
<td>203,069</td>
<td>19.1</td>
<td>203 (245)</td>
<td>108</td>
<td>229</td>
<td>6</td>
<td>60</td>
<td>138</td>
<td>31</td>
<td>165</td>
<td>4.79</td>
<td>4.44</td>
<td>24</td>
</tr>
<tr>
<td>4</td>
<td>128,567</td>
<td>10.7</td>
<td>114 (249)</td>
<td>59</td>
<td>120</td>
<td>3</td>
<td>32</td>
<td>86</td>
<td>2</td>
<td>71</td>
<td>4.91</td>
<td>5.10</td>
<td>16</td>
</tr>
<tr>
<td>5</td>
<td>280,845</td>
<td>20.9</td>
<td>230 (330)</td>
<td>118</td>
<td>240</td>
<td>7</td>
<td>68</td>
<td>156</td>
<td>16</td>
<td>172</td>
<td>3.98</td>
<td>4.7</td>
<td>41</td>
</tr>
<tr>
<td>6</td>
<td>302,182</td>
<td>21.9</td>
<td>234 (319)</td>
<td>130</td>
<td>258</td>
<td>7</td>
<td>78</td>
<td>162</td>
<td>18</td>
<td>198</td>
<td>4.36</td>
<td>4.21</td>
<td>28</td>
</tr>
<tr>
<td>7</td>
<td>267,968</td>
<td>19.8</td>
<td>206 (297)</td>
<td>139</td>
<td>450</td>
<td>28</td>
<td>147</td>
<td>272</td>
<td>31</td>
<td>186</td>
<td>8.27</td>
<td>9.56</td>
<td>66</td>
</tr>
<tr>
<td>8</td>
<td>282,518</td>
<td>21.4</td>
<td>209 (299)</td>
<td>125</td>
<td>277</td>
<td>11</td>
<td>80</td>
<td>188</td>
<td>9</td>
<td>191</td>
<td>4.84</td>
<td>4.51</td>
<td>23</td>
</tr>
<tr>
<td>9</td>
<td>344,478</td>
<td>22.3</td>
<td>277 (366)</td>
<td>153</td>
<td>287</td>
<td>7</td>
<td>89</td>
<td>187</td>
<td>11</td>
<td>237</td>
<td>3.98</td>
<td>4.85</td>
<td>23</td>
</tr>
<tr>
<td>10</td>
<td>326,926</td>
<td>19.3</td>
<td>263 (405)</td>
<td>143</td>
<td>268</td>
<td>6</td>
<td>80</td>
<td>179</td>
<td>9</td>
<td>180</td>
<td>4.41</td>
<td>4.30</td>
<td>27</td>
</tr>
<tr>
<td>11</td>
<td>387,591</td>
<td>19.0</td>
<td>311 (516)</td>
<td>167</td>
<td>365</td>
<td>7</td>
<td>97</td>
<td>244</td>
<td>24</td>
<td>239</td>
<td>4.28</td>
<td>4.21</td>
<td>47</td>
</tr>
<tr>
<td>12</td>
<td>457,373</td>
<td>16.6</td>
<td>367 (533)</td>
<td>190</td>
<td>343</td>
<td>8</td>
<td>114</td>
<td>213</td>
<td>16</td>
<td>253</td>
<td>3.38</td>
<td>3.72</td>
<td>30</td>
</tr>
<tr>
<td>13</td>
<td>539,419</td>
<td>19.2</td>
<td>469 (687)</td>
<td>221</td>
<td>407</td>
<td>10</td>
<td>89</td>
<td>293</td>
<td>25</td>
<td>270</td>
<td>3.43</td>
<td>4.11</td>
<td>36</td>
</tr>
<tr>
<td>14</td>
<td>510,063</td>
<td>15.5</td>
<td>468 (776)</td>
<td>253</td>
<td>420</td>
<td>7</td>
<td>133</td>
<td>279</td>
<td>8</td>
<td>244</td>
<td>4.12</td>
<td>4.33</td>
<td>35</td>
</tr>
<tr>
<td>Total</td>
<td>4,249,816</td>
<td>18.6</td>
<td>3,539 (5,401)</td>
<td>1,920</td>
<td>3,918</td>
<td>114</td>
<td>1,150</td>
<td>2,560</td>
<td>208</td>
<td>2,548</td>
<td>4.83</td>
<td>5.05</td>
<td>426</td>
</tr>
</tbody>
</table>

Chr., chromosome; sSNP, synonymous SNPs; nsSNP, non-synonymous SNPs; ncSNP, non-coding SNPs; MS, polymorphic microsatellites/indels; θ, average population mutation rate; π, average pairwise nucleotide diversity; rec. events, recombination events detected within each chromosome.

<sup>a</sup>Total bp sequenced for each of four isolates. <sup>b</sup>Numbers in parentheses are the number of predicted genes in 397.
Table 2-2. Summary of polymorphisms and substitutions in genes with significant McDonald-Kreitman $p$-values. $P_n$: number of nonsynonymous polymorphisms, $P_s$: number synonymous polymorphisms, $D_n$: number of nonsynonymous substitutions and $D_s$: number of synonymous substitutions; $N_c$: codon usage index.

<table>
<thead>
<tr>
<th>Gene</th>
<th>$P_n$</th>
<th>$P_s$</th>
<th>$D_n$</th>
<th>$D_s$</th>
<th>$p$-value</th>
<th>$N_c$</th>
</tr>
</thead>
<tbody>
<tr>
<td>PFL2395c</td>
<td>3</td>
<td>0</td>
<td>4.5</td>
<td>23.5</td>
<td>0.0078</td>
<td>40.28</td>
</tr>
<tr>
<td>PF14_0517</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.0095</td>
<td>34.48</td>
</tr>
<tr>
<td>PF08_0005</td>
<td>9</td>
<td>0</td>
<td>15.5</td>
<td>11.5</td>
<td>0.0184</td>
<td>36.51</td>
</tr>
<tr>
<td>PFE1150w</td>
<td>5</td>
<td>0</td>
<td>20.5</td>
<td>26.5</td>
<td>0.0226</td>
<td>33.56</td>
</tr>
<tr>
<td>PF07_0035</td>
<td>7</td>
<td>1</td>
<td>40.75</td>
<td>50.25</td>
<td>0.0279</td>
<td>39.57</td>
</tr>
<tr>
<td>MAL6P1.27</td>
<td>3</td>
<td>0</td>
<td>14</td>
<td>23</td>
<td>0.0302</td>
<td>37.05</td>
</tr>
</tbody>
</table>

Figure 2-1[1]. Physical maps showing distribution of polymorphic sites across chromosomes. Vertical bars represent SNPs (black, nonsynonymous substitutions; red, synonymous substitutions) or microsatellites (blue, under the horizontal lines). Only one nonsynonymous SNP and one synonymous SNP are presented if there was more than one SNP in a gene (noncoding SNPs were grouped with synonymous SNPs). Most of the chromosomal ends (green vertical bars) were excluded because of gene families such as *var*, *rifin* and *stevor*. 
Figure 2-2. Correlation between effective number of codons ($N_c$) and GC3s. GC3s is the fraction of codons, (of those that have a guanine or cytosine in the third position) that are synonymous at the third position.

Figure 2-3. Highly polymorphic genes grouped according to GO functional terms. Genes encoding antigens or proteins involved in invasion and adhesion are the most polymorphic groups.
Figure 2-4. Distribution of Tajima’s D across each chromosome for genes with at least one segregating site. Solid line is the genome-wide average and the dotted lines are the genome-wide average ± 1 standard deviation.
Figure 2-5. Distribution of $\alpha$ across each chromosome for genes that contain at least one synonymous polymorphism. Note: $\alpha$ values that are less than -1 have been set equal to -1.
Figure 2-6. Genes influenced by natural selection ($\alpha=1$) grouped by GO process terms. Genes involved in cell growth and maintenance and metabolism are the most abundant of the genes with known biological function.
Figure 2-7. Recombination matrices for each chromosome showing the minimum number of recombination events between each pair of SNPs. Chromosome positions are on the X and Y axes. Hot colors represent higher numbers of recombination events. The hottest colors are found in the sub-telomeric regions of each of the chromosomes.
CHAPTER 3

INFERRING THE DEMOGRAPHIC HISTORY OF *Plasmodium falciparum* USING POLYMORPHISM DATA
ABSTRACT

Demographic events often mimic the effects of natural selection on genetic variation making it necessary to take into account the demographic history of an organism when making inferences about selection. The accumulation of sequence data has sparked a debate over the history of the malaria parasite *Plasmodium falciparum* (see [1] for a review). We applied an approximate Bayesian computation approach to a chromosome-wide polymorphism data set for 99 isolates to investigate the demographic history of *P. falciparum*. Our approach allows us to explore a broad range of demographic models and to make inferences about several model parameters. This contrasts with other studies that have evaluated the Malaria’s Eve hypothesis based on only one or a few loci (e.g. [2, 3]). The introduction of agriculture and an increase in temperature and humidity in Africa 7,000 to 12,000 years ago resulted in an increase in the number of places for *Anopheles* mosquitoes, the major vector of *P. falciparum*, to breed[^4-6]. Hence, it is believed the expansion of malaria occurred within the last 10,000 years, coinciding with these events. However, our results support a model of global expansion and growth out of Africa more than 100,000 years ago.
INTRODUCTION

*Plasmodium falciparum* is the agent of the most lethal form of malaria, a disease that kills over 1 million people every year (WHO, 2005). As such, large efforts are being made to develop vaccines and new drugs to treat those affected by the disease. Identifying genes under selective constraints can lead to the identification of drug targets and loci that confer susceptibility/resistance to the disease. With the advent of new technologies that allow rapid, large-scale genotyping it has become a widespread practice to use population variation to make inferences about natural selection. Demographic events (e.g. population size change) often mimic the effects of selection on genetic variation making it difficult to distinguish between the two. It is therefore necessary to take into account the demographic history of an organism in order to make correct inferences about selection.

Microorganisms have the potential to be informative for making inferences about human demography, particularly migration events. *Helicobacter pylori*, a bacterium that colonizes the human gastrointestinal track, is believed to have a history that correlates with humans. High sequence diversity and increased recombination among strains of this bacteria provides more information for population genetic analysis[7] and can elucidate variability in closely related populations that human genetic variation cannot[8]. Differentiation among populations (as measured by $F_{st}$) for *H. pylori* is strongly correlated with human data and a decrease in diversity with distance from east Africa again agrees with observations in
humans\textsuperscript{[9]}. Humans are the main mode of long range migration of malaria, and therefore we should be able to draw information about human history from these parasites.

Historical documents indicate that malaria has been affecting humans for the last 10,000 years\textsuperscript{[4-6]}. Slash and burn agriculture and an increase in temperature and humidity in Africa 7,000 to 12,000 years ago resulted in an increase in the number of places for Anopheles mosquitoes, the major vector of \textit{P. falciparum}, to breed. The accumulation of sequence data has ignited a debate on the evolutionary history of extant worldwide populations of \textit{P. falciparum}. The “Malaria’s Eve” hypothesis proposed by Rich, et al\textsuperscript{[10]} suggests that a severe bottleneck occurred recently in the history of \textit{P. falciparum} and that current worldwide populations derived from only a few ancestral strains. This conclusion was reached after observing an absence of silent polymorphisms in 10 genes and discounting several other possible forces that would produce the same results (e.g. persistent low effective population size). Examining the DNA sequences of several genes encoding surface proteins Rich and Ayala\textsuperscript{[11]} again concluded that the absence of silent polymorphisms is a result of a demographic sweep (i.e. population bottleneck) and explained that recent dispersion across the world may be due to changes in human societies, genetic changes that altered compatibility of the host, parasite and vector, or changes in climate. Volkman, \textit{et al}\textsuperscript{[12]} re-sequenced 25 introns and again observed a lack of synonymous polymorphism and argued this was the result of a recent spread of the parasite from a single progenitor or of
constraints on codon usage. They estimated that the most recent common ancestor (MRCA) of *P. falciparum* lived between 3200 and 7700 years ago.

Conversely, other groups feel there is strong evidence against the Malaria’s Eve hypothesis. Hughes and Verra\[13\] surveyed variation at the CSP locus, which encodes peptides presented to T cells, to refute the findings of Rich, *et al*\[10\]. They estimated the number of synonymous substitutions between *P. falciparum* strains to be 0.015±0.011 and calculated the divergence of the CSP locus to be 2.1±1.5 million years ago (Mya). Using synonymous and non-coding SNPs from five isolates along the entire 3rd chromosome, Mu and colleagues\[14\] estimate that the time of divergence for *P. falciparum* strains is 122-170 thousand years ago (kya), predating the time frame of the Malaria’s Eve hypothesis. Finally, Joy, *et al*\[3\] concluded from mitochondria data that two migration events out of Africa occurred in the history of the species, the first taking place 50-100 kya.

We surveyed and analyzed chromosome-wide genetic variation data in four populations to explore the history of *P. falciparum* and to test the Malaria’s Eve hypothesis. We reject models with a constant population size and those that exclude migration between populations. Our results suggest the ancestral population size of the parasites is small and migration to other parts of the world likely began more than 100,000 years ago. This estimate predates the spread of agriculture, which resulted in an increase in the number of *Anopheles* mosquitoes (the major vector of *Plasmodium*) and subsequently in the prevalence of malaria among humans\[4-6\].
RESULTS

Summary Statistics and Power Analyses

To infer the demographic history of *P. falciparum* we obtained data for all non-redundant genes along chromosome 3 in 99 isolates from four worldwide populations (Figure 3-1). Population structure was assessed using the clustering algorithm implemented in Structure 2[^15], which assumes a model with *K* populations each of which is characterized by a set of allele frequencies. The partitioning of populations from allele frequencies generally follows geographical partitions (Figure 3-2)[^16]. We summarized the data with two estimators of the population mutation rate (*θ*=4*N*μ), *θπ* and *θW*, and Tajima’s *D*, a statistic that describes that allele frequency spectrum. We calculated the mean and variance of each statistic across genes for each population (Table 3-1a). We first examined four population models, essentially assuming that the four parasite populations have independent histories. Negative Tajima’s *D* values indicate an excess of rare variants, consistent with population growth, and the rejection of null models with a constant population size support this. Subsequently, an exponential growth model in which the ancestral population of size *N₀* begins to grow exponentially at time *t_onset* with rate *a*, such that *Nₜ* is equal to *N₀e⁻ᵃᵗₜₒⁿₛᵉᵗ*, where *N₀* is the current population size. A grid of growth rate and time parameters was explored for each population. The expected time to the most recent common ancestor (i.e. when all lineages coalesce into a single lineage) is 4*N*, and times associated with coalescent theory are therefore measured in terms of 4*N* generations. The ratio of the pre-growth
population size \((N_t)\) to the current population size \((N_0)\) is dependent on the product of \(\alpha\) and \(t_{\text{onset}}\), however, we found that the individual values of these parameters are more critical than their product.

We tested the power of our chosen summaries of the data (combined into a single statistic as described in [17]) to reject or distinguish between alternative models. For example, we tested whether we are able to reject constant population size models when growth has actually occurred in our simulations (Figure 3-3a). We assigned the average statistics calculated across 10,000 replicate simulations as our “observed” data and then tested our ability to reject other models. We did this for varying combinations of split time and growth rate parameters and were able to reject all models of a constant population size (Figure 3-3a) and to distinguish between alternative growth models (Figure 3-3b).

Most of the genes on chromosome 3 have only a few segregating sites, which might limit our power to detect the “true” model. Instead of simulating each gene independently, we simulated the entire chromosome as a single region but incorporating recombination inferred previously[16]. We found little difference in the power to distinguish between growth models when only a single region is simulated. We also examined the effects of two other factors on our ability to draw conclusions about demographic models: 1) the number of replicates simulated under each model, and 2) including/excluding recombination in the simulations. The likelihood surface generated from 10,000 replicate data sets and that generated from 250 replicate data sets are very similar. In addition, although there is little
difference between the likelihood of models inferred from data simulated with free recombination and data simulated with no recombination, the probability of a given model is higher for data generated with free recombination. Recombination generates independence, which increases statistical power; therefore it is intuitive that simulations with free recombination have higher probabilities.

Population History of Malaria

Population models reveal that the African population is growing at a faster rate than the other three populations and that it began expanding more than 10,000 years ago, predating the spread of agriculture and the increase in temperature and humidity in the region\(^{[4, 5, 18]}\) (Figure 3-4). Parameters values with the highest likelihood suggest that growth began approximately 9,000 to 19,000 years ago in Asia and America, assuming 2 to 4 generations per year, respectively. This also contradicts the findings of Joy, et al\(^{[3]}\) who inferred from mitochondrial DNA (mtDNA) that the population sizes of South America and Asia have remained constant over time. It is worth noting that the mtDNA shows no signs of recombining in Plasmodium\(^{[3]}\) and that any inferences made are effectively from a single gene.

In addition, we explored models with an instantaneous reduction in size (bottleneck) followed by immediate exponential growth for the African and Asian populations. We were unable to obtain an estimate for the severity of the bottleneck for Africa because all severities
were equally likely given any time. This is likely due to the fact that the extreme growth that has occurred in the African populations masks the signature of the severity of the bottleneck on variation. However, we can infer that a bottleneck likely occurred more than 0.1 \((4N)\) generations ago. We were able to resolve the parameters for Asia and found that the model parameters with the highest likelihood suggest that a bottleneck, which resulted in a ~75% reduction in population size, occurred 0.2 \((4N)\) generations ago, followed by exponential growth to its current size.

Global Expansions of Malaria

In an attempt to infer when the global expansion of \textit{P. falciparum} began, we evaluated global models in which all populations derived from a single population with an African origin. We were able to reject all models of no population growth and no migration between populations. Models of constant population size were unable to capture the observed allele frequency spectrum (as summarized by Tajima’s \(D\)), while models with no migration between populations had larger population differentiation than observed (Figure 3-6, \(4Nm=0\)) (i.e. \(Fst\) is too large).

We tested several models of expansion out of Africa that include population growth within each population and migration between pairs of populations (Figure 3-5). Model 1 contains a single dispersal event when \textit{P. falciparum} migrated to Asia, America and PNG simultaneously, while the other two models (Model 2 and Model 3) contain three distinct
events whereby parasites left Africa and inhabited other parts of the world. All models included equal, reciprocal migration between pairs of populations. In a coalescent framework, \( F_{st} \) can be related to coalescence times in the following way: 

\[
F_{st} = \frac{\bar{t} - \bar{t}_0}{\bar{t}},
\]

where \( \bar{t} \) is the mean coalescence time for two genes from a collection of populations and \( \bar{t}_0 \) is the mean coalescence time for two genes from the same population\(^{[19]} \). Therefore, we initially used observed \( F_{st} \) values (Table 3-1b) as a proxy for the three split times in Model 2, and subsequent sets of split times tested for this model were kept proportional to these values.

We explored grids of split times and population migration rates, \( 4Nm \), (where \( m \) is the fraction of migrants found in a population) for the three models discussed above (Figure 3-6). Initially, the growth rates for all populations were derived from the population models, but they did not perform well when shared ancestry and migration were included in the model. The growth rates that performed the best across models were relatively high (95, 80, 42, and 95 for Africa, Asia, America, and PNG, respectively). Although we are unable to distinguish between the three models, we were able to reject very recent growth and migration out of Africa \( (T=0.1) \), contradicting the findings of a common ancestor within the last 10,000 years\(^{[10, 12, 20]} \) and very small rates of migration \( (4Nm \leq 5) \). A population migration rate of 15 is the maximum likelihood estimate for all three models and was treated as a constant in subsequent simulations. As can be seen in Figure 3-6, the order of founding events out of Africa (i.e. the order in which parasites began inhabiting other parts of the
world) seems to have a negligible effect on current genetic variation. Our results suggest the founding events began 0.3 to 0.7 (4N) generations ago. Given the population size model 
\[ N_t = N_0 e^{-at_{moo}} \]
and assuming a per base pair mutation rate of 1.2x10^{-9}, we obtain a very small estimate for the pre-expansion population size, consistent with the severe bottleneck proposed by the “Malaria’s Eve” hypothesis[^10] and supported by Volkman, et al[^12], but not with the timing.

**DISCUSSION**

It is necessary to consider the history of multiple populations jointly in order to model the effects of ancestry, which can mimic processes such as selection. In addition, our results suggest inferences about parameters describing demographic events (e.g. time) is influenced by the type and amount of data analyzed. We applied an approximate Bayesian approach to a chromosome-wide data set in an attempt to reconstruct the demographic history of four worldwide populations of *P. falciparum*. Population models confirm that the populations have been expanding and although we were unable to resolve the order of migration events out of Africa, we rejected models with constant population size and isolation after divergence (i.e. no migration). Our inability to infer the order of migration events out of Africa may have to do with the fact that malaria did not expand recently from Africa, and may have existed in America previous to the slave trade (in preparation). We conclude that the parasite
began to inhabit other parts of the world more than 100,000 years ago, predating the timing of the Malaria’s Eve hypothesis.

It is critical to take demographic events into account when making inferences about patterns of LD and natural selection. Association studies rely on patterns of LD, which are affected by changes in population size, migration and other demographic forces. For example, a population with a constant population size is expected to exhibit more LD than a population that has experienced growth\textsuperscript{[21]}. Therefore, populations that have recently expanded are more applicable for mapping single disease genes resulting from newer mutations whereas populations of constant size are more appropriate for mapping genes that are caused by older mutations\textsuperscript{[22]}. Both natural selection and changes in population size can leave similar patterns on genetic variation. In order to make inferences about the significance of statistics that are often used to detect selection it is important to account for demography.

**METHODS**

*Data set*

The empirical data used as the template for our simulations consisted of SNP data from 98 genes across chromosome 3 for four worldwide populations (Africa, South and Central America, Asia and Papua New Guinea (PNG)) of *P. falciparum*. This data was
kindly provided by collaborators at the National Institutes of Health, National Institute of Allergy and Infectious Disease (NIH/NIAID).

Empirical measures

Genetic variation in each of the 98 genes was summarized by two measures of the population mutation rate ($\theta = 4N_e \mu$), Watterson’s estimator, $\theta_W$, \cite{23} based on the number of segregating sites and $\theta_\pi$ \cite{24} based on pairwise diversity. We also calculated Tajima’s $D$ statistic \cite{24} for each population, which tests for departures from the neutral model. Genetic distances for all pairs of populations were estimated by $F_{st}$, calculated as follows:

$$F_{st} = \frac{\theta_T - \theta_s}{\theta_T},$$

where $\theta_T$ is the population mutation rate for both populations combined and $\theta_s$ is the average population mutation rate of each of the two populations. The mean and variance of these measures was calculated across genes producing a total of 30 summary measures of the data.

Estimates of individual population recombination rates, $\rho = 4N_e r$, for each gene were based on estimates from [ref]. Global $\rho$ values for each gene were estimated using a weighted average of the individual population values based on the sample size as follows:

$$\rho_{global} = \frac{\sum_i n_i \cdot \rho_{ij}}{\sum_i n_i},$$

where $n_i$ is the sample size for population $i$ and $\rho_{ij}$ is the population recombination rate of gene $j$ for population $i$. The current effective population sizes, $N_0$, for
each population and the global population were estimated from the average $\theta_w = 4N_0\mu$ across genes, assuming a mutation rate of $10^{-9}$ per base pair per generation.

**Simulations**

Simulations were performed using R. R. Hudson’s coalescent simulation package, ms\textsuperscript{25}. 10,000 replicate data sets consisting of independent samples for each of the 98 genes were generated for each combination of parameter values. Simulations were conditioned on empirical values of $\rho$ and some product of the empirical $\theta$ values. Given that the effective population size of a population whose actual size has fluctuated is essentially a harmonic mean of size over time\textsuperscript{26} and given that the population size of *P. falciparum* has not remained constant over time, it is necessary to adjust input $\theta$ values for the simulations.

We implemented an exponential growth model in which the ancestral population of size $N_0$ begins to grow exponentially at time $t_{onset}$ with rate $\alpha$, such that $N_f$ is equal to $N_0 e^{-\alpha t_{onset}}$, where $N_0$ is the current population size. All times are measured in terms of $4N$ generations. Simulations were performed for each population, independently and with all four populations sharing a common ancestor. For the latter simulations (the full model), $N_0$ is the current global effective population size, $\theta$ is the global population mutation rate, and $\rho$ is a weighted average of the population recombination rate (as discussed above). Equal, reciprocal migration was simulated between all pairs of populations with the exception of Africa and PNG, and America and PNG. No migration was simulated between these two
pairs of populations in order to keep population differentiation consistent with true observations. Including migration from Africa and America to PNG caused $F_{st}$ estimates to become too large.

**Sampling from the posterior distribution**

For individual population models, the parameters we are interested in estimating are $t_{onset}$, the time when growth began, and $\alpha$, the growth rate. We choose a value for each parameter independently from the prior distributions and simulate samples as discussed above. We use a discrete uniform prior distribution for both parameters where $t_{onset} \in (0.1, 0.2, \ldots, 0.9, 1.0)$ and $\alpha \in (0, 5, 10, 20, 30, 40, 50)$. For each sample we calculate the summary statistics and determine the fit of the parameters using the method discussed below.

The same methodology was implemented for the global models. The parameters of interest for the global models are $T_1$, $T_2$, and $T_3$, the timing of the migration events out of Africa, and $m_{ij}$, the fraction of migrants from population $i$ that make up population $j$. Again, we use a discrete uniform prior distribution for each parameter where $T_1 \in (0.1, 0.2, \ldots, 0.7)$, $m_{ij} \in (0, 5, 10, \ldots, 30)$, and the range of $T_2$ and $T_3$ depend on the value chosen for $T_1$. 
Parameter calculation and fit

We used the method described by Voight, *et al* \(^{[17]}\) to determine if data generated under the simulated parameters are consistent with the empirical data. This approach allows the inclusion of multiple aspects of the data by estimating *p-values* for each of the summary statistics included (see *Empirical measures*) and combining those *p-values* into a single test statistic, $C$. The method was executed as follows:

1. Summary statistics are calculated for each replicate data set and the empirical data
2. A one-tailed *p-value* is calculated for each statistic, as the number of data sets with a value as or more extreme.
3. One tailed *p-values* are converted to two-tailed *p-values*
4. A $C$ value is calculated for each data set using the following formula: $C = \sum \ln p_i$
5. A one-tailed *p-value* is estimated as the number of simulated data sets with a $C$ values as or more extreme than the empirical $C$ value.

This method is essentially a likelihood approach, where the final one-tailed *p-value* represents the probability of the observed data (represented by a set of summary statistics) given a particular set of model parameters.
REFERENCES


Table 3-1. Summary statistics across Chromosome 3 in 98 isolates of *P. falciparum*. (a) Average and variance of $\theta_w$, $\theta_\pi$ and Tajima’s $D$ across 99 genes. (b) Matrix of pairwise $F_{st}$

(a)

<table>
<thead>
<tr>
<th></th>
<th>$\theta_w$</th>
<th>Var $\theta_w$</th>
<th>$\theta_\pi$</th>
<th>Var $\theta_\pi$</th>
<th>Taj $D$</th>
<th>Var Taj $D$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Africa</td>
<td>0.00132</td>
<td>1.27e-6</td>
<td>0.00083</td>
<td>1.02e-6</td>
<td>-0.7732</td>
<td>0.5744</td>
</tr>
<tr>
<td>Asia</td>
<td>0.00092</td>
<td>1.02e-6</td>
<td>0.00078</td>
<td>1.12e-6</td>
<td>-0.3815</td>
<td>0.9598</td>
</tr>
<tr>
<td>America</td>
<td>0.00091</td>
<td>8.87e-7</td>
<td>0.00077</td>
<td>7.10e-7</td>
<td>-0.2751</td>
<td>0.8716</td>
</tr>
<tr>
<td>PNG</td>
<td>0.00084</td>
<td>1.47e-6</td>
<td>0.00072</td>
<td>1.48e-6</td>
<td>-0.5199</td>
<td>0.8612</td>
</tr>
</tbody>
</table>

(b)

<table>
<thead>
<tr>
<th></th>
<th>Africa</th>
<th>Asia</th>
<th>America</th>
</tr>
</thead>
<tbody>
<tr>
<td>Asia</td>
<td>0.207</td>
<td></td>
<td></td>
</tr>
<tr>
<td>America</td>
<td>0.191</td>
<td>0.196</td>
<td></td>
</tr>
<tr>
<td>PNG</td>
<td>0.231</td>
<td>0.116</td>
<td>0.445</td>
</tr>
</tbody>
</table>
Figure 3-1. Sampling locations in Africa, Asia, Central and South America, and Papua New Guinea.

Figure 3-2. Inferred population structure of global parasite isolates. Population partitions using SNPs from Chromosome 3. Each vertical bar represents an individual isolate.
Figure 3-3. Power analyses. (a) Power to reject a constant size model when the population has actually experienced growth using all 30 statistics or only the average and variance of $\theta_x$. (b) Power to reject alternative growth models. The x-axis is the time and growth rate parameters. As before, all times are in $4N_e$ generations.
Figure 3-4. Likelihood surface for population models. The x-axis is the time since the growth began (measured in $4N_e$ generations), and the y-axis is the rate at which the population is growing ($\alpha$). Colors represent the probability of the model given the empirical data. We calculated $N_e$ for each population assuming a mutation rate of $1.2 \times 10^{-9}$. $N_{\text{Africa}} = 275,000$; $N_{\text{Asia}} = 191,700$; $N_{\text{America}} = 189,600$; $N_{\text{PNG}} = 175,000$. 
Figure 3-5. Schematic of global models explored. Arrows represent migration between populations. (a) Model 1, (b) Model 2, (c) Model 3
Figure 3-6. Likelihood surface for combinations of model parameters. The x-axis is the time, in $4N_e$ generations, of the first migration out of Africa, where $N_e$ is the global effective population size. The y-axis is the population migration rate, $4N_em$, where $m$ is the fraction of the population made up of migrants. Colors represent the probability of the model given the empirical data (as summarized by statistics).
CHAPTER 4

THE EVOLUTION OF POPULATION RECOMBINATION LANDSCAPES: IMPLICATIONS FOR DEMOGRAPHIC HISTORY AND ASSOCIATION STUDIES

Kate McGee and Philip Awadalla

Ste-Justine Hospital Research Centre, Faculty of Medicine, University of Montreal, Montreal, Quebec Canada
ABSTRACT

Recombination acts to create new combinations of alleles and to break up existing associations, and therefore plays an important role in shaping genetic variation and patterns of haplotype structure or linkage disequilibrium. Variation in recombination rates across genomic regions has been well characterized, particularly in model organisms including humans, but the stochastic and deterministic forces that shape rate variation across space and time remains the subject of active research. Discrepancies in hotspot number and location between humans and chimpanzees begin to suggest that recombination rates evolve more rapidly than DNA sequence. We studied how recombination rates evolve and investigated the effects of sample size, total recombination and time since a population split on inferences of recombination rate variation. Empirical observations made for human populations appear to share similar patterns of recombination; however, our simulations show that only a short amount of time need transpire before we expect little concordance in both the total amount of recombination and the pattern of recombination events among even closely related populations. Using patterns and overall rates of recombination is powerful for inferring population divergence times and, unlike other summary statistics, is robust to ascertainment bias. We used a simple method to infer population divergence times for human and malaria populations with an approximate Bayesian computation (ABC) procedure. We found three genes with a recombination landscape that is more conserved among African and European populations than is expected by chance. In addition, correlation in recombination landscapes of pairs of populations suggest that three human populations diverged from a common
ancestral population approximately 10,000 years ago, earlier than previous studies supporting
the “Out of Africa” hypothesis and pairs of non-African populations of malaria parasites
diverged 18,550 to 37,000 years ago.
INTRODUCTION

Recombination is a key factor affecting the evolution of DNA sequences. It influences genomic variation by creating new combinations of alleles and by breaking up existing allelic associations. Recently developed methods to estimate recombination rates based on patterns of linkage disequilibrium (LD) have made it practical to examine the fine-scale structure of recombination rates across an entire genome. Recombination events in humans and other organisms, such as yeast, are clustered together in regions known as recombination hotspots\cite{1}. These hotspots have recombination rates 10- to 100-fold greater than the background rate and it is estimated that more than 50% of recombination events occur in less than 10% of the sequence of the human genome\cite{1}. While the variation in recombination rates across genomic regions has been well characterized, the stochastic and deterministic forces that shape rate variation across time remains the subject of active research.

LD-based methods are often used to infer sex-averaged ancestral recombination rates from large samples of individuals, and experimental methods such as sperm typing estimate current rates of recombination in usually a smaller number of males. Several explanations have been proposed to explain the inconsistencies in the presence and/or magnitude of recombination hotspots between sperm-typing estimates and rates inferred from patterns of LD\cite{2-4} including the rapid or recent loss/gain of hotspots, sex-specific recombination events, demographic history, or natural selection\cite{5,6}. Characterizing the factors associated with how recombination patterns evolve over time is important in understanding the mechanisms that
underlie the recombination process, in developing new algorithms to more accurately estimate recombination rates from population genetic data and in the design of LD-based association studies.

While large-scale patterns of LD and recombination over the entire genome appear to be highly conserved\cite{7}, differences in fine-scale patterns at specific loci have been found within and among human populations\cite{3,8}. An attempt to identify subsets of common genetic variants that can be used to “tag” haplotypes based on patterns of LD has been made for use in genome-wide association studies. A recent study showed that a set of tag SNPs derived from continental populations, such as those in the HapMap Project, may not be adequate for isolated populations; such populations (e.g. founder populations) are of particular utility in mapping both simple and complex traits\cite{9}. In addition, de Bakker et al surveyed >7500 common SNPs across the MHC region and found clear differences in association with HLA types and SNPs among four populations\cite{10}. The disparities observed in these studies may reflect differences in demographic history among populations, or in fine-scale recombination rates, both of which influence patterns of LD. The ability to reproduce an association in multiple populations (i.e. the same SNP is associated with the same phenotype) is a gold standard in the field\cite{11}. If the extent of LD varies among populations, as several studies have shown (e.g. \cite{12,13}), then varying markers will be necessary to capture variation in different populations so that associations can be detected.

One goal of this study is to use patterns of recombination rate variation observed in simulated data to learn how recombination evolves over time and then use this information to
potentially make demographic inferences. Genetic differences among populations are important in understanding the evolutionary forces that shape variation and, as such, many measures have been developed to quantify these differences. The classic measure of genetic differentiation among populations, Wright’s $F_{st}$ [14], exploits allele frequencies to determine the proportion of total genetic variation that can be attributed to differences among populations. Another estimator that uses allele frequencies was introduced by Nicholson, et al [15] and measures the amount of genetic drift among individual populations relative to an inferred ancestral population. A variance or “drift” parameter, $c_j$, is estimated for each population, $j$, allowing inferences to be made about rates of genetic drift for each population as well as differences among populations. In addition, Hudson [16] developed a “nearest-neighbor” statistic, $S_{nn}$, which utilizes haplotype data rather than just individual SNP frequencies to estimate genetic differentiation among populations. Genetic differentiation provides insight into the amount of time populations have been separated and the rate of migration between populations.

The production of large polymorphism data sets has fueled the investigation of historical demographic events (e.g. changes in population size) of species and the estimation of critical parameter values associated with these events. Summaries of genetic variation data can be used to estimate $N_e$ (see [17] for a review), but different summaries provide information at varying points in the history of a population. For example, estimates based on the number of segregating sites reflect an average over a large time scale, whereas estimates from LD can provide information about population size changes during recent history [18].
Since population size has not remained constant in a number of organisms, such as humans and malaria, estimates of effective population size will vary across methods of inference. Several recent studies make use of a combination of summary statistics (e.g. Tajima’s $D$, $F_{st}$, $\theta$, etc.) to obtain estimates of multiple model parameters\cite{19, 20}. Wakeley and Hey (1997) introduced a set of statistics that describe the distribution of polymorphisms among populations (i.e. SNPs polymorphic in both populations, polymorphic in only one population, and fixed between populations), derived the expected values of each of these statistics, and solved for the time and population size parameters associated with an isolation model\cite{21}. Becquet and Przeworski also applied these summaries in a Markov chain Monte Carlo (MCMC) approach to estimate the parameters of the isolation-migration model for two closely related species\cite{22}. Studies have shown that these statistics capture much information about the parameters of the isolation and the isolation-migration model\cite{21, 23, 24}. Finally, Hey and Nielsen (2007) introduced a method that estimates the entire posterior density of the model parameters using MCMC simulations to integrate over genealogies\cite{25}. The expected time until a shared polymorphism is lost is 1.7$N$ generations\cite{24} therefore methods based on the sharing of alleles among populations, like those above, are limited in application to older population divergence times. Schaffner, et al used a large set of measures including two measures of LD ($r^2$ and $D'$) to create simulated data that matched observed data in three human populations. Although recombination is implicitly accounted for when using a measure of LD to infer model parameters, ours is the first study where the population
recombination rate and pattern of rate variation has been utilized as an explicit summary and we show that it is useful for making inferences of recent times.

Several studies exploring recombination hotspots in humans and chimpanzees have shown that despite greater than 95% sequence identity there is little concordance in the number, intensity, and location of hotspots in the two species \cite{26-28}. Given that primate populations share common histories spanning different timescales (among human populations or human and chimp ancestors), these observations beg straightforward questions: How long does the signature of ancestral recombination persist and what factors affect our ability to detect these ancient events? Is this informative with respect to population demographic inferences? To explore these questions we generated simulated data under a range of parameters and examined the effects of sample size, mutation rate, recombination map length and time of divergence of two populations on inferences of population recombination rates and locations, recombination heterogeneity, and correlations in recombination maps among populations. Our results suggest that sample size has a substantial impact on recombination rate inference and variation, and that mutation rates has greater impacts on inter-population correlations in recombination landscapes than the amount of time a pair of populations have diverged from each other. We also find that the recombination landscape is more informative with respect to demographic inferences among related pairs of populations than statistics based on the distribution of polymorphisms among populations when applied to genotyping data (e.g. HapMap data) or other chip-based methods.
The “Out of Africa” hypothesis of human origins postulates that all modern humans originate from Africa with migration to other parts of the world starting approximately 60,000 years ago. Correlation in recombination landscapes of human populations suggests that derivation from a single ancestral population occurred approximately 10,000 years ago, earlier than previously estimated. For malaria, the demographic history of \textit{P. falciparum} populations is under serious debate. The Malaria’s Eve hypothesis proposed by Rich, \textit{et al} \cite{29} and supported by others \cite{30, 31} contends that a severe bottleneck occurred recently in the history of \textit{P. falciparum} and that current worldwide populations derived recently from only a few ancestral strains, while others suggest sequence data supports a more ancient origin \cite{32-34}. Our results suggest global expansion from a common ancestral population for the three non-African populations 18,550 to 37,000 years ago.

**RESULTS/DISCUSSION**

*Persistence of recombination signals and the effects of drift*

A simple illustrative exercise for demonstrating the effects of random genetic drift on the signal of recombination is a sampling scheme where we simulate an ancestral population of size \(N\) and sample (with replacement) an equal number of sequences to form the first generation of a new population and continue to sample to form subsequent generations. We estimated the minimum number of recombination events \cite{35} needed to explain the SNP data (see Methods) to determine the proportion of ancestral recombination events detected in each generation. Not surprisingly, a higher rate of ancestral recombination results in the ability to detect a larger proportion of ancestral events in any given generation (Figure 4-1a).
However, differences in the number of events detected across regions with different recombination rates is greatly diminished when the mutation rate is increased ($\theta=0.005$) (Figure 4-1b). The number of events detected in a given generation converges for all rates after approximately 800 generations of repeated sampling (genetic drift). The signal of approximately 80 percent of the ancient recombination events is lost after 1000 generations suggesting that the majority of recombination events we observe in real populations (e.g. [1, 36]) will have occurred recently. The impact of drift on detecting ancestral recombination events will be increased in populations that fluctuate in size (e.g. [37, 38]).

Recombination landscape

Two groups observed that both the total recombination rate and location of recombination events in a region is not correlated among humans and chimpanzees [27, 28]. To address how variation in recombination rates along chromosomal segments changes over time, we compared population recombination rates ($\rho = 4N_e r$) of two isolated populations derived from a single ancestral population (see Methods). Shared ancestry increases the genetic similarities among populations; therefore comparing two populations derived from the same population provides information about the role of ancestral recombination events in patterns observed among populations. In diploid organisms, the expected time to the most recent common ancestor (MRCA) is $4N_e$. Therefore, measures of time associated are given in terms of $4N_e$ generations. Simulations of the coalescent were generated to explore different parameters, including split times between pairs of populations and values of the population

84
mutation rate ($\theta=4N\mu$) and $\rho$. Rate variation was inferred using the reverse jump MCMC (RJMCMC) method described in [39] and correlations (Pearson’s product moment correlation coefficient, $R$) were evaluated using estimates of $\rho$ for each interval between pairs of adjacent SNPs for each population.

As expected, the average correlation in the landscape of recombination decreases as the time since the populations split from a common ancestor ($T_{split}$) increases (Figure 4-2). However, even among independent populations ($T_{split}=1.0$), nearly eight percent of replicates produce landscapes that are highly correlated (Pearson’s $R \geq 0.9$) by chance (for $\rho = 4$ and $\theta = 0.001$ per base). The impact of $\theta$ is substantial (Figure 4-3) and as $\theta$ increases, the number of replicates with high correlation decreases rapidly. For example, with a recent split ($T_{split}=0.001$ and $\theta=0.001$), more than 30% of replicate data sets are highly correlated. In contrast, less than 20% exhibit the same correlation when $\theta=0.005$, and a further reduction (~6%) is observed when $\theta=0.01$. This is likely due to the number of informative sites from which to infer recombination. Fewer sites from which to estimate recombination rates results in the loss of resolution of rate variation. With a greater density of SNPs, there are more marker intervals among which recombination events can occur. In contrast, with fewer SNPs, regions appear more constant and similar among populations. For example, we generated a series of datasets with 50, 100 or 500 SNPs in a 10kb region. For each density we simulated replicate data sets with a single recombination hotspot (see Methods for details). We detected a recombination event in approximately 70-, 79-, and 90-percent of data sets simulated under
a single hotspot model with 50, 100, and 500 SNPs, respectively. The number of SNPs has a sizeable impact on the detection of recombination events.

The above observations were inferred using all polymorphic sites in either population; however, ancestral polymorphisms are often lost due to drift and may no longer segregate in both populations. This affects the number of shared marker intervals and decrease correlations between populations. However, when we compared recombination rates estimated only with sites segregating in both populations, it does not appear that shared versus all SNPs has a substantial impact on similarities in total map length or recombination landscape ($r^2$ and $R$ in Table 4-1, respectively) regardless of how long ago the split occurred. There is a small decrease in the average difference between the map lengths of two populations when only shared polymorphisms are utilized. This difference is greatest when $T_{split}$ is large (Table 4-1), and significant for $T_{split} = 0.01$ and 0.5 (Wilcoxon rank test: $p = 0.025$ and $p < 0.002$, respectively). Furthermore, the number of replicate pairs with significant correlations ($p < 0.05$) decreases with $T_{split}$ much more rapidly when only shared polymorphisms are used when $\theta$ is very large (0.08).

Finally, we evaluated the effects of sample size on overall map length and on differences in recombination landscapes between pairs of related populations. Correlations in total map length (Pearson’s $R$), for $n=20$, decreases rapidly with $T_{split}$ and is virtually lost when populations diverged longer than $0.1*4N_e$ generations ago across all values of $\theta$. When we sample $n=100$ chromosomes from each of the populations the correlation in map lengths between population pairs increases substantially when $T_{split}$ is recent (Figure 4-4a). The
correlation in recombination rate variation also significantly increases when a larger sample size is used for $T_{\text{split}}<0.1$ (Wilcoxon rank sum test, $p=3.25e-9$). It is clear that a sufficient number of chromosomes is required to obtain a true picture of variation in recombination rates, concordant with previous findings (e.g. [1]), particularly if comparisons between populations are to be made in order to make historical inferences.

**Recombination Rate Variation and Inferences of Population Demography**

Shared polymorphism among populations may occur either if the polymorphism was segregating in the ancestral population or if parallel mutations occurred after they diverged. The former is the most likely explanation for large genomes with low mutation rates (infinite sites model) [24] among populations that have recently diverged. Therefore, the number of shared polymorphisms segregating in two populations contains information about population divergence times and migration. We analyzed polymorphism data for several populations of the malaria parasite, *P. falciparum* [40], and humans [10, 41] to investigate when pairs of populations shared a common ancestor. Given what we have learned about recombination, it is possible to use this information to make inferences of population relatedness for populations that are recently diverged.

**Human Recombination Hotspots: Conservation and Demography**

We estimated recombination rates across three distinct, 3Mb regions on chromosome 1 (chosen randomly) with data obtained from the HapMap Project and looked at correlations
between all pairs of populations (Figure 4-5). We combined two summaries, correlation in recombination rates and the proportion of polymorphisms shared between pairs of populations, into a single statistic as described in [20] in an attempt to infer an estimate of the time since the populations diverged (see Methods). Results from our power analysis reveal that a combination of these two summaries gives us high power to distinguish between models (Figure 4-6). The ascertainment scheme of the Human HapMap Project[^42], which focuses on common alleles that are more likely to be shared among populations, makes drawing inferences from shared polymorphisms alone inappropriate - more SNPs are shared than expected in nature. Therefore, we also downloaded completely re-sequenced data for over 190 genes from the Seattle SNPs Project[^41] (see below).

Our simulations showed that we do not lose information about shared recombination histories between populations when only SNPs shared among populations are used in the inference of recombination rate variation and correlation. If we use only the correlation in recombination rates along HapMap regions, we inferred that human populations diverged from a single population no more than 0.01 \( (4N_e) \) generations ago or approximately 10,000 years if we assume an effective population size of 10,000 (e.g.[43]) and a generation time of 25 years (Figure 4-7).

Of the 193 genes examined from the Seattle SNPs Project, the proportion of shared polymorphisms relative to total among European- and African-American populations ranges from 0.05 to 0.7 (mean=0.39, SD=0.121). We were able to calculate the correlation in recombination landscape among these two populations for 86 genes having heterogeneous
recombination rates (the remaining genes showed no rate variation) (Figure 4-8). The distribution of the correlation coefficients across the 86 genes (Figure 4-9) is significantly different from that observed for simulations when $T_{\text{split}}$ is 0.001, 0.1, and 0.5 $4N$ generations (two-tailed Wilcoxon rank sum test, $p = 6.93\times10^{-7}, 7.10\times10^{-8},$ and $4.92\times10^{-13},$ respectively), but not when $T_{\text{split}}=0.01$ (two-tailed Wilcoxon rank sum test, $p = 0.316$). This split time suggests a divergence time of 0.04 $N$ generations (or approximately 10,000 years) ago for populations of European and African ancestry, supporting the inferences made from the HapMap data.

In addition, the correlation coefficient of one gene, IL22R, falls outside of the 97.5th percentile of the distributions of correlation coefficients for the simulated data (of $T_{\text{split}} = 0.01$) suggesting that the recombination landscape at this gene is perhaps more conserved than expected by chance. This high level of conservation could reflect the presence of a recombination hotspot in those genes. The factors influencing the persistence and location of recombination hotspots is largely unknown. Identifying regions with highly correlated recombination landscapes among populations may help determine the characteristics of a true hotspot.

Recombination Rate Variation in Neutral vs. Selected Regions

To compare the extent of shared polymorphism and recombination in a “neutral” region versus that of a region under selection we obtained polymorphism data for the 7.5Mb extended MHC locus analyzed in [10]. The probability of having a shared polymorphism at a neutral locus is very low when the time since two populations diverged is large [44], In
contrast, balancing selection acts to preserve genetic variation and therefore extend the life of a polymorphism\[^{45}\], making it more likely to observe shared polymorphism among populations at these loci. For example, the proportion of shared polymorphisms across the MHC locus which is known to be under balancing selection\[^{46}\], is high (90\%) when compared to the genes from the Seattle SNPs Project, where the highest proportion of SNPs shared among populations of European and African descent is 0.7. Takahata (1990) showed that long-standing balancing selection effectively lengthens local genealogies providing more time for recombination to occur\[^{47}\], thereby allowing local recombination patterns to evolve in different populations. In contrast, we found significant correlation in recombination landscapes among all pairs of African and non-African populations (Table 4-2, Figure 4-10). This clearly suggests that combinations of alleles at multiple loci are being maintained (LD) and is evidence for mutations acting epistatically.

**Malaria Recombination Hotspots and Demographic Inferences**

There are conflicting estimates for the age of worldwide *Plasmodium* populations (reviewed in [48]). The “Malaria’s Eve” hypothesis proposed by Rich, et al\[^{29}\] and supported by several other groups (e.g. [30, 31]) suggests that a severe bottleneck occurred recently (within the last 10,000 years) in the history of *P. falciparum* and that current global populations derived from only a few ancestral strains. Other groups present evidence against this hypothesis and support a more ancient origin (at least 50,000 years ago) for current populations\[^{32-34}\]. We repeated the above analysis with data obtained from re-sequencing of
all coding genes on the third chromosome in four populations (Africa, Asia, America and PNG) of *P. falciparum*. The average proportion of shared polymorphisms for all pairs of populations is 0.298, with the largest proportion being shared by Asia and PNG (0.351) and the smallest proportion shared by Africa and PNG (0.225). We combined the fraction of shared polymorphisms relative to the total and the correlation in recombination landscapes (Figure 4-11) into a single statistic and found that all pairs of non-African populations are derived from a single ancestral population approximately 0.1 \((4N_e)\) generations ago. This corresponds to approximately 18,550 to 37,000 years ago if we assume 2 to 4 generations per year and a per base pair mutation rate of \(1.2 \times 10^{-9}\). This estimate is only for non-African populations and is therefore earlier than the divergence time inferred in Chapter 3 for all four populations. We were unable to make an inference about pairs of African and non-African populations because none of the simulations capture the empirical patterns of allele- and recombination-sharing. Population recombination rates are very high, particularly for African parasites, and may exceed the limits of composite likelihood method implemented. To obtain better resolution we also estimated the minimum number of recombination events between pairs of sites. Correlation in recombination landscapes is high for all pairs of populations using these estimates indicative of continuous global expansion and migration perhaps from Africa.
CONCLUSIONS

We used coalescent simulations to examine the persistence of ancestral recombination and how changes in various parameters affect the relationship of recombination patterns among populations. We found that both the recombination landscape and the total amount of recombination along a chromosome evolves rapidly such that the correlation of these two measures in two populations of shared ancestry is nearly nonexistent after only a short time of divergence. The total amount of recombination in the region does not appear to have a significant impact on the correlation (Figure 4-4b), while the mutation rate greatly influences the similarity in our estimates of rate variation. These results suggest that recombination rates evolve quickly and inferences are influenced by many factors. Knowing what forces influence recombination rates and how they change over time will aid in the design of association studies, particularly in choosing appropriate marker regions, and may also be informative with respect to divergence times of closely related populations.

Previous methods for inferring demographic model parameters\cite{17,21,22} such as \( N_e \) and divergence times were limited in terms of parameter space and by their choice of summary statistics, which restricts estimation to either recent or ancient population parameters. Using recombination rate patterns overcomes SNP ascertainment bias on inferences of \( N_e \). We have shown that correlation in rates among populations has high power to make inferences of recent and ancient time parameters.
METHODS

Simulations of populations with shared histories

Simulated data for a 10kb region were generated using R.R. Hudson’s coalescent simulation program \( ms^{[49]} \). Two populations with a common ancestor were simulated under a model of constant population size, where the size of each of the derived populations is equal to that of the ancestral population. The population mutation rate \( \theta = 4N_e \mu \) and the time since the populations split from a common ancestor \( T_{\text{split}} \) were allowed to vary. Data were simulated for \( \theta = 0.001, 0.005, 0.01 \) and 0.08 and \( T_{\text{split}} = 0.001, 0.01, 0.1, 0.5, 0.9, \) and 1.0, where time is measured in \( 4N_e \) generations. (Note: \( T_{\text{split}} = 1.0 \) is essentially simulating two independent populations) No migration was allowed between the populations after the split. At least 500 replicate data sets were generated for each combination of parameters. The expected population recombination rate \( \rho = 4N_e r \) was equal to 4 across most simulations. This value was chosen because as \( \rho \) becomes larger, the composite likelihood surface flattens, making it impossible to determine the rate.

We performed a second set of simulations for a larger region. We simulated 1000 replicate simulations for a 500 kb region with a recombination length of 2500, which was derived from several regions of the human genome. Data was again simulated for two populations with a common ancestor at \( T = 0.001, 0.01, 0.1, 0.5 \) or 0.9 \( (4N_e) \) generations ago and recombination rates were estimated using LDHat with a penalty parameter of 20 (see below).
Evaluating SNP density and recombination hotspots

Marker density impacts the ability to detect recombination events. If the coverage differs in two populations it might impact our analyses of the correlation of the rates. Varying SNP densities were simulated by specifying three different numbers of segregating sites in a region (50, 100 or 500). Data with recombination hotspots were simulated using SNPsim\textsuperscript{[50]}. A single hotspot was simulated in a random location within a 10kb region. The background recombination rate over the entire region was set to zero and the hotspot recombination rate was 0.0012 per site based on the upper bound of estimates at the DNA3 hotspot found in the MHC locus\textsuperscript{[51]}. 100 chromosomes were sampled in each of 1000 replicate simulations for each of the three SNP densities. One thousand data sets were also generated under a model with no hotspots for each SNP density. These “null” data sets were used to assess the expected amount of recombination due to noise from the estimation process rather than actual recombination activity. The minimum number of recombination events needed to explain the data were estimated using the software RecMin\textsuperscript{[35]} (see below for details).

Estimating Minimum Number of Recombination Events

We estimated the minimum number of recombination events in the history of our sample using the program RecMin, which calculates the statistics ($R_h$ and $R_s$) described in \textsuperscript{[35]}. $R_h$ estimates the number of events based on the difference in the number of haplotypes observed and the number of segregating sites, while $R_s$ is based on approximating the history of the data by minimizing the number of recombination events.
Recombination rate landscape

We used the program LDHat\textsuperscript{[1, 39]} to estimate population recombination rates for each of the simulated data sets. Briefly, estimates of $4N_e \theta$ are calculated using an extension of Hudson’s composite likelihood approach\textsuperscript{[52]}. A reversible jump Markov Chain Monte Carlo (RJMCMC) approach is employed to compute the population recombination rate for the interval between each possible pair of markers, thereby allowing for variable recombination rates across the region of interest. For a complete description of the method implemented by LDHat see [1, 39].

A penalty parameter is introduced when calculating variable recombination rates to avoid over-fitting of the pairwise rates. In essence, large penalty parameters will result in the loss of detail, but will only identify positions where there is a large amount of evidence for a change in recombination rate. Conversely, a small penalty parameter will elucidate fine-scale recombination patterns, but it will also introduce noise [LDHat manual]. A block penalty parameter of 10 was used for the initial simulated data sets, a penalty of 20 was used for the data obtained for human HapMap and malaria parasite populations as well as the second set of simulated data, and a block penalty of 200 was used for the extended MHC region. The Markov chain was updated 1,000,000 times and estimates were sampled every 2000 iterations. The first 200,000 iterations were considered the burn-in period. Recombination rates were estimated using all polymorphic sites and only sites segregating in both populations.
**Correlations**

Correlations in recombination landscapes of two derived populations are summarized by Pearson’s product-moment coefficient, calculated using the statistical software package R (http://www.r-project.org/). Simulations with a constant recombination rate across the region were excluded from this analysis as correlation cannot be assessed. When using only shared polymorphisms to estimate recombination rates, only those data sets with 10 or more polymorphic sites were examined for correlation in recombination rate variation.

**Persistence of a signal of recombination**

To examine how genetic drift affects the ability to detect ancestral recombination events we implemented a simple sampling scheme using both backward and forward simulation methods. We simulated 100 replicate ancestral populations of size 10,000 and 100,000 using the coalescent\(^{[49]}\) for all combinations of \(\theta = 0.001, 0.005\) and \(\rho = 1, 4, 10\). We assumed a uniform distribution, where recombination was equally likely between any pairs of sites. We sampled 10,000 or 100,000 chromosomes with replacement from the ancestral population to form the first generation of a new, isolated population. To mimic the effect of genetic drift, we continued to sample from the previous generation for 1000 generations.
Estimating time since divergence

We applied an approximate Bayesian computation (ABC) procedure to determine the most likely timing of a common ancestor between a pair of populations. This method is essentially a likelihood approach so that

\[ L(R|T_{split} = t_{split}) \propto P(T_{split} = t_{split}|R), \]  

(1)

where \( R \) is the correlation in recombination landscapes of two populations and \( T_{split} \) is the time of divergence of two populations. To calculate the likelihood we simulated 1000 genealogies for each model and define

\[ P(T_{split} = t_{split}|R) = \frac{1}{1000} \sum_{g=1}^{1000} P_g, \]

(2)

where \( P_g \) equals 1 if \( |R_g| \geq |R_{obs}| \) and 0 otherwise. We simulated data under a model with two populations that split at a given time in the past (\( T_{split} \)). We generated approximately 1000 replicate data sets for a 500kb region for each of five split times (\( T_{split} = 0.001, 0.01, 0.1, 0.5, 0.9 \)), where time is measured in \( 4N \) generations. The population mutation rate (\( \theta = 0.001 \)) and the per base pair population recombination rate (\( \rho = 0.005 \)) was constant across all simulations.

We employed the technique described by Voight, et al.\textsuperscript{[20]} to obtain a single summary statistic, \( C \), for the *P. falciparum* data that combines the correlation in recombination landscapes and the proportion of polymorphisms shared among pairs of populations. We then applied the ABC approach discussed above to determine the maximum likelihood
estimator of $T_{split}$ for each pair of populations. Including multiple summary statistics slightly alters the above equations. Equation (1) becomes

$$L(R, P_s| T_{split} = t_{split}) \propto P(T_{split} = t_{split}| R, P_s),$$

where $P_s$ is the proportion of shared polymorphisms and Equation (2) becomes

$$P(T_{split} = t_{split}| R, P_s) = \frac{1}{1000} \sum_{g=1}^{1000} P_g,$$

where $P_g$ equals 1 if $|C_g| \geq |C_{obs}|$ and 0 otherwise.

Polymorphism data

We obtained SNP data from the HapMap Project for three, 3Mb regions across the human genome for samples from Africa, Asia and Europe. Region 1 spans 1-3Mb, region 2 spans 100-103Mb and region 3 spans 205-208 Mb. When calculating the recombination rates we only included SNPs that are segregating in all three populations.

Polymorphism data for chromosome three sampled in four worldwide populations of P. falciparum (Africa, Asia, America, and PNG) was obtained from collaborators at the National Institutes of Health, National Institute of Allergy and Infectious Disease (NIH/NIAID)
REFERENCES


Table 4-1. Summary of correlation in recombination rates among two simulated populations when all polymorphisms are used to estimate $\rho$ across region and only sites that are segregating in both populations are used to estimate $\rho$. $\theta=0.005$. $R$: average correlation of landscapes across replicate simulations; $|\rho_1-\rho_2|$: average difference between map lengths of two populations across replicate simulations; $r^2$: average correlation in map lengths of two populations across replicate simulations

| $T_{\text{split}}$ | All $R$ | Shared $R$ | All $r^2$ | Shared $r^2$ | All $|\rho_1-\rho_2|$ | Shared $|\rho_1-\rho_2|$ |
|-------------------|--------|------------|-----------|--------------|----------------|----------------|
| 0.001             | 0.592  | 0.602      | 0.504     | 0.514        | 1.43           | 1.32           |
| 0.01              | 0.381  | 0.379      | 0.206     | 0.224        | 2.07           | 1.80           |
| 0.1               | 0.040  | 0.066      | 0.001     | 0.005        | 3.10           | 2.97           |
| 0.5               | 0.013  | 0.045      | 0.003     | 0.086        | 3.61           | 2.62           |
Table 4-2. Proportion of shared polymorphisms and correlation in recombination landscapes for pairs of (a) human populations across the extended MHC region (b) human populations across three 3Mb regions on chromosome 1 (average across the four regions) and (c) *P. falciparum* populations. Top: correlation (Pearson’s $R$) in recombination landscapes for pairs of populations, Bottom: Fraction of SNPs (relative to the total) shared among pairs of populations.

(a)

<table>
<thead>
<tr>
<th>Proportion of shared SNPs</th>
<th>CEU</th>
<th>YRI</th>
<th>CHB</th>
<th>JPT</th>
</tr>
</thead>
<tbody>
<tr>
<td>CEU</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>YRI</td>
<td>0.9379</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CHB</td>
<td>0.9402</td>
<td>0.9417</td>
<td></td>
<td></td>
</tr>
<tr>
<td>JPT</td>
<td>0.9416</td>
<td>0.9434</td>
<td>0.9977</td>
<td></td>
</tr>
</tbody>
</table>

(b)

<table>
<thead>
<tr>
<th>Proportion of shared SNPs</th>
<th>CEU</th>
<th>YRI</th>
<th>CHB+JPT</th>
</tr>
</thead>
<tbody>
<tr>
<td>CEU</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>YRI</td>
<td>0.951</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CHB+JPT</td>
<td>0.960</td>
<td>0.960</td>
<td></td>
</tr>
</tbody>
</table>

(c)

<table>
<thead>
<tr>
<th>Proportion of shared SNPs</th>
<th>Africa</th>
<th>Asia</th>
<th>S Amer</th>
<th>PNG</th>
</tr>
</thead>
<tbody>
<tr>
<td>Africa</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Asia</td>
<td>0.3064</td>
<td></td>
<td>0.534</td>
<td>-0.877</td>
</tr>
<tr>
<td>S Amer</td>
<td>0.3132</td>
<td>0.2939</td>
<td></td>
<td>-0.447</td>
</tr>
<tr>
<td>PNG</td>
<td>0.2251</td>
<td>0.3508</td>
<td>0.2989</td>
<td></td>
</tr>
</tbody>
</table>
Figure 4-1. The persistence of ancestral recombination events over time. Data points represent the average proportion of ancestral recombination events detected in a given generation across 50 simulations for a population of size 10,000 and a population mutation rate of (a) $\theta=0.001$ (b) $\theta=0.005$.

Figure 4-2. Distribution of correlation in recombination landscapes between pairs of populations that diverge at different time intervals. Each point represents the mean correlation in the recombination landscapes of two populations across 500 simulations and bars are the standard error across simulations. $\theta=0.001$. 

106
Figure 4-3. Distribution of correlation in population recombination rates across values of $\theta$. Correlation in recombination rates in a 10kb region for two populations who split 0.001 ($4N_e$) generations ago. (a) Frequency distribution of correlation coefficients and (b) Mean correlation coefficient with standard error bars for 1000 simulations of $\theta=0.001$, $\theta=0.005$, $\theta=0.01$. 
Figure 4-4. Effects of sample size and total recombination map length on correlations of map length among populations. Each point represents the total map length for a single simulation. (a) Correlation in total map length increases with a larger sample size. $T_{split}=0.001$ (b) Correlation in total map length is not affected by increasing the value of $\rho = 4N_r\theta$. $\theta = 0.005$; $T_{split} = 0.01$. 

\[ R^2 = 0.222 \]

\[ R^2 = 0.494 \]

\[ R^2 = 0.205 \]

\[ R^2 = 0.296 \]

\[ R^2 = 0.208 \]
Figure 4-5. Estimates of population recombination rate variation across three 3Mb regions using the RJMCMC with a jump penalty of 20 for HapMap individuals from Europe (red), Africa (blue), and Asia (green).
Figure 4–6. Power analysis. Analysis to determine how well the proportion of shared polymorphisms combined with the correlation in recombination landscapes among two populations can predict the time of divergence. The highest likelihood should fall on the diagonal where the simulated divergence time is equal to the actual divergence time.
Figure 4-7. **Probability of the split time** ($T_{\text{split}}$) **given the observed correlation in recombination landscapes.** $P$-values were calculated as the number of simulations with an equal or greater correlation among population recombination landscapes and converted to two-tailed $P$-values. Each line represents a different 3Mb region across chromosome 1 (HapMap data).
Figure 4-8. Distribution of correlation coefficients across 86 genes sequenced in the Seattle SNPs Project. Each point represents the correlation in recombination landscape among two populations for a single gene.
Figure 4-9. Comparison of empirical and simulated correlation in recombination landscapes among two populations. Empirical distribution of correlations among European- and African American recombination for 86 genes. Simulated results for 500 replicates for $T_{split}=0.001, 0.01, 0.1, 0.5$. P-values represent significance of Wilcoxon rank test.
Figure 4-10. Distribution of recombination events along the first 4.5 Mb of the extended MHC region. (a) Estimates of the minimum number of recombination events between each pair of segregating sites scaled by the length of the sequence (b) Estimates of population recombination rate variation using the RJMCMC with a jump penalty of 50 for European (red), African (purple), Chinese (green) and Japanese (blue) individuals
Figure 4-11. The Distribution of Detectable Recombination Events on Chromosome 3 of *P. falciparum*. In (A) and (B), each panel shows, for two populations, a minimum number of recombination events (assuming an infinite-sites model) between each pair of segregating sites, scaled by physical distance to identify regions of high and low recombination. (A) African (upper) and American (lower) populations. (B) SE Asian (upper) and PNG (lower) populations. The color bar unit is recombination event/kb. (C) Estimates of population recombination rate variation for African (red line), SE Asian (blue), American (black), and PNG (purple) samples using the RJMCMC method with a jump penalty of five\(^{[40]}\).
CHAPTER 5

DETECTING SELECTION AND EXPLORING ASSOCIATIONS BETWEEN POLYMORPHISMS IN FUNCTIONALLY-RELATED GENES
**ABSTRACT**

Natural selection is potentially a major contributor shaping the nucleotide variation observed in the human genome. By applying measures of population differentiation and tests for deviations from the standard neutral model, we compared classes of mutations (silent and replacement) and subsets of functionally related versus non-related genes for SNP data from 128 genes sequenced in populations of African and European descent (the SeattleSNPs Project). Replacement polymorphisms exhibit a significantly higher rate of genetic drift from an inferred ancestral population relative to silent sites. Significant departures from neutrality were observed using Fay and Wu’s $H$ and the McDonald-Kreitman test suggesting that amino acid replacement mutations are under directional selection. Although several previous studies suggest that functionally related genes, such as those contained in the same biological pathway, exhibit significant linkage disequilibrium, our results suggest that they exhibit no more than expected by chance among a random group of genes.
INTRODUCTION

Characterizing evolutionary forces that contribute to the present landscape of genetic variation is a critical aspect of evolutionary genetics. Polymorphisms found in the human genome are key factors in understanding disease susceptibility and treatment response[1]. Natural selection is a major force contributing to genetic variation and investigating the process of local adaptations among populations will help identify mutations that have occurred in response to a change in environment or disease resistance[2]. However, the majority of diversity seen in humans occurs within populations[3] and estimates of genetic differentiation among human populations, such as $F_{st}$, are therefore small. Geographic differences in allele frequencies between populations are mainly the result of genetic drift[4], but it has been shown that genes under balancing selection exhibit a lower amount of diversity among populations than expected[5]. In contrast, it has been hypothesized that genes under positive selection, such as those responsible for population- or environment-specific phenotypes (e.g. ability to digest lactose, skin color), will have higher levels of diversity between populations than predicted from drift alone[6].

Selection is a locus-specific force and may cause differences in degrees of genetic differentiation at particular loci. For example, if a specific allele is favored in one population but not in another (i.e. population-specific directional selection), then the $F_{st}$ value for that locus may be higher than the $F_{st}$ value of a neutral locus[7]. Alternatively, balancing selection will reduce the amount of differentiation between populations, and therefore decrease estimates of $F_{st}$. In this way, the variance in $F_{st}$ estimates would be higher than the neutral case if selected loci are present in the sample[8]. Identifying genes that fall in the tails of the
genome-wide distribution of $F_{st}$ may help identify genes involved in population-specific
disease susceptibility or drug resistance.

Akey, et al\cite{Akey2009} examined estimates of population structure, as measured by $F_{st}$, for over
26,000 SNPs in three populations to detect genes that are outliers with respect to the genome.
They compared their empirical distribution of $F_{st}$ to a distribution simulated under a model
with no recombination to determine if their observed values were consistent with selective
neutrality. A significant difference between the overall empirical distribution of $F_{st}$ and the
simulated data was observed, suggesting the presence of selection. To identify the type of
selection, they split the SNPs into functional categories (coding, noncoding, and intronic) and
found the most significant difference in $F_{st}$ between coding and noncoding SNPs, which is
evidence of purifying selection. Hughes, et al\cite{Hughes2011} collected allele frequency data from 102
individuals from four populations in the United States at 1442 single nucleotide
polymorphism (SNP) loci. They observed a reduction in gene diversity at SNP loci
conferring amino acid changes, relative to sites at regulatory regions and intron-exon
boundaries, especially at those expected to disrupt protein structure. Sites causing a change
in an amino acid may be at low frequency in the population because they are mildly
deleterious, consistent with the observations of Akey, et al\cite{Akey2009}.

The extent and distribution of nonrandom allelic associations, or linkage
disequilibrium (LD), is greatly affected by recombination and is of critical importance in
disentangling interactions between multiple loci. Many evolutionary forces such as
demographic processes, migration, and selection can affect the amount of LD maintained
and/or detected in a sample. Recently, in mice, extensive LD was found between
functionally related genes, more so than was seen between genes with unrelated functions, suggesting that selection may promote the coinheritance of favorable alleles when genes interact epistatically\[10]\.

We assess the role of local adaptation on distinct classes of nucleotide sites (silent versus replacement) and genes (functionally related versus non-related) in humans with inter- and intra-specific data from 10,000 SNP loci. Standard and recently developed measures of population differentiation and departures from neutrality for African-American and European-American populations suggest that polymorphisms causing a change in amino acid are not selectively neutral. Specifically, a deficiency in nonsynonymous fixed differences between humans and chimpanzees and an increase in genetic differentiation at replacement sites supports directional selection at these sites. Results also suggest an increased rate of drift in European-Americans compared to African-American, which is especially apparent at silent sites and more genes exhibit signatures of selection in the European population than in the African population. We also tested the association of sites within functionally related genes, and, in contrast to a previous study\[10]\, an excess of linkage disequilibrium was not observed.

RESULTS AND DISCUSSION

Population differentiation among different classes of sites and genes

Analyses of population differentiation coupled with tests for departures from neutrality, were performed on SNPs categorized as silent or replacement. Silent sites include both intronic and synonymous polymorphisms. Genes were further classified as belonging to
a pathway or to a random set of genes. For many of our analyses, genes were partitioned into three groups: (1) genes found in the Protease-Activated Receptor (PAR) pathway, (2) genes found in the Clotting pathway (including three also found in the PAR pathway), and (3) genes not found in either pathway (see Methods). We will refer to the genes found in this last set as “non-pathway” genes.

We will observe differences in levels of population differentiation among replacement sites relative to silent sites if amino acid replacement sites are under selection. If we assume that silent mutations are selectively neutral then we can use patterns observed at these sites as a base line of what we should expect if there is no selection. Therefore, an increase or decrease in differentiation among populations (relative to neutral expectations) for other types of mutations will signal the action of selection. $F_{st}$ is a relative measure of population differentiation\textsuperscript{[11]} that estimates the proportion of genetic variation explained by differences among populations relative to the total amount of genetic variation, and can be used to test this theory. However, because $F_{st}$ is based on a ratio, differences in $F_{st}$ estimates for different classes of sites may be explained by a difference in within-population variation.

Nicholson, et al\textsuperscript{[14]} developed an absolute measure related to $F_{st}$ that infers the amount of genetic drift among populations relative to an inferred ancestral population. A variance parameter, $c_j$, is estimated for each population, $j$, allowing inferences to be made about rates of genetic drift for each population. In essence, this parameter estimates how far the allele frequencies of a population are from typical values. The distributions of $c_j$ are significantly different among the European and African populations (Figure 5-1 and Table 5-1b). The European population exhibits higher rates of drift relative to the African population, which is
not surprising given the known smaller effective population size\textsuperscript{[15-18]}. This effect is especially apparent when looking at $c_j$ for only silent sites. A significant difference in $c_j$ between the two classes of sites was also observed in each population (Wilcoxon rank test, $p<0.001$ for both populations). In contrast to the relative measure of population differentiation ($F_{st}$), the direct estimates ($c_j$) for replacement sites are larger relative to the silent sites for both populations, indicative of higher rates of genetic drift among replacement sites. The average difference between $c_j$ for silent and replacement sites is slightly larger in the African population than in the European population (0.253 and 0.203, respectively). Africa’s larger effective population size would increase the population selection rate ($4N_e s$) and amplify the effects of selection (as measured by the difference in $c_j$ among classes of sites) in the data.

\textit{Departures from the neutral model}

Using data from a closely related species, levels of polymorphism within species can be compared to substitutions between species. Under neutrality, the ratio of polymorphism to divergence should be constant across loci. McDonald and Kreitman\textsuperscript{[22]} (MK) developed a test to determine if the ratio of synonymous to nonsynonymous polymorphism is significantly different from the ratio of synonymous to nonsynonymous divergence within a locus. We first looked at African and European populations jointly and found 21 genes that show signatures of selection. When looking at each population separately, 20 genes are significant in the European-American population while only 13 genes are significant in the African-American population, 11 of which are significant in both (\textit{Figure 5-2}). In addition,
we observed a slightly larger proportion (~20%) of pathway genes with significant results compared to non-pathway genes (~16%). We also performed the MK test on groups of pathway genes (Clotting, PAR) under the assumption that these genes do not evolve independently of each other. The genes in the PAR pathway exhibit a significant departure from neutrality, while those in the Clotting pathway are borderline significant ($p<0.001$ and $p = 0.057$, respectively) due to a reduction in nonsynonymous fixed differences between humans and chimpanzees, similar to that observed in [23], suggesting negative selection on new mutations in these genes.

Fay and Wu’s $H^{[19]}$ statistic uses inter- and intra-species data to test for departures from selective neutrality. It exploits two estimates of the population mutation rate ($\theta_w$ and $\theta_H$) and unlike Tajima’s $D$ it takes into account the state (ancestral or derived) of an allele. Demographic events can leave signatures similar to those of natural selection on genetic variation. To minimize the effects of demography, coalescent simulations were performed for each gene under the standard neutral model and under the demographic model derived by Hudson and Adams$^{[20]}$ (see Methods for details). Although a larger number of genes have significant $H$ values in the African population under both models (Table 5-2), values (normalized by the number of segregating sites) are correlated between the two populations ($R^2=0.298$, Figure 5-3). However, it is important to note that for a number of genes in the African sample $H$ statistics are significant under the population expansion model, but not under the neutral model suggesting that, in concordance with Hudson and Adams$^{[20]}$, this model is not sufficient to explain the observed allele frequencies.
Linkage disequilibrium among functionally related loci

To further investigate the relationship between functionally related loci we tested whether a set of genes within a biological pathway evolve in a non-independent manner by assessing whether functionally related genes have more non-random associations among alleles relative to a set of unrelated genes. Zapata, et al[24] surveyed polymorphism at 15 protein-coding loci in *Drosophila melanogaster* and found that gametic disequilibrium was two times more frequent among loci with related function relative to loci with no functional relationship. They concluded that an epistatic relationship between functionally related loci was responsible for this observation. More recently, Petkov and colleagues[10] tested whether inbreeding during recombinant inbred line formation favors preexisting allelic combinations in linkage disequilibrium domains. If this were true, the number of lines showing recombination across these domains would be fewer than expected. There was little difference in recombination between LD domains and non-domain regions prior to inbreeding. After inbreeding there were strikingly fewer inbred lines with recombination across LD domains, but this was not seen in non-domain regions. They concluded that selection against unfavorable combinations of alleles is a major force generating LD.

We estimated levels of LD for the genes in the Clotting and PAR pathways (Figure 5-4) for the European-American sample only to avoid an artificial increase in levels of LD due to underlying population structure. Also, to eliminate detecting LD simply due to close physical proximity, we performed comparisons only between genes located on different chromosomes. The genes in the PAR pathway are spread on six chromosomes and those in the Clotting pathway are dispersed on 14 different chromosomes. We found 2735 (0.5%)
and 140 (0.4%) significant inter-locus comparisons for unlinked genes in the Clotting and PAR pathways, respectively (Table 5-3), which is no higher than levels seen in simulated data. To further investigate the evolution of functionally related genes we subsequently gathered HapMap data for the 14 genes in the MHC-1 pathway (Figure 5-5). We estimated composite LD for all possible comparisons between genes and found that less than 1% of all comparisons are significant, a large proportion of which are between SNPs located on the same chromosome (Table 5-5). It is possible that just being involved in the same pathway is not enough to increase associations between sites, but that being directly connected in a pathway is. We hypothesize there will be a greater amount of LD between pairs of sites in genes that are adjacent in the pathway. Surprisingly, we found little difference between the two analyses (Table 5-4) suggesting that the ordering of genes in a pathway is inconsequential. We also looked at LD for SNPs falling less than 5 Mb upstream of each gene in the pathway. There are three genes (PDIA3, HLA-A, and LTA) where at least 1% of upstream comparisons were significant in all four populations suggesting the presence of cis-regulating elements upstream of these genes. Despite their physical linkage (HLA-A and LTA are located in close proximity on chromosome 6 and much of the upstream regions overlap), only a small percentage of SNPs are in significant LD with both HLA-A and LTA suggesting that they are being regulated by different elements.

CONCLUSIONS

In contrast to demographic forces, natural selection acts on specific loci rather than the genome as a whole. This allows us to perform comparative analyses on multiple loci to
determine if the effects of selection are present. Using population differentiation statistics and basic tests of neutrality, we found evidence for natural selection in our data, specifically directional selection. This is best supported by our analysis of $c_j$, which produced significant differences between silent and replacement sites, as well as population differences.

**METHODS**

*SNP data*

Polymorphism data was obtained from the SeattleSNP Project\[^{25}\]. The data was collected from 24 African American individuals and 23 individuals of European descent (CEPH). For our analysis, we considered an insertion/deletion or a tri-allelic site as missing information. After eliminating these polymorphisms, a total of 14,213 single nucleotide polymorphisms (SNPs) found in 127 genes were used in our analyses (516 replacement sites and 13,697 silent sites).

These genes include the genes that make up the Clotting and PAR pathways. When performing analyses on the pathway genes we considered only 20 of the 21 genes found in the Clotting pathway. The omitted gene, PROC, has no replacement sites, and was therefore eliminated from analysis.

The NCBI accession numbers for the homologous chimpanzee sequences were obtained from the SeattleSNPs website. We aligned each human sequence with its homologous chimpanzee sequence using ClustalW\[^{26}\]. We then used the chimpanzee sequence to infer the ancestral state.
Population Differentiation Statistics

Estimates of $F_{st}$ were calculated using the population genetics software package Genepop (http://wbiomed.curtin.edu.au/genepop/). Genepop calculates $F_{st}$ values for each site as well as a single value for each gene based on a weighted analysis of variance. Given that negative $F_{st}$ has no biological significance all such values were set to zero for the purpose of our analyses. $F_{st}$ values were calculated separately for silent and replacement sites.

The data were analyzed in two ways: per gene and per site. The per-gene estimates were taken directly from the single $F_{st}$ value given by Genepop for each gene. The per-site values were simply calculated as the arithmetic mean of the $F_{st}$ values for every site in a gene. Variable numbers of sites in each gene will have no effect on our results, as the mean $F_{st}$ value for each gene was not correlated with the number of sites analyzed within that gene.

In addition to $F_{st}$ we calculated another measure of genetic drift, using a beta-binomial model. The model estimates a variance parameter, $c_j$, which measures how far the allele frequencies in each population deviate from the allele frequencies of an inferred ancestral population (see [14] for details). This measure is analogous to $F_{st}$, but estimates of a parameter value, $c_j$, for each population, $j$, rather than a single pairwise value. $c_j$ parameters were estimated using the R statistical software package as in Marchini, et al. [27]

A set of Monte Carlo randomizations was executed based on the number of genes or the number of sites in each pathway to test for significant differences between the average $F_{st}$ values for pathway and non-pathway genes. 10,000 randomizations were performed.
**Standard Tests of Selection**

Fay and Wu’s $H^{[19]}$ is the difference between two estimators of $\theta = 4N_e\mu$, the population mutation rate. Unlike estimators like Tajima’s $D^{[28]}$, $H$ considers the state (ancestral or derived) of the allele. The first estimate, is $\theta_\pi$, which measures average pairwise diversity and the second is

$$\theta_H = \frac{\sum 2Si^2}{n(n-1)}$$

where $Si$ is the number of derived alleles found $i$ times. PPARG was not included in this analysis because we were unable to obtain a high quality alignment between the human and chimpanzee sequences from which to determine ancestry.

HKA, a program developed by Jody Hey$^{[29]}$, was used to perform the Hudson-Kreitman-Aguade (HKA) test$^{[21]}$ on our data. The Fisher’s Exact Test function in the statistical software package R was used to assess significance for the McDonald-Kreitman (MK) test$^{[22]}$.

We manually filtered the chimpanzee-human alignments for some of the genes by eliminating starting and terminal regions with low similarity.

**Simulations**

We performed 10,000 simulations of the coalescent using Hudson’s ms program$^{[30]}$ to test for significance of Fay and Wu’s $H$. We employed two demographic models for each population. The first is the standard neutral model, which assumes no recombination,
random mating and constant population size. Next, we applied a separate model for each population’s unique demographic history based on the parameters described in [20]. For the European population we simulated a bottleneck followed by an exponential expansion. The bottleneck is estimated to have occurred 37,000 \((t = .0375 \, 4N_e \text{ generations})\) years ago and resulted in a decrease in population size to 15% of the ancestral population. Expansion to the present results in a current population size twice that of the ancestral population. The African model consisted of a much slower expansion starting 200,000 years ago and results in a current population size of 1.9 times that of the ancestral population.

**Linkage disequilibrium**

With only genotype information it is difficult or impossible to decipher between the two double heterozygotes \(AB/ab\) and \(Ab/aB\). To circumvent this problem we used the composite linkage disequilibrium method described by Weir \([31]\) to calculate LD in our data. This method takes advantage of the fact that, while the individual digenic frequencies of the heterozygotes are unknown, the sum for the frequencies is directly observable

\[
\begin{align*}
    p_{AB} + p_{A/B} &= 2p_{AB}^A + p_{AB}^A + \frac{1}{2}(p_{Ab}^A + p_{aB}^A)
\end{align*}
\]

The composite linkage disequilibrium coefficient

\[
\Delta_{AB} = p_{AB} + p_{A/B} - 2p_Ap_B
\]

is standardized by the allele frequencies giving the following \(\chi^2\)-distributed test statistic \([32]\)

\[
\chi^2 = \frac{n\Delta_{AB}}{p_A(1-p_A)p_B(1-p_B)}
\]

129
Pairs of sites with a test statistic greater than 10.83 (critical value for $p = 0.001$, d.f. = 1) were identified as exhibiting significant LD. We performed the LD analysis only on the CEPH population surveyed to avoid artificial LD caused by population structure, likely to be present in the admixed African-American sample. Sites with a minor allele frequency less than 0.05 were eliminated from the analysis and comparisons were made only between sites on distinct chromosomes.

Two simulation methods were used to determine the amount of LD expected to occur between a random set of sites throughout the genome. First, Hudson’s coalescent simulation program, *ms*[^30^], was used to simulate a number of sites across a chromosome with essentially free recombination ($4N_{e}r=100$). This allows each polymorphism to be independent and therefore can be used to mimic patterns seen between SNPs on different chromosomes. 100 replicate data sets were generated to produce approximately the same number of pairwise comparisons as in the empirical dataset for each pathway. An additional set of simulations was performed using the software package *cosi*[^33^]. An independent dataset was generated for each chromosome based on the number of segregating sites found in the empirical data. The independent data sets were then combined to mimic the SeattleSnps data and the composite LD statistic was computed for each pair of sites found on distinct chromosomes. 100 replicate data sets were simulated.
REFERENCES


23. BUSTAMANTE CD, F-AA, WILLIAMSON S, NIELSEN R, HUBISZ MT, GLANOWSKI S, TANENBAUM DM, WHITE TJ, SNINSKY JJ, HERNANDEZ RD, CIVELLO D, ADAMS MD,


29. HEY, J. *HKA*. http://lifesci.rutgers.edu/~heylab/HeylabSoftware.htm#HKA


Table 5-1. Summary of population differentiation. (a) $F_{st}$ and (b) $c_{ij}$. Values are the average estimates (per site or per gene) across all genes in a given category.

(a)

<table>
<thead>
<tr>
<th>Category</th>
<th>Silent (x10²)</th>
<th>Replacement (x10²)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non-pathway</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Per gene</td>
<td>9.88</td>
<td>5.84</td>
</tr>
<tr>
<td>Per site</td>
<td>5.27</td>
<td>4.36</td>
</tr>
<tr>
<td>PAR pathway</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Per gene</td>
<td>6.58</td>
<td>1.28</td>
</tr>
<tr>
<td>Per site</td>
<td>3.40</td>
<td>0.94</td>
</tr>
<tr>
<td>Clotting pathway</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Per gene</td>
<td>7.75</td>
<td>0.23</td>
</tr>
<tr>
<td>Per site</td>
<td>4.67</td>
<td>3.91</td>
</tr>
<tr>
<td>All</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Per gene</td>
<td>9.47</td>
<td>0.04</td>
</tr>
<tr>
<td>Per site</td>
<td>5.10</td>
<td>4.17</td>
</tr>
</tbody>
</table>

(b)

<table>
<thead>
<tr>
<th></th>
<th>AA</th>
<th>EA</th>
</tr>
</thead>
<tbody>
<tr>
<td>All</td>
<td>0.384</td>
<td>0.325</td>
</tr>
<tr>
<td>Silent</td>
<td>0.138</td>
<td>0.357</td>
</tr>
<tr>
<td>Replacement</td>
<td>0.290</td>
<td>0.595</td>
</tr>
</tbody>
</table>
Table 5-2. Genes with significant Fay and Wu’s $H$. Genes in this table have a significant $H$ under both a neutral model and under a growth model. (a) African-American population (b) European-American population

(a)

<table>
<thead>
<tr>
<th>Gene</th>
<th># Sites</th>
<th>$H$</th>
<th>neutral</th>
<th>growth</th>
</tr>
</thead>
<tbody>
<tr>
<td>alox15</td>
<td>85</td>
<td>-20.25</td>
<td>0.0252</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>c3</td>
<td>199</td>
<td>-28.54</td>
<td>0.0486</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>cyp4f2</td>
<td>139</td>
<td>-19.32</td>
<td>0.0435</td>
<td>0.0002</td>
</tr>
<tr>
<td>ephb6</td>
<td>94</td>
<td>-22.83</td>
<td>0.0022</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>f2</td>
<td>79</td>
<td>-23.6</td>
<td>0.0433</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>f2rl1</td>
<td>88</td>
<td>-23.84</td>
<td>0.043</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>f2rl3</td>
<td>38</td>
<td>-10.26</td>
<td>0.0294</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>f5</td>
<td>280</td>
<td>-65.32</td>
<td>0.0342</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>f8</td>
<td>20</td>
<td>-5.4</td>
<td>0.0293</td>
<td>0.0004</td>
</tr>
<tr>
<td>fgg</td>
<td>21</td>
<td>-5.29</td>
<td>0.027</td>
<td>0.0008</td>
</tr>
<tr>
<td>fgl2</td>
<td>17</td>
<td>-2.58</td>
<td>0.0465</td>
<td>0.006</td>
</tr>
<tr>
<td>habp2</td>
<td>262</td>
<td>-43.93</td>
<td>0.0337</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>il10ra</td>
<td>66</td>
<td>-10.04</td>
<td>0.0313</td>
<td>0.0003</td>
</tr>
<tr>
<td>il11ra</td>
<td>26</td>
<td>-7.74</td>
<td>0.0282</td>
<td>0.0001</td>
</tr>
<tr>
<td>il12b</td>
<td>38</td>
<td>-15.16</td>
<td>0.0056</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>il16</td>
<td>58</td>
<td>-11.17</td>
<td>0.0244</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>il1r2</td>
<td>166</td>
<td>-17.66</td>
<td>0.0381</td>
<td>0.0241</td>
</tr>
<tr>
<td>il24</td>
<td>36</td>
<td>-4.33</td>
<td>0.0167</td>
<td>0.0205</td>
</tr>
<tr>
<td>il4r</td>
<td>168</td>
<td>-26.04</td>
<td>0.0493</td>
<td>0.0001</td>
</tr>
<tr>
<td>il7r</td>
<td>112</td>
<td>-37.44</td>
<td>0.045</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>irak4</td>
<td>122</td>
<td>-17.87</td>
<td>0.0188</td>
<td>0.0001</td>
</tr>
<tr>
<td>jak3</td>
<td>92</td>
<td>-22.13</td>
<td>0.0153</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>kel</td>
<td>103</td>
<td>-20.35</td>
<td>0.001</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>nos3</td>
<td>86</td>
<td>-27.91</td>
<td>0.0472</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>pfc</td>
<td>23</td>
<td>-8.54</td>
<td>0.0011</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>plat</td>
<td>154</td>
<td>-22.37</td>
<td>0.033</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>plaur</td>
<td>150</td>
<td>-30.3</td>
<td>0.03</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>ppara</td>
<td>129</td>
<td>-33.39</td>
<td>0.0385</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>procr</td>
<td>16</td>
<td>-3.71</td>
<td>0.0435</td>
<td>0.0016</td>
</tr>
<tr>
<td>proz</td>
<td>77</td>
<td>-16.78</td>
<td>0.043</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>ptgs1</td>
<td>116</td>
<td>-18.53</td>
<td>0.0051</td>
<td>0.0001</td>
</tr>
</tbody>
</table>
Table 5-2 (continued)

(a)

<table>
<thead>
<tr>
<th>Gene</th>
<th># Sites</th>
<th>H</th>
<th>neutral</th>
<th>growth</th>
</tr>
</thead>
<tbody>
<tr>
<td>sele</td>
<td>78</td>
<td>-18.34</td>
<td>0.0257</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>sell</td>
<td>145</td>
<td>-27.22</td>
<td>0.0422</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>serpinc1</td>
<td>40</td>
<td>-12.15</td>
<td>0.0099</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>serpine1</td>
<td>73</td>
<td>-15.72</td>
<td>0.0204</td>
<td>0.0001</td>
</tr>
<tr>
<td>sftpa1</td>
<td>157</td>
<td>-21.96</td>
<td>0.0383</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>stat3</td>
<td>105</td>
<td>-14.43</td>
<td>0.027</td>
<td>0.0001</td>
</tr>
<tr>
<td>stat6</td>
<td>51</td>
<td>-9.46</td>
<td>0.0287</td>
<td>0.0002</td>
</tr>
<tr>
<td>thbd</td>
<td>18</td>
<td>-10.25</td>
<td>0.0134</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>tnfrsf1b</td>
<td>146</td>
<td>-32.13</td>
<td>0.0354</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>trpv5</td>
<td>157</td>
<td>-32.03</td>
<td>0.0019</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>tyk2</td>
<td>130</td>
<td>-32.6</td>
<td>0.0107</td>
<td>&lt;0.0001</td>
</tr>
</tbody>
</table>

(b)

<table>
<thead>
<tr>
<th>Gene</th>
<th># Sites</th>
<th>H</th>
<th>neutral</th>
<th>bn_growth</th>
</tr>
</thead>
<tbody>
<tr>
<td>aloxl5</td>
<td>60</td>
<td>-19.29</td>
<td>0.0436</td>
<td>0.0057</td>
</tr>
<tr>
<td>ephb6</td>
<td>45</td>
<td>-23.75</td>
<td>0.0445</td>
<td>0.0003</td>
</tr>
<tr>
<td>f13a1</td>
<td>41</td>
<td>-8.98</td>
<td>0.0441</td>
<td>0.0227</td>
</tr>
<tr>
<td>f2</td>
<td>53</td>
<td>-13.17</td>
<td>0.0253</td>
<td>0.0137</td>
</tr>
<tr>
<td>f2rl1</td>
<td>44</td>
<td>-11.53</td>
<td>0.0322</td>
<td>0.0127</td>
</tr>
<tr>
<td>f2rl3</td>
<td>25</td>
<td>-8.22</td>
<td>0.0436</td>
<td>0.0096</td>
</tr>
<tr>
<td>f5</td>
<td>193</td>
<td>-51.8</td>
<td>0.0442</td>
<td>0.0068</td>
</tr>
<tr>
<td>f8</td>
<td>13</td>
<td>-4.8</td>
<td>0.0499</td>
<td>0.0104</td>
</tr>
<tr>
<td>il11ra</td>
<td>11</td>
<td>-4.38</td>
<td>0.0371</td>
<td>0.0102</td>
</tr>
<tr>
<td>il12b</td>
<td>26</td>
<td>-12.81</td>
<td>0.0129</td>
<td>0.0016</td>
</tr>
<tr>
<td>il7r</td>
<td>70</td>
<td>-16.82</td>
<td>0.0202</td>
<td>0.013</td>
</tr>
<tr>
<td>jak3</td>
<td>53</td>
<td>-18.6</td>
<td>0.0471</td>
<td>0.004</td>
</tr>
<tr>
<td>nos3</td>
<td>47</td>
<td>-12.02</td>
<td>0.0205</td>
<td>0.0127</td>
</tr>
<tr>
<td>pfc</td>
<td>13</td>
<td>-10.23</td>
<td>0.0208</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>ppara</td>
<td>63</td>
<td>-16.4</td>
<td>0.0379</td>
<td>0.0114</td>
</tr>
<tr>
<td>sele</td>
<td>67</td>
<td>-19.26</td>
<td>0.0452</td>
<td>0.0074</td>
</tr>
<tr>
<td>serpinc1</td>
<td>26</td>
<td>-11.89</td>
<td>0.028</td>
<td>0.0024</td>
</tr>
<tr>
<td>sftp3b</td>
<td>17</td>
<td>-3.39</td>
<td>0.046</td>
<td>0.0406</td>
</tr>
<tr>
<td>thbd</td>
<td>11</td>
<td>-5.58</td>
<td>0.0047</td>
<td>0.0035</td>
</tr>
<tr>
<td>tnfrsf1b</td>
<td>106</td>
<td>-28.12</td>
<td>0.0459</td>
<td>0.0086</td>
</tr>
<tr>
<td>tyk2</td>
<td>82</td>
<td>-32.17</td>
<td>0.0426</td>
<td>0.0012</td>
</tr>
</tbody>
</table>

136
Table 5-3. Summary of LD results for the Clotting and PAR pathways.

<table>
<thead>
<tr>
<th></th>
<th>Clotting</th>
<th>Par</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Total</td>
<td>Signif</td>
</tr>
<tr>
<td>Actual ms sims</td>
<td>686632</td>
<td>0.004</td>
</tr>
<tr>
<td>Actual cosi sims</td>
<td>703023</td>
<td>0.0002-0.0052</td>
</tr>
</tbody>
</table>

Table 5-4. Summary of MHC-1 pathway LD results for all HapMap populations. (a) Pairwise comparisons for all genes in the pathway. (b) Pairwise comparisons for only genes that are adjacent to one another in the pathway.

(a)

<table>
<thead>
<tr>
<th></th>
<th>Total</th>
<th># signif</th>
<th>Prop</th>
<th>Total on same chr</th>
<th># signif on same chr</th>
<th>Prop</th>
</tr>
</thead>
<tbody>
<tr>
<td>YRI</td>
<td>8113</td>
<td>63</td>
<td>0.0078</td>
<td>878</td>
<td>16</td>
<td>0.2540</td>
</tr>
<tr>
<td>CEU</td>
<td>6750</td>
<td>47</td>
<td>0.0070</td>
<td>6</td>
<td>11</td>
<td>0.2340</td>
</tr>
<tr>
<td>CHB</td>
<td>4218</td>
<td>7</td>
<td>0.0017</td>
<td>515</td>
<td>0</td>
<td>0.0000</td>
</tr>
<tr>
<td>JPT</td>
<td>3500</td>
<td>9</td>
<td>0.0026</td>
<td>423</td>
<td>7</td>
<td>0.7780</td>
</tr>
</tbody>
</table>

(b)

<table>
<thead>
<tr>
<th></th>
<th>Total</th>
<th># signif</th>
<th>Prop</th>
<th>Total on same chr</th>
<th># signif on same chr</th>
<th>Prop</th>
</tr>
</thead>
<tbody>
<tr>
<td>YRI</td>
<td>1637</td>
<td>9</td>
<td>0.0055</td>
<td>260</td>
<td>1</td>
<td>0.1110</td>
</tr>
<tr>
<td>CEU</td>
<td>1383</td>
<td>10</td>
<td>0.0072</td>
<td>270</td>
<td>10</td>
<td>1.0000</td>
</tr>
<tr>
<td>CHB</td>
<td>778</td>
<td>0</td>
<td>NA</td>
<td>170</td>
<td>0</td>
<td>NA</td>
</tr>
<tr>
<td>JPT</td>
<td>704</td>
<td>2</td>
<td>0.0028</td>
<td>144</td>
<td>2</td>
<td>1.0000</td>
</tr>
</tbody>
</table>
Figure 5-1. Differences in $c_j$ estimates among populations.

Figure 5-2. Distribution of genes that show a significant departure from neutrality according to the McDonald-Kreitman test.
Figure 5-3. Correlation of $H/S$ among populations. Fay and Wu’s $H$ values for each gene are divided by the number of segregating sites in that gene to eliminate discrepancies due to differences in the number of polymorphisms.
Figure 5-4. Schematic of the (a) Clotting and (b) PAR pathways. These figures were obtained from the SeattleSNPs website.
Figure 5-5. Schematic of the MHC-1 pathway. (obtained from the KEGG pathway database)
CHAPTER 6

CONCLUSION
In an age where sequencing entire genomes and genotyping hundreds of thousands of SNPs is a reality, we are being inundated with genetic variation data. Evolutionary forces influence genetic variation in different ways. Both classic (e.g. Tajima’s $D$) and recently developed (e.g. MCMC) methods attempt to summarize this variation and apply it to make inferences about historical events. Chapters 2, 3, 4, and 5 present ways in which observed patterns of genetic variation in conjunction with simulated data can be used to make inferences about recombination, selection, and demography.

Simulations are a widely used and highly applicable approach in exploring the effects of evolutionary forces on sequence variation. They can help in making inferences of specific demographic parameters such as divergence times. With current computational efficiency it is possible to explore a large parameter space to find models that are most likely to produce the observed data. Evolutionary forces are not exclusive or unique in the way they imprint DNA sequences, and simulations can also be used to take into account one factor, while testing for the signature of another. For example, an increase in population size will skew the allele frequency spectrum toward higher numbers of alleles at low frequency. If we wanted to use Tajima’s $D$ to test for positive selection, which produces the same pattern, we could obtain a “null” distribution of $D$ values for data simulated under a growth model, and then look for empirical values that fall in the tails of the distribution. By accounting for the effects of an increase in population size, we are able to determine that natural selection is the most plausible cause of extreme $D$ values.
Identifying genes influenced by selective forces is an important preliminary step in finding genes underlying complex phenotypes (e.g. disease risk, drug response) and can be used to choose candidate genes for use in association studies. Association studies rely on linkage disequilibrium (LD) between a genetic marker and the causal variant. Recombination is a major force shaping patterns of LD along the genome and recombination patterns may vary across human populations. Therefore, markers genotyped for an association study in one population may not capture the same amount of variation, and therefore the same association, in another population. Exploring the behavior of recombination rates over time, the causes of recombination rate variation and patterns observed in natural populations will aid in the efficient design of association studies.

Malaria parasites have developed resistance to once commonly used anti-malarial drugs, rendering virtually useless. Association studies have been successful in indentifying several genes underlying this resistance and surveying genetic variation has led to the identification of antigens that were previously unknown. The catalog of polymorphisms in humans and other organisms is growing rapidly and future studies will involve large-scale genotyping of those polymorphisms in larger samples. Developing models and new methods to analyze this type of data is critical to correct interpretation. It is important to continue to characterize how evolutionary forces shape patterns of genetic variation and to incorporate these observations into analysis as they will be critical in addressing global health issues.