Genomic Analysis of Evolution in Plasmodium falciparum and Babesia microti

Contents

Abstract

Parasitic protozoan infections of the red blood cell are among the most widespread and devastating pathogens of vertebrates. In humans, two genera of pathogens cause disease: the *Plasmodia*, which cause malaria, and the *Babesia*, which cause babesiosis. In this thesis, we apply the tools of whole genome sequencing and evolutionary genetics to study factors contributing to the spread of these pathogens: in *P. falciparum*, the acquisition of multiple drug resistance, and in *B. microti* the development of azithromycin resistance and the population genetics of emergence.

In the first part of this work, we test whether an accelerated mutation rate predisposes to acquisition of drug resistance in *P. falciparum*. Epidemiologically, resistance tends to begin along the Thai-Cambodian border, and from there spreads to other parts of the world. Environmental conditions such as inadequate drug dosing likely facilitate drug resistance, but molecular evidence also suggests that parasites from the Thai-Cambodian border may harbor genetic traits that let them develop resistance to novel antimalarials at an elevated rate. Low-dose drug pressure has also been proposed to be mutagenic, since several antimalarial agents have known DNA binding properties and have been shown to impair DNA damage repair pathways in *P. falciparum*. To test these hypotheses, we directly assayed substitution rates in a parasite line from the Thai-Cambodian border and a South American isolate, with and without chloroquine pressure. Sampling parasite DNA over a total of 760 generations (~ 4.2 years), we identified 17 mutations, producing an estimate of the substitution rate at 1.065×10^{-9} substitutions per site, per generation. We find that chloroquine pressure does not alter the mutation rate. We further find that substitutions accrued at an approximately 3-fold rate in the lines from Southeast Asia, a result which trended toward but did not reach statistical significance ($p = 0.056$). We argue that this is insufficient by itself to account for the rapidly increased rate at which ARMD parasites acquire drug resistance. By sequencing intermediate timepoints, we also characterize the dynamics of allele substitution *in vitro*.

In the second part of this thesis, we characterize *Babesia microti* by sequencing clinical isolates and enzootic strains. Since the first case in 1969 [36], human babesiosis due *B. microti* has emerged as important infection in the Northeast USA [84]. In order to characterize natural selection, recent evolutionary history, and the genetic architecture of *Babesia microti* populations, we created a map of genetic diversity from clinical strains. We describe this map, and show that *B. microti* isolates from the Northeast USA possess a paucity of nucleotide diversity, consistent with very recent common ancestry of circulating strains. We describe how *B. microti* genomes display a predominance of rare alleles and a number of segregating sites in excess of pairwise nucleotide diversity, suggestive of a recent population expansion. Finally, we identify RPL4 as a candidate gene for azithromycin resistance based on a non-synonymous substitution that occurs in a highly conserved arginine in the azithromcyin binding region of the L4 component of the 50S ribosomal subunit in a patient with azithromycin failure.

Acknowledgments

A large number of people have contributed to this thesis. First and foremost, I would like to thank my supervisor, Dr. Pardis Sabeti, for her supervision, outstanding mentorship, and the opportunity to work on this project. It has been a great pleasure to work in her laboratory as an MD thesis student. I would also like to thank members of the Sabeti lab, who have been great colleagues. I have had the opportunity of working closely with Lisa Freimark, a research assistant in the lab. Lisa has provided invaluable technical assistance, particularly with parasite culture, DNA extraction, illumina library preparation, and has also been a pleasure to work with.

The *P. falciparum* work was done in collaboration with the Dr. Xin-zhuan Su, my former supervisor, at the National Institutes of Health. I would like to thank Dr. Su for sharing samples from his lab in a collaborative fashion that made this project possible, as well as for sharing his knowledge and enthusiasm for malaria research. Like Dr. Sabeti, his high standards for personal and professional conduct in research make his laboratory an inspiring place to work. I would also like to thank Dr. Xiaorong Feng and Dr. Richard Eastman, members of the Su laboratory, who performed much of the parasite culture and DNA extraction described in this thesis. Dr. Eastman deserves a special thank you for his good-natured willingness to answer questions, offer advice, and ship materials on numerous occasions.

Sequencing was performed at the Broad Institute, as a part of a consortium of malaria projects there. I would like to thank Dr. Daniel Neafsey and Kevin Galinsky for their assistance in getting the samples sequenced and mapped as a part of the Broad pipeline. I would also like to thank Dr. Dyann Wirth and Jon Herman, who have worked on closely a similar project which has run in parallel with this one, for helpful discussions and support during all phases of the project.

For the *Babesia* work, I worked closely with Dr. Eric Rosenberg, Sue Bazner, and Graham McGrath of the Clinical Microbiology lab at Massachusetts General Hospital,

as well as Dr. Heidi Goethert and Dr. Sam Telford at the Tufts School of Veterinary Medicine and Dr. Jeffrey Bailey and Alice Tran at the University of Massachusetts Medical School. They have all been fantastic, inspiring collaborators and I am thankful for the opportunity to work with them.

On several occasions, Avi Feller, a PhD student in the Harvard Department of Statistics, offered his thoughtful advice on some the statistical aspects of this work, for which he deserves a special thank you.

Finally, I would like to thank my parents, my sister, and my fiancee Amy, all of whom have offered endless support and encouragement throughout my training.

It has been a pleasure to work on this project with the help of all of these individuals, who have contributed in numerous ways to the quality of the final product. To reflect the contribution of multiple individuals, I use the pronoun 'we' in the remainder of the thesis. I specify in individual sections where tasks, such as parasite culture, were carried out by others and recognize in those sections the individual(s) that performed the work.

List of Abbreviations

ARMD - accelerated resistance to multiple drugs

BLAST - basic local alignment search tool

BER - base excision repair

CDC - Centers for Disease Control

DNA - deoxyribonucleic acid

GATK - genome analysis toolkit

IC15 - 15th percentile maximal inhibitory concentration

IGV - integrative genomics viewer

KAHRP - knob-associated histidine rich protein

MA - Massachusetts

MGH - Massachusetts General Hospital

MMR - mismatch repair

NER - nucleotide excision repair

NH - New Hampshire

NIH - National Institutes of Health

PCA - principal component analysis

RNA - ribonucleic acid

RBC - red blood cell

SNP - single nucleotide polymorphism

SP - sulfadoxine-pyramethamine

1 Introduction

Parasitic infection of the red cells are among the most widespread and oldest infections in vertebrates. Of the many parasites in existence, there are two that infect humans: malaria, caused by parasites of the genus *Plasmodia* and babesiosis, caused by the *Babesia spp.* Malaria is the much better known and more important, causing over two hundred million of cases of disease annually and approximately 800,000 deaths [7], but babesiosis is an often-forgotten cousin. Babesiosis is an ancient but emerging group of pathogens with widespread global reach, for which the true prevalence is not known [84], in large part because of its near-complete resemblance to malaria on diagnostic testing, typically a blood smear.

One of the most striking findings of the 21st century in parasitic infectious disease is that—from a species perspective—malaria is not always what it seems. Reports that *P. knowlesi*, a monkey parasite, frequently causes zoonotic infection in humans in Malaysia shocked the community of physicians and researchers [77]. It was called the "Fifth human malaria parasite" [91]; the surprise resulted not only from the confusion, but also from how widespread, persistent, and grave the error was. Zoonotic infection by *P. knowlesi* causes a life-threatening clinical phenotype, but had been confused for the relatively innocuous *P. malariae*. The title of a seminal paper said it best, "*Plasmodium knowlesi* malaria in humans is widely distributed and potentially life threatening" [14]. In retrospect, it is perhaps not surprising, as diagnosis of red cell parasites is based almost entirely on stained blood smears, and morphology can be similar among the various species, even to experienced parasitologists.

The same thing is probably happening to *Babesia* today. The morphology of several *Babesia spp.* resembles malaria. *Babesia* parasites are present around the globe, and recent reports detail the extent to which they are missed or misdiagnosed [85, 98].

Each of these parasitic protozoa merits study in its own right, but there is also much to be gained by studying them together. The early history of malaria and babesia immunol-

ogy and pathobiology were closely linked. In 1957, Maegraith and Gilles provided the first observation about the striking similarity between the pathologic process of babesiosis in dogs and malaria in humans [54]. Cox observed that infection with *P. berghei* or *P. vinckei* produced cross-protection with one another as well as *Babesia spp.*, but not with other species of malaria [12]. Most strikingly, he later showed that high parasitemia *P. vinckei* infections were cleared during acute infection with *B. microti* [13]. In the same study, Cox further showed that this protection did not depend upon antibody, and suggested that cytokines, in particular TNF, were responsible for the control and/or clearance of malaria and babesia. Thus comparison between the two parasites was important in establishing the cytokine theory of disease pathogenesis. Since that time, joint studies the two parasites have slowed, in part because an acceptance that much of the pathobiology of each illness results from excess cytokine production, and also because of the importance of cytoadherence and vascular congestion in the pathogenesis of *P. falciparum*, which appears not to occur in human babesiosis [9].

Here, we return to the approach of studying malaria and babesia side-by-side, and approach two related, yet distinct questions in protozoan genomics. The first is whether *Plasmodium falciparum* strains from Southeast Asia which are known to rapidly develop drug resistance, do so by using an increased mutational rate. The second question is what evolutionary forces are driving the emergence of human babesiosis in the Northeast USA. Along the way, we use what we know, and what we have learned, from each and apply it to the other. We also take the comparative approach because of convenience: if one wants to study the pathophysiology of malaria in human infections, it is difficult to do that outside of malaria endemic areas. We happen to be in the heart of an endemic area for another red cell parasitic protozoan, which affords a unique opportunity to study the pathophysiology of red cell parasitism in humans.

In this section, we have introduced the parasites and framed the larger context for the specific problems we study in this work. Introduction to the specific questions under study for each of the pathogens is provided in the introduction sections of Chapters 3

and 4.

2 Methods

2.1 Parasite Culture, DNA Collection, and Sequencing

In Vitro Culture of *P. falciparum*

Parasite culture was performed according to the standard methods of Trager and Jensen [80]. The culture was performed by Dr. Xiaorong Feng in the Dr. Xin-zhuan Su at the National Institute of Allergy and Infectious Disease (NIAID) in the National Institutes of Health (NIH) (we have recently set up the facilities to do this culture in the Sabeti lab, where the experiments described in Section 3.6.5 are being performed by me and Lisa). The experiment was performed in two stages. In the first stage, the chloroquine-resistant line 7G8 was cultured under no drug and in the presence of low-dose chloroquine. The goal of this experiment was to assess whether the inclusion of low-dose chloroquine altered the mutation rate. At day 285 of the experiment, it was decided to also include the line TM, a patient isolate from Indochina, similar to the W2 used in the original ARMD experiment [69].

Cultures were maintained at 4% hematocrit in 25mL flasks. Parasitemias less than 5% were maintained at all times, and the culture medium was changed once per week. Parasite pellets were harvested at regular internals, and DNA was extracted from these pellets. Low dose chloroquine, at IC15 concentrations, was added to the culture medium in flasks under chloroquine pressure. Both 7G8 and TM were resistant to chloroquine at the start of the experiment, enabling the growth of these lines in low concentrations of chloroquine without impairing parasite growth.

Isolation of *B. microti*-Infected RBC

Whole blood isolated from patients infected with *B. microti* was filtered through a filter consisting of a 5 centimeter cellulose column in a 10mL syringe. The filtrate contained infected and uninfected red blood cells, whereas human leukocytes were retained in the

column.

Genomic DNA Preparation

For *P. falciparum*, genomic DNA was isolated from parasite pellets following lysis with 0.1% saponin. Pellets were then processed using the Qiagen Whole Blood gDNA extraction kit, which consists of a proteinase K digestion step followed by column purification on a silica column. Elution was in 1X TE (10mM Tris-HCl pH 8.0, 1mM EDTA). DNA was stored at 4 °C.

For *B. microti*, genomic DNA was extracted from infected red blood cells purified on a cellulose column (described above) using a Qiagen DNeasy kit.

Sequencing

DNA sequencing was performed at the Broad Institute on an Illumina HiSeq 2000 instrument. Samples were multiplexed with barcoded adapters and run 10 per lane. Library construction was performed at the Broad institute initially, and then subsequently by me. Library construction consists of shearing 1μg of DNA at a concentration of 20μg/μL, to a mean fragment size of 400 base pairs on a Covaris sonicator, followed by end-repair, adapter ligation, and bead cleanup (formed on an automated IntegenX PrepX robotic system) of 15 μL of sheared material. Quality control of sheared and library material was evaluated using an Agilent Bioanalyzer with a High Sensitivity DNA kit.

2.2 Short Read Alignment and Variant Detection

Short reads were aligned to the malaria genome (3D7 reference) using BWA [52] to generate *.bam alignment files. The coverage of each nucleotide per site was extracted from alignment files using a custom python script, and this information was used to estimate allele frequency per site (if we let x_1, x_2, x_3, x_4 represent the coverage of A,C,G,T, respectively, then the estimate for the nucleotide frequency for a given site is estimated as $\hat{f}_i = \frac{x_i}{\sum x_i}$).

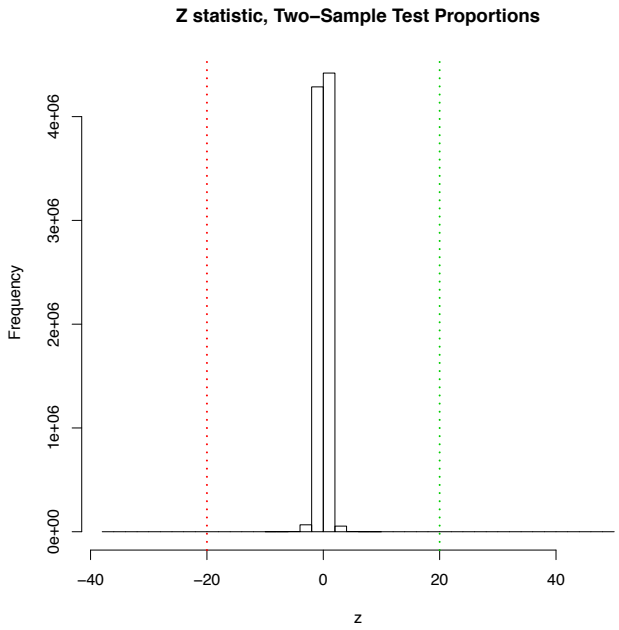


Figure 1: Histogram of values of Z statistics for the alleles in the beginning and end of the 7G8 line cultured in the presence of chloroquine. Selecting the outliers in this plot ($|z| > 20$) was used as a method to identify alleles that changed in allele frequency during the course of this study. These were then validated by manual inspection.

Nucleotides that had changed in frequency were calculated using a two-sample test of proportions,

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}},$$

with the histogram of z scores reported for the 7G8 line shown in Figure 1. A threshold of $z = 20$ was used to identify variants that were significantly different, a value chosen by empirical validation of the read alignments in an alignment browser (Integrative Genomics Viewer, Broad Institute [71]). An alternative approach was taken, using the Genome Analysis Toolkit (GATK) [57], requiring a minimum coverage of 20 reads and a genotype quality score of greater than 100, which yielded equivalent results.

2.3 Models and Model Fitting

Simulation of the Wright-Fisher Model

The Wright-Fisher model was simulated directly using a discrete-time, finite space Markov chain with a transition matrix, P, whose elements are given by

$$\binom{2N}{k} p^k q^{2N-k}.$$

The distribution of allele frequencies at time $t+1$, x_{t+1} was calculated by multiplying the allele in the previous generation by the transition matrix:

$$x_{t+1} = P x_t.$$

Logistic Regression to Model Changes in Allele Frequency

For each of the nucleotide substitutions, a logistic regression was fitted to the allele frequency data as a function of time using the general linear model function `glm()`, with the `family` parameter set to `binomial`.

Poisson Regression to Model Substitution Rate

A Poisson regression was fit to the data on substitution rate using the `glm()` function in R with the `family` parameter set to `Poisson`. The outcome variable of the regression was the number of substitutions and the predictor variables were genotype and drug pressure. An offset equal to the logarithm of the time in culture using the `offset` parameter was included in the regression in order to model rates.

2.4 Enrollment of Clinical Cases

Patients were identified through the MGH microbiology laboratory. Consent according to Partners IRB protocol 2014P000948 was obtained by study personnel (the author,

Sue Bazner, or Eric Rosenberg). A venous blood sample of 5-10mL in a heparin tube was obtained. This was stored at 4 °C until infected red blood cell isolation and genomic DNA extraction as above.

3 Monitoring *P. falciparum* Evolution *in vitro*

3.1 Introduction: Drug Resistance and Evolution in *P. falciparum*

There are three major classes of antimalarial drugs: folate antagonists, quinolones, and artemisinins. Folate antagonists, such as the commonly used combination sulfadoxine-pyramethamine (SP), interfere with DNA synthesis by blocking the enzymes dihdropteroate synthase and dihydrofolate reductase, respectively [27]. Quinolones, the class of drugs containing quinine and its derivatives, kill the malaria parasite by preventing the detoxification of heme in the parasitic digestive vacuole, likely through inhibition of the heme polymerase [78]. Artemisinins, a class which includes artesunate and artemether, are the third major category of antimalarials. This group of antiparasitic drugs has recently come recently into greater use because of widespread resistance to folate antagonists and quinolones, but the mechanism of action of artemisinins remains poorly understood.

From a historical standpoint, chloroquine resistance has been the most devastating. Resistance appeared in 1978 and rapidly spread to all the countries in Africa by the end of the 1980's [81]. During this period, malaria rates increased substantially, with hospitals reporting 2 - 3 fold excess of mortality from malaria deaths, an observation which, through careful epidemiological study, was conclusively linked to treatment failure with chloroquine [81, 82]. The reliance on chloroquine as first-line treatment, because of its cost, rapid parasite clearance when used against sensitive parasites, and widespread availability, proved disastrous when resistant mutants entered the population and rapidly spread throughout the world.

With the recent reports of resistance to artemisinin compounds [19, 62, 63], resistance has now emerged to every antimalarial drug used. The prospect of treatment refractory malaria seems a genuine possibility for the future.

How Resistance Arises

Drug resistance occurs because specific resistance alleles enter the population and are rapidly selected due to their fitness advantage in the face of drug pressure. The spread of resistance can be broken down into three steps:

1. Entry of the mutant allele into the population.
2. Selective advantage of the mutant allele and survival within an infected host.
3. Spread of the mutant allele through populations and into multiple hosts.

Much more is known about the third step, i.e. the molecular and cell biologic nature of alleles that confer resistance to antimalarial drugs, than about the first two. For example, a serine to asparagine substitution at position 108 of DHFR confers pyrimethamine resistance, and a lysine to threonine change at position 76 of the PfCRT gene results in chloroquine resistance [27, 90]. The dynamics of spread have also been studied: mutant alleles and their associated haplotypes can be detected sweeping through populations, tracing out the molecular epidemiologic paths of drug resistance [61, 93].

In contrast, the basic evolutionary mechanisms of the first phase of resistance, entry of the mutant allele into the parasite population, are poorly understood. The number of mutant alleles entering the population, at a particular locus, can be expressed as $N\mu$, where N gives the size of the population and μ represents the mutation rate, per-site, per-generation (see Section 3.4). This simple relation makes it clear that two key parameters determine the flux of mutant alleles into the population: 1) the mutation rate and 2) the population size. Population sizes can be estimated reasonably well, but the mutation rate is not known. Knowledge of μ is essential for predicting the likelihood that a specific resistance will arise, and therefore for understanding the resistance-free lifespan of antimalarial drugs alone or in combination. Furthermore, the extent to which mutation rate can be described as a single quantity, or whether it depends on the location within the genome, environmental and genetic factors, is unknown. The mutation rate is a quantitative phenotype likely to have different values among populations of distinct

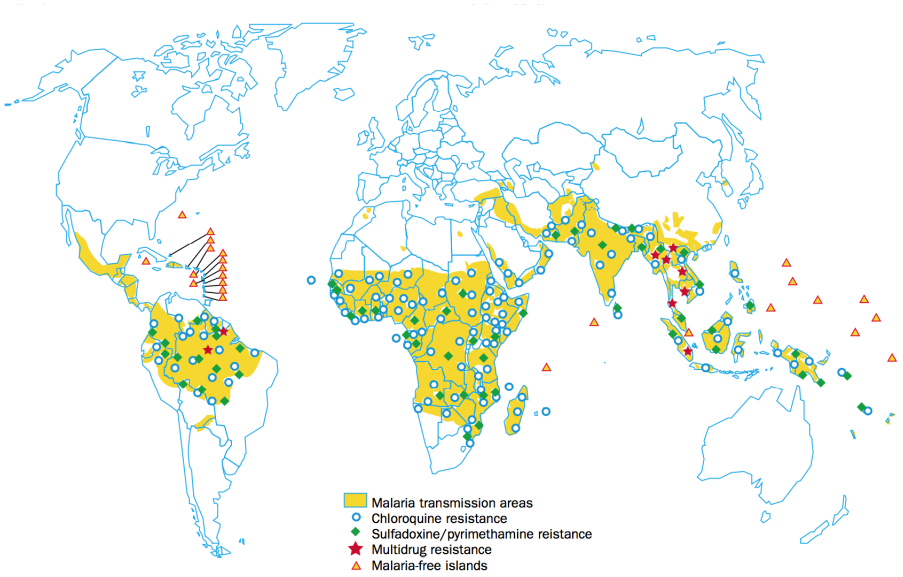


Figure 2: Map showing global distribution of drug-resistant parasites, as of 2002. Note the concentration of multiple drug resistant strains in Southeast Asia.

geographic origin. Do these values differ substantially enough to underlie the observation that resistance emerges at different rates in different locations?

The ARMD Phenotype

The geographic distribution of drug resistant parasites around the world is shown in Figure 2, taken from a review article by Wongsrichanalai [92]. When drug resistance first emerges, it tends to follow a predictable pattern. The first cases of sulfadoxine-pyrimethamine resistance, as well as the first cases of artemisinin resistance, appeared on the Thai-Cambodian border. From there, resistance alleles then spread to other parts of Asia, Africa, and South America [19, 92].

There are two main possibile reasons for this, and they are not mutually exclusive. The first is that environmental and epidemiological patterns create fertile conditions in which resistance can arise, such as sub-standard drug formulations and inadequate or incomplete treatment courses. The second is that parasites from this region harbor genetic factors that make it easier for them to acquire resistance. While the first explanation is a hallmark of situations that breed resistance and undoubtedly important, there has

Table 2. Frequency of resistance to 5-fluoroorotate

Initial size of parasite population per flask	Outcome of selection*				
	W2 (Indochina)	FCR3 ("The Gambia")	HB3 (Honduras)	3D7 ("Netherlands")	D6 (Sierra Leonne)
10^8	3 /3	0 /3	0 /3	0 /3	0 /3
10^7	3 /3	0 /3	0 /3	0 /3	0 /3
10^6	3 /3	0 /3	0 /3	0 /3	0 /3
10^5	0 /3	0 /3	0 /3	0 /3	0 /3

*The ratios represent the number of flasks that yielded resistant parasites within 2 months of culture, divided by the initial number of flasks setup.

Figure 3: Table showing the result of resistance selections against ARMD strains from reference [69]. The authors were repeatedly able to extract resistant parasites at starting parasitemias of as low as 1 million parasites in the W2 strain from Indochina, whereas they were unable to do so in the other isolates.

been increasing evidence from the scientific literature that the situation may be more complex, and that, in addition to environmental factors, genetics and gene-environment interactions also play a role.

In a now-classic series of experiments, parasites from around the world were exposed to previously unused antimalarial drugs to which they were all sensitive. Surprisingly, they did not evolve resistance at the same rate. In particular, the strain W2, a Southeast Asian strain from along the Thai Cambodian border, acquired resistance to new antimalarials approximately 100-1000 times faster than other strains [69]. A table, from the original article [69], is reproduced in Figure 3. This table shows the evolution of resistant parasites, as a function of starting parasitemia, in response to the drug 5-fluoroorotate, a novel antimalarial. A near-identical result was seen for the drug atovoquone, to which the parasites included in the study had also never been exposed.

For several years after publication, this phenotype, which was named accelerated resistance to multiple drugs (ARMD), remained in the literature as an intriguing finding that lacked support from other lines of evidence. More recently, however, evidence has begun to accumulate linking the ARMD phenotype to specific molecular correlates. Trotta and colleagues showed, using *in vitro* assays of plasmid repair after UV irridiation, that the

W2 strain from the original ARMD report was unable to as efficiently incorporate radiolabeled nucleotide into the damaged plasmid in a DNA repair reaction [83]. The key plots from their paper are reproduced in Figure 4. Furthermore, Trotta and colleagues also noted that antimalarial drugs inhibited repair kinetics [83], with a dose-dependent inhibition by chloroquine recorded, as well as inhibition by the other quinolones mefloquine and quinine (Figure 4). More recent studies have found similar results when examining particular DNA repair pathways, notably the mismatch repair pathway [5].

Repair of chemical damage to DNA base pairs is an essential and highly conserved aspect of prokaryotic and eukaryotic cell biology, underscoring the importance of this process to all of cellular life. Several sources of insult, including ultraviolet light, reactive oxygen species, mutagenic chemicals, and replication errors, can alter the helix structure and lead to errors in the genetic code. In order to mitigate this damage, three major pathways for repairing damaged DNA are employed: base excision repair (BER), nucleotide excision repair (NER), and mismatch repair (MMR). BER is responsible for removal of single-nucleotide molecular lesions requiring immediate attention, and involves generation of an abasic site, cleavage of the phosphodiester bond, followed by polymerase and ligase activates that resynthesizes the removed nucleotides. The NER pathway repairs larger lesions, such as bulky thymidine dimers induced by UV radiation, and consists of excision and resynthesis of a short sequence of nucleotides in a single strand. MMR is a 'proofreading' activity involved in evaluating, and maintaining, the integrity of newly synthesized DNA strands, correcting improperly paired bases (e.g. A/C, G/T pairs).

DNA damage repair pathways have been poorly characterized in *P. falciparum*, but the enzymatic activity of a BER pathway has been detected, and shown to be of the "long-patch" type, involving synthesis of a 2-10 nucleotide stretch [34]. MMR pathways have also been characterized in the parasite, and suggested to play a role in introducing variation into populations as a diversity-generating mechanism [3].

Why and how do antimalarials inhibit DNA damage repair? This is a question that goes back to the early history of chloroquine as an antimalarial drug and an chemical

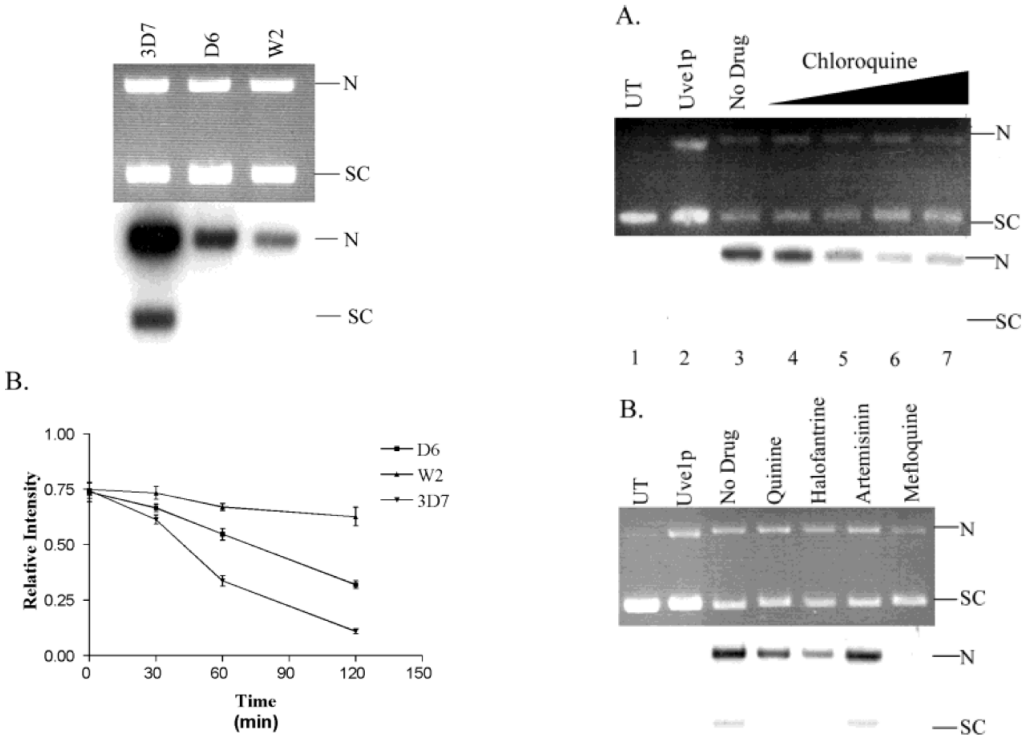


Figure 4: Left side: Figures from Trotta et al. [83] showing variability in kinetics of DNA damage repair, with the weakest repair response occuring in strain W2. Right side: Figure from Trotta et al. [83] demonstrating inhibition of DNA repair kinetics by antimalarial agents including chloroquine, quinine, mefloquine, and halofantrine. In these assays, repair efficiency is by the quantity of radiolabeled nucleotide incorporated into the N (nicked) form of the plasmid.

used in the research literature. Due to its rapid parasite clearance rates, low cost, and high therapeutic index, chloroquine was the first-line antimalarial drug for most of the second half of the twentieth century. Early investigations into its mechanism focused on its nucleic acid binding properties, and its corresponding effects in the cell. In 1965, Cohen and Yielding documented inhibition by chloroquine of DNA and RNA polymerase activity [10], and several workers demonstrated inhibition by chloroquine in DNA repair when mutagenesis was induced by alkylating agents [58] and UV light [97]. While the mechanism of action of chloroquine was later shown to be unrelated to its DNA binding activity [24, 78, 89], this secondary activity may result in an increase in the mutation rate under chloroquine pressure, and a corresponding acceleration in the acquisition of drug resistance to this and other drugs.

Therefore, in addition to the usual factors which accelerate the acquisition of drug resistance, the available scientific evidence suggests that two additional factors may play a role, accelerating the acquisition of resistance above and beyond that which is normally expected. The first is that parasites from this region may harbor specific genetic mutations, such as polymorphisms in DNA repair enzymes or lower levels of expression of these enzymes, which impair the kinetics of DNA repair. The second is that antimalarial agents themselves may be directly mutagenic, either by directly reacting with DNA bases or by reducing the efficiency of DNA repair mechanisms. A direct measurement is important because impaired kinetics, demonstrated in the artificial situation of an extract, may not translate into an elevated rate *in vivo*; the time frame and chemical conditions for DNA repair are different in replicating parasites, and the types of DNA damage encountered in the normal lifespan may not be accurately represented by brief, intense ultraviolet irradiation. For these reasons, we sought to compare the nucleotide substitution rates under conditions of chronic, low-dose drug pressure, and also in parasites from southeast Asia to parasites from South America, as a direct test of the hypothesis that these factors elevate the mutation rate and facilitate the evolution of drug resistant parasites.

The basic principle of the long-term culture experiment was to study the process of

mutation by investigating the variation accumulated in a population over time. In order to test the hypothesis that genetic factors and the environmental pressure of low-dose chloroquine elevate the mutation rate, two parasites, 7G8 and TM, were cultured in the laboratory of Dr. Xin-zhuan Su (NIH) under low-dose chloroquine pressure (IC15) or without drug pressure. We then performed whole genome sequencing on these isolates to obtain a census of genetic variation that had accumulated during the course of the experiment.

The 7G8 parasite, a South American line, was grown for a total of 510 days in each arm. With a generation time of 48 hours, this corresponds to 255 generations per arm. The TM isolate was grown for a shorter period of time. Growth was initiated at Day 60 and continued until Day 480; the chloroquine pressure was added at a later date (Day 230). Overall, 500 days (250 generations, or 125 generations per arm) of divergent evolutionary time separated the TM cultures with and without chloroquine pressure. The experiment is depicted schematically in Figure 5. The solid circles denote samples that were sequenced in time for this analysis; the unfilled circles show the full set of samples, reflecting the contribution of additional time points which are in the process of being sequenced.

3.2 Sequencing of Long Term Culture Isolates

We performed whole genome sequencing using the illumina platform. This generates short fragments ('reads') of sequence derived from a larger, sheared piece of DNA. In this case, we generated 101 bp reads from both sides of a 500 bp fragment. These reads were aligned to the *P. falciparum* genome, using the finished sequence of the 3D7 isolate, with the software tool MAQ [52]. Table 1 gives the total number of reads per library, the percentage of which could be assigned to the *P. falciparum* genome, and the number and percentage of successfully aligned reads. The libraries were sequenced in duplicate in two different lanes, and a plot showing the number of reads for each library in the two different lanes is presented in Figure 6.

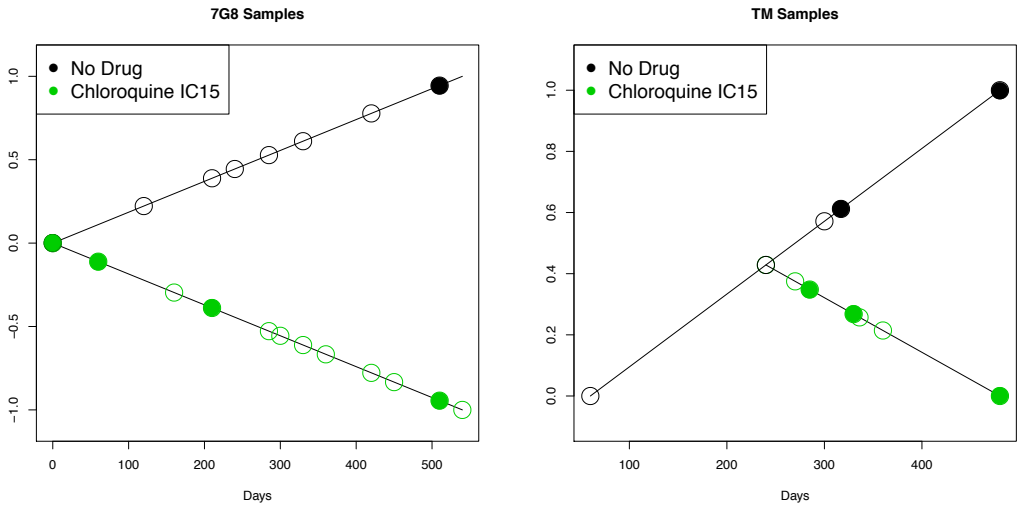


Figure 5: Schematic of 7G8 (left) and TM (right) samples used in this study. Samples depicted in solid color circles were sequenced in the first group; unfilled circles denote samples sequenced during the second round of sequencing.

A combined 746,413,184 reads were obtained from the two sequencing runs, producing a total of 74.6 gigabases. The read output from each lane was approximately equal, with one lane (lane 7) producing 372,968,692 reads and the other (lane 8) producing 373,444,492 reads. This resulted in 133-fold coverage, on average, for each of the 10 samples include in a single lane. The sequence and alignment statistics for individuals samples are provided in Table 1.

3.3 Survey of Variation in Cultured Populations

The results of whole genome sequencing generated a list of sequence variants that different between the strains. We first considered single nucleotide mutations. From a 23 MB genome, cultured for 255 generations, we identified a total of 7 nucleotide substitutions in the 7G8 line. These substitutions are catalogued in Table 2. Of these, 4 were identified in the line under chloroquine pressure, and 3 were in the line grown in the absence of chloroquine. In the TM line, which was cultured for 125 generations, we observed 9 substitutions, 4 of which took place in the presence of chloroquine, and 5 in its absence.

We identified large genomic deletions on chromosomes 2 and 13 in the 7G8 line without

ID	Lane	Total Reads	*P. falciparum* %	Pf Aligned Reads	Pf Aligned %
Day 2107G8 + CQ	7	27,609,386	91.28	22,610,897	89.72
Day 2107G8+ CQ	8	28,102,164	91.39	23,035,170	89.70
Day 510 7G8+ CQ	7	45,464,694	91.22	35,871,853	86.50
Day 510 7G8+ CQ	8	45,458,254	91.36	35,927,226	86.51
Day 336 TM + CQ	7	41,010,744	91.06	33,895,030	90.77
Day 336 TM + CQ	8	40,796,658	91.21	33,776,930	90.77
Day 480 TM + CQ	7	40,417,314	90.86	33,320,014	90.74
Day 480 TM + CQ	8	40,205,616	91.00	33,194,442	90.73
Day 5107G8	7	39,832,146	91.20	32,556,402	89.62
Day 5107G8	8	39,598,582	91.35	32,417,985	89.62
Day 285 TM + CQ	7	35,958,586	90.40	28,782,293	88.54
Day 285 TM + CQ	8	36,141,016	90.54	28,983,204	88.57
Day 317 TM	7	27,619,272	91.00	22,776,580	90.62
Day 317 TM	8	27,778,270	91.13	22,942,839	90.63
Day 480 TM	7	39,095,750	91.46	32,254,183	90.21
Day 480 TM	8	39,081,238	91.59	32,289,302	90.21
Day 7 7G8	7	39,522,184	91.14	31,236,075	86.72
Day 7 7G8	8	39,265,698	91.28	31,080,928	86.72
Day 60 7G8 + CQ	7	36,438,616	91.72	30,777,597	92.09
Day 60 7G8 + CQ	8	37,016,996	91.85	31,312,201	92.10

Table 1: Alignment statistics for each library sequenced in this experiment. Libraries were run in duplicates in two lanes (Lane 7 and Lane 8).

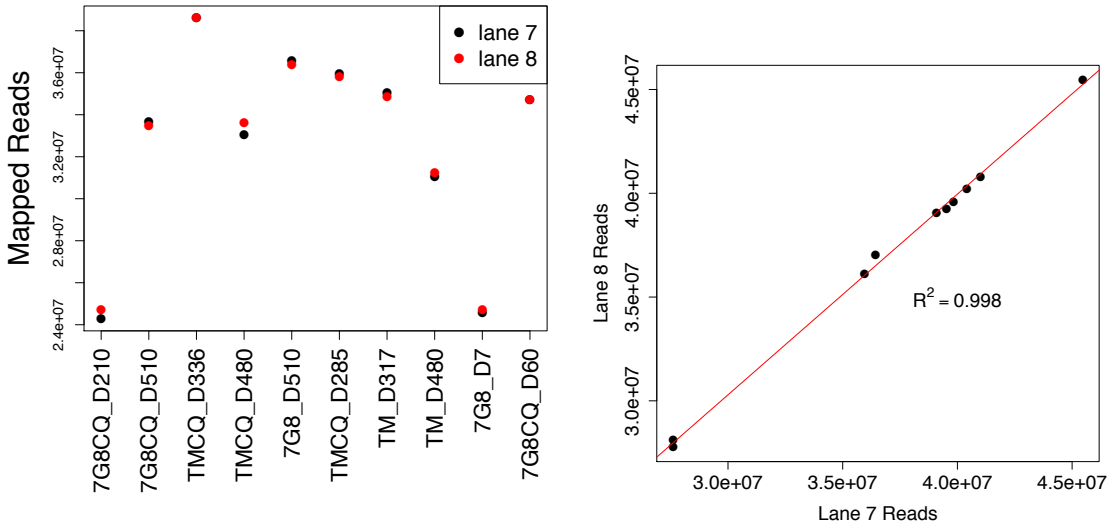


Figure 6: Comparison of sequence yield from each of the two lanes used for sequencing. In the top panel, the number of reads for each sample is shown in both lanes 7 and 8. The bottom panel shows the correlation between the number of reads from each lane. Variation in read numbers is mainly attributable factors in the library and not on the flowcell, since the same libraries sequenced in different lanes show almost identical numbers of reads. Fitting a linear model, we can estimate that only 0.2% of variation can be attributed to inter-lane variability. The remainder of variability likely results from the library construction and pooling procedures.

Strain	Drug	Position	Anc.	Der.	Effect	Protein	Description
7G8	-	Pf3D7_10_v3:124719	T	C	E → G	Pf3D7_1002500	conserved, unknown
7G8	-	Pf3D7_10_v3:993109	T	A	Y → N, +A	Pf3D7_1023700	conserved, unknown
7G8	-	Pf3D7_12_v3:1839598	G	T	T → K	Pf3D7_1243900	DOC2
7G8	CQ	Pf3D7_12_v3:415417	G	T	S → I	PF3D7_1208900	conserved, unknown
7G8	CQ	Pf3D7_12_v3:415418	C	T	S → I	PF3D7_1208900	conserved, unknown
7G8	CQ	Pf3D7_13_v3:2545538	A	T	L → H	PF3D7_1363400	conserved, unknown
7G8	CQ	Pf3D7_14_v3:2783901	T	G	I → S	PF3D7_1468000	conserved, unknown
TM	-	Pf3D7_06_v3:367812	C	A	F → L	Pf3D7_0608900	conserved, unknown
TM	-	Pf3D7_12_v3:433910	G	-	1bp deletion		
TM	-	Pf3D7_12_v3:433914	T	A	non-coding		
TM	-	Pf3D7_12_v3:433915	A	T	non-coding		
TM	-	Pf3D7_13_v3:1878801	A	C	non-coding		
TM	-	Pf3D7_13_v3:1878807	A	T	non-coding		
TM	CQ	Pf3D7_05_v3:1161298	A	T	L → I	Pf3D7_0528100	beta-adaptin
TM	CQ	Pf3D7_12_v3:1389331	A	C	N → H	Pf3D7_1233600	AARP1
TM	CQ	Pf3D7_13_v3:2002426	A	C	non-coding		
TM	CQ	Pf3D7_14_v3:1736213	G	A	E → K	Pf3D7_1442600	TRAP-like protein

Table 2: Nucleotide substitutions identified in this experiment. Anc. – the ancestral allele, Der. – the derived allele.

chloroquine pressure and on chromosome 8 in the W2 line (Table 3 and Figure 7). The deletion on chromosome 2 is a well-known deletion which removes the knob-associated histidine rich protein, KAHRP, and results in knobless parasites [67, 68]. Presumably, these parasites have a growth advantage *in vitro* but would be at a disadvantage *in vivo* since they cannot adhere to the peripheral vasculature. The deletion on chromosome 13 removes the EBA-140 invasion ligand, which may confer a selective advantage in culture or for invasion of the red blood cells used in the experiment.

Small insertions and deletions were identified using the variant detection algorithms in samtools [53]. However, unlike the case of SNPs, the sensitivity of these methods is probably quite low. This is due to the fundamental problem of attempting to call variants with short reads as well as to the relative lack of sophistication of algorithms for making such calls. Therefore, the insertion and deletion calls in the dataset likely do not represent a complete census, and will be worth revisiting when more sophisticated software packages are available in the future.

In contrast, the catalogue of large deletions, such as the $\sim$ 100 kb deletions seen on chromosomes 2, 8, and 13, probably represents a mostly complete annotation of the set of large deletions in the data. These deletions tend to occur in stereotypic locations [68], on the ends of chromosomes and in internal antigen repeat clusters, and could be assessed

Strain	Pressure	Position	Type	Size	Description
7G8		Pf3D7_02_v3:1-110000	deletion	110 kb	KAHRP deletion
7G8		Pf3D7_10_v3:993115	insertion	1bp	
7G8	CQ	Pf3D7_13_v3:1-125000	deletion	125 kb	Removes EBA-140, Pfg27
7G8	CQ	Pf3D7_11_v3:1774418	deletion	14 bp	
7G8	CQ	Pf3D7_11_v3:1774418	deletion	14 bp	
TM		Pf3D7_12_v3:433910	deletion	1bp	
TM		Pf3D7_13_v3:1878807	deletion		
TM	CQ	Pf3D7_08_v3:1-110000	deletion	110 kb	

Table 3: Insertions and deletions identified in this experiment.

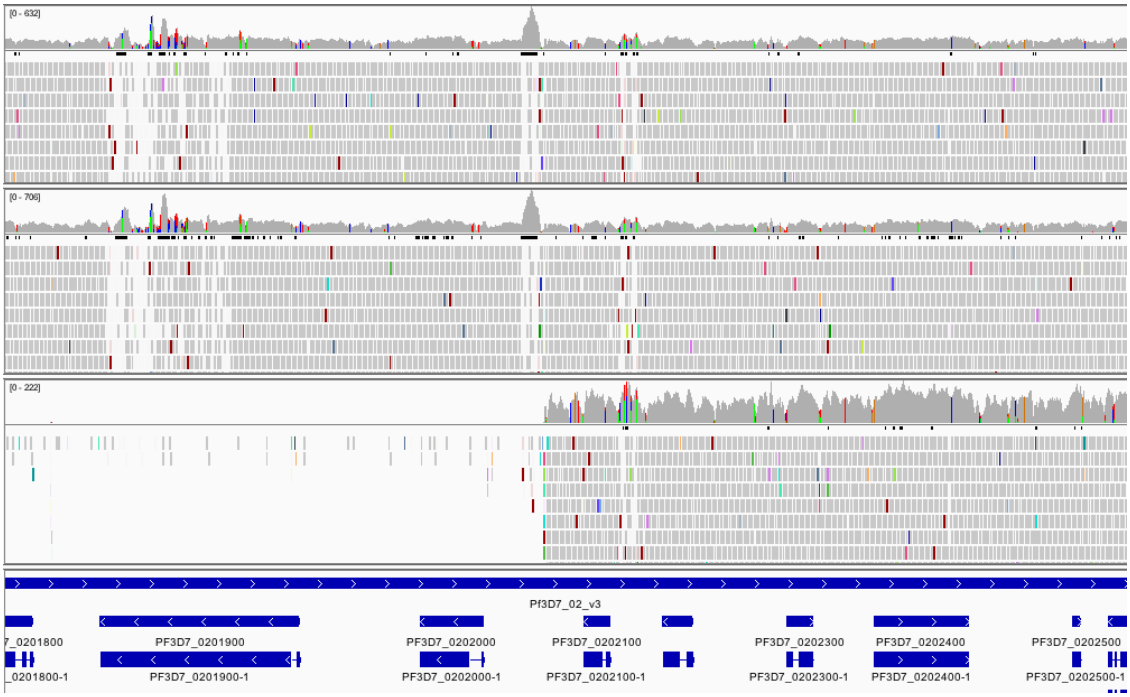


Figure 7: The left arm of chromosome 2 was deleted in the 510 day 7G8 line without chloroquine pressure. The deletion includes the KAHRP gene and extends $\sim$ 110 Kb into the chromosome.

by visual inspection of the coverage plot along each of the fourteen chromosomes. An example of a positive signal is shown in Figure 7, which shows the reads mapping to the left arm of chromosome 2, and associated coverage plots, in the time points of the the 7G8 line without chloroquine pressure. At the final time point, a loss in signal from the first 125 kilobases of chromosome 2 can be seen.

3.4 The Independent Sites, Fixed Rate Model

In order to interpret the variation identified in this experiment, we need to define what we mean by mutation rate. We do this in the context of a mathematical model which is powerful enough to provide a framework for interpreting the results of our experiment but straightforward enough to analyze easily.

We consider a simple model in which a site, x, mutates to x' with a probability p in a single generation.

$$P(X = x'|X = x) = p \tag{1}$$

We assume this probability is fixed. In the next generation, our allele is x with probability $1-p$ and x' with probability p. After t generations, the probability that x has not mutated to x' is $(1-p)^t$, and the probability that x mutates in generation t is given by $(1-p)^{t-1}p$, which is known as the geometric distribution. If we assume that the generation time is short compared to the length of the experiment, we can model this in continuous terms, using the continuous analogue of the geometric, known as the exponential distribution,

$$P(t, \lambda) = \lambda e^{-\lambda t},$$

which gives the time between mutations. The exponential (and in the discrete case, the geometric) distribution expresses a model of mutations in which they occur over time at a fixed rate, independent of the time of occurrence for other mutations; it is the only continuous probability distribution that satisfied these two conditions. Once this rate has been specified, the number of mutations, K, which have occurred by time t can be shown to follow a Poisson distribution,

$$P(K = k) = \frac{e^{-\lambda t}(\lambda t)^k}{k!}.$$

This is the definition of a Poisson Process [29]. Thus we define K to be the random variable which describes the number of times a site has mutated at time t. The expected

value of a Poisson distribution is λ, $E[K] = \lambda t$.

This model can be extended to multiple sites. We can calculate the probability of k mutations occurring at m sites if we assume that the sites are independent. Define Y as the random variable that describes the number of mutations for m sites at time t. Then

$$P(Y = y) = \prod_{i=1}^{m} \frac{e^{-\lambda_i t}(\lambda_i t)^k}{k!}, \tag{2}$$

and, if all the $\lambda_i = \lambda$, then

$$E[Y] = m\lambda t. \tag{3}$$

In other words, the average number of mutations in the fixed-rate, independent sites model is the number of sites multiplied by λ and also by the time of the experiment. We return repeatedly to this model, since it provides a natural setting for rate estimation and hypothesis testing about rate, and it also offers insight into the experimental process of mutation accumulation.

Before proceeding further, the limitations of this model are important to consider. The key relation, $E[Y] = m\lambda t$, is contingent on the validity of several assumptions. These are 1) that the mutation rate is a constant and the same for all sites, 2) that the time for any given mutation to occur is independent of the time for other mutations to occur, and 3) that the population consists of a single individual (or, equivalently, that drift is the only force driving substitution—see section 3.6). These assumptions are a reasonable starting point, and lead to a tractable model, but we cannot assume they will hold unconditionally in all situations. Whether these assumptions are met in practice, and if not, how they are violated, will be considered in later parts of this thesis.

We are now in a position to define what we mean by the mutation rate. In population genetics, μ, the mutation rate, which might better be called the mutation probability, is the probability with which a mutation will occur in replication of DNA in a single generation. In discrete-time models, it is the p defined in equation 1. In the continuous-time framework, μ is equivalent to λ. In genetics, a second quantity is also defined, termed

the substitution rate and denoted ρ, which is the rate at which substitutions become fixed, with allele frequency 1, in a population over time. This is mostly a quantity of convenience, since it is what is typically measured in an experimental sample alleles from a population over time. The two quantities, μ and ρ, are closely related, but they are not the same, except in certain cases. The parameter μ describes the number rate at which mutations enter the population, whereas ρ describes the rate at which mutations become substitutions, i.e. reach allele frequency 1. A trivial case in which the two quantities are equal occurs when the population consists of one individual; in this instance, all mutations entering the population are fixed with allele frequency 1. A more nuanced case occurs when genetic drift is the only force acting on populations. Under such conditions, it is a well-known result in population genetics that $\rho = \mu$ (see, for example [43] or [29]).

3.5 Estimation of Rate

The model described by equations 2 and 3 suggests a way to assess what covariates affect rate. We can use Poisson regression, which is the typical regression model for count data, and is further justified in this case because we expect, from theory, that the number of substitutions accumulated over time forms a Poisson process. Therefore, we model the mutation rate using Poisson regression,

$$E(Y|x) = e^{\theta^T x},$$

where x is a vector of input variables and θ are the estimated parameters. We test for the significance of genotype and chloroquine by testing whether these coefficients are significantly different from zero, using the hypothesis testing procedures for generalized linear models [56], which are implemented in the R.

The results of fitting a Poisson regression are shown in Table 4. The effect of chloroquine is not significant, indicating that the hypothesis that chloroquine alters substitution rate is not supported by the data. The effect of genotype is not significant at the $p = 0.05$

Parameter	**Estimate**	**Std. Error**	**Z**	$P(>\|z\|)$
Intercept	-4.289	0.5432	-9.463	$< 2 \times 10^{-16}$
Genotype	0.9643	0.5040	1.913	0.0557
Chloroquine	1.665×10^{-15}	0.500	0.000	1.000

Table 4: Model estimates for Poisson regression of mutation number.

level; however, the p-value in this case is $p = 0.0557$, suggesting a trend toward significance. Thus, from this experiment, it appears there is some evidence that genotype affects substitution rate, but it is not conclusive. One way to think about it is that there is an approximately 6% chance that, if the rates are not different, the observed level of variation is due to chance. The estimated substitution rate is estimated to be 2.9-fold higher the TM line, a fairly large effect, suggesting that lack of statistical significance may have been caused by low study power.

It is worth noting that two single nucleotide substitutions were adjacent to other single nucleotide variants (Pf3D7_12_v3:415417/8 in 7G8 and Pf3D7_12_v3:433914/15). All variants in Table 2 met quality control thresholds and appeared real upon visual inspection. Thus, the variants presented there represent our best estimate as to the true variants present in the sample. However, we considered the possibility that these were misclassified as insertions/deletions, and the effect this would have on the results of the model. Removing SNPs that had adjacent changes from the analysis did not materially change the results (in the model Table 4, the p-value for chloroquine remained at 1.0, and the p-value for genotype went to 0.073). Removing those variants plus any nearby SNPs, the p-value for chloroquine remains non-significant at $p = 0.53$ and the p-value for genotype shows a weaker trend toward significant ($p = 0.26$). In general, nearby variants present a challenge for the model, since they indicate a lack of independence between sites. Mutation of a variant may change a transcription factor binding site or alter a codon, exerting large selective pressure on adjacent sites. This, for example, appears to be the case for the mutations Pf3D7_12_v3:433910 and Pf3D7_12_v3:433914/5, which occur in the promoter region of KROX1, a predicted zinc-finger transcription factor, and

may affect its timing or expression levels. Alternatively, clusters of variants may result from a localized insult to DNA that results in damage at several adjacent sites. It is also possible that the process of repairing DNA damage may predispose to further mutations. Overall, the complication of clustered variants does not change the essential character of the results. Chloroquine is not significant, and genotype also does not reach significance, but appears to trend in that direction, an effect that may be attributable to low study power.

The independent sites model can be used to calculate the best estimate of λ. As per equation 3, the expected number of mutations under a Poisson process is $E[k] = m\lambda t$, and the maximum likelihood estimate of λ is

$$\hat{\lambda} = k/mt.$$

We detected a total of 17 mutations over 21 million base pairs and 760 generations, which leads to an estimated marginal rate of

$$\hat{\lambda} = 1.065 \times 10^{-9},$$

which we can interpret as the estimated substitution rate, $\hat{\rho}$, per-site, per-generation. Whether we can use this value to estimate the mutation rate, μ, is discussed in the next section.

3.6 Interpreting the Rate: Evolutionary Forces in Culture

This experiment involves a direct measurement of the substitution rate, defined as the average number of changes, with allele frequency = 1, that have occured per-generation, per-site, in a sample of alleles over time. As discussed in Section 3.4, this is closely, but not exactly, related to the mutation rate, which is the rate at which, in replication of a *single* genome, errors will be introduced, per-site, per-generation.

Under an idealized situation in which genetic drift dominates, and selection is negligible, for a haploid population of N alleles, we can expect $N\mu$ new mutations to enter the population in each generation. These mutant alleles with have frequency $= 1/N$ at the time of their introduction, and the probability of fixation under drift is equal to the allele frequency [29]. The fixation rate, ρ, can then be calculated as

$$\rho = (N\mu)\frac{1}{N} = \mu, \tag{4}$$

which is a famous result in population genetics [15]. The question, then is whether these assumptions are valid under the conditions of this experiment.

The relative strength of these competing evolutionary forces — genetic drift and selection — depends on population size. Drift acts as a force because in small populations there can be large swings in the frequency of alleles simply due to sampling; however, as populations get large, the likelihood of sizable shifts in allele frequency becomes very small, and selection dominates. Therefore, whether drift or selection is dominant depends critically on population size.

What is the population size in these experiments? The *in vitro* culture of malaria parasites in this experiment (and in virtually all malaria culture experiments) involved periodically "cutting" the culture, and we should account for this in the analysis. For a 10 mL culture at 4% hematocrit, one would expect a maximum population of 2×10^8, and a minimum population size of 4×10^6. Therefore, since the population size changes during the course of the experiment, it is important to account for this. There is a substantial literature on how to deal with fluctuating population sizes, and the notion of an "effective" population size, N_e, was introduced by Sewell Wright in 1938 [96]. Wright calculated that variation in population size could be handled by taking the harmonic mean ($N_e = (\frac{1}{n}\sum x_i^{-1})^{-1}$) of the fluctuating population sizes, which in this cases yields a value of approximately 7×10^7. Regardless of the exact number, the approximate calculations suggest a value for N which is large, suggesting that the role is likely to be

minimal. We make this more quantitative in the next section.

3.6.1 Modeling the Effect of Genetic Drift

The strength of genetic drift is proportional to the inverse of population size. Since the population size in our experimens is large, we expect the effect of drift to be small. Therefore it seems improbable that the substitutions we observe in the experiment are fixed by drift. How improbable? We can calculate this probability exactly using a quantitative model of genetic drift.

Figure 8 plots calculations for the probability of fixation at time t under the Wright-Fisher model [25, 95] of genetic drift (as discussed in methods, Section 2.3). This plot, which has the number of generations plotted in terms of the population size, makes it clear that the probability is essentially 0 that drift has fixed any of the mutations in this experiment. The probability that a mutation will fix does not become non-zero until approximately $N/3$, which for the population sizes in our experiment is approximately 63,000 years. Therefore, we cannot estimate the mutation rate by setting $\rho = \mu$.

The results of Figure 8 are consistent with the analytical calculations of Kimura and Ohta, who calculated that the expected time to fixation in a haploid population is $2N$ [43].

3.6.2 Directional Selection in Haploid Populations

The results above, while disappointing in that they indicate we cannot directly estimate the mutation rate, do conclusively rule out one evolutionary force, drift, as a force driving substitution in this experiment. And if drift is not acting, it must be another force. Selection seems like a good candidate, a hypothesis we can test by modeling the changes we would observe due to selection. Furthermore, since the populations are large, we can use continuous models, which are much easier to deal with than finite-state, finite-time models such as the Wright-Fisher simulation of genetic drift.

There is an extensive theory of selection which describes, quantitatively, the changes

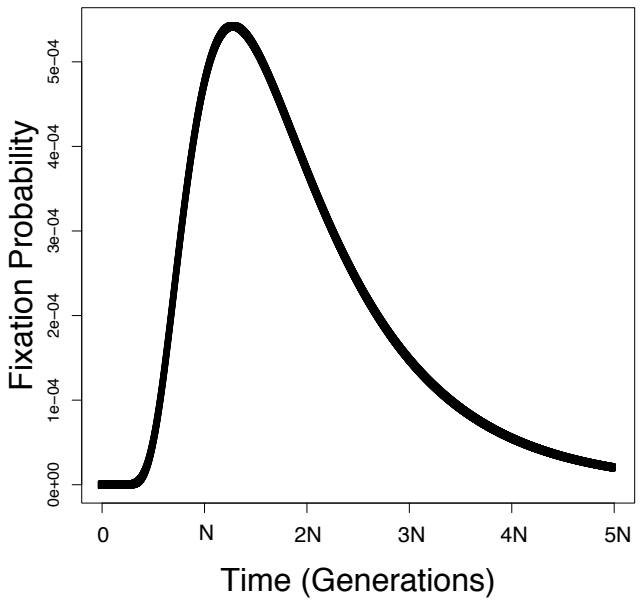


Figure 8: A simulation of the probability of fixation by time t under the Wright-Fisher model.

in allele frequency expected under various conditions. Since we have intermediate time points in this experiment, we can directly test this theory. In a classic paper from 1927, Haldane worked out the theory of directional selection in infinite populations [33]. While Haldane provided a family of results, the major one of importance for us is in a very large population, mutations under positive selection will change in frequency according to the logistic growth equation:

$$X(t) = \frac{1}{1 + Ce^{-rt}}. \tag{5}$$

$X(t)$ is interpreted as the allele frequency in the population over time, and the numerator sets the "carrying capacity" of the population = 1, consistent with the fact that allele frequencies should be contained in the interval $[0, 1]$. This equation results from solving the differential equation,

$$\frac{dX}{dt} = sX(1 - \frac{X}{K}).$$

The parameter in the exponent, r, gives the proportional change in allele frequency over time, which can be directly interpreted as the relative increase in fitness of the mutant allele, can be taken as the definition of the selection coefficient, s. The parameter K sets the carrying capacity (which, for allele frequency data, is set to 1). This differential equation describes the simplest system in which the rate of change decreases as the population reaches its capacity. It is a model for growth that begins exponentially but then slows down as it reaches a bound. In order to assess how well this model describes the alleles in our study, we can compare the trajectories of mutations in our samples to predicted trajectories under the model. To do so, we must fit the curve to our data, which is especially straightforward in this case because we can fit univariate logistic regression models to our data, which are of an identical form:

$$P(Y) = \frac{1}{1 + e^{\beta_0 + \beta_1 X_1}}.$$

Our estimate of the selection coefficient, s, is just the regression coefficient for the input variable, β_1. A benefit of this procedure is confidence intervals come automatically with standard software for fitting logistic regression models.

In this case, the regression predicts allele frequency as a function of the predictor variable, time, X_1. We performed this procedure for the allele frequency data from the 7G8 line, with the fitted timecourse data displayed in Figure 9 and the selection coefficients in Figure 10. The fit appears to be quite good, with the logistic regression curves closely tracking the mutation trajectories. Of course, we are fitting a curve of four datapoints with a model that contains two free parameters, so we expect a pretty good fit. Nevertheless, when compared to a fit using a smoothing spline[1] (red line), the two curves track closely together, suggesting that the logistic regression curves are in most cases not over-fitting the data. A more complete test of the quality of these fits will be possible when we have the additional timepoints that are in the process of being

[1]The spline fit is forced piecewise linear because there was not enough data to fit all the necessary parameters for a full cubic spline.

sequenced 5. When these data become available, we will also be able to fit models for all four lines of the experiment. Despite a small number of observed data points, the concordance between observation and prediction is an exciting validation of theoretical prediction, and underscores the predictive power of simple, quantitative arguments in population genetics.

3.6.3 Selection Coefficients

Quantitative modeling of selection and drift in our study strongly suggests that selection is the force driving substitution. Once we can quantify the strength of this selection using selection coefficients according to Haldane's theory, the next question becomes: How do the alleles actually confer the fitness advantage? While we can't answer this without performing more experiments, it is worthwhile to speculate, since some possibilities are more likely than others, and the discussion highlights some weaknesses of our model.

There are two very likely places in which the fitness advantage may lie: The first is in a simple growth advantage unique to the conditions of the assay, and the second is chloroquine pressure. As an example, let us consider the case of a large chromosomal deletion. The loss of a chromosomal end containing approximately 100 kilobases of material represents 0.5% of the total genome size. If we assume that DNA synthesis is a rate-limiting step replication, then a reduction in chromosomal size should result in an increase in fitness. How large an increase? The selection coefficients we measured were roughly 3%. This is larger than the 0.5% increase in fitness if it is assumed that a single cycle of replication sets the rate of replication, which suggests that a) eliminating DNA which is not necessary for *in vitro* growth does not account for the full, observed fitness advantage or or b) replication time depends non-linearly on genome size.

If the fitness advantage is a growth advantage, then this may reside in other places as well. As mentioned, we saw a large proportion of coding change in genes that code for proteins. These genes, such as Pf3D7_1002500, Pf3D7_1023700, and Pf3D7_1243900 (Table 2) for the most part have not been studied and do not have a function as-

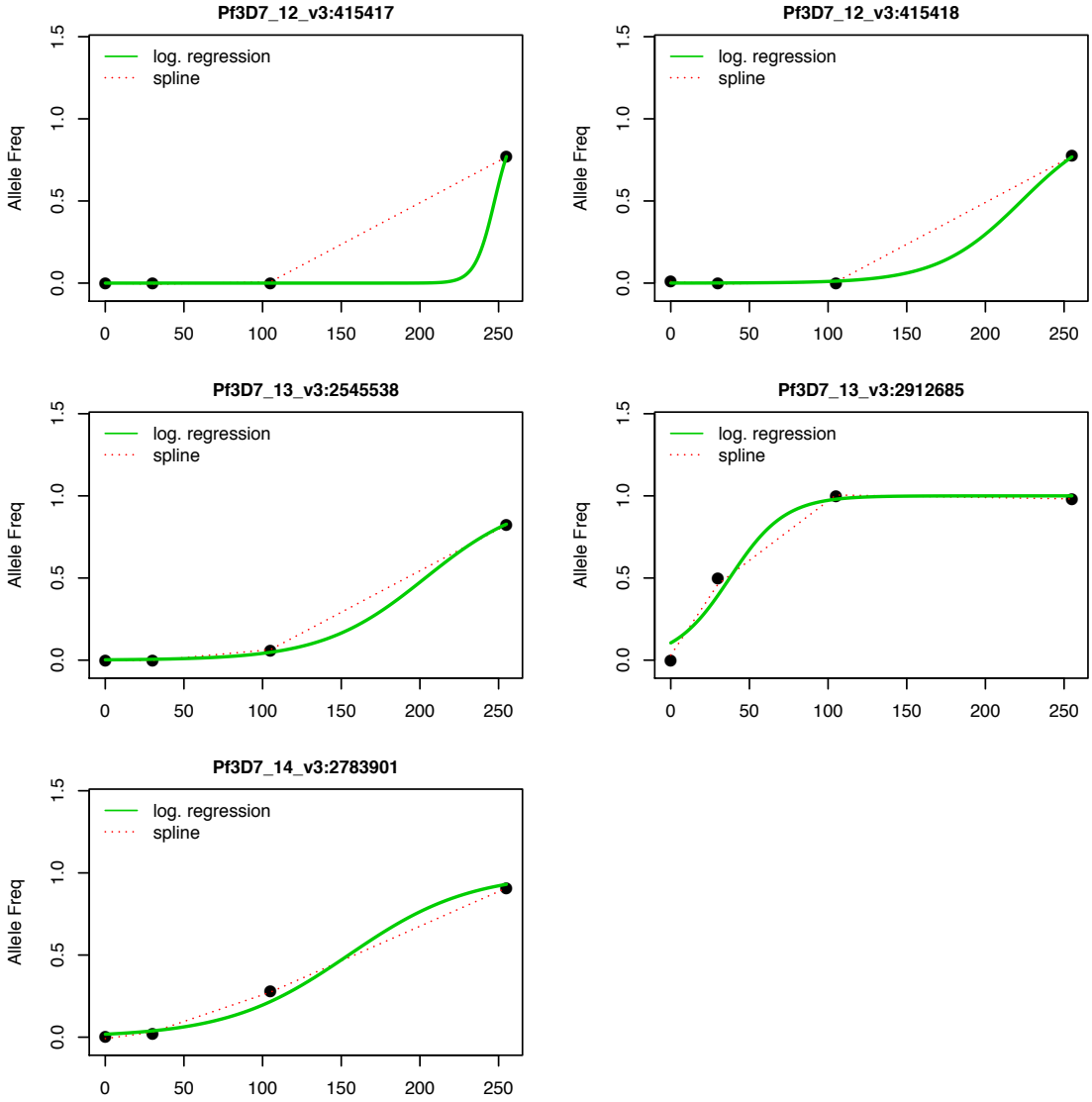


Figure 9: Fitted Allele Dynamics for the SNPs from 7G8 samples. The allele frequency data vs. time (measured in generations) have been fit with logistic regression models. A smoothing spline has also been fit (red line), but because of the number of data points, only a piecewise linear fit is possible.

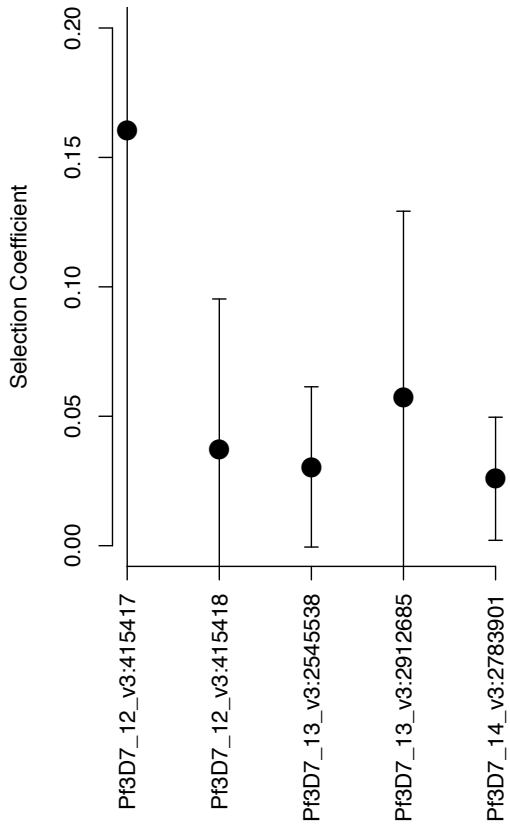


Figure 10: Estimated selection coefficients from the logistic regressions fitted in Figure 9. Error bars represent the standard errors of the regression coefficients.

signed through experiment or bioinformatic inference. Or, non-coding mutations, such as Pf3D7_12_v3:415417/8 in 7G8 and Pf3D7_12_v3:433914/15, may affect gene regulation. For the mutations detected under chloroquine pressure, an obvious hypothesis is that these substitutions confer additional drug resistance to the parasite. While 7G8 and TM are both already chloroquine resistant (which allowed us to grow them in low-dose chloroquine pressure), they may have acquired ability to grow in higher concentrations of drug. Whether any of the mutations that arose under chloroquine pressure yield further resistance can be tested by experimentally by measuring the sensitivity of these parasites to chloroquine, a test that we plan to carry out. Overall, identifying the mechanism by which the selective variants confer fitness advantages will be an exciting avenue of future study.

3.6.4 Periodic Selection and Passenger Mutations

Thus far we have estimated the rate of substitution and assessed evolutionary forces which are driving nucleotide substitution in our experiments. We observed that drift was very unlikely to fix mutations, and that therefore selection was likely to be playing a role, an observation that was consistent with the close fit of experimental measurements to theoretical prediction. There are, however, other ways for variants to increase in frequency, and we end by considering those.

The selection coefficients we estimate in Figure 10 are for a variant and all its linked alleles changing frequency in a population. For example, if a mutation occurs that confers a selective advantage and increases in frequency, all the alleles which are linked to that variant will also increase in frequency in the population. Therefore, a neutral or weakly selected variant can "hitchhike" to higher allele frequencies simply by being linked to a selected variant. Before recombination has had time to break down the association between the two alleles, they will be in 'linkage disequilibrium.' In recombining populations, the selected allele and linked variants will leave a signature of decreased genetic variation along a stretch of DNA, a signature which forms the basis of tests for recent

positive selection [73]. A haploid genome without recombination represents a particular extreme of this situation, because, with no recombination, linkage disequilibrium will persist indefinitely and does not decay along the length of the genome. In other words, the unit of selection is the haplotype and not the allele, and because there is no recombination, the length of haplotype is the length of the genome. This calls into question the notion of assigning a selection coefficient to a particular allele. It is probably more correct to consider that the genome has an effective selection coefficient, $\bar{s}$, which is the average of the selection allelic selection coefficients, s_i:

$$\bar{s} = \frac{\sum_i^n s_i}{n}.$$

Therefore, all we can say is that the genome which contains a variant is under positive selection, and, given a survey of variants at a single point in time, it is impossible to say which one is driving selection (or whether it is a combination of multiple variants). If the other mutations are neutral, i.e. all $s_i = 0$ except one, this phenomenon is known as 'periodic selection' in the bacteriological literature [2] or the concept of 'driver' and 'passenger' mutations in the cancer literature.[2]

Two things should distinguish between the 'driver' and 'passenger' mutations. The first is that driver mutations should precede, in time, passenger mutations, but that otherwise the changes in allele frequency should be correlated. This suggests that improved temporal resolution should help distinguish between these two classes of data.

A second difference is that passenger mutations should attain stable values intermediate fixation frequencies, i.e., equilibrium values of allele frequency which are not 0 or 1—the only stable allele frequencies for alleles under selection in a haploid population. This is in contrast to Haldane's classical theory of directional selection, which has stable equilibria solely at allele frequencies of 0 or 1. With periodic selection, stable equilibria exist for all values between 0 and 1, but not all values are equally likely. We can derive the

[2]Forces of this type are sometimes referred to, whimsically, with the term "genetic draft" (to emphasize the random nature of these forces, analogous to genetic drift) [29].

distribution of passenger mutation frequencies at equilibrium with a simple argument. A passenger mutation that begins earlier will rise to a higher allele frequency, and will stop changing in frequency once selective pressure has been removed from the driver mutation (i.e. it has reached allele frequency of 1). Thus, passenger mutations that occur soon after driver mutations will rise to high allele frequencies, whereas those that occur later on will rise to lower allele frequencies. Recall that Haldane's theory of selection yields the logistic function for directional selection in a haploid population (setting C and $r = 1$):

$$X(t) = \frac{1}{1 + e^{-t}}.$$

The likelihood that a passenger mutation occurs and reaches equilibrium at value x depends on the time that the driver mutation is in the population with frequency $1 - x$. The probability that a mutation does not fix at a value of $1 - x$ is thus proportional to the speed at which the allele frequency crosses this value, or the derivative of $X(t)$:

$$X'(t) = \frac{e^{t}}{(1 + e^{-t})^2}. \tag{6}$$

Thus, the probability density of allele frequency for driver mutations can is

$$Z(f) = d(1 - X'(f)), \tag{7}$$

with chosen d chosen so that $Z(f)$ has total probability mass of 1. This is shown in Figure 11. In other words, passenger mutations are very likely to have allele frequencies than 0.25 or greater than 0.75, but much less likely to fix with intermediate frequencies.

One might argue that the mutations observed in this study are likely to be driver mutations and not passenger mutations since they increase in frequency to high allele frequencies, at or almost at fixation. An understanding of the distribution of equilibrium allele frequencies for passenger mutations (Figure 11) shows why this line of reasoning is false. Mutations which fix at low frequency are extremely difficult to distinguish from

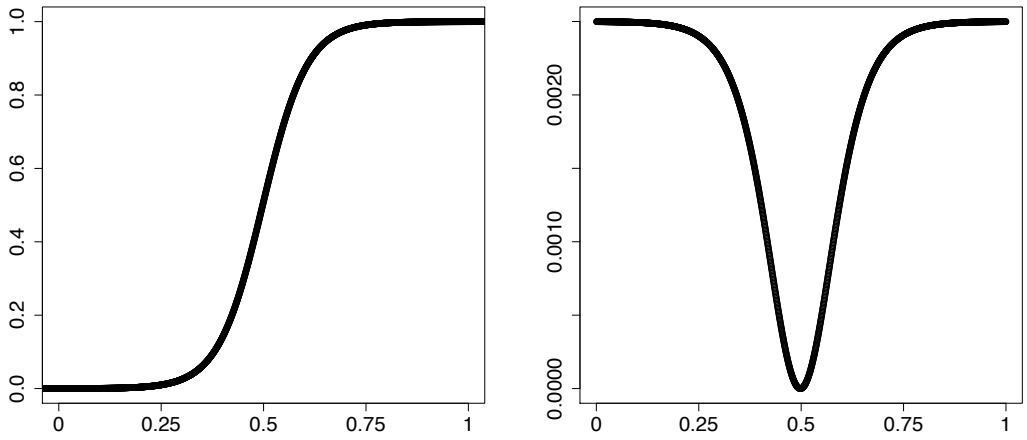


Figure 11: A plot of the logistic function (equation 5) is shown in the left panel, and the distribution of equilibrium allele frequencies of passenger mutations (equation 7) is shown in the right panel.

sequencing noise. Therefore, of passenger mutations which have risen above a certain threshold which can be differentiated from sequencing noise (e.g. 0.3), the vast majority will be at or near fixation (Figure 11). Thus, the fact that mutations have risen to a high frequency does not mean they are likely to be the targets of positive selection; they may do so simply by hitchhiking on selected variants.

In addition to clarifying the forces which drive substitution in our experiment, a second benefit of considering periodic selection as a force is that it suggests an alternative way to estimate the mutation rate. If an allele with a relatively large, positive selection coefficient, s_1, is added to a large population, then the rate of accumulation of passenger mutations should give an accurate assessment of the mutation rate. This selection coefficient of other alleles are likely to be substantially less than s_1, and therefore not alter the trajectory of fixation of the first allele; in other words, they will behave as approximately neutral.[3] The rate of accumulation of these mutations at allele frequencies greater than a given threshold should be proportional to the mutation rate, μ. We do not pursue this

[3]This may explain the close fit to Haldane's directional selection model, since trajectories of the form of equation 5 are not required under a model where the average selection coefficient can change with time as new mutations occur along the haplotype.

here; in order to derive an estimate of μ using this method, we would need to know the point at which the allele strongly selected allele appeared in culture. Nevertheless, such an approach may prove useful in a future experimental setting where the timing and initial allele frequency of a selected allele can be controlled.

3.6.5 Direct Estimation of the Mutation Rate

The experiments outlined in this thesis produce an estimate of the substitution rate, ρ, which, as discussed, is distinct from the mutation rate, μ. While this is relevant measurement to make because relative comparisons between substitution rates reflect relative mutation rates, the absolute value of the mutation rate in *P. falciparum* remains a quantity of importance because it defines the rate at which new alleles enter the population. The mutation rate is, as yet, unmeasured. In this section we describe a set of experiments, currently ongoing, aimed at estimating μ.

The number of mutations that enter a population in a single generation is the product of N, the number of alleles in the population (for haploid organisms), and μ, the mutation rate, typically given with units describing the rate per-site, per-generation. If we call X the random variable which describes the number of mutations per generation, then for m independent sites (bases), the average number of mutated alleles entering the population over t generations is

$$E[X] = mtN\mu. \tag{8}$$

If we consider a population of a single individual and use the value we observed for the substitution rate in this experiment, 1×10^{-9}, we expect an average of 0.023 mutations per generation. This is likely an underestimate, however, since we will only see substitution of positively selected alleles, and most new alleles are neutral or deleterious. Thus, we expect μ to be a little higher, perhaps in the neighborhood of 5×10^{-8}. In this case, we will see on average 0.115 mutations per generation.

From equation 8, we see that the rate at which mutations accumulate in a population

of 1 individual depends on the number of generations, t, the number of sites, m, and the rate mutation rate. The mutation rate is a fixed quantity (we assume), and thus there are two ways to increase the number of mutations observed: 1) to grow parasites for a larger number of generations, or 2) to increase the number of sites studied.

Recently, increasing the number of sites studied has become a straightforward thing to do with massively parallel sequencing. All that it requires is to sequence the genome of multiple different subclones of an isolate. Thus, in order to study mutation rate, it is not necessary to propogate a single individual for multiple generations (a time-consuming and technically difficult propostion); instead, one can propagate multiple individuals for a single generation, and perform the sequencing.

A large number of clones can be sequenced in a single lane, since allele frequencies will be only 0 or 1, all mutations which entered the population having fixed. With the introduction of massively parallel sequencing, this is now feasible to do. For example, with a baseline rate of 5×10^{-8}, if we sequenced two lanes of 48 samples / lane, we would expect to see 11 mutations; if the 2.7 fold increased rate for the Southeast Asian parasite studied holds true, we would expect to see 30 mutations. Sequencing on this scale is now relatively inexpensive to obtain and offers a direct estimate of the mutation rate as well as an increase in experimental power.

3.6.6 Clonal Interference

Figure 13 shows the allele frequencies through time of all the identified SNPs in this study. There are two notable features: the first is that only some of the allelic variants conform well to the directional selection model predicted by the Haldane theory, as evidenced by the relatively poor fit to the logistic regression line. The decreases in allele frequency of a positively selected variant are actually quite surprising in the context of this theory.

The second feature is the significant autocorrelation among SNPs (Figure 12). Changes in allele frequency of SNPs occur together and in concert because they are perfectly linked. There is assumed to be no recombination during asexual replication or chromo-

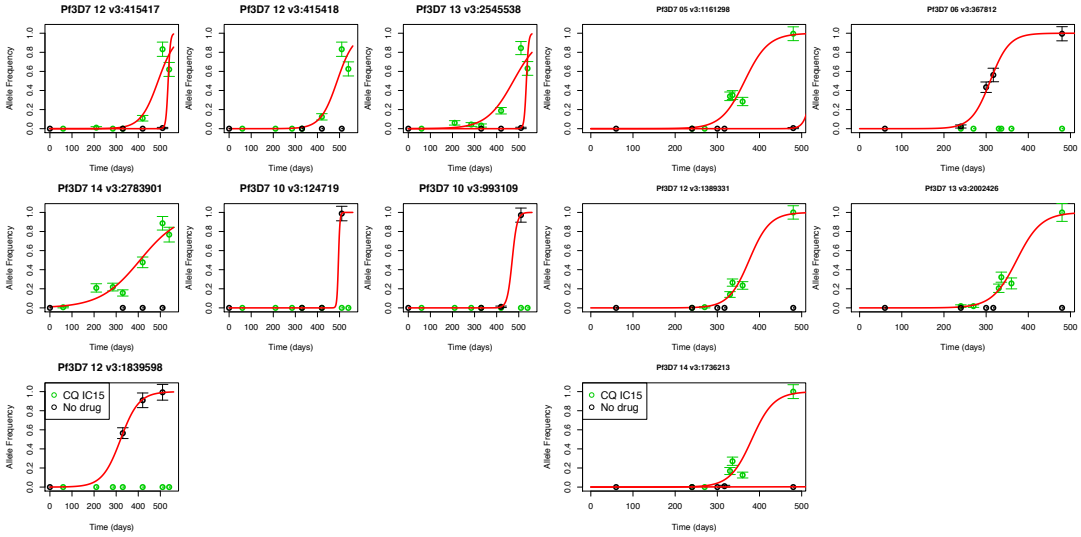


Figure 12: Allele Frequencies Through Time, 7G8 and TM284 strains

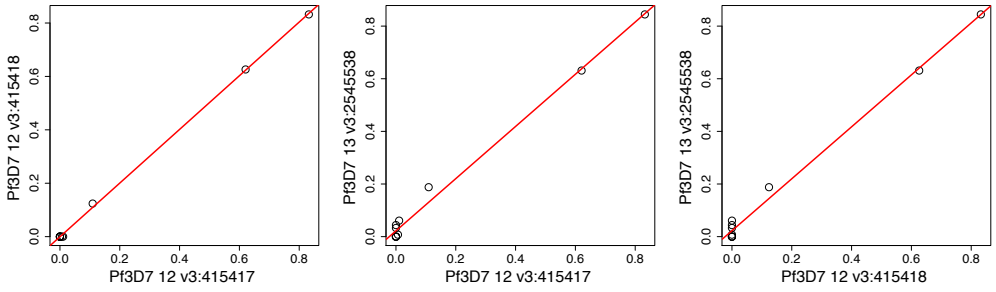


Figure 13: Autocorrelation among SNPs identified in this study

some reassortment, as discussed in the section on periodic selection. The most important feature is that the autocorrelation remains in tact when the SNPs are decreasing in allele frequency. Sampling "noise" introduced into the data will provide fluctuations in allele frequency and prevent a perfect fit from the directional selection lines even if there is perfect concordance with the Haldane model; however it is unlikely to have the same affect on all alleles in the sample, as is the case here. Therefore, the effect is not likely to be due to noise.

The most likely explanation for these two phenomena is clonal interference. Recall that the unit of selection is the genome of a single parasite, so that selective pressures acting on variants will be averaged for all variants in the genome. What is likely to have occurred is that a negatively selected variant has arisen on one of the parasite lineages which was initially undergoing positive selection and this variant and its progeny are in the process of being removed from the population. This process is known as clonal interference and has been observed to occur in several populations during experiments in which selection is monitored in real time. While clonal interference and periodic selection are related, the essential difference between clonal interference and periodic selection is that deleterious variants are *decreasing* the frequency of positive variants.

An alternate explanation is that there is indeed correspondence to the directional selection for an unlinked allele, and there has been sample mislabeling. This appears less likely as the phenomenon is consistent across multiple culture conditions.

The clone driving the dip in allele frequency–the interfering clone–should be detectable in theory, but it never rises to an appreciable frequency because it only causes a small decrement in allele frequency of the positively selected clone. Therefore, we were are unable to distinguish the interfering clone from noise at present, and clonal interference is strongly suggested by the dynamics of allele frequency changes but is not confirmed.

3.7 Can observed differences in substitution rate account for the ARMD phenotype?

The rate differences we observe in this experiment as a function of genetic background are on the order of three-fold; given the size and statistical power of the experiment, we cannot say that they are statistically different at all. Perhaps a better way to frame the hypothesis is the following: is there evidence that there is a rate difference large enough to account for the ARMD phenotype? In this case, the answer is more clearly that there is not. The acquisition of resistance occurs 100 - 1000 times faster in ARMD parasites; we see no evidence for rates this high, even in the presence of chloroquine pressure, a documented mutagen. Thus, the question remains unsettled as to whether there is a difference in rates at all, a more important question, whether the basis of the ARMD phenotype rests on differences in mutation rate, has been definitively settled by these experiments. It does not.

How then, does the ARMD phenotype work? Is it real? These are legitimate questions at this time. To some extent, the finding has not been compellingly replicated in the literature, and multiple lines of evidence are beginning to accumulate against it. These include this study showing that there is no difference in mutational rate substantial enough to account for the rate of increase in speed at which resistance evolves, as well as polymorphisms data from a large collection of geographically diverse *P. falciparum* isolates which show no relevant increase in polymorphism for southeast Asian parasites [4], and recent cloning experiments showing largely equal number of mutations collecting along clone trees as a function of genetic background [8].

3.8 Results Summary

We have used second generation sequencing methods to perform a census of variants accumulated during the course of the experiment, estimated the rate of substitutions per site, per generation, and tested whether genotype and drug pressure alter this rate.

We then examined the evolutionary forces that have caused substitution in our sample. Based on the dynamics of mutation accumulation, it is clear that 1) genetic drift is not acting at an appreciable rate in this experiment, 2) the alleles identified are under positive selection, and 3) the absence of recombination and therefore the presence of infinitely linked hitchhiking variants means we cannot localize the specific allele under selection, since the unit of selection is the entire genome; as a part of this, we demonstrate autocorrelated deviation from simple selection model, consistent with clonal interference. Finally, we argue that while the substitution rate may be slightly variable as a function of genetic background, any possible elevation in rate that is observed is not sufficient enough to account the ARMD phenotype in a lineage documented to possess it.

4 Genetic Variation in *Babesia microti*

4.1 Introduction: Human Babesiosis

Human babesiosis is a parasitic infection caused by protozoa of the genus *Babesia*, of which there multiple species that infect humans. Among these, *B. divergens* and *B. microti* cause the majority of clinical disease [31]. Both are zoonoses, transmitted to humans by ixodid ticks. *B. divergens* occurs predominantly in Europe, where it is enzootic in cattle and *I. ricinus* ticks and causes a rare but frequently fatal zoonotic infection in humans, particularly in asplenic individuals [39]. *B. microti* predominates in the United States, where it is enzootic among the rodent populations of the Northeast and Midwest, particularly *P. leucopus*, and *I scapularis* ticks. Infection with *B. microti* typically causes a milder illness marked by influenza-like syndromes and a hemolytic anemia [84]. Other *Babesia* species also infect humans, including the WA1 parasite endemic in the Pacific Northwest [66], the KO1 parasite in Korea [42], *B. venatorum* [85], a *B. divergens*-like organism [37], and several others [39]. Less is known about these parasites, and many have not yet undergone formal phylogenetic classification.

Like malaria, *Babesia spp.* are members of the phylum *Apicomplexa*, so named for their possession of a plastid organelle known as an "apicoplast". Babesiosis and malaria cause overlapping but distinct clinical syndromes. Shared features of pathobiology include clinical presentation with hemolytic anemia, fever, fatigue, malaise, thought to be secondary to massive cytokine release [47]; a spectrum of illness severity ranging from asympomatic infection to fulminant disease and a case fatality rate in immunocompetent hosts of 1-5% [84]; chronic infection involving persistent low-level parasitemia after resolution of the acute episode [44]; antigenic variation mediated by switching of multi-copy gene families [39]; and shared sensitivity to some classes of antiparasitic drugs [55]. Key differences include the apparent absence a liver stage in the *Babesia* parasites, subtypes of clinical disease in malaria—most notably cerebral malaria—and distinct mechanisms of organ injury. Cytoadherence, a major component of *P. falciparum* pathogenesis, has

not been documented in human babesiosis [9], though ocular manifestations of babesiosis with histological features of vasoocclusion raise this possibility [64]. The arthropod vector differs as well: anopheline mosquitos for malaria, ixodid ticks for babesia.

Human babesiosis due to *B. microti* has emerged rapidly over the past 50 years in the Northeast, with an accelerating spread over the past two decades [46, 49, 84]. While initial cases were predominantly limited to Nantucket and a few other coastal areas, infections are now reported throughout the Northeast from Maryland to Maine [1, 38]. The reasons for this are partly known: *Babesia* parasites possess a complex lifecycle directly involving both the *Ixodes* tick and the white-footed mouse, *Peromyscus leucopus*, and also indirectly involving the deer and Lyme disease spirochete, *Borrelia burgdorferi* [36]. The recent population expansion is often attributed to the increase in deer and tick populations, but other factors are important as well. Infection of *Ixodes* tick by *B. burgdorferi* increases the susceptibility of infection with *B. microti* [22], contributing to its spread. Also unknown is the effect of evolution by parasite itself, which may be adapting to spread in the Northeast or increasing its ability to infect the human red blood cell.

Recent years have brought not only an appreciation of an expanded geographic range for *B. microti*, but also an increased recognition of illness severity. Persistant parasitemia and relapsing disease are common, particularly among immunosuppressed individuals, where the case fatality rate is between 10 - 20% [48]. Resistance to first-line therapy is now commonly reported [94], and the second line therapy is a quinine-based regimen which is difficult to tolerate. Babesiosis is also a major threat to the blood supply, since donors are not screened and asymptomatic infections and transmission has been documented through donated blood products [18]. In 2012, babesiosis was the leading cause of transfusion-associated infection related fatalities, accounting for 38% of such deaths [26]. Co-infection with *Borrelia burgdorferi* and or *Ehrlichia chaffeensis* is common and results in illness of greater intensity and duration [45].

Due to its rapidly increasing case load and a greater appreciation of its severity,

human babesiosis is classified as an emerging infectious disease by the CDC [6]. Although there has been a consistent history of small case series and case reports since the 1960s documenting the initial cases of *Babesia microti* infection in the Northeast USA, little systemic study has been performed, and many questions remain unresolved. These include, for example, the mechanisms by which persistent parasitemia occurs, formal confirmation of suspected antigenic variation among *BMN* genes, the genetic and molecular mechanisms of drug resistance, and the processes by which the parasite causes organ-injury, including in the eye, lung, kidney, and liver.

In order to address the genetic basis of important *B. microti* phenotypes, we undertook a study of parasites sampled directly from cases of clinical infection. The goals of this work were to establish a map of genetic diversity and to use patterns of variation to characterize the population structure, understand demographic history of circulating strains, and identify genes under natural selection.

4.2 Clinical Description of Human Babesiosis Cases

Patients were enrolled between July 2014 and October 2014. Five cases in total were enrolled. Enrollment is ongoing; this report describes the first five clinical cases along with a cohort of enzootic samples from *I. scapularis* ($n = 2$) and *P. leucopus* ($n = 3$). Enzootic samples were contributed by the Telford laboratory at Tufts School of Veterinary Medicine. The clinical details of the cases are presented in Table 5. All enrolled patients were inpatients with moderate-to-severe babesiosis. The clinical courses and laboratory values were similar to those in reported case series [35, 70]: Patients were predominantly male, and presented with fever, anemia, hemolysis, thrombocytopenia, evidence of kidney and liver injury, and frequently also hyponatremia.

Patients were treated with one of two regimens: Clindamycin 600 mg intravenously every 6-8 hours with quinine orally every 8 hours, or atovaquone 750 mg orally twice daily with azithromycin 500mg on the first day and 250 mg orally once daily thereafter. All patients recovered from the acute illness. There was one relapse: Patient 5 had

Patient	%Pt	Sex	Age	Hct	Plt	Na	Cre	ALT	AST	Tbili	LDH
1	13.6	M	75	27.6	37	124	1.41	56	178	4.5	1406
2	40.0	M	53	27	28	136	0.97				
3	4.3	M	71	32	47	127	1.2	79	132	1.4	978
4	1.0	M	84	24.7	91	127	1	33	57	1.8	
5	21.7	F	49	29.3	113		1.68			575	
Mean	19.9		66.4	28.1	63.2	128.5	1.3	56	122	2.6	986

Table 5: Clinical characteristics of the patients upon initial presentation. %Pt - percent parasitemia, Hct - hematocrit, Plt - platelet count, Na - serum sodium, Cre - serum creatinine, ALT - alanine aminotransferase, AST - aspartate aminotransferase, LDH - lactate dehydrogenase.

presented in July, at which time she was initially treated with atovaquone/azithromycin. She was maintained on this regimen until October, when atovaquone was stopped and azithromycin was continued as monotherapy. She returned in November with severe babesia and a parasitemia of 21.7%, apparently resistant to azithromycin mono-therapy. At the time, she was successfully treated with clindamycin and quinine. Given the absence of blood transfusions during this interval and the lack of questing nymphal *Ixodes* forms during the late fall, this was almost certainly a recrudescence rather than a new infection.

4.3 Sequencing of *B. microti* Isolates and Variant Calling

Parasite genomic DNA was isolated (see Section 2) after human leukocyte depletion, and Illumina sequencing libraries were prepared and analyzed on a HiSeq 2500 instrument. Reads were aligned to the *B. microti* genome using BWA. Read and alignment statistics are summarized in the Table 6. We obtained high quality ($>$ 15X) coverage of all 6.5 Megabases in the *B. microti* genome for four of the five clinical isolates. One of the patient samples and the five enzootic samples had lower quality coverage, with a range of 0.5 - 2.5x.

SNPs were called using GATK's `UnifiedGenotyper` (Methods), with a quality threshold of 50. For all SNPs, a minimum of 70% of reads covering had to support the variant SNP in order for a call to be considered valid. Computationally annotated variants were then confirmed by manual inspection in IGV. From this pipeline, we identified a set of 52

Sample	Reads	Mapped	%Mapped	%Paired	MBases	Mean Depth
Bab01	102287300	164111	0.2	87.3	16.4	2.5
Bab02	108883706	3798440	3.5	95.3	379.8	58.4
Bab03	95302696	1792396	1.9	94.8	179.2	27.6
Bab04	85119820	676954	0.8	96.2	67.7	10.4
Bab05	117484738	36294502	30.9	96.5	3629.5	558.4
Tufts01	23070702	99290	0.4	72.0	9.9	1.5
Tufts02	15038188	108916	0.7	66.1	10.9	1.7
Tufts03	34643648	43713	0.1	84.4	4.4	0.7
Tufts04	30910094	37828	0.1	78.7	3.8	0.6
Tufts05	26077762	34266	0.1	82.8	3.4	0.5

Table 6: Read statistics for the samples in this study. Reads - the total number of reads that passed quality control filters; MBases - total *B. microti* sequence generated in megabases (1×10^6 bases); Mean Depth - mean coverage of the *B. microti* genome.

single nucleotide polymorphisms from the whole genome sequence of four babesia isolates for which full length genome sequence was available.

The SNPs are recorded in Table 7. Summary statistics, including the number of SNPs in non-coding and protein-coding regions of the genome, are given in Table 8. SNPs in protein-coding are annotated as those that do not change the amino acid sequence (synonymous) or those that result in an amino acid change (non-synonymous). The distribution of SNPs among the parasite chromosomes in given in Figure 14. There appears to be a uniform distribution of SNPs across the chromosomes, with the exception of the left arm of chromosome 1, which possesses an increase in the density of the number of SNPs. The increase in density in this region appears to be driven by an excess of substitution in the gene BBM_I00003 (see below; section 4.6), though it is also reminiscent of the increase nucleotide diversity seen in the telomeres of Apicomplexan parasites, including *P. falciparum* [86].

4.4 Analysis of Genetic Variation in *B. microti*

The whole genome sequence reveals an absence of genetic variation relative to other sequenced Apicomplexan parasites. The number of segregating sites, $S = 52$, is very

Chrom	Base	Major	Minor	Gene	Major AA	Minor AA	Type
mitochondria	2065	G	A				
mitochondria	2080	C	A,T				NC
mitochondria	3097	T	C	COX1	L		
mitochondria	5428	T	C	CYTB	L		
mitochondria	6311	G	T,A				NC
mitochondria	6326	C	T				NC
mitochondria	6574	T	C				NC
mitochondria	7317	C	T				NC
mitochondria	8598	A	G				NC
mitochondria	8840	G	A	COX3	L		
mitochondria	9210	C	T	COX3	H		
apicoplast	3566	G	A	LSU			RNA
apicoplast	4704	T	C	LSU			RNA
apicoplast	7864	G	A	RPL14	R		
chromosome 1	2081	A	T	BBM_I00003	F	I	NS
chromosome 1	2159	T	C	BBM_I00003	S	G	
chromosome 1	2207	G	A	BBM_I00003	L	F	NS
chromosome 1	2248	A	G	BBM_I00003	I	T	NS
chromosome 1	2347	A	G				NC
chromosome 1	7929	A	G				NC
chromosome 1	31817	T	C	BBM_I00095	D	D	?PG vs S
chromosome 1	66181	G	A	BBM_I00210			PG
chromosome 1	210944	G	T	BBM_I00580	F	L	NS
chromosome 1	345658	G	T	BBM_I00985	T	N	NS
chromosome 1	490358	G	A	BBM_I01355	A	A	S
chromosome 1	547252	T	C	BBM_I01510	Y	C	NS
chromosome 1	875249	C	T	BBM_I02415	E	K	NS
chromosome 1	1050798	G	A	BBM_I02895	L	L	S
chromosome 2	738797	G	T	BBM_II02040	L	F	NS
chromosome 2	878279	C	A	BBM_II02460	V	M	NS
chromosome 2	921987	T	C	BBM_II02600	V	A	NS
chromosome 2	949522	C	A				NC
chromosome 2	1000673	T	C	BBM_II02835	S	P	NS
chromosome 2	1403540	G	T	BBM_II04035	V	V	S
chromosome 3	227462	C	T				NC
chromosome 3	280886	C	A				NC
chromosome 3	809630	C	T				NC
chromosome 3	1146022	T	C	BBM_III03222			RNA
chromosome 3	1159413	G	T				NC
chromosome 3	1190774	G	A	BBM_III03370	A	V	NS
chromosome 3	1217919	C	T	BBM_III03455	G	D	NS
chromosome 3	1689654	C	T	BBM_III04695	V	I	NS
chromosome 3	1783033	A	T				NC
chromosome 3	1870912	C	T	BBM_III05175	V	V	S
chromosome 3	1942327	G	T	BBM_III05345	H	Q	NS
chromosome 3	2238421	C	A				NC
chromosome 3	2759627	G	T	BBM_IO7515	L	F	NS
chromosome 3	3133369	G	A	BBM_III08630	E	L	NS
chromosome 3	3133789	A	G	BBM_III08630	K	E	NS
chromosome 3	3304254	T	G	BBM_III09150	Y	D	NS
chromosome 3	3396025	G	T	BBM_III09465	R	M	NS
chromosome 3	3431866	T	G				NC

Table 7: SNPs identified in this study, classified by type. Abbreviations: NC - non-coding; S - synonymous; NS - non-synonymous; PG - pseudogene.

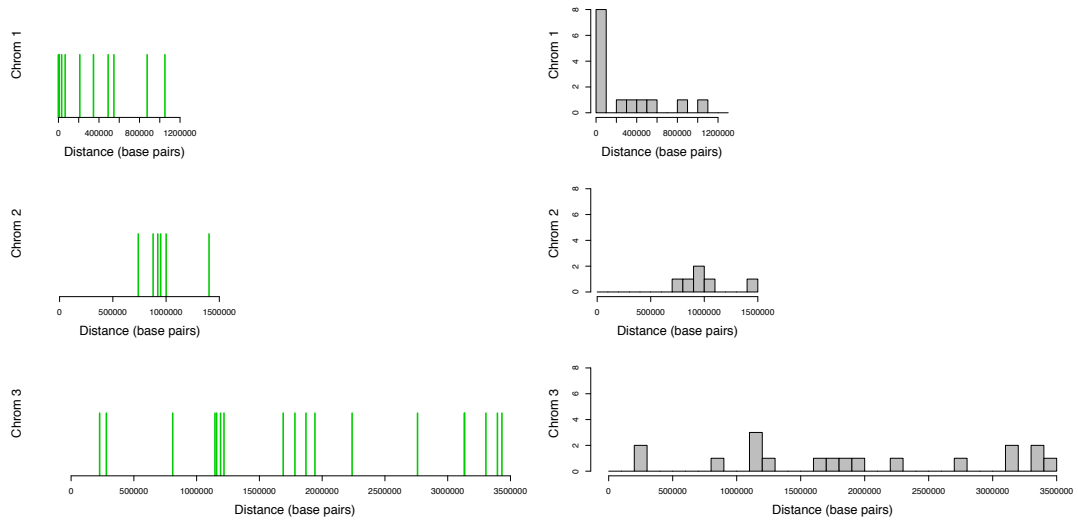


Figure 14: Genomic location (left) and density per in 100,000 bases (right) of chromosomal SNPs identified in this study.

small. S simply counts the number of sites in the sample where more than one variant is present; a disadvantage is that it depends on sample size. The mean number of pairwise differences between sequences in the sample, known as π, is also extremely small: We estimate $\pi = 3.1 \times 10^{-6}$. By comparison, π for *P. falciparum* is estimated to be approximately 0.003 [59, 60].

Limited genetic diversity can occur for several reasons, including population expansion, recent natural selection, population bottlenecks, population subdivision, and inadequate sampling. In this case, it is likely some combination of these, but the uniform nature of genetic variation throughout the genome, in concert with the epidemiological data [23, 46], makes it highly likely that the dominant force is recent common ancestry among the lineages sampled, consistent with a recent population expansion and/or population bottleneck.

We can evaluate the support for an emerging population more formally using population genetic models. The frequency of alleles in the population yields insight into demographic history. A rapidly growing population will have more alleles presents at low frequency, since variation is rapidly introduced into the population by mutation, but

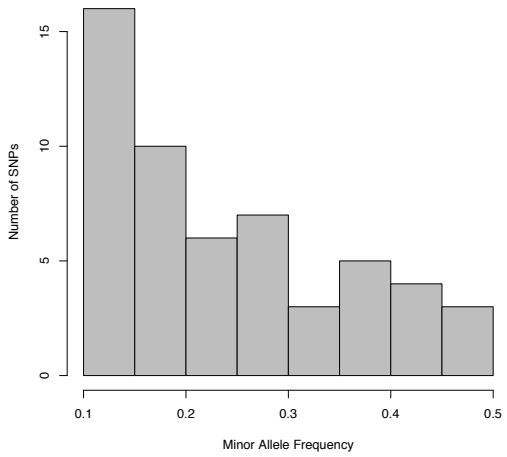


Figure 15: The folded site frequency spectrum for alleles in this study.

alleles take longer to reach fixation since the population size is growing [29]; the opposite will be true in a shrinking population. The site frequency spectrum is shown in Figure 15. By comparing the frequency of rare variants in the population, we can make statements about the neutral model and demographic history.

Watterson [87] showed that, in the absence of selection, the number of segregating sites, S, is related to the mutation rate, μ, and effective population size, N_E, by

$$\hat{\theta} = \frac{S}{\sum_i^{n-1} \frac{1}{i}},$$

where $\theta = 2N_E\mu$ for a haploid population. Tajima [79] noted that $E[\pi] = \theta$; under the neutral model, these estimates of $\hat{\theta}$ should be equal, but that deviations from neutrality will affect the two estimators differently. In the case of recent population growth with an excess of rare variants, the estimate of $\hat{\theta}$ using S will be increased; however, this will not affect the estimate of θ given by π. Tajima constructed a test statistic, D, which measures deviations from neutrality by comparing the estimates of θ obtained using π and S. The value of D is given in Table 8. In this case, Tajima's D statistic is -0.6 for the entire genome, indicating that there is an excess of rare alleles in the population,

	Chromosomal (n=4)	Mitochonrdial (n=10)	Apicoplast (n=4)	Genome (n=4)
Segregating Sites	38	11	3	52
Protein Coding	26	4	1	35
Synonymous	4	2	0	6
Non-synonymous	22	2	1	25
π	3.1E-6	3.55E-4	5.36E-5	4.19E-6
D	-0.53	0.08	-0.75	-0.569

Table 8: Summary statistics for the sequences in this study. The number of segregating sites, S, refers to the total number of positions at which a SNP was found in all sequences. An alternative measure of diversity is given by π, which measures the average fraction of sites which are discordant in pairwise comparisons between the sequences in the study. D is a test of neutrality.

suggesting recent population expansion or a selective sweep. The negative values of D are consistent with the epidemiologic and clinical data showing a great increase in case load; however, the study is underpowered with only four full-length genomes, and the values of D do not reach significance (a value of ± 2 is typically considered significant for D in this setting).

Sequence diversity can also be used to estimate divergence time between sequences if the mutation rate is known. Estimating sequence divergence times is a subtle and controversial proposition, and at present, the dataset we have available is not complete enough to justify a fully rigorous or complete model. Instead, we consider several straightforward approaches below to achieve an approximate calculation. The estimates should viewed as "ballpark" figures, but in this case, as the divergence is small, so is the ballpark, and the simple divergence estimates are useful.

At the most straightforward level, we can use an argument which is essentially dimensional analysis to derive a rough estimate. This also builds on the malaria experimental data above. If we assume that substitutions occur at a rate of 1×10^{-9} per site, per generation in Apicomplexan parasitic protozoa, then for parasites separated by, on average, 20 SNPs in the core chromosomal regions, (corresponding to a $\pi = 3.1 \times 10^{-6}$. We expect 0.0065 mutations to occur per genome per generation. For a generation time of 24 hours, then we expect 3100 generations to have separated the parasites in our study, or 8.5 years, leading to a mean estimate for divergence of 4.25 years ago.

This analysis admittedly ignores much complexity associated with sites under selec-

tion. However, the conceptual advantage of using an empirical substitution rate for the whole genome is that it avoids the breakdown into conditional mutational rates for selected sites, etc. This is one advantage of working with the substitution rate instead of the mutation rate. Much of the debate about using the mutation rate argues that it is only a good estimate of the substitution rate under the neutral model, therefore, one must focus on neutral sites. But if we empirically measure the substitution rate, then that discussion is considerably less relevant, because the substitution rate is what we use to calibrate our dimensional analysis. One still needs to justify the molecular clock assumption, and the estimate still rests on the (questionable) assumption of a similar rate among Apicomplexan protozoan parasites, but this is good enough to get us in the ballpark. Whether it is 4 or 40 years, there is no question that molecular sequence data place *Babesia* parasites sampled from this region within the last several decades.

An alternate estimate can be obtained by examining purely "neutral" sites such as those that occur in pseudogenes or non-coding regions. As expected, nucleotide diversity is greater in non-coding regions ($\pi_{nc} = 5.6 \times 10^{-6}$), consistent with functional constraint of protein coding sequences. If we assume a mutation rate, $\mu = 1 \times 10^{-9}$, and let $\mu = \rho$, assuming $N \gg t$ and neutral sites, then we expect $1.7 \times 10{-}3$ substitutions per generation. For the two most divergent sequences in the sample, Patients 3 and 4, this leads to divergence time of 19.0 years, or a time to most recent common ancestry of 9.5 years.

4.5 Population Structure

Several methods are available to compute the relatedness among samples. Most work by computing a distance measure between pairs of sequences. Pairwise distances are then stored in a matrix, D, which is hollow ($d_{ii} = 0$) and symmetric ($d_{ij} = d_{ji}$). The set of samples is displayed by one of several representations built from the distance matrix, e.g. as a collection of points in Euclidean space or as a tree where branch length is proportional to distance.

Multiple distance measures are available for sequences. We use the simplest, known as p distance, which involves counting the number of pairwise differences between sequences (equal to π for $n = 2$). This distance measure is appropriate for closely related sequences, but not for more divergent sequences where the possibility that single nucleotide positions have mutated more than once needs to be accounted for in the distance metric. Typically, p distance is appropriate for $p < 5\%$, a criterion that is clearly met for these samples. More complicated metrics that account for multiple substitutions at a single site or different transition probabilities between nucleotides are not necessary. These distances have been converted into a tree in Figure 16 by the neighbor-joining method [74]. "Best-fit" euclidean distance has also been calculated by Principal Component Analysis (PCA); technically, this amounts to a rank-2 approximation, $\hat{D}$ of the full-rank distance matrix D such that $\sum\sum_{ij}(d_{ij} - \hat{d}_{ij})^2$ is minimized.

The sequence distance data clearly shows stratification among populations in the Northeast. The samples from NH cluster together in the PCA, as do the samples from Millis, MA, and Bedford, MA, while the sample from Topsfield is more distantly related to both sites, mirroring the true geographic relationship . Thus, the sequence distance reflects geographic distance. This documents genetic structure in the population of sequences. It also raises important questions that we will aim to address in future work. For example, are zoonotic strains a random sampling of enzootic ones, or do some strains have an affinity for infecting humans? Are some some strains associated with increased virulence?

4.6 Azithromycin Resistance and Putative Sites of Selection

Recall that patient 5 (also referred to as Bab05) suffered a relapse of parasitemia three weeks after switching from atovaquone / azithromycin combination therapy to azithromycin monotherapy. While not proof that this parasite is "resistant" to azithromycin, *B. microti* parasites have shown exquisite sensitivity to azithromycin monotherapy in the hamster model [88], and azithromycin has been substituted with success for clindamycin in cases

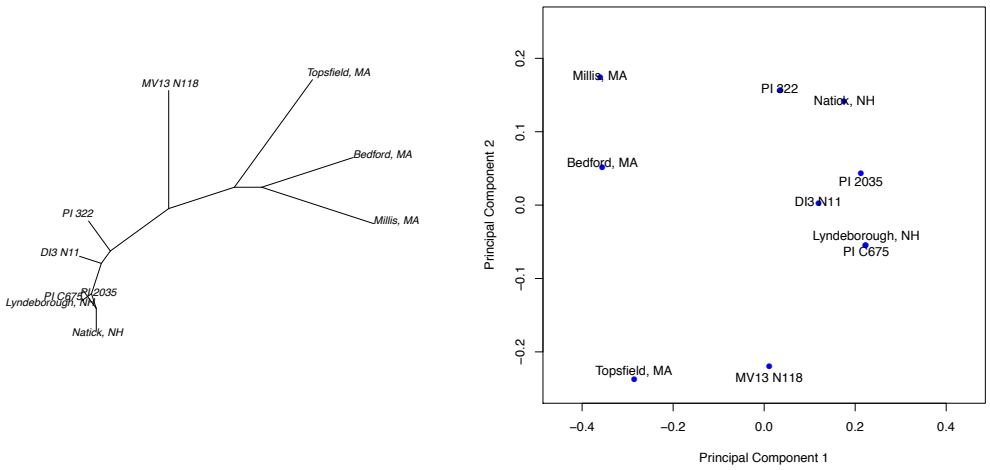


Figure 16: Unrooted, neighbor-joining tree (left) and principal component analysis (right) computed from pairwise distance matrix of mitochondrial sequences.

of quinine/clindamycin failure [75], suggesting that this efficacy as a monotherapy holds in humans. The alternative is that this treatment failure was due to subtherapeutic azithromycin levels. There is some evidence to support this: The initial animal trials did show a dose-dependent response to azithromycin, with prolonged hamster survival when higher doses of azithromycin monotherapy were used [88]. Furthermore, 250 mg orally once daily is a relatively low dose of azithromycin, with doses up to 1200 mg orally once daily given for toxoplasmosis, although trials of prophylaxis have shown 99% efficacy with the 250mg daily dose against *P. vivax*. Nevertheless, prolonged subtherapeutic dosing is in some ways just as interesting as outright high-level resistance, because sub-lethal dosing offers optimal conditions under which resistance can evolve: conferral a fitness advantage on those parasites able to grow optimally under of azithromycin pressure while ensuring that the flux of mutant alleles into the population remains high by maintaining population sizes. Thus, in either of these two scenarios — high level azithromycin resistance or subtherapeutic azithromycin dosing — we considered the possibility of having selected for variants that confer decreased sensitivity to azithromycin sufficiently high to search for such alleles.

Location	Base	Gene	Maj AA	Min AA	P1	P2	P3	P4	P5
apicoplast	7864	RPL4	R	H	G	G	G	G	A
chr1	2159	BBM_I00003	S	G	.	T	T	T	C
chr1	2207	BBM_I00003	L	F	.	G	G	G	A
chr1	2248	BBM_I00003	I	T	.	A	A	A	G
chr1	210944	BBM_I00580	F	L	.	G	G	G	T
chr4	1146022	BBM_III03222			T	T	T	T	C

Table 9: Non-synonymous mutations present only in the Bab05 isolate.

In order to identify potential genes involved in azithromycin failure, we filtered the SNP dataset for non-synonymous variants that were unique to Patient 5. This produced a list of six variants (Table 9). The first is in the RPL4 gene in the apicoplast. Four non-synonymous variants were present in BBM_I00003. The last variant was present in BBM_I03222, a coding RNA on chromosome 4 (this was counted as a non-synonymous change because the RNA was taken to be functional).

In RPL4, we identified a non-synonymous mutation unique to the Bab05 patient that replaces a highly conserved arginine with a histidine (Figure 17). This substitution was immediately striking because azithromycin resistance in *P. falciparum* malaria has been linked to RPL4, a component of the 50S ribosome [76]. Indeed, resistance in malaria is conferred by G→V substitution just 3 amino acids adjacent to the mutation in the Patient 5 isolate (Figure 17), a conserved region which functions as an azithromycin-binding pocket [76]. The likelihood that one of 6 non-synonymous substitutions would fall at the site of a highly-conserved, known azithromycin binding domain in a protein that has been experimentally confirmed to mediate azithromycin resistance in apicoplasts strongly suggests that RPL4 may be involved in azithromycin resistance in *B. microti*. Functional studies to test this hypothesis are underway at present.

We next examined the four non-synonymous mutations in the protein BBM_I00003. Three of these were unique to Patient 5. This protein is unannotated. It is located in the telomeric region of chromosome 1. BLASTp of the protein sequence shows homology H+/K+ exchangers as well as K+ transporters 18. The level of non-synonymous

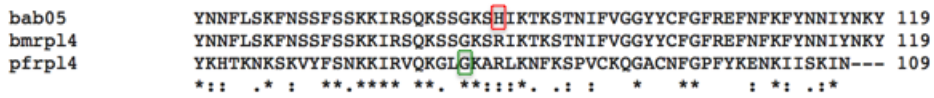
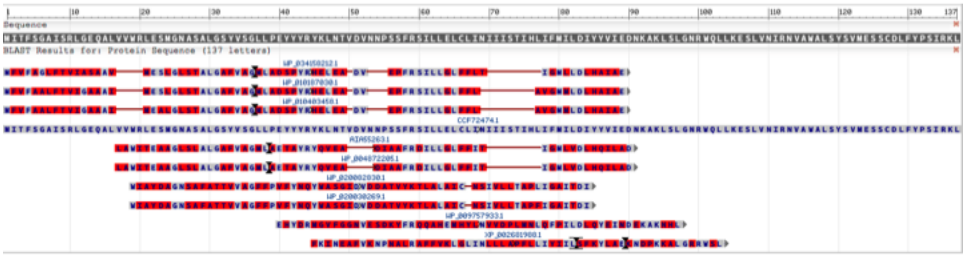


Figure 17: ClustalW alignment of RPL4 from *B. microti* and *P. falciparum*. The protein sequence from the mutated allele in the case of azithromycin failure is shown above. The R →H change is marked in red. A non-synonymous change alters a highly conserved residue in RPL4. This is 3 amino acids away from known azithromycin resistance mutations in *P. falciparum* (G → V; site marked in green) and *S. pneumoniae* (G→R, G→S; also, substitutions at the adjacent K,G,T – see [76]).

variation is quite striking relative to the rates of the genome in general, and this gene is responsible for the peak in diversity on the left hand side of chromosome 1 (Figure 14). In hypothesizing why this protein appeared to be under diversifying selection, it is important to note that Patient 5 had two unique characteristics. In addition to evidence of azithromycin inefficacy, this patient also experienced a chronic infection of five months duration, such as is typical in severely immunosuppressed or asplenic individuals [48]. Thus, two scenarios seem possible. BBM_I00003 may have be involved in azithromycin resistance, for example by facilitating export of the drug or altering the pH of a subcellular compartment so that the drug loses activity, and it may have evolved to improve this function under sustained azithromycin pressure. The clear homology to membrane proteins of diverse origin supports this. Alternatively, the protein may display excess nucleotide diversity because it is recognized by the immune system and evolves in order sustain chronic infection without immune clearance. This is a well-documented phenomenom among pathogens broadly [17]; in particular, the *Apicomplexa* frequently achieve this with multi-copy gene families located in telomeres [16,50], and the *BMN* family of proteins in *B. microti* is already suspected to be involved in this function [39]. The genomic location of BBM_I00003, the presence of homologous sequences within a single parasite (Figure 18), as well as the fact that one non-synonymous variant for BBM_I00003 is shared by an enzootic isolates, all argue in favor of this alternative scenario. A third possibility is that the variation in this gene is due to chance, but this seems unlikely given that the mean rate of polymorphism across genomic space is 1.25×10^{-6} bases.



Alignments Download GenPept Graphics Distance tree of results Multiple alignment

Description	Max score	Total score	Query cover	E value	Ident	Accession
unnamed protein product [Babesia microti strain RI]	275	275	100%	3e-92	100%	CCF72474.1
unnamed protein product [Babesia microti strain RI]	258	258	100%	1e-85	93%	CCF73165.1
unnamed protein product [Babesia microti strain RI]	215	215	83%	6e-69	94%	CCF73171.1
unnamed protein product [Babesia microti strain RI]	63.2	63.2	87%	6e-09	31%	CCF74728.1
autophagy 9 [Naegleria gruberi strain NEG-M]	42.4	42.4	43%	0.049	37%	XP_002681988.1
sodium:proton exchanger [Sphingomonas sp. ERG5]	37.7	37.7	64%	1.6	33%	WP_034158212.1
hypothetical protein [gamma proteobacterium SCGC AAA076-P09]	37.4	37.4	48%	2.1	29%	WP_020082830.1
sodium/hydrogen exchanger [Sphingomonas sp. PAMC 26605]	37.4	37.4	64%	2.3	32%	WP_010187030.1
hypothetical protein [gamma proteobacterium SCGC AAA076-P13]	36.2	36.2	48%	6.1	27%	WP_020030269.1
hypothetical protein [Ahrensia sp. R2A130]	35.8	35.8	42%	7.1	34%	WP_009757933.1
sodium/hydrogen exchanger [Sphingomonas echinoides]	35.8	35.8	64%	7.3	31%	WP_010403458.1
Glutathione-regulated potassium-efflux system protein KefB [Acidithiobacillus caldus ATCC 51756]	35.8	35.8	54%	8.0	32%	AIA55263.1
potassium transporter [Acidithiobacillus caldus]	35.8	35.8	54%	8.3	32%	WP_004872205.1

Figure 18: BLASTp search results for BBM_I00003 showing homology to Na+/H+ exchangers and K+ transporters. BBM_I00003 contains 4/26 (15%) of the total non-synonymous SNPs identified. Mutations occur only in the clone that failed azithromycin monotherapy.

A third genomic feature bears mention. We searched for large insertions and deletions such as those known to occur in malaria by performing a visual inspection of the coverage along chromosomes in each sample. We identified only one such large insertion, which is a three-fold amplification of a 15KB region on chromosome 2, starting at base 658,075 and extending to base 672,981, which appears in the isolate from Patient 5 (Figure 19). This region includes several unannotated genes, the largest of which is BBM_II01850, an ABC transporter with homology to multi drug resistance associated proteins (MRP) in other species (Figure 20). ABC transporters are frequently involved in drug resistance in malaria [65] and other organisms [51], and only three are annotated in the *B. microti* genome [11]. This insertion event is reminiscent of the amplifications in MDR1 in *P. falciparum* in response to multiple drugs [28, 40]. Amplification of this locus in Patient 5 also suggests a possible role in azithromycin resistance, though integration of low coverage in Patient 1 (the most closely patient isolate related to Patient 5; see Figure 16) suggested that the locus may also be present there in triplicate. One intriguing possibility is that *B. microti* employs a genomic strategy similar to *P. falciparum* to produce resistance, in which non-specific duplications of multiple loci are generated stochastically in response to initial selective pressure; higher level resistance then evolves through more precise head-to-tail amplification of the "founder" unit [32]. In this regard, copy number variation at the *B. microti* multidrug related protein (MRP) locus may represent one such substrate onto which higher level resistance could be crafted. As for the above two cases, RPL4, and BBM_I00003, the possibility also exists that this genomic variant is unrelated to azithromycin resistance, which seems particularly likely if this amplification in Patient 1 is confirmed. In that case, however, the presence of copy number variation in ABC transporters in the circulating reservoir of genetic polymorphism underscores the value of understanding the genetic structure of *Babesia* populations, and reminds us that clinical phenotype may depend on population differences in parasite genotype.

We close this section with a few words about unbiased screens for adaptive evolution. Natural selection leaves signatures in the genome that can be used for to detect loci puta-

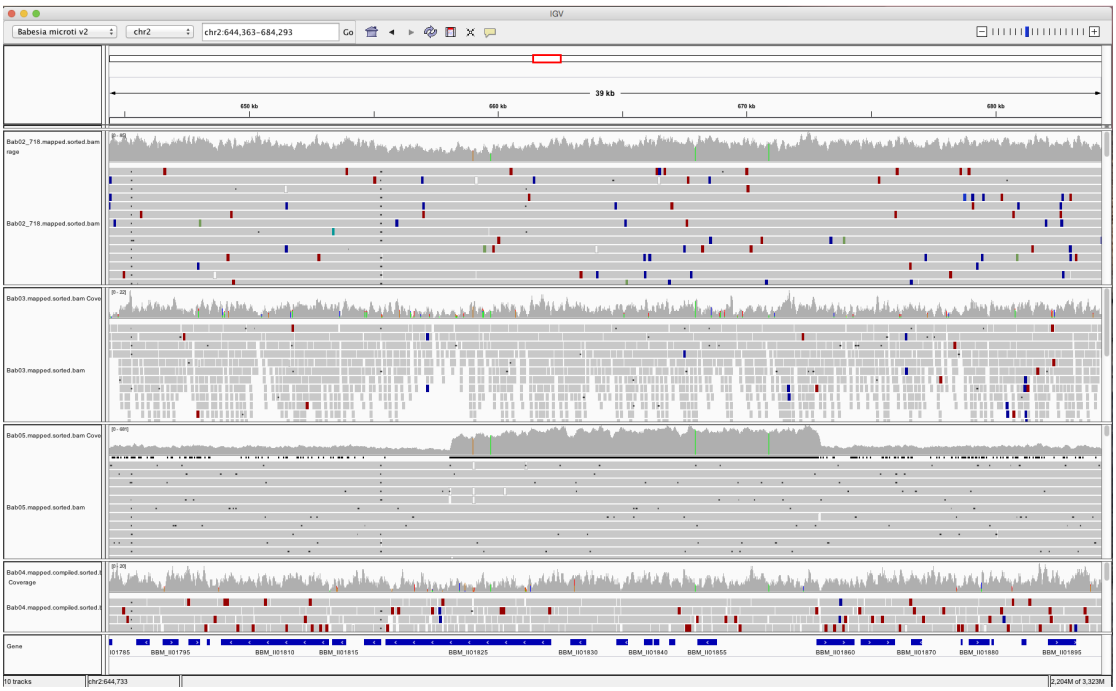


Figure 19: Read alignment and integrated coverage for patient samples 2 - 5 on chromosome 2, in the interval between 645 kilobases and 685 kilobases. There is a three-folf amplificaton of a 15KB fragment beginning at base 658,075 and extending to base 672,981. This interval contains an ABC transporter with homology to multidrug resistance related proteins.

Description	Score	Max Score	Query Cover	E	%Iden
unnamed protein product [Babesia microti strain RI]	2108	2108	100%	0.0	100%
ABC transporter, ATP-binding protein domain containing protein [Babesia bovis T2Bo]	256	467	88%	9e-67	29%
ABC transporter [Theileria annulata strain Ankara]	253	447	89%	6e-66	29%
ABC transporter family protein [Cryptosporidium muris RN66]	237	237	95%	6e-61	23%
ABC transporter, ATP-binding protein domain containing protein [Babesia equi]	231	495	76%	1e-58	30%
ABC transporter, ATP-binding protein domain containing protein, putative [Babesia bigen	221	427	88%	9e-56	29%
ABC transporter [Theileria orientalis strain Shintoku]	219	488	75%	5e-55	32%
hypothetical protein AMTR_s00065p00212850 [Amborella trichopoda]	206	206	89%	5e-51	24%
ABC transporter [Theileria parva strain Muguga]	202	489	81%	8e-50	30%
multidrug resistance-associated protein ABC domain protein [Medicago truncatula]	196	196	88%	6e-48	23%
multidrug resistance-associated protein ABC domain protein [Medicago truncatula]	196	196	88%	1e-47	23%
PREDICTED: multidrug resistance-associated protein 1-like [Ciona intestinalis]	193	193	89%	7e-47	22%
ABC transporter C family member [Medicago truncatula]	191	191	83%	3e-46	24%
hypothetical protein SDRG_00896 [Saprolegnia diclina VS20]	190	190	95%	5e-46	22%
ATP-binding cassette transporter sub-family C member 7 [Tigriopus japonicus]	190	190	88%	6e-46	23%
PREDICTED: ABC transporter C family member 13 isoform X2 [Populus euphratica]	187	187	88%	7e-45	22%
PREDICTED: ABC transporter C family member 13 isoform X1 [Populus euphratica]	187	187	88%	8e-45	22%
PREDICTED: ABC transporter C family member 13-like isoform X3 [Citrus sinensis]	185	185	83%	2e-44	22%

Figure 20: BLASTp search results for BBM_01850 showing homology to ABC transporters and multi drug resistance associated proteins.

tively under selection [72]. As has been discussed, pairwise nucleotide diversity is useful as a tool for detecting selection. Other summary statistics are useful as well: Among different populations, selection is expected to produce large allele frequency differences, and a high frequency of derived alleles. Finally, since recombination erodes genetic linkage among variants as a function of time, identifying linkage disequilibrium that extends less rapidly than expected is a marker of positive selection [73].

Applying these methods is a major goal for the future. At this time, we do not have enough samples to pursue these approaches in any generality. A map of linkage disequilibrium for the mitochondrial genome is shown in Figure 21. We have ten full-length mitochondrial sequences, so this is the only region for which we have sufficient sequence to even attempt it. However, as expected [41], there is no decay in linkage between markers as a function of distance, consistent with an absence of recombination in the mitochonrdial genome. The comparable plots for the apicoplast and core chromosomal regions are not shown because we only have full-length sequence for 4 genomes. As we accrue additional samples, we expect that methods based on linkage disequilibrium and long haplotypes will be powerful tools to detect selection at a population level in *B. microti*.

4.7 Sequencing Additional Strains: A Map of Genetic Diversity from 18 *Babesia microti* Isolates

Since submission of the first version of this thesis on February 15, 2015, we have completed sequencing of 13 additional isolates of *Babesia microti*. While analysis of these samples is ongoing, preliminary results from this expanded dataset are described here. By placing the already identified genetic diversity in its appropriate context, these new isolates yield great insight into US babesia genetic structure.

The initial sequencing of clinical strains raised several questions. First, how does the relatively small amount of genetic diversity identified in these strains compare to

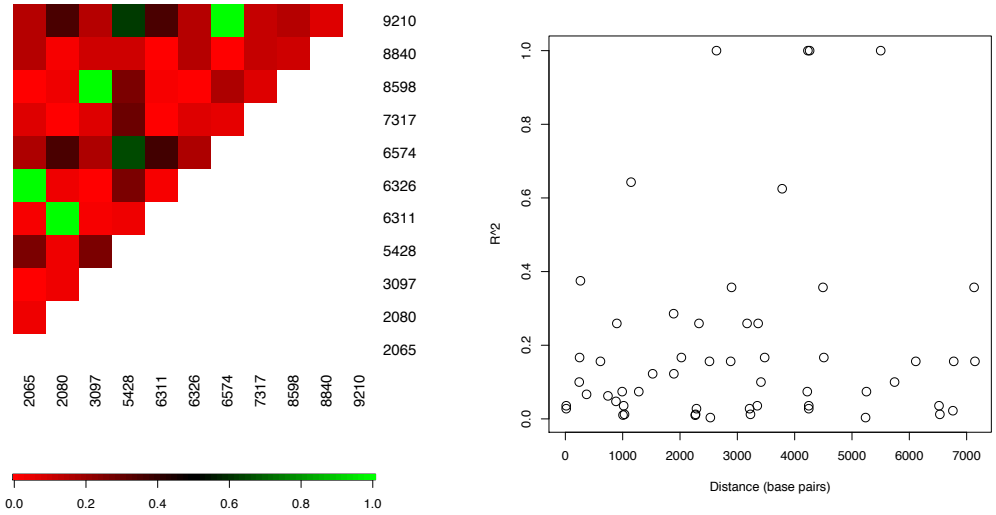


Figure 21: Matrix of pairwise values for R^2 for mitochondrial SNPs (left panel). Scatterplot of R^2 vs. distance between SNPs.

the diversity in New England and the United States? Second, how many babesia subpopulations are there in the US, and how long ago did they diverge? Third, do the variants unique to Patient 5 in the initial cohort — and therefore potentially related to azithromycin resistance — maintain their uniqueness when a larger set of samples is considered?

We sequenced 13 additional isolates in an effort to answer these questions. In collaboration with Dr. Sam Telford, we identified historical isolates that had been laboratory adapted that 1) represented *B. microti* diversity within New England the United States; 2) were isolated at different times over the past 50 years and/or grown continuously with passage between rodent and tick to estimate the rate of molecular evolution, and 3) represented members of established *B. microti* New England clades [30].

Sequencing of these additional isolates expanded the resolution of our existing map by adding an enormous quantity of genetic diversity. We identified a total of 2196 variants in the 18 strains. The vast majority of this diversity (1746 SNPs or 79.5%) was added by the single Midwest isolate, Minnesota strain (MN-1). The density of the genetic map has increased significantly, as shown in Figure 22. Nucleotide diversity among New England

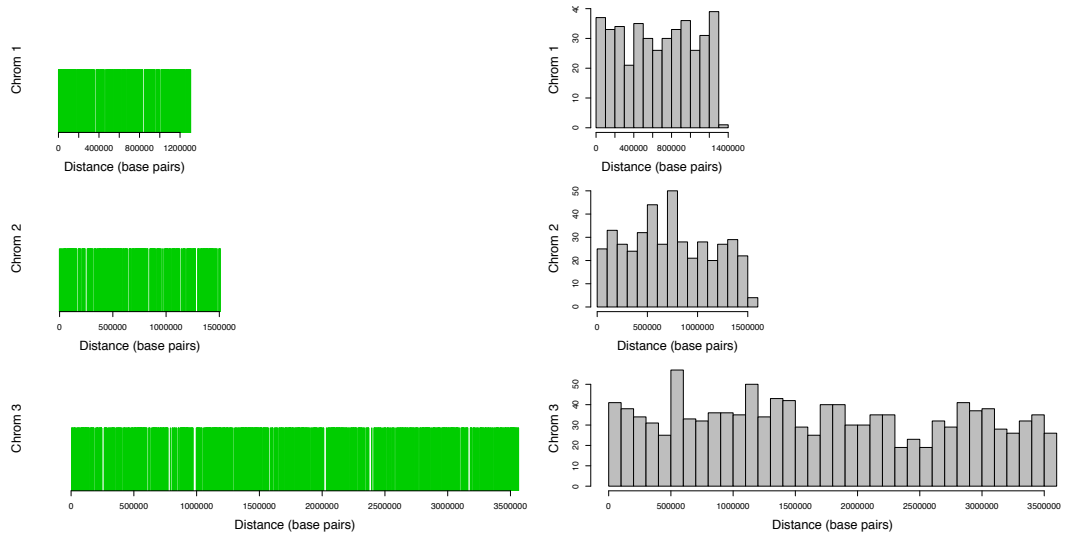


Figure 22: SNP locations (left) and SNP density (right) in version 2 of the genetic map encompassing all 18 *Babesia microti* isolates

isolates ($\pi = 1.2 \times 10^{-5}$) was over 10 times lower than between New England and Midwest isolates ($\pi = 3.3 \times 10^{-4}$), demonstrating the relatively ancient divergence of Midwest and Northeast populations along with the recent radiation of Northeast strains. The SNP location and SNP density is shown in Figure 22, demonstrating the dramatic increase in resolution from the initial version of the map, shown in Figure 14.

The identified genetic variation clusters into three lineages, shown in Figure 23. The most divergent of these contains the single Minnesota isolate. The other two represent the Nantucket and Connecticut/Rhode Island clades as identified by Goethert and Telford [30]. The geographic structure of isolates can also be seen in the PCA analysis in the top panel of Figure 24. Notably, all of the clinical isolates described in the initial sections and enrolled at MGH are members of the CT/RI lineage.

In addition to identifying geographic structure among the circulating isolates, we also tried to understand historical and demographic relationships. A widely used method is a Bayesian phylogenetic approach. In estimation of divergence times, tree topology is a nuisance parameter, and a Bayesian approach offers a way to deal with this by integrating over all possible tree topologies. The historical challenge has been numerical evaluation

of the high dimensional integrals that result; however, this has been addressed in recent years with markov chain monte carlo methods. This approach has been implemented in several software packages, the most recent and widely adopted of which is BEAST (Bayesian Evolutionary Analysis of Sampling Trees) [20, 21].

We use BEAST to estimate divergence times for the sample in our study. Dates of common ancestry are labeled in the Bayesian phylogenetic tree shown in the bottom of Figure 24. The rate of the molecular clock was estimated using samples collected at various times (Figure 25). A mean substitution rate was estimated at 4.8×10^{-8} mutations/site/year. The posterior distribution of substitution rate is summarized in Figure 25. Estimation of time to most recent common ancestry produced estimates of 70 years for CT/RI strains (95% HPD 37-113 years), 120 years for Nantucket isolates (95% HPD 70-185 years), 416 years (95% HPD 236-655 years) for New England strains and 2332 years (95% HPD 1313-3698 years) between New England clades and a Minnesota strain. The posterior distributions for these quantities are shown in Figure 26.

These data imply that the present-day geographic distribution of *B. microti* resulted from a divergence that occurred one to four thousand years ago followed by local expansion after seeding new geographic sites. Despite initial reports of human babesiosis from Nantucket, the vast majority of cases at clinical sites in Boston result from the Connecticut/ Rhode Island clade. This clade has emerged within the last 100 years and spread rapidly throughout New England.

Finally, the new data also shed light on the putative mechanisms of azithromycin resistance. Comparing against the complete set of 18 isolates, the arginine to histidine substitution in the RPL4 remains unique. In order to obtain adequate coverage from the babesia genome from Patient 1, which showed questionable evidence of copy number variation in MRP, we exhaustively sequenced this isolate using a whole lane of a HiSeq2500 instrument. This revealed a single copy of the MRP locus on chromosome 2, suggesting that possible evidence of copy number variation at MRP locus in this isolate from the initial round of sequencing was due to incorrectly decoded sequencing barcodes (due to

incorrect base calls in the barcoded portion of the read) in the context of overwhelming quantitites of DNA from Patient 5 relative to very limited quantitities of babesia DNA from the isolate from Patient 1. Thus, like the RPL4 mutation, this variant was unique to the azithromycin "resistant" case, supporting the possibility that the MRP locus may participate in azithromycin resistance. In future work, each of these candidates will need to be evaluated experimentally to test this hypothesis.

Overall, while the analysis of these new strains is still preliminary, the set of 18 isolates provides deep insight into the structure and evolution of *Babesia microti*. Genetic variation at a whole genome level in this larger set of isolates documents the presence of at least three major clades. These clades underwent separation into Midwestern and Northeastern populations between one- and four-thousand years ago; subdivision among Northeastern isolates was established between 250 and 700 years ago. While the Nantucket isolates appear to have remained relatively isolated, the CT/RI clade has spread quickly throughout New England, and this has occurred recently, with the ancestry of circulating strains dating to within the last 100 years.

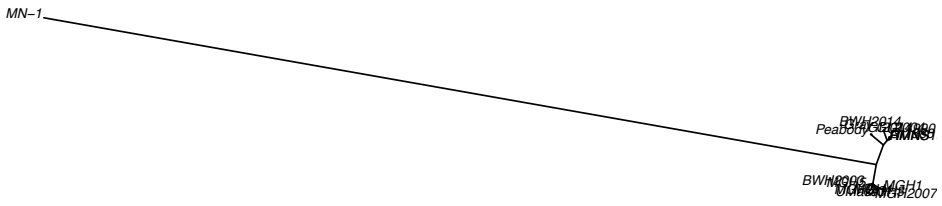


Figure 23: Unrooted tree of the 18 sequenced *Babesia microti* isolates, showing substantial divergence between the MN-1 isolate from Minnesota and the two East Coast clades, as well as the close relationship among samples in the East Coast clades, representing Nantucket strains (top cluster) and CT/RI (bottom cluster). All five clinical strains described above fall within the CT/RI clade.

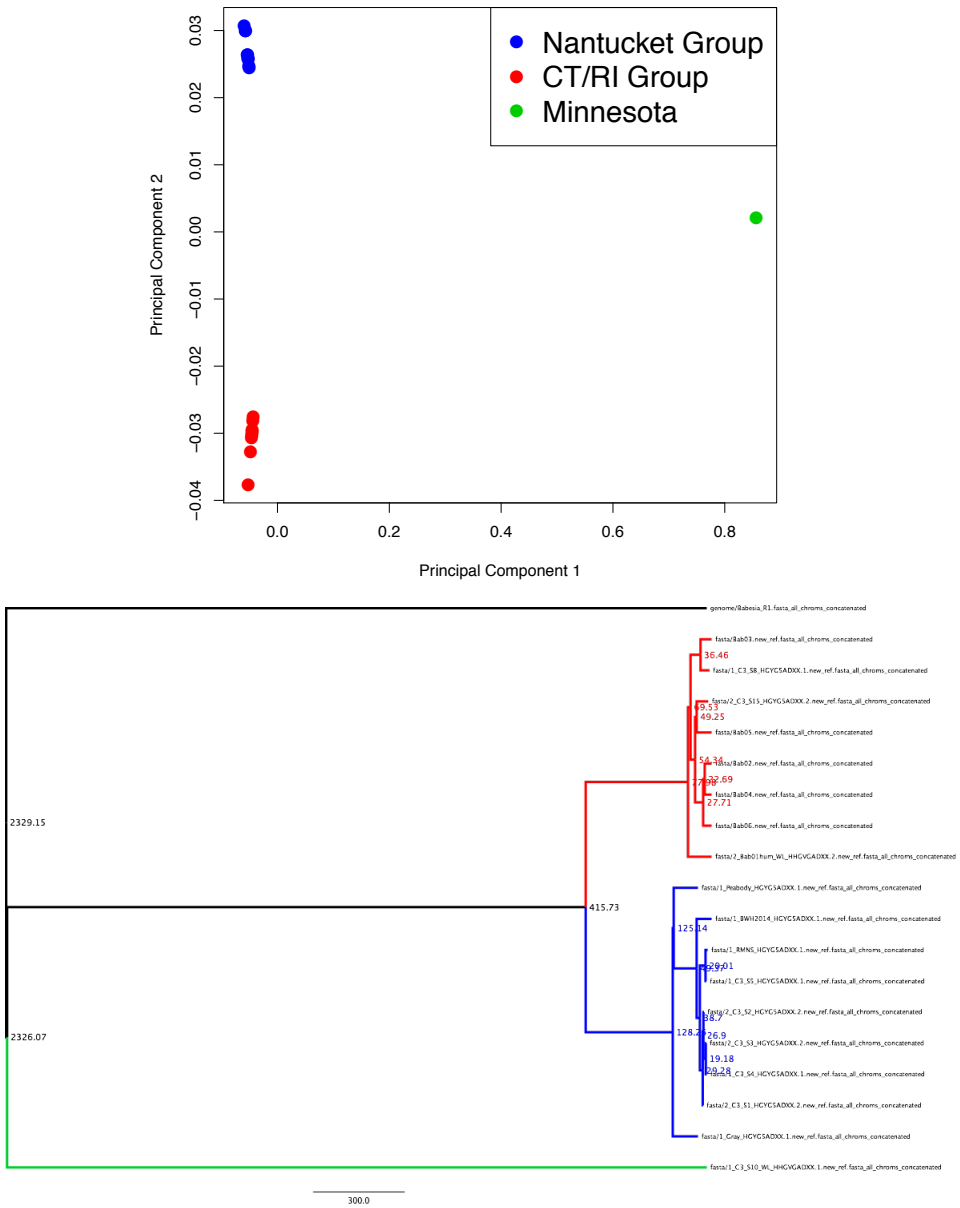


Figure 24: Top Panel: Principal Component Analysis of all 18 *Babesia microti* isolates showing the three major clades. Bottom Panel: Bayesian phylogenetic tree with nodes labeled with mean divergence time in years before present. CT/RI clade is in red, Nantucket clade is in blue, and Midwest clade is in green.

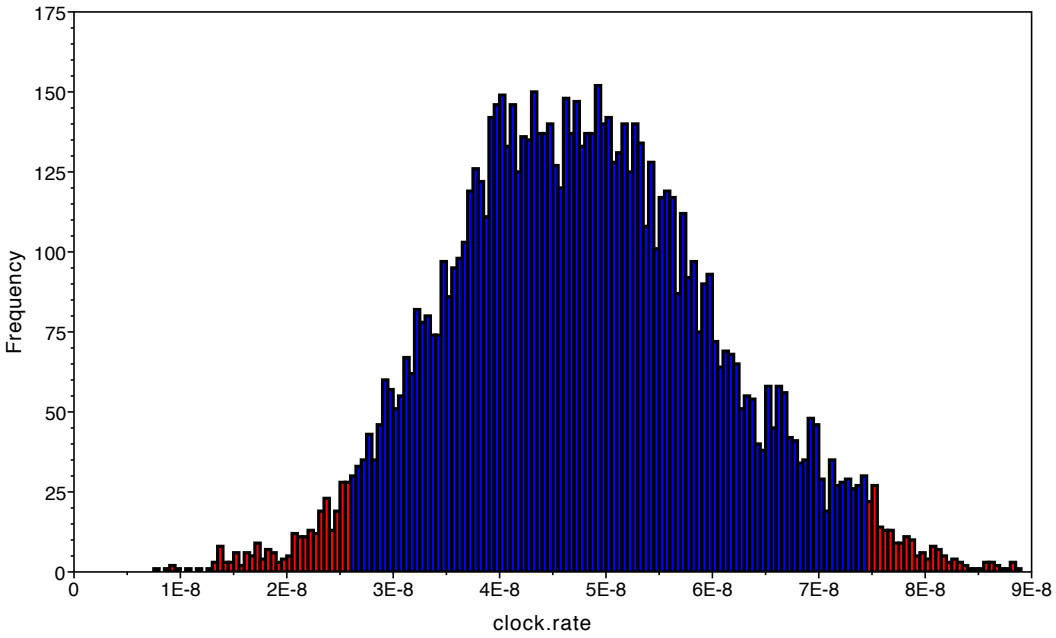


Figure 25: Posterior distribution for substitution rate in *Babesia microti*

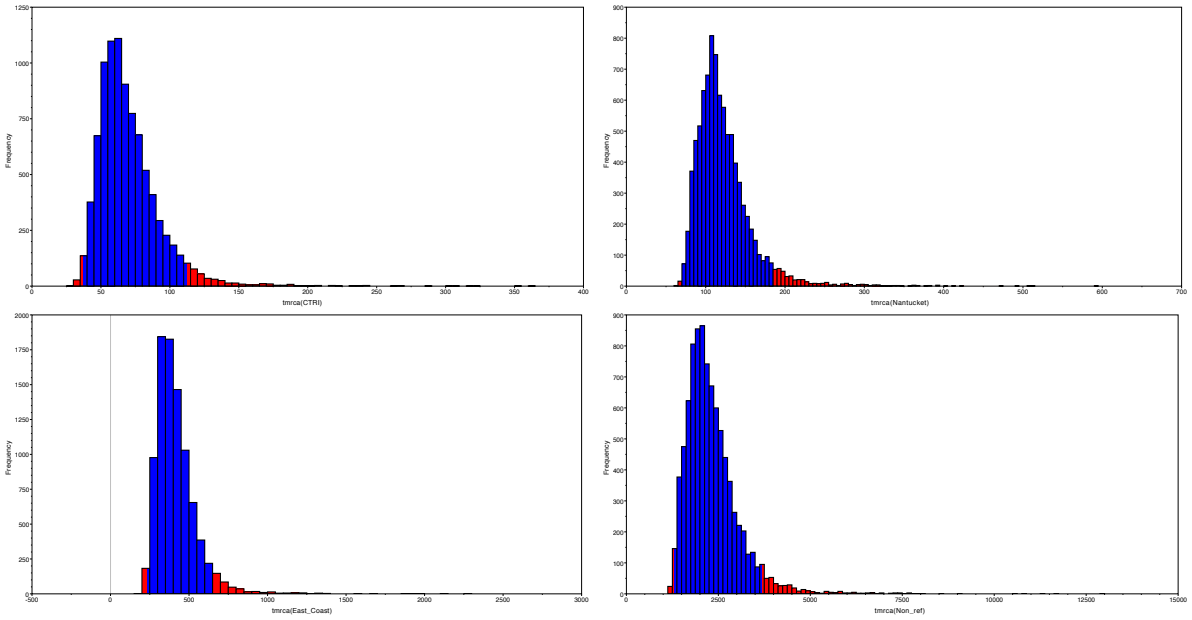


Figure 26: Posterior densities for time to most recent common ancestor (TMCRA) for Connecticut/RI strains (top), Nantucket Strains (second from top), East Coast Strains (second from bottom), and United States strains (bottom).

5 Conclusions

Malaria and babesia, the two parasitic protozoa of human red cells, account for an enormous burden of infectious disease worldwide. As we begin the twenty first century, the situation appears to be getting worse, not better. Babesiosis is emerging. Artemisinin resistance has now spread to multiple continents.

In this thesis, we have studied two related problems in red cell parasite biology, the evolutionary mechanisms driving acquisition of multiple drug resistance in malaria, and the emergence of *B. microti* and the development of azithromycin resistance. The common theme is evolution: the rate at which alleles are introduced into the population, and the dynamics of how they rise to appreciable frequency, ultimately resulting in fixed genetic substitutions.

We estimated the substation rate in *P. falciparum*, and also assessed factors that might increase or decrease this rate. We focused on two major variables: low-dose chloroquine and parasite genetic background. We chose these pressures because the experimental literature suggests that inherited defects in DNA damage repair, as well as known activity of antimalarial drugs in inhibiting DNA repair processes, may elevate the mutation rate in *P. falciparum* and foment the generation of drug resistant strains of *P. falciparum*. We hypothesized that these defects would result in increased rate of mutation under chloroquine pressure and in parasites sampled from Southeast Asia. To test these hypotheses, we cultured four isolates of *P. falciparum* for approximately 1.5 years each—two different strains in the presence and absence of low-dose chloroquine pressure. These experiments provided a direct test of the hypothesis that chloroquine and genetic background affect mutation rate.

We observed almost exactly equal numbers of substitutions in lines exposed to chloroquine as in lines grown without drug pressure, suggesting that chloroquine does not alter rate. Formal testing of this hypothesis through a Poisson regression model confirmed this result, with a non-significant p-value when a chloroquine coefficient was included in

the model. Thus we do not find evidence for the hypothesis that chloroquine increases rate in the context of the experiment conducted. We also considered whether genotype affects rate. We compared strains of two distinct genetic background, 7G8, a chloroquine resistant South American culture-adapted isolate, and TM284, a Southeast Asian isolate from along the Thai Cambodian border that is also chloroquine resistant. In this strain, we observed 10 mutations over 250 generations, as opposed to 7 mutations over 510 generations in the 7G8 strain, leading to a rate that appears to be elevated by a factor of 2.9. We formally tested the effect of genotype, when considered as a covariate in Poisson regression models, had a p-value of 0.056, with a trend toward significance but not reaching the typical threshold of $p = 0.05$. This appears likely to be an effect of study power. Despite a total of approximately 4.2 years total years in culture for the two strains, we detected a total of only 17 mutations. Nevertheless, this 3-fold difference in rate, even if it is confirmed, does not appear to be sufficient to account for the 10^{2-3}-fold increase in the rate at which ARMD parasites develop resistance to novel antimalarials. Our data do not support a model in which the ARMD phenotype is generated by a "hypermutator" parasite. It is possible that the parasite we studied did not possess the ARMD phenotype. TM284 is a recent clinical isolate from the Thai-Cambodian border, but it not directly related to W2 used in the original studies [69]. This is unlikely for several reasons: first, Dd2 (a subclone of W2) does not show elevated numbers of mutations in clone trees [8]. Second, studies of relatedness among global parasites show that the vast majority of genetic variability is between parasites on different continents, and that parasites from within a geographic area are closely related. Third, TM284 is a multi-drug resistant parasite, the exact sort which should be carrier the ARMD phenotype. Nevertheless, this remains a possible explanation for the data. A third possibility is that the genomic architecture in ARMD parasites exists to facilitate resistence in other ways aside from through the mutation rate: for example, the rate at which copy number variations develop or expand may be dramatically increased in these parasites. This seems the most likely possibility at this point. Alternatively, mutational pressure in the

form of ultraviolet light or a mutagen distinct from chloroquine is needed to bring out the phenotype. Finally, the ARMD phenotype may be an artifact of the experimental situation under which it was described.

Like all experimental studies, this one had drawbacks. The study has two major weaknesses. The first is low study power, and the second is a small number of strains and conditions considered. The power of the study was limited by several factors. The first is the rate at which evolution occurs. In order to identify as many mutations as possible, the cultures were grown for a long time—510 days in the case of the 7G8 strain, and 480 days in the case of the TM strain. Between the four culture conditions, there were a collective 4.2 years in culture, yet the experiment would have benefited from additional evolutionary time. Had more time separated the cultures, the effect of genotype on mutation rate may have achieved statistical significance. Additionally, the mutations which accumulated in this study were not sufficient to begin testing questions such as whether the rate varies throughout the genome. Such questions would be accessible with a larger, more powerful study.

Another reason for the low power of the experiment is that the substitution rate is likely lower than the mutation rate. For neutral mutations, the two quantities are equal, because the population size N, drops out of the relation in equation 4. For for mutations under selection, the probability of fixation is $\approx \frac{1-e^{2s}}{1-e^{2sN}}$ [29], which leads to complicated expressions for ρ. For example, in the case of weak positive selection, the following relation between the substitution rate and the mutation rate:

$$\rho = \frac{N\mu}{2s},$$

a difficult expression to work with because it depends on s, the selection coefficient, a random variable. In addition to confounding the estimation of mutation rate, the presence of selection also decreases study power because the distribution of selection coefficients will be skewed toward negative values (most mutations are selectively disadvantageous

or neutral).

Future work must focus on increasing the power of the experimental setup, and directly estimating μ rather than ρ. The most effective method is to decrease the study population size, which should be made as small as possible, since we would like to set the relative strength of selection, which is proportional to population size, large compared to drift, which is proportional to the inverse of population size. Based on this consideration, we are in the process of conducting a follow-up study in which the experiment is repeated, but the progeny from a cloning by limiting dilution are used. In this type of experiment, mutations that enter the population during the single-cell phase of the population will immediately fix. Thus, if the experiment is terminated before alleles can begin to be selected, all alleles will have frequency of 0 or 1. The results of this study suggest that window to be approximately 50 generations (Figure 9), since alleles with selection coefficients of $< 5\%$ do not appreciably rise in frequency over this window. Since only a single generation of evolutionary time will pass in such experiments, it will be necessary to generate and sequence a large number of clones. Nevertheless, the cost of this experiment would not be as high as it might seem, since the clones need only be sequenced to low coverage because the allele frequency will be either 0 or 1. Coverage in the 5 - 10 fold range is adequate, and at those level, approximately 50-100 strains can be multiplexed in a single lane of illumina sequencing, bringing costs down.

We also attempted to characterize the emergence of *B. microti* at a molecular level. We generated a census of genetic variants in this parasite, and use this to provide molecular evidence of a growing population and recent common ancestry of clinical isolates and enzootic strains. We demonstrate substructure to the parasite populations circulating in New England, and show that this is associated with geographic distance between sites. Finally, we provide preliminary genetic and bioinformatic evidence that substitutions in codon 86 of RPL4, encoding a highly conserved arginine adjacent to a known azithromycin binding pocket, confer resistance to azithromycin in *B. microti*. We also identify two other genomic changes—hypervariability in a putative H+/Na+ exchanger

and copy number variation in ABC transporter—that may be involved important *B. microti* phenotypes.

The possibility that *B. microti* evolves resistance to azithromycin in the same manner is *P. falciparum* and *S. pneumoniaie* is an intriguing and potentially important result. If true, this would permit the development of molecular assays to detect drug resistance before it leads catastrophic clinical scenarios. A priority is confirming substitutions in the azithromycin-binding pocket of RPL4 mediate resistance. We are in the process of trying to attempting to evolve azithromycin resistance in *B. microti* in the lab and testing the functional properties of the allele identified in this study. This is not straightforward as there is no *in vitro* culture system for *B. microti*, but the experiments can be performed in SCID mice.

The population genetic results in this study are preliminary as a result of the limited sample size. Future work will increase the power of the study in two ways. First, we are planning to enroll additional cases at multiple sites in the upcoming years. Second, we are planning to develop a hydrid capture library to enrich for *B. microti* DNA from enzootic samples, since the proportion of *B. microti* DNA in infected *I. scapularis* ticks and *P. leucopus* rodents is quite small (Table 6). A larger sample size will increase the resolution of the map of genetic variation and facilitate unbiased screens of natural selection, resolve the demographic history of emergence, and test whether subsets of enzootic isolates are more likely to infect humans.

Overall, this report makes several key contributions to the literature. It is the first to estimate the substitution rate in *P. falciparum*. We provide evidence that chloroquine, previously hypothesized to elevate the mutation rate, does not significantly elevate substitution rate in live *P. falciparum*. We document a possible small increase in substitution rate in parasites of Southeast Asian origin, but argue that this is sufficiently small to refute the hypothesis that the ARMD phenotype is mediated at the level of mutational rate. This work also provides the first real-time view of the process of selection in malaria samples. The trajectories of mutations in our sample validate Haldane's theory of natu-

ral selection in infinite populations, but also demonstrate evidence of clonal interference. We derive the distribution of equilibrium allele frequencies for passenger mutations, and propose a novel approach to estimating mutation rate through the accumulation of such mutations. We also use evolutionary genomic approaches to study clinical infections, focusing on *B. microti*. By leveraging similar techniques of whole-genome sequencing and evolutionary genetic analysis, we provide molecular evidence supporting a recent population expansion of *B. microti*, documenting recent common ancestry and geographic substructure among circulating strains, and identify candidate genes for azithromycin resistance. Collectively, our data provide insight into the evolution and pathogenesis of protozoan infections of human red cells. A clear understanding of the mechanisms by which parasite spread and evolve resistance to drugs will inform control measures and help to preserve the efficacy of antiparasitic drugs.

References

[1] Marcela E Perez Acosta, Peter T. Ender, Erin M. Smith, and Jeffrey A. Jahre. Babesia microti infection, eastern pennsylvania, usa. *Emerg Infect Dis*, 19(7):1105–1107, Jul 2013.

[2] K. C. Atwood, L. K. Schneider, and F. J. Ryan. Periodic selection in escherichia coli. *Proc Natl Acad Sci U S A*, 37(3):146–155, Mar 1951.

[3] Lara Bethke, Susan Thomas, Kerone Walker, Ronak Lakhia, Radha Rangarajan, and Dyann Wirth. The role of dna mismatch repair in generating genetic diversity and drug resistance in malaria parasites. *Mol Biochem Parasitol*, 155(1):18–25, Sep 2007.

[4] Tyler S. Brown, Christopher G. Jacob, Joana C. Silva, Shannon Takala-Harrison, Abdoulaye Djimd, Arjen M. Dondorp, Mark Fukuda, Harald Noedl, Myaing Myaing Nyunt, Myat Phone Kyaw, Mayfong Mayxay, Tran Tinh Hien, Christopher V. Plowe, and Michael P. Cummings. Plasmodium falciparum field isolates from areas of repeated emergence of drug resistant malaria show no evidence of hypermutator phenotype. *Infect Genet Evol*, 30C:318–322, Dec 2014.

[5] Meryl A. Castellini, Jeffrey S. Buguliskis, Louis J. Casta, Charles E. Butz, Alan B. Clark, Thomas A. Kunkel, and Theodore F. Taraschi. Malaria drug resistance is associated with defective dna mismatch repair. *Mol Biochem Parasitol*, 177(2):143–147, Jun 2011.

[6] CDC. Babesiosis surveillence - 18 states, 2011. *MMWR*, Jul 13;61(27):505–9, 2012.

[7] Richard E. Cibulskis, Maru Aregawi, Ryan Williams, Mac Otten, and Christopher Dye. Worldwide incidence of malaria in 2009: estimates, time trends, and a critique of methods. *PLoS Med*, 8(12):e1001142, Dec 2011.

[8] Antoine Claessens, William L. Hamilton, Mihir Kekre, Thomas D. Otto, Adnan Faizullabhoy, Julian C. Rayner, and Dominic Kwiatkowski. Generation of antigenic diversity in plasmodium falciparum by structured rearrangement of var genes during mitosis. *PLoS Genet*, 10(12):e1004812, Dec 2014.

[9] Ian A. Clark, Alison C. Budd, Gunther Hsue, Bret R. Haymore, Alina J. Joyce, Richard Thorner, and Peter J. Krause. Absence of erythrocyte sequestration in a case of babesiosis in a splenectomized human patient. *Malar J*, 5:69, 2006.

[10] S. N. Cohen and K. L. Yielding. Inhibition of dna and rna polymerase reactions by chloroquine. *Proc Natl Acad Sci U S A*, 54(2):521–527, Aug 1965.

[11] Emmanuel Cornillot, Kamel Hadj-Kaddour, Amina Dassouli, Benjamin Noel, Vincent Ranwez, Benot Vacherie, Yoann Augagneur, Virginie Brs, Aurelie Duclos, Sylvie Randazzo, Bernard Carcy, Franoise Debierre-Grockiego, Stphane Delbecq, Karina Moubri-Mnage, Hosam Shams-Eldin, Sahar Usmani-Brown, Frdric Bringaud, Patrick Wincker, Christian P. Vivars, Ralph T. Schwarz, Theo P. Schetters, Peter J.

Krause, Andr Gorenflot, Vincent Berry, Valrie Barbe, and Choukri Ben Mamoun. Sequencing of the smallest apicomplexan genome from the human pathogen babesia microti. *Nucleic Acids Res*, 40(18):9102–9114, Oct 2012.

[12] F. E. Cox. Protective immunity between malaria parasites and piroplasms in mice. *Bull World Health Organ*, 43(2):325–336, 1970.

[13] F. E. Cox. Heterologous immunity between piroplasms and malaria parasites: the simultaneous elimination of plasmodium vinckei and babesia microti from the blood of doubly infected mice. *Parasitology*, 76(1):55–60, Feb 1978.

[14] Janet Cox-Singh, Timothy M E. Davis, Kim-Sung Lee, Sunita S G. Shamsul, Asmad Matusop, Shanmuga Ratnam, Hasan A. Rahman, David J. Conway, and Balbir Singh. Plasmodium knowlesi malaria in humans is widely distributed and potentially life threatening. *Clin Infect Dis*, 46(2):165–171, Jan 2008.

[15] J.F. Crow and M. Kimura. *An introduction to population genetics theory*. New York, Evanston and London: Harper & Row, Publishers, 1970.

[16] K. W. Deitsch, E. R. Moxon, and T. E. Wellems. Shared themes of antigenic variation and virulence in bacterial, protozoal, and fungal infections. *Microbiol Mol Biol Rev*, 61(3):281–293, Sep 1997.

[17] Kirk W. Deitsch, Sheila A. Lukehart, and James R. Stringer. Common strategies for antigenic variation by bacterial, fungal and protozoan pathogens. *Nat Rev Microbiol*, 7(7):493–503, Jul 2009.

[18] J. Dobroszycki, B. L. Herwaldt, F. Boctor, J. R. Miller, J. Linden, M. L. Eberhard, J. J. Yoon, N. M. Ali, H. B. Tanowitz, F. Graham, L. M. Weiss, and M. Wittner. A cluster of transfusion-associated babesiosis cases traced to a single asymptomatic donor. *JAMA*, 281(10):927–930, Mar 1999.

[19] Arjen M. Dondorp, Franois Nosten, Poravuth Yi, Debashish Das, Aung Phae Phyo, Joel Tarning, Khin Maung Lwin, Frederic Ariey, Warunee Hanpithakpong, Sue J. Lee, Pascal Ringwald, Kamolrat Silamut, Mallika Imwong, Kesinee Chotivanich, Pharath Lim, Trent Herdman, Sen Sam An, Shunmay Yeung, Pratap Singhasivanon, Nicholas P J. Day, Niklas Lindegardh, Duong Socheat, and Nicholas J. White. Artemisinin resistance in plasmodium falciparum malaria. *N Engl J Med*, 361(5):455–467, Jul 2009.

[20] Alexei J. Drummond, Simon Y W. Ho, Matthew J. Phillips, and Andrew Rambaut. Relaxed phylogenetics and dating with confidence. *PLoS Biol*, 4(5):e88, May 2006.

[21] Alexei J. Drummond and Andrew Rambaut. Beast: Bayesian evolutionary analysis by sampling trees. *BMC Evol Biol*, 7:214, 2007.

[22] Jessica M. Dunn, Peter J. Krause, Stephen Davis, Edouard G. Vannier, Meagan C. Fitzpatrick, Lindsay Rollend, Alexia A. Belperron, Sarah L. States, Andrew Stacey,

Linda K. Bockenstedt, Durland Fish, and Maria A. Diuk-Wasser. Borrelia burgdorferi promotes the establishment of babesia microti in the northeastern united states. *PLoS One*, 9(12):e115494, 2014.

[23] E. S. Eskow, P. J. Krause, A. Spielman, K. Freeman, and J. Aslanzadeh. Southern extension of the range of human babesiosis in the eastern united states. *J Clin Microbiol*, 37(6):2051–2052, Jun 1999.

[24] D. A. Fidock, T. Nomura, A. K. Talley, R. A. Cooper, S. M. Dzekunov, M. T. Ferdig, L. M. Ursos, A. B. Sidhu, B. Naud, K. W. Deitsch, X. Z. Su, J. C. Wootton, P. D. Roepe, and T. E. Wellems. Mutations in the p. falciparum digestive vacuole transmembrane protein pfcrt and evidence for their role in chloroquine resistance. *Mol Cell*, 6(4):861–871, Oct 2000.

[25] R.A. Fisher. *The genetical theory of natural selection.* Dover Publications, Inc., New York & Constable & Co., Ltd., London., 1958.

[26] Food and Drug Administration. Fatalities reported to fda following blood collection and transfusion: Annual summary for fiscal year 2012. http://www.fda.gov/BiologicsBloodVaccines/SafetyAvailability/ReportaProblem/TransfusionDon 2014.

[27] S. J. Foote and A. F. Cowman. The mode of action and the mechanism of resistance to antimalarial drugs. *Acta Trop*, 56(2-3):157–171, Mar 1994.

[28] S. J. Foote, D. E. Kyle, R. K. Martin, A. M. Oduola, K. Forsyth, D. J. Kemp, and A. F. Cowman. Several alleles of the multidrug-resistance gene are closely linked to chloroquine resistance in plasmodium falciparum. *Nature*, 345(6272):255–258, May 1990.

[29] J.H. Gillespie. *Population genetics: a concise guide.* Johns Hopkins University Press, 2004.

[30] Heidi K. Goethert and Sam R. Telford. Not "out of nantucket": Babesia microti in southern new england comprises at least two major populations. *Parasit Vectors*, 7(1):546, Dec 2014.

[31] A. Gorenflot, K. Moubri, E. Precigout, B. Carcy, and T. P. Schetters. Human babesiosis. *Ann Trop Med Parasitol*, 92(4):489–501, Jun 1998.

[32] Jennifer L. Guler, Daniel L. Freeman, Vida Ahyong, Rapatbhorn Patrapuvich, John White, Ramesh Gujjar, Margaret A. Phillips, Joseph DeRisi, and Pradipsinh K. Rathod. Asexual populations of the human malaria parasite, plasmodium falciparum, use a two-step genomic strategy to acquire accurate, beneficial dna amplifications. *PLoS Pathog*, 9(5):e1003375, 2013.

[33] J.B.S. Haldane. A mathematical theory of natural and artificial selection, part v: selection and mutation. *Mathematical Proceedings of the Cambridge Philosophical Society*, 23(7):838–844, 1927.

[34] B. M. Haltiwanger, Y. Matsumoto, E. Nicolas, G. L. Dianov, V. A. Bohr, and T. F. Taraschi. Dna base excision repair in human malaria parasites is predominantly by a long-patch pathway. *Biochemistry*, 39(4):763–772, Feb 2000.

[35] J. C. Hatcher, P. D. Greenberg, J. Antique, and V. E. Jimenez-Lucho. Severe babesiosis in long island: review of 34 cases and their complications. *Clin Infect Dis*, 32(8):1117–1125, Apr 2001.

[36] G. R. Healy, A. Speilman, and N. Gleason. Human babesiosis: reservoir in infection on nantucket island. *Science*, 192(4238):479–480, Apr 1976.

[37] Barbara L. Herwaldt, Guy de Bruyn, Norman J. Pieniazek, Mary Homer, Kathryn H. Lofy, Susan B. Slemenda, Thomas R. Fritsche, David H. Persing, and Ajit P. Limaye. Babesia divergens-like infection, washington state. *Emerg Infect Dis*, 10(4):622–629, Apr 2004.

[38] Barbara L. Herwaldt, Paul C. McGovern, Michal P. Gerwel, Rachael M. Easton, and Rob Roy MacGregor. Endemic babesiosis in another eastern state: New jersey. *Emerg Infect Dis*, 9(2):184–188, Feb 2003.

[39] M. J. Homer, E. S. Bruinsma, M. J. Lodes, M. H. Moro, S Telford, 3rd, P. J. Krause, L. D. Reynolds, R. Mohamath, D. R. Benson, R. L. Houghton, S. G. Reed, and D. H. Persing. A polymorphic multigene family encoding an immunodominant protein from babesia microti. *J Clin Microbiol*, 38(1):362–368, Jan 2000.

[40] G. S. Humphreys, I. Merinopoulos, J. Ahmed, C J M. Whitty, T. K. Mutabingwa, C. J. Sutherland, and R. L. Hallett. Amodiaquine and artemether-lumefantrine select distinct alleles of the plasmodium falciparum mdr1 gene in tanzanian children treated for uncomplicated malaria. *Antimicrob Agents Chemother*, 51(3):991–997, Mar 2007.

[41] Deirdre A. Joy, Xiaorong Feng, Jianbing Mu, Tetsuya Furuya, Kesinee Chotivanich, Antoniana U. Krettli, May Ho, Alex Wang, Nicholas J. White, Edward Suh, Peter Beerli, and Xin-zhuan Su. Early origin and recent expansion of plasmodium falciparum. *Science*, 300(5617):318–321, Apr 2003.

[42] Jung-Yeon Kim, Shin-Hyeong Cho, Hyun-Na Joo, Masayoshi Tsuji, Sung-Ran Cho, Il-Joong Park, Gyung-Tae Chung, Jung-Won Ju, Hyeng-Il Cheun, Hyeong-Woo Lee, Young-Hee Lee, and Tong-Soo Kim. First case of human babesiosis in korea: detection and characterization of a novel type of babesia sp. (ko1) similar to ovine babesia. *J Clin Microbiol*, 45(6):2084–2087, Jun 2007.

[43] M. Kimura and T. Ohta. The average number of generations until fixation of a mutant gene in a finite population. *Genetics*, 61 (3):763–771, 1969.

[44] P. J. Krause, A. Spielman, SR Telford, 3rd, V. K. Sikand, K. McKay, D. Christianson, R. J. Pollack, P. Brassard, J. Magera, R. Ryan, and D. H. Persing. Persistent parasitemia after acute babesiosis. *N Engl J Med*, 339(3):160–165, Jul 1998.

[45] P. J. Krause, S Telford, 3rd, A. Spielman, R. Ryan, J. Magera, T. V. Rajan, D. Christianson, T. V. Alberghini, L. Bow, and D. Persing. Comparison of pcr with blood smear and inoculation of small animals for diagnosis of babesia microti parasitemia. *J Clin Microbiol*, 34(11):2791–2794, Nov 1996.

[46] P. J. Krause, SR Telford, 3rd, R. Ryan, A. B. Hurta, I. Kwasnik, S. Luger, J. Niederman, M. Gerber, and A. Spielman. Geographical and temporal distribution of babesial infection in connecticut. *J Clin Microbiol*, 29(1):1–4, Jan 1991.

[47] Peter J. Krause, Johanna Daily, Sam R. Telford, Edouard Vannier, Paul Lantos, and Andrew Spielman. Shared features in the pathobiology of babesiosis and malaria. *Trends Parasitol*, 23(12):605–610, Dec 2007.

[48] Peter J. Krause, Benjamin E. Gewurz, David Hill, Francisco M. Marty, Edouard Vannier, Ivo M. Foppa, Richard R. Furman, Ellen Neuhaus, Gail Skowron, Shaili Gupta, Carlo McCalla, Edward L. Pesanti, Mary Young, Donald Heiman, Gunther Hsue, Jeffrey A. Gelfand, Gary P. Wormser, John Dickason, Frank J. Bia, Barry Hartman, Sam R Telford, 3rd, Diane Christianson, Kenneth Dardick, Morton Coleman, Jennifer E. Girotto, and Andrew Spielman. Persistent and relapsing babesiosis in immunocompromised patients. *Clin Infect Dis*, 46(3):370–376, Feb 2008.

[49] Peter J. Krause, Kathleen McKay, Joseph Gadbaw, Diane Christianson, Linda Closter, Timothy Lepore, Sam R Telford, 3rd, Vijay Sikand, Raymond Ryan, David Persing, Justin D. Radolf, Andrew Spielman, and Tick-Borne Infection Study Group . Increasing health burden of human babesiosis in endemic sites. *Am J Trop Med Hyg*, 68(4):431–436, Apr 2003.

[50] S. Kyes, P. Horrocks, and C. Newbold. Antigenic variation at the infected red cell surface in malaria. *Annu Rev Microbiol*, 55:673–707, 2001.

[51] Hermann Lage. Abc-transporters: implications on drug resistance from microorganisms to human cancers. *Int J Antimicrob Agents*, 22(3):188–199, Sep 2003.

[52] Heng Li and Richard Durbin. Fast and accurate short read alignment with burrows-wheeler transform. *Bioinformatics*, 25(14):1754–1760, Jul 2009.

[53] Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, Richard Durbin, and 1000 Genome Project Data Processing Subgroup . The sequence alignment/map format and samtools. *Bioinformatics*, 25(16):2078–2079, Aug 2009.

[54] B. MAEGRAITH, H. M. GILLES, and K. DEVAKUL. Pathological processes in babesia canis infections. *Z Tropenmed Parasitol*, 8(4):485–514, Dec 1957.

[55] S. E. Marley, M. L. Eberhard, F. J. Steurer, W. L. Ellis, P. B. McGreevy, and TK Ruebush, 2nd. Evaluation of selected antiprotozoal drugs in the babesia microti-hamster model. *Antimicrob Agents Chemother*, 41(1):91–94, Jan 1997.

[56] P. McCullagh and J.A. Nelder. *Generalized linear models*. Chapman & Hall/CRC, 1989.

[57] Aaron McKenna, Matthew Hanna, Eric Banks, Andrey Sivachenko, Kristian Cibulskis, Andrew Kernytsky, Kiran Garimella, David Altshuler, Stacey Gabriel, Mark Daly, and Mark A. DePristo. The genome analysis toolkit: a mapreduce framework for analyzing next-generation dna sequencing data. *Genome Res*, 20(9):1297–1303, Sep 2010.

[58] RO Michael and GM Williams. Choloroquine inhibition of repair of dna damage induced in mammalian cells by methyl methanesulfonate. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*, 25 (3):391–396, 1974.

[59] Jianbing Mu, Philip Awadalla, Junhui Duan, Kate M. McGee, Deirdre A. Joy, Gilean A T. McVean, and Xin-zhuan Su. Recombination hotspots and population structure in plasmodium falciparum. *PLoS Biol*, 3(10):e335, Oct 2005.

[60] Jianbing Mu, Junhui Duan, Kateryna D. Makova, Deirdre A. Joy, Chuong Q. Huynh, Oralee H. Branch, Wen-Hsiung Li, and Xin-Zhuan Su. Chromosome-wide snps reveal an ancient origin for plasmodium falciparum. *Nature*, 418(6895):323–326, Jul 2002.

[61] Shalini Nair, Jeff T. Williams, Alan Brockman, Lucy Paiphun, Mayfong Mayxay, Paul N. Newton, Jean-Paul Guthmann, Frank M. Smithuis, Tran Tinh Hien, Nicholas J. White, Franois Nosten, and Tim J C. Anderson. A selective sweep driven by pyrimethamine treatment in southeast asian malaria parasites. *Mol Biol Evol*, 20(9):1526–1536, Sep 2003.

[62] Harald Noedl, Youry Se, Kurt Schaecher, Bryan L. Smith, Duong Socheat, Mark M. Fukuda, and Artemisinin Resistance in Cambodia 1 (A. R. C1) Study Consortium . Evidence of artemisinin-resistant malaria in western cambodia. *N Engl J Med*, 359(24):2619–2620, Dec 2008.

[63] Harald Noedl, Duong Socheat, and Wichai Satimai. Artemisinin-resistant malaria in asia. *N Engl J Med*, 361(5):540–541, Jul 2009.

[64] J. M. Ortiz and RC Eagle, Jr. Ocular findings in human babesiosis (nantucket fever). *Am J Ophthalmol*, 93(3):307–311, Mar 1982.

[65] S. A. Peel. The abc transporter genes of plasmodium falciparum and drug resistance. *Drug Resist Updat*, 4(1):66–74, Feb 2001.

[66] D. H. Persing, B. L. Herwaldt, C. Glaser, R. S. Lane, J. W. Thomford, D. Mathiesen, P. J. Krause, D. F. Phillip, and P. A. Conrad. Infection with a babesia-like organism in northern california. *N Engl J Med*, 332(5):298–303, Feb 1995.

[67] L. G. Pologe and J. V. Ravetch. A chromosomal rearrangement in a p. falciparum histidine-rich protein gene is associated with the knobless phenotype. *Nature*, 322(6078):474–477, 1986.

[68] L. G. Pologe and J. V. Ravetch. Large deletions result from breakage and healing of p. falciparum chromosomes. *Cell*, 55(5):869–874, Dec 1988.

[69] P. K. Rathod, T. McErlean, and P. C. Lee. Variations in frequencies of drug resistance in plasmodium falciparum. *Proc Natl Acad Sci U S A*, 94(17):9389–9393, Aug 1997.

[70] TK Reubush, 2nd, P. B. Cassaday, H. J. Marsh, S. A. Lisker, D. B. Voorhees, E. B. Mahoney, and G. R. Healy. Human babesiosis on nantucket island. clinical features. *Ann Intern Med*, 86(1):6–9, Jan 1977.

[71] James T. Robinson, Helga Thorvaldsdttir, Wendy Winckler, Mitchell Guttman, Eric S. Lander, Gad Getz, and Jill P. Mesirov. Integrative genomics viewer. *Nat Biotechnol*, 29(1):24–26, Jan 2011.

[72] P. C. Sabeti, S. F. Schaffner, B. Fry, J. Lohmueller, P. Varilly, O. Shamovsky, A. Palma, T. S. Mikkelsen, D. Altshuler, and E. S. Lander. Positive natural selection in the human lineage. *Science*, 312(5780):1614–1620, Jun 2006.

[73] Pardis C. Sabeti, David E. Reich, John M. Higgins, Haninah Z P. Levine, Daniel J. Richter, Stephen F. Schaffner, Stacey B. Gabriel, Jill V. Platko, Nick J. Patterson, Gavin J. McDonald, Hans C. Ackerman, Sarah J. Campbell, David Altshuler, Richard Cooper, Dominic Kwiatkowski, Ryk Ward, and Eric S. Lander. Detecting recent positive selection in the human genome from haplotype structure. *Nature*, 419(6909):832–837, Oct 2002.

[74] N. Saitou and M. Nei. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol*, 4(4):406–425, Jul 1987.

[75] C. M. Shih and C. C. Wang. Ability of azithromycin in combination with quinine for the elimination of babesial infection in humans. *Am J Trop Med Hyg*, 59(4):509–512, Oct 1998.

[76] Amar Bir Singh Sidhu, Qingan Sun, Louis J. Nkrumah, Michael W. Dunne, James C. Sacchettini, and David A. Fidock. In vitro efficacy, resistance selection, and structural modeling studies implicate the malarial parasite apicoplast as the target of azithromycin. *J Biol Chem*, 282(4):2494–2504, Jan 2007.

[77] Balbir Singh, Lee Kim Sung, Asmad Matusop, Anand Radhakrishnan, Sunita S G. Shamsul, Janet Cox-Singh, Alan Thomas, and David J. Conway. A large focus of naturally acquired plasmodium knowlesi infections in human beings. *Lancet*, 363(9414):1017–1024, Mar 2004.

[78] A. F. Slater and A. Cerami. Inhibition by chloroquine of a novel haem polymerase enzyme activity in malaria trophozoites. *Nature*, 355(6356):167–169, Jan 1992.

[79] F. Tajima. Statistical method for testing the neutral mutation hypothesis by dna polymorphism. *Genetics*, 123(3):585–595, Nov 1989.

[80] W. Trager and J. B. Jensen. Human malaria parasites in continuous culture. *Science*, 193(4254):673–675, Aug 1976.

[81] J. F. Trape. The public health impact of chloroquine resistance in africa. *Am J Trop Med Hyg*, 64(1-2 Suppl):12–17, 2001.

[82] J. F. Trape, G. Pison, M. P. Preziosi, C. Enel, A. Desgres du Lo, V. Delaunay, B. Samb, E. Lagarde, J. F. Molez, and F. Simondon. Impact of chloroquine resistance on malaria mortality. *C R Acad Sci III*, 321(8):689–697, Aug 1998.

[83] Richard F. Trotta, Matthew L. Brown, James C. Terrell, and Jeanne A. Geyer. Defective dna repair as a potential mechanism for the rapid development of drug resistance in plasmodium falciparum. *Biochemistry*, 43(17):4885–4891, May 2004.

[84] Edouard Vannier and Peter J. Krause. Human babesiosis. *N Engl J Med*, 366(25):2397–2407, Jun 2012.

[85] Edouard Vannier and Peter J. Krause. Babesiosis in china, an emerging threat. *Lancet Infect Dis*, Dec 2014.

[86] Sarah K. Volkman, Pardis C. Sabeti, David DeCaprio, Daniel E. Neafsey, Stephen F. Schaffner, Danny A Milner, Jr, Johanna P. Daily, Ousmane Sarr, Daouda Ndiaye, Omar Ndir, Soulyemane Mboup, Manoj T. Duraisingh, Amanda Lukens, Alan Derr, Nicole Stange-Thomann, Skye Waggoner, Robert Onofrio, Liuda Ziaugra, Evan Mauceli, Sante Gnerre, David B. Jaffe, Joanne Zainoun, Roger C. Wiegand, Bruce W. Birren, Daniel L. Hartl, James E. Galagan, Eric S. Lander, and Dyann F. Wirth. A genome-wide map of diversity in plasmodium falciparum. *Nat Genet*, 39(1):113–119, Jan 2007.

[87] G. A. Watterson. On the number of segregating sites in genetical models without recombination. *Theor Popul Biol*, 7(2):256–276, Apr 1975.

[88] L. M. Weiss, M. Wittner, S. Wasserman, H. S. Oz, J. Retsema, and H. B. Tanowitz. Efficacy of azithromycin for treating babesia microti infection in the hamster model. *J Infect Dis*, 168(5):1289–1292, Nov 1993.

[89] T. E. Wellems, L. J. Panton, I. Y. Gluzman, V. E. do Rosario, R. W. Gwadz, A. Walker-Jonah, and D. J. Krogstad. Chloroquine resistance not linked to mdr-like genes in a plasmodium falciparum cross. *Nature*, 345(6272):253–255, May 1990.

[90] T. E. Wellems and C. V. Plowe. Chloroquine-resistant malaria. *J Infect Dis*, 184(6):770–776, Sep 2001.

[91] N. J. White. Plasmodium knowlesi: the fifth human malaria parasite. *Clin Infect Dis*, 46(2):172–173, Jan 2008.

[92] Chansuda Wongsrichanalai, Amy L. Pickard, Walther H. Wernsdorfer, and Steven R. Meshnick. Epidemiology of drug-resistant malaria. *Lancet Infect Dis*, 2(4):209–218, Apr 2002.

[93] John C. Wootton, Xiaorong Feng, Michael T. Ferdig, Roland A. Cooper, Jianbing Mu, Dror I. Baruch, Alan J. Magill, and Xin-Zhuan Su. Genetic diversity and chloroquine selective sweeps in plasmodium falciparum. *Nature*, 418(6895):320–323, Jul 2002.

[94] Gary P. Wormser, Aakanksha Prasad, Ellen Neuhaus, Samit Joshi, John Nowakowski, John Nelson, Abraham Mittleman, Maria Aguero-Rosenfeld, Jeffrey Topal, and Peter J. Krause. Emergence of resistance to azithromycin-atovaquone in immunocompromised patients with babesia microti infection. *Clin Infect Dis*, 50(3):381–386, Feb 2010.

[95] S. Wright. Evolution in mendelian populations. *Genetics*, 16 (2):1–97, 1931.

[96] S. Wright. Size of a population and breeding structure in relation to evolution. *Science*, 87:430–431, 1938.

[97] K.L. Yielding, L. Yielding, and D. Gaudin. Inhibition by chloroquine of uv repair in e. coli b. *Proceedings of the Society for Experimental Biology and Medicine*, 133(3):999–1001, 1970.

[98] Xia Zhou, Sheng-Guo Li, Shen-Bo Chen, Jia-Zhi Wang, Bin Xu, He-Jun Zhou, Hong-Xiang Zhu Ge, Jun-Hu Chen, and Wei Hu. Co-infections with babesia microti and plasmodium parasites along the china-myanmar border. *Infect Dis Poverty*, 2(1):24, 2013.